PEIT: Bridging the Modality Gap with Pre-trained Models for End-to-End Image Translation

Shaolin Zhu[*], **Shangjie Li**[*], **Yikun Lei, Deyi xiong**[†]
College of Intelligence and Computing, Tianjin University, Tianjin, China
{zhushaolin, sj_li, yikunlei, dyxiong}@tju.edu.cn

Abstract

Image translation is a task that translates an image containing text in the source language to the target language. One major challenge with image translation is the modality gap between visual text inputs and textual inputs/outputs of machine translation (MT). In this paper, we propose PEIT, an end-to-end image translation framework that bridges the modality gap with pre-trained models. It is composed of four essential components: a visual encoder, a shared encoder-decoder backbone network, a vision-text representation aligner equipped with the shared encoder and a cross-modal regularizer stacked over the shared decoder. Both the aligner and regularizer aim at reducing the modality gap. To train PEIT, we employ a two-stage pre-training strategy with an auxiliary MT task: (1) pre-training the MT model on the MT training data to initialize the shared encoder-decoder backbone network; and (2) pre-training PEIT with the aligner and regularizer on a synthesized dataset with rendered images containing text from the MT training data. In order to facilitate the evaluation of PEIT and promote research on image translation, we create a large-scale image translation corpus ECOIT containing 480K image-translation pairs via crowd-sourcing and manual post-editing from real-world images in the e-commerce domain. Experiments on the curated ECOIT benchmark dataset demonstrate that PEIT substantially outperforms both cascaded image translation systems (OCR+MT) and previous strong end-to-end image translation model, with fewer parameters and faster decoding speed. Codes are available at https://github.com/lishangjie1/PEIT.

1 Introduction

Image translation (IT), transforming an image containing text in the source language to an image containing the target translation of the text (Mansimov et al., 2020; Jain et al., 2021), has recently attracted interest (Calixto et al., 2017a; Song et al., 2021). Traditional approaches to IT usually combine optical character recognition (OCR) with machine translation (MT) in a cascaded manner, e.g., Google Translate's Instant Camera[1] and Google Lens[2]. Such pipeline suffers from error propagation and high latency. To address this issue, end-to-end (E2E) image translation, analogous to E2E speech translation that directly translates speech in one language into speech/text in another, has been studied recently (Jain et al., 2021; Mansimov et al., 2020).

As a cross-modal task, a major challenge of IT is the representation discrepancy across the textual and visual modality. The text contained in an image is in its visual modality, unlike the text input for text-only machine translation. Its meaning also correlates with the visual context in the image. Such visual modality and text-vision correlation make it difficult for IT models to capture the meaning of the text in the context of the image and hence deteriorate translation quality.

Previous efforts to E2E IT, e.g., the method presented in (Jain et al., 2021), use ResNet as the visual encoder to encode the latent semantic representations of images, and a pre-trained text-only decoder to generate target translations. Such framework may not be able to sufficiently leverage cross-modal knowledge as it only uses convolutional neural networks (CNN) to model both image and visual text contained in the image and do not explicitly deal with the modality gap issue.

To mitigate this problem, we propose PEIT that bridges the modality gap with pre-trained models for end-to-end IT. The PEIT is composed of four

[*]Equal contribution.
[†]Corresponding author.

[1]https://blog:google/products/translate/googletranslates-instant-camera-translation-gets-upgrade/
[2]https://ai:googleblog:com/2019/09/giving-lens-new-readingcapabilities-in:html

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 13433–13447

essential components: a visual encoder, a shared encoder-decoder backbone network, a vision-text representation aligner and a cross-modal regularizer. We use a two-stage pre-training strategy to pre-train PEIT. In the first pre-training stage, we pre-train an NMT model on a huge amount of MT training data, which is used to initialize the shared encoder-decoder network, transfer knowledge to E2E IT and unify cross-modal representations. Following previous E2E IT practice (Jain et al., 2021; Mansimov et al., 2020), we also pre-train the shared encoder-decoder backbone network on a synthesized dataset with rendered images containing sentences from the MT training data after the network has been initialized by the pre-trained MT model. During the second pre-training stage, the aligner equipped with the shared encoder is jointly trained to align vision-text input representations in the same semantic space via contrastive learning. The regularizer stacked over the shared decoder is also optimized to force the decoder to generate the same translation for the same input in different modalities.

To the best of our knowledge, there is no public dataset available for IT task. We hence curate a large-scale image translation dataset in e-commerce domain, ECOIT, containing product images automatically crawled from a Chinese e-commerce website[3] paired with post-edited target translations (480K sentences with 3.64M source tokens). We fine-tune PEIT on the constructed ECOIT to perform the IT task.

The main contributions of this work are summarized as follows:

- We build the first large-scale benchmark dataset ECOIT to facilitate the training and evaluation of E2E image translation. The dataset will be released soon.
- We propose PEIT that bridges the vision-text modality gap and transfers knowledge from the MT task to E2E IT as MT has a huge amount of training data.
- To well align visual and textual representations in the unified semantic space so as to bridge the modality gap, we propose a vision-text representation aligner equipped with the shared encoder and a cross-modal regularizer stacked over the shared decoder.
- Experiments on the ECOIT dataset show that our model achieves the state-of-the-art results compared to previous strong E2E IT models and cascaded IT systems and demonstrate the robustness of the proposed model in real-world image translation scenarios.

[3]https://www.taobao.com/.

2 Related Work

Recent years have witnessed increasing attention on multimodal machine translation (MMT) that translates a source sentence into the target language accompanied with an additional modality (Sulubacak et al., 2020). Given the additional modality and its relation to the source sentence, MMT can be roughly divided into image-guided translation (Calixto et al., 2017b; Song et al., 2021), video-guided translation (Wang et al., 2019), speech translation (Han et al., 2021; Fang et al., 2022), IT (Jain et al., 2021). Image-guided MMT aims to leverage visual context to aid textual machine translation (Yang et al., 2020; Li et al., 2022a). The significant difference between image-guided translation and image translation is that the latter embeds the source sentence in its visual modality in the image while the former has the image and the source sentence separated and the image is used to provide additional information for translating the source sentence.

In contrast to image-guided translation, IT has not yet been fully explored in the literature probably due the lack of publicly available datasets for IT. Both Jain et al. (2021) and Mansimov et al. (2020) propose end-to-end approaches to it. Jain et al. (2021) uses a convolutional encoder to encode the image and Transformer decoder to generate target translation. The end-to-end IT model is able to locate characters in image, performs implicit tokenization on the source text, and then extracts latent semantic representations from them. This model can extract the latent token representations of image and text, and map into a shared space to implement the E2E IT. While they provide an initial definition of the IT task, they neither consider the modality gap nor verify the effect of the proposed models on real-world images.

For speech translation (ST), recent efforts have shifted towards end-to-end speech-to-text translation that directly translates a speech in the source language into a text in the target language (Babu et al., 2022; Ao et al., 2022). This is because end-to-end ST is of less error propagation and low latency compared with traditional cascaded ST (Inaguma

ECOIT	#Sentences	#Images	#Source Tokens(characters)	#Target Tokens
Training	477,490	477,490	3,626,371	2,338,888
Validation	2,000	2,000	12,875	8,316
Test	1,020	1,020	7,534	5,006

Table 1: Data statistics of ECOIT

et al., 2021; Fang et al., 2022). However, E2E ST suffers from the high cost of speech-to-text parallel data creation. Pre-training and multitask learning strategies have been explored to mitigate this data scarcity issue (Dong et al., 2021; Yang et al., 2022). In addition, similar to E2E IT, E2E ST is also confronted with the cross-modality issue, which can be mitigated by sharing the same semantic space for audio and text representations (Han et al., 2021). Partially motivated by E2E ST, we propose an end-to-end framework for IT from the perspectives of pre-training with data of the MT task, sharing parameters across modalities, knowledge transfer via multitask learning, attempting to address the data scarcity and modality gap issues in IT.

3 Large-Scale Parallel Image Translation Dataset: ECOIT

In order to facilitate the training and evaluation of E2E IT and hence promote its research, we build a large-scale E-COmmerce parallel IT dataset, ECOIT, based on the Taobao[4] e-commerce platform. The reason for building the dataset in the e-commerce domain is that product descriptions and advertising slogans are often contained in the images of products to attract shoppers and promote sales. In other words, e-commerce provides a huge amount of images containing text from a wide range of domains, which much fits into the motivation of IT. To build this dataset, we first crawl $\sim$ 600,000 images that contain Chinese texts. We then use an OCR detector[5],with a high accuracy of 90%, to automatically recognize the Chinese texts in images. Recognized texts are manually scrutinized: those with over 3 incorrectly recognized Chinese characters are removed while those with less than 3 wrong characters are manually corrected. After this manual scrutinization, we have 479,490 image-sentence pairs. We automatically translate these Chinese texts into English with Google translate API. To guarantee translation quality, we hire crowd-sourced workers who are Chinese-English bilingual speakers to manually post-edit English translations to ensure both fluency and adequacy. More than 80% of automatic English translations have been post-edited. 2,000 image-translation pairs are selected as the validation set while 1,020 pairs are selected as the test set. Table 1 displays the statistics of the dataset. The entire dataset will be released soon.

[4]https://www.taobao.com/

[5]https://github.com/JaidedAI/EasyOCR

4 Methodology

This section starts with the task formulation of IT, followed by an overview of the model architecture of PEIT and the two-stage pre-training strategy that leverages MT knowledge from both the encoder and decoder to reduce the modality gap.

4.1 Task Formulation

Similar to corpora for E2E ST, e.g., MUST-C (Cattoni et al., 2021), the created ECOIT is composed of triplets, each of which consists of an image containing text, the text extracted from the image, the target translation of the text. The corpus can be denoted as $D = \{(\mathbf{v}, \mathbf{x}, \mathbf{y})\}$ where $\mathbf{v}$ denotes the image, $\mathbf{x} = \{\mathrm{x}_1, ..., \mathrm{x}_\mathrm{N}\}$ is the text contained in the image, and $\mathbf{y} = \{\mathrm{y}_1, ..., \mathrm{y}_\mathrm{M}\}$ is the translation in the target language. N and M are the length of the source and target text, respectively. The goal of E2E IT is to find the best $\mathbf{y}$ given the input image:

$$\mathcal{L}_{\mathrm{IT}} = -\sum_{t=1}^{\mathrm{M}} \log \mathrm{p}(\mathrm{y}_t|\mathrm{y}_{<t}, \mathbf{v}; \theta) \quad (1)$$

4.2 Model Architecture

The model architecture of PEIT is illustrated in Figure 1. It consists of four essential modules: a visual encoder, a shared encoder-decoder backbone network, a vision-text representation aligner via contrastive learning and a cross-modal regularizer. The aligner is equipped with the shared encoder, which attempts to unify the representations of the same input with different modalities (vision vs. text) in the same semantic space. The reguarizer is deployed at the shared decoder, which forces the decoder to yield the same translation for the same input in different modalities. The visual encoder encodes the input image $\mathbf{v}$ to its semantic representation $\mathbf{V}$, which is then fed into the shared encoder-decoder backbone network (a standard Transformer) for translation.

In order to obtain the semantic representation of the text contained in an image, we adopt two strong visual encoder architectures: ResNet (He et al., 2016) and CRNN (Shi et al., 2017). In order

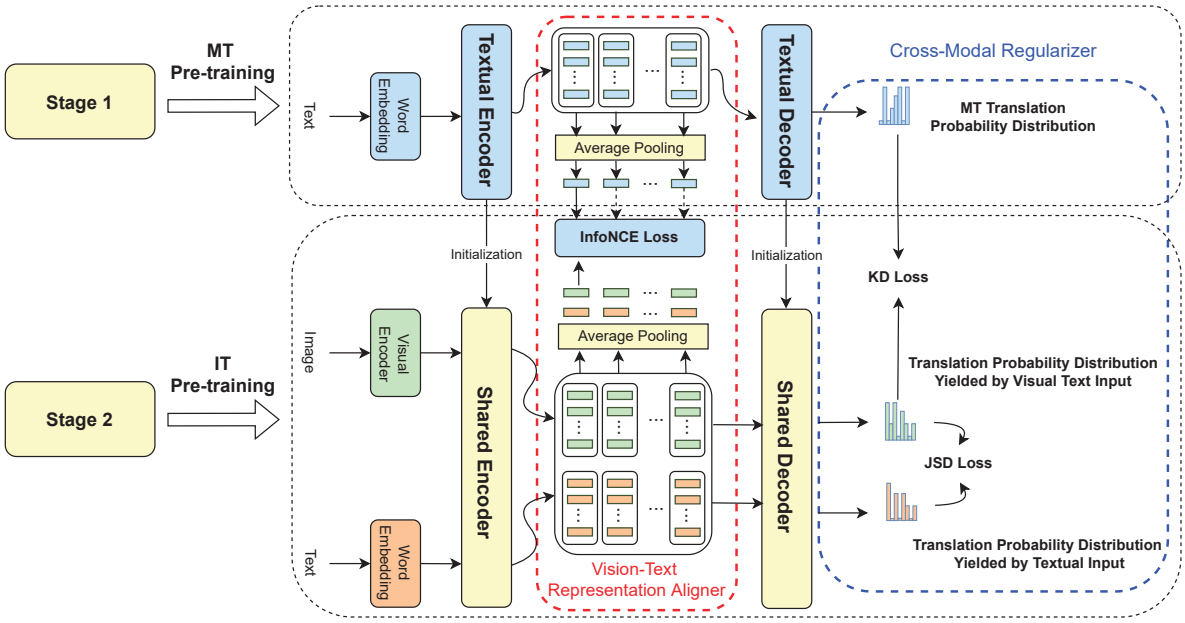


Figure 1: The diagram of the proposed PEIT with the visual encoder, shared encoder-decoder backbone network, vision-text representation aligner and cross-modal regularizer. The text-only MT model is pre-trained in the first stage to initialize PEIT that is pre-trained in the second stage together with the aligner and regularizer.

to match the length of encoded image features with that of the corresponding text, following (Ye et al., 2021), we stack two additional layers of 2-stride 1-dimensional convolutional layers with the GELU activation on the top of the visual encoder, which reduces the time dimension by a factor of 4. Given an input image $\mathbf{v}$, we can get its feature vectors $\mathbf{V} = \{\mathbf{V}_1, ..., \mathbf{V}_K\}$ by :

$$\mathbf{V} = \mathrm{E}_{\mathrm{img}}(\mathbf{v}) \in \mathbb{R}^{\mathrm{K}\times\mathrm{d}} \tag{2}$$

where K denotes the number of embedded feature vectors, and d is the dimension of feature vectors. $\mathrm{E}_{\mathrm{img}}$ denotes the visual encoder. For an input text sentence, we use an embedding layer to transform $\mathbf{x}$ into vectors $\mathbf{X} = \{\mathbf{X}_1, ..., \mathbf{X}_N\}$ by $\mathbf{X} = \mathrm{E}_{\mathrm{txt}}(\mathbf{x}) \in \mathbb{R}^{\mathrm{N}\times\mathrm{d}}$.

4.3 Two-Stage Pre-Training

Due to the lack of IT training data, we take a two-stage pre-training strategy to transfer knowledge from the auxiliary MT task to E2E IT with a huge amount of MT training data and synthesized data with rendered images. In pre-training stage 1, we pre-train a vanilla Transformer MT model on a large-scale textual parallel corpus. The pre-trained MT model is used to initialize the shared encoder-decoder backbone network and to train the aligner and regularizer for modality unification. Let $\mathbf{h}^{\mathrm{t}}_{\mathrm{enc}}$ be the output of the pre-trained MT encoder and $\mathbf{h}^{\mathrm{t}}_{\mathrm{dec}}$ be the output of the MT decoder.

In pre-training stage 2, we pre-train the shared encoder-decoder backbone network on the synthesized data (see Section 5.1) created from the MT training data with alternating visual and textual inputs ($\mathbf{v}$ and $\mathbf{x}$) . For textual pre-training, the shared encoder takes the representation $\mathbf{X}$ of $\mathbf{x}$ as input to generate $\mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{X})$. The shared decoder is optimized to generate the corresponding translation $\mathbf{y}$ with the maximum likelihood estimation as follows:

$$\mathcal{L}_t = -\sum_{\mathrm{n}=1}^{\mathrm{M}} \log \mathrm{p}(\mathrm{y}_\mathrm{n}|\mathrm{y}_{<\mathrm{n}}, \mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{X})) \tag{3}$$

where M is the length of $\mathbf{y}$.

For visual pre-training, the shared encoder takes the representation $\mathbf{V}$ of $\mathbf{v}$ as input to generate $\mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{V})$. The shared decoder is optimized to generate the corresponding translation $\mathbf{y}$ with the maximum likelihood estimation as follows:

$$\mathcal{L}_v = -\sum_{\mathrm{n}=1}^{\mathrm{M}} \log \mathrm{p}(\mathrm{y}_\mathrm{n}|\mathrm{y}_{<\mathrm{n}}, \mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{V})) \tag{4}$$

4.4 Vision-Text Representation Aligner

As shown by Eq. (3) and Eq. (4), the shared decoder is supposed to yield the same translation for the same input in different modalities (i.e., $\mathbf{v}$ and $\mathbf{x}$). However, the actual results are not as expected (see Section 5.4). The main reasons are that (1) The

position embeddings of words in the textual input is of great importance to translation (Vaswani et al., 2017), while the representation of an image cannot provide effective position information of words in the text contained in the image; (2) There is no effective mechanism to align the cross-modal representations to capture the modality-invariant information in the shared encoder. Due to the two issues, it is difficult for the shared encoder-decoder backbone network to capture and convey the underlying semantic information of the text contained in an image into the target language by alternatively optimizing $\mathcal{L}_v$ and $\mathcal{L}_t$. Partially inspired by the application of two pre-training configurations (Tang et al., 2022) and contrastive learning (Li et al., 2022b) in natural language processing, we propose the vision-text representation aligner (see the red box in Figure 1) to unify visual and textual modality into the shared semantic space of the encoder. In detail, we use the MT encoder pre-trained in stage 1 to guide the training of the shared encoder via contrastive learning. We analyse the reason why using the pre-trained MT encoderin Appendix 5.7. Specifically, we simultaneously input image representations $\mathbf{V}$ and word embeddings $\mathbf{X}$ into the shared encoder and the pre-trained MT encoder, respectively. $\mathbf{V}$ and $\mathbf{X}$ are from a mini-batch $\{\mathbf{v}_i, \mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{\mathrm{N_b}} \in \mathfrak{B}$, where $\mathrm{N_b}$ is the size of the mini-batch $\mathfrak{B}$. After performing average-pooling on the output hidden state sequence of the shared encoder and pre-trained MT encoder, we can obtain sentence-level representations, $\{\mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{V}_i), \mathbf{h}^{\mathrm{t}}_{\mathrm{enc}}(\mathbf{X}_i)\}$ of $\{\mathbf{v}_i, \mathbf{x}_i\}$ from the mini-batch, which forms a positive pair, while other samples $\{\mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{V}_i), \mathbf{h}^{\mathrm{t}}_{\mathrm{enc}}(\mathbf{X}_j)\}, i \neq j$ from the same mini-batch, are treated as negative pairs. The contrastive loss is computed as follows:

$$\mathcal{L}^{v}_{\mathrm{enc}} = -\log \frac{\exp(s(\mathbf{v}_i, \mathbf{x}_i)/\tau)}{\sum_{\mathbf{x}_j \in \mathfrak{B}} \exp(s(\mathbf{v}_i, \mathbf{x}_j)/\tau)} \quad (5)$$

$$s(\mathbf{v}_i, \mathbf{x}_j) = \frac{\mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{V}_i)\mathbf{h}^{\mathrm{t}}_{\mathrm{enc}}(\mathbf{X}_j)^{\top}}{\|\mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{V}_i)\|_2 \left\|\mathbf{h}^{\mathrm{t}}_{\mathrm{enc}}(\mathbf{X}_j)\right\|_2} \quad (6)$$

$s(.)$ is a similarity function, $\|.\|_2$ is the L2 regularization as defined in (Li et al., 2022b), τ is a temperature hyperparameter. $\mathcal{L}^{v}_{\mathrm{enc}}$ is a InfoNCE loss function, which only leverages text data to model visual information.

To further align visual and textual representations in the shared encoder, we then simultaneously input a text $\mathbf{X}$ into the pre-trained MT encoder and the shared encoder. The pre-trained MT encoder is used to guide the training of the shared encoder, similarly via contrastive learning as follows:

$$\mathcal{L}^{t}_{\mathrm{enc}} = -\log \frac{\exp(s(\mathbf{x}_i, \mathbf{x}_i)/\tau)}{\sum_{\mathrm{x}_j \in \mathfrak{B}} \exp(s(\mathbf{x}_i, \mathbf{x}_j)/\tau)} \quad (7)$$

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{X}_i)\mathbf{h}^{\mathrm{t}}_{\mathrm{enc}}(\mathbf{X}_j)^{\top}}{\|\mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{X}_i)\|_2 \left\|\mathbf{h}^{\mathrm{t}}_{\mathrm{enc}}(\mathbf{X}_j)\right\|_2} \quad (8)$$

Obviously, when $\mathcal{L}^{t}_{\mathrm{enc}}$ and $\mathcal{L}^{v}_{\mathrm{enc}}$ are trained simultaneously, the output $\mathbf{h}^{\mathrm{t}}_{\mathrm{enc}}(\mathbf{X})$ of the pre-trained MT encoder acts as a pivot that links $\mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{X})$ and $\mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{V})$. Therefore, the visual ($\mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{V})$) and textual ($\mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{X})$) representations can be aligned by simultaneously optimizing $\mathcal{L}^{t}_{\mathrm{enc}}$ and $\mathcal{L}^{v}_{\mathrm{enc}}$.

4.5 Cross-Modal Regularizer

In addition to the aligner equipped with the shared encoder, we introduce the cross-modal regularizer (see the blue box in Figure 1), to further transfer knowledge from the auxiliary MT task to E2E IT and to reduce the modality gap at the decoder side. To transfer knowledge from the pre-trained MT decoder to the shared decoder, we employ the knowledge distillation (KD) method presented in (Liu et al., 2019) and define the KD loss $\mathcal{L}_{\mathrm{KD}}$ as follows:

$$\mathcal{L}_{\mathrm{KD}} = -\sum_{n=1}^{\mathrm{M}} \sum_{k=1}^{|\mathrm{C}|} \log \mathrm{p}(\mathrm{y}_n = k|\mathrm{y}_{<n}, \mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{V})) \times \mathrm{p}(\mathrm{y}_n = k|\mathrm{y}_{<n}, \mathbf{h}^{\mathrm{t}}_{\mathrm{enc}}(\mathbf{X})) \quad (9)$$

|C| is the vocabulary size of the output target text.

As mentioned before, we alternatively feed visual inputs ($\mathbf{V}$) and text inputs ($\mathbf{X}$) into the shared encoder-decoder backbone network. Since $\mathbf{V}$ and $\mathbf{X}$ are actually the same text in different modalities, they are supposed to be translated into the same target translation. In order to achieve this goal, we regularize the output predictions for the visual and textual input by minimizing the Jensen-Shannon Divergence (JSD) between the two output distributions as follows:

$$\mathcal{L}_{\mathrm{JSD}} = \sum_{n=1}^{\mathrm{M}} \mathrm{JSD}\{\mathrm{p}(\mathrm{y}_n|\mathrm{y}_{<n}, \mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{V}))|| \mathrm{p}(\mathrm{y}_n|\mathrm{y}_{<n}, \mathbf{h}^{\mathrm{s}}_{\mathrm{enc}}(\mathbf{X}))\} \quad (10)$$

Due to the lack of sufficient image translation training data, we use a multi-stage training strategy to transfer knowledge from other tasks (e.g., MT) that have a large amount of training data. And also due to the multimodality gap between vision and text, we propose to use multiple losses to attempt to reduce it. As the vision-text aligner and cross-modal regularizer are jointly trained with the shared encoder-decoder backbone network in pre-training stage 2 on the synthesized data, the pre-training loss in stage 2 can be formulated as follows:

$$\mathcal{L} = \mathcal{L}_t + \mathcal{L}_v + \mathcal{L}^v_{\text{enc}} + \mathcal{L}^t_{\text{enc}} + \mathcal{L}_{\text{KD}} + \mathcal{L}_{\text{JSD}} \quad (11)$$

4.6 Fine-Tuning

After the two-stage pre-training, we continue to fine-tune our PEIT on the curated image translation dataset ECOIT so as to endow PEIT with the ability to translate real-world images containing text into the target language. For fine-tuning and inference, we only keep the visual encoder and the shared encoder-decoder backbone network, removing the pre-trained MT module together with the vision-text aligner and cross-modal regularizer. The kept components are fine-tuned with the cross-entropy loss $\mathcal{L}_v$ on ECOIT.

5 Experiments

We conducted extensive experiments with a large MT training corpus, a synthesized dataset with rendered images based on the MT training corpus and the curated image translation data to examine the effectiveness of the proposed PEIT against previous end-to-end and cascaded baselines.

5.1 Dataset

For pre-training the MT task in stage 1, we extracted a subset from the United Nations Parallel Corpus[6] as our Chinese-English MT training dataset, which contains 15M parallel sentences. For pre-training PEIT components in stage 2, we extracted sentences whose length is less than 20 words from the Chinese-English MT training data used in stage 1, producing a Chinese-English corpus C with 3M parallel sentences. We then synthesized an image translation corpus that consists 10M pairs of rendered images (with different backgrounds, font sizes, font styles, etc.) containing sentences from C. To make synthesized images lifelike, we used a set of backgrounds randomly extracted from the ECOIT dataset and the font sizes are ranging from 30 pixels to 60 pixels. We fixed the size of image at 64x600 resolution. In order to synthesize images for pre-training, we randomly extract a sentence from a text corpus and randomly select the font style/font size/color for the sentence. This allows us to know the region size (height and width) required to put this sentence in the image. We then randomly extract an image from ECOIT as the background image and try to find a suitable image sub-area for writing the sentence without overlapping the existing text (i.e., product descriptions) in the image. After writing, we cut the writing area as a synthesized image by matrix slicing. In doing so, we have real-world background images, which allows us to train a strong encoder to extract semantic representations of texts embedded in real-world backgrounds. For fine-tuning (see Section 4.6), we used the curated ECOIT dataset. The development and test sets of ECOIT were used to evaluate our fine-tuned model and baselines.

5.2 Settings and Baselines

Model Configuration The shared encoder-decoder backbone network contains 6 Transformer encoder blocks and 6 Transformer decoder blocks, where the model dimension is 256, and the number of attention heads is 8. In the pre-training stages, we used polynomial decay learning rate schedule with a learning rate of 1e-4. We trained models with at most 33K input tokens per batch for 100K steps. During fine-tuning, the learning rate was set to 3e-5, and the maximum number of training step was 30K. We early-stopped fine-tuning if the loss on the dev set did not decrease for ten epochs. For both pre-training and fine-tuning, we used an Adam optimizer with $\beta 1 = 0.9$, $\beta 2 = 0.98$. The vaule of temperature hyperparameter τ was set as 0.1. More detailed experimental settings are in Appendix A.1

Baselines We compared our method against two strong image translation systems:

- Cascaded System: This system first uses a text detector to extract the text from an image. The extracted text is then fed into the pre-trained MT model to yield the target translation. We tried three different text recognition models (CRNN (VGG+BiLSTM) from easyocr[7]; DenseNet (Huang et al., 2017) from cnocr[8]; PP-OCR (Du et al., 2020)) in the cascaded system.

[6] https://conferences.unite.un.org/UNCorpus/

[7] https://github.com/JaidedAI/EasyOCR

[8] https://github.com/breezedeus/cnocr

model	#param	speed (tokens/s)	Pre-training		Fine-tuning	
			BLEU	METEOR	BLEU	METEOR
text-only NMT	28.8M	3,580	20.3	43.5	50.3	71.9
Cascaded						
Cascade (CRNN)	-	936	16.6	38.6	41.9	64.6
Cascade (DenseNet)	-	920	17.5	38.3	44.1	64.1
Cascade (PP-OCR)	-	943	18.4	39.2	45.0	66.2
End-to-End						
ItNet	60.6M	2,143	9.6	27.2	39.3	61.1
PEIT (ResNet)	71.6M	2,031	13.9	30.3	46.1	68.3
PEIT (CRNN)	**33.2M**	**3,383**	13.7	30.1	**47.2**	**69.2**

Table 2: Results of different image translation models on the ECOIT test set.

- ItNet (Jain et al., 2021): This is an end-to-end image translation system. It first pre-trains a standard Transformer on a text-only parallel dataset. ResNet is used as the image encoder to encode the latent semantic representations of images. The combination of the pre-trained decoder and image encoder is then fine-tuned on a synthetic dataset. We reimplemented this model and pre-trained & fine-tuned it on our datasets.

5.3 Main Results

For evaluating translation performance, we used two automatic evaluation metrics sacreBLEU[9] and METEOR[10] (Denkowski and Lavie, 2014).

Comparison with End-to-End Baselines In order to examine the effectiveness of our proposed pre-training method, we evaluated both pre-trained models (pre-trained on the MT/synthesized data) and fine-tuned models (fine-tuned on the training data of ECOIT after being pre-trained) on the ECOIT test set. As shown in Table 2, while our reimplemented ItNet is a strong end-to-end image translation baseline, our best model PEIT (CRNN) achieves a substantial improvement of 7.9 BLEU over it even though PEIT with CRNN has fewer parameters than ItNet, demonstrating the effectiveness of the proposed method, especially the aligner and regularizer that are absent in ItNet. In order to fairly compare with ItNet, we also used ResNet as the visual encoder. We observe that our model based on ResNet still significantly outperforms ItNet. Although CRNN (VGG+BiLSTM) has fewer parameters than ResNet, our experiments show that it is more effective than ResNet in IT task. The reason for this may be that CRNN is more specialized in OCR than ResNet.

Model		BLEU	METEOR
PEIT		45.9	67.5
Aligner	w/o $\mathcal{L}^{v}_{\text{enc}}$	43.9	65.2
	w/o $\mathcal{L}^{t}_{\text{enc}}$	45.7	67.2
	w/o $\mathcal{L}^{v}_{\text{enc}} + \mathcal{L}^{t}_{\text{enc}}$	43.7	65.0
Regularizer	w/o $\mathcal{L}_{\text{KD}}$	44.1	66.0
	w/o $\mathcal{L}_{\text{JSD}}$	44.8	66.7
	w/o $\mathcal{L}_{\text{KD}} + \mathcal{L}_{\text{JSD}}$	43.5	65.2
	w/o all	43.0	64.5

Table 3: Ablation study results of PEIT which is pre-trained with 3M image-translation pairs during the pre-training stage 2.

Comparison with Cascaded Baselines We also implemented three strong cascaded systems with different OCR components. As shown in Table 2, although PEIT is worse than cascaded systems in the case of pre-training, it significantly outperforms all three cascaded systems after being fine-tuned. It is better than the best cascaded system by 2.2+ BLEU and 3.0+ METEOR. The reason for worse performance in the pre-training stage is that we used the United Nations Parallel Corpus to pre-train PEIT in pre-training stage, the domain of which is far different from the e-commerce domain of ECOIT. Additionally, our end-to-end PEIT benefits from low latency, translating images containing text over three times faster than the cascaded systems.

5.4 Ablation Study

To investigate the effect of the proposed vision-text representation aligner and the cross-modal regularizer, both of which aim at bridging the modality

[9]BLEU+case.mixed+numrefs.1+smooth.none+tok.13a +version.2.2.1

[10]http://www.cs.cmu.edu/ alavie/METEOR/

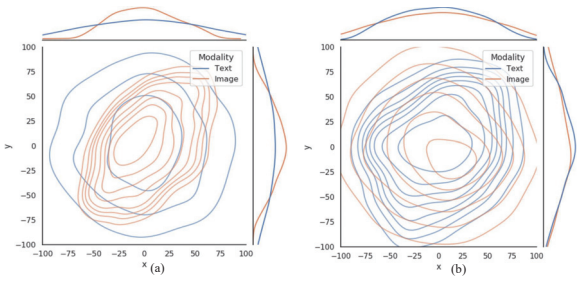


Figure 2: The bivariate density estimation visualization of the averaged representations of images and corresponding texts input to the shared encoder (a) and output from the shared encoder (b).

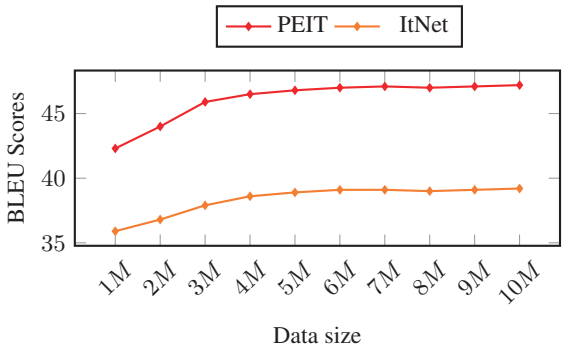


Figure 3: BLEU score curves on the ECOIT test set against the size of synthesized image-translation data used during the pre-training stage 2.

Model	BLEU	METEOR
SC	45.3	67.0
CC	45.9	67.5

Table 4: The BLEU and METEOR scores on Chinese-English in-image translation task. "SC" is self-contrastive learning. "CC" is cross-contrastive learning.

gap for image translation, we conducted ablation study by removing the losses associated with the two components. Results are reported in Table 3, from which we observe that:

- Both the aligner and regularizer are beneficial to PEIT as removing either of them completely or partially results in performance degradation.
- The vision-text representation aligner equipped with the shared encoder is as effective as the cross-modal regulazrizer as removing the former leads to a similar performance drop as removing the latter in terms of both BLEU and METEOR.
- Simultaneously removing both leads to the largest performance drop compared with discarding either of them.

5.5 Analysis on the Effect of the Vision-Text Representation Aligner

To examine whether the aligner is able to alleviate the modality gap in learned representations, we investigated the learned representations of images containing text and those of the corresponding texts input to the shared encoder and output from the shared encoder. We visualize the averaged representations of images and texts in Figure 2. In Figure 2, we average the sequential representations of the image and text sequences over the sequence dimension, and apply the T-SNE dimensionality reduction algorithm to reduce the 256 dimensions to two dimensions. We then plot the bivariate kernel density estimation based on the reduced 2-dim representations. Clearly, we observe that the shared encoder equipped with the vision-text representation aligner significantly improves the modality alignment between visual and textual representations in the semantic space.

5.6 Analysis on the Effect of the Size of the Synthesized Data in the Pre-training Stage 2

The two pre-training stages contribute a lot to PEIT (e.g., transferring knowledge from MT to image translation, aligning the vision and text modality in the shared semantic space), especially the pre-training stage 2. We hence want to investigate the impact of the amount of training data used in pre-training stage 2 on PEIT. For this, we varied the amount of synthesized image-translation data from 1M to 10M pairs. The results are illustrated in Figure 3, which suggests that PEIT is steadily superior to ItNet.

5.7 Self-Contrastive or Cross-Contrastive Learning

Apart from Speech translation (Ao et al., 2022; Fang et al., 2022), which use a shared encoder to leverage larges-cale unlabeled text data, our model use a external encoder rather than a shared encoder to get the fine-grained cross-modal representations. Since PEIT is a multimodal translation model that can accept image or text input, a natural choice is to use the output representations of PEIT for contrastive learning between image and text, rather than using an additional pre-trained

Model	En-Fr	En-Ru
text-only NMT	31.4	21.7
ItNet	19.7	12.6
PEIT	24.8	16.8

Table 5: The BLEU scores on English-French (En-Fr) and English-Russian (En-Ru) image translation.

textual encoder for text representation extraction. We refer to the former as self-contrastive learning, and the latter as cross-contrastive learning. We conduct Chinese-English translation experiment to examine the effect of self-contrastive and cross-contrastive learning, the results are shown in Table 4. The cross-contrastive learning is slightly better than self-contrastive learning, We infer that the poor quality of image representation at the beginning of self-contrastive training degrades the performance of text representation and eventually stabilizes at a relatively poor level. As the text representations in cross-contrastive training come from an external pre-trained text translation model whose representations are constant, therefore we adopt cross-contrastive learning in PEIT.

5.8 Evaluation on Other Languages

We further evaluated our PEIT on English-French and English-Russian image translation. Following ItNet, the visual encoder of PEIT is ResNet with Xavier initialization. We applied a reshape operation to each 2D feature map from the output of the visual encoder, converting them to a 1D vector sequence. We used a parallel MT corpus from UNv1.0-6way[11] and constructed a synthetic corpus with 23M rendered images containing texts from the MT corpus via the same method as described in Section 5.1, except that we ranged the font size from 20 pixels to 30 pixels, and fixed the size of images at 320x480 resolution. We limited the number of lines of text in each rendered image to less than 10 lines as there are no images with > 7 text lines in the test set. We used WMT newstest2013 En-Fr and newstest2016 En-Ru as the validation sets, newstest2014 En-Fr and newstest2017 En-Ru as the test sets to construct the corresponding image translation validation and test sets. Table 5 shows the results of our proposed PEIT and ItNet on these two language pairs. Again we observe that our model substantially outperforms ItNet by 5.1 and 4.2 BLEU on English-French and English-Russian IT, respectively.

[11] https://conferences.unite.un.org/UNCorpus/Home/DownloadOverview

6 Conclusion

In this paper, we have presented PEIT, an end-to-end image translation framework that attempts to bridge the modality gap with pre-trained models, as well as ECOIT, a large-scale high-quality Chinese-English image translation benchmark dataset with real-world e-commerce images containing text, which facilitates future research on this emerging direction. PEIT, containing the vision-text representation aligner and cross-modal regularizer for modality bridging, is pre-trained in two stages and fine-tuned on the curated dataset. Experiments and in-depth analyses demonstrate that PEIT is significantly better than both cascaded image translation systems and previous end-to-end image translation models.

Limitations

Although PEIT is an end-to-end approach to image translation, in the current form, it needs to be pre-trained in two stages with MT and synthesized data and fine-tuned on the curated image translation data. The training procedure is longer than the standard MT task due to the lack of training data and the cross-modality challenge. For the created ECOIT dataset, we used online MT to automatically generate translations and then manually post-edited translations via crowd-sourcing. This significantly reduces the cost of building a large-scale image translation dataset from scratch but may introduce translation noise and "machine translationese" (Vanmassenhove et al., 2021) in comparison to professional human translation.

Acknowledgments

The present research was supported by the Key Research and Development Program of Yunnan Province (Grant No. 202203AA080004). We would like to thank the anonymous reviewers for their insightful comments.

References

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 5723–5738. Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 2278–2282. ISCA.

Iacer Calixto, Daniel Stein, Evgeny Matusov, Sheila Castilho, and Andy Way. 2017a. Human evaluation of multi-modal neural machine translation: A case-study on e-commerce listing titles. In *Proceedings of the Sixth Workshop on Vision and Language, VL@EACL 2017, Valencia, Spain, April 4, 2017*, pages 31–37. Association for Computational Linguistics.

Iacer Calixto, Daniel Stein, Evgeny Matusov, Pintu Lohar, Sheila Castilho, and Andy Way. 2017b. Using images to improve machine-translating e-commerce product listings. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 637–643. Association for Computational Linguistics.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Comput. Speech Lang.*, 66:101155.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation, WMT@ACL 2014, June 26-27, 2014, Baltimore, Maryland, USA*, pages 376–380. The Association for Computer Linguistics.

Qianqian Dong, Mingxuan Wang, Hao Zhou, Shuang Xu, Bo Xu, and Lei Li. 2021. Consecutive decoding for speech-to-text translation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12738–12748. AAAI Press.

Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing Dang, and Haoshuang Wang. 2020. PP-OCR: A practical ultra lightweight OCR system. *CoRR*, abs/2009.09941.

Qingkai Fang, Rong Ye, Lei Li, Yang Feng, and Mingxuan Wang. 2022. STEMM: self-learning with speech-text manifold mixup for speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7050–7062. Association for Computational Linguistics.

Chi Han, Mingxuan Wang, Heng Ji, and Lei Li. 2021. Learning shared semantic space for speech-to-text translation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 2214–2225. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society.

Hirofumi Inaguma, Tatsuya Kawahara, and Shinji Watanabe. 2021. Source and target bidirectional knowledge distillation for end-to-end speech translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 1872–1881. Association for Computational Linguistics.

Puneet Jain, Orhan Firat, Qi Ge, and Sihang Liang. 2021. Image translation network.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and Jingbo Zhu. 2022a. On vision features in multimodal machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 6327–6337. Association for Computational Linguistics.

Yaoyiran Li, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulic. 2022b. Improving word translation via two-stage contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4353–4374. Association for Computational Linguistics.

Yuchen Liu, Hao Xiong, Zhongjun He, Jiajun Zhang, Hua Wu, Haifeng Wang, and Chengqing Zong. 2019. End-to-end speech translation with knowledge distillation. *arXiv preprint arXiv:1904.08075*.

Elman Mansimov, Mitchell Stern, Mia Xu Chen, Orhan Firat, Jakob Uszkoreit, and Puneet Jain. 2020. Towards end-to-end in-image neural machine translation. *CoRR*, abs/2010.10648.

Baoguang Shi, Xiang Bai, and Cong Yao. 2017. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2298–2304.

Yuqing Song, Shizhe Chen, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. 2021. Product-oriented machine translation with cross-modal cross-lingual pretraining. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pages 2843–2852. ACM.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, 34(2–3):97–147.

Yun Tang, Hongyu Gong, Ning Dong, Changhan Wang, Wei-Ning Hsu, Jiatao Gu, Alexei Baevski, Xian Li, Abdelrahman Mohamed, Michael Auli, and Juan Pino. 2022. Unified speech-text pre-training for speech translation and recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1488–1499. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4580–4590. IEEE.

Huiyun Yang, Huadong Chen, Hao Zhou, and Lei Li. 2022. Enhancing cross-lingual transfer by manifold mixup. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Pengcheng Yang, Boxing Chen, Pei Zhang, and Xu Sun. 2020. Visual agreement regularized training for multi-modal machine translation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 9418–9425. AAAI Press.

Rong Ye, Mingxuan Wang, and Lei Li. 2021. End-to-end speech translation via cross-modal progressive training. *arXiv preprint arXiv:2104.10380*.

A Appendix

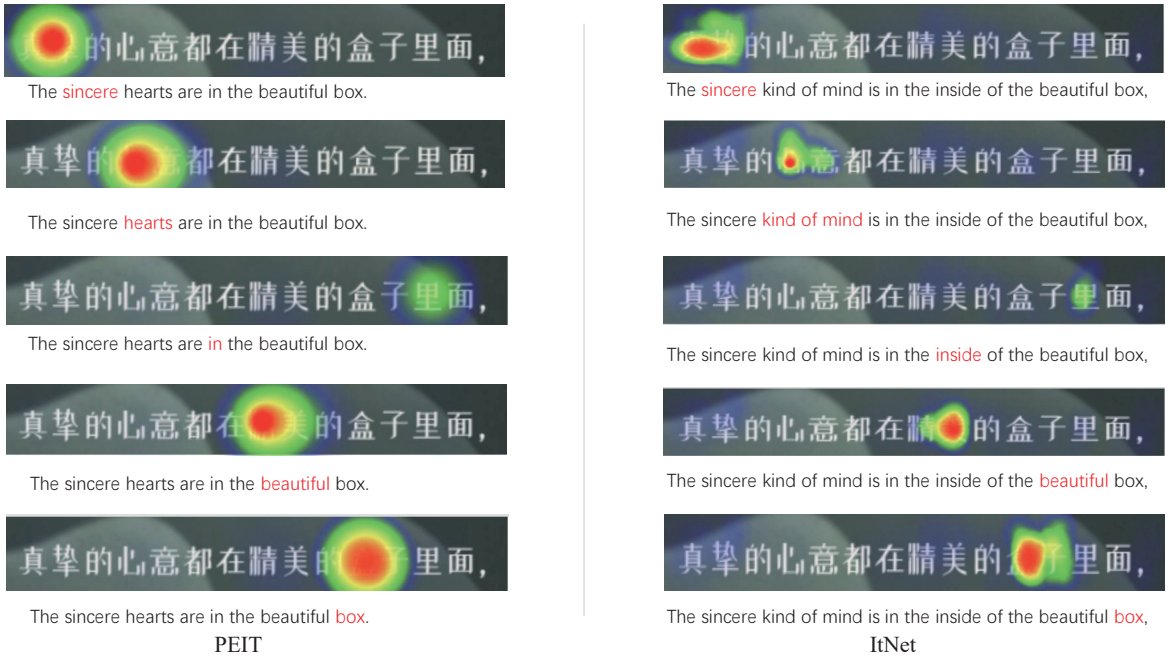
A.1 Setting Details

For Chinese-English, we segment Chinese data using characters. We limit the maximum sentence length to 20 tokens. For English-French and English-Russian, we do not filter out the sentence length. We apply byte pair encoding to segment all sentences with merge operations of 32K. All out-of-vocabulary words are mapped to a distinct token <UNK>. We use the schedule strategy with 4,000 warmup steps. The training batch consist of approximately 25,000 source tokens and 25,000 source and target tokens. Label smoothing of the value of 0.1 is used for training. We trained our models for 100k steps on 8 NVIDIA TITAN RTX GPUs. For evaluation, we use beam search with a width of 5. We do not apply checkpoint averaging on the parameters for evaluation. We adopted two strong visual encoder architecture, ResNet-101 and CRNN (VGG+BiLSTM). For ResNet-101, we used Xavier initialization to initialize parameters. For CRNN, we use a pre-trained text recognition model from easyocr[12] to initialize parameters.

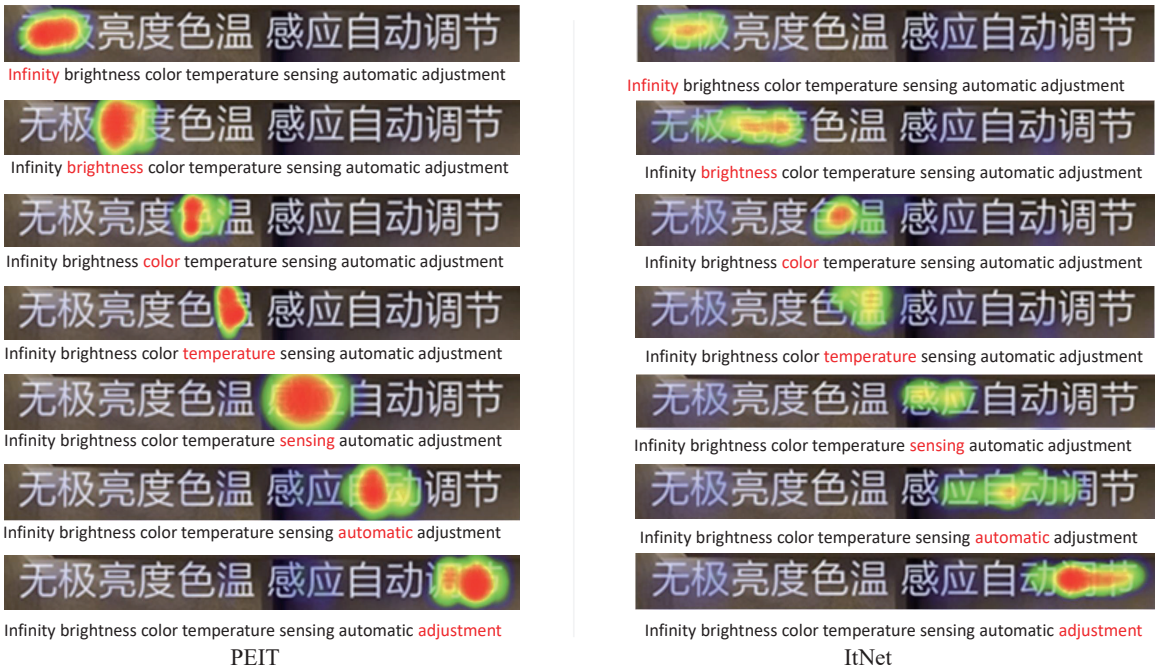
A.2 Visualization

In Section 5.5, we demonstrate that the the proposed PEIT could significantly improve the similarity of word representations across modalities. We also show the visualization of two examples (a) and (b) in Figure 4. The visualization shows the translations and the cross-attention assignment probabilities for visual information of PEIT and ItNet. It demonstrates the proposed PEIT can enhance shared fine-grained latent translation information. We can observe that our method can gets better translations than ItNet. It means that our method makes target word translations get more reasonable visual information compared to ItNet.

[12]https://github.com/JaidedAI/EasyOCR



(a)



(b)

Figure 4: Translation cases and Visualization. Colored words represent the cross-attention assignment probabilities for visual information.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
7

☑ A2. Did you discuss any potential risks of your work?
7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

☐ A4. Have you used AI writing assistants when working on this paper?
Not applicable. Left blank.

B ☑ Did you use or create scientific artifacts?

5

☑ B1. Did you cite the creators of artifacts you used?
5

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
5

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
3

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.