

Doctoral Thesis

**Multimodal Neural Machine Translation based on
Image-Text Semantic Correspondence**

Yuting Zhao

March, 2023

Tokyo Metropolitan University
Graduate School of Systems Design
Department of Computer Science

A Doctoral Thesis
submitted to Graduate School of Systems Design,
Tokyo Metropolitan University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Yuting Zhao

Thesis Committee:

Mamoru Komachi (Professor, Tokyo Metropolitan University)
Yasufumi Takama (Professor, Tokyo Metropolitan University)
Eri Shimokawara (Associate Professor, Tokyo Metropolitan University)
Takashi Ninomiya (Professor, Ehime University)

Acknowledgements

First of all, I am immensely grateful to my supervisor Professor Mamoru Komachi for his lead, guidance, and support from my master's course until my Ph.D. course. I will always be grateful to him for welcoming me into the Komachi laboratory, letting me find my research interests in the NLP field, and making me who I am today. Without him, it is not possible for me to start my Ph.D. journey and my research career. His wisdom and sense of responsibility will always influence me in every aspect of my research career and will always be a role model for me to learn from.

I am very grateful to Dr. Chenhui Chu. Thanks for giving me guidance, support, and encouragement in many aspects of my research career. His knowledge and attitude to research will always influence me in my future study and career.

I am very grateful to Dr. Tomoyuki Kajiwara for his kind guidance and suggestion during our collaborative research. I will always keep every moment that he supports me and encourages me in mind.

I am very grateful to my mentor Ioan Calapodescu and Professor Laurent Besacier during my internship at NAVER LABS. I am very grateful to you for welcoming me to join the NLP team. Thank you for your professional guidance and kind support. I would also like to thank all the scientists in the NLP group, especially the STAG project, I thoroughly enjoyed our time working together and learned so much.

I sincerely thank my master's thesis committee members: Professor Toru Yamaguchi and Professor Yasufumi Takama and my doctoral thesis committee members: Professors Yasufumi Takama, Takashi Ninomiya, and Associate Professor Eri Shimokawara. I was able to complete this thesis thanks to your insightful advice and comments.

Many thanks to the seniors who have graduated and classmates from the Komachi laboratory for mentoring me, helping me, and supporting me during these five years.

Last but not least, thanks to all my friends. Most importantly, great thanks to my parents and my family.

Publication List

Journal Papers

1. Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, Chenhui Chu. Word-Region Alignment-Guided Multimodal Neural Machine Translation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Volume 30, Pages 244-259, Jan 2022.
2. Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, Chenhui Chu. Region-Attentive Multimodal Neural Machine Translation. *Neurocomputing*, Volume 476, Pages 1-13, Mar 2022.

International Conference Papers

1. Yuting Zhao, Ioan Calapodescu. Multimodal Robustness for Neural Machine Translation. *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8505-8516, 2022.
2. Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, Chenhui Chu. Double Attention-based Multimodal Neural Machine Translation with Semantic Image Regions. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pp. 105-114, 2020.
3. Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, Chenhui Chu. TMEKU System for the WAT 2021 Multimodal Translation Task System. *Proceedings of the 8th Workshop on Asian Translation*, pp. 174-180, 2021.
4. Longtu Zhang, Yuting Zhao, Mamoru Komachi. TMU Japanese-Chinese Un-supervised NMT System for WAT 2018 Translation Task. *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 5th Workshop on Asian Translation*, pp. 981-987, 2018.

Domestic Conference Papers

1. Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, Chenhui Chu. Neural Machine Translation with Semantically Relevant Image Regions. In the 27th Annual Meeting of the Language Processing Society of Japan, Vol. 2, pp. c3, 2021.

2. Yuting Zhao, Longtu Zhang, Mamoru Komachi. Application of Unsupervised NMT Technique to Japanese-Chinese Machine Translation. In The 33rd Annual Conference of the Japanese Society for Artificial Intelligence, pp. 3B4E204–3B4E204, 2019.

3. Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, Chenhui Chu. Double Attention-based Multimodal Neural Machine Translation with Semantic Image Regions. In The 241st Meeting of Special Interest Group of Information Processing Society of Japan Natural Language Processing, Vol.2019–NL–241, 2019.

4. Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, Chenhui Chu. Double Attention-based Multimodal Neural Machine Translation with Semantic Image Regions. In The 14th Symposium of Young Researcher Association for NLP Studies, 2019.

Multimodal Neural Machine Translation based on Image-Text Semantic Correspondence*

Yuting Zhao

Abstract

Multimodal neural machine translation (MNMT) extends the conventional text-to-text neural machine translation (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.

Recently, researchers in this field have established a shared task called multimodal neural machine translation (MNMT), which consists of translating a target sentence from a source language description into another language using information from the image described by the source sentence. The training resource for MNMT is called Multi30K, which is a triple dataset containing images, image descriptions, and multi-lingual translations. The research topic of this work focuses on exploiting the effective integration of image information based on image-text semantic correspondence to improve the translation performance of MNMT.

In this study, I propose two methods of the MNMT task to enhance the translation of the text by leveraging image-text semantic correspondence to the images effectively: one is named region-attentive MNMT model and the other is named word-region alignment-guided MNMT model. These two methods have been implemented on two mainstream architectures of NMT: the recurrent neural network (RNN) and the

*Doctoral Thesis, Department of Computer Science, Graduate School of Systems Design, Tokyo Metropolitan University, March, 2023.

Transformer. Experimental results on English–German and English–French translation tasks using the public Multi30k dataset prove that my methods can achieve significant improvement with respect to their competitive baselines and outperform most of the existing MNMT methods across BLEU and METEOR evaluation metrics. Further analysis corresponding to each method demonstrates that the proposed methods can achieve better translation performance because of their better image information use, respectively.

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 thesis and describes future directions.
- Chapter 6 introduces the social impacts of this work.

Keywords:

Neural machine translation (NMT), Multimodal neural machine translation (MNMT), Multi30k, Semantic correspondence.

Contents

Acknowledgements	i
Publication List	ii
1 Introduction	2
2 Existing Works	6
2.1 How to integrate image modality?	6
2.2 How to correlate text and image modalities?	7
3 Region-Attentive MNMT	9
3.1 Introduction	9
3.2 Methodology	11
3.2.1 RA-RNN: Region-Attentive Multimodal RNN	11
Sentence Encoder	12
Image Encoder	12
Decoder	12
Text-attention mechanism.	13
Image-attention mechanism.	13
Generation.	14
3.2.2 RA-TRANS: Region-Attentive Multimodal Transformer	14
Encoder	14
Decoder	16
Double Cross-Attentions	16
3.3 Experiments	19
3.3.1 Dataset	19
3.3.2 Evaluation Metrics	20

3.3.3	Baselines	20
	RNN.	20
	Grid-Attentive Multimodal RNN (GA-RNN).	20
	TRANS.	20
	Grid-Attentive Multimodal Transformer (GA-TRANS).	21
3.3.4	Setup	21
	Settings of the RNN-Based Models	21
	Settings of the Transformer-Based Models	21
3.4	Results	22
3.4.1	Results within RNN-Based Models	22
3.4.2	Results within Transformer-Based Models	23
3.4.3	Comparison of Proposed Model and Existing Ones	23
3.5	Analyses	26
3.5.1	Pairwise Evaluation	26
3.5.2	Qualitative Analysis	28
	Visualization within RNN-Based Models	28
	Visualization within Transformer-Based Models	29
	Visualization between RA-RNN and RA-TRANS	30
3.6	Summary	32
4	WRA-Guided MNMT	33
4.1	Introduction	33
4.2	Methodology	35
4.2.1	WRA Generation	35
	Soft WRA	36
	Hard WRA	36
4.2.2	WRA-Guided RNN-Based MNMT Model	37
	Textual Encoder	37
	Visual Encoder	37
	Word-to-Region (W2R)	39
	Doubly Attentive Decoder	40
4.2.3	WRA-Guided Transformer-Based MNMT Model	41
	Textual Encoder	41
	Visual Encoder	42

	Word-to-Region (W2R)	42
	Doubly Attentive Decoder	43
4.3	Experiments	44
4.3.1	Datasets	44
4.3.2	Evaluation	44
4.3.3	Setup	44
	Settings of the RNN-Based Models	45
	Settings of the Transformer-Based Models	45
4.3.4	Further Experimental Comparison	46
4.4	Results	47
4.4.1	Results on the En→De Task	47
	Results within RNN-Based Models	47
	Results within Transformer-Based Models	49
	Comparison of Proposed Model and Existing Ones	49
4.4.2	Results on the En→Fr Task	51
4.5	Analyses	51
4.5.1	Ablation Study	51
	Different Integration Strategies of WRA	51
	Different Intermodal Fusion Operations	54
4.5.2	Visualization	57
4.5.3	Case Study	59
4.6	Summary	62
5	Conclusions and Future Directions	63
6	Social Impacts	64
	Bibliography	65

List of Figures

3.1	Overview of region-attentive multimodal neural machine translation (RA-NMT).	9
3.2	RA-RNN: Region-Attentive Multimodal RNN.	11
3.3	RA-TRANS: Region-Attentive Multimodal Transformer.	15
3.4	Left: Text Scaled Dot-Product Attention; Right: Image Scaled Dot-Product Attention.	17
3.5	Left: Text Multi-Head Attention; Right: Image Multi-Head Attention.	17
3.6	Examples for text-only RNN, grid-attentive multimodal RNN (GA-RNN), and region-attentive multimodal RNN (RA-RNN). Red and blue words indicate incorrect and correct, respectively.	29
3.7	Examples for text-only Transformer (TRANS), grid-attentive multimodal Transformer (GA-TRANS), and region-attentive multimodal Transformer (RA-TRANS). Red and blue words indicate incorrect and correct, respectively.	30
3.8	Examples for region-attentive multimodal RNN (RA-RNN) and region-attentive multimodal Transformer (RA-TRANS). Red and blue words indicate incorrect and correct, respectively. In each example, left/right figures correspond to the visualization of RA-RNN/RA-TRANS.	31
4.1	Example of WRA-guided MNMT. The WRA builds semantic relevance between the vision and language. Specifically, each region-level visual feature is annotated using a visual concept that is used to create a relationship with every source word. When generating the “rouge,” similar attention weights (“att” in the figure) are assigned to both the corresponding source word “red” and image region “red shirt.”	34

4.2	WRA generation. Each region-level visual feature was annotated using a visual concept consisting of an attribute class, followed by an object class. Subsequently, it was used to create a relationship with each source word based on semantic similarity. The WRA represents the semantic correlation between each regional visual feature and all words in a sentence.	35
4.3	WRA-Guided RNN-Based MNMT Model.	38
4.4	WRA-Guided Transformer-Based MNMT Model.	42
4.5	Representation visualization for textual features, independent visual features, and enriched visual features with soft/hard WRA-guided textual features. Representations are learned by RNN-based/Transformer-based $MNMT_{W2R(sa)}$ and $MNMT_{W2R(ha)}$ on the En→De and En→Fr tasks, respectively. <i>Text (blue)</i> : the textual representations generated by the textual encoder. <i>Image_(independent) (orange)</i> : the visual representations generated by visual encoder before conducting W2R, which are independent of textual representations. <i>Image_(MNMT_{W2R(sa/ha)}) (green)</i> : the enriched visual representations generated by W2R, which have been related with textual features by leveraging soft/hard WRA as a bridge.	58
4.6	Improved examples for the case study.	60

List of Tables

3.1	The experimental results of RNN-based architectures. The best performance is highlighted in bold. † indicates that the result is significantly better than the text-only RNN baseline at a p-value of < 0.05	22
3.2	The experimental results of Transformer-based architectures. The best performance is highlighted in bold. † and ‡ indicate that the result is significantly better than the text-only TRANS and GA-TRANS baselines at p-value < 0.05 , respectively.	23
3.3	Comparison with existing methods. Among all the results, I highlight the best performance in bold. All the experimental results of my proposal are the average scores over three runs.	25
3.4	Pairwise evaluation. I counted the number and proportion of various categories among 50 random examples.	27
4.1	BLEU and METEOR scores on Multi30k En→De task. The results are significantly better than those of NMT (†) and MNMT _R (‡) with p-value < 0.05 . The best performance in my models and existing MNMT models appear in bold. All my results are the average scores over three runs.	48
4.2	BLEU and METEOR scores on Multi30k En→Fr task. The results are significantly better than those of NMT (†) and MNMT _R (‡) with p-value of < 0.05 . The best performance in my models and existing MNMT models appear in bold. All my results are the average scores over three runs.	50
4.3	Ablation study on different integration strategies of WRA. BLEU and METEOR scores on En→De task using the Multi30k dataset. The best performance is shown in bold. All results are the average scores over three runs.	52

4.4	Ablation study on different integration strategies of WRA. BLEU and METEOR scores on En→Fr task using the Multi30k dataset. The best performance is shown in bold. All results are the average scores over three runs.	53
4.5	Ablation study on different intermodal fusion operations. BLEU and METEOR scores on En→De task using the Multi30k dataset. The best performance is shown in bold. All results are the average scores over three runs.	55
4.6	Ablation study on different intermodal fusion operations. BLEU and METEOR scores on En→Fr task using the Multi30k dataset. The best performance is shown in bold. All results are the average scores over three runs.	56

1 Introduction

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 [Sutskever et al., 2014, Cho et al., 2014b, Bahdanau et al., 2015, Vaswani et al., 2017]. 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. Computational language semantic understanding 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-to-text NMT [Sutskever et al., 2014, Bahdanau et al., 2015] 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 [Specia et al., 2016, Elliott et al., 2017, Barrault et al., 2018].

The study of the potential for improving translation quality using images was pioneered by [Elliott et al., 2015]. Subsequent studies have integrated image information using a single global visual feature vector extracted by convolutional neural networks (CNNs). For example, some models use the global visual feature in the following ways: initializing the encoder/decoder hidden states [Elliott et al., 2015, Huang et al., 2016, Calixto and Liu, 2017]; performing element-wise multiplication with target word embeddings [Caglayan et al., 2017a]; impacting the text encoder by learning an image representation jointly [Elliott and Kádár, 2017, Helcl et al., 2018]. In addition, some models use the global visual feature to interact between the sources through a latent variable [Calixto et al., 2019], a shared space [Zhou et al., 2018], or a universal representation [Zhang et al., 2020]. Although they aim to combine text and image sources to generate a good translation, the effect of the image cannot be fully exerted because the single global visual features of an entire image are complex.

To effectively utilize an image, other studies represent image information with a sequence of equally sized grid local visual feature vectors extracted by CNN. These grid features are used to preserve the spatial correspondence with the input image. For example, a joint representation is generated by combining visual and textual representations [Fukui et al., 2016], compute a multimodal context vector using a multimodal or filtered attention mechanism [Caglayan et al., 2016b, Caglayan et al., 2016a, Caglayan et al., 2018], and focus on textual and visual annotations independently by different strategies on attention mechanisms [Calixto et al., 2016, Calixto et al., 2017, Libovický and Helcl, 2017, Delbrouck and Dupont, 2017]. As these equally sized grid-based local visual features do not convey specific semantics, the role of visual features is dispensable in translation.

To overcome the above difficulties, current studies attempt to represent an image using multiple object-level regional features [Tan and Bansal, 2019, Zhao et al., 2020]. [Huang et al., 2016], for example, integrated regional features followed by the text sequence. [Toyama et al., 2016] proposed a transformation to mix global visual features and regional features. [Grönroos et al., 2018] and [Ive et al., 2019] generated a single representation of regional features to initialize the encoder or target word embeddings. Furthermore, [Yang et al., 2020] proposed a multi-head co-attention upon regional features. [Yin et al., 2020] used a unified multimodal graph to capture seman-

tic relationships between words and objects. As proved in [Caglayan et al., 2019], MNMT models disregard visual features because the quality of the image features or the manner in which they are integrated into the model is not satisfactory. Sequentially, some recent works tried to explore the correlations within visual and textual modalities [Zhao et al., 2021a, Zhao et al., 2021b]. Thus far, a significant challenge in the MNMT task is how to enhance the translation of the text by leveraging their semantic correspondence to the images.

In this work, two methods are proposed to cope with this significant challenge of the 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: recurrent neural network (RNN) and Transformer. Experimental results on English–German and English–French translation tasks using the Multi30k dataset show that the proposed method improves over baselines and outperforms most of the existing MNMT methods. 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 methods. Further analysis demonstrates that this model can achieve better translation performance by integrating WRA, leading to better visual information use.

The remainder of this thesis is organized as follows:

- 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 thesis and describes future directions.
- Chapter 6 introduces the social impacts of this work.

2 Existing Works

2.1 How to integrate image modality?

Early MNMT models integrated visual information using a single global visual feature extracted by a convolutional neural network (CNN). They used the global visual feature to contextualize textual representations in the following ways: (1) appending them at the head/tail to the original textual sequence [Huang et al., 2016]; (2) initializing the textual encoder and/or decoder RNN hidden states with them [Calixto and Liu, 2017]; (3) interacting with them elementwise using textual annotations or target word embeddings [Caglayan et al., 2017a]; and (4) influencing the textual encoder by learning the visual representation alongside them [Elliott and Kádár, 2017]. Although these models were designed to enrich the textual context using sufficient visual information to improve the translation, it is difficult to summarize all the semantic information of an entire image into a single global visual feature.

To address this issue, subsequent researchers represented visual information using a set of convolutional local features that are equally sized grid local features. These features were used to preserve spatial correspondence with the image. The following integration methods were investigated: (1) computing a multimodal context using a multimodal attention mechanism that simultaneously focuses on an image and its source description [Caglayan et al., 2016b, Caglayan et al., 2016a]; (2) conjecturing a learnable masking operation over the convolutional feature maps to help the attention mechanism filter out local features that are irrelevant to translation and focus on the most important part of the visual inputs [Caglayan et al., 2018]; (3) focusing on textual and visual features independently using different attention strategies [Calixto et al., 2017, Libovický and Helcl, 2017, Delbrouck and Dupont, 2017]; (4) attending to local features by setting an additional attention sublayer after self-attention [Helcl et al., 2018, Libovický et al., 2018]. However, in the aforementioned approaches, the

attention mechanism cannot easily distinguish equally sized local features. As proved in [Elliott, 2018], attending to specific regions of the image is crucial to improving the translation.

Consequently, in recent studies, images are represented using multiple object-level regional features to solve the aforementioned limitations by attempting the following integration strategies: [Huang et al., 2016] integrates regional features followed by the text sequence. [Toyama et al., 2016] proposes a transformation to mix global visual features and regional features. [Grönroos et al., 2018] and [Ive et al., 2019] generate a single representation of regional features to initialize the encoder or target word embeddings. Furthermore, [Yang et al., 2020] proposed a multi-head co-attention upon regional features. [Yin et al., 2020] used a unified multimodal graph to capture semantic relationships between words and objects. So far, how to effectively integrate image information for the MNMT model still remains an open question.

2.2 How to correlate text and image modalities?

Although regional features aid object localization or semantic information presentation [Zhao et al., 2020], the manner in which they are integrated into the model still needs to be improved. Based on [Caglayan et al., 2019], if the textual modality is sufficient to accomplish the translation task, the visual modality should be integrated to play a complementary role rather than a redundant role.

Toward this end, an emerging trend of exploiting correlations between modalities has been considered promising. Some strategies have been developed: (1) jointly learning a shared vision-language embedding space and a translator [Zhou et al., 2018]; (2) modeling the interaction between visual and textual features using a latent variable model alongside a translation model [Calixto et al., 2019]; (3) training multi-head co-attention to capture the interaction between visual and textual features in multiple subspaces [Yang et al., 2020]; (4) learning a universal visual representation by retrieving associated images for words in a source sentence [Zhang et al., 2020]; (5) utilizing manually annotated datasets to train supervised visual attention [Nishihara et al., 2020]; (6) integrating multimodal graph neural networks [Yin et al., 2020] and dynamic context-guided capsule networks [Lin et al., 2020] into the MNMT. Although these researchers successfully prove the effectiveness of relating textual and visual in-

formation for MNMT, there are lingering concerns. First, jointly learning visual and textual representations with latent space requires large-scale training data that MNMT lacks. Second, in multimodal tasks, different modalities do not usually have equal importance. It is suggested that texts are obviously more important than images [Yao and Wan, 2020]. Likewise, the impact of textual predominance has been revealed by [Chowdhury and Elliott, 2019]. Therefore, how to effectively correlate multimodal inputs is a lingering challenge for MNMT.

Meanwhile, some methods have been developed for correlating modalities in other multimodal tasks that focus on image-to-text one-way operation: (1) Correlating textual and visual modalities by a multimodal embedding space [Karpathy and Li, 2015, Gupta et al., 2017]. Rather than learning a joint space, a pre-processed facility WRA using visual concepts as an intermediary to build semantic relevance between words and regions is proposed in this paper. (2) Aligning textual and visual features by different attention mechanisms, such as a mutual attention mechanism [Liu et al., 2019], a stack of co-attention layers [Nguyen and Okatani, 2018], and self-attention [Huang et al., 2020]. In contrast, in my study, double attention is collocated for learning alignment between source words and target words and between image regions and target words. (3) Jointly training visual and textual attention mechanisms [Nam et al., 2017] or jointly learning word-tag-region triple embeddings [Li et al., 2020] to find shared semantics between images and sentences. Although previous efforts have explored different strategies to mix vision and language, the efficient integration of multimodal information still remains a challenging task though.

3 Region-Attentive MNMT

3.1 Introduction

In this proposed method, as shown in Figure 3.1, I attempt to combine object detection with an additional region-dependent attention mechanism for fully exploiting semantic image region features upon NMT architectures, which is called region-attentive multi-modal neural machine translation (RA-NMT). In RA-NMT, it is possible to focus on different parts of the source sentence and different object-level regions of the image at the same time. The main motivation behind this is that I expect the proposed method to take advantage of useful visual information by attending to specific regions of the image to assist in translating source words.

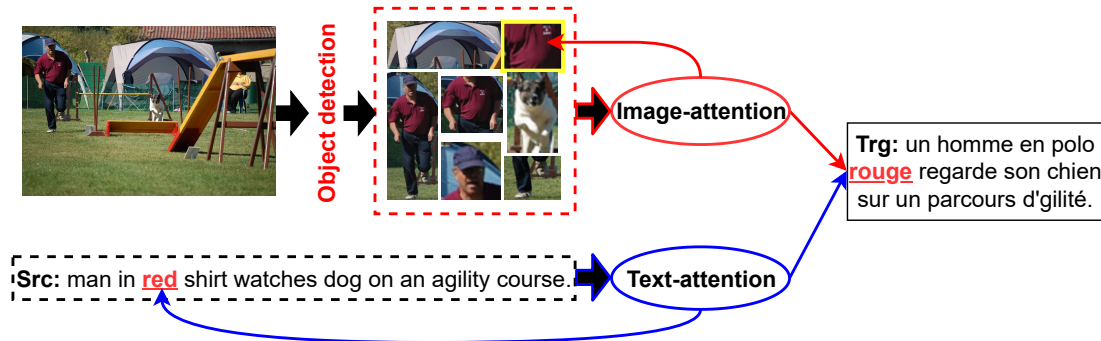


Figure 3.1: Overview of region-attentive multimodal neural machine translation (RA-NMT).

Technically, rather than equally sized grid local visual features, I present that semantic image region features containing object attributes and relationships are essential to MNMT. Furthermore, inspired by previous studies [Caglayan et al., 2016b, Calixto et al., 2017, Libovický and Helcl, 2017, Delbrouck and Dupont, 2017] on the investigation of the attention mechanism for multi-source learning, I introduce that a

region-dependent attention mechanism is a promising way to make MNMT attend to the salient regions of an image. Therefore, instead of utilizing regional features to initialize/contextualize language representations [Huang et al., 2016, Ivey et al., 2019], I propose integrating semantic image region features into MNMT with two modality-dependent attention mechanisms, one for text and the other for the semantic image regions, which is significantly different from the previous studies.

In this study, I implemented and verified the proposed method on not only the RNN-based architecture but also Transformer-based architecture, which are called the region-attentive multimodal RNN (RA-RNN) method and region-attentive multimodal transformer (RA-TRANS) method, respectively. Experimental results on different language pairs of the Multi30k dataset show that my proposed method improves over baselines and outperforms most of the state-of-the-art MNMT methods. Further analysis demonstrates that the proposed method can achieve better translation performance because of its better visual feature use.

The main contributions are as follows:

- I propose a multimodal method that combines object detection with an additional region-dependent attention mechanism to fully exploit semantic image region features on NMT architectures, which is called RA-NMT. This proposal is implemented and verified on two types of NMT architectures: RNN and Transformer.
- Extensive experimental results show that my proposed method improves over baselines on both RNN and Transformer architectures. The further experimental comparison shows that my proposed method outperforms most existing MNMT methods.
- Further analysis demonstrates that the proposed method can make better use of visual information by attending to specific semantic image regions with an additional region-dependent attention mechanism.

3.2 Methodology

3.2.1 RA-RNN: Region-Attentive Multimodal RNN

As shown in Figure 3.2, the proposed RA-RNN, based on [Calixto et al., 2017], comprises three parts: sentence encoder, image encoder, and decoder.

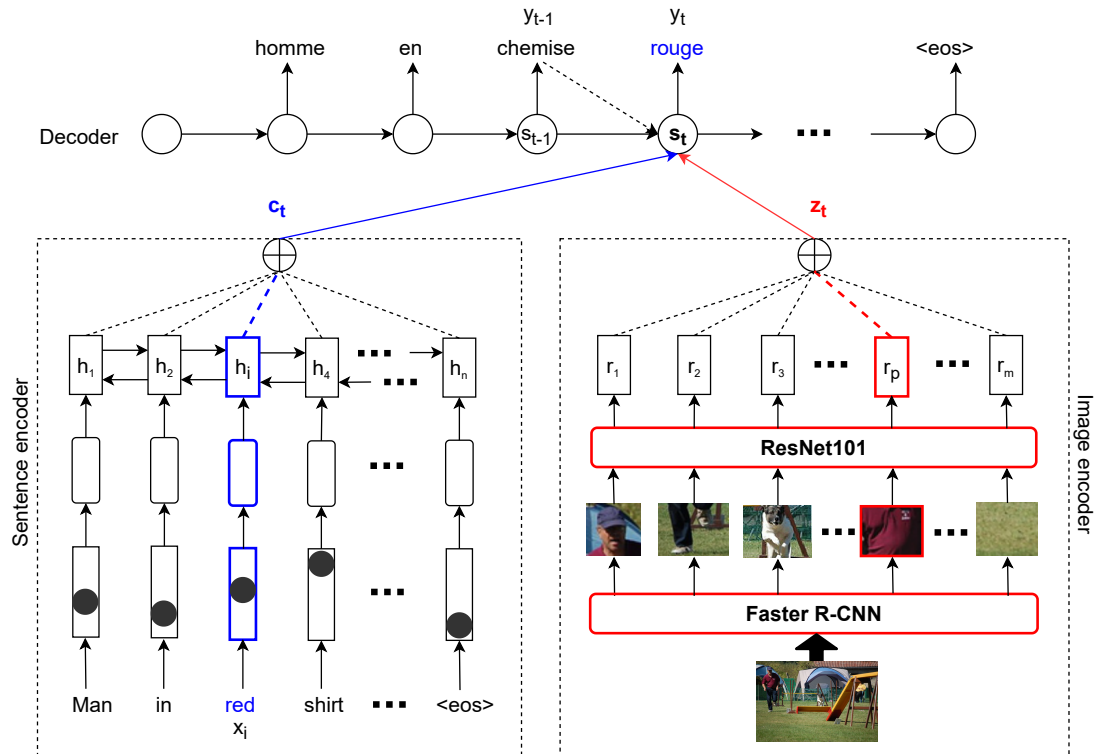


Figure 3.2: RA-RNN: Region-Attentive Multimodal RNN.

I integrate the visual features using an additional attention mechanism. From the source sentence $X = (x_1, x_2, x_3, \dots, x_n)$ to the target sentence $Y = (y_1, y_2, y_3, \dots, y_g)$, the image attention mechanism focuses on all semantic image region features to calculate the image context vector z_t , whereas the text-attention mechanism computes the text context vector c_t . The decoder is an RNN with a conditional gated recurrent unit (cGRU) to generate the current hidden state s_t and target word y_t on two attention mechanisms.

At time step t , a hidden state proposal \hat{s}_t is initially computed in cGRU, and then

the image context vector z_t and text context vector c_t are calculated.

$$\begin{aligned}\hat{\xi}_t &= \sigma(W_\xi E_Y[y_{t-1}] + U_\xi s_{t-1}) \\ \hat{\gamma}_t &= \sigma(W_\gamma E_Y[y_{t-1}] + U_\gamma s_{t-1}) \\ \dot{s}_t &= \tanh(W E_Y[y_{t-1}] + \hat{\gamma}_t \odot (U s_{t-1})) \\ \hat{s}_t &= (1 - \hat{\xi}_t) \odot \dot{s}_t + \hat{\xi}_t \odot s_{t-1}\end{aligned}$$

where W_ξ , U_ξ , W_γ , U_γ , W , and U are trainable parameters; E_Y is the target word vector.

Sentence Encoder

The sentence encoder is a bi-directional RNN with GRU [Cho et al., 2014a]. Given a sentence $X = (x_1, x_2, x_3, \dots, x_n)$, the encoder updates the forward hidden states with annotation vectors $(\vec{h}_1, \vec{h}_2, \vec{h}_3, \dots, \vec{h}_n)$, and updates the backward with annotation vectors $(\overleftarrow{h}_1, \overleftarrow{h}_2, \overleftarrow{h}_3, \dots, \overleftarrow{h}_n)$. By concatenating the forward and backward vectors $h_i = [\vec{h}_i; \overleftarrow{h}_i]$, each h_i encodes the entire sentence while focusing on the x_i word, and all words in a sentence are denoted as $C = (h_1, h_2, \dots, h_n)$.

Image Encoder

The image encoder is an object-detection-based approach following [Anderson et al., 2018], acting as a feature extractor in the object-level image region.

As shown in Figure 3.2, when given an input image, the image encoder first employs an object detection method, which is Faster R-CNN [Ren et al., 2015] pre-trained on Visual Genome [Krishna et al., 2017], to propose m object-level image regions from each image. Then, based on the detected object-level image regions, a ResNet101 [He et al., 2016] pre-trained on ImageNet [Russakovsky et al., 2015] is utilized to extract semantic image region features. Finally, each semantic image region feature is represented as a vector r with dimensions d_r , and all of these features in each image are denoted as $R = (r_1, r_2, r_3, \dots, r_m)$.

Decoder

The decoder comprises three parts: the text-attention mechanism, the image-attention mechanism, and generation.

Text-attention mechanism. At time step t , the text context vector c_t is generated as follows:

$$\begin{aligned} e_{t,i}^{\text{text}} &= (V^{\text{text}})^T \tanh(U^{\text{text}} \hat{s}_t + W^{\text{text}} h_i) \\ \alpha_{t,i}^{\text{text}} &= \text{softmax}(e_{t,i}^{\text{text}}) \\ c_t &= \sum_{i=1}^n \alpha_{t,i}^{\text{text}} h_i \end{aligned}$$

where V^{text} , U^{text} , and W^{text} are trainable parameters; $e_{t,i}^{\text{text}}$ is the attention energy; $\alpha_{t,i}^{\text{text}}$ is the attention weight matrix of the source sentence.

Image-attention mechanism. At time step t , the image-attention mechanism focuses on the m semantic image region features and computes the image context vector z_t .

I initially calculate the attention energy $e_{t,p}^{\text{img}}$, which scores the degree of output matching between the inputs around position p and the output at position t , as follows:

$$e_{t,p}^{\text{img}} = (V^{\text{img}})^T \tanh(U^{\text{img}} \hat{s}_t + W^{\text{img}} r_p)$$

where V^{img} , U^{img} , and W^{img} are trainable parameters.

Then, the weight matrix $\alpha_{t,p}^{\text{img}}$ of each r_p is computed as follows:

$$\alpha_{t,p}^{\text{img}} = \text{softmax}(e_{t,p}^{\text{img}})$$

At time step t , the image-attention mechanism dynamically focuses on the m semantic image region feature vectors and computes the image context vector z_t , as follows:

$$z_t = \beta_t \sum_{p=1}^m \alpha_{t,p}^{\text{img}} r_p$$

For z_t , at each decoding time step t , a gating scalar $\beta_t \in [0, 1]$ [Xu et al., 2015] was used to adjust the proportion of the image context vector according to the previous hidden state s_{t-1} .

$$\beta_t = \sigma(W_\beta s_{t-1} + b_\beta)$$

where W_β and b_β are trainable parameters.

Generation. At time step t of the decoder, the new hidden state s_t is generated in the cGRU, as follows:

$$\begin{aligned}\xi_t &= \sigma(W_\xi^{\text{text}}c_t + W_\xi^{\text{img}}z_t + \bar{U}_\xi\hat{s}_t) \\ \gamma_t &= \sigma(W_\gamma^{\text{text}}c_t + W_\gamma^{\text{img}}z_t + \bar{U}_\gamma\hat{s}_t) \\ \bar{s}_t &= \tanh(W^{\text{text}}c_t + W^{\text{img}}z_t + \gamma_t \odot (\bar{U}\hat{s}_t)) \\ s_t &= (1 - \xi_t) \odot \bar{s}_t + \xi_t \odot \hat{s}_t\end{aligned}$$

where W_ξ^{text} , W_ξ^{img} , \bar{U}_ξ , W_γ^{text} , W_γ^{img} , \bar{U}_γ , W^{text} , W^{img} , and \bar{U} are model parameters; ξ_t and γ_t are the output of the update/reset gates; \bar{s}_t is the proposed updated hidden state.

Finally, the output probability is computed as follows:

$$\text{softmax}(L_o \tanh(L_s s_t + L_c c_t + L_z z_t + L_w E_Y[y_{t-1}]))$$

where L_o , L_s , L_c , L_z , and L_w are trainable parameters.

3.2.2 RA-TRANS: Region-Attentive Multimodal Transformer

As shown in Figure 3.3, RA-TRANS comprises three parts: encoder, decoder, and image encoder. I propose RA-TRANS based on Transformer architecture [Vaswani et al., 2017]. In the decoder, I implement two modality-dependent cross-attention mechanisms over the multi-source (image, text). The image encoder follows the method described in Section 3.2.1.

Encoder

To represent source sentences, an input embedding layer acts as a lookup table to map each word to a vector representation. Because the encoder in the transformer has no recurrence like that in RNN, it is necessary to inject positional information into the input embeddings, which is done using positional encoding.

The encoder comprises a stack of N identical layers. Each layer has self-attention and feed-forward sublayers. The self-attention sub-layer is a multi-head attention mechanism that allows the model to jointly attend to information from different representation subspaces. The feed-forward sub-layer is a basic, position-wise, fully connected feed-forward network, which is applied to each position separately and identically.

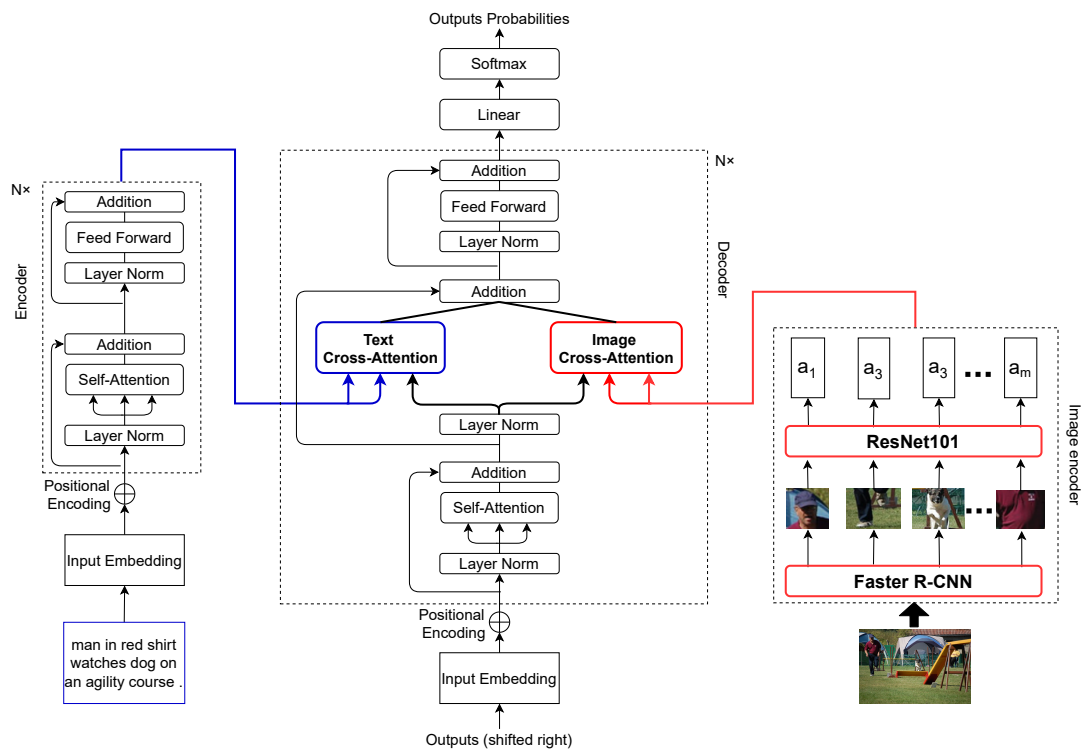


Figure 3.3: RA-TRANS: Region-Attentive Multimodal Transformer.

In addition to the two sub-layers described above, the residual connection [He et al., 2016] and layer normalization [Ba et al., 2016] are also key components of the transformer. There is a residual connection around every one of the two sublayers and a layer normalization inside the residual connection in my model. Therefore, the output of each sublayer is defined as $(x + \text{Sublayer}(\text{LayerNorm}(x)))$, where $\text{Sublayer}()$ is the function implemented by the sublayer itself. To encourage these residual connections, all sublayers and embedding layers produce outputs of dimension d_{model} .

Decoder

The decoder comprises a stack of N identical layers. In addition to the two sub-layers similar to the encoder, the decoder inserts two cross-attention mechanisms between them. One is text cross-attention, which performs multi-head attention on encoder output features. The other is image cross-attention, which performs multi-head attention over semantic image region features. There is also a residual connection around every sublayer and a layer normalization inside the residual connection, similar to the encoder.

When generating a target word at a time step t , the attention from one of the sources may be strong or weak from the other, and thus, summing two cross-attention outputs would help learn the better translation. Therefore, the summarized output from two cross-attentions is fed into the feed-forward network sub-layer, which consists of two linear transformations with a ReLU activation in between:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

While the linear transformations are the same across different positions, they use different parameters from layer to layer, where W_1 , W_2 , b_1 , and b_2 are trainable parameters. In this equation, the dimensions of the input and output are d_{model} , and the inner feed-forward neural network layer has dimensions d_{ff} . Finally, the decoder is capped off with a linear layer that acts as a classifier and a softmax layer to obtain the target word probabilities.

Double Cross-Attentions

As illustrated on the left of Figures 3.4 and 3.5, conventional cross-attention in the transformer acts as a query mapping of key-value sets to an output, which is multi-head

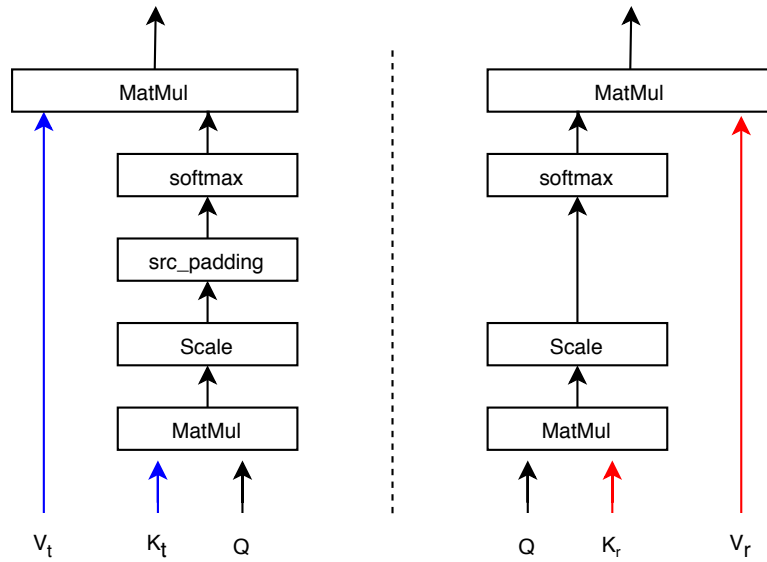


Figure 3.4: Left: Text Scaled Dot-Product Attention; Right: Image Scaled Dot-Product Attention.

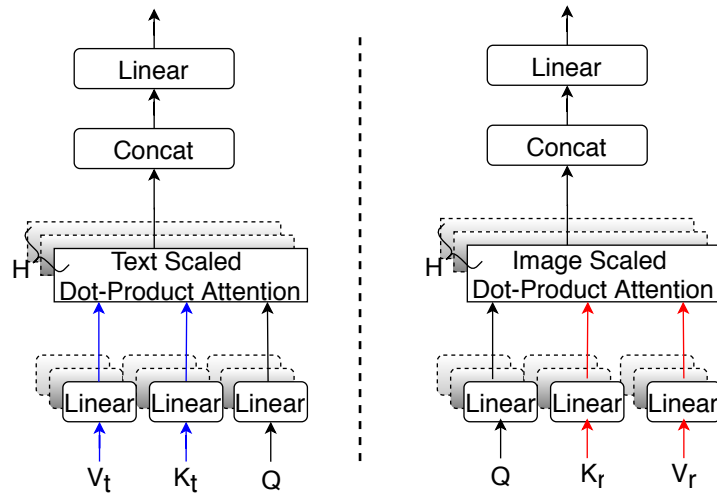


Figure 3.5: Left: Text Multi-Head Attention; Right: Image Multi-Head Attention.

attention that performs the attention function on the encoder output features using H heads in parallel. Each scaled dot-product attention process is called one head. Each head produces an output vector that is concatenated into a single vector before passing through the final linear layer.

The input involves queries and keys of dimension d_k and values of dimension d_v . Each query is multiplied with all keys by dot product multiplication and scaled by $\sqrt{d_k}$; then, there is an src_padding on padding source text input into the maximum length. Finally, the softmax function is applied to obtain the weights of the values. The final output of the scaled dot-product attention is computed as the weighted sum of the values. The weight assigned to each value is calculated using the compatibility function of the query with the corresponding key.

The cross-attention is simultaneously calculated on a set of queries, keys, and values and packed together into a matrix Q, K_t, V_t . The output matrix is computed as follows:

$$\text{Attention}(Q, K_t, V_t) = \text{softmax}\left(\frac{QK_t^T}{\sqrt{d_k}}\right)V_t$$

$$\text{MultiHead}(Q, K_t, V_t) = \text{Concat}(\text{head}_t^1, \dots, \text{head}_t^H)W^O$$

where $\text{head}_t^{i \in [1, H]} = \text{Attention}(QW_i^Q, K_tW_i^K, V_tW_i^V)$.

The projections are parameter matrices:

$$W_i^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}$$

$$W_i^K \in \mathbb{R}^{d_{\text{model}} \times d_k}$$

$$W_i^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$$

$$W^O \in \mathbb{R}^{Hd_v \times d_{\text{model}}}$$

In the proposed RA-TRANS, I use the conventional cross-attention in the transformer as the text cross-attention mechanism. I implemented an additional image cross-attention mechanism, which is multi-head attention, that performs the attention function on m semantic image region features using H heads in parallel.

The image cross-attention mechanism is illustrated on the right side of Figures 3.4 and 3.5. Unlike text-scaled dot-product attention, image-scaled dot-product attention has no source input padding because the number of semantic image regions is fixed.

The image cross-attention mechanism is defined as:

$$\text{Attention}(Q, K_r, V_r) = \text{softmax}\left(\frac{QK_r^T}{\sqrt{d_k}}\right)V_r$$

$$\text{MultiHead}(Q, K_r, V_r) = \text{Concat}(\text{head}_r^1, \dots, \text{head}_r^H)W^o$$

where $\text{head}_r^{i \in [1, H]} = \text{Attention}(QW_i^q, K_rW_i^k, V_rW_i^v)$.

The projections are parameter matrices:

$$W_i^q \in \mathbb{R}^{d_{\text{model}} \times d_k}$$

$$W_i^k \in \mathbb{R}^{d_{\text{model}} \times d_k}$$

$$W_i^v \in \mathbb{R}^{d_{\text{model}} \times d_v}$$

$$W^o \in \mathbb{R}^{Hd_v \times d_{\text{model}}}$$

3.3 Experiments

3.3.1 Dataset

I experimented on English→German (En→De) and English→French (En→Fr) tasks of the Multi30k dataset [Elliott et al., 2016]. The dataset contained 29k training images and 1,014 validation images. For testing, I used the 2016 test set, which included 1,000 images. Each image was paired with its English descriptions as well as human translations of German and French. I used Moses [Koehn et al., 2007] toolkit* to normalize and tokenize all sentences. Then, I converted the space-separated tokens into sub-word units using the byte pair encoding (BPE) model [Sennrich et al., 2016].† With 10k merge operations, the resulting vocabulary sizes of each language pair were 5,202→7,065 tokens for En→De and 5,833→6,575 tokens for En→Fr. The number of tokens in the sentence was limited to a maximum of 100. I trained models to translate from English to German/French and report the evaluation of cased, tokenized sentences with punctuation.

*<https://github.com/moses-smt/mosesdecoder>

†<https://github.com/rsennrich/subword-nmt>

3.3.2 Evaluation Metrics

I evaluated the quality of translation according to the token-level BLEU [Papineni et al., 2002] and METEOR [Denkowski and Lavie, 2014] metrics.

I trained all models three times and calculated the BLEU and METEOR scores. Finally, I reported the average over three runs. Moreover, I reported the statistical significance of BLEU using bootstrap resampling [Koehn, 2004] over a merger of three test translation results. I defined the statistical significance test threshold as 0.05, and reported only when the p-value was less than the threshold.

3.3.3 Baselines

RNN.

I trained a text-only RNN model using the OpenNMT [Klein et al., 2017] toolkit[‡] as a baseline. The RNN was trained on En→De and En→Fr, wherein only the textual part of Multi30k was used. This architecture comprises a 2-layer bidirectional GRU encoder and a 2-layer cGRU decoder with an attention mechanism.

Grid-Attentive Multimodal RNN (GA-RNN).

I trained a GA-RNN [Calixto et al., 2017] model[§] as another baseline, which was extended from OpenNMT. This architecture comprises a 2-layer bidirectional GRU encoder and a 2-layer cGRU decoder with two attention mechanisms. I trained this model with 7×7 equally sized grid local visual features from each image extracted by a ResNet101 pre-trained on ImageNet. Each grid-based local visual feature was represented as a 2,048-dimension vector.

TRANS.

I trained a text-only Transformer model using the transformer’s settings in the OpenNMT toolkit as a baseline. The TRANS was also trained on only the textual part of Multi30k on En→De and En→Fr tasks.

[‡]<https://github.com/OpenNMT/OpenNMT-py>

[§]<https://github.com/iacercalixto/MultimodalNMT>

Grid-Attentive Multimodal Transformer (GA-TRANS).

I trained a GA-TRANS model based on the GA-RNN model by modifying the transformer’s settings in the OpenNMT toolkit as another baseline. An image cross-attention mechanism was implemented on the grid-based local visual features in the transformer’s settings. This architecture was also trained with 7×7 grid-based local visual features from each image extracted by a ResNet101 pre-trained on ImageNet, and each feature was represented as a 2,048-dimension vector.

3.3.4 Setup

I implemented my proposed RA-RNN and RA-TRANS based on GA-RNN and GA-TRANS baselines, respectively, by modifying the image attention mechanism to focus on m semantic image region feature vectors generated from the image encoder. For the image encoder in both the RA-RNN and RA-TRANS methods, the number of semantic image region features was set to $m = 100$ and the dimension of regional feature vectors was set to $d_r = 2,048$.

Settings of the RNN-Based Models

I set the hidden state dimension of the bi-directional GRU encoder and cGRU decoder to 500, source word embedding dimension to 500, sentence-minibatches to 40, beam size to 5, text dropout to 0.3, and image region dropout to 0.5. I trained the model using stochastic gradient descent with ADAM [Kingma and Ba, 2015] and a learning rate of 0.002 for 25 epochs. Finally, after both the validation perplexity and accuracy converged, the model with the highest BLEU score of the validation set was selected to evaluate the test set.

Settings of the Transformer-Based Models

I set $N = 6$ layers for the encoder and decoder. The number of dimensions of all the input and output layers was set to $d_{\text{model}} = 512$. The inner feed-forward neural network layer was set to $d_{\text{ff}} = 2,048$. The heads of all the multi-head modules were set to $H = 8$ in both the encoder and decoder layers. I applied linear projection on visual features to reduce the dimensions from 2,048 to 512 to have the same size as

word embeddings. I applied a dropout of 0.3 on linear projection. During training, the sentence-minibatches were set to 40, the value of label smoothing was set to 0.1, and the attention dropout and residual dropout were 0.3. An Adam optimizer was used to tune the model parameters. The learning rate was set to two with a warm-up step of 8,000. I trained the model up to 100 epochs, and the model with the highest BLEU score of the validation set was selected to evaluate the test set.

3.4 Results

3.4.1 Results within RNN-Based Models

Table 3.1 presents the experimental results of RNN-based architectures, showing that the proposed RA-RNN achieves better performance than both the text-only RNN baseline and the GA-RNN baseline in all translation tasks. In particular, the results of the RA-RNN are significantly better than those of the text-only RNN baseline with a p-value of < 0.05 on both language pairs. This illustrates that integrating semantic image region visual features is capable of promoting translation performance, and my proposed method can make better use of visual information.

Methods	En→De		En→Fr	
	BLEU	METEOR	BLEU	METEOR
RNN	34.8	53.4	56.5	71.9
GA-RNN	36.5	54.8	57.8	72.8
<i>RA-RNN</i>	36.9[†]	55.5	58.1[†]	73.2
v.s. RNN	(↑ 2.1)	(↑ 2.1)	(↑ 1.6)	(↑ 1.3)
v.s. GA-RNN	(↑ 0.4)	(↑ 0.7)	(↑ 0.3)	(↑ 0.4)

Table 3.1: The experimental results of RNN-based architectures. The best performance is highlighted in bold. † indicates that the result is significantly better than the text-only RNN baseline at a p-value of < 0.05 .

3.4.2 Results within Transformer-Based Models

Table 3.2 presents the experimental results of the Transformer-based architectures, showing that the proposed RA-TRANS outperforms the baselines on both the En→De and En→Fr tasks.

Methods	En→De		En→Fr	
	BLEU	METEOR	BLEU	METEOR
TRANS	35.4	52.8	57.4	72.2
GA-TRANS	37.5	55.6	59.5	74.4
<i>RA-TRANS</i>	38.0^{†‡}	56.0	60.1^{†‡}	74.8
v.s. TRANS	(↑ 2.6)	(↑ 3.2)	(↑ 2.7)	(↑ 2.6)
v.s. GA-TRANS	(↑ 0.5)	(↑ 0.4)	(↑ 0.6)	(↑ 0.4)

Table 3.2: The experimental results of Transformer-based architectures. The best performance is highlighted in bold. † and ‡ indicate that the result is significantly better than the text-only TRANS and GA-TRANS baselines at p-value < 0.05, respectively.

It is worth noting that my RA-TRANS results are significantly better than not only the text-only TRANS baseline but also the GA-TRANS baseline with a p-value of < 0.05 on both tasks. This demonstrates that the proposed method is universal, which can result in consistent improvements in performance on different NMT architectures. Thus, I confirm the effectiveness and generality of the proposed method.

3.4.3 Comparison of Proposed Model and Existing Ones

To further verify the merit of the proposed method, I also implemented the proposed method on the state-of-the-art text-only NMT baseline mentioned in [Calixto et al., 2019] and the state-of-the-art transformer baseline mentioned in [Yin et al., 2020], respectively. Furthermore, I compared the experimental results of my proposed method with the following state-of-the-art MNMT methods:

Parallel RCNNs [Huang et al., 2016]: The encoder of RNN is composed of multiple encoding threads. In each thread, a regional visual feature is followed by a text

sequence.

NMT_{SRC+IMG} [Calixto et al., 2017]: Integrates two separate attention mechanisms over the source words and conventional grid local visual features in a cGRU decoder.

IMG_D [Calixto and Liu, 2017]: Integrates global visual features as additional data to initialize the decoder’s hidden state.

Imagination [Elliott and Kádár, 2017]: Jointly learns a translation model and visually grounded representations.

{Soft, Stochastic} Attention + Grounded Image (GI) [Delbrouck and Dupont, 2017]: Employs two kinds of attention mechanisms, which are superimposed by an additional grounding attention method, for considering visual annotations of image feature maps to generate context vectors.

VMMT_F [Calixto et al., 2019]: An MNMT model that incorporates image context through a latent variable model.

Del+Obj [Ive et al., 2019]: A transformer-based deliberation model enriched with object-level features.

MTF [Yao and Wan, 2020]: A transformer-based NMT model with multimodal self-attention to integrate text and image features.

GMFE-NMT [Yin et al., 2020]: A transformer-based NMT model integrated with a multimodal graph neural network (GNN) encoder on the grounding-based correspondences between phrase-level words and regions.

ImagiT [Long et al., 2021]: An MNMT method via visual imagination.

As shown in Table 3.3, all the existing methods are divided into two groups: RNN-based methods and Transformer-based methods. Then, I display the experimental results of the proposed method and the state-of-the-art methods’ results for the respective group. Note that previous methods mainly report the results on the En→De language pair of the Multi30k 2016 test set, and hence, the existing results on the En→Fr task are fewer than those of the En→De task.

[¶]The results of my proposal reported here are implemented on the state-of-the-art text-only NMT baseline mentioned in [Calixto et al., 2019] using the OpenNMT toolkit. The experimental settings are consistent with the setup described in Section 3.3.4.

^{¶¶}The results of my proposal reported here are implemented on the state-of-the-art transformer baseline mentioned in [Yin et al., 2020] using the Nmtpytorch toolkit [Caglayan et al., 2017b]. The experimental settings are consistent with the setup in Section 3.3.4, except that the learning rate was tuned to 0.03 and the model was trained up to 300 epochs.

Methods	En→De		En→Fr	
	BLEU	METEOR	BLEU	METEOR
<i>RNN-Based Methods</i>				
Text-only NMT	35.0	54.9	56.5	71.9
Parallel RCNNs	36.5	54.1	N/A	N/A
NMT _{SRC+IMG}	36.5	55.0	57.8	72.8
IMG _D	37.3	55.1	N/A	N/A
Imagination	36.8	55.8	N/A	N/A
Soft Attention + GI	37.6	55.3	N/A	N/A
Stochastic Attention + GI	38.2	55.4	N/A	N/A
VMMT _F	37.7	56.0	N/A	N/A
<i>Proposed Method</i> _{(OpenNMT)[¶]}	36.9	55.5	58.1	73.2
<i>Transformer-Based Methods</i>				
Text-only transformer	38.4	56.5	59.5	73.7
Del+Obj	38.0	55.6	59.8	74.4
MTF	38.7	55.7	N/A	N/A
GMFE-NMT	39.8	57.6	60.9	74.9
ImagiT	38.5	55.7	59.7	74.0
<i>Proposed Method</i> _(Nmtpytorch)	38.6	57.7	60.1	75.0

Table 3.3: Comparison with existing methods. Among all the results, I highlight the best performance in bold. All the experimental results of my proposal are the average scores over three runs.

By comparing the performance of the proposed method with the state-of-the-art methods, I draw two conclusions as follows:

First, the proposed method outperforms the state-of-the-art text-only baselines on different basic neural architectures. For instance, the proposed method in the respective group outperforms the text-only NMT baseline and the transformer baseline by 1.6 and 0.6 BLEU scores, respectively, on the En→Fr task. Therefore, I can confirm the effectiveness and generality of the proposed method.

Second, the evaluation results of the proposed method outperform most of the existing MNMT methods. Among the Transformer-based methods, my proposed method achieves the best performance evaluated by the METEOR score on both language pairs; furthermore, the results of the proposed method outperform most of the METEOR scores in the existing RNN-based methods on different language pairs as well. This demonstrates that my proposed method is competitive among all the state-of-the-art MNMT methods.

3.5 Analyses

3.5.1 Pairwise Evaluation

To further analyze the translation performance of my proposed method, I performed a pairwise evaluation and statistical analysis. The results of the pairwise evaluation of the En→Fr language pair are summarized in Table 3.4.

Based on two kinds of NMT architectures, I conducted three groups of comparisons. Specifically, I compared the proposed RA-RNN/RA-TRANS translations with their corresponding baselines' translations to identify improvement or deterioration of translation performance, and I compared the translations of RA-RNN and RA-TRANS to identify which architecture can achieve better translation performance. For each group, I randomly selected 50 examples for evaluation and categorized 50 investigated examples into various categories by counting the number and proportion.

After statistical analysis, I find that almost half of the investigated examples show that my RA-RNN performs better than at least one baseline model. Similarly, half of the investigated examples show that my RA-TRANS outperforms at least one baseline model. It further verified the effectiveness and generality of my proposed method.

<i>RA-RNN v.s. RNN-based baselines</i>		
Better than both baselines	8	(16%)
Better than GA-RNN baseline	6	(12%)
Better than RNN baseline	10	(20%)
No change	24	(48%)
Deteriorated	2	(4%)
<i>RA-TRANS v.s. Transformer-based baselines</i>		
Better than both baselines	10	(20%)
Better than GA-TRANS baseline	4	(8%)
Better than TRANS baseline	11	(22%)
No change	24	(48%)
Deteriorated	1	(2%)
<i>RA-TRANS v.s. RA-RNN</i>		
RA-TRANS is better than RA-RNN	8	(16%)
RA-RNN is better than RA-TRANS	2	(4%)
No change	40	(80%)

Table 3.4: Pairwise evaluation. I counted the number and proportion of various categories among 50 random examples.

Moreover, the number of examples in which my RA-TRANS is better than both baselines is slightly improved compared with a similar case of the RA-RNN. By comparing the translation performance of my RA-TRANS and RA-RNN, I find that the number of examples where RA-TRANS is better than RA-RNN is four times larger than the opposite cases. This illustrates that my RA-TRANS can achieve a better translation performance compared with RA-RNN.

3.5.2 Qualitative Analysis

For qualitative analysis, I analyze translation performance by comparing the translation results of the proposed method and its baselines, along with visualizing the semantic image regions that are attended by the image-attention mechanism at every time step.

According to the attention weight assigned to each region, the semantic image regions are shown with deep or shallow transparency in the image at every time step. As the weight increases, the image region becomes more transparent. Considering the number of 100 bounding boxes in one image and the overlapping areas, I visualized the top five weighted semantic image regions. In the image, a blue bounding box indicates the most weighted image region, and the red text along with the bounding box shows the target word generated at that time step. Then, I analyze whether the semantic image regions have a positive or negative effect at the time step when a target word is generated.

To distinguish the translation quality, I highlight the better translation with blue and the worse translation with red.

Visualization within RNN-Based Models

In Figure 3.6, I present two examples to analyze the effect of semantic image regions on translation quality within RNN-based models. The first is an example of a positive effect, whereas the second is the opposite.

For the first example, it illustrates that the semantic image regions of the proposed method can play a positive role in providing object attributes.

In detail, by comparing the translation result of my RA-RNN and its baselines, I find that the RA-RNN translates “striped beach chairs” better, which is a phrase made up of an adjective and a noun. From the visualization of the most weighted semantic


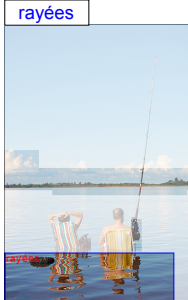


		<p>Src (En) two people are sitting fishing on <u>striped beach chairs</u> in a body of water .</p> <p>Ref (Fr) deux personnes sont assises dans <u>des fauteuils de plage rayés</u> , pêchant dans une étendue d's eau .</p> <p>RNN deux personnes sont assises sur une structure de plage rayée (a striped beach structure) dans un plan d's eau .</p> <p>GA-RNN deux personnes sont assises , pêchent sur une plage de sable (a sandy beach) dans un plan d's eau .</p> <p>RA-RNN deux personnes sont assises à pêcher sur des chaises rayées (striped chairs) dans un plan d's eau .</p>
		<p>Src (En) men playing volleyball , with one player missing the ball but hands still <u>in the air</u> .</p> <p>Ref (Fr) des hommes jouant au volleyball , avec un joueur ratant le ballon mais avec les mains toujours <u>en l's air</u> .</p> <p>RNN des hommes jouant au volleyball , un joueur à l's attraper , mais les autres mains ayant toujours dans les airs (in the air) .</p> <p>GA-RNN des hommes jouant au volley-ball , avec un joueur qui le regarde dans les airs (in the air) .</p> <p>RA-RNN des hommes jouant au volleyball , avec un joueur qui passer le ballon mais les mains du vol (of the flight) .</p>

Figure 3.6: Examples for text-only RNN, grid-attentive multimodal RNN (GA-RNN), and region-attentive multimodal RNN (RA-RNN). Red and blue words indicate incorrect and correct, respectively.

image region, I can identify the semantics of “chairs” and “striped,” respectively.

For the second example, it presents that attending to the semantic image regions that are not related to the text’s semantics is not helpful for translation performance.

As shown in the example, “air” is correctly translated by baselines. However, the RA-RNN translates “in the air” into “du vol (of the flight).” I observe that the transparent semantic image regions with the top five weights in the image are scattered and unconnected. I can not understand any semantic information in the visualized image regions. I speculate that the word “air” is challenging to interpret depending on visual features. Furthermore, the proposed method translates it into “vol (flight),” which is close to another meaning of the polysemous “air,” not completely different from the original meaning.

Visualization within Transformer-Based Models

In Figure 3.7, I present two examples to analyze the effect of semantic image regions on translation quality within Transformer-based models. The first is an example of a positive effect, whereas the second is the opposite.


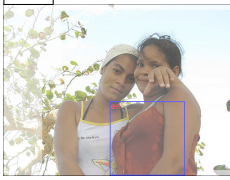
manipule		<p>Src (En) the woman in blue is <u>operating</u> a camera in front of two other women .</p> <p>Ref (Fr) la femme en bleu <u>manipule</u> un appareil photo devant deux autres femmes .</p> <p>TRANS la femme en bleu manie (wields) une caméra en face de deux autres femmes .</p> <p>GA-TRANS la femme en bleu fait fonctionner (function) un appareil photo devant deux autres femmes .</p> <p>RA-TRANS la femme en bleu manipule (manipulate) un appareil photo devant deux autres femmes .</p>
en		<p>Src (En) two women <u>wearing</u> tank tops are looking at the camera .</p> <p>Ref (Fr) deux femmes <u>portant</u> des débardeurs regardent l's objectif .</p> <p>TRANS deux femmes vêtues (wearing) de débardeurs regardent l's objectif .</p> <p>GA-TRANS deux femmes portant (wearing) des débardeurs regardent l's objectif .</p> <p>RA-TRANS deux femmes en (in) débardeurs regardent l's objectif .</p>

Figure 3.7: Examples for text-only Transformer (TRANS), grid-attentive multimodal Transformer (GA-TRANS), and region-attentive multimodal Transformer (RA-TRANS). Red and blue words indicate incorrect and correct, respectively.

For the first example, it shows that the semantic image regions of the proposed method can play a positive role in providing verb attributes.

In this example, compared with baselines’ translations, the RA-TRANS translates “operating” better, which is a verb. By visualizing the most weighted semantic image region, I can identify the semantics of “operate.”

For the second example, I find that the semantic image regions of the proposed method have no effect on distinguishing synonyms.

As illustrated in the example, “wearing” is correctly translated by baselines. However, the RA-RNN translates the verb into “in,” which is a preposition. Although I can identify the semantic of “wearing” from the most weighted semantic image region, it can also be understood as “in.”

Visualization between RA-RNN and RA-TRANS

In Figure 3.8, I present two examples to analyze the effect of semantic image regions on translation quality between RA-RNN and RA-TRANS architectures. The first is an example of a case where RA-TRANS is better than RA-RNN, whereas the second is the opposite case.

For the first example, it reflects that the performance of the image attention mech-





		<p>Src (En) a construction worker is driving heavy <u>equipment</u> at a work site .</p> <p>Ref (Fr) un ouvrier du bâtiment conduit un gros <u>engin</u> sur un chantier .</p> <p>RA-RNN un ouvrier du bâtiment conduit un gros plan (plan) d's chantier .</p> <p>RA-TRANS un ouvrier du bâtiment conduit un gros engin (machine) sur un chantier .</p>
		<p>Src (En) a father-figure and two children outside their home doing <u>yard work</u> such as using a hoe on the grass and planting a tree .</p> <p>Ref (Fr) une figure paternelle et deux enfants devant leur maison , faisant des <u>activités de jardinage</u> comme utiliser une binette dans l's herbe et planter un arbre .</p> <p>RA-RNN un père et deux enfants à l's extérieur de chez leur maison , faisant du jardinage (gardening) tandis qu's ils utilisent une binette sur l's herbe et plantant un arbre .</p> <p>RA-TRANS un mannequin et deux enfants devant leur maison , faisant des travaux (work) de travail en utilisant une binette tandis qu's ils utilisant une binette et un arbre .</p>

Figure 3.8: Examples for region-attentive multimodal RNN (RA-RNN) and region-attentive multimodal Transformer (RA-TRANS). Red and blue words indicate incorrect and correct, respectively. In each example, left/right figures correspond to the visualization of RA-RNN/RA-TRANS.

anism is also crucial to the translation quality of the proposed method. In another word, the semantic image region features and the effectiveness of the image attention mechanism are indispensable.

As shown in this example, the RA-TRANS translates “equipment” better than RA-RNN. From the top five weighted semantic image regions, I can identify the semantics of “equipment,” either in RA-RNN or RA-TRANS. However, as the most weighted semantic image region by the image attention mechanism in RA-RNN does not provide any relevant semantic information to the text’s semantics, it eventually leads to worse translation.

For the second example, it demonstrates that the improvement in translation performance benefits from attending to the specific semantic image region features.

As shown in this example, the RA-RNN translates “yard work” better than RA-TRANS. I find that the RA-RNN focuses on a potted plant in a small garden from the most weighted semantic image region, however, the RA-TRANS focuses on a boy’s work activities. Moreover, I observe that the top five weighted semantic image regions on which the two architectures focus are quite different. The RA-RNN mainly focuses on the garden, whereas the RA-TRANS focuses on the action.

3.6 Summary

In this proposal, I proposed a multimodal NMT method, namely, RA-NMT, with semantic image regions. The proposed method was implemented on two types of NMT architectures: RNN and Transformer. Experimental results showed that the proposed method achieved a significant improvement above the text-only NMT baseline and grid-attentive multimodal NMT baseline based on either of the two neural architectures. In addition, the proposed method implemented on the state-of-the-art NMT baselines can not only achieve better performance than the baselines but can also outperform most of the existing MNMT methods, which verifies its effectiveness and competitiveness. Further analysis demonstrated that the proposed method effectively improves translation performance, and the improvement benefits from attending to specific semantic image region features, leading to better use of visual information.

4 WRA-Guided MNMT

4.1 Introduction

An overview of my proposal is shown in Figure 4.1. This study proposes a novel facility named word-region alignment (WRA) that explicitly correlates source words with image regions as an additional input in the proposed MNMT model.

Unlike existing MNMT models, I design the WRA as an intermediate component to bridge multimodal inputs. Specifically, as visual concepts can summarize the semantics of image regions, I utilize these visual concepts as a medium to pre-process the semantic relevance between source words and image regions. In terms of architecture, I propose a novel integration strategy word-to-region (W2R) that leverages the WRA, facilitating the interaction between semantically relevant textual and visual features. Under the integration strategy W2R, the pre-processed WRA is leveraged as a bridge to link textual and visual inputs, acting as an auxiliary cue to guide textual features to interact with semantically relevant regional visual features.

Additionally, two modality-dependent attention mechanisms are utilized to generate textual and visual contexts for decoding target words. By advancing the correlation between textual and visual modalities by integrating WRA, the textual and visual context can provide semantically relevant information to generate accurate translations.

Overall, the main contributions are as follows:

- I proposed WRA, an intermediate component as an additional input to bridge multimodal inputs based on semantic relevance.
- I proposed a novel integration strategy W2R of the MNMT model that leverages the WRA to guide the model to translate certain source words into target words while attending to semantically relevant image regions.

- I implemented my proposal on both RNN-based and Transformer-based architectures and evaluated it on English–German and English–French tasks using the Multi30k dataset [Elliott et al., 2016]. Extensive experiments validated the consistent efficacy of the proposed method and revealed that it significantly improved baselines based on different evaluation metrics and outperformed most of the existing methods.
- I conducted detailed analyses to prove the effectiveness of the proposed method and demonstrate that my method can lead to better visual information use.

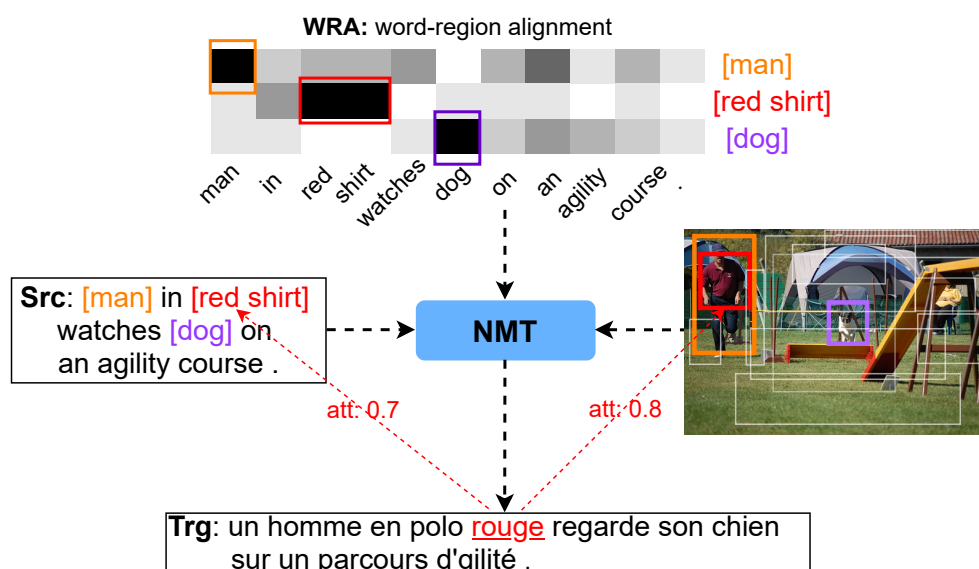


Figure 4.1: Example of WRA-guided MNMT. The WRA builds semantic relevance between the vision and language. Specifically, each region-level visual feature is annotated using a visual concept that is used to create a relationship with every source word. When generating the “rouge,” similar attention weights (“att” in the figure) are assigned to both the corresponding source word “red” and image region “red shirt.”

4.2 Methodology

In this section, I describe my methodology as follows: (1) I introduce the proposed WRA in Section 4.2.1, according to Figure 4.2; (2) The details of the integration of WRA into the RNN-based MNMT model shown in Figure 4.3 and the Transformer-based MNMT model shown in Figure 4.4 are presented in Section 4.2.2 and Section 4.2.3, respectively.

4.2.1 WRA Generation

As shown in Figure 4.2, I propose a WRA containing explicit semantic interactions between the source words and image regions. The WRA is pre-processed; it acts as an auxiliary input to guide interactions between the textual and visual information inside the entire model.

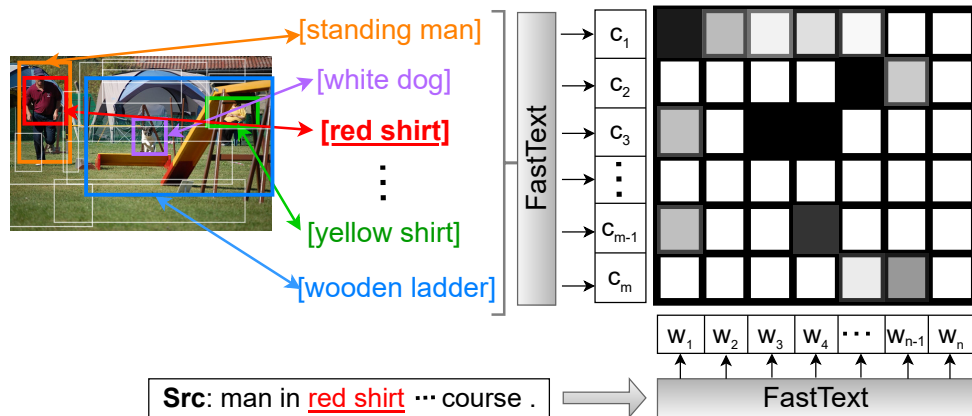


Figure 4.2: WRA generation. Each region-level visual feature was annotated using a visual concept consisting of an attribute class, followed by an object class. Subsequently, it was used to create a relationship with each source word based on semantic similarity. The WRA represents the semantic correlation between each regional visual feature and all words in a sentence.

For regions, I use the object detection method in [Anderson et al., 2018] to extract object-level regions for each image. Specifically, each image region is not only denoted by a bounding box in the image but also detected along with a visual concept consisting of an attribute class followed by an object class (see Figure 4.2). I extract m

image regions along with visual concepts for each image that are used to annotate the semantics of the corresponding regions. Then, I convert the source words and visual concepts into subword units.

I identify two types of explicit WRA:

Soft WRA

The soft WRA is generated as a cosine similarity matrix that presents the semantic similarity score between source words and image regions.

To calculate cosine similarity scores between the source words and image regions, first, I utilize fastText* to learn subword representations of the source words and visual concepts. I use a pre-trained model† containing two million word vectors trained on subword information on Common Crawl (600B tokens). The subword embeddings of source words can be generated directly, whereas the subword embeddings of visual concepts should take an average of all the constituent subwords because they are phrases. Then, these embeddings provide a mapping function from a subword to a 300-dim vector, where semantically similar subwords are embedded close to each other. As shown in Figure 4.2, the source words are represented by $W = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \dots, \mathbf{w}_n\}$ and visual concepts are represented by $C = \{\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \dots, \mathbf{c}_m\}$, where n denotes the source length and m denotes the region amount.

I define the soft WRA by computing the similarity score between the source words and visual concepts as follows:

$$g_{i,j} = \frac{\mathbf{c}_i^T \cdot \mathbf{w}_j}{\|\mathbf{c}_i\| \cdot \|\mathbf{w}_j\|}, i \in [1, m], j \in [1, n]$$

Here, $g_{i,j}$ represents the similarity score between the i -th region and the j -th word.

Finally, I define the semantic relevance between the i -th region and the whole source sentence using a similarity vector \mathbf{g}_i . Then, the soft WRA is represented as $G_{sa} = \{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3, \dots, \mathbf{g}_m\}$.

Hard WRA

The hard WRA is generated as a maximum similarity matrix along the source word axis based on the soft WRA.

*<https://github.com/facebookresearch/fastText>

†<https://fasttext.cc/docs/en/english-vectors.html>

I make a hard choice to pair each region with only one word in the whole sentence by aligning the most semantically relevant words to each region based on similarity score $g_{i,j}$:

$$g'_{i,j} = \begin{cases} 1, & \text{if } \arg \max_{j \in [1,n]}(\mathbf{g}_i) = j, \\ 0, & \text{otherwise} \end{cases}$$

Here, $g'_{i,j}$ represents the replacement of the maximum similarity score between the i -th region and the j -th word with 1, and the replacement of others with 0.

Finally, the hard WRA can be represented using m one-hot vectors by $G_{\text{ha}} = \{\mathbf{g}'_1, \mathbf{g}'_2, \mathbf{g}'_3, \dots, \mathbf{g}'_m\}$.

4.2.2 WRA-Guided RNN-Based MNMT Model

As illustrated in Figure 4.3, based on text-to-text RNN architecture [Bahdanau et al., 2015], my WRA-guided RNN-based MNMT model comprises four parts: textual encoder, visual encoder, word-to-region (W2R), and decoder. The most unique stage is the W2R, where the soft/hard WRA are integrated to guide interactions between textual and visual representations.

Textual Encoder

The textual encoder is a bi-directional RNN with a gated recurrent unit (GRU) [Cho et al., 2014a]. Given a source sentence $X = (x_1, x_2, x_3, \dots, x_n)$, the encoder updates the forward hidden states with annotations $(\vec{h}_1, \vec{h}_2, \vec{h}_3, \dots, \vec{h}_n) \in \mathbb{R}^{d_s}$ and updates the backward with annotations $(\overleftarrow{h}_1, \overleftarrow{h}_2, \overleftarrow{h}_3, \dots, \overleftarrow{h}_n) \in \mathbb{R}^{d_s}$. By concatenating the forward and backward annotations, the textual representation is denoted as $H = (\mathbf{h}_1, \mathbf{h}_2, \mathbf{h}_3, \dots, \mathbf{h}_n) \in \mathbb{R}^{d_h}$.

Visual Encoder

The visual encoder is an object-detection-based approach [Anderson et al., 2018] to regional feature extraction. Given an input image, the visual encoder employs the faster R-CNN [Ren et al., 2015] in conjunction with ResNet-101 [He et al., 2016] as its backbone, which is pre-trained on the Visual Genome [Krishna et al., 2017] dataset to extract m regional visual features from each image. Each regional feature is represented

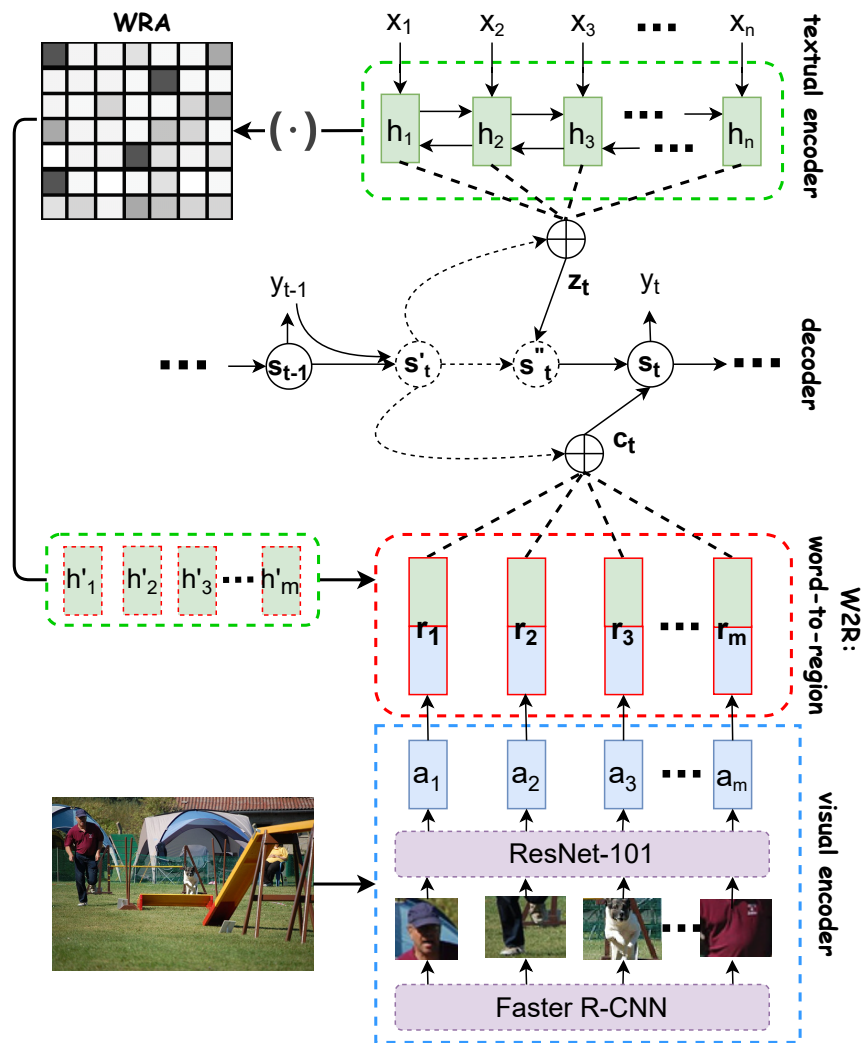


Figure 4.3: WRA-Guided RNN-Based MNMT Model.

as a vector \mathbf{a} and the visual representation is denoted by $A = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \dots, \mathbf{a}_m) \in \mathbb{R}^{d_r}$.

Word-to-Region (W2R)

After generating textual and visual representations independently, the WRA is integrated as an additional input to bridge them, acting as an auxiliary cue to guide source words to interact with regional features.

The formulation, which entails two stages, is illustrated in Figure 4.3.

In the first stage, intermediate textual representations $H' = (\mathbf{h}'_1, \mathbf{h}'_2, \mathbf{h}'_3, \dots, \mathbf{h}'_m) \in \mathbb{R}^{d_r}$ are associated with each image region under the guidance of the WRA. Subsequently, two definitions are used to calculate the WRA-guided textual feature \mathbf{h}'_i , with respect to the i -th region.

- Under the guidance of the soft WRA:

$$\mathbf{h}'_i = T\left(\frac{1}{n} \odot (\mathbf{g}_i \cdot H)\right)$$

- Under the guidance of the hard WRA:

$$\mathbf{h}'_i = T(\mathbf{g}'_i \cdot H)$$

where \mathbf{g}_i is from the G_{sa} ; \mathbf{g}'_i is from the G_{ha} ; H is the textual representation; n is the source length; and a linear transformation function is defined by $T : \mathbb{R}^{d_h} \rightarrow \mathbb{R}^{d_r}$.

In the second stage, the WRA-guided textual representations $H' = (\mathbf{h}'_1, \mathbf{h}'_2, \mathbf{h}'_3, \dots, \mathbf{h}'_m) \in \mathbb{R}^{d_r}$ are combined with the visual representations $A = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \dots, \mathbf{a}_m) \in \mathbb{R}^{d_r}$ through concatenation, to enrich each image region using semantically relevant textual features. Through this transformation, the interaction between the independently represented textual and visual features is effectively facilitated.

$$\mathbf{r}_i = \text{CONCAT}(\mathbf{h}'_i, \mathbf{a}_i) = \begin{bmatrix} \mathbf{h}'_i \\ \mathbf{a}_i \end{bmatrix} \quad (4.1)$$

Thus, the visual representations are semantically enhanced by combining WRA-guided textual features and advanced to multimodal representations, denoted as $R = (\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \dots, \mathbf{r}_m) \in \mathbb{R}^{d_r}$.

Doubly Attentive Decoder

The doubly attentive decoder inspired by [Calixto et al., 2017] functions in the textual and multimodal contexts based on two independent attention mechanisms, then predicts the probability of a target word. It is a deepGRU [Delbrouck and Dupont, 2018] with three stacked GRUs derived from the convoluted gated recurrent units (cGRUs).[‡]

To generate the target word y_t at time step t , a hidden state proposal \mathbf{s}'_t is computed in the first GRU (f_{gru_1}) using the previous target word y_{t-1} and hidden state \mathbf{s}_{t-1} as follows:

$$\begin{aligned}\mathbf{s}'_t &= f_{\text{gru}_1}(y_{t-1}, \mathbf{s}_{t-1}) \\ \mathbf{s}'_t &= (1 - \hat{\xi}_t) \odot \dot{\mathbf{s}}_t + \hat{\xi}_t \odot \mathbf{s}_{t-1} \\ \dot{\mathbf{s}}_t &= \tanh(W E_Y[y_{t-1}] + \hat{\gamma}_t \odot (U \mathbf{s}_{t-1})) \\ \hat{\gamma}_t &= \sigma(W_\gamma E_Y[y_{t-1}] + U_\gamma \mathbf{s}_{t-1}) \\ \hat{\xi}_t &= \sigma(W_\xi E_Y[y_{t-1}] + U_\xi \mathbf{s}_{t-1})\end{aligned}$$

where $W_\xi, U_\xi, W_\gamma, U_\gamma, W$, and U are the training parameters and E_Y is the target word embedding.

Textual attention focuses on every textual representation \mathbf{h}_i in H by assigning an attention weight, following which the textual context vector \mathbf{z}_t is generated as follows:

$$\begin{aligned}e_{t,i}^{\text{txt}} &= (V^{\text{txt}})^T \tanh(U^{\text{txt}} \mathbf{s}'_t + W^{\text{txt}} \mathbf{h}_i) \\ \alpha_{t,i}^{\text{txt}} &= \text{softmax}(e_{t,i}^{\text{txt}}) \\ \mathbf{z}_t &= \sum_{i=1}^n \alpha_{t,i}^{\text{txt}} \mathbf{h}_i,\end{aligned}$$

where $V^{\text{txt}}, U^{\text{txt}}, W^{\text{txt}}$ are the training parameters; $e_{t,i}^{\text{txt}}$ is the attention energy; $\alpha_{t,i}^{\text{txt}}$ is the attention weight matrix.

Likewise, visual attention focuses on every multimodal representation \mathbf{r}_i in R by assigning an attention weight. Then, the multimodal context vector \mathbf{c}_t is generated as follows:

[‡]<https://github.com/nyu-dl/dl4mt-tutorial/blob/master/docs/cgru.pdf>

$$\begin{aligned}
e_{t,j}^{\text{img}} &= (V^{\text{img}})^T \tanh(U^{\text{img}} \mathbf{s}'_t + W^{\text{img}} \mathbf{r}_i) \\
\alpha_{t,j}^{\text{img}} &= \text{softmax}(e_{t,j}^{\text{img}}) \\
\mathbf{c}_t &= \sum_{j=1}^m \alpha_{t,j}^{\text{img}} \mathbf{r}_i,
\end{aligned}$$

where V^{img} , U^{img} , W^{img} are the training parameters; $\alpha_{t,j}^{\text{img}}$ is a weight matrix; $e_{t,j}^{\text{img}}$ is the attention energy.

Based on the textual context vector \mathbf{z}_t and previous hidden state proposal \mathbf{s}'_t , a hidden state proposal \mathbf{s}''_t is computed in the second GRU (f_{gru_2}). Similarly, based on the multimodal context vector \mathbf{c}_t and the \mathbf{s}''_t , the final hidden state \mathbf{s}_t is generated in the third GRU (f_{gru_3}). Because the calculation of f_{gru_2} and f_{gru_3} are similar to the function of f_{gru_1} , I organize them as follows:

$$\begin{aligned}
\mathbf{s}_t &= f_{\text{gru}_3}([\mathbf{c}_t, y_{t-1}], \mathbf{s}''_t) \\
\mathbf{s}''_t &= f_{\text{gru}_2}(\mathbf{z}_t, \mathbf{s}'_t)
\end{aligned}$$

I ensure that both representations have their own projections to compute the candidate probabilities by obtaining textual and visual GRU blocks as follows:

$$\begin{aligned}
\mathbf{b}_t^{\text{v}} &= f_{\text{ght}}(W_{\text{b}}^{\text{v}} \mathbf{s}_t) \\
\mathbf{b}_t^{\text{t}} &= f_{\text{ght}}(W_{\text{b}}^{\text{t}} \mathbf{s}''_t) \\
y_t \sim p_t &= \text{softmax}(W_{\text{proj}}^{\text{t}} \mathbf{b}_t^{\text{t}} + W_{\text{proj}}^{\text{v}} \mathbf{b}_t^{\text{v}}).
\end{aligned}$$

where W_{b}^{v} , W_{b}^{t} , $W_{\text{proj}}^{\text{t}}$, $W_{\text{proj}}^{\text{v}}$ are training parameters, and f_{ght} is a gated hyperbolic tangent activation [Teney et al., 2018] substituted for the tanh function.

4.2.3 WRA-Guided Transformer-Based MNMT Model

As illustrated in Figure 4.4, based on text-to-text Transformer architecture [Vaswani et al., 2017], my proposed model also comprises four parts: textual encoder, visual encoder, word-to-region (W2R), and decoder. The W2R is the core stage to leverage WRA.

Textual Encoder

In the Transformer [Vaswani et al., 2017], a source sentence is encoded by a textual encoder with multiple layers. The encoder is composed of a stack of N identical lay-

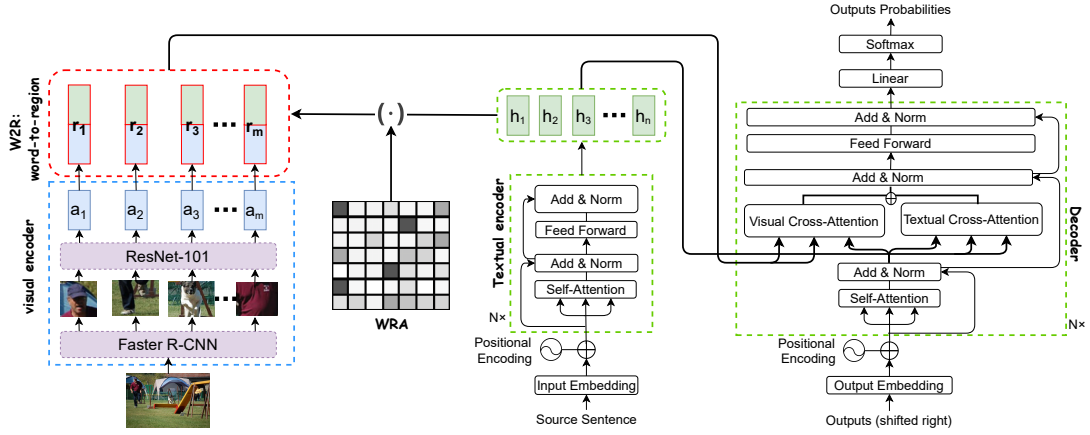


Figure 4.4: WRA-Guided Transformer-Based MNMT Model.

ers, each of which included two sublayers. The first and second sublayers are the multi-head attention and position-wise fully connected feed-forward network (FFN). Residual connection and layer normalization are used between sublayers. Formally, the output of each sublayer is defined as $\text{LayerNorm}(x + \text{Sublayer}(x))$, where $\text{Sublayer}()$ is the function implemented by the sublayer itself. To encourage these residual connections, all the sublayers and embedding layers produce outputs of dimension d_m . Each source word is encoded as a vector \mathbf{h} and the textual representation is denoted by $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n) \in \mathbb{R}^{d_m}$.

Visual Encoder

The internal structure of the visual encoder is the same as that introduced in Section 4.2.2. Similarly, the visual representation is denoted by $\mathbf{A} = (\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \dots, \mathbf{a}_m) \in \mathbb{R}^{d_r}$.

Word-to-Region (W2R)

Based on the two stages of the WRA integration method introduced in Section 4.2.2, I similarly combine semantically relevant textual features into visual features guided by soft/hard WRA. Consequently, the visual representations are combined with the WRA-guided textual representations and converted into semantically enhanced multimodal representations, which are denoted as $\mathbf{R} = (\mathbf{r}_1, \mathbf{r}_2, \mathbf{r}_3, \dots, \mathbf{r}_m) \in \mathbb{R}^{d_r}$.

Doubly Attentive Decoder

The decoder also comprises a stack of N identical layers. In addition to the two sub-layers, similar to the encoder, a decoder with double cross-attention mechanisms over the sources inspired by [Libovický et al., 2018] is implemented. The proposed model attends to textual representations and multimodal representations simultaneously to generate textual and multimodal contexts. Each of them is attended to using the same set of queries, that is, the output of the self-attention sublayer. A residual connection link is used between the queries and summed context vectors from the parallel double cross-attentions.

The conventional textual cross-attention used in the Transformer, called the multi-head scaled dot-product attention, is simultaneously calculated on a set of queries, keys, and values with M heads in parallel. Then, the generation of each head is packed together into a matrix Q, K_t, V_t . The output matrix is computed as follows:

$$\begin{aligned} \text{MultiHead}(Q, K_t, V_t) &= \text{Concat}(\text{head}_t^1, \dots, \text{head}_t^M)W^O \\ \text{where } \text{head}_t^{i \in [1, M]} &= \text{Attention}(QW_i^Q, K_tW_i^K, V_tW_i^V) \\ &= \text{softmax}\left(\frac{QW_i^Q(K_tW_i^K)^T}{\sqrt{d_m}}\right)V_tW_i^V \end{aligned}$$

where W_i^* and W^O are learnable parameter matrices.

Similarly, the visual cross-attention mechanism performs the multi-head scaled dot-product attention on multimodal representations with M heads in parallel, as follows:

$$\begin{aligned} \text{MultiHead}(Q, K_r, V_r) &= \text{Concat}(\text{head}_r^1, \dots, \text{head}_r^M)W^o \\ \text{where } \text{head}_r^{i \in [1, M]} &= \text{Attention}(QW_i^q, K_rW_i^k, V_rW_i^v) \\ &= \text{softmax}\left(\frac{QW_i^q(K_rW_i^k)^T}{\sqrt{d_m}}\right)V_rW_i^v \end{aligned}$$

where W_i^* and W^o are learnable parameter matrices.

Therefore, the summarized output from the two cross-attentions is fed into the residual connection and layer normalization. Then, the output is fed into the FFN sublayer, where the dimensions of the input and output are d_m and d_{ff} . Finally, the decoder is capped using a linear layer and a softmax layer to predict the probability of a target word.

4.3 Experiments

4.3.1 Datasets

I experimented on English→German (En→De) and English→French (En→Fr) tasks using the Multi30k dataset [Elliott et al., 2016]. The dataset contained 29k training images and 1,014 validation images. For testing, I used three public test sets to evaluate my models: the Test2016 set and the Test2017 set, each containing 1k images. Each image was paired with image descriptions expressed by both the original English sentences and the German and French translations. I lowercased and tokenized the English, German, and French descriptions and English visual concepts using the script in the Moses Toolkit.[§] I converted space-separated tokens into subword units using a byte pair encoding (BPE) model.[¶] With 10k merge operations, the resulting vocabulary sizes of each language pair were 5,202→7,065 tokens for En→De and 5,833→6,575 tokens for En→Fr.

4.3.2 Evaluation

I evaluated the quality of the translation according to the token level BLEU [Papineni et al., 2002] and METEOR [Denkowski and Lavie, 2014] metrics and reported the average score over three runs.

I conducted a statistical significance test with bootstrap resampling [Koehn, 2004] for the merger of three test translations using the script in Moses Toolkit. I reported a statistically significant improvement in BLEU if the p -value is < 0.05 .^{||}

4.3.3 Setup

In my experiments, I split two branches based on the architecture of the models: RNN-based models and Transformer-based models. Each branch includes the following types of models for comparison.

[§]<https://github.com/moses-smt/mosesdecoder>

[¶]<https://github.com/rsennrich/subword-nmt>

^{||}I did not report on METEOR due to the statistical significance test for METEOR is not implemented in the Moses script.

- **NMT**: the text-to-text NMT model, wherein only the textual sentences were used.
- **MNMT_R**: the doubly attentive MNMT model [Zhao et al., 2020] using regional visual features, without integrating WRA to process W2R strategy.
- **MNMT_{W2R(sa)}**: the proposed MNMT model incorporating soft WRA to guide W2R stage.
- **MNMT_{W2R(ha)}**: the proposed MNMT model incorporating hard WRA to guide W2R stage.

All the models were implemented with Nmtpytorch [Caglayan et al., 2017b].**

In the visual encoder for all the MNMT models, the number of regional features was set to $m = 36$, and the dimensions of the regional feature vectors were set to $d_r = 2,048$.

Settings of the RNN-Based Models

I set the dimensions of the encoder and decoder hidden states at $d_s = 256$, the textual representation dimension at $d_h = 512$, word embedding at 128-dim, batch size at 46, textual dropout at 0.3, visual dropout at 0.5, model dropout at 0.5, and both blocks \mathbf{b}_t^t and \mathbf{b}_t^y at 0.5. I used the Adam optimizer [Kingma and Ba, 2015] with a learning rate of 0.0004 for all the models. I consistently stopped training when the METEOR score did not improve over 10 evaluations on the validation set, and one validation evaluation was performed every 1,000 iterations.

Settings of the Transformer-Based Models

I set the encoder and decoder to contain $N = 6$ layers. The dimensions of all the input and output layers were set to $d_m = 512$. The textual representation dimension was $d_m = 512$ and the dimensions of the inner feed-forward neural network layer were $d_{ff} = 2,048$. The number of all the multi-head modules in the encoder and decoder layers was set to $M = 8$.

**<https://github.com/lium-lst/nmtpytorch>

For training En→De and En→Fr tasks, the sentence-minibatch size was set to 64, the label smoothing value was set at 0.1, and the attention dropout and residual dropout were 0.3. I used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$. The learning rate was tuned to 0.05, with a warm-up step of 4,000 for the NMT model; 0.03 with a warm-up step of 4,000 for the MNMT_R, MNMT_{W2R(sa)}, and MNMT_{W2R(ha)} models. I stopped training when the METEOR score did not improve over 10 evaluations on the validation set and one validation evaluation was performed every 1,000 iterations.

4.3.4 Further Experimental Comparison

To empirically verify the merit of my proposed model, I also presented the performance of the following state-of-the-art MNMT models for comparison, namely:

- **VAG-NMT** [Zhou et al., 2018]: Jointly optimizes a translation model and learns a shared vision-language space.
- **VMMT_F** [Calixto et al., 2019]: An MNMT model that incorporates image context learned by a latent variable model.
- **Del+Obj** [Ive et al., 2019]: A Transformer-based deliberation model enriched using object-level features.
- **Trans+VR** [Zhang et al., 2020]: A Transformer model with universal visual representation by a topic-image lookup table.
- **VAR-{S2S, TF} (hard)** [Yang et al., 2020]: Jointly trains the source-to-target and target-to-source translation models through hard visual agreement regularization.
- **MNMT+SVA** [Nishihara et al., 2020]: A Transformer-based MNMT model with the supervised visual attention mechanism.
- **GMFE-NMT** [Yin et al., 2020]: A graph-based multimodal fusion encoder to conduct graph encoding for NMT.
- **MTF** [Yao and Wan, 2020]: A Transformer-based NMT model with multimodal self-attention to integrate text and image features.

- **OVC+L_m** [Wang and Xiong, 2021]: An MNMT model with grounding translation on desirable visual objects by masking irrelevant objects in the visual modality.
- **ImagiT** [Long et al., 2021]: An MNMT method via visual imagination.

4.4 Results

4.4.1 Results on the En→De Task

Results within RNN-Based Models

As shown in Table 4.1, compared with the text-to-text NMT, the MNMT_R consistently improved the translation performance, benefiting from integrating regional features. Nevertheless, the improvements were less significant across different metrics on all the test sets. From this point, I observed that even if high-quality regional features are fused, the role of the visual feature is limited by the integration method and is not fully realized. In contrast, the proposed MNMT_{W2R(sa)} and MNMT_{W2R(ha)} models yielded significantly improved translation results, compared to the NMT baseline, and consistently obtained a larger margin than the MNMT_R.

The key difference between the MNMT_{W2R} and MNMT_R was the integration of WRA. It was verified that integrating the WRA enabled better use of the visual features; therefore, the performance was better than that of the model without the WRA. I think that the significant improvements could be attributed to two aspects of the proposed model:

- The WRA bridges vision and language well.
- Integrating WRA-guided textual features with visual features can promote visual feature utilization.

In general, both the MNMT_{W2R(sa)} and MNMT_{W2R(ha)} models performed well, and there was almost no gap in translation results between the integration of soft WRA and hard WRA. In more detail, the integration of soft WRA could help visual attention focus on regional visual features by considering all textual features according to the semantic relevance of each image region. On the other hand, the integration of hard

Multi30k En→De				
Models	Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR
Existing MNMT Models				
VAG-NMT [Zhou et al., 2018]	N/A	N/A	31.6	52.2
VMMT _F [Calixto et al., 2019]	37.7	56.0	30.1	49.9
Del+Obj [Ive et al., 2019]	38.0	55.6	N/A	N/A
Trans+VR [Zhang et al., 2020]	36.9	N/A	28.6	N/A
VAR-S2S (hard) [Yang et al., 2020]	N/A	N/A	29.3	51.2
VAR-TF (hard) [Yang et al., 2020]	N/A	N/A	29.3	50.2
MNMT+SVA [Nishihara et al., 2020]	39.9	58.1	N/A	N/A
GMFE-NMT [Yin et al., 2020]	39.8	57.6	32.2	51.9
MTF [Yao and Wan, 2020]	38.7	55.7	N/A	N/A
OVC+L _m [Wang and Xiong, 2021]	N/A	N/A	32.3	53.4
ImagiT [Long et al., 2021]	38.5	55.7	32.1	52.4
RNN-Based Models				
NMT	37.4	57.5	29.6	51.3
MNMT _R	37.5	57.7	30.1	51.6
MNMT _{W2R(sa)}	38.4 ^{††}	58.1	30.2 [†]	51.9
MNMT _{W2R(ha)}	38.4 ^{††}	58.0	31.2 ^{††}	52.2
Transformer-Based Models				
NMT	38.4	57.5	31.5	51.9
MNMT _R	38.4	57.6	31.1	51.5
MNMT _{W2R(sa)}	39.3 ^{††}	58.3	32.3 ^{††}	52.8
MNMT _{W2R(ha)}	39.0 ^{††}	58.2	31.8 [†]	52.6

Table 4.1: BLEU and METEOR scores on Multi30k En→De task. The results are significantly better than those of NMT (†) and MNMT_R (‡) with p -value < 0.05 . The best performance in my models and existing MNMT models appear in bold. All my results are the average scores over three runs.

WRA could assist visual attention to focus on regional visual features by indicating the most semantically relevant textual features for each of them. Therefore, the integration of either soft WRA or hard WRA could help visual attention pay attention to regional visual features that were semantically related to the textual features, leading to better visual information use.

Results within Transformer-Based Models

As shown in Table 4.1, based on the Transformer architecture, the $MNMT_R$ model cannot outperform the state-of-the-art text-to-text NMT model. This may be attributed to two factors:

- When the primary modality (text) is sufficient to accomplish the translation task, the visual context cannot play a supplementary role; however, it may interfere with the effect of the textual context.
- When encoding source words, the Transformer considers the association between words and the entire sentence. However, there was no relationship between the regional features. Therefore, the visual context may not be as useful as the textual context.

In contrast, the proposed $MNMT_{W2R(sa)}$ and $MNMT_{W2R(ha)}$ models consistently improved the translation performance over the text-to-text NMT model. The significant improvements show that the proposed $MNMT_{W2R(sa)}$ and $MNMT_{W2R(ha)}$ models overcome both problems. Specifically, in my proposed models, I maintained the textual context while enriching the visual features with WRA-guided textual features to generate a multimodal context such that the multimodal context can play a more effective role than the pure visual context.

Comparison of Proposed Model and Existing Ones

I conducted early stopping on the METEOR metric; therefore, I mainly compared my METEOR results with existing models. As shown in Table 4.1, the METEOR scores of the Transformer-based $MNMT_{W2R(sa)}$ and $MNMT_{W2R(ha)}$ models surpassed most of the state-of-the-art MNMT results. The best performance was yielded by the proposed Transformer-based $MNMT_{W2R(sa)}$ model. This demonstrates that my proposed models are competitive among all the existing MNMT models.

Multi30k En→Fr				
Models	Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR
Existing MNMT Models				
VAG-NMT [Zhou et al., 2018]	N/A	N/A	53.8	70.3
Del+Obj [Ive et al., 2019]	59.8	74.4	N/A	N/A
VAR-S2S (hard) [Yang et al., 2020]	N/A	N/A	52.6	69.9
VAR-TF (hard) [Yang et al., 2020]	N/A	N/A	53.3	70.4
Trans+VR [Zhang et al., 2020]	57.5	N/A	48.5	N/A
GMFE-NMT [Yin et al., 2020]	60.9	74.9	53.9	69.3
OVC+L_m [Wang and Xiong, 2021]	N/A	N/A	54.1	70.5
ImagiT [Long et al., 2021]	59.7	74.0	52.4	68.3
RNN-Based Models				
NMT	59.3	74.6	51.6	69.2
MNMT_R	59.5	74.7	51.6	69.0
MNMT_{W2R(sa)}	59.7	75.0	52.2 ^{†‡}	69.6
MNMT_{W2R(ha)}	60.3^{†‡}	75.5	52.3^{†‡}	69.6
Transformer-Based Models				
NMT	60.7	75.2	53.1	69.6
MNMT_R	60.6	75.4	52.7	69.2
MNMT_{W2R(sa)}	61.7 ^{†‡}	76.3	54.1^{†‡}	70.6
MNMT_{W2R(ha)}	61.8^{†‡}	76.3	54.0 ^{†‡}	70.4

Table 4.2: BLEU and METEOR scores on Multi30k En→Fr task. The results are significantly better than those of NMT (†) and MNMT_R (‡) with p -value of < 0.05 . The best performance in my models and existing MNMT models appear in bold. All my results are the average scores over three runs.

4.4.2 Results on the En→Fr Task

As shown in Table 4.2, the results of the En→Fr task showed a consistent trend with the results of the En→De task. Thus, the generality of the proposed framework was established.

First, based on either the RNN or Transformer, the proposed $\text{MNMT}_{\text{W2R}(\text{sa})}$ model and $\text{MNMT}_{\text{W2R}(\text{ha})}$ model consistently surpassed the NMT model by a significant margin. In contrast, the MNMT_{R} without the WRA failed to achieve a unified improvement over NMT and the improvement was less significant. This validates the effectiveness of my postulation that translation performance can be effectively improved through the integration of WRA.

Second, the results of the proposed Transformer-based $\text{MNMT}_{\text{W2R}(\text{sa})}$ model and $\text{MNMT}_{\text{W2R}(\text{ha})}$ model surpassed all the state-of-the-art MNMT results based on both the BLEU and METEOR metrics. The best performance was achieved by the Transformer-based $\text{MNMT}_{\text{W2R}(\text{sa})}$ model, which is consistent with the result of the En→De task.

4.5 Analyses

4.5.1 Ablation Study

To further verify the effectiveness of the different components in my proposed model, I also showed the performance of the following ablated versions. All the ablated versions were implemented on both RNN-based and Transformer-based models with soft/hard WRA.

Different Integration Strategies of WRA

In the proposed $\text{MNMT}_{\text{W2R}(\text{sa})}$ and $\text{MNMT}_{\text{W2R}(\text{ha})}$ models, I integrate WRA to guide textual features into corresponding visual features to generate multimodal context during the W2R stage. In the ablation study, I conduct extensive experiments on different integration strategies of WRA to confirm the effectiveness of the integration method of WRA.

Region-to-Word (R2W): Unlike the $\text{MNMT}_{\text{W2R}(\text{sa})}$ and $\text{MNMT}_{\text{W2R}(\text{ha})}$ models, I implemented a region-to-word (R2W) stage instead of the W2R stage introduced in Sec-

<i>Different Strategies \Rightarrow WRA Integration</i>				
<i>Models</i>	Multi30k En\rightarrowDe			
	Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR
RNN-Based Models				
<i>MNMT</i> _{W2R(sa)}	38.4	58.1	30.2	51.9
<i>MNMT</i> _{W2R(ha)}	38.4	58.0	31.2	52.2
<i>MNMT</i> _{R2W(sa)}	34.6	55.6	25.6	48.0
<i>MNMT</i> _{R2W(ha)}	31.5	51.9	22.4	44.2
<i>MNMT</i> _{R\RightarrowW(sa)}	35.3	56.1	26.8	49.0
<i>MNMT</i> _{R\RightarrowW(ha)}	31.7	52.1	22.6	44.8
Transformer-Based Models				
<i>MNMT</i> _{W2R(sa)}	39.3	58.3	32.3	52.8
<i>MNMT</i> _{W2R(ha)}	39.0	58.2	31.8	52.6
<i>MNMT</i> _{R2W(sa)}	37.7	56.5	30.3	50.4
<i>MNMT</i> _{R2W(ha)}	36.1	55.5	27.9	48.9
<i>MNMT</i> _{R\RightarrowW(sa)}	37.8	56.8	30.9	51.3
<i>MNMT</i> _{R\RightarrowW(ha)}	36.5	55.7	29.0	49.2

Table 4.3: Ablation study on different integration strategies of WRA. BLEU and METEOR scores on En \rightarrow De task using the Multi30k dataset. The best performance is shown in bold. All results are the average scores over three runs.

<i>Different Strategies \Rightarrow WRA Integration</i>				
<i>Models</i>	Multi30k En\rightarrowFr			
	Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR
RNN-Based Models				
<i>MNMT</i> _{W2R(sa)}	59.7	75.0	52.2	69.6
<i>MNMT</i> _{W2R(ha)}	60.3	75.5	52.3	69.6
<i>MNMT</i> _{R2W(sa)}	57.0	73.2	48.0	65.9
<i>MNMT</i> _{R2W(ha)}	52.5	69.2	43.2	62.2
<i>MNMT</i> _{R\RightarrowW(sa)}	57.3	73.3	48.3	66.3
<i>MNMT</i> _{R\RightarrowW(ha)}	52.9	69.7	44.4	63.2
Transformer-Based Models				
<i>MNMT</i> _{W2R(sa)}	61.7	76.3	54.1	70.6
<i>MNMT</i> _{W2R(ha)}	61.8	76.3	54.0	70.4
<i>MNMT</i> _{R2W(sa)}	60.6	75.4	52.3	68.9
<i>MNMT</i> _{R2W(ha)}	57.8	73.1	49.6	66.9
<i>MNMT</i> _{R\RightarrowW(sa)}	61.4	75.8	53.4	69.8
<i>MNMT</i> _{R\RightarrowW(ha)}	58.9	74.0	51.3	68.2

Table 4.4: Ablation study on different integration strategies of WRA. BLEU and METEOR scores on En \rightarrow Fr task using the Multi30k dataset. The best performance is shown in bold. All results are the average scores over three runs.

tions 4.2.2 and 4.2.3. In R2W, I maintained the pure visual context and integrated the WRA-guided visual features with corresponding textual features to enrich the textual context and realize a multimodal context.

W2R and R2W ($R \rightleftharpoons W$): I implemented W2R together with R2W to achieve a bi-directional integration strategy called $R \rightleftharpoons W$. In $R \rightleftharpoons W$, both the visual context enriched by WRA-guided textual features and the textual context enriched by WRA-guided visual features became multimodal contexts.

As shown in Tables 4.3 and 4.4, the proposed $\text{MNMT}_{\text{W2R(sa)}}$ and $\text{MNMT}_{\text{W2R(ha)}}$ models achieved the best performance among all the integration strategies. When the WRA was integrated with the R2W and $R \rightleftharpoons W$ strategies, the results were worse than the W2R integration strategy because it might disturb the textual context. It was suggested that maintaining the textual context while enriching the visual context into a multimodal context by WRA-guided textual features was the most appropriate integration strategy for WRA.

Moreover, compared with the R2W strategy, the $R \rightleftharpoons W$ strategy was slightly better. I conjectured that when interfering with the textual context in the text-to-text task, enriching the visual context using WRA-guided textual features, instead of the pure visual context, enabled better visual information use in the image-to-text task. These results validated the effectiveness of the proposed WRA integration strategy.

Different Intermodal Fusion Operations

I explored the impact of different intermodal fusion operations during the generation of a multimodal context on the overall performance.

In the proposed W2R stage, I combined the WRA-guided textual features and visual features to generate a multimodal context with CONCAT as the fusion operator, which is defined in Equation 4.1. Instead of CONCAT, I investigated the SUM and MULTIPLY operations to fuse modalities for generating multimodal contexts.

From Tables 4.5 and 4.6, it can be observed that the CONCAT operation was the most effective fusion operation to generate the multimodal context in my proposal. The results were slightly worse when the fusion was realized using the SUM and MULTIPLY operators. This difference could be attributed to the fact that concatenation could make use of a linear layer that learned how to integrate the modality-specific activations into the multimodal context vector, as demonstrated in [Caglayan et al., 2016b].

<i>Different Intermodal Fusion Operations</i>				
<i>Variants</i>	Multi30k En→De			
	Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR
RNN-Based MNMT _{W2R(sa)}				
<i>CONCAT</i>	38.4	58.1	30.2	51.9
<i>MULTIPLY</i>	36.1	56.5	28.1	50.3
<i>SUM</i>	37.9	57.8	30.3	51.7
RNN-Based MNMT _{W2R(ha)}				
<i>CONCAT</i>	38.4	58.0	31.2	52.2
<i>MULTIPLY</i>	37.5	57.7	29.8	51.2
<i>SUM</i>	37.5	57.7	30.1	51.7
My Transformer-Based MNMT _{W2R(sa)}				
<i>CONCAT</i>	39.3	58.3	32.3	52.8
<i>MULTIPLY</i>	39.2	58.3	31.6	52.1
<i>SUM</i>	38.4	57.7	31.4	51.8
Transformer-Based MNMT _{W2R(ha)}				
<i>CONCAT</i>	39.0	58.2	31.8	52.6
<i>MULTIPLY</i>	39.0	58.1	31.4	51.9
<i>SUM</i>	38.2	57.6	31.6	52.2

Table 4.5: Ablation study on different intermodal fusion operations. BLEU and METEOR scores on En→De task using the Multi30k dataset. The best performance is shown in bold. All results are the average scores over three runs.

<i>Different Intermodal Fusion Operations</i>				
<i>Variants</i>	Multi30k En→Fr			
	Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR
RNN-Based MNMT _{W2R(sa)}				
<i>CONCAT</i>	59.7	75.0	52.2	69.6
<i>MULTIPLY</i>	58.3	74.0	50.9	68.3
<i>SUM</i>	60.0	75.6	51.9	69.1
RNN-Based MNMT _{W2R(ha)}				
<i>CONCAT</i>	60.3	75.5	52.3	69.6
<i>MULTIPLY</i>	59.8	74.8	51.7	68.8
<i>SUM</i>	59.6	74.8	52.2	69.4
Transformer-Based MNMT _{W2R(sa)}				
<i>CONCAT</i>	61.7	76.3	54.1	70.6
<i>MULTIPLY</i>	61.5	76.0	53.5	69.5
<i>SUM</i>	61.6	76.1	54.1	70.6
Transformer-Based MNMT _{W2R(ha)}				
<i>CONCAT</i>	61.8	76.3	54.0	70.4
<i>MULTIPLY</i>	60.9	75.3	52.6	69.0
<i>SUM</i>	61.1	75.8	54.0	70.3

Table 4.6: Ablation study on different intermodal fusion operations. BLEU and METEOR scores on En→Fr task using the Multi30k dataset. The best performance is shown in bold. All results are the average scores over three runs.

4.5.2 Visualization

In Figure 4.5, I visualized the learned textual and visual representations to further analyze the proposed method.

In detail, I visualized the textual and visual representations, which are learned by RNN-based/Transformer-based $\text{MNMT}_{\text{W2R}(\text{sa})}$ and $\text{MNMT}_{\text{W2R}(\text{ha})}$ on different language pairs, respectively. For the Test2016 set of the En→De and En→Fr tasks, I generated textual and visual representations as follows:

- Text: the hidden representations for textual features generated by the textual encoder.
- Image_(independent): representations of regional visual features generated by the visual encoder, independent of textual features.
- Image_($\text{MNMT}_{\text{W2R}(\text{sa/ha})}$): the learned representations for regional visual features generated by the W2R integration strategy using soft/hard WRA, which are enriched with semantically relevant textual features guided by WRA.

I took the average of word/region representations to obtain the representations for each sentence and image and visualized them using the T-SNE toolkit.

As shown in Figure 4.5, the representations learned by the RNN-based/Transformer-based $\text{MNMT}_{\text{W2R}(\text{sa})}$ model and $\text{MNMT}_{\text{W2R}(\text{ha})}$ model are visualized, respectively. The representation distribution of different proposed models conveys the same commonality that the distribution of $\text{image_}(\text{MNMT}_{\text{W2R}(\text{sa/ha})})$ is always in the middle of the text and $\text{image_}(\text{independent})$. It can be observed that the distribution of $\text{image_}(\text{MNMT}_{\text{W2R}(\text{sa/ha})})$ is always closer to the text than the $\text{image_}(\text{independent})$. Furthermore, although the $\text{image_}(\text{MNMT}_{\text{W2R}(\text{sa/ha})})$ is close to the text, the distribution of the text is not disturbed by the $\text{image_}(\text{MNMT}_{\text{W2R}(\text{sa/ha})})$.

Visualizations in Figure 4.5 further prove the contributions of each key component of the proposed method as follows:

First, the utilization of WRA is a key component that causes the distribution of the $\text{image_}(\text{independent})$ and $\text{image_}(\text{MNMT}_{\text{W2R}(\text{sa/ha})})$ to be different. It can be found that, after enriching the independent visual features with related textual features guided by WRA, the proposed method can bring the visual features closer to the textual features.

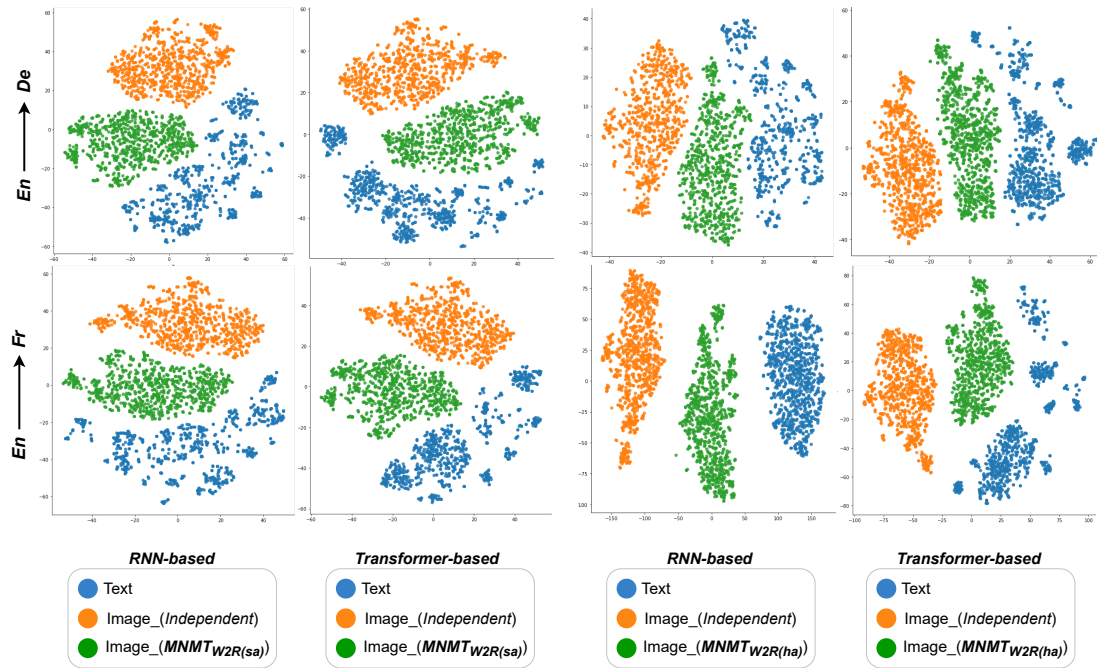


Figure 4.5: Representation visualization for textual features, independent visual features, and enriched visual features with soft/hard WRA-guided textual features. Representations are learned by RNN-based/Transformer-based $MNMT_{W2R(sa)}$ and $MNMT_{W2R(ha)}$ on the $En \rightarrow De$ and $En \rightarrow Fr$ tasks, respectively. *Text* (blue): the textual representations generated by the textual encoder. *Image_(independent)* (orange): the visual representations generated by visual encoder before conducting W2R, which are independent of textual representations. *Image_($MNMT_{W2R(sa/ha)}$)* (green): the enriched visual representations generated by W2R, which have been related with textual features by leveraging soft/hard WRA as a bridge.

This demonstrates that the intermediate facility WRA plays a crucial role as a bridge, connecting independent textual and visual features that are far from each other.

Second, the W2R integration strategy is another key component that brings the image_($MNMT_{W2R(sa/ha)}$) close to the text without disturbing the distribution of the text. By concatenating visual features with WRA-guided textual features in the W2R strategy, the proposed method can encourage visual features to interact with semantically relevant textual features to help them be closer without disturbing the textual features. Therefore, it can be demonstrated that the W2R strategy plays a crucial role in advancing interactions between textual and visual modalities without disturbing the textual features.

4.5.3 Case Study

To further analyze the effectiveness of the proposed model, I showed two cases generated by different models to analyze the translation quality.

I performed the visualization as follows: (1) I visualized the source-target alignment of textual attention. (2) I visualized the region-target alignment of visual attention at a time step that generated a certain target word while attending to the most weighted region. The region was denoted by a bounding box along with the target word. (green indicated $MNMT_{W2R}$ and gray indicated $MNMT_R$).

In Figure 4.6, I showed two cases to analyze the translation quality. The upper case shows the results from the RNN-based models and the lower case shows the results from the Transformer-based models.

In the upper example, the $MNMT_{W2R(ha)}$ correctly translated “backyard” to a compound noun of “arrière-cour.” However, the NMT and $MNMT_R$ without the WRA models mistranslated it as “cour,” which means “yard” in English. Through visualization, I observed that the regional visual feature utilized by the $MNMT_{W2R(ha)}$ model provided more helpful information for generating more accurate translations than the $MNMT_R$ model. This showed that the proposed model can fully utilize visual information to complement textual information to learn more accurate translations.

In the lower example, the $MNMT_{W2R(sa)}$ correctly translated the source phrase “tank tops” to the target word “débardeurs,” whereas the NMT and $MNMT_R$ without the WRA models failed. From the visualization, I observed that the textual attention weights for the source tokens “tank” and “tops” to the target word “débardeurs” were



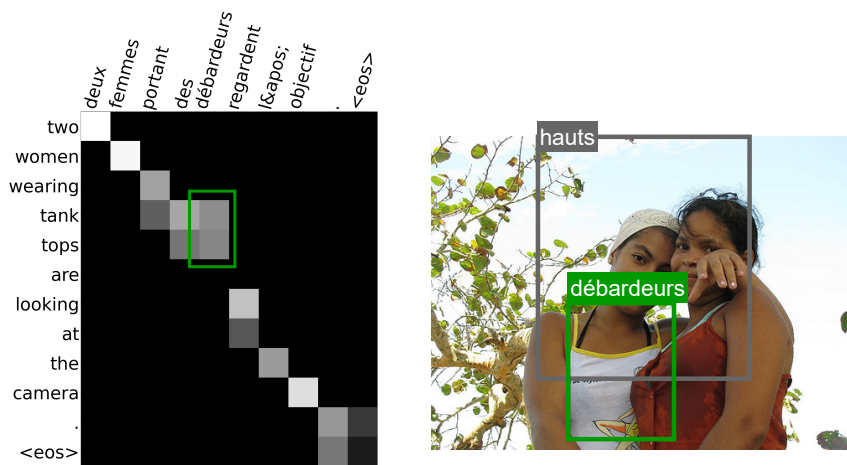
Src (En): a man is grilling out in his *backyard* .

Ref (Fr): un homme fait un barbecue dans son *arriere-cour* .

MNMT_{w2R}: un homme fait griller quelque chose dans sa *arriere-cour* .

MNMT_R: un homme fait griller quelque chose dans sa *cour (yard)* .

NMT: un homme fait griller quelque chose dans sa *cour (yard)* .



Src (En): two women wearing *tank tops* are looking at the camera .

Ref (Fr): deux femmes portant des *débardeurs* regardent l'objectif .

MNMT_{w2R}: deux femmes portant des *débardeurs* regardent l'objectif .

MNMT_R: deux femmes vêtues de *hauts (tops)* regardent l'objectif .

NMT: deux femmes portant des *hauts (tops)* regardent l'objectif .

Figure 4.6: Improved examples for the case study.

the highest. At the same time, the region with the highest visual attention weight provided semantically relevant information about “tank tops” to help generate the target word “débardeurs.” This showed that in the proposed model, the textual and visual context provided semantically relevant information interactively to generate a more accurate translation.

These cases revealed that my proposal can lead to the better visual information use and improved translation accuracy.

4.6 Summary

In this proposal, I proposed WRA to link textual and visual features based on semantic relevance. To facilitate the semantic correlation between textual and visual contexts, I proposed a novel integration strategy W2R. The W2R integration method guided by WRA effectively maintained the textual context while transforming the visual context into a multimodal one by enriching it using WRA-guided textual features. Extensive experimental results showed that the proposed model significantly outperformed the competitive baselines on the En→De and En→Fr language pairs consistently. Moreover, the performance of the proposed model surpassed most of the existing MNMT methods. Further analysis demonstrated that the proposed models superior translation performance was attributable to better visual information utilization.

5 Conclusions and Future Directions

This thesis presents two effective solutions to the difficulties faced by the MNMT task: (1) The region-attentive MNMT aims to generate target words by attending to specific semantic image regions with an additional region-dependent attention mechanism; (2) The WRA-guided MNMT aims to guide textual features to interact with semantically relevant regional visual features by incorporating an auxiliary facility WRA as a bridge. Both these two methods have been implemented on two mainstream architectures of NMT: the RNN and the Transformer. Extensive experiments on English–German and English–French translation tasks using the Multi30k dataset have been conducted to verify the effectiveness of the proposed methods. Consistent results on different language pairs and different architectures show that the proposed methods can improve over baselines and outperforms most of the state-of-the-art MNMT methods. Further analyses demonstrate that both of the proposed methods can achieve better translation performance because of their better image information use.

In the future, I suggest exploring new solutions to overcome the challenges faced by machine translation research from the following research directions:

- I suggest using much finer visual information, such as instance semantic segmentation, to improve the quality of visual features.
- I recommend taking into account parsing word attributes in the sentence structure to avoid redundant visual information use, such as some non-visual words.
- As data sparsity issues are still a limitation of the MNMT task and the data collection is really expensive, I think that exploiting how to effectively integrate retrieval and reranking pipelines for MNT or exploring the unsupervised way using the pseudo dataset are more promising for the next research step.

6 Social Impacts

Computational language understanding is at the heart of MT which requires inferring the meaning of a sentence in one language and transferring that meaning to another language. For human beings, semantic understanding comes from perceiving multiple modalities including linguistic, visual, and auditory to help understand semantic tasks. Similar to human perception, MT can also benefit from incorporating auxiliary modalities. Therefore, this work is inspired by the semantic understanding achieved by human multimodal perception, and the following societal impacts that are closely related to human technologies and applications will be driven and stimulated.

From a technical perspective, this work combines the two major research areas of computer vision (CV) and natural language processing (NLP). Therefore, many other multimodal tasks can benefit from the findings of this work, e.g. Visual Question Answering (VQA), multimodal Dialog, and video-guided MT. In addition, as human semantic understanding comes from multimodal perceptions of language, images, sounds, and videos, multimodal learning systems inspired by the human perceptual system can be extended to different realistic downstream fields, such as AI Agents, User Interfaces, Robotics Industry, and so on.

From the application perspective, this work visually enriches purely linguistic understanding to improve machine translation as a multimodal task. Communication in the world is inseparable from language, and language communication is based on effective machine translation. In the future, multimodal machine translation systems will facilitate semantic communication between languages, 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.

Bibliography

- [Anderson et al., 2018] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., and Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 6077–6086.
- [Ba et al., 2016] Ba, L. J., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *CoRR*, abs/1607.06450.
- [Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, pages 1–15.
- [Barrault et al., 2018] Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. (2018). Findings of the third shared task on multimodal machine translation. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., N ev ol, A., Neves, M. L., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 304–323.
- [Caglayan et al., 2017a] Caglayan, O., Aransa, W., Bardet, A., Garc a-Mart nez, M., Bougares, F., Barrault, L., Masana, M., Herranz, L., and van de Weijer, J. (2017a). LIUM-CVC submissions for WMT17 multimodal translation task. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., and Kreutzer, J., editors, *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 432–439.

- [Caglayan et al., 2016a] Caglayan, O., Aransa, W., Wang, Y., Masana, M., García-Martínez, M., Bougares, F., Barrault, L., and van de Weijer, J. (2016a). Does multimodality help human and machine for translation and image captioning? In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 627–633.
- [Caglayan et al., 2018] Caglayan, O., Bardet, A., Bougares, F., Barrault, L., Wang, K., Masana, M., Herranz, L., and van de Weijer, J. (2018). LIUM-CVC submissions for WMT18 multimodal translation task. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., Névéol, A., Neves, M. L., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 597–602.
- [Caglayan et al., 2016b] Caglayan, O., Barrault, L., and Bougares, F. (2016b). Multimodal attention for neural machine translation. *CoRR*, abs/1609.03976.
- [Caglayan et al., 2017b] Caglayan, O., García-Martínez, M., Bardet, A., Aransa, W., Bougares, F., and Barrault, L. (2017b). NMT-PY: A flexible toolkit for advanced neural machine translation systems. *Prague Bull. Math. Linguistics*, 109:15–28.
- [Caglayan et al., 2019] Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. (2019). Probing the need for visual context in multimodal machine translation. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4159–4170.
- [Calixto et al., 2016] Calixto, I., Elliott, D., and Frank, S. (2016). DCU-UvA multimodal MT system report. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 634–638.
- [Calixto and Liu, 2017] Calixto, I. and Liu, Q. (2017). Incorporating global visual features into attention-based neural machine translation. In Palmer, M., Hwa, R.,

- and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 992–1003.
- [Calixto et al., 2017] Calixto, I., Liu, Q., and Campbell, N. (2017). Doubly-attentive decoder for multi-modal neural machine translation. In Barzilay, R. and Kan, M., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1913–1924.
- [Calixto et al., 2019] Calixto, I., Rios, M., and Aziz, W. (2019). Latent variable model for multi-modal translation. In Korhonen, A., Traum, D. R., and Màrquez, L., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6392–6405.
- [Cho et al., 2014a] Cho, K., van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014a). On the properties of neural machine translation: Encoder-decoder approaches. In Wu, D., Carpuat, M., Carreras, X., and Vecchi, E. M., editors, *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111.
- [Cho et al., 2014b] Cho, K., van Merriënboer, B., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014b). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Moschitti, A., Pang, B., and Daelemans, W., editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.
- [Chowdhury and Elliott, 2019] Chowdhury, K. D. and Elliott, D. (2019). Understanding the effect of textual adversaries in multimodal machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP 2019*, pages 35–40.
- [Delbrouck and Dupont, 2017] Delbrouck, J. and Dupont, S. (2017). An empirical study on the effectiveness of images in multimodal neural machine translation. In

- Palmer, M., Hwa, R., and Riedel, S., editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 910–919.
- [Delbrouck and Dupont, 2018] Delbrouck, J. and Dupont, S. (2018). Bringing back simplicity and lightness into neural image captioning. *CoRR*, abs/1810.06245.
- [Denkowski and Lavie, 2014] Denkowski, M. J. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014, June 26-27, 2014, Baltimore, Maryland, USA*, pages 376–380.
- [Elliott, 2018] Elliott, D. (2018). Adversarial evaluation of multimodal machine translation. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2974–2978.
- [Elliott et al., 2017] Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017). Findings of the second shared task on multimodal machine translation and multilingual image description. In Bojar, O., Buck, C., Chatterjee, R., Federmann, C., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., and Kreutzer, J., editors, *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 215–233.
- [Elliott et al., 2015] Elliott, D., Frank, S., and Hasler, E. (2015). Multi-language image description with neural sequence models. *CoRR*, abs/1510.04709.
- [Elliott et al., 2016] Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, 2016, August 12, Berlin, Germany*, pages 70–74.
- [Elliott and Kádár, 2017] Elliott, D. and Kádár, Á. (2017). Imagination improves multimodal translation. In Kondrak, G. and Watanabe, T., editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP*

2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers, pages 130–141.

- [Fukui et al., 2016] Fukui, A., Park, D. H., Yang, D., Rohrbach, A., Darrell, T., and Rohrbach, M. (2016). Multimodal compact bilinear pooling for visual question answering and visual grounding. In Su, J., Carreras, X., and Duh, K., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 457–468.
- [Grönroos et al., 2018] Grönroos, S., Huet, B., Kurimo, M., Laaksonen, J., Mérialdo, B., Pham, P., Sjöberg, M., Sulubacak, U., Tiedemann, J., Troncy, R., and Vázquez, R. (2018). The MeMAD submission to the WMT18 multimodal translation task. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., Névéal, A., Neves, M. L., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 603–611.
- [Gupta et al., 2017] Gupta, T., Shih, K. J., Singh, S., and Hoiem, D. (2017). Aligned image-word representations improve inductive transfer across vision-language tasks. In *2017 IEEE International Conference on Computer Vision, ICCV 2017*, pages 4223–4232.
- [He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778.
- [Helcl et al., 2018] Helcl, J., Libovický, J., and Varis, D. (2018). CUNI system for the WMT18 multimodal translation task. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Monz, C., Negri, M., Névéal, A., Neves, M. L., Post, M., Specia, L., Turchi, M., and Verspoor, K., editors, *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 616–623.
- [Huang et al., 2016] Huang, P., Liu, F., Shiang, S., Oh, J., and Dyer, C. (2016). Attention-based multimodal neural machine translation. In *Proceedings of the First*

Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany, pages 639–645.

- [Huang et al., 2020] Huang, Q., Wei, J., Cai, Y., Zheng, C., Chen, J., Leung, H., and Li, Q. (2020). Aligned dual channel graph convolutional network for visual question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020*, pages 7166–7176.
- [Ive et al., 2019] Ive, J., Madhyastha, P., and Specia, L. (2019). Distilling translations with visual awareness. In Korhonen, A., Traum, D. R., and Màrquez, L., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6525–6538.
- [Karpathy and Li, 2015] Karpathy, A. and Li, F. (2015). Deep visual-semantic alignments for generating image descriptions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 3128–3137.
- [Kingma and Ba, 2015] Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, pages 1–15.
- [Klein et al., 2017] Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT: Open-source toolkit for neural machine translation. In Bansal, M. and Ji, H., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, System Demonstrations*, pages 67–72.
- [Koehn, 2004] Koehn, P. (2004). Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A meeting of SIGDAT, a Special Interest Group of the ACL, held in conjunction with ACL 2004, 25-26 July 2004, Barcelona, Spain*, pages 388–395.
- [Koehn et al., 2007] Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O.,

- Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In Carroll, J. A., van den Bosch, A., and Zaenen, A., editors, *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*, pages 177–180.
- [Krishna et al., 2017] Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. (2017). Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73.
- [Li et al., 2020] Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., and Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the 2020 European Conference on Computer Vision*, pages 121–137.
- [Libovický and Helcl, 2017] Libovický, J. and Helcl, J. (2017). Attention strategies for multi-source sequence-to-sequence learning. In Barzilay, R. and Kan, M., editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 196–202.
- [Libovický et al., 2018] Libovický, J., Helcl, J., and Marecek, D. (2018). Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation, WMT 2018*,, pages 253–260.
- [Lin et al., 2020] Lin, H., Meng, F., Su, J., Yin, Y., Yang, Z., Ge, Y., Zhou, J., and Luo, J. (2020). Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1320–1329.
- [Liu et al., 2019] Liu, F., and Xuancheng Ren, Y. L., He, X., and Sun, X. (2019). Aligning visual regions and textual concepts for semantic-grounded image representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*, pages 6847–6857.
- [Long et al., 2021] Long, Q., Wang, M., and Li, L. (2021). Generative imagination elevates machine translation. In *Proceedings of the 2021 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5738–5748.

- [Nam et al., 2017] Nam, H., Ha, J., and Kim, J. (2017). Dual attention networks for multimodal reasoning and matching. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pages 2156–2164.
- [Nguyen and Okatani, 2018] Nguyen, D. and Okatani, T. (2018). Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*.
- [Nishihara et al., 2020] Nishihara, T., Tamura, A., Ninomiya, T., Omote, Y., and Nakayama, H. (2020). Supervised visual attention for multimodal neural machine translation. In *Proceedings of the 2020 International Conference on Computational Linguistics*, pages 4304–4314.
- [Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318.
- [Ren et al., 2015] Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 91–99.
- [Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Li, F. (2015). ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252.
- [Sennrich et al., 2016] Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th An-*

nual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers, pages 1715–1725.

- [Specia et al., 2016] Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation, WMT 2016, collocated with ACL 2016, August 11-12, Berlin, Germany*, pages 543–553.
- [Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.
- [Tan and Bansal, 2019] Tan, H. and Bansal, M. (2019). LXMERT: learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, EMNLP 2019*, pages 5099–5110.
- [Teney et al., 2018] Teney, D., Anderson, P., He, X., and van den Hengel, A. (2018). Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018*, pages 4223–4232.
- [Toyama et al., 2016] Toyama, J., Misono, M., Suzuki, M., Nakayama, K., and Matsuo, Y. (2016). Neural machine translation with latent semantic of image and text. *CoRR*, abs/1611.08459.
- [Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

- [Wang and Xiong, 2021] Wang, D. and Xiong, D. (2021). Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *AAAI Conference on Artificial Intelligence*, pages 2720–2728.
- [Xu et al., 2015] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In Bach, F. R. and Blei, D. M., editors, *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 2048–2057.
- [Yang et al., 2020] Yang, P., Chen, B., Zhang, P., and Sun, X. (2020). Visual agreement regularized training for multi-modal machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9418–9425.
- [Yao and Wan, 2020] Yao, S. and Wan, X. (2020). Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 4346–4350.
- [Yin et al., 2020] Yin, Y., Meng, F., Su, J., Zhou, C., Yang, Z., Zhou, J., and Luo, J. (2020). A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3025–3035.
- [Zhang et al., 2020] Zhang, Z., Chen, K., Wang, R., Utiyama, M., Sumita, E., Li, Z., and Zhao, H. (2020). Neural machine translation with universal visual representation. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, pages 1–11.
- [Zhao et al., 2020] Zhao, Y., Komachi, M., Kajiwara, T., and Chu, C. (2020). Double attention-based multimodal neural machine translation with semantic image regions. In *Proceedings of the 22nd Annual Conference of the European Associa-*

tion for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020, pages 105–114.

[Zhao et al., 2021a] Zhao, Y., Komachi, M., Kajiwara, T., and Chu, C. (2021a). Neural machine translation with semantically relevant image regions. page c3.

[Zhao et al., 2021b] Zhao, Y., Komachi, M., Kajiwara, T., and Chu, C. (2021b). Tmeku system for the wat2021 multimodal translation task. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 174–180.

[Zhou et al., 2018] Zhou, M., Cheng, R., Lee, Y. J., and Yu, Z. (2018). A visual attention grounding neural model for multimodal machine translation. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 3643–3653.