A Strong and Robust Baseline for Text-Image Matching

Fangyu Liu
University of Waterloo
fangyu.liu@uwaterloo.ca

Rongtian Ye
Aalto University
rongtian7@gmail.com

Abstract

We review the current schemes of text-image matching models and propose improvements for both training and inference. First, we empirically show limitations of two popular loss (sum and max-margin loss) widely used in training text-image embeddings and propose a trade-off: a kNN-margin loss which 1) utilizes information from hard negatives and 2) is robust to noise as all K-most hardest samples are taken into account, tolerating *pseudo* negatives and outliers. Second, we advocate the use of Inverted Softmax (IS) and Cross-modal Local Scaling (CSLS) during inference to mitigate the so-called hubness problem in high-dimensional embedding space, enhancing scores of all metrics by a large margin.

1 Introduction

In recent years, deep eural models have gained a significant edge over *shallow*[1] models in cross-modal matching tasks. Text-image matching has been one of the most popular ones among them. Most methods involve two phases: 1) training: two neural networks (one image encoder and one text encoder) are learned end-to-end, mapping texts and images into a joint space, where vectors (either texts or images) with similar meanings are close to each other; 2) inference: for a query in modality A, after being encoded into a vector, a nearest neighbor search is performed to match the vector against all vector representations of items[2] in modality B. As the embedding space is learned through jointly modeling vision and language, it is often referred to as *Visual Semantic Embeddings* (VSE).

While the state-of-the-art architectures being consistently advanced (Nam et al., 2017; You et al., 2018; Wehrmann et al., 2018; Wu et al., 2019), few works have focused on the more fundamental problem of text-image matching - that is, the optimization objectives during training and inference. And that is what this paper focuses on. In the following of the paper, we will discuss 1) the optimization objective during training, i.e., loss function, and 2) the objective used in inference (how should a text-image correspondence graph be predicted).

Loss function. Faghri et al. (2018) brought the most notable improvement on loss function used for training VSE. They proposed a max-margin triplet ranking loss that emphasizes on the hardest negative sample within a min-batch. The max-margin loss has gained significant popularity and is used by a big set of recent works (Engilberge et al., 2018; Faghri et al., 2018; Lee et al., 2018; Wu et al., 2019). We, however, point out that the max-margin loss is very sensitive to label noise and encoder performance, and also easily overfits. Through experiments, we show that it only achieves the best performance under a careful selection of model architecture and dataset. Before Faghri et al. (2018), a pairwise ranking loss has been usually adopted for text-image model training. The only difference is that, instead of only using the hardest negative sample, it sums over all negative samples (we thus refer to it as the sum-margin loss). Though sum-margin loss yields stable and consistent performance under all dataset and architecture conditions, it does not make use information from hard samples but treats all samples equally by summing the margins up. Both Faghri et al. (2018) and our own experiments point to a clear trend that, more and cleaner data there is, the higher quality the encoders have, the better performance the max-margin loss has; while the smaller and less clean the data is, the less powerful the encoders are, the better sum-margin loss

[1] *shallow* means non-neural methods.

[2] In this paper, we refer to vectors used for searching as "queries" and vectors in the searched space as "items".

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 169–176
Florence, Italy, July 28 - August 2, 2019.

would perform (and max-margin would fail).

In this paper, we propose the use of a trade-off: a kNN-margin loss that sums over the k hardest sample within a mini-batch. It 1) makes sufficient use of hard samples and also 2) is robust across different model architectures and datasets. In experiments, the kNN-margin loss prevails in (almost) all data and model configurations.

Inference. During text-image matching inference, a nearest-neighbor search is usually performed to obtain a ranking for each of the queries. It has been pointed out by previous works (Radovanović et al., 2010; Dinu et al., 2015; Zhang et al., 2017) that *hubs* will emerge in such high-dimensional space and nearest neighbor search can be problematic for this need. Qualitatively, the hubness problem means a small portion of queries becoming "popular" nearest neighbor in the search space. Hubs harm model's performance as we already know that the predicted text-image correspondence should be a *bipartite matching*[3]. In experiments, we show that the hubness problem is the primary source of error for inference. Though has not attracted enough attention in text-image matching, hubness problem has been extensively studied in Bilingual Lexicon Induction (BLI) which aims to find a matching between two sets of bilingual word vectors. We thus propose to use similar tools during the inference phase of text-image matching. Specifically, we experiment with Inverted Softmax (IS) (Smith et al., 2017) and Cross-modal Local Scaling (CSLS) (Lample et al., 2018) to mitigate the hubness problem in text-image embeddings.

Contributions. The major contributions of this work are

- analyzing the shortcomings of sum and max-margin loss, proposing a kNN-margin loss as a trade-off (for training);
- proposing the use of Inverted Softmax and Cross-modal Local Scaling to replace naive nearest neighbor search (for inference).

2 Method

We first introduce the basic formulation of text-image matching model and sum/max-margin loss in 2.1. Then we propose our intended kNN-margin loss in Section 2.2 and the use of IS and CSLS for inference in Section 2.3.

2.1 Basic Formulation

The bidirectional text-image retrieval framework consists of a text encoder and an image encoder. The text encoder is composed of word embeddings, a GRU (Chung et al., 2014) or LSTM (Hochreiter and Schmidhuber, 1997) layer and a temporal pooling layer. The image encoder is a VGG19 (Simonyan and Zisserman, 2014) or ResNet152 (He et al., 2016) pre-trained on ImageNet (Deng et al., 2009) and a linear layer. We denote them as functions f and g which map text and image to two vectors of size d respectively.

For a text-image pair (t, i), the similarity of t and i is measured by cosine similarity of their normalized encodings:

$$s(i,t) = \left\langle \frac{f(t)}{\|f(t)\|_2}, \frac{g(i)}{\|g(i)\|_2} \right\rangle : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}. \tag{1}$$

During training, a margin based triplet ranking loss is adopted to cluster positive pairs and push negative pairs away from each other. We list the both the sum-margin loss used in Frome et al. (2013); Kiros et al. (2015); Nam et al. (2017); You et al. (2018); Wehrmann et al. (2018):

$$\begin{aligned} \min_\theta \sum_{i \in I} \sum_{\bar{t} \in T \setminus \{t\}} [\alpha - s(i,t) + s(i,\bar{t})]_+ \\ + \sum_{t \in T} \sum_{\bar{i} \in I \setminus \{i\}} [\alpha - s(t,i) + s(t,\bar{i})]_+; \end{aligned} \tag{2}$$

and the max-margin loss used by Engilberge et al. (2018); Faghri et al. (2018); Lee et al. (2018); Wu et al. (2019):

$$\begin{aligned} \min_\theta \sum_{i \in I} \max_{\bar{t} \in T \setminus \{t\}} [\alpha - s(i,t) + s(i,\bar{t})]_+ \\ + \sum_{t \in T} \max_{\bar{i} \in I \setminus \{i\}} [\alpha - s(t,i) + s(t,\bar{i})]_+, \end{aligned} \tag{3}$$

where $[\cdot]_+ = \max(0, \cdot)$; α is a preset margin (we use $\alpha = 0.2$); T and I are all text and image encodings in a mini-batch; t is the descriptive text for image i and vice versa; $\bar{t}$ denotes non-descriptive texts for i while $\bar{i}$ denotes non-descriptive images for t.

2.2 kNN-margin Loss

We propose a simple yet robust strategy for selecting negative samples: instead of counting all

[3] In Graph Theory, a set of edges is said to be a **matching** if none of the edges share a common endpoint.

(Eq. 2) or hardest (Eq. 3) sample in a mini-batch, we take the k-hardest samples. We first define a function $\texttt{kNN}(x, M, k)$ to return the k closest points in point set M to x. Then the kNN-margin loss is formulated as:

$$\begin{aligned}\min_{\theta} &\sum_{i\in I}\sum_{\bar{t}\in K_1}[\alpha - s(i,t) + s(i,\bar{t}))]_+ \\ &+\sum_{t\in T}\sum_{\bar{i}\in K_2}[\alpha - s(t,i) + s(t,\bar{i}))]_+\end{aligned} \tag{4}$$

where

$$K_1 = \texttt{kNN}(i, T\backslash\{t\}, k), K_2 = \texttt{kNN}(t, I\backslash\{i\}, k).$$

In max-margin loss, when the hardest sample is misleading or incorrectly labeled, the wrong gradient would be imposed on the network. We call it a *pseudo* hard negative. In kNN-margin loss, though some *pseudo* hard negatives might still generate false gradients, they are likely to be canceled out by the negative samples with correct information. As only the k hardest negatives are considered, the selected samples are still hard enough to provide meaningful supervision to the model. In experiments, we show that kNN-margin loss indeed demonstrates such characteristics.

2.3 Hubness Problem During Inference

The standard procedure for inference is performing a naive nearest neighbor search. This, however, leads to the hubness problem which is the primary source or error as we will show in Section 3.5. We thus leverage the prior that "one query should not be the nearest neighbor for multiple items" to improve the text-image matching. Specifically, we use two tools introduced in BLI: Inverted Softmax (IS) (Smith et al., 2017) and Cross-modal Local Scaling (CSLS) (Lample et al., 2018).

2.3.1 Inverted Softmax (IS)

The main idea of IS is to estimate the confidence of a prediction $i \rightarrow t$ not merely by similarity score $s(i,t)$, but the score reweighted by t's similarity with other queries:

$$s'(i,t) = \frac{e^{\beta s(i,t)}}{\sum_{\bar{i}\in I\backslash\{i\}} e^{\beta s(\bar{i},t)}} \tag{5}$$

where β is a temperature (we use $\beta = 30$). Intuitively, it scales down the similarity if t is also very close to other queries.

2.3.2 Cross-modal Local Scaling (CSLS)

CSLS aims to decrease a query vector's similarity to item vectors lying in *dense* areas while increase similarity to *isolated*[4] item vectors. It punishes the occurrences of an item being the nearest neighbor to multiple queries. Specifically, we update the similarity scores with the formula:

$$\begin{aligned}s'(i,t) = 2s(i,t) &- \frac{1}{k}\sum_{i_t\in K_1} s(i_t,t) \\ &- \frac{1}{k}\sum_{t_i\in K_2} s(i,t_i)\end{aligned} \tag{6}$$

where $K_1 = \texttt{kNN}(t, I, k)$ and $K_2 = \texttt{kNN}(i, T, k)$ (we use $k = 10$).

3 Experiments

In this section we introduce our experimental setups (Section 3.1, 3.2, 3.3) and quantitative results (Section 3.4, 3.5).

3.1 Dataset

dataset	# train	# validation	# test
Flickr30k	30,000	1,000	1,000
MS-COCO 1k	113,287	5,000	1,000
MS-COCO 5k	113,287	5,000	5,000

Table 3: Train-validation-test splits of used datasets.

We use Flickr30k (Young et al., 2014) and MS-COCO (Lin et al., 2014) as our experimental datasets. We list their splitting protocols in Table 3. For MS-COCO, there has been several different splits used by the research community. In convenience of comparing to a wide range of results reported by other works, we use two protocols and they are referred as MS-COCO 1k and 5k where 1k and 5k differs only in the test set used (1k's test set is a subset of 5k's). Notice that MS-COCO 5k computes the average of 5 folds of 1k images. Also, in both Flickr30k and MS-COCO, 1 image has 5 captions - so 5 (text,image) pairs are used for every image.

3.2 Evaluation Metrics

We use R@Ks (recall at K), Med r and Mean r to evaluate the results:

[4] *Dense* and *isolated* are in terms of query.

#	architecture	loss	image→text					text→image				
			R@1	R@5	R@10	Med r	Mean r	R@1	R@5	R@10	Med r	Mean r
1.1	GRU+VGG19	sum-margin	30.2	58.7	70.4	4.0	33.0	22.9	50.6	61.4	**5.0**	49.5
1.2		max-margin	30.7	58.7	69.6	4.0	30.3	22.4	48.4	59.3	6.0	39.0
1.3		kNN-margin ($k = 3$)	**34.1**	**61.7**	69.9	**3.0**	**24.7**	**25.1**	**52.5**	64.6	**5.0**	34.3
1.4		kNN-margin ($k = 5$)	33.4	61.6	**71.1**	**3.0**	26.7	24.2	51.8	**64.8**	**5.0**	**32.7**
1.5		kNN-margin ($k = 10$)	33.3	59.4	69.4	**3.0**	28.4	23.4	50.6	63.5	**5.0**	33.8

Table 1: Quantitative results on Flickr30k (Young et al., 2014).

#	architecture	loss	image→text					text→image				
			R@1	R@5	R@10	Med r	Mean r	R@1	R@5	R@10	Med r	Mean r
2.1	GRU+VGG19	sum-margin	48.9	79.9	89.0	1.8	5.6	38.3	73.5	85.3	**2.0**	**8.4**
2.2		max-margin	**51.8**	**81.1**	90.5	**1.0**	**5.5**	**39.0**	73.9	84.7	**2.0**	12.0
2.3		kNN-margin	50.6	**81.1**	**90.6**	1.4	**5.5**	38.7	**74.0**	**85.5**	**2.0**	11.8
2.4	GRU+ResNet152	sum-margin	53.2	85.0	93.0	**1.0**	3.9	41.9	77.2	88.0	**2.0**	8.7
2.5		max-margin	**58.7**	**88.2**	94.0	**1.0**	**3.2**	**45.0**	78.9	88.6	**2.0**	8.6
2.6		kNN-margin	57.8	87.6	**94.4**	**1.0**	3.4	43.9	**79.0**	**88.8**	**2.0**	**8.1**

Table 2: Quantitative results on MS-COCO (Lin et al., 2014). Using the 5k test set.

- R@K: the ratio of "# of queries that the ground-truth item is ranked in top K" to "total # of queries" (we use $K = \{1, 5, 10\}$; the higher the better);
- Med r: the median of the ground-truth ranking (the lower the better);
- Mean r: the mean of the ground-truth ranking (the lower the better).

We compute all metrics for both text→image retrieval and image→text matching. We follow the convention of taking the model with maximum R@Ks sum (both text→image and image→text) on the validation set as the best model for testing.

3.3 Hyperparameters

Training. For max-margin models, we follow the configuration specified in Faghri et al. (2018). For all other models, we start with a learning rate of 0.001 and decay it by 10 times after every 10 epochs. We train all models for 30 epochs with a batch size of 128. All models are optimized using an Adam optimizer (Kingma and Ba, 2015).

Model. We use 300-d word embeddings and 1024 internal states for GRU text encoders (all randomly initialized with Xavier init. (Glorot and Bengio, 2010); $d = 1024$ for both text and image embeddings. All image encoders are fixed (with no finetuning) for fair comparison.

3.4 Loss Function Performance

Table 1 and 2 show quantitative results on Flickr30k and MS-COCO respectively.

Flickr30k. kNN-margin loss achieves significantly better performance on all metrics than all other loss. It is worth noticing that max-margin loss fails on this dataset (even much worse than sum-margin). kNN-margin loss with $k = \{3, 5\}$ get the highest scores. We use $k = 3$ for the following experiments unless explicitly specified.

MS-COCO. Max-margin loss performs much better on MS-COCO, especially on R@1 - it has the best R@1 across both configurations. kNN-margin is comparable to max-margin. Specifically, it produces slightly worse R@1s, almost identical R@5s, and slightly better R@10s. Sum-margin, however, performs poorly on MS-COCO. It is worth noting that here we are using the 5k test set, which is a superset of the widely adopted 1k test set. We will compare with quantitative results reported on the 1k test set in the next section.

3.5 Hubs during Inference

To show hubness is indeed a major source of error, we select one of the text-image embeddings to do statistics. We use the model on Table 2 line 2.1 to generate embeddings on MS-COCO's test set. Among the $25,000$ (query, item) pairs, only $1,027$ (4.1%) items are the nearest neighbor (NN) of solely 1 query; there are, however, $19,805$ (79.2%) items that are NN to 0 query and $3,007$ (12.0%) items that are NN to ≥ 5 queries, indicating wide existence of hubs. Moreover, the most "popular" item is NN to 51 queries. We know that one item ought to be NN to only one query

#	dataset	model	inference	image→text					text→image				
				R@1	R@5	R@10	Med r	Mean r	R@1	R@5	R@10	Med r	Mean r
3.1	Flickr30k	GRU+VGG19 kNN-margin	naive	34.1	61.7	69.9	**3.0**	24.7	25.1	52.5	64.6	5.0	34.3
3.2			IS	**36.0**	**64.5**	**72.9**	**3.0**	**20.1**	25.2	52.6	64.4	5.0	31.1
3.3			CSLS	**36.0**	64.4	72.5	**3.0**	20.3	**26.7**	**54.3**	**65.7**	**4.0**	**30.8**
3.4	MS-COCO 5k	GRU+ResNet152 kNN-margin	naive	57.8	87.6	94.4	**1.0**	3.4	43.9	79.0	88.8	**2.0**	8.1
3.5			IS	**64.2**	**89.4**	95.0	**1.0**	3.2	46.7	80.1	89.3	**2.0**	7.8
3.6			CSLS	62.4	89.3	**95.4**	**1.0**	**3.0**	**47.2**	**80.7**	**89.9**	**2.0**	**7.7**
3.7	MS-COCO 1k	(Kiros et al., 2015) (ours[5])		49.9	79.4	90.1	2.0	5.2	37.3	74.3	85.9	**2.0**	10.8
3.8		(Vendrov et al., 2016)		46.7	-	88.9	2.0	5.7	37.9	-	85.9	**2.0**	8.1
3.9		(Huang et al., 2017)		53.2	83.1	91.5	**1.0**	-	40.7	75.8	87.4	**2.0**	-
3.10		(Liu et al., 2017)		56.4	85.3	91.5	-	-	43.9	78.1	88.6	-	-
3.11		(You et al., 2018)		56.3	84.4	92.2	**1.0**	-	45.7	81.2	90.6	**2.0**	-
3.12		(Faghri et al., 2018)		58.3	86.1	93.3	**1.0**	-	43.6	77.6	87.8	**2.0**	
3.13		(Faghri et al., 2018) (ours)		60.5	89.6	94.9	**1.0**	3.1	46.1	79.5	88.7	**2.0**	8.5
3.14		(Wu et al., 2019)		64.3	89.2	94.8	**1.0**	-	48.3	81.7	**91.2**	**2.0**	-
3.15		GRU+ResNet152 kNN-margin	naive	58.3	89.2	95.4	**1.0**	3.1	45.0	80.4	89.6	**2.0**	7.2
3.16			IS	**66.4**	91.8	96.1	**1.0**	2.7	48.6	81.5	90.3	**2.0**	7.3
3.17			CSLS	65.4	**91.9**	**97.1**	**1.0**	**2.5**	**49.6**	**82.7**	**91.2**	**2.0**	**6.5**

Table 4: Quantitative results of different inference methods across different datasets and models. Line 3.1-3.3 are using the model from Table 1 line 1.3 and line 3.4-3.6, 3.15-3.17 are using the model from Table 2 line 2.9. Line 3.7-3.14 are results reported by previous works which all adopted naive nearest neighbor search for inference.

in the ground-truth query-item matching. So, we can spot errors even before ground-truth labels are revealed - for instance, the most "popular" item with 51 NNs must be the *false* NN for at least 50 queries. Table 5 shows the brief statistics.

	$k=0$	$k=1$	$k\geq 2$	$k\geq 5$	$k\geq 10$
#	19,805	1,026	4,169	3,007	500
percentage	79.2%	4.1%	16.7%	12.0%	2.0%

Table 5: Statistics of # items being NN to k queries in the embeddings of Table 2, line 2.1, text→image. There are in total 25,000 (text,image) paris in this embedding.

Both IS and CSLS demonstrate compelling empirical performance in mitigating the hubness problem. Table 4 shows the quantitative results. R@Ks and also Med r, Mean r are improved by a large margin with both methods. In most configurations, CSLS is slightly better than IS on improving text→image inference while IS is better at image→text. The best results (line 3.8, 3.9) are even better than the recently reported state-of-the-art (Wu et al., 2019) (Table 4 line 3.14), which performs a naive nearest neighbor search. This suggests that the hubness problem deserves much more attention and careful selection of inference methods is vital for text-image matching.

[5]"ours" means our implementation.

4 Limitations and Future Work

This paper brings up a baseline with excellent empirical performance. We plan to contribute more theoretical and technical novelty in follow up works for both the training and inference phase of text-image matching models.

Loss function. Though the kNN-margin loss has superior empirical performance, it is leveraging the prior knowledge we hardcoded in it - it relies on a suitable k to maximize its power. Flickr30k and MS-COCO are relatively clean and high-quality datasets while the real world data is usually not. With the kNN-margin loss being a strong baseline, we plan to bring a certain form of self-adaptiveness into the loss function to help it automatically decide what to learn based on the distribution of data points.

Also, to further validate the robustness of loss functions, we plan to experiment models on more *noisy* data. The reason for max-margin's failure on Flikr30k is more likely that the training set is too small - so the model easily overfits. However, the dataset (Flikr30k) itself is rather clean and accurate. It makes more sense to experiment with a noisy dataset with *weak* text-image correspondence or even false labels. We have two types of candidates for this need: 1) academic datasets that contain "foil" (Shekhar et al., 2017) or adversarial samples (Shi et al., 2018); 2) a real-world text-image dataset such as a news article-image

dataset (Elliott and Kleppe, 2016; Biten et al., 2019).

Inference. Both IS and CSLS are *soft* criteria. If we do have the strong prior that the final text-image correspondence is a bipartite matching, we might as well make use of that information and impose a *hard* constraint on it. The task of text-image matching, after all, is also a form of assignment problem in Combinatorial Optimization (CO). We thus plan to investigate tools from the CO literature such as the Hungarian Algorithm (Kuhn, 1955), which is the best-known algorithm for producing a maximum weight bipartite matching; the Murty's Algorithm (Murty, 1968), which generalizes the Hungarian Algorithm into producing the K-best matching - so that rankings are available for computing R@K scores.

5 Related Work

In this section, we introduce works from two fields which are highly-related to our work: 1) text-image matching and VSE; 2) Bilingual Lexicon Induction (BLI) in the context of cross-modal matching.

5.1 Text-image Matching

Since the dawn of deep learning, works have emerged using a two-branch structure to connect both language and vision. Frome et al. (2013) brought up the idea of VSE, which is to embed pairs of (text, image) data and compare them in a joint space. Later works extended VSE for the task of text-image matching (Hodosh et al., 2013; Kiros et al., 2015; Gong et al., 2014; Vendrov et al., 2016; Hubert Tsai et al., 2017; Faghri et al., 2018; Wang et al., 2019), which is also our task of interest. It is worth noting that there are other lines of works which also jointly model language and vision. The closest one might be image captioning (Lebret et al., 2015; Karpathy and Fei-Fei, 2015). But image captioning aims to generate novel captions while text-image matching retrieves existing descriptive texts or images in a database.

5.2 Bilingual Lexicon Induction (BLI)

We specifically talk about BLI as the tools we used to improve inference performance come from this literature. BLI is the task of inducing word translations from monolingual corpora in two languages (Irvine and Callison-Burch, 2017). Words are usually represented by vectors trained from Distributional Semantics, eg. Mikolov et al. (2013). So, the word translation problem converts to finding the appropriate matching among two sets of vectors which makes it similar to our task of interest. Smith et al. (2017); Lample et al. (2018) proposed to first conduct a direct Procrustes Analysis (Schönemann, 1966) between two sets of vectors, then use criteria that heavily punish hubs during inference to avoid the hubness problem. We experimented with both methods in our task.

6 Conclusion

We discuss the pros and cons of prevalent loss functions used in text-image matching and propose a kNN-margin loss as a trade-off which yields strong and robust performance across different model architectures and datasets. Instead of using naive nearest neighbor search, we advocate to adopt more polished inference strategies such as Inverted Softmax (IS) and Cross-modal Local Scaling (CSLS), which can significantly improve scores of all metrics.

We also analyze the limitations of this work and indicate the next step for improving both the loss function and the inference method.

7 Acknowledgement

We thank the reviewers for their careful and insightful comments. We thank our family members who have both spiritually and financially supported our independent research. The author Fangyu Liu thanks Rémi Lebret for introducing him this interesting problem and many of the related works.

References

Ali Furkan Biten, Lluis Gomez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good news, everyone! context driven entity-aware captioning for news images. *CVPR*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 248–255. IEEE.

Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. Improving zero-shot learning by mitigating the hubness problem. *ICLR worshop*.

Desmond Elliott and Martijn Kleppe. 2016. 1 million captioned dutch newspaper images.

Martin Engilberge, Louis Chevallier, Patrick Pérez, and Matthieu Cord. 2018. Finding beans in burgers: Deep semantic-visual embedding with localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3984–3993.

F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler. 2018. Vse++: Improving visual-semantic embeddings with hard negatives.

A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

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 (ECCV)*, pages 529–545. Springer.

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.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

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.

Yan Huang, Wei Wang, and Liang Wang. 2017. Instance-aware image and sentence matching with selective multimodal lstm. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2310–2318.

Yao-Hung Hubert Tsai, Liang-Kang Huang, and Ruslan Salakhutdinov. 2017. Learning robust visual-semantic embeddings. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3571–3580.

Ann Irvine and Chris Callison-Burch. 2017. A comprehensive analysis of bilingual lexicon induction. *Computational Linguistics*, 43(2):273–310.

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. 2015. Adam: A method for stochastic optimization. *ICLR*.

R. Kiros, R. Salakhutdinov, and R. S. Zemel. 2015. Unifying visual-semantic embeddings with multimodal neural language models. *Transactions of the Association for Computational Linguistics (TACL)*.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Herv Jgou. 2018. Word translation without parallel data. In *International Conference on Learning Representations*.

Rémi Lebret, Pedro O Pinheiro, and Ronan Collobert. 2015. Phrase-based image captioning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning-Volume 37 (ICML)*, pages 2085–2094. JMLR. org.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216.

T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and L. Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision (ECCV)*, pages 740–755. Springer.

Yu Liu, Yanming Guo, Erwin M Bakker, and Michael S Lew. 2017. Learning a recurrent residual fusion network for multimodal matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4107–4116.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Katta G Murty. 1968. Letter to the editoran algorithm for ranking all the assignments in order of increasing cost. *Operations research*, 16(3):682–687.

Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 299–307.

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. Hubs in space: Popular nearest neighbors in high-dimensional data. *Journal of Machine Learning Research*, 11(Sep):2487–2531.

Peter H Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurelie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. "foil it! find one mismatch between image and language caption". In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 1: Long Papers)*, pages 255–265.

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*, pages 3715–3727.

K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. 2017. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *ICLR*.

I. Vendrov, R. Kiros, S. Fidler, and R/ Urtasun. 2016. Order-embeddings of images and language. *ICLR*.

Liwei Wang, Yin Li, Jing Huang, and Svetlana Lazebnik. 2019. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):394–407.

Jônatas Wehrmann et al. 2018. Bidirectional retrieval made simple. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7718–7726.

Hao Wu, Jiayuan Mao, Yufeng Zhang, Yuning Jiang, Lei Li, Weiwei Sun, and Wei-Ying Ma. 2019. Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Quanzeng You, Zhengyou Zhang, and Jiebo Luo. 2018. End-to-end convolutional semantic embeddings. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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.

Li Zhang, Tao Xiang, and Shaogang Gong. 2017. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2021–2030.