Efficient Object-Level Visual Context Modeling for Multimodal Machine Translation: Masking Irrelevant Objects Helps Grounding

Dexin Wang and Deyi Xiong[*]
College of Intelligence and Computing, Tianjin University, Tianjin, China 300350
{dexinwang, dyxiong}@tju.edu.cn

Abstract

Visual context provides grounding information for multimodal machine translation (MMT). However, previous MMT models and probing studies on visual features suggest that visual information is less explored in MMT as it is often redundant to textual information. In this paper, we propose an **O**bject-level **V**isual **C**ontext modeling framework (OVC) to efficiently capture and explore visual information for multimodal machine translation. With detected objects, the proposed OVC encourages MMT to ground translation on desirable visual objects by masking irrelevant objects in the visual modality. We equip the proposed OVC with an additional object-masking loss to achieve this goal. The object-masking loss is estimated according to the similarity between masked objects and the source texts so as to encourage masking source-irrelevant objects. Additionally, in order to generate vision-consistent target words, we further propose a vision-weighted translation loss for OVC. Experiments on MMT datasets demonstrate that the proposed OVC model outperforms state-of-the-art MMT models and analyses show that masking irrelevant objects helps grounding in MMT.[1]

Introduction

Multimodal Machine Translation aims at translating a sentence paired with an additional modality (e.g. audio modality in spoken language translation or visual modality in image/video-guided translation) into the target language (Elliott et al. 2016), where the additional modality, though closely semantically related to the text, provides an alternative and complementary view to it. By contrast to text-only neural machine translation (NMT), MMT characterizes with the assumption that the additional modality helps improve translation by either grounding the meaning of the text or providing multimodal context information (Lee et al. 2018). Hence, MMT exhibits pronounced reliance on language-vision/speech interaction.[2]

However, effectively integrating visual information and language-vision interaction into machine translation has been regarded as a big challenge (Yang et al. 2020) for years since Multi30K (Elliott et al. 2016) is proposed as a benchmark dataset for MMT. Many previous MMT studies on Multi30K, which exploit complete source texts during both training and inference, have found that visual context is needed only in special cases, e.g., translating sentences with incorrect or ambiguous source words, by both human and machine translation, and is hence marginally beneficial to multimodal machine translation (Lala et al. 2018; Ive, Madhyastha, and Specia 2019).

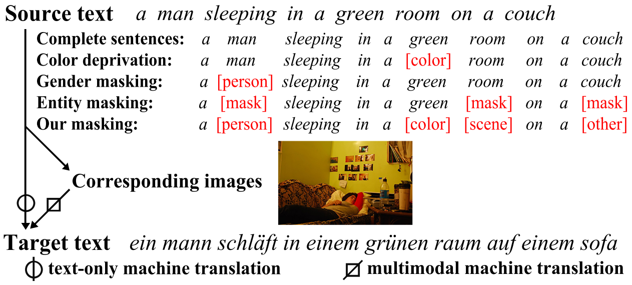
Source text *a man sleeping in a green room on a couch*
Complete sentences: a man sleeping in a green room on a couch
Color deprivation: a man sleeping in a [color] room on a couch
Gender masking: a [person] sleeping in a green room on a couch
Entity masking: a [mask] sleeping in a green [mask] on a [mask]
Our masking: a [person] sleeping in a [color] [scene] on a [other]
Corresponding images
Target text *ein mann schläft in einem grünen raum auf einem sofa*
ϕ text-only machine translation
multimodal machine translation

Figure 1: Word masking in multimodal machine translation.

In this paper, we hypothesize that visual context can be efficiently exploited to enhance MMT, instead of being ignored as a redundant input, from three aspects as follows:

- *Source text processing and encoding*. In most cases, source texts provide sufficient information for translation, which makes visual context redundant. Therefore, weakening the input signal from the textual modality may force MMT to pay more attention to the visual modality.
- *Visual feature learning tailored for translation*. Not all parts in visual images are useful for translation. Learning visual features that are not only linked but also complementary to source texts is desirable for MMT.
- *Target word generation and decoding*. Visual representations can be used to not only initialize the decoder (Zhou et al. 2018) but also guide target word prediction (e.g., rewarding target prediction consistent with visual context).

Regarding the first aspect, we have witnessed that pioneering efforts (Caglayan et al. 2019; Ive, Madhyastha, and Specia 2019), different from previous methods, mask specific words (e.g. gender-neutral words) in source texts, forc-

[*] Corresponding author.

[1] Source code is available at https://github.com/tjunlp-lab/OVC.

[2] In this paper, we focus on multimodal machine translation with both visual and textual modality.

ing MMT to distill visual information into text generation, as shown in Figure 1. In addition to the source text masking, in this paper, we attempt to explore all the three aforementioned aspects in a unified framework for MMT. Specifically, we propose an efficient object-level visual context modeling framework (OVC) to capture desirable visual features and to reward vision-consistent target predictions for MMT. In this framework, we first detect a bag of objects from images. Inspired by the word masking method in source texts (Caglayan et al. 2019), we also encourage OVC to mask visual objects that are not relevant to source texts by computing object-text similarity in a preprocessing step. For this, we propose an object-masking loss that calculates the cross-entropy loss difference between original translation and translations generated with the relevant-object-masked image vs. irrelevant-object-masked image. This is to reward masking irrelevant objects in visual context while masking relevant objects is penalized.

In order to force the decoder to generate vision-consistent target words, we change the traditional cross-entropy translation loss into a vision-weighted loss in OVC, which tends to reward the generation of vision-related words or rare but vision-consistent words.

To examine the effectiveness of the proposed OVC in visual feature learning, we test OVC against the baselines in both standard and source-degradation setting with word masking as shown in Figure 1.

The contributions of this work can be summarized as follows:

- We propose a new approach to MMT, which masks both objects in images and specific words in source texts for better visual feature learning and exploration.
- We propose two additional training objectives to enhance MMT: an object-masking loss to penalize undesirable object masking and a vision-weighted translation loss to guide the decoder to generate vision-consistent words.
- We conduct experiments and in-depth analyses on existing MMT datasets, which demonstrate that our model can outperform or achieve competitive performance against state-of-the-art MMT models.

Related Work

MMT without Text Masking

Since the release of the Multi30K dataset, a variety of different approaches have been proposed for multimodal machine translation. Efforts for the MMT modeling mechanism can be categorized into RNN-based sequence-to-sequence models and attention-based ones. Elliott and Kádár (2017) and Caglayan et al. (2017) employ GRU/LSTM-based encoder-decoder models to encode source texts and integrate a single image vector into the model. The image vector is either used to initialize the encoder or decoder (Zhou et al. 2018; Ive, Madhyastha, and Specia 2019) or to fuse with word embeddings in the embedding layer of the encoder (Caglayan et al. 2017). Attention-based sequence-to-sequence approaches have been proposed for MMT (Huang et al. 2016), which compute either spatially-unaware image-to-texts attention (Zhang et al. 2020) or spatially-aware object-to-text to capture vision-text interaction so as to enhance the encoder and decoder of MMT (Yang et al. 2020).

We also have witnessed two proposed categories for MMT from the perspective of cross-modal learning approaches, which either explicitly transform visual features and textual embeddings from one modality to the other at both training and inference (Caglayan et al. 2017; Yin et al. 2020), or implicitly align the visual and textual modalities to generate vision-aware textual features at training. Unlike the explicit approaches, the implicit cross-modal learning methods do not require images as input at inference, taking the image features as latent variables across different languages (Elliott and Kádár 2017; Calixto, Rios, and Aziz 2019; Hirasawa et al. 2019), which also serves as a latent scheme for unsupervised MMT (Lee et al. 2018). Despite the success of plenty of models on Multi30K, an interesting finding is that the visual modality is not fully exploited and only marginally beneficial to machine translation (Caglayan et al. 2017; Ive, Madhyastha, and Specia 2019).

Text-Masked MMT

To probe the real need for visual context in MMT, several researchers further explore new settings where visual features are not explicitly expressed by source texts on purpose. In other words, specific source words that are linked to visual features are purposely masked. In particular, Ive, Madhyastha, and Specia (2019) focus on three major linguistic phenomena and mask ambiguous, inaccurate and gender-neutral (e.g., player) words in source texts on Multi30K. Their experiment results suggest that the additional visual context is important for addressing these uncertainties. Caglayan et al. (2019) propose more thoroughly masked schemes on Multi30K by applying color deprivation, whole entity masking and progressive masking on source texts. They find that MMT is able to integrate the visual modality when the available visual features are complementary rather than redundant to source texts.

Although masking source words forces MMT models to pay more attention to and therefore exploit the visual modality for translation, there is a big performance gap between the standard setting (without text masking) and source-degradation setting (purposely masking specific words). For example, in the experiments reported by Ive, Madhyastha, and Specia (2019), the best METEOR on WMT 2018 MMT EN-DE test set for the standard setting is 46.5 while the highest METEOR score for the source-degradation setting is only 41.6. Although specific words are masked in source texts, visual features that are semantically linked to these words are available in the visual modality provided for MMT. This indicates that the visual modality is not fully exploited by current MMT models even though the available information is complementary to source texts.

Efficient Object-Level Visual Context Modeling

In this section, we elaborate the proposed OVC model. The backbone of the model is a GRU-based encoder-decoder

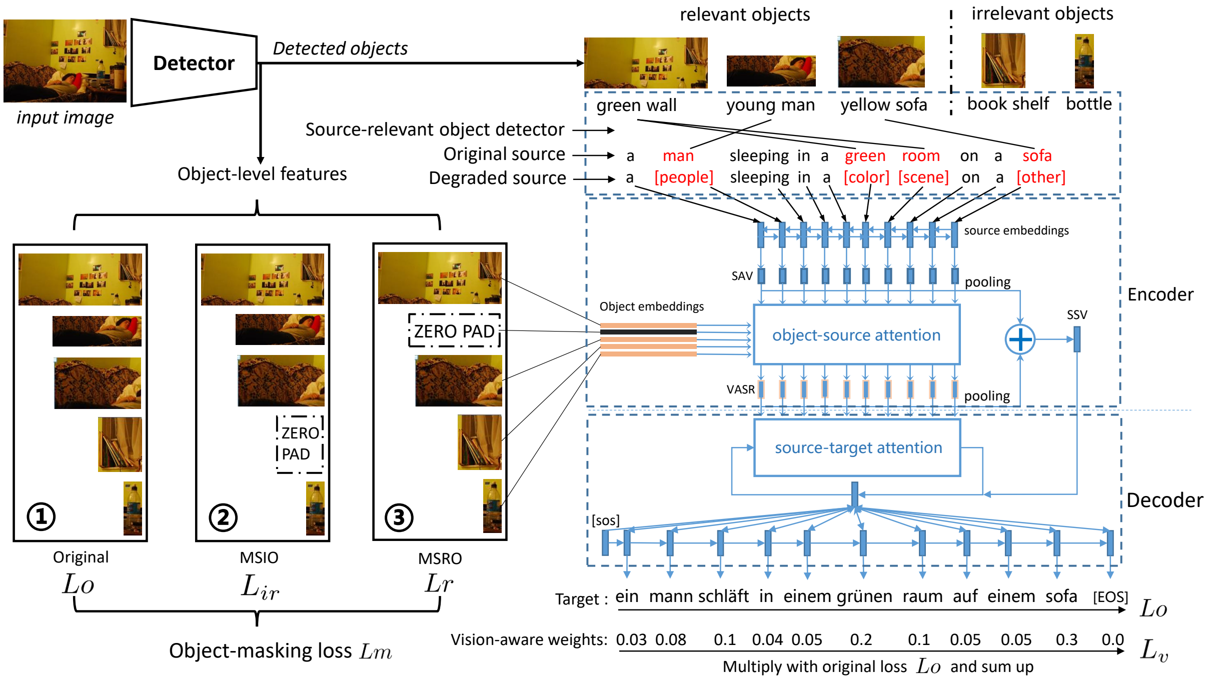

Figure 2: The architecture of the proposed OVC framework. MSIO: masking source-irrelevant objects. MSRO: masking source-relevant objects. SAV denotes source annotation vectors. VASR is the vision-aware source representation of the source sentence.

neural network with two multihead attention layers that model the attention between source tokens and detected objects in the input image as well as the attention between target and source tokens. The architecture of OVC is shown in Figure 2. The source input to OVC can be either an original source sentence or the degradation of the source sentence (see Section 'Experiment' for more details on how we degrade source sentences by masking specific words). The visual modality is integrated into the model through the object-source multihead attention, which is also explored in two additional training objectives: the object-masking loss and the vision-weighted translation loss.

Encoder

The encoder of OVC consists of a bidirectional GRU module and an object-source attention layer that performs the fusion of textual and visual modalities. The inputs to the encoder include token embeddings of source texts and object-level visual features from the paired image. Let $W_s^n = \{w_s^1, w_s^2, ..., w_s^n\}$ denote the token embedding matrix of the source sentence, where n is the number of tokens. The object-level features are a set of vector embeddings of objects detected by a pretrained object detector. Each detected object is labeled with its predicted object category and attribute (e.g., "young man", "green wall"). In our case, we use Resnet101 (He et al. 2016) as the object detector which compresses each object into a 2048-dimension vector. We denote the object embedding matrix as $O^m = \{o^1, o^2, ..., o^m\}$, where m is the number of all detected objects. During training, some objects from the paired image are randomly selected and masked, which we'll discuss in the following subsection in detail. The representation for a masked object is set to a zero vector.

The bidirectional GRU transforms the sequence of source token embeddings into a sequence of annotation vectors (SAV):

$$H_s^n = (h_s^1, h_s^2, ..., h_s^n) \tag{1}$$

We then adopt a multihead attention layer over H_s^n and O^m to obtain a vision-aware source representation (VASR) as follows:

$$\text{VASR} = \text{MultiHead}_1(H_s^n, O^m, O^m) \tag{2}$$

where MultiHead(Q, K, V) is a multihead attention function taking a query matrix Q, a key matrix K, and a value matrix V as inputs. After that, we aggregate VASR and H_s^n into a mixed-modality source sentence vector (SSV) by applying average-pooling (AP) on both VASR and H_s^n to get two separate vectors and then adding the two vectors as follows:

$$\text{SSV} = \text{AP}(\text{VASR}) + \text{AP}(H_s^n) \tag{3}$$

Decoder

The decoder of OVC also consists of a multihead attention layer to compute source-target attention and a GRU module to update hidden states. SSV is fed into the GRU layer to initialize the decoder as follows:

$$H_t^0 = \text{GRU}(w_t^{[\text{sos}]}, \text{SSV}) \tag{4}$$

where $w_t^{[sos]}$ is the embedding of the start symbol. At each time step, the multihead attention layer computes the source-target attention as follows:

$$T^{i+1} = \text{MultiHead}_2(H_t^i, \text{VASR}, \text{VASR}) \tag{5}$$

where H_t^i is the hidden state at i-th time step of the decoder. The GRU module aggregates previous word embedding and T^{i+1} to update the hidden state as follows:

$$H^{i+1} = \text{GRU}(w_t^i, T^{i+1}) \tag{6}$$

where w_t^i denotes the embedding of the i-th target word. Finally, we project H_t into logit vectors for target word prediction over the vocabulary.

Training Objectives

In order to facilitate our model to capture the deep interaction between the textual and visual modalities, in OVC, we propose two additional translation objectives to better integrate visual features into MMT: an object-masking loss and a vision-weighted translation loss.

Object-Masking Loss. The object-masking loss (denoted as L_m) is to optimize MMT to discriminate good grounding of source tokens to the visual modality from bad grounding by telling the model the difference between masking source-relevant objects and masking those irrelevant. If an object is masked, the corresponding o^i is set to a zero vector. Specifically, the goals of using this objective are two-folds:

- forcing the model to penalize masking objects (e.g., the masked object in ③ of Figure 2) on which source words (or tags in degraded source sentences) can be grounded.
- rewarding masking schemes where irrelevant objects (e.g., the masked object in ② of Figure 2) are masked so as to avoid the negative impact from them.

Before we define the object-masking loss, let's discuss how we detect source-relevant objects from those irrelevant. Generally, we compute the degree of the relevance of an object to the source sentence by semantic similarity with the aid of a pretrained language model.[3] In particular, we first compute a cosine similarity matrix (denoted as S^{m*n}) for all possible object-word pairs (w_{op}^i, w_{sp}^j) for each object, where w_{op}^i is the word embedding for the category word of the i-th object, w_{sp}^j is the word embedding for the j-th source token. Both embeddings are from the same pretrained language model. Notice that $W_{sp}^n = \{w_{sp}^1, w_{sp}^2, ..., w_{sp}^n\}$ is different from W_s^n in that the former is from the pretrained language model and only used for source-relevant object detection in the preprocessing step while the latter is initialized randomly and trained with the model. We perform max-pooling over the corresponding row of the similarity matrix S to obtain the similarity score of the object to the entire source sentence. In this way, we collect a vector of similarity scores OSS (object-to-sentence similarity) for all objects as follows:

$$\text{OSS}_i = \max S_{i,1:n}, \;\; i = 1, 2, ..., m \tag{7}$$

We then define an indicator d to indicate whether an object is source-relevant or not as follows:

$$d_i = 1 \;\text{ if }\; \text{OSS}_i > \gamma \;\text{ otherwise }\; 0, \;\; i = 1, 2, ..., m \tag{8}$$

where γ is a predefined similarity threshold hyper-parameter.[4]

With d, we calculate the object-masking loss as follows:

$$L_r = L(O_{\varnothing_i}^m, W_s^n) \;\text{ if }\; d_i = 1 \tag{9}$$

$$L_{ir} = L(O_{\varnothing_i}^m, W_s^n) \;\text{ if }\; d_i = 0 \tag{10}$$

$$L_m = -(L_r - L_o) + (L_{ir} - L_o)^2 \tag{11}$$

where L denotes the cross-entropy translation loss of OVC fed with different visual features, $O_{\varnothing_i}^m$ denotes O^m where the i-th object is masked (i.e, $o_i = \mathbf{0}$), L_o denotes the original cross-entropy loss of OVC where no objects are masked, L_r calculates the new cross-entropy loss if a source-relevant object is masked while L_{ir} is the new loss if a source-irrelevant object is masked. Therefore, minimizing L_m will force the model to reward masking irrelevant objects and penalize masking relevant objects. For each training instance, OVC randomly samples source-irrelevant objects for computing L_{ir} and source-relevant objects for generating L_r. For each masked instance, we make sure that all masked objects are either source-relevant or source-irrelevant. No mixed cases are sampled.

Vision-Weighted Translation Loss. Partially inspired by VIFIDEL (Madhyastha, Wang, and Specia 2019) which checks whether the generated translations are consistent with the visual modality by evaluating the visual fidelity of them, we introduce a vision-weighted translation loss. Similar to OSS, we first compute a target-to-source semantic similarity matrix S'^{r*n} where r is the number of target tokens. In order to allow the model to pay more attention to vision-related tokens[5] in source texts (e.g., "man", "green" in Figure 2), we further set elements that are not vision-related in S' to $\mathbf{0}$. Then we compute a target-to-vision-related-source similarity vector TVS as follows:

$$\text{TVS}_j = \max S'_{j,1:n}, \;\; j = 1, 2, ..., r \tag{12}$$

After that, we calculate a weight for each target word to estimate how much the target word is consistent with the visual modality as follows:

$$q_j = \frac{\text{TVS}_j / f_j}{\sum_{a=1}^{r} \text{TVS}_a / f_a}, \;\; j = 1, 2, ..., r \tag{13}$$

where f_j is the frequency of the j-th token in the training data. f_j is applied to de-bias rare vision-related words. Then the vision-weighted loss L_v can be computed as follows:

$$L_v = \sum_{j=1}^{r} q_j * Lo_j \tag{14}$$

where Lo_j is the cross-entropy loss of the j-th target word. Generally, L_v favors target words that are vision-consistent. Rare words can be encouraged to generate if they are related to the visual modality through the de-biasing factor f_j.

[3] We use multilingual-cased-base-BERT which is only used in the preprocessing step to determine source-relevant objects. The reason of using the multilingual BERT is that object category words and source words may be in two different languages. Other multilingual word embeddings can be used here too.

[4] In our case, we set γ to 0.48 after randomly checking 100 samples in the training data to select a suitable threshold.

[5] Vision-related tokens are marked and provided by Flickr30K-Entities (Plummer et al. 2015).

Overall Objective of OVC. We aggregate the basic translation loss L_o, the object-masking loss L_m and the vision-weighted loss L_v for each sample into three forms as follows:

$$L_{ovc_m} = (Lo + L_r + L_{ir})/3 + \alpha * L_m \tag{15}$$

$$L_{ovc_v} = Lo + \beta * L_v \tag{16}$$

$$L_{ovc_{full}} = (Lo + L_r + L_{ir})/3 + \alpha * L_m + \beta * L_v \tag{17}$$

where α and β are two hyper-parameters to control the two additional training objectives. L_{ovc_m} is the loss objective of OVC with the additional object-masking loss where we optimize the average of three translation losses (Lo, L_r and L_{ir}) instead of Lo itself, otherwise the L_{ir} and L_r terms will be uncontrollable during optimization. L_{ovc_v} is the loss objective of OVC with the additional vision-weighted translation loss while $L_{ovc_{full}}$ is the full-fledged objective of OVC with the two additional losses.

Experiments

In order to evaluate the proposed OVC framework for MMT, we conducted a series of experiments on MMT datasets and compared OVC with state-of-the-art MMT models.

Dataset

We used three datasets:

- Multi30K (Elliott et al. 2016). This is a widely-used benchmark dataset for MMT, which contains English captions for images from Flickr30K (Young et al. 2014) and corresponding translations into German, French and Czech. We conducted experiments with English-to-French (En-Fr) and English-to-German (En-De) and adopted the default split of Multi30K in WMT 2017 MMT shared task, which consists of 29,000 samples for training and 1,014 for validation, and 1,000 for test. We used sentences with subwords preprocessed by the implementation of VAG-NMT. For these splits,[6] the vocabulary contains 8.5K sub-words for English, 9.4K for German and 8.7K for French.
- WMT17 MMT test set (Elliott et al. 2017). This test set contains 1,000 unduplicated images manually selected from 7 different Flickr groups.
- Ambiguous COCO. This is an out-of-domain test set of WMT 2017 with 461 images whose captions are selected to contain ambiguous verbs.

Experiment Settings

Following previous works (Ive, Madhyastha, and Specia 2019; Yin et al. 2020), we evaluated OVC in the following two settings.

- **Standard setting**: For this setting, we retain all words in source texts and feed them as textual input into all MMT models for both training and inference.
- **Source-degradation setting**: In this setting, we mask words in source texts according to Flickr30K-Entities (Plummer et al. 2015), which manually categorizes words in English captions in Multi30K into 9 classes:'people', 'scene', 'clothing', 'instruments', 'animals', 'bodyparts', 'vehicles', 'other' and '`notvisual`'. We did not mask the '`notvisual`' category as words in this category cannot been grounded in the corresponding image. Except for the '`notvisual`' words, we replaced vision-related words with their corresponding category tags. Besides, we replaced color-related words as an identical 'color' category in the remaining source texts, as shown in Figure 1. 20.9% of words (79,622 out of 380,793) in the training set and 21.0% of words (2,818 out of 13,419) in the validation set were masked in this way. As Flickr30K-Entities do not provide tags for the re-sampled images in the WMT17 MMT test set, we only evaluated MMT models on the development set in this experiment setting. We fed all MMT models with masked source texts as textual input during both training and inference.

Baselines

We compared our proposed OVC against 6 different strong baselines:

- Transformer (Vaswani et al. 2017): state-of-the-art neural machine translation architecture with self-attention.
- Imagination (Elliott and Kádár 2017): an RNN-based sequence-to-sequence MMT system which implicitly aligns images and their corresponding source texts.
- VAG-NMT (Zhou et al. 2018): an RNN-/Attention-mixed MMT system using vision-text attention to obtain a vision-aware context representation as the initial state of its decoder.
- VMMT (Calixto, Rios, and Aziz 2019): a GRU-based MMT approach that imposes a constraint on the KL term to explore non-negligible mutual information between inputs and a latent variable.
- GMMT (Yin et al. 2020): a stacked graph-based and transformer-based MMT model using object-level features and a textual graph parser for modeling semantic interactions.
- VAR-MMT (Yang et al. 2020): an attention-based MMT model that employs visual agreement regularization on visual entity attention via additional word aligners.

For fairness, all the models were trained using Multi30K. No extra resource was used. In the standard setting, we compared OVC against these baselines whose performance on the WMT17 MMT test set are directly reported from their corresponding papers. Note that the performance of Transformer is taken from (Yin et al. 2020). For the source-degradation setting, we only compared OVC with different objectives as this is a new setting where no results of existing models are available.

[6] https://drive.google.com/drive/folders/1G645SexvhMsLPJhPAPBjc4FnNF7v3N6w

Models	WMT17 MMT test set				Ambiguous COCO			
	En⇒Fr		En⇒De		En⇒Fr		En⇒De	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing MMT Models								
(T) Transformer‡	52.0	68.0	30.6	50.4	-	-	27.3	46.2
(R) Imagination_i	-	-	30.2	51.2	-	-	26.4	45.8
(R) VAG-NMT_i‡	53.5±0.7	70.0±0.7	31.6±0.5	52.2±0.3	44.6±0.6	64.2±0.5	27.9±0.6	47.8±0.6
(R) VAG-NMT_i	53.8±0.3	70.3±0.5	31.6±0.3	52.2±0.3	45.0±0.4	**64.7±0.4**	28.3±0.6	48.0±0.5
(R) VAR-MMT_o	52.6	69.9	29.3	51.2	-	-	-	
(T) VAR-MMT_o	53.3	70.4	29.5	50.3	-	-	-	
(R) LIUMCVC_i	52.7±0.9	69.5±0.7	30.7±1.0	52.2±0.4	43.5±1.2	63.2±0.9	26.4±0.9	47.4±0.3
(R) VMMT_i	-	-	30.1±0.3	49.9±0.3	-	-	25.5±0.5	44.8±0.2
(T) GMMT_o	53.9	69.3	32.2	51.9	-	-	28.7	47.6
Our Proposed Models								
OVC	53.5±0.2	70.2±0.3	31.7±0.3	51.9±0.4	44.7±0.6	64.1±0.3	28.5±0.5	47.8±0.3
OVC+L_m	54.1±0.7	**70.5±0.5**	32.3±0.6	**52.4±0.3**	**45.3±0.5**	64.6±0.5	**28.9±0.5**	**48.1±0.5**
OVC+L_v	**54.2±0.4**	70.5±0.5	**32.4±0.4**	52.3±0.5	45.2±0.4	64.6±0.3	28.6±0.5	48.0±0.6
OVC+L_m+L_v	54.0±0.4	70.4±0.4	32.4±0.6	52.2±0.3	45.1±0.6	64.5±0.5	28.8±0.4	48.0±0.4

Table 1: Results of standard experiments. ‡ denotes text-only models. _i denotes models using image-level features while _o object-level features. Prefix R denotes RNN-based approaches while T Transformer-based approaches. OVC+L_m, OVC+L_v and OVC+L_m+L_v denote OVC trained with L_{ovc_m}, L_{ovc_v} and $L_{ovc_{full}}$ respectively.

En⇒De		
Metrics	BLEU	METEOR
OVC_t	21.02	40.61
OVC_i	22.02	41.91
OVC_o	21.98	41.57
OVC_o+HM	25.31	43.85
OVC_o+L_m	26.30	45.37
OVC_o+L_v	22.18	42.01
OVC_o+L_m+L_v	22.57	42.24
En⇒Fr		
OVC_t	37.01	55.35
OVC_i	37.40	55.68
OVC_o	36.94	54.92
OVC_o+HM	37.39	55.38
OVC_o+L_m	39.31	57.28
OVC_o+L_v	37.25	55.79
OVC_o+L_m+L_v	37.63	56.14

Table 2: Results for the source-degradation setting on the WMT17 MMT development set. _t denotes text-only models. HM is a hard masking scheme where irrelevant objects are masked in a hard way according to the predefined threshold.

Results in Standard Setting

Model Setting for OVC

In order to avoid the influence of the increasing number of parameters on the comparison, we limited the number of parameters in our OVC models to be comparative to that in (Zhou et al. 2018) (16.0M parameters). In order to achieve this, we set the size of word embeddings in OVC to 256. The encoder of source texts has one bidirectional-GRU layer and one multihead object-text attention layer. The hidden state sizes of all modules in the encoder were set to 512. The decoder has one multihead attention layer and two stacked GRU layers, of which the hidden sizes were set to 512 and the input sizes 256 and 512 for the two GRU layers, respectively. We used Adam as the optimizer with a scheduled learning rate and applied early-stopping with a patient step of 10 during training. With these settings, our proposed OVC of its full form has 11.3M parameters. All models were trained in the teacher-forcing manner. Other settings were kept the same as in (Zhou et al. 2018). All implementations were built based upon Pytorch and models were both trained and evaluated on one 2080Ti GPU. We performed a grid search on the WMT17 MMT development set to obtain the hyper-parameters: α was set to 0.1 and β was set to 0.1.

For image-level visual features, we used the pool5 outputs of a pretrained Resnet-50, released by WMT 2017. For object-level visual features, we first took the pool5 outputs of a pretrained Resnet101 detector[7] as candidates. We then selected objects of the highest 20 object confidences as our object-level features.

To make our experiments more statistically reliable, for the proposed model, we run each experiment for three times and report the average results over the three runs. The results in the standard setting are listed in Table 1. OVC trained with the two additional losses either outperforms existing Transformer-based and RNN-based MMT models with an average improvement of 0.25 BLEU and 0.10 METEOR, or achieves competitive results to them. The basic OVC shows no advantage over existing image-level MMT models. For example, in most cases, the basic OVC is not better than VAG-NMT_i on the WMT17 MMT test set and Ambiguous COCO. We conjecture that the object-level visual features may contain irrelevant information for machine trans-

[7] https://github.com/peteanderson80/bottom-up-attention

lation. And since the Multi30K training data is small and textually repetitive, this makes it hard for object-level MMT models to learn fine-grained grounding alignments. However, after being equipped with the two proposed additional objectives, OVC is superior to both image- and object-level MMT models. It gains an average improvement of 0.4∼0.6 BLEU and 0.3∼0.5 METEOR using the additional L_m, while 0.1∼0.7 BLEU and 0.2∼0.5 METEOR using the additional L_v, which indicate that our proposed objectives enhance the visual grounding capability of OVC. Additionally, we visualize the object-source attention of OVC trained with different objectives in Appendix to support this hypothesis.

Results in Source-Degradation Setting and Ablation Study

In this setting, we compared different OVC variants using different objectives, which is also the ablation study of our proposed OVC. We also trained OVC in a text-only setting by dropping the object-to-source attention layer in its encoder, where VASR is replaced by the annotation vectors and SSV is directly the average-pooling result of the annotation vectors.

The results are shown in Table 2. Under the source-degradation setting, with image-level features, OVC is better than its text-only version, which is consistent with previous multimodal machine translation findings (Caglayan et al. 2019). With object-level features, the performance of OVC is generally worse than that with image-level features and even worse than the text-only OVC on English-to-French translation. This again confirms our finding with the basic OVC under the standard setting. Besides, it can be seen that the improvements of both L_m and L_v in the source-degradation setting are generally larger than those in the standard setting. Particularly, L_m gains an average improvement of 3.35 BLEU and 3.08 METEOR while L_v achieves an average improvement of 0.255 BLEU of 0.655 METEOR over the basic OVC.

For a deep understanding on the impact of object masking, we further compared a hard masking scheme where source-irrelevant objects are compulsively masked in a hard way instead of using the training objective in a soft way according to the predefined similarity threshold. The stable improvement of behavior of OVC_o+HM vs. OVC_o and OVC_o+L_m vs. OVC_o+HM suggest that masking irrelevant objects helps grounding in MMT as vision-related words are all masked in the degraded source sentences. Since the only difference between L_m and HM is that L_m penalizes masking source-relevant objects and encourages masking source-irrelevant objects simultaneously in a soft way, the improvements of L_m over HM indicate that the proposed object-masking loss is a more efficient way for grounding in MMT.

Results in Mixed Setting

Finally, we trained MMT models in a mixed setting where source-degradation and standard texts are mixed together for training and evaluation is done on the source-degradation data. Specifically, we trained OVC with the source-degradation & standard mixed training set of Multi30K and evaluated it on the source-degradation samples of the WMT17 MMT En⇒De development set to investigate the potential ability of the source-degraded framework in helping standard MMT. The results are shown in Table 3 with different proportions of mixed standard samples and degraded samples.

ST:SD	BLEU	METEOR
1.0:0.0	22.43	41.64
1.0:0.2	22.66	41.53
1.0:0.4	**23.01**	**42.08**
1.0:0.5	22.75	41.73
1.0:0.6	22.68	41.68
1.0:0.8	22.82	42.00
1.0:1.0	22.05	41.03

Table 3: Results under the mixed setting on the WMT17 MMT En⇒De development set. ST:SD denotes the ratio between the number of standard samples and the number of source-degradation samples in the mixed setting.

It is interesting to find that the performance of OVC does not consistently rise as the number of sampled source-degradation samples increase. The best proportion of additional source-degradation data is 1.0:0.4. We assume that a certain amount of source-degradation samples can improve the grounding ability of MMT models, which offsets the information loss in source-degradation samples. However, more source-degradation sample may undermine the ability of MMT in conveying the meaning of source sentences to target translations.

Conclusion and Future Work

In this paper, to efficiently model the language-vision interaction and integrate visual context into multimodal machine translation, we have presented OVC, an object-level visual context modeling framework. In OVC, we model the interaction between the textual and visual modality through the object-text similarity and object-source multihead attention on the source side as well as the vision-weighted loss on the target side. In order to tailor the visual feature learning for multimodal machine translation, the additional object-masking loss is proposed to force OVC to be aware of whether the masked objects are relevant to source texts and to perform desirable masking in a soft way. The presented vision-weighted translation loss is to guide the decoder to generate vision-consistent target words. Experiment results show that our proposed framework achieves competitive performance against several existing state-of-the-art MMT models in the standard setting. Experiments and analyses on the source-degradation settings suggest that the proposed two additional training objectives, especially the object-masking loss, helps grounding in MMT.

In the future, we plan to improve the proposed OVC in grounding via other mechanisms (e.g., cross-modality pretraining). And we are also interested in extending our OVC framework to the video-guided MMT (Wang et al. 2019).

Acknowledgments

The present research was supported by the National Key Research and Development Program of China (Grant No. 2019QY1802). We would like to thank the anonymous reviewers for their insightful comments.

Appendix

Due to the page limit, please refer to our arxiv version[8] for details on the visualization of source-object attention and case study.

References

Caglayan, O.; Aransa, W.; Bardet, A.; García-Martínez, M.; Bougares, F.; Barrault, L.; Masana, M.; Herranz, L.; and van de Weijer, J. 2017. LIUM-CVC Submissions for WMT17 Multimodal Translation Task. In *Proceedings of the Second Conference on Machine Translation*, 432–439. Copenhagen, Denmark: Association for Computational Linguistics. doi:10.18653/v1/W17-4746. URL https://www.aclweb.org/anthology/W17-4746.

Caglayan, O.; Madhyastha, P.; Specia, L.; and Barrault, L. 2019. Probing the Need for Visual Context in Multimodal Machine Translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4159–4170. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-1422. URL https://www.aclweb.org/anthology/N19-1422.

Calixto, I.; Rios, M.; and Aziz, W. 2019. Latent Variable Model for Multi-modal Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6392–6405. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1642. URL https://www.aclweb.org/anthology/P19-1642.

Elliott, D.; Frank, S.; Barrault, L.; Bougares, F.; and Specia, L. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, 215–233. Copenhagen, Denmark: Association for Computational Linguistics. doi:10.18653/v1/W17-4718. URL https://www.aclweb.org/anthology/W17-4718.

Elliott, D.; Frank, S.; Sima'an, K.; and Specia, L. 2016. Multi30K: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, 70–74. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/W16-3210. URL https://www.aclweb.org/anthology/W16-3210.

Elliott, D.; and Kádár, Á. 2017. Imagination Improves Multimodal Translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 130–141. Taipei, Taiwan: Asian Federation of Natural Language Processing. URL https://www.aclweb.org/anthology/I17-1014.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hirasawa, T.; Yamagishi, H.; Matsumura, Y.; and Komachi, M. 2019. Multimodal Machine Translation with Embedding Prediction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, 86–91. Minneapolis, Minnesota: Association for Computational Linguistics. doi:10.18653/v1/N19-3012. URL https://www.aclweb.org/anthology/N19-3012.

Huang, P.-Y.; Liu, F.; Shiang, S.-R.; Oh, J.; and Dyer, C. 2016. Attention-based Multimodal Neural Machine Translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, 639–645. Berlin, Germany: Association for Computational Linguistics. doi:10.18653/v1/W16-2360. URL https://www.aclweb.org/anthology/W16-2360.

Ive, J.; Madhyastha, P.; and Specia, L. 2019. Distilling Translations with Visual Awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6525–6538. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1653. URL https://www.aclweb.org/anthology/P19-1653.

Lala, C.; Madhyastha, P. S.; Scarton, C.; and Specia, L. 2018. Sheffield Submissions for WMT18 Multimodal Translation Shared Task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, 624–631. Belgium, Brussels: Association for Computational Linguistics. doi:10.18653/v1/W18-6442. URL https://www.aclweb.org/anthology/W18-6442.

Lee, J.; Cho, K.; Weston, J.; and Kiela, D. 2018. Emergent Translation in Multi-Agent Communication. In *Proceedings of the International Conference on Learning Representations*.

Madhyastha, P.; Wang, J.; and Specia, L. 2019. VIFIDEL: Evaluating the Visual Fidelity of Image Descriptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 6539–6550. Florence, Italy: Association for Computational Linguistics. doi:10.18653/v1/P19-1654. URL https://www.aclweb.org/anthology/P19-1654.

Plummer, B. A.; Wang, L.; Cervantes, C. M.; Caicedo, J. C.; Hockenmaier, J.; and Lazebnik, S. 2015. Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2641–2649.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. *ArXiv* abs/1706.03762.

Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.; and Wang, W. Y. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* 4580–4590.

[8] https://arxiv.org/pdf/2101.05208.pdf

Yang, P.; Chen, B.; Zhang, P.; and Sun, X. 2020. Visual Agreement Regularized Training for Multi-Modal Machine Translation. In *AAAI*, 9418–9425.

Yin, Y.; Meng, F.; Su, J.; Zhou, C.; Yang, Z.; Zhou, J.; and Luo, J. 2020. A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3025–3035. Online: Association for Computational Linguistics. doi:10.18653/v1/2020.acl-main.273. URL https://www.aclweb.org/anthology/2020.acl-main.273.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2: 67–78. doi:10.1162/tacl_a_00166. URL https://www.aclweb.org/anthology/Q14-1006.

Zhang, Z.; Chen, K.; Wang, R.; Utiyama, M.; Sumita, E.; Li, Z.; and Zhao, H. 2020. Neural Machine Translation with Universal Visual Representation. In *ICLR*.

Zhou, M.; Cheng, R.; Lee, Y. J.; and Yu, Z. 2018. A Visual Attention Grounding Neural Model for Multimodal Machine Translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 3643–3653. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/D18-1400. URL https://www.aclweb.org/anthology/D18-1400.