Fusion of Image-text attention for Transformer-based Multimodal Machine Translation

Junteng Ma, Shihao Qin, Lan Su, Xia Li✉
School of Information Science and Technology
Guangdong University of Foreign Studies
Guangzhou, China
{juntengma, shihao_qin, destinyofloveing}@126.com
xiali@mail.gdufs.edu.cn

Lixian Xiao
Faculty of Asian Languages and Cultures
Guangdong University of Foreign Studies
Guangzhou, China
200110732@oamail.gdufs.edu.cn

***Abstract*—In recent years, multimodal machine translation has become one of the hot research topics. In this paper, a machine translation model based on self-attention mechanism is extended for multimodal machine translation. In the model, an Image-text attention layer is added in the end of encoder layer to capture the relevant semantic information between image and text words. With this layer of attention, the model can capture the different weights between the words that is relevant to the image or appear in the image, and get a better text representation that fuses these weights, so that it can be better used for decoding of the model. Experiments are carried out on the original English-German sentence pairs of the multimodal machine translation dataset, Multi30k, and the Indonesian-Chinese sentence pairs which is manually annotated by human. The results show that our model performs better than the text-only transformer-based machine translation model and is comparable to most of the existing work, proves the effectiveness of our model.**

***Index Terms*—Multimodal Machine Translation; Image-text attention; Transformer-based; Self-attention.**

I. Introduction

Multimodal neural machine translation [1,2] (MNMT) aims to use several modalities, such as image modal information, to help solve the problem of semantic ambiguity in plain text, so as to improve the quality of machine translation.

Previous work of multimodal machine translation can be roughly divided into recurrent neural network based (RNN-based) framework and Transformer-based framework. In the work of RNN-based architecture, the main idea is to integrate different forms of image features into different parts of the model, such as the work of Huang et al. [3], Calixto et al. [4-5] and Caglayan et al. [6]. Elliott et al. [7] proposed a way of "imagination" which decompose the multimodal machine translation task into two subtasks, one is a regular translation task, and the other task is to predict the corresponding visual representation by the encoded sentence representation, that is making the distance between the text and the image representation closer.

With the advantages of self-attention mechanism for text-only machine translation, which is proposed by Vaswani et al. [8], some work began to extend Transformer-based framework for multimodal machine translation. For example, Helcl et al. [9] used Transformer to build MNMT model. They proposed two ideas, one is to modify the structure of decoder by adding a visual cross-attention layer, the other is to use the imagination [7] method. Inspired by Caglayan et al. [6], Grönroos et al. [10] regarded the image features as pseudo words and used a gating procedure to process the image feature.

Different from the previous Transformer-based MNMT model, our model is mainly to change the internal structure of the encoder. We argue that if the visual information can be applied to the source words at the end of encoding layer, then the semantic information of those words that are more related to the image could be enhanced. Based on this motivation, we propose to add an image-text attention layer in the end of encoder layer, so that the model can receive two modalities in the end of encoder and capture the relationship between visual and text information, refining the sentence representation. In this way, the output of the encoder may contain both image and textual representations and the key information related to image is strengthened to the representation. The contribution of our work are as follows:

(1) In order to capture the different weights between the words that is relevant to the image or appear in the image, we propose to add an image-text attention layer in the end of encoder layer, so we can get a better text representation that fuses the semantic relationship between image and text words.

(2) We carried several experiments on the Multi30k dataset, not only on English to German sentence pairs but also on Indonesian to Chinese sentence pairs (The validation data and test data of Indonesian to Chinese sentence pairs are annotated by human.) We will show that our extension of multimodal machine translation model performs better than the text-only transformer-based machine translation model.

II. Model

A. The Architecture of Our Model

Followed by the encoder-decoder scheme, the architecture of our model is based on Transformer NMT [8]. Transformer contains self-attention layer and a feed-forward network layer. Both the end of encoder and decoder consists of several layers to get better text representation by deepening the network. Since Transformer does not have the ability to capture sequence information of source words like RNN, position embedding is added to the word embedding to give position information to each word.

✉ corresponding author: xiali@mail.gdufs.edu.cn

In order to enhance semantic information of the text words and improve the translation quality of text-only NMT model, we extend the Transformer-based machine translation model for multimodal machine translation by fusing with the attention of image vision and text words. The architecture is shown in Figure 1.

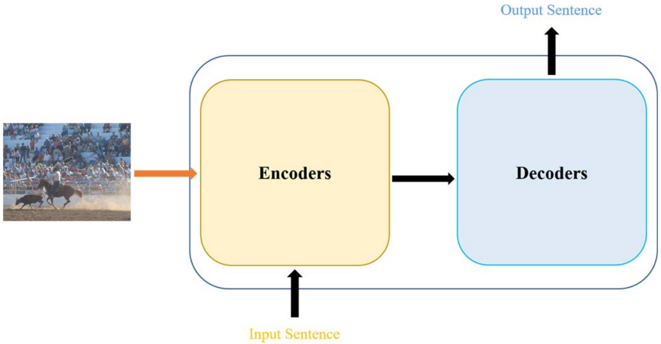


Figure 1. The architecture of our model.

B. *Fusion of Image-text Attention*

The main method of image fusion in previous work is to change the structure of the decoder layer [9] or to apply a gating procedure to the image feature in the output of encoder or decoder [10]. Different from previous work, we introduced an image-text attention layer in the model, which is between the self-attention layer and the feed-forward network layer. The model is shown as Figure 2.

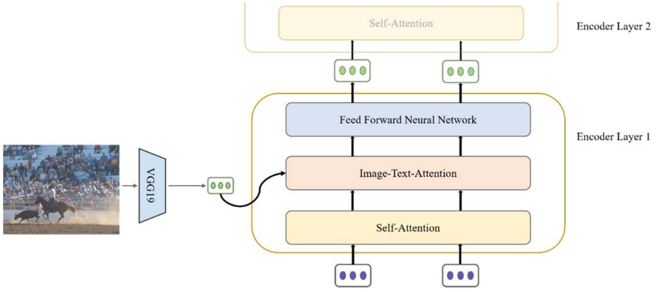


Figure 2. Encoder of our model.

In the original Transformer NMT, each encoder layer only contains a self-attention layer and a feed-forward network layer. In self-attention layer, each input vector is converted into three different vectors, which are the Query vector, the Key vector, and the Value vector. They are obtained by multiplying the word embedding and three different weight matrices W^Q, W^K, W^V, where the shapes of these three matrices are the same. For a given Q, the similarity function is first used to calculate and compare with each K. Then, the result are normalized to obtain weights to calculate the context vector, which is a weighted sum of weights and V.

The similarity function of self-attention in Transformer uses scaled dot-product, so the results of self-attention is as Equation (1), where d is the dimension of Q, K, V.

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \qquad (1).$$

In order to improve the performance of the self-attention layer, the Multi-Head attention mechanism is used, which not only expands the ability of the model to focus on different positions, but also provides multiple representative subspaces of the self-attention layer. The specific approach is to project the input Q, K, V into multiple subspaces, that is, Multi-Head, performs self-attention calculation, and concatenate the output of each head and then feed to a fully connection layer. The Equation is as (2), Where W_i^Q, W_i^K, $W_i^V \in \mathbb{R}^{d_{model} \times d}$, $W^O \in \mathbb{R}^{hd \times d_{model}}$, d_{model} is the model's dimension, h is the number of heads.

$$\begin{aligned}\mathrm{MultiHead}(\mathrm{Q},\mathrm{K},\mathrm{V}) &= Concat(head_1, head_2, \dots, head_h)W^O,\\ \text{where } head_i &= Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \qquad (2).\end{aligned}$$

In our model, as shown in Figure 2, the Q in the image-text attention layer is regarded as the image feature corresponding to the sentence, which is extracted by a pre-trained VGG-19 [11]. K, V are derived from the self-attention layer's output. The context vector obtained by the Multi-Head attention mechanism thus contains the interaction information between the image and the source language sentence. Supposed that the extracted image features as I, and the output of image-text attention layer is expressed as Equation (3):

$$\begin{aligned}Context_{img} &= \mathrm{MultiHead}(\mathrm{I},\mathrm{K},\mathrm{V})\\ &= Concat(head_1, head_2, \dots, head_h)W'^O,\\ \text{where } head_i &= ImageTextAttention(IW_i^I, KW_i'^K, VW_i'^V) (3).\end{aligned}$$

All layers are interconnected with residual connections and their outputs are normalized by layer normalization [12].

C. *Decoder Layer of Our Model*

Followed the work of Vaswani et al. [8], the decoder layer of our model is shown in Figure 3. In the decoder layer, self-attention is masked to prevent the decoder to process the "future" states. Different from the encoder layer, there is an additional sub-layer called encoder-decoder-attention layer after self-attention layer, which attends to the final output of the encoder and the output of self-attention layer. It allows every position in the decoder to attend over all positions in the input sequence.

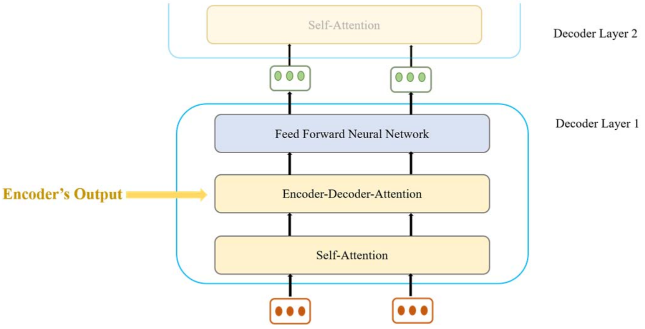


Figure 3. Decoder of our model.

D. *Model Training*

Followed previous work, we define multimodal translation as the task of producing target language translation y, given the source language x and an image v. Multimodal neural machine translation can be formed as minimizing the negative log-likelihood of a translation model that is additionally conditioned on the corresponding image. The loss function is shown in Equation (4).

$$\mathrm{J}(\theta) = -\sum_{j} logp(y_j \mid y_{<j}, x, v) \qquad (4).$$

Where x is the source sentence description of the image v, y_j is the *j-th* token of the target sentence y.

III. Experiments

A. *Dataset*

We use Multi30k [13] dataset in our experiments to test the performance of our model. Multi30k is a multilingual expansion of the original Flickr30k [14] dataset for image description generation task. Each image in Multi30k consists of an English sentence description and a German sentence translated by professional translators. The data of Multi30k is divided into training set, validation set and test set. The image size of training set, validation set and test set are 29,000, 1,014 and 1,000 respectively in Multi30k, each image has one sentence pair: the original English description and its German translation.

As for the preprocess of the dataset, we use Moses [15] scripts to tokenize, normalize-punctuation and true case for both English and German sentences. And we use byte pair encoding compression algorithm [16] to convert the words into sub-words. We use the training set of Multi30k to train the model, and validation set for model selection and test set for evaluation. We use BLEU4 [17] and METEOR [18] to measure the quality of the translation result.

B. *Experiment Setup*

As for the image features, we use a VGG-19 network which is pre-trained on ImageNet [19] to extract visual representations. We feed the images in Multi30k and use the 4096D activations of the penultimate fully connected layer FC7 as our global image features.

For the configuration of Transformer, we use the Transformer network with 6 layers and 8 heads, the model dimension with 512 and feed-forward network dimension with 2048. We use Adam optimizer [20] with initial learning rate 0.01, and Noam learning rate decay scheme [8] with $\beta_1 = 0.9$, $\beta_2 = 0.997$, $\epsilon = 10^{-9}$ and 16,000 warm-up steps. In the experiment, dropout rate is 0.1, beam search size is 10. We train the model on training set and select the best model according to BLEU4 scores performed on validation set. We report the results on test set with the best model.

C. *Experimental Results*

In this section, we will introduce the baselines we used in our experiments and present the results of our model on different datasets.

1) *Baselines*

In order to verify the performance of our model, we use several baseline models for comparison, they are: Transformer Text-only NMT, Huang et al., 2016[3], Calixto et al.,2017 [4], Calixto et al.,2017[5], Caglayan et al.,2017[6], and Helcl et al.,2018[9]. All these methods can be roughly classified as RNN-based methods and Transformer-based methods.

RNN-based. We use several RNN-based models as our baselines, they are Huang et al., 2016[3], Calixto et al.,2017 [4], Calixto et al.,2017[5] and Caglayan et al.,2017[6]. The work of Huang et al. [3] proposed a MNMT model that fuses regional images which is extracted by RCNN [21], together with the entire image as pseudo words. Calixto et al. [4] used two independent attention mechanisms for text and images. Calixto et al. [5] incorporated global image feature to initialize the decoder's hidden state. And the work of Caglayan et al. [6] modulated each target embedding with global image feature, which is extracted by ResNet[22], using element-wise multiplication.

Transformer-based. We also use Transformer-based model as our baselines. We use text-only Transformer-based model and multimodal as our two baselines. Text-only Transformer-based NMT model is trained on Multi30k's sentence pairs, and does not use the visual information. Our model is an extension of it. We also use the work of Helcl et al. [9] to be another baseline. Helcl et al. [9] proposed a Transformer network with Imagination method [7] which decompose the MNMT into two sub-tasks: translation task and use the encoded textual representation to predict the corresponding image feature, to bring their distance closer.

2) *Results on English to German Sentence Pairs.*

Table 1 shows the results on English-German sentence pairs of Multi30k, the bold number is the results of our model. As we can see, our proposed model performs better than the Text-only NMT and improves 0.25 BLEU scores and 0.5 METEOR scores. That is to say, the image features do have positive effect on the model. Compared with most of the baseline model, like Huang et al. [3] and Calixto et al. [4][5], Image-text attention still performs better in BLEU and METEOR metrics, and is comparable to the work of Caglayan ta al. [6]. But there still has a gap between our model and Helcl et al. [9]'s imagination system.

TABLE I. The results on English to German sentence pairs.

	Models	BLEU4	METEOR
RNN-based	Huang et al. (parallel RCNNs) [3]	36.5	54.1
	Calixto et al. (SRC+IMG) [4]	36.5	55.0
	Calixto et al. (IMG$_D$) [5]	37.3	55.1
	Caglayan et al. (trg-mul) [6]	37.8	57.7
Transformer-based	Helcl et al.(Imagination) [9]	38.8	56.4
	Text-only	37.14	55.4
Our model	Image-Text-Attention	**37.39**	**55.9**

3) *Results on Indonesia to Chinese Sentence Pairs.*

In this paper, we also evaluated our model on low-resource language data. We manually translated the English and German sentence pairs of the validation and test set into the Indonesian and Chinese respectively. As for the training set of Multi30k, we use Google Translation[1] to translate the English-German sentence pairs into the corresponding Indonesian and Chinese.

TABLE II. THE RESULTS ON INDONESIAN TO CHINESE SENTENCE PAIRS.

	Models	BLEU4
RNN-based	Text-only	27.48
Transformer-based	Text-only	28.69
Our method	Image-Text-Attention	**29.02**

Table 2 shows the results of our model and the baseline models text-only NMT base on RNN and Transformer. From Table 2, we can see that Transformer-based NMT model does outperform RNN-based NMT model, and our proposed model has the best result and 0.33 BLEU scores higher (METEOR metric does not support Chinese) than Transformer Text-only NMT. These results again demonstrate the effectiveness of our model.

4) *Discussion.*

According to the results shown above, we can see the effectiveness and advantages of our model. That means the Image-text attention layer does help the model to consider the relevance between visual and the text representation, to get better semantic information of encoder's output. And the image can provide supplemental information to the model.

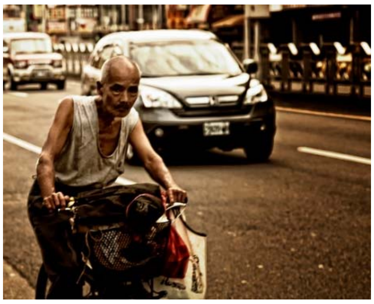
But there is still a gap between our model and the model proposed by Helcl et al. [9]. We think it is possible that our MNMT's final sentence representation in the end of encoder layer is more or less affected by image features, while imagination model in the work of Helcl et al. [9] can learn a better textual representation by using visual information, which can help the model improve the translation quality. And the selection of image features may cause the difference.

IV. CASE STUDY

In order to better demonstrate our model, we conducted a case study of the results on the test set. Due to the better understanding of the Chinese, we select the Indonesian to Chinese translation for case study. Figure 4 and Figure 5 are the results of two cases from our model **Image-Text-Attention** and Transformer-based **Text-only NMT**.

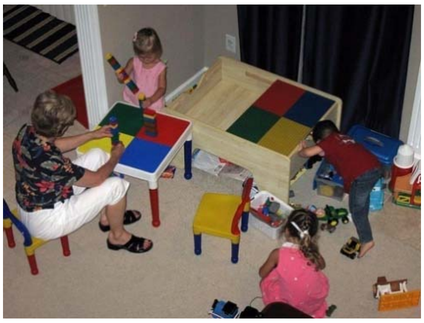
From Figure 4, we can see that our Image-Text-Attention model performs much better than that of Text-only NMT model. The phrase "肮脏的(dirty)" is generated in the target sentence in our model while Text-only NMT does not. And the whole translated sentence is more accurately and smoothly. We also can see that the adjective to describe the old man in the image "秃顶的(bald)" is generated which even not appears in the reference sentence which seems that our model consider the visual information as well.

In Figure 5, **Text-only NMT** translates "积木(blocks)" into "灌木丛(bushes)" by mistake. Maybe it is because both phrases has the same character "木" and leads the model to deviate from the correct result during prediction. But our proposed model translates it correctly, which means that visual information can help the model back on track.



Indonesian source sentence.	Seorang lelaki yang tua dan kurus mengenakan kemeja yang putih dan kotor sedang mengendarai sepeda di jalan.
Chinese reference.	穿着**肮脏的**白色衬衣的一个皮包骨头的老人在街道上骑着自行车。
English sentence	An old skinny man wearing the dirty white shirt riding on a bicycle on the street.
Model	Translation Result
Text-only NMT	一个年长的男人和一个瘦小的家伙穿着一件白色的衬衫，骑着自行车在大街上。(An old man and a skinny guy are wearing a white shirt and riding a bicycle on the street.)
Image-Text-Attention	一个年长，**秃顶的**男人穿着白色和**肮脏的**衬衫骑在街上。(An old, **bald** man was riding down the street in a white and **dirty** shirt)

Figure 4. Translations of different models on Case 1.



Indonesian source sentence	Seorang wanita yang lebih tua dan seorang anak kecil dengan kemeja merah muda bermain dengan balok warna-warni.
Chinese reference	一个老妇人和一个穿着粉红色的衬衫的小孩子，玩五颜六色的**积木块**。
English sentence	An older woman and a young child in a pink shirt playing with multicolored blocks.
Model	Translation Result
Text-only NMT	一个年长的女人和一个穿着粉红色衬衫的小孩玩五颜六色的**灌木丛**。(An older woman and a child in a pink shirt played with colorful **bushes**)
Image-Text-Attention	一个年长的女人和一个穿着粉红色衬衫的小男孩正在玩五颜六色的**积木**。(An old woman and a little boy in a pink shirt were playing with colorful **blocks**)

Figure 5. Translations of different models on Case 2.

[1] https://translate.google.com/ Translated on March 2, 2019.

V. Related Work

Machine translation has made a great progress in recent years, from statistical methods [23-28] to neural-network based machine translation (NMT) methods. Kalchbrenner et al. [29] proposed a neural machine translation model based on distributed continuous representation. This neural network applies an end-to-end fashion and is an early research in the field of machine translation that propose the concept of NMT. Cho et al. [30] and Sutskever et al. [31] improved it in 2014, which better promoted neural-network based machine translation model. In subsequent studies, attention mechanism is used to NMT and achieve better results. Bahdanau et al. [32] introduced attention mechanism based on the work of Cho et al. [30]. Luong et al. [33] improved Bahdanau et al. [32]'s work and proposed a new local and global attention mechanism for NMT. Besides using RNN to implement seq2seq model, researchers began to use CNN architecture for NMT model [34-35]. Gehring et al. [36] proposed an encoder-decoder architecture totally based on CNN. Vaswani et al. [8] first proposed to use Transformer for seq2seq machine translation. Because Transformer abandoned the traditional RNN structure and only use self-attention for feature extraction, the model achieved good results in text-only NMT.

In recent years, multimodal machine translation has become a hot research topic in machine translation. Vinyals et al. [37] proposed an IDG model, which use pre-trained CNN as encoder of the seq2seq model for image caption task. Huang et al. [3] proposed to use regional and global image features from VGG-19 to be fused into the model by regarding them as pseudo text words. Calixto et al. [5] used the global image features and incorporated them in different ways into the NMT model.

Elliott et al. [7] proposed a new solution called Imagination. It decomposes the MNMT task into two subtasks, one is to train a "imaginet decoder" to predict the corresponding visual representation with a margin-based objective, the other task is the regular translation task. Caglayan et al. [6] tried to fuse image features in different ways: (1) compute a new context vector through the regional image features and the target word and concatenate to the original context vector. (2) Modulate the encoder or decoder's output with global image feature using element-wise multiplication.

Besides RNN-based seq2seq architecture, Helcl et al. [9] used Transformer to build MNMT model. They proposed two ideas, one is to modify the structure of decoder by adding a visual cross-attention layer, the other is that use the imagination [7] method. Grönroos et al. [10] regarded image feature as a pseudo words and use a gating procedure to process the image feature.

Different from the previous Transformer MNMT model, our model is mainly to change the internal structure of the encoder layer. In each encoder layer, an Image-text attention layer is added between the self-attention layer and the feed-forward network layer, capturing the relationship between source words and image. So that the semantic information of important words that is related to image can be enhanced, thus improve the translation quality.

VI. Conclusion

This paper extends a Transformer network for multimodal machine translation. We introduced an Image-text attention layer in the end of encoder layer to capture the relationships between source sentence words and the corresponding image by receiving image features as input and constructing Q, K, V to calculate image and text attention like Self-Attention. So that the semantic information of those words that are more related to image could be enhance.

Several experiments are carried on original English to German sentence pairs of Multi30k dataset and Indonesian to Chinese sentence pairs which is manually annotated by human. The results show that compared to the Text-only NMT, our model has a better performance and is comparable to most of the existing work.

In the future, we will try to visualize the attention weights and see how visual information affects the performance of the model. And we will explore some of new fusion ways to incorporate different forms of image features, like global or local image features into our model.

Acknowledgment

This work is supported by National Nature Science Foundation of China (61976062) and the Science and Technology Program of Guangzhou, China (201904010303).

References

[1] T. Baltrusaitis, C. Ahuja, L. Morency, "Multimodal Machine Learning: A Survey and Taxonomy", in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, pp. 423-443, 2019.

[2] L. Specia, S. Frank, K. Sima'an, D. Elliott, "A Shared Task on Multimodal Machine Translation and Crosslingual Image Description", in Proceedings of the First Conference on Machine Translation (WMT), 2016, pp. 543-553.

[3] P. Huang, F. Liu, S. Shiang, J. Oh, C. Dyer, "Attention-based Multimodal Neural Machine Translation", in Proceedings of the First Conference on Machine Translation, Berlin, Germany, 2016, pp. 639–645.

[4] I. Calixto, Q. Liu, N. Campbell, "Doubly-Attentive Decoder for Multi-modal Neural Machine Translation", in Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017, pp. 1913-1924.

[5] I. Calixto, Q. Liu, N. Campbell, "Incorporating Global Visual Features into Attention-Based Neural Machine Translation", in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2017, pp. 992-1003.

[6] O. Caglayan, W. Aransa, A. Bardet, M. Garcia-Martinez, F. Bougares, L. Barrault, "LIUM-CVC Submissions for WMT17 Multimodal Translation Task", in Proceedings of the Conference on Machine Translation(WMT), 2017, pp. 432-439.

[7] D. Elliott, A. Kadar, "Imagination Improves Multimodal Translation", in Proceedings of the 8th International Joint Conference on Natural Language Processing, 2017, pp. 130-141.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L.Kaiser, I. Polosukhin, "Attention Is All You Need", in 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 2017, pp. 71-81.

[9] J. Helcl, J. Libovicky, D. Varis, "CUNI System for the WMT18 Multimodal Translation Task", in Proceedings ofthe Third Conference on Machine Translation (WMT), Volume 2: Shared Task Papers, 2018, pp. 622-629.

[10] S. Grönroos, B.Huet, M. Kurimo, J. Laaksonen, B. Merialdo, P. Pham, M. Sjöberg, U. Sulubacak, J. Tiedemann, R. Troncy, R. Vazqquez, "The MeMAD Submission to the WMT18 Multimodal", in Proceedings of the Third Conference on Machine Translation (WMT) Volume 2: Shared Task Papers, 2018, pp. 609-617.

[11] K. Simonyan, A. Zisserman, “Very Deep Convolutional Networks For Large-Scale Image Recognition”, in International Conference on Learning Representations (ICLR), 2015, pp. 1–14.

[12] L. Ba, R. Kiros, G. Hinton, “Layer Normalization”, arXiv, vol. abs/1607.06450, 2016.

[13] D. Elliott, S. Frank, K. Sima’an, L. Specia, “Multi30K: Multilingual English-German Image Descriptions”, in Proceedings of the 5th Workshop on Vision and Language, 2016, pp. 70-74.

[14] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions”, in Transactions of the Association for Computational Linguistics, 2014, pp. 67-78.

[15] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, E. Herbst, “Moses: Open Source Toolkit for Statistical Machine Translation”, in Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, 2007, pp. 177-180.

[16] R. Sennrich, B. Haddow, A. Birch, “Neural Machine Translation of Rare Words with Subword Units”, in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016, pp. 1715-1725.

[17] K. Papineni, S. Roukos, T. Ward, W. Zhu, “BLEU: a Method for Automatic Evaluation of Machine Translation”, in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 2002, pp. 311-318.

[18] M. Denkowski, A. Lavie, “Meteor Universal: Language Specific Translation Evaluation for Any Target Language”, Proceedings of the Ninth Workshop on Statistical Machine Translation, 2014, pp. 376-380.

[19] O. Russakovsky, J. Deng, H. Su, J Krause, S Satheesh, S Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg · L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge”, in International Journal of Computer Vision, 2015, pp. 211-252.

[20] D. P. Kingma, J. L. Ba, “Adam: A Method for Stochastic Optimization”, in International Conference on Learning Representations (ICLR), 2015.

[21] S. Ren, K. He, R. Girshick, J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks”, in Advances in Neural Information Processing Systems (NIPS), 2015, pp. 91-99.

[22] K. He, X. Zhang, S. Ren, J. Sun, “Deep Residual Learning for Image Recognition”, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[23] P. F. Brown, J. Cocke, S. A. Della Pietra, V. J. Della Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, P. S. Roossin, “A Statistical Approach to Machine Translation”, Computational Linguistics, vol. 16, pp. 79-85, 1990.

[24] P. F. Brown, V. J. Della Pietra, S. A. Della Pietra, R. L. Mercer, “The Mathematics of Statistical Machine Translation : Parameter Estimation”, Computational Linguistics - Special issue on using large corpora: II, vol. 19, pp. 263-311, 1993.

[25] F. Josef Och, H. Ney “Discriminative Training and Maximum Entropy Models for Statistical Machine Translation”, in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, pp. 295-302.

[26] Y. Liu, K. Wang, C. Zong, K. Su, “A unified framework and models for integrating translation memory into phrase-based statistical machine translation”, Computer Speech & Language, vol. 55, pp. 176-206, 2019.

[27] J. Zhang, C. Zong, “Learning a Phrase-based Translation Model from Monolingual Data with Application to Domain Adaptation”, in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, 2013, pp. 1425-1434.

[28] Z. Tu, Y. Liu, Y. Hwang, Q. Liu, S. Lin, “Dependency Forest for Statistical Machine Translation”, in Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), 2010, pp. 1092-1100.

[29] N. Kalchbrenner, P. Blunsom, “Recurrent Continuous Translation Models”, in Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2013, pp.1700-1709.

[30] K. Cho, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation”, in Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1724-1734.

[31] I. Sutskever, O.Vinyals, Q. V. Le, “Sequence to Sequence Learning with Neural Networks”, in Advances in Neural Information Processing Systems, 2014, pp. 3104-3112.

[32] D. Bahdanau, K. Cho, Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate”, in International Conference on Learning Representations (ICLR), 2015.

[33] M. Luong, H. Pham, C. D. Manning, “Effective Approaches to Attention-based Neural Machine Translation”, in Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2015, pp. 1412-1421

[34] J. Bradbury, S. Merity, C. Xiong, R. Socher, “Quasi-Recurrent Neural Networks”, in International Conference on Learning Representations (ICLR), 2017.

[35] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. van den Oord, A. Graves, K. Kavukcuoglu, “Neural Machine Translation in Linear Time”, arXiv, vol. 1610.10099, 2017.

[36] J. Gehring, M. Auli, D. Grangier, D. Yarats, Y. N. Dauphin, “Convolutional Sequence to Sequence Learning”, in Proceeding ICML'17 Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1243-1252.

[37] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, “Show and Tell: A Neural Image Caption Generator”, in 2015 IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, pp. 3156-3164.