Probing Multi-modal Machine Translation with Pre-trained Language Model

Yawei Kong, Kai Fan
Alibaba DAMO Academy
{yawei.kyw,k.fan}@alibaba-inc.com

Abstract

Multi-modal machine translation (MMT) aimed at using images to help disambiguate the target during translation and improving robustness, but some recent works showed that the contribution of visual features is either negligible or incremental. In this paper, we show that incorporating pre-trained (vision) language model (VLP) on the source side can improve the multi-modal translation quality significantly. Motivated by BERT, VLP aims to learn better cross-modal representations that improve target sequence generation. We simply adapt BERT to a cross-modal domain for the vision language pre-training, and the downstream multi-modal machine translation can substantially benefit from the pre-training. We also introduce an attention based modality loss to promote the image-text alignment in the latent semantic space. Ablation study verifies that it is effective in further improving the translation quality. Our experiments on the widely used Multi-30K dataset show increased BLEU score up to 6.2 points compared with the text-only model, achieving the state-of-the-art results with a large margin in the semi-unconstrained scenario and indicating a possible direction to rejuvenate the multi-modal machine translation.

1 Introduction

Joint models of language and vision have achieved remarkable results, such as in image caption (Karpathy and Fei-Fei, 2015) and visual question answering (Antol et al., 2015). Multi-modal machine translation (MMT) was first introduced as a shared competition task at the 2016 Conference on Machine Translation (WMT16) (Specia et al., 2016) as an interdisciplinary study to incorporate a visual element into the multilingual translation task. This task continued for three years until WMT18, and the findings presented by the organizers suggest that the text-only systems remain competitive, and that the contribution of visual modality is not entirely convincing (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). Moreover, the experiments in (Elliott, 2018) find that a publicly available MMT system produces great translations with random, incongruent images, further undermining the importance of visual features. The empirical results have so far raised doubts about whether the visual features can really help MMT, and there is evidence pointing to a negative answer.

Data used	img	src	tgt	examples
Multi-30K	✓	✓	✓	most works
+external data	✓	✓		(Grönroos et al., 2018)
		✓	✓	(Helcl et al., 2018)
	✓	✓		(Yin et al., 2020), ours

Table 1: Different unconstrained scenarios in MMT.

We hypothesize that one reason is the data scale of the benchmarking Multi-30K (Elliott et al., 2016) – it is likely insufficient for a deep model to learn better cross-modality or cross-lingual representations. However, the pre-training techniques such as BERT (Devlin et al., 2019) or cross-lingual language model (XLM) (Conneau and Lample, 2019) can capture rich representations of the inputs from languages and be applied to various downstream tasks by providing context-aware embeddings, leading to remarkable improvements even on small datasets. Furthermore, the pre-trained vision and language model LXMERT (Tan and Bansal, 2019) pioneers the cross-modality pre-training and sets an influential record in vision and language reasoning tasks. These advances lead us to believe that a better cross-modality representation can help multi-modal machine translation as well.

In this work, we discuss the unconstrained scenario of MMT, but unlike previous setting in most WMT 2018 submissions (Grönroos et al., 2018; Helcl et al., 2018), we did not include any external data of the parallel source and target textual corpus. Since we want to incorporate a pre-trained (vision)

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3689–3699

language model as an encoder backbone into the transformer architecture (Vaswani et al., 2017) for neural machine translation, our used external data only contains the images and the source texts.

Particularly, our model is initialized with the widely used BERT, and pre-trained on large scale image-text dataset (about six million pairs), expecting to learn a better cross-modality representation between the image and the source language. Next, we stack a regular transformer decoder on top of the pre-trained (vision) language model and proceed to the task of MMT. Meanwhile, we design another modality loss in addition to the traditional sequential cross entropy loss. The modality loss is to minimize the difference between source-target cross attention and image-target cross attention. Intuitively, minimizing this loss function can promote the modality alignment among the three possible pairwise configurations in the latent semantic space. In other words, differences among (source, target), (source, image), and (target, image) alignments can be reduced. Our experimental section also presents a detailed analysis of how each factor separately contributes to the overall gains.

In summary, this paper makes the following contributions. **(1)** We propose to integrate a pre-trained vision language model into multi-modal machine translation, aiming at learning and utilizing better cross-modality representations. **(2)** We address the importance of the modality loss which can further boost the model performance. **(3)** We conduct extensive experiments on the benchmark Multi-30K dataset, and our results outperform strong baselines by a large margin.

2 Related Works

Constrained Scenario Most works like (Calixto et al., 2017; Zhou et al., 2018; Ive et al., 2019; Yao and Wan, 2020) in MMT prefer to use Multi-30K dataset alone. For example, a standard paradigm of MMT explored by many previous works is to simultaneously learn the vision language interaction and the target language generation (Calixto et al., 2017; Zhou et al., 2018; Ive et al., 2019; Yang et al., 2020). However, training on such a limited dataset, the benefits provided by visual features of these methods are quantitatively marginal w.r.t. automatic evaluation metrics BLEU and METEOR.

Unconstrained Scenario In the submissions of WMT 2018 (Grönroos et al., 2018; Helcl et al., 2018) as shown in Table 1, either images / source texts or the source / target texts parallel dataset (or back-translation) are added to improve the model performance. However, as they discovered, training with the large scale parallel textual corpus will shift the machine translation model towards the pure textual domain, further weakening the effect of visual features. The additional target data will also make the fair comparison difficult. A special unconstrained scenario by (Su et al., 2019b) leverages large monolingual language data to pre-train an unsupervised translation model. It considers the cross representation of the source-target in an unsupervised manner, but the image domain is still isolated without proper training.

We will discuss another unconstrained scenario that only allows to use additional images and source texts. Zhu et al. (2019) investigates the representation from pre-trained BERT by feeding it into all layers of a text-only translation model. This work, to a large extent, encourages us to explore how the (vision) language pre-trained model can benefit the MMT. However, we found that a direct architecture of feeding cross-modality representations (from LXMERT) to multi-modal translation model does not work well.

To our best knowledge, Yin et al. (2020) currently achieves the state-of-the-art on Multi-30K. It employed a common encoder-decoder framework by hard-encoding a multi-modal graph to guide the learning of the image-text cross attention, where the graph structure is annotated by a pre-trained visual grounding model (Yang et al., 2019). The external data is not explicitly used in this work, but the pre-trained visual grounding model uses BERT as part of its backbone. Instead of relying on a pre-defined graph to prevent the attention between the word and visual feature without connection, we obtain a soft cross attention from large-scale vision-language data pre-training. It is also worth mentioning that we make the BERT based visual grounding and multi-modal machine translation into an end-to-end trainable architecture.

3 Our Method

3.1 Initial Trial

The overall architecture of our proposed approach is based on the commonly used transformer (Vaswani et al., 2017), which is the basic unit of most pre-trained (vision) language model. Our initial experiment is to adopt pre-trained (vision) language model as the encoder. The baseline is to train

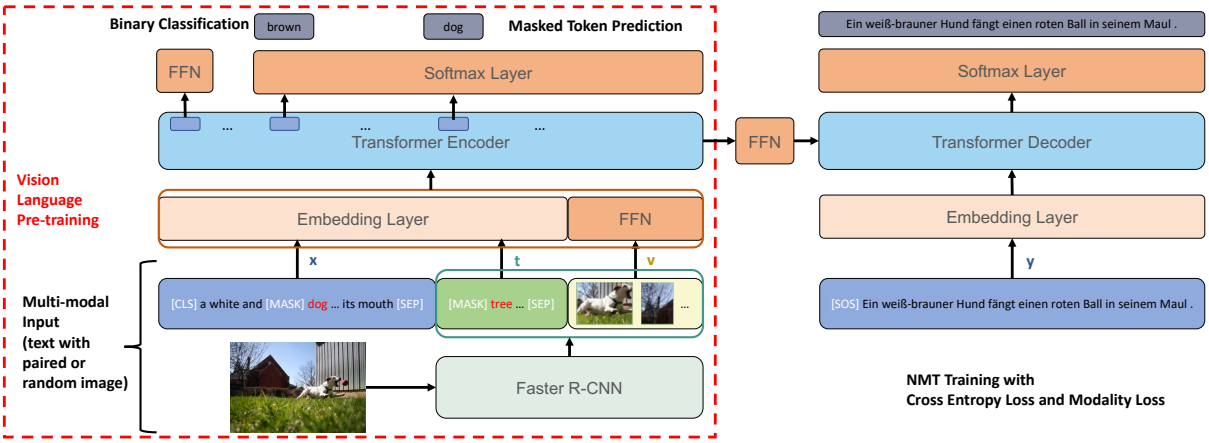

Figure 1: The overall architecture of our proposed multi-modal NMT with pre-trained vision language model. Note that the [MASK] tokens and random images are merely applied during vision language pre-training.

Encoder	visual feature	Test2016 EnDe		Test2016 EnFr	
		BLEU	Meteor	BLEU	Meteor
Transformer	-	38.3	56.6	59.6	74.6
BERT	-	39.1	57.1	61.0	75.3
LXMERT	✓	37.4	55.2	57.7	68.6

Table 2: BERT/LXMERT are frozen.

a transformer NMT from scratch. The first competitive system is simply BERT, and the second one is the pre-trained vision language model LXMERT. LXMERT claimed that the initialization with pre-trained BERT will harm the performance of their downstream tasks. Table 2 shows the preliminary results indicating that the pre-trained LXMERT as the encoder performs surprisingly worse than text-only BERT. Does the table suggest that the visual features are equally marginalized in MMT equipped with pre-trained language model? However, since BERT encoder can bring more improvements, we can abandon LXMERT's conclusion and return to the paradigm with BERT initialization.

3.2 Vision Language Pre-training (VLP)

Ive et al. (2019) finds that integrating both object-based embedding features and image features into the NMT model results better performance in human evaluation on comprehensibility. We therefore favor the object-semantics alignment whose interaction is composed of text embedding, object tag embedding and object image features.

We visualize the training rationale of the VLP in the red dashed box of Figure 1. Suppose that an image and its description $\mathbf{x}$ are presented as the input, where $\mathbf{x}$ represents a sequence of n tokens $(x_1, ..., x_n)$, i.e., the sentence of the source language in our following NMT system. We first process the image with the efficient object detection model Faster-RCNN (Ren et al., 2015) to detect the object regions, box positions, object tags and attribute tags. Particularly, two sets of features are extracted. One is the image visual features of all detected objects, denoted as $\mathbf{v}$. The other is the classification tags of the corresponding objects, denoted as $\mathbf{t}$, as textual features.

Since the backbone of our transformer encoder is pre-trained BERT, the input text $\mathbf{x}$ and object tags $\mathbf{t}$ are both language tokens that can be easily concatenated. However, there is a dimensionality mismatch between the BERT embedding layer and the visual features. For dimension reduction, a fully-connected layer is necessary with input $\mathbf{v}$, and its task is to learn cross modality transferring. The final input fed into the multiple transformer layers of BERT can be written as follows.

$$\text{Cat}\left[\text{Emb}\left(\text{Cat}[\mathbf{x}, \mathbf{t}]\right), \text{FFN}(\mathbf{v})\right] \quad (1)$$

We now face two similar tasks as BERT.

Task 1: Masked LM Same as the standard BERT, our training objective employs the masking token prediction, where 15% of the input text tokens are randomly selected and replaced with the special token [MASK]. Then, only the masked token will be predicted.

Task 2: Paired Image Prediction Analogous to the standard BERT, we pre-train the binarized paired image prediction task that mimics predicting the next sentence, where the training data can be trivially generated for each batch. Specifically, for a given text input, we choose its paired image or

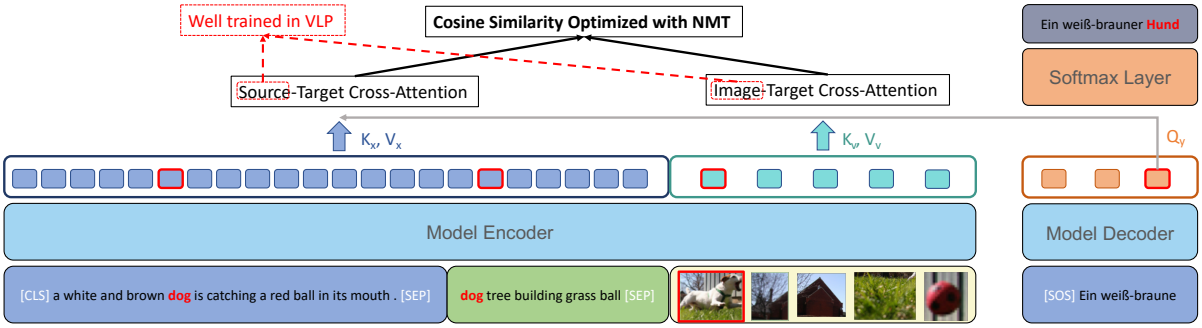

Figure 2: The visualization of modality loss for an input sentence-image pair. It exemplifies the computational flow of the modality loss w.r.t. the last layer of the decoder when decoding "Hund" in German.

a random image each with probability 50%. The output vector of the first special token [CLS] is used as the aggregate multi-modal representation for this classification task.

3.3 Multi-modal NMT

Once the vision language model has been fully trained on a large paired image-text dataset, it is reasonable to assume that the obtained cross-modality representations between the source text and the image are more powerful than those training on the limited Multi-30K. The (key, value) pairs of both the textual and visual features participate in the dot-product attention of the transformer decoder. But there is another dimensionality mismatch between the BERT output and the decoder hidden size. To close this gap, we append an additional fully connected layer after the last layer of BERT. In this section, we also introduce a novel modality loss that is potential to benefit the multi-modal representation learning while but incurs only a few extra model parameters.

Modality Loss To train a multi-modal machine translation task, i.e., generating the tokens in the target language $\mathbf{y} = (y_1, ..., y_m)$, a common objective is the sequential cross entropy loss $\mathcal{L}_{\text{XENT}} = -\sum_{j=1}^{m} \log p(y_j|\mathbf{y}_{<j}\mathbf{x}, \mathbf{v})$, which is the sum of the negative log-likelihoods of the auto-regressive text generation task. Our proposed auxiliary modality loss can be intuitively depicted as Figure 2.

Concretely, when generating the j-th token in the target, the output textual and visual (key, value) pairs from the encoder are separately used to compute the cross-lingual and cross-modality attention with the query vector of the l-th layer in the decoder. The derived vectors can be written as follows.

$$\mathbf{h}_{x,j}^{(l)} = \text{Softmax}\left(K_x \mathbf{q}_j^{(l)} / \sqrt{d}\right) V_x \qquad (2)$$

where d is the hidden size of the model decoder, and similar attention holds for visual features $\mathbf{h}_{v,j}^{(l)} = \text{Softmax}\left(K_v \mathbf{q}_j^{(l)} / \sqrt{d}\right) V_v$. Thus, the modality loss can be represented as

$$\mathcal{L}_{\text{M}}^{(l)} = \sum_{j=1}^{m} (1 - \cos(\mathbf{h}_{x,j}^{(l)}, \mathbf{h}_{v,j}^{(l)})) \qquad (3)$$

where the cosine similarity is defined as $\cos(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \mathbf{b}}{\|\mathbf{a}\|, \|\mathbf{b}\|}$. Consequently, the overall training objective is a weighted combination of two loss functions.

$$\mathcal{L} = \mathcal{L}_{\text{XENT}} + \sum_{l=1}^{L} \lambda^{(l)} \mathcal{L}_{\text{M}}^{(l)} \qquad (4)$$

where L is total number of transformer layers in decoder. Empirically, we found that only using the modality loss of the last layer is sufficient to improve the model performance. Intuitively, the query vector will be directly fed into the softmax layer for decoding the target tokens, making the last layer more informative than other remote layers.

A common method of choosing the weighting parameter λ is to run cross validation on the held-out development data. For the task at hand, this is a time-consuming process. We instead discard the layer-wise $\lambda^{(l)}$ in Eq. (4) and introduce a self-tuning module with respect to the generation process of every single target token. Mathematically, the refined modality loss can be formulated as,

$$\widetilde{\mathcal{L}}_{\text{M}}^{(l)} = \sum_{j=1}^{m} \lambda_j^{(l)} (1 - \cos(\mathbf{h}_{x,j}^{(l)}, \mathbf{h}_{v,j}^{(l)})). \qquad (5)$$

where the token level λ_j is learnable and derived from a feedforward neural network.

$$\lambda_j^{(l)} = \text{Sigmoid}(\mathbf{w}_y^\top \text{Emb}(y_j) + \mathbf{w}_x^\top \mathbf{h}_{x,j}^{(l)} + \mathbf{w}_v^\top \mathbf{h}_{v,j}^{(l)})$$

where $\mathbf{w}_y, \mathbf{w}_x, \mathbf{w}_v$ are three d-dimensional vectors shared cross different decoder layers and required

Algorithm 1 Training Pipeline

Require: Image, source text paired data $\mathcal{D}_{\text{VLP}}$; Image, source/target text triple data $\mathcal{D}_{\text{MMT}}$.

1: Initialize the transformer encoder of NMT with pre-trained BERT.
2: Pre-train the transformer encoder on $\mathcal{D}_{\text{VLP}}$ with masked language model task and pair image prediction task.
3: Extract the image, source text paired from $\mathcal{D}_{\text{MMT}}$.
4: Continue the vision language pre-training on above extracted data.
5: Freeze the transformer encoder, and optimize other parameters on $\mathcal{D}_{\text{MMT}}$ with cross entropy loss and modality loss until convergence.
6: Optimize all model parameters on $\mathcal{D}_{\text{MMT}}$ with cross entropy loss and modality loss until convergence.

to jointly optimize with the model parameters, but useless during inference. We expect the model to dynamically adjust the weight parameters of the tokens with different importance. For example, there is a good chance that the content words also appear as detected objects by the Faster R-CNN. If the term $\mathbf{w}_v^\top \mathbf{h}_{v,j}^{(l)}$ can positively increase its scale for such words, the corresponding λ_js become larger and therefore reinforce the maximization of the cosine similarity. In contrast, although it happens that a mapping exists between the source and target functional words, the image-target cross attention may become weak, making it less necessary to promote the similarity. The term $\mathbf{w}_x^\top \mathbf{h}_{x,j}^{(l)}$ is intended to model the importance of the source contribution. Our results, however, show that in the current experiment setup its effect is not quite as significant.

3.4 Two-Stage Training

When BERT is applied to the downstream tasks, the task-specific module parameters are usually plugged into BERT and all the trainable parameters are simultaneously fine-tuned (Devlin et al., 2019). However, we found this is not the optimal strategy of training our downstream task – multi-modal machine translation. The large number of untrained parameters in the transformer decoder almost account for half of the model size. We conjecture that the encoder parameters have already reached a flat plateau after the pre-training, and it is difficult to set the consistent optimization hyper-parameters (such as learning rate, decay rate or warm-up steps) for both the encoder and decoder.

Therefore, we adopt a two-stage training schedule. In the first stage, the encoder parameters are frozen and only the decoder parameters are optimized w.r.t. the cross entropy and modality loss. In the second stage, all model parameters become trainable and are updated concurrently. This step simulates the regular BERT fine-tuning procedure, and its convergence is expected to lead to a better performance. To this end, we have elaborated the key ideas of our proposed method and summarize the training pipeline of the entire model training process in Algorithm 1.

4 Experiments

In this section, we describe the datasets, the detailed settings as well as the compared baselines.

4.1 Datasets and Settings

Multi-30K We conduct experiments on the Multi-30K dataset (Elliott et al., 2016), where each image is paired with one English(En) description and human translations of German(De) and French(Fr). It has 29,000 instances for training and 1,014 instances for development. Besides, we evaluate our model on various testing sets, including the Multi-30K 2016 test set, the WMT17 test set and the ambiguous MSCOCO test set, which contain 1,000, 1,000 and 461 instances, respectively.

External Data We use about 6 million image and English text paired data for our vision language model pre-training, including MSCOCO (Lin et al., 2014), Im2text (Ordonez et al., 2011), visual7w (Zhu et al., 2016), VQA 2.0 (Goyal et al., 2017), Conceptual captions (Sharma et al., 2018), GQA (Hudson and Manning, 2019). We first process the image with a popular off-the-shelf Faster-RCNN toolkit[1] (Ren et al., 2015; Anderson et al., 2018; Wu et al., 2019). The Faster R-CNN (Ren et al., 2015) network is pre-trained on the MSCOCO dataset and fine-tuned on the Visual Genome (Krishna et al., 2017) dataset to detect salient visual objects, where the number of visual objects ranges from 10 to 100 with the highest prediction probability and 2048 is the dimension of the flattened last pooling layer in the ResNet (He et al., 2016) backbone. Then, we obtain the position-sensitive

[1] https://github.com/airsplay/py-bottom-up-attention

Model	En⇒De						Notes on external resources
	Test2016		Test2017		MSCOCO		
	BLEU	Meteor	BLEU	Meteor	BLEU	Meteor	
Our text-only	38.3	56.6	30.3	51.0	28.6	47.7	Our own implemented transformer
Doubly-Att	36.5	55.0	-	-	-	-	
Fusion-conv	37.0	57.0	29.8	51.2	25.1	46.0	
Trg-mul	37.8	57.7*	30.7	52.2*	26.4	47.4	
VAG	31.6	52.2	-	-	-	-	Constrained methods
VMMT	37.7	56.0	30.1	49.9	25.5	44.8	ResNet features only
DNetwork	38.0	55.6	-	-	-	-	
Multimodal-Att	38.7	55.7	-	-	-	-	
Semi-unconstrained methods							
VMMT	38.4	58.3	-	-	-	-	few Back-translation data
Multimodal-Att	39.5	56.9	-	-	-	-	few Back-translation data
Graph-Fusion	39.8*	57.6	32.2*	51.9	28.7*	47.6*	BERT(en), visual grounding tool
Our Model	**42.7**	**60.7**	**35.5**	**54.9**	**32.8**	**52.2**	BERT(en), images-en
WMT 2018 unconstrained methods							
MeMAD	45.1	-	40.8	-	36.9	-	images-en, OpenSub en-de/fr
CUNI	42.7	59.1	-	-	-	-	images-en, Bookshop en-de/fr, Back-translation

Table 3: Experimental results on the En⇒De MMT. Our results are highlighted in bold. ∗ indicates previous SOTA. B will be short for BLEU and M will be short for Metoer in other tables.

Model	En⇒Fr			
	Test2016		Test2017	
	B	M	B	M
Our Text-only	59.6	74.6	52.7	69.1
Doubly-Att	59.9	74.1	52.4	68.1
Fusion-conv	53.5	70.4	51.6	68.6
Trg-mul	54.7	71.3	52.7	69.5*
VAG	53.8	70.3	-	-
DNetwork	59.8	74.4	-	-
Semi-unconstrained methods				
Graph-Fusion	60.9*	74.9*	53.9*	69.3
Our Model	**65.8**	**79.1**	**58.2**	**73.5**
WMT 2018 unconstrained methods				
MeMAD	68.3	-	62.5	-
CUNI	62.8	77.0	-	-

Table 4: Experimental results on the En⇒Fr MMT.

visual features by concatenating the region features and the corresponding positions.

For the English text, we follow the same pre-processing as the open-source BERT toolkit[2]. The BERT base model with hidden size 768 is utilized as initialization. Note that unlike (Grönroos et al., 2018; Helcl et al., 2018), we never include any external data related to the target languages for both vision language pre-training and machine translation training. For notation simplicity and differentiating their setting, we define our scenario as **semi-constrained**.

[2] https://github.com/huggingface/transformers

4.2 Baselines

We mainly compare with the following representative and competitive frameworks. The constrained methods include **Doubly-Att** (Calixto et al., 2017), **Fusion-conv / Trg-mul** (Caglayan et al., 2017), **VAG** (Zhou et al., 2018), **VMMT** (Calixto et al., 2019) and **Multimodal-Att** (Yao and Wan, 2020). **MeMAD** and **CUNI** (Grönroos et al., 2018; Helcl et al., 2018) mainly discussed the unconstrained scenario of MMT. In addition, VMMT and Multimodal-Att attempted to adding in-domain back-translation data. We prefer to include them into semi-unconstrained methods as well. **Graph-Fusion** (Yin et al., 2020) uses BERT based visual ground model to hard-code a unified multi-modal graph and performs semantic interactions by graph fusion layers, achieving the current state-of-the-art performance.

4.3 Main Results

In Table 3 and 4, we report the main experimental results of our proposed method with previous research works. All reported numbers of our approach are evaluated on the best performed model for the validation set. Note that when optimizing the parameters, we only use the modality loss calculated from the last layer with learnable token level $\lambda_j^{(6)}$. In other words, the reported numbers are obtained by minimizing $\mathcal{L}_{\text{XENT}} + \widetilde{\mathcal{L}}_{\text{M}}^{(6)}$. In the ablation study, we demonstrate this simplification not only reduces the computational complexity, but also achieves better result than our initial proposal.

Both tables show that our multi-modal translation outperforms the existing models and baselines, especially the recent state-of-the-art algorithm Graph Fusion, which also leveraged the pre-trained BERT based visual grounding model from large scale paired image-text data. However, it only hard-coded the inferred multi-modal graph by visual grounding to construct the mask matrix of cross modality attention in the transformer encoder. One advantage of our work is that we directly build our NMT model on top of the pre-trained vision language BERT, making the most of pre-trained cross modality attention. Another advantage is that our end-to-end trainable model can spontaneously avoid the error accumulation.

Since our multi-modal translation model is implemented based on the text-only transformer, we also report the text-only results with our own implemented transformer for a fair comparison. Our text-only transformer is a surprisingly strong baseline and very competitive with most cited works. For English to German translation task, our text-only baseline almost beats all previous works on the ambiguous MSCOCO test set, and is only inferior to two systems on Multi-30K test sets with less than 2 BLEU score difference. For English to French translation task, only the Graph Fusion algorithm significantly outperforms our text-only transformer. In contrast, on the three test sets of English to German, our final multi-modal translation model can on average achieve approximately +4.6 BLEU and +4.2 METEOR over the text-only baseline. On the two test sets of English to French, the averaged gains of our model are about +5.85 and +4.45 on BLEU and METEOR.

Model	Test2016		Test2017		MSCOCO	
	B	M	B	M	B	M
Text-only En⇒De Model						
Transformer	38.3	56.6	30.3	51.0	28.6	47.7
BERT-NMT	39.4	56.6	29.7	48.6	27.9	46.2
BERT-enc 1st	39.1	57.1	31.8	51.1	29.5	47.9
BERT-enc 2nd	40.0	58.7	34.7	53.8	30.6	51.2
Multi-modal En⇒De Model						
Our Model	**42.7**	**60.7**	**35.5**	**54.9**	**32.8**	52.2
- $\mathcal{L}_{\text{M}}$	41.8	60.0	34.7	54.6	32.3	**52.3**

Table 5: Comparison with variants of text-only models. 1st and 2nd means the 1st and 2nd stage of training.

4.4 Probing Textual Language Model

Our implemented text-only transformer only uses the source-target parallel corpus extracted from Multi-30K, which overlooks the power of the pre-training on the source side. Because our multi-modal encoder has been fully pre-trained, we systematically compare it with another two text-only baselines. The first baseline virtually has the same architecture as multi-modal framework but without vision language pre-training, denoted as BERT-enc. The second one is **BERT-NMT** (Zhu et al., 2019) by incorporating the output of BERT into the attention module of the transformer. We directly run the experiments with their released codebase[3]. Without image data, all text-only models only optimize the cross entropy loss, so we also present the result of our model without the modality loss.

As shown in Table 5, the BERT-NMT is sometimes even worse than the regular transformer. We hypothesize that the existence of too many untrained parameters in the encoder makes the model difficult to optimize on the limited Multi-30K dataset. When we directly use the pre-trained BERT as the encoder and train the model with two-stage schedule, we observe a consistent improvement on the metrics over the regular transformer, i.e., +1.1 BLEU at 1st-stage and +2.7 BLEU at 2nd-stage. Thus, we argue that with the proper 2-stage training strategy, the pre-trained BERT can account for one half of the overall gains in our final model.

4.5 Ablation Study

To validate the contribution of each component in our approach, we conduct a series of incremental experiments to observe the model performances in different scenarios, summarized in Table 6.

[3] https://github.com/bert-nmt/bert-nmt

Multi-modal Model	Test2016		Test2017		MSCOCO		Average	
	B	M	B	M	B	M	ΔB	ΔM
End2End	38.7	58.3	31.6	53.1	29.1	50.0	-	-
1st-Stage	40.0	57.5	32.4	51.5	30.8	49.8	+1.27	-0.87
+ 2nd-Stage	41.8	60.0	34.7	54.6	32.3	**52.3**	+3.13	+1.83
+ Last Layer Modality loss $\widetilde{\mathcal{L}}_{\mathrm{M}}^{(6)}$	**42.7**	**60.7**	**35.5**	**54.9**	**32.8**	52.2	**+3.90**	**+2.13**
or + All Layers Modality loss $\sum_{i=1}^{6} \widetilde{\mathcal{L}}_{\mathrm{M}}^{(6)}$	41.7	59.9	34.8	54.7	32.0	51.7	+3.07	+1.63
or + Last Layer Modality loss $\mathcal{L}_{\mathrm{M}}^{(6)}(\lambda^{(6)} = 0.4)$	42.1	59.9	34.9	54.6	31.8	51.3	+3.17	+1.46

Table 6: Ablation study of MMT training on the En⇒De dataset after VLP. Different modality losses are exclusive.

Two-Stage Training In previous analysis, we've seen how the 2-stage training can benefit the text-only model. In Table 6, we present the metrics of different multi-modal models. The end-to-end training, similar to the traditional fine-tuning strategy in (Devlin et al., 2019), optimizes all model parameters of the downstream task once the VLP is finished. We found it leads even worse result than optimizing the decoder alone (i.e., 1st-stage training) on the metric BLEU. In addition, the result after the 2nd-stage fine-tuning produces significant performance increase. We also plot the learning curve of BLEU on development dataset in Figure 3. The apparent gap between two curves confirms the contribution of 2-stage training.

Modality Loss Note that the results in the first three lines of Table 6 are achieved by optimizing the cross entropy loss alone. In this study, we will verify the effectiveness of the modality loss in 3 different setups. We found only optimizing the modality loss of the last layer can achieve the best performance. As we discussed before, the query vector of the last layer will directly and maximally influence the generation of the target token, while the vectors from remote layers seem not important. We can use the statistics of the learnable λ to avoid the time-consuming cross-validation. For example, we set λ as the approximate mean 0.4 in the original modality loss Eq. (4). Although a slightly performance drop appears, we can get rid of 3 trainable vectors.

4.6 Case Studies

Actually, the translation performance of the MMT with vision language model only exceeds about 2 BLEU scores compared with the NMT with BERT language model. So we cannot guarantee that all sentences in the testsets can be better translated by MMT with VLP. We only exemplify two cases with better translation quality for MMT with VLP, to indicate the potential benefits.

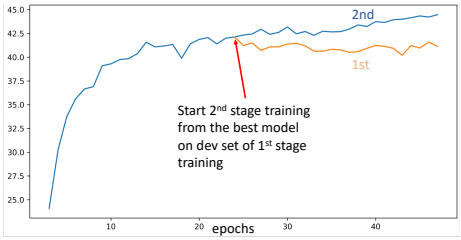

Figure 3: Learning curve of two-stage training w.r.t. BLEU on development set.

In the first case, German words "personen" and "leute" both mean "people", where leute is a general expression and can't be in singular, and "personen" is a formal expression when stating how many people. In object detection model, the tag "person" possibly enhances the NMT model to produce a similar German word "personen". In addition, person is also a German word.

The second case comes from the Ambiguous COCO testset. The NMT with BERT language model cannot miss the translation of the word pizza. The detected object "pizza" may also emphasize the word and help the MMT, though MMT translated the rectangular pizza to stein-pizza (stone-pizza).

4.7 Discussion

The major limitation of our method is that the training pipeline cannot easily generalize to other source languages other than English, because the image-text paired data is unavailable in other languages. Liu et al. (2020) presented a sequence-to-sequence denoising auto-encoder pre-trained on large-scale monolingual corpora in many languages, and successfully applied to multi-lingual translation. Hopefully, we can explore the similar unsupervised cross-lingual or zero-shot transfer learning techniques, which help adapt the multi-lingual BERT

src	four people relaxing on a grassy hill overlooking a rocky valley .
ref	vier personen entspannen auf einem grasbewachsenen hügel mit ausblick auf ein felsiges tal .
brt	vier leute entspannen sich auf einem grasbewachsenen hügel mit blick auf ein steiniges tal .
vlp	vier personen entspannen sich auf einem grasbewachsenen hügel mit blick auf ein steiniges tal .
src	a girl with arms crossed leaning on counter over a rectangular pizza , by a wall calendar and containers .
ref	ein mädchen mit gekreuzten armen stützt sich auf eine theke mit einer rechteckigen pizza , neben einem wandkalender und behältern .
brt	ein mädchen mit verschränkten armen lehnt sich mit überkreuzten armen an einer theke neben einer wand und kartons .
vlp	ein mädchen mit gekreuzten armen lehnt sich über eine theke neben einer wand , auf der sich ein stein-pizza und behälter steht .

Table 7: Case Studies

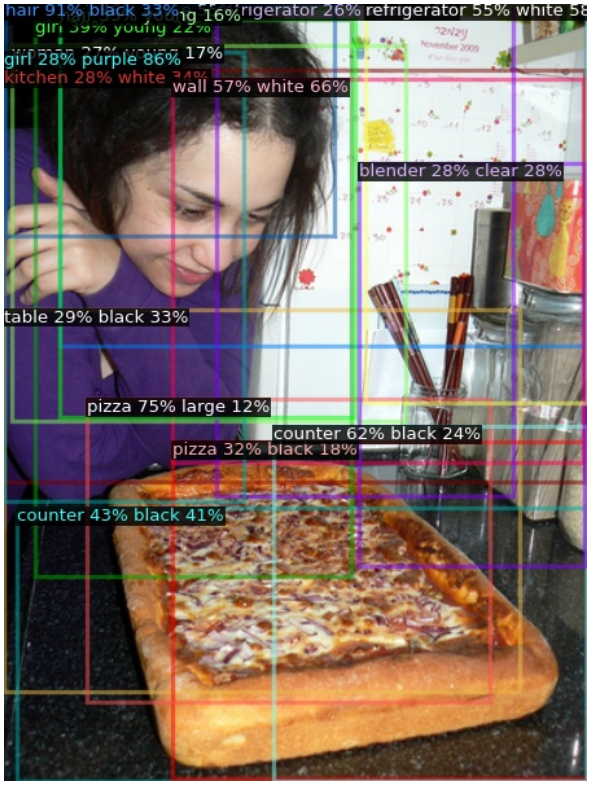

Figure 4: The image for the second case

to a vision multi-lingual model. We will leave this direction as our future work. The main purpose is not to design a better vision language model for other downstream tasks such as VQA. Note that the contemporary works including ViLBERT (Lu et al., 2019) and Oscar (Li et al., 2020) may share the same idea to utilize pre-trained BERT. Our idea is mostly enlighten by (Ive et al., 2019). Another different approach is VL-BERT (Su et al., 2019a), which required to mask sub-regions of the image and introduced masked ROI classification loss, rather than mimicking the NSP loss in traditional BERT.

5 Conclusion

In this paper, we found the vision language pre-training on the source side can significantly improve the multi-modal machine translation, even without additional target corpus. Although the model architecture is as simple as the regular encoder-decoder transformer, our proposed training pipeline can help the MMT system outperform previous works by a large margin on the Multi-30K dataset. The success of the source-image cross-modality representation learning encourages us to design the modality loss that aims at transferring the pre-trained representations to the target-image pair. The quantitative analysis also demonstrates its effectiveness.

Impact Statement

Vision language pre-training has achieved great success in many NLP tasks. We believe it would definitely benefit the multi-modal translation and expect this work can indicate a new unconstrained scenario.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

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.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439, Copenhagen, Denmark. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7059–7069.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 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.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, et al. 2018. The memad submission to the wmt18 multimodal translation task. *arXiv preprint arXiv:1808.10802*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Jindřich Helcl, Jindřich Libovickỳ, and Dušan Variš. 2018. Cuni system for the wmt18 multimodal translation task. *arXiv preprint arXiv:1811.04697*.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6700–6709.

Julia Ive, Pranava Swaroop Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. *arXiv preprint arXiv:2004.06165*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.

Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, pages 1143–1151.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Lucia Specia, Stella Frank, Khalil Sima'an, 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.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019a. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.

Yuanhang Su, Kai Fan, Nguyen Bach, C-C Jay Kuo, and Fei Huang. 2019b. Unsupervised multi-modal neural machine translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10482–10491.

Hao Tan and Mohit Bansal. 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)*, pages 5103–5114.

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 *Advances in neural information processing systems*, pages 5998–6008.

Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Pengcheng Yang, Boxing Chen, Pei Zhang, and Xu Sun. 2020. Visual agreement regularized training for multi-modal machine translation. In *AAAI*, pages 9418–9425.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4683–4693.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 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*, pages 3025–3035.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tieyan Liu. 2019. Incorporating bert into neural machine translation. In *International Conference on Learning Representations*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.