Doctoral Dissertation

Broadening the Coverage of Scenes and Descriptions towards Versatile Image Captioning Systems

Ukyo Honda

Program of Information Science and Engineering
Graduate School of Science and Technology
Nara Institute of Science and Technology

Supervisor: Professor Taro Watanabe
Natural Language Processing Lab. (Division of Information Science)

Submitted on March 17, 2023

A Doctoral Dissertation
submitted to Graduate School of Science and Technology,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of Engineering

Ukyo Honda

Thesis Committee:

Supervisor	Taro Watanabe (Professor, Division of Information Science)
Co-supervisor	Satoshi Nakamura (Professor, Division of Information Science)
Co-supervisor	Hiroyuki Shindo (Associate Professor, Division of Information Science)
Co-supervisor	Yuji Matsumoto (Doctor, RIKEN AIP)

Broadening the Coverage of Scenes and Descriptions towards Versatile Image Captioning Systems[1]

Ukyo Honda

Abstract

Image captioning plays a fundamental role in the vision-and-language research, which lies in the intersection of computer vision and natural language processing, by converting the information in images into natural language descriptions. To be versatile pipelines for downstream vision-and-language tasks, captioning systems should be able to describe various types of scenes with extensive information. However, current captioning systems are limited in their coverage of scenes and descriptions: they can handle limited types of scenes, and their output descriptions tend to be overly generic.

The goal of this dissertation is to broaden the limited coverage. The first half of this dissertation is devoted to addressing the limitation of describable scenes by introducing unsupervised image captioning methods. Captioning models require a large number of image–sentence pairs to learn how to describe images, but the coverage of those pairs is limited even in the recent large-scale datasets. Collecting the image–sentence pairs for every type of scene incurs intensive costs and, thus, is not scalable. To broaden the coverage of describable scenes without the heavy reliance on costly data collection, unsupervised image captioning has been studied to train captioning systems without any supervision of the image–sentence pairs. Previous work focused on aligning images with their *pseudo-captions*, *i.e.*, sentences that contain object labels detected from

[1]Doctoral Dissertation, Graduate School of Science and Technology, Nara Institute of Science and Technology, March 17, 2023.

input images. However, the sentence-level alignment forces word-level spurious alignments because the pseudo-captions have many words that are irrelevant to a given image. We propose a gating mechanism and pseudo-labels on it to remove the word-level spurious alignment and show that our method significantly enhances the captioning performance, outperforming the previous sentence-level alignment methods.

In the second half of this dissertation, we address the limitation of obtainable descriptions. Current image captioning systems tend to output overly generic captions and ignore the characteristic details of each image. To investigate the cause of this limitation, we first analyze the outputs of current captioning systems. Our analysis shows that reinforcement learning (RL), the *de-facto* standard training for captioning models, shifts the probability mass from low-frequency words to high-frequency words, decreasing the output vocabulary. Based on this analysis, we introduce lightweight fine-tuning methods that are hinted by long-tail classification and debiasing methods to alleviate the side effect of RL. Experimental results show that the methods significantly increase the output vocabulary and enable RL captioning models to describe information distinctive from other images.

Finally, we discuss the remaining problems and future directions for image captioning research.

Keywords:

image captioning, text generation, unsupervised learning, reinforcement learning, long-tail classification, debiasing, vision and language, natural language processing, neural networks

Acknowledgements

まずはじめに，奈良先端科学技術大学院大学（奈良先端大）教授 渡辺太郎先生と，理化学研究所 松本裕治先生に感謝いたします．博士課程を通して両先生に指導いただけたことは，とても幸運で，光栄なことでした．渡辺先生のコメントは常に具体的で，的確に研究を修正していただきました．松本先生は修士から博士課程一年目までの指導教官であり，その後は理研の所属チームリーダーとして指導していただきました．法学専攻から自然言語処理に飛び込んだ自分を受け入れ，ここまで指導してくださったことに，心から感謝申し上げます．

奈良先端大教授 中村哲先生と，奈良先端大特任准教授 進藤裕之先生には，お忙しいなか副査を引き受けていただきました．中村先生との中間審査でのミーティングは，自分の研究を客観的に見つめ直すきっかけになりました．進藤先生からは，コメントからはもちろんですが，常に机に向かう姿や，応用を見据えた研究テーマ設定から，研究への向き合い方を学ばせていただきました．

研究室秘書の北川裕子さんには，学生生活全般にわたる事務で大変お世話になりました．気さくに声をかけていただいたことも，研究室生活のいい思い出になっています．奈良先端大事務室の方々にも，休学をはさむイレギュラーな審査に対応していただいたことに感謝申し上げます．

研究室の博士課程同期は，お互い辛いときに気晴らしに行ける貴重な存在でした．修士同期との議論は，博士進学後も自分の研究に大きな影響を与えてくれました．今でも，自然言語処理分野で活躍する姿に刺激を受けています．優秀な先輩後輩にも恵まれて，楽しい研究室生活を送ることができました．ここに名前は挙げきれませんが，皆様大変お世話になりました．

大学外では，オムロンサイニックエックス株式会社の牛久祥孝さん，橋本敦史さんに大変お世話になりました．両氏のもとでの二度のインターンがなければ，自分が Vision and Language 分野で研究を進めることはできなかったと思います．NTT コミュニケーション科学基礎研究所の平尾努さんには，インターンから最初の国際会議論文採録まで指導していただき，研究の進め方を学ばせていただきま

した．皆様に感謝申し上げます．
　最後に，ここまで支えてくれた家族に感謝を捧げます．

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Natural language is a principal tool for humans to obtain and convey information. This importance of language has led to extensive research in natural language processing (NLP) so that computers can process information as we humans do. However, language is not the only channel for humans to process information. For example, human communication relies on non-verbal information in addition to verbal information [126, 125]; common sense is implicitly shared and not always mentioned explicitly (*reporting bias*) [171]; we do not verbalize all the information we obtain through the five senses. Language use in the real world is supported by information that is not expressed in the form of language. Thus, NLP systems have to utilize not only the verbalized information but also the information not verbalized yet to achieve human-like information processing in the real world.

Vision and Language lies in the challenging research fields to achieve multimodal information processing with a special focus on language and vision. The tasks include image captioning [46], visual question answering [9], visual dialogue [35], vision-and-language navigation [6], multimodal machine translation [44], text-to-image generation [149], and so on.

Among these tasks, image captioning plays a fundamental role in connecting multimodal information by converting visual information into natural language descriptions. Generated captions can be used in various downstream tasks, such as image indexing for detailed image searching [19, 179], aiding visually impaired users [60], visual question answering on images and videos [49, 88, 71, 195, 17], visual dialogue [192], news

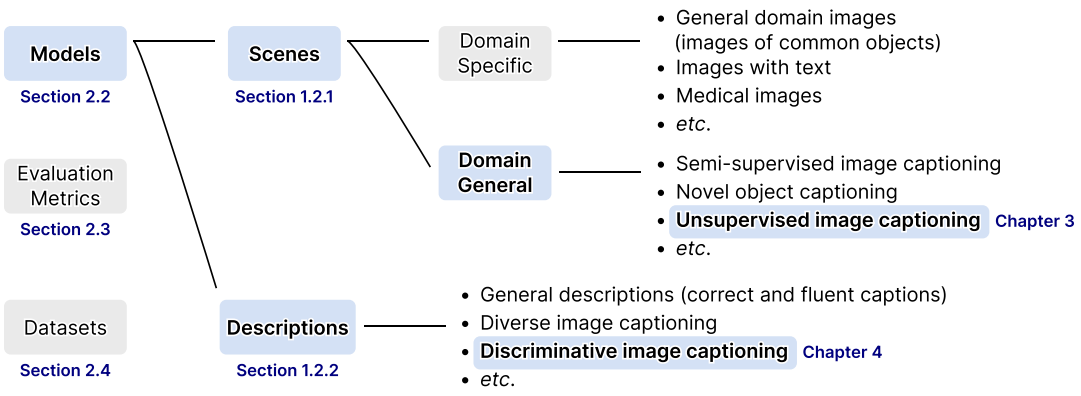


Figure 1.1: Task structure of image captioning. Tasks in our focus are colored with the blue background.

generation [206], image synthesis [51] or editing [139], and object interaction anticipation [140]. Video captioning is a more advanced extension of image captioning, as videos are sequences of images [177, 176, 200].

1.2 Challenges in Image Captioning

Major components of image captioning can be categorized into **models** to generate captions, **evaluation metrics** to evaluate the captions, and **datasets** to conduct the training and evaluation. In this dissertation, we focus on the models. Based on our priority, the following sections introduce challenges in developing the models. See Sections 2.3 and 2.4 for the overview of the studies regarding evaluation metrics and datasets, respectively.

The most popular challenge in model development is to improve the correctness and fluency of captions for images in the general domain. A surge of studies have pursued generating more correct and fluent captions for images of common objects [181, 198, 146, 151, 5, 105]. However, that is not the only challenge with the goal of versatile image captioning systems. Based on the narrow scope of the most popular challenge, we categorize the rest challenges as those done to improve the "**coverage**" of the models. In this section, we provide a brief overview of the rest challenges in two broad categories, scenes and descriptions, which correspond to the inputs and outputs of cap-

tioning models.

See Figure 1.1 for the structure of tasks in image captioning. The tasks in our focus are on blue background. Although out of the focus of this dissertation, we also overview the other tasks in the corresponding sections.

1.2.1 Coverage of Scenes

Scenes to be described significantly vary depending on the use. For example, unseen objects are more likely to interest users than common objects. Photos taken by visually impaired users tend to be low-quality and contain text [60]. The domain of images for visual question answering ranges from clear images of common objects [9, 54] to low-quality images with text [59, 199], clipart images [9], and medical images [98, 12, 64, 78]. However, current captioning systems can handle limited types of scenes [1, 60, 163]. In this dissertation, we refer to the challenge of increasing describable scenes as broadening the **coverage of scenes**.

Domain-specific image captioning has been studied to handle the diverse domains of scenes: photos taken by visually impaired people [60], images with text [163], mobile user interfaces [106], medical images [79, 43, 142, 130], and so on [166].

Although effective, domain-specific approaches require the cost of collecting domain-specific image–sentence[1] pairs for training. Another line of approaches seeks domain-general methods that do not require the intensive collection of training data. Semi-supervised image captioning augments image–sentence pairs by automatically annotating unlabeled images [112, 87]. Novel object captioning tries to describe unseen objects by utilizing automatically detected object labels [66, 175, 1]. Unsupervised image captioning learns to describe images without any pair of images and text [48].

1.2.2 Coverage of Descriptions

Images carry a large amount of information. Thus, it is not enough to correctly and fluently describe general content from images; it is important to pick up and describe information that users may be interested in. For example, the granularity of the image search query and index varies from coarse to fine, depending on the

[1] Captions are sometimes annotated in the form of a noun phrase, not necessarily in a complete sentence. In this dissertation, the term "sentence" includes those noun phrases.

use [19]. Visual question answering, visual dialogue, and news generation need not only salient information but also details users are interested in [49, 71, 195]. In addition, news generation requires proper names of entities beyond their abstract category names [206]. Regardless of the demand for extensive descriptions, outputs of current captioning systems tend to be overly generic, ignoring characteristic details of each image [34, 33, 187, 190]. We refer to the challenge of increasing the variety of descriptions as broadening the **coverage of descriptions**.

Previous studies have tried to enrich image descriptions to cover extensive information about images. Discriminative image captioning, also called distinctive image captioning or descriptive image captioning, seeks to generate captions that are informative enough to distinguish input images from other images [154, 117, 112]. Diverse image captioning focuses on generating a set of unique captions for each input image [191]. Dense image captioning exhaustively describes all the object regions of images [80]. Controllable image captioning specifies object regions of images and then generates captions that are faithful to the selected regions [31]. Change captioning takes pairs of similar images as input and describes the difference within each pair [74, 138]. Personalized image captioning changes text style of captions depending on users [28, 162]. Aesthetic image captioning returns feedback to photos taken by users [16, 52]. Entity-aware image captioning incorporates external knowledge to describe proper names of entities [170, 113, 14].

Aside from enriching, reducing social biases from captions is also important for suitable descriptions for users [207, 208, 65]. Cross-lingual image captioning transfers captioning models trained in pivot language to other languages with no image–sentence pairs [55, 165].

1.3 Research Objective and Contributions

The objective of this dissertation is to achieve more versatile image captioning systems by broadening the coverage of both sides: scenes and descriptions. In particular, we address the task of unsupervised image captioning and discriminative image captioning.

We address unsupervised image captioning to handle edge cases in the scene coverage, where scenes have no corresponding image–sentence pairs during training. Cap-

tioning models require a large number of image–sentence pairs to learn how to describe images, but the coverage of those pairs are limited even in the recent large-scale datasets [1, 60, 170]. Collecting the image–sentence pairs for every types of scenes incurs intensive costs and thus is not scalable. To broaden the coverage of scenes without the heavy reliance on costly data collection, unsupervised image captioning has been studied to train captioning systems without any supervision of the image–sentence pairs. Previous work focused on aligning images with their *pseudo-captions*, *i.e.*, sentences that contain object labels detected from an input images [48, 97]. However, those pseudo-captions have many words that are irrelevant to a given image, thus, the sentence-level alignment results in the word-level spurious alignment. We propose methods to remove the word-level spurious alignment between the images and their pseudo-captions. Our methods significantly enhance the captioning performance and outperform the previous sentence-level alignment methods.

Regarding the description coverage, our priority is given to discriminative image captioning, that is, describing the characteristic details that can distinguish input images from other images. This is because such details should generally be of interest to users. As mentioned above, however, current captioning systems output overly generic captions [34, 33, 187, 190]. We first investigate the cause of the overly-generic captions. Our analysis shows that reinforcement learning (RL), which is the *de-facto* standard training for captioning models, shifts the probability mass from low-frequency words to high-frequency words and consequently decreases the output vocabulary. We introduce lightweight fine-tuning methods hinted by long-tail classification and debiasing methods to alleviate the side effect of RL. Experimental results show that the methods significantly increase the output vocabulary and enable the current captioning models to describe the contents distinctive from other images. This is the contribution to reveal that the limited coverage of descriptions has been caused by the side effect of RL and to introduce the practical, lightweight fine-tuning to mitigate the side effect.

1.4 Structure of the Dissertation

The reminder of this dissertation is organized as follows.

Chapter 2: Preliminaries provides the basic knowledge about image captioning: task setting, models, datasets, and evaluation metrics in image captioning.

Chapter 3: Unsupervised Image Captioning is devoted to broadening the coverage of scenes. We explain the task setting of unsupervised image captioning and its difficulty: the spurious alignment between the words in pseudo-captions and their images. Experimental results demonstrate that our proposed models outperform previous models by removing the spurious alignment, which is an important contribution towards mitigating the limitation of describable scenes.

Chapter 4: Discriminative Image Captioning addresses the limited coverage of descriptions. We first analyze the outputs of current captioning systems and show that RL decreases the output vocabulary. Based on this finding, we propose lightweight fine-tuning methods to increase the output vocabulary so that captions will subsequently include the information specific to each image. Experimental results confirm that our methods successfully contribute to broadening the description coverage from overly-generic information to distinctive details.

Chapter 5: Conclusion discusses the remaining problems and future directions towards versatile image captioning systems.

Chapter 2

Preliminaries

2.1 Task Setting

Image captioning is a task of describing images in the natural language. In the typical supervised setting, captioning models are trained on the pairs of images $\mathcal{I}$ and captions $\mathcal{Y}$. Let $\mathcal{D} = \{(\boldsymbol{I}^{(i)}, \boldsymbol{y}^{(i)})\} \mid i = 1, ..., N\}$ be a training data, where $\boldsymbol{I}^{(i)} \in \mathcal{I}$ is an image and $\boldsymbol{y}^{(i)} = (y_1^{(i)}, ..., y_T^{(i)}) \in \mathcal{Y}$ is its corresponding caption[1]. The last token $y_T^{(i)}$ is a special token $\langle \mathrm{eos} \rangle$ that indicates the end of a sentence.

Given the training data $\mathcal{D}$, the goal of training is generally to find the optimal parameters in the following objective function:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(i)} \mid \boldsymbol{I}^{(i)}), \tag{2.1}$$

where $\boldsymbol{\theta}$ are parameters of captioning models. See Eq. (2.8) for the exception.

2.2 Models

This section provides a brief overview of model architectures in image captioning and how to train the models.

[1] Typically, each image is paired with multiple captions. This notation assigns different values of i to the same images with different captions.

2.2.1 Models before Deep Learning

Early captioning studies decompose the probability $p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{I})$ as follows:

$$p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{I}) = p_{\boldsymbol{\psi}}(\boldsymbol{y} \mid \boldsymbol{m}) p_{\boldsymbol{\phi}}(\boldsymbol{m} \mid \boldsymbol{I}), \tag{2.2}$$

where $\boldsymbol{m}$ is the intermediate text representation of image contents such as triplets consist of objects, attributes, and relation, e.g., $\langle\langle brown, dog\rangle, near, person\rangle$. In the training stage, models learn to find the optimal parameters for $p_{\boldsymbol{\phi}}(\boldsymbol{m} \mid \boldsymbol{I})$, separately from $p_{\boldsymbol{\psi}}(\boldsymbol{y} \mid \boldsymbol{m})$:

$$\boldsymbol{\phi}^* = \arg\max_{\boldsymbol{\phi}} \sum_{i=1}^{N} \log p_{\boldsymbol{\phi}}(\boldsymbol{m}^{(i)} \mid \boldsymbol{I}^{(i)}). \tag{2.3}$$

Then, the trained models do the prediction as follows:

$$\boldsymbol{y}^* = \arg\max_{\boldsymbol{y}} p_{\boldsymbol{\psi}}(\boldsymbol{y} \mid \boldsymbol{m}^*), \tag{2.4}$$

$$\boldsymbol{m}^* = \arg\max_{\boldsymbol{m}} p_{\boldsymbol{\phi}^*}(\boldsymbol{m} \mid \boldsymbol{I}). \tag{2.5}$$

$p_{\boldsymbol{\psi}}(\boldsymbol{y} \mid \boldsymbol{m}^*)$ is modeled by tree or n-gram matching scores between $\boldsymbol{m}^*$ and $\boldsymbol{y}$ to retrieve the closest sentence $\boldsymbol{y}^*$ [46], n-gram language model probability on modified permutations of $\boldsymbol{m}^*$, or slot filling on templates with linguistic constraints [94].

2.2.2 Neural Captioning Models

Deep neural networks and large-scale captioning datasets have brought a remarkable progress in image captioning [181, 166]. Neural captioning models have an advantage of directly modeling the probability $p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{I})$ without the decomposition of Eq. (2.2).

Encoder–Decoder is a common architecture of neural captioning models [181, 41, 123, 91], which is inspired by neural machine translation models [26, 167]. The encoder maps input images into the feature space, and the decoder generates captions given the encoded image features.

The encoder employs pre-trained image-processing models such as image classifiers [181, 62], object detectors [5, 150], and image–text alignment models [158, 145].

The encoded image features can be either a single feature vector or multiple feature vectors of grids or bounding boxes. In the latter case, the multiple feature vectors are dynamically aggregated into a single feature vector using **attention mechanism** [198, 10].

The decoder typically employs auto-regressive decoding models such as Long Short-Time Memory (LSTM) [69] and Transformer [172]. During decoding, the probability $p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{I})$ is auto-regressively factorized as:

$$p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{I}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{I}). \tag{2.6}$$

The encoder–decoder can directly model the probability $p_{\boldsymbol{\theta}}(\boldsymbol{y} \mid \boldsymbol{I})$, allowing end-to-end training: the entire parameters of the model can be updated jointly. In practice, however, the encoder has often been fixed during training for computational efficiency. Recent transformer-only models enable fully end-to-end training and have achieved state-of-the-art performance [186, 135].

Maximum Likelihood Estimation (MLE) is the basic method to train neural captioning models. Given the factorization of Eq. (2.6) and one-hot encoding of target words y_t, the loss for each pair of $(\boldsymbol{I}^{(i)}, \boldsymbol{y}^{(i)})$ is computed by the following **cross-entropy (CE)** loss function[2]:

$$\mathcal{L}_{\mathrm{CE}}(\boldsymbol{\theta}) = -\frac{1}{T} \sum_{t=1}^{T} \log p_{\boldsymbol{\theta}}(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{I}). \tag{2.7}$$

Captioning models learn to minimize the loss to achieve the optimal parameters $\boldsymbol{\theta}^*$ of Eq. (2.1).

Reinforcement Learning (RL) is an alternative training method to the MLE. Although the MLE can maximize the log-likelihood of ground-truth captions given their paired images, that objective does not necessarily correspond to the test-time objective. The goal of RL is to directly maximize the test-time evaluation scores by minimizing the following negative expected reward:

$$\mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta}) = -\mathbb{E}_{\tilde{\boldsymbol{y}} \sim p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{y}} \mid \boldsymbol{I})}[r(\tilde{\boldsymbol{y}})], \tag{2.8}$$

[2]Hereafter, we sometimes omit the instance index (i) for brevity. Unless otherwise noted, loss functions represent calculations for a single training instance.

where $\tilde{\boldsymbol{y}} = (\tilde{y}_1, ..., \tilde{y}_T)$ is a sequence sampled from a policy $p_{\boldsymbol{\theta}}$ and $r : \mathcal{Y} \rightarrow \mathbb{R}$ is a reward function. Typically, the reward function r is an evaluation metric, CIDEr [174]. However, minimizing $\mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta})$ has difficulty in that the reward is non-differentiable.

To compute the gradients with the non-differentiable reward, [146] proposed a method to approximate the gradients using RL. They applied the REINFORCE algorithm [193] to text generation. In practice, the gradients for updating parameters are approximated by S samples per each image as follows.

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta}) \approx -\frac{1}{S} \sum_{s=1}^{S} (r(\tilde{\boldsymbol{y}}^s) - b) \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{y}}^s \mid \boldsymbol{I}). \tag{2.9}$$

Here, b is a baseline reward that reduces the variance in the gradients. Self-Critical Sequence Training (SCST) [151] is a variant of this RL method where the baseline reward b is a reward for a sequence sampled with greedy decoding.

RL, especially SCST, is the *de facto* standard training method for state-of-the-art captioning models because it significantly improves the performance in various evaluation metrics [166].

2.3 Evaluation

Typically, captioning performance is evaluated by computing the similarity between output captions and ground-truth captions. This section explains how the similarity is measured by introducing evaluation metrics used in image captioning.

2.3.1 Exact-Matching Metrics

Conventionally, captioning experiments have employed exact, surface-form matching scores between output captions and ground-truth captions. BLEU [137] counts the precision in n-gram matching, ROUGE [107] counts the recall in n-gram matching, METEOR [37] computes the F1 score in n-gram matching, CIDEr [174] computes n-gram similarity weighted by TF-IDF, and SPICE [3] utilizes scene-graph parser to compute the matching between the dependency graph of output captions and ground-truth captions.

The problem of the exact-matching metrics is that they cannot evaluate the words not included in reference captions. Although each image has around five reference cap-

tions in MS COCO, those captions are not enough to cover all contents pictured in each image. Actually, the correlation between human judgment and CIDEr scores continues to improve as the number of reference captions increases to around fifty [174].

2.3.2 Soft-Matching Metrics

To compensate for the insufficient coverage of reference captions, more studies have focused on soft-matching evaluation.

Text-Based Soft-Matching Metrics consider the semantic similarity beyond the difference in the surface form by employing word embeddings. WMD [95, 86] computes the distance between static word embeddings of target and reference captions. Although beneficial, static word-embedding methods, *e.g.*, CBOW, skip-gram [131, 132], and GloVe [143], have a disadvantage in that they cannot embed contextual information into each word embedding.

With the advent of large pre-trained language models such as ELMo [144] and BERT [39], evaluation metrics have started to utilize contextual word embeddings. BERTScore [204] calculates the distance between target and reference captions using BERT-encoded contextual word embeddings. Improved BERTScore (BERTS++) [201] considers variance of reference captions when penalizing mismatches between target and reference captions.

Text-and-Image-Based Soft-Matching Metrics further compensate for missing reference captions by considering image features as additional references. TIGEr [76] and REO [75] employ pre-trained image-text matching models to ground captions to image regions and then use grounded image features to calculate the similarity between target and reference captions. VIFIDEL [119] uses object detectors to extract additional reference from images; FAIEr [188] additionally employed scene graph generators to consider relation between objects. [32] trains evaluation models on captioning data given outputs of text encoder and image classifier. Recently, end-to-end outputs of large pre-trained cross-modal models are reliable enough to compute image-text similarity without the tailored score computation or training above. UMIC [99] uses UNITER [24], ViLBERTScore [100] uses VilBERT [114], and CLIPScore (CLIPS)

and RefCLIPScore (RefCLIPS) [67] uses CLIP [145] to compute the image-text similarity directly from image and text inputs.

Text-and-image-based soft-matching metrics are reported to correlate with human judgments much better than exact-matching metrics and text-based soft-matching metrics [84]. The current best-performing metric is RefCLIPScore [67, 84].

2.4 Datasets

This section introduces widely-used data to train and evaluate captioning models.

2.4.1 Captioning Datasets

Captioning datasets consist of the pairs of images and corresponding captions, *e.g.*, Pascal VOC 2008 [46], Frickr8k [147], Flickr30k [202], and MS COCO [109, 23]. Among them, MS COCO is the largest and most popular dataset for image captioning. It consists of 123,287 images and around five captions to each image. Following the previous work, we use this dataset for all of our experiments. In all experiments, the images are split into the training/validation/test set of 113,287/5,000/5,000 images [83].

Google's Conceptual Captions (GCC) [157] is much larger dataset of web-crawled 3.3M images and single caption to each image. LAION-5B [155], which followed LAION-400M [156, 145], is the current largest image–text pair dataset of web-crawled 5B images and single caption to each image. Although the size is large, these datasets are noisy and mostly used for pre-training before captioning training [105, 158].

Visual Genome [93] annotated short descriptions to each region of images, not to the entire images. Similar to the noisy image–text pairs above, this dataset is often utilized for the encoder pre-training rather than for captioning training.

2.4.2 Datasets for Image Backbones

Besides the captioning datasets, other image-processing datasets are utilized to pre-train the image encoder as we described in Section 2.2.2. ImageNet [153] is the standard dataset to train image classifiers.

Visual Genome [93] provides the information about image regions and corresponding objects and attributes, in addition the the short description of the regions. This annotation is utilized to pre-train object detectors to encode richer image features than image classifiers [5].

Recently, large-scale image–text pairs were crawled from the web to pre-train image–text alignment models [145, 156]. The pre-trained model is utilized as an alternative image-feature extractor to image classifiers and object detectors [158].

Chapter 3

Unsupervised Image Captioning with Careful Word-Level Alignment to Broaden the Scene Coverage

3.1 Introduction

Image captioning is the task of describing images in natural languages. This is a fundamental challenge with regard to automatically retrieving and summarizing the visual information in a human-readable form. Recently, considerable progress has been made [181, 198, 146, 151, 5, 105] owing to the development of deep neural networks and a large number of annotated image–sentence pairs [202, 109, 23, 93, 157].

However, those image–sentence pairs are limited in their coverage of scenes. For example, the standard captioning dataset MS COCO [109, 23] covers only approximately 100 object categories out of 500 object categories defined in an object detection dataset [1]. In addition to objects, attributes and relations are also not covered well owing to the small vocabulary size: 8791 words [83]. Although increasing the image–sentence pairs is a straightforward way to address this problem, it is difficult due to the cost of manual annotation. Crawling image–text pairs from the web can also compensate for the lack of data [145, 156, 155, 157], but there are still missing scenes. For example, even a 400M image–text pairs ignores the words that occur less than 100 times in English Wikipedia [145].

Unsupervised image captioning [48] aims to describe scenes that have no corresponding image–sentence pairs, without requiring any annotation of the pairs. The

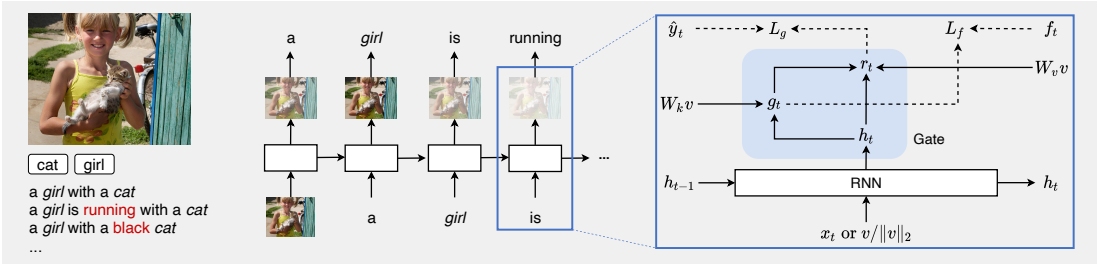


Figure 3.1: Overview of our model. The input is listed on the left-hand side: an image, its detected object labels, and its pseudo-captions. The model learns to generate the pseudo-captions while considering the correspondence between the image and each word being generated. The detailed process is shown in the blue box on the right-hand side. The base encoder–decoder model output $\boldsymbol{h}_t$, a gate value g_t, and a pseudo-label f_t on the gate are described in Sections 3.2.1, 3.2.2, and 3.2.3, respectively. The dashed arrows indicate the processes conducted only during training.

only available resources are images and sentences drawn from different sources and object labels detected from the images. Although it is highly challenging, unsupervised image captioning has the potential to cover a broad range of scenes by exploiting a large number of images and sentences that are not paired by expensive manual annotation.

To train a captioning model in this setting, previous work [48, 97] employed sentences that contained the object labels detected from given images. We refer to these sentences as *pseudo-captions*. However, pseudo-captions are problematic in that they are likely to contain words that are irrelevant to the given images. Assume that an image contains two objects *cat* and *girl* (Figure 3.1). This situation could give rise to various possible pseudo-captions, *e.g.*, "a girl with a cat," "a girl is running with a cat," "a girl with a black cat." Although the first sentence is the correct caption of the image in Figure 3.1, the words *running* and *black* of the other sentences are irrelevant to the image. As the detected object labels provide insufficient information to judge which sentence corresponds to the image, many pseudo-captions containing such mismatched words can be produced.

Regardless of the problem in pseudo-captions, previous work [48, 97] did not ex-

plicitly remove word-level mismatches. They tried aligning the features of images and their pseudo-captions at the sentence level. Although this line of approach can potentially align the images and sentences correctly if there are sentences that exactly describe each image, it is not likely to hold for the images and sentences retrieved from different sources.

To shed light on the problem of word-level spurious alignment in the previous work, we focus on removing mismatched words from image–sentence alignment. To this end, we introduce a simple gating mechanism that is trained to exclude image features when generating words other than the most reliable words in pseudo-captions: the detected objects. The experimental results show that the proposed method outperforms previous methods without introducing complex sentence-level learning objectives. Combined with the sentence-level alignment method of previous work, our method further improves its performance. These results confirm the importance of careful alignment in word-level details.

3.2 Method

Our model comprises a sequential encoder–decoder model, a gating mechanism on the encoder–decoder model, a pseudo-label on the gating mechanism, and a decoding rule to avoid the repetition of object labels, as presented in Figure 3.1.[1]

3.2.1 Base Encoder–Decoder Model

As seen in Eq. (2.1), supervised neural captioning models typically learn to maximize the following objective function to achieve the alignment between images and captions:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}^{(i)} \mid \boldsymbol{I}^{(i)}),$$

where $\boldsymbol{\theta}$ are the parameters of captioning models, $\boldsymbol{I}^{(i)}$ is an input image, and $\boldsymbol{y}^{(i)} = (y_1^{(i)}, \ldots, y_T^{(i)})$ is its corresponding caption. The last token $y_T^{(i)}$ is a special end-of-sentence token $\langle \mathrm{eos} \rangle$.

[1] The code is available at https://github.com/ukyh/RemovingSpuriousAlignment

In unsupervised image captioning, however, the corresponding caption $\boldsymbol{y}$ is not available. Instead, object labels in given images are provided by pre-trained object detectors. Previous work utilized the detected object labels to assign a roughly corresponding caption $\hat{\boldsymbol{y}}$, *i.e.*, a pseudo-caption, to the given image. Following the previous work, we define pseudo-captions of an image as sentences containing the object labels detected from the image. Given the pairs of images and theirs pseudo-caption, $\hat{\mathcal{D}} = \{(\boldsymbol{I}^{(i)}, \hat{\boldsymbol{y}}^{(i)}) \mid i = 1, \ldots, N\}$, our base encoder–decoder model maximizes the following objective function:

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}}^{(i)} \mid \boldsymbol{I}^{(i)}). \tag{3.1}$$

Following Eq. (2.6), the probability $p_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}} \mid \boldsymbol{I})$ is auto-regressively factorized as:

$$p_{\boldsymbol{\theta}}(\hat{\boldsymbol{y}} \mid \boldsymbol{I}) = \prod_{t=1}^{T} p_{\boldsymbol{\theta}}(\hat{y}_t \mid \hat{\boldsymbol{y}}_{<t}, \boldsymbol{I}). \tag{3.2}$$

Let the softmax function $\pi : \mathbb{R}^{|\mathcal{W}|} \to \mathbb{R}$ be

$$\pi_{w_i}(\boldsymbol{z}) = \frac{\exp(z_{w_i})}{\sum_{w_j \in \mathcal{W}} \exp(z_{w_j})}, \tag{3.3}$$

where z_{w_i} indicates the element of any vector $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{W}|}$ at the index of a word $w_i \in \mathcal{W}$. $\mathcal{W}$ is the entire vocabulary. Then, our base encoder–decoder model computes the $p_{\boldsymbol{\theta}}(\hat{y}_t \mid \hat{\boldsymbol{y}}_{<t}, \boldsymbol{I})$ as follows:

$$p_{\boldsymbol{\theta}}(\hat{y}_t \mid \hat{\boldsymbol{y}}_{<t}, \boldsymbol{I}) = \pi_{y_t}(\boldsymbol{W}_p \boldsymbol{h}_t + \boldsymbol{b}), \tag{3.4}$$

$$\boldsymbol{h}_t = \begin{cases} \mathrm{Dec}_{\boldsymbol{\psi}}\left(\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}, \boldsymbol{h}_0\right), & \text{if } t = 1; \\ \mathrm{Dec}_{\boldsymbol{\psi}}(\boldsymbol{x}_t, \boldsymbol{h}_{t-1}), & \text{otherwise}, \end{cases} \tag{3.5}$$

$$\boldsymbol{v} = \boldsymbol{W}_a \mathrm{Enc}_{\boldsymbol{\phi}}(\boldsymbol{I}), \tag{3.6}$$

$$\boldsymbol{x}_t = \mathrm{Emb}_{\boldsymbol{\omega}}(\hat{y}_{t-1}), \tag{3.7}$$

where $\boldsymbol{W}_p \in \mathbb{R}^{|\mathcal{W}| \times d}$ is a vocabulary-size transformation matrix, $\boldsymbol{b} \in \mathbb{R}^{|\mathcal{W}|}$ is a bias term, $\mathrm{Dec}_{\boldsymbol{\psi}} : \mathcal{I} \times \mathcal{Y} \to \mathbb{R}^d$ is a decoder, $\boldsymbol{h}_0 \in \mathbb{R}^d$ is a zero vector, and $\mathrm{Enc}_{\boldsymbol{\phi}} : \mathcal{I} \to \mathbb{R}^{d'}$ is a pre-trained image encoder with a transformation matrix $\boldsymbol{W}_a \in \mathbb{R}^{d \times d'}$ on top of it. $\mathrm{Emb}_{\boldsymbol{\omega}} : \mathcal{W} \to \mathbb{R}^d$ is an embedding function to convert input words into their word embeddings. That is, $\mathrm{Emb}_{\boldsymbol{\omega}}(\hat{y}_{t-1})$ returns $\boldsymbol{\omega}_{\hat{y}_{t-1}} \in \mathbb{R}^d$, which is a transposed row vector of $\boldsymbol{\omega} \in \mathbb{R}^{|\mathcal{W}| \times d}$ at the index of a word $\hat{y}_{t-1}$. See Section 3.3.3 for the details of the encoder and decoder.

3.2.2 Gating Mechanism to Consider Word-Level Correspondence

As indicated in Eq. (3.1), our base encoder–decoder model learns to generate all of the words in pseudo-captions from the images. However, pseudo-captions are highly likely to contain words that are irrelevant to the given images. Forcing a model to generate the pseudo-captions in their entirety from the images might be more disadvantageous than beneficial due to the spurious alignments at the word level.

To enable our model to handle word-level mismatches, we introduce a simple gating mechanism. Our model, which is equipped with this gating mechanism, takes an image representation at each t-th time step. We design the gate to control the amount of image representation used to generate the t-th word. In other words, we expect the gate to determine the extent to which the given image corresponds to the t-th word. With a slight modification to Eq. (3.4), we define our model with the gating mechanism as follows:

$$p^g_{\boldsymbol{\theta}}(\hat{y}_t \mid \hat{\boldsymbol{y}}_{<t}, \boldsymbol{I}) = \pi_{y_t}(\boldsymbol{W}_p \boldsymbol{r}_t + \boldsymbol{b}), \tag{3.8}$$

$$\boldsymbol{r}_t = g_t \frac{\boldsymbol{W}_v \boldsymbol{v}}{\|\boldsymbol{W}_v \boldsymbol{v}\|_2} + (1 - g_t)\boldsymbol{h}_t, \tag{3.9}$$

$$g_t = \mathrm{sigmoid}(\tanh(\boldsymbol{W}_k \boldsymbol{v})^\top \boldsymbol{h}_t), \tag{3.10}$$

where $\boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d \times d}$ are the transformation matrices for computing the gate value $g_t \in [0, 1]$ and the output of the gate $\boldsymbol{r}_t \in \mathbb{R}^d$. When g_t is close to one, it forces the model to use more information from the image ($\boldsymbol{v}$) to generate the t-th word; when g_t is close to zero, it forces the model to do the opposite.

We compute the gate value g_t by the inner product of the context feature $\boldsymbol{h}_t$ and the image feature $\boldsymbol{v}$ transformed by $\boldsymbol{W}_k$. This computation encourages models to generate words from image features only when the two features $\boldsymbol{h}_t$ and $\boldsymbol{v}$ are similar, that is, when the words predicted from each feature are similar. In pseudo-captions, the words that can be predicted relatively easily from image features are object labels detected in images. When contexts indicate that the next word is likely to be an object label, *e.g.*, preceding articles, the words predicted from $\boldsymbol{h}_t$ and $\boldsymbol{v}$ are expected to be similar. In this case, g_t is expected to be larger. Conversely, when contexts indicate that the next word is not likely to be an object label, *e.g.*, preceding object labels, the words predicted from $\boldsymbol{h}_t$ and $\boldsymbol{v}$ are expected not to be similar, and g_t is expected to be smaller. This design of the gating mechanism allows models to increase g_t only for the words that

correspond to images, giving room for preventing the word-level spurious alignment.

The fed image representation $\boldsymbol{W}_v\boldsymbol{v}$ is kept constant at every time step t. Thus, even when the t-th word is correctly pictured in the image $\boldsymbol{I}$, $\boldsymbol{W}_v\boldsymbol{v}$ itself cannot determine which specific object in the image should be generated according to the current context in the output caption. Therefore, we apply L2 normalization to the image representation in Eq. (3.9) to ensure that a relatively greater amount of the contextual information ($\boldsymbol{h}_t$) is used.

To train our model with the gating mechanism, we minimize the following CE loss, *cf.* Eq. (2.7), for each pair of images and their pseudo-captions:

$$\mathcal{L}_g(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T} \log p^g_{\boldsymbol{\theta}}(\hat{y}_t \mid \hat{\boldsymbol{y}}_{<t}, \boldsymbol{I}). \tag{3.11}$$

Following the previous work [48, 97], the parameters $\boldsymbol{\phi}$ of pre-trained image encoders are fixed during training.

3.2.3 Pseudo-Labels on Gate to Remove Word-Level Spurious Alignment

The above gating mechanism is expected to reflect the correspondence between images and words in pseudo-captions. However, learning to reflect the correspondence correctly is difficult for the gate under the noisy and weak supervision of pseudo-captions.

In this work, our focus is to remove the spurious alignment between images and words in pseudo-captions. Consequently, we apply the following rule to the gate that largely suppresses image representations to use: g_t should be close to one if the t-th word to generate is a detected object label; otherwise, it should be close to zero. This is based on the assumption that, given an image and its pseudo-caption, the reliable words in the pseudo-caption are only the detected object labels, and the others are likely to be irrelevant to the image.

We assign a pseudo-label $f_t \in \{0, 1\}$ on the gate: $f_t = 1$ if a word $\hat{y}_t$ corresponds to any of the object labels detected from a given image; otherwise, $f_t = 0$. The gate then learns the correspondence by minimizing the following loss function:

$$\mathcal{L}_f(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T} \Big[\alpha f_t \log g_t + (1 - f_t)\log(1 - g_t)\Big], \tag{3.12}$$

where α is the weight to emphasize the loss when $f_t = 1$. A relatively large value is recommended for α to prevent g_t from always being zero because the number of detected object labels (where $f_t = 1$) in pseudo-captions is generally smaller than the number of the other words (where $f_t = 0$). See Section 3.3.3 for how to determine the value of α.

Combined with the loss function of Eq. (3.11), the final loss function is defined as follows:

$$\mathcal{L}(\boldsymbol{\theta}) = \mathcal{L}_g(\boldsymbol{\theta}) + \mathcal{L}_f(\boldsymbol{\theta}). \tag{3.13}$$

3.2.4 Unique-Object Decoding

Another problem of our gating mechanism is repetition. When $g_t = 1$, Eq. (3.9) forces models to predict the next word only by the image representation $\frac{\boldsymbol{W}_v \boldsymbol{v}}{\|\boldsymbol{W}_v \boldsymbol{v}\|_2}$, ignoring the context representation $\boldsymbol{h}_t$. Thus, when the value of g_t is quite high, the model always outputs the most salient object label in the given image. This ignorance of contexts causes the repetition of the most salient object label.

To avoid this repetition, we applied a simple decoding rule during the evaluation. Given that the model generates a word y_t at t-th time step, our decoding rule checks whether y_t is in predefined object categories, *i.e.*, object categories defined for object detectors. If y_t is found in the object categories, the rule forces the probability of generating y_t to be zero in the subsequent time steps. See Algorithm 1 for details.

This decoding constraint is based on the strong assumption that the same object name never appears more than once in a single caption. To test whether this assumption generally holds for captions, we calculated the percentage of captions in which the same object label appears twice or more among the ground-truth captions on the development set of popular captioning datasets: MS COCO [109, 23], GCC [157], and Flickr30k [202]. The results were 2.2% for MS COCO, 1.9% for GCC, and 3.9% for Flickr30k, indicating that the percentage of the captions with the same object label appearing multiple times is very low for all datasets. These results suggest that the proposed decoding constraint is generally effective for caption datasets. The generation of captions in which the same object label appears multiple times is a subject for future work.

Algorithm 1: Unique-Object Decoding (Greedy)

Input: Captioning model parameters $\boldsymbol{\theta}$, an image $\boldsymbol{I}$, a maximum decoding length T, a set of object categories $\mathcal{O}$, and the special token $\langle \text{eos} \rangle$

$\boldsymbol{y}_{\text{prev}} \leftarrow$ empty list $[]$ ▷ a list of output words
$\mathcal{O}_{\text{prev}} \leftarrow \emptyset$ ▷ a set of output objects
for $t = 1, \ldots, T$ **do**
  $\boldsymbol{p}_{t,\cdot} = p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{y}_{\text{prev}}, \boldsymbol{I}) \in \mathbb{R}^{|\mathcal{W}|}$
  for $w \in \mathcal{W}$ **do**
    if $w \in \mathcal{O}_{\text{prev}}$ **then**
      $p_{t,w} \leftarrow 0$
    end
  end
  $y_t = \arg\max_{w \in \mathcal{W}} \boldsymbol{p}_{t,\cdot}$
  $\boldsymbol{y}_{\text{prev}} \leftarrow \boldsymbol{y}_{\text{prev}} + [y_t]$
  if $y_t == \langle \text{eos} \rangle$ **then**
    break
  end
  if $y_t \in \mathcal{O}$ **then**
    $\mathcal{O}_{\text{prev}} \leftarrow \mathcal{O}_{\text{prev}} \cup y_t$
  end
end
return $\boldsymbol{y}_{\text{prev}}$

	Training Text	Object Detector	Image Encoder	Text Decoder
[48]	SS	Faster-RCNN trained on OpneImages-v2	Inception-v4	1-layer LSTM of 512 dimensions
[97]	GCC	Faster-RCNN trained on OpneImages-v4	ResNet-101	1-layer GRU of 200 dimensions

Table 3.1: Summary of the difference in the experimental settings.

3.3 Experiments

We ran the experiments under two different settings, [48] and [97], for a fair comparison with each. The difference of the settings is summarized in Table 3.1.

3.3.1 Datasets

Evaluation Set. To evaluate our proposed method, we used the MS COCO dataset [109, 23] with the validation and test split defined by [83]. Each split has 5,000 images and five reference captions for each image.

Training Images. We used the remaining images of the MS COCO dataset for training (113,286 images). Note that the we did not use the captions of these training images.

Object Labels. Following the previous work [48, 97], we used pre-traind object detectors [72] to retrieve the object labels found in the images[2]. The training data of the object detectors differs depending on the previous work: OpenImages-v2 [92] in [48] and OpenImages-v4 [96] in [97]. Thus, we used the Faster-RCNN object detector [150] trained on OpenImages-v2[3] to compare with [48] and that trained on OpenImages-v4[4] to compare with [97]. Note that these object detectors were not

[2]Although these pre-trained object detectors require bounding box and semantic label annotations, they can be replaced with any multi-label image classifiers, which can be trained on image-tag pairs that are largely and freely available on the web. To ensure this compatibility, bounding box features are not used in unsupervised image captioning.

[3]http://download.tensorflow.org/models/object_detection/faster_rcnn_inception_resnet_v2_atrous_oid_2018_01_28.tar.gz

[4]http://download.tensorflow.org/models/object_detection/faster_rcnn_inception_resnet_v2_atrous_oid_v4_2018_12_12.tar.gz

trained on MS COCO images. Also note that we refrained from using the detected bounding boxes and their features for a fair comparison with the previous work.

Training Text. Following the previous work, we used the Shutterstock image description corpus (SS) [48] to compare with [48] and the training split captions of GCC [157] to compare with [97]. SS consists of 2.3M image descriptions crawled from Shutterstock, an online stock photography website; GCC consists of 3.3M image descriptions crawled from the web. Note that these sentences are not the descriptions of the images in MS COCO and we did not use the images associated with these sentences.

3.3.2 Evaluation

In the evaluation, we set the maximum decoding length to 20. Our model decoded captions by using greedy search and unique-object decoding, described in Section 3.2.4. Following the previous work, we employed the evaluation metrics as follows: BLEU [137], ROUGE [107], METEOR [37], CIDEr [174] and SPICE [3].

3.3.3 Implementation Details

Image Encoder. For a fair comparison with the previous work, we employed different image encoders depending on the compared method: Inception-v4 [168] in the settings of [48] and ResNet-101 [62, 63] in the settings of [97]. Both image encoders were pre-trained on ImageNet [153] and are publicly available[5]. The parameters of the image encoder were fixed during training and prediction.

Text Decoder. We used different recurrent neural networks (RNN) decoders for a fair comparison: LSTM [69] in the settings of [48] and gated recurrent units (GRU) [26] in the settings of [97]. Following the previous work, the number of hidden layers' dimensions was set to 512 for LSTM and 200 for GRU. The number of the RNN layer was set to one. Word embeddings were randomly initialized and had the same dimensions as the RNN hidden layer.

[5] `https://github.com/tensorflow/models/tree/master/research/slim`; specifically, `inception_v4_2016_09_09` and `resnet_v2_101_2017_04_14` models.

Pseudo-Captions. Captions tend to describe salient objects, not all detected objects. For example, the frequent object *person* often co-occurs with *face* and *clothing* in images, but these three are not always the salient objects to be described in a caption. To avoid collecting the pseudo-captions that only contain these frequent objects, we retrieved pseudo-captions that contain single detected object or pair of them, rather than those contain all detected objects. In this retrieval, we converted object labels to their plural forms using a dictionary used in [48] so that the pseudo-captions could also cover the plural forms of the objects.

For each pair of objects, we selected sentences where $1 < n \leq 4$ words existed between the objects (n is the number of words). $n > 1$ is to collect neither the objects' compound words nor the sentences omitted articles, *e.g.*, "*plant* on *table*"; $n \leq 4$ is to pick up the sentences likely to describe the relations of the target objects. For each object, we selected sentences wherein $n \leq 2$ words were in between the object and its dependent adjective to pick up the sentences likely to describe the object in detail. We used spaCy[6] `en_core_web_lg` model for parsing.

Value of α. As described above, each pseudo-caption contains only one or two detected objects, which is very few compared with the average sentence lengths of the text corpora: 12.0 in SS and 10.7 in GCC. To balance the label imbalance of f_t, we searched the value for α of Eq. (3.12) at a power of 2 and found that $\alpha = 16$, which roughly equals the quotient of $\frac{\text{Sentence Length}}{\text{Detected Objects =1 or 2}}$, worked well across the settings.

Training Iteration. Algorithm 2 shows the detail of training iteration. After collecting the pseudo-captions, we created a set of the pairs of object labels that were used to collect the pseudo-captions. The training is iterated over the pairs in this set, rather than over each image, to avoid overfitting for the most frequent object labels. On each iteration of the pairs of objects, we randomly sampled the image and pseudo-caption, wherein both of the objects were contained. Likewise, we did the same sampling on each object in the pairs.

The number of the object pairs was 11,607 and 10,612 in the settings of [48] and [97], respectively. We set the batch size to eight and terminated the training when the best validation score (specifically, the CIDEr score) did not exceed for 20 epochs. For

[6] `https://spacy.io`

the optimizer, we used Adam with the recommended hyperparameters [89].

Algorithm 2: Training Iteration

Input: Captioning model parameters $\boldsymbol{\theta}$, a set of object categories $\mathcal{O}$, a set of pseudo-captions $\hat{\mathcal{Y}}$, and $\mathcal{D}' = \{(\boldsymbol{I}^{(i)}, \mathcal{O}^{(i)}) \mid i = 1, \ldots, N\}$, a set of pairs of an image $\boldsymbol{I}^{(i)}$ and its detected objects $\mathcal{O}^{(i)}$

/* $\mathcal{M}$ denotes a collection of key-value pairs */

$\mathcal{M}_{\text{image}} \leftarrow \{w : \emptyset\}_{w \in \mathcal{O}}$

$\mathcal{M}_{\text{caption}} \leftarrow \{w : \emptyset\}_{w \in \mathcal{O}}$

for $w \in \mathcal{O}$ **do**

  for $(\boldsymbol{I}^{(i)}, \mathcal{O}^{(i)}) \in \mathcal{D}'$ **do**

    if $w \in \mathcal{O}^{(i)}$ **then**

      $\mathcal{M}_{\text{image}}[w] \leftarrow \mathcal{M}_{\text{image}}[w] \cup \boldsymbol{I}^{(i)}$

      $\mathcal{M}_{\text{caption}}[w] \leftarrow \mathcal{M}_{\text{caption}}[w] \cup \{\hat{\boldsymbol{y}}^{(i)} \mid w \in \hat{\boldsymbol{y}}^{(i)}, \hat{\boldsymbol{y}}^{(i)} \in \hat{\mathcal{Y}}\}$

    end

  end

end

$\mathcal{C} \leftarrow \{(w_i, w_j) \mid (w_i, w_j) \in \texttt{Combination}(\mathcal{O}, 2)\}$

for $epoch = 1, \ldots, M$ **do**

  $\mathcal{D}_{epoch} \leftarrow \emptyset$

  for $(w_i, w_j) \in \mathcal{C}$ **do**

    $\mathcal{D}_{epoch} \leftarrow \mathcal{D}_{epoch} \cup (\texttt{Sample}(\mathcal{M}_{\text{image}}[w_i]), \texttt{Sample}(\mathcal{M}_{\text{caption}}[w_i]))$

    $\mathcal{D}_{epoch} \leftarrow \mathcal{D}_{epoch} \cup (\texttt{Sample}(\mathcal{M}_{\text{image}}[w_j]), \texttt{Sample}(\mathcal{M}_{\text{caption}}[w_j]))$

    $\mathcal{D}_{epoch} \leftarrow \mathcal{D}_{epoch} \cup (\texttt{Sample}(\mathcal{M}_{\text{image}}[w_i] \cap \mathcal{M}_{\text{image}}[w_j]), \texttt{Sample}(\mathcal{M}_{\text{caption}}[w_i] \cap \mathcal{M}_{\text{caption}}[w_j]))$

  end

  Train $\boldsymbol{\theta}$ on $\mathcal{D}_{epoch}$

end

return $\boldsymbol{\theta}$

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
[48]	41.0	22.5	11.2	5.6	12.4	28.7	28.6	8.1
Ours	**49.5** $\pm$ **0.7**	**27.3** $\pm$ **1.2**	**13.1** $\pm$ **0.8**	**6.3** $\pm$ **0.5**	**14.0** $\pm$ **0.1**	**34.5** $\pm$ **0.3**	**31.9** $\pm$ **1.0**	**8.6** $\pm$ **0.2**
[97]				6.5	12.9	35.1	22.7	
Ours	50.4 $\pm$ 1.5	29.5 $\pm$ 0.8	14.4 $\pm$ 0.5	**7.6** $\pm$ **0.4**	**13.5** $\pm$ **0.3**	**37.3** $\pm$ **0.2**	**31.8** $\pm$ **0.7**	8.4 $\pm$ 0.1

Table 3.2: Comparison with the previous models. The experimental settings are different above [48] and below [97] the double line. See Table 3.1 for details. The scores of our model are the *mean* $\pm$ *standard deviation* of five runs. The scores obtained for BLEU-1 to 3 and SPICE are not provided in the original paper of [97].

3.3.4 Comparison with the Previous Models

Table 3.2 lists the results of our model compared with the previous models. We computed the mean and standard deviation of five results obtained with different seeds. Clearly, our method outperformed the previous approaches in all the evaluation metrics. These results confirm the effectiveness of our simple method.

3.3.5 Ablation Study

Table 3.3 lists the results of our model obtained in the ablation studies. We tested the ablation of the gating mechanism (*gate*; Section 3.2.2), pseudo-labels on the gating mechanism (*pseudoL*; Section 3.2.3), unique-object decoding (*unique*; Section 3.2.4), and image features (*image*). The pseudo-labels cannot be implemented without the base gating mechanism. Thus, the model "w/o *gate* w/ *pseudoL*" is not applicable. The model w/o *image* is the same as Ours (full) except that it only uses the word embeddings of detected object labels, rather than image features. It encodes detected object labels into word embeddings and then takes their mean[7] and replaces the image feature $\boldsymbol{v}$ with it. All models here were trained in the same manner as described in Section 3.3.3.

The results show that the pseudo-labels on the gating mechanism significantly con-

[7]The number of detected objects was 3.0 in the setting of [48] and 4.0 in the setting of [97] on average. Thus, taking the mean does not break the detected information significantly.

	gate	*pseudoL*	*unique*	*image*	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
[48] Setup												
Ours (full)	✓	✓	✓	✓	**49.5**	**27.3**	**13.1**	6.3	14.0	34.5	**31.9**	**8.6**
w/o *pseudoL*	✓		✓	✓	0.0	0.0	0.0	0.0	0.0	0.5	0.9	0.3
w/o *gate*			✓	✓	40.9	21.5	10.1	4.8	12.7	32.1	17.6	6.0
w/o *unique*	✓	✓		✓	47.2	26.2	13.0	**6.4**	**14.1**	**34.9**	28.3	8.5
w/o *image*	✓	✓	✓		43.3	23.3	10.8	5.1	13.1	31.7	25.5	7.8
[97] Setup												
Ours (full)	✓	✓	✓	✓	**50.4**	**29.5**	**14.4**	**7.6**	**13.5**	**37.3**	**31.8**	**8.4**
w/o *pseudoL*	✓		✓	✓	44.5	25.4	12.2	6.2	12.4	36.7	29.2	7.5
w/o *gate*			✓	✓	44.5	24.2	12.0	6.2	11.6	34.2	19.4	5.8
w/o *unique*	✓	✓		✓	47.9	27.1	13.0	6.4	12.6	36.3	26.9	7.4
w/o *image*	✓	✓	✓		47.1	26.0	12.8	6.6	13.1	34.7	29.7	8.0

Table 3.3: Ablation studies. The experimental settings are different above [48] and below [97] the double line. See Table 3.1 for details. The scores of Ours (full) are the mean of five runs; those of the other ablated models are the results of a single run.

tribute to the performance; the scores degraded significantly from Ours (full) to w/o *pseudoL* in all the metrics. In contrast, the base gating mechanism does not function well by itself; not all scores of w/o *gate* were lower than those of w/o *pseudoL*. These results demonstrate that explicitly removing the word-level spurious alignment contributes the most to the relatively high performance of our model. Although it is a relatively low contribution compared with the pseudo-labels, unique-object decoding also enhanced performance.

We found that the scores of w/o *pseudoL* in the setting of [48] were quite low because most of the outputs of the model were empty: the model output $\langle \mathrm{eos} \rangle$ at the first time step of decoding. Further analysis revealed that the model used image features only and predicted high-frequency words. That is, the gate value at the first time step was almost always $g_1 = 1$ and top-predicted y_1 were high-frequency words such as $\langle \mathrm{eos} \rangle$, "," (comma), "in", "a", and "the". This is because the text used in [48] is complex[8].

[8]GCC used in [97] collected the text from the web and filtered by removing sentences containing low-frequency words and converting proper nouns to superlatives. The vocabulary size of GCC after our preprocessing was 15,412 words, and the percentage of $\langle \mathrm{unk} \rangle$ (a special token representing unknown words) in the text was about 0.3%. SS used in [48] also collected the text from the web. However, it did not apply the filtering described above. As a result, the vocabulary size of SS was 18,670 words, and the percentage of $\langle \mathrm{unk} \rangle$ in the text was 0.9%, both of them were larger than those of GCC. Moreover, the average sentence length of GCC was 10.7 words, while that of SS was 12.0

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
[48]	41.0	22.5	11.2	5.6	12.4	28.7	28.6	8.1
Ours	49.5 ± 0.7	$\mathbf{27.3 \pm 1.2}$	$\mathbf{13.1 \pm 0.8}$	6.3 ± 0.5	$\mathbf{14.0 \pm 0.1}$	34.5 ± 0.3	31.9 ± 1.0	8.6 ± 0.2
Ours + [48]	$\mathbf{50.9 \pm 0.1}$	$\mathbf{28.0 \pm 0.1}$	$\mathbf{14.0 \pm 0.1}$	$\mathbf{7.1 \pm 0.0}$	$\mathbf{14.1 \pm 0.0}$	$\mathbf{35.2 \pm 0.1}$	$\mathbf{35.7 \pm 0.1}$	$\mathbf{9.2 \pm 0.0}$

Table 3.4: Results of combining our method with previous methods [48]. The scores of our model and the combined model are the *mean* $\pm$ *standard deviation* of five runs. We marked in bold the scores within the standard deviation of the best scores.

High complexity of text makes it difficult for the gating mechanism to automatically identify the words corresponding to the images. Due to the difficulty of automatic word-level alignment, the model learned a simple but incorrect alignment that maps image features to only the high-frequency words $\langle \mathrm{eos} \rangle$, which appear in every pseudo-caption.

The degraded performance of w/o *image* suggests that object labels themselves are insufficient to describe images correctly. We observed that this model was vulnerable to errors propagated through object detectors. See Section 3.3.9 for the examples.

3.3.6 Combining with Previous Methods

Our method focuses on removing word-level spurious alignment between images and pseudo-captions, whereas the previous methods focus on aligning images and pseudo-captions at the sentence level. To utilize the strength of each, we combined our method with the previous method of [48].

We first trained our model on the setting of [48] and generated captions for the images in training data. We then paired the generated captions with the training images to make the pairs of images and pseudo-captions[9]. With these pairs, the caption generator of the previous work initialized its parameters by learning to generate the pseudo-captions from the images. After the initialization, we fine-tuned the caption generator

words.

[9]To avoid assigning obviously incorrect pseudo-captions, we omitted the pseudo-captions that contained fewer than one detected object for the images with more than two detected objects. For the images with fewer than one detected object, we omitted the pseudo-captions that contained no detected objects.

using the publicly available code of the previous work[10]. We used the same hyperparameters as the previous work except for the learning rate: 1e-5 for the generator and 1e-8 for the discriminator.

Table 3.4 shows that the combined model further improves the performance from our model and [48]. In particular, the improvement from [48] was much larger than that from our model. These results suggest that removing the word-level spurious alignment is critical for the subsequent sentence-level alignment.

3.3.7 Effects of Removing Spurious Alignment

To further investigate the effects of removing the spurious alignment, we evaluated our model on the performance in predicting *noisier words*: the words other than the detected object labels. Our method suppresses the alignment of those noisier words with images because they are likely to be irrelevant to the given images, while previous methods force the alignment. Consequently, the largest performance gap should occur in the prediction of those noisier words. To test the difference, we evaluated the following set-of-words matching on the MS COCO test set.

Let $\mathcal{S}$ be a set of words of a caption generated from an image $\boldsymbol{I}$ and $\mathcal{T}_k$ be a set of words taken from k-th reference caption of $\boldsymbol{I}$. Given a set of detected object labels $\mathcal{O}$ of $\boldsymbol{I}$, we took the intersections $\mathcal{S}_{\text{det}} = \mathcal{S} \cap \mathcal{O}$ and $\mathcal{S}_{\text{other}} = \mathcal{S} \cap \overline{\mathcal{O}}$ for $\mathcal{S}$. Similarly, we took the intersections $\mathcal{T}_{k,\text{det}} = \mathcal{T}_k \cap \mathcal{O}$ and $\mathcal{T}_{k,\text{other}} = \mathcal{T}_k \cap \overline{\mathcal{O}}$ for $\mathcal{T}_k$. We define the precision (P), recall (R) and F1 score (F) of $\mathcal{S}$ against $\mathcal{T}_k$ as follows:

$$P = \frac{|\mathcal{S} \cap \mathcal{T}_k|}{|\mathcal{S}|}, \tag{3.14}$$

$$R = \frac{|\mathcal{S} \cap \mathcal{T}_k|}{|\mathcal{T}_k|}, \tag{3.15}$$

$$F = 2\frac{PR}{P+R}. \tag{3.16}$$

We then define the precision, recall, and F1 score of $\mathcal{S}_{\text{det}}$ against $\mathcal{T}_{k,\text{det}}$ by replacing $\mathcal{S}$ with $\mathcal{S}_{\text{det}}$ and $\mathcal{T}_k$ with $\mathcal{T}_{k,\text{det}}$, and likewise for those of $\mathcal{S}_{\text{other}}$ against $\mathcal{T}_{k,\text{other}}$. We calculated the above scores for each pair of a generated captions and their reference captions, and then subsequently averaged the scores across the pairs. We excluded the

[10] https://github.com/fengyang0317/unsupervised_captioning

		Precision	Recall	F1
Objects	[48]	**56.6**	57.4	**55.4**
	Ours	51.0	56.7	51.6
	Ours + [48]	54.0	**61.8**	**55.4**
Others	[48]	22.3	17.0	18.8
	Ours	27.8	**21.9**	23.4
	Ours + [48]	**29.9**	**21.9**	**24.2**

Table 3.5: Set-of-words matching scores with respect to detected object labels and the other words.

pairs with empty $\mathcal{T}_{k,\mathrm{det}}$ or $\mathcal{T}_{k,\mathrm{other}}$ as the score of those pairs is always zero or none for any model.

Table 3.5 shows the results. Overall, the scores on detected object labels (Objects) were about two times higher than those on the other words (Others), indicating the difficulty of learning the alignment of the noisier words. Our model performed better in predicting the noisier words, outperforming [48] in all the metrics. These results indicate that refraining from the alignment works better than forcing it for the noisier words.

In contrast, our model performed worse in predicting detected object labels. This is because our method trusts all detected object labels and aligns them with images without any constraints used in previous work. Combined with the previous method (Ours + [48]), our model improved the performance on Objects.

Another possible solution to the lower performance on Objects is to raise the threshold applied to the object detector's confidence. Object detectors output object labels if their confidence is higher than the threshold. Table 3.6 shows the percentage of detected object labels that were contained in ground-truth captions (Precision) and the number of images in which one or more objects were detected (Valid Images). We tested the object detector used in [48]. The table shows that Precision increases as the threshold value increases. Thus, raising the threshold will prevent our model from aligning incorrectly detected object labels with images. The threshold was set at 0.3 in both the previous work and this study[11], so this solution is feasible. However, rais-

[11] [97] did not provide details on the threshold value, so we used the default threshold value of 0.3 in

Threshold	Precision	Valid Images
0.3	24.7	**4,953**
0.5	29.5	4,837
0.7	35.4	4,088
0.9	**47.2**	1,960

Table 3.6: The percentage of detected object labels that were contained in ground-truth captions (Precision) and the number of images in which one or more objects were detected (Valid Images) when we varied the threshold applied to the object detector. We used the MS COCO validation set. The maximum number of Valid Images is 5,000.

ing the threshold reduces the number of valid images that can be used for training. If additional images are collected on the web and added to the training, it is possible to raise the threshold without reducing the number of images used for training. To keep the number of images consistent with the previous studies, we did not conduct experiments using additional images in this study.

3.3.8 Properties of Output Words

In this section, we examine the properties of the output words. By assigning the pseudo-label f_t, our method encourages models to align detected object labels with the image representation $\boldsymbol{v}$ and the other words with the contextual representation $\boldsymbol{h}_t$. Thus, our model is likely to predict the other words mostly based on the previous output sequences, as language models do. Language models are known to predict words that occur frequently in text data [36, 70]. If this is the case, then the other words predicted by our model tend to be the frequent words in the training text corpus.

Based on the above hypothesis, we counted how many times an output {object label (Objects), the other word (Others)}[12] occurs in the training text corpus, SS. Table 3.7 presents the results. Although there are no significant differences in Objects, we observe substantial difference in Others. Our outputs' vocabulary in Others is about five

the setting of [97], too.

[12]We analyzed *each unique word across all the output captions* in the MS COCO test set, so we roughly divided the words into object labels and the others, not into *detected* object labels and the others.

		Word Type	Frequency
Objects	[48]	205	20,013
	Ours	306	15,052
	Ours + [48]	239	18,226
Others	[48]	827	24,865
	Ours	169	83,693
	Ours + [48]	121	110,358

Table 3.7: Analysis of generated captions with respect to object labels (Object) and the other words (Others). Word Type is the number of unique words, and Frequency is the mean of the frequency of the words in the training text corpus.

times smaller than that of [48], and the words are highly frequent in the training text corpus.

The results also show that a model performs better if it has the smaller and more frequent vocabulary of the words other than object labels (*cf.* Table 3.4). This correlation is convincing considering the coverage of frequent words. For example, a general caption such as "a man *with* a bike" can correctly describe various scenes in which a man is {riding, sitting on, leaning on, standing near, *etc.*} a bike. This positive effect of frequency suggests that firstly aligning the frequent words and gradually extending them can be a promising approach.

3.3.9 Qualitative Analysis of Outputs

Figure 3.2 shows the captions generated by our model, its ablated models, [48], and the combined model. All of the models were trained on the setting of [48]. Our model generated correct captions for images (a) and (b). It successfully generated object labels that were not even detected by the object detector: *bat* in (a) and *mirror* in (b). In contrast, errors of the object detector directly propagated to the output captions of w/o *image* model: the model generated an incorrect object *a bottle of wine*, owing to the missing object *bat* in (a).

Captions of the other images are negative results of our model. We observe that our model tends to repeat similar objects: *cat* and *dog* in (c), and *elephant* and *elephants*

in (f). Without unique-object decoding, this tendency got worse: w/o *unique* model repeated *cat* in (c) and (e), and *elephant* in (f). Ours + [48] model did not change much of the prediction of our model, as we set the learning rate low (see Section 3.3.6). However, it allowed the partial correction seen in (c): the combined model modified *dog* to *suitcase*.

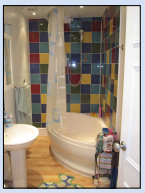
In our outputs, words other than object labels tended to be frequent words and composed short phrases. On the contrary, [48] tended to generate less frequent words (*savuti* and *kenya* in (f)) and longer phrases (*portrait of a happy young* in (a) and *young couple in love* in (d)), which were incorrect predictions in these examples.

Figure 3.3 shows output captions of our model and the gate values for each word. Overall, the gate values were high for object labels and low for the other words. Although our model was correct on the words other than object labels in these examples, these words were generated mostly by contextual features, thus heavily relied on contextual frequency. This heavy reliance on contexts resulted in generating the same word after an object label without considering images: *is sitting on* followed *cat* in both (c) and (e), but it is not correct in the image of (e).

(a)

Objects	man, footware, uniform
Gold	a person on a field with a baseball bat
[48]	portrait of a happy young man in uniform
Ours	young man in a white uniform holds a baseball bat
w/o *filter*	N/A
w/o *gate*	cook with serious face in burgundy uniform holds vegetables in wicker basket
w/o *unique*	young man in a white uniform and hat with a backpack and a backpack on the background of a mountain
w/o *image*	young man in a white uniform is holding a bottle of wine
+ [48]	young man in a white uniform holds a baseball bat

(b)



Objects	bathtub, curtain, sink
Gold	a bath tub sitting next to a sink in a bathroom
[48]	interior of a modern bathroom with bathtub and toilet
Ours	white bathtub with white sink and a mirror
w/o *filter*	N/A
w/o *gate*	white bathtub with tile trim and black trim
w/o *unique*	a white bathtub with a white bathtub
w/o *image*	bathroom interior with white bathtub and shower
+ [48]	a white bathtub with a sink and a mirror

(c)

Objects	cat
Gold	a cat sitting inside a small, packed suitcase
[48]	a cat in a hat and a cat
Ours	a cat is sitting on a white dog
w/o *filter*	N/A
w/o *gate*	a cat in a white helmet and a blue jacket is sitting on a wooden floor
w/o *unique*	a cat is sitting on a cat
w/o *image*	a cute cat is sleeping on a wooden floor
+ [48]	a cat is sitting on a suitcase

(d)

Objects	person, man, clothing, furniture
Gold	a man standing on top of a wooden skateboard ramp
[48]	young couple in love sitting on a bench in the park
Ours	young man in a white bench with a skateboard
w/o *filter*	N/A
w/o *gate*	a young man in a black jacket and a black helmet is sitting on a bench in a park
w/o *unique*	a young man in a white shirt and a hat with a bench in the park
w/o *image*	young man in a white shirt and black tie standing with a confident smile and smiling
+ [48]	young man in a jeans jacket and a skateboard in the park

(e)

Objects	cat
Gold	a cat standing on a toilet with an open lid
[48]	the cat sits on the toilet
Ours	a cat is sitting on a toilet
w/o *filter*	N/A
w/o *gate*	a cat is sitting on a wooden bench in the park
w/o *unique*	a cat is sitting on a cat
w/o *image*	a cute cat is sleeping on a wooden floor
+ [48]	a cat is sitting on a toilet in the bathroom

(f)

Objects	elephant
Gold	two elephants at a zoo with their trunks touching
[48]	elephant walking through the river in the savuti , kenya
Ours	a elephant in a elephants
w/o *filter*	N/A
w/o *gate*	a young african elephant in a safari park
w/o *unique*	a young elephant in a elephant in the zoo
w/o *image*	a lone elephant in the nature habitat , europe
+ [48]	a elephant in a elephants

Figure 3.2: Sample captions of six input images taken from the MS COCO validation set. Our model generated correct captions for the images in the blue background and wrong captions for the images in the red background.

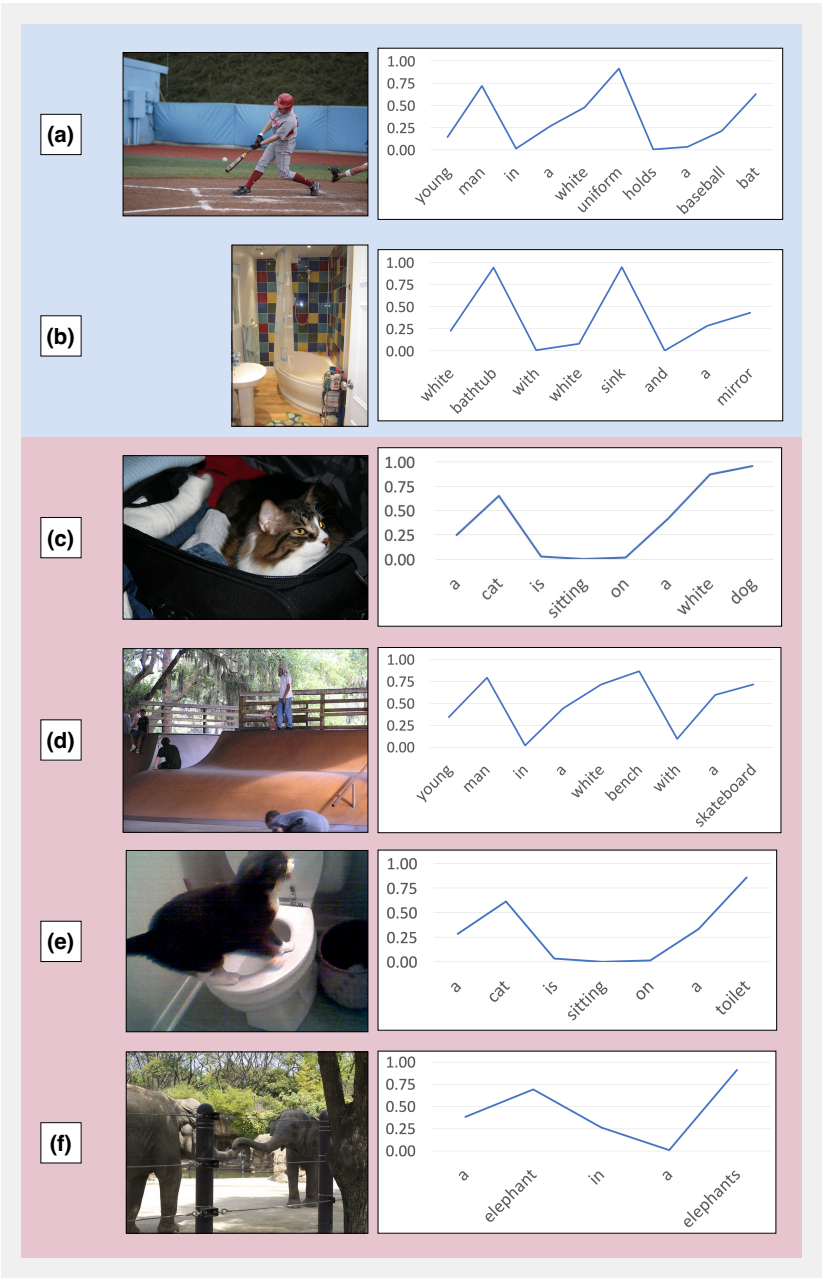


Figure 3.3: Sample captions with gate values. The plot represents the values of g_t for each predicted word. The value of g_t becomes high when the word is predicted using mainly image representation. Our model generated correct captions for the images in the blue background and wrong captions for the images in the red background.

3.4 Related Work

There has been considerable research with different settings and approaches to describe scenes that have no image–sentence pairs. Novel object captioning [66, 175, 1] attempted describing unseen objects in captions. They incorporated an image classifier or object detector trained on objects not included in image–sentence pairs. Other variants of novel object captioning use object labels extracted from reference captions [4] or a small number of image–text pairs containing unseen object labels [122]. [116] tested captioning models on the generation of unseen combinations of objects, and [136] extended this to the unseen combinations of objects, attributes, and relations. In both settings, only the combinations were unseen, but each word in the combinations appeared in the training data. Semi-supervised approaches utilized caption retrieval models to automatically collect the corresponding captions for unannotated images to augment image–sentence pairs [112, 87].

The above work was evaluated on the scenes where correct descriptions partially overlapped with those in the training image–sentence pairs. However, there can be scenes with no such overlap due to the limited coverage of the currently available image–sentence pairs. Taking a step further, unsupervised image captioning [48] aims to describe scenes that have no overlap with the image–sentence pairs, without the annotation of the pairs. To test in that situation, the task does not allow to use any image–sentence pairs. The only available resources are images and sentences drawn from different sources and object labels detected from the images.

[48] first trained an encoder–decoder model that takes object labels in a sentence as its input and outputs the original sentence. After training, this model took the object labels detected from each image and outputted a sentence to pair with the image as its pseudo-caption. These pairs were then used to initialize a caption generator for the subsequent image–sentence alignment: bi-directional (image-to-sentence and sentence-to-image) feature reconstruction and GAN training [53] to ensure fluency in generated captions. In the work of [97], pseudo-captions were sentences that contained object labels detected from a given image. They employed metric learning and GAN training to minimize the difference between images and pseudo-captions in their latent space, as well as to maximize the difference between images and sentences wherein no detected object label was included. [15] introduced an additional attention network into the model of [48]. They pre-trained the attention network on extra image

resources annotated with objects and relations of the objects so that it could consider the interactions between the objects.

Our approach is different from them in that it focuses on removing the mismatched words of pseudo-captions to take reliable supervision only, rather than forcing the use of the entire pseudo-captions for image–sentence alignment. Although the previous work additionally ensured to align detected object labels to images, they did not prevent the spurious alignment between images and words.

[58] is a contemporaneous work that proposed a memory network to generate natural sentences from detected object labels. They focused on filling the gap between a set of discrete words and natural sentences. Our work and theirs are the same in that the focus is not on image–sentence alignment at the sentence level; the difference is that we focus on investigating the effect of removing word-level spurious alignment. We designed our method simply and explicitly for the objective and provided in-depth analyses on the effect.

As an eased setting of unsupervised image captioning, unpaired image captioning has also been explored [48, 97, 56, 110]. The major difference from unsupervised image captioning is that images and sentences are drawn from image–sentence pairs, rather than from different sources. That is, every image has completely matched captions in pseudo-captions, which is not the case in unsupervised image captioning. As correct captions exist for each image, previous approaches focused on matching images and sentences at the sentence level. Contrary to these approaches, we focus on employing unsupervised image captioning and devising a method to remove word-level spurious alignment in the much noisier pseudo-captions.

Another variation of unpaired image captioning is the generation of captions in one language that has no image–sentence pairs, using paired images and captions in another language [55, 165]. However, this line of research is beyond the scope of our work, as it requires image–sentence pairs to be at least in one language.

Our gating mechanism borrowed the idea of adaptive attention [115, 116]. Adaptive attention serves to control when generating words from image representations. Although these methods assume that the control is automatically learned from image–sentence pairs, this is not the case in an unsupervised setting. Our method is different from theirs in that we add heuristic pseudo-labels to train the gate when using image representations.

3.5 Limitations

Our experiments are limited in domains. The domain of images is limited to images of common objects; the text domain is restricted to crawled captions of images. Experimentation with images from other domains and text from the general domain will be the subject of future work. Although the language is also limited to English, a subsequent study has shown that our methods work robustly in Chinese and French [128].

Our method improves caption correctness at the expense of a diversity of words other than object labels (Section 3.3.8). While this is an effective method in the current situation where scores are low, enrichment in the vocabulary is needed to increase scores further. As discussed in Section 3.3.8, an effective approach will be progressing gradually from high-frequency word alignment to low-frequency word alignment.

Although related to object labels, it is difficult for unsupervised image captioning to describe higher-level objects composed of multiple object labels: *e.g.*, *people* (a number of *persons*) and *party* (*persons*, *drink*, *food*, *etc.*). Given the component objects are described correctly, paraphrasing output captions with pre-trained text paraphrasing models will be one future approach to address this problem without using image–sentence pairs.

We have shown that sentence-level alignment on top of careful word-level alignment will be a promising direction to improve the performance, including words other than object labels (Sections 3.3.6 and 3.3.7). However, the upper bound of the performance on words other than objects is not clear yet. Examining the performance of unsupervised image captioning methods using ground-truth object labels is an important future work to assess the limitations of this task.

3.6 Conclusion

We investigated the importance of removing word-level spurious alignment between images and pseudo-captions in the task of unsupervised image captioning. For this purpose, we introduced a simple gating mechanism trained to align image features with only the most reliable words in pseudo-captions. The experimental results showed that our proposed method outperformed the previous methods without the sentence-level learning objectives used in the previous methods. Moreover, our method improved the

performance further by combining with the previous methods. These results confirm the importance of careful alignment in word-level details.

Chapter 4

Discriminative Image Captioning by Relieving a Bottleneck of RL to Broaden the Description Coverage

4.1 Introduction

Image captioning plays a fundamental role at the intersection of computer vision and natural language processing by converting the information in images into natural language descriptions. Generated captions can be used in various downstream tasks: aiding visually impaired users [60], visual question answering on images and videos [49, 88, 71, 195, 17], visual dialogue [192], news generation [206], and so on.

For those downstream tasks, captions should be **discriminative**: captions should describe the characteristic and important details of the input images [154]. However, current captioning models tend to generate overly generic captions [34, 33, 187, 190]. In particular, models trained with the standard **RL** [151], which is the *de facto* standard training method in current image captioning [166], unexpectedly perform poorly in discriminativeness despite the significant advantages in various other criteria [111, 182]. For example, a high-performing Transformer [172] captioning model trained with RL generates exactly the same caption for the four different images shown in Figure 4.1, ignoring the other salient details of each image.

To address the problem of overly generic captions, studies have been intensely conducted on **discriminative image captioning**, which is also called **distinctive** image captioning or **descriptive** image captioning. Previous research has created new RL

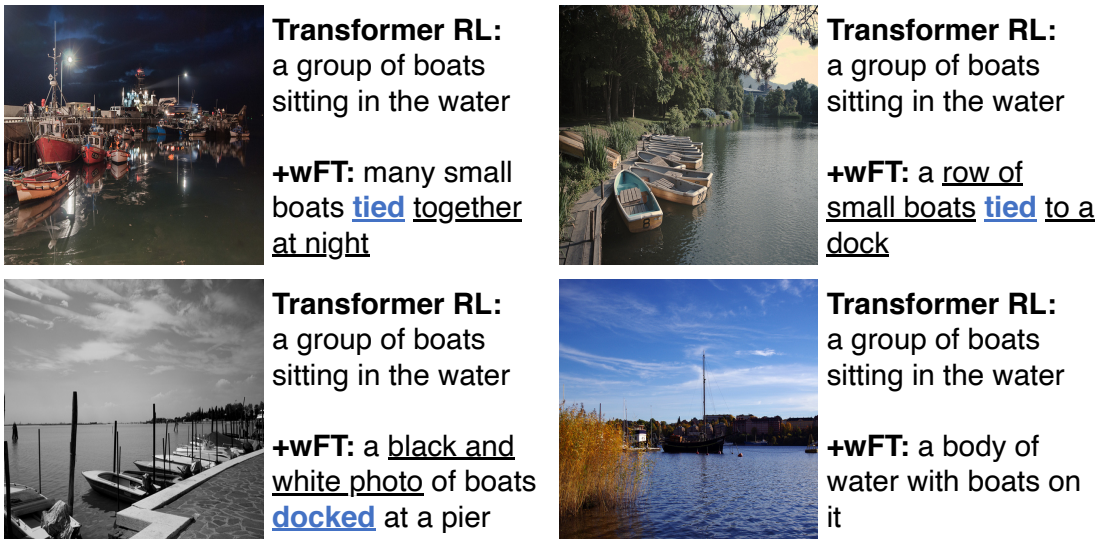


Figure 4.1: Caption examples in the MS COCO validation set. **Transformer RL** is a Transformer captioning model trained with RL and **wFT** is our fine-tuning method. Transformer RL generates exactly the same caption for the four images. The underlined words indicate the characteristic information that are not mentioned by Transformer RL, and the blue words are those that have never appeared in the outputs of the model. See Appendix B.1 for more examples.

rewards regarding discriminativeness or new model architectures to enhance discriminativeness. These approaches improved the discriminativeness; however, their models come with additional computations, require retraining from scratch, and do not shed light on the cause of *existing models*' low discriminativeness.

Instead of creating or paying those computational costs, we first analyze the cause of the unexpectedly low discriminativeness of *off-the-shelf RL models*, *i.e.*, pre-trained, existing RL models, to explore ways to improve their discriminativeness. Our first contribution is the identification of a deeply rooted side effect in RL that limits output words to high-frequency words. The limited vocabulary is a severe bottleneck for discriminativeness as it is difficult for a model to describe the details beyond its vocabulary.

Given this identification of the bottleneck, now we can directly address the bottle-

neck by simply encouraging the generation of low-frequency words. This task relaxation allows us to introduce long-tail classification and debiasing methods to discriminative image captioning for the first time. Our second contribution is our effective and efficient methods that switch any off-the-shelf RL models to discriminativeness-aware models with only a single-epoch fine-tuning on the part of the parameters. Unlike previous approaches, our methods do not require any discriminativeness rewards, new model architectures, or retraining from scratch.

Extensive experiments demonstrate that increasing low-frequency words in outputs significantly boosts discriminativeness from off-the-shelf RL models and even outperforms previous discriminativeness-aware models with much smaller computational costs. These results verify that the limited vocabulary of RL models has been the major cause of their low discriminativeness. Detailed analysis and human evaluation also show that our methods enhance the discriminativeness without sacrificing the overall quality. We believe that our novel findings on the cause of low discriminativeness and the practical solutions to it will significantly impact future research on discriminative image captioning.

4.2 Discriminativeness and a Bottleneck of RL

Currently, RL is the *de facto* standard training method for models used in image captioning because it significantly improves the performance in various evaluation metrics [166]. However, it does not improve discriminativeness and may even decrease it [111, 182]. In this section, we examine the cause of the unexpectedly low discriminativeness.

4.2.1 RL in Image Captioning

In this section, we briefly review RL described in Section 2.2.2. The goal of RL is to directly optimize non-differentiable test-time metrics by minimizing the negative expected reward:

$$\mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta}) = -\mathbb{E}_{\tilde{\boldsymbol{y}} \sim p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{y}}|\boldsymbol{I})}[r(\tilde{\boldsymbol{y}})],$$

where $\tilde{\boldsymbol{y}} = (\tilde{y}_1, ..., \tilde{y}_T)$ is a sequence sampled from a policy $p_{\boldsymbol{\theta}}$, $\boldsymbol{I}$ is an input image, and r is a reward function. To compute the gradients of $\mathcal{L}(\boldsymbol{\theta})$, [146] applied the RE-

INFORCE algorithm [193] to text generation. In practice, the algorithm approximates the gradients for updating parameters by S samples per image as follows.

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\mathrm{RL}}(\boldsymbol{\theta}) \approx -\frac{1}{S} \sum_{s=1}^{S} (r(\tilde{\boldsymbol{y}}^s) - b) \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\tilde{\boldsymbol{y}}^s \mid \boldsymbol{I}).$$

Here, b is a baseline reward that reduces the variance in the gradients. Typically, the reward function r is CIDEr [174], and the baseline reward b is a reward for a sequence sampled with greedy decoding [151].

4.2.2 RL Limits Vocabulary

Despite its effectiveness, RL has been found not to improve discriminativeness and somehow decrease the number of unique n-grams in output captions [111, 182]. As the relation between RL and these two negative effects is not obvious, it has been just considered a curious case.

We elucidate for the first time the relation between *RL and limited vocabulary* by combining two recent findings.

- RL has been shown to make the output distribution peaky [27, 85]. RL samples sequences from policy $p_{\boldsymbol{\theta}}$ as described in Section 4.2.1. Typically, $p_{\boldsymbol{\theta}}$ is initialized with a text-generation model pre-trained with the CE loss on ground-truth text. In text generation, however, the initialized $p_{\boldsymbol{\theta}}$ outputs peaky distributions. Consequently, RL samples and rewards the words at the peak only, shaping more peaky distributions [27]. Then, where does $p_{\boldsymbol{\theta}}$ tend to be peaky?

- Text-generation models have been theoretically and empirically shown to output distributions peaky at high-frequency words in the training corpus [134, 148, 36, 70]. That is, the initialized $p_{\boldsymbol{\theta}}$ is peaky at high-frequency words.

These two findings conclude that *RL shifts the probability mass from low-frequency words to high-frequency words* by only sampling and rewarding the latter.

Figure 4.2 confirms the above by plotting the relative frequency of the words sampled for the training images. The words are sorted by their frequency in ground-truth captions and divided into 200 bins. Compared to the ground-truth captions and the

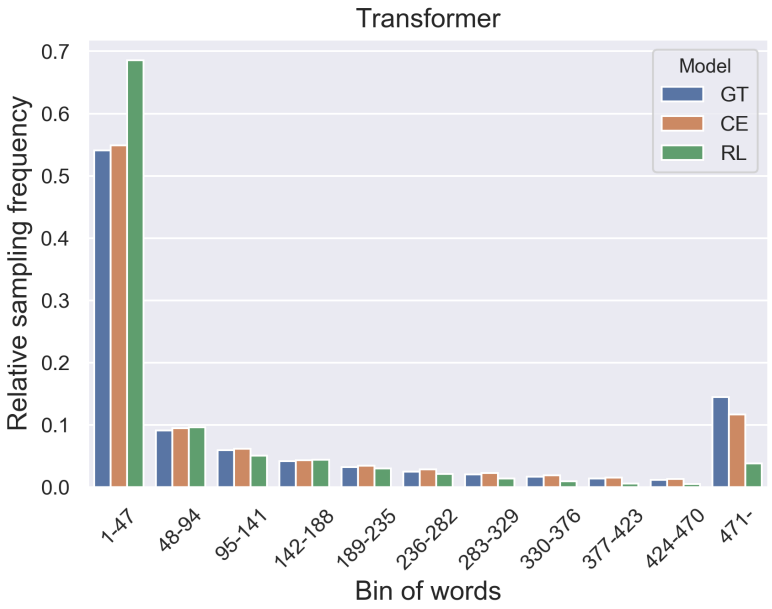


Figure 4.2: Relative frequency of the words in the sequences sampled for the images in MS COCO training set. Five sequences were sampled for each image. The words (9,486 unique words excluding an out-of-vocabulary token $\langle \mathrm{unk} \rangle$) are sorted by their frequency in ground-truth captions and divided into 200 bins. We show the first 10 bins and the sum of the rest. GT is the ground-truth caption of the training images, CE is the output of a captioning model trained with the CE loss, and RL is the output of a captioning model trained with RL. Here, we used the Transformer model.

sequences sampled with a CE model, the sequences sampled with an RL model are clearly limited to the high-frequency words, forming a peaky distribution[1].

4.2.3 Vocabulary Limits Discriminativeness

Neural captioning models typically generate captions using sequential vocabulary-size classification [181]. However, the actual vocabulary a model can generate is much smaller than the entire vocabulary as the output distribution is highly skewed towards

[1] Although Figure 4.2 shows only the results obtained with the Transformer captioning model, we also confirmed that other models output peaky distributions [151, 5]. See Appendix B.2 for the details.

high-frequency words. If the actual vocabulary cannot cover the details of an image, the model is forced to avoid those details and output only the information that high-frequency words can describe. For example, the blue words in Figure 4.1 are not in the actual vocabulary of the RL model; these words have never been generated during evaluation. As a result, the RL model had to ignore the characteristic relations *tied* and *docked* and ended up describing exactly the same for all four images.

Based on the observations, now we can hypothesize that the unexpectedly low discriminativeness of RL models has been rooted in the limited vocabulary. This identification of the bottleneck is a key contribution as it allows us to address the low discriminativeness directly at the root.

4.3 Methods to Relieve the Bottleneck

We have shown that RL results in the limited vocabulary as it steals the probability mass from low-frequency words. Thus, increasing those low-frequency words is the easy yet critical solution to the bottleneck. One way to achieve this is to jointly optimize both the RL loss and the CE loss on ground-truth captions so that the low-frequency words in ground-truth captions would be more likely to be sampled during RL training [187]. However, this approach still relies on the sampling from a skewed policy and requires retraining from scratch.

To increase the actual vocabulary more effectively and efficiently, we refine the mapping from encoded features to low-frequency words. This refinement can be applied to any RL models and can be achieved by modifying only the mapping function parameters with a single-epoch fine-tuning.[2]

4.3.1 Simple Fine-Tuning (sFT)

The first method is a **simple fine-tuning (sFT)**. It is based on a decoupled two-stage training [81], which is a current strong baseline model for long-tail classification [169, 129, 189]. [81] decoupled the learning procedure into representation learning and classification, and then found that classification, *i.e.*, the mapping from representations to label distributions, is critical for long-tail classification. They decoupled the classifi-

[2]The code is available at https://github.com/ukyh/switch_disc_caption

cation model $f_{\boldsymbol{\theta}} : \mathcal{X} \to \mathbb{R}^c$ into an encoder $g_{\boldsymbol{\psi}} : \mathcal{X} \to \mathbb{R}^d$ and a classifier consisting of weight and bias parameters, $\boldsymbol{W} \in \mathbb{R}^{d \times c}$ and $\boldsymbol{b} \in \mathbb{R}^c$. Given an input $x \in \mathcal{X}$, $f_{\boldsymbol{\theta}}(x) = \boldsymbol{W}^\top g_{\boldsymbol{\psi}}(x) + \boldsymbol{b}$. Here, c is the number of classes and d is the dimension of encoder outputs.

Representation learning is the first stage of training, where they trained the entire classification model $f_{\boldsymbol{\theta}}$ on a full training dataset. The second stage is classification, where they fixed the encoder parameters $\boldsymbol{\psi}$ and adjusted only the classifier parameters. For the second-stage adjustment, they applied class-balanced sampling to encourage learning on low-frequency labels.

Following [81], we decouple a captioning model into an encoder and a classifier. In image captioning, the first-stage training of [81] corresponds to RL training on the full training dataset. The second-stage training corresponds to adjusting the classifier parameters on the *vocabulary-balanced* sequences. However, sampling from the skewed policy of text-generation models cannot provide sequences containing low-frequency words (Section 4.2.2). Thus, we use ground-truth captions as relatively vocabulary-balanced samples. sFT simply fine-tunes the *classifier parameters* of a pre-trained RL captioning model by minimizing the CE loss on ground-truth captions. The loss for each pair of images and ground-truth captions is as follows:

$$\mathcal{L}_{\mathrm{CE}}(\hat{\boldsymbol{\theta}}) = -\frac{1}{T}\sum_{t=1}^{T} \log p_{\hat{\boldsymbol{\theta}}}(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{I}), \tag{4.1}$$

where $\boldsymbol{y} = (y_1, ..., y_T)$ is a ground-truth caption of image $\boldsymbol{I}$ and $\hat{\boldsymbol{\theta}}$ denotes the model parameters $\boldsymbol{\theta}$ that are *initialized with RL training*. During this fine-tuning, *only the classifier parameters* $\{\boldsymbol{W}, \boldsymbol{b}\} \in \hat{\boldsymbol{\theta}}$ *are updated* with the gradients $\nabla_{\boldsymbol{W}} \mathcal{L}_{\mathrm{CE}}(\hat{\boldsymbol{\theta}})$ and $\nabla_{\boldsymbol{b}} \mathcal{L}_{\mathrm{CE}}(\hat{\boldsymbol{\theta}})$, respectively.

Let the softmax function $\pi : \mathbb{R}^{|\mathcal{W}|} \to \mathbb{R}$ be

$$\pi_{w_i,\beta}(\boldsymbol{z}) = \frac{\exp(\beta z_{w_i})}{\sum_{w_j \in \mathcal{W}} \exp(\beta z_{w_j})}, \tag{4.2}$$

where z_{w_i} indicates the element of a vector $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{W}|}$ at the index of a word $w_i \in \mathcal{W}$. $\mathcal{W}$ is the entire vocabulary. β is an inverse-temperature hyperparameter that controls the steepness of the softmax distribution. Then, the conditional probability $p_{\boldsymbol{\theta}}(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{I})$

is computed as follows:

$$p_{\boldsymbol{\theta}}(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{I}) = \pi_{y_t,\beta}(s^t_{\boldsymbol{\theta}}(\boldsymbol{y}_{<t}, \boldsymbol{I})), \tag{4.3}$$

$$s^t_{\boldsymbol{\theta}}(\boldsymbol{y}_{<t}, \boldsymbol{I}) = \boldsymbol{W}^\top g_{\boldsymbol{\psi}}(\boldsymbol{y}_{<t}, \boldsymbol{I}) + \boldsymbol{b}, \tag{4.4}$$

where $\boldsymbol{W} \in \mathbb{R}^{d \times |\mathcal{W}|}$ and $\boldsymbol{b} \in \mathbb{R}^{|\mathcal{W}|}$. Again, d is the output dimension of the encoder $g_{\boldsymbol{\psi}}$. We use LSTM [69] or Transformer [172] for $g_{\boldsymbol{\psi}}$.

4.3.2 Weighted Fine-Tuning (wFT)

Ground-truth captions contain more low-frequency words than sampled sequences, but some low-frequency words are still difficult to learn because of their low frequency. Our second method is **weighted fine-tuning (wFT)**, which further pursues vocabulary balance by rebalancing the loss of high-frequency words and low-frequency words in ground-truth captions.

To rebalance the loss, we exploit the frequency bias of RL models: RL models overly assign the probability to high-frequency words but not to low-frequency words. Given the properties of the frequency bias, fine-tuning for discriminativeness should focus more on the words that an RL model is *not* confident of but should be avoided on the words that an RL model is confident of. wFT incorporates these heuristics by modifying the probability $p_{\boldsymbol{\theta}}$ of $\mathcal{L}_{\mathrm{CE}}$ to the **bias product (BP)** [29, 61, 68] probability, $p_{\boldsymbol{\theta},\boldsymbol{\theta}'}$:

$$p_{\boldsymbol{\theta},\boldsymbol{\theta}'}(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{I}) = \pi_{y_t,1}\Bigg[\log \frac{\pi_{\cdot,\beta}(s^t_{\boldsymbol{\theta}}(\boldsymbol{y}_{<t}, \boldsymbol{I}))}{p_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{y}_{<t}, \boldsymbol{I})} + \log \frac{\pi_{\cdot,\beta'}(s^t_{\boldsymbol{\theta}'}(\boldsymbol{y}_{<t}, \boldsymbol{I}))}{p_{\boldsymbol{\theta}'}(\cdot \mid \boldsymbol{y}_{<t}, \boldsymbol{I})}\Bigg], \tag{4.5}$$

where $\pi_{\cdot,\beta}(\boldsymbol{z}) \in \mathbb{R}^{|\mathcal{W}|}$. By inserting $p_{\boldsymbol{\theta},\boldsymbol{\theta}'}$ into $\mathcal{L}_{\mathrm{CE}}$, we define the objective function of wFT as follows:

$$\mathcal{L}_{\mathrm{BP}}(\hat{\boldsymbol{\theta}}) = -\frac{1}{T}\sum_{t=1}^{T} \log p_{\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\theta}}'}(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{I}). \tag{4.6}$$

Similar to sFT, the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are *initialized with the same RL model to be* $\hat{\boldsymbol{\theta}}$ *and* $\hat{\boldsymbol{\theta}}'$. The difference is that, although the classifier parameters of $\hat{\boldsymbol{\theta}}$ are updated, *all the parameters of* $\hat{\boldsymbol{\theta}}'$ *are fixed* during fine-tuning[3]. Figure 4.3 shows the change in the BP loss compared to the CE loss. The BP severely suppresses the loss when the

[3] [22] also utilized fixed pre-trained models to reweight their loss for stylized image captioning. How-

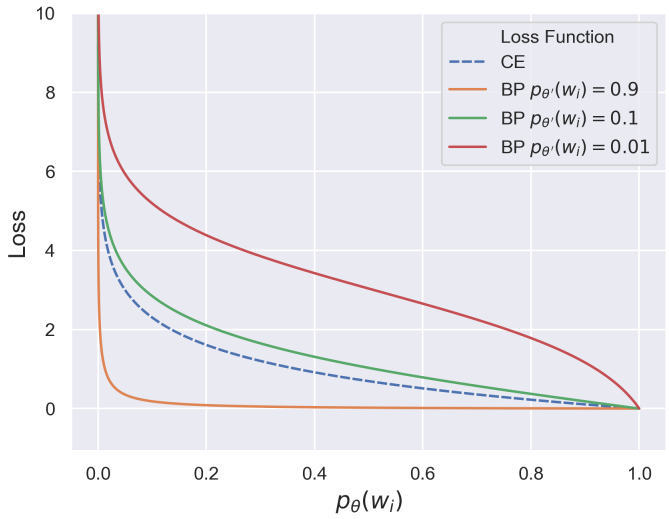


Figure 4.3: Visualization of the CE loss $-\log p_{\boldsymbol{\theta}}(w_i)$ and BP loss $-\log p_{\boldsymbol{\theta},\boldsymbol{\theta}'}(w_i)$. To compute the BP loss, we need the entire distribution of $\{p_{\boldsymbol{\theta}}(w_i)\}_{w_i \in \mathcal{W}}$ and $\{p_{\boldsymbol{\theta}'}(w_i)\}_{w_i \in \mathcal{W}}$. Here, we set the index i to 1 and assigned $\frac{1}{5}(1 - p_{\boldsymbol{\theta}}(w_1))$ to the words of the next five indices, $w_2, ..., w_6$. This is because we observed that the five most probable words occupied 99% of the probability in the output distribution of the RL models. We assumed that the five most probable words were the same between $p_{\boldsymbol{\theta}}$ and $p_{\boldsymbol{\theta}'}$ as the parameters were initialized with the same RL model. Thus, we assigned $\frac{1}{5}(1 - p_{\boldsymbol{\theta}'}(w_1))$ to the words of the next five indices, $w_2, ..., w_6$, likewise $p_{\boldsymbol{\theta}}$. Here, β and β' were set to 1.

frequency-biased policy $p_{\boldsymbol{\theta}'}$ is confident, and largely increases the loss when $p_{\boldsymbol{\theta}'}$ is not confident. In this way, the BP allows models to unlearn the frequency bias learned with RL. As with sFT, only the classifier parameters $\{\boldsymbol{W}, \boldsymbol{b}\} \in \hat{\boldsymbol{\theta}}$ are updated with the gradients $\nabla_{\boldsymbol{W}} \mathcal{L}_{\mathrm{BP}}(\hat{\boldsymbol{\theta}})$ and $\nabla_{\boldsymbol{b}} \mathcal{L}_{\mathrm{BP}}(\hat{\boldsymbol{\theta}})$, respectively.

The previous BP methods used the probability $p_{\boldsymbol{\theta}}$ during evaluation to avoid incor-

ever, their method is designed to train new models from scratch and is not applicable to refining pre-trained models; their loss function (Eq. (6) in [22]) is stuck at zero when we initialize the parameters with the same pre-trained model. This requirement for retraining from scratch is a fundamental deviation from our goal of improving the discriminativeness of off-the-shelf RL models.

porating the bias of $p_{\boldsymbol{\theta}'}$ into the predictions [29, 61]. Although it worked well in their classification tasks, we found this train–test gap makes the decoding unstable in text generation. To mitigate the train–test gap, we use two variants of decoding: (1) decode with $p_{\boldsymbol{\theta}}$ but use a small β' for $p_{\boldsymbol{\theta}'}$ during training to ease the gap between $p_{\boldsymbol{\theta}}$ and $p_{\boldsymbol{\theta},\boldsymbol{\theta}'}$, or (2) use $p_{\boldsymbol{\theta},\boldsymbol{\theta}'}$ during both training and decoding (**BP decoding**) as $p_{\boldsymbol{\theta},\boldsymbol{\theta}'}$ itself is already less biased than $p_{\boldsymbol{\theta}'}$.

4.4 Experiments

4.4.1 Setup

Dataset and Metrics. We used the MS COCO captioning dataset[4] [109, 23] with Karpathy splitting [83]. After preprocessing, the entire vocabulary size $|\mathcal{W}|$ was 9,487[5]. In the evaluation, the captions were decoded using a beam search of size 5 and evaluated using various evaluation metrics. Specifically, we used CIDEr [174] SPICE [3], BERTS++ [201], TIGEr [76], CLIPS, and RefCLIPS [67]. Note that the correlation with human judgments increases in the above order, with RefCLIPS indicating the state-of-the-art correlation [67, 84]. Following the previous studies [111, 182, 161], we evaluated discriminativeness with **R@K** scores: the percentage of captions with which a pre-trained image–text retrieval model [45] could correctly retrieve the original images from the entire validation/test images within the rank of K $\in \{1, 5, 10\}$. A higher R@K indicates that the model generates more discriminative captions with characteristic information of images. Evaluation was conducted in a single run for each model. See Appendix B.3 for the libraries and settings we used for these evaluations.

Comparison Models. Following [182], we used Att2in [151], UpDown [5], and Transformer [172] as the baseline models. The models were pre-trained with the standard RL [151] and are publicly available[6]. In addition to the baseline models, we compared our models with discriminativeness-aware models, which were state-

[4]Each split of training/validation/test contained 113,287/5,000/5,000 images, and each image had around five ground-truth captions.

[5]The words that occur less than five times in the training captions were converted to $\langle \mathrm{unk} \rangle$ token.

[6]`https://github.com/ruotianluo/self-critical.pytorch`: {`Att2in`, `UpDown`, `Transformer`}+`self_critical`

of-the-art at the time of the submission of the work on which this chapter is based: **CIDErBtw** [182], **NLI** [161], **DiscCap** [117], and **Visual Paraphrase** [111]. The first three created new discriminativeness rewards to be optimized with RL. Visual Paraphrase introduced a new model architecture to paraphrase simpler captions to more complex captions. See Section 4.5 for more details of these models. As we mentioned in the beginning of Section 4.3, the CE loss on ground-truth captions can be utilized in a different way from our methods. We report the results of jointly optimizing the RL loss and CE loss (**Joint CE** [187, 42]). It optimizes $\mathcal{L}_{\text{Joint}}(\boldsymbol{\theta}) = \lambda\mathcal{L}_{\text{RL}}(\boldsymbol{\theta}) + (1-\lambda)\mathcal{L}_{\text{CE}}(\boldsymbol{\theta})$ during RL training. We also tested **Only CE**, which sets $\lambda = 0$ to solely optimize the CE loss, as the baseline without RL. See Appendices B.10 and B.11 for the comparisons with more recent models [203, 105, 25].

Hyperparameters. Our models used the same hyperparameters as the baseline models, except for the epoch size, learning rate, and β' in Eq. (4.5). We set the epoch size for fine-tuning to 1 and searched for the best learning rate from {1e-3, 1e-4, 1e-5, 1e-6}. For BP in Eq. (4.5), we set $\beta = 1$ and searched for the best β' from {0.1, 1}. As with our models, we set all hyperparameters of the CE-based models to the same as the baseline models except for the $\lambda \in \{0, 0.2, 0.5, 0.8\}$. We disabled scheduled sampling [13] for our fine-tuning and the CE loss to separate them from the RL loss strictly. We took the best hyperparameters according to the R@1 scores in the validation set. Note that *we used different hyperparameters for the wFT with different decoding methods* (See Section 4.3.2). Appendix B.4 shows the best hyperparameters. We followed the previous work for the hyperparameters of the other models.

All the models except Visual Paraphrase had the same size of trainable parameters as their baselines. See Appendix B.5 for the exact number of parameters. Our fine-tuning was completed in around 10 minutes using a single GPU of 16 GB memory. See Appendix B.6 for the exact time for training and comparison with other methods.

		Vocabulary			*Standard Evaluation*						*Discriminativeness*		
		Unique-1	Unique-S	Length	CIDEr	SPICE	BERTS++	TIGEr	CLIPS	RefCLIPS	R@1	R@5	R@10
Att2in	**Att2in RL**	445	2,524	9.3	117.4	20.5	43.6	73.9	73.0	79.7	16.3	41.9	57.2
	+ sFT	880	3,156	9.0	115.4	20.4	**43.9**	74.3	73.7	**80.3**	20.1	48.0	62.8
	+ wFT	**1,197**	**3,732**	8.9	104.3	19.5	43.1	74.2	73.9	80.2	20.6	49.7	64.5
	+ wFT (BP decoding)	1,102	3,615	9.4	109.3	20.1	43.7	**74.4**	**74.0**	80.2	21.1	**50.5**	64.8
	CIDErBtw	470	2,630	9.3	**119.0**	20.7	43.8	74.1	73.1	79.8	17.2	44.1	58.7
	NLI	465	2,626	9.2	118.9	20.6	43.8	74.1	73.2	79.9	17.6	44.4	59.8
	DiscCap†		3,093	9.3	114.2	**21.0**					**21.6**	50.3	**65.4**
	Joint CE	700	2,907	9.1	111.7	19.9	43.5	74.0	73.3	80.0	19.1	46.7	61.5
	Only CE	689	2,845	9.2	110.7	20.1	43.5	74.0	73.3	79.9	19.0	46.6	61.1
	Visual Paraphrase†		4,576	12.9	86.9	21.1					26.3	57.2	70.8
UpDown	**UpDown RL**	577	3,103	9.5	122.7	**21.5**	44.2	74.6	74.0	80.5	21.1	49.9	64.6
	+ sFT	1,190	3,788	9.2	115.9	21.0	44.2	**74.9**	74.8	**80.9**	25.0	56.8	71.2
	+ wFT	**1,479**	**4,268**	9.1	101.8	19.5	43.1	74.6	74.9	80.7	26.0	57.6	72.2
	+ wFT (BP decoding)	1,275	4,177	9.6	110.0	20.6	44.1	**74.9**	**75.0**	80.8	**26.7**	**58.7**	**72.4**
	CIDErBtw	582	3,108	9.4	**123.0**	**21.5**	**44.4**	74.6	74.2	80.7	21.9	50.9	65.9
	NLI	575	3,144	9.4	122.4	21.4	**44.4**	74.6	74.1	80.6	21.5	50.7	65.6
	Joint CE	857	3,120	9.4	111.8	20.5	43.7	74.3	73.8	80.2	21.8	51.2	65.2
	Only CE	878	3,126	9.4	109.2	20.1	43.4	74.2	73.6	80.0	21.8	49.9	64.5
Transformer	**Transformer RL**	753	3,433	9.2	127.7	22.5	45.1	75.0	75.0	81.3	26.6	56.2	70.5
	+ sFT	1,458	3,959	9.1	118.7	21.7	44.8	75.2	75.6	**81.5**	30.6	62.3	75.7
	+ wFT	1,776	4,274	9.1	103.1	20.0	43.3	74.8	75.8	81.2	32.5	64.5	77.1
	+ wFT (BP decoding)	**1,964**	**4,373**	9.4	107.3	21.1	44.2	75.2	**76.1**	**81.5**	**33.5**	**65.9**	**78.2**
	CIDErBtw	837	3,609	9.5	128.2	22.6	45.1	75.2	75.0	81.2	27.7	57.6	71.6
	NLI	876	3,744	9.5	**129.1**	**23.0**	**45.4**	**75.3**	75.5	**81.5**	29.8	59.9	73.4
	Joint CE	1,083	3,491	9.3	123.8	21.9	45.0	74.8	75.0	81.2	27.3	57.2	70.8
	Only CE	935	3,599	9.4	112.2	20.8	44.0	74.5	74.8	80.9	26.5	55.8	69.7

Table 4.1: Comparison of **baseline models**, our models, and discriminativeness-aware models. Automatic evaluation results on the MS COCO test set. *Unique-1* and *Unique-S* indicate the number of unique unigrams and sentences, respectively. *Length* is the average length of the output captions. Scores with † were reported in [111]. Other scores were reproduced by us.

4.4.2 Comparison with Baseline Models and Discriminativeness-Aware Models

Table 4.1 shows the results compared to those obtained with the baseline models and discriminativeness-aware models.

Vocabulary. First, we observe that our methods (sFT and wFT) successfully increase the actual vocabulary size: both of them considerably increased Unique-1 compared to all the baseline models. wFT increased the vocabulary more than sFT, indicating that

rebalancing the loss further encouraged low-frequency word generation. The increased vocabulary resulted in more specific captions to each image: Unique-S also increased significantly. Consistent with previous studies [187, 111, 182], the models trained with the CE loss (Joint CE and Only CE) achieved the larger vocabulary than the baseline RL models. However, the improvement of our methods was even larger than these CE-based models. Despite the significant increase in the vocabulary size, our method kept the captions concise: the average sentence length was close to those of the baseline models.

Discriminativeness. Our goal is to enhance the discriminativeness of RL models by addressing their limited vocabulary. As expected, our methods successfully improved the discriminativeness: the R@K scores of our models were considerably higher than those of the baselines. Corresponding to the better improvement in vocabulary size, wFT increased discriminativeness more than sFT. *These results confirm our hypothesis that the limited vocabulary of RL models has been a major bottleneck for discriminativeness*.

Among the Att2in-based models, Visual Paraphrase achieved the highest discriminativeness. However, this model is not directly comparable to the others because it increases the trainable parameters for its specialized model architecture. Moreover, its improvement in discriminativeness was achieved at the expense of conciseness, which is another desirable property for discriminative image captions [154]: its sentence length was substantially longer than the other models. DiscCap performed comparably with our models, but its reward requires high computational costs. CIDErBtw and NLI proposed more lightweight rewards to be applicable to larger models, but they still need retraining from scratch. Among the larger models (UpDown and Transformer), our models achieved the highest discriminativeness despite the small computational cost.

Standard Evaluation. As our methods increase low-frequency words in outputs, the outputs are likely to include the words that are **out-of-references (OOR)**. That is, low-frequency words may not be covered by reference captions regardless of their correctness due to the low frequency. These low-frequency OOR words unfairly decrease scores in conventional evaluation metrics because those metrics count *exact matches*

in the surface form of text[7].

To fairly evaluate the OOR words, recent metric research has focused on *soft matching* metrics [77, 67]. Soft-matching metrics can evaluate the semantic similarity between target captions and reference captions beyond the surface form of text by utilizing pre-trained language models (PLMs) [201, 204] or pre-trained cross-modal models (PCMs) [76, 100, 67]. Their correlation with human judgments is significantly higher than that of exact-matching metrics in both precision and recall [84]. In particular, PCM-based metrics, which can utilize image features in addition to reference captions, have substantially enhanced the evaluation performance and have achieved the state-of-the-art correlation with human judgments [67, 84].

Given the above advantages, we employed soft-matching metrics in addition to conventional exact-matching metrics. Not surprisingly, our models decreased the scores in the exact-matching metrics (CIDEr and SPICE). However, our models scored comparably with the baselines in the PLM-based metric (BERTS++) and rather outperformed them in the state-of-the-art PCM-based metrics (TIGEr, CLIPS, and RefCLIPS). The higher performance in the superior soft-matching metrics indicates that our methods do not degrade the overall quality of captions. To further validate the overall quality of our output captions, the following Section 4.4.3 analyzes the cause of this performance gap in more detail.

4.4.3 Analysis of the Performance Gap

Properties of OOR Words. The critical difference between the conventional exact-matching metrics and the recent soft-matching metrics is the (in)ability to evaluate OOR words[8]. Based on the difference, we hypothesize that the performance gap is caused by a difference in the properties of OOR words. We analyzed the OOR words of our models, comparing with those of RL baselines and Only CE, which scores similarly to our models in exact-matching metrics but decreases soft-matching scores

[7] Some metrics use stemming, lemmatization, and/or WordNet synsets to evaluate synonyms but their coverage is limited.

[8] Note that this difference does not mean that exact-matching metrics represent precision, and soft-matching metrics represent recall. Exact-matching metrics cannot represent precision because the reference captions do not cover all correct descriptions. That is, exact-matching metrics can only represent the flawed precision with false negatives. Actually, exact-matching metrics correlate with human judgments worse than soft-matching metrics not only in recall but also in precision [84].

	Repetition	*OOR*		*Text-Based*			*Text-and-Image-Based*		
				Exact-Matching			*Soft-Matching*		
	Rep (%) ↓	Number ↓	Rank ↑	CIDEr	SPICE	BERTS++	TIGEr	CLIPS	RefCLIPS
Att2in RL	4.1	**8,665**	79.4	**117.4**	**20.5**	43.6	73.9	73.0	79.7
+ sFT	3.8	8,813	164.0	115.4	20.4	**43.9**	74.3	73.7	**80.3**
+ wFT	**3.2**	10.454	**237.9**	104.3	19.5	43.1	74.2	73.9	80.2
+ wFT (BP decoding)	3.6	10,386	204.7	109.3	20.1	43.7	**74.4**	**74.0**	80.2
Only CE	3.9	9,913	133.1	110.7	20.1	43.5	74.0	73.3	79.9
UpDown RL	3.9	**8,463**	100.1	**122.7**	**21.5**	**44.2**	74.6	74.0	80.5
+ sFT	3.6	9,252	225.8	115.9	21.0	**44.2**	**74.9**	74.8	**80.9**
+ wFT	**3.0**	11,478	**301.0**	101.8	19.5	43.1	74.6	74.9	80.7
+ wFT (BP decoding)	3.4	11,065	236.9	110.0	20.6	44.1	**74.9**	**75.0**	80.8
Only CE	3.7	10,874	152.9	109.2	20.1	43.4	74.2	73.6	80.0
Transformer RL	3.6	**7,824**	129.8	**127.7**	**22.5**	**45.1**	75.0	75.0	81.3
+ sFT	3.2	9,397	296.0	118.7	21.7	44.8	**75.2**	75.6	**81.5**
+ wFT	**2.6**	11,930	379.7	103.1	20.0	43.3	74.8	75.8	81.2
+ wFT (BP decoding)	2.9	11,673	**461.0**	107.3	21.1	44.2	**75.2**	**76.1**	**81.5**
Only CE	3.3	10,661	165.6	112.2	20.8	44.0	74.5	74.8	80.9
Human	**2.4**	17,963	**815.6**	88.4	21.2	42.9	73.3	**77.7**	**82.0**

Table 4.2: Comparison of OOR words and the resulting difference in exact-matching and soft-matching metrics. We report the results on the MS COCO test set. A higher value in *Rank* indicates a lower frequency rank of the OOR words. We also report the rate of repetition.

in contrast to our models. Table 4.2 shows the number of OOR words and their average **frequency rank**. The frequency rank refers to the order of words when sorted by their frequency in training captions; the most frequent word ranks 1st, and the value of rank increases as the frequency decreases. Although our models and Only CE output the similar number of OOR words, the significant difference in the frequency rank indicates that the properties of our OOR words are different from those of Only CE; that is, the OOR words of our models consist of much more low-frequency words than those of Only CE. Low-frequency words are likely to be OOR by the nature of their frequency, regardless of their correctness.

The soft-matching metrics could tell this difference and scored our models higher than Only CE models and even higher than baseline RL models. Especially, this tendency was more clear in the state-of-the-art PCM-based metrics (TIGEr, CLIPS, and

RefCLIPS). On the contrary, the exact-matching metrics (CIDEr and SPICE) could not tell the difference by definition and decreased the scores roughly in proportion to the number of OOR words. Appendix B.7 shows the qualitative analysis of the underrated captions.

Comparison with Human-Annotated Captions. Manually annotated captions are known to show low exact-matching scores although they achieve substantially higher scores in manual evaluation [84, 33]. In Table 4.2, we observe that human-annotated captions (*Human*)[9] have similar properties to ours: a large number of low-frequency OOR words, low exact-matching scores, but high scores in the state-of-the-art metrics (CLIPS and RefCLIPS).

Repetition. We also confirmed that the decrease in exact-matching scores was not caused by repetition, which is a typical side effect of heavily maximizing discriminativeness rewards [187, 178]. Table 4.2 shows that our models' repetition rates[10] were rather lower than those of baselines.

Conclusion. From the above results, we conclude that the lower exact-matching scores of our models are caused by the nature of low-frequency words and the deficiency of exact-matching metrics, not by the degeneration of our models. The results of the human evaluation in the following Section 4.4.4 further support this conclusion.

4.4.4 Human Evaluation

As discussed in Sections 4.4.2 and 4.4.3, automatic evaluation of our models has difficulty due to the OOR words caused by the low frequency. To further validate the performance of our models, we conducted human evaluations using Amazon Mechanical Turk (AMT) on three criteria: discriminativeness, correctness, and fluency. Correctness and fluency are *absolute scores*: we instructed workers to give a maximum score 5

[9] Following [111, 33], we randomly sampled one reference caption for each image and evaluated the similarity against the rest of the references.

[10] Let $\mathcal{C}$ be a set of captions; $f^n(\cdot)$ and $u^n(\cdot)$ be the functions to return n-grams and unique n-grams, respectively. We computed the repetition rate (*Rep*) by $\frac{1}{|\mathcal{C}|N}\sum_{i=1}^{|\mathcal{C}|}\sum_{n=1}^{N} 1 - \frac{|u^n(\mathcal{C}_i)|}{|f^n(\mathcal{C}_i)|}$, where we set $N = 4$.

	Discriminativeness	Correctness	Fluency
Transformer RL	3.00	4.42	4.83
+ wFT	**3.34****	4.45	**4.84**
NLI	3.18**	**4.54**	4.76

Table 4.3: Human evaluation results on the subset of the MS COCO test set. The discriminativeness score of Transformer RL was fixed at 3.00 because we set it as the baseline. */** indicates that a score is statistically significantly different from that of the baseline model (t-test with $p < 0.05/0.01$); one-sample t-test for discriminativeness and independent two-sample t-test for the other criteria.

to the captions that *did not* contain incorrect information (ungrammatical or unnatural expressions) in terms of correctness (fluency). In contrast, discriminativeness is designed as a *relative score* because it is difficult to set an absolute standard for discriminativeness; unlike correctness or fluency, we cannot define the perfectly discriminative captions. Following [182], we instructed the workers to determine the discriminativeness of a caption by comparing the caption with that of a baseline model[11].

We evaluated the Transformer-based models, which performed the best in the automatic evaluation. Although wFT with BP decoding performed better, here we picked up wFT with $p_{\boldsymbol{\theta}}$ decoding to set the total number of parameters for decoding strictly the same across the models. Following [182], we randomly selected 50 images from the MS COCO test set and assigned five workers to each image. See Appendix B.8 for more details on the AMT instruction. Table 4.3 shows the results. wFT, which had the highest R@K scores, also achieved the highest discriminativeness here. wFT achieved the same or higher correctness and fluency than the baseline model, in contrast to the exact-matching scores in Table 4.2. These results are consistent with the results of the state-of-the-art soft-matching metrics, confirming again that our methods do not degrade the quality of captions.

[11] If a target caption describes the same information as a baseline caption, the workers give the target caption a score of 3; if the target caption describes more (less) characteristic information than the baseline caption, the workers give the target caption a score of 4 or 5 (1 or 2).

4.5 Related Work

Image Captioning is the task of describing images in natural languages. The quality of captions has been remarkably improved by recent advances such as the encoder–decoder captioning model [181], attention mechanism [198], RL training [146, 151], attention over bounding box features [5], large-scale pre-training [105], and large-scale captioning datasets [202, 109, 23, 93, 157]. Despite these advancements, current captioning models generate overly generic captions [34, 33, 187, 190].

Discriminative Image Captioning has been explored to generate more informative captions. [154] was the first to study this task. They defined the more informative captions as the captions that *concisely* describe the information discriminative from *distractor images*, *i.e.*, images similar to an input image. [7] proposed neural listener and speaker models that cooperate to generate discriminative captions for abstract scenes. [133] adapted the models to single-colored images. [173] and [30] extended the domain to real images and improved inference efficiency. [183] proposed a memory attention network to describe unique objects among distractor images. [124] introduced a dataset with harder distractor images.

These approaches require selecting distractor images for inference. [117] and [112] proposed the methods that do not require this step. Their models learn to generate discriminative captions by maximizing the R@K scores for sampled captions using RL [151]. The R@K scores are computed with a pre-trained image–text retrieval model [45] over images in a mini-batch. [178] proposed a method to jointly train the image–text retrieval model and captioning model. Despite their effectiveness, R@K scores are associated with high computational costs and require a large batch size. CIDErBtw [182] and NLI [161] achieved state-of-the-art discriminativeness at this work's submission time with more lightweight rewards. They weighted the contribution of ground-truth captions for the CIDEr reward according to their differences from similar but different captions [182] or their entailment scores against other ground-truth captions [161]. Another approach exploited unrelated captions as negative examples and trained caption generators with contrastive learning [34] or GAN [34, 53].

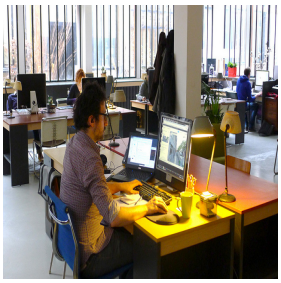
Visual Paraphrase [111] and [194] are related to our work in that they exploited low-frequency n-grams to enhance discriminativeness. [111] divided ground-truth captions into two subsets according to n-gram TF-IDF scores and proposed a new model to

paraphrase low TF-IDF captions into high TF-IDF ones. [194] proposed the use of n-gram TF-IDF scores as an additional reward to a variant of R@K reward.

Different from above approaches, our objective is set to remedy the low discriminativeness of existing RL models. Our models can be achieved with single-epoch fine-tuning of pre-trained RL models, without requiring either drastic changes in the model architecture [111], additional computational costs of rewards [194], or retraining from scratch.

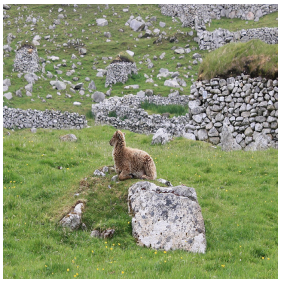
Diverse Image Captioning is the task of generating a set of diverse captions for a given image [191]. Diverse image captioning is aimed at enumerating various pieces of information with a set of captions, whereas discriminative image captioning aims to concisely describe the most characteristic information with a single caption. Similar to this study, some studies utilized captions that contained more low-frequency words, such as ground-truth captions [187, 118] or captions sampled from CE models [160]. Their models learn to generate these captions in addition to the captions sampled from RL models. However, these approaches still rely on sampling from skewed policies and require retraining of a model from scratch. Other approaches have adapted GAN [159], conditional VAE [184, 8, 18, 121, 164, 90, 152], flow-based generative models [120, 40], RL [21], POS tag sequences [38], and beam search [180, 73] to enhance the diversity within a set of captions.

Long-Tail Classification has been studied extensively in various tasks as label imbalance is prevalent across datasets [205, 104]. In text-generation tasks, label imbalance exists in the frequency of words. Previous approaches have addressed the imbalance by normalizing classifier weights [134, 148] or using variants of Focal loss [148, 57, 77, 197, 108]. In contrast to these approaches, we adapted long-tail classification to mitigate the side effects of RL in the context of discriminative image captioning. Appendix B.9 shows that our methods outperformed these approaches.



Transformer RL:
a man sitting at a desk with a computer

+wFT:
a person sitting at a desk with multiple computers



Transformer RL:
a sheep laying on the grass in a field

+wFT:
an animal that is laying down on some grass

Transformer RL:
a living room with a couch and a table

+wFT:
a living room filled with white furniture and red walls

Figure 4.4: Examples of the limitation of our methods. All the examples are from the MS COCO validation set. The underlined words are relatively low-frequency hypernyms.

4.6 Limitations

Our experiments were limited to the MS COCO dataset, although it is the standard dataset for image captioning. The images belong to the general domain (images of common objects), and the captions are in English only. To compensate for the limitation, we have demonstrated the effectiveness of our methods with the multiple baseline models.

Our current methods have a limitation in that they cannot select discriminative ones among low-frequency words. Although discriminative in general, low-frequency words do not always describe more specific information than others. Figure 4.4 shows

the examples. Our model output relatively low-frequency hypernyms such as *person*, *animal*, and *furniture* instead of the more frequent but more specific hyponyms: *man*, *sheep*, and *couch*. Utilizing thesauruses like WordNet [47] will be a promising approach to reduce those relatively low-frequency hypernyms from outputs.

The MS COCO dataset contains social biases, and captioning models have the risk of amplifying those biases [207, 208, 65]. Our methods are also not free from the risk, as they are not designed to reduce those social biases from existing models.

4.7 Conclusion

We have investigated the cause of overly generic captions of RL models and found out that RL decreases the discriminativeness by limiting the output words to high-frequency words. We propose the lightweight fine-tuning methods to address the bottleneck directly and achieve significantly higher discriminativeness with only the slight modification on off-the-shelf RL models. Our identification of the bottleneck and practical solutions will significantly impact future research on discriminative image captioning.

As an additional practical advantage, our models can control the granularity of descriptions from coarse to fine by just switching the off-the-shelf/fine-tuned classifier parameters. In terms of broader impact, our methods can be easily applied to the RL models in other text generation tasks, such as machine translation [196], summarization [141], and dialogue generation [103] to enrich the output vocabulary. In terms of broader impact, our methods can be easily applied to the RL models in other text generation tasks, such as machine translation [196], summarization [141], and dialogue generation [103] to enrich the output vocabulary.

Chapter 5

Conclusion

5.1 Summary

This dissertation aims to improve image captioning systems with the goal of making them more versatile. Towards this goal, we have proposed novel methods to broaden the coverage of image captioning systems on both sides of scenes and descriptions.

Chapter 3 has addressed unsupervised image captioning. In this chapter, we have broadened the scene coverage by handling edge cases in the scene coverage, where scenes have no corresponding image–sentence pairs during training. Based on the observation that pseudo-captions often contain words that are irrelevant to images, we have proposed the gating mechanism and pseudo-labels to remove word-level spurious alignment between images and pseudo-captions. Experimental results have shown that our models significantly outperform previous models.

Chapter 4 has addressed discriminative image captioning. In this chapter, we have broadened the description coverage by enriching the output vocabulary of off-the-shelf RL models. First, we have investigated the outputs of current captioning systems and show that RL decreases the output vocabulary. Then, based on this finding, we have proposed lightweight fine-tuning methods to increase the output vocabulary so that captions will subsequently include the information specific to each image. Extensive experiments have demonstrated that our methods substantially enhance the vocabulary size and discriminativeness of output captions.

These results indicate our sound contribution towards versatile image captioning systems on both sides of coverage.

5.2 Limitations and Future Directions

Aside from the limitations specific to each study (Sections 3.5 and 4.6), we discuss overall limitations of this dissertation.

Combination of Methods

Our studies have found that the small vocabulary is a crucial issue to be addressed in the future, both in supervised and unsupervised image captioning. Thus, the next important goal is to make our discriminative image captioning methods applicable to various task settings.

The first limitation is, however, the difficulty of combining our methods. Although our work covers both sides of coverage, the proposed methods are difficult to combine due to the difference in task settings. Our discriminativeness-aware methods require ground-truth captions, but they are unavailable in unsupervised image captioning. A future direction of our work is to combine and apply our methods to semi-supervised image captioning tasks where at least a few ground-truth captions are available [87].

Extrinsic Evaluation

The second limitation is the lack of extrinsic evaluation in downstream tasks such as image searching, visual question answering, and visual dialogue generation. Evaluating captioning systems on performance in downstream tasks will give a more thorough comparison in terms of versatility.

Coverage of Tasks

The third limitation is the small coverage of tasks on each side: scenes and descriptions. As shown in Section 1.2, other important challenges remain besides unsupervised image captioning and discriminative image captioning. A unified sequence-to-sequence learning framework [186] will be a promising research direction to address all these challenges at once.

[186] trained a single model on a variety of tasks with prompts designed for each task. This framework can be applied to handle various image captioning tasks with a

single model. For example, one can train a model on each task with the concatenation of scene-aware prompts (*e.g.*, images of common objects, images with text, *etc.*) and description-aware prompts (*e.g.*, general descriptions, discriminative descriptions, *etc.*), then change the combination of the prompts at the inference stage according to the use.

Major Components other than Models

The last limitation is that our work does not address challenges in evaluation metrics and datasets, which are the rest of the major components in image captioning other than model development (Section 1.2). While the focus of this dissertation is not on these components, we provide future directions in them as follows.

Evaluation Metrics. As seen in Section 4.4.3, exact-matching evaluation metrics unfairly penalize captions with correct-but-OOR words. Recent studies have also shown that exact-matching metrics do not correlate well with human evaluation compared with soft-matching metrics [84, 67]. However, exact-matching metrics are still the most widely-used metrics, even in recent studies [105, 203, 186, 135]. This ignorant of soft-matching metrics is problematic as it might encourage the creation of systems that generate less informative captions.

To facilitate the use of soft-matching metrics, future work should re-evaluate the evaluation metrics to identify their strengths and weaknesses. Extending our analysis of Section 4.4.3 by manually evaluating correct-but-OOR words will be a promising direction to validate the superior performance of soft-matching metrics further[1].

Datasets. Increasing data is a simple yet highly effective way to broaden the coverage of scenes and descriptions. Although web-crawled image–sentence pairs have not been used for training captioning models due to their noise, filtering the pairs will enable their direct use for training. Recent studies apply cross-modal matching models to compute image–sentence similarity and filter out noisy pairs [102, 82]. Another line of recent approaches utilizes noisy image–sentence pairs to train adapters to

[1]Low reference coverage has also been reported as a problem in machine translation evaluations [50]. Therefore, our analysis might be extended to other text-generation evaluations, including machine translation.

fuse pre-trained unimodal models, thereby leveraging the knowledge of each modality [20, 2, 101].

Challenges in General Image Captioning

Recent advances in image captioning have achieved human-level performance in the task of general image captioning: the task of outputting correct and fluent descriptions given images of common objects. Recent captioning models considerably outperformed humans in the standard automatic evaluation metrics [111] and performed close to humans even in manual evaluation [84]. Thus, the remaining challenges have been shifted to broadening the coverage of scenes [1, 163] and descriptions [187, 84], which we have focused on in this dissertation.

However, there is still a performance gap from humans in terms of correctness (not fluency), and even the state-of-the-art models output incorrect descriptions in rare cases [84]. While improving the correctness is a straightforward contribution, it is also practically helpful to notify users of the images that models cannot handle. To this end, calibrating the captioning model's certainty is a promising direction for future research. In terms of application, improving the efficiency of captioning models is also important for future work [185].

We hope this dissertation serves as a milestone for these future studies.

References

[1] H. Agrawal, K. Desai, Y. Wang, X. Chen, R. Jain, M. Johnson, D. Batra, D. Parikh, S. Lee, and P. Anderson. nocaps: novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8947–8956, 2019.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[3] P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: Semantic propositional image caption evaluation. In *Proceedings of the European Conference on Computer Vision*, pages 382–398, 2016.

[4] P. Anderson, S. Gould, and M. Johnson. Partially-supervised image captioning. In *Advances in Neural Information Processing Systems*, 2018.

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[6] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

[7] J. Andreas and D. Klein. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, 2016.

[8] J. Aneja, H. Agrawal, D. Batra, and A. Schwing. Sequential latent spaces for modeling the intention during diverse image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4261–4270, 2019.

[9] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[10] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, 2015.

[11] S. Basu, G. S. Ramachandran, N. S. Keskar, and L. R. Varshney. Mirostat: A neural text decoding algorithm that directly controls perplexity. In *Proceedings of the International Conference on Learning Representations*, 2021.

[12] A. Ben Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller. VQA-Med: Overview of the medical visual question answering task at imageclef 2019. In *CLEF2019 Working Notes*, 2019.

[13] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, 2015.

[14] A. F. Biten, L. Gomez, M. Rusinol, and D. Karatzas. Good news, everyone! context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12466–12475, 2019.

[15] S. Cao, G. An, Z. Zheng, and Q. Ruan. Interactions guided generative adversarial network for unsupervised image captioning. *Neurocomputing*, 417:419–431, 2020.

[16] K.-Y. Chang, K.-H. Lu, and C.-S. Chen. Aesthetic critiques generation for photos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3534–3543, 2017.

[17] S. Changpinyo, D. Kukliansy, I. Szpektor, X. Chen, N. Ding, and R. Soricut. All you may need for VQA are image captions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1947–1963, 2022.

[18] F. Chen, R. Ji, J. Ji, X. Sun, B. Zhang, X. Ge, Y. Wu, F. Huang, and Y. Wang. Variational structured semantic inference for diverse image captioning. In *Advances in Neural Information Processing Systems*, 2019.

[19] H. Chen, A. Trouve, K. J. Murakami, and A. Fukuda. An intelligent annotation-based image retrieval system based on rdf descriptions. *Computers and Electrical Engineering*, 58:537–550, 2017.

[20] J. Chen, H. Guo, K. Yi, B. Li, and M. Elhoseiny. VisualGPT: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040, 2022.

[21] J. Chen and Q. Jin. Better captioning with sequence-level exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10890–10899, 2020.

[22] T. Chen, Z. Zhang, Q. You, C. Fang, Z. Wang, H. Jin, and J. Luo. "Factual" or "Emotional": Stylized image captioning with adaptive learning and attention. In *Proceedings of the European Conference on Computer Vision*, pages 519–535, 2018.

[23] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft COCO Captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[24] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. UNITER: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision*, pages 104–120, 2020.

[25] J. Cho, S. Yoon, A. Kale, F. Dernoncourt, T. Bui, and M. Bansal. Fine-grained image captioning with CLIP reward. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 517–527, 2022.

[26] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1724–1734, 2014.

[27] L. Choshen, L. Fox, Z. Aizenbud, and O. Abend. On the weaknesses of reinforcement learning for neural machine translation. In *Proceedings of the International Conference on Learning Representations*, 2020.

[28] C. Chunseong Park, B. Kim, and G. Kim. Attend to you: Personalized image captioning with context sequence memory networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 895–903, 2017.

[29] C. Clark, M. Yatskar, and L. Zettlemoyer. Don ' t take the easy way out: Ensemble based methods for avoiding known dataset biases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 4069–4082, 2019.

[30] R. Cohn-Gordon, N. Goodman, and C. Potts. Pragmatically informative image captioning with character-level inference. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 439–443, 2018.

[31] M. Cornia, L. Baraldi, and R. Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8307–8316, 2019.

[32] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie. Learning to evaluate image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5804–5812, 2018.

[33] B. Dai, S. Fidler, R. Urtasun, and D. Lin. Towards diverse and natural image

descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979, 2017.

[34] B. Dai and D. Lin. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, 2017.

[35] A. Das, S. Kottur, J. M. F. Moura, S. Lee, and D. Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979, 2017.

[36] D. Demeter, G. Kimmel, and D. Downey. Stolen probability: A structural weakness of neural language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2191–2197, 2020.

[37] M. Denkowski and A. Lavie. METEOR Universal: Language specific translation evaluation for any target language. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 376–380, 2014.

[38] A. Deshpande, J. Aneja, L. Wang, A. G. Schwing, and D. Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10695–10704, 2019.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[40] L. Dinh, D. Krueger, and Y. Bengio. NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014.

[41] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2625–2634, 2015.

[42] S. Edunov, M. Ott, M. Auli, D. Grangier, and M. Ranzato. Classical structured prediction losses for sequence to sequence learning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 355–364, 2018.

[43] C. Eickhoff, I. Schwall, A. García Seco de Herrera, and H. Müller. Overview of ImageCLEFcaption 2017 - the image caption prediction and concept extraction tasks to understand biomedical images. In *CLEF2017 Working Notes*, 2017.

[44] D. Elliott, S. Frank, K. Sima'an, and L. Specia. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the Workshop on Vision and Language*, pages 70–74, 2016.

[45] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. In *Proceedings of the British Machine Vision Conference*, 2018.

[46] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Proceedings of the European Conference on Computer Vision*, pages 15–29, 2010.

[47] C. Fellbaum. *WordNet: An Electronic Lexical Database*. 1998.

[48] Y. Feng, L. Ma, W. Liu, and J. Luo. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134, 2019.

[49] A. Fisch, K. Lee, M.-W. Chang, J. H. Clark, and R. Barzilay. CapWAP: Captioning with a purpose. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 8755–8768, 2020.

[50] M. Freitag, D. Grangier, and I. Caswell. BLEU might be guilty but references are not innocent. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 61–71, 2020.

[51] Y. Ge, J. Xu, B. N. Zhao, L. Itti, and V. Vineet. DALL-E for detection: Language-driven context image synthesis for object detection. *arXiv preprint arXiv:2206.09592*, 2022.

[52] K. Ghosal, A. Rana, and A. Smolic. Aesthetic image captioning from weakly-labelled photographs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 4550–4560, 2019.

[53] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[54] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6325–6334, 2017.

[55] J. Gu, S. Joty, J. Cai, and G. Wang. Unpaired image captioning by language pivoting. In *Proceedings of the European Conference on Computer Vision*, pages 519–535, 2018.

[56] J. Gu, S. Joty, J. Cai, H. Zhao, X. Yang, and G. Wang. Unpaired image captioning via scene graph alignments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10322–10331, 2019.

[57] S. Gu, J. Zhang, F. Meng, Y. Feng, W. Xie, J. Zhou, and D. Yu. Token-level adaptive training for neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1035–1046, 2020.

[58] D. Guo, Y. Wang, P. Song, and M. Wang. Recurrent relational memory network for unsupervised image captioning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 920–926, 2020.

[59] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617, 2018.

[60] D. Gurari, Y. Zhao, M. Zhang, and N. Bhattacharya. Captioning images taken by people who are blind. In *Proceedings of the European Conference on Computer Vision*, pages 417–434, 2020.

[61] H. He, S. Zha, and H. Wang. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 132–142, 2019.

[62] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[63] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision*, pages 630–645, 2016.

[64] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020.

[65] L. A. Hendricks, K. Burns, K. Saenko, T. Darrell, and A. Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European Conference on Computer Vision*, pages 771–787, 2018.

[66] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, K. Saenko, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2016.

[67] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, 2021.

[68] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[69] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[70] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi. The curious case of neural text degeneration. In *Proceedings of the International Conference on Learning Representations*, 2020.

[71] Y. Hu, H. Hua, Z. Yang, W. Shi, N. A. Smith, and J. Luo. PromptCap: Prompt-guided task-aware image captioning. *arXiv preprint arXiv:2211.09699*, 2022.

[72] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3296–3297, 2017.

[73] D. Ippolito, R. Kriz, J. Sedoc, M. Kustikova, and C. Callison-Burch. Comparison of diverse decoding methods from conditional language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3752–3762, 2019.

[74] H. Jhamtani and T. Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4024–4034, 2018.

[75] M. Jiang, J. Hu, Q. Huang, L. Zhang, J. Diesner, and J. Gao. REO-relevance, extraness, omission: A fine-grained evaluation for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 1475–1480, 2019.

[76] M. Jiang, Q. Huang, L. Zhang, X. Wang, P. Zhang, Z. Gan, J. Diesner, and J. Gao. TIGEr: Text-to-image grounding for image caption evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 2141–2152, 2019.

[77] S. Jiang, P. Ren, C. Monz, and M. de Rijke. Improving neural response diversity with frequency-aware cross-entropy loss. In *Proceedings of the World Wide Web Conference*, pages 2879–2885, 2019.

[78] Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, C. Tan, M. Chen, S. Huang, X. Liu, and S. Yu. Biomedical question answering: A survey of approaches and challenges. *ACM Computing Surveys*, 55(2):1–36, 2022.

[79] B. Jing, P. Xie, and E. Xing. On the automatic generation of medical imaging reports. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2577–2586.

[80] J. Johnson, A. Karpathy, and L. Fei-Fei. DenseCap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

[81] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *Proceedings of the International Conference on Learning Representations*, 2020.

[82] W. Y. Kang, J. Mun, S. Lee, and B. Roh. Noise-aware learning from web-crawled image-text data for image captioning. *arXiv preprint arXiv:2212.13563*, 2022.

[83] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[84] J. Kasai, K. Sakaguchi, L. Dunagan, J. Morrison, R. Le Bras, Y. Choi, and N. A. Smith. Transparent human evaluation for image captioning. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3464–3478, 2022.

[85] S. Kiegeland and J. Kreutzer. Revisiting the weaknesses of reinforcement learning for neural machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1673–1681, 2021.

[86] M. Kilickaya, A. Erdem, N. Ikizler-Cinbis, and E. Erdem. Re-evaluating automatic metrics for image captioning. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 199–209, 2017.

[87] D.-J. Kim, J. Choi, T.-H. Oh, and I. S. Kweon. Image captioning with very scarce supervised data: Adversarial semi-supervised learning approach. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*, pages 2012–2023, 2019.

[88] H. Kim, Z. Tang, and M. Bansal. Dense-caption matching and frame-selection gating for temporal localization in VideoQA. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 4812–4822, 2020.

[89] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

[90] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[91] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[92] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. OpenImages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017.

[93] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bernstein, and L. Fei-Fei. Visual Genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

[94] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1601–1608, 2011.

[95] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger. From word embeddings to document distances. In *Proceedings of the International Conference on Machine Learning*, pages 957–966, 2015.

[96] A. Kuznetsova, H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig, and V. Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

[97] I. Laina, C. Rupprecht, and N. Navab. Towards unsupervised image captioning with shared multimodal embeddings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7414–7424, 2019.

[98] J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):1–10, 2018.

[99] H. Lee, S. Yoon, F. Dernoncourt, T. Bui, and K. Jung. UMIC: An unreferenced metric for image captioning via contrastive learning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 220–226, 2021.

[100] H. Lee, S. Yoon, F. Dernoncourt, D. S. Kim, T. Bui, and K. Jung. ViLBERTScore: Evaluating image caption using vision-and-language BERT. In *Proceedings of the Workshop on Evaluation and Comparison of NLP Systems*, pages 34–39, 2020.

[101] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[102] J. Li, D. Li, C. Xiong, and S. Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, pages 12888–12900, 2022.

[103] J. Li, W. Monroe, A. Ritter, D. Jurafsky, M. Galley, and J. Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, 2016.

[104] X. Li, X. Sun, Y. Meng, J. Liang, F. Wu, and J. Li. Dice loss for data-imbalanced nlp tasks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 465–476, 2020.

[105] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the European Conference on Computer Vision*, pages 121–137, 2020.

[106] Y. Li, G. Li, L. He, J. Zheng, H. Li, and Z. Guan. Widget captioning: Generating natural language description for mobile user interface elements. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 5495–5510, 2020.

[107] C.-Y. Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004.

[108] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[109] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pages 740–755, 2014.

[110] F. Liu, M. Gao, T. Zhang, and Y. Zou. Exploring semantic relationships for image captioning without parallel data. In *Proceedings of the IEEE International Conference on Data Mining*, pages 439–448, 2019.

[111] L. Liu, J. Tang, X. Wan, and Z. Guo. Generating diverse and descriptive image captions using visual paraphrases. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4240–4249, 2019.

[112] X. Liu, H. Li, J. Shao, D. Chen, and X. Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European Conference on Computer Vision*, pages 338–354, 2018.

[113] D. Lu, S. Whitehead, L. Huang, H. Ji, and S.-F. Chang. Entity-aware image caption generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4013–4023, 2018.

[114] J. Lu, D. Batra, D. Parikh, and S. Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, 2019.

[115] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3242–3250, 2017.

[116] J. Lu, J. Yang, D. Batra, and D. Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7219–7228, 2018.

[117] R. Luo, B. Price, S. Cohen, and G. Shakhnarovich. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6964–6974, 2018.

[118] R. Luo and G. Shakhnarovich. Analysis of diversity-accuracy tradeoff in image captioning. *arXiv preprint arXiv:2002.11848*, 2020.

[119] P. Madhyastha, J. Wang, and L. Specia. VIFIDEL: Evaluating the visual fidelity of image descriptions. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 6539–6550, 2019.

[120] S. Mahajan, I. Gurevych, and S. Roth. Latent normalizing flows for many-to-many cross-domain mappings. In *Proceedings of the International Conference on Learning Representations*, 2020.

[121] S. Mahajan and S. Roth. Diverse image captioning with context-object split latent spaces. In *Advances in Neural Information Processing Systems*, 2020.

[122] J. Mao, X. Wei, Y. Yang, J. Wang, Z. Huang, and A. L. Yuille. Learning like a child: Fast novel visual concept learning from sentence descriptions of images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2533–2541, 2015.

[123] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). In *Proceedings of the International Conference on Learning Representations*, 2015.

[124] Y. Mao, L. Chen, Z. Jiang, D. Zhang, Z. Zhang, J. Shao, and J. Xiao. Rethinking the reference-based distinctive image captioning. In *Proceedings of the ACM International Conference on Multimedia*, pages 4374–4384, 2022.

[125] A. Mehrabian and S. R. Ferris. Inference of attitudes from nonverbal communication in two channels. *Journal of Consulting Psychology*, 31(3):248–252, 1967.

[126] A. Mehrabian and M. Wiener. Decoding of inconsistent communications. *Journal of Personality and Social Psychology*, 6(1):109–114, 1967.

[127] C. Meister, T. Pimentel, G. Wiher, and R. Cotterell. Typical decoding for natural language generation. *arXiv preprint arXiv:2202.00666*, 2022.

[128] Z. Meng, D. Yang, X. Cao, A. Shah, and S.-N. Lim. Object-centric unsupervised image captioning. In *Proceedings of the European Conference on Computer Vision*, pages 219–235, 2022.

[129] A. K. Menon, S. Jayasumana, A. S. Rawat, H. Jain, A. Veit, and S. Kumar. Long-tail learning via logit adjustment. In *Proceedings of the International Conference on Learning Representations*, 2020.

[130] P. Messina, P. Pino, D. Parra, A. Soto, C. Besa, S. Uribe, M. Andía, C. Tejos, C. Prieto, and D. Capurro. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Computing Surveys*, 54(10s):1–40, 2022.

[131] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations Workshop*, 2013.

[132] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 2013.

[133] W. Monroe, R. X. Hawkins, N. D. Goodman, and C. Potts. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338, 2017.

[134] T. Q. Nguyen and D. Chiang. Improving lexical choice in neural machine translation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 334–343, 2018.

[135] V.-Q. Nguyen, M. Suganuma, and T. Okatani. GRIT: Faster and better image captioning transformer using dual visual features. In *Proceedings of the European Conference on Computer Vision*, pages 167–184, 2022.

[136] M. Nikolaus, M. Abdou, M. Lamm, R. Aralikatte, and D. Elliott. Compositional generalization in image captioning. In *Proceedings of the Conference on Computational Natural Language Learning*, pages 87–98, 2019.

[137] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[138] D. H. Park, T. Darrell, and A. Rohrbach. Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4624–4633, 2019.

[139] G. Parmar, K. K. Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023.

[140] R.-G. Pasca, A. Gavryushin, Y.-L. Kuo, O. Hilliges, and X. Wang. Summarize the past to predict the future: Natural language descriptions of context boost multimodal object interaction. *arXiv preprint arXiv:2301.09209*, 2023.

[141] R. Pasunuru and M. Bansal. Multi-reward reinforced summarization with saliency and entailment. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 646–653, 2018.

[142] J. Pavlopoulos, V. Kougia, and I. Androutsopoulos. A survey on biomedical image captioning. In *Proceedings of the Workshop on Shortcomings in Vision and Language*, pages 26–36, 2019.

[143] J. Pennington, R. Socher, and C. Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[144] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2227–2237, 2018.

[145] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021.

[146] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations*, 2016.

[147] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147, 2010.

[148] V. Raunak, S. Dalmia, V. Gupta, and F. Metze. On long-tailed phenomena in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3088–3095, 2020.

[149] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *Proceedings of the International Conference on Machine Learning*, pages 1060–1069, 2016.

[150] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015.

[151] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017.

[152] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the International Conference on Machine Learning*, pages 1278–1286, 2014.

[153] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[154] A. Sadovnik, Y.-I. Chiu, N. Snavely, S. Edelman, and T. Chen. Image description with a goal: Building efficient discriminating expressions for images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2791–2798, 2012.

[155] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022.

[156] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki. LAION-400M: Open dataset of

CLIP-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[157] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual Captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 2556–2565, 2018.

[158] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer. How much can CLIP benefit vision-and-language tasks? In *Proceedings of the International Conference on Learning Representations*, 2022.

[159] R. Shetty, M. Rohrbach, L. Anne Hendricks, M. Fritz, and B. Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144, 2017.

[160] J. Shi, Y. Li, and S. Wang. Partial off-policy learning: Balance accuracy and diversity for human-oriented image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2187–2196, 2021.

[161] Z. Shi, H. Liu, and X. Zhu. Enhancing descriptive image captioning with natural language inference. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, pages 269–277, 2021.

[162] K. Shuster, S. Humeau, H. Hu, A. Bordes, and J. Weston. Engaging image captioning via personality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12516–12526, 2019.

[163] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh. TextCaps: A dataset for image captioning with reading comprehension. In *Proceedings of the European Conference on Computer Vision*, pages 742–758, 2020.

[164] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, 2015.

[165] Y. Song, S. Chen, Y. Zhao, and Q. Jin. Unpaired cross-lingual image caption generation with self-supervised rewards. In *Proceedings of the ACM International Conference on Multimedia*, pages 784–792, 2019.

[166] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara. From show to tell: A survey on image captioning. *arXiv preprint arXiv:2107.06912*, 2021.

[167] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, 2014.

[168] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4278–4284, 2017.

[169] K. Tang, J. Huang, and H. Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. In *Advances in Neural Information Processing Systems*, pages 1513–1524, 2020.

[170] K. Tran, X. He, L. Zhang, J. Sun, C. Carapcea, C. Thrasher, C. Buehler, and C. Sienkiewicz. Rich image captioning in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 434–441, 2016.

[171] B. D. Van Durme. *Extracting implicit knowledge from text*. 2009.

[172] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017.

[173] R. Vedantam, S. Bengio, K. Murphy, D. Parikh, and G. Chechik. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260, 2017.

[174] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

[175] S. Venugopalan, L. Anne Hendricks, M. Rohrbach, R. Mooney, T. Darrell, and K. Saenko. Captioning images with diverse objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1170–1178, 2017.

[176] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence – video to text. In *Proceedings of the IEEE International Conference on Computer Vision*, 2015.

[177] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1494–1504, 2015.

[178] G. Vered, G. Oren, Y. Atzmon, and G. Chechik. Joint optimization for cooperative image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8898–8907, 2019.

[179] G. Verma, V. Vinay, S. Bansal, S. Oberoi, M. Sharma, and P. Gupta. Using image captions and multitask learning for recommending query reformulations. In *Advances in Information Retrieval: European Conference on IR Research*, pages 681–696, 2020.

[180] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. In *Proceedings of the International Conference on Learning Representations*, 2017.

[181] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[182] J. Wang, W. Xu, Q. Wang, and A. B. Chan. Compare and reweight: Distinctive image captioning using similar images sets. In *Proceedings of the European Conference on Computer Vision*, pages 370–386, 2020.

[183] J. Wang, W. Xu, Q. Wang, and A. B. Chan. Group-based distinctive image captioning with memory attention. In *Proceedings of the ACM International Conference on Multimedia*, pages 5020–5028, 2021.

[184] L. Wang, A. G. Schwing, and S. Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems*, 2017.

[185] N. Wang, J. Xie, H. Luo, Q. Cheng, J. Wu, M. Jia, and L. Li. Efficient image captioning for edge devices. *arXiv preprint arXiv:2212.08985*, 2022.

[186] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the International Conference on Machine Learning*, pages 23318–23340, 2022.

[187] Q. Wang and A. B. Chan. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4195–4203, 2019.

[188] S. Wang, Z. Yao, R. Wang, Z. Wu, and X. Chen. FAIEr: Fidelity and adequacy ensured image caption evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14050–14059, 2021.

[189] X. Wang, L. Lian, Z. Miao, Z. Liu, and S. Yu. Long-tailed recognition by routing diverse distribution-aware experts. In *Proceedings of the International Conference on Learning Representations*, 2020.

[190] Z. Wang, B. Feng, K. Narasimhan, and O. Russakovsky. Towards unique and informative captioning of images. In *Proceedings of the European Conference on Computer Vision*, pages 629–644, 2020.

[191] Z. Wang, F. Wu, W. Lu, J. Xiao, X. Li, Z. Zhang, and Y. Zhuang. Diverse image captioning via grouptalk. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2957–2964, 2016.

[192] J. White, G. Poesia, R. Hawkins, D. Sadigh, and N. Goodman. Open-domain clarification question generation without question examples. In *Proceedings of*

the Conference on Empirical Methods in Natural Language Processing, pages 563–570, 2021.

[193] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[194] J. Wu, T. Chen, H. Wu, Z. Yang, G. Luo, and L. Lin. Fine-grained image captioning with global-local discriminative objective. *IEEE Transactions on Multimedia*, 23:2413–2427, 2021.

[195] J. Wu, Z. Hu, and R. Mooney. Generating question relevant captions to aid visual question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 3585–3594, 2019.

[196] L. Wu, F. Tian, T. Qin, J. Lai, and T.-Y. Liu. A study of reinforcement learning for neural machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 3612–3621, 2018.

[197] Q. Wu, L. Li, H. Zhou, Y. Zeng, and Z. Yu. Importance-aware learning for neural headline editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9282–9289, 2020.

[198] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*, pages 2048–2057, 2015.

[199] Y. Xu, L. Chen, Z. Cheng, L. Duan, and J. Luo. Open-ended visual question answering by multi-modal domain adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 367–376, 2020.

[200] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4507–4515, 2015.

[201] Y. Yi, H. Deng, and J. Hu. Improving image captioning evaluation by considering inter references variance. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 985–994, 2020.

[202] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

[203] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[204] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. BERTScore: Evaluating text generation with bert. In *Proceedings of the International Conference on Learning Representations*, 2020.

[205] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021.

[206] Z. Zhang, Y. Gu, and B. A. Plummer. Show and write: Entity-aware news generation with image information. *arXiv preprint arXiv:2112.05917*, 2021.

[207] D. Zhao, A. Wang, and O. Russakovsky. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14830–14840, 2021.

[208] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017.

Appendix A

Unsupervised Image Captioning with Careful Word-Level Alignment to Broaden the Scene Coverage – Supplementary Material

A.1 Evaluation using Soft-Matching Metrics

We have observed that our unsupervised image captioning methods enhance the performance in exact-matching metrics by increasing high-frequency words and decreasing low-frequency words (Section 3.3.8). On the other hand, however, we also have observed that exact-matching metrics wrongly penalize correct but low-frequency words (Section 4.4.3). Therefore, it is possible that the exact-matching metrics unfairly favored our proposed models.

To provide more thorough experimental results, we evaluated unsupervised image captioning models using soft-matching metrics. Table A.1 shows the results. The results are consistent with the exact-matching evaluation we reported in Table 3.4. Our model outperformed the previous model in both exact- and soft-matching scores; the combined model achieved the highest scores across all the metrics. These results further confirm the superiority of our methods.

Unlike supervised image captioning, there is no score gap between exact- and soft-matching metrics in unsupervised image captioning. This is because the scores are drastically lower than those of supervised image captioning models. The score gap

	Text-Based			*Text-and-Image-Based*		
	Exact-Matching		*Soft-Matching*			
	CIDEr	SPICE	BERTS++	TIGEr	CLIPS	RefCLIPS
[48]	28.6	8.1	30.9	63.2	59.2	64.9
Ours	32.4	8.4	32.0	63.5	61.8	65.9
Ours + [48]	**35.7**	**9.2**	**32.4**	**63.7**	**62.9**	**68.0**

Table A.1: Evaluation across exact- and soft-matching metrics. We show the results of single run. The highest scores are marked in bold.

arises when there are correct-but-OOR words (Section 4.4.3), but OOR words are likely to be incorrect in the low score range of unsupervised image captioning models.

Appendix B

Discriminative Image Captioning by Relieving a Bottleneck of RL to Broaden the Description Coverage – Supplementary Material

B.1 Further Output Examples

Figure B.1 shows caption examples in the MS COCO validation set. The blue words are those that have never appeared in the output captions of the baseline model. We observe that these blue words express various types of characteristic information of the images. Here, *weather vane* and *flamingos* are characteristic objects of the images (a) and (b); *shallow*, *funny*, and *staring straight ahead* are characteristic attributes of the images (b) and (c); and *racing* and *sniffing* are characteristic relations in the images (d) and (e). These examples further support our hypothesis that the limited vocabulary of RL models hinders discriminativeness.

(a)

Transformer RL: a tower with a clock on top of it
+wFT: a clock tower with a **weather vane** on top
NLI: a tower with a clock on the top of it
Human: a weather vane atop a cathedral clock tower

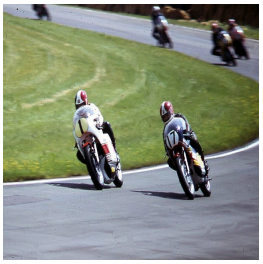
(b)

Transformer RL: a group of birds standing in the water
+wFT: a large group of **flamingos** stand in **shallow** water
NLI: a group of pink umbrellas are standing in the water
Human: a flock of pink flamingos standing in shallow water

(c)

Transformer RL: a black cat wearing a hat on top of a table
+wFT: a cat wears a **funny** hat while **staring straight ahead**
NLI: a black cat wearing a hat sitting on a table
Human: the cute black cat is wearing a bee's hat

(d)



Transformer RL: a group of people riding motorcycles on a road
+wFT: a group of people **racing** motorcycles on a race track
NLI: a group of people riding motorcycles on a race track
Human: people are racing motorcycles on a race track

(e)

Transformer RL: a dog next to a cup of coffee
+wFT: a dog is **sniffing** a cup of coffee
NLI: a dog standing next to a coffee cup on a table
Human: a squinting dog on a brick patio sniffs a cup of coffee

Figure B.1: Caption examples in the MS COCO validation set. The blue words are those that have never appeared in the output captions of the baseline model (Transformer RL). *Human* shows a ground-truth caption of each image.

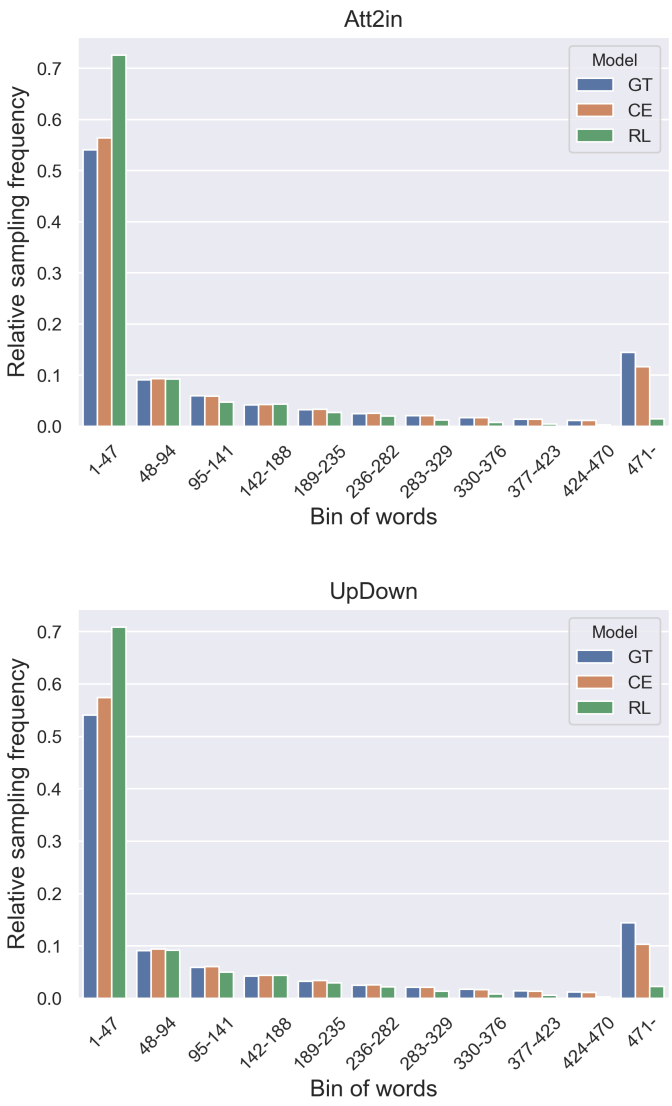


Figure B.2: Relative frequency of the words in the sequences sampled for the training images. Five sequences were sampled for each image. The words (9,486 unique words excluding an out-of-vocabulary token $\langle \mathrm{unk} \rangle$) are sorted by their frequency in ground-truth captions and divided into 200 bins. We show the first 10 bins and the sum of the rest. GT is the ground-truth caption of the training images, CE is the output of a captioning model trained with the CE loss, and RL is the output of a captioning model trained with RL.

B.2 Peaky Distributions in Other Models

Figure B.2 shows the results of the relative frequency of the words sampled for the training images by the LSTM-based models: Att2in [151] and UpDown [5]. Similar to the Transformer model, the sequences sampled with the LSTM-based RL models are clearly limited to high-frequency words, forming the peaky distributions.

B.3 Libraries for Evaluation

We used the following libraries for evaluation with all the hyperparameters set to the default values.

- **CIDEr, SPICE, CLIPS, and RefCLIPS**
  `https://github.com/jmhessel/pycocoevalcap`
- **BERTS++**
  `https://github.com/ck0123/improved-bertscore-for-image-captioning-evaluation`
- **TIGEr**
  `https://github.com/SeleenaJM/CapEval`
- **R@K**
  `https://github.com/fartashf/vsepp`
  Following [111], we used a publicly available model, `coco_vse++_resnet_restval_finetune`.

B.4 Best Hyperparameters

We searched for the best hyperparameters for the learning rate from {1e-3, 1e-4, 1e-5, 1e-6}, and the inverse-temperature hyperparameter β' of Eq. (4.5) from {0.1, 1}. The best learning rate was 1e-5 for Transformer models and 1e-4 for the other models (Att2in and UpDown). The best β' was 0.1 for wFT with $p_{\boldsymbol{\theta}}$ decoding and 1 for wFT with BP decoding. Note that sFT does not use β'.

The best learning rate was the same in CE-based models (Joint CE and Only CE): 1e-5 for Transformer and 1e-4 for the others. The best $\lambda \in \{0.2, 0.5, 0.8\}$ for Joint CE was 0.8 for Transformer and 0.2 for the others.

B.5 The Number of Parameters

The exact number of parameters was 14,451,985 for Att2in, 52,125,025 for UpDown, and 57,474,832 for Transformer. Note that the parameters $\boldsymbol{\theta}'$ are not included because they are not trainable and fixed through the entire training and evaluation; rather, the actual trainable parameters are decreased to the classifier parameters in our models. Visual Paraphrase has double decoders of Att2in; thus, it increases the number of trainable parameters and requires training of the specialized model from scratch.

	Epoch	Batch	Hour/Epoch	Total Hour
Att2in RL	20	10	0.68	13.54
+ sFT	1	10	**0.08**	**0.08**
+ wFT	1	10	0.12	0.12
CIDErBtw	50	10	0.70	35.11
NLI	50	16	0.87	43.55
Joint CE	20	10	1.15	22.97
UpDown RL	20	10	0.71	14.16
+ sFT	1	10	**0.09**	**0.09**
+ wFT	1	10	0.14	0.14
CIDErBtw	50	10	0.76	38.09
NLI	50	16	0.87	43.74
Joint CE	20	10	1.08	21.67
Transformer RL	25	10	3.23	80.66
+ sFT	1	10	**0.11**	**0.11**
+ wFT	1	10	0.18	0.18
CIDErBtw	25	10	3.27	81.76
NLI	25	16	2.74	68.54
Joint CE	25	10	4.06	101.43

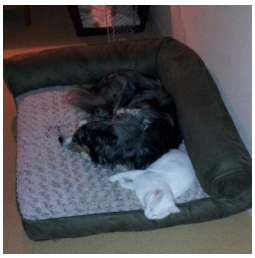
Table B.1: Time to train discriminativeness-aware captioning models. Note that we excluded the time for initialization before RL because there is not much difference among the methods. Results for the baseline RL models are shown in gray text because we did not train these models but used publicly-available pre-trained models.

B.6 Comparison of Computational Cost

Table B.1 shows the time to train discriminativeness-aware captioning models. We used a single GPU of 16 GB memory for all training. Clearly, our methods require far less time for training. This is because our methods do not require retraining from scratch but only require a single-epoch fine-tuning to publicly-available pre-trained RL models.

B.7 Qualitative Analysis of Underrated Captions

Figure B.3 shows caption examples, automatic evaluation scores, and reference captions. Clearly, our wFT model correctly described all five images with diverse vocabulary. However, the CIDEr scores for our captions were considerably lower than those for the baseline model captions. The cause of this underrating is the small coverage of the reference captions: the reference captions rarely include the low-frequency words colored in blue due to their low frequency. Conventional exact-matching metrics such as CIDEr cannot evaluate those correct-but-OOR words by the definition of exact-matching. In contrast, RefCLIPS, the current best-performing metric, can consider the information not covered by reference captions by incorporating image features and soft-matching. Figure B.3 shows that RefCLIPS evaluated the correct-but-OOR words more correctly and gave more plausible scores to our captions. These examples further support our conclusion that the lower exact-matching scores of our models are caused by the nature of low-frequency words and the deficiency of exact-matching metrics, not by the degeneration of our models.



Transformer RL:
a dog laying on top of a couch
[CIDEr: 133.3, RefCLIPS: 77.7]
+wFT:
a dog **curled** up **asleep** on a **cushion**
[CIDEr: 38.7, RefCLIPS: 79.2]
Reference Coverage: 0/5
N/A

Transformer RL:
a person flying a kite in the ocean
[CIDEr: 60.2, RefCLIPS: 79.8]
+wFT:
a man **kiteboarding** on top of a body of water
[CIDEr: 3.5, RefCLIPS: 79.2]
Reference Coverage: 0/5
N/A

Transformer RL:
a vase filled with yellow flowers on a table
[CIDEr: 216.7, RefCLIPS: 78.5]
+wFT:
a **clear** vase filled with **multi colored** flowers
[CIDEr: 94.0, RefCLIPS: 82.0]
Reference Coverage: 1/5
an arrangement of flowers in a **clear** glass canning jar haging on a wall

Transformer RL:
a yellow and blue airplane sitting on a runway
[CIDEr: 104.1, RefCLIPS: 78.2]
+wFT:
a yellow and blue jet **airliner** on a runway
[CIDEr: 58.8, RefCLIPS: 78.3]
Reference Coverage: 1/5
a brown lot **airliner** sitting on the tarmac

Transformer RL:
a herd of cows grazing in a field of grass
[CIDEr: 172.2, RefCLIPS: 81.6]
+wFT:
a herd of cattle grazing on a **dry** grass field
[CIDEr: 74.0, RefCLIPS: 84.4]
Reference Coverage: 2/5
a bunch of cows grazing in a **dry** field together
cows wandering in a **dry** grass filled meadow

Figure B.3: Underrated captions in the MS COCO validation set. The blue words are those that have never appeared in the output captions of the baseline model (Transformer RL). *Reference Coverage* shows the number of reference captions (out of five) that cover at least one of the blue words.

B.8 Details of Human Evaluation

We show our AMT interface in Figure B.4. Each image was evaluated with the five questions in the discrete 5-point scale. We required workers to satisfy the following qualifications: being an AMT Master and living in the U.S. Workers were notified that this experiment was intended to evaluate caption quality. We paid \$0.1 for each image, and the median of the actual working time was 41 seconds per image. The hourly reward was estimated as \$8.78, which is higher than the minimum wage in the U.S., \$7.25 per hour.

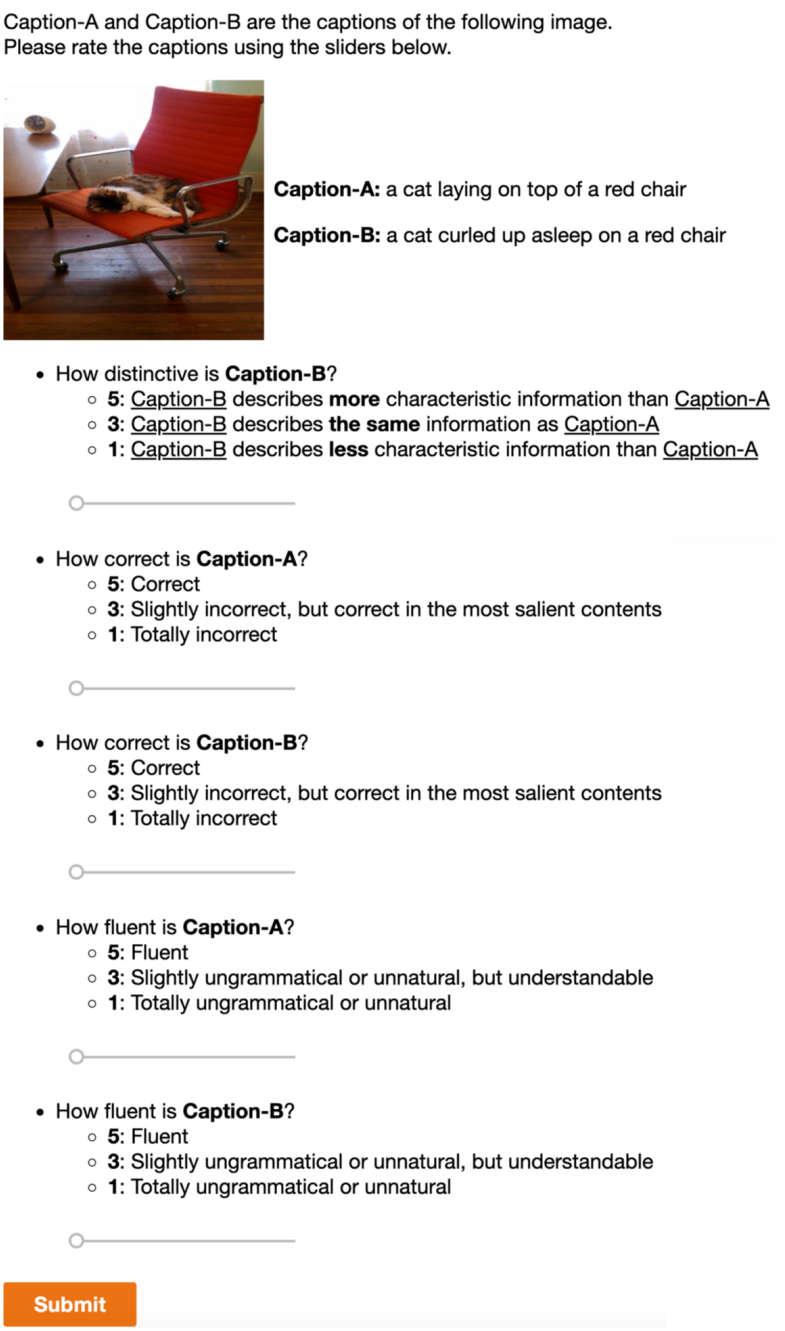


Figure B.4: A screenshot of our AMT interface.

B.9 Comparison with Other Long-Tail Classification Methods

We adapted the long-tail classification method of [81] to relieve the bottleneck of RL and proposed sFT and wFT. Both methods were carefully designed for RL models, but these were not the only way to employ long-tail classification methods. In this section, we discuss the other possible adaptations based on [148].

[148] explored ways to employ long-tail classification methods for machine translation. Their first method was τ-normalization (τ-norm), which directly adopted the method of [81]. Based on an observation that the norm of classifier parameters correlates with the frequency of the classes, they normalized the classifier weight $\boldsymbol{W}$ as follows:

$$\widetilde{\boldsymbol{W}}_{w_i} = \frac{\boldsymbol{W}_{w_i}}{\|\boldsymbol{W}_{w_i}\|^{\tau}}, \tag{B.1}$$

where $\boldsymbol{W}_{w_i} \in \mathbb{R}^d$ indicates a vector at the index of a word w_i and τ is a temperature hyperparameter that controls the degree of the normalization.

The other methods of [148] were Focal loss (FL) and Anti-Focal loss (AFL). AFL is a variant of FL [108], which was aimed at reweighting the loss according to the confidence of the model predictions. Let $p^t_{\boldsymbol{\theta}} = p_{\boldsymbol{\theta}}(y_t \mid \boldsymbol{y}_{<t}, \boldsymbol{I})$. FL and AFL for each training data are then written as follows:

$$\mathcal{L}_{\mathrm{FL}}(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T}(1 - p^t_{\boldsymbol{\theta}})^{\gamma}\log p^t_{\boldsymbol{\theta}}, \tag{B.2}$$

$$\mathcal{L}_{\mathrm{AFL}}(\boldsymbol{\theta}) = -\frac{1}{T}\sum_{t=1}^{T}(1 + \alpha p^t_{\boldsymbol{\theta}})^{\gamma}\log p^t_{\boldsymbol{\theta}}, \tag{B.3}$$

where γ and α are hyperparamters that control the degree of the reweighting. Other work also explored ways to employ long-tail classification methods for text generation, but those approaches are categorized as either τ-norm [134] or variants of FL [57, 77, 197], which we already explored above.

We compared our methods (sFT and wFT) with τ-norm, FL, and AFL. In our experiments, we normalized the bias term $\boldsymbol{b}$[1] in addition to the weight term $\boldsymbol{W}$ as we found

[1] $\tilde{\boldsymbol{b}} = \frac{\boldsymbol{b}}{\|\boldsymbol{b}\|^{\tau}}$, where the value of the hyperparameter τ was set to the same as that of $\widetilde{\boldsymbol{W}}$.

it performed better than normalizing the weight term only. We applied FL and AFL as the alternative weighting to BP for a fair comparison with our methods. That is, we fine-tuned the classifier parameters by optimizing $\mathcal{L}_{\mathrm{FL}}(\hat{\boldsymbol{\theta}})$ or $\mathcal{L}_{\mathrm{AFL}}(\hat{\boldsymbol{\theta}})$, where $\hat{\boldsymbol{\theta}}$ were initialized with the pre-trained RL models. We used the best hyperparameters reported in [148]: $\tau = 0.2$, $\gamma = 1$, and $\alpha = 1$. Similar to our models, other hyperparameters were set to the same values as the baseline models, except for the epoch size and learning rate. We explored the same values for these hyperparameters as our models: we set the epoch size for fine-tuning to 1 and searched for the best learning rates from {1e-3, 1e-4, 1e-5, 1e-6}. We selected the best learning rate according to the R@1 scores in the validation set. The best learning rate was 1e-4 for Att2in RL + FL/AFL, 1e-4 for UpDown RL + FL/AFL, and 1e-5 for Transformer RL + FL/AFL. Note that we did not explore the learning rate for τ-norm because it does not require training.

In open-ended text generation tasks, *e.g.*, story generation and text generation after prompts, stochastic sampling methods are used instead of beam search to increase the diversity in output text [70, 11, 127]. Although image captioning does not fall in the category of open-ended text generation as input images tightly scope the correctness of captions, we additionally test whether the randomness in stochastic sampling can increase the output vocabulary. We used Nucleus sampling [70] with a hyperparameter $p = 0.95$, which is the best hyperparameter reported [70, 127].

	Vocabulary			*Standard Evaluation*						*Discriminativeness*		
	Unique-1	Unique-S	Length	CIDEr	SPICE	BERTS++	TIGEr	CLIPS	RefCLIPS	R@1	R@5	R@10
Att2in RL	445	2,524	9.3	**117.4**	**20.5**	43.6	73.9	73.0	79.7	16.3	41.9	57.2
+ sFT	880	3,156	9.0	115.4	20.4	**43.9**	74.3	73.7	**80.3**	20.1	48.0	62.8
+ wFT	**1,197**	**3,732**	8.9	104.3	19.5	43.1	74.2	73.9	80.2	20.6	49.7	64.5
+ wFT (BP decoding)	1,102	3,615	9.4	109.3	20.1	43.7	**74.4**	**74.0**	80.2	**21.1**	**50.5**	**64.8**
+ τ-norm	437	2,414	9.1	117.3	20.4	43.5	73.8	72.9	79.7	15.4	40.7	55.8
+ FL	903	3,217	9.0	114.8	20.4	43.8	74.3	73.7	**80.3**	20.1	48.1	63.2
+ AFL	886	3,116	9.0	115.3	20.4	43.8	74.3	73.7	**80.3**	19.7	47.6	62.7
+ Nucleus sampling	475	2,726	9.3	116.5	20.3	43.5	73.9	72.9	79.7	16.5	41.9	57.1
UpDown RL	577	3,103	9.5	**122.7**	**21.5**	**44.2**	74.6	74.0	80.5	21.1	49.9	64.6
+ sFT	1,190	3,788	9.2	115.9	21.0	**44.2**	**74.9**	74.8	**80.9**	25.0	56.8	71.2
+ wFT	**1,479**	**4,268**	9.1	101.8	19.5	43.1	74.6	74.9	80.7	26.0	57.6	72.2
+ wFT (BP decoding)	1,275	4,177	9.6	110.0	20.6	44.1	**74.9**	**75.0**	80.8	**26.7**	**58.7**	**72.4**
+ τ-norm	576	2,967	9.3	122.6	21.3	**44.2**	74.4	73.8	80.5	19.6	48.1	63.4
+ FL	1,201	3,830	9.2	114.9	20.9	44.1	**74.9**	74.7	**80.9**	25.2	57.0	70.9
+ AFL	1,171	3,760	9.2	116.4	20.9	**44.2**	**74.9**	74.7	**80.9**	24.9	56.6	70.7
+ Nucleus sampling	592	3,339	9.5	120.7	21.3	**44.2**	74.6	73.9	80.4	20.9	49.7	64.4
Transformer RL	753	3,433	9.2	**127.7**	**22.5**	**45.1**	75.0	75.0	81.3	26.6	56.2	70.5
+ sFT	1,458	3,959	9.1	118.7	21.7	44.8	**75.2**	75.6	81.5	30.6	62.3	75.7
+ wFT	1,776	4,274	9.1	103.1	20.0	43.3	74.8	75.8	81.2	32.5	64.5	77.1
+ wFT (BP decoding)	**1,964**	**4,373**	9.4	107.3	21.1	44.2	**75.2**	**76.1**	81.5	**33.5**	**65.9**	**78.2**
+ τ-norm	1,027	3,483	9.2	124.4	22.1	44.9	74.8	74.9	81.2	26.1	55.8	69.7
+ FL	1,523	4,018	9.1	116.5	21.4	44.6	**75.2**	75.7	81.5	31.2	63.1	76.3
+ AFL	1,402	3,908	9.1	120.5	21.9	44.8	**75.2**	75.6	**81.6**	30.0	62.1	75.9
+ Nucleus sampling	1,053	3,751	9.3	123.7	22.0	44.8	74.9	75.0	81.2	26.9	55.8	70.4

Table B.2: Comparison with the other long-tail classification methods. Automatic evaluation results on the MS COCO test set. *Unique-1* and *Unique-S* indicate the number of unique unigrams and sentences, respectively. *Length* is the average length of output captions.

Table B.2 shows the results. τ-norm and Nucleus sampling showed the similar results. Both methods slightly increased the output vocabulary but the performance generally remained the same as the baseline models. These results indicate that the output vocabulary cannot be significantly increased while maintaining the relative probability of words: Nucleus sampling samples according to the original output distributions and τ-norm changes the distribution only by the difference in the norm, basically flattening the distribution. In contrast, FL and AFL drastically change the relative probability of words by refining the mapping from encoded features to low-frequency words, as with sFT and wFT. They successfully increased the vocabulary size and discriminativeness. However, the gains were smaller than those of wFT.

To analyze the cause of the difference between FL, AFL, and the BP loss (wFT), we

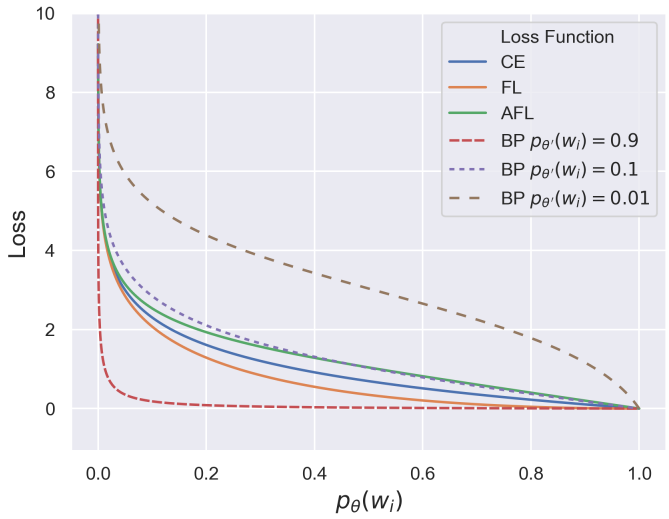


Figure B.5: Visualization of the losses: CE $-\log p_{\boldsymbol{\theta}}(w_i)$, BP $-\log p_{\boldsymbol{\theta},\boldsymbol{\theta}'}(w_i)$, FL $(1 - p_{\boldsymbol{\theta}}(w_i))^\gamma \log p_{\boldsymbol{\theta}}(w_i)$, and AFL $(1 + \alpha p_{\boldsymbol{\theta}}(w_i))^\gamma \log p_{\boldsymbol{\theta}}(w_i)$. We set $\beta = \beta' = 1$, $\gamma = 1$, and $\alpha = 1$.

visualized the losses in Figure B.5. FL suppresses the loss when a model is confident, whereas AFL increases the loss when a model is moderately confident. Compared with these losses, BP changes the loss more drastically. When the frequency-biased policy $p_{\boldsymbol{\theta}'}$ is highly confident, BP strictly suppresses the loss to prevent further learning on that word; when $p_{\boldsymbol{\theta}'}$ is not confident, BP highly increases the loss to encourage the learning on that word. This drastic rebalancing of the loss resulted in wFT's larger vocabulary size and higher discriminativeness.

B.10 Effectiveness on More Recent Models

To further demonstrate the effectiveness of our methods, we tested our fine-tuning methods on a more recent captioning model, **VinVL** [203, 105]. VinVL boosts its performance through large-scale cross-modal pre-training. The significant performance improvements have made VinVL a popular captioning model and one of the most advanced captioning models available today [166, 186, 135].

	Vocabulary			*Standard Evaluation*						*Discriminativeness*		
	Unique-1	Unique-S	Length	CIDEr	SPICE	BERTS++	TIGEr	CLIPS	RefCLIPS	R@1	R@5	R@10
VinVL RL	1,126	4,298	10.0	**140.9**	**25.2**	**46.1**	**75.7**	77.6	**83.3**	36.1	68.5	80.2
+ sFT	1,834	4,649	10.0	126.0	23.8	45.5	75.6	**78.2**	**83.3**	39.2	**72.1**	83.8
+ wFT	**1,852**	4,652	10.0	124.9	23.7	45.5	75.6	**78.2**	**83.3**	39.2	72.0	83.9
+ wFT (BP decoding)	1,734	**4,717**	9.8	122.4	23.5	45.2	**75.7**	**78.2**	**83.3**	**39.6**	**72.1**	**84.6**

Table B.3: Test on the more recent captioning model. Automatic evaluation results on the MS COCO test set. *Unique-1* and *Unique-S* indicate the number of unique unigrams and sentences, respectively. *Length* is the average length of output captions.

We used the best-performing pre-trained model as our baseline[2]. The model is publicly available[3]. Note that this model was trained using the standard RL [151].

As in the previous experiments, we applied our fine-tuning methods for one epoch only. We searched for the best learning rates for fine-tuning from {1e-5, 1e-6}, and the inverse-temperature hyperparameter β' of Eq. (4.5) from {0.01, 0.1, 1}. Other hyperparameters were set to the same as the baseline model. The best learning rate was 1e-5. The best β' was 0.01 for wFT with $p_{\boldsymbol{\theta}}$ decoding and 1 for wFT with BP decoding. Note that sFT does not use β'.

Table B.3 shows similar results as Table 4.1. Our methods significantly increased the vocabulary size from the baseline and accordingly enhanced the discriminativeness. The standard evaluation metrics also showed the same tendency. Although our models scored lower than the baseline in the conventional exact-matching metrics (CIDEr and SPICE), the gap became smaller in the more advanced soft-matching metrics (BERTS++ and TIGEr). In the state-of-the-art soft-matching metrics (CLIPS and RefCLIPS), our models achieved the same or even higher scores than the baseline. These results show that our methods are also effective on the more recent model. Moreover, these results further validate that our methods can switch any off-the-shelf RL models to discriminativeness-aware models while maintaining the overall quality of captions.

[2] `coco_captioning_large_scst`

[3] https://github.com/microsoft/Oscar/blob/master/VinVL_MODEL_ZOO.md#Image-Captioning-on-COCO

B.11 Comparison and Combination with More Recent Discriminativeness-Aware Models

Contemporaneous to our work, [25] showed that maximizing reference-free CLIPS-based reward enhanced discriminativeness significantly. In this section, we clarify the advantages of our methods over the CLIPS-based RL by comparing and combining our methods with it.

The pre-trained models of [25] are publicly available[4]. We used the transformer model trained with the standard CIDEr reward, **Transformer* RL (CIDEr)**[5], and the one trained with the reward proposed by [25], **Transformer* RL (CLIPS + Grammar)**[6]. The proposed reward is computed by the weighted sum of CLIPS and grammaticality scores. We also included **Transformer* Only CE**[7] in the comparison as the baseline without RL[8].

As in the previous experiments, we applied our fine-tuning methods for one epoch only. We searched for the best learning rates for fine-tuning from {1e-5, 1e-6, 1e-7}, and the inverse-temperature hyperparameter β' of Eq. (4.5) from {0.01, 0.1, 1}. Other hyperparameters were set to the same as the baseline model. The best learning rate for Transformer* RL (CIDEr) was 1e-5; the best β' was 0.1 for wFT with $p_{\boldsymbol{\theta}}$ decoding and 1 for wFT with BP decoding. The best learning rates for Transformer* RL (CLIPS + Grammar) were 1e-6 for wFT with BP decoding and 1e-7 for the others; the best β' was 1 for wFT with both decoding methods. Note that sFT does not use β'.

Table B.4 shows the results. Similar to the previous results, our methods significantly enhanced the vocabulary size and discriminativeness from the RL models while maintaining or even increasing the scores in the state-of-the-art soft-matching metrics. The CLIPS + Grammar reward also achieved the high discriminativeness compared with the standard CIDEr reward.

However, the improvement of the CLIPS-based RL came at the expense of the *conciseness* and overall quality of captions in contrast to our methods: compared to Trans-

[4] https://github.com/j-min/CLIP-Caption-Reward
[5] clipRN50_cider
[6] clipRN50_clips_grammar
[7] clipRN50_mle
[8] Note that clipRN50 does not mean that the model used the CLIPS-based reward. It denotes that the model used CLIP [145] as the image encoder, unlike the other models tested in this paper.

	Vocabulary			*Standard Evaluation*						*Discriminativeness*		
	Unique-1	Unique-S	Length	CIDEr	SPICE	BERTS++	TIGEr	CLIPS	RefCLIPS	R@1	R@5	R@10
Transformer* RL (CIDEr)	691	3,650	9.5	**126.0**	**22.8**	**45.2**	74.6	75.8	81.6	27.1	57.2	70.6
+ sFT	1,265	4,071	9.1	122.9	22.2	**45.2**	**74.8**	76.4	**82.0**	31.4	62.0	75.0
+ wFT	**1,546**	4,337	9.0	111.3	21.0	44.2	74.5	76.5	81.8	31.6	63.3	75.7
+ wFT (BP decoding)	1,543	**4,471**	9.5	112.3	21.7	44.9	**74.8**	**76.9**	81.9	**34.0**	**65.4**	**78.4**
Transformer* RL (CLIPS + Grammar)	952	4,847	13.0	74.1	19.8	43.6	**75.0**	**79.2**	81.2	44.2	77.0	86.9
+ sFT	969	4,848	12.8	76.4	20.1	43.8	**75.0**	**79.2**	81.2	44.6	**77.3**	87.0
+ wFT	969	4,847	12.9	76.4	20.1	43.8	**75.0**	**79.2**	81.2	44.8	77.2	**87.1**
+ wFT (BP decoding)	**1,001**	**4,853**	12.2	**82.5**	**20.6**	**44.1**	**75.0**	**79.2**	**81.3**	**45.5**	77.2	**87.1**
Transformer* Only CE	1,174	3,637	9.4	113.8	20.9	44.1	74.0	75.1	81.1	26.2	55.2	68.6

Table B.4: Test on the more recent discriminativeness-aware model. Transformer* used a different image encoder than the other transformer models tested in this paper. Automatic evaluation results on the MS COCO test set. *Unique-1* and *Unique-S* indicate the number of unique unigrams and sentences, respectively. *Length* is the average length of output captions.

former* RL (CIDEr), Transformer* RL (CLIPS + Grammar) significantly increased the sentence length and decreased scores in the standard evaluation metrics, including the current best-performing metric, RefCLIPS. Although increasing the sentence length is one way to describe images in detail, concise description is more desirable to convey the most characteristic information clearly and efficiently [154]. Despite the longer sentence length, the side effect was still observed: CLIPS-based RL decreased the output vocabulary from the Only CE baseline.

These results indicate that our methods and the CLIPS-based RL increased discriminativeness by different factors: more specific vocabulary and longer descriptions, respectively. In other words, the contribution of our methods is orthogonal to that of the CLIPS-based RL. To utilize the strength of each, we applied our methods to the CLIPS-based RL model. Although the CLIPS-based RL achieved the high discriminativeness and relatively large vocabulary size due to the longer sentences, our methods further enhanced the discriminativeness and vocabulary size. Surprisingly, our methods also improved the standard evaluation scores, including exact-matching scores. This result suggests that our fine-tuning with ground-truth captions restored the overall quality of captions, which was degraded by over-optimization for reference-free CLIPS.

Another critical advantage of our methods is computational efficiency. Training of CLIPS-based RL took *one day using eight GPUs* [25], while ours only took *40 minutes using a single GPU*.

The above results conclude that our methods are orthogonal to the more recent discriminative image captioning method and have important advantages in conciseness and efficiency.

List of Publications

Journals

1. 本多右京, 橋本敦史, 渡辺太郎, 松本裕治. 擬似教師ありキャプション生成における部分的不一致の除去. 人工知能学会論文誌, Vol. 37, No. 2, pages H-L82_1–12, 2022.

2. Yohei Momoki, Akimichi Ichinose, Yutaro Shigeto, Ukyo Honda, Keigo Nakamura, and Yuji Matsumoto. Characterization of Pulmonary Nodules in Computed Tomography Images Based on Pseudo-Labeling Using Radiology Reports. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 32, No. 5, pages 2582–2591, 2021.

International Conferences (Refereed)

1. Ukyo Honda, Taro Watanabe, and Yuji Matsumoto. Switching to Discriminative Image Captioning by Relieving a Bottleneck of Reinforcement Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1124–1134, 2023.

2. Ukyo Honda, Yoshitaka Ushiku, Atsushi Hashimoto, Taro Watanabe, and Yuji Matsumoto. Removing Word-Level Spurious Alignment between Images and Pseudo-Captions in Unsupervised Image Captioning. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 3692–3702, 2021.

3. Ukyo Honda, Tsutomu Hirao, and Masaaki Nagata. Pruning Basic Elements for Better Automatic Evaluation of Summaries. In *Proceedings of the Conference of*

the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 661–666, 2018.

4. Jungmin Choi, Ukyo Honda, Taro Watanabe, Kentaro Inui, and Hiroki Ouchi. Law Retrieval with Supervised Contrastive Learning using the Hierarchical Structure of Law. In *Proceedings of the Pacific Asia Conference on Language, Information and Computation (PACLIC)*, 10 pages, 2022.

Domestic Conferences (Non-Refereed)

1. 本多右京, 渡辺太郎, 松本裕治. 強化学習における画像キャプションの低識別性問題と Long-Tail 分類手法を用いた対処. 言語処理学会第 28 回大会, pages 146–151, 2022.
2. 本多右京, 牛久祥孝, 橋本敦史, 渡辺太郎, 松本裕治. 画像と単語の不一致を考慮した疑似教師ありキャプション生成. 言語処理学会第 27 回大会, pages 1507–1512, 2021.
3. 本多右京, 平尾努, 永田昌明. 冗長な Ngram/Basic Elements を除いた自動要約評価指標. 言語処理学会第 24 回大会, pages 17–20, 2018.
4. チェ ジョンミン, 本多右京, 渡辺太郎, 大内啓樹. 法律の階層構造を利用した教師あり対照学習による法律検索. 第 36 回人工知能学会全国大会, pages 1–4, 2022.

Awards

1. 最優秀賞 (主著論文). 言語処理学会第 28 回大会. 強化学習における画像キャプションの低識別性問題と Long-Tail 分類手法を用いた対処. 2022.
2. 若手奨励賞 (主著論文). 言語処理学会第 27 回大会. 画像と単語の不一致を考慮した疑似教師ありキャプション生成. 2021.
3. 優秀賞 (共著論文). 第 36 回人工知能学会全国大会, インタラクティブセッション発表部門. 法律の階層構造を利用した教師あり対照学習による法律検索. 2022.