NEURAL MACHINE TRANSLATION WITH UNIVERSAL VISUAL REPRESENTATION

Zhuosheng Zhang\textsuperscript{1,2,3}, Kehai Chen\textsuperscript{4}, Rui Wang\textsuperscript{4,*}, Masao Utiyama\textsuperscript{4}, Eiichiro Sumita\textsuperscript{4}, Zuchao Li\textsuperscript{1,2,3,*}, Hai Zhao\textsuperscript{1,2,3,*}  
\textsuperscript{1}Department of Computer Science and Engineering, Shanghai Jiao Tong University  
\textsuperscript{2}Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China  
\textsuperscript{3}MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University  
\textsuperscript{4}National Institute of Information and Communications Technology (NICT), Kyoto, Japan  
zhangzs@sjtu.edu.cn, charlee@sjtu.edu.cn, zhachai@cs.sjtu.edu.cn, \{khchen, wangrui, mutiyama, eiichiro.sumita\}@nict.go.jp

ABSTRACT

Though visual information has been introduced for enhancing neural machine translation (NMT), its effectiveness strongly relies on the availability of large amounts of bilingual parallel sentence pairs with manual image annotations. In this paper, we present a universal visual representation learned over the monolingual corpora with image annotations, which overcomes the lack of large-scale bilingual sentence-image pairs, thereby extending image applicability in NMT. In detail, a group of images with similar topics to the source sentence will be retrieved from a light topic-image lookup table learned over the existing sentence-image pairs, and then is encoded as image representations by a pre-trained ResNet. An attention layer with a gated weighting is to fuse the visual information and text information as input to the decoder for predicting target translations. In particular, the proposed method enables the visual information to be integrated into large-scale text-only NMT in addition to the multimodel NMT. Experiments on four widely used translation datasets, including the WMT’16 English-to-Romanian, WMT’14 English-to-German, WMT’14 English-to-French, and Multi30K, show that the proposed approach achieves significant improvements over strong baselines.

1 INTRODUCTION

Visual information has been introduced for neural machine translation in some previous studies (NMT) (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018; Iype et al., 2019) though the contribution of images is still an open question (Elliott, 2018; Caglayan et al., 2019). Typically, each bilingual (or multilingual) parallel sentence pair is annotated manually by one image describing the content of this sentence pair. The bilingual parallel corpora with manual image annotations are used to train a multimodel NMT model by an end-to-end framework, and results are reported on a specific data set, Multi30K (Calixto & Liu, 2017; Calixto et al., 2017).

One strong point of the multimodel NMT model is the ability to use visual information to improve the quality of the target translation. However, the effectiveness heavily relies on the availability of bilingual parallel sentence pairs with manual image annotations, which hinders the image applicability to the NMT. As a result, the visual information is only applied to the translation task over a small and specific multimodel data set Multi30K (Elliott et al., 2016), but not to large-scale text-only NMT (Bahdanau et al., 2014; Gehring et al., 2017; Vaswani et al., 2017) and low-resource

\*Corresponding author. Zhuosheng Zhang and Zuchao Li were internship research fellows at NICT when this work was done. Hai Zhao was partially supported by Key Projects of National Natural Science Foundation of China (No. U1836222 and No. 61733011). Rui Wang was partially supported by JSPS grant-in-aid for early-career scientists (19K20354): “Unsupervised Neural Machine Translation in Universal Scenarios” and NICT tenure-track researcher startup fund “Toward Intelligent Machine Translation”. 
text-only NMT (Fadaee et al., 2017; Lample et al., 2018; Ma et al., 2019; Zhou et al., 2019). In addition, because of the high cost of annotation, the content of one bilingual parallel sentence pair is only represented by a single image, which is weak in capturing the diversity of visual information. The current situation of introducing visual information results in a bottleneck in the multimodal NMT, and is not feasible for text-only NMT and low-resource NMT.

In this paper, we present a universal visual representation (VR) method\(^1\) relying only on image-monolingual annotations instead of the existing approach that depends on image-bilingual annotations, thus breaking the bottleneck of using visual information in NMT. In detail, we transform the existing sentence-image pairs into topic-image lookup table from a small-scale multimodal data set \textit{Multi30K}. During the training and decoding process, a group of images with similar topic to the source sentence will be retrieved from the topic-image lookup table learned by the term frequency-inverse document frequency, and thus is encoded as image representations by a pre-trained ResNet (He et al., 2016). A simple and effective attention layer is then designed to fuse the image representations and the original source sentence representations as input to the decoder for predicting target translations. In particular, the proposed approach can be easily integrated into the text-only NMT model without annotating large-scale bilingual parallel corpora. The proposed method was evaluated on four widely-used translation datasets, including the WMT’16 English-to-Romanian, WMT’14 English-to-German, WMT’14 English-to-French, and \textit{Multi30K} which are standard corpora for NMT and multi-modal machine translation (MMT) evaluation. Experiments and analysis show effectiveness. In summary, our contributions are primarily three-fold:

1. We present a universal visual representation method that overcomes the shortcomings of the bilingual (or multilingual) parallel data with manual image annotations for MMT.
2. The proposed method enables the text-only NMT to use the multimodality of visual information without annotating the existing large scale bilingual parallel data.
3. Experiments on different scales of translation tasks verified the effectiveness and generality of the proposed approach.

2 RELATED WORK

Building fine-grained representation with extra knowledge is an important topic in language modeling (Li et al., 2020a;b; Zhang et al., 2020b;a), among which adopting visual modality could potentially benefit the machine with more comprehensive perception of the real world. Inspired by the studies on the image description generation (IDG) task (Mao et al., 2014; Elliott et al., 2015; Venugopalan et al., 2015; Xu et al., 2015), a new shared translation task for multimodal machine translation was addressed by the machine translation community (Specia et al., 2016). In particular, the released dataset \textit{Multi30K} (Elliott et al., 2016) includes 29,000 multilingual (English, German, and French) parallel sentence pairs with image annotations (Elliott et al., 2017; Barrault et al., 2018). Subsequently, there has been a rise in the number of studies (Caglayan et al., 2016; 2017; Calixto et al., 2016; Huang et al., 2016; Libovicky & Helcl, 2017; Helcl et al., 2018). For example, Calixto et al. (2017) proposed a doubly-attentive multi-modal NMT model to incorporate spatial visual features, improving the translation performance. Compared with spatial-visual features, Calixto & Liu (2017) further incorporated global image features as words in the source sentence and to enhance the encoder or decoder hidden state. In contrast, some recent studies indicated that the visual modality is either unnecessary (Zhang et al., 2017) or only marginally beneficial (Grönroos et al., 2018). More recently, Iye et al. (2019) showed that visual information is only needed in particular cases, such as for ambiguous words where the textual context is not sufficient.

However, these approaches only center around a small and specific \textit{Multi30K} data set to build multimodal NMT model, which hinders image applicability to NMT. The reason would be the high cost of image annotations, resulting potentially in the image information not being adequately discovered. We believe that the capacity of MMT has not yet been excavated sufficiently and there is still a long way to go before the potential of MMT is fully discovered. In this work, we seek to break this constraint and enable visual information to benefit NMT, especially text-only NMT.

\(^1\)The code is publicly available at https://github.com/cooelf/UVR-NMT
3 Universal Visual Retrieval

**Algorithm 1** Topic-image Lookup Table Conversion Algorithm

**Require:** Input sentences, $S = \{X_1, X_2, \ldots, X_i\}$ and paired images $E = \{e_1, e_2, \ldots, e_I\}$

**Ensure:** Topic-image lookup table $Q$ where each word is associated with a group of images

```plaintext
1: Obtain the TF-IDF dictionary $F = \text{TF-IDF}(S)$
2: Transform sentence-image pair to topic-image lookup table $Q = \text{LookUp}(S, E, F)$
3: procedure $\text{TF-IDF}(S)$
4: for each sentence in $S$ do
5: Filter stop-words in the sentence
6: Calculate the TF-IDF weight for each word
7: end for
8: return TF-IDF dictionary $F$
9: end procedure
10: procedure $\text{LookUp}(S, E, F)$
11: for each pair $\{T_i, e_i\} \in \text{zip}(S, E)$ do
12: Rank and pick out the top-$w$ “topic” words in the sentence according to the TF-IDF score in the dictionary $F$, and each sentence is reformed as $T = \{t_1, t_2, \ldots, t_w\}$
13: Pair the $w$ words with the corresponding image $e_i$
14: for each word $t_j$ in $T$ do
15: if $e_i$ not in $Q[t_j]$ then
16: Add $e_j$ to the corresponding image set $Q[t_j]$ for word $t_j$
17: end if
18: end for
19: end for
20: return Topic-image lookup table $Q$
21: end procedure
```

In this section, we will introduce the proposed universal visual representation method. Generally, the default input setting of the MMT is a sentence-image pair. Our basic intuition is to transform the existing sentence-image pairs into topic-image lookup table, which assumes the topic words in a sentence should be relevant to the paired image. Consequently, a sentence can possess a group of images by retrieving the topic-image lookup table.

**Topic-image Lookup Table Conversion** To focus on the major part of the sentence and suppress the noise such as stopwords and low-frequency words, we design a filtering method to extract the “topic” words of the sentence through the term frequency-inverse document frequency (TF-IDF) inspired by [Chen et al., 2019]. Specifically, given an original input sentence $X = \{x_1, x_2, \ldots, x_I\}$ of length $I$ and its paired image $e$, $X$ is first filtered by a stopword list and then the sentence is treated as a document $g$. We then compute TF-IDF $TI_{i,j}$ for each word $x_i$ in $g$.

$$TI_{i,j} = \frac{a_{i,j}}{\sum_k a_{k,j}} \times \log \frac{|G|}{1 + |\{j : x_i \in g\}|},$$

(1)

where $a_{i,j}$ represents the number of occurrences of the word $x_i$ in the input sentence $g$, $|G|$ the total number of source language sentences in the training data, and $|\{j : x_i \in g\}|$ the number of source sentences including word $x_i$ in the training data. We then select the top-$w$ high TF-IDF words as the new image description $T = \{t_1, t_2, \ldots, t_w\}$ for the input sentence $X$. After preprocessing, each filtered sentence $T$ is paired with an image $e$, and each word $t_i \in T$ is regarded as the topic word for image $e$. After processing the whole corpus (i.e., Multi30K), we form a topic-image lookup table $Q$ as described in Algorithm, in which each topic word $t_i$ would be paired with dozens of images.

**Image Retrieval** For input sentence, we first obtain its topic words according to the text preprocessing method described above. Then we retrieve the associated images for each topic word

---

3 We use the training set of the Multi30K dataset to build the topic-image lookup table.

4 We describe our methods by regarding the processing unit as word though this method can also be applied to a subword-based sentence for which the subword is regarded as the processing unit.

https://github.com/stopwords-iso/stopwords-en
from the lookup table $Q$ and group all the retrieved images together to form an image list $G$. We observe that an image might be associated with multiple topic words so that it would occur multiple times in the list $G$. Therefore, we sort the images according to the frequency of occurrences in $G$ to maintain the total number of images for each sentence at $m$.

Figure 1 illustrates the retrieval process. In the left block, we show six examples of sentence-image pairs in which the topic words are in boldface. Then we process the corpus using the topic-image transformation method demonstrated above and obtain the topic-image lookup table. For example, the word *dog* is associated with 1,512 images. For an input source sentence, we obtain the topic words (in boldface) using the same preprocessing. Then we retrieve the corresponding images from the lookup table for each topic word. Now we have a list of images, and some images appear multiple times as they have multiple topics (like the boxed image in Figure 1). So we sort the retrieved image list by the count of occurrence to pick out the top-$m$ images that cover the most topics of the sentence.

At test time, the process of getting images is done using the image lookup table built by the training set, so we do not need to use the images from the dev and test sets in *Multi30K* dataset. Intuitively, we do not strictly require the manual alignment of the word (or concept) and image, but rely on the co-occurrence of topic word and image, which is simpler and more general. In this way, we call our method as universal visual retrieval.

## 4 NMT with Universal Visual Representation

In this section, we introduce the proposed universal visual representation (VR) method for NMT. The overview of the framework of our proposed method is shown in Figure 2.

### 4.1 Source Representation for Neural Machine Translation

In the state-of-the-art Transformer-based NMT [Vaswani et al., 2017], source information is encoded as source representation by an SAN-based encoder with multiple layers. Specifically, the encoder is composed of a stack of $L$ identical layers, each of which includes two sub-layers. The first sub-layer is a self-attention module, whereas the second is a position-wise fully connected feed-forward network. A residual connection [He et al., 2016] is applied between the two sub-layers, and then

---

1. More examples are provided in the Appendix A.1.
2. The lookup table can be easily adapted to a wide range of other NLP tasks even without any paired image, and therefore opens our proposed model to generalization.
Figure 2: Overview of the framework of our proposed method.

a layer normalization \cite{Ba et al., 2016} is performed. Formally, the stack of learning the source representation is organized as follows:

\[
\begin{align*}
\mathbf{H}_l &= \text{LN}(\text{ATT}_l(\mathbf{Q}_{l-1}, \mathbf{K}_{l-1}, \mathbf{V}_{l-1}) + \mathbf{H}_{l-1}), \\
\mathbf{H}_l &= \text{LN}(\text{FFN}_l(\mathbf{H}_l) + \mathbf{H}_l),
\end{align*}
\]

(2)

where \text{ATT}_l(\cdot), \text{LN}(\cdot), \text{and FFN}_l(\cdot) are the attention module, layer normalization, and the feed-forward network for the \(l\)-th identical layer, respectively. \(\{\mathbf{Q}_{l-1}, \mathbf{K}_{l-1}, \mathbf{V}_{l-1}\}\) are query, key, and value vectors that are transformed from the \((l-1)\)-th layer \(\mathbf{H}_{l-1}\). For example, \(\{\mathbf{Q}^0, \mathbf{K}^0, \mathbf{V}^0\}\) are packed from the summation \(\mathbf{H}^0\) of the positional embeddings and word embeddings. Finally, the output of the stack of \(L\) identical layers \(\mathbf{H}^L\) is the final source sentence representation.

4.2 Aggregation for Text and Image Representations

After retrieval as described in Section 3, each original sentence \(X = \{x_1, x_2, \ldots, x_I\}\) is paired with \(m\) images \(E = \{e_1, e_2, \ldots, e_m\}\) retrieved from the topic-image lookup table \(Q\). First, the source sentence \(X = \{x_1, x_2, \ldots, x_I\}\) is fed into the encoder (Eq. 2) to learn the source sentence representation \(\mathbf{H}^L\). Second, the images \(E = \{e_1, e_2, \ldots, e_m\}\) are the inputs to a pre-trained ResNet \cite{He et al., 2016} followed by a feed forward layer to learn the source image representation \(\text{textM} \in \mathbb{R}^{m \times 2048}\). Then, we apply an attention mechanism\footnote{We used single head here for simplicity.} to append the image representation to the text representation:

\[
\overline{\mathbf{H}} = \text{ATT}_M(\mathbf{H}^L, \mathbf{K}_M, \mathbf{V}_M),
\]

(3)

where \(\{\mathbf{K}_M, \mathbf{V}_M\}\) are packed from the learned source image representation \(M\).
Intuitively, NMT aims to produce a target word sequence with the same meaning as the source sentence rather than a group of images. In other words, the image information may play an auxiliary effect during the translation prediction. Therefore, we compute $\lambda \in [0, 1]$ to weight the expected importance of source image representation for each source word:

$$\lambda = \text{sigmoid}(W_\lambda \mathbf{H} + U_\lambda \mathbf{H}^L),$$  \hspace{1cm} (4)

where $W_\lambda$ and $U_\lambda$ are model parameters. We then fuse $\mathbf{H}^L$ and $\mathbf{H}$ to learn an effective source representation:

$$\mathbf{H} = \mathbf{H}^L + \lambda \mathbf{H}.$$

Finally, $\mathbf{H}$ is fed to the decoder to learn a dependent-time context vector for predicting target translation. Note that there is a single aggregation layer to fuse image and text information.

5 Experiments

5.1 Data

The proposed method was evaluated on four widely-used translation datasets, including WMT'16 English-to-Romanian (EN-RO), WMT'14 English-to-German (EN-DE), WMT'14 English-to-French (EN-FR), and Multi30K which are standard corpora for NMT and MMT evaluation.

1) For the EN-RO task, we experimented with the officially provided parallel corpus: Europarl v7 and SETIMES2 from WMT’16 with 0.6M sentence pairs. We used newsdev2016 as the dev set and newstest2016 as the test set.

2) For the EN-DE translation task, 4.43M bilingual sentence pairs of the WMT14 dataset were used as training data, including Common Crawl, News Commentary, and Europarl v7. The newstest2013 and newstest2014 datasets were used as the dev set and test set, respectively.

3) For the EN-FR translation task, 36M bilingual sentence pairs from the WMT14 dataset were used as training data. Newstest12 and newstest13 were combined for validation and newstest14 was used as the test set, following the setting of Gehring et al. (2017).

4) The Multi30K dataset contains 29K English→{German, French} parallel sentence pairs with visual annotations. The 1,014 English→{German, French} sentence pairs visual annotations are as dev set. The test sets are test2016 and test2017 with 1,000 pairs for each.

5.2 System Setting

Image Retrieval Implementation We used 29,000 sentence-image pairs from Multi30K to build the topic-image lookup table. We segmented the sentences using the same BPE vocabulary as that for each source language. We selected top-$8$ ($w = 8$) high TF-IDF words, and the default number of images $m$ was set 5. The detailed case study is shown in Section 6.1. After preprocessing, we had about 3K topic words, associated with a total of 10K images for retrieval. Image features were extracted from the averaged pooled features of a pre-trained ResNet50 CNN (He et al., 2016). This led to feature maps $V \in \mathbb{R}^{2048}$.

Baseline Our baseline was text-only Transformer (Vaswani et al., 2017). We used six layers for the encoder and the decoder. The number of dimensions of all input and output layers was set to 512 and 1024 for base and big models. The inner feed-forward neural network layer was set to 2048. The heads of all multi-head modules were set to eight in both encoder and decoder layers. For Multi30K dataset, we further evaluated a multimodal baseline (denoted as MMT) where each source sentence was paired with an original image. The other settings were the same as our proposed model.

Model Implementation The byte pair encoding algorithm was adopted, and the size of the vocabulary was set to 40,000. In each training batch, a set of sentence pairs contained approximately 4096×4 source tokens and 4096×4 target tokens. During training, the value of label smoothing was set to 0.1, and the attention dropout and residual dropout were $p = 0.1$. We used Adam optimizer (Kingma & Ba, 2014) to tune the parameters of the model. The learning rate was varied under a warm-up strategy with 8,000 steps. For evaluation, we validated the model with an interval
Table 1: Results on EN-RO, EN-DE, and EN-FR for the NMT tasks. Trans. is short for transformer. N/A denotes that those numbers are not reported in the corresponding literature. “++/+” after the BLEU score indicate that the proposed method was significantly better than the corresponding baseline Transformer (base or big) at significance level $p<0.01/0.05$.

<table>
<thead>
<tr>
<th>System</th>
<th>Architecture</th>
<th>EN-RO</th>
<th>EN-DE</th>
<th>EN-FR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>#Param</td>
<td>BLEU</td>
</tr>
<tr>
<td><strong>Existing NMT systems</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vaswani et al. (2017)</td>
<td>Trans. (base)</td>
<td>N/A</td>
<td>27.3</td>
<td>N/A</td>
</tr>
<tr>
<td></td>
<td>Trans. (big)</td>
<td>N/A</td>
<td>28.4</td>
<td>N/A</td>
</tr>
<tr>
<td>Lee et al. (2018)</td>
<td>Trans. (base)</td>
<td>32.40</td>
<td>N/A</td>
<td>24.37</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Our NMT systems</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This work</td>
<td>Trans. (base)</td>
<td>32.66</td>
<td>61.54M</td>
<td>27.31</td>
</tr>
<tr>
<td></td>
<td>+VR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>33.78++</td>
<td>63.04M</td>
<td><strong>28.14++</strong></td>
</tr>
<tr>
<td></td>
<td>Trans. (big)</td>
<td>33.85</td>
<td>207.02M</td>
<td>28.45</td>
</tr>
<tr>
<td></td>
<td>+VR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>34.46+</td>
<td>211.02M</td>
<td><strong>29.14++</strong></td>
</tr>
</tbody>
</table>

Table 1 shows the translation results for the WMT’14 EN-DE, EN-FR, and WMT’16 EN-RO text-only translation task. Our implemented Transformer (base/big) models showed similar BLEU scores with the original Transformer (Vaswani et al., 2017), ensuring that the proposed method can be evaluated over strong baseline NMT systems. As seen, the proposed +VR significantly outperformed the baseline Transformer (base), demonstrating the effectiveness of modeling visual information for text-only NMT. In particular, the effectiveness was adapted to the translation tasks of the three language pairs which have different scales of training data, verifying that the proposed approach is a universal method for improving the translation performance.

Our method introduced only 1.5M and 4.0M parameters for base and big transformers, respectively. The number is less than 3% of the baseline parameters as we used the fixed image embeddings from the pre-trained ResNet feature extractor. Besides, the training time was basically the same as the baseline model (Section 6.3).

In addition, the proposed method was also evaluated for MMT on the multimodal dataset, *Multi30K*. Results in Table 2 show that our model also outperformed the transformer baseline. Compared with the results in text-only NMT, we find that the image presentation gave marginal contribution, which was consistent with the findings in previous work (Zhang et al., 2017; Grönroos et al., 2018; Caglayan et al., 2019). The most plausible reason might be that the sentences in *Multi30K* are so simple, short, and repetitive that the source text is sufficient to perform the translation (Caglayan et al., 2019; Iye et al., 2019). This verifies our assumption of the current bottleneck of MMT due to the limitation of *Multi30K* and shows the necessity of our new setting of transferring multimodality into more standard and mature text-only NMT tasks.

---

[https://github.com/moses-smt/mosesdecoder/tree/RELEASE-4.0/scripts/generic/multi-bleu.perl](https://github.com/moses-smt/mosesdecoder/tree/RELEASE-4.0/scripts/generic/multi-bleu.perl)

[https://github.com/pytorch/fairseq](https://github.com/pytorch/fairseq)
### Table 2: Results from the test2016 and test2017 for the MMT task. Del denotes the deliberation network in [Ive et al., 2019]. Elliott et al. (2017) is the official baseline (text-only NMT) on WMT17-Multi30K 2017 test data. Trans. is short for transformer and MMT is the multimodal baseline described in Section 5.2. Because we used the same model for test2016 and test2017 evaluation, the numbers of parameters are the same.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Existing NMT systems</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calixto et al. (2017)</td>
<td>RNN</td>
<td>33.7</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Elliott et al. (2017)</td>
<td>RNN</td>
<td>N/A</td>
<td>19.3</td>
<td>N/A</td>
<td>N/A</td>
<td>44.3</td>
<td>N/A</td>
</tr>
<tr>
<td>Elliott &amp; Kadar (2017)</td>
<td>Imagination</td>
<td>36.8</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>[Ive et al., 2019]</td>
<td>Trans. (big)</td>
<td>36.4</td>
<td>N/A</td>
<td>N/A</td>
<td>59.0</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td><strong>Our MMT systems</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>This work</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MMT. (base)</td>
<td>35.09</td>
<td>27.10</td>
<td>50.72M</td>
<td>57.40</td>
<td>48.02</td>
<td>50.65M</td>
<td></td>
</tr>
<tr>
<td>MMT. (big)</td>
<td>35.60</td>
<td>28.02</td>
<td>190.58M</td>
<td>57.87</td>
<td>49.63</td>
<td>190.43M</td>
<td></td>
</tr>
<tr>
<td>Trans. (base)</td>
<td>35.39</td>
<td>26.31</td>
<td>49.13M</td>
<td>37.88</td>
<td>48.35</td>
<td>49.07M</td>
<td></td>
</tr>
<tr>
<td>+VR</td>
<td>35.72</td>
<td>26.87</td>
<td>50.72M</td>
<td>58.32</td>
<td>48.69</td>
<td>50.65M</td>
<td></td>
</tr>
<tr>
<td>Trans. (big)</td>
<td>36.86</td>
<td>27.62</td>
<td>186.38M</td>
<td>56.97</td>
<td>48.17</td>
<td>186.23M</td>
<td></td>
</tr>
<tr>
<td>+VR</td>
<td><strong>36.94</strong></td>
<td><strong>28.63</strong></td>
<td><strong>190.58M</strong></td>
<td><strong>57.53</strong></td>
<td><strong>48.46</strong></td>
<td><strong>190.43M</strong></td>
<td></td>
</tr>
</tbody>
</table>

### 6 Analysis

#### 6.1 Influence of the Number of Images

To evaluate the influence of the number of paired images $m$, we constrained $m$ in $\{0, 1, 3, 5, 7, 9, 15, 20, 30\}$ for experiments on the EN-RO test set, as shown in Figure 4. When $m = 0$, the model is the baseline NMT model, whose BLEU score was lower than all the models with images. As the number of images increases, the BLEU score also increased at the beginning (from 32.66 to 33.78) and then slightly decreased when $m$ exceeds 5. The reason might be that too many images for a sentence would have greater chance of noise. Therefore, we set $m = 5$ in our models.

The number of sentence-image pairs to create the lookup table could also make effects. We randomly split the pairs of Multi30K into the proportion in $\{0.1, 0.3, 0.5, 0.7, 0.9\}$, the corresponding BLEU scores for EN-RO are $\{33.07, 33.44, 34.01, 34.06, 33.80\}$. Furthermore, we also evaluate the performance by adding external sentence-pairs from the training set of MS COCO image caption dataset (Lin et al., 2014). The BLEU scores are 33.55 and 33.71 respectively for COCO only and Multi30K+COCO. These results indicate that a modest number of pairs would be beneficial.

#### 6.2 The Influence of Gating Weight $\lambda$

![Figure 4: Influence of the number of images on the BLEU score.](image)

![Figure 5: Quantitative study of the gating weight $\lambda$.](image)

In our model, the weight $\lambda$ of the gated aggregation method was learned automatically to measure the importance of the visual information. We compared by manually setting the weight $\lambda$ into scalar...
values in \{0.1, 0.3, 0.5, 0.7, 0.9\} for experiments on the EN-RO test set. Figure 5 shows that all models with manual \(\lambda\) outperformed the baseline Trans. (base), indicating the effectiveness of image information. In contrast, they were inferior to the performance of our model. This means that the degree of dependency for image information varies for each source sentence, indicating the necessity of automatically learning the gating weights of image representations.

6.3 Extra Computation Time

There are mainly two extra computation cost using our method, including 1) obtaining image data for sentences and 2) learning image representations, which are negligible compared with training a NMT model. The time of obtaining image data for MT sentences for EN-RO dataset is less than 1 minute using GPU. The lookup table is formed as the mapping of token (only topic words) index to image id. Then, the retrieval method is applied as the tensor indexing from the sentence token indices (only topic words) to image ids, which is the same as the procedure of word embedding. The retrieved image ids are then sorted by frequency. Learning image representations takes about 2 minutes for all the 29,000 images in Multi30K using 6G GPU memory for feature extraction and 8 threads of CPU for transforming images. The extracted features are formed as the “image embedding layer” with the size of (29000, 2400) for quick accessing in neural network.

7 Conclusion

This work presents a universal visual representation method for neural machine translation relying on monolingual with image annotations, which breaks the restraint of heavy dependence on bilingual sentence-image pairs in the current multimodel NMT setting. In particular, this method enables visual information to be applied to large-scale text-only NMT through a topic-image lookup. We hope this work sheds some light for future MMT research. In the future, we will try to adopt the proposed method to other tasks.

References


A Appendix

A.1 Examples of retrieved images

Figure 5: Examples of the topic-image lookup table and retrieved images for sentences in Multi30K dataset. We only show six images for each topic or sentence for instance. The topics in each sentence are in boldface.
Published as a conference paper at ICLR 2020

Retrieved Images for Sentences (WMT)

The old system of **private arbitration courts** is off the **table**

This **issue** is **shaping** as a potential early rift with the **business community**

He said he then **heard his friend**, Hamza **calling** to him

The **red flag** has been risen

The **character attempts** to **pass human** smugglers and then **border police** on his way to a refugee **centre** in the European Union.

Figure 6: Examples of the retrieved images for sentences in WMT datasets. We only show six images for each sentence for instance. The topics in each sentence are in boldface.