Feature-level Incongruence Reduction for Multimodal Translation

Zhifeng Li and **Yu Hong**✉ and **Yuchen Pan** and **Jianmin Yao** and **Guodong Zhou**
Institute of Artificial Intelligence, Soochow University
School of Computer Science and Technology, Soochow University
No.1, Shizi ST, Suzhou, China, 215006
{lizhifeng0915,tianxianer,yuchenpan59419}@gmail.com
johnnytang1120@gmail.com, {gdzhou,jyao}@suda.edu.cn

Abstract

Caption translation aims to translate image annotations (captions for short). Recently, Multimodal Neural Machine Translation (MNMT) has been explored as the essential solution. Besides of linguistic features in captions, MNMT allows visual (image) featuresto be used. The integration of multimodal features reinforces the semantic representation and considerably improves translation performance. However, MNMT suffers from the incongruence between visual and linguistic features. To overcome the problem, we propose to extend MNMT architecture with a harmonization network, which harmonizes multimodal features (linguistic and visual features) by unidirectional modal space conversion. It enables multimodal translation to be carried out in a seemingly monomodal translation pipeline. We experiment on the golden Multi30k-16 and 17. Experimental results show that, compared to the baseline, the proposed method yields the improvements of 2.2% BLEU for the scenario of translating English captions into German (En→De) at best, 7.6% for the case of English-to-French translation (En→Fr) and 1.5% for English-to-Czech (En→Cz). The utilization of harmonization network leads to the competitive performance to the-state-of-the-art.

1 Introduction

Caption translation is required to translate a source-language caption into target-language, where a caption refers to the sentence-level text annotation of an image. As defined in the shared multimodal translation task[1] in WMT, caption translation can be conducted over both visual features in images and linguistic features of the accompanying captions. The question of how to opportunely utilize images for caption translation motivates the study of multimodality, including not only the extraction of visual features but the cooperation between visual and linguistic features. In this paper, we follow the previous work (Specia et al., 2016) to boil caption translation down to a problem of multimodal machine translation.

So far, a large majority of previous studies tend to develop a neural network based multimodal machine translation model (viz., MNMT), which consists of three basic components:

- **Image encoder** which characterizes a captioned image as a vector of global or multi-regional *visual features* using a convolutional neural network (CNN) (Huang et al., 2016).
- **Neural translation network** (Caglayan et al., 2016; Sutskever et al., 2014; Bahdanau et al., 2014) which serves both to encode a source-language caption and to generate the target-language caption by decoding, where the latent information that flows through the network is referred to *linguistic feature*.
- **Multimodal learning network** which uses visual features to enhance the encoding of linguistic semantics (Ngiam et al., 2011). Besides of the concatenation and combination of linguistic and visual features, vision-to-language attention mechanisms serve as the essential operations for cross-modality learning. Nowadays, they are implemented with single-layer attentive (Caglayan et al., 2017a; Calixto et al., 2017b), doubly-attentive (Calixto et al., 2017a), interpolated (Hitschler et al., 2016) and multi-task (Zhou et al., 2018) neural networks, respectively.

Multimodal learning networks have been successfully grounded with different parts of various neural translation networks. They are proven effective in enhancing translation performance. Nevertheless, the networks suffer from incongruence between visual and linguistic features because:

- Visual and linguistic features are projected into incompatible semantic spaces and therefore fail to be corresponded to each other.

[1] http://www.statmt.org/wmt16/

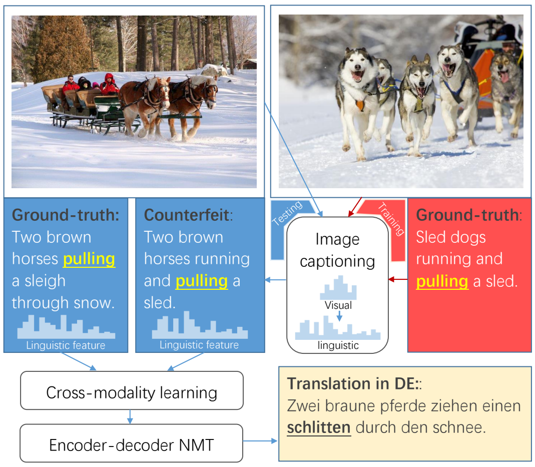

Ground-truth:
Two brown horses pulling a sleigh through snow.
Linguistic feature
Counterfeit:
Two brown horses running and pulling a sled.
Linguistic feature
Testing
Training
Image captioning
Visual
linguistic
Ground-truth:
Sled dogs running and pulling a sled.
Cross-modality learning
Encoder-decoder NMT
Translation in DE::
Zwei braune pferde ziehen einen schlitten durch den schnee.

Figure 1: An example in which image captioning contributes to the reduction of incongruence.

- Linguistic features are sequence-dependent. This is attributable to pragmatics, syntax or even rhetoric. On the contrary, visual features are sequence-independent but position-sensitive. This is attributable to spatial relationships of visual elements. Thus, a limited number of visual features can be directly used to improve the understanding of linguistic features and translation.

Considering the Figure 1(“*Counterfeit*” means Image Captioning output), the visual features enable a image processing model to recognize “*two horses*” as well as their position relative to a “*sleigh*”. However, such features are obscure for a translation model and useful for translating a verb, such as “*pulling*” in the caption. In this case, incongruence of heterogeneous features results from the unawareness of the correspondence between spatial relationship (“*running horses*” ahead of “*sleigh*”) and linguistic semantics (“*pulling*”).

To ease the incongruence, we propose to equip the current MNMT with a harmonization network, in which visual features are not directly introduced into the encoding of linguistic semantics. Instead, they are transformed into linguistic features before absorbed into semantic representations. In other words, we tend to make a detour during the cross-modality understanding, so as to bypass the modality barrier (Figure 2). In our experiments, we employ a captioning model to conduct harmonization. The hidden states it produced for decoding caption words are intercepted and involved into the representation learning process of MNMT.

The rest of the paper is organized as follows: Section 2 presents the motivation and methodological framework. Section 3 gives the NMT model we use. In Section 4, we introduce the captioning model that is trainable for cross-modality feature space transformation. Section 5 presents the captioning based harmonization networks as well as the resultant MNMT models. We discuss test results in Section 6 and overview the related work in Section 7. We conclude the paper in section 8.

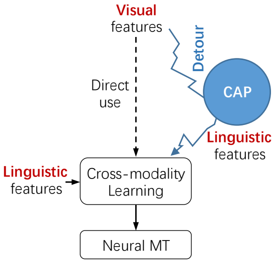


Figure 2: Bypassing modality barrier by captioning.

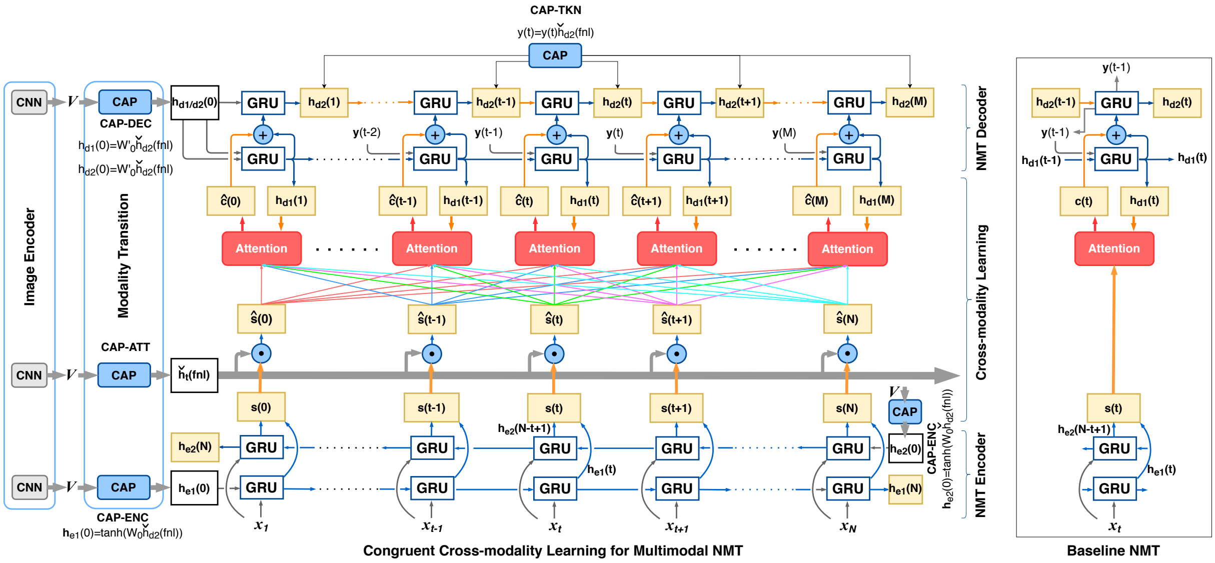
2 Fundamentals and Methodological Framework

We utilize Anderson et al (2018)’s image captioning (CAP for short) to guide the cross-modality feature transformation, converting visual features into linguistic. CAP is one of the generation models which are specially trained to generate language conditioned on visual features of images. Ideally, during training, it learns to perceive the correspondence between visual and linguistic features, such as that between the spatial relationship of “*running dogs ahead of a sled*” in Figure 1 and the meaning of the verb “*pulling*”. This allows CAP to produce appropriate linguistic features during testing in terms of similar visual features, such as that in the case of predicting the verb “*pulling*” for the scenario of “*running horses ahead of a sleigh*”.

Methodologically speaking, we adopt the linguistic features produced by the encoder of CAP instead of the captions generated by the decoder of CAP. On the basis, we integrate both the linguistic features of the original source-language caption and those produced by CAP into Calixto et al (2017b)’s attention-based cross-modality learning model (see Figure 3). Experimenal results show that the learning model substantially improves Bahdanau et al (2014)’s encoder-decoder NMT system.

3 Preliminary 1: Attentive Encoder-Decoder NMT (Baseline)

We take Bahdanau et al. (2014)’s attentive encoder-decoder NMT as the baseline. It is constructed



Congruent Cross-modality Learning for Multimodal NMT

Baseline NMT

Figure 3: The overall architecture of MNMT.

with a BiGRU encoder and a Conditional GRU (CGRU) decoder (Firat and Cho, 2016; Caglayan et al., 2017a). Attention mechanism is used between BiGRU and CGRU. The diagram at the right side of Figure 3 shows the baseline framework.

For a source-language caption, we represent it with a sequence of randomly-initialized (Kalchbrenner and Blunsom, 2013) word embeddings X=$(x_1, ..., x_N)$, where each x_t is uniformly specified as a k-dimensional word embedding. Conditioned on the embeddings, Chung et al (2014)'s BiGRU is used to compute the bidirectional hidden states S=$(s_1, ..., s_N)$, where each s_t is obtained by combining the t-th hidden state of forward GRU and that of backward GRU: s_t=$[\overrightarrow{GRU}^e(x_t), \overleftarrow{GRU}^e(x_t)]$. Padding (Libovickỳ and Helcl, 2018) and dynamic stabilization (Ba et al., 2016) are used.

Firat and Cho (2016)'s CGRU is utilized for decoding, which comprises two forward GRU units, i.e., $\overrightarrow{GRU}^{d1}$ and $\overrightarrow{GRU}^{d2}$ respectively. $\overrightarrow{GRU}^{d1}$ plays the role of producing the inattentive decoder hidden states H^{d1}=$(h_1^{d1}, ..., h_M^{d1})$, where each h_t^{d1} is computed based on the output state h_{t-1}^{d1} and prediction y_{t-1} at the previous time step: h_t^{d1}=$\overrightarrow{GRU}^{d1}(h_{t-1}^{d1}, y_{t-1})$ (Note: the prediction y_t denotes the k-dimensional embedding of the predicted word at the t-th decoding step). By contrast, $\overrightarrow{GRU}^{d2}$ serves to produce the attentive decoder hidden states H^{d2}=$(h_1^{d2}, ..., h_M^{d2})$, where each h_t^{d2} is computed conditioned on the previous attentive state h_{t-1}^{d2}, the current inattentive state h_t^{d1}, as well as the current attention-aware context c_t: h_t^{d2}=$\overrightarrow{GRU}^{d2}(h_{t-1}^{d2}, h_t^{d1} \oplus c_t)$. The context c_t is obtained by the attention mechanism over the global encoder hidden states S: $c_t = \alpha_t S$, where α_t denotes the attention weight at the t-th time step. Eventually, the prediction of each target-language word is carried out as follows (where, W_h, W_c, W_y, b_o and b_y are trainable parameters):

$$\mathcal{D}_t(y_{t-1}, h_t^{d2}, c_t) \sim \begin{cases} o_t = tanh(y_{t-1} + \mathrm{W}_h h_t^{d2} + \\ \qquad \mathrm{W}_c c_t + b_o) \\ P(y_t|o_t) = \mathrm{softmax}(\mathrm{W}_y^{\top} o_t \\ \qquad + b_y) \end{cases} \tag{1}$$

4 Preliminary 2: Image-dependent Linguistic Feature Acquisition by CAP

For an image, captioning models serve to generate a sequence of natural language (caption) that describes the image. Such kind of models are capable of transforming visual features into linguistic features by encoder-decoder networks. We utilize Anderson et al. (2018)'s CAP to obtain the transformed linguistic features.

4.1 CNN based Image Encoder

What we feed into CAP is a full-size image which needs to be convolutionally encoded beforehand. He et al (He et al., 2016a)'s CNNs (known as ResNet) with deep residual learning mechanism (He et al., 2016b) is capable of encoding images. In our experiments, we employ the recent version of ResNet, i.e., ResNet-101 , which is constructed with 101 convolutional layers. It is pretrained on ImageNet (Russakovsky et al., 2015) in the scenario of 1000-class image classification.

Using ResNet-101, we characterize an image as a convolutional feature matrix: $V \in \mathbb{R}^{k\times 2048} = \{v_1, ..., v_k\}$, in which each element $v_i \in \mathbb{R}^{2048}$ is a real-valued vector and corresponds to an image region in the size of 14×14 pixels.

4.2 Top-down Attention-based CAP

CAP learns to generate a caption over V. It is constructed with two-layer RNNs with LSTM (Anderson et al., 2018), LSTM1 and LSTM2 respectively. LSTM1 (in layer-1) computes the current first-layer hidden state $\check{h}_t^{d1}$ conditioned on the current first-layer input $\check{x}_t^{d1}$ and previous hidden state $\check{h}_{t-1}^{d1}$: $\check{h}_t^{d1}$=LSTM1($\check{x}_t^{d1}$, $\check{h}_{t-1}^{d1}$). The input $\check{x}_t^{d1}$ is obtained by concatenating the previous hidden state $\check{h}_{t-1}^{d1}$ and previous prediction $\check{y}_{t-1}$, as well as the condensed global visual feature $\bar{v}$: $\check{x}_t^{d1}$=[$\bar{v}$, $\check{h}_{t-1}^{d1}$, $\check{y}_{t-1}$], where $\bar{v}$ is calculated by the normalized accumulation of overall convolutional features in V: $\bar{v} = \frac{1}{k}\sum_i v_i$ ($\forall v_i \in V$). We specify the first-layer hidden state as the initial image-dependent linguistic features.

Attention mechanism (Sennrich et al., 2015) is used for highlighting the attention-worthy image context, so as to produce the attention-aware vector of image context $\check{v}_t$: $\check{v}_t = \sum \check{\alpha}_t V$. The attention weight $\check{\alpha}_t$ is obtained by aligning the current image-dependent hidden state $\check{h}_t^{d1}$ with every convoluted visual feature v_i: $\check{\alpha}_t = \text{softmax}.f(\check{h}_t^{d1}, v_i)$, where $f(*)$ is the non-linear activation function.

LSTM2 (in layer-2) serves as a neural language model (viz., language-oriented generation model). It learns to encode the current second-layer hidden state $\check{h}_t^{d2}$ conditioned on the current second-layer input $\check{x}_t^{d2}$ and previous hidden state $\check{h}_{t-1}^{d2}$: $\check{h}_t^{d2}$=LSTM1($\check{x}_t^{d2}$, $\check{h}_{t-1}^{d2}$). The input $\check{x}_t^{d2}$ is obtained by concatenating the current first-layer hidden state $\check{h}_t^{d1}$ (emitted from layer-1) and current attention-aware image context $\check{v}_t$: $\check{x}_t^{d2}$=[$\check{v}_t$, $\check{h}_t^{d1}$]. We specific a second-layer hidden state $\check{h}_t^{d2}$ as the image-dependent attention-aware linguistic features. Towards the image captioning task, CAP generally decodes the second-layer hidden states $\check{h}_t^{d2}$ to predict caption words. In our case, we tend to integrate them into multimodal NMT by cross-modality learning (see the next section).

5 Harmonization for MNMT

In the previous work of multimodal NMT, visual features in V are directly used for cross-modality learning. By contrast, we transform visual features into image-dependent attention-aware linguistic features (i.e., second-layer hidden states $\check{h}_t^{d2}$ emitted by CAP) before use. We provide four-class variants of cross-modality learning to improve NMT. They absorb image-dependent attention-aware linguistic features in different ways, including a variant that comprises attentive feature fusion (CAP-ATT) and three variants (CAP-ENC, CAP-DEC and CAP-TKN) which carry out reinitialization and target-language embedding modulation. Figure 3 shows the positions in the baseline NMT where the variants come into play.

CAP-ATT intends to improve NMT by conducting joint representation learning across the features of the source-language caption and that of the accompanying image. On one side, CAP-ATT adopts the encoder hidden state s_t (emitted by the BiGRU encoder of the baseline NMT) and uses it as the language-dependent linguistic feature. On the other side, it takes the image-dependent attention-aware linguistic feature $\check{h}_t^{d2}$ (produced by CAP). We suppose that the two kinds of features (i.e., $\check{h}_t^{d2}$ and s_t) are congruent with each other. On the basis, CAP-ATT blends $\check{h}_t^{d2}$ into s_t to form the joint representation $\widehat{s}_t$. Element-wise feature fusion (Cao and Xiong, 2018) is used to compute $\widehat{s}_t$: $\widehat{s}_t = s_t \odot \check{h}_t^{d2}$. Using the joint representation $\widehat{s}_t$, CAP-ATT updates the attention-aware context c_t which is fed into the CGRU decoder of the baseline NMT: $\widehat{c}_t$=$\alpha_t\widehat{S}$, $\forall\widehat{s} \in \widehat{S}$. By substituting the updated context $\widehat{c}_t$ into the computation of the CGRU decoder, CAP-ATT further refines the decoder hidden state $h_t^{d_2}$ and prediction of target-language words. Equation 2 formulates the decoding process, where $\mathcal{D}_t$ is the shorthand of equation (1).

$$\widehat{\mathcal{D}}_t(y_{t-1}, \widehat{h}_t^{d2}, \widehat{c}_t) \sim \begin{cases} \widehat{h}_t^{d_2} = \overrightarrow{GRU}^{d2}(\widehat{h}_{t-1}^{d_2}, h_t^{d_1} \\ \qquad \oplus\, \widehat{c}_t) \\ y_t \Leftarrow \mathcal{D}_t(y_{t-1}, \widehat{h}_t^{d_2}, c_t) \end{cases} \tag{2}$$

CAP-ENC reinitializes the BiGRU encoder of the baseline NMT with the final image-dependent attention-aware linguistic feature $\check{h}_t^{d2}$ (t=N) (produced by CAP): $\overleftarrow{h}_0 = \overrightarrow{h}_0 = \tanh(\mathrm{W}_0\check{h}_t^{d2})$, where $\overleftarrow{h}_0$ and $\overrightarrow{h}_0$ are the initial states of BiGRU, and W_0 refers to the trainable parameter. **CAP-**

DEC uses $\check{h}_t^{d2}$ (t=N) to reinitialize the CGRU decoder of the baseline NMT: $h_0^{d1} = h_0^{d2} = \tanh(\mathrm{W}_0'\check{h}_t^{d2})$, where h_0^{d1} and h_0^{d2} are the initial decoder hidden states of CGRU. Using $\check{h}_t^{d2}$ (t=N), **CAP-TKN** modulates the predicted target-language word embedding y_t at each decoding step: $y_t = y_t \odot \tanh(\mathrm{W}_{tkn}\check{h}_t^{d2})$, where W_{tkn} is the trainable parameter. **CAP-ALL** equips a MNMT system with all the variants.

6 Experimentation

6.1 Resource and Experimental Datasets

We perform experiments on Multi30k-16 and Multi30k-17[2], which are provided by WMT for the shared tasks of multilingual captioning and multimodal MT (Elliott et al., 2016). The corpora are used as the extended versions of Flichr30k (Young et al., 2014), since they contain not only English (En) image captions but their translations in German (De), French (Fr) and Czech (Cz). Hereinafter, we specify an example in Multi30k as an image which is accompanied by three En→De, En→Fr and En→Cz caption-translation pairs. Each of Multi30k-*16* and Multi30k-*17* contains about 31K examples. We experiment on the corpora separately, and as usual divide each of them into training, validation and test sets, at the scale of 29K, 1,014 and 1K examples, respectively.

In addition, we carry out a complementary experiment on the ambiguous COCO which contains 461 examples (Elliott et al., 2017). Due to the inclusion of ambiguous verbs, the examples in ambiguous COCO can be used for the evaluation of visual sense disambiguation in a MNMT scenario.

6.2 Training and Hyperparameter Settings

For preprocessing, we apply Byte-Pair Encoding (BPE) (Sennrich et al., 2015) for tokenizing all the captions and translations in Multi30k and COCO, and use the open-source toolkit[3] of Moses (Koehn et al., 2007) for lowercasing and punctuation normalization. It reproduces the neural network architecture of Anderson et al (Anderson et al., 2018)'s top-down attentive CAP. The only difference is that it merely utilizes ResNet-101 in generating the input set of visual features V, without the use of Faster R-CNN (Ren et al., 2015). This CAP has been trained on MSCOCO captions dataset (Lin et al., 2014) using the same hyperparameter settings as that in Anderson et al. (2018)'s work.

Besides of the baseline NMT (Bahdanau et al., 2014) mentioned in section 2, we compare our model with Caglayan et al (Caglayan et al., 2017a)'s convolutional visualfeature based MNMT. In this paper, we follow Caglayan et al (Caglayan et al., 2017a)'s practice to implement and train our model. First of all, we implement our model with the nmtpy framework (Caglayan et al., 2017b) using Theano v0.9. During training, ADAM with a learning rate of 4e-4 is used and the batch size is set as 32. We initialize all the parameters (i.e., transformation matrices and biases) using Xavier and clip the total gradient norm to 5. We drop out the input embeddings, hidden states and output states with the probabilities of (0.3, 0.5, 0.5) for En→De MT, (0.2, 0.4, 0.4) for En→Fr and (0.1, 0.3, 0.3) for En→Cz. In order to avoid overfitting, we apply a L_2 regularization term with a factor of $1e$-5. We specify the dimension as 128 for all token embeddings (k = 128) and 256 for hidden states.

6.3 Comparison to the Baseline

We carry out 5 independent experiments (5 runs) for each of the proposed MNMT variants. In each run, any of the variants is retrained and redeveloped under cold-start conditions using a set of randomly-selected seeds by MultEval[4]. Eventually, the resultant models are evaluated on the test set with BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014) and TER (Snover et al., 2006).

For each variant, we report not only the comprehensive performance (denoted as $ensemble$) which is obtained using ensemble learning (Garmash and Monz, 2016) but that without ensemble learning. In the latter case, the average performance (μ) and deviations (σ) in the 5 runs are reported.

6.3.1 Performance on Multi30k

Tables 1 and 2 respectively show the performance of our models on Multi30k-16 and Multi30k-17 for the translation scenarios of En→De, En→Fr and En→Cz. Each of our MNMT models in the tables is denoted with a symbol "+", which indicates that a MNMT model is constructed with the baseline and one of our cross-modality learning models. The baseline is specified as the monomodal NMT model which is developed by Bahdanau et al. (2014) (as

[2] https://github.com/multi30k/dataset/tree/master/data-/task1/raw

[3] https://github.com/moses-smt/mosesdecoder/tree/-master/scripts/tokenizer

[4] https://github.com/jhclark/multeval

En→De	Multi30k-*16* ($\mu \pm \sigma$/ensemble)			Multi30k-*17* ($\mu \pm \sigma$/ensemble)		
	BLEU	METEOR	TER	BLEU	METEOR	TER
Baseline	38.1±0.8/40.7	57.3±0.5/59.2	N/A	30.8±1.0/33.7	51.6±0.5/53.8	N/A
+CAP−ATT	39.2±0.8/41.3	57.5±0.6/59.4	40.9±0.8/39.5	32.1±0.9/33.6	51.0±0.7/52.9	48.7±0.8/47.3
+CAP−ENC	39.1±0.8/41.2	57.6±0.7/59.2	40.9±0.8/39.3	32.5±0.8/33.8	**52.2±0.7/54.5**	**48.5±0.8/46.3**
+CAP−DEC	38.9±0.8/41.0	57.4±0.7/59.3	**41.3±0.8/39.1**	**33.0±0.8/34.3**	51.6±0.7/53.2	48.6±0.8/47.1
+CAP−TKN	39.1±0.8/40.9	57.3±0.6/58.6	**41.3±0.8/39.1**	32.2±0.8/33.9	51.3±0.7/53.5	48.5±0.8/47.0
+CAP−ALL	**39.6±0.9/42.1**	**57.5±0.7/59.9**	41.1±0.8/39.4	31.6±0.8/33.9	51.6±0.7/53.7	49.7±0.7/47.1

En→Fr	Multi30k-*16* ($\mu \pm \sigma$/ensemble)			Multi30k-*17* ($\mu \pm \sigma$/ensemble)		
	BLEU	METEOR	TER	BLEU	METEOR	TER
Baseline	52.5±0.3/54.3	69.6±0.1/71.3	N/A	50.4±0.9/53.0	67.5±0.7/69.8	N/A
+CAP−ATT	**60.1±0.8/63.3**	**74.3±0.6/77.1**	**25.1±0.7/22.7**	52.5±0.9/56.1	68.2±0.7/71.2	31.5±0.7/28.4
+CAP−ENC	59.3±0.9/62.8	73.5±0.6/76.4	26.2±0.7/23.3	52.2±0.8/55.8	68.1±0.7/71.1	31.5±0.7/28.5
+CAP−DEC	60.1±0.9/62.6	74.2±0.7/76.3	25.6±0.6/23.0	51.9±0.9/55.7	67.6±0.7/71.3	**31.6±0.7/28.1**
+CAP−TKN	60.3±0.8/63.0	74.5±0.6/76.6	25.2±0.6/23.0	52.7±0.9/56.0	**68.3±0.6/71.3**	31.5±0.7/28.6
+CAP−ALL	60.1±0.8/62.7	74.3±0.6/76.4	25.0±0.4/23.1	**52.8±0.9/56.1**	68.6±0.6/71.1	31.2±0.7/28.9

Table 1: Performance for both En→De and En→Fr on Multi30k-16 and Multi30k-17.

En→Cz	Multi30k(2016) ($\mu \pm \sigma$ /ensemble)		
	BLEU	METEOR	TER
Baseline	30.5±0.8/32.6	29.3±0.4/31.4	N/A
+CAP−ATT	31.8±0.9/33.4	**30.2±0.4/32.6**	46.1±0.8/43.6
+CAP−ENC	31.7±0.8/33.3	29.9±0.4/32.1	46.3±0.8/43.5
+CAP−DEC	31.6±0.9/33.3	30.0±0.4/32.3	45.6±0.8/43.6
+CAP−TKN	**32.0±0.9/33.9**	30.1±0.4/32.3	45.7±0.8/43.3
+CAP−ALL	31.8±0.9/33.6	29.9±0.4/31.5	**45.3±0.8/43.3**

Table 2: Performance for En→Cz on Multi30k-16

mentioned in section 2) and redeveloped as the baselines in a variety of research studies on multimodal NMT (Calixto et al., 2017a,b; Caglayan et al., 2017a). We quote the results reported in Caglayan et al. (2017a)'s work as they were better.

It can be observed that our MNMT models outperform the baseline. They benefits from the performance gains yielded by the variants of CAP based cross-modality learning, which are no less than 1.5% BLEU when ensemble learning is used, and 0.6% when not to use it. In particular, +CAP-ATT obtains a performance increase of up to 7.6% BLEU (μ) in the scenario of En→Fr MT. The gains in METEOR score we obtain are less obvious than that in BLEU, which is about 5.3% (μ) at best.

We follow Clark et al. (2011) to perform significance test. The test results show that +CAP-ATT, +CAP-DEC and +CAP-TKN achieve a p-value of 0.02, 0.01 and 0.007, respectively. Clark et al. (2011) have proven that the performance improvements are significant only if the p-value is less than 0.05. Therefore, the proposed method yields statistically significant performance gains.

6.3.2 Performance on Ambiguous COCO

Table 3 shows the translation performance. It can be found that our models yield a certain amount of gains (in BLEU scores) for En→De translation, and raise both BLEU and METEOR scores for En→Fr. The METER scores for En→De are comparable to that the baseline achieved. However, the improvement is less significant compared to that obtained on Multi30k-16&17 (see Table 1). Considering that the ambiguous COCO contains a larger number of ambiguous words than Multi30k-16&17, we suggest that our method fails to largely shield the baseline from the misleading of ambiguous words.

Nevertheless, our method doesn't result in a two-fold error propagation, but on the contrary it alleviates the negative influences of the errors because:

- Error propagation, in general, is inevitable when a GRU or LSTM unit is used. Both are trained to predict a sequence of words one by one. Appropriate prediction of previous words is crucial for ensuring the correctness of subsequent words. Thus, once a mistake is made at a certain decoding step, the error will be propagated forward, and mislead the prediction of subsequent words.

- The baseline is equipped with a GRU decoder and therefore suffers from error propagation. More seriously, ambiguous words increase the risk of error propagation. This causes a significant performance reduction on Ambiguous COCO. For example, the BLEU score for En→De is 28.7% at best. It is far below that (40.7%) obtained on Multi30k-16&17.

- Two-fold error propagation is suspected to occur when LSTM-based CAP is integrated with the baseline. Though the opposite is actually true. After CAP is used, the translation performance is improved instead of falling down.

6.4 Comparison to the state of the art

We survey the state-of-the-art research activities in the field of MNMT, and compare them with ours

Ambiguous coco (2017)	En→De ($\mu \pm \sigma$/ensemble) BLEU	METEOR	TER	En→Fr ($\mu \pm \sigma$/ensemble) BLEU	METEOR	TER
Helcl et al (2017)	25.7	45.6	N/A	43.0	62.5	N/A
Caglayan et al (2017)	29.4$^\diamond$	49.2$^\diamond$	N/A	46.2$^\diamond$	66.0$^\diamond$	N/A
Zhou et al (2018)	28.3	48.0	N/A	45.0	64.7	N/A
Baseline	26.4±0.2/28.7	46.8±0.7/48.9	N/A	41.2±1.2/43.3	61.3±0.9/63.3	N/A
+CAP-ATT	27.1±1.2/29.3	47.7±0.9/48.8	**53.0±1.1/50.7**	43.8±1.2/46.8	62.2±0.9/65.0	36.5±1.0/34.5
+CAP-ENC	27.1±1.1/29.4	47.5±0.9/48.7	54.1±1.1/51.2	42.8±1.2/46.3	60.8±0.9/65.3	**38.1±1.0/33.4**
+CAP-DEC	**27.8±1.1/29.9**	**47.8±1.0/49.3**	53.8±1.1/50.8	43.2±1.2/46.1	61.5±0.9/65.3	37.3±1.0/34.3
+CAP-TKN	27.3±1.2/29.6	46.4±0.9/48.9	54.2±1.2/51.1	44.5±1.2/46.8	62.4±0.9/65.3	37.8±1.0/34.0
+CAP-ALL	27.6±1.1/29.8	46.4±0.9/48.9	54.4±1.2/50.8	**44.3±1.2/47.1**	**62.6±0.9/65.4**	36.4±1.0/33.5

Table 3: Performance on Amb-COCO (Note: "◇" is the sign of the performance when ensemble learning is used.)

En→De	Multi30k-*16* BLEU	METEOR	Multi30k-*17* BLEU	METEOR
Huang et al (2016)	36.5	54.1		
Calixto et al (2017a)	36.5	55.0		
Calixto et al (2017b)	41.3$^\diamond$	59.2$^\diamond$		
Elliott et al (2017)	40.2$^\diamond$	59.3$^\diamond$		
Helcl et al (2017)	34.6	51.7	28.5	49.2
Caglayan et al (2017a)	41.2$^\diamond$	59.4$^\diamond$	33.5$^\diamond$	53.8$^\diamond$
Helcl et al (2018)	38.7	57.2		
Zhou et al (2018)			31.6	52.2
Ours (μ)	39.6	57.5	33.0	52.2
Ours (ensemble)	**42.1**	**59.9**	**34.3**	**54.5**

En→Fr	Multi30k-*16* BLEU	METEOR	Multi30k-*17* BLEU	METEOR
Helcl et al (2017)			50.3	67.0
Caglayan et al (2017a)	56.7$^\diamond$	73.0$^\diamond$	55.7$^\diamond$	71.9$^\diamond$
Helcl et al (2018)	60.8	75.1		
Zhou et al (2018)			53.8	70.3
Ours (μ)	60.1	74.3	52.8	68.6
Ours (ensemble)	**63.3**	**77.1**	**56.1**	**71.1**

En→Cz	Multi30k-*16* BLEU	METEOR	Multi30k-*17* BLEU	METEOR
Helcl et al (2018)	31.0	29.9		
Ours (μ)	32.0	30.2		
Ours (ensemble)	**33.9**	**32.6**		

Table 4: Comparison results on Multi30k (Note: "◇" is the sign indicating the use of ensemble learning).

(as shown in Table 4). Comparison are made for all the WMT translation scenarios (En→De, Fr and Cz) on Multi30k-16&17 but merely for En→De and En→Fr on ambiguous COCO (as shown in Table 3). To our best knowledge, there is no previous attempt to evaluate the performance of an En→Cz translation model on ambiguous COCO, and thus a precise comparison for that is not available. It is noteworthy that some of the cited work reports the ensemble learning results for MNMT, others make no mention of it. We label the former with a symbol of "◇" in Tables 3 and 4 to ease the comparison.

It can be observed that our best model outperforms the state of the art for most scenarios over different corpora except the En→Fr case on Multi30k-17. The performance increases are most apparent in the case of En→Fr on Multi30k-16 when ensemble learning is used, where the BLEU and METEOR scores reach the levels of more than 63% and 77%, with the improvements of 6.6% and 4.1%.

We regard the work of Caglayan et al. (2017a) and Calixto et al. (2017a) as the representatives in our systematic analysis. Caglayan et al. (2017a) directly use raw visual features (i.e., V mentioned in section 3.1) to enhance NMT at different stages, including that of initialization, encoding and decoding. Calixto et al. (2017a) develop a doubly-attentive decoder, where both visual features of images and linguistic features of captions are used for computing the attention scores during decoding.

- **Caglayan et al. (2017a)'s model**: Caglayan et al. (2017a)'s model integrates visual features V into the decoding process. By contrast, we conduct the integration using linguistic features which are transformed from visual features. It is proven that our integration approach leads to considerable performance increases. Accordingly, we suppose that reducing incongruence between visual and linguistic features contributes to cross-modality learning in MNMT.

- **Calixto et al. (2017a)'s model**: Our CAP-ATT is similar to Calixto et al. (2017a)'s model due to the use of attention mechanisms during decoding. The difference is that CAP-ATT transforms visual features into linguistic features before attention computation. This operation leads to the increases of both BLEU (2.7%) and METEOR (2%) on Multi30k-16. The results demonstrate that attention scores can be computed more effectively between features of the same type.

6.5 Performance in Adversarial Evaluation

We examine the use efficiency of images for MNMT using Elliott's adversarial evaluation (Elliott, 2018). Elliott suppose that if a model efficiently uses images during MNMT, its performance would degrade when it is cheated by some incongruent images. Table 5 shows the test results, where "C" is specified as a METEOR score which is evaluated when there is not any incongruent image in the test set, while "I" is that when some incogruent images are used to replace the original images.

If the value of "C" is larger than "I", a positive $\triangle_E$-Awareness can be obtained. It illustrates an acceptable use efficiency. On the contrary, a negative $\triangle_E$-Awareness is a warning of low efficiency.

Table 5 shows the test results. It can be observed that our +CAP-ATT and +CAP-ALL models achieve positive $\triangle_E$-Awareness for all the translation scenarios on Multi30k-16. In addition, the models obtain higher values of $\triangle_E$-Awareness than Caglayan et al. (2017a)'s models of `decinit` and `hierattn`. As mentioned above, Caglayan et al directly use visual features to enhance the MNMT, while we use the image-dependent linguistic features that are transformed from visual features. Therefore, we suppose that modality transformation leads to a higher use efficiency of images.

En→De	Multi30k (2016) ($\mu \pm \sigma$)		
	C	I	$\triangle_E$-Awareness
+CAP-ATT	58.5	58.5±0.2	0.001 ±0.002
+CAP-ENC	57.8	58.5±0.1	-0.007 ±0.001
+CAP-DEC	58.3	58.0±0.0	0.020 ±0.001
+CAP-TKN	58.7	58.6±0.1	0.001 ±0.001
+CAP-ALL	**59.0**	**58.5**±**0.2**	**0.005** ±**0.002**
Caglayan et al′s trgmul	N/A	N/A	-0.001 ±0.002
Caglayan et al′s decinit	N/A	N/A	0.003 ±0.001
Helcl et al′s hierattn	N/A	N/A	0.019 ±0.003

En→Fr	Multi30k (2016) ($\mu \pm \sigma$)		
	C	I	$\triangle_E$-Awareness
+CAP-ATT	**74.8**	**74.2**±**0.1**	**0.005** ±**0.001**
+CAP-ENC	73.8	74.2±0.1	-0.004 ±0.001
+CAP-DEC	74.3	74.3±0.1	-0.001 ±0.001
+CAP-TKN	74.9	74.6±0.1	0.003 ±0.001
+CAP-ALL	74.8	74.5±0.1	0.003 ±0.001

En→Cz	Multi30k (2016) ($\mu \pm \sigma$)		
	C	I	$\triangle_E$-Awareness
+CAP-ATT	35.2	34.7±0.2	0.005 ±0.002
+CAP-ENC	34.7	34.4±0.1	0.003 ±0.001
+CAP-DEC	34.8	34.4±0.1	0.004 ±0.001
+CAP-TKN	34.9	35.1±0.1	-0.002 ±0.001
+CAP-ALL	**34.6**	**33.8**±**0.1**	**0.007** ±**0.001**

Table 5: Test results in Elliott's utility test.

7 RELATED WORK

We have mentioned the previous work of MNMT in section 1, where the research interest has been classified into image encoding, encoder-decoder NMT construction and cross-modality learning. Besides, we present the methods of Caglayan et al. (2017a) and Calixto et al. (2017a) in the section 4.4.2, along with the systematic analysis. Besides, many scholars within the research community have made great efforts upon the development of sophisticated NMT architectures, including multi-source (Zoph and Knight, 2016), multi-task (Dong et al., 2015) and multi-way (Firat et al., 2016) NMT, as well as those equipped with attention mechanisms (Sennrich et al., 2015). The research activities are particularly crucial since they broaden the range of cross-modality learning strategies.

Current research interest has concentrated on the incorporation of visual features into NMT (Lala et al., 2018), by means of visual-linguistic context vector concatenation (Libovickỳ et al., 2016), doubly-attentive decoding (Calixto et al., 2017a), hierarchical attention combination (Libovickỳ and Helcl, 2017), cross-attention network (Helcl et al., 2018), gated attention network (Zhang et al., 2019), joint (Zhou et al., 2018) and ensemble (Zheng et al., 2018) learning . In addition, image attention optimization (Delbrouck and Dupont, 2017) and monolingual data expansion (Hitschler et al., 2016) have been proven effective in this field. Ive et al. (2019) use an off-shelf object detector and an additional image dataset (Kuznetsova et al., 2018) to form a bag of category-level object embeddings. Conditioned on the embeddings, Ive et al. (2019) develop a sophisticated MNMT model which integrates self-attention and cross-attention mechanisms into the encoder-decoder based deliberation architecture.

This paper also touches on the research area of image captioning. Mao et al. (2014) provide an interpretable image modeling method using multimodal RNN. Vinyals et al. (2015) design a caption generator (IDG) by Seq2Seq framework. Further, Xu et al. (2015) propose an attention-based IDG.

8 CONCLUSION

We demonstrate that the captioning based harmonization model reduces incongruence between multimodal features. This contributes to the performance improvement of MNMT. It is proven that our method increases the use efficiency of images.

The interesting phenomenon we observed in the experiments is that modality incongruence reduction is more effective in the scenario of En→Fr translation than that of En→De and En→Cz. This raises a problem of adaptation to languges. In the future, we will study on the distinct grammatical and syntactic principles of target languages, as well as their influences on the adaptation. For example, the syntax of French can be considered as most strict. Thus, a sequence-dependent feature vector may be more adaptive to MNMT towards French. Accordingly, we will attempt to develop a generative adversarial network based adaptation enhancement model. The goal is to refine the generated linguistic features by learning to detect and eliminate the features of less adaptability.

Acknowledgements

We thank the reviewers for their insightful comments. The idea is proposed by the corresponding author (Yu Hong). Jian Tang provides an effective assistance for conducting the experiments. We thank the colleagues for their help.

This work was supported by the national Natural Science Foundation of China (NSFC) via Grant Nos. 62076174, 61773276, 61836007.

References

P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, volume 3, page 6.

J. L. Ba, J. R. Kiros, and G. E. Hinton. 2016. Layer normalization. *arXiv:1607.06450*.

D. Bahdanau, K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv:149.0473*.

O. Caglayan, W. Aransa, A. Bardet, M. García-Martínez, F. Bougares, L. Barrault, M. Masana, L. Herranz, and J. Van de Weijer. 2017a. Lium-cvc submissions for wmt17 multimodal translation task. pages 450–457.

O. Caglayan, W. Aransa, Y. Wang, M. Masana, M. García-Martínez, F. Bougares, L. Barrault, and J. Van de Weijer. 2016. Does multimodality help human and machine for translation and image captioning? *arXiv:1605.09186*.

O. Caglayan, M. García-Martínez, A. Bardet, W. Aransa, F. Bougares, and L. Barrault. 2017b. Nmtpy: A flexible toolkit for advanced neural machine translation systems. *The Prague Bulletin of Mathematical Linguistics*, 109(1):15–28.

I. Calixto, Q. Liu, and N. Campbell. 2017a. Doubly-attentive decoder for multi-modal neural machine translation. *arXiv:1702.01287*.

I. Calixto, Q. Liu, and N. Campbell. 2017b. Incorporating global visual features into attention-based neural machine translation. *arXiv:1701.06521*.

Q. Cao and D. Xiong. 2018. Encoding gated translation memory into neural machine translation. In *EMNLP*, pages 3042–3047.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv:1412.3555*.

J. H. Clark, C. Dyer, A. Lavie, Alon, and N. A. Smith. 2011. Better hypothesis testing for statistical machine translation: Controlling for optimizer instability. In *ACL*, pages 176–181. ACL.

J. Delbrouck and S. Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. *arXiv:1707.00995*.

M. Denkowski and A. Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *ICML*, pages 376–380.

D. Dong, H. Wu, W. He, D. Yu, and H. Wang. 2015. Multi-task learning for multiple language translation. In *ACL*, pages 1723–1732.

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

D. Elliott, S. Frank, L. Barrault, F. Bougares, and L. Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. *arXiv:1710.07177*.

D. Elliott, S. Frank, K. Sima'an, and L. Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv:1605.00459*.

O. Firat and K. Cho. 2016. Conditional gated recurrent unit with attention mechanism. *System BLEU baseline*, 31.

O. Firat, K. Cho, and Y. Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. *arXiv:1601.01073*.

E. Garmash and C. Monz. 2016. Ensemble learning for multi-source neural machine translation. In *COLING*, pages 1409–1418.

K. He, X. Zhang, S. Ren, and J. Sun. 2016a. Deep residual learning for image recognition. In *CVPR*, pages 770–778.

K. He, X. Zhang, S. Ren, and J. Sun. 2016b. Identity mappings in deep residual networks. In *ECCV*, pages 630–645. Springer.

J. Helcl, J. Libovickỳ, and D. Varis. 2018. Cuni system for the wmt18 multimodal translation task. In *WMT*, pages 616–623.

J. Hitschler, S. Schamoni, and S. Riezler. 2016. Multimodal pivots for image caption translation. *arXiv:1601.03916*.

P. Huang, F. Liu, S. Shiang, J. Oh, and C. Dyer. 2016. Attention-based multimodal neural machine translation. In *WMT*, pages 639–645.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. *arXiv preprint arXiv:1906.07701*.

N. Kalchbrenner and P. Blunsom. 2013. Recurrent continuous translation models. In *EMNLP*, pages 1700–1709.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, and et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL*, pages 177–180. Association for Computational Linguistics.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. 2018. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*.

C. Lala, P. S. Madhyastha, C. Scarton, and L. Specia. 2018. Sheffield submissions for wmt18 multimodal translation shared task. In *ICML*, pages 624–631.

J. Libovickỳ and J. Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. *arXiv:1704.06567*.

J. Libovickỳ and J. Helcl. 2018. End-to-end non-autoregressive neural machine translation with connectionist temporal classification. *arXiv:1811.04719*.

J. Libovickỳ, J. Helcl, M. Tlustỳ, P. Pecina, and O. Bojar. 2016. Cuni system for wmt16 automatic post-editing and multimodal translation tasks. *arXiv:1606.07481*.

T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer.

J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. 2014. Explain images with multimodal recurrent neural networks. *arXiv:1410.1090*.

J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. 2011. Multimodal deep learning. In *ICML*, pages 689–696.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318. Association for Computational Linguistics.

S. Ren, K. He, R. Girshick, and J. Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99.

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

R. Sennrich, B. Haddow, and A. Birch. 2015. Neural machine translation of rare words with subword units. *arXiv:1508.07909*.

M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *ICML*, volume 200.

L. Specia, S. Frank, K. Sima'an, and D. Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *ICML*, volume 2, pages 543–553.

I. Sutskever, O. Vinyals, and Q. V. Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164.

K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057.

P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *ACL*, 2:67–78.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2019. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

R. Zheng, Y. Yang, M. Ma, and L. Huang. 2018. Ensemble sequence level training for multimodal mt: Osu-baidu wmt18 multimodal machine translation system report. *arXiv:1808.10592*.

M. Zhou, R. Cheng, Y. J. Lee, and Z. Yu. 2018. A visual attention grounding neural model for multimodal machine translation. *arXiv:1808.08266*.

B. Zoph and K. Knight. 2016. Multi-source neural translation. *arXiv:1601.00710*.