Generative Imagination Elevates Machine Translation

Quanyu Long*, Mingxuan Wang, Lei Li
ByteDance AI Lab, China
oscarqylong@gmail.com, {wangmingxuan.89, lileilab}@bytedance.com

Abstract

There are thousands of languages on earth, but visual perception is shared among peoples. Existing multimodal neural machine translation (MNMT) methods achieve knowledge transfer by enforcing one encoder to learn shared representation across textual and visual modalities. However, the training and inference process heavily relies on well-aligned bilingual sentence - image triplets as input, which are often limited in quantity. In this paper, we hypothesize that visual imagination via synthesizing visual representation from source text could help the neural model map two languages with different symbols, thus helps the translation task. Our proposed end-to-end imagination-based machine translation model (ImagiT) first learns to generate semantic-consistent visual representation from source sentence, and then generate target sentence based on both text representation and imagined visual representation. Experiments demonstrate that our translation model benefits from visual imagination and significantly outperforms the text-only neural machine translation (NMT) baseline. We also conduct analyzing experiments, and the results show that imagination can help fill in missing information when performing the degradation strategy.

Introduction

Visual foundation has been introduced in a novel multimodal Neural Machine Translation (MNMT) task (Specia et al. 2016; Elliott et al. 2017; Barrault et al. 2018), which uses bilingual (or multilingual) parallel corpora annotated by images describing sentences’ contents (see Figure 1(a)). The superiority of MNMT lies in its ability to use visual information to improve the quality of translation, but its effectiveness largely depends on the availability of data sets, especially the quantity and quality of annotated images. In addition, because the cost of manual image annotation is relatively high, at this stage, MNMT is mostly applied on a small and specific dataset, Multi30K (Elliott et al. 2016), and is not suitable for large-scale text-only Neural Machine Translation (NMT) (Bahdanau, Cho, and Bengio 2014; Vaswani et al. 2017). Such limitations hinder the applicability of visual information in NMT.

To solve the bottlenecks mentioned above, Zhang et al. (2019) propose to build a lookup table from an image dataset and then using the search-based method to retrieve pictures that match the source language keywords. However, the lookup table is built from Multi30K, which leads to a relatively limited coverage of the pictures, and introduce much irrelevant noise. It does not always find the exact image corresponding to the text or the image may not even exist in the database. Elliott and Kadar (2017) present a multitask learning framework to ground visual representation to a shared space. Their architecture shares an encoder between a primary NMT task and an auxiliary task of ranking the visual features for image retrieval. However, neither the image is explicitly generated, nor the visual feature is directly leveraged by the translation decoder. Based on other researchers’ earlier exploration, we hypothesize that the potential of vision in conventional text-only NMT has not been fully dis-
covered. *A picture is worth a thousand words.* Imagining the picture of a sentence is the instinctive reaction of a human being who is learning bilingualism since an image is much more informative than a sentence.

In this paper, we propose a novel end-to-end machine translation model that is embedded in visual semantics with generative imagination (ImagiT) (see Figure 1(b)). Given a source language sentence, ImagiT first encodes it and transforms the word representations into visual features through an attentive generator, which can effectively capture the semantics of both global and local levels, and the generated visual representations can be considered as semantic-equivalent reconstructions of sentences. A simple yet effective integration module is designed to aggregate the textual and visual modalities. In the final stage, the model learns to generate the target language sentence based on the joint features. To train the model in an end-to-end fashion, we apply a visual realism adversarial loss and a text-image pair-aware adversarial loss, as well as text-semantic reconstruction loss and target language translation loss based on cross-entropy. Furthermore, we also incorporate a triplet ranking loss to make the generated visual feature close to the matched text in the latent visual semantic space.

In contrast with prior MNMT work, our proposed ImagiT model does not require images as input during the inference time but can leverage visual information through imagination, making it an appealing method in low-resource scenario. Moreover, ImagiT is much more flexible and general, accepting external parallel text data or non-parallel image captioning data. We evaluate our Imagination modal on the Multi30K dataset. The experiment results show that our proposed method significantly outperforms the text-only NMT baseline. The analysis demonstrates that imagination helps the translation.

To summarize, the paper has the following contributions:
1. We propose generative imagination, a new setup for machine translation assisted by synthesized visual representation, without annotated images as input;
2. We propose the ImagiT method, which shows advantages over the conventional MNMT model and gains significant improvements over the text-only NMT baseline;
3. We conduct experiments to verify and analyze how imagination helps the translation.

**Related work**

**MNMT** As a language shared by people worldwide, visual modality may help machines have a more comprehensive perception of the real world. Multimodal neural machine translation (MNMT) is a novel machine translation task proposed by the machine translation community, which aims to design multimodal translation frameworks using context from the additional visual modality (Specia et al. 2016). The shared task release the dataset Multi30K (Elliott et al. 2016), which is an extended German version of Flickr30K (Young et al. 2014), then expanded to French and Czech (Elliott et al. 2017) (Barrault et al. 2018). In the three versions of tasks, scholars have proposed many multimodal machine translation models and methods. Huang et al. (2016) encodes word sequences with regional visual objects, while Calixto, Liu, and Campbell (2017b) study the effects of incorporating global visual features to initialize the encoder/decoder hidden states of RNN. Caglayan et al. (2017) models the image-text interaction by leveraging element-wise multiplication. Elliott and Kadar (2017) proposed a multitask learning framework to ground visual representation to a shared space and learn with the auxiliary triplet alignment task. The common practice is to use convolutional neural networks to extract visual information and then using attention mechanisms to extract visual contexts (Caglayan, Barrault, and Bougares 2016; Calixto, Elliott, and Frank 2016; Libovicky and Heil 2017). Ives, Madhyastha, and Specia (2019) propose a translate-and-refine approach using two-stage decoder. Calixto, Rios, and Aziz (2018) put forward a latent variable model to capture the multimodal interactions between visual and textual features. Caglayan et al. (2019) show that visual content is more critical when the textual content is limited or uncertain in MMT. Recently, Yao and Wan (2020) propose multi-modal self-attention in Transformer to avoid encoding irrelevant information in images, and Yin et al. (2020) propose a graph-based multimodal fusion encoder to capture various relationships.

**Text-to-image synthesis** Traditional Text-to-image (T2I) synthesis mainly uses keywords to search for small image regions, and finally optimizes the entire layout (Zhu et al. 2007). After generative adversarial networks (GANs) (Goodfellow et al. 2014) were proposed, scholars have presented a variety of GAN-based T2I models. Reed et al. (2016) propose DC-GAN and design a direct and straightforward network and a training strategy for T2I generation. Zhang et al. (2017) propose stackGAN, which contains multiple cascaded generators and discriminators, and the higher stage generates better quality pictures. In previous work, scholars only considered global semantics. Xu et al. (2018) proposed AttnGAN to apply the attention mechanism to capture fine-grained word-level information. MirrorGAN (Qiao et al. 2019) employs a mirror structure, which reversely learns from the inverse task of T2I to further validate whether generated images are consistent with the input texts. The inverse task is also known as image captioning.

**ImagiT model**

As shown in Figure 2, ImagiT embodies the encoder-decoder structure for end-to-end machine translation. Between the encoder and the decoder, there is an imagination step to generate semantic-equivalent visual representation. Technically, our model is composed of following modules: source text encoder, generative imagination network, redescription, aligning VSE space, and decoder for translation. We will elaborate on each of them in the rest of this section.

**Source text encoder**

Vaswani et al. (2017) propose the state-of-art Transformer-based machine translation framework, which can be written
as follows:

$$H^l = LN(Att^l(Q^{l-1}, K^{l-1}, V^{l-1}) + H^{l-1}),$$

$$H^l = LN(FFN^l(H^l) + H^{l-1}),$$

Where $Att^l$, $LN$, and $FFN^l$ are the self-attention module, layer normalization, and the feed-forward network for the $l$-th identical layer respectively. The core of the Transformer is the multi-head self-attention, in each attention head, we have:

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V),$$

$$\alpha_{ij} = softmax\left(\frac{(x_i W^Q)(x_j W^K)}{\sqrt{d}}\right),$$

$W^V, W^Q, W^K$ are layer-specific trainable parameter matrices. For the output of final stacked layer, we use $w = \{w_0, w_1, ..., w_{L-1}\}, w \in \mathbb{R}^{d \times L}$ to represent the source word embedding, $L$ is the length of the source sentence. Besides, we add a special token to each source language sentence to obtain the sentence representation $s \in \mathbb{R}^d$.

**Generative imagination network**

Generative Adversarial Network (Goodfellow et al. 2014) has been applied to synthesis images similar to ground truth (Zhang et al. 2017; Xu et al. 2018; Qiao et al. 2019). We follow the common practice of using the conditioning augmentation (Zhang et al. 2017) to enhance robustness to small perturbations along the conditioning text manifold. $F^{ca}$ represents the conditioning augmentation function, and $s^{ca}$ represents the enhanced sentence representation.

$$s^{ca} = F^{ca}(s),$$

$\{F_0, F_1\}$ are two visual feature converters, including perceptron, residual and upsampling layers. Furthermore, we define $\{f_0, f_1\}$ are the visual features after two transformations. Owing to space constraints, for detailed layer structure, please refer to (Xu et al. 2018).

$$f_0 = F_0(z, s^{ca}),$$

$$f_1 = F_1(f_0, F^{attn}(f_0, s^{ca})),$$

$z$ is the noise vector, sampled from the standard normal distribution, and it will be concatenated with $s^{ca}$. $f_0 \in \mathbb{R}^{M_0 \times N_0}$. Each column of $f_0$ is a feature vector of a sub-region of the image, which can also be treat as a pseudo-token. To generate fine-grained details at different sub-regions of the image by paying attention to the relevant words in the source language, we use image vector in each sub-region to query word vectors by leveraging attention strategy. $F^{attn}$ is an attentive function to obtain word-context feature, then we have:

$$F^{attn}(f_0, s^{ca}) = \sum_{l=0}^{L-1}(U_0 w_l)(softmax(f_0^T(U_0 w_l)))^T,$$

Figure 2: Overview of the framework of the proposed ImagiT. $F_0$ and $F_1$ are text-to-image converters, comprising of perceptron, residual, and unsampling blocks. $L \times$ represents L identical layers.
Word feature $w_l$ is firstly converted into the common semantic space of the image feature, $U_0$ is a perceptron layer. Then it will be multiplied with visual feature $f_0$ to acquire the attention score. $f_j$ is the output of the imagination network, capturing multiple levels (word level and sentence level) of semantic meaning. It will be utilized directly for target language generation, and it will also be passed to the discriminator for adversarial training.

**Redescription**

Redescription, also known as image captioning, can be regarded as the inverse problem of text-to-image generation, which generates the given image’s description. If an imagined image is semantic equivalent to the source sentence, then its description should be almost identical to the given text. Thus we leverage the redescription to translate the imagined visual representation back to the source language (Qiao et al. 2019), and this symmetric structure can make the imagined visual feature act like a mirror, effectively enhancing the semantic consistency of the imagined visual feature and precisely reflect the underlying semantics. Following Qiao et al. (2019), we utilize the widely used encoder-decoder image captioning framework (Vinyals et al. 2015), and fix the parameter of the pre-trained image captioning framework when end-to-end training other modules in ImagiT.

$$p_t = \text{Decoder}(h_{t-1}), t = 0, 1, ..., L - 1. \quad (9)$$

$$\mathcal{L}_{\text{redes}} = - \sum_{t=0}^{L-1} \log p_t(T_1), \quad (10)$$

$p_t$ is the predicted probability distribution over the words at $t$th decoding step, and $T_1$ is the $T_1$-th entry of the probability vector.

**Aligning the VSE space**

In this section, we align a shared visual-semantic embedding (VSE) space to find a good distributed representation that can capture the semantic meaning across two modalities. Drawing the generated visual representation and the corresponding sentence closer contributes to better semantic alignment in the language latent space, and it can also encourage the imagination network to validate the semantic consistency further.

Specifically, we leverage the contrastive triplet loss in cross-modal retrieval to encourage the textual-visual alignment in the VSE space (Kiros, Salakhutdinov, and Zemel 2014; Lee et al. 2018). Furthermore, we utilize fine-grained and bidirectional (textual-to-visual, visual-to-textual) attention mechanisms to learn to value the imagined visual representation with close semantic relatedness to the source sentence. We denote $h^f_j$ to represent object-level visual feature, and $h^w$ to represent the token-level textual feature. The visually-attend textual representation $h^{fw}_j$ is the weighted combination of $h^w$. Similarly, $h^{wf}_i$ represents textually-attend visual representation. For visual-aware attention, we have:

$$h^{fw}_j = \sum_{i=0}^{L-1} \alpha_{ij} h^w_i, \quad (11)$$

$$\alpha_{ij} = \text{softmax}_i(\cos(h^w_i, h^{f}_j)), \quad (12)$$

Where $j = 0, ..., M - 1$, $L$ is the sentence length, and $M$ is the visual feature maps. Then we get:

$$S(f, w) = \frac{1}{2M} \sum_{j=0}^{M-1} \cos(h^{fw}_j, h^f_j) \quad (13)$$

$$+ \frac{1}{2L} \sum_{i=0}^{L-1} \cos(h^{wf}_i, h^w_i)$$

$$\mathcal{L}_c(f, w) = \max_{f} \gamma - S(f, w) + S(\tilde{f}, \tilde{w})_+ \quad (14)$$

$$+ \max_{\tilde{w}} \gamma - S(f, w) + S(f, \tilde{w})_+$$

$S(f, w)$ is the cosine similarity between a sentence and an image, $\mathcal{L}_c(f, w)$ is the contrastive triplet loss. $\tilde{f}$ and $\tilde{w}$ are the contrastive non-paired example with respect to the selected source text and the selected imagined visual representation. $|+|$ is the hinge loss function. When the loss decreases, the distance between a paired image and sentence will drop while the distance between an unpaired image and sentence increases.

**Multimodal aggregation**

After obtaining the imagined visual representation, we aggregate two modalities for the translation decoder. Although the vision carries richer information, it also contains irrelevant noise. Comparing to encoding and integrating visual feature directly, a more elegant method is to induce the hidden representation under the guide of image-aware attention and graph perspective of transformer (Yao and Wan 2020), since each local spatial regions of the image can also be considered as pseudo-tokens, which can be added to the source fully-connected graph. In the multimodal self-attention layer, we add the spatial feature of the generated feature map in the source sentence, getting $\tilde{x} \in \mathbb{R}^{(L+M) \times d}$, then perform image-aware attention, we have:

$$c_i = \sum_{j=0}^{L-1} \tilde{\alpha}_{ij} (w_j W^K V^T) \quad (15)$$

$$\tilde{\alpha}_{ij} = \text{softmax}_i \left( \frac{(\tilde{x}_i W^Q)(w_j W^K)}{\sqrt{d}} \right) \quad (16)$$

**Objective function**

During the translation phase, similar to equation [10] we have:

$$\mathcal{L}_{\text{trans}} = - \sum_{t} \log p_t(T_t) \quad (17)$$
The function of the generator is defined as: the input text and the generated image. The final objective is to guarantee the semantic consistency between the generator generates visually realistic images. The second term is to distinguish real and fake, ensuring that the generator generates visually realistic images, and the real image as input, we employ two adversarial losses: a visual realism adversarial loss, and a text-image pair-aware adversarial loss computed by the discriminator. Especially, as shown in Figure 3, the discriminator takes source language sentence, and the real image as input, and we employ two adversarial training to alternatively train the generator and the discriminator. Especially, as shown in Figure 3, the discriminator takes source language sentence, and the real image as input, and we employ two adversarial losses: a visual realism adversarial loss, and a text-image pair-aware adversarial loss computed by the discriminator (Zhang et al. 2017; Xu et al. 2018; Qiao et al. 2019).

\[
L_{G_0} = \frac{1}{2} \mathbb{E}_{I_i \sim p_{I_i}} \log(D(I_i)) - \frac{1}{2} \mathbb{E}_{I_i \sim p_{I_i}} \log(1 - D(I_i, s)) \tag{18}
\]

The first term is to distinguish real and fake, ensuring that the generated visual feature is 128 × 128. The upsampling and residual block in visual feature transformers consist of 3 × 3 stride 1 convolution, batch normalization, and ReLU activation. The training is early-stopped if the dev set BLEU score do not improve for 10 epochs, since the translation is the core task. The batch size is 64, and the learning rate is initialized to be 2e−4 and decayed to half of its previous value every 100 epochs. A similar learning schedule is adopted in Zhang et al. (2017). The margin

**Datasets**

We evaluate our proposed ImagIT model on two datasets, Multi30K (Elliott et al. 2016) and Ambiguous COCO (Elliott et al. 2017). To show its ability to train with external out-of-domain datasets, we adopt MS COCO (Lin et al. 2014) in the next analyzing section.

Multi30K is the largest existing human-labeled collection for MNMT, containing 31K images and consisting of two multilingual expansions of the original Flickr30K (Young et al. 2014) dataset. The first expansion has five English descriptions and five German descriptions, and they are independent of each other. The second expansion has one of its English description manually translated to German by a professional translator, then expanded to French and Czech in the following shared task (Elliott et al. 2017; Barrault et al. 2018). We only apply the second expansion in our experiments, which has 29,000 instances for training, 1,014 for development, and 1,000 for evaluation. We present our results on English-German (En-De) Test2016 and Test2017.

Ambiguous COCO is a small evaluation dataset collected in the WMT2017 multimodal machine translation challenge (Elliott et al. 2017), which collected and translated a set of image descriptions that potentially contain ambiguous verbs. It contains 461 images from the MS COCO (Lin et al. 2014) for 56 ambiguous vers in total.

MS COCO is the widely used non-parallel text-image paired dataset in T2I and I2T generation. It contains 82,783 training images and 40,504 validation images with 91 different object types, and each image has 5 English descriptions.

**Settings**

Our baseline is the conventional text-only Transformer (Vaswani et al. 2017). Specifically, each encoder-decoder has a 6-layer stacked Transformer network, eight heads, 512 hidden units, and the inner feed-forward layer filter size is set to 2048. The dropout is set to \( p = 0.1 \), and we use Adam optimizer (Kingma and Ba 2014) to tune the parameter. The learning rate increases linearly for the warmup strategy with 8,000 steps and decreases with the step number's inverse square root. We train the model up to 10,000 steps, the early-stop strategy is adopted. We use the same setting as Vaswani et al. (2017). We use the metrics BLEU (Papineni et al. 2002) and METEOR (Denkowski and Lavie 2014) to evaluate the translation quality.

For the imagination network, the noise vector’s dimension is 100, and the generated visual feature is 128 × 128. The upsampling and residual block in visual feature transformers consist of 3 × 3 stride 1 convolution, batch normalization, and ReLU activation. The training is early-stopped if the dev set BLEU score do not improve for 10 epochs, since the translation is the core task. The batch size is 64, and the learning rate is initialized to be 2e−4 and decayed to half of its previous value every 100 epochs. A similar learning schedule is adopted in Zhang et al. (2017). The margin
Table 1: Main result from the Test2016, Test2017, and Ambiguous COCO for the En⇒De MNMT task. The first category collects the existing MNMT systems, which take both source sentences and paired images as input. The second category illustrates the systems that do not require images as input. Since our method falls into the second group, our baseline are the text-only transformer (Vaswani et al. 2017) and the aforementioned works (Zhang et al. 2019; Elliott and Kádár 2017).

| Model                  | Test2016 BLEU | Test2017 BLEU | Test2017 METEOR | Ambiguous COCO BLEU | Ambiguous COCO METEOR |
|------------------------|---------------|---------------|-----------------|---------------------|-----------------------|
| IMGD (Calixto, Liu, and Campbell 2017b) | 37.3          | 55.1          | N/A             | N/A                 | N/A                   |
| NMT_SRC+IMG (Calixto, Liu, and Campbell 2017a) | 36.5          | 55.0          | N/A             | N/A                 | N/A                   |
| trg-mul (Caglayan et al. 2017) | 37.8          | 57.7          | 30.7            | 52.2                | 27.1                  |
| VMMT (Calixto, Rios, and Aziz 2018) | 37.6          | 56.0          | N/A             | N/A                 | N/A                   |
| VAG-NMT (Zhou et al. 2018) | N/A           | N/A           | 31.6            | 52.2                | 28.3                  |
| Transformer+Att (Ive, Madhyastha, and Specia 2019) | 38.0          | 55.6          | N/A             | N/A                 | N/A                   |
| Multimodal transformer (Yao and Wan 2020) | 38.7          | 55.7          | N/A             | N/A                 | N/A                   |
| Transformer baseline (Vaswani et al. 2017) | 37.6          | 55.3          | 31.7            | 52.1                | 27.9                  |
| Multitask (Elliott and Kádár 2017) | 36.8          | 55.8          | 30.2            | 51.2                | N/A                   |
| Lookup table (Zhang et al. 2019) | 36.9          | N/A           | 28.6            | N/A                 | N/A                   |
| Our ImagiT             | 38.4          | 55.7          | 32.1            | 52.4                | 28.7                  |

Results

Table 1 illustrates the results for the Test2016, Test2017, and the Ambiguous COCO MNMT tasks. Our text-only Transformer baseline (Vaswani et al. 2017) has similar results compared to most prior MNMT works, which is consistent with the previous findings (Caglayan et al. 2019), that is, textual modality is good enough to translate for Multi30K dataset. This finding helps to explain that it is already tricky for a MNMT model to ground visual modality even with the presence of annotated images. However, Our ImagiT gains extensive improvements over the text-only Transformer baseline on three evaluation datasets, demonstrating that our model can effectively embed the visual semantics during the training time and guide the translation through imagination with the absence of annotated images during the inference time. We assume much of the performance improvement is due to ImagiT’s strong ability to capture the interaction between text and image, generate semantic-consistent visual representations, and incorporate information from visual modality properly.

We also observe that our approach surpasses the results of most MNMT systems by a noticeable margin in terms of BLEU score and METEOR score on three evaluation dataset, and is also competitive with the state-of-the-art MNMT framework (Yao and Wan 2020) on Test2016. Especially for Ambiguous COCO, which was purposely curated such that verbs have ambiguous meaning, demands more visual contribution for guiding the translation and selecting correct words. Our ImagiT benefits from visual imagination and substantially outperforms previous works on ambiguous COCO.

Analysis

Can ImagiT generate visual grounded representations?

Since the proposed model does not require images as input, one may ask how it uses visual information and where the information comes? We claim that ImagiT has already been embedded with visual semantics during the training phase, and in this section, we validate that ImagiT is able to generate visual grounded representation by performing the image retrieval task.

For each source sentence, we generate the intermediate visual representation. Furthermore, we query the ground truth image features for each generated representation to find the closest image vectors around it based on the cosine similarity. Then we can measure the $R@K$ score, which computes the recall rate of the matched image in the top K nearest neighborhoods.

Table 2: Image retrieval task. We evaluate on Multi30K and MS COCO.

| Model         | R@1 | R@5 | R@10 |
|---------------|-----|-----|------|
| ImagiT on Multi30K | 64.7 | 88.7 | 94.2 |
| ImagiT on MS COCO   | 64.3 | 89.5 | 94.7 |

Some previous studies on VSE perform sentence-to-image retrieval and image-to-sentence retrieval, but their results can not be directly compared with ours, since we are performing image-to-image retrieval in practical. However, from Table 2, the results demonstrate that our generated representation has the excellent quality of learned shared semantics and have been grounded with visual semantic-consistency.
How does the imagination help the translation?

Although we have validated the effectiveness of ImagiT on three widely used MNMT evaluation datasets. A natural question to ask is how does the imagination guide the translation, and to which extent? When human beings confronting with complicate sentences and obscure words, we often resort to mind-picturing and mental visualization to assist us to auto-complete and fill the whole imagination. Thus we hypothesis that imagination could help recover and retrieve the missing and implicate textual information.

Inspired by [Ive, Madhyastha, and Specia (2019); Caglayan et al. (2019)], we apply degradation strategy to the input source language, and feed to the trained Transformer baseline, MNMT baseline, and ImagiT respectively, to validate if our proposed approach could recover the missing information and obtain better performance. And we conduct the analysing experiments on En-De Test2016 evaluation set.

Color deprivation is to mask the source tokens that refers to colors, and replace them with a special token [M]. Under this circumstance, text-only NMT model have to rely on source-side contextual information and biases, while for MNMT model, it can directly utilize the paired color-related information-rich images. But for ImagiT, the model will turn to imagination and visualization.

Table 3: Color deprivation. s represents the original source sentence, while ñ is the degraded sentence.

| Model       | S     | S     |
|-------------|-------|-------|
| text-only Transformer | 37.6  | 36.3  |
| MNMT        | 38.2  | 37.7  |
| ImagiT      | 38.4  | 37.9  |

Table 3 demonstrates the results of color deprivation. We implement a simple transformer-based MNMT baseline model using the multimodal self-attention approach [Yao and Wan 2020]. Thus the illustrated three models in Table 3 can be compared directly. We can observe that the BLEU score of text-only NMT decreases 1.3, whereas MNMT and ImagiT system only decreases 0.5. This result corroborates that our ImagiT has a similar ability to recover color compared to MNMT, but our ImagiT achieves the same effect through its own efforts, i.e., imagination.

Visually depictable entity masking [Plummer et al. (2015)] extend Flickr30K with coreference chains to tag mentions of visually depictable entities. Similar to color deprivation, we randomly replace 0%, 15%, 30%, 45%, 60% visually depictable entities with a special token [M].

Figure 4 is the result of visually depictable entity masking. We observe a large BLEU score drop of text-only Transformer baseline with the increasing of masking proportion, while MNMT and ImagiT are relatively smaller. This result demonstrates that our ImagiT model can much more effectively infer and imagine missing entities compared to text-only Transformer, and have comparable capability over the MNMT model.

Will better imagination with external data render better translation?

Our ImagiT model also accepts external parallel text data or non-parallel image captioning data, and we can easily modify the objective function to train with out-of-domain non-triple data. To train with text-image paired image captioning data, we can pre-train our imagination model by ignoring $L_{trans}$ term. In other words, the T2I synthesis module can be solely trained with MS COCO dataset. We randomly split MS COCO in half, and use COCO$_{half}$ and COCO$_{full}$ to pre-train ImagiT. The MS COCO is processed using the same pipeline as in [Zhang et al. (2017)]. Furthermore, the training setting of COCO$_{half}$ and COCO$_{full}$ are the same with batch size 64 and maximum epoch 600. The results are:

Table 4: Translation results when using out-of-domain non-parallel image captioning data.

| Model             | BLEU | METEOR |
|-------------------|------|--------|
| ImagiT            | 38.4 | 55.7   |
| ImagiT + COCO$_{half}$ | 38.6 | 56.3   |
| ImagiT + COCO$_{full}$ | 38.7 | 56.7   |

Table 4 demonstrates the results of color deprivation. We implement a simple transformer-based MNMT baseline model using the multimodal self-attention approach [Yao and Wan 2020]. Thus the illustrated three models in Table 4 can be compared directly. We can observe that the BLEU score of text-only NMT decreases 1.3, whereas MNMT and ImagiT system only decreases 0.5. This result corroborates that our ImagiT has a similar ability to recover color compared to MNMT, but our ImagiT achieves the same effect through its own efforts, i.e., imagination.

Visually depictable entity masking [Plummer et al. (2015)] extend Flickr30K with coreference chains to tag mentions of visually depictable entities. Similar to color deprivation, we randomly replace 0%, 15%, 30%, 45%, 60% visually depictable entities with a special token [M].

Figure 4 is the result of visually depictable entity masking. We observe a large BLEU score drop of text-only Transformer baseline with the increasing of masking proportion, while MNMT and ImagiT are relatively smaller. This result demonstrates that our ImagiT model can much more effectively infer and imagine missing entities compared to text-only Transformer, and have comparable capability over the MNMT model.

As is shown in Table 4, our ImagiT model pre-trained with half MS COCO gain 0.6 METEOR increase, and the improvement becomes more apparent when training with the whole MS COCO. We can contemplate that large-scale external data may further improve the performance of ImagiT, and we have not utilized parallel text data (e.g., WMT), even image-only and monolingual text data can also be adopted to enhance the model capability, and we leave this for future work.

Conclusion

This work presents generative imagination-based machine translation model (ImagiT), which can effectively capture the source semantics and generate semantic-consistent visual representations for imagination-guided translation. Without annotated images as input, out model gains significant improvements over text-only NMT baselines and is comparable with the SOTA MNMT model. We analyze how imagination elevates machine translation and show improvement using external image captioning data. Further work may center around introducing more parallel and non-parallel, text, and image data for different training schemes.
References
Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Barrault, L.; Bougares, F.; Specia, L.; Lala, C.; Elliott, D.; and Frank, S. 2018. Findings of the Third Shared Task on Multimodal Machine Translation. In Proceedings of the Third Conference on Machine Translation: Shared Task Papers, 304–323.

Caglayan, O.; Aransa, W.; Bardet, A.; García-Martínez, M.; Bougares, F.; Barrault, L.; Masana, M.; Herranz, L.; and Van de Weijer, J. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. arXiv preprint arXiv:1707.04481.

Caglayan, O.; Barrault, L.; and Bougares, F. 2016. Multimodal attention for neural machine translation. arXiv preprint arXiv:1609.03976.

Caglayan, O.; Madhyastha, P.; Specia, L.; and Barrault, L. 2019. Probing the need for visual context in multimodal machine translation. arXiv preprint arXiv:1903.08678.

Calixto, I.; Elliott, D.; and Frank, S. 2016. DCU-UvA multimodal MT system report. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, 634–638.

Calixto, I.; Liu, Q.; and Campbell, N. 2017a. Doubly-attentive decoder for multi-modal neural machine translation. arXiv preprint arXiv:1702.01287.

Calixto, I.; Liu, Q.; and Campbell, N. 2017b. Incorporating global visual features into attention-based neural machine translation. arXiv preprint arXiv:1701.06521.

Calixto, I.; Rios, M.; and Aziz, W. 2018. Latent variable model for multi-modal translation. arXiv preprint arXiv:1811.00357.

Denkowski, M.; and Lavie, A. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In Proceedings of the Ninth Workshop on Statistical Machine Translation, 376–380.

Elliott, D.; Frank, S.; Barrault, L.; Bougares, F.; and Specia, L. 2017. Findings of the second shared task on multimodal machine translation and multilingual image descriptionfind. arXiv preprint arXiv:1710.07177.

Elliott, D.; Frank, S.; Simaan, K.; and Specia, L. 2016. Multi30k: Multilingual english-german image descriptions. arXiv preprint arXiv:1605.00459.

Elliott, D.; and Kádár, A. 2017. Imagination improves multimodal translation. arXiv preprint arXiv:1705.04350.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In Advances in neural information processing systems, 2672–2680.

Huang, P.-Y.; Liu, F.; Shiang, S.-R.; Oh, J.; and Dyer, C. 2016. Attention-based multimodal neural machine translation. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, 639–645.

Ive, J.; Madhyastha, P.; and Specia, L. 2019. Distilling translations with visual awareness. arXiv preprint arXiv:1906.07701.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In Proceedings of the European Conference on Computer Vision (ECCV), 201–216.

Libovický, J.; and Helcl, J. 2017. Attention strategies for multi-source sequence-to-sequence learning. arXiv preprint arXiv:1704.06567.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In European conference on computer vision, 740–755. Springer.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, 311–318.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE international conference on computer vision, 2641–2649.

Qiao, T.; Zhang, J.; Xu, D.; and Tao, D. 2019. Mirrorgan: Learning text-to-image generation by redescription. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1505–1514.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396.

Specia, L.; Frank, S.; Simaan, K.; and Elliott, D. 2016. A shared task on multimodal machine translation and crosslingual image description. In Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers, 543–553.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In Advances in neural information processing systems, 5998–6008.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In Proceedings of the IEEE conference on computer vision and pattern recognition, 3156–3164.

Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1316–1324.
Yao, S.; and Wan, X. 2020. Multimodal Transformer for Multimodal Machine Translation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 4346–4350.

Yin, Y.; Meng, F.; Su, J.; Zhou, C.; Yang, Z.; Zhou, J.; and Luo, J. 2020. A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation. arXiv preprint arXiv:2007.08742.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics 2: 67–78.

Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision, 5907–5915.

Zhang, Z.; Chen, K.; Wang, R.; Utiyama, M.; Sumita, E.; Li, Z.; and Zhao, H. 2019. Neural Machine Translation with Universal Visual Representation. In International Conference on Learning Representations.

Zhou, M.; Cheng, R.; Lee, Y. J.; and Yu, Z. 2018. A visual attention grounding neural model for multimodal machine translation. arXiv preprint arXiv:1808.08266.

Zhu, X.; Goldberg, A. B.; Eldawy, M.; Dyer, C. R.; and Strock, B. 2007. A text-to-picture synthesis system for augmenting communication. In AAAI, volume 7, 1590–1595.