Harvesting Deep Models for Cross-Lingual Image Annotation

Qijie Wei, Xiaoxu Wang, Xirong Li*
Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China
Multimedia Computing Lab, School of Information, Renmin University of China

ABSTRACT

This paper considers cross-lingual image annotation, harvesting deep visual models from one language to annotate images with labels from another language. This task cannot be accomplished by machine translation, as labels can be ambiguous and a translated vocabulary leaves us limited freedom to annotate images with appropriate labels. Given non-overlapping vocabularies between two languages, we formulate cross-lingual image annotation as a zero-shot learning problem. For cross-lingual label matching, we adapt zero-shot by replacing the current monolingual semantic embedding space by a bilingual alternative. In order to reduce both label ambiguity and redundancy we propose a simple yet effective approach called label-enhanced zero-shot learning. Using three state-of-the-art deep visual models, i.e., ResNet-152, GoogleNet-Shuffle and OpenImages, experiments on the test set of Flickr8k-CN demonstrate the viability of the proposed approach for cross-lingual image annotation.

KEYWORDS

Cross-lingual image annotation, English-Chinese, zero-shot learning

ACM Reference format:
Qijie Wei, Xiaoxu Wang, Xirong Li. 2017. Harvesting Deep Models for Cross-Lingual Image Annotation. In *Proceedings of CBMI, Florence, Italy, June 19-21, 2017,* 5 pages.
DOI: 10.1145/3095713.3095751

1 INTRODUCTION

By assigning relevant labels to visual content, image auto-annotation enables a semantic level access to the increasing amounts of unlabeled images and videos [9]. Different from existing works that focus on image annotation in a monolingual setting, mostly English, this paper studies *cross-lingual* image annotation. The topic is interesting because image training examples associated with non-English labels are in short supply in the public literature, making it difficult to directly train models for another language. Koochali *et al.* [5] report that the majority of user provided tags on Flickr are in English. As a showcase, we consider Chinese as a target language, studying how to annotate images with Chinese labels by harvesting deep visual models originally trained for predicting English labels.

*Corresponding author (xirong@ruc.edu.cn).

CBMI, Florence, Italy

978-1-4503-5333-5/17/06…$15.00
DOI: 10.1145/3095713.3095751

There are some initial efforts on describing images in a bilingual setting [7, 13, 15]. Given candidate annotations predicted by an English model and a Chinese model respectively, Xue *et al.* [15] model the two sources of annotations by an n-partite graph, refining the annotations by reinforcement learning on the graph. More recently, Li *et al.* [7] and Miyazaki *et al.* [13] investigate the possibility of training neural image captioning models for generating Chinese and Japanese captions, respectively. While their target language differs, they take a similar approach by first extending an existing image sentence corpus to a bilingual version, using crowd sourcing to gather captions written in the target language. In contrast to [7, 13, 15], we assume zero availability of labeled examples nor image annotation model in the target language.

One might consider word-by-word (machine) translation as a solution for cross-lingual image annotation. However, words can be ambiguous, *e.g.,* 'shutter' may refer to 'camera shutter' or 'window shutter'. Moreover, the translated vocabulary is subject to the predefined vocabulary of the source model, giving us limited freedom to annotate images in the target language.

As English words are comprised of alphabets while Chinese words are written using Chinese characters, no overlap exists between Chinese and English labels. As such, zero-shot learning, which aims for predicting novel labels [8, 14], seems to be well fit. Existing models for zero-shot learning are developed in a monolingual setting. For the cross-lingual setting, we need to adapt the zero-shot models, replacing their monolingual word embedding spaces by a bilingual alternative. Consequently, Chinese labels closest to a given image are chosen as the final annotation. As illustrated in Fig. 1, this solution tends to generate redundant labels such as 百叶窗 (window shutter) and 窗户 (window).

We propose in this paper *label-enhanced zero-shot* that reduces both label ambiguity and redundancy. Next, we detail the proposed approach.

2 OUR APPROACH

2.1 Problem Statement

Cross-lingual image annotation is to annotate a given image with labels from a target language by reusing image annotation models trained for a source language. As aforementioned, a typical source language is *English*, due to the large availability of training images associated with English labels, *e.g.,* ImageNet [2] and YFCC100m [5]. In this work we consider *Chinese* as the showcase of the target language.

To make our description more formal, we introduce some notation. Let $\mathcal{Y}_s$ be a vocabulary in the source language, and $\mathcal{Y}_t$ be a vocabulary in the target language. Let x be a given image. We have access to an image annotation model that for each label $w_s \in \mathcal{Y}_s$ there is $p(w_s|x)$ predicting the probability of w_s being relevant w.r.t. the given image. By definition, there is no overlap between the

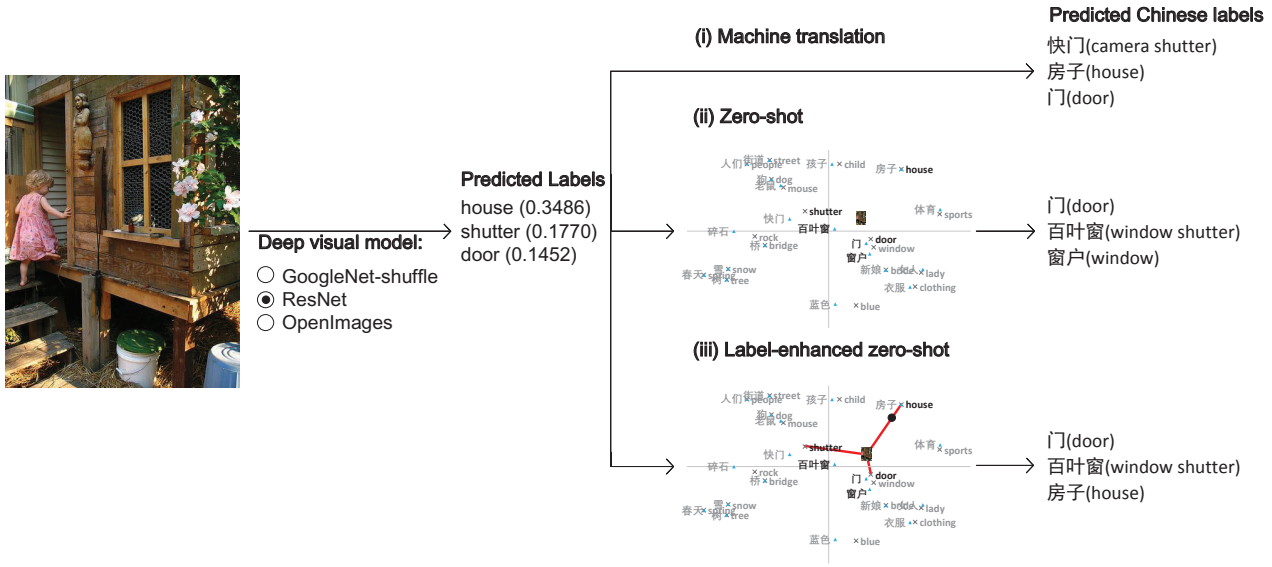


Figure 1: A conceptual illustration of three approaches to cross-lingual image annotation. Compared with (i) Machine translation which incorrectly translates the ambiguous label 'shutter' as 快门 (camera shutter) and (ii) the zero-shot approach [8, 14] that picks up Chinese labels that are nearest to the test image yet redundant, such as 百叶窗 (window shutter) and 窗户 (window), the proposed *label-enhanced* zero-shot approach resolves the ambiguity and predicts Chinese labels that are relevant and more diverse.

two vocabularies, *i.e.,* $\mathcal{Y}_s \cap \mathcal{Y}_t = \emptyset$. This means we cannot directly employ $p(w_s|x)$ to annotate the image with labels from $\mathcal{Y}_t$. So the goal of this paper is to approximately compute $p(w_t|x)$ for each $w_t \in \mathcal{Y}_t$ by exploiting the existing image annotation model $p(w_s|x)$ in a cross-lingual setting.

The fact that $\mathcal{Y}_s \cap \mathcal{Y}_t = \emptyset$ allows us to formulate cross-lingual image annotation as the zero-shot learning problem. However, previous approaches to zero-shot learning [8, 14] work in a monolingual label embedding space, making them not directly applicable. To overcome this obstacle, we describe in Section 2.2 how to construct a bilingual word2vec space wherein the semantic similarity between w_s and w_t can be computed as the cosine similarity between the corresponding embedding vectors. Later in Section 2.3 we describe the proposed label-enhanced zero-shot approach that performs image annotation in the bilingual space.

2.2 Bilingual Label Embedding

For bilingual label embedding, we adopt the BiSkip model recently developed by Thang *et al.* [10]. It extends the skip-gram model [12] to enable learning from a bilingual corpus. Let l_s and l_t be paired sentences in a bilingual corpus. Given the aligned words w_s in l_s and w_t in l_t, BiSkip uses w_s to predict its nearby words in l_s and the nearby words of w_t in l_t and vice versa. So in essence BiSkip learns simultaneously two monolingual skip-gram models, *i.e.,* l_s to l_s and l_t to l_t, and two cross-lingual skip-gram models, *i.e.,* l_s to l_t and l_t to l_s.

Note that the BiSkip model from [10] was trained on an English-German corpus. To the best of our knowledge, no corpus of paired English-Chinese sentences is publicly available. Luckily, there are several English-learning websites that provides many examples in both English and Chinese. We implement a web crawler to collect such examples, obtaining 170k pairs of bilingual sentences in total. Table 1 shows examples of our bilingual corpus.

Table 1: Pairs of bilingual sentences, randomly chosen from our training corpus.

English sentence	Chinese Sentence
They diverted the plane to another airport because of the weather.	由于天气原因,他们把飞机导航到另一机场。
There is a garage in the southeast corner of his house.	他房子的东南角有一个车房。
The pollution has already turned vast areas into a wasteland.	污染已经使大片地区沦为不毛之地。

2.3 Label-enhanced Zero-shot Learning

Our solution is built on the top of the word2vec based zero-shot models [8, 14]. To make the paper self contained, we describe in brief the two models in the new cross-lingual context. Let $\phi(w)$ be the embedding vector of word $w \in \mathcal{Y}_s \cup \mathcal{Y}_t$ in the bilingual space. Given an image x, we use $w_s(x,k)$ to indicate the k-th most likely label predicted by $p(w_s|x)$. The ConSE model [14] embeds the image as a convex combination of the embedding vectors of the

predicted labels, *i.e.,*

$$\phi(x) = \frac{1}{Z}\sum_{k=1}^{K} p(w_s(x,k)|x) \cdot \phi(w_s(x,k)), \quad (1)$$

where K is the number of labels used, and Z is the normalization factor that sums over $p(w_s(x,k)|x)$ with $k = 1, \ldots, K$.

HierSE [8] improves ConSE by exploiting the WordNet hierarchy to resolve label ambiguity, and thus generates better label and image embeddings. Given a specific label w_s, let $super(w_s)$ be all its ancestors defined by WordNet. The HierSE version of $\phi(w_s)$ is computed as

$$\phi_{hi}(w_s) = \frac{1}{Z'}\sum_{w' \in \{w_s\} \cup super(w_s)} w(w'_s|w_s) \cdot \phi(w'_s), \quad (2)$$

where $w(w'_s|w_s)$ is a weight subject to exponential delay with respect to the minimal path length from w_s to w'_s, and Z' is the normalization factor summing over $w(w'_s|w_s)$. Consider again the label 'shutter', which corresponds to two WordNet nodes. One node (WordNet-id n04211528) means camera shutter, while the other node (WordNet-id n04211356) refers to window shutter. Their $\phi(w_s)$ is identical. As their $super(w_s)$ differs, the two nodes will have distinct $\phi_{hi}(w_s)$. The HierSE version of the image embedding vector is obtained by substituting $\phi_{hi}(w_s)$ for $\phi(w_s)$ in Eq. 1.

Now, with the image mapped into the bilingual space, the probability of a target label w_t being relevant w.r.t. the image, *i.e.,* $p(w_t|x)$, can be estimated by computing the cosine similarity between $\phi(w_t)$ and $\phi_{hi}(x)$. However, because labels of similar meanings are forced to stay close, simply finding labels nearest to the image tends to generate redundant annotations, *e.g.,* 百叶窗 (window shutter) and 窗户 (window). To overcome this drawback, we propose to adjust the image embedding vector according to each of the top K predicted source labels. In particular, we define the adjustment w.r.t. the k-th label as

$$\phi_k(x) \leftarrow \theta \cdot \phi(x) + (1-\theta) \cdot \phi(w_s(x,k)), \quad (3)$$

where θ is a weight ranging from 0 to 1. Consider two special cases where θ is set to 1 and 0, respectively. The former is the previous zero-shot learning model, while the latter uses the embedding vector of each predicted source label to represent the image. The weight θ strikes a balance between the two cases. Note that when the source vocabulary $\mathcal{Y}_s$ is not derived from WordNet and consequently HierSE is not applicable, adding the image embedding vector to the label embedding vector helps label disambiguation. We term the proposed method label-enhanced zero-shot learning.

After obtaining the K label-enhanced image embedding vectors $\{\phi_1(x), \ldots, \phi_K(x)\}$, we select target labels in a sequential manner. At the k-th step, we select a novel w_t that maximizes the cosine similarity between $\phi(w_t)$ and $\phi_k(x)$. As such, we annotate the given image with relevant and more diverse labels.

3 EVALUATION

3.1 Experimental Setup

Deep visual models. To generate English labels, we experiment with three state-of-the-art models that have been pretrained and publicly accessible, *i.e.,* ResNet-152 [3], GoogleNet-Shuffle [11], and OpenImages [6]. The vocabularies of ResNet-152 and GoogeNet-Shuffle contain 1k and 13k labels, respectively, all taken from ImageNet [2]. By contrast, the vocabulary of OpenImages consits of 6k labels, which are middle-level concepts sampled from the Google knowledge graph. For details of the individual models we refer interested readers to the original papers. We try both ConSE and HierSE semantic embeddings on ResNet-152 and GoogleNet-Shuffle, while ConSE only on OpenImages as no hierarcy is provided for this model.

Bilingual space. We use the MultiVec toolkit [1] to train the BiSkip model. As a Chinese sentence does not contain markers as word boundaries, it needs to be segmented into a sequence of meaningful words. We utilize boson[1] for Chinese text segmentation. The learned bilingual space contains 26,255 English words and 27,853 Chinese words. For ResNet-152, GoogleNet-Shuffle and OpenImages, the number of labels that can be embedded into the space is 834, 9,808 and 4,399 respectively.

Test set. We use the test set of Flickr8k-CN [7] which contains 1000 test images. Each image is originally associated with five English sentences from Flickr8k [4], which have been manually translated into five Chinese sentences. We extract ground-truth labels from these Chinese sentences as follows. We again employ boson for Chinese text segmentation and part-of-speech tagging. For each image, noun (*e.g.,* 自行车), verb (*e.g.,* 爬), adjective (*e.g.,* 荒芜), and locality (*e.g.,* 街上) words that appear in at least two of the five sentences are preserved. This results in 1,012 distinct Chinese words in total, among which 935 can be mapped to the bilingual space. The number of ground truth words per image ranges from 1 to 13, with an average value of 5.8.

Baseline methods. Machine translation is a natural baseline. Following [7] we use Baidu translation. As our approach is developed on the basis of ConSE and HierSE, they need to be compared.

Performance metrics. We report $hit@n$, the percentage of test images that have at least one correct label covered by the top n predict labels, $n \in \{1, 5, 10\}$. As the sum of $hit@1$, $hit@5$ and $hit@10$ reflects ranking quality of relevant labels, we use this value to measure the overall performance.

3.2 Experiments

Through the experiments we aim to understand the influence of the three major factors on the proposed approach. The factors are deep visual models (ResNet-152, GoogleNet-Shuffle, or OpenImages), zero-shot models (ConSE or HierSE), and the weight θ.

Properties of the proposed approach. Fig. 2 shows the performance curves of the proposed approach, given varied θ and specific deep visual models. With the increase of θ, the performance increases first and decreases later. Notice that the left end ($\theta = 0$) corresponds to a special case of the proposed approach, where for each predicted English label, its nearest Chinese label in the bilingual space is selected. Because θ is set to be 0, the distance between the Chinese label and the image is not considered. The selected Chinese label might be irrelevant w.r.t. the image if the English label is ambiguous. Taking the image into account by increasing the value of θ helps semantic disambiguation. Also notice that the performance

[1] http://bosonnlp.com/

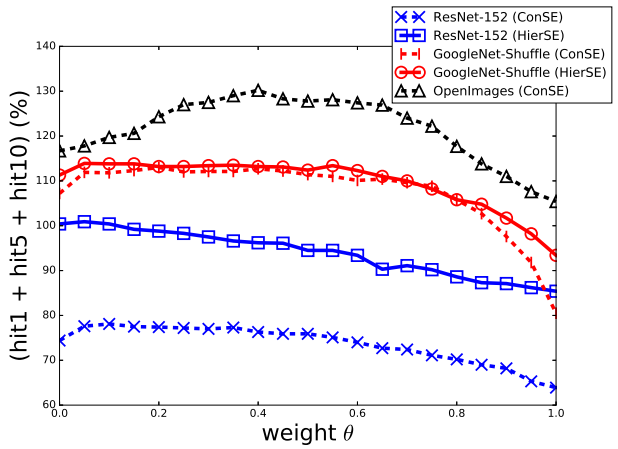


Figure 2: The influence of the weight θ on the proposed method, given three different deep models.

Table 2: Comparing the three approaches to cross-lingual image annotation. The proposed label-enhanced zero-shot learning performs the best.

Deep visual model	Cross-language approaches	hit@k(%)			sum(%)
		1	5	10	
ResNet-152	Machine translation	6.4	11.8	15.6	33.8
	Zero-shot	17.2	30.8	37.4	85.4
	Proposed approach	**19.6**	**37.9**	**43.4**	**100.9**
GoogleNet-Shuffle	Machine translation	6.7	17.6	26.5	50.8
	Zero-shot	**19.0**	34.5	39.9	93.4
	Proposed approach	17.8	**43.3**	**52.8**	**113.9**
OpenImages	Machine translation	9.4	44.6	59.8	113.8
	Zero-shot	**20.8**	38.0	44.6	105.4
	Proposed approach	17.1	**50.8**	**62.3**	**130.2**

gain for HierSE is relatively smaller than for ConSE, because HierSE has performed label disambiguation already. This also explains the larger gain for OpenImages because HierSE is inapplicable for this model and thus its English labels have not been semantically disambiguated. All these results support the effectiveness of the proposed approach for reducing label ambiguity.

Recall that the right end ($\theta = 1$) corresponds to the baseline zero-shot approaches. Better performance as compared to the baselines verifies the effectiveness of label-enhanced zero-shot learning.

Comparing the three visual models, OpenImages performs the best. We find that its vocabulary contains many more common concepts than the two alternatives. Common concepts tend to have more meaningful embeddings in the bilingual space, while at the same time have more overlap with words that common users use to describe images. Consequently, the English labels predicted by OpenImges are not only better embedded, but also more close to Chinese labels describing the test images.

Comparing three approaches. Table 2 summarizes the performance of the three approaches, *i.e.,* machine translation, zero-shot, and the proposed label-enhanced zero-shot. Notice that for ResNet-152 and GoogleNet-Shuffle, HierSE is used for it performs better than ConSE. While zero-shot performs relatively well on *hit*@1, the redundancy in its predictions, as exemplified in Table 3, results in relatively lower *hit*@5 and *hit*@10. By contrast, the proposed approach reduces such redundancy, scoring much higher *hit*@5 and *hit*@10.

Limitations. As shown in Fig. 2, the optimal value of θ depends on the visual models. Due to the limited availability of images associated with Chinese ground truth, we cannot check how well a specific value or range of θ generalizes over distinct datasets. So a future work is to construct another test set independent of Flickr8k for cross-lingual image annotation. When evaluating the relevance of predicted labels, word semantics such as synonyms and hypernyms are not taken int account. Consider the last row of Table 3 for instance. Although 体育 (sports) is relevant w.r.t. the given image, the prediction is treated as incorrect as it is not in the ground truth. Constructing a Chinese label hierarchy might help improve the evaluation.

4 CONCLUSIONS

We study cross-lingual image annotation in a novel setting where no labeled examples or image annotation model is available for the target language. Experiments on the test set of Flickr8k-CN support conclusions as follows. Equipped the pretrained OpenImages model, the proposed label-enhanced zero-shot learning approach performs the best. Compared to two baselines, *i.e.,* machine translation and zero-shot learning by ConSE / HierSE, the new approach annotates images with relevant and more diverse Chinese labels.

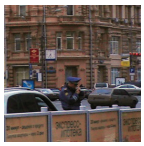
ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China (No. 61672523, No. 71531012). The Titan X Pascal used for this research was donated by the NVIDIA Corporation.

REFERENCES

[1] Alexandre Bérard, Christophe Servan, Olivier Pietquin, and Laurent Besacier. 2016. MultiVec: a multilingual and multilevel representation learning toolkit for nlp. In *LREC*.
[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *CVPR*.
[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
[4] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR* 47 (2013), 853–899.
[5] Alireza Koochali, Sebastian Kalkowski, Andreas Dengel, Damian Borth, and Christian Schulze. 2016. Which Languages do People Speak on Flickr? A Language and Geo-Location Study of the YFCC100m Dataset. In *MMCommons*.
[6] I Krasin, T Duerig, N Alldrin, A Veit, S Abu-El-Haija, S Belongie, D Cai, Z Feng, V Ferrari, V Gomes, and others. 2016. OpenImages: A public dataset for large-scale multi-label and multiclass image classification. https://github.com/openimages. (2016).
[7] Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016. Adding Chinese captions to images. In *ICMR*.
[8] Xirong Li, Shuai Liao, Weiyu Lan, Xiaoyong Du, and Gang Yang. 2015. Zero-shot image tagging by hierarchical semantic embedding. In *SIGIR*.
[9] Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees G. M. Snoek, and Alberto Del Bimbo. 2016. Socializing the Semantic Gap: A Comparative

Table 3: Top-5 predicted Chinese labels by different approaches, with correct predictions highlighted by underlines. Text in parentheses is English translation, provided for non-Chinese readers. Notice that given a test image, Chinese labels predicted by the baseline zero-shot approach are similar and thus redundant. Chinese labels predicted by the proposed approach are more diverse.

Image	Ground truth	Predicted English labels	Machine translation	Zero-shot	Label-enhanced zero-shot
	站	transport	运输	大桥 (bridge)	交通 (transport)
	汽车	vehicle	车辆	公路 (road)	车辆 (vehicle)
	城市	downtown	市中心	交通 (transport)	大桥 (bridge)
	警察	public_transport	公共交通	坡道 (rampway)	街道 (street)
	街道	urban_area	城市地区	停车场 (parking lot)	城市 (city)
		water	水	湖中 (in the lake)	水 (water)
	男人	bird	鸟	海水 (sea water)	鸟 (bird)
	水边	reflection	反射	小溪 (creek)	湖中 (in the lake)
	钓鱼	wildlife	野生动物	水花 (spray)	河 (river)
		fauna	动物	河里 (in the river)	动物 (animal)
	蓝色 房间	wood	木材	木头 (wood)	木头 (wood)
	桌子 人	man_made_object	人造物体	木地板 (wooden floor)	家具 (furniture)
	木	furniture	家具	木质 (wooden)	木地板 (wooden floor)
	坐	iron	铁	砌 (build by laying bricks or stones)	房间 (room)
	男孩	room	房间	瓷砖 (tile)	艺术 (art)
	队伍 黄色	sports	体育	篮球 (sports)	体育 (sports)
	篮球 跳	ball_game	球游戏	曲棍球 (hockey)	篮球 (basketball)
	得分 比赛	team_sport	团队运动	垒球 (softball)	曲棍球 (hockey)
	绿色 男人	player	运动员	足球 (soccer)	垒球 (softball)
	运动员 人	tournament	赛	网球 (tennis)	足球 (soccer)

Survey on Image Tag Assignment, Refinement, and Retrieval. *CSUR* 49, 1 (2016), 14:1–14:39.

[10] Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Bilingual word representations with monolingual quality in mind. In *NAACL workshop*.

[11] Pascal Mettes, Dennis Koelma, and Cees Snoek. 2016. The ImageNet Shuffle: Reorganized Pre-training for Video Event Detection. In *ICMR*.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

[13] Takashi Miyazaki and Nobuyuki Shimizu. 2015. Cross-lingual image caption generation. In *ACL*.

[14] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean. 2014. Zero-Shot Learning by Convex Combination of Semantic Embeddings. In *ICLR*.

[15] Xiaoguang Rui, Nenghai Yu, Mingjing Li, and Lei Wu. 2009. On cross-language image annotations. In *ICME*.