Forward and Backward Multimodal NMT for Improved Monolingual and Multilingual Cross-Modal Retrieval

Po-Yao Huang
Carnegie Mellon University
Pittsburgh, PA
poyaoh@cs.cmu.edu

Xiaojun Chang
Monash University
Melbourne, VIC
cxj273@gmail.com

Alexander Hauptmann
Carnegie Mellon University
Pittsburgh, PA
alex@cs.cmu.edu

Eduard Hovy
Carnegie Mellon University
Pittsburgh, PA
hovy@cs.cmu.edu

ABSTRACT

We explore methods to enrich the diversity of captions associated with pictures for learning improved visual-semantic embeddings (VSE) in cross-modal retrieval. In the spirit of "A picture is worth a thousand words", it would take dozens of sentences to parallel each picture's content adequately. But in fact, real-world multimodal datasets tend to provide only a few (typically, five) descriptions per image. For cross-modal retrieval, the resulting lack of diversity and coverage prevents systems from capturing the fine-grained inter-modal dependencies and intra-modal diversities in the shared VSE space. Using the fact that the encoder-decoder architectures in neural machine translation (NMT) have the capacity to enrich both monolingual and multilingual textual diversity, we propose a novel framework leveraging multimodal neural machine translation (MMT) to perform forward and backward translations based on salient visual objects to generate additional text-image pairs which enables training improved monolingual cross-modal retrieval (English-Image) and multilingual cross-modal retrieval (English-Image and German-Image) models. Experimental results show that the proposed framework can substantially and consistently improve the performance of state-of-the-art models on multiple datasets. The results also suggest that the models with multilingual VSE outperform the models with monolingual VSE.

KEYWORDS

Cross-modal Retrieval, Joint Embedding, Multilingual Multimodal Representation, Multimodal Machine Translation.

ACM Reference Format:
Po-Yao Huang, Xiaojun Chang, Alexander Hauptmann, and Eduard Hovy. 2020. Forward and Backward Multimodal NMT for Improved Monolingual and Multilingual Cross-Modal Retrieval. In *2020 International Conference on Multimedia Retrieval (ICMR'20), June 8–11, 2020, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3372278.3390674

ICMR '20, June 8–11, 2020, Dublin, Ireland

ACM ISBN 978-1-4503-7087-5/20/06…$15.00
https://doi.org/10.1145/3372278.3390674

1 INTRODUCTION

With the exploding amount of visual and textual content available nowadays, bridging the gap between vision and language has become an important but long-standing challenge given the richness and diversity within the two modalities. Cross-modal retrieval models, which learn to encode the textual and visual perspectives of instances into the joint visual-semantic embedding (VSE) space, serve as the important bridges connecting the two worlds.

Learning a robust joint embedding space requires an abundant amount of parallel data [17, 45]. However, "A picture is worth a thousand words", in most existing multimodal datasets [18, 30, 46], there are typically at most five image-text pairs available due to the cost of human annotation. The lack of coverage and diversity in the text descriptions thus inevitably constraints model performance in cross-modal tasks such as text-to-image retrieval and captioning.

As analyzed by Dai *et al.* [8, 9], captioning models tend to overfit the ground-truth annotations and produce less distinctive descriptions while discouraging other reasonable alternatives. For cross-modal retrieval, Huang *et al.* [17] quantifies the impact of limited text content for learning the VSE space. Similar issues have been observed in neural machine translation (NMT) [3, 7] which relies on a plethora of parallel corpora for training. Sennrich *et al.* [36] propose to back translate additional monolingual target sentences as additional synthetic pairs. The back-translation technique alleviates overfitting, improves fluency and has thus become a common practice in many later MT systems [14, 35].

Inspired by the progress in NMT, we propose to incorporate multi-modal neural machine translation (MMT) [11, 19] which generates additional multilingual descriptions conditioned on both the image and the caption to improve the textual diversity required for training cross-modal retrieval models. Specially, we explore MMT as the generative approach to perform forward (English + Image → German) and backward (German + Image → English) translation to synthesize additional multilingual or monolingual text-image pairs for training cross-modal retrieval models.

There are three advantages of incorporating MMT into cross-modal retrieval models: First, the increased multilingual or monolingual diversity provides vital variation among image-text pairs to learn an improved monolingual or multilingual VSE space for cross-modal retrieval. Second, from the perspective of multi-task learning, the model learns two tasks in parallel (English-Image and German-Image matching) which help regularize the model for learning a more robust VSE space. Third, the proposed framework inherently enables learning multilingual VSE (forward pass) for multilingual cross-modal retrieval. Our framework sheds light on an important research question: For multilingual cross-modal search, (*e.g.* Search image with German and English queries), is it sufficient to simply translate queries into English than search? Or it is preferred to learn a multilingual VSE for retrieval? In this paper, we investigate and show that learning multilingual VSE enabled by the proposed framework without additional image-text annotation outperforms the translate-search approach.

As shown in [20] that the VSE space is prone to be object-oriented, in this paper, we leverage the shared regional features for the MMT and the cross-modal retrieval model. Furthermore, we design a new objective emphasizing on multilingual multimodal harder negatives to address the new multilingual multimodal scenario enabled by the proposed framework. In the experiments, we demonstrate the superiority of the proposed framework incorporating regional MMT for cross-modal retrieval in the monolingual image-text matching (English-Image) tasks on Flickr30K [46] and MS-COCO [30]. It is worth noting that, instead of tweaking even fancier architectures, we show that an (arguably) simplest attention network (SAN) with regional visual features yields competitive performance at a factor of computation cost. Furthermore, we showcase that the proposed framework can handle the multilingual image-text matching tasks (English-Image and German-Image) on Multi30K [11] where it yields new state-of-the-art performance.

To summarize, we make the following contributions:

- We promote a novel MMT-powered cross-modal retrieval framework which generates diverse multilingual and monolingual descriptions via forward and backward translations to learn robust VSE for cross-modal retrieval.
- Our framework consistently and substantially improves baseline models and achieves state-of-the-art performance in the monolingual cross-modal retrieval tasks on Flickr30K and MS-COCO. We further show that even the simplest attention network (SAN) yields competitive results with comparably low computation cost.
- The learned multilingual VSE can generalize to multilingual cross-modal retrieval tasks on Multi30K where our framework achieves state of the art.
- We demonstrate that learning multilingual VSEs for multilingual cross-modal retrieval outperforms the translate-search approach with monolingual VSE models.

2 RELATED WORK

Cross-modal retrieval models: A rich line of research has explored learning cross-view representations [2, 10, 12, 27, 42, 45] with various types of learning objectives and encoders. For the objective, Kiros *et al.* [27] set the tone of using a hinge-based triplet ranking loss to learn the VSE space where the distance between the matched image-caption pair is closer than the mismatched ones. Vendrov *et al.* [42] propose an objective that could preserve the partial order of the visual-semantic hierarchy. In [44, 45], Wang *et al.* extend the ranking loss with structure-preserving constraints by mining the inter- and intra-modal unpaired instances. In VSE++ [12], Faghri *et al.* empirically show that emphasizing hard negative examples within the sampled mini-batch results in robust embeddings.

Numerous encoder architectures and attention mechanisms have been explored for learning VSE. Convolutional neural networks (CNN) and recurrent neural networks (RNN) and their variants have been widely used to encode the visual and textual content. Recently, attention mechanisms have been addressed for cross-modal retrieval. Broadly, there are two attention types: The first type is intra-modal attention [17, 32, 39], where the content in each modality is encoded independently. In [39], Song *et al.* use multi-head self-attention to compute multiple representations for an instance. The second type is inter-modal attention[21, 29] where the representations depend on the interaction of multimodal contents. For example, text-to-image and image-to-text attention are performed in In SCAN [29]. In [20], Huang *et al.* combine two types of attention mechanisms. In this paper, we incorporate MMT for these two attention types and show that a model with simple intra-modal attention and regional visual features can yield competitive performance compared to the models with complicated architectures.

An emerging research trend studies cross-modal retrieval under a practical multilingual setup [13, 16] (*e.g.*German-Image and English-Image). In [13], Gella *et al.* show that learning with multilingual instances improves the robustness of VSE. In [16], Huang *et al.* propose a diversity objective over multi-head attention to further improve the robustness of multilingual multimodal representations.

The availability of abundant monolingual or multilingual multimodal parallel corpora is critical for learning VSE. Recent advancement in cross-modal retrieval explores augmentation methods to enrich the English-Image corpora. In [38], Shi *et al.* augment the text part with adversarial samples to improve the robustness of cross-modal retrieval. The low resource setup is studied in [17] where the textual semantics of regional objects are utilized as the additional sentences. Our methods with MMT can be considered as an augmentation approach but we additionally enable new multilingual cross-modal retrieval with monolingual multimodal corpora.

Multimodal Machine Translation: Machine translation with deep learning has achieved great success with the encoder-decoder architecture. The sequential nature of sentences can be effectively handled by RNN [4, 40]. Fully convolutional NMT was proposed in [23, 24], followed by more advanced architectures such as Transformers [41] with self-attention and positional encoding. On the other hand, MMT was introduced in [11] and considers additional images as the complementary information source for MT. MMT aims at leveraging either global [6], grid-like [5] or regional [19] visual information to generate better translations.

The back-translation technique [36], which generates synthetic parallel corpus with additional target-side monolingual corpora, has been shown effective for mitigating the prerequisite of high-quality parallel corpora. Although being asserted in [33] that there are still many unknown factors, the improved language modeling with pseudo multilingual pairs has become the common practice in many later NMT systems [14, 35].

Inspired by back-translation in NMT, we extend the concept to a multimodal version. For cross-modal retrieval, we propose to incorporate MMT as the multimodal generator to synthesize additional text sentences in the target language conditioned on both the images and the source sentences. We then use the synthesized

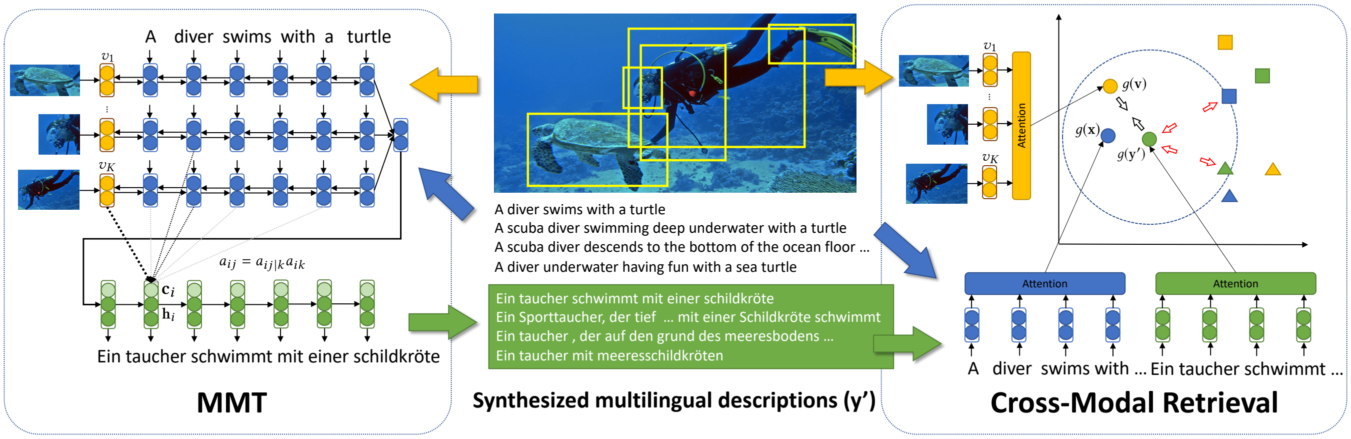

A diver swims with a turtle
v_1
v_K
$a_{ij} = a_{ij|k}a_{ik}$
$\mathbf{c}_i$
$\mathbf{h}_i$
Ein taucher schwimmt mit einer schildkröte
MMT
A diver swims with a turtle
A scuba diver swimming deep underwater with a turtle
A scuba diver descends to the bottom of the ocean floor ...
A diver underwater having fun with a sea turtle
Ein taucher schwimmt mit einer schildkröte
Ein Sporttaucher, der tief ... mit einer Schildkröte schwimmt
Ein taucher , der auf den grund des meeresbodens ...
Ein taucher mit meeresschildkröten
Synthesized multilingual descriptions (y')
v_1
v_K
Attention
$g(\mathbf{v})$
$g(\mathbf{x})$
$g(\mathbf{y}')$
Attention
Attention
A diver swims with ... Ein taucher schwimmt ...
Cross-Modal Retrieval

Figure 1: The proposed MMT + cross-modal retrieval framework for learning the multilingual VSE space. The additional multilingual descriptions (in green) are synthesized by MMT with the paired monolingual data (in blue) and the image. Regional visual features of salient visual objects (in yellow) are shared over MMT and cross-modal retrieval models.

sentence and the original image to form additional monolingual or multilingual image-text pairs for learning the VSE space.

3 PROPOSED FRAMEWORK

Figure 1 illustrates the proposed framework which incorporates MMT for improved monolingual and multilingual cross-modal retrieval. Our framework is composed of an MMT module and a multilingual cross-modal retrieval model that share the regional visual features. In the following we detail the main components.

3.1 Regionally-Attended Multimodal Neural Machine Translation

Multimodal neural machine translation (MMT) [11] is the multimodal extension of NMT [7] which incorporates visual information into machine translation (MT). In this work, we use MMT as a multimodal generator $\mathbf{y} = \text{MMT}(\mathbf{x}, \mathbf{v})$ where both textual ($\mathbf{x}$) and visual content ($\mathbf{v}$) are used to generate new text content ($\mathbf{y}$) to form additional diverse image-text pairs (German-Image or English-Image) to improve cross-modal retrieval and enable multilingual search.

Salient visual objects are shown to be effective for various multimodal tasks [1, 19]. Inspired by [19], we design an MMT backbone to encode attend-able salient visual objects with the text sequences in parallel threads. Let us denote by $\mathbf{x} = [x_1, \cdots, x_N]$ the source sentence (*e.g.*English) and $\mathbf{y} = [y_1, \cdots, y_M]$ the target sentence (*e.g.*German), and $\mathbf{v} = [v_1, \cdots, v_K]$ the salient regional visual features of the corresponding image. We map and parallelize k visual objects along with the source text sequence to form K multimodal sequences $\mathbf{u}_k = [v_k, x_1, ...x_N]$. The source multimodal sequence is then feed to the encoder which is a bi-directional long short-term memory (Bi-LSTM) network generating visually-encoded textual hidden states $\mathbf{s}_{j,k,j=1\cdots N+1}$.

During decoding after mean pooling over $\mathbf{s}_{-1,k}$, the decoder first estimates the attentional state $\tilde{\mathbf{h}}_i$ then predicts i-th word token in the target language. The attentional state $\tilde{\mathbf{h}}_i$ is calculated with the attention mechanism:

$$\tilde{\mathbf{h}}_i = \tanh(\mathbf{W}_c[\mathbf{c}_i;\mathbf{h}_i]) \quad (1)$$

Unlike the flattened global attention in [19], we consider localizing and attending to the (local) visually-encoded sequence k important. This can be regarded as the multimodal analogy of the local attention proposed in MT [31]. Specifically, we factorize the local attention over multimodal threads:

$$a_{ij} = a_{ij|k}a_{ik} \quad (2)$$

where

$$\begin{aligned} a_{ij|k} &= \text{softmax}(\mathbf{h}_i, \mathbf{s}_{j,k}) \\ a_{ik} &= \text{softmax}(\mathbf{h}_i, \mathbf{s}_{:,k}) \end{aligned} \quad (3)$$

and $\mathbf{s}_{:,k}$ is the mean-pooled states of k-th thread. Intuitively, $a_{ij|k}$ presents the attention to individual encoder states and a_{ik} is the regional attention to k-th multimodal sequence with k-th visual object. The final attended context can therefore be formulated as:

$$\mathbf{c}_i = \sum_j \sum_k \frac{a_{tj}}{|a_{tj}|}\mathbf{s}_{jk} \quad (4)$$

For decoding, the decoder then predicts target tokens y_i by evaluating the following probability:

$$p(y_i|\mathbf{y}_{<i}, \mathbf{x}, \mathbf{v}) = \text{softmax}(\mathbf{W}_o\tilde{\mathbf{h}}_i) \quad (5)$$

With a parallel corpus $\mathbf{x} \in X, \mathbf{y} \in Y, \mathbf{v} \in V, (\mathbf{x}, \mathbf{y}, \mathbf{v}) \in \mathcal{D}^{MMT}$, the maximum likelihood estimation (MLE) can be adopted to optimize the (source to target language) MMT model by minimizing the following objective:

$$\mathcal{L}^{MMT} = \mathbb{E}_{(\mathbf{x},\mathbf{y},\mathbf{v})\in\mathcal{D}^{MMT}}\left[-\log\prod_{i=1}^{M} p(y_i|\mathbf{y}_{<i}, \mathbf{x}, \mathbf{v})\right] \quad (6)$$

In practice, we first pre-train MMT on a larger text-only parallel corpus by removing the leading visual object then fine-tune on a smaller multimodal parallel corpus.

3.2 Attention-based Cross-Modal Retrieval and Simplest Attention Network

In conventional cross-modal retrieval, a monolingual multimodal corpus $(\mathbf{x}, \mathbf{v}) \in \mathcal{D}$ is available for learning the VSE space. In the proposed framework, we leverage MMT ($\mathbf{y}' = \text{MMT}(\mathbf{x}, \mathbf{v})$) to synthesize additional sentences in the target language, resulting in an MMT-enhanced corpus with $(\mathbf{x}, \mathbf{v}), (\mathbf{y}', \mathbf{v}) \in \mathcal{D}'$. In practice, forward MMT (English + Image → German) generates additional German-Image pairs. Another backward MMT (German + Image → English) then generates additional English-Image pairs with the forwarded outputs. We use an attention network and an improved multilingual triplet ranking loss to learn the multilingual VSE space for cross-modal retrieval.

Many inter-modal and intra-modal attention mechanisms have been recently proposed for cross-modal retrieval. In this paper, we show that most of them would be further improved with the proposed framework. Instead of building even more complicated attention mechanisms, we investigate the essential and possibly the simplest attention form with *intra-modal* attention. We term it simplest attention network (SAN). We show that, with MMT, SAN can still achieve comparable performance to current state-of-the-art models with the proposed framework.

Let $T \in X \cup Y'$ denotes the union of the original and the MMT translated sentences. Let us denoted by $\mathbf{z}^{(\mathbf{t})}$ the Bi-LSTM encoded text tokens and $\mathbf{z}^{(\mathbf{v})}$ the linear transformed regional features. We train a textual context vector $\mathbf{c}_T$ to attend and pool salient text tokens and a visual context vector $\mathbf{c}_V$ for the visual part. The attention weight is exponentially proportional to the closeness measured by a simple dot-product of the encoded context followed by a one-layer perceptron denoted by $f(.)$. Formally,

$$\alpha_i = \frac{\exp(f_c(\mathbf{c})^\top f_z(\mathbf{z}_i))}{\sum_j \exp(f_c(\mathbf{c})^\top f_z(\mathbf{z}_i))}. \tag{7}$$

The fixed-length representation is computed by $\mathbf{z}' = \sum_i \alpha_i \mathbf{z}_i$.

SAN and other attention-based networks in the proposed framework consider the regional visual features shared with MMT as well as the synthesized sentences from MMT and the original sentences. The regional visual features are extracted through visual object proposals by Faster R-CNN [34]. In the case of odd-passed MMT (*e.g.* one forward pass, Image + English→German), the multilingual text encoder and the context vector takes both English and German inputs. Naturally, the proposed framework learns multilingual multimodal representation and is capable of handling multilingual cross-modal retrieval tasks.

3.3 Harder Negative Multilingual Multimodal Example Mining for Learning the VSE space

Triplet ranking loss is a widely used objective for learning the VSE space in cross-modal retrieval. We extend the original triplet loss in consideration of synthesized sentences and generalize it for the multilingual multimodal (English-German-Image) scenario enabled by MMT with harder negative mining.

Let $s(.,.)$ denotes a similarity measure (*e.g.* cosine similarity) in the VSE space. For clarity in notation, let us denote by $g(\mathbf{z})$ the fixed-length representation of an image or a sentence after attention in the VSE space. And let $\mathbf{t} \in T$ represents the sentence in the source or synthesized languages. In order to preserve the underlying inter-modal structure in the joint embedding space, Kiros *et al.* [27] proposed a hinge-based $[.]_+ = \max(., 0)$ triplet loss:

$$\begin{aligned}\mathcal{L} = \sum_{(\mathbf{v},\mathbf{t})} \Bigg\{ & \sum_{\hat{\mathbf{t}}} \left[\alpha - s\left(g(\mathbf{v}), g(\mathbf{t})\right) + s(g(\mathbf{v}), g(\hat{\mathbf{t}}))\right]_+ \\ & + \sum_{\hat{\mathbf{v}}} \left[\alpha - s(g(\mathbf{v}), g(\mathbf{t})) + s(g(\hat{\mathbf{v}}), g(\mathbf{t}))\right]_+ \Bigg\},\end{aligned} \tag{8}$$

where $\hat{\mathbf{v}}$ and $\hat{\mathbf{t}}$ are the unpaired negatives and α is the margin. For robustness, it was empirically shown in [12] that taking only the harder negatives achieves better performance. Additionally, as the synthesized descriptions via forward/backward translation are comparably noisy to the original manually annotated description, we empirically find it beneficial to distinguish them. Therefore, we propose the following multilingual triplet objective:

$$\begin{aligned}\mathcal{L}^{VSE} = \sum_{\mathbf{v},\mathbf{t}} \Bigg\{ & \left[\alpha - s(g(\mathbf{v}), g(\mathbf{t})) + \sum_l \beta_l s(g(\mathbf{v}), g(\hat{\mathbf{t}}^{(l)}_{\text{hard}}))\right]_+ \\ & + \left[\alpha - s(g(\mathbf{v}), g(\mathbf{t})) + s(g(\hat{\mathbf{v}}_{\text{hard}}), g(\mathbf{t}))\right]_+ \Bigg\}\end{aligned} \tag{9}$$

where l indexes each language type and $\beta_l, \sum_l \beta_l = 1$ is the hyper parameter to weight the harder sentence $\hat{\mathbf{t}}_h^{(l)_{\text{hard}}}$ for a image $\mathbf{v}$ in language type l. The intuition behind Eq. 9 is to encourage the correct image-text pairs to locate closer up to a margin α to the closest unpaired one. Note that different from the objective proposed in [13], we jointly consider the hard unpaired sentences in *ALL* language types[1]. Let $I(\mathbf{d}) = l, l \in \{\text{English}, \text{German}_{\text{MMT}}, \text{English}_{\text{MMT}}\}$ denotes the language type indicator function for an instance $\mathbf{d}$. We define the harder negatives as:

$$\begin{aligned}\hat{\mathbf{t}}^{(l)}_{\text{hard}} &= \underset{\mathbf{d} \neq \mathbf{t}, I(\mathbf{d})=l}{\text{argmax}}\ s(g(\mathbf{v}), g(\mathbf{d})) \\ \hat{\mathbf{v}}_{\text{hard}} &= \underset{\mathbf{q} \neq \mathbf{v}}{\text{argmax}} s(g(\mathbf{q}), g(\mathbf{t}))\end{aligned} \tag{10}$$

In the training phase, we sample monolingual or multilingual mini-batches (both with an equal portion of original sentences and MMT-synthesized sentences) and minimize Eq. 9. In the testing phase, we directly use the encoder and the attention module to generate the image and sentence representations. We estimate and rank the similarities between the query and the instances in the dataset for cross-modal retrieval.

4 EXPERIMENTS

We conduct extensive experiments to confirm the superiority of the proposed framework. We focus on two tasks: 1) Monolingual text-image matching on Flickr30K [46] and MS-COCO [30]. 2) Multilingual text-image matching on Multi30K [11].

4.1 Dataset and Evaluation Metric

Flickr30K: There are 31,783 images in the Flickr30K dataset. Five English descriptions are available for each image. For experiments on Flickr30K, we use the standard split defined in [25] with 29,000 training, 1,000 validation, and 1,000 testing images.

[1] In this paper we consider English and German only. However, Eq. 9 can be generalized to more languages.

Table 1: MMT improves various cross-modal retrieval models in the 1K testing set of Flickr30K. f and b stand for forward and backward translation with MMT which generates additional German and English descriptions respectively. R@k is recall at k. We re-implemented DAN* with a Faster R-CNN as the visual encoder. Both intra-modal attention models (SAN, DAN) and inter-modal attention model (SCAN) are improved with the proposed framework.

	Text-to-Image Retrieval			Image-to-Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
SAN	45.8	73.7	81.8	60.0	87.0	93.1
MT (f) + SAN	46.5	74.2	82.5	60.6	87.5	93.2
MT (f+b) + SAN	46.1	73.9	82.1	60.3	87.3	93.0
MMT (f) + SAN	**47.1**	**74.6**	83.0	**62.1**	**87.9**	**93.8**
MMT (f+b) + SAN	46.7	74.2	**83.2**	61.2	86.9	92.9
DAN [32] (ResNet)	39.4	69.2	79.1	55.0	81.8	89.0
DAN* (FRCNN-ResNet)	46.4	74.2	83.1	61.5	86.6	93.5
MMT (f) + DAN*	47.7	**75.3**	**83.7**	**64.3**	**88.5**	**94.4**
MMT (f+b) + DAN*	**47.9**	74.8	83.3	63.8	87.8	93.9
SCAN [29]	45.8	73.7	83.0	61.8	87.5	93.7
MMT (f) + SCAN	**47.3**	**75.3**	83.4	**64.5**	88.3	**94.3**
MMT (f+b) + SCAN	47.1	75.1	**83.5**	64.0	**88.5**	94.1

MS-COCO: MS-COCO contains 123,287 images where each image is annotated with five English descriptions. For the splits we follow the split defined in [25] to move originally left 30,504 validation images to the training set, resulting in a training set of 113,287 training images. We report either the average of metrics on five folds of 1,000 testing set or the metrics on the whole 5,000 testing set.

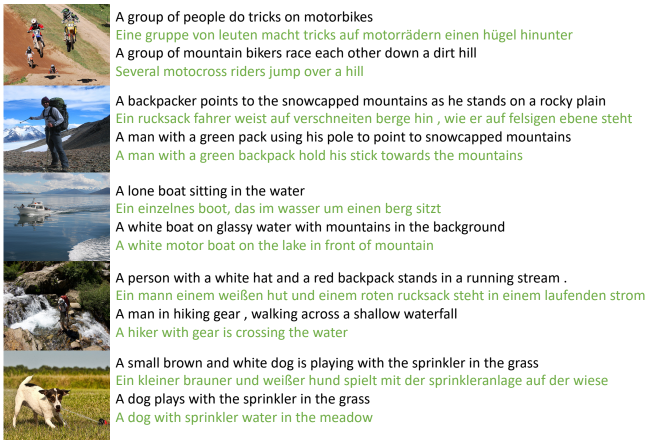
Multi30K: Multi30K is the multilingual extension of Flickr30K with additional multilingual (German) descriptions. The train, validation and testing split contains 29K, 1K, and 1K images, respectively. There are two types of annotations available: 1) One parallel German translation for each image. 2) Five independently collected German descriptions for each image. We use type 1 annotation for MMT pre-training and type 2 English annotations for the multilingual cross-modal retrieval experiment.

Evaluation Metric: As in most prior work on cross-modal retrieval tasks [12, 29, 32, 47], we measure rank-based performance by recall at K (R@k). Given a query, recall at k (R@k) calculates the percentage of test instances for which the correct one can be found in the top-K retrieved instances. In MS-COCO and Flickr30K, for image-to-text retrieval, there are 5 sentences considered correct. For text-to-image retrieval, there is only 1 correct image for each text query. We report R@1, R@5, and R@10.

4.2 Experiment Setup

For monolingual cross-modal (English-Image) retrieval tasks, we first pre-train text-only NMT with WMT German-English corpus in [31] then expand the network with the MMT visual part for model fine-tuning on the Multi30K training split with 29K type 1 multilingual pairs. We use Moses [28] pre-processing scripts and BPE [37] for text pre-processing.

Please note that models under our framework are trained with *NO* additional images while we double parallel text descriptions: MMT(f) models contain original English and MMT-translated German descriptions whereas MMT(f + b) models contain English and MMT-translated English descriptions. The encoder in the cross-modal retrieval will then take the randomly shuffled English descriptions (original or MMT synthesized) or the German descriptions (via forward MMT) for training. For validation and testing, only the original English descriptions are used.



A group of people do tricks on motorbikes
Eine gruppe von leuten macht tricks auf motorrädern einen hügel hinunter
A group of mountain bikers race each other down a dirt hill
Several motocross riders jump over a hill

A backpacker points to the snowcapped mountains as he stands on a rocky plain
Ein rucksack fahrer weist auf verschneiten berge hin , wie er auf felsigen ebene steht
A man with a green pack using his pole to point to snowcapped mountains
A man with a green backpack hold his stick towards the mountains

A lone boat sitting in the water
Ein einzelnes boot, das im wasser um einen berg sitzt
A white boat on glassy water with mountains in the background
A white motor boat on the lake in front of mountain

A person with a white hat and a red backpack stands in a running stream .
Ein mann einem weißen hut und einem roten rucksack steht in einem laufenden strom
A man in hiking gear , walking across a shallow waterfall
A hiker with gear is crossing the water

A small brown and white dog is playing with the sprinkler in the grass
Ein kleiner brauner und weißer hund spielt mit der sprinkleranlage auf der wiese
A dog plays with the sprinkler in the grass
A dog with sprinkler water in the meadow

Figure 2: Additional sentences generated by MMT. The sentences in green are generated by either forward MMT (Image + English → German) or forward + backward MMT (Image + German → English). Our framework incorporates the generated image-text pairs for learning the VSE space.

To identify and vectorize salient visual regions in images, we use the code in [1] to detect 36 salient objects in each image and extract their corresponding regional features. Each image is thus represented by a 36×2048 array. For text pre-processing, we truncate

Table 2: Comparison with other state-of-the-art models and their visual encoders on the 1K testing set of Flickr30K.

	Text-to-Image Retrieval			Image-to-Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
DVSA [25] (AlexNet)	15.2	37.7	50.5	22.2	48.2	61.4
SPE [45] (VGG)	29.7	60.1	72.1	40.3	68.9	79.9
VSE++ [12] (ResNet)	39.6	-	79.5	52.9	-	87.2
DPC [47] (ResNet)	39.1	69.2	80.9	55.6	81.9	89.5
SCO [22] (ResNet)	41.1	70.5	80.1	55.5	82.0	89.3
DAN [32] (ResNet)	39.4	69.2	79.1	55.0	81.8	89.0
SCAN [29] (FRCNN-ResNet)	45.8	74.4	83.0	61.8	87.5	93.7
Ours (SAN) (FRCNN-ResNet)	45.8	73.7	81.8	60.0	87.0	93.1
Ours (MMT(f) + SAN) (FRCNN-ResNet)	47.1	74.6	83.0	62.1	87.9	93.8
Ours (MMT(f) + DAN^{*2}) (FRCNN-ResNet)	**47.7**	**75.3**	**83.7**	64.3	**88.5**	**94.4**
Ours (MMT(f) + SCAN) (FRCNN-ResNet)	47.3	**75.3**	83.4	**64.5**	88.3	94.3
SCAN (Fusion)	48.6	77.7	85.2	67.4	90.3	95.8
Ours (MMT(f) + SCAN) (Fusion)	**50.3**	**78.7**	**86.0**	**69.0**	**91.1**	**96.3**

maximum sentence length to 100 and make tokens appear less than 4 times unknown.

We train SAN, re-implement dual attention network (DAN) [32], and train stacked-cross attention network (SCAN) [29] with the additional data from MMT to validate the effectiveness of the proposed framework. DAN and SCAN are chosen because their network architectures are representative for intra- and inter-modal attention. SCAN is already with regional features. For DAN, after we verify that the performance is comparable for the original DAN and our implementation, we swap the original ResNet [15] features with the regional features. For training, we set all hidden dimensions to 512. The dimension of randomly initialized word embeddings is 300. We use Adam [26] optimizer with $5e^{-4}$ learning rate then $5e^{-5}$ after 15th epoch. We train with 128 batch size for 24 epochs on all datasets. Models with the best validation performance are selected for testing.

For multilingual cross-modal retrieval tasks (English-German-Image), we follow the same MMT pre-training procedure. We discard the manually annotated German description in the type 2 annotation and synthesize the German descriptions with MMT for type 2 English descriptions. We use both English and MMT-translated German descriptions for training to learn multilingual multimodal representation. Please note that other compared baselines in the multilingual cross-modal retrieval are trained with original type 2 manually annotated German descriptions while we use the synthetic descriptions from MMT. In the testing phase, we use the original type 2 German and English annotations for a fair comparison with the baseline models.

4.3 Ablation Studies

First, we validate the proposed regional MMT + attention-based cross-modal retrieval framework on Flickr30K. Quantitatively, the proposed MMT model achieves 37.5 in BLEU-4 and 55.1 in Meteor scores in the Multi30K testing set after pre-training. The lexical diversity is improved from 8, 830 to 9, 128 (English tokens) with forward then backward MMT and 19, 960 (English + German tokens) with forward MMT. As qualitatively shown in Figure 2, the additional diverse German/English descriptions synthesized by MMT and the corresponding images are then included for training cross-modal retrieval models.

We then conduct the ablation studies to quantify and compare the contribution with different model configurations. As shown in Table 1, models with additional text diversity, either from MT or MMT, achieve better performance in all metrics compared with their vanilla versions. These consistent and substantial improvements over various baseline models validate the effectiveness of the proposed MMT-powered framework. MMT outperforms MT as the additional regional visual information is considered visually pertinent when generating additional text descriptions and form better image-text pairs for training cross-modal retrieval models. Surprisingly, even for monolingual tasks, the additional diversity from multilingual content benefits significantly. Although both additional monolingual and multilingual descriptions improve model performance, in most cases, we observe that it is preferred to have multilingual diversity (additional German descriptions by MMT(f)) than monolingual diversity (additional English descriptions by MMT($f + b$)).

Regularization by multi-task learning in the English-Image and German-Image tasks may be a feasible explanation for the superior MMT(f) performance. The models are forced to fulfill two tasks at the same time, and thus result in the well-generalized textual and visual encoders and the corresponding inter-/intra-modal attention. For mini-batch construction among original English descriptions and synthesized German descriptions, we also investigate various sampling strategies but does not observe significant difference among them. A simple random sampling strategy delivers stable and typically better results.

Comparing the difference in performance between MMT(f) and MMT($f + b$), we manually investigate the MMT results and observe that, in some cases, after forward then backward translations, some smaller objects in the image are missed in the description after

Table 3: Comparison with other state-of-the-art models on MS-COCO.

	Text-to-Image Retrieval			Image-to-Text Retrieval		
	R@1	R@5	R@10	R@1	R@5	R@10
			1K Testing Images			
DVSA [25] (AlexNet)	27.4	60.2	74.8	38.4	69.9	80.5
Order-embeddings [43] (VGG)	37.9	-	85.9	46.7	-	88.9
VSE++ [12] (ResNet)	52.0	-	92.0	64.6	-	95.7
DPC [47] (ResNet)	47.1	79.9	90.0	65.6	89.8	95.5
SCO [22] (ResNet)	56.7	87.5	**94.8**	69.9	92.9	97.5
DAN [32] (Our re-impl, FRCNN-ResNet)	56.4	86.7	93.4	70.6	94.0	97.2
SCAN [29] (FRCNN-ResNet)	56.4	87.0	93.9	70.9	94.5	97.8
Ours (SAN, FRCNN-ResNet)	56.0	86.5	93.6	69.9.	93.1	97.0
Ours (MMT(f) + SAN, FRCNN-ResNet)	56.7	87.2	93.6	70.6	93.8	97.5
Ours (MMT(f) + DAN, FRCNN-ResNet)	**57.2**	**87.8**	93.7	**71.7**	**95.0**	98.0
Ours (MMT(f) + SCAN, FRCNN-ResNet)	57.1	87.5	94.4	71.6	94.9	**98.2**
			5K Testing Images			
DVSA [25] (AlexNet)	10.7	29.6	42.2	16.5	39.2	52.0
Order-embeddings [43] (VGG)	31.7	-	74.6	23.3	-	84.7
VSE++ [12] (ResNet)	30.3	-	72.4	41.3	-	81.2
DPC [47] (ResNet)	25.3	53.4	66.4	41.2	70.5	81.1
SCO [22] (ResNet)	33.1	62.9	75.5	42.8	72.3	83.0
DAN [32] (Our re-impl, FRCNN-ResNet)	34.4	65.0	75.5	45.8	76.8	86.8
SCAN (FRCNN-ResNet)	34.4	63.7	75.7	46.4	77.4	87.2
Ours (SAN) (FRCNN-ResNet)	34.0	63.7	74.5	44.8	75.6	85.6
Ours (MMT(f) + SAN) (FRCNN-ResNet)	34.7	64.4	75.0	46.5	77.1	86.5
Ours (MMT(f) + DAN) (FRCNN-ResNet)	**35.3**	**65.6**	76.0	**47.6**	77.8	87.5
Ours (MMT(f) + SCAN) (FRCNN-ResNet)	35.0	64.4	**76.2**	47.1	**78.5**	**87.7**
SCAN (Fusion)	38.6	**69.3**	80.4	50.4	82.2	90.0
Ours (MMT(f) + SCAN) (Fusion)	**39.4**	69.1	**80.7**	**51.2**	**82.7**	**90.3**

two-step translations. We also try three-hop (forward-backward-forward) and four-hop translation but do not observe meaningful improvement. An explanation for the saturated performance after the multi-hop generation process by MMT (also true for MT) is that if the forward and the backward translation share the same parallel training corpus without additional data, the lexical distribution may still be similar and thus limits the textual diversity. One possible enhancement is to use additional images and back translate to text via other generative models such as image captioning.

4.4 Monolingual Cross-Modal Retrieval Results

We compare the MMT-enhanced models with other baselines on two commonly used cross-modal retrieval datasets: the Flickr30K dataset and the MS-COCO dataset. Table 2 summarizes the results on Flickr30K. Models with more advanced visual features generally perform better. We observed that the regional Faster RCNN features of visual objects are important to improve inter-modal dependency for matching text and image. With the improved text coverage and diversity through the proposed MMT framework, we achieving new state-of-the-art performance in image-text and text-image retrieval with DAN and SCAN backbone. It is noteworthy that, with MMT and regional visual features, even SAN yields competitive performance with roughly 140 times speedup in comparison to SCAN and 2 times to DAN.

Table 3 summarizes the results on MS-COCO. The proposed MMT framework along with vanilla baseline models achieves new state-of-the-art performance in most metrics. Importantly, the results show that MMT trained on out-of-domain images (Multi30K) still improves cross-modal retrieval models on MS-COCO, thus validate the improved textual diversity for learning the VSE space. SAN also achieves comparable performance to current state-of-the-art models on MS-COCO. From the monolingual cross-modal experiments on Flickr30K and MS-COCO, the improvement with the proposed MMT-based framework is substantial and can be generalized to multiple datasets.

4.5 Multilingual Cross-Modal Retrieval Results

Table 4 presents the German-Image and English-Image retrieval results on the Multi30K testing set. We compare our model with the results reported by [13]. VSE [27] and OE [43] models are trained for

Table 4: Multilingual cross-modal retrieval (German-Image and English-Image) on the 1K testing set of Multi30K. Visual encoders:VGG†. Monolingual models*.

Method	German-to-Image			Image-to-German			English-to-Image			Image-to-English		
	R@1	R@5	R10	R@1	R@5	R10	R@1	R@5	@10	R@1	R@5	R10
VSE$^{\dagger *}$ [27]	20.3	47.2	60.1	29.3	58.1	71.8	23.3	53.6	65.8	31.6	60.4	72.7
OE$^{\dagger *}$ [43]	21.0	48.5	60.4	26.8	57.5	70.9	25.8	56.5	67.8	34.8	63.7	74.8
PIVOT† [13]	22.5	49.3	61.7	28.2	61.9	73.4	26.2	56.4	68.4	33.8	62.8	75.2
Ours (MMT(f) + SAN†)	25.2	54.5	63.8	36.5	64.5	76.2	30.9	63.0	73.5	36.6	67.5	80.5
Ours (MT(f) + SAN)	35.8	63.6	71.5	52.5	80.5	85.1	46.5	74.1	81.9	61.6	87.6	93.1
Ours (MMT(f) + SAN)	36.5	64.0	72.9	54.8	81.2	86.8	46.9	74.5	83.0	63.4	88.0	93.7
Ours (MMT(f) + DAN)	36.8	**64.3**	**73.5**	**55.5**	**81.8**	**88.2**	**47.6**	75.2	**83.7**	64.2	**88.5**	**94.4**
Ours (MMT(f) + SCAN)	**37.0**	64.0	73.1	53.0	81.5	87.4	47.3	**75.4**	83.5	**64.6**	88.2	94.2

Table 5: Multilingual models vs. Monolingual models with translate-search in the Multi30K testing set.

Method	German-to-Image			Image-to-German		
	R@1	R@5	R10	R@1	R@5	R10
SAN(English)	35.5	63.2	71.5	53.5	80.5	86.1
MMT(f) + SAN	36.5	64.0	72.9	54.8	81.2	86.8
DAN (English)	35.7	63.5	72.3	54.5	80.9	87.4
MMT(f) + DAN	36.8	64.3	73.5	55.5	81.8	88.2

two languages separately. In contrast, PIVOT [13] and our model are capable of handling multilingual queries (German and English) for cross-modal retrieval with the same model. For a fair comparison with PIVOT, we also downgrade our visual backbone from ResNet of Faster-RCNN to VGG.

The proposed framework successfully yields state-of-the-art performance in the multilingual cross-modal retrieval tasks, outperforming previous baselines by a large margin. MMT(f)+DAN achieves the best result which is slightly better than MMT(f)+SCAN. With MMT, even SAN with VGG will outperform PIVOT. Comparing MMT and MT, MMT again outperforms MT as it generates more visually pertinent descriptions. The improvement over baselines also validates the proposed multilingual harder negative objective function can effectively learn the underlying multilingual structure in the multilingual VSE space.

While our original intuition is to utilize MMT to improve multilingual text diversity for monolingual retrieval tasks, we surprisingly find out that the synthetic multilingual data can also help to learn better multilingual multimodal representations. The results suggest that a back-translation-like strategy for other multilingual multimodal contexts may be feasible to learn more robust visual-semantic embeddings. We leave this direction as our future work.

4.6 Translate-Search vs. Multilingual VSE

For cross-modal search with queries in languages other than English, an interesting question would be: Is it sufficient to just translate the query into English than search with the English-Image model or it is preferred to have a model handling multilingual queries? We conduct experiments to answer this question. For monolingual models, in the testing phase, we use the pre-trained MT model to translate the German queries/descriptions into English for retrieval with the model trained with English-only annotation. In contrast, for multilingual models enabled by the proposed MMT-powered framework, we directly encode the German queries in the multilingual VSE space for retrieval.

Table 5 shows the testing results on the Multi30K testing set. As can be observed, the monolingual models with translate-search already provide a surprisingly strong baseline. Nevertheless, the results imply that multilingual models enabled by the proposed framework outperform the translate-search approach. We consider this the effect that the multilingual models benefit from the additional synthetic training pairs in another language and thus improve the generalizability in the VSE space. Note that the comparison is legit as for both types of models the same original English-image annotations are provided.

5 CONCLUSION AND FUTURE WORK

We present a novel approach for improving cross-modal retrieval models with monolingual and multilingual textual diversities via forward and backward multimodal neural machine translation (MMT). Our approach enables learning with two (English-Image and MMT-Image) cross-modal retrieval tasks in parallel and produces more robust visual-semantic embeddings with a new multilingual multimodal triplet objective considers synthetic image-text pairs. The proposed framework improves various state-of-the-art models and the simplest attention network (SAN) for monolingual and multilingual image-text matching tasks on Flickr30K, MS-COCO, and Multi30K. The models with improved multilingual diversity from forward MMT achieves better performance than monolingual ones from forward+backward MMT in monolingual cross-modal retrieval. The proposed framework can easily be generalized to the multilingual multimodal scenario, where our models outperform multilingual and translate-search monolingual baselines.

ACKNOWLEDGEMENTS

This work is supported by the DARPA grant FA8750-18-2-0018 funded under the AIDA program and the GAILA project (award HR00111990063). The authors would like to thank Google Cloud for providing research credits. The authors would also like to thank the anonymous reviewers for their suggestions.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*.

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *International Conference on Machine Learning*. 1247–1255.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR* abs/1409.0473 (2014).

[5] Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes Garcia-Martinez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost Van de Weijer. 2017. LIUM-CVC Submissions for WMT17 Multimodal Translation Task. In *SECOND CONFERENCE ON MACHINE TRANSLATION*, Vol. 2. 432–439.

[6] 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*. 992–1003.

[7] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the Properties of Neural Machine Translation: Encoder–Decoder Approaches. *Syntax, Semantics and Structure in Statistical Translation* (2014), 103.

[8] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards Diverse and Natural Image Descriptions via a Conditional GAN. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2989–2998.

[9] Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*. 898–907.

[10] Aviv Eisenschtat and Lior Wolf. 2017. Linking Image and Text with 2-Way Nets. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 1855–1865.

[11] 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*. Association for Computational Linguistics, 70–74. https://doi.org/10.18653/v1/W16-3210

[12] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. (2018). https://github.com/fartashf/vsepp

[13] Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image Pivoting for Learning Multilingual Multimodal Representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2839–2845. https://doi.org/10.18653/v1/D17-1303

[14] Di He, Yingce Xia, Tao Qin, Liwei Wang, Nenghai Yu, Tieyan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*. 820–828.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[16] Po-Yao Huang, Xiaojun Chang, and Alexander G. Hauptmann. 2019. Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Association for Computational Linguistics, 1461–1467. https://doi.org/10.18653/v1/D19-1154

[17] Po-Yao Huang, Guoliang Kang, Wenhe Liu, Xiaojun Chang, and Alexander G. Hauptmann. 2019. Annotation Efficient Cross-Modal Retrieval with Adversarial Attentive Alignment. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery, New York, NY, USA, 1758–1767. https://doi.org/10.1145/3343031.3350894

[18] Po-Yao Huang, Junwei Liang, Jean-Baptiste Lamare, and Alexander G. Hauptmann. 2018. Multimodal Filtering of Social Media for Temporal Monitoring and Event Analysis. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval (ICMR '18)*. Association for Computing Machinery, New York, NY, USA, 450–457. https://doi.org/10.1145/3206025.3206079

[19] 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*, Vol. 2. 639–645.

[20] Po-Yao Huang, Vaibhav, Xiaojun Chang, and Alexander G. Hauptmann. 2019. Improving What Cross-Modal Retrieval Models Learn through Object-Oriented Inter- and Intra-Modal Attention Networks. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval (ICMR '19)*. Association for Computing Machinery, New York, NY, USA, 244–252. https://doi.org/10.1145/3323873.3325043

[21] Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-Aware Image and Sentence Matching with Selective Multimodal LSTM. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 7254–7262.

[22] Yan Huang, Qi Wu, and Liang Wang. 2017. Learning semantic concepts and order for image and sentence matching. *arXiv preprint arXiv:1712.02036* (2017).

[23] Lukasz Kaiser and Samy Bengio. 2016. Can Active Memory Replace Attention?. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. 3774–3782.

[24] Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *arXiv preprint arXiv:1610.10099* (2016).

[25] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3128–3137.

[26] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[27] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models. *NIPS Workshop* (2014).

[28] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, 177–180.

[29] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked Cross Attention for Image-Text Matching. *arXiv preprint arXiv:1803.08024* (2018).

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[31] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Lisbon, Portugal, 1412–1421. http://aclweb.org/anthology/D15-1166

[32] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 2156–2164.

[33] Alberto Poncelas, Dimitar Shterionov, Andy Way, Gideon Maillette de Buy Wenniger, and Peyman Passban. 2018. Investigating Backtranslation in Neural Machine Translation. *arXiv preprint arXiv:1804.06189* (2018).

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[35] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891* (2016).

[36] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 86–96.

[37] Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 1715–1725.

[38] Haoyue Shi, Jiayuan Mao, Tete Xiao, Yuning Jiang, and Jian Sun. 2018. Learning Visually-Grounded Semantics from Contrastive Adversarial Samples. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*. 3715–3727. https://aclanthology.info/papers/C18-1315/c18-1315

[39] Yale Song and Mohammad Soleymani. 2019. Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 1979–1988. https://doi.org/10.1109/CVPR.2019.00208

[40] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 3104–3112.

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[42] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-Embeddings of Images and Language. *CoRR* abs/1511.06361 (2015).

[43] Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *arXiv preprint arXiv:1511.06361* (2015).

[44] Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2018. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).

[45] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning Deep Structure-Preserving Image-Text Embeddings. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE, 5005–5013.

[46] 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 (2014), 67–78.

[47] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, and Yi-Dong Shen. 2017. Dual-Path Convolutional Image-Text Embedding. *CoRR* abs/1711.05535 (2017). arXiv:1711.05535 http://arxiv.org/abs/1711.05535