| | |
|---|---|
| 氏　　名 | ZHAO YUTING（チョウ ウテイ） |
| 所　　属 | システムデザイン研究科 システムデザイン専攻 |
| 学位の種類 | 博士（情報科学） |
| 学位記番号 | シス博　第 179 号 |
| 学位授与の日付 | 令和５年３月 25 日 |
| 課程・論文の別 | 学位規則第４条第１項該当 |
| 学位論文題名 | Multimodal Neural Machine Translation based on Image-Text Semantic Correspondence<br>（画像とテキストの意味対応に基づくマルチモーダルニューラル機械翻訳） |
| 論文審査委員 | 主査　教　授　小町　守<br>委員　教　授　高間　康史<br>委員　准教授　下川原　英理<br>委員　教　授　二宮　崇（愛媛大学） |

**【論文の内容の要旨】**

Machine translation (MT) is a task to automatically translate text from one language to another. Neural machine translation (NMT) is a prominent approach to MT in the field both actively researched and also deployed in many online translation services such as Google Translator. NMT has achieved state-of-the-art translation performance. The strength of NMT lies in its ability to learn directly, in an end-to-end fashion, mapping from the input text to the associated output text. Linguistic is at the heart of NMT, which requires representing the meaning of a source sentence in one language and predicting that to a target sentence in another language by training with large amounts of parallel sentences.

In contrast, humans are able to handle semantic tasks by making use of complex combinations of linguistic, visual, and auditory multimodalities simultaneously to improve the quality of perception and understanding. From a computational perspective, NMT also can benefit from incorporating auxiliary modalities, too, in order to approach human-level understanding in various aspects. As a consequence, multimodal NMT is a better reflection of how humans acquire and process language, with many theoretical advantages in language grounding and understanding over text-based NMT in the presence of multimodal content.

Multimodal neural machine translation (MNMT) extends the conventional text-based NMT by exploiting an auxiliary source modality, specifically images, to translate source sentences paired with images into a target language. The main motivation behind this is that the translation is expected to be more accurate than textual translation because there are numerous situations in which textual context alone is insufficient for correct translation such as for ambiguous words and grammatical gender. Therefore, many studies have focused on incorporating image modality to aid the interpretation of language for improving translation performance.

To effectively utilize an image, some studies have contextualized textual representations using global visual features extracted from an entire image to initialize textual encoder/decoder recurrent neural network (RNN) hidden states. However, the effect of the image cannot be fully exerted because the single global visual features of an entire image are complex. Other studies represent image information with a sequence of equally sized grid local visual feature vectors extracted by CNNs. These grid features are used to preserve the spatial correspondence with the input image. As these equally sized grid-based local visual features do not convey specific semantics, the role of visual features is dispensable in translation.

Consequently, current studies have incorporated richer local visual features for MNMT, such as DenseCap. However, their efforts have not convincingly demonstrated that visual features can improve the translation quality. It has been proved in the latest work when the textual context is limited, visual features can help generate better translations. MNMT disregards visual features because the quality of the image features or the way in which they are integrated is not satisfactory. Therefore, a significant challenge in MNMT task is how to enhance the translation of the text by leveraging their semantic correspondence to the images effectively.

In this work, two methods are proposed to cope with this significant challenge of MNMT task. One is named region-attentive MNMT and the other is named word-region alignment-guided MNMT (WRA-guided MNMT). The main contributions of this work are as follows:

- For the region-attentive MNMT method, I propose to utilize semantic image regions extracted by object detection for MNMT and integrate visual and textual features using two modality-dependent attention mechanisms. The main motivation behind this method is to exploit the effect of semantic information captured inside the visual features. The proposed method was implemented and verified on two neural architectures of NMT: the RNN and the Transformer. Experimental results on English-German and English-French translation tasks

using the public Multi30k dataset show that the proposed method improves over baselines and outperforms most of the existing MNMT models. Further analysis demonstrates that the proposed method can achieve better translation performance because of its better visual information use.

- For the WRA-guided MNMT method, I propose a novel facility named word-region alignment (WRA) for linking the semantic correlation between text and image modalities in MNMT as a bridge. The main motivation behind this method is to leverage the semantic relevance between the two modalities for improving translation with image guidance. The proposed method also has been implemented on two mainstream architectures of NMT: the RNN and the Transformer. Experimental results on English-German and English-French translation tasks using the Multi30k dataset prove that the proposed method has a significant improvement with respect to the competitive baselines and outperforms most of the existing MNMT models. Further analysis demonstrates that this model can achieve better translation performance by integrating WRA, leading to better visual information use.

This thesis is organized as follows: Chapter 1 introduces the background and overview of this work. Chapter 2 describes existing works of the MNMT task. Chapter 3 details the proposed method of the region-attentive MNMT model. Chapter 4 details the proposed method of the word-region alignment-guided MNMT model. Chapter 5 makes a conclusion of this dissertation and describes future directions.