Neural Machine Translation with Phrase-Level Universal Visual Representations

Qingkai Fang[1,2]**, Yang Feng**[1,2*]
[1] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] University of Chinese Academy of Sciences, Beijing, China
{fangqingkai21b,fengyang}@ict.ac.cn

Abstract

Multimodal machine translation (MMT) aims to improve neural machine translation (NMT) with additional visual information, but most existing MMT methods require paired input of source sentence and image, which makes them suffer from shortage of sentence-image pairs. In this paper, we propose a phrase-level retrieval-based method for MMT to get visual information for the source input from existing sentence-image data sets so that MMT can break the limitation of paired sentence-image input. Our method performs retrieval at the phrase level and hence learns visual information from pairs of source phrase and grounded region, which can mitigate data sparsity. Furthermore, our method employs the conditional variational auto-encoder to learn visual representations which can filter redundant visual information and only retain visual information related to the phrase. Experiments show that the proposed method significantly outperforms strong baselines on multiple MMT datasets, especially when the textual context is limited.

1 Introduction

Multimodal machine translation (MMT) introduces visual information into neural machine translation (NMT), which assumes that additional visual modality could improve NMT by grounding the language into a visual space (Lee et al., 2018). However, most existing MMT methods require additional input of images to provide visual representations, which should match with the source sentence. Unfortunately, in practice it is difficult to get this kind of pairwise input of text and images which hinders the applications of MMT. What is worse, to train an MMT model, the training data still involves the target sentence besides the source

sentence and the image, which is costly to collect. As a result, the MMT model is usually trained on a small Multi30K (Elliott et al., 2016) data set, which limits the performance of MMT. Therefore, it is necessary to utilize the separate image data set to obtain visual representations to break the constraints of pairwise input.

Towards this end, some researchers (Zhang et al., 2020; Wu et al., 2021) propose to integrate a retrieval module into NMT, which retrieve images related to the source sentence from existing sentence-image pairs as complementary input, and then use a pre-trained convolutional neural network (CNN) to encode the images. However, such sentence-level retrieval usually suffers from sparsity as it is difficult to get the images that properly match with the source sentence. Besides, visual features outputted by the CNN contain richer information (*e.g.*, color, size, shape, texture, and background) than the source text, thus encoding them in a bundle without any filtering will introduce noise into the model.

To solve these problems, we propose a novel retrieval-based method for MMT to learn phrase-level visual representations for the source sentence, which can mitigate the aforementioned problems of sparse retrieval and redundant visual representations. For the sparsity problem, our method retrieves the image at the phrase level and only refers to the grounded region in the image related with the phrase. For the redundancy problem, our method employs the conditional variational auto-encoder to force the learned representations to properly reconstruct the source phrase so that the learned representations only retain the information related to the source phrase . Experiments on Multi30K (Elliott et al., 2016) show that the proposed method gains significant improvements over strong baselines. When the textual context is limited, it achieves up to 85% gain over the text-only baseline on the BLEU score. Further analysis demonstrates that

---

the proposed method can obtain visual information that is more related to translation quality.

2 Phrase-Guided Visual Representation

We use phrase-level visual representation to improve NMT. In this section, we will introduce our proposed *phrase-guided visual representation*. We first build a phrase-level image set, and then introduce a latent-variable model to learn a phrase-guided visual representation for each image region.

2.1 Phrase-Level Image Set

Our phrase-level image set is built from the training set of Multi30K, which contains about 29K bilingual sentence-image pairs. We only use the images $\mathbf{e}$ and source descriptions $\mathbf{x}$ from them, which is denoted as $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{e}_i)\}_{i=1}^N$. We extract <noun phrase, image region> pairs from <sentence, image> pairs in $\mathcal{D}$ to build our phrase-level image set, which is denoted as $\mathcal{D}_p$.

For each sentence $\mathbf{x}_i$, we use an open-source library *spaCy*[1] to identify the noun phrases, which is denoted as $\mathbf{P}_i = (\mathbf{p}_1^i, \mathbf{p}_2^i, ..., \mathbf{p}_{t_i}^i)$, where t_i is the number of noun phrases in $\mathbf{x}_i$. For each noun phrase $\mathbf{p}_j^i$, we detect the corresponding region $\mathbf{r}_j^i$ from the paired image $\mathbf{e}_i$ using the visual grounding toolkit (Yang et al., 2019). Then $(\mathbf{p}_j^i, \mathbf{r}_j^i)$ is added to our phrase-level image set $\mathcal{D}_p$. Figure 1 illustrates an example.

Finally, we obtain the phrase-level image set $\mathcal{D}_p = \{(\mathbf{p}_i, \mathbf{r}_i)\}_{i=1}^T$, where $T = \sum_{i=1}^N t_i$. It contains about 102K pairs in total.

2.2 Latent-Variable Model

For an image region $\mathbf{r}$, we can obtain the visual features $\mathbf{v}$ with a pre-trained ResNet-101 Faster R-CNN (He et al., 2016; Ren et al., 2015), which contains rich visual information (*e.g.*, color, size, shape, texture, and background). However, we should not pay much attention to the visual information not mentioned in the corresponding phrase, which will introduce too much noise and even be harmful to NMT. Therefore, we further introduce a continuous latent variable to explicitly model the semantic information of image regions under the guidance of phrases. We adopt the framework of conditional variational auto-encoder (CVAE) (Kingma and Welling, 2014; Sohn et al., 2015) to maximize the conditional marginal log-likelihood
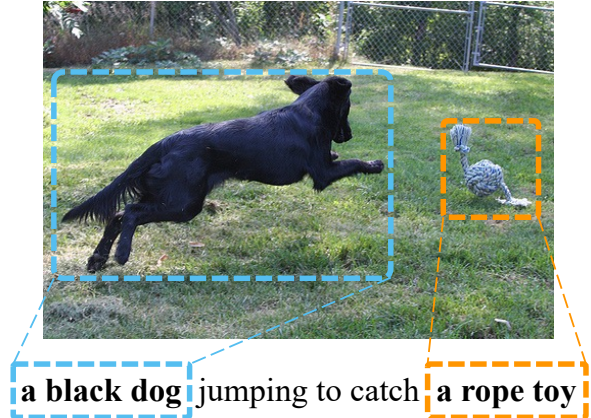
---

[1] https://spacy.io

Figure 1: Example of extracting <noun phrase, image region> pairs from existing <sentence, image> pairs.

$\log p(\mathbf{p}|\mathbf{v}) = \log \int_z p(\mathbf{p}|z, \mathbf{v})p(z|\mathbf{v})dz$ by maximizing the evidence lowerbound (ELBO):

$$\mathcal{L}_{cvae}(\omega, \phi, \theta) = \mathbb{E}_{z \sim q_\phi(\mathbf{z}|\mathbf{p},\mathbf{v})}[\log p_\theta(\mathbf{p}|\mathbf{z}, \mathbf{v})] \\ - \text{KL}[q_\phi(\mathbf{z}|\mathbf{p}, \mathbf{v}) \| p_\omega(\mathbf{z}|\mathbf{v})], \quad (1)$$

where $p_\omega(\mathbf{z}|\mathbf{v})$ is the prior, $q_\phi(\mathbf{z}|\mathbf{p}, \mathbf{v})$ is an approximate posterior and $p_\theta(\mathbf{p}|\mathbf{z}, \mathbf{v})$ is the decoder. The prior p_ω is modeled as a Gaussian distribution:

$$p_\omega(\mathbf{z}|\mathbf{v}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_p(\mathbf{v}), \boldsymbol{\sigma}_p(\mathbf{v})^2\mathbf{I}), \quad (2)$$

$$\boldsymbol{\mu}_p(\mathbf{v}) = \text{Linear}(\mathbf{v}), \quad (3)$$

$$\boldsymbol{\sigma}_p(\mathbf{v}) = \text{Linear}(\mathbf{v}), \quad (4)$$

where $\text{Linear}(\cdot)$ denotes linear transformation. The approximate posterior q_ϕ is also modeled as a Gaussian distribution:

$$q_\phi(\mathbf{z}|\mathbf{p}, \mathbf{v}) = \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_q(\mathbf{p}, \mathbf{v}), \boldsymbol{\sigma}_q(\mathbf{p}, \mathbf{v})^2\mathbf{I}), \quad (5)$$

$$\boldsymbol{\mu}_q(\mathbf{p}, \mathbf{v}) = \text{Linear}([\text{RNN}(\mathbf{p}), \mathbf{v}]), \quad (6)$$

$$\boldsymbol{\sigma}_q(\mathbf{p}, \mathbf{v}) = \text{Linear}([\text{RNN}(\mathbf{p}), \mathbf{v}]), \quad (7)$$

where $\text{RNN}(\cdot)$ denotes a single-layer unidirectional recurrent neural network (RNN). The final hidden state of RNN is used to compute the mean and variance vectors.

To be able to update the parameters using backpropagation, we use the reparameterization trick (Kingma and Welling, 2014) to sample $\mathbf{z}$ from q_ϕ:

$$\mathbf{z} = \boldsymbol{\mu}_q + \boldsymbol{\sigma}_q \odot \boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \quad (8)$$

The decoder $p_\theta(\mathbf{p}|\mathbf{z}, \mathbf{v})$ is also implemented by a single-layer unidirectional RNN. The initial hidden state of decoder RNN is defined as:

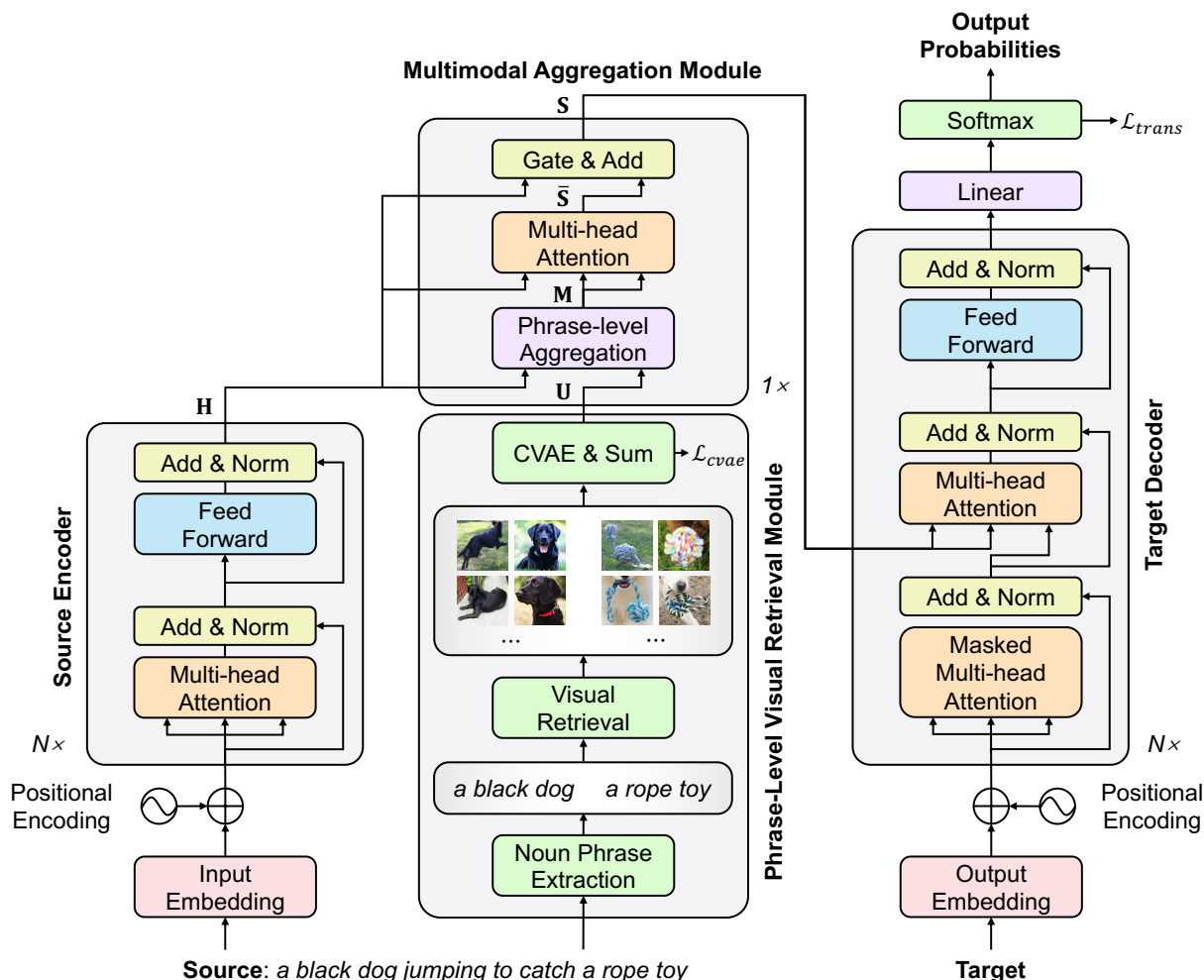$$\mathbf{s} = \text{Linear}([\mathbf{z}, \mathbf{v}]), \quad (9)$$

Figure 2: Overview of our proposed method.

and then the decoder will reconstruct the phrase $\mathbf{p}$ based on $\mathbf{s}$. We refer to $\mathbf{s}$ as *phrase-guided visual representation*, since it pays more attention to the semantic information mentioned in the phrase and filters out irrelevant information. We will describe how to incorporate it into NMT in the next section.

3   NMT with Phrase-Level Universal Visual Representation

In this section, we will introduce our retrieval-based MMT method. Specifically, we obtain visual context through our proposed phrase-level visual retrieval, and then learn a universal visual representation for each phrase in the source sentence, which is used to improve NMT. Figure 2 shows the overview of our proposed method, which is composed of four modules: source encoder, phrase-level visual retrieval module, multimodal aggregation module, and target decoder. The source encoder and target decoder are the same as the encoder and decoder of conventional text-only Trans-

former (Vaswani et al., 2017). Therefore, we will introduce the phrase-level visual retrieval module and multimodal aggregation module in detail in the rest of this section.

We denote the input source sentence as $\mathbf{x} = (x_1, x_2, ..., x_n)$, the ground truth target sentence as $\mathbf{y}^* = (y_1^*, y_2^*, ..., y_m^*)$ and the generated translation as $\mathbf{y} = (y_1, y_2, ..., y_m)$. The input source sentence $\mathbf{x}$ will be encoded with the source encoder to obtain source sentence representation, which is denoted as $\mathbf{H} = (\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_n)$.

3.1   Phrase-Level Visual Retrieval Module

To obtain the visual context of the source sentence without input paired images, we design a phrase-level visual retrieval module. Specifically, for the input sentence $\mathbf{x} = (x_1, x_2, ..., x_n)$, we identify the noun phrases $\bar{\mathbf{P}} = (\bar{\mathbf{p}}_1, \bar{\mathbf{p}}_2, ..., \bar{\mathbf{p}}_t)$ in $\mathbf{x}$. Each phrase $\bar{\mathbf{p}}_i = (x_{l_i}, x_{l_i+1}, ..., x_{l_i+d_i-1})$ is a continuous list of tokens, where l_i is the index of the first token and d_i is the length of $\bar{\mathbf{p}}_i$. For each noun

phrase $\bar{\mathbf{p}}_i$, we will retrieve several relevant <noun phrase, image region> pairs from our phrase-level image set $\mathcal{D}_p$ according to the semantic similarity between phrases, and then use the image regions as visual context. We design a phrase encoder to compute the phrase embedding, which is used to measure the semantic similarity between phrases.

Phrase Encoder   Our phrase encoder $\text{Enc}_{\text{p}}(\cdot)$ is based on a pre-trained BERT (Devlin et al., 2019). For a phrase $\mathbf{p} = (p_1, p_2, ..., p_l)$, we first use BERT to encode it into contextual embeddings:

$$\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_l = \text{BERT}(p_1, p_2, ..., p_l), \qquad (10)$$

then the phrase embedding is the average embedding of all tokens:

$$\text{Enc}_{\text{p}}(\mathbf{p}) = \frac{1}{l} \sum_{i=1}^{l} \mathbf{c}_i. \qquad (11)$$

Visual Retrieval   For a given phrase $\bar{\mathbf{p}}$, we retrieve top-K relevant <noun phrase, image region> pairs from $\mathcal{D}_p$. For $(\mathbf{p}_i, \mathbf{r}_i) \in \mathcal{D}_p$, the relevance score with given phrase $\bar{\mathbf{p}}$ can be defined as the cosine distance between their phrase embeddings:

$$\text{RS}(\bar{\mathbf{p}}, (\mathbf{p}_i, \mathbf{r}_i)) = \frac{\text{Enc}_{\text{p}}(\bar{\mathbf{p}}) \cdot \text{Enc}_{\text{p}}(\mathbf{p}_i)}{\|\text{Enc}_{\text{p}}(\bar{\mathbf{p}})\| \|\text{Enc}_{\text{p}}(\mathbf{p}_i)\|}, \qquad (12)$$

then we retrieve top-K relevant pairs for $\bar{\mathbf{p}}$:

$$\{(\mathbf{p}_{i_k}, \mathbf{r}_{i_k})\}_{k=1}^{K} = \underset{i=1..T}{\text{top-}K}(\text{RS}(\bar{\mathbf{p}}, (\mathbf{p}_i, \mathbf{r}_i))). \qquad (13)$$

Universal Visual Representation   For every pair $(\mathbf{p}_{i_k}, \mathbf{r}_{i_k})$, we can obtain the phrase-guided visual representation $\mathbf{s}_{i_k}$ through our latent-variable model as described in Section 2.2. Finally, the phrase-level *universal visual representation* of $\bar{\mathbf{p}}$ is defined as the weighted sum of all $\{\mathbf{s}_{i_k}\}$:

$$\mathbf{u} = \frac{1}{K} \sum_{k=1}^{K} \text{RS}(\bar{\mathbf{p}}, (\mathbf{p}_{i_k}, \mathbf{r}_{i_k})) \cdot \mathbf{s}_{i_k}. \qquad (14)$$

Our universal visual representation considers multi-view visual information from several image regions, which avoids the bias caused by a single image region. Finally, for all phrases $\bar{\mathbf{P}} = (\bar{\mathbf{p}}_1, \bar{\mathbf{p}}_2, ..., \bar{\mathbf{p}}_t)$ in $\mathbf{x}$, we obtain the corresponding universal visual representation $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_t)$.

3.2 Multimodal Aggregation Module

Inspired by the recent success of modality fusion in multimodal machine translation (Yin et al., 2020; Zhang et al., 2020; Fang et al., 2022), we design a simple *multimodal aggregation module* to fuse the source sentence representation $\mathbf{H}$ and phrase-level universal visual representation $\mathbf{U}$. At first, we perform a phrase-level aggregation. For each phrase $\bar{\mathbf{p}}_i = (x_{l_i}, x_{l_i+1}, ..., x_{l_i+d_i-1})$, we fuse the universal visual representation $\mathbf{u}_i$ and the textual representation of corresponding tokens $(\mathbf{h}_{l_i}, \mathbf{h}_{l_i+1}, ..., \mathbf{h}_{l_i+d_i-1})$:

$$\mathbf{m}_i = \text{LayerNorm}(\mathbf{u}_i + \sum_{j=l_i}^{l_i+d_i-1} \mathbf{o}_{ij} \odot \mathbf{h}_j), \quad (15)$$

$$\mathbf{o}_{ij} = \text{sigmoid}(\mathbf{W}_1 \mathbf{u}_i + \mathbf{W}_2 \mathbf{h}_j), \qquad (16)$$

where $\odot$ denotes element-wise product. Now we obtain the multimodal phrase representation $\mathbf{M} = (\mathbf{m}_1, \mathbf{m}_2, ..., \mathbf{m}_t)$. Afterwards, we apply a multi-head attention mechanism to append $\mathbf{M}$ to the source sentence representation:

$$\bar{\mathbf{S}} = \text{MultiHead}(\mathbf{H}, \mathbf{M}, \mathbf{M}). \qquad (17)$$

We then fuse $\bar{\mathbf{S}}$ and $\mathbf{H}$ with a gate mechanism:

$$\mathbf{S} = \mathbf{H} + \boldsymbol{\lambda} \odot \bar{\mathbf{S}}, \qquad (18)$$

$$\boldsymbol{\lambda} = \text{sigmoid}(\mathbf{W}_3 \mathbf{H} + \mathbf{W}_4 \bar{\mathbf{S}}). \qquad (19)$$

Finally, $\mathbf{S}$ is fed into our target decoder for predicting the translation. The translation model is trained with a cross-entropy loss:

$$\mathcal{L}_{trans} = -\sum_{i=1}^{m} \log p(y_i^* | \mathbf{y}_{<i}, \mathbf{x}). \qquad (20)$$

4   Experiments

We conduct experiments on the following datasets:

Multi30K   Multi30K dataset contains bilingual parallel sentence pairs with image annotations, where each image is paired with one English description and the translations in German and French. Training, validation and test sets contain 29,000, 1,014, and 1,000 instances, respectively. We also report the results on the WMT17 test set and the ambiguous MSCOCO test set, which contain 1,000 and 461 instances respectively.

WMT16 EN-DE   WMT16 EN-DE dataset contains about 4.5M sentence pairs. We choose *newstest2013* for validation and *newstest2014* for test.

Models	EN-DE			EN-FR		
	Test2016	Test2017	MSCOCO	Test2016	Test2017	MSCOCO
Transformer (Vaswani et al., 2017)	39.87	31.78	29.36	60.51	52.44	42.49
Imagination (Elliott and Kádár, 2017)	$39.70^{-0.17}$	$32.15^{+0.37}$	$29.76^{+0.40}$	$60.88^{+0.37}$	$52.89^{+0.45}$	$42.87^{+0.38}$
UVR-NMT (Zhang et al., 2020)	$38.19^{-1.68}$	$31.85^{+0.07}$	$28.55^{-0.81}$	$60.02^{-0.49}$	$51.50^{-0.94}$	$43.22^{+0.73}$
Ours	$\mathbf{40.30^{+0.43}}$	$\mathbf{33.45^{+1.67}}$**	$\mathbf{30.28^{+0.92}}$	$\mathbf{61.31^{+0.80}}$*	$\mathbf{53.15^{+0.71}}$*	$\mathbf{43.65^{+1.16}}$*

Table 1: BLEU scores on Multi30K dataset. * and ** mean the improvements over Transformer (Vaswani et al., 2017) baseline is statistically significant ($p < 0.05$ and $p < 0.01$, respectively).

WMT16 EN-RO WMT16 EN-RO dataset contains about 0.6M sentence pairs. We choose *news-dev2016* for validation and *newstest2016* for test.

For all the above datasets, all sentences are tokenized and segmented into subwords units using byte-pair encoding (BPE) (Sennrich et al., 2016). The vocabulary is shared for source and target languages, with a size of 10K for Multi30K, and 40K for WMT16 EN-DE and WMT16 EN-RO.

4.1 System Settings

Model Implementation For the latent-variable model, the image region is encoded with a pre-trained ResNet101 Faster-RCNN (He et al., 2016; Ren et al., 2015). Both the phrase encoder and decoder are implemented using a single-layer uni-directional RNN with 512 hidden states. The size of the latent variable is set to 64. The batch size is 1024, and the learning rate is 5e-5. We train the model up to 200 epochs with Adam optimizer (Kingma and Ba, 2015). We adopt *KL cost annealing* and *word dropout* tricks to alleviate the posterior collapse problem following Bowman et al. (2016). The annealing step is set to 20000 and the word dropout is set to 0.1. Note that the phrases are segmented using the same BPE vocabulary as that for each source language.

For the translation model, we use Transformer (Vaswani et al., 2017) as our baseline. Both encoder and decoder contain 6 layers. The number of attention heads is set to 4. The dropout is set to 0.3, and the value of label smoothing is set to 0.1. For the visual retrieval module, we retrieve top-5 image regions for each phrase. We use Adam optimizer (Kingma and Ba, 2015) to tune the parameters. The learning rate is varied under a warm-up strategy with 2,000 steps. We train the model up to 8,000, 20,000, and 250,000 steps for Multi30K, WMT16 EN-RO, and WMT16 EN-DE, respectively. We average the checkpoints of last 5 epochs for evaluation. We use beam search with a beam size of 4. Different from previous work,

we use sacreBLEU[2] (Post, 2018) to compute the BLEU (Papineni et al., 2002) scores and the statistical significance of translation results with paired bootstrap resampling (Koehn, 2004) for future standard comparison across papers. Specifically, we measure case-insensitive detokenized BLEU for Multi30K (sacreBLEU signature: nrefs:1 | bs:1000 | seed:12345 | case:lc | eff:no | tok:13a | smooth:exp | version:2.0.0)[3] and case-sensitive detokenized BLEU for WMT datasets (sacreBLEU signature: nrefs:1 | bs:1000 | seed:12345 | case:mixed | eff:no | tok:13a | smooth:exp | version:2.0.0).

All models are trained and evaluated using 2 RTX3090 GPUs. We implement the translation model based on *fairseq*[4] (Ott et al., 2019). We train latent-variable model and translation model individually.

4.2 Baseline Systems

Our baseline is the text-only Transformer (Vaswani et al., 2017). Besides, we implement **Imagination** (Elliott and Kádár, 2017) and **UVR-NMT** (Zhang et al., 2020) based on Transformer, and compare our method with them. The details of these methods can be found in Section 6. We use the same configuration for all baseline systems as our model.

5 Results and Analysis

5.1 Results on Multi30K Dataset

Table 1 shows the results on Multi30K. Our proposed method significantly outperforms the Transformer (Vaswani et al., 2017) baseline, demonstrating that our proposed phrase-level universal visual representation can be helpful to NMT. Our method also surpass **Imagination** (Elliott and Kádár, 2017) and **UVR-NMT** (Zhang et al., 2020). We consider

---

[2]https://github.com/mjpost/sacrebleu
[3]This is because the official pre-processing script of Multi30K dataset lowercases the corpus, see https://github.com/multi30k/dataset/blob/master/scripts/task1-tokenize.sh
[4]https://github.com/pytorch/fairseq
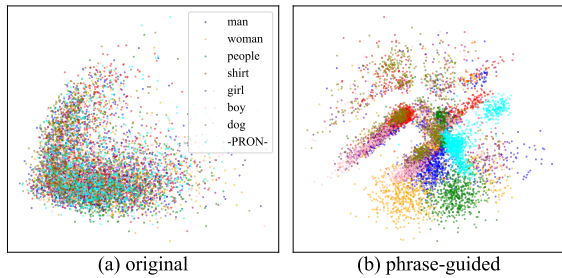
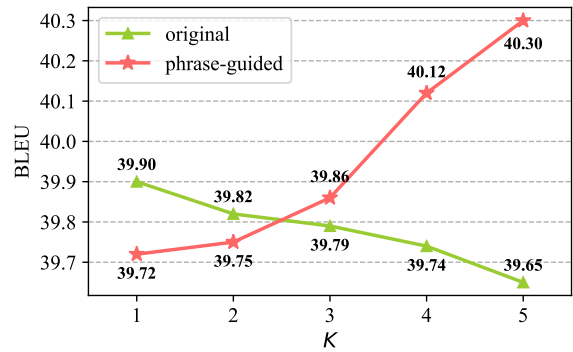Figure 3: Visualization of different visual representations.

Figure 4: BLEU scores with different number of retrieved image regions K. Phrase-guided visual representations achieve better performance as K increases.

Models	Test2016	Test2017	MSCOCO
Transformer	10.42	8.59	7.08
Ours	$19.41^{+8.99}$	$13.67^{+5.08}$	$12.23^{+5.15}$

Table 2: BLEU scores on Multi30K En-De under source-degradation setting.

it is mainly due to the following reasons. First, our phrase-level visual retrieval can obtain strongly correlated image regions instead of weakly correlated whole images. Second, our phrase-level universal visual representation considers visual information from multiple image regions and pays more attention to the semantic information mentioned in the phrases. Last, our phrase-level aggregation module makes it easier for the translation model to exploit the visual information.

5.2 Effects of Latent-Variable Model

In Section 2.2, we introduce a latent-variable model to learn a phrase-guided visual representation for each image region. To understand how it improves the model performance compared with original visual features, we visualize the representations by reducing the dimension with Principal Component Analysis (PCA). Specifically, for all <noun phrase, image region> pairs in $\mathcal{D}_p$, we cluster the image regions by the head[5] of noun phrases. We select top-8 clusters according to their size, and randomly sample 1000 image regions for each cluster. As shown in Figure 3, the original visual features of different clusters are mixed together, indicating that they contains too much irrelevant information. In contrast, our proposed phrase-guided visual representations, which pay more attention to the semantic information, form several clusters according to their heads.

Combined with our visual retrieval module, we found that as the number of retrieved image regions K increases, the BLEU score keeps decreasing when we use original visual features, while increasing when we use our proposed phrase-guided visual representations, which is shown in Figure 4. We believe the decrease of BLEU score is due to the

irrelevant information in original visual features, and thus directly sum them together will introduce too much noise. Our method filters out those irrelevant information, and multiple image regions could avoid the bias caused by a single one, which leads to the increase of BLEU score. However, we don't observe further improvements when using more image regions.

5.3 Source-Degradation Setting

We further conduct experiments under source-degradation setting, to verify the effectiveness of our method when the source textual context is limited. Following Wu et al. (2021), we mask the visually grounded tokens in the source sentence, which affects around 43% of tokens in Multi30K. As shown in Table 2, our method achieves almost 85% improvements over the text-only Transformer baseline. It means our proposed phrase-level universal visual representation can fill in the missing information effectively.

5.4 Phrase-Level vs. Sentence-Level Retrieval

To prove the effectiveness of phrase-level retrieval, we implement a sentence-level variant of our method. In this variant, we switch the latent-variable model, retrieval module and aggregation module from phrase-level to sentence-level. In this way, we retrieve several images as visual con-

---

[5]https://en.wikipedia.org/wiki/Head_(linguistics)

a person is driving a black car (#136 in Test2017)

(a) Phrase-level retrieval

(b) Sentence-level retrieval

Figure 5: Example of different levels of retrieval. We denote the index of retrieved image (regions) in the training set of Multi30K with #id.

Models	Test2016	Test2016 (Mask)
Transformer	39.87	10.42
Ours-sentence	$40.02^{+0.15}$	$11.52^{+1.10}$*
Ours	$\mathbf{40.30^{+0.43}}$	$\mathbf{19.41^{+8.99}}$**

Table 3: BLEU scores on Multi30K En-De Test2016. (Mask) indicates source-degradation setting. * and ** mean the improvements over Transformer (Vaswani et al., 2017) baseline is statistically significant ($p < 0.05$ and $p < 0.01$, respectively).

text to help the translation. As shown in Table 3, the sentence-level variant *Ours-sentence* performs worse than *Ours*, especially in the case of source-degradation setting. We believe it is because phrase-level retrieval can obtain more relevant image regions as visual context, which contain less noise and can be integrated into textual representations more precisely. In contrast, sentence-level retrieval leads to images with much irrelevant information, and makes it difficult for the model to capture the fine-grained semantic correspondences between images and descriptions. To understand this difference more intuitively, we give an example in Figure 5. As we can see, for the input sentence, phrase-level retrieval can obtain closely related image regions for noun phrases *a person* and *a black car*, while the results of sentence-level retrieval are actually weakly related with the input sentence.

5.5 Results on WMT News Datasets

Finally, we conduct experiments on WMT16 EN-DE and WMT16 EN-RO datasets. As shown in Table 4, we observe that both Zhang et al. (2020) and our method only achieve marginal improvements compared with text-only Transformer baseline. We consider that there are two main reasons. On the one hand, most of tokens in such news text are not naturally related to specific visual contents. We found that the percentage of visual grounded tokens in the training set of WMT16 EN-DE is only 7% (vs. 43% in Multi30K), so the contribution of visual information is indeed limited. On the other hand, the news text is far from the descriptive text in Multi30K. In this way, the retrieved image regions are actually weakly correlated with the source phrase. We did some analysis to verify our hypotheses. As described in Section 3.1, we retrieve top-K pairs for each phrase according to the relevance scores. We define the average relevance scores (ARS) as follows:

$$\text{ARS}(k) = \mathbb{E}_{\mathbf{p} \in \mathcal{D}_{\text{val}}} \text{RS}(\mathbf{p}, (\mathbf{p}_{i_k}, \mathbf{r}_{i_k})), \quad (21)$$

which means the average relevance scores for all phrases in the validation set. As shown in Figure 6, ARS on WMT news datasets are much lower than that on Multi30K, which proves that the gap between news text and descriptive text does exists.

Models	EN-DE	EN-RO
Transformer	26.54	32.67
UVR-NMT	26.89[+0.35]	32.93[+0.26]
Ours	**26.97[+0.43]**	**33.18[+0.51]**

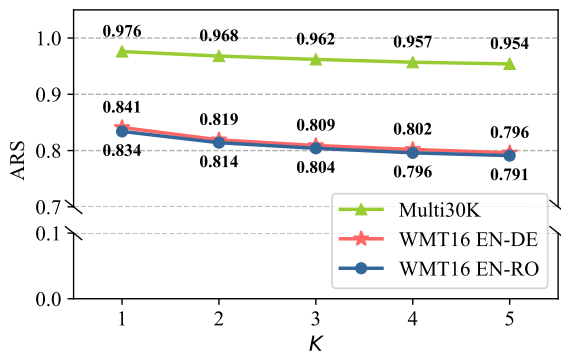Table 4: BLEU scores on WMT16 EN-DE and WMT16 EN-RO dataset.

Figure 6: Average relevance scores (ARS) during visual retrieval for all phrases in the validation set.

6 Related Work

Multimodal machine translation (MMT) aims to enhance NMT (Vaswani et al., 2017; Zhang et al., 2019; Li et al., 2021) with additional visual context. Since the release of Multi30K (Elliott et al., 2016) dataset, researchers have proposed many MMT methods. Early methods (Huang et al., 2016; Calixto and Liu, 2017; Caglayan et al., 2016; Calixto et al., 2016; Caglayan et al., 2017; Libovický and Helcl, 2017; Delbrouck and Dupont, 2017b,a; Zhou et al., 2018; Calixto et al., 2017; Helcl et al., 2018; Caglayan et al., 2018) are mainly based on the RNN-based encoder-decoder architecture with attention (Bahdanau et al., 2015). Recent methods based on Transformer (Vaswani et al., 2017) achieve better performance. Yao and Wan (2020); Yin et al. (2020); Liu et al. (2021) design multimodal encoder to fuse the textual and visual information during encoding. Ive et al. (2019); Lin et al. (2020) enhance the decoder with deliberation networks (Xia et al., 2017) or capsule networks (Sabour et al., 2017) to better utilize visual information during decoding. Caglayan et al. (2021) propose a cross-lingual visual pre-training method and fine-tuned for MMT. It is worth noting that some of previous works (Ive et al., 2019; Lin et al., 2020; Yin et al., 2020; Wang and Xiong, 2021; Nishihara et al., 2020; Zhao et al., 2021) adopt regional

visual information like us, which shows effectiveness compared with global visual features. The major difference between our method and theirs is that our method is a retrieval-based method, which breaks the reliance on bilingual sentence-image pairs, Therefore, our method is still applicable when the input is text only (without paired images), which is unfortunately not available with those previous methods.

In addition to focusing on model design, Yang et al. (2020); Nishihara et al. (2020); Wang and Xiong (2021) propose auxiliary loss to allow the model to make better use of visual information. Caglayan et al. (2019); Wu et al. (2021) conduct systematic analysis to probe the contribution of visual modality. Caglayan et al. (2020); Ive et al. (2021) focus on improving simultaneous machine translation with visual context.

All of the above methods require a specific image as input to provide visual context, which heavily restricts their applicability. To break this bottleneck, Hitschler et al. (2016) propose target-side image retrieval to help the translation. Elliott and Kádár (2017) propose a multitask learning framework **Imagination** to decomposes the multimodal translation into learning translation and learning visually grounded representation. Calixto et al. (2019) introduce a latent variable and estimate a joint distribution over translations and images. Long et al. (2020) predict the translation with visual representation generated by a generative adversarial network (GAN) (Goodfellow et al., 2014). The most closely related work to our method is **UVR-NMT** (Zhang et al., 2020), which breaks the reliance on bilingual sentence-image pairs. Like some retrieval-enhanced MT (Feng et al., 2017; Gu et al., 2017) methods, they build a topic-image lookup table from Multi30K, and then retrieve images related to the source sentence as visual context based on the topic words. The central differences between Zhang et al. (2020) and our method are as follows:

- First, their method depends on the weak correlation between words and images, which leads to much noise in the retrieved images, while our approach relies on the strong correlation between noun phrases and image regions.

- Second, our phrase-level retrieval can obtain more related visual context than their sentence-level retrieval (Section 5.4).

- Last, their method directly uses visual features extracted by ResNet (He et al., 2016), which may introduce too much noise. We adopt a latent-variable model to filter out irrelevant information and obtain a better representation.

7 Conclusion

In this paper, we propose a retrieval-based MMT method, which learns a phrase-level universal visual representation to improve NMT. Our method not only outperforms the baseline systems and most existing MMT systems, but also breaks the restrictions on input that hinder the development of MMT in recent years. Experiments and analysis demonstrate the effectiveness of our proposed method. In the future, we will explore how to apply our method to other tasks.

Acknowledgements

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 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*.

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew Dai, Rafal Jozefowicz, and Samy Bengio. 2016. Generating sentences from a continuous space. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, Berlin, Germany. Association for Computational Linguistics.

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.

Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer. 2018. LIUM-CVC submissions for WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 597–602, Belgium, Brussels. Association for Computational Linguistics.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *CoRR*, abs/1609.03976.

Ozan Caglayan, Julia Ive, Veneta Haralampieva, Pranava Madhyastha, Loïc Barrault, and Lucia Specia. 2020. Simultaneous machine translation with visual context. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2350–2361, Online. Association for Computational Linguistics.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1317–1324, Online. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 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, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Iacer Calixto, Desmond Elliott, and Stella Frank. 2016. DCU-UvA multimodal MT system report. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 634–638, Berlin, Germany. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, 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, Vancouver, Canada. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017a. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–

919, Copenhagen, Denmark. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017b. Multimodal compact bilinear pooling for multimodal neural machine translation. *CoRR*, abs/1703.08084.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

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, Berlin, Germany. Association for Computational Linguistics.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. STEMM: Self-learning with Speech-text Manifold Mixup for Speech Translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Yang Feng, Shiyue Zhang, Andi Zhang, Dong Wang, and Andrew Abel. 2017. Memory-augmented neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1390–1399, Copenhagen, Denmark. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor O. K. Li. 2017. Search engine guided non-parametric neural machine translation. *CoRR*, abs/1705.07267.

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 Libovický, and Dušan Variš. 2018. CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 616–623, Belgium, Brussels. Association for Computational Linguistics.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2399–2409, Berlin, Germany. Association for Computational Linguistics.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Julia Ive, Andy Mingren Li, Yishu Miao, Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2021. Exploiting multimodal reinforcement learning for simultaneous machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3222–3233, Online. Association for Computational Linguistics.

Julia Ive, Pranava 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, Florence, Italy. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 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*.

Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. 2018. Emergent translation in multi-agent communication. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net.

Jicheng Li, Pengzhi Gao, Xuanfu Wu, Yang Feng, Zhongjun He, Hua Wu, and Haifeng Wang. 2021. Mixup decoding for diverse machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 312–320, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1320–1329, New York, NY, USA. Association for Computing Machinery.

Pengbo Liu, Hailong Cao, and Tiejun Zhao. 2021. Gumbel-attention for multi-modal machine translation. *arXiv preprint arXiv:2103.08862*.

Quanyu Long, Mingxuan Wang, and Lei Li. 2020. Generative imagination elevates machine translation. *CoRR*, abs/2009.09654.

Tetsuro Nishihara, Akihiro Tamura, Takashi Ninomiya, Yutaro Omote, and Hideki Nakayama. 2020. Supervised visual attention for multimodal neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4304–4314, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

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*, volume 28. Curran Associates, Inc.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. 2017. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. *CoRR*, abs/2101.05208.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166, Online. Association for Computational Linguistics.

Yingce Xia, Fei Tian, Lijun Wu, Jianxin Lin, Tao Qin, Nenghai Yu, and Tie-Yan Liu. 2017. Deliberation networks: Sequence generation beyond one-pass decoding. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Pengcheng Yang, Boxing Chen, Pei Zhang, and Xu Sun. 2020. Visual agreement regularized training for multi-modal machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):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 *ICCV*.

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, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy. Association for Computational Linguistics.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2021. Neural machine translation with semantically relevant image regions. *red*, 2:c3.

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, Brussels, Belgium. Association for Computational Linguistics.