A Novel Graph-based Multi-modal Fusion Encoder for Neural Machine Translation

Yongjing Yin[1]*, **Fandong Meng**[2], **Jinsong Su**[1]†, **Chulun Zhou**[1], **Zhengyuan Yang**[3], **Jie Zhou**[2], **Jiebo Luo**[3]
[1]Xiamen University, Xiamen, China
[2]Pattern Recognition Center, WeChat AI, Tencent Inc, Beijing, China
[3]Department of Computer Science, University of Rochester, Rochester NY 14627, USA
yinyongjing@stu.xmu.edu.cn fandongmeng@tencent.com
jssu@xmu.edu.cn

Abstract

Multi-modal neural machine translation (NMT) aims to translate source sentences into a target language paired with images. However, dominant multi-modal NMT models do not fully exploit fine-grained semantic correspondences between semantic units of different modalities, which have potential to refine multi-modal representation learning. To deal with this issue, in this paper, we propose a novel graph-based multi-modal fusion encoder for NMT. Specifically, we first represent the input sentence and image using a unified multi-modal graph, which captures various semantic relationships between multi-modal semantic units (words and visual objects). We then stack multiple graph-based multi-modal fusion layers that iteratively perform semantic interactions to learn node representations. Finally, these representations provide an attention-based context vector for the decoder. We evaluate our proposed encoder on the Multi30K datasets. Experimental results and in-depth analysis show the superiority of our multi-modal NMT model.

1 Introduction

Multi-modal neural machine translation (NMT) (Huang et al., 2016; Calixto et al., 2017) has become an important research direction in machine translation, due to its research significance in multi-modal deep learning and wide applications, such as translating multimedia news and web product information (Zhou et al., 2018). It significantly extends the conventional text-based machine translation by taking images as additional inputs. The assumption behind this is that the translation is expected to be more accurate compared to purely text-based translation, since the visual context helps to resolve ambiguous multi-sense words (Ive et al., 2019).

Apparently, how to fully exploit visual information is one of the core issues in multi-modal NMT, which directly impacts the model performance. To this end, a lot of efforts have been made, roughly consisting of: (1) encoding each input image into a global feature vector, which can be used to initialize different components of multi-modal NMT models, or as additional source tokens (Huang et al., 2016; Calixto et al., 2017), or to learn the joint multi-modal representation (Zhou et al., 2018; Calixto et al., 2019); (2) extracting object-based image features to initialize the model, or supplement source sequences, or generate attention-based visual context (Huang et al., 2016; Ive et al., 2019); and (3) representing each image as spatial features, which can be exploited as extra context (Calixto et al., 2017; Delbrouck and Dupont, 2017a; Ive et al., 2019), or a supplement to source semantics (Delbrouck and Dupont, 2017b) via an attention mechanism.

Despite their success, the above studies do not fully exploit the fine-grained semantic correspondences between semantic units within an input sentence-image pair. For example, as shown in Figure 1, the noun phrase "*a toy car*" semantically corresponds to the blue dashed region. The neglect of this important clue may be due to two big challenges: 1) how to construct a unified representation to bridge the semantic gap between two different modalities, and 2) how to achieve semantic interactions based on the unified representation. However, we believe that such semantic correspondences can be exploited to refine multi-modal representation learning, since they enable the representations within one modality to incorporate cross-modal information as supplement during multi-modal semantic interactions (Lee et al., 2018; Tan and Bansal, 2019).

*This work is done when Yongjing Yin was interning at Pattern Recognition Center, WeChat AI, Tencent Inc, China.
†Corresponding author.

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

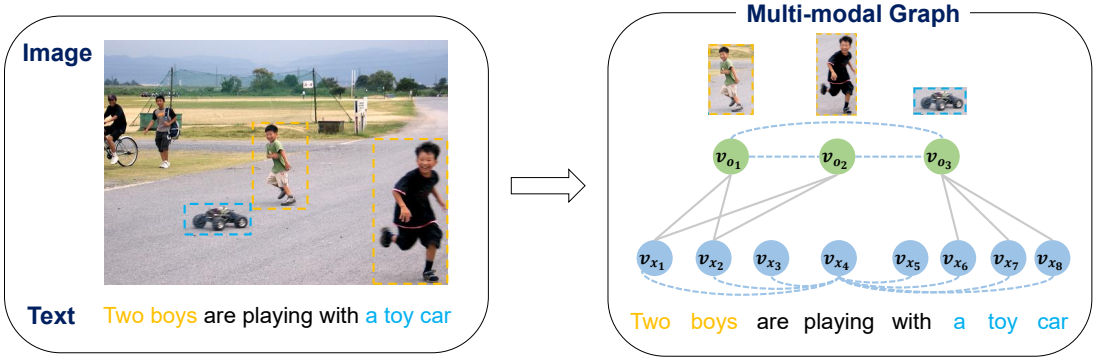


Figure 1: The multi-modal graph for an input sentence-image pair. The blue and green solid circles denote textual nodes and visual nodes respectively. An intra-modal edge (dotted line) connects two nodes in the same modality, and an inter-modal edge (solid line) links two nodes in different modalities. Note that we only display edges connecting the textual node "*playing*" and other textual ones for simplicity.

In this paper, we propose a novel graph-based multi-modal fusion encoder for NMT. We first represent the input sentence and image with a unified multi-modal graph. In this graph, each node indicates a semantic unit: *textual word* or *visual object*, and two types of edges are introduced to model semantic relationships between semantic units within the same modality (*intra-modal edges*) and semantic correspondences between semantic units of different modalities (*inter-modal edges*) respectively. Based on the graph, we then stack multiple graph-based multi-modal fusion layers that iteratively perform semantic interactions among the nodes to conduct graph encoding. Particularly, during this process, we distinguish the parameters of two modalities, and sequentially conduct intra- and inter-modal fusions to learn multi-modal node representations. Finally, these representations can be exploited by the decoder via an attention mechanism.

Compared with previous models, ours is able to fully exploit semantic interactions among multi-modal semantic units for NMT. Overall, the major contributions of our work are listed as follows:

- We propose a unified graph to represent the input sentence and image, where various semantic relationships between multi-modal semantic units can be captured for NMT.
- We propose a graph-based multi-modal fusion encoder to conduct graph encoding based on the above graph. To the best of our knowledge, our work is the first attempt to explore multi-modal graph neural network (GNN) for NMT.
- We conduct extensive experiments on Multi30k datasets of two language pairs. Experimental results and in-depth analysis indicate that our encoder is effective to fuse multi-modal information for NMT. Particularly, our multi-modal NMT model significantly outperforms several competitive baselines.
- We release the code at https://github.com/DeepLearnXMU/GMNMT.

2 NMT with Graph-based Multi-modal Fusion Encoder

Our multi-modal NMT model is based on attentional encoder-decoder framework with maximizing the log likelihood of training data as the objective function.

2.1 Encoder

Essentially, our encoder can be regarded as a multi-modal extension of GNN. To construct our encoder, we first represent the input sentence-image pair as a unified multi-modal graph. Then, based on this graph, we stack multiple multi-modal fusion layers to learn node representations, which provides the attention-based context vector to the decoder.

2.1.1 Multi-modal Graph

In this section, we take the sentence and the image shown in Figure 1 as an example, and describe how to use a multi-modal graph to represent them. Formally, our graph is undirected and can be formalized as G=(V,E), which is constructed as follows:

In the node set V, each node represents either a textual word or a visual object. Specifically, we adopt the following strategies to construct these two kinds of nodes: (1) We include all words as separate **textual nodes** in order to fully exploit textual

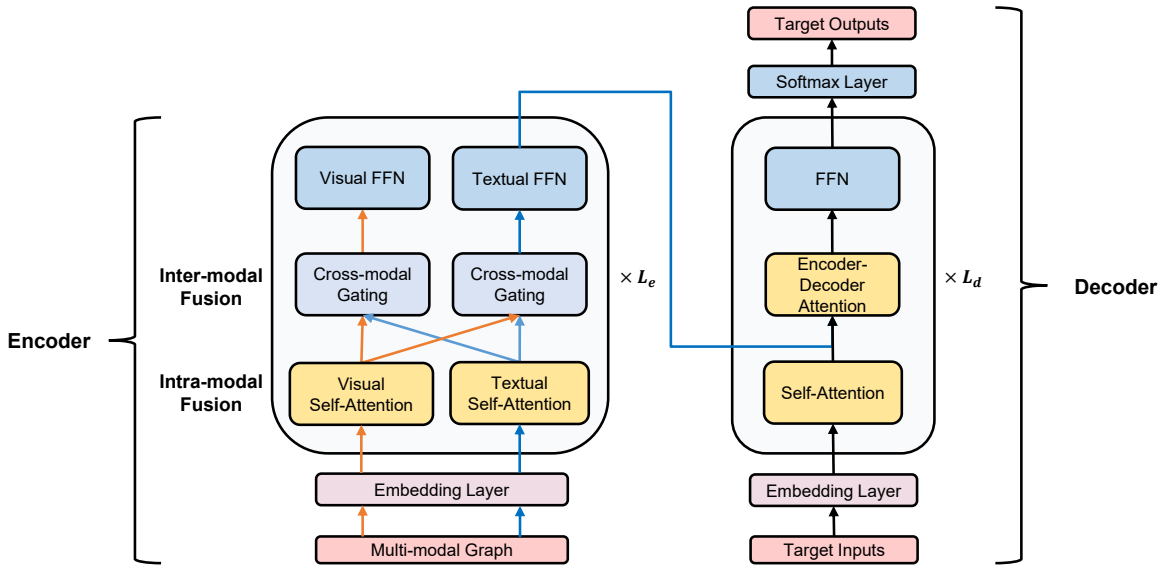


Figure 2: The architecture of our NMT model with the graph-based multi-modal fusion encoder. Note that we actually do not apply a Visual FFN to the last layer in the encoder.

information. For example, in Figure 1, the multi-modal graph contains totally eight textual nodes, each of which corresponds to a word in the input sentence; (2) We employ the Stanford parser to identify all noun phrases in the input sentence, and then apply a visual grounding toolkit (Yang et al., 2019) to detect bounding boxes (visual objects) for each noun phrase. Subsequently, all detected visual objects are included as independent **visual nodes**. In this way, we can effectively reduce the negative impact of abundant unrelated visual objects. Let us revisit the example in Figure 1, where we can identify two noun phrases "*Two boys*" and "*a toy car*" from the input sentence, and then include three visual objects into the multi-modal graph.

To capture various semantic relationships between multi-modal semantic units for NMT, we consider two kinds of edges in the edge set E: (1) Any two nodes in the same modality are connected by an **intra-modal edge**; and (2) Each textual node representing any noun phrase and the corresponding visual node are connected by an **inter-modal edge**. Back to Figure 1, we can observe that all visual nodes are connected to each other, and all textual nodes are fully-connected. However, only nodes v_{o_1} and v_{x_1}, v_{o_1} and v_{x_2}, v_{o_2} and v_{x_1}, v_{o_2} and v_{x_2}, v_{o_3} and v_{x_6}, v_{o_3} and v_{x_7}, v_{o_3} and v_{x_8} are connected by inter-modal edges.

2.1.2 Embedding Layer

Before inputting the multi-modal graph into the stacked fusion layers, we introduce an embedding layer to initialize the node states. Specifically, for each textual node v_{x_i}, we define its initial state $H_{x_i}^{(0)}$ as the sum of its word embedding and position encoding (Vaswani et al., 2017). To obtain the initial state $H_{o_j}^{(0)}$ of the visual node v_{o_j}, we first extract visual features from the fully-connected layer that follows the ROI pooling layer in Faster-RCNN (Ren et al., 2015), and then employ a multi-layer perceptron with ReLU activation function to project these features onto the same space as textual representations.

2.1.3 Graph-based Multi-modal Fusion Layers

As shown in the left part of Figure 2, on the top of embedding layer, we stack L_e *graph-based multi-modal fusion layers* to encode the above-mentioned multi-modal graph. At each fusion layer, we sequentially conduct intra- and inter-modal fusions to update all node states. In this way, the final node states encode both the context within the same modality and the cross-modal semantic information simultaneously. Particularly, since visual nodes and textual nodes are two types of semantic units containing the information of different modalities, we apply similar operations but with different parameters to model their state update process, respectively.

Specifically, in the l-th fusion layer, both updates of textual node states $\mathbf{H}_x^{(l)}=\{H_{x_i}^{(l)}\}$ and visual node states $\mathbf{H}_o^{(l)}=\{H_{o_j}^{(l)}\}$ mainly involve the following steps:

Step1: Intra-modal fusion. At this step, we employ **self-attention** to generate the contextual representation of each node by collecting the message from its neighbors of the same modality.

Formally, the contextual representations $\mathbf{C}_x^{(l)}$ of all textual nodes are calculated as follows: [1]

$$\mathbf{C}_x^{(l)} = \text{MultiHead}(\mathbf{H}_x^{(l-1)}, \mathbf{H}_x^{(l-1)}, \mathbf{H}_x^{(l-1)}), \quad (1)$$

where MultiHead(**Q**, **K**, **V**) is a multi-head self-attention function taking a query matrix **Q**, a key matrix **K**, and a value matrix **V** as inputs. Similarly, we generate the contextual representations $\mathbf{C}_o^{(l)}$ of all visual nodes as

$$\mathbf{C}_o^{(l)} = \text{MultiHead}(\mathbf{H}_o^{(l-1)}, \mathbf{H}_o^{(l-1)}, \mathbf{H}_o^{(l-1)}). \quad (2)$$

In particular, since the initial representations of visual objects are extracted from deep CNNs, we apply a simplified multi-head self-attention to preserve the initial representations of visual objects, where the learned linear projects of values and final outputs are removed.

Step2: Inter-modal fusion. Inspired by studies in multi-modal feature fusion (Teney et al., 2018; Kim et al., 2018), we apply a **cross-modal gating** mechanism with an element-wise operation to gather the semantic information of the cross-modal neighbours of each node.

Concretely, we generate the representation $M_{x_i}^{(l)}$ of a text node v_{x_i} in the following way:

$$M_{x_i}^{(l)} = \sum_{j \in A(v_{x_i})} \alpha_{i,j} \odot C_{o_j}^{(l)}, \quad (3)$$

$$\alpha_{i,j} = \text{Sigmoid}(\mathbf{W}_1^{(l)} C_{x_i}^{(l)} + \mathbf{W}_2^{(l)} C_{o_j}^{(l)}), \quad (4)$$

where $A(v_{x_i})$ is the set of neighboring visual nodes of v_{x_i}, and $\mathbf{W}_1^{(l)}$ and $\mathbf{W}_2^{(l)}$ are parameter matrices. Likewise, we produce the representation $M_{o_j}^{(l)}$ of a visual node v_{o_j} as follows:

$$M_{o_j}^{(l)} = \sum_{i \in A(v_{o_j})} \beta_{j,i} \odot C_{x_i}^{(l)}, \quad (5)$$

$$\beta_{j,i} = \text{Sigmoid}(\mathbf{W}_3^{(l)} C_{o_j}^{(l)} + \mathbf{W}_4^{(l)} C_{x_i}^{(l)}), \quad (6)$$

where $A(v_{o_j})$ is the set of adjacent textual nodes of v_{o_j}, and $\mathbf{W}_3^{(l)}$ and $\mathbf{W}_4^{(l)}$ are also parameter matrices.

The advantage is that the above fusion approach can better determine the degree of inter-modal fusion according to the contextual representations of each modality. Finally, we adopt position-wise feed forward networks FFN($*$) to generate the textual node states $\mathbf{H}_x^{(l)}$ and visual node states $\mathbf{H}_o^{(l)}$:

$$\mathbf{H}_x^{(l)} = \text{FFN}(\mathbf{M}_x^{(l)}), \quad (7)$$

$$\mathbf{H}_o^{(l)} = \text{FFN}(\mathbf{M}_o^{(l)}), \quad (8)$$

where $\mathbf{M}_x^{(l)} = \{M_{x_i}^{(l)}\}$, $\mathbf{M}_o^{(l)} = \{M_{o_j}^{(l)}\}$ denote the above updated representations of all textual nodes and visual nodes respectively.

2.2 Decoder

Our decoder is similar to the conventional Transformer decoder. Since visual information has been incorporated into all textual nodes via multiple graph-based multi-modal fusion layers, we allow the decoder to dynamically exploit the multi-modal context by only attending to textual node states.

As shown in the right part of Figure 2, we follow Vaswani et al. (2017) to stack L_d identical layers to generate target-side hidden states, where each layer l is composed of three sub-layers. Concretely, the first two sub-layers are a masked self-attention and an encoder-decoder attention to integrate target- and source-side contexts respectively:

$$\mathbf{E}^{(l)} = \text{MultiHead}(\mathbf{S}^{(l-1)}, \mathbf{S}^{(l-1)}, \mathbf{S}^{(l-1)}), \quad (9)$$

$$\mathbf{T}^{(l)} = \text{MultiHead}(\mathbf{E}^{(l)}, \mathbf{H}_x^{(L_e)}, \mathbf{H}_x^{(L_e)}), \quad (10)$$

where $\mathbf{S}^{(l-1)}$ denotes the target-side hidden states in the l-1-th layer. In particular, $\mathbf{S}^{(0)}$ are the embeddings of input target words. Then, a position-wise fully-connected forward neural network is uesd to produce $\mathbf{S}^{(l)}$ as follows:

$$\mathbf{S}^{(l)} = \text{FFN}(\mathbf{T}^{(l)}). \quad (11)$$

Finally, the probability distribution of generating the target sentence is defined by using a softmax layer, which takes the hidden states in the top layer as input:

$$P(Y|X, I) = \prod_t \text{Softmax}(\mathbf{W}\mathbf{S}_t^{(L_d)} + b), \quad (12)$$

where X is the input sentence, I is the input image, Y is the target sentence, and $\mathbf{W}$ and b are the parameters of the softmax layer.

3 Experiment

We carry out experiments on multi-modal English⇒German (En⇒De) and English⇒French (En⇒Fr) translation tasks.

[1] For simplicity, we omit the descriptions of layer normalization and residual connection.

3.1 Setup

Datasets We use the Multi30K dataset (Elliott et al., 2016), where each image is paired with one English description and human translations into German and French. Training, validation and test sets contain 29,000, 1,014 and 1,000 instances respectively. In addition, we evaluate various models on the WMT17 test set and the ambiguous MSCOCO test set, which contain 1,000 and 461 instances respectively. Here, we directly use the preprocessed sentences [2] and segment words into subwords via byte pair encoding (Sennrich et al., 2016) with 10,000 merge operations.

Visual Features We first apply the Stanford parser to identify noun phrases from each source sentence, and then employ the visual ground toolkit released by Yang et al. (2019) to detect associated visual objects of the identified noun phrases. For each phrase, we keep the visual object with the highest prediction probability, so as to reduce negative effects of abundant visual objects. In each sentence, the average numbers of objects and words are around 3.5 and 15.0 respectively. [3] Finally, we compute 2,048-dimensional features for these objects with the pre-trained ResNet-100 Faster-RCNN (Ren et al., 2015).

Settings We use Transformer (Vaswani et al., 2017) as our baseline. Since the size of training corpus is small and the trained model tends to be over-fitting, we first perform a small grid search to obtain a set of hyper-parameters on the En⇒De validation set. Specifically, the word embedding dimension and hidden size are 128 and 256 respectively. The decoder has L_d=4 layers[4] and the number of attention heads is 4. The dropout is set to 0.5. Each batch consists of approximately 2,000 source and target tokens. We apply the Adam optimizer with a scheduled learning rate to optimize various models, and we use other same settings as (Vaswani et al., 2017). Finally, we use the metrics BLEU (Papineni et al., 2002) and METEOR (Denkowski and Lavie, 2014) to evaluate the quality of translations. Particularly, we run all models three times for each experiment and report the average results.

[2] http://www.statmt.org/wmt18/multimodal-task.html

[3] There is no parsing failure for this dataset. If no noun is detected for a sentence, the object representations will be set to zero vectors and the model will degenerate to Transformer.

[4] The encoder of the text-based Transformer also has 4 layers.

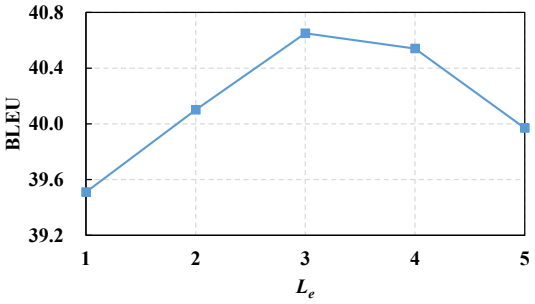


Figure 3: Results on the En⇒De validation set regarding the number L_e of graph-based multi-modal fusion layers.

Baseline Models In addition to the text-based Transformer (Vaswani et al., 2017), we adapt several effective approaches to Transformer using our visual features, and compare our model with them[5]:

- **ObjectAsToken(TF)** (Huang et al., 2016). It is a variant of the Transformer, where all visual objects are regarded as extra source tokens and placed at the front of the input sentence.
- **Enc-att(TF)** (Delbrouck and Dupont, 2017b). An encoder-based image attention mechanism is incorporated into Transformer, which augments each source annotation with an attention-based visual feature vector.
- **Doubly-att(TF)** (Helcl et al., 2018). It is a doubly attentive Transformer. In each decoder layer, a cross-modal multi-head attention sublayer is inserted before the fully connected feed-forward layer to generate the visual context vector from visual features.

We also display the performance of several dominant multi-modal NMT models such as **Doubly-att(RNN)** (Calixto et al., 2017), **Soft-att(RNN)** (Delbrouck and Dupont, 2017a), **Stochastic-att(RNN)** (Delbrouck and Dupont, 2017a), **Fusion-conv(RNN)** (Caglayan et al., 2017), **Trg-mul(RNN)** (Caglayan et al., 2017), **VMMT(RNN)** (Calixto et al., 2019) and **Deliberation Network(TF)** (Ive et al., 2019) on the same datasets.

3.2 Effect of Graph-based Multi-modal Fusion Layer Number L_e

The number L_e of multi-modal fusion layer is an important hyper-parameter that directly determines

[5] We use suffixes "(**RNN**)" and "(**TF**)" to represent RNN- and Transformer-style NMT models, respectively.

Model	En⇒De					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Existing Multi-modal NMT Systems						
Doubly-att(RNN) (Calixto et al., 2017)	36.5	55.0	-	-	-	-
Soft-att(RNN) (Delbrouck and Dupont, 2017a)	37.6	55.3	-	-	-	-
Stochastic-att(RNN) (Delbrouck and Dupont, 2017a)	38.2	55.4	-	-	-	-
Fusion-conv(RNN) (Caglayan et al., 2017)	37.0	57.0	29.8	51.2	25.1	46.0
Trg-mul(RNN)(Caglayan et al., 2017)	37.8	**57.7**	30.7	**52.2**	26.4	47.4
VMMT(RNN) (Calixto et al., 2019)	37.7	56.0	30.1	49.9	25.5	44.8
Deliberation Network(TF) (Ive et al., 2019)	38.0	55.6	-	-	-	-
Our Multi-modal NMT Systems						
Transformer (Vaswani et al., 2017)	38.4	56.5	30.6	50.4	27.3	46.2
ObjectAsToken(TF) (Huang et al., 2016)	39.0	57.2	31.7	51.3	28.4	47.0
Enc-att(TF) (Delbrouck and Dupont, 2017b)	38.7	56.6	31.3	50.6	28.0	46.6
Doubly-att(TF) (Helcl et al., 2018)	38.8	56.8	31.4	50.5	27.4	46.5
Our model	**39.8**	57.6	**32.2**	51.9	**28.7**	**47.6**

Table 1: Experimental results on the En⇒De translation task.

the degree of fine-grained semantic fusion in our encoder. Thus, we first inspect its impact on the EN⇒DE validation set.

Figure 3 provides the experimental results using different L_e and our model achieves the best performance when L_e is 3. Hence, we use L_e=3 in all subsequent experiments.

3.3 Results on the En⇒De Translation Task

Table 1 shows the main results on the En⇒De translation task. Ours outperforms most of the existing models and all baselines, and is comparable to Fusion-conv(RNN) and Trg-mul(RNN) on METEOR. The two results are from the state-of-the-art system on the WMT2017 test set, which is selected based on METEOR. Comparing the baseline models, we draw the following interesting conclusions:

First, our model outperforms ObjectAsToken(TF), which concatenates regional visual features with text to form attendable sequences and employs self-attention mechanism to conduct inter-modal fusion. The underlying reasons consist of two aspects: explicitly modeling semantic correspondences between semantic units of different modalities, and distinguishing model parameters for different modalities.

Second, our model also significantly outperforms Enc-att(TF). Note that Enc-att(TF) can be considered as a single-layer semantic fusion encoder. In addition to the advantage of explicitly modeling semantic correspondences, we conjecture that multi-layer multi-modal semantic interactions are also beneficial to NMT.

Third, compared with Doubly-att(TF) simply using an attention mechanism to exploit visual information, our model achieves a significant improvement, because of sufficient multi-modal fusion in our encoder.

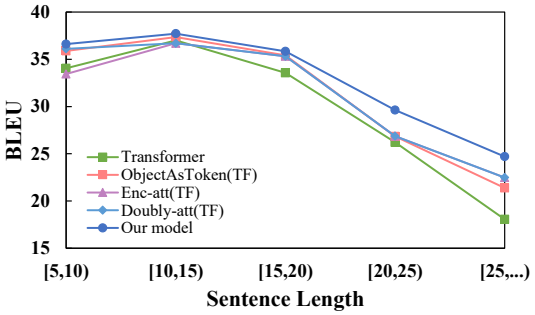


Figure 4: BLEU scores on different translation groups divided according to source sentence lengths.

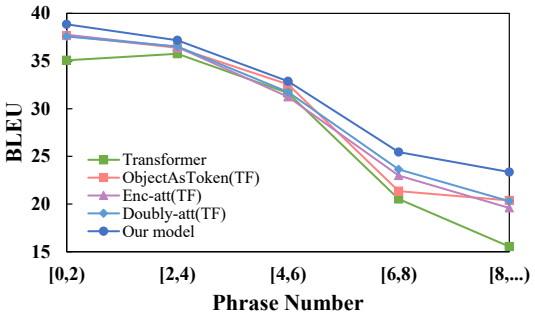


Figure 5: BLEU scores on different translation groups divided according to source phrase numbers.

Besides, we divide our test sets into different groups based on the lengths of source sentences and the numbers of noun phrases, and then compare the performance of different models in each group. Figures 4 and 5 report the BLEU scores on these groups. Overall, our model still consistently achieves the best performance in all groups. Thus, we confirm again the effectiveness and gen-

Model	En⇒De					
	Test2016		Test2017		MSCOCO	
	BLEU	METEOR	BLEU	METEOR	BLEU	METEOR
Our model	39.8	57.6	32.2	51.9	28.7	47.6
w/o inter-modal fusion	38.7	56.7	30.7	50.6	27.0	46.7
visual grounding ⇒ fully-connected	36.4	53.4	28.3	47.0	24.4	42.9
different parameters ⇒ unified parameters	39.2	57.3	31.9	51.4	27.7	47.4
w/ attending to visual nodes	39.6	57.3	32.0	51.3	27.9	46.8
attending to textual nodes ⇒ attending to visual nodes	30.9	48.6	22.3	41.5	20.4	38.7

Table 2: Ablation study of our model on the EN⇒DE translation task.

Model	En⇒Fr			
	Test2016		Test2017	
	BLEU	METEOR	BLEU	METEOR
Existing Multi-modal NMT Systems				
Fusion-conv(RNN) (Caglayan et al., 2017)	53.5	70.4	51.6	68.6
Trg-mul(RNN)(Caglayan et al., 2017)	54.7	71.3	52.7	**69.5**
Deliberation Network(TF) (Ive et al., 2019)	59.8	74.4	-	-
Our Multi-modal NMT Systems				
Transformer (Vaswani et al., 2017)	59.5	73.7	52.0	68.0
ObjectAsToken(TF) (Huang et al., 2016)	60.0	74.3	52.9	68.6
Enc-att(TF) (Delbrouck and Dupont, 2017b)	60.0	74.3	52.8	68.3
Doubly-att(TF) (Helcl et al., 2018)	59.9	74.1	52.4	68.1
Our model	**60.9**	**74.9**	**53.9**	69.3

Table 3: Experimental results on the En⇒Fr translation task.

Model	Training	Decoding	Parameter
Transformer	2.6K	17.8	3.4M
ObjectAsToken(TF)	1.6K	17.2	3.7M
Enc-att(TF)	1.3K	16.9	3.6M
Doubly-att(TF)	1.0K	12.9	3.8M
Our model	1.1K	16.7	4.0M

Table 4: Training speed (tokens/second), decoding speed (sentences/second) and the number of parameters of different models on the En⇒De translation task.

erality of our proposed model. Note that in the sentences with more phrases, which are usually long sentences, the improvements of our model over baselines are more significant. We speculate that long sentences often contain more ambiguous words. Thus compared with short sentences, long sentences may require visual information to be better exploited as supplementary information, which can be achieved by the multi-modal semantic interaction of our model.

We also show the training and decoding speed of our model and the baselines in Table 4. During training, our model can process approximately 1.1K tokens per second, which is comparable to other multi-modal baselines. When it comes to decoding procedure, our model translates about 16.7 sentences per second and the speed drops slightly compared to Transformer. Moreover, our model only introduces a small number of extra parameters and achieves better performance.

3.4 Ablation Study

To investigate the effectiveness of different components, we further conduct experiments to compare our model with the following variants in Table 2:

(1) *w/o inter-modal fusion*. In this variant, we apply two separate Transformer encoders to learn the semantic representations of words and visual objects, respectively, and then use the doubly-attentive decoder (Helcl et al., 2018) to incorporate textual and visual contexts into the decoder. The result in line 3 indicates that removing the inter-modal fusion leads to a significant performance drop. It suggests that semantic interactions among multi-modal semantic units are indeed useful for multi-modal representation learning.

(2) *visual grounding ⇒ fully-connected*. We make the words and visual objects fully-connected to establish the inter-modal correspondences. The result in line 4 shows that this change causes a significant performance decline. The underlying reason is the fully-connected semantic correspondences introduce much noise to our model.

(3) *different parameters ⇒ unified parameters*. When constructing this variant, we assign unified parameters to update node states in different modalities. Apparently, the performance drop reported in line 5 also demonstrates the validity of our ap-

proach using different parameters.

(4) *w/ attending to visual nodes*. Different from our model attending to only textual nodes, we allow our decoder of this variant to consider both two types of nodes using doubly-attentive decoder. From line 6, we can observe that considering all nodes does not bring further improvement. The result confirms our previous assumption that visual information has been fully incorporated into textual nodes in our encoder.

(5) *attending to textual nodes* $\Rightarrow$ *attending to visual nodes*. However, when only considering visual nodes, the model performance drops drastically (line 7). This is because the number of visual nodes is far fewer than that of textual nodes, which is unable to produce sufficient context for translation.

3.5 Case Study

Figure 6 displays the 1-best translations of a sampled test sentence generated by different models. The phrase "*a skateboarding ramp*" is not translated correctly by all baselines, while our model correctly translates it. This reveals that our encoder is able to learn more accurate representations.

3.6 Results on the En⇒Fr Translation Task

We also conduct experiments on the EN⇒Fr dataset. From Table 3, our model still achieves better performance compared to all baselines, which demonstrates again that our model is effective and general to different language pairs in multi-modal NMT.

4 Related Work

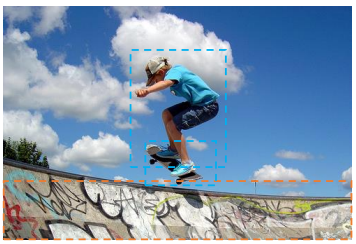
Multi-modal NMT Huang et al. (2016) first incorporate global or regional visual features into attention-based NMT. Calixto and Liu (2017) also study the effects of incorporating global visual features into different NMT components. Elliott and Kádár (2017) share an encoder between a translation model and an image prediction model to learn visually grounded representations. Besides, the most common practice is to use attention mechanisms to extract visual contexts for multi-modal NMT (Caglayan et al., 2016; Calixto et al., 2017; Delbrouck and Dupont, 2017a,b; Barrault et al., 2018). Recently, Ive et al. (2019) propose a translate-and-refine approach and Calixto et al. (2019) employ a latent variable model to capture the multi-modal interactions for multi-modal NMT. Apart from model design, Elliott (2018) reveal that visual information seems to be ignored by the multi-modal NMT models. Caglayan et al. (2019) conduct a systematic analysis and show that visual information can be better leveraged under limited textual context.

Different from the above-mentioned studies, we first represent the input sentence-image pair as a unified graph, where various semantic relationships between multi-modal semantic units can be effectively captured for multi-modal NMT. Benefiting from the multi-modal graph, we further introduce an extended GNN to conduct graph encoding via multi-modal semantic interactions.

Note that if we directly adapt the approach proposed by Huang et al. (2016) into Transformer, the model (ObjectAsToken(TF)) also involves multi-modal fusion. However, ours is different from it in following aspects: (1) We first learn the contextual representation of each node within the same modality, so that it can better determine the degree of inter-modal fusion according to its own context. (2) We assign different encoding parameters to different modalities, which has been shown effective in our experiments.

Additionally, the recent study LXMERT (Tan and Bansal, 2019) also models relationships between vision and language, which differs from ours in following aspects: (1) Tan and Bansal (2019) first apply two transformer encoders for two modalities, and then stack two cross-modality encoders to conduct multi-modal fusion. In contrast, we sequentially conduct self-attention and cross-modal gating at each layer. (2) Tan and Bansal (2019) leverage an attention mechanism to implicitly establish cross-modal relationships via large-scale pretraining, while we utilize visual grounding to capture explicit cross-modal correspondences. (3) We focus on multi-modal NMT rather than vision-and-language reasoning in (Tan and Bansal, 2019).

Graph Neural Networks Recently, GNNs (Marco Gori and Scarselli, 2005) including gated graph neural network (Li et al., 2016), graph convolutional network (Duvenaud et al., 2015; Kipf and Welling, 2017) and graph attention network (Velickovic et al., 2018) have been shown effective in many tasks such as VQA (Teney et al., 2017; Norcliffe-Brown et al., 2018; Li et al., 2019), text generation (Gildea et al., 2018; Becky et al., 2018; Song et al., 2018b, 2019) and text representation (Zhang et al., 2018; Yin et al., 2019; Song et al.,



Source:	A boy riding a skateboard on a skateboarding ramp .
Reference:	Ein junge fährt skateboard auf einer skateboardrampe .
Tranformer:	Ein junge fährt auf einem skateboard auf einer rampe .
Doubly-att(TF):	Ein junge fährt mit einem skateboard auf einer rampe .
Enc-att(TF):	Ein junge fährt ein skateboard auf einer rampe .
ObjectAsToken(TF):	Ein junge fährt auf einem skateboard auf einer rampe .
Our model:	Ein junge fährt auf einem skateboard auf einer skateboardrampe .

Figure 6: A translation example of different multi-modal NMT models. The baseline models do not accurately understand the phrase "*a skateboarding ramp*" (orange), while our model correctly translate it.

2018a; Xue et al., 2019).

In this work, we mainly focus on how to extend GNN to fuse multi-modal information in NMT. Close to our work, Teney et al. (2017) introduce GNN for VQA. The main difference between their work and ours is that they build an individual graph for each modality, while we use a unified multi-modal graph.

5 Conclusion

In this paper, we have proposed a novel graph-based multi-modal fusion encoder, which exploits various semantic relationships between multi-modal semantic units for NMT. Experiment results and analysis on the Multi30K dataset demonstrate the effectiveness of our model.

In the future, we plan to incorporate attributes of visual objects and dependency trees to enrich the multi-modal graphs. Besides, how to introduce scene graphs into multi-modal NMT is a worthy problem to explore. Finally, we will apply our model into other multi-modal tasks such as multi-modal sentiment analysis.

Acknowledgments

This work was supported by the Beijing Advanced Innovation Center for Language Resources (No. TYR17002), the National Natural Science Foundation of China (No. 61672440), and the Scientific Research Project of National Language Committee of China (No. YB135-49).

References

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 WMT 2018*, pages 304–323.

Daniel Becky, Gholamreza Haffariz, and Trevor Cohn. 2018. Graph-to-sequence learning using gated graph neural networks. In *Proceedings of ACL 2018*, pages 273–283.

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 WMT 2017*, pages 432–439.

Ozan Caglayan, Loïc Barrault, and Fethi Bougares. 2016. Multimodal attention for neural machine translation. *CoRR*, abs/1609.03976.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of NAACL-HLT 2019*, pages 4159–4170.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of ACL 2017*, pages 992–1003.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of ACL 2017*, pages 1913–1924.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of ACL 2019*, pages 6392–6405.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017a. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of EMNLP 2017*, pages 910–919.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017b. Modulating and attending the source image during encoding improves multimodal translation. *CoRR*, abs/1712.03449.

Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of WMT 2014*, pages 376–380.

David K. Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P. Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Proceedings of NeurIPS 2015*, pages 2224–2232.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of EMNLP 2018*, pages 2974–2978.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of ACL 2016*, pages 70–74.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of IJCNLP 2017*, pages 130–141.

Daniel Gildea, Zhiguo Wang, Yue Zhang, and Linfeng Song. 2018. A graph-to-sequence model for amr-to-text generation. In *Proceedings of ACL 2018*, pages 1616–1626.

Jindrich Helcl, Jindrich Libovický, and Dusan Varis. 2018. CUNI system for the WMT18 multimodal translation task. In *Proceedings of WMT 2018*, pages 616–623.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of WMT 2016*, pages 639–645.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of ACL 2019*, pages 6525–6538.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. In *Proceedings of NeurIPS 2018*, pages 1571–1581.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of ICLR 2017*.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of ECCV 2018*, pages 212–228.

Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-aware graph attention network for visual question answering. In *Proceedings of ICCV 2019*, pages 10312–10321.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. 2016. Gated graph sequence neural networks. In *Proceedings of ICLR 2016*.

Gabriele Monfardini Marco Gori and Franco Scarselli. 2005. A new model for learning in graph domains. In *Proceedings of IJCNN 2005*, pages 729–734.

Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. 2018. Learning conditioned graph structures for interpretable visual question answering. In *Proceedings of NeurIPS 2018*, pages 8344–8353.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of ACL 2002*, pages 311–318.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: towards real-time object detection with region proposal networks. In *Proceedings of NeurIPS 2015*, pages 91–99.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of ACL 2016*, pages 1715–1725.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using amr. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Linfeng Song, Zhiguo Wang, Mo Yu, Yue Zhang, Radu Florian, and Daniel Gildea. 2018a. Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks. *arXiv preprint arXiv:1809.02040*.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018b. A graph-to-sequence model for amr-to-text generation. In *Proceedings of ACL 2018*, pages 1616–1626.

Hao Tan and Mohit Bansal. 2019. LXMERT: learning cross-modality encoder representations from transformers. In *Proceedings of EMNLP 2019*, pages 5099–5110.

Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of CVPR 2018*, pages 4223–4232.

Damien Teney, Lingqiao Liu, and Anton van den Hengel. 2017. Graph-structured representations for visual question answering. In *Proceedings of CVPR 2017*, pages 3233–3241.

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 *Proceedings of NeurIPS 2017*, pages 4831–4836.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of ICLR 2018*.

Mengge Xue, Weiming Cai, Jinsong Su, Linfeng Song, Yubin Ge, Yubao Liu, and Bin Wang. 2019. Neural collective entity linking based on recurrent random walk network learning. In *Proceedings of IJCAI 2019*, pages 5327–5333.

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 *Proceedings of ICCV 2019*, pages 4682–4692.

Yongjing Yin, Linfeng Song, Jinsong Su, Jiali Zeng, Chulun Zhou, and Jiebo Luo. 2019. Graph-based neural sentence ordering. In *Proceedings of IJCAI 2019*, pages 5387–5393.

Yue Zhang, Qi Liu, and Linfeng Song. 2018. Sentence-state lstm for text representation. In *Proceedings of ACL 2018*, pages 317–327.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of EMNLP 2018*, pages 3643–3653.