Entity-level Cross-modal Learning Improves Multi-modal Machine Translation

Xin Huang[1,2], **Jiajun Zhang**[1,2] and **Chengqing Zong**[1,2,3]
[1] National Laboratory of Pattern Recognition, Institute of Automation, CAS, Beijing, China
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] CAS Center for Excellence in Brain Science and Intelligence Technology, Beijing, China
{xin.huang, jjzhang, cqzong}@nlpr.ia.ac.cn

Abstract

Multi-modal machine translation (MMT) aims at improving translation performance by incorporating visual information. Most of the studies leverage the visual information through integrating the global image features as auxiliary input or decoding by attending to relevant local regions of the image. However, this kind of usage of visual information makes it difficult to figure out how the visual modality helps and why it works. Inspired by the findings of (Caglayan et al., 2019) that entities are most informative in the image, we propose an explicit entity-level cross-modal learning approach that aims to augment the entity representation. Specifically, the approach is framed as a reconstruction task that reconstructs the original textural input from multi-modal input in which entities are replaced with visual features. Then, a multi-task framework is employed to combine the translation task and the reconstruction task to make full use of cross-modal entity representation learning. The extensive experiments demonstrate that our approach can achieve comparable or even better performance than state-of-the-art models. Furthermore, our in-depth analysis shows how visual information improves translation.

1 Introduction

Multi-modal machine translation (MMT) aims at improving the translation performance with the help of visual information such as image (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018; Zhang et al., 2020). The assumption behind this is that images consist of relatively complete information compared with textual description and can provide complementary knowledge to guide translation (Elliott et al., 2016).

Previous studies mainly focus on integrating the visual information into neural machine translation as a global feature or as attention-based local features. Benefiting from the similar representation between visual features (He et al., 2016) and textual hidden states (Bahdanau et al., 2015), several attempts have been made to incorporate image features as an auxiliary input to exploit its global semantics (Calixto and Liu, 2017; Elliott and Kádár, 2017; Zhou et al., 2018). Some works leverage the spatial information in the decoding stage by attending to relevant local regions of the image (Calixto et al., 2017; Caglayan et al., 2017, 2018; Libovický and Helcl, 2017; Libovický et al., 2018; Yao and Wan, 2020; Ive et al., 2019).

However, these sentence-level approaches which implicitly incorporate image features make it extremely difficult to figure out how visual features affect the representation of source-side sentences or the decision when generating a target-side word. Furthermore, results from (Elliott, 2018) have shown that visual information maybe not the reason why MMT models were promoted, and it is observed that irrelevant images can improve translation unexpectedly.

Inspired by the findings of (Caglayan et al., 2019) that entities are most informative in the image, we propose an entity-level cross-modal learning approach for multi-modal machine translation (EMMT). Different from sentence-level cross-modal semantics fusion approaches, our approach aims to augment the entity representation explicitly. We frame the entity-level cross-modal learning approach as a reconstruction task that reconstructs the original textual sentence from a degraded multi-modal input (Lewis et al., 2020). The multi-modal input is a mixture of a degraded sentence and related visual objects. The degraded sentence is generated by erasing the visually depictable entity words as done by (Caglayan et al., 2019) and filling the erased position with corresponding visual objects. Reconstructed from this kind of input, entity words are learned in a cross-modal way. Then, a multi-task framework is employed to combine the translation task and the reconstruction task. Thanks

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1067–1080
November 7–11, 2021. ©2021 Association for Computational Linguistics

to the shared parameters from the reconstruction model, the translation could make full use of cross-modal entity representation learning and significant gains are obtained.

We further take an in-depth analysis to figure out why the approach works by contrasting the translation correctness of entity words with several MMT models. The results show that the translation accuracy of entity words significantly increases with the help of visual information.

The major contributions of our work are listed as follows:

- We propose an entity-level cross-modal learning approach that explicitly enhances the entity representation.
- We present a multi-task method to implicitly make full use of visually enhanced entity representation to improve text translation.
- Our approach significantly improves the translation performance compared with strong baselines and performs on par with or outperforms the state-of-the-art methods. The in-depth analysis demonstrates why visual modality helps obtain better translations and contributes to a better understanding of multi-modal machine translation.

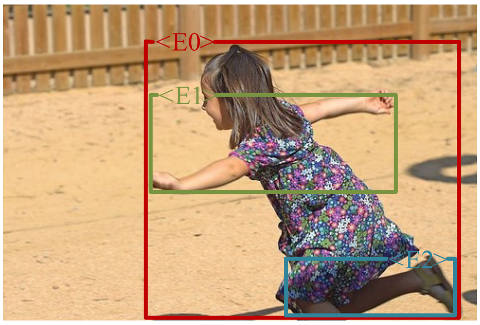
2 Our Approach

In this section, we first introduce how our entity-level cross-modal learning approach explicitly incorporates visual information into entity words. Then, we frame this approach as a reconstruction task and combine it with the translation task in a multi-task framework. Finally, we provide three parameter-sharing schemes to fully exploit the advantage of this multi-task learning approach.

2.1 Explicit Entity-level Cross-modal Fusion

As the cross-modal learning method is applied to the entity words, we define the linguistic entity in two granularities: `phrase entity` and `word entity`.

Phrase Entity A `phrase entity` is a visually depictable phrase which is a full description of a visual object image. For example, in Figure 1 the person in the red bounding box is described as "A girl" in the sentences X_0 and X_2. The "girl" is the object itself. "A" is an adjunct word that quantifies the "girl". Both words are meaningful components to describe a visual object image.



$\boldsymbol{X_0}$: A girl in a flower dress is running on sand.

$\boldsymbol{X_1}$: The young girl is standing on one leg.

$\boldsymbol{X_2}$: A girl running with outstretched arms.

$\boldsymbol{X_{2,MMw}}$: A <E0> running with outstretched <E1>.

$\boldsymbol{X_{2,MMp}}$: <E0> <E0> running with <E1> <E1>.

Figure 1: An example of a described picture with captions from three people. It shows how we replace the entity words with visual objects. The noun phrases parsed by a NLP toolkit are marked with colors. The replaced words are marked by "⟨E0⟩ " and "⟨E1⟩ ".

Word Entity A `word entity` is the nouns in a `phrase entity`. For different people, the visual object image could be described from any aspect. As shown in Figure 1, the visual object "⟨E0⟩ " is described as "The young girl" in X_1 which is different from X_0 and X_2. To eliminate the influence of different adjunct words, we only take the nouns as the entity words.

Explicit Multi-modal Input Fusion As there exist two kinds of linguistic entities, we set two replacement rules to the explicit cross-modal fusion method: the phrase-level replacement and the word-level replacement. The phrase-level replacement rule erases all words in the `phrase entity` and fills the positions with visual object images. For example, in Figure 1, X_2 is the original sentence in which "A girl" corresponds with the visual object marked with "⟨E0⟩ ". In its degraded version $X_{2,MMp}$, both "A" and "girl" are erased and replaced with entity "⟨E0⟩ ".

The word-level replacement rule works similarly to the phrase-level. As illustrated in Figure 1, only the `word entity` "girl" and "arms" are erased and replaced. The final input is a mixture of a degraded sentence with several visual object images.

2.2 Cross-modal Learning as Reconstruction

To fully exploit information from both modalities for entity words, we frame the entity-level cross-modal learning approach as a reconstruction task.

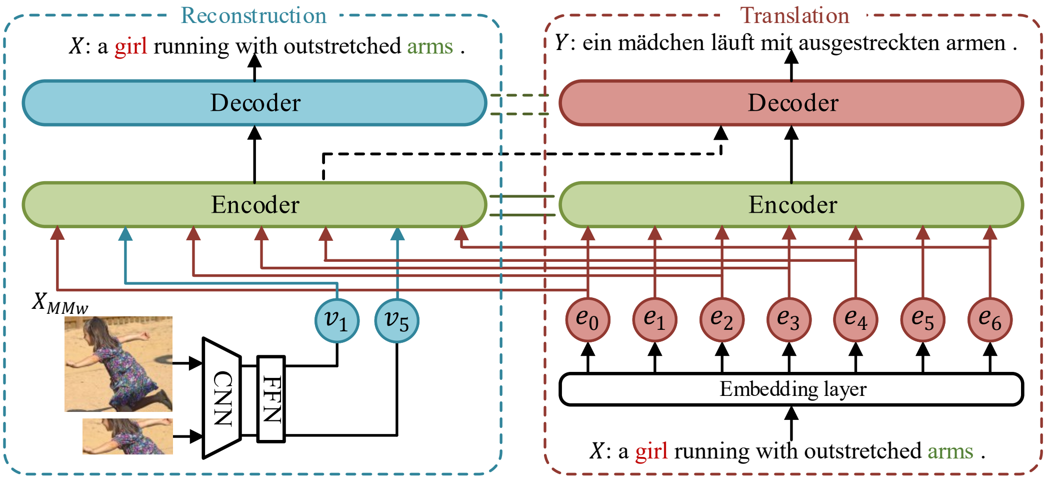

Reconstruction
X: a girl running with outstretched arms .
Decoder
Encoder
X_{MMw}
v_1
v_5
CNN
FFN
Translation
Y: ein mädchen läuft mit ausgestreckten armen .
Decoder
Encoder
e_0
e_1
e_2
e_3
e_4
e_5
e_6
Embedding layer
X: a girl running with outstretched arms .

Figure 2: The EMMT model learns a better representation of the entity word by reconstructing source text from a degraded multi-modal input. The encoders with the same color are parameter-shared. The two green-dashed lines between decoders indicate that we can also share the decoder parameters by merge the source and target vocabularies. The black dashed arrow represents that the reconstructing target language text is also feasible.

As shown in Figure 2, the left model named "Reconstruction" is in a sequence-to-sequence learning framework. The multi-modal sequence input X_{MMw} is a mixture of an entity-level degraded textual sentence and several visual object images. The vector representations of the sequence $\{e_0, v_1, e_2, e_3, e_4, v_5, e_6\}$ are from different feature spaces. The task reconstructs the original textual sentences $X = \{x_0, x_1, \ldots, x_N\}$ from the degraded multi-modal sequences X_{MMw}. The model can learn entity word information from both the visual feature space in the encoding stage and the linguistic feature space in the decoding stage. The reconstruction model is trained to minimise the negative log-likelihood function:

$$L_R(\theta, \psi) = -\sum_{i}^{N} log\ p(x_i | x_{<i}, X_{MM}) \qquad (1)$$

where X_{MM} is X_{MMw} or X_{MMp}, θ is the parameters of the shared encoder, and ψ is the parameters of the reconstruction decoder.

We also consider reconstructing the target language text $Y = \{y_0, y_1, \ldots, y_M\}$. As shown in Figure 2, the black dash line points to decoder of the translation model. To reconstruct the target text Y, we modify the reconstruction objective function to:

$$L_R(\theta, \psi) = -\sum_{j}^{M} log\ p(y_j | y_{<j}, X_{MM}) \qquad (2)$$

2.3 Multi-task Framework

As illustrated in Figure 2, the architecture of the reconstruction model is basically the same as the translation model. The objective function of translation model is also similar to $L_R(\theta, \psi)$:

$$L_T(\theta, \varphi) = -\sum_{i}^{N} log\ p(y_i | y_{<i}, X) \qquad (3)$$

where φ is the decoder parameters of the translation model. To combine the reconstruction task with the translation task, we mix their objective function with the parameter w (Elliott and Kádár, 2017):

$$L(\theta, \varphi, \psi) = wL_T(\theta, \varphi) + (1 - w)L_R(\theta, \psi) \quad (4)$$

where w is the probability of updating translation model parameters in current minibatch. For the reconstruction task, its probability is $1 - w$.

2.4 Parameter Sharing Schemes

As described in previous sections and illustrated in Figure 2, with the help of shared parameters from the reconstruction model, the translation model could make full use of cross-modal entity representation learning and obtain significant gains. Benefiting from the similar design in the model architecture of the two tasks, we investigate two reconstruction directions which are introduced in subsection 2.2 and design three parameter sharing schemes as follows.

Reconstruct Source Text with Respective Decoders Among all parameter sharing schemes, the encoder parameters are shared between the reconstruction model and the translation model. The decoder parameters are optional. In this scheme, we utilize respective decoders in two models which means that the decoder parameters are exclusive to each model. The joint objective is Equation 4. We use the identifier “SR” to refer to this scheme in our experiments.

Reconstruct Source Text with Shared Decoder By merging the source-side and the target-side vocabularies, and sharing embedding layers between the encoder and the decoder, the parameter-shared decoder for both reconstruction and translation is feasible. To distinguish reconstruction from translation, we provide an additional language identification token as the first output word during decoding.[1] With this setting, ψ is the same as φ which means all parameters are shared between the translation model and the reconstruction model. Therefore, we adjust the objective function as:

$$L(\theta, \varphi) = wL_T(\theta, \varphi) + (1 - w)L_R(\theta, \varphi) \quad (5)$$

We use the identifier “SS” to refer to this scheme.

Reconstruct Target Text with Shared Decoder Unlike source text reconstruction, the decoder parameters can be shared easily if the output is in the same target language. Within this scheme, the reconstruction task works more like a multi-modal translation task, as shown by the black dashed line arrow in Figure 2. The objective function of multi-task learning is the same with Equation 5. We use the identifier “T” to refer to this scheme.

3 Experimental Setup

We test our approach on both RNN-based and Transformer-based models and carry out experiments on English to German (En → De) translation task.

Dataset We test our approach on the Multi30K dataset (Elliott et al., 2016) in which each image is paired with one English description and one translated German description. Multi30k was split into three parts: training, validation, and test, containing 29,000, 1,014, and 1,000 pairs of sentences respectively. We also evaluate our model in the Multi30k 2017 test set and the ambiguous MSCOCO test set which contains 1,000 and 461 pairs of sentences respectively. To figure out the upper bound of our approaches, we also incorporate the ground truth bounding boxes of entities. It is reached by using Flickr30K Entities dataset (Plummer et al., 2015, 2017) which was built from Flickr30K (Young et al., 2014).

Entity Extraction To extract visually depictable phrases and detect the corresponding visual objects, we apply an approach similar to the work (Yin et al., 2020). First, we employ an advanced natural language processing toolkit `spaCy` to extract noun phrases in the source-side sentences. For word-level replacement mentioned in subsection 2.1, we keep the nouns in a phrase as the entity word. This affects 32.6% of the words in both the training and the test set for word-level replacement and 45.1% for phrase-level replacement. We measure the medians of entity word frequency in the word-level and the phrase-level replacement. Both of the medians are 2 which means most of the entity words are low frequency. Then, we employ the visual grounding toolkit released by Yang et al. (2019) to detect the visual objects which are related to the extracted noun phrases. Theoretically, only visually depictable phrases are detectable. The ground truth visual bounding boxes and the entity phrases are given by Flickr30K Entities dataset (Plummer et al., 2015, 2017). Finally, we apply the ResNet-50 (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015) to extract 2,048D global features for the visual objects.

RNN-based Model For RNN-based models, the baseline model is an encoder-decoder-based neural machine translation model with attention (Luong et al., 2015). The encoder is a single layer 500D bidirectional RNN with LSTM (Hochreiter and Schmidhuber, 1997), both decoders in the reconstruction model and the translation model are single layer 500D LSTMs, and the embedding layers are 500D. The dropout is set to 0.3 for the encoder, the decoder, and the attention layer. All model parameters are initialized sampling from a uniform distribution $u(-0.1, +0.1)$ and bias vectors are set to 0. The RNN-based models are trained with the Adam optimizer with an initial learning rate of 0.002. We set the minibatch size to 40. Models are selected based on BLEU4 (Papineni et al., 2002) results of the translation task on the validation data.

[1] In our experiments, we use “$\langle$en_sos$\rangle$ ” as the language identification token for English reconstruction decoding and “$\langle$de_sos$\rangle$ ” for German translation decoding.

	RNN-based	Test2016		Test2017		MSCOCO	
		BLEU	**METEOR**	**BLEU**	**METEOR**	**BLEU**	**METEOR**
	NMT	35.9 (0.1)	54.9 (0.1)	28.8 (0.6)	49.5 (0.2)	25.9 (1.0)	45.7 (0.7)
	pRCNNs (Huang et al., 2016)	36.5 (0.8)	54.1 (0.7)	-	-	-	-
	DATT (Calixto et al., 2017)	36.5	55.0	-	-	-	-
	Imagination (Elliott and Kádár, 2017)	36.8 (0.8)	55.8 (0.4)	-	-	-	-
	VMMT$_{\mathrm{C}}$ (Calixto et al., 2019)	37.5 (0.3)	55.7 (0.1)	26.1 (6.6)	45.4 (7.3)	21.8 (5.6)	41.2 (6.3)
	VMMT$_{\mathrm{F}}$ (Calixto et al., 2019)	37.7 (0.4)	56.0 (0.1)	30.0 (0.3)	49.9 (0.3)	25.5 (0.5)	44.8 (0.2)
word	EMMT$_{\mathrm{SR}}$	37.8 (0.2)	56.1 (0.2)	30.1 (0.7)	**50.3** (0.1)	**27.0** (0.1)	**46.4** (0.2)
	EMMT$_{\mathrm{SS}}$	**38.0** (0.5)	56.2 (0.2)	30.3 (0.5)	50.1 (0.1)	26.1 (0.7)	45.6 (0.7)
	EMMT$_{\mathrm{T}}$	36.3 (0.5)	55.0 (0.1)	28.4 (0.1)	48.6 (0.2)	25.3 (0.1)	44.3 (0.4)
phrase	EMMT$_{\mathrm{SR}}$	**38.0** (0.1)	**56.5** (0.3)	30.2 (0.8)	**50.3** (0.4)	26.8 (0.5)	46.1 (0.6)
	EMMT$_{\mathrm{SS}}$	37.8 (0.1)	56.1 (0.2)	**30.5** (0.5)	50.1 (0.3)	26.0 (0.1)	45.5 (0.4)
	EMMT$_{\mathrm{T}}$	36.8 (0.1)	55.0 (0.4)	29.4 (0.2)	49.0 (0.1)	26.3 (0.6)	45.3 (0.7)

Table 1: Experiment results of RNN-based EMMT on the Multi30K 2016/2017 test set and the Ambiguous MSCOCO 2017 test set. For each model, we report the mean and the standard deviation over 3 independent runs. Best overall results are bold.

The training procedure is halted if the model does not improve BLEU4 scores on the validation set for 10 epochs. We translate test data on the last saved model.

Transformer-based Model For Transformer-based models, we set it up with a 128D word embedding layer and 256D hidden size. The embedding layer is shared between source and target vocabularies. Both the encoder and the decoder have $L_d = 4$ layers, and the number of heads is 4. We set the dropout to 0.2 which gets a similar baseline model with (Yin et al., 2020). Adam optimizer is applied in the same way with the original transformer model (Vaswani et al., 2017). Each training batch contained 2,000 source tokens and corresponding target sentences and images. The training was halted after 80,000 steps. All above Transformer-based settings are basically the same as the set up in the publication of (Yin et al., 2020) which we will compare with.

Other Settings We train our models by randomly selecting from the translation task and the reconstruction task. The parameter w is the probability of updating the translation model in the current minibatch. It is set according to the ratio of the amount of data used in the translation task and the reconstruction task. For the Multi30K dataset, we set 0.5 to keep the balance between two tasks. we report mean and standard deviation over 3 independent runs for all models. Finally, we evaluate translation quality using the metrics of BLEU4 (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014).

4 Experimental Results

4.1 Baselines

We compare the proposed models against the following MMT systems. RNN-based models:

- **NMT:** It is the text-only RNN-based attentional NMT system (Luong et al., 2015) with default setting.
- **pRCNNs** (Huang et al., 2016): Visual objects are respectively encoded with the source sentence. In the decoding phase, the decoder chooses to attend mostly to the relevant words in the sequence encoded with the relevant visual object.
- **DATT** (Calixto et al., 2017): It is an NMT model with a doubly attentive decoder. One of the decoders attends to the relevant region of the image to help to predict a word.
- **Imagination** (Elliott and Kádár, 2017): It is an NMT model with an auxiliary task that imagines the image from the source sentence description.
- **VMMT** (Calixto et al., 2019): The VMMT$_{\mathrm{C}}$ and VMMT$_{\mathrm{F}}$ are latent variable models that interact between visual and textual features.

Transformer-based models:

- **Transformer** (Vaswani et al., 2017): It is the text-only Transformer system with default setting in section 3.

	Transformer-based	Test2016		Test2017		MSCOCO	
		BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
	Transformer	38.5 (0.7)	57.5 (0.3)	31.0 (1.0)	51.9 (0.4)	27.5 (0.6)	47.4 (0.1)
	DelMMT (Ive et al., 2019)	38.0	55.6	-	-	-	-
	MMT-TF (Yao and Wan, 2020)	38.7	55.7	-	-	-	-
	GAMMT (Liu et al., 2021)	39.2	**57.8**	31.4	51.2	26.9	46.0
	GMMT (Yin et al., 2020)	**39.8**	57.6	32.2	51.9	28.7	**47.6**
word	$\mathrm{EMMT_{SR}}$	39.7 (0.3)	57.5 (0.1)	**32.9** (0.2)	51.7 (0.4)	**29.1** (0.5)	47.5 (0.2)
	$\mathrm{EMMT_{SS}}$	39.4 (0.6)	**57.8** (0.5)	32.4 (0.4)	**52.1** (0.3)	28.3 (0.7)	47.5 (0.4)
	$\mathrm{EMMT_{T}}$	38.7 (0.3)	56.2 (0.5)	31.0 (0.4)	49.6 (0.7)	26.5 (0.7)	44.9 (0.3)
phrase	$\mathrm{EMMT_{SR}}$	39.3 (0.3)	57.4 (0.7)	32.7 (0.9)	51.8 (0.4)	28.7 (0.9)	47.5 (0.6)
	$\mathrm{EMMT_{SS}}$	39.0 (0.7)	57.3 (0.5)	32.4 (0.7)	51.6 (0.4)	28.3 (0.2)	47.2 (0.0)
	$\mathrm{EMMT_{T}}$	38.5 (0.7)	56.2 (0.2)	30.5 (0.7)	49.7 (0.1)	26.8 (0.8)	45.6 (0.5)

Table 2: Experiment results of Transformer-based EMMT.

	Model	RNN-based				Transformer-based			
		BLEU		METEOR		BLEU		METEOR	
word	$\mathrm{EMMT_{SR}}$	**38.0** (0.1)	↑ 0.2	56.1 (0.4)	- 0.0	**39.9** (0.5)	↑ 0.2	**58.0** (0.3)	↑ 0.3
	$\mathrm{EMMT_{SS}}$	38.0 (0.0)	- 0.0	55.9 (0.2)	↓ 0.3	**39.5** (0.6)	↑ 0.1	57.2 (0.3)	↓ 0.6
	$\mathrm{EMMT_{T}}$	**36.7** (0.2)	↑ 0.4	**55.5** (0.4)	↑ 0.5	38.0 (0.3)	↓ 0.7	**56.9** (0.4)	↑ 0.7
phrase	$\mathrm{EMMT_{SR}}$	**38.1** (0.7)	↑ 0.1	**56.6** (0.3)	↑ 0.1	**39.4** (0.1)	↑ 0.1	57.3 (0.2)	↓ 0.1
	$\mathrm{EMMT_{SS}}$	37.8 (0.3)	- 0.0	**56.2** (0.4)	↑ 0.1	**39.3** (0.1)	↑ 0.3	57.1 (0.1)	↓ 0.2
	$\mathrm{EMMT_{T}}$	**36.9** (0.2)	↑ 0.1	**55.3** (0.4)	↑ 0.3	**38.8** (0.6)	↑ 0.3	**56.6** (0.5)	↑ 0.4

Table 3: Results of applying ground truth bounding boxes for visual objects which are provided by Flickr30K Entities. The bolded results exceed the results of applying detected bounding boxes which are reported in Table 1 and Table 2. We highlight in green/red the improvement.

- **DelMMT** (Ive et al., 2019): The images are applied in the second decoding stage that refines translations from the first drafts with the help of visual information.
- **MMT-TF** (Yao and Wan, 2020): This work designed a multi-modal self-attention that links the source sentence representations with the image feature sequence as the query in the self-attention.
- **GAMMT** (Liu et al., 2021): A Gumbel-attention was proposed to integrate visual information by the Gumbel-Attention score matrix which selects the text-related parts of the image features.
- **GMMT** (Yin et al., 2020): A graph-based and transformer-based multi-modal encoder takes the object-level image features and source sentences as graph inputs.

4.2 Results on the En→De Translation Task

As introduced in subsection 2.1 and 2.4, we have two entity replacement rules and three parameter sharing schemes in total to set up our models. We use “word/phrase” as the identifier to mark whether we apply the word-level replacement or the phrase-level replacement to the text degradation. The identifiers “SR/SS/T” are parameter sharing schemes introduced in subsection 2.4. For example, if we apply a RNN-based EMMT to reconstruct source sentences from a degraded multi-modal input in which its phrases are replaced by visual objects, and use respective decoders for two models, the model should be named as $\mathrm{EMMT_{SR}}$ and be displayed in the “phrase” rows.

Results of RNN-based Models Table 1 shows the main results of our RNN-based models on the En→De translation task. We compare our models with five RNN-based MMT models. Most of our models outperform the best RNN-based MMT model $\mathrm{VMMT_{F}}$ and achieve great improvement compared with the text-only baseline model.

Results of Transformer-based Models Table 2 shows the main results of our Transformer-based models. We compare our models with 4 transformer-based MMT models in which GMMT (Yin et al., 2020) is the state-of-the-art MMT model.

Our best model is comparable with or superior to GMMT. Note that GMMT is graph-based model with more complicated than ours and our approach has another advantage that it does not rely on image during test inference. Similar to the results of RNN-based models, models with target sentence reconstruction direction are not able to reach up to best results. We speculate that it is because reconstructing the target language text from multi-modal input is much more difficult than reconstructing the original source language text.

Results on Gold Flickr30K Entities Table 3 shows the results on Flickr30K Entities data set. Most of the models outperform the models applying detected visual objects. These results also suggest that models trained on the detected visual objects approximate the models trained on the ground truth visual objects.

Overall, the results displayed in Table 1 to Table 3 suggest that the word-level replacement and reconstructing source text with the respective decoders are the best settings of our approach.

4.3 Adversarial Evaluation and Ablation Study

RNN-based		BLEU			
		vo	ro	rw	mlm
word	EMMT_{SR}	**37.8**	37.4	37.1	36.8
	EMMT_{SS}	**38.0**	37.8	37.6	37.6
	EMMT_{T}	**36.3**	**36.3**	35.0	35.6
phrase	EMMT_{SR}	**38.0**	37.2	37.3	36.9
	EMMT_{SS}	**37.8**	37.5	37.6	37.6
	EMMT_{T}	**36.8**	36.2	35.9	35.3

Table 4: Adversarial evaluation and ablation study results on Multi30K 2016 test set. The best results are bold, and the worst are underlined.

As pointed out by previous studies that noise is the major part of visual features in the image-to-text task. It is necessary to find out whether our model can eliminate noise and learn useful information from visual features. We suppose that our models benefit from the visual object information, the multi-task scheme, and the de-noising ability. To investigate the effectiveness of these components, we conduct several experiments to compare our models with the following variants:

(1) *randomized inputs*. In this variant, we apply a random visual object ("ro") or a random word ("rw") (Lewis et al., 2020) to replace the original visual object in the training stage. The noise in this scheme is from the feature space of images or the textual representation space.

(2) *masked language model*. We replace all entity words with a special token "$\langle$mask$\rangle$ ". In this way, the reconstruction model degenerates to a masked language model ("mlm").

We apply these schemes to our RNN-based models on the Multi30K test2016 data. We use "vo" to represent our models on the detected visual objects which were displayed in Table 1.

As shown in Table 4, most of our models outperform the noise input models. It indicates that our models learn valuable information from visual objects for improving translation performance. The results of "mlm" show that the entity-masked multi-task scheme brings limited benefit to translation quality.

5 Entity Word Analysis

In this section, we take an in-depth analysis to find out why our entity-level cross-modal learning approach works. We intuitively assume that the approach provides an extra gain to the translation correctness of entity words. Therefore, we measure the translation `accuracy` for different types of words and subtract the result of the baseline model from MMT models as the extra gain which we call the `increment`. We split all words into two parts: the entity words which were mentioned as word entity in subsection 2.1 and the other words which correspond to no visual object. Different from sentence-level approaches, our entity-level MMT models are expected to obtain more `increment` for the entity words. It is represented as lowering the `increment difference` between the other words and the entity words.

The measurement is based on the sentence-level translation results of various MMT models. To get the word-level translation, we employ the fast-align (Dyer et al., 2013) toolkit which aligns tokens from source-side to target-side and concatenates the training set and the test set to train better alignments. The aligned target-side words are considered to be the translation of the source-side words. We take the alignment results of reference parallel data as the correct translation and compare it with the results of translated data from the MMT models. We pose a contrast among four kinds of MMT models: 6 of our RNN-based models, 12 of adversarial models in Table 4, 6 of MLM models

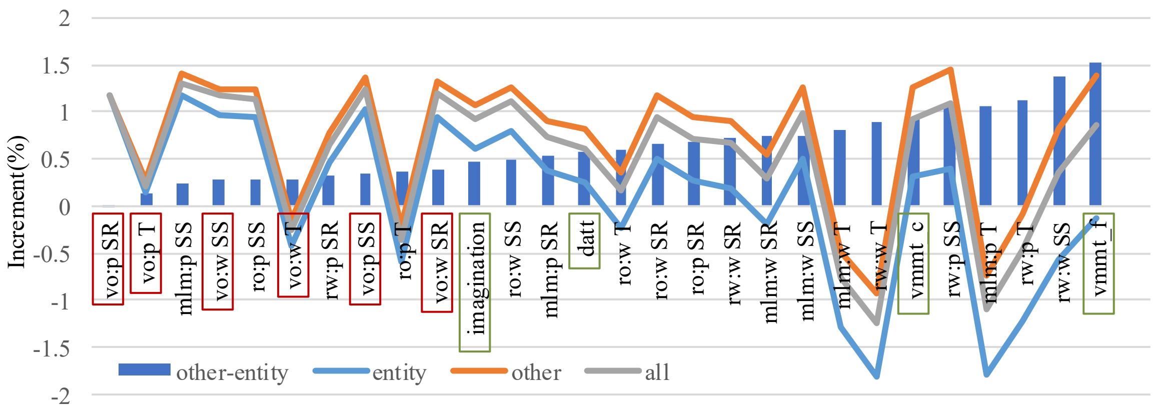


Figure 3: Experiment results of entity word analysis. "all" words in the Multi30K 2016 test are split into two parts. The "entity" represents the entity words in the word-level replacement rule. The "other" represents all other words. Our models are in the red bounding boxes, and former MMT models are in the green bounding boxes.

in Table 4, and 4 of other MMT models in Table 1.

Results on the Detected Visual Objects The results of models "vo" are shown in Figure 3 with red bounding boxes. A notable aspect of this figure is that most of the `differences` are positive. As mentioned in subsection 3, entity words are low-frequency which makes it harder to learn a better representation. We sort all results in ascending order, based on their `differences`. It shows that all our models come out among the lowest `differences` which indicates image information helps to narrow the gap of translation quality between entity words and other words.

Results of Our Multi-task Models As shown in Figure 3, neither adversarial models ("ro" and "rw") nor MLM models get stable lower `differences`. No evidence was found that the de-noising ability or the MLM of our multi-task scheme was a guarantee for helping the translations of entity words. It further proves that our entity-level cross-modal learning approach learns valuable visual information from visual objects.

Results of Previous Works The translation results of "DATT", "VMMT_C", and "VMMT_F" are generated from three times independent runs of their released codes. The "Imagination" is reproduced by ourselves. The results in Figure 3 show no advantage in lowering the `increment differences`. The overall contrastive analysis results indicate that our entity-level cross-modal learning approach is effective in optimizing the translation quality of the entity words.

6 Related Work

Previous studies mainly focused on fusing the multi-modal information into the sentence-level semantics (Huang et al., 2016; Calixto and Liu, 2017; Calixto et al., 2017; Libovický and Helcl, 2017; Delbrouck and Dupont, 2017) in the RNN-based architecture (Bahdanau et al., 2015). Besides above approaches, Toyama et al. (2016); Calixto et al. (2019) proposed to apply latent variables as the unified semantic representations. Ive et al. (2019) proposed a translate-and-refine approach to generate a good translation from the first draft by making better use of the target language and visual context. There are also works (Wang et al., 2018a,b; Zhao et al., 2020) show that extra modality information is useful in a more fine-grained way.

Recently, Yin et al. (2020) proposed a fine-grained method that employs a graph-based multi-modal fusion encoder to fuse image and source text in the entity level. The input sentence and image are represented as a unified graph. The encoder can capture the relations among visual objects and linguistic entities. Similar to their model, our model also explores the fine-grained multi-modal semantics at the entity-level. However, the differences lie in two aspects: (1) our entity-level cross-modal learning task is framed as a reconstruction problem that is simpler than their graph-based model, and we also investigate word-level and phrase-level entities. (2) Our model does not rely on images during the inference stage. Furthermore, our results are comparable or better than theirs.

Our work is mainly inspired by the image probing work (Caglayan et al., 2019). This work degrades the textual input by replacing some depictable words with visual information and then tests whether the image is helpful. The authors find that the entities are most informative in the image. In our work, we utilize a similar degradation method to prepare the multi-modal input. We go a step further to learn an entity-level cross-modal representation that is proved to significantly improve translation performance.

Wang and Xiong (2021) proposed several loss functions that help their model to capture more relevant information from visual objects. They also tried to weaken the functionality of textual modality to exploit visual information. The differences are as follows: (1) Their approach needs entity labels manually annotated from the Flickr30K Entities dataset. (2) Their model works well in the textual degradation situation. However, this strategy is not able to facilitate their model in the general intact-textual scenarios.

Elliott and Kádár (2017) also applied a multi-task framework to perform multi-modal translation. The differences are in the following aspects: (1) Their work fuse visual information in the sentence-level semantics. Our models benefit from entity-level cross-modal fusion. (2) The image is grounded from source sentences in their model. In our approach, the model performs an image-to-text reconstruction task. (3) Benefiting from the similar frameworks of the reconstruction task and the translation task, our multi-task method is flexible enough to provide three parameter-sharing schemes.

Our adversarial evaluation in the experiments is inspired by the work (Elliott, 2018) in which the authors proposed an adversarial approach to measure the utility of the image in multi-modal translation. In this work, a random image is fed into the model instead of the paired one. Then the difference in performance reflects the importance of visual information. We apply a similar adversarial evaluation to our models by randomizing the visual object images in the training stage.

To fully exploit visual modality, we degrade the linguistic context and reconstruct the original text from both modalities. This strategy leads to a similar model framework to the vision-language pre-trained models (Li et al., 2019; Lu et al., 2019; Su et al., 2020). However, vision-language pre-trained models can only initialize the encoder while our methods provide various models to learn cross-modal representations in the encoder-decoder architecture.

7 Conclusion

In this paper, we have proposed an entity-level cross-modal learning approach that explicitly incorporates visual information into linguistic entities and is combined with the text-only translation task in a multi-task framework. Our extensive results show that our models can achieve comparable or even better performance than state-of-the-art models. Furthermore, we take an in-depth analysis to figure out why the approach works by contrasting the translation correctness of entity words with multiple adversarial models and former MMT models. The results show that the translation accuracy of entity words significantly increases with the help of entity-level visual information. Our findings suggest that images can be utilized explicitly in an MMT model and better approaches are favored to leverage the fine-grained object information in the image.

Acknowledgements

The research work descried in this paper has been supported by the Natural Science Foundation of China under Grant No. U1836221.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439, Copenhagen, Denmark. Association for Computational Linguistics.

Ozan Caglayan, Adrien Bardet, Fethi Bougares, Loïc Barrault, Kai Wang, Marc Masana, Luis Herranz, and Joost van de Weijer. 2018. LIUM-CVC submissions for WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 597–602, Belgium, Brussels. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 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)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada. Association for Computational Linguistics.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405, Florence, Italy. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919, Copenhagen, Denmark. Association for Computational Linguistics.

Michael 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*, pages 376–380, Baltimore, Maryland, USA. Association for Computational Linguistics.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 644–648, Atlanta, Georgia. Association for Computational Linguistics.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

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

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 130–141, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9:1735–80.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Jindřich Libovický, Jindřich Helcl, and David Mareček. 2018. Input combination strategies for multi-source transformer decoder. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 253–260, Brussels, Belgium. Association for Computational Linguistics.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870.

Bryan A. Plummer, Liwei Wang, Christopher M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *IJCV*, 123(1):74–93.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael and Bernstein. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: pre-training of generic visual-linguistic representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Joji Toyama, Masanori Misono, Masahiro Suzuki, Kotaro Nakayama, and Yutaka Matsuo. 2016. Neural machine translation with latent semantic of image and text. *CoRR*, abs/1611.08459.

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.

Dexin Wang and Deyi Xiong. 2021. Efficient object-level visual context modeling for multimodal machine translation: Masking irrelevant objects helps grounding. 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 2720–2728. AAAI Press.

Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2018a. Associative multichannel autoencoder for multimodal word representation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 115–124, Brussels, Belgium. Association for Computational Linguistics.

Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2018b. Learning multimodal word representation via dynamic fusion methods. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5973–5980. AAAI Press.

Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. 2019. A fast and accurate one-stage approach to visual grounding. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4682–4692. IEEE.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3025–3035, Online. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2020. Double attention-based multimodal neural machine translation with semantic image regions. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 105–114, Lisboa, Portugal. European Association for Machine Translation.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium. Association for Computational Linguistics.

A Adversarial Evaluation and Ablation Study for Reconstruction

RNN-based		BLEU			
		vo	ro	rw	mlm
word	EMMT$_{\mathrm{SR}}$	**44.5**	43.9	42.4	40.0
	EMMT$_{\mathrm{SS}}$	**45.2**	43.6	2.6	25.1
	EMMT$_{\mathrm{T}}$	**16.6**	15.8	15.7	16.0
phrase	EMMT$_{\mathrm{SR}}$	**35.4**	23.7	26.5	24.9
	EMMT$_{\mathrm{SS}}$	**35.2**	27.6	2.1	24.2
	EMMT$_{\mathrm{T}}$	**13.3**	10.6	10.3	10.5

Table 5: Adversarial evaluation and ablation study reconstruction results on Multi30K 2016 test set. The best results are bold, and the worst are underlined.

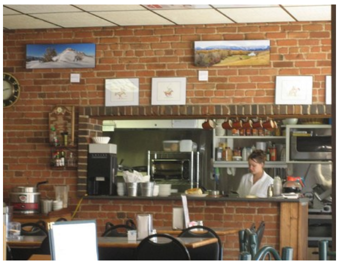
As shown in Table 5, all our models outperform the noise input models for the reconstruction task. These results further support the evidence from the translation task and indicate that the reconstruction task is effective. And the gains seem to be affected by the proportion of visual features we input. As we can see that the phrase-level replacement schemes seem to enlarge the quality difference between our models and adversarial models. It indicates that the less textual information is, the greater the role of visual information it plays.

The large performance gap between the word-level replacement rule and the phrase-level is caused by the unpredictable adjunct words which were mentioned in subsection 2.1. It is extremely difficult to predict the adjunct words from the visual objects. It is the reason why the phrase-level replacement models get much lower reconstruction BLEU than word-level. Besides the adjunct words, the word entities are also included in the phrase entities. It makes sure that phrase-level schemes get similar translation performance to word-level.

pigeons → tauben

Src: a man crouches while doing chores, as pigeons wander in the background.
Ref: ein mann bückt sich und macht hausarbeiten während im hintergrund tauben vorbeilaufen.
NMT: ein mann hockt und macht dabei mit dem maul im hintergrund.
vo.wSR: ein mann bückt sich, während er mit der weihnachtszeit im hintergrund mit tauben.
ro.wSR: ein mann kauert beim salto im hintergrund beim salto im hintergrund.
rw.wSR: ein mann hockt während eines festumzugs im hintergrund.

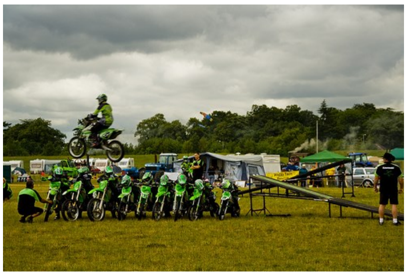
Figure 4: Case #4631909374. An example of under-translated entity word "pigeon". Our model "vo.wSR" correctly translates the "pigeon" to "tauben".



shirt → hemd

Src: a woman in a white shirt works behind the counter at a cafe.
Ref: eine frau in einem weißen hemd arbeitet hinter dem tresen in einem café.
NMT: eine frau in einem weißen oberteil arbeitet hinter der theke in einem café.
vo.wSR: eine frau in weißem hemd arbeitet hinter der theke in einem café.
ro.wSR: eine frau in einem weißen oberteil arbeitet hinter der theke in einem café.
rw.wSR: eine frau in einem weißen oberteil arbeitet hinter der theke in einem café.

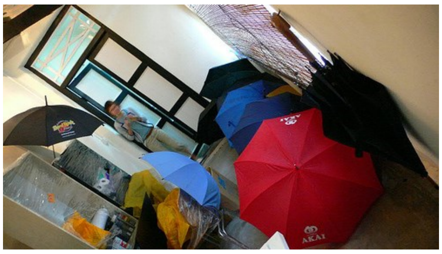
Figure 5: Case #3884010975. An example of entity word "shirt". Our model "vo.wSR" correctly translates the "shirt" to "hemd".



motorbike → motorrad, motorbikes→motorräder

Src: a man in green jumps serveral motorbikes on his own motorbike.
Ref: ein mann in grün springt auf seinem motorrad über mehrere andere motorräder.
NMT: ein grün gekleideter mann springt mit seinem motorrad auf seinem motorrad.
vo.wSR: ein mann in grüner kleidung springt mit dem motorräder auf seinem motorrad.
ro.wSR: ein grün gekleideter mann springt auf seinem eigenen motorrad.
rw.wSR: ein mann in grün springt mit seinem skateboard auf der nase.

Figure 6: Case #3646927481. An example of entity words "motorbike" and "motorbikes". Our model "vo.wSR" successfully separates similar words.



Umbrellas → regenschirmen(as reference), schirmen

Src: a woman texts on her phone while surrounded by umbrellas.
Ref: eine frau tippt auf ihrem handy, umgeben von regenschirmen.
NMT: eine frau schreibt auf ihrem handy, während sie umgeben von schirmen umgeben ist.
vo.wSR: eine frau textet mit regenschirmen auf ihrem handy.
ro.wSR: eine frau schreibt auf ihrem handy, während er von schirmen umgeben ist.
rw.wSR: eine frau telefoniert mit ihrem handy, während er von schirmen sie ansieht.

Figure 7: Case #280007961. An example of entity word "umbrellas". Our model "vo.wSR" gets the same result as the reference.