Increasing Visual Awareness in Multimodal Neural Machine Translation from an Information Theoretic Perspective

Baijun Ji[1], **Tong Zhang**[1], **Yicheng Zou**[2], **Bojie Hu**[1*], **Si Shen**[3]
[1] Tencent Minority-Mandarin Translation, Beijing, China
[2] School of Computer Science, Fudan University
[3] Research Base on Interdisciplinary Terminology and Translation, Nanjing University
[1]{baijunji,zatozhang, bojiehu}@tencent.com [2] yczou18@fudan.edu.cn

Abstract

Multimodal machine translation (MMT) aims to improve translation quality by equipping the source sentence with its corresponding image. Despite the promising performance, MMT models still suffer the problem of input degradation: models focus more on textual information while visual information is generally overlooked. In this paper, we endeavor to improve MMT performance by increasing visual awareness from an information theoretic perspective. In detail, we decompose the informative visual signals into two parts: source-specific information and target-specific information. We use mutual information to quantify them and propose two methods for objective optimization to better leverage visual signals. Experiments on two datasets demonstrate that our approach can effectively enhance the visual awareness of MMT model and achieve superior results against strong baselines.

1 Introduction

Multimodal machine translation (MMT) typically improves text translation quality by introducing an extra visual modality input. This task hypothesizes that the corresponding visual context is helpful for disambiguation and omission completion when the source sentence is ambiguous or even incorrect. Compared to the purely text-based machine translation, the critical point of MMT is to find out an effective approach to integrating images into a MT model that makes the most of visual information.

Most of the existing studies have dedicated their efforts to the way of extracting multi-granularity visual features for integration (Calixto and Liu, 2017; Delbrouck and Dupont, 2017; Ive et al., 2019; Zhao et al., 2022; Li et al., 2022; Fang and Feng, 2022) or designing model architectures for better message passing across various modalities (Calixto et al., 2017, 2019; Yao and Wan, 2020; Yin et al., 2020; Lin et al., 2020; Liu et al., 2021), which achieve promising performances in most of the multimodal scenarios.

Despite the success of these methods, some researchers reveal that visual modality is underutilized by the modern MMT models (Grönroos et al., 2018; Elliott, 2018). Elliott (2018) leverage an adversarial evaluation method to appraise the awareness of visual contexts of MMT models. An interesting observation is that the disturbed image input generally shows no significant effects on the performance of public MMT models, indicating the insensitivity to the visual contexts. Recently, Wu et al. (2021) revisit the need for visual context in MMT and find that their strong MMT model tends to ignore the multimodal information. They suggest that the gains from the multimodal signals over text-only counterparts are, in fact, due to the regularization effect.

In practice, the existing approaches mainly focus on a better interaction method between textual- and visual- information. In training, they generally employ maximum likelihood estimation (MLE) to optimize the entire model. In most cases, the source sentence carries sufficient context for translation, leading to the neglect of visual modality under such training paradigm. However, the visual information is still crucial when the source sentence faces the problem of incorrection, ambiguity and gender-neutrality . Accordingly, we argue that the visual modality remains an excellent potential to facilitate translation, which inspire us to explore taking full advantage of the visual modality in MMT.

To this end, we propose increasing visual awareness from an information theoretic perspective. Specifically, we quantify the amount of visual information with an information-theoretic metric, namely Mutual Information (MI). The quantification is then divided into two parts: 1) the source-specific part, which means the information between source texts and images; 2) the target-specific part,

*Corresponding author

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 6755 - 6764
December 7-11, 2022 ©2022 Association for Computational Linguistics

which means the mutual information between target texts and images given source texts, namely conditional mutual information (CMI). For the source-specific part, we maximize the mutual information by resorting to a lower bound of MI. For the target-specific part, we propose a novel optimization method for maximizing the conditional mutual information (See details in 3.2). We evaluate our approach on two public datasets. The experimental results and in-depth analysis indicate that our method can effectively enhance the MMT model's sensitivity to visual context, which manifests that MMT has great potential when the utilization of visual context is adequate.

To conclude, the main contributions of our paper are three-fold:

1) We propose a novel MI-based learning framework to improve multimodal machine translation by increasing visual awareness. This framework is interpretable and can quantify how much visual information is utilized.

2) To optimize the mutual information objective, we divide the corresponding visual information into source/target-specific part. Furthermore, we propose a novel method to maximize the conditional mutual information in the target-specific part.

3) Comprehensive experiments demonstrate that our method can effectively encourage MMT models to keep sensitive to visual context and significantly outperform strong baselines on two public MMT datasets.

2 Background

2.1 Multimodal Machine Translation

We start with the formulation of a regular MMT task. Given a triplet dataset of $\{(x^{(i)}, y^{(i)}, z^{(i)})\}_1^N$, the problem naturally turns into the likelihood maximization:

$$\mathcal{L}_{MMT} = -\frac{1}{N}\sum_{i=1}^{N} \log p(y^{(i)}|x^{(i)}, z^{(i)}), \quad (1)$$

where z denotes the image itself and x, y denote the description of the image in two different languages. Recently, MMT models are usually consisted of a Transformer-based encoder-decoder model and an additional image encoder. There exist several lines of works to integrate convolutional features into the NMT model. We adopt the global feature, which is in the form of a single vector of an image. This integration method has been proven beneficial to some extent in prior works.

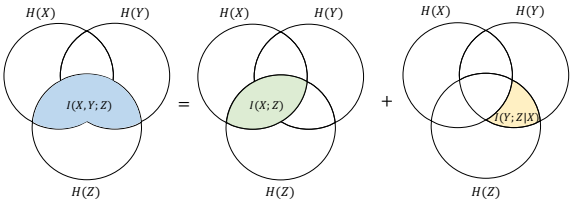

Figure 1: Chain rule for multual information. We decompose visual awareness into two parts: 1) mutual information between source text and image and 2) conditional mutual information between target text and image.

2.2 Mutual Information (MI)

Mutual Information (MI) relates two random variables X and Y and measures the amount of information obtained from X by observing the random variable Y, which is defined as:

$$I(\mathrm{X};\mathrm{Y}) = H(\mathrm{X}) - H(\mathrm{X}|\mathrm{Y}), \quad (2)$$

where $H(\cdot)$ denotes information entropy. Mutual information is symmetric and non-negative. High mutual information shows a large reduction of the random variable X's uncertainty given another variable Y.

3 Methodology

In this section, we mathematically describe the basic notion of our proposed method from an information theoretic perspective. Let X denotes a random variable over source sentences, Y denotes a random variable over target sentences, and Z is a random variable over images. They are jointly distributed according to some true p.m.f. $p(x, y, z)$. We formulate the **Visual Awareness** of a MMT model as a joint mutual information $I(\mathrm{X}, \mathrm{Y}; \mathrm{Z})$, which indicates how much the visual modality contributes to the overall system. A high joint mutual information suggests a strong sensitivity of the MMT model to visual information. However, it is intractable to directly maximize this joint mutual information in training. Thus, in this work we decompose $I(\mathrm{X}, \mathrm{Y}; \mathrm{Z})$ into two parts by the chain rule of mutual information and optimize these two

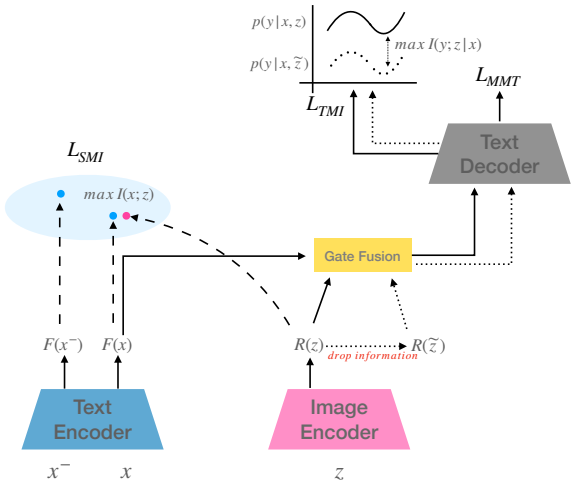

Figure 2: An overview of our method, where x and z represent the source text and image, x^- is a contrastive sample of x, and $\widetilde{z}$ is a deteriorated version of z by removing visual information. L_{SMI} aims to maximize MI between x and z to learn grounded representation. L_{TMI} aims to maximize between y and z to strengthen the effect of visual context in generation.

parts separately:

$$I(\mathrm{X},\mathrm{Y};\mathrm{Z}) = \underbrace{I(\mathrm{X};\mathrm{Z})}_{source-specific} + \underbrace{I(\mathrm{Y};\mathrm{Z}|\mathrm{X})}_{target-specific}. \tag{3}$$

As shown in Figure 1, the green part $I(\mathrm{X};\mathrm{Z})$ is the source-specific mutual information (SMI), which indicates the mutual information between source texts and images; the yellow part $I(\mathrm{Y};\mathrm{Z}|\mathrm{X})$ is the target-specific conditional mutual information (TMI), which indicates the mutual information between target texts and images given source texts. Figure 2 illustrates the overview of our proposed method. In the following, we describe how to optimize the source-specific and target-specific mutual information in detail.

3.1 Source-Specific Mutual Information

Source-specific mutual information $I(\mathrm{X};\mathrm{Z})$ determines how much information is shared between the source texts and images. Inspired by van den Oord et al. (2018), we minimize the **InfoNCE** (Logeswaran and Lee, 2018) loss $\mathcal{L}_{SMI}$ to maximize a lower bound on $I(\mathrm{X};\mathrm{Z})$.

$$\begin{aligned} I(\mathrm{X};\mathrm{Z}) \geq & -\mathbb{E}_{p(x,z)}\Big[f(\mathcal{F}(x),\mathcal{R}(z)) - \\ & \mathbb{E}_{q(x^-)}\log\Big[\sum_{x^-}\exp f(\mathcal{F}(x^-),\mathcal{R}(z))\Big]\Big] \\ = & \mathcal{L}_{SMI}, \end{aligned} \tag{4}$$

where $f(\cdot)$ is the cosine similarity function between $\mathcal{F}(x)$ and $\mathcal{R}(z)$. $\mathcal{F}(x)$ denotes the pooled textual representation of a source sentence x and $\mathcal{R}(z)$ denotes the pooled visual representation from a pre-trained vision model (e.g., ResNet or Vision Transformer). x^- is a negative source sentence drawn from the proposal distribution $q(x^-)$. This contrastive learning objective aligns the textual and visual representations into a unified semantic space, and encourages the text encoder to generate grounded representations, which has shown effectiveness in various scenarios (Elliott and Kádár, 2017; Kádár et al., 2017).

3.2 Target-Specific Conditional Mutual Information

In MMT, $I(\mathrm{Y};\mathrm{Z}|\mathrm{X})$ represents how much information the image Z provides corresponding to the target Y given the source X. Our goal is to maximize the conditional mutual information as follows:

$$I(\mathrm{Y};\mathrm{Z}|\mathrm{X}) = H(\mathrm{Y}|\mathrm{X}) - H(\mathrm{Y}|\mathrm{X},\mathrm{Z}). \tag{5}$$

Note that the term $H(\mathrm{Y}|\mathrm{X})$ refers to the conditional probability $P(Y|X)$, which is non-accessible in a MMT model. To remedy this issue, we instead introduce a pseudo conditional mutual information $I(\mathrm{Y};\widetilde{\mathrm{Z}}|\mathrm{X})$, which is formulated as:

$$I(\mathrm{Y};\widetilde{\mathrm{Z}}|\mathrm{X}) = H(\mathrm{Y}|\mathrm{X}) - H(\mathrm{Y}|\mathrm{X},\widetilde{\mathrm{Z}}), \tag{6}$$

where $\widetilde{\mathrm{Z}}$ represents a set of **deteriorated images** $(\widetilde{z_1},\widetilde{z_2},\cdots,\widetilde{z_n})$. The visual information of these deteriorated images are artificially interfered by some means. For a well-trained visual-aware MMT, we suppose a deteriorated image would not significantly reduce the uncertainty of target sentence prediction compared to its text-only counterpart. Thus our goal is to minimize the pseudo conditional mutual information $I(\mathrm{Y};\widetilde{\mathrm{Z}}|\mathrm{X})$. Accordingly, we have the training objective as follows:

$$\begin{aligned} & I(\mathrm{Y};\mathrm{Z}|\mathrm{X}) - I(\mathrm{Y};\widetilde{\mathrm{Z}}|\mathrm{X}) \\ & = -H(\mathrm{Y}|\mathrm{X},\mathrm{Z}) + H(Y|\mathrm{X},\widetilde{\mathrm{Z}}) \\ & = \sum_{x,y,z} p(x,y,z)\log p(y|x,z) \\ & \quad - \sum_{x,y,\widetilde{z}} p(x,y,\widetilde{z})\log p(y|x,\widetilde{z}). \end{aligned} \tag{7}$$

Minimizing $I(\mathrm{Y};\widetilde{\mathrm{Z}}|\mathrm{X})$ means that the MMT can correctly discriminate the useless visual information. Moreover, maximizing $I(\mathrm{Y};\mathrm{Z}|\mathrm{X})$ means that

the MMT model is able to make the most of the grounding information from images.

However, the above training objective cannot be directly calculated since the probability mass function $p(x, y, z)$ is unobtainable. Therefore, we follow Bugliarello et al. (2020); Fernandes et al. (2021) to use a Monte Carlo estimator to approximate it. Under sufficient held-out training data of $\{(x^i, y^i, z^i)\}_{i=1}^{N}$, we can estimate the true cross-entropy as follows:

$$H(\mathrm{Y}|\mathrm{X},\mathrm{Z}) \approx -\frac{1}{N}\sum_{i=1}^{N}\log p(y^{(i)}|x^{(i)}, z^{(i)}). \quad (8)$$

This estimation will get more accurate when the training data set is large enough. We note that $\widetilde{z}$ is an deteriorated version of the image z and is also drawn from the same true probability distribution. We can employ the same estimation method to approximate the cross-entropy. Finally, we have an estimation for the optimization objective:

$$\mathcal{L}_{TMI} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{p(y^{(i)}|x^{(i)}, z^{(i)})}{p(y^{(i)}|x^{(i)}, \widetilde{z}^{(i)})}. \quad (9)$$

Moreover, the training process can be unstable when the term $p(y^{(i)}|x^{(i)}, \widetilde{z}^{(i)})$ becomes very small, ignoring the fact that x has already contained the majority of information for translation. Following Huang et al. (2018), we perform back propagation only when $\log\frac{p(y^{(i)}|x^{(i)},z^{(i)})}{p(y^{(i)}|x^{(i)},\widetilde{z}^{(i)})}$ is lower than a margin value m. The above objective can be converted into:

$$\mathcal{L}_{TMI} = \frac{1}{N}\sum_{i=1}^{N}\max\{0, m - \log\frac{p(y^{(i)}|x^{(i)}, z^{(i)})}{p(y^{(i)}|x^{(i)}, \widetilde{z}^{(i)})}\}. \quad (10)$$

3.3 Deteriorated Image Generation

In this section we describe how to generate a deteriorated view of the image z. One straightforward approach is masking relevant objects in the visual modality by using an object detector. This method can reduce the amount of information contained in a picture. However, using an external object detector is time-consuming, making the MMT task more complex. In this work, we adopt a more efficient and feasible dropout-based method to generate the deteriorated images at the representation level. Specifically, we randomly select the units in the visual vector and mask them with zeros by following a Bernoulli distribution of probability ρ. In this way, the majority of information in the image representations will be corrupted.

3.4 Model Architecture

We use a simple gated mechanism for multimodal fusion, which has achieved competitive results versus other complicated models. Following the recent work (Wu et al., 2021), the source sentences and the images are separately encoded by using a transformer-based encoder and a pre-trained Resnet model. Formally, the textual and image representations are combined by a learnable gate vector λ as:

$$\lambda = \mathrm{Sigmoid}(W_\lambda \cdot \mathrm{Concat}[\mathcal{F}(x); \mathcal{R}(z)]),$$
$$\mathrm{H}_{\mathrm{fusion}} = \mathcal{F}(x) + \lambda \odot \mathcal{R}(z). \quad (11)$$

W_λ is a trainable matrix and $\odot$ denotes element-wise product. The decoder takes the modality input $\mathrm{H}_{\mathrm{fusion}}$ in the recurrent fashion as the text-only NMT does. The deteriorated image input is fused in the same way by replacing $\mathcal{R}(z)$ with $\mathcal{R}(\widetilde{z})$.

3.5 Training Objective

Finally, the two MI-based objectives are jointly optimized with MMT loss from scratch:

$$\mathcal{L} = \mathcal{L}_{MMT} + \alpha|s|\mathcal{L}_{SMI} + \beta|s|\mathcal{L}_{TMI}, \quad (12)$$

where α and $\beta \in [0, 1]$ are the hyperparameters for balancing MI maximizing objective and MT loss, and $|s|$ denotes the average length of the sequence since the MI-based loss is sentence-level (Pan et al., 2021).

4 Experiments

4.1 Data

We evaluate our methods on two standard MMT datasets, including **AmbigCaps** (Li et al., 2021) and **Multi30K** (Elliott et al., 2016).

AmbigCaps contains 81K Turkish-English parallel sentence pairs with visual annotations. We use 1,000 sentences provided by Li et al. (2021) for validation and testing, respectively. This dataset is processed into a gender-ambiguous one, in which sentences containing the gender of the entity and professions with gender implications removed. Moreover, AmbigCaps is designed to translate a gender-neutral language (Turkish) into a gender-specified language (English), which is suitable for the MMT

Row	Method	Validation		Test	
		BLEU	METEOR	BLEU	METEOR
Only Text					
1	Transformer♠(Li et al., 2021)	-	-	35.71	-
2	Transformer♢	37.46	69.31	37.60	68.64
Existing MMT Systems					
3	Imagination♢(Elliott and Kádár, 2017)	37.83	68.92	38.11	69.25
4	Gated Fusion♠(Li et al., 2021)	-	-	36.68	-
5	Selective Attn♢(Li et al., 2022)	38.47	68.91	38.30	69.31
Our MMT System					
6	Gated Fusion♢	38.29	69.62	38.06	69.13
7	Our Method	**39.24**	**70.30**	**39.40**	**70.22**

Table 1: Tr ⇒ En translation results on AmbigCaps dataset. ♠ indicates the results come from the original papers and ♢ indicates the results are from our implementation. All models are based on the transformer-tiny configuration.

task since the source textual information is not naturally sufficient.

Multi30K is a widely used dataset for MMT, which contains 29K images with one English description and a German manual translation. We follow a standard split for experiments and use 1,014 sentences for validation and 1,000 sentences for testing (Test2016). In addition, we also evaluate the WMT17 test set (Test2017) and the ambiguous MSCOCO test set, which include 1,000 and 461 triplets, respectively.

4.2 System Setting

We use the Transformer-Tiny (Vaswani et al., 2017) configuration to conduct all of our experiments, which can even obtain better performance than those large models (Wu et al., 2021). The tiny transformer model consists of 4 encoder and decoder layers, with 1024 embedding/hidden units, 4096 feed-forward filter size and 4 heads per layer. We employ the Adam optimizer with lr $= 0.005$, $t_{\text{warm\_up}} = 2000$ and dropout $= 0.3$ for optimization. At the training time, each training batch includes 4096 source tokens and target tokens. For a fair comparison with previous works, we use the early stopping strategy if the performance on validation does not gain improvements for ten epochs on Multi30K. Empirically, $1e^{-3}$ for α and $1e^{-4}$ for β can get the best results. The model for AmbigCaps is trained on a single Tesla P40 GPU and the one for Multi30k is trained on two GPUs for a fair comparison with previous work.

In particular, the number of patience on AmbigCaps is set to twenty epochs since the converge speed on this dataset is slow. We average the results of the last ten checkpoints during inference. At decoding time, we generate with $beam\_size = 5$ and length penalty is equal to 1.0.

For visual features, we use a ResNet-50 (He et al., 2016) model pre-trained on ImageNet as the image encoder. The dimension of the global feature is 2048 and we use a projection matrix to convert the shape of image features into that of text features. For data processing, we generate subwords following Sennrich et al. (2016) with 10,000 merging operations jointly to segment words into subwords, which generates a vocabulary of 6,765 and 9,712 tokens for AmbigCaps and Multi30K, respectively.

Finally, we report BLEU[1] (Papineni et al., 2002) and METEOR[2] (Denkowski and Lavie, 2014) scores to evaluate the quality of different translation system. Our implementation are implemented on Fairseq (Ott et al., 2019).

4.3 Baseline Methods

We compare our method with baselines as follows:

- **Vanilla Transformer** (Vaswani et al., 2017). We report the Transformer-tiny results.
- **Imagination** (Elliott and Kádár, 2017; Helcl and Libovický, 2018). This method adopts a margin loss as a regularizer to the text encoder in order to learn grounded representations. We reimplement this method in our

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

[2] https://github.com/facebookresearch/vizseq/blob/main/vizseq/scorers/meteor.py

Method	Test2016		Test2017		MSCOCO	
	B	M	B	M	B	M
Transformer (Vaswani et al., 2017)	41.02	68.22	33.36	62.05	29.88	56.64
Imagination (Elliott and Kádár, 2017)	41.31	68.06	32.89	61.29	29.9	56.57
UVR (Zhang et al., 2020)	40.79	-	32.16	-	29.02	-
DCCN (Lin et al., 2020)	39.7	-	31.0	-	26.7	-
Multimodal Graph(Yin et al., 2020)	39.8	-	32.2	-	28.7	-
Gated Fusion (Wu et al., 2021)	41.55	68.34	33.59	61.94	29.04	56.15
Selective Attn (Li et al., 2022)	**41.93**	68.55	33.62	61.61	29.72	**56.94**
Our Method	41.77	**68.60**	**34.58**	**62.40**	**30.61**	56.70

Table 2: En ⇒ De results on Multi30K dataset. Some results are from Wu et al. (2021) and Li et al. (2022). B and M denote BLEU and METEOR score.

Transformer-based models instead of RNNs for a fair comparison.

- **Gated Fusion** (Wu et al., 2021). It uses a gate vector to combine textual representations and image representations and then feds them into the MMT decoder. This is a simple yet effective method and can outperform most of MMT systems.

- **Selective Attn** (Li et al., 2022). This method has the same model architecture with **Gated Fusion** and supports more fine-grained features extracted from images.

Since **AmbigCaps** is a new MMT dataset, we reproduce the above baseline methods and report the corresponding results.

4.4 Results on Tr⇒En Translation Task

The evaluation results of different MMT systems on the Tr ⇒ En translation task are presented in Table 1. By observing the reported results, we draw the following interesting conclusions:

First, we observe that our implementation for the text-only transformer and Gated Fusion outperform the results reported in the original paper. This is due to the larger number of patience. It also indicates our method is more competitive and still achieves improvement against the strong baselines.

Second, all of the MMT models beat the text-only transformer, which indicates that AmbigCaps is a suitable dataset and verifies the benefit of the visual modality. Gated Fusion gains 0.83 BLEU points and 0.31 METEOR points compared to the text-only model. Selective Attn shows further improvement due to the more-grained features. Our model significantly outperforms the text-only transformer by 1.8 on BLEU and 1.58 on METEOR.

Third, our model also beats Imagination, Selective Attn and Gated Fusion. The gap between different baseline results is not significant. Note that the Imagination model does not use the image during the inference time. Our model uses the same model architecture with Selective Attn and Gated Fusion and consistently improves the BLEU score, 1.30 and 1.34 points more than Selective Attn and Gated Fusion. It indicates that our method can significantly utilize the image context and improve translation by using the same model configuration.

4.5 Results on En⇒De Translation Task

Table 2 shows the main results on the En⇒De translation task. We find that the text-only transformer can act as a solid baseline with the aid of the elaborate model configuration and training settings, which even beats DCCN and Multimodal Graph models. The BLEU score of our method on Test2016 is quite marginal compared to the base model Gated Fusion, which is the point that we attempt to improve. We argue that the Test2016 test set may contain less textual ambiguity. Besides, we observe significant BLEU/METEOR gains on Test2017 and MSCOCO, outperforming the text-only baseline by 1.22/0.73 points and 0.99/1.57 points. Moreover, our method beats all robust baseline systems on the two test sets. These results further indicate the necessity of visual context for multimodal translation.

System	B	M	ID ↑	GA ↑
$\mathcal{L}_{MMT}$	38.06	69.13	3.32	79.73%
+ $\mathcal{L}_{SMI}$	38.72	69.50	3.35	80.23 %
+ $\mathcal{L}_{TMI}$	38.74	69.93	4.37	79.93%
Full Model	39.40	70.22	8.31	80.65%

Table 3: Ablation study on AmbigCaps. B and M denote BLEU and METEOR score. ↑ indicates the visual awareness is better as the value is larger.

5 Model Analysis

5.1 Ablation Study

To make a better understanding of the influence of the training objective, we perform ablation studies to validate the impact of each part on translation quality. Column 2 to column 3 in Table 3 shows both $\mathcal{L}_{SMI}$ and $\mathcal{L}_{TMI}$ results in a close gain in BLEU score, 0.66 and 0.68 points more than the vanilla MMT model. We find $\mathcal{L}_{TMI}$ can perform better than $\mathcal{L}_{SMI}$ in METEOR score. Combining the two training objectives can further boost the translation performance, which means $\mathcal{L}_{SMI}$ and $\mathcal{L}_{TMI}$ can be beneficial to each other. It indicates that our method improves the image utilization at both the source and target sides.

Besides, we also investigate how the visual signals help to translate and whether the gains on translation come from the increasing visual awareness . To achieve this goal, we introduce two new metrics to evaluate the visual awareness of the MMT model, which are *Incongruent Decoding* and *Gender Accuracy*. We will first explain how these metrics are calculated.

Incongruent Decoding (ID) We follow Elliott et al. (2018) to use an adversarial evaluation method to test if our method is more sensitive to the visual context. We randomly replace the image with an incongruent image or set the congruent image's feature to all zeros. Then we observe the value $\triangle$BLEU by calculating the difference between the congruent data and the incongruent one. A larger value means the model is more sensitive to the image context.

Gender Accuracy (GA) Following Li et al. (2021), we divide the translation sentences into three categories: *male, female* and *undetermined*. *Male* means that the target sentences contains at least one of the male pronouns {'he', 'him', 'his', 'himself'}. The definition of *Female* is similar to *Male*. The sentence is *undetermined* when it contains both male and female pronouns or neither, and will not be considered for the calculation of gender accuracy. It is designed to determine whether the ambiguous words are indeed translated correctly.

Column 4 to column 5 in Table 3 show the results of visual awareness evaluation. When given the incongruent image, $\mathcal{L}_{MMT}$ drops more than 3 points, indicating that this simple method can also utilize the visual context to some extent. When $\mathcal{L}_{SMI}$ or $\mathcal{L}_{TMI}$ is used, the dropped value will become larger. Combining the two objective can lead to a great decline. It means the visual part plays a vital role in our system. There is also a significant increase in gender accuracy with our method. It also proves that the translation improvement partly benefits from the correct translation of ambiguous words.

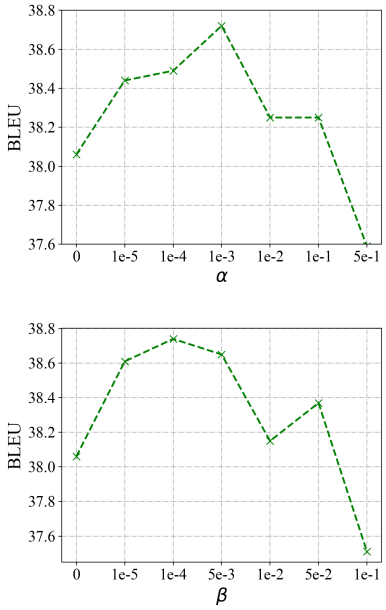

Figure 3: Illustration of BLEU score with different α (left) and β (right).

5.2 The Influence of Weight α and β

To evaluate the effects of the coefficients of α and β, we observe the translation performance over various values by setting the corresponding coefficient to zero. Figures 3 shows the performance fluctuates greatly with a low rate of α and β. With the coefficients increasing slightly, the BLEU score improves, which probes the effectiveness of increasing visual awareness by applying our method. The translation performance drops slightly as the coefficients increase since the MI-based loss will dominate the optimization and ignore the MLE

training objective. In particular, the best performance is achieved when $\alpha = 1e^{-3}$ or $\beta = 1e^{-4}$ respectively.

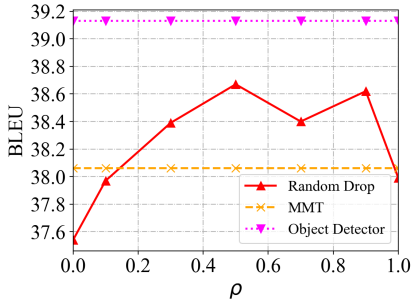

Figure 4: Illustration of BLEU score on AmbigCaps with various methods. The orange dash line denotes the model that only uses $\mathcal{L}_{MMT}$. The red solid line denotes the one that masks features with different probabilities ρ. The fuchsia dotted line denotes the one that uses an object detector for masking.

5.3 The Influence of Deteriorated Image Generation.

First, we compare the effects of different methods to generate the deteriorated image in Figure 4. We find that using a object detector for masking relevant objects can generate the best translation results. It is not surprising since the picture in AmbigCaps is not complex and the detector model is competent to detect the most objects correctly. We should note that our method is a unsupervised manner but can still generate the competitive results. We also explore the different influence of $\mathcal{L}_{TMI}$ with the various dropout rate ρ. When the value of ρ is very small, the BLEU score is lower than the baseline. This is due to that a small value of ρ means that most image information is preserved and the term $I(\mathrm{Y}; \widetilde{\mathrm{Z}}|\mathrm{X})$ should be as large as possible instead of being minimized. Meanwhile, $\rho = 1$ means that the image is totally unused, and the global vector is all zeros. This conspicuous feature induces the model to take a shortcut to discriminate the useless image easily. Only the appropriate value of $\rho \geq 0.5$ can get stable and good results. It indicates a well-designed method to generate deteriorated image is critical for better translation.

6 Related Work

6.1 Mutual Information Maximization

Mutual information has been widely explored in recent NLP work. Bugliarello et al. (2020) proposed cross-mutual information to measure the neural translation difficulty. Fernandes et al. (2021) used conditional mutual information as the metric to estimate the context usage in document NMT. Our work differs in that we employ MI as a training objective to enhance image usage instead of a pure metric. Wang et al. (2022) maximized MI to solve the OOV problem in NER. Kong et al. (2020) improved language model pre-training from a mutual information maximization perspective. To our best knowledge, we are the first to employ mutual information to increase image awareness in MMT generation task.

6.2 Efficient Visual Context Modeling

To equip the MMT model with the visual context modeling ability, Caglayan et al. (2021) and Song et al. (2021) used cross-modal pre-training to learn visually-grounded representations. This point is similar to our source-related MI method, which aligns the textual and visual representations in a shared semantic space. Wang and Xiong (2021) forced the model to reward masking irrelevant objects and penalize masking relevant objects. It also makes the model sensitive to the changes in the visual modality and enhances visual awareness. But detecting the relevant objects and correlating the regions with the vision-consistent target words are rather complex.

6.3 Contrastive Learning

Our method is highly related to contrastive learning. The source-specific loss is optimized by using the **InfoNCE** bound, and the target-specific loss can also be viewed as using a contrastive image form and lowering the probability of generating the target. Contrastive learning has been recently used in translation tasks. Pan et al. (2021) closed the gap among representations of different languages to improve the multilingual translation. Zhang et al. (2021) meliorated word representation space to solve the low-frequency word prediction. Hwang et al. (2021) used coreference information for generating the contrastive samples to train the model to be sensitive to coreference inconsistency. Different from this work, our method does not need external knowledge.

7 Conclusion

In this paper, we summarize that only using the MLE objective is unable to ensure the image in-

formation is fully utilized, and there still exists tremendous potential to explore. We propose a method based on mutual information to increase visual awareness in multimodal machine translation. The information from images is divided into two parts(source-specific and target-specific). We quantify this contribution from the image and put forward the corresponding optimization solutions. Experiments on Multi30K and AmbigCaps show the superiority of our methods. Detailed analysis experiments conducted probes that our methods are quite interpretable. In the feature, we would like to apply our MI-based method to visual language pre-training and other unsupervised tasks to improve the downstream tasks.

8 Limitations

Since we conduct all experiments on the machine translation task, it is unclear whether our approach can benefit other multimodal tasks, e.g. Visual QA or multimodal Dialog. Another thing worth noting is that the image-source-target triplet data is not easily available in reality. Thus how to effectively use the unpaired image in the unsupervised way is a more promising research.

References

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of english than into it: Measuring neural translation difficulty by cross-mutual information. In *ACL*.

Ozan Caglayan, Menekse Kuyu, Mustafa Sercan Amac, Pranava Swaroop Madhyastha, Erkut Erdem, Aykut Erdem, and Lucia Specia. 2021. Cross-lingual visual pre-training for multimodal machine translation. *ArXiv*, abs/2101.10044.

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

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

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

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *EMNLP*.

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

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *EMNLP*.

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

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

Qingkai Fang and Yang Feng. 2022. Neural machine translation with phrase-level universal visual representations. In *ACL*.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André F. T. Martins. 2021. Measuring and increasing context usage in context-aware machine translation. In *ACL*.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma T. Laaksonen, Bernard Mérialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The memad submission to the wmt18 multimodal translation task. In *WMT*.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Jindřich Helcl and Jindřich Libovický. 2018. Cuni system for the wmt18 multimodal translation task. In *WMT*.

Jiaji Huang, Yi Li, Wei Ping, and Liang Huang. 2018. Large margin neural language model. In *EMNLP*.

Julia Ive, Pranava Swaroop Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. *ArXiv*, abs/1906.07701.

Ákos Kádár, Grzegorz Chrupała, and A. Alishahi. 2017. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43:761–780.

Yong keun Hwang, Hyungu Yun, and Kyomin Jung. 2021. Contrastive learning for context-aware neural machine translation using coreference information. In *WMT*.

Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama. 2020. A mutual information maximization perspective of language representation learning. *ArXiv*, abs/1910.08350.

Bei Li, Chuanhao Lv, Zefan Zhou, Tao Zhou, Tong Xiao, Anxiang Ma, and Jingbo Zhu. 2022. On vision features in multimodal machine translation. *ArXiv*, abs/2203.09173.

Jiaoda Li, Duygu Ataman, and Rico Sennrich. 2021. Vision matters when it should: Sanity checking multimodal machine translation models. In *EMNLP*.

Huan Lin, Fandong Meng, Jinsong Su, Yongjing Yin, Zhengyuan Yang, Yubin Ge, Jie Zhou, and Jiebo Luo. 2020. Dynamic context-guided capsule network for multimodal machine translation. *Proceedings of the 28th ACM International Conference on Multimedia*.

Peng Liu, Hailong Cao, and Tiejun Zhao. 2021. Gumbel-attention for multi-modal machine translation. *ArXiv*, abs/2103.08862.

Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. *ArXiv*, abs/1803.02893.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li. 2021. Contrastive learning for many-to-many multilingual neural machine translation. *ArXiv*, abs/2105.09501.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *ArXiv*, abs/1508.07909.

Yuqing Song, Shizhe Chen, Qin Jin, Wei Luo, Jun Xie, and Fei Huang. 2021. Product-oriented machine translation with cross-modal cross-lingual pretraining. *Proceedings of the 29th ACM International Conference on Multimedia*.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. In *AAAI*.

Xiao Wang, Shihan Dou, Li Xiong, Yicheng Zou, Qi Zhang, Tao Gui, Liang Qiao, Zhanzhan Cheng, and Xuanjing Huang. 2022. Miner: Improving out-of-vocabulary named entity recognition from an information theoretic perspective. *ArXiv*, abs/2204.04391.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Benjamin C.M. Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. *ArXiv*, abs/2105.14462.

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

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

Tong Zhang, Wei Ye, Baosong Yang, Long Zhang, Xingzhang Ren, Dayiheng Liu, Jinan Sun, Shikun Zhang, Haibo Zhang, and Wen Zhao. 2021. Frequency-aware contrastive learning for neural machine translation. *ArXiv*, abs/2112.14484.

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

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2022. Region-attentive multimodal neural machine translation. *Neurocomputing*, 476:1–13.