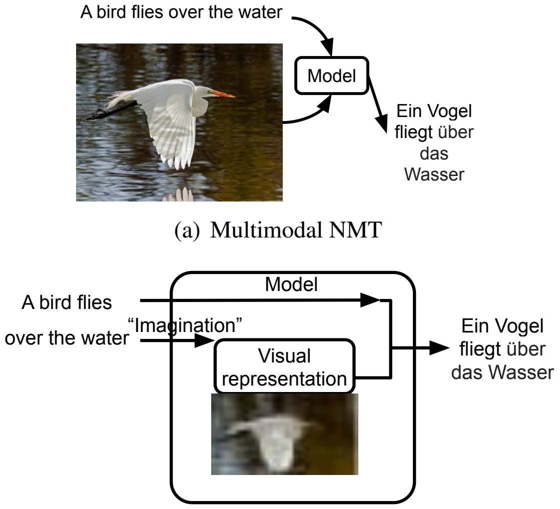
Generative Imagination Elevates Machine Translation

Quanyu Long[1], **Mingxuan Wang**[2], **and Lei Li**[2]
[1]Nanyang Technological University, Singapore
[2]ByteDance AI Lab, China
quanyu001@e.ntu.edu.sg;
{wangmingxuan.89, lileilab}@bytedance.com

Abstract

There are common semantics shared across text and images. Given a sentence in a source language, whether depicting the visual scene helps translation into a target language? Existing multimodal neural machine translation methods (MNMT) require triplets of bilingual sentence - image for training and tuples of source sentence - image for inference. In this paper, we propose ImagiT, a novel machine translation method via visual imagination. ImagiT first learns to generate visual representation from the source sentence, and then utilizes both source sentence and the "imagined representation" to produce a target translation. Unlike previous methods, it only needs the source sentence at the inference time. Experiments demonstrate that ImagiT benefits from visual imagination and significantly outperforms the text-only neural machine translation baselines. Further analysis reveals that the imagination process in ImagiT helps fill in missing information when performing the degradation strategy.


A bird flies over the water
Model
Ein Vogel fliegt über das Wasser
A bird flies over the water
Model
"Imagination"
Visual representation
Ein Vogel fliegt über das Wasser

(a) Multimodal NMT

(b) ImagiT

Figure 1: The problem setup of our proposed ImagiT is different from existing multimodal NMT. A multimodal NMT model takes both text and paired image as the input, while ImagiT takes only sentence in the source language as the usual NMT task. ImagiT synthesizes an image and utilize the internal visual representation to assist translation.

1 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 et al., 2015; Vaswani et al., 2017). Such limitations hinder the applicability of visual information in NMT.

To address the bottlenecks mentioned above, Zhang et al. (2020) 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 potentially introduces 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 Kádár (2017) present a multitask learning framework to ground visual representation to a shared space. Their architecture called "imagination" shares an encoder between a primary NMT task and an auxiliary task of ranking the vi-

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5738–5748
June 6–11, 2021. ©2021 Association for Computational Linguistics

sual features for image retrieval. However, neither the image is explicitly generated, nor the visual feature is directly leveraged by the translation decoder, the model simply learns the visual grounded shared encoder. Based on other researchers' earlier exploration, we hypothesize that the potential of vision in conventional text-only NMT has not been fully discovered. Different with Elliott and Kádár (2017) implicit approach, we understand "imagination" to be more like "picturing", since it is similar to humans who can visually depict figures in the mind from an utterance. Our approach aims to explicitly imagine a "vague figure" (see Figure 1(b)) to guide the translation, since *A picture is worth a thousand words*, and imagining the picture of a sentence is the instinctive reaction of a human being who is learning bilingualism.

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.

In contrast with most 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 also flexible, 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 help the model complete the missing information in the sentence when we perform degradation masking, and we also see improvements in translation quality by pre-training the model with an external non-parallel image captioning dataset.

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.

2 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 releases 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 and Liu (2017) 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 Kádár (2017) propose 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 et al., 2016; Calixto et al., 2016; Libovický and Helcl, 2017). Ive et al. (2019) propose a translate-and-refine approach using two-stage decoder. Calixto et al. (2019) 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 multimodal 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.

3 ImagiT model

As shown in Figure2, 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, image captioning, multimodal aggregation and decoder for translation. We will elaborate on each of them in the rest of this section.

3.1 Source text encoder

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

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

$$\mathbf{H}^{l} = LN(FFN^{l}(\overline{\mathbf{H}}^{l}) + \overline{\mathbf{H}}^{l}), \quad (2)$$

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:

$$z_i = \sum_{j=1}^{n} \alpha_{ij}(x_j W^V), \quad (3)$$

$$\alpha_{ij} = softmax(\frac{(x_i W^Q)(x_j W^K)^\top}{\sqrt{d}}). \quad (4)$$

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 $s \in \mathbb{R}^d$.

3.2 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 and improve the diversity of generated samples.[1] F^{ca} represents the conditioning augmentation function, and s^{ca} represents the enhanced sentence representation.

$$s^{ca} = F^{ca}(s), \quad (5)$$

$\{F_0, F_1\}$ are two visual feature converters, sharing similar architecture. F_0 contains a fully connected layer and four deconvolution layers (Noh et al., 2015) to obtain image-sized feature vectors. Furthermore, we define $\{f_0, f_1\}$ are the visual features after two transformations with different resolution. For detailed layer structure and block design, please refer to (Xu et al., 2018).

$$f_0 = F_0(z, s^{ca}), \quad (6)$$

$$f_1 = F_1(f_0, F^{attn}(f_0, s^{ca})), \quad (7)$$

[1] Zhang et al. (2017) also mentions that the randomness in the Conditioning Augmentation is beneficial for modeling text to image semantic translation as the same sentence usually corresponds to objects with various poses and appearances.

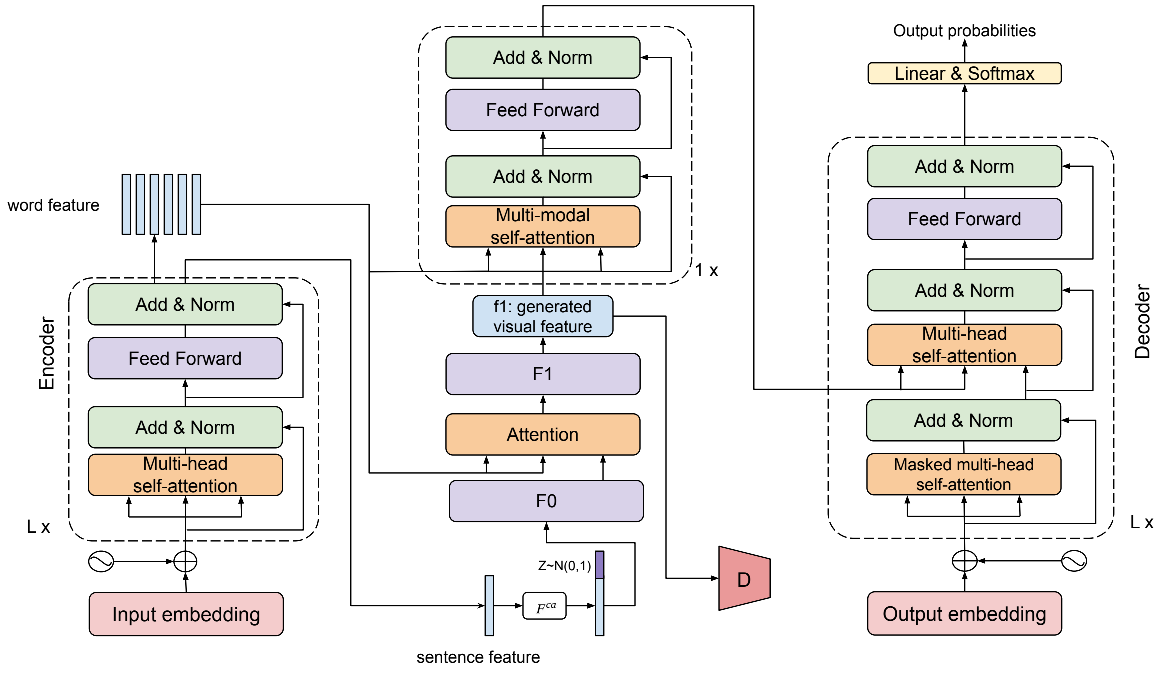


Figure 2: Overview of the framework of the proposed ImagiT. F_0 and F_1 are text-to-image converters, sharing similar structures, comprising of perceptron, residual, and unsampling blocks. L× represents L identical layers. Noting that we only need to obtain the generated visual feature to guide the translation, for the whole pipeline, up-sampling this feature to image is redundant.

Where $f_0 \in \mathbb{R}^{M_0 \times N_0}$, z is the noise vector, sampled from the standard normal distribution, and it will be concatenated with s^{ca}. Each column of f_i 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 subregions 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:

$$F^{attn}(f_0, s^{ca}) = \sum_{l=0}^{L-1} (U_0 w_l)(softmax(f_0^T(U_0 w_l)))^\top, \tag{8}$$

Word feature w_l is firstly converted into the common semantic space of the visual feature, U_0 is a perceptron layer. Then it will be multiplied with f_0 to acquire the attention score. f_1 is the output of the imagination network, capturing multiple levels (word level and sentence level) of semantic meaning. f_1 is denoted as the blue block "generated visual feature" in Figure2. It will be utilized directly for target language generation, and it will also be passed to the discriminator for adversarial training. Note that for the whole pipeline, upsampling f_1 to an image is redundant.

Comparing to T2I synthesis works which use cascaded generators and disjoint discriminators(Zhang et al., 2017; Xu et al., 2018; Qiao et al., 2019), we only use one stage to reduce the model size and make our generated visual feature f_1 focus more on text-mage consistency, but not the realism and authenticity.

3.3 Image captioning

Image captioning (I2T) can be regarded as the inverse problem of text-to-image generation, generating 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 image captioning 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 parameters of the pre-trained image captioning framework when end-to-end training other modules in

ImagiT.

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

$$\mathcal{L}_{I2T} = -\sum_{t=0}^{L-1} \log p_t(T_t). \quad (10)$$

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

3.4 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, that is, the attention query vector is the combination of text and visual embeddings, getting $\tilde{x} \in \mathbb{R}^{(L+M)\times d}$. Then perform image-aware attention, the key and value vectors are just text embeddings, we have:

$$c_i = \sum_{j=0}^{L-1} \tilde{\alpha}_{ij}(w_j W^V), \quad (11)$$

$$\tilde{\alpha}_{ij} = softmax(\frac{(\tilde{x}_i W^Q)(w_j W^K)^\top}{\sqrt{d}}). \quad (12)$$

3.5 Objective function

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

$$\mathcal{L}_{trans} = -\sum_{t} \log p_t(T_t), \quad (13)$$

To train the whole network end-to-end, we leverage adversarial training to alternatively train the generator and the discriminator. Especially, as shown in Figure 3, the discriminator take the imagined visual representation, 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).

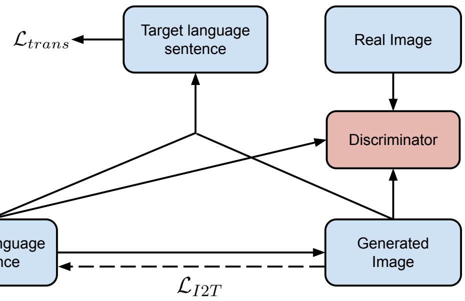


Figure 3: Training objective. The discriminator takes source language sentences, generated images, and real images as input, then computes two adversarial loss: realism loss and text-image paired loss. $\mathcal{L}_{I2T}$ is designed to guarantee the semantic consistency, and $\mathcal{L}_{trans}$ is the core loss function to translate integrated embedding to the target language.

$$\begin{aligned}\mathcal{L}_{G_0} = &-\frac{1}{2}\mathbb{E}_{f_1\sim p_G}[\log(D(f_1)] \\ &-\frac{1}{2}\mathbb{E}_{f_1\sim p_G}[\log(D(f_1, s)],\end{aligned} \quad (14)$$

f_1 is the generated visual feature computed by equation 7 from the model distribution p_G, s is the global sentence vector. The first term is to distinguish real and fake, ensuring that the generator generates visually realistic images. The second term is to guarantee the semantic consistency between the input text and the generated image. $\mathcal{L}_{G_0}$ jointly approximates the unconditional and conditional distributions. The final objective function of the generator is defined as:

$$\mathcal{L}_G = \mathcal{L}_{G_0} + \lambda_1 \mathcal{L}_{I2T} + \lambda_2 \mathcal{L}_{trans}. \quad (15)$$

Accordingly, the discriminator D is trained by minimizing the following loss:

$$\begin{aligned}\mathcal{L}_D = &-\frac{1}{2}\mathbb{E}_{I\sim p_{data}}[\log(D(I)] \\ &-\frac{1}{2}\mathbb{E}_{f_1\sim p_G}[\log(1 - D(f_1)] \\ &-\frac{1}{2}\mathbb{E}_{I\sim p_{data}}[\log(D(I, s)] \\ &-\frac{1}{2}\mathbb{E}_{f_1\sim p_G}[\log(1 - D(f_1, s)].\end{aligned} \quad (16)$$

Where I is from the true image distribution p_{data}. The first two items are unconditional loss, the latter two are conditional loss.

Model	**En⇒De**				**En⇒Fr**			
	Test2016		**Test2017**		**Test2016**		**Test2017**	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Multimodal Neural Machine Translation Systems								
IMG$_D$ (Calixto and Liu, 2017)	37.3	55.1	N/A	N/A	N/A	N/A	N/A	N/A
NMT$_{SRC+IMG}$ (Calixto et al., 2017)	36.5	55.0	N/A	N/A	N/A	N/A	N/A	N/A
fusion-conv (Caglayan et al., 2017)	37.0	57.0	29.8	51.2	53.5	70.4	51.6	68.6
trg-mul (Caglayan et al., 2017)	37.8	**57.7**	30.7	52.2	54.7	71.3	52.7	69.5
VAG-NMT (Zhou et al., 2018)	N/A	N/A	31.6	52.2	N/A	N/A	**53.8**	**70.3**
Transformer+Att (Ive et al., 2019)	38.0	55.6	N/A	N/A	59.8	**74.4**	N/A	N/A
Multimodal (Yao and Wan, 2020)	**38.7**	55.7	N/A	N/A	N/A	N/A	N/A	N/A
ImagiT + ground truth	38.6	55.7	**32.4**	**52.5**	**59.9**	74.3	52.8	68.6
Text-only Neural Machine Translation Systems								
Transformer (Vaswani et al., 2017)	37.6	55.3	31.7	52.1	59.0	73.6	51.9	68.3
Multitask (Elliott and Kádár, 2017)	36.8	55.8	N/A	N/A	N/A	N/A	N/A	N/A
VMMT$_F$ (Calixto et al., 2019)	37.6	56.0	N/A	N/A	N/A	N/A	N/A	N/A
Lookup table (Zhang et al., 2020)	36.9	N/A	28.6	N/A	57.5	N/A	48.5	N/A
ImagiT	38.5	55.7	32.1	52.4	59.7	74.0	52.4	68.3

Table 1: Main result from the Test2016, Test2017 for the En⇒De and En⇒Fr MNMT task. The first category (Multimodal Neural Machine Translation Systems) 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, the baselines are the text-only Transformer (Vaswani et al., 2017) and the aforementioned works (Zhang et al., 2020; Elliott and Kádár, 2017).

4 Experiments

4.1 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) English-French (En-Fr) 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.

4.2 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, 2015) 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

Model	En⇒De		En⇒Fr	
	Ambiguous COCO		**Ambiguous COCO**	
	BLEU	METEOR	BLEU	METEOR
Multimodal Neural Machine Translation Systems				
fusion-conv (Caglayan et al., 2017)	25.1	46.0	43.2	63.1
trg-mul (Caglayan et al., 2017)	26.4	47.4	43.5	63.2
VAG-NMT (Zhou et al., 2018)	28.3	48.0	45.0	64.7
ImagiT + ground truth	**28.8**	**48.9**	**45.3**	**65.1**
Text-only Neural Machine Translation Systems				
Transformer baseline (Vaswani et al., 2017)	27.9	47.8	44.9	64.2
ImagiT	28.7	48.8	**45.3**	65.0

Table 2: Experimental results on the Ambiguous COCO En⇒De and En⇒Fr translation task.

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 size γ is set to 0.1, the balance weight $\lambda_1 = 20$, $\lambda_2 = 40$.

4.3 Results

Table 1 illustrates the results for the En-De Test2016, En-De Test2017, En-Fr Test2016 and En-Fr Test2017 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 improvements over the text-only Transformer baseline on four 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 four evaluation datasets. Our ImagiT is also competitive with ImagiT + ground truth, which is our translation decoder taking ground truth visual representations instead of imagined ones, and can be regarded as the upper boundary of imagiT. This proves imaginative ability of ImagiT.

Table 2 shows results for the En-De En-Fr Ambiguous COCO. 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. and even gets the same performance as ImagiT + ground truth (45.3 BLEU).

4.4 Ablation studies

The hyper-parameter λ_1 in equation 15 is important. When $\lambda_1 = 0$, there is no image captioning component, the BLEU score drops from 38.5 to 37.9, while this variant still outperforms the Transformer baseline. This indicates the effectiveness of image captioning module, since it will potentially prevent visual-textual mismatching, thus helps generator achieve better performance. When λ_1 increases from 5 to 20, the BLEU and METEOR increase accordingly. Whereas λ_1 is set to equal to λ_2, the BLEU score falls to 38.3. That's reasonable because $\lambda_2 \mathcal{L}_{trans}$ is the main task of the whole model.

Evaluation metric	BLEU	METEOR
ImagiT, $\lambda_1 = 0$	37.9	55.3
ImagiT, $\lambda_1 = 5$	38.2	55.5
ImagiT, $\lambda_1 = 10$	38.4	55.7
ImagiT, $\lambda_1 = 20$	38.5	55.7
ImagiT, $\lambda_1 = 40$	38.3	55.6

Table 3: Ablation studies of ImagiT with different weight settings

5 Analysis

5.1 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.

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

Table 4: Image retrieval task. We evaluate on Multi30K and MS COCO.

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 4, especially for $R@10$, the results demonstrate that our generated representation has excellent quality of shared semantics and have been grounded with visual semantic-consistency.

5.2 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 that 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 et al. (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.

Model	S	$\overline{S}$
text-only Transformer	37.6	36.3
MNMT	38.2	37.7
ImagiT	38.4	37.9

Table 5: Color deprivation. s represents the original source sentence, while $\overline{s}$ is the degraded sentence.

Table 5 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 5 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. One possible explanation is that ImagiT could learn the correlation and co-occurrence of the color and specific entities during the training phase, thus imagiT could infer the color from the context and recover it by visualization.

Visually depictable entity masking. Plummer et al. (2015) extend Flickr30K with cereference 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.

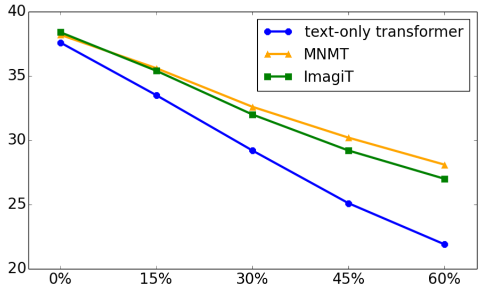


Figure 4: Visually depictable entity masking. From top to bottom is MNMT, ImagiT, text-only transformer.

5.3 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 $\mathcal{L}_{trans}$ term (Yang et al., 2020). 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:

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

Table 6: Translation results when using out-of-domain non-parallel image captioning data.

As is shown in Table 6, 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.

6 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.

7 Broader Impact

This work brings together text-to-image synthesis, image captioning, and neural machine translation (NMT) for an adversarial learning setup, advancing the traditional NMT to utilize visual information. For multimodal neural machine translation (MNMT), which possesses annotated images and can gain better performance, manual image annotation is costly, so MNMT is only applied on a small and specific dataset. This work tries to extend the applicability of MNMT techniques and visual information in NMT by imagining a semantic equivalent picture and making it appropriately utilized by visual-guided decoder. Compared to the previous multimodal machine translation approaches, this technique takes only sentences in the source languages as the usual machine translation task, making it an appealing method in low-resource scenarios. However, the goal is still far from being achieved, and more efforts from the community are needed for us to get there. One pitfall of our proposed model is that trained ImagiT is not applicable to larger-scale text-only NMT tasks, such as WMT'14, which is mainly related to economies and politics, since those texts are not easy to be visualized, containing fewer objects and visually depictable entities. We advise practitioners who apply visual information in large-scale text-to-text translation to be aware of this issue. In addition, the effectiveness of MNMT model largely depends on the quantity and quality of annotated images, likewise, our model performance also depends on the quality of generated visual representations. We will need to carefully study how the model balance the contribution of different modality and response to ambiguity and bias to avoid undesired behaviors of the learned models.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and S. Frank. 2018. Findings of the third shared task on multimodal machine translation. In *WMT*.

Ozan Caglayan, Walid Aransa, Adrien Bardet, M. García-Martínez, Fethi Bougares, Loic Barrault, Marc Masana, L. Herranz, and Joost van de Weijer. 2017. Lium-cvc submissions for wmt17 multimodal translation task. *ArXiv*, abs/1707.04481.

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

Ozan Caglayan, P. Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *NAACL-HLT*.

Iacer Calixto, Desmond Elliott, and S. Frank. 2016. Dcu-uva multimodal mt system report. In *WMT*.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *EMNLP*.

Iacer Calixto, Qun Liu, and N. Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. *ArXiv*, abs/1702.01287.

Iacer Calixto, Miguel Rios, and W. Aziz. 2019. Latent variable model for multi-modal translation. In *ACL*.

Michael J. Denkowski and A. Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *WMT@ACL*.

Desmond Elliott, S. Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. *ArXiv*, abs/1710.07177.

Desmond Elliott, S. Frank, K. Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *ArXiv*, abs/1605.00459.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *IJCNLP*.

Ian J. Goodfellow, Jean Pouget-Abadie, M. Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *NIPS*.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *WMT*.

J. Ive, P. Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *ACL*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. *ArXiv*, abs/1704.06567.

Tsung-Yi Lin, M. Maire, Serge J. Belongie, James Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. 2014. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312.

Hyeonwoo Noh, Seunghoon Hong, and B. Han. 2015. Learning deconvolution network for semantic segmentation. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1520–1528.

Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Bryan A. Plummer, Liwei Wang, C. Cervantes, Juan C. Caicedo, J. Hockenmaier, and S. Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *International Journal of Computer Vision*, 123:74–93.

Tingting Qiao, J. Zhang, Duanqing Xu, and D. Tao. 2019. Mirrorgan: Learning text-to-image generation by redescription. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1505–1514.

S. Reed, Zeynep Akata, Xinchen Yan, L. Logeswaran, B. Schiele, and H. Lee. 2016. Generative adversarial text to image synthesis. In *ICML*.

Lucia Specia, S. Frank, K. Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *WMT*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Oriol Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

T. Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and X. He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1316–1324.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9378–9385.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *ACL*.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, J. Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *ACL*.

P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. 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.

Han Zhang, Tao Xu, and Hongsheng Li. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5908–5916.

Zhuosheng Zhang, Kehai Chen, Rui Wang, M. Utiyama, Eiichiro Sumita, Z. Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *ICLR*.

Mingyang Zhou, Runxiang Cheng, Y. Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *EMNLP*.

Xiaojin Zhu, A. Goldberg, M. Eldawy, C. Dyer, and Bradley Strock. 2007. A text-to-picture synthesis system for augmenting communication. In *AAAI*.