Aligning Multilingual Word Embeddings for Cross-Modal Retrieval Task

Alireza Mohammadshahi
IDIAP Research Inst.
EPFL
alireza.mohammadshahi@epfl.ch

Rémi Lebret
EPFL
remi.lebret@epfl.ch

Karl Aberer
EPFL
karl.aberer@epfl.ch

Abstract

In this paper, we propose a new approach to learn multimodal multilingual embeddings for matching images and their relevant captions in two languages. We combine two existing objective functions to make images and captions close in a joint embedding space while adapting the alignment of word embeddings between existing languages in our model. We show that our approach enables better generalization, achieving state-of-the-art performance in text-to-image and image-to-text retrieval task, and caption-caption similarity task. Two multimodal multilingual datasets are used for evaluation: Multi30k with German and English captions and Microsoft-COCO with English and Japanese captions.

1 Introduction

In recent years, there has been a huge and significant amount of research in text and image retrieval tasks which needs the joint modeling of both modalities. Further, a large number of image-text datasets have become available (Elliott et al., 2016; Hodosh et al., 2013; Young et al., 2014; Lin et al., 2014), and several models have been proposed to generate captions for images in the dataset (Lu et al., 2018; Bernardi et al., 2016; Anderson et al., 2017; Lu et al., 2016; Mao et al., 2014; Rennie et al., 2016). There has been a great amount of research in learning a joint embedding space for texts and images in order to use the model in sentence-based image search or cross-modal retrieval task (Frome et al., 2013; Kiros et al., 2014; Donahue et al., 2014; Lazaridou et al., 2015; Socher et al., 2013; Hodosh et al., 2013; Karpathy et al., 2014).

Previous works in image-caption task and learning a joint embedding space for texts and images are mostly related to English language, however, recently there is a large amount of research in other languages due to the availability of multilingual datasets (Funaki and Nakayama, 2015; Elliott et al., 2016; Rajendran et al., 2015; Miyazaki and Shimizu, 2016; Lucia Specia and Elliott, 2016; Young et al., 2014; Hitschler and Riezler, 2016; Yoshikawa et al., 2017). The aim of these models is to map images and their captions in a single language into a joint embedding space (Rajendran et al., 2015; Calixto et al., 2017).

Related to our work, Gella et al. (2017) proposed a model to learn a multilingual multimodal embedding by utilizing an image as a pivot between languages of captions. While a text encoder is trained for each language in Gella et al. (2017), we propose instead a model that learns a shared and language-independent text encoder between languages, yielding better generalization. It is generally important to adapt word embeddings for the task at hand. Our model enables tuning of word embeddings while keeping the two languages aligned during training, building a task-specific shared embedding space for existing languages.

In this attempt, we define a new objective function that combines a pairwise ranking loss with a loss that maintains the alignment in multiple languages. For the latter, we use the objective function proposed in Joulin et al. (2018) for learning a linear mapping between languages inspired by cross-domain similarity local scaling (CSLS) retrieval criterion (Conneau et al., 2017) which obtains the state-of-the-art performance on word translation task.

In the next sections, the proposed approach is called Aligning Multilingual Embeddings for cross-modal retrieval (AME). With experiments on two multimodal multilingual datasets, we show that AME outperforms existing models on text-image multimodal retrieval tasks. The code we used to train and evaluate the model is available at https://github.com/alirezamshi/AME-CMR

Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), pages 11–17
Hong Kong, China, November 3, 2019.

2 Datasets

We use two multilingual image-caption datasets to evaluate our model, Multi30k and Microsoft COCO (Elliott et al., 2016; Lin et al., 2014).

Multi30K is a dataset with 31'014 German translations of English captions and 155'070 independently collected German and English captions. In this paper, we use independently collected captions which each image contains five German and five English captions. The training set includes 29'000 images. The validation and test sets contain 1'000 images.

MS-COCO (Lin et al., 2014) contains 123'287 images and five English captions per image. Yoshikawa et al. (2017) proposed a model which generates Japanese descriptions for images. We divide the dataset based on Karpathy and Li (2014). The training set contains 113'287 images. Each validation and test set contains 5'000 images.

3 Problem Formulation

3.1 Model for Learning a Multilingual Multimodal Representation

Assume image i and captions c_{X_i} and c_{Y_i} are given in two languages, X and Y respectively. Our aim is to learn a model where the image i and its captions c_{X_i} and c_{Y_i} are close in a joint embedding space of dimension m. AME consists of two encoders f_i and f_c, which encode images and captions. As multilingual text encoder, we use a recurrent neural network with gated recurrent unit (GRU). For the image encoder, we use a convolutional neural network (CNN) architecture. The similarity between a caption c and an image i in the joint embedding space is measured with a similarity function $P(c, i)$. The objective function is as follows (inspired by Gella et al. (2017)):

$$L_R = \sum_{(c_{S_i}, i)} \Big(\sum_{c_{S_j}} max\{0, \alpha - P(c_{S_i}, i) + P(c_{S_j}, i)\} + \sum_{j} max\{0, \alpha - P(c_{S_i}, i) + P(c_{S_i}, j)\} \Big) \tag{1}$$

Where S stands for both languages, and α is the margin. c_{S_j} and j are irrelevant caption and image of the gold-standard pair (c_{S_i}, i).

3.2 Alignment Model

Each word k in the language X is defined by a word embedding $\mathrm{x}_k \in \mathbb{R}^d$ ($\mathrm{y}_k \in \mathbb{R}^d$ in the language Y respectively). Given a bilingual lexicon of N pairs of words, we assume the first n pairs $\{(\mathrm{x}_i, \mathrm{y}_i)\}_{i=1}^{n}$ are the initial seeds, and our aim is to augment it to all word pairs that are not in the initial lexicons. Mikolov et al. (2013) proposed a model to learn a linear mapping $\boldsymbol{W} \in \mathbb{R}^{d \times d}$ between the source and target languages:

$$min_{\boldsymbol{W} \in \mathbb{R}^{d \times d}} \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{W} x_i, y_i | x_i, y_i) \\ \ell(\boldsymbol{W} x_i, y_i | x_i, y_i) = (\boldsymbol{W} x_i - y_i)^2 \tag{2}$$

Where ℓ is a square loss. One can find the translation of a source word in the target language by performing a nearest neighbor search with Euclidean distance. But, the model suffers from a "hubness problem": some word embeddings become uncommonly the nearest neighbors of a great number of other words (Doddington et al., 1998; Dinu and Baroni, 2014).

In order to resolve this issue, Joulin et al. (2018) proposed a new objective function inspired by CSLS criterion to learn the linear mapping:

$$L_A = \frac{1}{n} \sum_{i=1}^{n} -2\mathrm{x}_i^T \boldsymbol{W}^T \mathrm{y}_i + \frac{1}{k} \sum_{\mathrm{y}_j \in \mathcal{N}_Y(\boldsymbol{W}\mathrm{x}_i)} \mathrm{x}_i^T \boldsymbol{W}^T \mathrm{y}_j + \frac{1}{k} \sum_{\boldsymbol{W}\mathrm{x}_j \in \mathcal{N}_X(\mathrm{y}_i)} \mathrm{x}_j^T \boldsymbol{W}^T \mathrm{y}_i \tag{3}$$

Where $\mathcal{N}_X(\mathrm{y}_i)$ means the k-nearest neighbors of y_i in the set of source language X. They constrained the linear mapping $\boldsymbol{W}$ to be orthogonal, and word vectors are $l2$-normalized.

The whole loss function is the equally weighted summation of the aforementioned objective functions:

$$L_{total} = L_R + L_A \tag{4}$$

The model architecture is illustrated in Figure 1. We observe that updating the parameters in (3) every T iterations with learning rate lr_{align} obtains the best performance.

We use two different similarity functions, symmetric and asymmetric. For the former, we use the cosine similarity function and for the latter, we use the metric proposed in Vendrov et al. (2015), which encodes the partial order structure of the visual-semantic hierarchy. The metric similarity is defined as:

$$S(a, b) = -||max(0, b - a)||^2 \tag{5}$$

Where a and b are the embeddings of image and caption.

	Image to Text				Text to Image				
	R@1	R@5	R@10	Mr	R@1	R@5	R@10	Mr	Alignment
symmetric									
Parallel (Gella et al., 2017)	31.7	62.4	74.1	3	24.7	53.9	65.7	5	-
UVS (Kiros et al., 2014)	23.0	50.7	62.9	5	16.8	42.0	56.5	8	-
EmbeddingNet (Wang et al., 2017)	40.7	69.7	79.2	-	29.2	59.6	71.7	-	-
sm-LSTM (Huang et al., 2016)	42.5	71.9	81.5	2	30.2	60.4	72.3	3	-
VSE++ (Faghri et al., 2017)	**43.7**	71.9	82.1	2	32.3	60.9	72.1	3	-
Mono	41.4	74.2	84.2	2	32.1	63.0	73.9	3	-
FME	39.2	71.1	82.1	2	29.7	62.5	74.1	3	76.81%
AME	43.5	**77.2**	**85.3**	**2**	**34.0**	**64.2**	**75.4**	**3**	66.91%
asymmetric									
Pivot (Gella et al., 2017)	33.8	62.8	75.2	3	26.2	56.4	68.4	4	-
Parallel (Gella et al., 2017)	31.5	61.4	74.7	3	27.1	56.2	66.9	4	-
Mono	47.7	77.1	86.9	2	35.8	66.6	76.8	3	-
FME	44.9	76.9	86.4	2	34.2	66.1	77.1	3	76.81%
AME	**50.5**	**79.7**	**88.4**	**1**	**38.0**	**68.5**	**78.4**	**2**	73.10%

Table 1: Image-caption ranking results for English (Multi30k)

	Image to Text				Text to Image				
	R@1	R@5	R@10	Mr	R@1	R@5	R@10	Mr	Alignment
symmetric									
Parallel (Gella et al., 2017)	28.2	57.7	71.3	4	20.9	46.9	59.3	6	-
Mono	34.2	67.5	79.6	3	26.5	54.7	66.2	4	-
FME	36.8	69.4	80.8	2	26.6	56.2	68.5	4	76.81%
AME	**39.6**	**72.7**	**82.7**	**2**	**28.9**	**58.0**	**68.7**	**4**	66.91%
asymmetric									
Pivot (Gella et al., 2017)	28.2	61.9	73.4	3	22.5	49.3	61.7	6	-
Parallel (Gella et al., 2017)	30.2	60.4	72.8	3	21.8	50.5	62.3	5	-
Mono	**42.0**	72.5	83.0	2	29.6	58.4	69.6	4	-
FME	40.5	73.3	83.4	2	29.6	59.2	**72.1**	3	76.81%
AME	40.5	**74.3**	**83.4**	**2**	**31.0**	**60.5**	70.6	**3**	73.10%

Table 2: Image-caption ranking results for German (Multi30k)

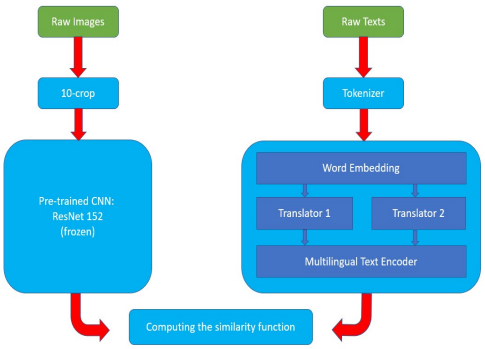

Figure 1: The AME - model architecture

4 Experiment and Results

4.1 Details of Implementation [1]

We use a mini-batch of size 128. We use Adam optimizer with learning rate 0.00011 (0.00006) and with early stopping on the validation set. We set the dimensionality of joint embedding space and the GRU hidden layer to $m = 1024$. We utilize the pre-trained aligned word vectors of FastText for the initial word embeddings. For Japanese word embedding, we use pre-trained word vectors of FastText[2], then align it to the English word embedding with the same hyper-parameters used for MS-COCO. We set the margin $\alpha = 0.2$ and $\alpha = 0.05$ for symmetric and asymmetric similarity functions respectively.

We assign k-nearest neighbors to be 5 (4). We set $T = 500$, and $lr_{align} = 2$ (5). We tokenize English and German captions with Europarl tokenizer (Koehn, 2005). For the Japanese caption, we use Mecab analyzer (Kudo et al., 2004). We train the model for 30 (20) epochs with updating the learning rate (divided by 10) on epoch 15 (10).

To extract features of images, we use a ResNet152 (He et al., 2015) CNN architecture pre-trained on Imagenet and extract the image features from FC7, the penultimate fully connected layer. We use average features from 10-crop of the re-scaled images.

For the metric of alignment, we use bilingual lexicons of Multilingual Unsupervised and Super-

[1] In this section, the hyper-parameters in parentheses are related to the model trained on MS-COCO.

[2] Available at https://fasttext.cc/docs/en/crawl-vectors.html, and https://fasttext.cc/docs/en/aligned-vectors.html.

	Image to Text				Text to Image				
	R@1	R@5	R@10	Mr	R@1	R@5	R@10	Mr	Alignment
symmetric									
UVS (Kiros et al., 2014)	43.4	75.7	85.8	2	31.0	66.7	79.9	3	-
EmbeddingNet (Wang et al., 2017)	50.4	79.3	89.4	-	39.8	75.3	86.6	-	-
sm-LSTM (Huang et al., 2016)	53.2	83.1	91.5	1	40.7	75.8	87.4	2	-
VSE++ (Faghri et al., 2017)	**58.3**	**86.1**	93.3	1	**43.6**	77.6	87.8	2	-
Mono	51.8	84.8	93.5	1	40.0	77.3	89.4	2	-
FME	42.2	76.6	91.1	2	31.2	69.2	83.7	3	92.70%
AME	54.6	85	**94.3**	**1**	42.1	**78.7**	**90.3**	**2**	82.54%
asymmetric									
Mono	53.2	87.0	94.7	1	42.3	78.9	90	2	-
FME	48.3	83.6	93.6	2	37.2	75.4	88.4	2	92.70%
AME	**58.8**	**88.6**	**96.2**	**1**	**46.2**	**82.5**	**91.9**	**2**	84.99%

Table 3: Image-caption ranking results for English (MS-COCO)

	Image to Text				Text to Image				
	R@1	R@5	R@10	Mr	R@1	R@5	R@10	Mr	Alignment
symmetric									
Mono	42.7	77.7	88.5	2	33.1	69.8	84.3	3	-
FME	40.7	77.7	88.3	2	30.0	68.9	83.1	3	92.70%
AME	**50.2**	**85.6**	**93.1**	**1**	**40.2**	**76.7**	**87.8**	**2**	82.54%
asymmetric									
Mono	49.9	83.4	93.7	2	39.7	76.5	88.3	**2**	-
FME	48.8	81.9	91.9	2	37.0	74.8	87.0	**2**	92.70%
AME	**55.5**	**87.9**	**95.2**	**1**	**44.9**	**80.7**	**89.3**	**2**	84.99%

Table 4: Image-caption ranking results for Japanese (MS-COCO)

	EN → DE			DE → EN		
	R@1	R@5	R@10	R@1	R@5	R@10
FME	51.4	76.4	84.5	46.9	71.2	79.1
AME	**51.7**	**76.7**	**85.1**	**49.1**	**72.6**	**80.5**

Table 5: Textual similarity scores (asymmetric, Multi30k).

vised Embeddings (MUSE) benchmark (Lample et al., 2017). MUSE is a large-scale high-quality bilingual dictionaries for training and evaluating the translation task. We extract the training words of descriptions in two languages. For training, we combine "full" and "test" sections of MUSE, then filter them to the training words. For evaluation, we filter "train" section of MUSE to the training words. [3]

For evaluating the benefit of the proposed objective function, we compare AME with monolingual training (Mono), and multilingual training without the alignment model described in Section 3.2. For the latter, the pre-aligned word embeddings are frozen during training (FME). We add Mono since the proposed model in Gella et al. (2017) did not utilize pre-trained word embeddings for the initialization, and the image encoder is different (ResNet152 vs. VGG19).

We compare models based on two retrieval metrics, recall at position k (R@k) and Median of ranks (Mr).

4.2 Multi30k Results

In Table 1 and 2, we show the results for English and German captions. For English captions, we see 21.28% improvement on average compared to Kiros et al. (2014). There is a 1.8% boost on average compared to Mono due to more training data and multilingual text encoder. AME performs better than FME model on both symmetric and asymmetric modes, which shows the advantage of fine-tuning word embeddings during training. We have 25.26% boost on average compared to Kiros et al. (2014) in asymmetric mode.

For German descriptions, The results are 11.05% better on average compared to (Gella et al., 2017) in symmetric mode. AME also achieves competitive or better results than FME model in German descriptions too.

4.3 MS-COCO Results[4]

In Table 3 and 4, we show the performance of AME and baselines for English and Japanese captions. We achieve 10.42% improvement on aver-

[3] You can find the code for building bilingual lexicons on the Github link.

[4] To compare with baselines, scores are measured by averaging 5 folds of 1K test images.

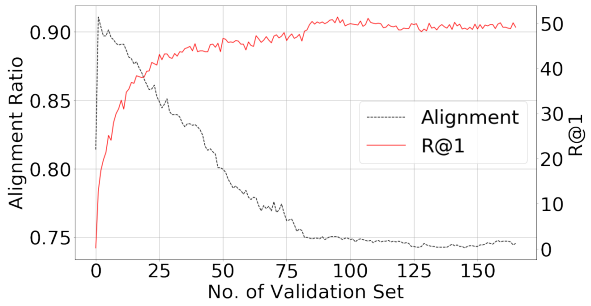

Figure 2: Alignment ratio in each validation step (asymmetric mode - image-to-text - Multi30k dataset)

age compared to Kiros et al. (2014) in the symmetric manner. We show that adapting the word embedding for the task at hand, boosts the general performance, since AME model significantly outperforms FME model in both languages.

For the Japanese captions, AME reaches 6.25% and 3.66% better results on average compared to monolingual model in symmetric and asymmetric modes, respectively.

4.4 Alignment results

In Tables 1 and 2, we can see that the alignment ratio for AME is 6.80% lower than FME which means that the translators can almost keep languages aligned in Multi30k dataset. In MS-COCO dataset, the alignment ratio for AME is 8.93% lower compared to FME.

We compute the alignment ratio and recall at position 1 (R@1) in each validation step. Figure 2 shows the trade-off between alignment and retrieval tasks. At the first few epochs, the model improves the alignment ratio since the retrieval task hasn't seen enough number of instances. Then, the retrieval task tries to fine-tune word embeddings. Finally, they reach an agreement near the half of training process. At this point, we update the learning rate of retrieval task to improve the performance, and the alignment ratio preserves constant.

Additionally, we also train AME model without adding the alignment objective function, and the model breaks the alignment between the initial aligned word embeddings, so it's essential to add the alignment objective function to the retrieval task.

4.5 Caption-Caption Similarity Scores

Given the caption in a language, the task is to retrieve the related caption in another language. In Table 5, we show the performance on Multi30k dataset in asymmetric mode. AME outperforms the FME model, confirming the importance of word embeddings adaptation.

5 Conclusion

We proposed a multimodal model with a shared multilingual text encoder by adapting the alignment between languages for image-description retrieval task while training. We introduced a loss function which is a combination of a pairwise ranking loss and a loss that maintains the alignment of word embeddings in multiple languages. Through experiments with different multimodal multilingual datasets, we have shown that our approach yields better generalization performance on image-to-text and text-to-image retrieval tasks, as well as caption-caption similarity task.

In the future work, we can investigate on applying self-attention models like Transformer (Vaswani et al., 2017) on the shared text encoder to find a more comprehensive representation for descriptions in the dataset. Additionally, we can explore the effect of a weighted summation of two loss functions instead of equally summing them together.

6 Acknowledgement

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2017. Bottom-up and top-down attention for image captioning and VQA. *CoRR*, abs/1707.07998.

Raffaella Bernardi, Ruket Çakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *CoRR*, abs/1601.03896.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Multilingual multi-modal embeddings for natural language processing. *CoRR*, abs/1702.01101.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *CoRR*, abs/1710.04087.

Georgiana Dinu and Marco Baroni. 2014. Improving zero-shot learning by mitigating the hubness problem. *CoRR*, abs/1412.6568.

George Doddington, Walter Liggett, Alvin Martin, Mark Przybocki, and Douglas Reynolds. 1998. Sheep, goats, lambs and wolves a statistical analysis of speaker performance in the nist 1998 speaker recognition evaluation. In *INTERNATIONAL CONFERENCE ON SPOKEN LANGUAGE PROCESSING*.

Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2014. Long-term recurrent convolutional networks for visual recognition and description. *CoRR*, abs/1411.4389.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *CoRR*, abs/1605.00459.

Fartash Faghri, David J. Fleet, Ryan Kiros, and Sanja Fidler. 2017. VSE++: improved visual-semantic embeddings. *CoRR*, abs/1707.05612.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129. Curran Associates, Inc.

Ruka Funaki and Hideki Nakayama. 2015. Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 585–590. Association for Computational Linguistics.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. *CoRR*, abs/1707.07601.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Julian Hitschler and Stefan Riezler. 2016. Multimodal pivots for image caption translation. *CoRR*, abs/1601.03916.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. In *J. Artif. Intell. Res.*

Yan Huang, Wei Wang, and Liang Wang. 2016. Instance-aware image and sentence matching with selective multimodal LSTM. *CoRR*, abs/1611.05588.

Armand Joulin, Piotr Bojanowski, Tomas Mikolov, Hervé Jégou, and Edouard Grave. 2018. Loss in translation: Learning bilingual word mapping with a retrieval criterion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1889–1897. Curran Associates, Inc.

Andrej Karpathy and Fei-Fei Li. 2014. Deep visual-semantic alignments for generating image descriptions. *CoRR*, abs/1412.2306.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In *In Proc. of EMNLP*, pages 230–237.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. *CoRR*, abs/1501.02598.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2016. Knowing when to look: Adaptive attention via A visual sentinel for image captioning. *CoRR*, abs/1612.01887.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. *CoRR*, abs/1803.09845.

Khalil Simaan Lucia Specia, Stella Frank and Desmond Elliott. 2016. A shared task on multimodal machine translation and cross-lingual image description.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. 2014. Deep captioning with multimodal recurrent neural networks (m-rnn). *CoRR*, abs/1412.6632.

Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790. Association for Computational Linguistics.

Janarthanan Rajendran, Mitesh M. Khapra, Sarath Chandar, and Balaraman Ravindran. 2015. Bridge correlational neural networks for multilingual multimodal representation learning. *CoRR*, abs/1510.03519.

Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2016. Self-critical sequence training for image captioning. *CoRR*, abs/1612.00563.

Richard Socher, Andrej Karpathy, Quoc V. Le, Chris D. Manning, and Andrew Y. Ng. 2013. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. *CoRR*, abs/1511.06361.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2017. Learning two-branch neural networks for image-text matching tasks. *CoRR*, abs/1704.03470.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR captions: Constructing a large-scale japanese image caption dataset. *CoRR*, abs/1705.00823.

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