公益財団法人矢崎科学技術振興記念財団
奨励研究助成　成果報告書

公益財団法人矢崎科学技術振興記念財団
　理事長 殿

研究助成期間終了にあたり、下記の通り成果を報告します。

2025 年 4 月 30 日
氏名　ZHAO YUTING
所属　九州大学
職位　助教

1. 申請研究の題目
多言語およびマルチモーダルな機械翻訳に関する研究

2. 研究の目的
In recent years, with the advent of Transformer and high-quality image recognition models, many multimodal machine translation (MMT) models have been proposed and have shown advanced translation performance that exceeds conventional MT models that use only sentences in the source language. However, the practicability and efficiency of the MMT models are still limited to the weak correlation between text and image. This project will model a novel architecture of MMT to enforce the correlation between text and image concepts (words and objects) through learning a cross modal shared embedding space on both semantic and spatial levels. The motivation behind this is to innovate previous MMT methods, which represent textual and visual features in different space independently, to represent multimodal information in a shared embedding space.

3. 研究の内容
We construct multimodal, visual and linguistic scene graphs, which can be used to represent visual-linguistic relationships in a shared space with the patten of graph nodes and edges.

We first generate image scene graphs from images using multimodal large-scale language models, which efficiently extracts structured image spatial relationships. To obtain coherent and well-structured image relationships, we set the sampling temperature to 0 for deterministic generations and raise it to 0.4 for more challenging cases requiring exploratory outputs. Our prompt engineering comprises specifying how to formulate relations and produce structured scene graph content; illustrating common errors in output formatting and how to rectify them; and providing abstract but well-formed scene graph templates without using specific object names, thus preventing prompt contamination.

Secondly, we generate a text scene graphs by applying the textual parser, which efficiently extracts structured textual semantic relationships. The resulting text scene graphs encode entities and their relations.

Then we connect image scene graph and text scene graph into a multimodal scene graph through the super node, which utilizes the M-CLIP image encoder to generate visual embeddings of the entire image as super node features and establishing fully connected relationships with all nodes from both image and text scene graphs. This architecture enables effective multimodal information flow through the graph structure. For node and edge feature generation, we leverage the M-CLIP text encoder to process all nodes and edges in multimodal scene graphs.

Finally, we employ multi-layer graph neural network to represent multimodal information from image and text based on semantic-level and spatial-level relationships. Then a pre-trained mBART decoder on aligned multimodal scene graph using the Multi30K dataset, enabling the model to capture structured multimodal relationship patterns for decoding target translation.

4. 研究の成果と結論、今後の課題

We evaluate our model on Multi30K. Multi30K is a widely used MMT benchmark, serving as a multilingual extension of the Flickr30k dataset. Evaluation is conducted on three standard test splits: Test2016, Test2017, and MSCOCO for EN->{DE, FR} translation tasks. The MSCOCO test split comprises sentences with ambiguous verbs and out-of-domain data points from the MSCOCO dataset, representing a generally challenging setting for MMT. The implementation is based on PyTorch, leveraging the Huggingface. All evaluations are conducted using the BLEU, computed with SacreBLEU, which is recognised as the standard for early stopping. We report results based on the checkpoint that achieves the highest BLEU score on the validation set. We also benchmark our model using the METEOR metric, calculated with the evaluate library.

Finally, beneficial from multimodal deep relationship learning based on scene graph patterns, empirical experiments show significant improvements of this method in both BLEU and METEOR over all baseline models on EN-> {DE, FR} translation tasks.

In this work, we limited our experiments to English as the source language. This constraint stems from our reliance on the graph parser, which currently only supports English language rxsprocessing. Whilst this design choice allows us to effectively extract structural information from English inputs, extending the approach to non-English source languages would require developing comparable parsing capabilities for each source language – a significant undertaking we leave for future work.

5. 成果の価値

5－1. 学術的価値

From an academic perspective, this work paves the way for research on integrating multimodal information for solving downstream tasks, such as image captioning, visual question answering, and etc. The integration method of multimodal information still important for inspiring further studies on developing multimodal large-scale language models and visual-language models.

5－2. 社会的価値

From a social perspective, this work comprehensively breaks through the existing barriers of multimodal translation, aiming to achieve high-performance and novel machine translation. Communication in the world is inseparable from language, and language communication is based on effective machine translation. Realizing a diversified machine translation system with multiple languages and diverse medium is the potential way for the next generation of machine translation applications. In the future, multilingual and multimodal machine translation systems will facilitate semantic communication between multiple languages with taking advantage of diverse modalities, it will become a popular way to allow people to communicate freely without having to learn multiple languages, such as travel, cultural exchange, and trade.

6. 研究成果

The experimental result of this work is summarized in a long paper, which is under submission. The model implementation and the constructed multimodal scene graph datasets will be published.

以上