Multimodal Transformer for Multimodal Machine Translation

Shaowei Yao, Xiaojun Wan
Center for Data Science, Peking University
Wangxuan Institute of Computer Technology, Peking University
The MOE Key Laboratory of Computational Linguistics, Peking University
{yaosw,wanxiaojun}@pku.edu.cn

Abstract

Multimodal Machine Translation (MMT) aims to introduce information from other modality, generally static images, to improve the translation quality. Previous works propose various incorporation methods, but most of them do not consider the relative importance of multiple modalities. In MMT, equally treating text and images may encode too much irrelevant information from images which may introduce noise. In this paper, we propose the multimodal self-attention in Transformer to solve the issues above. The proposed method learns the representations of images based on the text, which avoids encoding irrelevant information in images. Experiments and visualization analysis demonstrate that our model benefits from visual information and substantially outperforms previous works and competitive baselines in terms of various metrics.

1 Introduction

Multimodal machine translation (MMT) is a novel machine translation (MT) task which aims at designing better translation systems using context from an additional modality, usually images (See Figure 1). It initially organized as a shared task within the First Conference on Machine Translation (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). Current works focus on the dataset named Multi30k (Elliott et al., 2016), a multilingual extension of Flickr30k dataset with translations of the English image descriptions into different languages.

Previous works propose various incorporation methods. Calixto and Liu (2017) utilize global image features to initialize the encoder/decoder hidden states of RNN. Elliott and Kádár (2017) model the source sentence and reconstruct the image representation jointly via multi-task learning. Recently, Ive et al. (2019) propose a translate-and-refine approach using two-stage decoder based on Transformer (Vaswani et al., 2017). Calixto et al. (2019) put forward a latent variable model to learn the interaction between visual and textual features. However, in multimodal tasks the different modalities usually are not equally important. For example, in MMT the text is obviously more important than images. Although the image carries richer information, it also contains more irrelevant content. If we directly encode the image features, it may introduce a lot of noise.

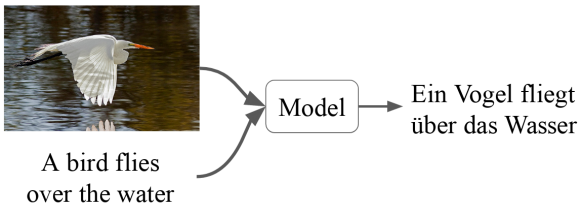


Figure 1: An Example for Multimodal Machine Translation.

To address the issues above, we propose the multimodal Transformer. The proposed model does not directly encode image features. Instead, the hidden representations of images are induced from the text under the guide of image-aware attention. Meanwhile, we introduce a better way to incorporate information from other modality based on a graph perspective of Transformer. Experimental results and visualization show that our model can make good use of visual information and substantially outperforms the current state of the art.

2 Methodology

Our model is adapted from Transformer and it is also an encoder-decoder architecture, consisting of stacked encoder and decoder layers. The focus of our work is to build a powerful encoder to incorporate the information from other modality. Thus, we will first begin with an introduction to the incorpo-

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4346–4350

ration method. Then we will detail the multimodal self-attention. The final representations of text and images are sent to the sequence decoder to generate the target text.

2.1 Incorporating Method

The method of incorporating information from other modality is based on a graph perspective of Transformer. The core of Transformer is self-attention which employs the multi-head mechanism. Each attention head operates on an input sequence $x = (x_1, ..., x_n)$ of n elements where $x_i \in \mathbb{R}^d$, and computes a new sequence $z = (z_1, ..., z_n)$ of the same length where $z \in \mathbb{R}^d$:

$$z_i = \sum_{j=1}^{n} \alpha_{ij} \left(x_j W^V\right) \tag{1}$$

where α_{ij} is weight coefficient computed by a softmax function:

$$\alpha_{ij} = \text{softmax}\left(\frac{\left(x_i W^Q\right)\left(x_j W^K\right)^T}{\sqrt{d}}\right) \tag{2}$$

$W^V, W^Q, W^V \in \mathbb{R}^{d\times d}$ are layer-specific trainable parameter matrices.

Thus we can see that each word representation is induced from all the other words. If we consider every word to be a node, then Transformer can be regarded as a variant of GNN which treats each sentence as a fully-connected graph with words as nodes (Battaglia et al., 2018; Yao et al., 2020). In traditional MT tasks, the source sentence graph only contains nodes with text information. If we want to incorporate information from other modality, we should add the nodes with other modality information into the source graph. Therefore, as the words are local semantic representations of the sentence, we extract the spatial features which are the semantic representations of local spatial regions of the image. We add the spatial features of the image as pseudo-words in the source sentence and feed it into the multimodal self-attention layer.

2.2 Multimodal Self-attention

As stated before, in MMT the text and images are not equally important. Directly encoding images which contain a lot of irrelevant content may introduce noise. Therefore, we propose the multimodal self-attention to encode multimodal information. In multimodal self-attention, the hidden representations of the image are induced from text under

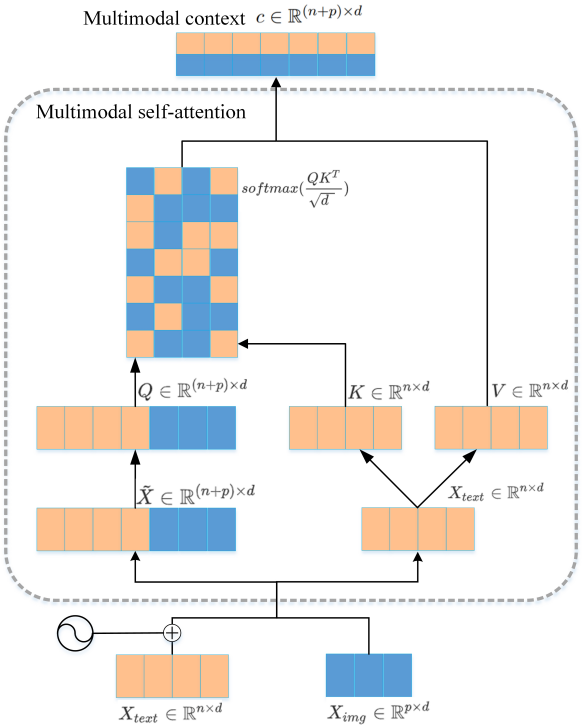


Figure 2: Multimodal self-attention

the guide of image-aware attention which provides a latent adaptation from the text to the image. A visual representation is illustrated in Figure 2.

Formally, we consider two modalities $text$ and img, with two entries from each of them denoted by $x^{\text{text}} \in \mathbb{R}^{n\times d}$ and $x^{\text{img}} W^{\text{img}} \in \mathbb{R}^{p\times d}$, respectively. The output of multimodal self-attention is computed as follows:

$$c_i = \sum_{j=1}^{n} \tilde{\alpha}_{ij} \left(x_j^{\text{text}} W^V\right) \tag{3}$$

where $\tilde{\alpha}_{ij}$ is weight coefficient computed by a softmax function:

$$\tilde{\alpha}_{ij} = \text{softmax}\left(\frac{\left(\tilde{x}_i W^Q\right)\left(x_j^{\text{text}} W^K\right)^T}{\sqrt{d}}\right) \tag{4}$$

where $c \in \mathbb{R}^{(n+p)\times d}$ is the hidden representation of words and the image. At last layer, c is fed into sequence decoder to generate target sequence. We can see that the hidden representations of the image is only induced from words but under the guide of image-aware attention. The extracted spatial features of the image are not directly encoded in the model. Instead, they adjust the attention of each word to compute the hidden representations of the image. In each encoder layer we also employ

residual connections between each layer as well as layer normalization. And the decoder are followed the standard implemention of Transformer.

3 Experiment

3.1 Baselines and Metrics

We compare the performance of our model with previous kinds of models: (1) sequence-to-sequence model only trained on text data (LSTM, Transformer). (2) Previous works trained on both text and image data. We evaluated the translation quality of our model in terms of BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014), which have been used in most previous works.

3.2 Datasets

We build and test our model on the Multi30k dataset (Elliott et al., 2016), which consists of two multilingual expansions of the original Flickr30k dataset referred to as M30k$_T$ and M30K$_C$, respectively. Multi30k contains 30k images, and for each of the images, M30k$_T$ has one of its English description manually translated into German by a professional translator. M30K$_C$ has five English descriptions and five German descriptions, but the German descriptions were crowdsourced independently from their English versions. The training, validation, test sets of Multi30k contain 29k, 1014 and 1k instances respectively. We use M30k$_T$ as the original training data and M30k$_C$ for building additional back-translated training data following Calixto et al. (2019). We present our experiment results on English-German (En-De) Test2016. We use LSTM trained on the textual part of the M30KT dataset (De-En, the original 29k sentences) without images to build a back-translation model (Sennrich et al., 2016), and then apply this model to translate 145k monolingual German description in M30k$_C$ into English as additional training data. This part of data we refer to as back-translated data.

3.3 Settings

We preprocess the data by tokenizing and lowercasing. Word embeddings are initialized using pretrained 300-dimensional Glove vectors. we extract spatial image features from the last convolutional layer of ResNet-50. The spatial features are $7 \times 7 \times 2048$-dimensional vectors which are representations of local spatial regions of the image.

Our encoder and decoder have both 6 layers with 300-dimensional word embeddings and hidden states. We employ 10 heads here and dropout=0.1. We used Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and minibatches of size 32 or 128 (depends on if add the back-translated data). Meanwhile, we increase learning rate linearly for the first $warmup\_steps$, and decrease it thereafter proportionally to the inverse square root of the step number. We used $warmup\_steps = 8000$. The similar learning rate schedule is adopted in (Vaswani et al., 2017).

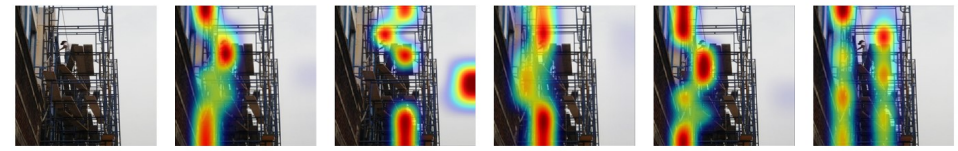
Model	BLEU4	METEOR
LSTM	36.8	54.9
Transformer	37.8	55.3
IMG$_D$ (Calixto and Liu, 2017)	37.3	55.1
NMT$_{SRC+IMG}$ (Calixto et al., 2017)	36.5	55.0
Transformer+Att (Ive et al., 2019)	36.9	54.5
Del+obj (Ive et al., 2019)	<u>38.0</u>	55.6
VMMT$_F$ (Calixto et al., 2019)	37.6	<u>56.0</u>
Ours	**38.7**	55.7
+ back-translated data		
IMG$_D$ (Calixto and Liu, 2017)	38.5	55.9
NMT$_{SRC+IMG}$ (Calixto et al., 2017)	37.1	54.5
VMMT$_F$ (Calixto et al., 2019)	38.4	<u>56.3</u>
Ours	**39.5**	**56.9**

Table 1: Comparison results on the Multi30k test set. The best baseline results are underlined. Bold highlights statistically significant improvements.

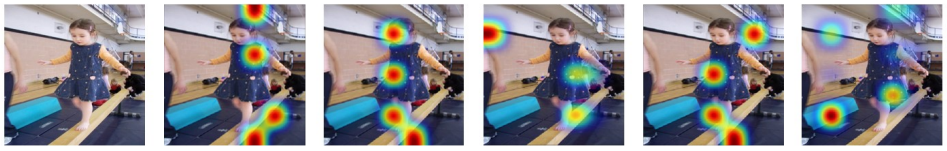
3.4 Results

The results of all methods are shown in Table 1. We can see our Transformer baseline has comparable results compared to most previous works, When trained on the original data, our model substantially outperforms the SoTA according to BLEU and gets a competitive result according to METEOR. Moreover, we note that our model surpasses the text-only baseline by above 1 BLEU points. It demonstrates that our model benefits a lot from the visual modality.

To further investigate our model performance on more data, we also train the models with additional back-translated data, and the comparison results are shown in the lower part of Table 1. We can see that almost all models get improved with the additional training data, but our model obtains the most improvements and achieving new SoTA results on all metrics. It demonstrates that our model will perform better on the larger dataset.



source: a guy works on a building .
baseline: ein mann steht an einem gebäude .
ours: ein mann arbeitet an einem gebäude .
ours$_{back}$: ein typ arbeitet an einem gebäude .
reference: ein typ arbeitet an einem gebäude .



source: a girl in a jean dress is walking along a raised balance beam .
baseline: ein mädchen in einem jeansjacke geht eine erhöhte plattform entlang .
ours: ein mädchen in einem jeansanzug läuft auf einem schwebebalken .
ours$_{back}$: ein mädchen in einem jeanskleid läuft mit einem angehobenen balken entlang .
reference: ein mädchen in einem jeanskleid läuft über einen erhöhten schwebebalken .

Figure 3: Translation cases and Visualization. Colored words represent some of the improvements.

3.5 Visualization Analysis

Figure 3 depicts translations for two cases in the test set. Colors highlight improvement. Furthermore, we visualize the contributions of different local regions of the image in different attention heads, which shows our model can focus on the appropriate regions of the image. For example, our model pays more attention to the building and the person in the first case, and thus the model understands that the person is working on the building rather than just standing there. In the second case, most attention heads attend to the balance beam and the jean dress of the girl, avoiding errors in the translation.

3.6 Ablation Study

To further study the influence of the individual components in our model, we conduct ablation experiments to better understand their relative importance. The results are presented in Table 2. Firstly, we investigate the effect of multimodal self-attention. As shown in the second columns (replace with self-attention) in Table 2. If we simply concatenate the word vectors with the image features and then perform self-attention, we will lose 0.6 BLEU score and 0.4 METEOR score. Inspired by Elliott (2018), we further examine the utility of the image by the adversarial evaluation. When we replace all input images with a blank picture, the performance of the model drops a lot. When we replace all input images with a random image (the context of image does not match the description in the sentence pair), the model performs even worse than the text-only model. The image here is actually a noise which distracts the translation.

	BLEU4	MEMTEOR
Full Model	**38.7**	**55.7**
- replace with self-attention	38.1	55.3
- replace with blank images	37.1	54.8
- replace with random images	36.7	54.5

Table 2: Influence of different components in our model.

4 Conclusion

In this paper, we propose the multimodal self-attention to consider the relative importance between different modalities in the MMT task. The hidden representations of less important modality (image) are induced from the important modality (text) under the guide of image-aware attention. The experiments and visualization show that our model can make good use of multimodal information and get better performance than previous works.

There are various multimodal tasks where multiple modalities have different relative importance. In future work, we would like to investigate the effectiveness of our model in these tasks.

Acknowledgments

This work was supported by National Natural Science Foundation of China (61772036), MSRA Collaborative Research Program, and Key Laboratory of Science, Technology and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). We thank the anonymous reviewers for their helpful comments. Xiaojun Wan is the corresponding author.

References

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323.

Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. 2018. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *EMNLP*, pages 992–1003.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *ACL*, pages 1913–1924.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *ACL*, pages 6392–6405.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *EMNLP*, pages 2974–2978.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 130–141.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *ACL*, pages 6525–6538.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *ICLR*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *ACL*, pages 86–96.

Lucia Specia, Stella Frank, Khalil Simaan, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Shaowei Yao, Tianming Wang, and Xiaojun Wan. 2020. Heterogeneous graph transformer for graph-to-sequence learning. In *Proceedings of the Association for Computational Linguistics (ACL)*.