Unsupervised Multimodal Neural Machine Translation with Pseudo Visual Pivoting

Po-Yao Huang[1], Junjie Hu[1], Xiaojun Chang[2], Alexander Hauptmann[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Faculty of Information Technology, Monash University
`{poyaoh, junjieh, alex}@cs.cmu.edu, cxj273@gmail.com`

Abstract

Unsupervised machine translation (MT) has recently achieved impressive results with monolingual corpora only. However, it is still challenging to associate source-target sentences in the latent space. As people speak different languages biologically share similar visual systems, the potential of achieving better alignment through visual content is promising yet under-explored in unsupervised multimodal MT (MMT). In this paper, we investigate how to utilize visual content for disambiguation and promoting latent space alignment in unsupervised MMT. Our model employs multimodal back-translation and features pseudo visual pivoting in which we learn a shared multilingual visual-semantic embedding space and incorporate visually-pivoted captioning as additional weak supervision. The experimental results on the widely used Multi30K dataset show that the proposed model significantly improves over the state-of-the-art methods and generalizes well when images are not available at the testing time.

1   Introduction

Neural machine translation (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014) has achieved near human-level performance (Wu et al., 2016). However, its effectiveness strongly relies on the availability of large-scale parallel corpora. Unfortunately, preparing the parallel data remains a challenge as there are more than 6,500 languages in the world, and recruiting translators with bilingual or multilingual knowledge to cover all those languages is impractical.

As a result, developing methods alleviating the need of well-annotated large parallel corpora has recently attracted increasing attention in the community. These methods fall into two broad categories. The first type of methods use a third language as the pivot (Firat et al., 2016; Chen et al., 2017; Cheng et al., 2017; Johnson et al., 2017)

to enable zero-resource translation. Although the progress is encouraging, pivoting with a third language still demands bilingual knowledge for collecting large-scale parallel source-pivot and pivot-target corpora. The second type of methods explore unsupervised approaches (Conneau et al., 2018a; Artetxe et al., 2018; Lample et al., 2018a) have recently achieved impressive translation quality. These methods rely only on monolingual data and back-translation (Sennrich et al., 2016a). However, as discussed in (Lample et al., 2018b), the alignment of source-target sentences is uncertain and highly subject to proper initialization.

Using visual content for unsupervised MT (Chen et al., 2018; Su et al., 2019) is a promising solution for pivoting and alignment based on its availability and feasibility. Abundant multimodal content in various languages are available online (*e.g.* Instagram and YouTube). It is also easier to recruit monolingual annotators to describe an image than to find multilingual translators to translate sentences. Importantly, visual content is eligible to improve the alignments in the language latent spaces since the physical visual perception is similar among people speaking different languages (*e.g.* similar "blue car" for a German and a French).

Based on these insights, we propose a novel unsupervised multimodal MT framework incorporating images as pseudo pivots promoting latent space alignment. In addition to use features of visual objects for multimodal back-translation, we align a shared multilingual visual-semantic embedding (VSE) space via leveraging disjoint image-sentence pairs in different languages. As illustrated in Figure 2, for sentences approximately pivoted by similar images (src-*img*-tgt), drawing embeddings of corresponding image-sentence pairs closer results in better alignments of semantically equivalent sentences in the language latent spaces. Inspired by back-translation, we further explore another pseudo pivoting strategy, which approximates multilingual

sentence pairs (*src*-img-*tgt*) conditioned on a real image via captioning. Instead of using annotation of images for pivoting as in (Chen et al., 2018), we generate sentences in two languages pivoted on the real image, and then approximately pairing them as weak supervision for training unsupervised MT system. This approach is analogous to a cross-modal version of back-translation.

We make the following contributions: (1) Building a unified view of employing visual content for pseudo pivoting. (2) We learn and improve the alignments in the shared multilingual multimodal embedding space for unsupervised MMT with disjoint image-text pairs in different languages. (3) Our model achieves state of the art on Multi30K and generalizes well to the text-only scenario.

2 Background

Neural Machine Translation Typical NMT models are based on the encoder-decoder framework with attention (Bahdanau et al., 2015). Let $\mathbf{x} = (x_1, \cdots, x_N)$ denotes a source sentence and $\mathbf{y} = (y_1, \cdots, y_M)$ denotes a target sentence, where $(\mathbf{x}, \mathbf{y}) \in (\mathcal{X}, \mathcal{Y})$. The encoder-decoder model learns to estimate the following likelihood from the source sentence to the target sentence:

$$p_{x \to y}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{M} p(y_i|\mathbf{y}_{<i}, \mathbf{x}) \qquad (1)$$

When a parallel corpus is available, the maximum likelihood estimation (MLE) is usually adopted to optimize the (source to target language) NMT model by minimizing the following loss:

$$\mathcal{L}_{x \to y}^{MT} = \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim (\mathcal{X}, \mathcal{Y})} \left[-\log p_{x \to y}(\mathbf{y}|\mathbf{x}) \right] \quad (2)$$

Among all encoder-decoder models, the Transformer (Vaswani et al., 2017) architecture recently achieves state-of-the-art translation quality. Instead of using recurrent or convolutional operations, it facilitates multi-head self-attention (Lin et al., 2017). In this paper, we choose the Transformer as the underlying architecture for both the translation and the captioning modules.

Unsupervised Machine Translation While conventional MT systems rely on the availability of a large parallel corpus, translation with zero-resource (unsupervised MT) (Lample et al., 2018a; Artetxe et al., 2018; Lample et al., 2018b) has drawn increasing research attention. Only monolingual sentences are presented at the training and validation phase, *i.e.*, only $\mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$ are available.

Successful unsupervised MT systems share several common principles. First, they require the pre-training step to initialize the model and establish strong monolingual language models properly. For example, XLM (Conneau and Lample, 2019) utilizes the masked language model objective in BERT (Devlin et al., 2019). MASS (Song et al., 2019) utilizes a span-based sequence-to-sequence masking objective for language model pre-training.

Second, these systems transform the unsupervised problem into a weakly or self-supervised one by automatically generating pseudo sentence pairs via back-translation (Sennrich et al., 2016a). The idea behind can be analogous to the cycle-consistency objective in CycleGAN (Zhu et al., 2017) for image-image translation with unpaired data. Specifically, let us denote by $h^*(\mathbf{y}) = (\hat{x}_1, \cdots, \hat{x}_N)$ the sentence in the source language inferred from $\mathbf{y} \in \mathcal{Y}$ such that $h^*(\mathbf{y}) = $ argmax $p_{y \to x}(\mathbf{x}|\mathbf{y})$. Similarly, let us denote by $g^*(\mathbf{x}) = (\hat{y}_1, \cdots, \hat{y}_M)$ the sentence in the target language inferred from $\mathbf{x} \in \mathcal{X}$ such that $g^*(\mathbf{x}) = $ argmax $p_{x \to y}(\mathbf{y}|\mathbf{x})$. Then the "pseudo" parallel sentences $(h^*(\mathbf{y}), \mathbf{y})$ and $(\mathbf{x}, g^*(\mathbf{x}))$ can be further used to train two two MT models ($\mathcal{X} \to \mathcal{Y}$ and $\mathcal{Y} \to \mathcal{X}$) by minimizing the following back-translation loss:

$$\mathcal{L}_{x \leftrightarrow y}^{BT} = \mathbb{E}_{\mathbf{x} \sim \mathcal{X}} \left[-\log p_{y \to x}(\mathbf{x}|g^*(\mathbf{x})) \right]$$
$$+ \mathbb{E}_{\mathbf{y} \sim \mathcal{Y}} \left[-\log p_{x \to y}(\mathbf{y}|h^*(\mathbf{y})) \right] \qquad (3)$$

Although reinforcement learning-based approaches (He et al., 2016a) and Gumbel-softmax reparametrization (Maddison et al., 2017) have been used to handle back-propagation thorough non-differentiable "argmax" predictions. in this paper, we do not back-propagate through $h^*(\mathbf{y})$ and $g^*(\mathbf{x})$ to simplify the training process.

3 Unsupervised Multimodal Machine Translation

As illustrated in Figure 1, our model is composed of seven modules: Two encoder-decoder pairs for translation, two decoders for captioning, and one shared visual encoder. In this section, we first detail our basic MMT model architecture and the unsupervised setup. Then we introduce pseudo visual pivoting: learning multilingual VSE and pivoted captioning.

3.1 Multimodal MT

Multimodal machine translation (Specia et al., 2016) (MMT) considers additional images as a
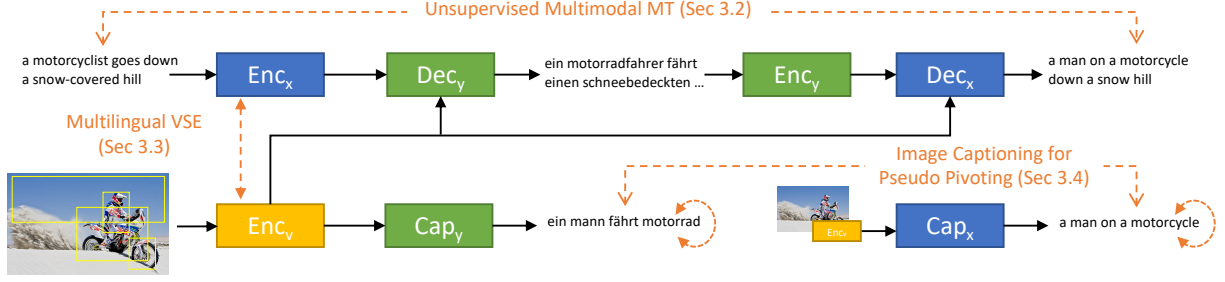
Figure 1: The proposed model structure (English↔German). We incorporate visual objects for unsupervised multimodal MT and improve the language latent space alignment with *pseudo visual pivoting* (§3.3-§3.4).

complementary information source for MT. An image $\mathbf{z}$ and the description in two languages form a triplet $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in (\mathcal{X}, \mathcal{Y}, \mathcal{Z})$. The Transformer encoder reads the source sentence and encodes it with hierarchical self-attention into $\mathbf{h}^x = \{\mathbf{h}_1^x, \cdots, \mathbf{h}_N^x\}, \mathbf{h}_i^x \in \mathbb{R}^d$, where d is the dimension of the embedding space. The visual encoder encodes the image into $\mathbf{h}^z = \{\mathbf{h}_1^z, \cdots, \mathbf{h}_K^z\}, \mathbf{h}_i^z \in \mathbb{R}^d, K_{\max} = 36$. Most previous work (Chen et al., 2018; Su et al., 2019) use 2D ($K = 14 \times 14$) feature maps of ImageNet pre-trained ResNet (He et al., 2016b). In contrast, we utilize the regional features of K salient visual objects in an image extracted by Faster-RCNN (Ren et al., 2015) and a 1-layer MLP as the encoder to encode visual objects.

Various attention strategies for sequence-to-sequence learning have been addressed in (Libovický and Helcl, 2017). Our model employs the hierarchical multi-head multimodal attention for decoding. For decoding at time stamp i, the textual attention $\text{Attn}(\mathbf{h}_i^y, \mathbf{h}^x)$ computes the context vector $\mathbf{c}_i = \sum_j \alpha_j \mathbf{h}_j^x$ via a attention-based alignment $\alpha_j = \text{Align}(\mathbf{h}_i^y, \mathbf{h}_j^x)$, where $\sum_j \alpha_j = 1$ and $\mathbf{h}_i^y$ is the decoder state. Essentially, the one-head attention in Transformer is implemented as $\mathbf{c}_i = \text{softmax}(\mathbf{Q}_i(\mathbf{K}^x)^\top / \sqrt{d})\mathbf{V}^x$ where $\{\mathbf{Q}, \mathbf{K}^x, \mathbf{V}^x\}$ are the packed d-dimensional *Query, Key, Value* vectors, which are the mapped and packed version of $\{\mathbf{h}_i^y, \mathbf{h}^x, \mathbf{h}^x\}$. For decoding with encoded visual and textual inputs, we utilize multimodal attention to compute the context vector $\mathbf{c}_i$:

$$\mathbf{c}_i^x = \text{Attn}(\mathbf{h}_{i-1}^y, \mathbf{h}^x) + \lambda_v \text{Attn}(\mathbf{h}_{i-1}^y, \mathbf{h}^z) \quad (4)$$

In practice we set $\lambda_v = 1.0$. Our multimodal decoder models the likelihood to predict the next token as:

$$p(y_i|\mathbf{y}_{<i}, \mathbf{x}, \mathbf{z}) = \text{softmax}(f(\mathbf{c}_i, y_{i-1}, \mathbf{h}_{i-1}^y)), \quad (5)$$

where $f(.)$ denotes the aggregated non-linear feature mapping in Transformer.

3.2 Unsupervised Learning

Unsupervised multimodal MT (Nakayama and Nishida, 2017; Chen et al., 2018; Su et al., 2019) poses a new yet challenging problem. On both the source and target sides, only non-overlapping monolingual multimodal data are presented for training and validation. Specifically, the data available are: $(\mathbf{x}, \mathbf{z}_x) \in (\mathcal{X}, \mathcal{Z}), (\mathbf{y}, \mathbf{z}_y) \in (\mathcal{Y}, \mathcal{Z})$, such that $\{\mathbf{x}\} \cap \{\mathbf{y}\} = \phi, \{\mathbf{z}_x\} \cap \{\mathbf{z}_y\} = \phi$. Note that there are no parallel translation pairs available (unsupervised), and the images are mutually exclusive for different languages.

For multimodal back-translation, the generated pseudo target sentence conditioned on the source sentence and image can be re-written as $g^*(\mathbf{x}, \mathbf{z}_x) = \text{argmax}\, p_{xz \to y}(\mathbf{y}|\mathbf{x}, \mathbf{z}_x)$, where $p_{xz \to y}(\mathbf{y}|\mathbf{x}, \mathbf{z}) = \prod_{i=1}^M p(y_i|\mathbf{y}_{<i}, \mathbf{x}, \mathbf{z})$. Similar for $p_{yz \to x}(\mathbf{x}|\mathbf{y}, \mathbf{z})$ and $h^*(\mathbf{y}, \mathbf{z}_y)$. For unsupervised multimodal MT, the multimodal back-translation objective can be extended as:

$$\mathcal{L}_{x \leftrightarrow y}^{MBT} = \mathbb{E}_{(\mathbf{x}, \mathbf{z}_x)}\Big[\text{-log}\, p_{yz \to x}\left(\mathbf{x}|g^*(\mathbf{x}, \mathbf{z}_x), \mathbf{z}_x\right) \Big]$$
$$+ \mathbb{E}_{(\mathbf{y}, \mathbf{z}_y)}\Big[\text{-log}\, p_{xz \to y}\big(\mathbf{y}|h^*(\mathbf{y}, \mathbf{z}_y), \mathbf{z}_y)\big) \Big]$$
$$(6)$$

We simplify the notation of expectation for clarity.

Aligning the latent spaces of the source and target languages without supervision is challenging, as discussed in (Lample et al., 2018b). However, as people speak different languages biologically share similar visual systems, we envision that the shared visual space can serve as the pivot for alignment. Unlike most previous work (Chen et al., 2018; Su et al., 2019) treating images merely as a feature, we propose two visual pivoting approaches: (1) Aligning the multilingual VSE space; (2) Image pseudo pivoting via captioning. As illustrated in Figure 2, for (1), we use images as the approximate pivots connecting real non-parallel sentences. (src-*img*-tgt.) In (2), for each pivoting real image, we gener-
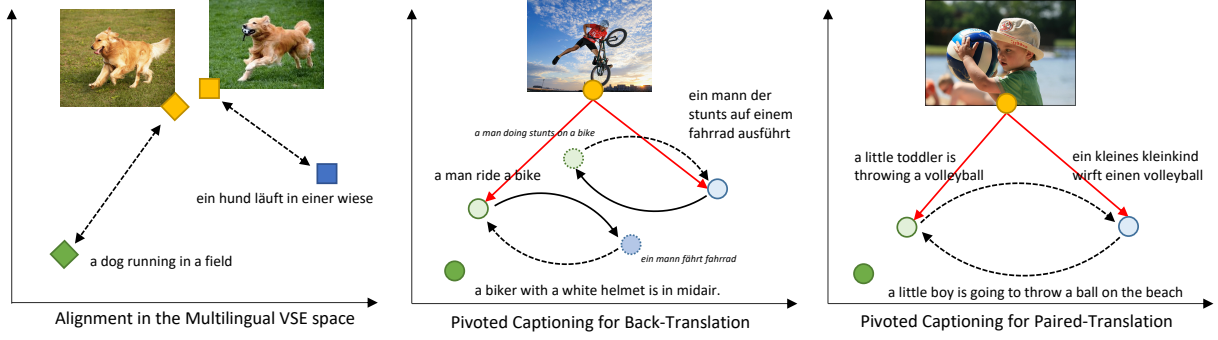
Figure 2: Pseudo visual pivoting: (1) multilingual VSE (src-*img*-tgt, in fact src-img$_1$, tgt-img$_2$), and (2) pivoted captioning (*src*-img-*tgt*). The *italic* items do not exist and are approximated (pseudo). (src, img, tgt) is colored in (green, yellow, blue). Solid red and black lines indicate captioning and translation without updates. Encoder-decoder are updated with dashed lines to improve the alignments in the multilingual multimodal embedding space.

ate captions in both languages to construct "pseudo" source-target sentence pairs. (*src*-img-*tgt*), where the *italic* item is "pseudo". We collectively term the proposed approach *pseudo visual pivoting*.

3.3 Multilingual Visual-Semantic Embedding

We posit that for $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, the two language spaces $\mathcal{X}, \mathcal{Y}$ could be properly associated by respectively aligning two monolingual VSE spaces $\mathcal{X} \leftrightarrow \mathcal{Z}$ and $\mathcal{Y} \leftrightarrow \mathcal{Z}$. We leverage the contrastive objective in cross-modal retrieval (Kiros et al., 2014; Huang et al., 2019b) for aligning multimodal inputs in the shared VSE space where the embeddings are close if they are semantically associated or paired.

Specifically, we generalize the fine-grained (object-level and token-level), monolingual textual-to-visual, and visual-to-textual attention (Lee et al., 2018; Huang et al., 2019c) into the multilingual setup. For fine-grained image-sentence alignment, let $s_{ij} = \cos(\mathbf{h}_i^x, \mathbf{h}_j^z)$ denotes the cosine similarity between the i-th encoded token and the j-th encoded visual object. The image-sentence similarity can be measured by averaging the cosine similarities between the visually-attend sentence embeddings and the visual embeddings of the objects. The visually-attended sentence embeddings $\mathbf{h}^{zx}$ are the weighted combination of the encoded tokens $\mathbf{h}^x$. Precisely, we compute $\mathbf{h}_j^{zx} = \sum_{i=1}^{N} \alpha_{ij} \mathbf{h}_i^x$, where $j = 1 \cdots K$ and $\alpha_{ij} = \text{softmax}_i(s_{ij})$. Let us denote by $S(\mathbf{x}, \mathbf{z}) = \frac{1}{2K} \sum_{j=1}^{K} \cos(\mathbf{h}_j^{zx}, \mathbf{h}_j^z) + \frac{1}{2N} \sum_{i=1}^{N} \cos(\mathbf{h}_i^{xz}, \mathbf{h}_i^x)$ as the image-sentence similarity, the contrastive triplet loss encouraging image-sentence alignment in the VSE space can be written as:

$$
\begin{aligned}
\mathcal{L}_c(\mathbf{x}, \mathbf{z}) = \max_{\tilde{\mathbf{x}}} & \left[\gamma - S(\mathbf{x}, \mathbf{z}) + S(\tilde{\mathbf{x}}, \mathbf{z}) \right]_+ \\
& + \max_{\tilde{\mathbf{z}}} \left[\gamma - S(\mathbf{x}, \mathbf{z}) + S(\mathbf{x}, \tilde{\mathbf{z}}) \right]_+,
\end{aligned} \quad (7)
$$

where $[.]_+$ is the hinge function, and $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{z}}$ are the non-paired (negative) instances for $\mathbf{x}$ and $\mathbf{z}$. Intuitively, when the loss decreases, the matched images and sentences will be drawn closer down to a margin γ than the hardest non-paired ones. Formally, we minimizing the following objective for cross-modal alignments in the two VSE spaces:

$$
\mathcal{L}_{x,y,z}^{VSE} = \mathbb{E}_{(\mathbf{x}, \mathbf{z}_x)} \left[\mathcal{L}_c(\mathbf{x}, \mathbf{z}_x) \right] + \mathbb{E}_{(\mathbf{y}, \mathbf{z}_y)} \left[\mathcal{L}_c(\mathbf{y}, \mathbf{z}_y) \right] \quad (8)
$$

3.4 Image Captioning for Pseudo Pivoting

Inspired by back-translation with monolingual corpora, we propose a novel cross-modal approach to generate weakly-supervised pairs to guide language space alignment for unsupervised MMT. Precisely, we leverage image captioning to synthesize pseudo sentence pairs (pivoted and conditioned on the image) for back-translation and paired-translation.

Image Captioning Image captioning models are akin to MT models besides the non-sequential visual encoder. For example, an image-to-source captioning model estimates the likelihood as $p_{z \to x}(\mathbf{x}|\mathbf{z}) = \prod_{i=1}^{N} p(x_i|\mathbf{x}_{<i}, \mathbf{z})$, where $\mathbf{z}$ is the encoded image. Essentially, the captioning model learns to minimize the following loss:

$$
\mathcal{L}_{z \to x}^{CAP} = \mathbb{E}_{(\mathbf{z}_x, \mathbf{x})} \left[-\log p_{z \to x}(\mathbf{x}|\mathbf{z}_x) \right] \quad (9)
$$

As illustrated in Figure 2, we incorporate two captioning models $\mathcal{Z} \to \mathcal{X}$ and $\mathcal{Z} \to \mathcal{Y}$ to generate additional "pseudo" parallel sentences pivoted on the image as additional weak supervision to better align language latent spaces in unsupervised MMT. For example, with Image $\to$ English and Image $\to$ German, the generated pseudo (English, German)

pair is then pivoted on the Image. Learning captioning models is practical as it is easier to collect large-scale image-text pairs than translation pairs. We pre-train these captioning models and use them to generate sentences in two languages depicting the same image, *i.e.*, $c_x^*(\mathbf{z}_x) = \text{argmax} p_{z \to x}(\mathbf{x}|\mathbf{z}_x)$ and $c_y^*(\mathbf{z}_x) = \text{argmax} p_{z \to y}(\mathbf{y}|\mathbf{z}_x)$. The pivoted captions then enable the following two objectives:

Pivoted Captioning for Back-Translation We utilize the synthetic multilingual captions (*i.e.*, $c_x^*(\mathbf{z}_x)$, $c_y^*(\mathbf{z}_x)$ from the source images and $c_x^*(\mathbf{z}_y)$, $c_y^*(\mathbf{z}_y)$ from the target images) to reversely reconstruct the synthetic captions from their translations in both directions. Formally, we compute the following caption-based back-translation loss:

$$\mathcal{L}_{x \leftrightarrow y}^{CBT} = \mathbb{E}_{\mathbf{z}_x} \Big[\text{-log} \, p_{yz \to x} \big(c_x^*(\mathbf{z}_x) | g^*(c_x^*(\mathbf{z}_x), \mathbf{z}_x), \mathbf{z}_x \big)$$
$$\text{-log} \, p_{xz \to y} \big(c_y^*(\mathbf{z}_x) | g^*(c_y^*(\mathbf{z}_x), \mathbf{z}_x), \mathbf{z}_x \big) \Big]$$
$$+ \mathbb{E}_{\mathbf{z}_y} \Big[\text{-log} \, p_{yz \to x} \big(c_x^*(\mathbf{z}_y) | h^*(c_x^*(\mathbf{z}_y), \mathbf{z}_y), \mathbf{z}_y \big)$$
$$\text{-log} \, p_{xz \to y} \big(c_y^*(\mathbf{z}_y) | h^*(c_y^*(\mathbf{z}_y), \mathbf{z}_y), \mathbf{z}_y \big) \Big]$$
$$(10)$$

Pivoted Captioning for Paired-Translation With the synthetic "pseudo" paired (source, target) captions pivoted on a image (*e.g.* $(c_y^*(\mathbf{z}_x), c_x^*(\mathbf{z}_x))$), the caption-based paired-translation loss is defined as:

$$\mathcal{L}_{x \leftrightarrow y}^{CPT} = \mathbb{E}_{\mathbf{z}_x} \Big[\text{-log} \, p_{xz \to y}(c_y^*(\mathbf{z}_x) | c_x^*(\mathbf{z}_x), \mathbf{z}_x) \Big]$$
$$+ \mathbb{E}_{\mathbf{z}_y} \Big[\text{-log} \, p_{yz \to x}(c_x^*(\mathbf{z}_y) | c_y^*(\mathbf{z}_y), \mathbf{z}_y) \Big]$$
$$(11)$$

Note that similar to the text back-translation, for $\mathcal{L}_{x \leftrightarrow y}^{CPT}$ and $\mathcal{L}_{x \leftrightarrow y}^{CBT}$, we do not back-prop through the captioning step. For optimization, we sample mini-batches and minimizing the following loss:

$$\mathcal{L} = \mathcal{L}_{x \leftrightarrow y}^{MBT} + \mathcal{L}_{x,y,z}^{VSE} + \mathcal{L}_{x \leftrightarrow y}^{CBT} + \mathcal{L}_{x \leftrightarrow y}^{CPT} \quad (12)$$

Here we drop the weights w of each loss for clarity. In practice, all the weights are set to 1.0 except for w_{CPT} where we employ a decreasing learning scheduler specified in the next section.

4 Experiments and Results

We first describe the implementation details and the experimental setup. Then we compare our approach with baselines with detailed analysis.

4.1 Dataset and Preprocessing

We conduct experiments on the Multi30K (Elliott et al., 2016) dataset, the benchmark dataset for mul-

timodal MT. It contains 29K training, 1K validation, and 1K testing images. Each image has three descriptions in English/German/French, which are translations of each other.

To ensure the model never learn from parallel sentences, we randomly split Multi30K training and validation sets in half for one language and use the complementary half for the other. The resulting M30k-half are two corpora with non-overlapping 14,500 training and 507 validation image-sentence pairs, respectively.

For text pre-processing, we use Moses (Koehn et al., 2007) scripts for tokenization and apply the Byte Pair Encoding (BPE) (Sennrich et al., 2016b) from XLM. To identify and extract features of visual objects in images, we use the Faster-RCNN (Ren et al., 2015) model in (Anderson et al., 2018) to detect up to 36 salient visual objects per image and extract their corresponding 2048-dim regional features.

4.2 Implementation

We use Transformer as the underlying architecture for the translation and captioning modules. Each encoder/decoder of the translator is with 6-layer stacked Transformer network, 8 heads, 1024 hidden units, and 4096 feed-forward filter size. The captioner is a 6-layer Transformer decoder with the same configuration. The visual encoder is a 1-layer MLP which maps visual feature to the shared 1,024-dim embedding space then adds the positional encoding to encode spatial locations (normalized top-left and bottom-right coordinates) of visual objects. Our implementation is based on the codebase of XLM and MASS.

4.3 Experimental Details

We respectively conduct unsupervised MMT experiments on Multi30K-half for two language pairs: English-French and English-German.

Pre-Training Pre-training is a critical step for unsupervised MT. We follow the setup in UMMT (Su et al., 2019) for a fair comparison. For each language, we create a text-only pre-training set by combining the shuffled first 10 million sentences of the WMT News Crawl datasets from 2007 to 2017 with 10 times of M30k-half, resulting in a text-only dataset with 10.145 million unparalleled sentences in English, French, German respectively.

For text pre-training, we leverage the script and the masked seq-to-seq objective proposed in

MASS, which randomly masks a span in a sentence then encourages the model to decode and reconstruct the masked sequence as the monolingual language model pre-training. More details can be found in the original paper. Note that there is no fine-tuning (back-translation) on WMT for a fair comparison with other baselines.

For multimodal pre-training of the captioning modules, we use the out-of-domain MS-COCO (Lin et al., 2014) dataset. We randomly split the training set into two disjoint subsets. Each set contains 56,643 images and 283,215 sentences. We use the translate-train strategy as in XNLI (Conneau et al., 2018b). We leverage Google Translate to translate one set of English sentences into French and German. We pre-train the captioning modules with Eq. 9 and fix them during fine-tuning to avoid overfitting. Note that the captioning modules are trained on non-parallel sentences with disjoint image subsets, which implies no overlap between English-German or English-French sentences.

Fine-tuning on Multi30K-half We fine-tune on the training set of Multi30K-half for 18 epochs. We train our model with the Adam optimizer (Kingma and Ba, 2014) with a linear warm-up and a learning rate varying from 10^{-7} to 10^{-5}. We apply a linearly decreasing weight from 1.0 to 0.1 at 10-th epoch for w^{CPT} as we empirically observe that the generated captions are relatively too noisy to serve as good pseudo pairs in the later stage of training. The margin γ in VSE is set to 0.1. Other hyper-parameters in Transformer follow the default setting in MASS. We use 4 Titan Xp GPUs with 1,000 tokens in each mini-batch for training.

Evaluation and Model selection For evaluation, we report BLEU scores by multi-bleu.pl[1] in Moses and METEOR[2] scorea on the Multi30K testing set.

For model selection without a parallel validation corpus, we consider the unsupervised criterion proposed in (Lample et al., 2018a) based on the BLEU scores of "round-trip" translations (source → target → source and target → source → target) which have been empirically shown to correlate well with the testing metrics.

4.4 Baseline Models

We compare recent unsupervised text-only and multimodal MT baselines listed in the following: (1) MUSE (Conneau et al., 2018a) is a word-to-word

MT model with pre-trained Wikipedia embeddings. (2) UNMT (Lample et al., 2018a) sets the tone of using denoising autoencoder and back-translation for unsupervised MT. (3) XLM (Conneau and Lample, 2019) deploys masked language model from BERT. (4) MASS (Song et al., 2019) uses a masked seq-to-seq pre-training objective, achieves the current state-of-the-art performance in text-only unsupervised MT. (5) Game-MMT (Chen et al., 2018) is a reinforcement learning-based unsupervised MMT. (6) UMMT (Su et al., 2019) use visual feature for denoising autoencoder and back-translation. UMMT is the current state of the art in unsupervised MMT. We either use the reported scores in the original papers or use their best scripts with their pre-trained language models publicly available for fine-tuning on Multi30K-half.

4.5 Main Results: Unsupervised MMT

4.5.1 Comparison with the Baseline Models

Table 1 presents the benchmark results with other state-of-the-art unsupervised MT and MMT models on the Multi30K testing set. The first four rows show the results of the recent text-only MT models. Game-MMT and UMMT are MMT models using both image and text inputs. Our full model (T+V+VSE+CBT+CPT) yields new state-of-the-art performance in BLEU and METEOR, outperforming the text-only and multimodal baseline model by a large margin. Notably, our full model outperforms UMMT by +5.5~12.5 BLEU scores, sets a new state of the art in unsupervised MMT.

Although pre-training plays a vital role in unsupervised MT, comparing Ours-Text only and Ours-Full, the results suggest that multimodal content can further boost the performance for unsupervised MT. Images provide +2.7~3.7 BLEU score improvement across four tasks. Note that our model uses different monolingual pre-training corpora to MASS and XLM for the fair comparison with UMMT. With a similar pre-training objective, our text-only model is worse than MASS, while Ours-Full outperforms MASS by +2.3~3.7 in BLEU.

Comparing the multimodal models trained with and without visual content (UMMT-T vs. UMMT-Full and Ours-T vs. Ours-Full), our model achieves +2.5~3.7 improvements in BLEU while +1.4~2.5 for UMMT. The results imply that, even with a higher text-only baseline (*e.g.* 49.5 vs. 37.2 in en →fr), the proposed model incorporates visual

---

[1]https://github.com/moses-smt/mosesdecoder/blob/master-/scripts /generic/multi-bleu.perl

[2]https://github.com/cmu-mtlab/meteor

Model	en→fr BLEU	en→fr METEOR	fr→en BLEU	fr→en METEOR	en→de BLEU	en→de METEOR	de→en BLEU	de→en METEOR
MUSE[†] (Conneau et al., 2018a)	8.5	-	16.8	-	15.7	-	5.4	-
UNMT[†] (Lample et al., 2018a)	32.8	-	32.1	-	22.7	-	26.3	-
XLM[†] (Conneau and Lample, 2019)	46.3	64.3	42.0	38.1	27.4	48.7	30.7	31.0
MASS[†] (Song et al., 2019)	49.8	65.8	43.7	38.7	30.2	51.3	32.5	33.4
Game-MMT (Chen et al., 2018)	-	-	-	-	16.6	-	19.6	-
UMMT-T[†] (Su et al., 2019)	37.2	33.7[*]	38.5	36.4	21.0	25.4[*]	25.0	28.4
UMMT-Full (Su et al., 2019)	39.8	35.5[*]	40.5	37.2	23.5	26.1[*]	26.4	29.7
Ours-Text only[†]	49.5	65.7	43.5	38.5	30.1	51.5	32.4	33.0
Ours-Full	**52.3**	**67.6**	**46.0**	**39.8**	**33.9**	**54.1**	**36.1**	**34.7**

Table 1: **Results on unsupervised MT**. Comparison with benchmarks on the Multi30K testing set. Our full model is with T+V+VSE+CBT+CPT. The best score is marked bold. [†] means text-only. [*] is the METEOR score shown in the UMMT paper.

content more effectively.

In Figure 3, we provide some qualitative results on the Multi30K testing set. We observe a consistent improvement of unsupervised translation quality with our full model to the text-only one. Without parallel translation pairs as the vital supervision, the proposed pseudo visual pivoting successfully disambiguates the word semantics in the similar syntactic category and results in improved cross-lingual word alignment; for instance, "cafe" vs. "soda" machine in the third French example, and "felsigen" (rocky) vs. "verschneiten" (snowy) in the first German example.

4.5.2 Ablation Studies

To quantify module-wise contribution in pseudo visual pivoting, we summarize our ablation studies in Table 2. Comparing the performance improvement from text-only to the model with regional visual features (T+V), the features of salient visual objects contribute +0.6~0.9 BLEU score over a much higher text-only baseline compared to UMMT.

In pseudo visual pivoting, +VSE promotes the alignments in the monolingual VSE spaces and results in an additional +1.3~2.0 gain in BLEU. This improvement validates our hypothesis that the visual space can effectively serve as the bridge connecting the source and target language latent spaces. Also, synthesizing image-pivoted pseudo caption pairs effectively provides weak supervision for aligning the cross-lingual latent space in unsupervised MMT. We observe that the pivoted captions for paired translation (CPT) is more effective than treating them as back-translation pairs (CBT). Utilizing generated image-pivoted captions is shown to be a promising approach for weakly supervised

Model (Ours)	en→fr	fr→en	en→de	de→en
Text only	49.52	43.48	30.10	32.35
T+V	50.43	44.10	31.01	32.95
T+V+VSE	51.72	45.73	32.67	34.94
T+V+CPT	51.64	45.55	33.04	35.02
T+V+CBT	51.23	45.21	32.51	33.87
T+V+VSE+CBT	51.81	45.83	33.01	34.38
T+V+CPT+CBT	51.85	45.65	33.61	35.85
T+V+VSE+CPT	52.19	**46.10**	33.73	35.60
Full Model	**52.29**	45.98	**33.85**	**36.07**

Table 2: Ablation studies. BLEU comparison of different training objectives.

or unsupervised MMT. The full model which employs VSE, CBT, and CPT achieves +1.9~3.1 improvements compared to our multimodal baseline (row two, visual feature only).

4.5.3 Generalizability

How does our unsupervised MMT model generalize when images are not available at the testing time? Table 3 shows the testing results *without* images. As can be observed, our model generalizes well. The differences are mostly less than 1.0 in BLEU. As our model, when being tested without visual content, still outperforms other unsupervised text-only or multimodal MT models listed in Table 1, the minor drop in BLEU implies that the improved cross-lingual latent space alignment via pseudo visual pivoting is likely to be more critical than using images as an input feature for decoding. Luckily, such alignment is already preserved in the training phase with the proposed approach.

An interesting question is: How much does the visual content (as a feature) contribute? As in leave-one-feature-out cross-validation, we compare
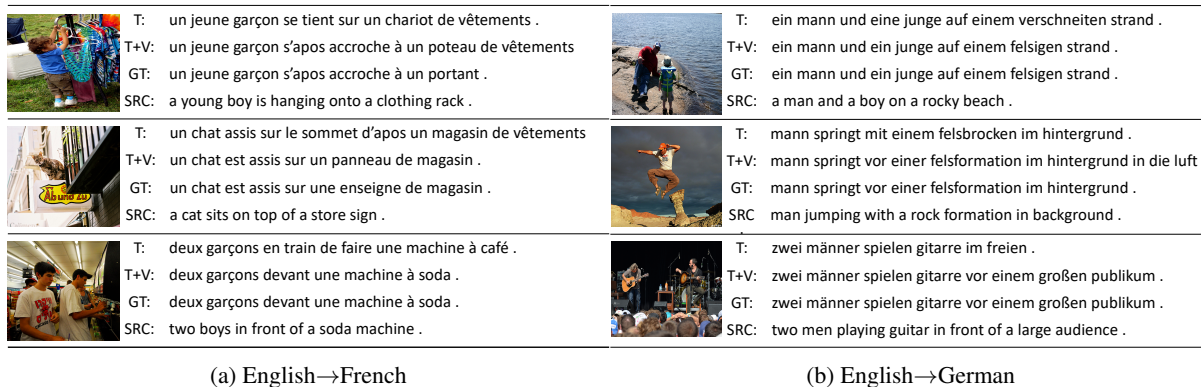
	(a) English→French		(b) English→German

English→French panel:

T:	un jeune garçon se tient sur un chariot de vêtements .
T+V:	un jeune garçon s'apos accroche à un poteau de vêtements
GT:	un jeune garçon s'apos accroche à un portant .
SRC:	a young boy is hanging onto a clothing rack .

T:	un chat assis sur le sommet d'apos un magasin de vêtements
T+V:	un chat est assis sur un panneau de magasin .
GT:	un chat est assis sur une enseigne de magasin .
SRC:	a cat sits on top of a store sign .

T:	deux garçons en train de faire une machine à café .
T+V:	deux garçons devant une machine à soda .
GT:	deux garçons devant une machine à soda .
SRC:	two boys in front of a soda machine .

English→German panel:

T:	ein mann und eine junge auf einem verschneiten strand .
T+V:	ein mann und ein junge auf einem felsigen strand .
GT:	ein mann und ein junge auf einem felsigen strand .
SRC:	a man and a boy on a rocky beach .

T:	mann springt mit einem felsbrocken im hintergrund .
T+V:	mann springt vor einer felsformation im hintergrund in die luft
GT:	mann springt vor einer felsformation im hintergrund .
SRC:	man jumping with a rock formation in background .

T:	zwei männer spielen gitarre im freien .
T+V:	zwei männer spielen gitarre vor einem großen publikum .
GT:	zwei männer spielen gitarre vor einem großen publikum .
SRC:	two men playing guitar in front of a large audience .

(a) English→French   (b) English→German

Figure 3: Qualitative results of the proposed model. GT: ground truth. T+V: Our full model.

Model	en→fr	fr→en	en→de	de→en
UMMT	$39.44_{-0.35}$	$40.30_{-0.23}$	$23.18_{-0.34}$	$25.47_{-0.92}$
Ours-no VSE	$51.60_{-0.25}$	$45.39_{-0.26}$	$33.25_{-0.36}$	$\mathbf{35.15}_{-0.70}$
Ours-Full	$\mathbf{51.64}_{-0.65}$	$\mathbf{45.48}_{-0.50}$	$\mathbf{33.32}_{-0.53}$	$35.04_{-1.03}$

Table 3: BLEU with text-only inputs at the testing time. Subscripts are the differences to testing with T+V.

Img overlap % (# imgs/sents)	en→fr	fr→en	en→de	de→en
0% (14.5K/14.5K)	52.29	45.98	33.85	36.07
50% (22K/22K)	55.13	47.54	34.61	37.01
100% (29K/29K)	**58.34**	**50.57**	**35.45**	**38.55**
0% (T only/14.5K)	49.52	43.48	30.10	32.35
100% (T only/29K)	53.35	46.27	31.35	34.06
0% (3.0K/3.0K)	31.48	27.91	23.94	26.60
0% (T only/3.0K)	30.33	26.95	21.65	23.47

Table 4: Testing BLEU of the full T+V model and the text-only model trained with overlapped images or low-resource unpaired corpora.

the difference of performance between inferencing with and without images. The larger the difference (the subscripts in Table 3) implies a model better utilizes visual content. Compared with UMMT, our model has better utilization. We observe that the key to such difference is the VSE objective. Our model trained without the VSE objective results in worse utilization (smaller difference at the testing time), possibly because the source text-image pairs are distant in the multilingual VSE space.

4.5.4 Real-pivoting & Low-resource Corpora

Will our model benefit from "real" pivoting (src-img_1, img_1-tgt, overall src-img_1-tgt)? We train our models with overlapped images while leaving sentences in the source and target languages un-paralleled (use *no* translation pairs). From the first three rows in Table 4, the performance is improved when training with the overlapped images and their corresponding sentences. Comparing the improvement from 0% to 100% of the text-only model and the full model, a larger gain is observed with the proposed pseudo visual pivoting which aligns and reduces uncertainty in the language latent spaces.

Furthermore, under the low-resource setting (3.0K non-parallel data, row six and seven), a substantial improvement over the text-only model is still observed. These results suggest that the proposed pseudo visual pivoting is likely to generalize to the semi-supervised and the low-resource setting, which we consider as our future work.

4.5.5 Supervised Case

Although the proposed pseudo visual pivoting targets unsupervised MMT, we are also interested in its performance under the fully supervised setup. To gain insights, we conduct supervised MMT experiments by changing the back-translation objective for unsupervised MT (Eq. 6) to the supervised MT objective (Eq. 2) with additional visual inputs. We benchmark with recent supervised MMT models, including Imagination (Elliott and Kádár, 2017), LIUM-CVC (Caglayan et al., 2017), and VAG (Zhou et al., 2018) on Multi30K.

Table 5 shows the testing results. Our model significantly outperforms other baselines and achieves state-of-the-art performance. Comparing to the unsupervised model trained with full Multi30K (Table 4,100% (29K/29K)), the direct supervision from parallel translation pairs results in a +6.5∼7.1 gain in BLEU. Notably, images provide a minor improvement with full supervision from translation pairs. This result implies that, compared to serving as a complementary feature, visual information likely contributes more to improving cross-lingual alignment via pseudo visual pivoting for MMT with limited supervision.

	en→fr		en→de	
Model	BLEU	METEOR	BLEU	METEOR
Imagination	-	-	30.2	51.2
LIUM-CVC	52.7	69.5	30.7	52.2
VAG	53.8	70.3	31.6	52.2
Ours (T)	65.2	**79.3**	42.0	60.5
Ours (T+V)	**65.5**	79.1	**42.3**	**60.6**

Table 5: Supervised MMT results on Multi30K

5 Related Work

Unsupervised MT For pivoting with a third language, Firat et al. (2016) pre-train a multi-way multilingual model to generate pseudo pairs to improve zero-shot translation. Chen et al. (2017) use a teacher-student framework and assume parallel sentences share a similar likelihood for generating sentences in the third language while Cheng et al. (2017) maximize the expected likelihood. Our model does not rely on a third language. Our framework is along the line of research in (Lample et al., 2018a,b; Conneau and Lample, 2019), which aims at learning an aligned latent space between the two languages to translate by reconstruction. Nevertheless, we focus on the multimodal setup where the visual space is dissimilar to the language spaces with challenging asymmetric interactions between modalities.

Supervised MMT Supervised MMT is introduced in (Specia et al., 2016) as a multi-encoder single-decoder framework with additional image inputs. Huang et al. (2016) encode word sequences with regional visual objects while Calixto and Liu (2017) leverage global visual feature. LIUM-CVC (Caglayan et al., 2017) uses element-wise multiplication to model the image-text interaction. Imagination (Elliott and Kádár, 2017) and VAG (Zhou et al., 2018) learns with the auxiliary image reconstruction and source-image-target triplet alignment tasks, respectively. While these methods achieve improvements, their advantage over the text-only models is still minor under the supervised scenario. As analyzed in (Caglayan et al., 2019), visual content is more critical when the textual content is limited or uncertain in MMT. We study the more challenging unsupervised MMT.

Unsupervised MMT To our best knowledge, three recent works have generalized MMT to the unsupervised setting. Nakayama and Nishida (2017) learn modal-agnostic fixed length image/sentence embeddings. In contrast, our model promotes fine-grained (object-token) varying-length embedding,

which better aligns VSE space. Game-MMT (Chen et al., 2018) use a captioning and a translation model maximizing the likelihood of translated captions to original sentences. We synthesize captions for symmetric back-translation and considers no ground truth image annotation in the loop. Empirically, it is preferred to separate real and generated captions. UMMT (Su et al., 2019) uses Transformers, autoencoder loss, and multimodal back-translation. We do not use autoencoder. Our model leverages object detection for multimodal back-translation and equips pseudo visual pivoting.

Image Captioning and VSE Our method draws inspiration from captioning and cross-modal retrieval. Recent progress in captioning aims at using reinforcement learning to improve diversity (Dai et al., 2017) or maximize metric (Rennie et al., 2017). We use a vanilla MLE objective. For learning VSE, we leverage the contrastive loss (Kiros et al., 2014) from cross-modal retrieval, which is shown more robust than maximizing canonical correlation among modalities as in (Andrew et al., 2013; Huang et al., 2018). For encoding image and text, we generalize the cross-modality attention from SCAN (Lee et al., 2018) to the multilingual scenario for learning a multilingual VSE space (Gella et al., 2017; Huang et al., 2019a).

6 Conclusion

We have presented a novel approach: *pseudo visual pivoting* for unsupervised multimodal MT. Beyond features, we use visual content to improve the cross-lingual alignments in the shared latent space. Precisely, our model utilizes the visual space as the approximate pivot for aligning the multilingual multimodal embedding space. Besides, it synthesizes image-pivoted pseudo sentences in two languages and pairs them to translate by reconstruction without parallel corpora. The experiments on Multi30K show that the proposed model generalizes well and yields new state-of-the-art performance.

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*.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

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*.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes Garcia-Martinez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost Van de Weijer. 2017. Lium-cvc submissions for wmt17 multimodal translation task. In *SECOND CONFERENCE ON MACHINE TRANSLATION*, volume 2, pages 432–439.

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 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.

Yun Chen, Yang Liu, Yong Cheng, and Victor O.K. Li. 2017. A teacher-student framework for zero-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1925–1935, Vancouver, Canada. Association for Computational Linguistics.

Yun Chen, Yang Liu, and Victor OK Li. 2018. Zero-resource neural machine translation with multi-agent communication game. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Yong Cheng, Qian Yang, Yang Liu, Maosong Sun, and Wei Xu. 2017. Joint training for pivot-based neural machine translation. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3974–3980.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 7057–7067.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018a. Word translation without parallel data. In *International Conference on Learning Representations (ICLR)*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018b. XNLI: evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2475–2485.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional gan. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2989–2998. IEEE.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

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. 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, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 130–141. Asian Federation of Natural Language Processing.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas. Association for Computational Linguistics.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2839–2845. Association for Computational Linguistics.

Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016a. Dual learning for machine translation. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 820–828.

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

Po-Yao Huang, Xiaojun Chang, and Alexander G. Hauptmann. 2019a. Multi-head attention with diversity for learning grounded multilingual multimodal representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 1461–1467. Association for Computational Linguistics.

Po-Yao Huang, Guoliang Kang, Wenhe Liu, Xiaojun Chang, and Alexander G. Hauptmann. 2019b. Annotation efficient cross-modal retrieval with adversarial attentive alignment. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 1758–1767. ACM.

Po-Yao Huang, Junwei Liang, Jean-Baptiste Lamare, and Alexander G. Hauptmann. 2018. Multimodal filtering of social media for temporal monitoring and event analysis. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ICMR 2018, Yokohama, Japan, June 11-14, 2018*, pages 450–457. ACM.

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, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 639–645. The Association for Computer Linguistics.

Po-Yao Huang, Vaibhav, Xiaojun Chang, and Alexander G. Hauptmann. 2019c. Improving what cross-modal retrieval models learn through object-oriented inter- and intra-modal attention networks. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval, ICMR 2019, Ottawa, ON, Canada, June 10-13, 2019*, pages 244–252. ACM.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent continuous translation models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1700–1709.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *NIPS Workshop*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised machine translation using monolingual corpora only. In *International Conference on Learning Representations (ICLR)*.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5039–5049, Brussels, Belgium. Association for Computational Linguistics.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. *arXiv preprint arXiv:1803.08024*.

Jindrich Libovický and Jindrich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pages 196–202. Association for Computational Linguistics.

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.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2017. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Hideki Nakayama and Noriki Nishida. 2017. Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. *Machine Translation*, 31(1-2):49–64.

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.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1179–1195.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016a. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 86–96.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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)*, volume 1, pages 1715–1725.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2019. Mass: Masked sequence to sequence pre-training for language generation. In *International Conference on Machine Learning*, pages 5926–5936.

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, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pages 543–553.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

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.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2242–2251.