Make the Blind Translator See The World: A Novel Transfer Learning Solution for Multimodal Machine Translation

Minghan Wang wangminghan@huawei.com
Jiaxin Guo guojiaxin1@huawei.com
Yimeng Chen chenyimeng@huawei.com
Chang Su suchang8@huawei.com
Min Zhang zhangmin186@huawei.com
Shimin Tao taoshimin@huawei.com
Hao Yang yanghao30@huawei.com
Huawei Translation Service Center, Beijing, China

Abstract

Based on large-scale pretrained networks, the liability to be easily overfitting with limited labelled training data of multimodal translation (MMT) is a critical issue in MMT. To this end, we propose a transfer learning solution. Specifically, 1) A vanilla Transformer is pre-trained on massive bilingual text-only corpus to obtain prior knowledge; 2) A multimodal Transformer named VLTransformer is proposed with several components incorporated visual contexts; and 3) The parameters of VLTransformer are initialized with the pre-trained vanilla Transformer, then being fine-tuned on MMT tasks with a newly proposed method named cross-modal masking which forces the model to learn from both modalities. We evaluated on the Multi30k en-de and en-fr dataset, improving up to 8% BLEU score compared with the SOTA performance. The experimental result demonstrates that performing transfer learning with monomodal pretrained NMT model on multimodal NMT tasks can obtain considerable boosts.

1 Introduction

Transformer-based models using large-scale parallel corpora have significantly improved the performance of neural machine translation (NMT), marking an important milestone (Vaswani et al., 2017). Additionally, multimodal machine translation (MMT) incorporating image signals into RNN-based encoder-decoder shows improvements on translation quality due to the forceful disambiguation (Specia et al., 2016a). In this paper, we aim to investigate, on top of Transformer, whether the paradigm of first pretraining and then fine-tuning can be effectively applied to MMT, concretely transferring from monomodal to multimodal tasks.

Constant attention has been paid on MMT task (Specia et al., 2016a) in the Conference of Machine Translation (WMT) in recent years (2016-2018). Formally, it aims to learn a function mapping: $\mathcal{X} \times \mathcal{I} \rightarrow \mathcal{Y}$, which takes source text and an image as input and translate them into the target text as shown in Figure 1. Additional modality is to disambiguate the source sentence, with the reference of image. However, the effectiveness of the visual context has been questioned by prior work (Specia et al., 2016b; Elliott et al., 2017; Barrault et al., 2018; Caglayan et al., 2019). They show that visual context is not convincingly useful and the marginal gain is pretty modest, which is speculated to be resulted from the limitation of available datasets — the

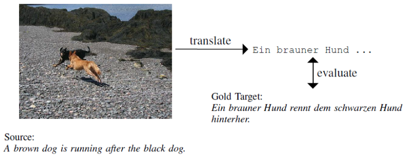
translate
Ein brauner Hund ...
evaluate
Gold Target:
Ein brauner Hund rennt dem schwarzen Hund hinterher.
Source:
A brown dog is running after the black dog.

Figure 1: An example of the multimodal translation. (Specia et al., 2016b)

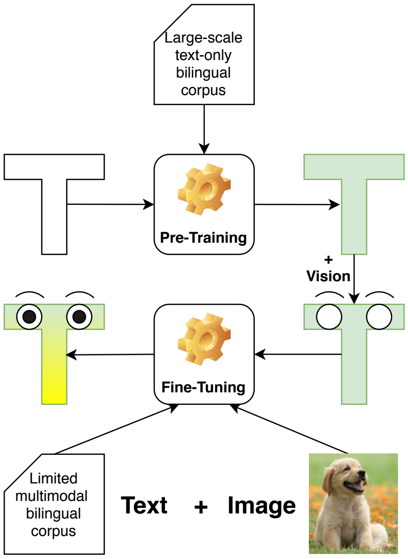
Large-scale text-only bilingual corpus
Pre-Training
+ Vision
Fine-Tuning
Limited multimodal bilingual corpus
Text
+
Image

Figure 2: The multimodal transfer learning solution. 1) Initialize a vanilla Transformer. 2) Train the model with large-scale parallel corpus. 3) Add visual related components. 4) Fine-tune the model on limited multimodal corpus.

scale of parallel dataset of MMT task is not enough to train a robust MMT model. Compared with the translation corpus on news such as Common Crawl and UN corpus, commonly-used MMT dataset Multi30k (Elliott et al., 2016) is too small to train large-capacity models with millions of parameters. Therefore, it is imperative to put efforts on methods in low-resource MMT.

For the text-only NMT tasks, the Transformer (Vaswani et al., 2017) provides a novel architecture on language generation which supersedes RNN architectures rapidly with enhanced parallelizability. Meanwhile, the framework of pre-training and fine-tuning becomes a standard pipeline since BERT (Devlin et al., 2019) achieved the SOTA performances over a bunch of natural language understanding tasks. This to some extent suggests that transfer learning could effectively solve NLP tasks which requires deep understanding on the semantics but have limited size of in-domain data.

Therefore, in this paper, we will investigate whether it's feasible to apply transfer learning to MMT task, i.e. transferring the prior knowledge learned from monomodal task into a multimodal task, as shown in Figure 2. The contribution of our work can be summarized as follows:

- We propose the Visual Language Transformer (VLTransformer) which is compatible for both monomodal and multimodal inputs. The model achieves competitive results on

Multi30k En-De and En-Fr tasks.

- We present a method of fine-tuning a pretrained monomodal MT model in the multimodal MT task, which is implemented by appropriately masking elements in both modalities to encourage the model to make full use of the input information.

2 Related Work

There are a spectrum of prior works investigating MMT. (Caglayan et al., 2016; Calixto and Liu, 2017) used standard RNN encoder-decoder with attention (Bahdanau et al., 2015) to fuse textual and visual features. Both of them employed pretrained image classification models like VGG and ResNet to extract visual features and combine with textual features with different schemes of attentions. Imaginet is proposed to predict the visual feature conditioned on textual inputs, which is used to improve the quality of the representation of contexts (Elliott and Kádár, 2017), where they decompose the MMT task into two sub-tasks where each can be trained separately with large external corpus. Hirasawa et al. (2019) extends the work of Imagination by converting the decoding process into a similarity based searching between the predicted embedding and the embedding of the vocabulary, which is achieved by optimizing a marginal loss on pre-trained word embeddings with predicted word embeddings.

Besides, (Specia et al., 2016b; Elliott et al., 2017; Barrault et al., 2018) make comprehensive summaries on the MMT tasks from MMT 2016 to 2018, which shows two major findings from the task: 1). The effectiveness of the additional modality is still questionable or limited, which encourages researchers to go further on the usage of visual information. 2). Fine-grained evaluation metrics have to be adopted to evaluate the true impact of the multimodality.

There are still some impressive works built upon Transformer-based architecture. MeMAD (Grönroos et al., 2018) achieves the best performance on flickr16 and flickr17 test sets with a multimodal Transformer model, which is pre-trained on massive out of domain data including OpenSubtitles and MS-COCO captions. They perform comprehensive experiments on the model with different data and model settings. (Zhang et al., 2020) proposes the method named universal visual retrieval which builds a look up table from topic word and image with TF-IDF. Before translation, m images are retrieved from the image set. Then, visual features will be aggregated with textual features to produce the hidden states. The UMNMT proposed in (Su et al., 2019) makes it possible to train a MMT model with bilingual but non-paired corpus and images. In their work, each language has an encoder and a decoder but shares one image encoder. They use the cycle-consistency loss to train the model by translating the text into target language, then, recover it back.

In summary, many approaches are proposed to tackle the MMT task from following two direction:

- Improve the architecture of the model to make better use of visual modality.

- Leveraging external resources, monolingual or monomodal resources to enhance the performance.

However, we find that the pre-training and fine-tuning framework is under-investigated for MMT tasks, especially the cross-modal pre-training, which motivates us to explore in this work.

3 VLTransformer

First of all, we briefly review the architecture of Transformer (Vaswani et al., 2017). In the transformer, source texts are fed into the encoder and transformed into vectors with the word embedding and positional embedding, then, N layers of multi-head attention blocks are applied

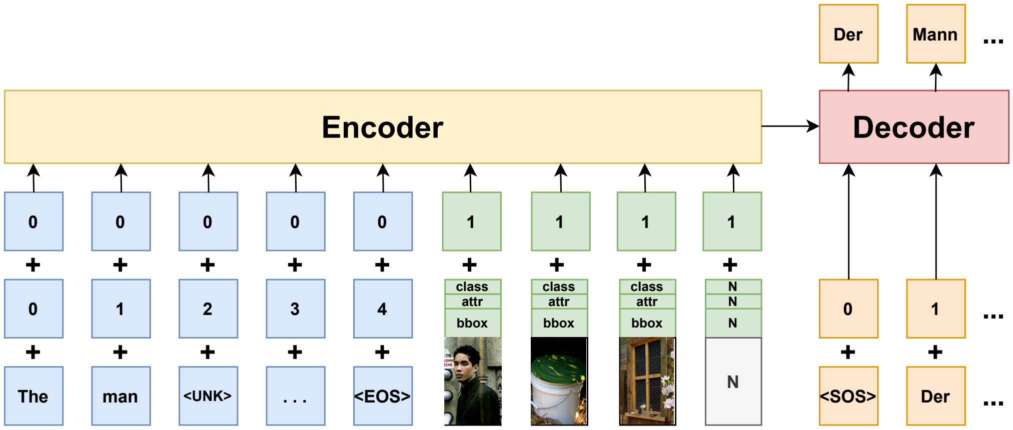

Figure 3: This figure shows the architecture of the proposed VLTransformer. For the textual inputs, three rows (from bottom to top) represent for word embedding, positional encoding and type embedding, respectively. For image inputs, two rows represent for the summation of 4 groups of transformed visual features (pooled ROI, bounding box, attributes and class) and the type embedding respectively. The decoder remains unchanged comparing with original Transformer. The $<$ unk $>$ and the N feature vector are cross-modal masks, being exclusively appears in one modality and are controlled by τ and p.

to produce the hidden states $\mathbf{H}$. For the decoder, the previously generated tokens until step t will be fed into the decoder to interact with the context $\mathbf{H}$ to predict the token of step $t + 1$. More formally, the encoding and decoding process is denoted as follows:

$$\begin{aligned}
\mathbf{E}_S &= \mathrm{We}_S(\mathbf{X}) + \mathrm{Pe}_S(\mathbf{X}) && (1)\\
\mathbf{H}_S &= \mathrm{MHA}_{\text{encoder}}(\mathbf{E}_S) && (2)\\
\mathbf{E}_T &= \mathrm{We}_T(\mathbf{Y}_{[:t]}) + \mathrm{Pe}_T(\mathbf{Y}_{[:t]}) && (3)\\
\mathbf{H}_T &= \mathrm{MHA}_{\text{decoder}}(\mathbf{E}_T, \mathbf{H}_S) && (4)\\
y_{t+1} &= g(h_{T,t}) && (5)
\end{aligned}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are source and target tokens, $\mathbf{E}_S, \mathbf{H}_S$ and $\mathbf{E}_T, \mathbf{H}_T$ represent for embeddings and hidden states of source and target texts respectively. We and Pe are word embeddings and positional embeddings. MHA represents for the Multi-head Attention blocks. y_{t+1} is the predicted token comes from the transformation of the last hidden state $h_{T,t}$.

3.1 Image Embedding

To create high quality visual features, we use the Bottom-Up and Top-Down Attention (BUTD) (Anderson et al., 2018) to extract image features. Specifically, the Bottom Up attention of BUTD is based on Faster R-CNN (Ren et al., 2015) for object detection. They pre-train the model on the Visual Genome (Krishna et al., 2017) dataset which has fine-grained labels of objects with 1600 object classes and 400 object attributes. The extracted features are used as follows in the MMT model:

$$\mathbf{V} = \phi_{\text{ROI}}(\mathbf{V}_{\text{ROI}}) + \phi_c(\mathbf{V}_c) + \phi_a(\mathbf{V}_a) + \phi_{\text{bbox}}(\mathbf{V}_{\text{bbox}}) \qquad (6)$$

where the pooled ROI features are represented by $\mathbf{V}_{\text{ROI}} \in \mathbb{R}^{m \times d_{\text{ROI}}}$, $d_{\text{ROI}} = 2048$ in the experiment, m is the number of detected objects. $\mathbf{V}_c \in \mathbb{R}^{m \times 1600}$ are predicted class one-hot vectors which will be multiplied with an embedding matrix in the experiment. $\mathbf{V}_a \in \mathbb{R}^{m\times 400}$ are attribute class one-hot vectors, and the bounding boxes $\mathbf{V}_{\text{bbox}} \in \mathbb{R}^{m\times 4}$ represents for normalized coordinates (x_0, y_0, x_1, y_1) of detected objects. Coordinates are normalized into $[0, 1]$ with the size of the image, i.e. $x/x_{\text{img}}, y/y_{\text{img}}$. ϕ represents for linear transformations to scale the dimensionality along with the original Transformer d_{model}. The summation of 4 types of features simultaneously encodes most of necessary visual information, which is more fine-grained and informative comparing with previous works (Elliott and Kádár, 2017; Zhou et al., 2018; Caglayan et al., 2016) which only uses pooled ResNet (He et al., 2016) features or pooled object embeddings (Grönroos et al., 2018).

3.2 Fusion of Image and Text

In order to take the advantage of pre-trained NMT models and avoid overfitting using large-capacity network with limited multimodal labelled training data, we introduce parameters that needs to be trained from scratch as few as possible into the model. Therefore, instead of using architectures like LXMERT (Tan and Bansal, 2019) and the model proposed in (Zhang et al., 2020), where large sets of newly initialized parameters will be introduced into an independent image encoder, we share the original encoder layers of the Transformer to encode both modalities by directly concatenating the visual and the textual features. More specifically:

$$\mathbf{E}_S = \text{We}_S(\mathbf{X}) + \text{Pe}_S(\mathbf{X}) + \text{Te}(\mathbf{X}) \tag{7}$$

$$\mathbf{V} = \mathbf{V} + \text{Te}(\mathbf{V}) \tag{8}$$

$$\mathbf{E}_{S,V} = [\mathbf{E}_S; \mathbf{V}] \tag{9}$$

where the Te represents for newly introduced type embedding inspired by the Next sentence prediction (NSP) of BERT (Devlin et al., 2019), which uses 0 for text and 1 for vision. $\mathbf{E}_S$ is the replacement of Eq. 1. Finally, we concatenate embeddings of tokens and objects along the length dimension, as described in Figure 3. The sequence length becomes the summation of token number and detected objects number, $|\mathbf{E}_{S,V}| = |\mathbf{V}| + |\mathbf{E}_S|$.

In such case, we only introduce a few amount of parameters to incorporate vision features, which reduces the perturbation on the Transformer Encoder and Decoder. In the experiment, we find that this can significantly improve the training efficiency on the small dataset. In addition, compared with the cross-attention method (i.e. H=SelfAttn(Token, Vision, Vision) which maps visual information onto token representations), concatenation reserves complete contexts in both modalities for the decoder, which is not limited by the length of source sentence.

3.3 Cross Modal Masking

In experiment, compared to using text-only inputs, we find that directly fine-tuning the pre-trained transformer on multimodality inputs can't obtain extra performance boosts, which motivates us to investigate the reason behind that. Observing the attention map of encoder-decoder attention weights, we find that the model only assigns weights to text representations and entirely ignores visual information.

To force the model fully exploit both two modalities: text and image, we propose a cross modal masking (CMM) method to train the model with complementary information by partially masking out some inputs in one of any modality. Specifically, we randomly choose a modality to mask following the Bernoulli distribution, and then, randomly mask q tokens or q objects within specific modality. The masked token will be replaced by special token "¡unk¿" and the masked image region will be replaced by a noisy vector sampled from the standard normal distribution. This method is inspired by the masked language model (Devlin et al., 2019) and (Chen et al., 2020). Differently, they use the masking for unsupervised pre-training, while we use it directly in the translation task without predicting the masked place. Thus, masking here

Method	test 2016				test 2017			
	En-De		En-Fr		En-De		En-Fr	
	B	M	B	M	B	M	B	M
WMT16_MMT_Winner (Specia et al., 2016b)	34.2	53.2	-	-	-	-	-	-
WMT17_MMT_Winner (Elliott et al., 2017)	-	-	-	-	33.4	54	55.9	72.1
Imagination (Elliott and Kádár, 2017)	36.8	55.8	-	-	-	-	-	-
NMTUVR (Zhang et al., 2020)	36.94	-	57.53	-	28.63	-	48.46	-
UMONS (Delbrouck and Dupont, 2018)	40.34	59.58	62.49	76.83	32.57	53.6	55.13	71.52
MeMAD (Grönroos et al., 2018)	45.09	-	68.30	-	40.81	-	62.45	-
Pretrained Trans (baseline)	41.2	59.69	46.3	65.9	37.9	56.3	48.3	65.8
Fine-tuned Trans	45.6	62.9	65.7	79.2	42.7	60.1	60.8	75.9
VLTransformer (ours)	46.2	63.5	65.4	78.8	43.6	60.4	62.0	76.3
VLTransformer + CMM (ours)	**48.1**	**64.7**	**68.7**	**81.5**	**44.0**	**61.3**	**63.5**	**77.3**

Table 1: The experimental result of the Multi30k dataset on test-2016 and test-2017 En-De and En-Fr tasks. First six rows are results of previous works including the 2016 and 2017 winner, widely used Imagination, the 2018 MMT task participants MeMAD and UMONS, as well as the newly proposed Transformer based model NMTUVR. Last four rows are our ablation studies including the un-fine-tuned Transformer, fine-tuned Transformer and the VLTransformer with and without cross-modal masking (CMM). We can see that on 4 test sets, the VLTransformer with CMM is consistently better than the text-only model and the model trained without CMM. Note that B and M represents for BLEU and METEOR, Trans represents for Transformer.

only acts like the noise introduced in denoising autoencoder, it forces the model to learn by predicting unknown tokens and recover the corrupted vectors. We find it effectively prevents the model from neglecting visual contexts by CMM in training. See Figure 3 for more intuitive details.

4 Experiment

4.1 Dataset

In the experiment, we use the Multi30k (Elliott et al., 2016) dataset to evaluate our method. The sizes of the dataset are 29000:1014:1000:1000 for training, validation, test2016 and test2017 set, each instance in form of triples (source, target, image). English descriptions are provided as source texts, German and French corpus are provided as target texts. All corresponding images are from Flickr30k (Young et al., 2014) dataset. We use the Moses toolkit (Hoang and Koehn, 2008) to pre-process the data with lowercasing, tokenizing and punctuation normalization.

For image features, we use BUTD (Anderson et al., 2018) to extract 4 groups of features for each object, including pooled ROI feature vector, object class, object attribute and bounding box. Maximum of 36 detected objects are reserved with the prediction probability higher than 0.5. The BUTD model is not fine-tuned in the translation task.

4.2 Setup

We use the pre-trained transformer model provided by fairseq (Ott et al., 2019) which is implemented with PyTorch (Paszke et al., 2019). The En-De model (Transformer-Large) is trained on WMT'19 corpus and En-Fr (Transformer-Big) model is trained on WMT'14 corpus. Both models share the vocabulary between source and target language, resulting in sizes of 42020 and 44508 for En-De and En-Fr vocabularies. The parameters of the embedding layer as well as the output projection layer are also shared for the encoder and the decoder in both models. The BPE (Sennrich et al., 2016) is applied to create the vocabulary. The model of En-De is slightly larger (270M) than the En-Fr (222M) model, because of the difference of the dimen-

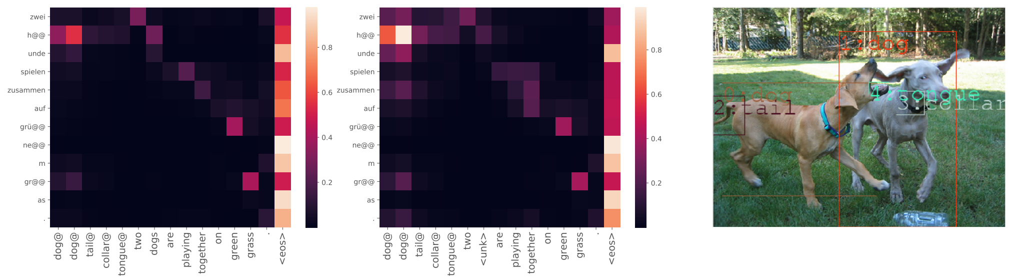

Figure 4: This figure shows an example of the attention map between source inputs and target tokens in En-De MMT translation. The X axis for left 2 plots are source inputs, where visual features are represented by the object class (token with an @ in the end, only 5 high score objects are preserved as shown in the right plot. The order for visual and text inputs are changed for more clearance). The difference between the left and the middle plot is that the **cross-modal masking is performed on the middle one** where ”dog” is deliberated replaced by $< \text{unk} >$. We can see that when the ”dog” is masked, the model pays more attention on the visual features of two detected dogs.

sionality of the FFN block (8192 for En-De and 4096 for En-Fr). Apart from that, the En-De and En-Fr model have exactly same architectures with hidden size of 1024, 6 × encoders and 6 × decoders. The parameter size of the vision related components are 6M for both model, thereby makes the VLTransformer to have 276M and 228M parameters for En-De and En-Fr, respectively.

During fine-tuning, we use the learning-rate of 1e-4 with 4000 steps of warm-up and inverse-sqrt warm-up strategy. We use 0.3 for dropout probability, 0.1 for label smoothing (Pereyra et al., 2017), Adam (Kingma and Ba, 2015) is used as the optimizer. For the VL-Transformer, we use the parameter of fairseq pre-trained Transformer to initialize the backbone and text related embeddings, vision related parameters are initialized randomly. The model is fine-tuned on a Tesla V100 GPU with fp16 enabled and converges in less than 20 minutes for 10 epochs.

The baseline method is the pre-trained Transformer without fine-tuning. We use BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) as evaluation metrics with lowercased text.

5 Analysis

We compare our results with another six latest methods in Table 1. As the goal of newly-proposed NMTUVR (Zhang et al., 2020) is to improve universal NMT with multimodality, direct comparison with ours is unfair. As expectation, the pre-trained Transformer set a very strong baseline, which demonstrate that a well-trained text-only NMT model has been able to produce satisfying translations in the absence of word and phrase ambiguitity. At the same time, the profit of fine-tuning the Transformer is significant, even with only textual inputs. For the VLTransformer, the model trained without CMM is already better than the text-only method, which could demonstrates the effectiveness of visual contexts, in addition, the model trained with CMM is consistently better than the model without CMM, which demonstrates that CMM is a key point to improve the cross-modal interaction. Comparing with the MeMAD (Grönroos et al., 2018) which uses massive of external multimodal corpus (OpenSubtitles and MS-COCO), we only use the officially published training set for fine-tuning which is more efficiency.

Figure 4 is an example of the En-De translation from a VLTransformer model trained with

CMM. We filter 5 high score objects to investigate the alignment between target tokens and source inputs. There is evidence showing that the model is able to attend correct objects (i.e. two dogs) no matter the word ”dog” is appeared in source texts or not (replaced by the $< \text{unk} >$ or not), which means it could translate the sentence with both modality.

Although the attention map looks good, we actually manually amplify the score of visual features, in the experiment, we find that the model is more inclined to get contextual information from text instead of image although we have already used cross-modal masking. Some reasons can be speculated: 1) The size of training data is relatively small which means the newly initialized visual related parameters can not be fully trained. 2) We investigate the extracted detected objects and find out that there are mistakes in the detection which actually leads noise into the model.

6 Conclusion

We propose a cross-modal transfer learning solution to take full advantage of pre-trained monomodal model in the multimodal task. The approach of CMM to incorporate visual information into translation achieves remarkable results in the MMT tasks evaluated on Multi30k dataset, which reveals that prior knowledge of monomodal data can be transferred in a multimodal model even if fine-tuning on limited multimodal data. Furthermore, the shared encoder demonstrates perfect compatibility with the newly introduced visual features, which encourages us to dig into methods for visual and textual alignment with Transformer architectures. To sum, we show the evidence that our model is able to decode from both modalities after fine-tuning with the cross-modal masking method.

References

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.

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*.

Banerjee, S. and Lavie, A. (2005). METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72.

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 *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 304–323.

Caglayan, O., Barrault, L., and Bougares, F. (2016). Multimodal attention for neural machine translation. *CoRR*, abs/1609.03976.

Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. (2019). Probing the need for visual context in multimodal machine translation. In *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, I. and Liu, Q. (2017). 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, Copenhagen, Denmark, September 9-11, 2017*, pages 992–1003.

Chen, Y., Li, L., Yu, L., Kholy, A. E., Ahmed, F., Gan, Z., Cheng, Y., and Liu, J. (2020). UNITER: universal image-text representation learning. In Vedaldi, A., Bischof, H., Brox, T., and Frahm, J., editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Delbrouck, J. and Dupont, S. (2018). UMONS submission for WMT18 multimodal translation task. *CoRR*, abs/1810.06233.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *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 4171–4186.

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 *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pages 215–233.

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*, pages 70–74.

Elliott, D. and Kádár, Á. (2017). Imagination improves multimodal translation. In *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.

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 *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 603–611.

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.

Hirasawa, T., Yamagishi, H., Matsumura, Y., and Komachi, M. (2019). Multimodal machine translation with embedding prediction. In *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 3-5, 2019, Student Research Workshop*, pages 86–91.

Hoang, H. and Koehn, P. (2008). Design of the moses decoder for statistical machine translation. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing@ACL 2008, Columbus, Ohio, USA, June 20, 2008*, pages 58–65.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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. *International Journal of Computer Vision*, 123(1):32–73.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

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.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In Wallach, H., Larochelle, H., Beygelzimer, A., d' Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, L., and Hinton, G. E. (2017). Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*.

Ren, S., He, K., Girshick, R. B., and Sun, J. (2015). Faster R-CNN: towards real-time object detection with region proposal networks. In *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.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016a). A shared task on multimodal machine translation and crosslingual image description. In *First Conference on Machine Translation, Volume 2: Shared Task Papers*, WMT, pages 540–550, Berlin, Germany.

Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016b). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 543–553.

Su, Y., Fan, K., Bach, N., Kuo, C. J., and Huang, F. (2019). Unsupervised multi-modal neural machine translation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10482–10491.

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 and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110.

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 *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.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.

Zhang, Z., Chen, K., Wang, R., Utiyama, M., Sumita, E., Li, Z., and Zhao, H. (2020). Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

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