Unsupervised Multi-modal Neural Machine Translation

Yuanhang Su*
University of Southern California
suyuanhang@hotmail.com

Kai Fan*†
Alibaba Group Inc. (US)
interfk@gmail.com

Nguyen Bach
Alibaba Group Inc. (US)
nguyenbh@gmail.com

C.-C. Jay Kuo
University of Southern California
cckuo@sipi.usc.edu

Fei Huang
Alibaba Group Inc. (US)
feirhuang@gmail.com

Abstract

Unsupervised neural machine translation (UNMT) has recently achieved remarkable results [20] with only large monolingual corpora in each language. However, the uncertainty of associating target with source sentences makes UNMT theoretically an ill-posed problem. This work investigates the possibility of utilizing images for disambiguation to improve the performance of UNMT. Our assumption is intuitively based on the invariant property of image, i.e., the description of the same visual content by different languages should be approximately similar. We propose an unsupervised multi-modal machine translation (UMNMT) framework based on the language translation cycle consistency loss conditional on the image, targeting to learn the bidirectional multi-modal translation simultaneously. Through an alternate training between multi-modal and uni-modal, our inference model can translate with or without the image. On the widely used Multi30K dataset, the experimental results of our approach are significantly better than those of the text-only UNMT on the 2016 test dataset.

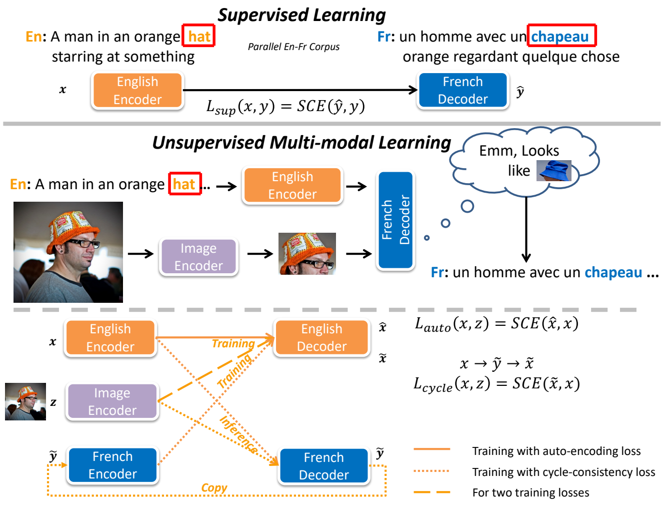
Supervised Learning
En: A man in an orange hat starring at something
Parallel En-Fr Corpus
Fr: un homme avec un chapeau orange regardant quelque chose
x
English Encoder
$L_{sup}(x,y) = SCE(\hat{y},y)$
French Decoder
$\hat{y}$
Unsupervised Multi-modal Learning
En: A man in an orange hat ..
English Encoder
Image Encoder
French Decoder
Emm, Looks like
Fr: un homme avec un chapeau ...
x
English Encoder
Training
English Decoder
$\hat{x}$
$\tilde{x}$
$L_{auto}(x,z) = SCE(\hat{x},x)$
$x \to \tilde{y} \to \tilde{x}$
$L_{cycle}(x,z) = SCE(\tilde{x},x)$
z
Image Encoder
Training
Inference
$\tilde{y}$
French Encoder
French Decoder
$\tilde{y}$
Copy
Training with auto-encoding loss
Training with cycle-consistency loss
For two training losses

Figure 1: Illustration of our proposed approach. We leverage the designed loss function to tackle a supervised task with the unsupervised dataset only. SCE means sequential cross-entropy.

1. Introduction

Our long-term goal is to build intelligent systems that can perceive their visual environment and understand the linguistic information, and further make an accurate translation inference to another language. Since image has become an important source for humans to learn and acquire knowledge (e.g. video lectures, [1, 18, 32]), the visual signal might be able to disambiguate certain semantics. One way to make image content easier and faster to be understood by humans is to combine it with narrative description that can be self-explainable. This is particularly important for many natural language processing (NLP) tasks as well, such as image caption [27] and some task-specific translation–sign language translation [6]. However, [24] demonstrates that most multi-modal translation algorithms are not significantly better than an off-the-shelf text-only machine translation (MT) model for the Multi30K dataset [12]. There remains an open question about how translation models should take advantage of visual context, because from the perspective of information theory, the mutual information of two random variables $I(X,Y)$ will always be no greater than $I(X;Y,Z)$, due to the following fact $I(X;Y,Z) - I(X;Y) = KL(p(X,Y,Z)\|p(X|Y)p(Z|Y)p(Y))$, where the Kullback-Leibler (KL) divergence is non-negative. This conclusion makes us believe that the visual content will hopefully help the translation systems.

[*] indicates equal contribution.
[†] corresponding author.

Since the standard paradigm of multi-modal translation always considers the problem as a supervised learning task, the parallel corpus is usually sufficient to train a good translation model, and the gain from the extra image input is very limited. Moreover, the scarcity of the well formed dataset including both images and the corresponding multilingual text descriptions is also another constraint to prevent the development of more scaled models. In order to address this issue, we propose to formulate the multi-modal translation problem as an unsupervised learning task, which is closer to real applications. This is particularly important given the massive amounts of paired image and text data being produced everyday (e.g., news title and its illustrating picture).

Our idea is originally inspired by the text-only unsupervised MT (UMT) [8, 19, 20], investigating whether it is possible to train a general MT system without any form of supervision. As [20] discussed, the text-only UMT is fundamentally an ill-posed problem, since there are potentially many ways to associate target with source sentences. Intuitively, since the visual content and language are closely related, the image can play the role of a pivot "language" to bridge the two languages without paralleled corpus, making the problem "more well-defined" by reducing the problem to supervised learning. However, unlike the text translation involving word generation (usually a discrete distribution), the task to generate a dense image from a sentence description itself is a challenging problem [21]. High quality image generation usually depends on a complicated or large scale neural network architecture [23, 13, 30]. Thus, it is not recommended to utilize the image dataset as a pivot "language" [7]. Motivated by the cycle-consistency [31], we tackle the unsupervised translation with a multi-modal framework which includes two sequence-to-sequence encoder-decoder models and one shared image feature extractor. We don't introduce the adversarial learning via a discriminator because of the non-differentiable $\arg\max$ operation during word generation. With five modules in our framework, there are multiple data streaming paths in the computation graph, inducing the auto-encoding loss and cycle-consistency loss, in order to achieve the unsupervised translation.

Another challenge of unsupervised multi-modal translation, and more broadly for general multi-modal translation tasks, is the need to develop a reasonable multi-source encoder-decoder model that is capable of handling multi-modal documents. Moreover, during training and inference stages, it is better to process the mixed data format including both uni-modal and multi-modal corpora.

First, this challenge highly depends on the attention mechanism across different domains. Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) are naturally suitable to encode the language text and visual image respectively; however, encoded features of RNN has autoregressive property which is different from the local dependency of CNN. The multi-head self-attention transformer [26] can mimic the convolution operation, and allow each head to use different linear transformations, where in turn different heads can learn different relationships. Unlike RNN, it reduces the length of the paths of states from the higher layer to all states in the lower layer to one, and thus facilitates more effective learning. For example, the BERT model [9], that is completely built upon self-attention, has achieved remarkable performance in 11 natural language tasks. Therefore, we employ transformer in both the text encoder and decoder of our model, and design a novel joint attention mechanism to simulate the relationships among the three domains. Besides, the mixed data format requires the desired attention to support the flexible data stream. In other words, the batch fetched at each iteration can be either uni-modal text data or multi-modal text-image paired data, allowing the model to be adaptive to various data during inference as well.

Succinctly, our contributions are three-fold: **(1)** We formuate the multi-modal MT problem as unsupervised setting that fits the real scenario better and propose an end-to-end transformer based multi-modal model. **(2)** We present two technical contributions: successfully train the proposed model with auto-encoding and cycle-consistency losses, and design a controllable attention module to deal with both uni-modal and multi-modal data. **(3)** We apply our approach to the Multilingual Multi30K dataset in English↔French and English↔German translation tasks, and the translation output and the attention visualization show the gain from the extra image is significant in the unsupervised setting.

2. Related Work

We place our work in context by arranging several prior popular topics, along the the axes of UMT, image caption and multi-modal MT.

Unsupervised Machine Translation Existing methods in this area [2, 19, 20] are mainly modifications of encoder-decoder schema. Their key ideas are to build a common latent space between the two languages (or domains) and to learn to translate by reconstructing in both domains. The difficulty in multi-modal translation is the involvement of another visual domain, which is quite different from the language domain. The interaction between image and text are usually not symmetric as two text domains. This is the reason why we take care of the attention module cautiously.

Image Caption Most standard image caption models are built on CNN-RNN based encoder-decoder framework [17, 27], where the visual features are extracted from CNN and then fed into RNN to output word sequences as captions. Since our corpora contain image-text paired data, our method also draws inspiration from image caption modeling. Thus, we also embed the image-caption model within our computational graph, whereas the transformer architecture is adopted as a substitution for RNN.

Multi-modal Machine Translation This problem is first proposed by [24] on the WMT16 shared task at the intersection of natural language processing and computer vision. It can be considered as building a multi-source encoder on top of either MT or image caption model, depending on the definition of extra source. Most Multi-modal MT research still focuses on the supervised setting like [5], while [7, 22], to our best knowledge, are the two pioneering works that consider generalizing the Multi-modal MT to an unsupervised setting. However, their setup puts restrictions on the input data format. For example, [7] requires the training data to be image text pair but the inference data is text-only input, and [22] requires image text pair format for both training and testing. These limit the model scale and generalization ability, since large amount of monolingual corpora is more available and less expensive. Thus, in our model, we specifically address this issue with controllable attention and alternative training scheme.

3. Methodology

In this section we first briefly describe the main MT systems that our method is built upon and then elaborate on our approach.

3.1. Neural Machine Translation

If a bilingual corpus is available, given a source sentence $\mathbf{x} = (x_1, ..., x_n)$ of n tokens, and a translated target sentence $\mathbf{y} = (y_1, ..., y_m)$ of m tokens, where $(\mathbf{x}, \mathbf{y}) \in \mathcal{X} \times \mathcal{Y}$, the NMT model aims at maximizing the likelihood,

$$p(\mathbf{y}|\mathbf{x}) = \sum_{t=1}^{m} p(y_t|\mathbf{y}_{<t}, \mathbf{x}). \tag{1}$$

The attention based sequence-to-sequence encoder-decoder architecture [3, 29, 14, 26] is usually employed to parameterize the above conditional probability.

The encoder reads the source sentence and outputs the hidden representation vectors for each token, $\{\mathbf{h}_1^e, ..., \mathbf{h}_n^e\} = \text{Enc}_x(\mathbf{x})$. The attention based decoder is defined in a recurrent way. Given the decoder has the summarized representation vector $\mathbf{h}_t^d = \text{Dec}_y(\mathbf{y}_{<t}, \mathbf{x})$ at time stamp t, the model produces a context vector $\mathbf{c}_t = \sum_{j=1}^{n} \alpha_i \mathbf{h}_j^e$ based on an alignment model, $\{\alpha_1, ..., \alpha_n\} = \text{Align}(\mathbf{h}_t^d, \{\mathbf{h}_1^e, ..., \mathbf{h}_n^e\})$, such that $\sum_{j=1}^{n} \alpha_j = 1$. Therefore, the conditional probability to predict the next token can be written as,

$$p(y_t|\mathbf{y}_{<t}, \mathbf{x}) = \text{softmax}(g(\mathbf{c}_t, y_{t-1}, \mathbf{h}_{t-1}^d)). \tag{2}$$

in which $g(\cdot)$ denotes a non-linear function extracting features to predict the target. The encoder and decoder model described here is in a general formulation, not constrained to be RNN [3] or transformer architecture [26].

3.2. Multi-modal Neural Machine Translation

In this task, an image $\mathbf{z}$ and the description of the image in two different languages form a triplet $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{I}$. The problem naturally becomes maximizing the new likelihood $p(\mathbf{y}|\mathbf{x}, \mathbf{z})$. Though the overall framework of such a translation task is still the encoder-decoder architecture, the detailed feature extractor and attention module can vary greatly, due to the extra source image. The traditional approach [24, 10] is to encode the source text and the image separately and combine them at the high level features, where the image feature map can be represented as $\{\mathbf{h}_1^i, ..., \mathbf{h}_k^i\} = \text{Enc}_z(\mathbf{z})$ and Enc_z is usually a truncated image classification model, such as Resnet [16]. Unlike the number of the text features which is exactly the number of tokens in the source, the number of the image features depends on the last layer in the truncated network. We propose to compute the context vector via an attention module,

$$\mathbf{c}_t = \text{Attention}(\mathbf{h}_t^d, \{\mathbf{h}_1^e, ..., \mathbf{h}_n^e\}, \{\mathbf{h}_1^i, ..., \mathbf{h}_k^i\}) \tag{3}$$

Since three sets of features appear in Eq (3), there are more options of the attention mechanism than text-only NMT. The decoder can remain the same in the recurrent fashion.

3.3. Unsupervised Learning

The unsupervised problem requires a new problem definition. On both the source and the target sides, only monolingual documents are presented in the training data, i.e., the data comes in the paired form of $(\mathbf{x}, \mathbf{z}) \in \mathcal{X} \times \mathcal{I}$ and $(\mathbf{y}, \mathbf{z}) \in \mathcal{Y} \times \mathcal{I}$. The triplet data format is no longer available. The purpose is to learn a multi-modal translation model $\mathcal{X} \times \mathcal{I} \rightarrow \mathcal{Y}$ or a text-only one $\mathcal{X} \rightarrow \mathcal{Y}$. Note there is no explicit paired information cross two languages, making it impossible to straightforwardly optimize the supervised likelihood. Fortunately, motivated by the CycleGAN [31] and the dual learning in [15], we can actually learn the translation model for both directions between the source and the target in an unsupervised way. Additionally, we can even make the multi-modal and uni-modal inference compatible with deliberate fine-tuning strategy.

3.4. Auto-Encoding Loss

As Figure 2 illustrates, there are five main modules in the overall architecture, two encoders and two decoders for the source and target languages, and one extra image encoder. Since the lack of triplet data, we can only build the first two following denoised auto-encoding losses without involving the paired $\mathbf{x}$ and $\mathbf{y}$,

$$\mathcal{L}_{\text{auto}}(\mathbf{x}, \mathbf{z}) = SCE(\text{Dec}_x(\text{Enc}_x(\mathbf{x}), \text{Enc}_z(\mathbf{z})), \mathbf{x}) \tag{4}$$

$$\mathcal{L}_{\text{auto}}(\mathbf{y}, \mathbf{z}) = SCE(\text{Dec}_y(\text{Enc}_y(\mathbf{y}), \text{Enc}_z(\mathbf{z})), \mathbf{x}) \tag{5}$$

where $SCE(\cdot, \cdot)$ represents sequential cross-entropy loss. We use "denoised" loss here, because the exact auto-encoding structure will likely force the language model

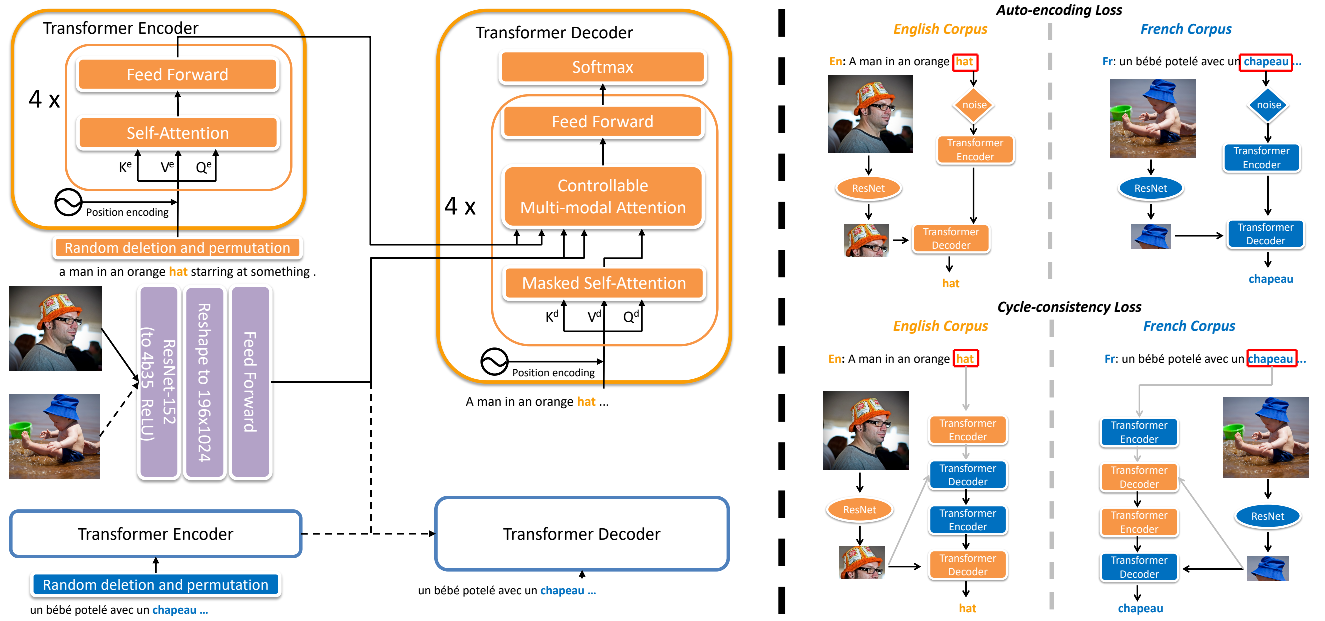

Figure 2: Model overview. Left Panel: The detailed unsupervised multi-modal neural machine translation model includes five modules, two transformer encoder, two transformer decoder and one ResNet encoder. Some detailed network structures within the transformer, like skip-connection and layer normalization, are omitted for clarity. Right Panel: The entire framework consists of four training paths: the gray arrows in the paths for cycle-consistency loss indicate the model is under inference mode. *E.g.*, the time step decoding for token "hat" is illustrated.

learning a word-to-word copy network. The image is seemingly redundant since the text input contains the entire information for recovery. However, it is not guaranteed that our encoder is lossless, so the image is provided as an additional supplement to reduce the information loss.

3.5. Cycle-Consistency Loss

The auto-encoding loss can, in theory, learn two functional mappings $\mathcal{X} \times \mathcal{I} \to \mathcal{X}$ and $\mathcal{Y} \times \mathcal{I} \to \mathcal{Y}$ via the supplied training dataset. However, the two mappings are essentially not our desiderata, even though we can switch the two decoders to build our expected mappings, e.g., $\mathcal{X} \times \mathcal{I} \to \mathcal{Y}$. The crucial problem is that the transferred mappings achieved after switching decoders lack supervised training, since no regularization pushes the latent encoding spaces aligned between the source and target.

We argue that this issue can be tackled by another two cycle-consistency properties (note that we use the square brackets [] below to denote the inference mode, meaning no gradient back-propagation through such operations),

$$\mathrm{Dec}_x(\mathrm{Enc}_y(\mathrm{Dec}_y[\mathrm{Enc}_x(\mathbf{x}), \mathrm{Enc}_z(\mathbf{z})]), \mathrm{Enc}_z(\mathbf{z})) \approx \mathbf{x} \quad (6)$$

$$\mathrm{Dec}_y(\mathrm{Enc}_x(\mathrm{Dec}_x[\mathrm{Enc}_y(\mathbf{y}), \mathrm{Enc}_z(\mathbf{z})]), \mathrm{Enc}_z(\mathbf{z})) \approx \mathbf{y} \quad (7)$$

The above two properties seem complicated, but we will decompose them step-by-step to see its intuition, which are also the key to make the auto-encoders translation models across different languages. Without loss of generality, we use Property (6) as our illustration, where the same idea is applied to (7). After encoding the information from source and image as the high level features, the encoded features are fed into the decoder of another language (i.e. target language), thus obtaining an inferred target sentence,

$$\tilde{\mathbf{y}} = F_{xz \to y}(\mathbf{x}, \mathbf{z}) \triangleq \mathrm{Dec}_y[\mathrm{Enc}_x(\mathbf{x}), \mathrm{Enc}_z(\mathbf{z})]. \quad (8)$$

Unfortunately, the ground truth $\mathbf{y}$ corresponding to the input $\mathbf{x}$ or $\mathbf{z}$ is unknown, so we cannot train $F_{xz \to y}$ at this time. However, since $\mathbf{x}$ is the golden reference, we can construct the pseudo supervised triplet $(\mathbf{x}, \tilde{\mathbf{y}}, \mathbf{z})$ as the augmented data to train the following model,

$$F_{yz \to x}(\tilde{\mathbf{y}}, \mathbf{z}) \triangleq \mathrm{Dec}_x(\mathrm{Enc}_y(\tilde{\mathbf{y}}), \mathrm{Enc}_z(\mathbf{z})). \quad (9)$$

Note that the pseudo input $\tilde{\mathbf{y}}$ can be considered as the corrupted version of the unknown $\mathbf{y}$. The noisy training step makes sense because injecting noise to the input data is a common trick to improve the robustness of model even for traditional supervised learning [25, 28]. Therefore, we incentivize this behavior using the cycle-consistency loss,

$$\mathcal{L}_{\text{cyc}}(\mathbf{x}, \mathbf{z}) = SCE(F_{yz \to x}(F_{xz \to y}(\mathbf{x}, \mathbf{z}), \mathbf{z}), \mathbf{x}). \quad (10)$$

This loss indicates the cycle-consistency (6), and the mapping $\mathcal{Y} \times \mathcal{I} \to \mathcal{X}$ can be successfully refined.

3.6. Controllable Attention

In additional to the loss function, another important interaction between the text and image domain should focus on the decoder attention module. In general, we proposal to extend the traditional encoder-decoder attention to a multi-domain attention.

$$\mathbf{c}_t = \text{Att}(\mathbf{h}_t^d, \mathbf{h}^e) + \lambda_1 \text{Att}(\mathbf{h}_t^d, \mathbf{h}^i) + \lambda_2 \text{Att}(\mathbf{h}_t^d, \mathbf{h}^e, \mathbf{h}^i) \tag{11}$$

where λ_1 and λ_2 can be either 1 or 0 during training, depending on whether the fetched batch includes image data or not. For example, we can easily set up a flexible training scheme by alternatively feeding the monolingual language data and text-image multimodal data to the model. A nice byproduct of this setup allows us to successfully make a versatile inference with or without image, being more applicable to real scenarios.

In practice, we utilize the recent developed self-attention mechanism [26] as our basic block, the hidden states contain three sets of vectors Q, K, V, representing queries, keys and values. Therefore, our proposed context vector can be rewritten as,

$$\begin{aligned}\mathbf{c}_t = & \text{softmax}\left(\frac{Q_t^d (K^e)^\top}{\sqrt{d}}\right) V^e + \lambda_1 \text{softmax}\left(\frac{Q_t^d (K^i)^\top}{\sqrt{d}}\right) V^i \\ & + \lambda_2 \text{softmax}\left(\frac{Q_t^d (K^{ei})^\top}{\sqrt{d}}\right) V^{ei} \\ & + \lambda_2 \text{softmax}\left(\frac{Q_t^d (K^{ie})^\top}{\sqrt{d}}\right) V^{ie} \end{aligned} \tag{12}$$

where d is the dimensionality of keys, and $[K^{ei}, V^{ei}] = \text{FFN}\left(\text{softmax}\left(\frac{Q^e (K^i)^\top}{\sqrt{d}}\right) V^i\right)$ means the attention from text input to image input, and $[K^{ie}, V^{ie}]$ represents the symmetric attention in the reverse direction. Note the notation Q^e has no subscript and denotes as a matrix, indicating the softmax is row-wise operation. In practice, especially for Multi30K dataset, we found λ_2 is less important and $\lambda_2 = 0$ brings no harm to the performance. Thus, we always set it as 0 in our experiments, but non-zero λ_2 may be helpful in other cases.

4. Experiments

4.1. Training and Testing on Multi30K

We evaluate our model on Multi30K [12] 2016 test set of English↔French (En↔Fr) and English↔German (En↔De) language pairs. This dataset is a multilingual image caption dataset with 29000 training samples of images and their annotations in English, German, French [11] and Czech [4]. The validation set and test set have 1014 and 1000 samples respectively. To ensure the model never sees any paired sentences information (which is an unlikely scenario in practice), we randomly split half of the training and validation sets for one language and use the complementary half for the other. The resulting corpora is denoted as *M30k-half* with 14500 and 507 training and validation samples respectively.

To find whether the image as additional information used in the training and/or testing stage can bring consistent performance improvement, we train our model in two different ways, each one has train with text only (-txt) and train with text+image (-txt-img) modes. We would compare the best performing training method to the state-of-the-art, and then do side-by-side comparison between them:

Pre-large (P): To leverage the controllable attention mechanism for exploring the linguistic information in the large monolingual corpora, we create text only pre-training set by combining the first 10 million sentences of the WMT News Crawl datasets from 2007 to 2017 with 10 times M30k-half. This ends up in a large text only dataset of 10145000 unparalleled sentences in each language. **P-txt**: We would then pre-train our model without the image encoder on this dataset and use the M30k-half validation set for validation. **P-txt-img**: Once the text-only model is pre-trained, we then use it for the following fine-tuning stage on M30k-half. Except for the image encoder, we initialize our model with the pre-trained model parameters. The image encoder uses pre-trained ResNet-152 [16]. The error gradient does not back-propagate to the original ResNet network.

Scratch (S): We are also curious about the role of image can play when no pre-training is involved. We train from scratch using text only (**S-txt**) and text with corresponding image (**S-txt-img**) on M30k-half.

4.2. Implementation Details and Baseline Models

The text encoder and decoder are both 4 layers transformers with dimensionality 512, and for the related language pair, we share the first 3 layers of transformer for both encoder and decoder. The image encoder is the truncated ResNet-152 with output layer res4b35_relu, and the parameters of ResNet are freezing during model optimization. Particularly, the feature map $14 \times 14 \times 1024$ of layer res4b35_relu is flattened to 196×1024 so that its dimension is consistent with the sequential text encoder output. The actual losses (4) and (5) favor a standard denoising auto-encoders: the text input is perturbed with deletion and local permutation; the image input is corrupted via dropout. We use the same word preprocessing techniques (Moses tokenization, BPE, binarization, fasttext word embedding on training corpora, etc.) as reported in [20], please refer to the relevant readings for further details.

We would like to compare the proposed UMNMT model to the following UMT models.

- MUSE [8]: It is an unsupervised word-to-word transla-

Models	En→ Fr	Fr→ En	En→ De	De→ En
MUSE	8.54	16.77	15.72	5.39
Game-NMT	-	-	16.6	19.6
UNMT-text	32.76	32.07	22.74	26.26
S-txt	6.01	6.75	6.27	6.81
S-txt-img	9.40	10.04	8.85	9.97
P-txt	37.20	38.51	20.97	25.00
P-txt-img	**39.79**	**40.53**	**23.52**	**26.39**

Table 1: BLEU benchmarking. The numbers of baseline models are extracted from the corresponding references.

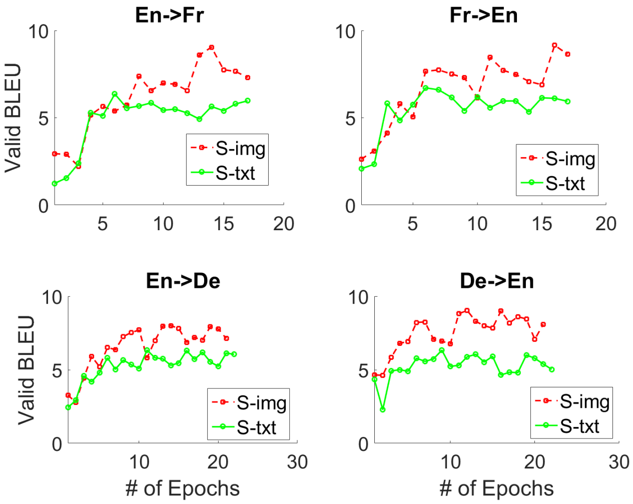

Figure 3: Validation BLEU comparison between text-only and text+image.

tion model. The embedding matrix is trained on large scale wiki corpora.

- Game-NMT [7]: It is a multi-modal zero-source UMT method trained using reinforcement learning.
- UNMT-text [19]: It is a mono-modal UMT model which only utilize text data and it is pretrained on synthetic paired data generated by MUSE.

4.3. Benchmarking with state-of-the-art

In this section, we report the widely used BLEU score of test dataset in Table 1 for different MT models. Our best model has achieved the state-of-the-art performance by leading more than 6 points in En→Fr task to the second best. Some translation examples are shown in Figure 4. There is also close to 1 point improvement in the En→De task. Although pre-training plays a significant role to the final performance, the image also contributes more than 3 points in case of training from scratch (S-txt vs. S-txt-img), and around 2 points in case of fine tuning (P-txt vs. P-txt-img). Interestingly, it is observed that the image contributes less performance improvement for pre-training than training from scratch. This suggests that there is certain information overlap between the large monolingual corpus and the M30k-half images. We also compare the Meteor, Rouge, CIDEr score in Table 2 and validation BLEU in Figure 3 to show the consistent improvement brought by using images.

GT	un homme avec un chapeau orange regardant quelque chose (a man in an orange hat starring at something)
P-txt	un homme en orange maettant quelque chose au loin (a man in orange putting something off)
P-txt-img	un homme en chapeau orange en train de filmer quelque chose (a man in an orange hat filming something)
GT	une femme en t-shirt bleu et short blanc jouant au tennis (a woman in a blue shirt and white shorts playing tennis)
P-txt	une femme en t-shirt bleu et short blanc jouant au tennis (a woman in blue t-shirt and white shorts playing tennis)
P-txt-img	une femme en t-shirt bleu et short blanc jouant au tennis (a woman in blue t-shirt and white shorts playing tennis)
GT	un chien brun ramasse une brindille sur un revêtement en pierre (a brown dog picks up a twig from stone surface)
P-txt	un chien marron retrouve un twig de pierre de la surface (a brown dog finds a twig of stone from the surface)
P-txt-img	un chien brun accède à la surface d' un étang (a brown dog reaches the surface of a pond)
GT	un garçon saisit sa jambe tandis il saute en air (a boy grabs his leg as he jumps in the air)
P-txt	un garçon se met à sa jambe devant lui (a boy puts his leg in front of him)
P-txt-img	un garçon installe sa jambe tandis il saute en air (a boy installs his leg while he jumps in the air)

Figure 4: Translation results from different models (GT: ground truth)

4.4. Analysis

In this section, we would shed more light on how and why images can help for unsupervised MT. We would first visualize which part of the input image helps the translation by showing the heat map of the transformer attention. We then show that image not only helps the translation by providing more information in the testing stage, it can also act as a training regularizer by guiding the model to converge to a better local optimal point in the training stage.

4.4.1 Attention

To visualize the transformer's attention from regions in the input image to each word in the translated sentences, we use the scaled dot-production attention of the transformer decoder's multi-head attention block as shown in Figure 2, more specifically, it is the softmax$\left(\frac{Q_t^d (K^i)^{\intercal}}{\sqrt{d}}\right)$. This is a matrix of shape $l_T \times l_S$, where l_T is the translated sentence length and l_S is the source length. Since we flatten the 14×14 matrix from the ResNet152, the $l_S = 196$. A heat map for the jth word in the translation is then generated by mapping the value of kth entry in $\{c_i[j, k]\}_{k=1}^{196}$ to their receptive field in the original image, averaging the value in the overlapping area and then low pass filtering. Given this heat map, we would visualize it in two ways: **(1)** We overlay the contour of the heat-map with the original image as shown

Models	En→Fr			Fr→En			En→De			De→En		
	Meteor	Rouge	CIDEr	Meteor	Rouge	CIDEr	Meteor	Rouge	CIDEr	Meteor	Rouge	CIDEr
S-txt	0.137	0.325	0.46	0.131	0.358	0.48	0.116	0.306	0.35	0.128	0.347	0.47
S-txt-img	**0.149**	**0.351**	**0.65**	**0.155**	**0.401**	**0.75**	**0.138**	**0.342**	**0.59**	**0.156**	**0.391**	**0.70**
P-txt	0.337	0.652	3.36	0.364	0.689	3.41	0.254	0.539	1.99	0.284	0.585	2.20
P-txt-img	**0.355**	**0.673**	**3.65**	**0.372**	**0.699**	**3.61**	**0.261**	**0.551**	**2.13**	**0.297**	**0.597**	**2.36**

Table 2: UMNMT shows consistent improvement over text-only model across normalized Meteor, Rouge and CIDEr metrics.

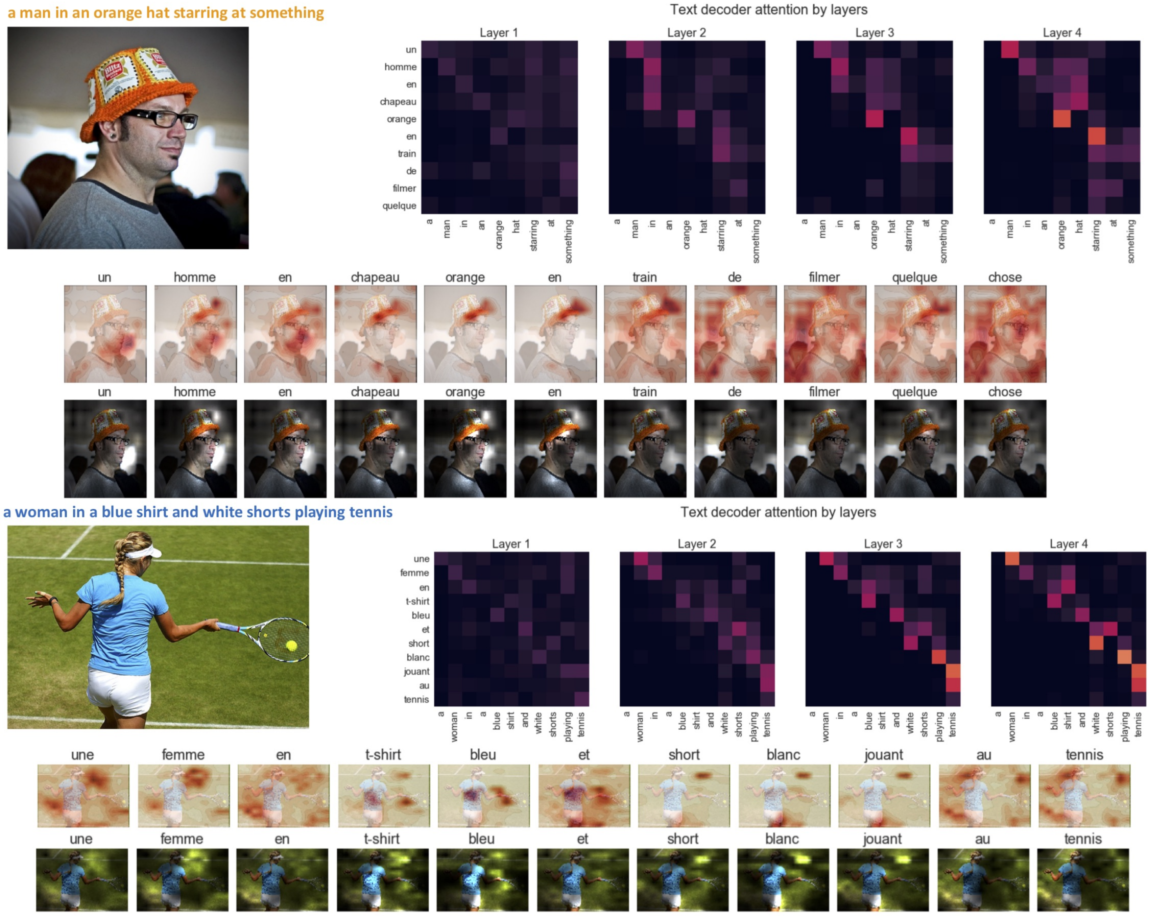

Figure 5: Correct attention for {"humme", "chapeau", "orange", "chose"} and {"bleu", "t-shirt", "blanc", "short"}.

in the second, and fifth rows of Figure 5 and the second row of Figure 6; (**2**) We normalize the heat map between 0 and 1, and then multiply it with each color channel of the input image pixel-wise as shown in the third and sixth rows of Figure 5 and in the third row of Figure 6.

We visualize the text attention by simply plotting the text attention matrix softmax$\left(\frac{Q_t^d (K^e)^{\mathrm{T}}}{\sqrt{d}}\right)$ in each transformer decoder layer as shown in "Text decoder attention by layers" in these two figures.

Figure 5 shows two positive examples that when transformer attends to the right regions of the image like "orange", "chapeau", or "humme" (interestingly, the nose) in the upper image or "bleu", "t-shirt", "blanc" or "short" in the lower image. Whereas in Figure 6, transformer attends to the whole image and treat it as a pond instead of focusing on the region where a twig exists. As a result, the twig was mistook as pond. More visualization results can see the detailed version[1]. For the text attention, we can see the text heat map becomes more and more diagonal as the decoder layer goes deeper in both figures. This indicates the text attention gets more and more focused since the English and French have similar grammatical rules.

[1] https://arxiv.org/pdf/1811.11365.pdf

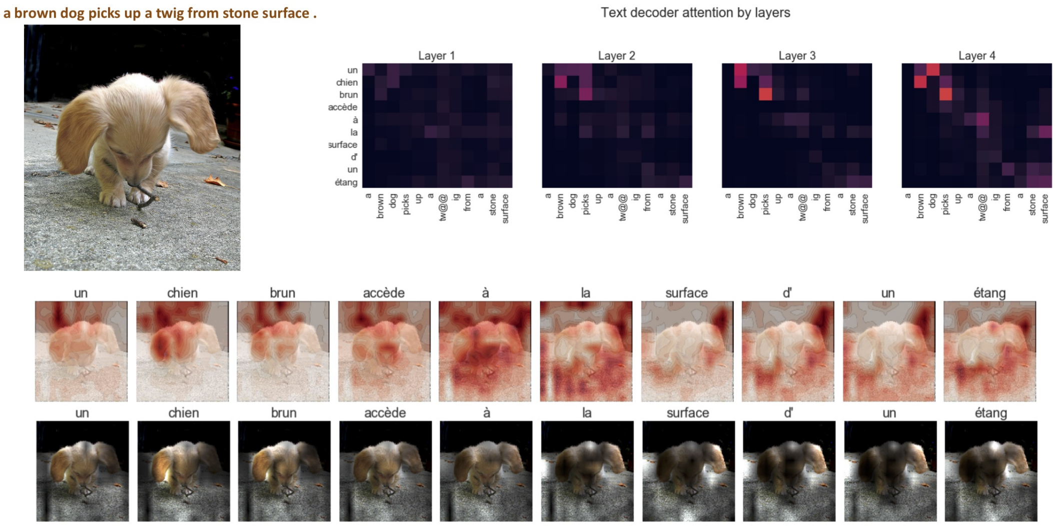

Figure 6: Correct attention for {“chien”, “brun”, “accède” and “surface”}, but missed “twig” for “étang”.

Models	En→ Fr	Fr→ En	En→ De	De→ En
S-txt	6.01	6.75	6.27	6.81
S-txt-img	**7.55**	**7.66**	**7.70**	**7.53**
P-txt	37.20	38.51	20.97	25.00
P-txt-img	**39.44**	**40.30**	**23.18**	**25.47**

Table 3: BLEU for testing with **TEXT ONLY** input

Models	En→ Fr	Fr→ En	En→ De	De→ En
S-txt	13.26 ↑	11.37 ↑	4.15 ↑	6.14 ↑
S-txt-img	**16.10** ↑	**13.30** ↑	**6.40** ↑	**7.91** ↑
P-txt	1.19 ↑	1.70 ↑	1.39 ↑	2.00 ↑
P-txt-img	**5.52** ↑	**2.46** ↑	**1.72** ↑	**3.12** ↑

Table 4: BLEU **INCREASE** (↑) UMNMT model trained on full Multi30k over UMNMT model trained on M30k-half (Table 1 Row 5-8).

4.4.2 Generalizability

As mentioned in the introduction, the model would certainly get more information when image is present in the inferencing stage, but can images be helpful if they are used in the training stage but not readily available during inferencing (which is a very likely scenario in practice)? Table 3 shows that even when images are not used, the performance degradation are not that significant (refer to Row 6-8 in Table 1 for comparison) and the trained with image model still outperforms the trained with text only model by quite a margin. This suggests that images can serve as additional information in the training process, thus guiding the model to converge to a better local optimal point. Such findings also verify the proposed controllable attention mechanism. This indicates the requirement of paired image and monolingual text in the testing stage can be relaxed to feeding the text-only data if paired image or images are not available.

4.4.3 Uncertainty Reduction

To show that images help MT by aligning different languages with similar meanings, we also train the UMNMT model on the whole Multi30K dataset where the source and target sentences are pretended unparalleled (i.e., still feed the image text pairs to model). By doing this, we greatly increase the sentences in different languages of similar meanings, if images can help align those sentences, then the model should be able to learn better than the model trained with text only. We can see from Table 4 that the performance increase by using images far outstrip the model trained on text only data, in the case of En→ Fr, the P-txt-img has more than 4 points gain than the P-txt.

5. Conclusion

In this work, we proposed a new unsupervised NMT model with multi-modal attention (one for text and one for image) which is trained under an auto-encoding and cycle-consistency paradigm. Our experiments showed that images as additional information can significantly and consistently improve the UMT performance. This justifies our hypothesis that the utilization of the multi-modal data can increase the mutual information between the source sentences and the translated target sentences. We have also showed that UMNMT model trained with images can still achieve better performance than trained with text-only model even if images are not available in the testing stage. Overall, our work pushes unsupervised machine translation more applicable to the real scenario.

References

[1] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4575–4583, 2016.

[2] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *International Conference on Learning Representations (ICLR)*, 2018.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[4] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, 2018.

[5] Iacer Calixto, Miguel Rios, and Wilker Aziz. Latent visual cues for neural machine translation. *arXiv preprint arXiv:1811.00357*, 2018.

[6] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793, 2018.

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

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

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[10] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. *arXiv preprint arXiv:1710.07177*, 2017.

[11] Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[12] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics, 2016.

[13] Kai Fan, Qi Wei, Wenlin Wang, Amit Chakraborty, and Katherine Heller. Inversenet: Solving inverse problems with splitting networks. *arXiv preprint arXiv:1712.00202*, 2017.

[14] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. Convolutional sequence to sequence learning. In *International Conference on Machine Learning*, pages 1243–1252, 2017.

[15] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, pages 820–828, 2016.

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

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[18] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

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

[20] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[21] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. In *International Conference on Learning Representations (ICLR)*, 2016.

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

[23] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48*, pages 1060–1069. JMLR. org, 2016.

[24] Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 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*, volume 2, pages 543–553, 2016.

[25] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[27] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[28] Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. Switchout: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, 2018.

[29] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[30] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018.

[31] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2017.

[32] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1004–1013, 2018.