Image Pivoting for Learning Multilingual Multimodal Representations

Spandana Gella   Rico Sennrich   Frank Keller   Mirella Lapata
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
{spandana.gella, rico.sennrich}@ed.ac.uk
{keller,mlap}@inf.ed.ac.uk

Abstract

In this paper we propose a model to learn multimodal multilingual representations for matching images and sentences in different languages, with the aim of advancing multilingual versions of image search and image understanding. Our model learns a common representation for images and their descriptions in two different languages (which need not be parallel) by considering the image as a pivot between two languages. We introduce a new pairwise ranking loss function which can handle both symmetric and asymmetric similarity between the two modalities. We evaluate our models on image-description ranking for German and English, and on semantic textual similarity of image descriptions in English. In both cases we achieve state-of-the-art performance.

1 Introduction

In recent years there has been a significant amount of research in language and vision tasks which require the joint modeling of texts and images. Examples include text-based image retrieval, image description and visual question answering. An increasing number of large image description datasets has become available (Hodosh et al., 2013; Young et al., 2014; Lin et al., 2014) and various systems have been proposed to handle the image description task as a generation problem (Bernardi et al., 2016; Mao et al., 2015; Vinyals et al., 2015; Fang et al., 2015). There has also been a great deal of work on sentence-based image search or cross-modal retrieval where the objective is to learn a joint space for images and text (Hodosh et al., 2013; Frome et al., 2013; Karpathy et al., 2014; Kiros et al., 2015; Socher et al., 2014; Donahue et al., 2015).

Previous work on image description generation or learning a joint space for images and text has mostly focused on English due to the availability of English datasets. Recently there have been attempts to create image descriptions and models for other languages (Funaki and Nakayama, 2015; Elliott et al., 2016; Rajendran et al., 2016; Miyazaki and Shimizu, 2016; Specia et al., 2016; Li et al., 2016; Hitschler et al., 2016; Yoshikawa et al., 2017).

Most work on learning a joint space for images and their descriptions is based on Canonical Correlation Analysis (CCA) or neural variants of CCA over representations of image and its descriptions (Hodosh et al., 2013; Andrew et al., 2013; Yan and Mikolajczyk, 2015; Gong et al., 2014; Chandar et al., 2016). Besides CCA, a few others learn a visual-semantic or multimodal embedding space of image descriptions and representations by optimizing a ranking cost function (Kiros et al., 2015; Socher et al., 2014; Ma et al., 2015; Vendrov et al., 2016) or by aligning image regions (objects) and segments of the description (Karpathy et al., 2014; Plummer et al., 2015) in a common space. Recently Lin and Parikh (2016) have leveraged visual question answering models to encode images and descriptions into the same space.

However, all of this work is targeted at monolingual descriptions, i.e., mapping images and descriptions in a single language onto a joint embedding space. The idea of pivoting or bridging is not new and language pivoting is well explored for machine translation (Wu and Wang, 2007; Firat et al., 2016) and to learn multilingual multimodal representations (Rajendran et al., 2016; Calixto et al., 2017). Rajendran et al. (2016) propose a
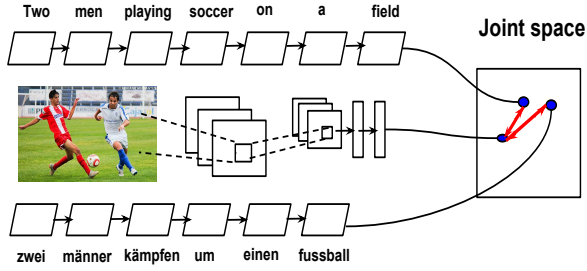
Figure 1: Our multilingual multimodal model with image as pivot

model to learn common representations between M views and assume there is parallel data available between a pivot view and the remaining $M-1$ views. Their multimodal experiments are based on English as the pivot and use large parallel corpora available between languages to learn their representations.

Related to our work Calixto et al. (2017) proposed a model for creating multilingual multimodal embeddings. Our work is different from theirs in that we choose the image as the pivot and use a different similarity function. We also propose a single model for learning representations of images and multiple languages, whereas their model is language-specific.

In this paper, we learn multimodal representations in multiple languages, i.e., our model yields a joint space for images and text in multiple languages using the image as a pivot between languages. We propose a new objective function in a multitask learning setting and jointly optimize the mappings between images and text in two different languages.

2   Dataset

We experiment with the Multi30k dataset, a multilingual extension of Flickr30k corpus (Young et al., 2014) consisting of English and German image descriptions (Elliott et al., 2016). The Multi30K dataset has 29k, 1k and 1k images in the train, validation and test splits respectively, and contains two types of multilingual annotations: (i) a corpus of one English description per image and its translation into German; and (ii) a corpus of five independently collected English and German descriptions per image. We use the independently collected English and German descriptions to train our models. Note that these descriptions are not

translations of each other, i.e., they are not parallel, although they describe the same image.

3   Problem Formulation

Given an image i and its descriptions c_1 and c_2 in two different languages our aim is to learn a model which maps i, c_1 and c_2 onto same common space $\mathbb{R}^N$ (where N is the dimensionality of the embedding space) such that the image and its gold-standard descriptions in both languages are mapped close to each other (as shown in Figure 1). Our model consists of the embedding functions f_i and f_c to encode images and descriptions and a scoring function S to compute the similarity between a description–image pair.

In the following we describe two models: (i) the PIVOT model that uses the image as pivot between the description in both the languages; (ii) the PARALLEL model that further forces the image descriptions in both languages to be closer to each other in the joint space. We build two variants of PIVOT and PARALLEL with different similarity functions S to learn the joint space.

3.1   Multilingual Multimodal Representation Models

In both PIVOT and PARALLEL we use a deep convolutional neural network architecture (CNN) to represent the image i denoted by $f_i(i) = W_i \cdot CNN(i)$ where W_i is a learned weight matrix and $CNN(i)$ is the image vector representation. For each language we define a recurrent neural network encoder $f_c(c_k) = GRU(c_k)$ with gated recurrent units (GRU) activations to encode the description c_k.

In PIVOT, we use monolingual corpora from multiple languages of sentences aligned with images to learn the joint space. The intuition of this model is that an image is a universal representation across all languages, and if we constrain a sentence representation to be closer to image, sentences in different languages may also come closer. Accordingly we design a loss function as follows:

$$loss_{pivot} = \sum_k \left[\sum_{(c_k,i)} \left(\sum_{c_k'} \max\{0, \alpha - S(c_k,i) + S(c_k',i)\} \right. \right.$$
$$\left. \left. + \sum_{i'} \max\{0, \alpha - S(c_k,i) + S(c_k,i')\} \right) \right] \tag{1}$$

where k stands for each language. This loss function encourages the similarity $S(c_k, i)$ between gold-standard description c_k and image i to be greater than any other irrelevant description c'_k by a margin α. A similar loss function is useful for learning multimodal embeddings in a single language (Kiros et al., 2015). For each minibatch, we obtain invalid descriptions by selecting descriptions of other images except the current image of interest and vice-versa.

In PARALLEL, in addition to making an image similar to a description, we make multiple descriptions of the same image in different languages similar to each other, based on the assumption that these descriptions, although not parallel, share some commonalities. Accordingly we enhance the previous loss function with an additional term:

$$
\begin{aligned}
loss_{para} = loss_{pivot} + \sum_{(c_1,c_2)} \Bigg(\sum_{c'_1} \max\{0, \alpha - S(c_1, c_2) \\
+ S(c'_1, c_2)\} + \sum_{c'_2} \max\{0, \alpha - S(c_1, c_2) + S(c_1, c'_2)\} \Bigg)
\end{aligned}
\tag{2}
$$

Note that we are iterating over all pairs of descriptions (c_1, c_2), and maximizing the similarity between descriptions of the same image and at the same time minimizing the similarity between descriptions of different images.

We learn models using two similarity functions: symmetric and asymmetric. For the former we use cosine similarity and for the latter we use the metric of Vendrov et al. (2016) which is useful for learning embeddings that maintain an order, e.g., dog and cat are more closer to pet than animal while being distinct. Such ordering is shown to be useful in building effective multimodal space of images and texts. An analogy in our setting would be two descriptions of an image are closer to the image while at the same time preserving the identity of each (which is useful when sentences describe two different aspects of the image). The similarity metric is defined as:

$$
S(a, b) = -||max(0, b - a)||^2
\tag{3}
$$

where a and b are embeddings of image and description.

We call the symmetric similarity variants of our models as PIVOT-SYM and PARALLEL-SYM, and the asymmetric variants PIVOT-ASYM and PARALLEL-ASYM.

4 Experiments and Results

We test our model on the tasks of image-description ranking and semantic textual similarity. We work with each language separately. Since we learn embeddings for images and languages in the same semantic space, our hope is that the training data for each modality or language acts complementary data for the another modality or language, and thus helps us learn better embeddings.

Experiment Setup We sampled minibatches of size 64 images and their descriptions, and drew all negative samples from the minibatch. We trained using the Adam optimizer with learning rate 0.001, and early stopping on the validation set. Following Vendrov et al. (2016) we set the dimensionality of the embedding space and the GRU hidden layer N to 1024 for both English and German. We set the dimensionality of the learned word embeddings to 300 for both languages, and the margin α to 0.05 and 0.2, respectively, to learn asymmetric and symmetric similarity-based embeddings.[1] We keep all hyperparameters constant across all models. We used the L2 norm to mitigate over-fitting (Kiros et al., 2015). We tokenize and truecase both English and German descriptions using the Moses Decoder scripts.[2]

To extract image features, we used a convolutional neural network model trained on 1.2M images of 1000 class ILSVRC 2012 object classification dataset, a subset of ImageNet (Russakovsky et al., 2015). Specifically, we used VGG 19-layer CNN architecture and extracted the activations of the penultimate fully connected layer to obtain features for all images in the dataset (Simonyan and Zisserman, 2015). We use average features from 10 crops of the re-scaled images.[3]

Baselines As baselines we use monolingual models, i.e., models trained on each language separately. Specifically, we use Visual Semantic Embeddings (VSE) of Kiros et al. (2015) and Order Embeddings (OE) of Vendrov et al. (2016). We

---

[1] We constrain the embeddings of descriptions and images to have non-negative entries when using asymmetric similarity by taking their absolute value.

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

[3] We rescale images so that the smallest side is 256 pixels wide, we take 224×224 crops from the corners, center, and their horizontal reflections to get 10 crops for the image.

System	Text to Image				Image to Text			
	R@1	R@5	R@10	Mr	R@1	R@5	R@10	Mr
VSE (Kiros et al., 2015)	23.3	53.6	65.8	5	31.6	60.4	72.7	3
OE (Vendrov et al., 2016)	25.8	**56.5**	67.8	4	**34.8**	**63.7**	74.8	3
PIVOT-SYM	23.5	53.4	65.8	5	31.6	61.2	73.8	3
PARALLEL-SYM	24.7	53.9	65.7	5	31.7	62.4	74.1	3
PIVOT-ASYM	26.2	56.4	**68.4**	4	33.8	62.8	**75.2**	3
PARALLEL-ASYM	**27.1**	56.2	66.9	4	31.5	61.4	74.7	3

Table 1: Image-description ranking results of English on Flickr30k test data.

System	Text to Image				Image to Text			
	R@1	R@5	R@10	Mr	R@1	R@5	R@10	Mr
VSE (Kiros et al., 2015)	20.3	47.2	60.1	6	29.3	58.1	71.8	4
OE (Vendrov et al., 2016)	21.0	48.5	60.4	6	26.8	57.5	70.9	4
PIVOT-SYM	20.3	46.4	59.2	6	26.9	56.6	70.0	4
PARALLEL-SYM	20.9	46.9	59.3	6	28.2	57.7	71.3	4
PIVOT-ASYM	**22.5**	49.3	61.7	6	28.2	**61.9**	**73.4**	**3**
PARALLEL-ASYM	21.8	**50.5**	**62.3**	**5**	**30.2**	60.4	72.8	**3**

Table 2: Image-description ranking results of German on Flickr30k test data.

Image	Descriptions	Image Rank		
		OE	PIVOT	PARALLEL
	2 Menschen auf der Straße mit Megafon	141	37	6
	two people in blue shirts are outside with a bullhorn	85	7	3
	ein Verkäufer mit weißem Hut und blauem Hemd , verkauft Kartoffeln oder ähnliches an Männer und Frauen	36	1	3
	at an outdoor market , a small group of people stoop to buy potatoes from a street vendor , who has his goods laid out on the ground	24	2	2

Table 3: The rank of the gold-standard image when using each German and English descriptions as a query on models trained using asymmetric similarity.

use a publicly available implementation to train both VSE and OE.[4]

4.1 Image-Description Ranking Results

To evaluate the multimodal multilingual embeddings, we report results on an image-description ranking task. Given a query in the form of a description or an image, the task its to retrieve all images or descriptions sorted based on the relevance. We use the standard ranking evaluation metrics of recall at position k (R@K, where higher is better) and median rank (Mr, where lower is better) to evaluate our models. We report results for both English and German descriptions. Note that we have one single model for both languages.

In Tables 1 and 2 we present the ranking results of the baseline models of Kiros et al. (2015) and Vendrov et al. (2016) and our proposed PIVOT and PARALLEL models. We do not compare our image-description ranking results with Calixto et al. (2017) since they report results on half of validation set of Multi30k whereas our results are on the publicly available test set of Multi30k. For English, PIVOT with asymmetric similarity is either competitive or better than monolingual models

and symmetric similarity, especially in the R@10 category it obtains state-of-the-art. For German, both PIVOT and PARALLEL with the asymmetric scoring function outperform monolingual models and symmetric similarity. We also observe that the German ranking experiments benefit the most from the multilingual signal. A reason for this could be that the German description corpus has many singleton words (more than 50% of the vocabulary) and English description mapping might have helped in learning better semantic embeddings. These results suggest that the multilingual signal could be used to learn better multimodal embeddings, irrespective of the language. Our results also show that the asymmetric scoring function can help learn better embeddings. In Table 3 we present a few examples where PIVOT-ASYM and PARALLEL-ASYM models performed better on both the languages compared to baseline order embedding model even using descriptions of very different lengths as queries.

4.2 Semantic Textual Similarity Results

In the semantic textual similarity task (STS), we use the textual embeddings from our model to compute the similarity between a pair of sen-

---

[4] https://github.com/ivendrov/order-embedding

Model	VF	2012	2014	2015
Shared Task Baseline	–	29.9	51.3	60.4
STS Best System	–	87.3	83.4	86.4
GRAN (Wieting et al., 2017)	–	83.7	84.5	85.0
MLMME (Calixto et al., 2017)	VGG19	–	72.7	79.7
VSE (Kiros et al., 2015)	VGG19	80.6	82.7	89.6
OE (Vendrov et al., 2016)	VGG19	82.2	84.1	90.8
PIVOT-SYM	VGG19	80.5	81.8	89.2
PARALLEL-SYM	VGG19	82.0	81.4	90.4
PIVOT-ASYM	VGG19	83.1	83.8	90.3
PARALLEL-ASYM	VGG19	**84.6**	**84.5**	**91.5**

Table 4: Results on Semantic Textual Similarity Image datasets (Pearson's $r \times 100$). Our systems that performed better than best reported shared task scores are in **bold**.

S1	S2	GT	Pred
Black bird standing on concrete.	Blue bird standing on green grass.	1.0	4.2
Two zebras are playing.	Zebras are socializing.	4.2	1.2
Three goats are being rounded up by a dog.	Three goats are chased by a dog	4.6	4.5
A man is folding paper.	A woman is slicing a pepper.	0.6	0.6

Table 5: Example sentences with gold-standard semantic textual similarity score and the predicted score using our best performing PARALLEL-ASYM model.

tences (image descriptions in this case). We evaluate on video task from STS-2012 and image tasks from STS-2014, STS-2015 (Agirre et al. 2012, Agirre et al. 2014, Agirre et al. 2015). The video descriptions in the STS-2012 task are from the MSR video description corpus (Chen and Dolan, 2011) and the image descriptions in STS-2014 and 2015 are from UIUC PASCAL dataset (Rashtchian et al., 2010).

In Table 4, we present the Pearson correlation coefficients of our model predicted scores with the gold-standard similarity scores provided as part of the STS image/video description tasks. We compare with the best reported scores for the STS shared tasks, achieved by MLMME (Calixto et al., 2017), paraphrastic sentence embeddings (Wieting et al., 2017), visual semantic embeddings (Kiros et al., 2015), and order embeddings (Vendrov et al., 2016). The shared task baseline is computed based on word overlap and is high for both the 2014 and the 2015 dataset, indicating that there is substantial lexical overlap between the STS image description datasets. Our models outperform both the baseline system and the best system submitted to the shared task. For the 2012 video paraphrase corpus, our multilingual methods performed better than the monolingual methods showing that similarity across paraphrases can be learned using multilingual signals. Similarly, Wieting et al. (2017) have reported to learn better paraphrastic sentence embeddings with multilingual signals. Overall, we observe that models learned using the asymmetric scoring function outperform the state-of-the-art on these datasets, suggesting that multilingual

sharing is beneficial. Although the task has nothing to do German, because our models can make use of datasets from different languages, we were able to train on significantly larger training dataset of approximately 145k descriptions. Calixto et al. (2017) also train on a larger dataset like ours, but could not exploit this to their advantage. In Table 5 we present the example sentences with the highest and lowest difference between gold-standard and predicted semantic textual similarity scores using our best performing PARALLEL-ASYM model.

5 Conclusions

We proposed a new model that jointly learns multilingual multimodal representations using the image as a pivot between languages. We introduced new objective functions that can exploit similarities between images and descriptions across languages. We obtained state-of-the-art results on two tasks: image-description ranking and semantic textual similarity. Our results suggest that exploiting multilingual and multimodal resources can help in learning better semantic representations.

Acknowledgments

References

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. Association for Computational Linguistics.

Eneko Agirre, Carmen Baneab, Claire Cardiec, Daniel Cerd, Mona Diabe, Aitor Gonzalez-Agirrea, Weiwei Guof, Inigo Lopez-Gazpioa, Montse Maritxalara, Rada Mihalceab, 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, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of 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*, pages 385–393. Association for Computational Linguistics.

Galen Andrew, Raman Arora, Jeff A. Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 1247–1255.

Raffaella Bernardi, Ruket Cakici, 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. *Journal of Artifical Intelligence Research*, 55:409–442.

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

Sarath Chandar, Mitesh M. Khapra, Hugo Larochelle, and Balaraman Ravindran. 2016. Correlational neural networks. *Neural Computation*, 28(2):257–285.

David L Chen and William B 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-Volume 1*, pages 190–200. Association for Computational Linguistics.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634.

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, hosted by the 54th Annual Meeting of the Association for Computational Linguistics, VL@ACL 2016, August 12, Berlin, Germany*.

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482.

Orhan Firat, Baskaran Sankaran, Yaser Al-Onaizan, Fatos T. Yarman-Vural, and Kyunghyun Cho. 2016. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 268–277.

Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. 2013. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129.

Ruka Funaki and Hideki Nakayama. 2015. Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal*.

Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*, pages 529–545. Springer.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Andrej Karpathy, Armand Joulin, and Fei Fei F Li. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897.

Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2015. Unifying visual-semantic embeddings with multimodal neural language models. *Transactions of the Association for Computational Linguistics*.

Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016. Adding chinese captions to images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 271–275. ACM.

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 *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755.

Xiao Lin and Devi Parikh. 2016. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*, pages 261–277. Springer.

Lin Ma, Zhengdong Lu, Lifeng Shang, and Hang Li. 2015. Multimodal convolutional neural networks for matching image and sentence. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2623–2631.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-rnn). *International Conference on Learning Representations*.

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.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649.

Janarthanan Rajendran, Mitesh M. Khapra, Sarath Chandar, and Balaraman Ravindran. 2016. Bridge correlational neural networks for multilingual multimodal representation learning. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, 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.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of Association of Computational Linguistics*, 2:207–218.

Lucia Specia, Stella Frank, Khalil Simaan, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation, Berlin, Germany. Association for Computational Linguistics*.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. *International Conference on Learning Representations*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164.

John Wieting, Jonathan Mallinson, and Kevin Gimpel. 2017. Learning paraphrastic sentence embeddings from back-translated bitext. *arXiv preprint arXiv:1706.01847*.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. *Machine Translation*, 21(3):165–181.

Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3441–3450.

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: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.