論 文 内 容 の 要 旨

博士論文題目

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

(汎用的なキャプション生成システムに向けたシステム入出力範囲の拡張)

氏　　名　本多　右京

(論文内容の要旨)

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 addresses 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 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.

氏　名	本多　右京

（論文審査結果の要旨）

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 research in this thesis focuses on broadening the limited coverage of scenes and descriptions. In the first study, the limitation of describable scenes is addressed by introducing unsupervised image captioning, in which images are automatically aligned with their captions using the detected object labels. The proposed method significantly enhances the captioning performance by a gating mechanism with pseudo-labels associated with it in order to remove word-level spurious alignment in the pseudo image-sentence pairs. In the second study, the limitation of diversity in descriptions is addressed by treating it as a long-tail classification and a debiasing problem, in which the probability mass is shifted from low-frequent words into high-frequent words. The proposed solution employs a simple fine-tuning method with a small modification to the model structure and generates distinctive captions by significantly increasing the output vocabulary.

The research in this thesis demonstrates that the limitations in scenes and descriptions are resolved by the careful analysis of the task settings and by simple yet effective methods to directly broadening the coverages. The proposed methods are sound and potentially applicable to other wider areas, e.g., text generation. Their effectiveness is demonstrated by the systematic experimental design followed by detailed discussion with human analysis. The studies are published in one high quality peer-reviewed journal paper and one peer-reviewed international conference paper. The research would have an impact not only to the intersection areas of computer vision and natural language processing, but to the relevant fields of text generation, e.g., machine translation and summarization. As a result, the thesis is sufficiently qualified as a Doctoral thesis of Engineering.