Multi-Head Attention with Diversity for Learning Grounded Multilingual Multimodal Representations

Po-Yao Huang[1], Xiaojun Chang[2], Alexander Hauptmann[1]
[1]Language Technologies Institute, Carnegie Mellon University
[2]Faculty of Information Technology, Monash University
poyaoh@cs.cmu.edu, cxj273@gmail.com, alex@cs.cmu.edu

Abstract

With the aim of promoting and understanding the multilingual version of image search, we leverage visual object detection and propose a model with diverse multi-head attention to learn grounded multilingual multimodal representations. Specifically, our model attends to different types of textual semantics in two languages and visual objects for fine-grained alignments between sentences and images. We introduce a new objective function which explicitly encourages attention diversity to learn an improved visual-semantic embedding space. We evaluate our model in the German-Image and English-Image matching tasks on the Multi30K dataset, and in the Semantic Textual Similarity task with the English descriptions of visual content. Results show that our model yields a significant performance gain over other methods in all of the three tasks.

1 Introduction

Joint visual-semantic embeddings (VSE) are central to the success of many vision-language tasks, including cross-modal search and retrieval (Kiros et al., 2014; Karpathy and Fei-Fei, 2015; Gu et al., 2018), visual question answering (Antol et al., 2015; Goyal et al., 2017), multimodal machine translation (Huang et al., 2016; Elliott and Kádár, 2017), etc. Learning VSE requires extensive understanding of the content in individual modalities and an in-depth alignment strategy to associate the complementary information from multiple views.

With the availability of large-scale parallel English-Image corpora (Lin et al., 2014; Young et al., 2014), a rich line of research has advanced learning VSE under the monolingual setup. Most recent works (Kiros et al., 2014; Vendrov et al., 2015; Karpathy and Fei-Fei, 2015; Klein et al., 2015; Wang et al., 2016, 2018; Huang et al.,

2019b) leverage triplet ranking losses to align English sentences and images in the joint embedding space. In VSE++ (Faghri et al., 2018), Faghri *et al.* improve VSE by emphasizing hard negative samples. Recent advancement in VSE models explores methods to enrich the English-Image corpora. Shi et al. (2018) propose to augment dataset with textual contrastive adversarial samples to combat adversarial attacks. Recently, Huang et al. (2019a) utilize textual semantics of regional objects and adversarial domain adaptation for learning VSE under low-resource constraints.

An emerging trend generalizes learning VSE in the multilingual scenario. Rajendran et al. (2016) learn M-view representations when parallel data is available only between one pivot view and the rest of views. PIVOT (Gella et al., 2017) extends the work from Calixto et al. (2017) to use images as the pivot view for learning multilingual multimodal representations. Kádár et al. (2018) further confirms the benefits of multilingual training.

Our work is motivated by Gella et al. (2017) but has important differences. First, to disentangle the alignments in the joint embedding space, we employ visual object detection and multi-head attention to selectively align salient visual objects with textual phrases, resulting in visually-grounded multilingual multimodal representations. Second, as multi-head attention (Vaswani et al., 2017) is appealing for its efficiency and ability to jointly attend to information form different perspectives, we propose to further encourage the diversity among attention heads to learn an improved visual-semantic embedding space. Figure 1 illustrates our gradient updates promoting diversity. The proposed model achieves state-of-the-art performance in the multilingual sentence-image matching tasks in Multi30K (Elliott et al., 2016) and the semantic textual similarity task (Agirre et al., 2012, 2014, 2015).
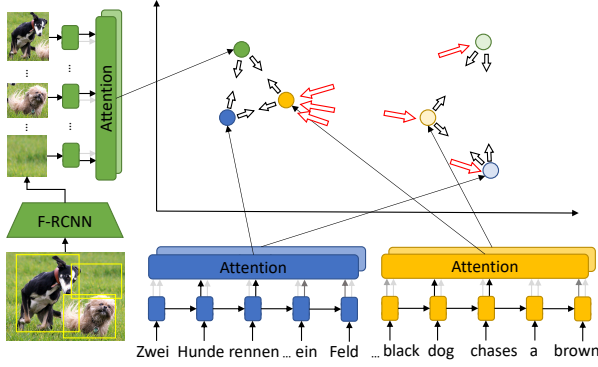
Figure 1: Multi-head attention with diversity for learning grounded multilingual multimodal representations. (A two-headed example with a part of diversity loss l_θ^D colored in red.)

2 Related Works

Classic attention mechanisms have been addressed for learning VSE. These mechanisms can be broadly categorized by the types of {*Query*, *Key*, *Value*} as discussed in Vaswani et al. (2017). For intra-modal attention, {*Query*, *Key*, *Value*} are within the same modality. In DAN (Nam et al., 2017), the content in each modalities is iteratively attended through multiple steps with intra-modal attention. In SCAN (Lee et al., 2018), inter-modal attention is performed between regional visual features from Anderson et al. (2018) and text semantics. The inference time complexity is $O(MN)$ (for generating M query representations for a size N datatset). In contrast to the prior works, we leverage intra-modal multi-head attention, which can be easily parallelized compared to DAN and is with a preferred $O(M)$ inference time complexity compared to SCAN.

Inspired by the idea of attention regularization in Li et al. (2018), for learning VSE, we propose a new margin-based diversity loss to encourage a margin between attended outputs over multiple attention heads. Multi-head attention diversity within the same modality and across modalities are jointly considered in our model.

3 The Proposed Model

Figure 1 illustrates the overview of the proposed model. Given a set of images as the pivoting points with the associate English and German[1] descriptions or captions, the proposed VSE model aims to

---

[1]For clarity in notation, we discuss only two languages. The proposed model can be intuitively generalized to more languages by summing additional terms in Eq. 4 and Eq. 6-7.

learn a multilingual multimodal embedding space in which the encoded representations (v, e, g) of a paired instance (i, x^e, x^g) are closely aligned to each other than non-paired ones.

Encoders: For a sampled (i, x^e, x^g) pair, we first encode the tokens in the English sentence $x^e = \{x_1^e, \ldots, x_N^e\}$ and the tokens in the German sentence $x^g = \{x_1^g, \ldots, x_N^g\}$ through the word embedding matrices followed by two bi-directional LSTMs. The outputs of the textual encoders are $e = \{e_1, \ldots, e_N\}, e_n \in \mathbb{R}^H$ for English and $g = \{g_1, \ldots, g_N\}, g_n \in \mathbb{R}^H$ for German, where N is the max sentence length and H is the dimension of the shared embedding space. For the image, we leverage a Faster-RCNN (Ren et al., 2015) network with a ResNet (He et al., 2016) backbone to detect and encode salient visual objects in the image. With a trainable one-layered perceptron to transform visual features into the shared embedding space, we encode the image as $v = \{v_1, \ldots, v_M\}, v_m \in \mathbb{R}^H$, where M is the maximum amount of visual objects in an image.

Multi-head attention with diversity: We employ K-head attention networks to attend to the visual objects in an image as well as the textual semantics in a sentence then generate fixed-length image/sentence representations for alignment. Specifically, the k-th attended German sentence representation g^k is computed by:

$$a_i^k = \text{softmax}(\tanh(W_{c_g}^k c_g^k)^\top \tanh(W_g^k g_i)) \quad (1)$$

$$g^k = \sum_{i=1}^N a_i^k g_i, \quad (2)$$

where a_i^k is the attention weight, $W_g^k \in \mathbb{R}^{H \times H_{attn}}, W_{c_g}^k \in \mathbb{R}^{H_c \times H_{attn}}$ is the learnable transformation matrix for German. $c_g^k \in \mathbb{R}^{H_c}$ is the learnable H_c-dimensional contextual vector for distilling important semantics from German sentences. The final German sentence representation is the concatenation of K-head attention outputs $g = [g^0 \| g^1 \| \ldots \| g^K]$. Similar for encoding the English sentence $e = [e^0 \| e^1 \| \ldots \| e^K]$ and the image $v = [v^0 \| v^1 \| \ldots \| v^K]$.

With $\{V, E, G\}$ where $v \in V, e \in E, g \in G$ as the set of attended fixed-length image and sentence representations in a sampled batch, we use the widely-used hinge-based triplet ranking loss with hard negative mining (Faghri et al., 2018) to align instances in the visual-semantic embedding space. Taking Image-English instances $\{V, E\}$ as

an example, we leverage the triplet correlation loss defined as:

$$l_\theta(V, E) = \sum_p \big[\alpha - s(v_p, e_p) + s(v_p, \hat{e}_p)\big]_+$$
$$+ \sum_q \big[\alpha - s(v_q, e_q) + s(\hat{v}_q, e_q)\big]_+,$$

$$(3)$$

where α is the correlation margin between positive and negative pairs, $[.]_+ = \max(0, .)$ is the hinge function, and $s(a, b) = \frac{a^T b}{\|a\|\|b\|}$ is the cosine similarity. p and q are the indexes of the images and sentences in the batch. $\hat{e}_p = \text{argmax}_q s(v_p, e_{q \neq p})$ and $\hat{v}_q = \text{argmax}_p s(v_{p \neq q}, e_q)$ are the hard negatives. When the triplet loss decreases, the paired images and German sentences are drawn closer down to a margin α than the hardest non-paired ones. Our model aligns $\{V, E\}$, $\{V, G\}$ and $\{E, G\}$ in the joint embedding space for learning multilingual multimodal representations with the sampled $\{V, E, G\}$ batch. We formulate the overall triplet loss as:

$$l_\theta(V, E, G) = l_\theta(V, G) + l_\theta(V, E) + \gamma l_\theta(G, E).$$
$$(4)$$

Note that the hyper-parameter γ controls the contribution of $l_\theta(G, E)$ since (e, g) may not be a translation pair even though (e, v) and (g, v) are image-caption pairs.

One of the desired properties of multi-head attention is its ability to jointly attend to and encode different information in the embedding space. However, there is no mechanism to support that these attention heads indeed capture diverse information. To encourage the diversity among K attention heads for instances within and across modalities, we propose a new simple yet effective margin-based diversity loss. As an example, the multi-head attention diversity loss between the sampled images and the English sentences (*i.e.* diversity across-modalities) is defined as:

$$l_\theta^D(V, E) = \sum_p \sum_k \sum_r \big[\alpha_D - s(v_p^k, e_p^{k \neq r})\big]_+$$
$$(5)$$

As illustrated with the red arrows for update in Figure 1, the merit behind this diversity objective is to increase the distance (up to a diversity margin α_D) between attended embeddings from different attention heads for an instance itself or its cross-modal parallel instances. As a result, the diversity objective explicitly encourage multi-head

attention to concentrate on different aspects of information sparsely located in the joint embedding space to promote fine-grained alignments between multilingual textual semantics and visual objects. With the fact that the shared embedding space is multilingual and multimodal, for improving both intra-modal/lingual and inter-modal/lingual diversity, we model the overall diversity loss as:

$$l_\theta^D(V, E, G) = l_\theta^D(V, V) + l_\theta^D(G, G) + l_\theta^D(E, E)$$
$$+ l_\theta^D(V, E) + l_\theta^D(V, G) + l_\theta^D(G, E),$$
$$(6)$$

where the first three terms are intra-modal/lingual and the rest are cross-modal/lingual. With Eq. 4 and Eq. 6, we formalize the final model loss as:

$$l_\theta^{All}(V, E, G) = l_\theta(V, E, G) + \beta l_\theta^D(V, E, G),$$
$$(7)$$

where β is the weighting parameter which balances the diversity loss and the triple ranking loss. We train the model by minimizing $l_\theta^{All}(V, E, G)$.

4 Experiments

Following Gella et al. (2017), we evaluate on the multilingual sentence-image matching tasks in Multi30K (Elliott et al., 2016) and the semantic textual similarity task (Agirre et al., 2012, 2015).

4.1 Experiment Setup

We use the model in Anderson et al. (2018) which is a Faster-RCNN (Ren et al., 2015) network pretrained on the MS-COCO (Lin et al., 2014) dataset and fine-tuned on the Visual Genome (Krishna et al., 2016) dataset to detect salient visual objects and extract their corresponding features. 1,600 types of objects are detectable. We then pack and represent each image as a 36×2048 feature matrix where 36 is the maximum amount of salient visual objects in an image and 2048 is the dimension of the flattened last pooling layer in the ResNet (He et al., 2016) backbone of Faster-RCNN.

For the text processing, we lower-case, tokenize, and then truncate the maximum sentence length to 100. We use 300-dim word embedding matrices initialized either randomly or with pre-trained multilingual embeddings. (We use the multilingual version of FastText (Mikolov et al., 2018)). We also experiment incorporating the last layer of contextualized multilingual BERT embeddings (Devlin et al., 2018) to replace the word embedding matrices as the textual input features for the bi-directional LSTMs.

Method	German to Image			Image to German			English to Image			Image to English		
	R@1	R@5	R10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R10
VSE[†][*] (Kiros et al., 2014)	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[†][*] (Vendrov et al., 2015)	21.0	48.5	60.4	26.8	57.5	70.9	25.8	56.5	67.8	34.8	63.7	74.8
DAN[*] (Nam et al., 2017)	31.0	60.9	71.0	46.5	77.5	83.0	39.4	69.2	69.1	55.0	81.8	89.0
VSE++[*] (Faghri et al., 2018)	31.3	62.2	70.9	47.5	78.5	84.5	39.6	69.1	79.8	53.1	82.1	87.5
SCAN[*] (Lee et al., 2018)	35.7	64.9	74.6	52.3	81.8	88.5	45.8	74.4	83.0	61.8	87.5	93.7
Pivot[†] (Gella et al., 2017)	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[†] (Random, VGG19)	25.8	54.9	65.1	34.1	65.5	76.5	30.1	62.5	71.6	36.4	68.0	80.9
Ours (Random, No diversity)	36.3	65.3	74.7	53.1	82.3	88.8	46.2	74.7	82.9	63.3	87.0	93.3
Ours (Random)	39.2	67.5	76.7	55.0	84.7	91.2	48.7	77.2	85.0	66.4	88.3	93.4
Ours (w/ FastText)	40.3	70.1	**79.0**	**60.4**	**85.4**	**92.0**	**50.1**	78.1	85.7	**68.0**	88.8	94.0
Ours (w/ BERT)	**40.7**	**70.5**	78.8	56.5	84.6	91.3	48.9	**78.3**	**85.8**	66.5	**89.1**	**94.1**

Table 1: Comparison of multilingual sentence-image retrieval/matching (German-Image) and (English-Image) results on Multi30K. (Visual encoders:VGG[†] otherwise ResNet or Faster-RCNN(ResNet).) (Monolingual models[*].)

Method	MS-Vid (2012)	Pascal (2014)	Pascal (2015)
STS Baseline	29.9	51.3	60.4
STS Best System	86.3	83.4	86.4
GRAN (Wieting et al., 2017)	83.7	84.5	85.0
VSE (Kiros et al., 2014)	80.6	82.7	89.6
OE (Vendrov et al., 2015)	82.2	84.1	90.8
DAN (Nam et al., 2017)	84.1	84.3	90.8
VSE++ (Faghri et al., 2018)	84.5	84.8	91.2
SCAN (Lee et al., 2018)	84.0	83.9	90.7
PIVOT (Gella et al., 2017)	84.6	84.5	91.5
Ours (w/ Random)	85.8	87.8	91.5
Ours (w/ FastText)	86.2	**88.3**	**91.8**
Ours (w/ BERT)	**86.4**	88.0	91.7

Table 2: Results on the image and video datasets of Semantic Textual Similarity task. (Pearson's $r \times 100$).

For training, we sample batches of size 128 and train 20 epochs on the training set of Multi30K. We use the Adam (Kingma and Ba, 2014) optimizer with 2×10^{-4} learning rate then 2×10^{-5} after 15-*th* epoch. Models with the best summation of validation R@1,5,10 are selected to generate image and sentence embeddings for testing. Weight decay is set to 10^{-6} and gradients larger than 2.0 are clipped. We use 3-head attention ($K = 3$) and the embedding dimension $H = 512$. The same dimension is shared by all the context vectors in the attention modules. Other hyperparameters are set as follows: $\alpha = 0.2, \alpha_D = 0.1, \beta = 1.0$ and $\gamma = 0.6$.

4.2 Multilingual Sentence-Image Matching

We evaluate the proposed model in the multilingual sentence-image matching (retrieval) tasks on Multi30K: (i) Searching images with text queries (Sentence-to-Image). (ii) Ranking descriptions with image queries (Image-to-Sentence). English and German are considered.

Multi30K (Elliott et al., 2016) is the multilingual extension of Flickr30K (Young et al., 2014).

The training, validation, and testing split contain 29K, 1K, and 1K images respectively. Two types of annotations are available in Multi30K: (i) One parallel English-German translation for each image and (ii) five independently collected English and five German descriptions/captions for each image. We use the later. Note that the German and English descriptions are *not* translations of each other and may describe an image differently.

As other prior works, we use recall at k (R@k) to measure the standard ranking-based retrieval performance. Given a query, R@k calculates the percentage of test instances for which the correct one can be found in the top-k retrieved instances. Higher R@k is preferred.

Table 1 presents the results on the Multi30K testing set. The VSE baselines in the first five rows are trained with English and German descriptions independently. In contrast, PIVOT (Gella et al., 2017) and the proposed model are capable of handling multilingual input queries with single model. For a fair comparison with PIVOT, we also report the result of swapping Faster-RCNN with VGG as the visual feature encoder in our model.

As can be seen, the proposed models successfully obtain state-of-the-art results, outperforming other baselines by a significant margin. German-Image matching benefit more from joint training with English-Image pairs. The models with pre-trained multilingual embeddings and contextualized embeddings achieve better performance in comparison to randomly initialized word embeddings, especially for German. One explanation is that the degradation from German singletons is alleviated by the multi-task training with English and the pre-trained embeddings. While the model with BERT performs better in English, FastText is preferred for German-Image matching.

The person in the striped shirt is mountain climbing .

Zwei personen sitzen auf einem rasen vor einem gebäude .

A man is repelling down a cliff next to a lake .

Figure 2: Qaulitative text-to-image matching results on Multi30K. Correct (colored in green) if ranked first.

The man with pierced ears is wearing glasses and an orange hat
Der mann trägt eine orange wollmütze .

Three children in football uniforms of two different teams are playing football on a football field .
3 kinder am sportplatz , zwei im blauen dress , einer im schwarz-weißen mit blauen schutzhelmen rangeln .

A woman midair vaulting over a bar .
Die frau springt über die stange auf die matte .

Figure 3: t-SNE visualization and grounding of the learned multilingual multimodal embeddings on Multi30K. Note the sentences are *not* translation pairs.

4.3 Semantic Textual Similarity Results

For semantic textual similarity (STS) tasks, we evaluate on the video task from STS-2012 (Agirre et al., 2012) and the image tasks from STS-2014-15 (Agirre et al., 2014, 2015). The video descriptions are from the MSR video description corpus (Chen and Dolan, 2011) and the image descriptions are from the PASCAL dataset (Rashtchian et al., 2010). In STS, a system takes two input sentences and output a semantic similarity ranging from [0,5]. Following Gella et al. (2017), we directly use the model trained on Multi30K to generate sentence embeddings then scaled the cosine similarity between the two embeddings as the prediction.

Table 2 lists the standard Pearson correlation coefficients r between the system predictions and the STS gold-standard scores. We report the best scores achieved by paraphrastic embeddings (Wieting et al., 2017) (text only) and the VSE models in the previous section. Note that the compared VSE models are all with RNN as the text encoder and *no* STS data is used for training. Our models achieve the best performance and the pre-trained word embeddings are preferred.

4.4 Qualitative Results and Grounding

In Figure 2 we samples some qualitative multilingual text-to-image matching results. In most cases our model successfully retrieve the one and only one correct image. Figure 3 depicts the t-SNE visualization of the learned visually grounded multilingual embeddings of the (v, e, g) pairs pivoted on v in the Multi30K testing set. As evidenced,

although the English and German sentences describe different aspects of the image, our model correctly aligns the shared semantics (*e.g.* ("*man*", "*mann*"), ("*hat*", "*wollmtze*")) in the embedding space. Notably, the embeddings are visually-grounded as our model associate the multilingual phrases with exact visual objects (*e.g. glasses* and *ears*). We consider learning grounded multilingual multimodal dictionary as the promising next step.

As limitations we notice that actions and small objects are harder to align. Additionally, the alignments tends to be noun-phrase/object-based whereas spatial relationships (*e.g.* "*on*", "*over*") and quantifiers remain not well-aligned. Resolving these limitations will be our future work.

5 Conclusion

We have presented a novel VSE model facilitating multi-head attention with diversity to align different types of textual semantics and visual objects for learning grounded multilingual multimodal representations. The proposed model obtains state-of-the-art results in the multilingual sentence-image matching task and the semantic textual similarity task on two benchmark datasets.

Acknowledgement

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 252–263.

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 81–91.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, volume 1, pages 385–393.

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*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Multilingual multi-modal embeddings for natural language processing. *arXiv preprint arXiv:1702.01101*.

David Chen and William Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 190–200.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics.

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

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives.

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*, pages 2839–2845. Association for Computational Linguistics.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jiuxiang Gu, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang. 2018. Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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*, pages 770–778.

Po-Yao Huang, Guoliang Kang, Liu Wenhe, Xiaojun Chang, and Alexander G. Hauptmann. 2019a. Annotation efficient cross-modal retrieval with adversarialattentive alignment. In *2019 ACM Multimedia Conference on Multimedia Conference*. ACM.

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

Po-Yao Huang, Vaibhav, Xiaojun Chang, and Alexander G. Hauptmann. 2019b. 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, pages 244–252, New York, NY, USA. ACM.

Ákos Kádár, Desmond Elliott, Marc-Alexandre Côté, Grzegorz Chrupała, and Afra Alishahi. 2018. Lessons learned in multilingual grounded language learning. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 402–412, Brussels, Belgium. Association for Computational Linguistics.

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*, pages 3128–3137.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *NIPS Workshop*.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4437–4446.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations.

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*.

Jian Li, Zhaopeng Tu, Baosong Yang, Michael R Lyu, and Tong Zhang. 2018. Multi-head attention with disagreement regularization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2897–2903.

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*, pages 740–755. Springer.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

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*, pages 2156–2164. IEEE.

Janarthanan Rajendran, Mitesh M Khapra, Sarath Chandar, and Balaraman Ravindran. 2016. Bridge correlational neural networks for multilingual multimodal representation learning. In *Proceedings of NAACL-HLT*, pages 171–181.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics.

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*, pages 91–99.

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*, pages 3715–3727.

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*, pages 5998–6008.

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

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*.

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*, pages 5005–5013. IEEE.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285.

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