COCO-CN for Cross-Lingual Image Tagging, Captioning and Retrieval

Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, Jieping Xu

***Abstract*—This paper contributes to cross-lingual image annotation and retrieval in terms of data and baseline methods. We propose *COCO-CN*, a novel dataset enriching MS-COCO with manually written Chinese sentences and tags. For more effective annotation acquisition, we develop a recommendation-assisted collective annotation system, automatically providing an annotator with several tags and sentences deemed to be relevant with respect to the pictorial content. Having 20,342 images annotated with 27,218 Chinese sentences and 70,993 tags, COCO-CN is currently the largest Chinese-English dataset that provides a unified and challenging platform for cross-lingual image tagging, captioning and retrieval. We develop conceptually simple yet effective methods per task for learning from cross-lingual resources. Extensive experiments on the three tasks justify the viability of the proposed dataset and methods. Data and code are publicly available at https://github.com/li-xirong/coco-cn.**

***Index Terms*—COCO-CN, Chinese language, cross-lingual learning, image tagging, image captioning, image retrieval**

I. Introduction

AUTOMATED description of multimedia content, let it be image or video, is among the core themes for multimedia analysis and retrieval. For a long time, efforts on describing multimedia have limited the form of description to be individual words [1], *e.g., car*, *road*, and *people*, due to the semantic gap between low-level features and high-level semantics. Only recently we have witnessed noticeable advances in producing natural language-like descriptions, *e.g., cars racing on a road surrounded by lots of people*, that indicate not only the presence of specific concepts but also their interactions [2]. While these advances can certainly be attributed to technical breakthroughs, mostly deep learning, the importance of well-labeled datasets shall not be underestimated. Exemplars are ImageNet [3] for visual class recognition, NUS-WIDE [4] for image tagging, MS-COCO [5] for image captioning, and more recently Twitter100k [6] for cross-media retrieval, to name a few. Notice that annotations of these datasets are all written in the English language. To facilitate research that goes beyond the monolingual setting, this paper presents a new dataset called *COCO-CN*. It extends MS-COCO with manually written Chinese sentences and tags, see Table I.

TABLE I
EXAMPLES FROM THE PROPOSED COCO-CN DATASET. THE NEW DATASET EXTENDS MS-COCO [5] WITH MANUALLY WRITTEN CHINESE SENTENCES AND TAGS. TEXT IN PARENTHESES IS ENGLISH TRANSLATION, PROVIDED FOR NON-CHINESE READERS.

Image	MS-COCO text	COCO-CN	
		Sentences	*Tags*
	man is flying a kite in a field	一个男人在河边的草地上放风筝 (A man flying a kite on the riverside grass)	蓝天 (bule sky) 天空 (sky) 多云 (cloudy) 年轻人 (youth) 青草 (grass) 草地 (lawn)
	young man serving a tennis ball to his opponent	在红土场上，一个身穿蓝色运动服的男人正跳起发球 (On a clay court, a man wearing a blue sportswear is jumping to serve)	红土球场 (clay court) 网球 (tennis) 男人 (man) 球拍 (racket) 网球运动员 (tennis player) 红土 (red clay)
	zebras run on the grass near the trees	一群野生动物在草地上奔跑 (A herd of wild animals are running on the grass)	灌木丛 (bush) 斑马 (zebra) 草地 (grass) 大自然 (nature)

Although the importance of multimedia annotation and retrieval in a *cross-lingual* setting has been recognized early [7], only recently has the topic gained increasing attention [8]–[15]. Cross-lingual image tagging is important as it studies how to exploit data labeled in a source language to obtain image tagging models for a target language. Consequently, common users are provided with the possibility to have their images automatically tagged in a native language even when training examples in that language are in short supply. In [10], for instance, Miyazaki and Shimizu exploit a transfer learning strategy, initializing the visual embedding matrix of a Japanese captioning model using its counterpart from a trained English captioning model. Tsutsui and Crandall propose to generate multilingual sentences in a single network, using artificial tokens to control the language [8]. Lan *et al.* demonstrate the possibility of training an image captioning model by exploiting English resources to generate fluent Chinese sentences [14]. Elliott *et al.* show that using images as a side information improves multi-modal machine translation [15]. Cross-lingual applications are crucial for the majority of the world's population who do not speak English.

To support this line of research, several image datasets with non-English annotations have been developed. Depending on their applications, the target languages of these datasets differ, including German for multi-modal machine translation [9],

Manuscript received July 01, 2018; revised October 27, 2018 and December 25, 2018; accepted January 13, 2019.

This work was supported by NSFC (No. 61672523, No. 61771468), the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 18XNLG19). The associate editor coordinating the review of this manuscript and approving it for publication was Dr Wolfgang Hürst. (*Corresponding author: Jieping Xu*).

The authors are with the Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China, and the AI & Media Computing Lab, School of Information, Renmin University of China, Beijing 100872, China (e-mail: {xirong,xjieping}@ruc.edu.cn)

[11], German and French for image retrieval [12], Japanese for cross-lingual document retrieval [16] and image captioning [10], [17], and Chinese for image captioning [14], [18], [19]. Except for IAPR TC-12 [7] and AIC-ICC [19] which consist of images collected from the Internet, each of these datasets is built on top of an existing English dataset, with MS-COCO as the most popular choice.

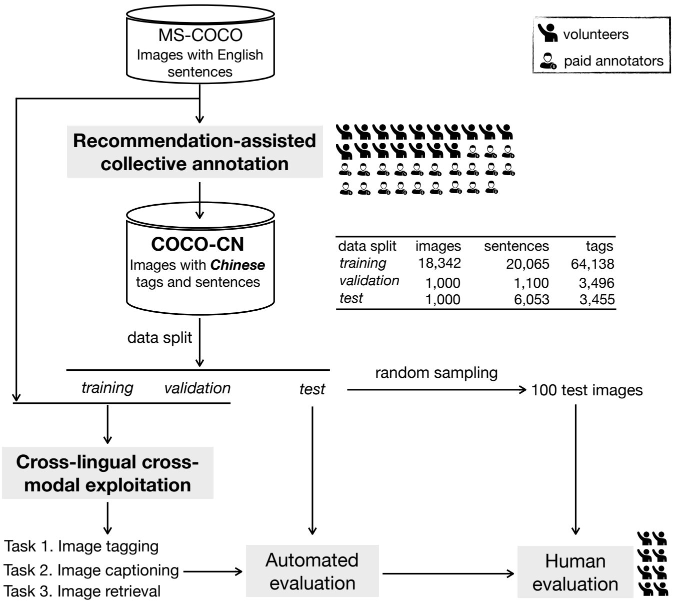
Despite the encouraging progress, we see two deficiencies in the current research in terms of datasets and methods. While Chinese is the most spoken language in the world, existing datasets for this language are either small in scale (Flickr8k-CN [18]) or strongly biased for describing human activities and monolingual (AIC-ICC [19]). As for methods, there is a paucity of literature on cross-lingual image tagging. Initially set up for cross-lingual purposes, the ImageCLEF series have factually become monolingual [20]. The only work published in the last five years[1] is by Wei *et al.* [13], where the authors assume zero-availability of training images in their target language. Clarifai [21], an online image tagging service, is capable of predicting tags in multiple languages. But the non-English tags appear to be simply machine translation of the predicted English tags. *How to effectively perform cross-lingual image tagging is unclear*.

Works on cross-lingual image captioning consistently report that for training an image captioning model for a target language, machine-translated sentences are less effective than manually written sentences [10], [14], [17], [18]. No attempt is made to answer *whether and in what ways machine-translated data can be effectively exploited together with manually-written sentences in the target language* for better performance. For matching images and sentences in two different languages, Gella *et al.* propose a neural model to learn multilingual multi-modal representations with image as pivot [22]. A shortcoming of that model is it requires many triplets for training, where each image is associated with bilingual descriptions. Cross-lingual image retrieval requires not only cross-modal matching between a query and an image, but also cross-lingual matching between the same query and the description associated with the image, but in another language.

To attack the above two deficiencies, this paper makes three contributions as follows:

• We present *COCO-CN*, a new dataset extending MS-COCO with manually written Chinese sentences and tags. It thus differs from previous cross-lingual efforts on MS-COCO that target Japanese [10], [17], German [11] or French [12]. With over 20k images described and tagged, COCO-CN is the largest Chinese-English dataset for cross-lingual image annotation and retrieval. The development of COCO-CN and its usage in the varied tasks are illustrated in Fig. 1. Data and code are available at https://github.com/li-xirong/coco-cn.

• For COCO-CN construction a novel recommendation-assisted annotation system is developed. While much progress has been made for image auto-tagging [23]–[25] and caption retrieval [26]–[28], whether these techniques can be used to assist dataset construction is unexplored. The new system allows us to assess the possibility of exploring the state-of-the-art for reducing annotation workload. Moreover, with its source code released, the system provides a good starting point for peers who are interested in building similar datasets.



data split	images	sentences	tags
training	18,342	20,065	64,138
validation	1,000	1,100	3,496
test	1,000	6,053	3,455

Fig. 1. **The development of COCO-CN and its usage in three tasks**.

• We show the applicability of COCO-CN for three tasks, *i.e.,* cross-lingual image tagging, captioning and retrieval, and its superiority over alternate datasets. Moreover, we develop models specifically for the cross-lingual setting, *i.e., Cascading MLP* for tagging, *Sequential Learning* for captioning and *Enhanced Word2VisualVec* for retrieval. While conceptually simple, these models effectively learn from bilingual resources and consequently advance the state-of-the-art in the cross-lingual context.

II. Related Work

We summarize in Table II an (incomplete) list of publicly available image datasets that are associated with manually annotated, non-English descriptions. The datasets closely related to ours in terms of the target language are Flickr8k-CN [18], Flickr30k-CN [14] and AIC-ICC [19], all focusing on Chinese.

To the best of our knowledge, Flickr8k-CN is the first dataset for image captioning in Chinese [18], adopting Flickr8k [31] as its data source. Each image has been re-annotated with five Chinese sentences written by native speakers via a local crowd sourcing service. Probably because no annotation guideline or quality control was provided, the Flickr8k-CN descriptions tend to be short, containing on average 8 Chinese characters per sentence. As for Flickr30k-CN [14], its manual annotation covers only the test set of Flickr30k [30], by collectively translating 5,000 test sentences from English to Chinese. AIC-ICC is a big step forward, consisting of 240k images collected from Internet search engines and 1.2 million crowd-sourced sentences [19]. With an average number of 21.6 Chinese characters per sentence, the texts appear to be more descriptive than their Flickr8k-

[1] Paper [13] was the only hit when querying Google Scholar with "cross-lingual image annotation" or "cross-lingual image tagging".

TABLE II
PUBLIC IMAGE DATASETS WITH MANUALLY ANNOTATED, NON-ENGLISH DESCRIPTIONS. WITH OVER 20K IMAGES DESCRIBED AND TAGGED, COCO-CN IS CURRENTLY THE LARGEST CHINESE-ENGLISH DATASET APPLICABLE FOR CROSS-LINGUAL IMAGE TAGGING, CAPTIONING AND RETRIEVAL.

Dataset	Release	Data source	Languages	Images	Sentences	Tags	Applications
IAPR TC-12 [7]	2006	Internet	English / German	20,000	70,460	no	image retrieval
Japanese Pascal sentences [16]	2015	Pascal sentences [29]	Japanese / English	1,000	5,000	no	cross-lingual document retrieval
YJ Captions [10]	2016	MS-COCO [5]	Japanese / English	26,500	131,740	no	image captioning
STAIR Captions [17]	2017	MS-COCO	Japanese / English	164,062	820,310	no	image captioning
MIC test data [12]	2016	MS-COCO	French / German / English	1,000	5,000	no	image retrieval
Bilingual caption [11]	2016	MS-COCO	German / English	1,000	1,000	no	machine translation / image captioning
Multi30k [9]	2016	Flickr30k [30]	German / English	31,014	186,084	no	machine translation / image captioning
Flickr8k-CN [18]	2016	Flickr8k [31]	Chinese / English	8,000	45,000	no	image captioning
AIC-ICC [19]	2017	Internet	Chinese	240,000	1,200,000	no	image captioning
Flickr30k-CN [14]	2017	Flickr30k	Chinese / English	1,000	5,000	no	image captioning
COCO-CN (this paper)	2018	MS-COCO	Chinese / English	20,342	27,218	yes	image tagging / captioning / retrieval

CN counterpart. Nevertheless, AIC-ICC has two downsides. First, it is a monolingual dataset, inapplicable in a cross-lingual setting. Second, all its images are about human activities, requiring human figures to be clearly visible. Thereby the AIC-ICC sentences are tailored to human activity description. As a consequence, an image captioning model trained on AIC-ICC always generates sentences with a pattern of someone doing something, even if no person is present in test images.

COCO-CN naturally inherits the visual diversity of MS-COCO. Its cross-lingual applicability is granted by jointly exploiting the new Chinese and the existing English annotations. Moreover, we collect tags to further enrich the annotations. The tags can be used for multi-label image classification. They can also be applied to multi-modal tasks, *e.g.,* image captioning with tags. All this makes COCO-CN a versatile dataset for cross-lingual image tagging, captioning and retrieval.

III. COCO-CN CONSTRUCTION

The complexity of pictorial content makes it non-trivial to write a fluent sentence to fully describe an image. Consider the image in the first row of Table I for instance. The gist of the scene is a man flying a kite outdoor. If one tries to cover the entire context, *e.g.,* sky, cloud, river, tree, and grass field, it will be a lengthy sentence that does not read smoothly. So we collect tags to complement natural-language descriptions.

A. Recommendation-assisted Collective Annotation

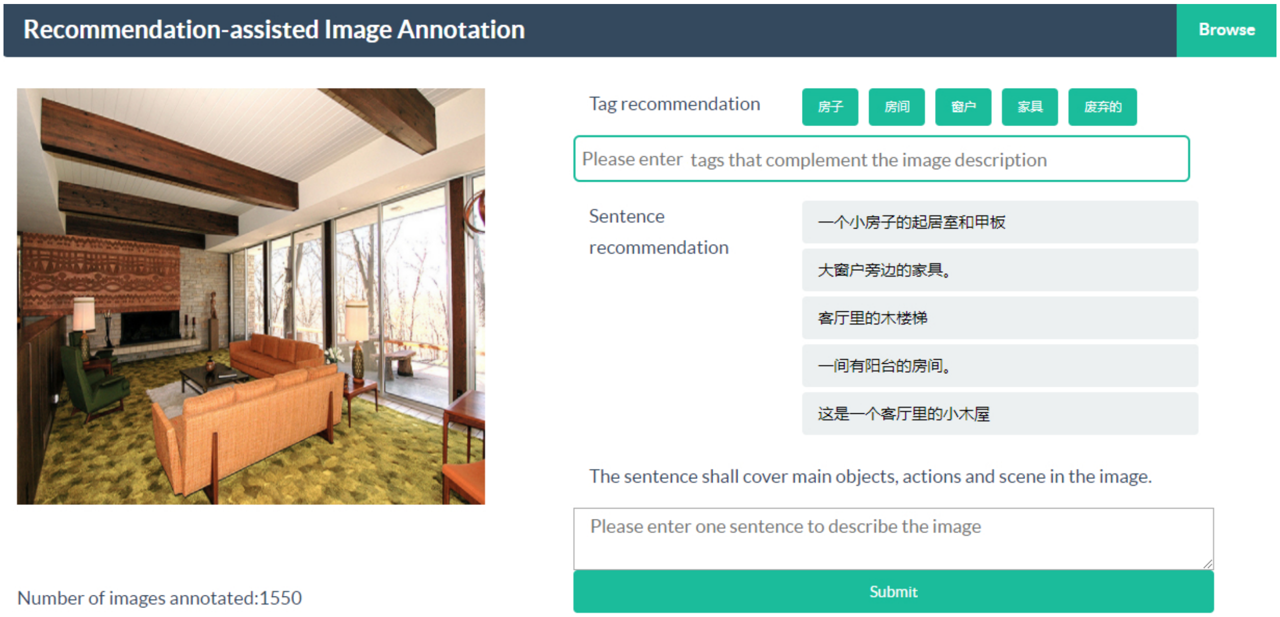
1) User Interface: For the ease of participation, we develop a web based image annotation system that allows a user to annotate images remotely and independently. Fig. 2 provides a snapshot of the system. Our novel design is the deployment of two content-based recommendation modules, one for sentences and the other for tags, that assist manual annotation on the fly. When clicked, a sentence or tag from the recommendation board will instantly appear in the corresponding editable text form. User operations including clicking and editing are logged for data analytics.

Since there is no off-the-shelf model available for recommending Chinese sentences or tags based on the pictorial content, we build our own models, as described in Section III-A2 and III-A3. Notice that we started the COCO-CN project in 2016 with the different modules developed in different periods, so the choice of image features was subject to the availability of the corresponding CNN models.

2) Chinese Sentence Recommendation: Given an image to be annotated, our sentence recommendation module retrieves the sentence most likely to be relevant from a large Chinese sentence pool. The pool, consisting of 120k sentences, is obtained by performing machine translation on the MS-COCO English sentences. Li *et al.* make an empirical comparison between two public machine translation services, *i.e.,* Google and Baidu, reporting the better performance of Baidu for translating image descriptions from English to Chinese [18]. So we use Baidu translation.

Given that each image is associated with five machine-translated sentences, a straightforward strategy would be to simply allow the annotators to select the most appropriate one and edit it as they see fit. We do not choose this strategy for the following considerations. First, we encourage the annotators, while being assisted, to describe the images from their (personal) viewpoints, rather than just make a good translation. Learning from images with unaligned bilingual descriptions is more challenging yet more practical. Second, we want to keep the system flexible to handle novel images without any descriptions. Last but not least, to justify the usability of sentence retrieval for manual annotation.

For sentence retrieval we adopt Word2VisualVec (W2VV) [28], which demonstrates state-of-the-art performance on multiple datasets. W2VV is designed to predict deep visual features from just an (English) textual input. In particular, the mapping from a given text to a visual feature vector is achieved by first encoding the text into a fixed-length vector. The vector then goes through a Multilayer Perceptron (MLP) to produce the visual feature vector. Consequently, the cross-modal relevance between a given image and a given sentence is computed as cosine similarity in the visual feature space. While W2VV is originally developed for retrieving English captions, it can be easily extended to the Chinese scenario by substituting its English based encoding layer for a Chinese counterpart. Paired images and Chinese sentences are required to train the Chinese version of W2VV. To that end, we split at random the pool of 120k sentences into two disjoint subsets, with 90% for training and 10% for validation. We train a


Recommendation-assisted Image Annotation

Tag recommendation

Please enter tags that complement the image description

Sentence recommendation

一个小房子的起居室和甲板

大窗户旁边的家具。

客厅里的木楼梯

一间有阳台的房间。

这是一个客厅里的小木屋

The sentence shall cover main objects, actions and scene in the image.

Please enter one sentence to describe the image

Number of images annotated:1550

Fig. 2. **A snapshot of our web based image annotation system**. Rather than starting from scratch, an annotator is provided with five tags and five sentences automatically recommended by the system based on the pictorial content. When clicked, the tags and sentences will instantly appear in the editable text forms to assist manual annotation. Source code of the annotation system is publicly available at https://tinyurl.com/cococn-system.

W2VV model that uses the bag-of-words based encoding and predicts a 2,048-dim ResNet-152 feature. As shown in Fig. 2, for a given image we compute its relevance to each sentence in the pool on the fly, and recommend the top-5 retrieved sentences to an annotator.

3) Chinese Tag Recommendation: Our tag recommendation module is to predict multiple Chinese tags based on the visual content. Since the MS-COCO images are quite diverse, a large vocabulary is required. We need a large-scale training set wherein each image is labeled with multiple Chinese tags. Since a dataset as such is not publicly available, we consider Flickr images associated with user tags in different languages. However, we observe that the number of Flickr images labeled with Chinese tags is much less than its English counterpart. This is in line with [32], where Koochali *et al.* report that the majority of the Flickr tags are in English. Also we note that the quality of Chinese tags is much lower. Therefore, we use Flickr images but manually translate the tag vocabulary from English to Chinese. In particular, we adopt one million Flickr images provided by [23]. To construct a Chinese tag vocabulary, we first sort all the English tags in the 1M set in descending order by the number of distinct users who have used them for image tagging. We manually go through the ranked tag list, performing one-by-one manual translation. Tags lacking correspondence to specific objects, scenes or events are excluded. Note that synonyms are grouped. For instance, 'car', 'auto' and 'automobile' are all translated into the Chinese word '汽车'. Consequently, images labeled with 'car', 'auto' or 'automobile' are considered as positive training examples of '汽车'. A word with multiple meanings is translated to a list of Chinese words. For instance, 'glass' is mapped to {'玻璃', '玻璃杯', '眼镜'}. We let the annotators decide which translation is the most appropriate. Finally we obtain a Chinese vocabulary of 1,951 classes, covering 2,085 English and 2,367 Chinese words.

As Flickr data is known to be noisy [1], [23], [33], we de-noise with the following rule of thumb. We remove overly-tagged images (associated with more than 20 tags) and images having less than two tags from the constructed vocabulary. This results in a training set of 800k images. For image representation we employ a pre-trained GoogLeNet model [34], taking the output of its pool5 layer as a 1,024-dim visual feature. An MLP with a network structure of 1024-1024-2048-1951 is trained on the 800k images. The top-5 Chinese tags predicted by the MLP model are presented to each annotator.

B. COCO-CN Annotation Process

Sentences of existing datasets such as MS-COCO, STAIR captions, Flickr8k-CN and AIC-ICC were gathered by distributing the annotation task on a specific crowd-sourcing platform. However, how to measure and thus control the quality of crowd-sourced sentences remains unresolved. Consider MS-COCO for instance. During manual translation we found a number of typos and misspellings in its sentences, *e.g.,* "ready to hi [hit] a ball", "smile son [on] their faces", "airplane with a striped tale [tail]", and "a giraffe standing by a palm team [tree]". In order to gather high-quality annotations, we did not fully rely on crowd sourcing. Our annotation team was comprised of 17 volunteers (staff and graduate students in our lab) and 22 paid undergraduate students from our department.

Before they started, the team were instructed with the following guidelines. A sentence shall cover the main objects, actions and scene in a given image. The annotators were asked to provide at least one tag that complements the image description. For a better understanding of the guidelines, we

TABLE III
STATISTICS OF THE **COCO-CN** SENTENCES AND TAGS. THE SENTENCES CONTRIBUTE A VOCABULARY OF 7,096 DISTINCT WORDS, WHERE THE PERCENTAGES OF NOUNS, VERBS AND ADJECTIVES ARE 59.5%, 22.6% AND 5.1%. THE NUMBER OF DISTINCT TAGS IS 4,867. THE NUMBER OF TAGS OUTSIDE THE SENTENCE VOCABULARY IS 1,918.

Chars per sentence			**Words per sentence**			**Tags per image**		
min	*max*	*mean*	*min*	*max*	*mean*	*min*	*max*	*mean*
4	59	16.8	2	43	10.9	1	15	3.5

provided a number of well / badly labeled examples. For quality control, an inspection committee comprised of paper authors regularly sampled a small proportion of the annotation results by distinct annotators for a manual check. Annotators responsible for unqualified results were put into a watch list and informed to review and re-edit their annotations. An annotator entering the watch list three times would be excluded from the team, with his/her annotations removed as well.

For effective acquisition of labeled data, our goal is to have more images annotated than to collect more sentences per image. This strategy is inspired by [28] where the authors empirically show that given the same number of image-sentence pairs for training, having more images results in a better model for cross-modal matching. To that end, we generated for each annotator a unique list of images by randomly shuffling the MS-COCO images. Consequently, the chance of two annotators labeling the same image was low.

C. Data Analytics

1) Progress Overview: We have 20,342 images annotated with 22,218 sentences and 70,993 tags in total. Notice that we did not select these 20k images in advance. They were obtained as a consequence of manual annotation. Since images were randomly assigned to each annotator, the 20k images in COCO-CN are factually a representative subset of MS-COCO. In addition, we have 5,000 MS-COCO English sentences *manually* translated to Chinese for evaluating image captioning (c.f. Section IV-C). This adds the number of manually written Chinese sentences up to 27,218. Different from an English sentence where words are separated by whitespace characters, a Chinese sentence has no marker as word boundaries. It has to be segmented into a sequence of meaningful words before further analysis. We employ the BosonNLP toolkit [35] for sentence segmentation and part-of-speech tagging. Basic statistics of the sentences and tags are listed in Table III. The average number of annotators per image is 1.1, while the average number of images annotated per person is 521.6.

Chinese tag vocabulary. The sentences contribute a vocabulary of 7,096 distinct words, while the manual tags contribute a vocabulary of 4,867 distinct words, among which 1,918 words are outside the sentence vocabulary. At the image level, on average nearly 45% of the manual tags are not covered by the associated sentences. It is clear that only tags with a reasonable amount of training images can be effectively modeled. So we expand the manual tags with words automatically extracted from Chinese sentences as follows. We consider only nouns, adjectives and verbs. Adjectives and verbs consisting of single Chinese characters, *e.g.,* '满', '亮' and '穿', are over generic and thus removed. We further exclude rare tags that occur less than 20 times in the COCO-CN training set (see Fig. 1 and Section IV-A for data partition). We then manually go through the remaining tags, deleting those we consider unsuited for describing the pictorial content. Consequently, we obtain a vocabulary of 655 Chinese tags. The number of positive training images ranges from 20 to 1,876, with an averaged number of 123.

COCO-CN tags *versus* MS-COCO object classes. The images in the MS-COCO dataset are already annotated with 80 object classes such as *person*, *bicycle* and *toothbrush*. After manually translating these classes into Chinese tags, we find that there are 70 tags in common. In other words, we have enriched the MS-COCO annotations with $655 - 70 = 585$ novel concepts.

COCO-CN sentences *versus* MS-COCO sentences. For the ease of cross-lingual comparison, for each image in COCO-CN we employ machine translation to convert its five English captions from MS-COCO to Chinese. Given a COCO-CN sentence we investigate the overlap between its keywords, *i.e.,* nouns, verbs and adjectives, and those in the translated sentences. On average a COCO-CN sentence brings in 45.3% novel keywords, with 39.9% nouns, 51.9% verbs, and 69.5% adjectives. The usage of adjectives is more subjective and personalized. Consider for instance describing a dog as *small* or *big* and judging if a cat is *cute*.

Annotation quality. Concerning the annotation team, no one entered the watch list three times. The annotators were asked to perform the annotation task independently, which helped improve the diversity of human annotations. The regular inspection mechanism, with nearly 10% of the images double-checked, helped maintain the overall annotation quality. Concerning the variety of description for each image and the vocabulary for the description, we evaluate in two aspects as follows. First, for each image we use its five machine translated sentences as references, and compute four metrics, i.e., BLEU-4, METEOR, ROUGE-L and CIDEr, that have been *de facto* for evaluating caption quality. The corresponding scores are 16.1, 27.6, 42.9, and 60.0, respectively, suggesting a substantial divergence between human annotations and machine translations. Second, we compare their vocabulary size. Not surprisingly, the human vocabulary, with a size of 7,096, is smaller than the machine vocabulary which has 14,219 distinct words. Notice however that the number of machine-translated sentences is five times as large as the number of manually written sentences. So the human vocabulary is reasonably diverse. Moreover, we checked low-level typos in each manually written sentence by matching each of its words with a vocabulary of nearly 350k words from the Jieba Chinese text segmentation tool. For 942 out of the 22,218 sentences, they contain words outside the vocabulary. A manual check shows that only 56 sentences have true typos, *e.g.,* incorrectly typing '西兰花' (broccoli) as '西蓝花' or '牛仔裤' (jeans) as '牛子裤'. For the other 886 sentences, their mismatched words such as '矮柜' (low cabinet) and '幼象' (baby elephant) are actually valid. The accuracy of the human annotations will be further verified by image captioning experiments, where

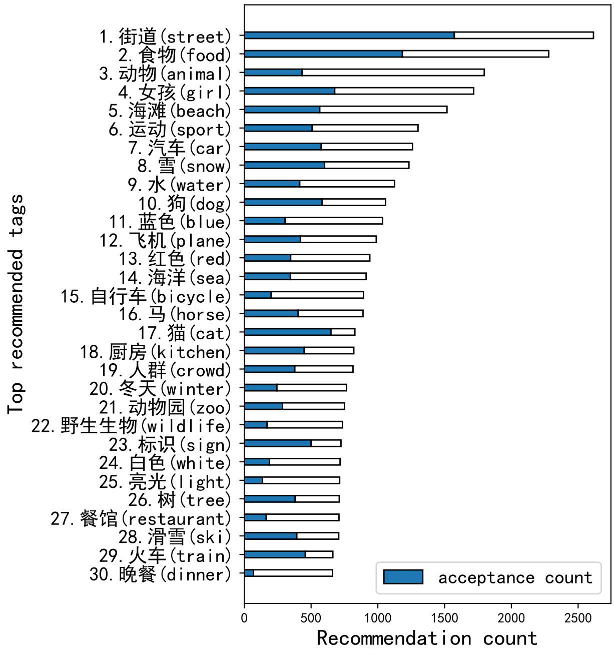


Fig. 3. **The top-30 recommended tags**. Their English translations are provided in parentheses for non-Chinese readers.

COCO-CN is compared against current alternatives including Flickr8k-CN and AIC-ICC as training data. We also compare statistics of the sentences respectively written by the volunteer group and the paid group, and find no considerable difference.

2) The Influence of Recommendation: We now analyze if the two recommendation modules are helpful during the annotation process.

Tag recommendation. The annotators accepted approximately 2.7 out of the five suggested tags per image, with an acceptance rate of 54%. Fig. 3 shows the top 30 recommended tags and their acceptance counts. The best-predicted tag is '猫' (*cat*), got accepted 649 out of 828 times.

Sentence recommendation. Among the 22,218 sentences, 31% of them were typed by the annotators after clicking (and optionally re-editing) the recommendations. The number of edited sentences is 5,796. To measure the manual effort saved per image, we propose to calculate the Levenshtein distance [36] between the recommended sentence clicked by an annotator and the final submitted sentence. As a string metric, the Levenshtein distance between two Chinese sentences is the minimum number of single Chinese character edits (insertions, deletions or substitutions) required to change one sentence into the other. Our baseline is sentence annotation without recommendation, where the distance between the annotation and the recommendation (which is null) is actually the length of the annotation. The baseline has an average distance of 16.8, while the recommendation counterpart is 15.0. Fig. 4 shows the distance distribution curves with and without recommendation. These results suggest that sentence recommendation reduces the annotation workload to some extent. Nevertheless, sentence recommendation is less effective than tag recommendation.

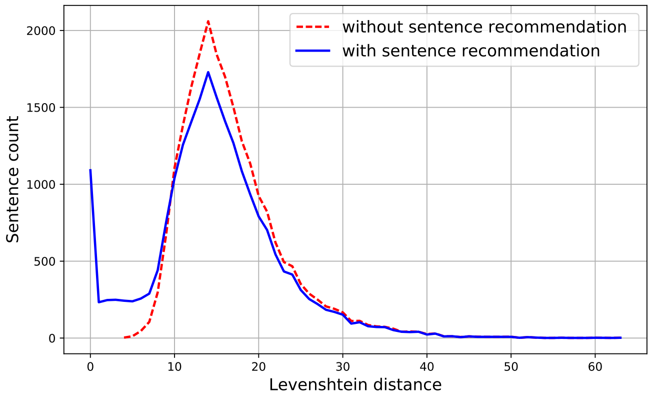


Fig. 4. **Distribution of the Levenshtein distance with and without sentence recommendation**. Sentence recommendation decreases the distance, and thus reduces annotation workload.

IV. Applications

A. *Common setup*

Before detailing the individual tasks, we first describe an experimental setup shared by the tasks.

Data partition. We randomly select 1,000 images from COCO-CN as a held-out test set, and another set of 1,000 images as a validation set for hyper parameter optimization. The remaining 18,342 images are used for training. When evaluating image tagging and captioning in an automatic manner, limitations exist due to vocabulary discrepancy and incomplete ground truth. Therefore, we randomly select 100 images from the test set, which we term COCO-CN test100, for human evaluation. In addition, for human evaluation we construct another test set of 100 images, which are randomly sampled from NUS-WIDE other than MS-COCO.

Image feature. For the three tasks, we employ the state-of-the-art ResNeXt-101 model [37] with input size of 224×224. The model, provided by the MediaMill team at University of Amsterdam, is pre-trained on the full ImageNet dataset with a bottom-up reorganization [34] of the 22k ImageNet classes. The reorganization balances the classes by grouping or deleting over-specific and rare classes and down-sampling training examples for over-frequent classes, and consequently reduces the number of classes to 12k. We take the output of the last pooling layer, obtaining a 2,048-dim feature vector. For each image we resize it to 256×256, and extract its feature using an oversampling strategy. That is, we extract features from its 10 sub images, obtained by clipping the image and its horizontal flip with a window of 224×224 at their center and four corners. The 10 features are averaged as the image feature. Note that the ResNeXt-101 feature is better than the GoogLeNet and ResNet-152 features previously used in our annotation system, see the appendix.

B. *Task I: Cross-Lingual Image Tagging*

We describe how to exploit the Chinese annotations from COCO-CN and the English annotations from MS-COCO for building an image tagging model that predicts Chinese tags for a given image.

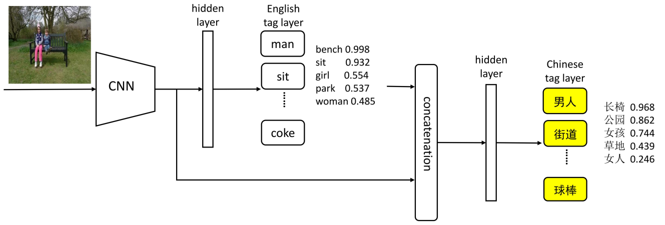


Fig. 5. **The proposed Cascading MLP for cross-lingual image tagging**. We train the first MLP that predicts English tags. The output of this MLP is concatenated with the image CNN feature to form the input for the second MLP, which predicts Chinese tags.

1) Setup: **Tagging model**. For learning from the cross-lingual resources, we propose Cascading MLP. As illustrated in Fig. 5, the model consists of two MLPs. The first MLP is trained on MS-COCO for English tag prediction. A vocabulary of 512 English tags is constructed from the original MS-COCO captions by sorting nouns, verbs and adjectives in descending order in terms of their occurrence frequency. As such, this MLP essentially transforms the visual feature of a given image into a 512-dim semantic feature, with each dimension corresponding to a specific English tag. The visual and semantic features are concatenated to form a semantic-enhanced input of the second MLP. This MLP is trained using COCO-CN to predict the probability of each Chinese tag being relevant with respect to a given image. As shown in Fig. 5, each MLP has one hidden layer consisting of 1,024 neurons. A rectified linear unit (ReLU) is followed to increase the nonlinearity of the model. As an image can be described by multiple tags, we use sigmoid (rather than softmax commonly used for visual recognition) as the activation function of the two output layers. The loss function is binary cross entropy. During training we apply a common dropout trick (with a rate of 0.5) to reduce the risk of overfitting.

Baselines. In order to justify the effectiveness of the proposed model, we consider the following four alternates.
1) COCO-MT. An MLP is trained using tags extracted from the machine-translated Chinese sentences of MS-COCO. Leaving out 2k images covered by the COCO-CN validation and test sets, COCO-MT has 121,287 training images.
2) COCO-CN. An MLP is trained on COCO-CN alone.
3) Clarifai. An online image tagging service [21] that predicts for a given image up to 200 tags in multiple languages including Chinese. The service claims to be a market leader since winning the top five places in image classification at the ImageNet 2013 competition, and is accessible via its API.
4) Multi-task MLP. A multi-task MLP is trained on COCO-CN with bi-lingual annotations to simultaneously predict English and Chinese tags. The model is a noncascaded version of Fig. 5, where the sum of the English and Chinese binary cross-entropy losses is minimized during training.

Evaluation criteria. We report the popular Precision, Recall and F-measure at top 5. Each metric is computed per image and averaged over all test images.

2) Results: **Automated evaluation**. Table IV shows automated evaluation results of the five tagging models. The proposed Cascading MLP tops the performance. Although the number of training images in COCO-MT is 6.6 times as large as COCO-CN, the MLP trained on COCO-CN outperforms its COCO-MT counterpart with a clear margin. This result shows the importance of high-quality annotation for training. Some image tagging results are presented in Table VII. In the first row, the tag *umbrella* is not predicted by MLP trained on COCO-CN, while in the second row, MLP trained on COCO-MT incorrectly predicts the tag *keyboard*. Learning from the two complementary datasets, Cascading MLP makes better prediction in general.

TABLE IV
AUTOMATED EVALUATION OF DIFFERENT MODELS FOR IMAGE TAGGING. CASCADING MLP LEARNED FROM CROSS-LINGUAL DATA IS THE BEST.

Model	**Precision**	**Recall**	**F-measure**
Clarifai [21]	0.217	0.261	0.228
MLP trained on COCO-MT	0.432	0.525	0.456
MLP trained on COCO-CN	0.477	0.576	0.503
Multi-task MLP	0.482	0.583	0.508
Cascading MLP	**0.491**	**0.594**	**0.517**

Multi-task MLP is better than MLP trained on COCO-CN, suggesting a positive effect of multi-task regularization. Nonetheless, it is less effective than Cascading MLP. Moreover, in contrast to Multi-task MLP, Cascading MLP, by learning two monolingual MLPs sequentially, does not require aligned bi-lingual annotations per training image, and is thus more flexible to exploit partially annotated yet larger training data. Therefore, the cascading architecture is more suited for learning from unaligned cross-lingual resources.

We also compare Multi-task MLP and Cascading MLP on Flickr8k-CN, where the result shows that the latter is better. For more details we refer to https://tinyurl.com/tag-flickr8kcn.

Human evaluation. It is possible that the lower performance of Clarifai is caused by a discrepancy between the Chinese vocabulary of the online service and our ground-truth vocabulary. To resolve this uncertainty, we performed a user study as follows. For each of the 100 pre-specified images, we collected the top 5 predicted tags by each model. Eight subjects participated in the user study. Each image together with the collected tags was shown to two subjects, who independently rated each of the tags as *relevant*, *irrelevant* or *unsure*. To avoid bias, the tags were randomly shuffled in advanced. Only tags rated as relevant by both subjects were preserved. Table V shows the tagging performance given the new ground truth. As the ground truth is more complete, all the scores improve. Cascading MLP again performs the best. Moreover, the qualitative conclusion concerning which model performs better remains the same.

To prevent dataset bias, we repeated the human study on another test set of 100 images randomly sampled from NUS-WIDE [4], which is independent of all the training sets, *i.e.,* Flickr8k-CN, AIC-ICC, COCO-MT and COCO-CN, used in this work. We term this additional test set NUS-WIDE100. As Table VI shows, Cascading MLP is again the best.

TABLE V
HUMAN EVALUATION OF DIFFERENT MODELS FOR IMAGE TAGGING ON **COCO-CN TEST100**. CASCADING MLP AGAIN PERFORMS THE BEST.

Model	**Precision**	**Recall**	**F-measure**
Clarifai	0.634	0.358	0.451
MLP trained on COCO-MT	0.778	0.453	0.563
MLP trained on COCO-CN	0.836	0.488	0.607
Cascading MLP	**0.858**	**0.501**	**0.623**

TABLE VI
HUMAN EVALUATION OF IMAGE TAGGING ON **NUS-WIDE100**, A RANDOM SUBSET OF 100 IMAGES FROM NUS-WIDE.

Model	**Precision**	**Recall**	**F-measure**
Clarifai	0.636	0.448	0.515
MLP trained on COCO-MT	0.668	0.472	0.542
MLP trained on COCO-CN	0.694	0.487	0.561
Cascading MLP	**0.726**	**0.509**	**0.588**

C. Task II: Cross-Lingual Image Captioning

Naturally, COCO-CN is suitable for research on image captioning in Chinese and in a cross-lingual setting. To demonstrate the necessity of COCO-CN, we present the following experiments, comparing models trained on different datasets.

1) Setup: **Test set**. Each of the 1,000 test images in COCO-CN is associated with one manually written Chinese sentence at least. To enrich the ground truth, we manually translated the existing five English sentences per test image. Consequently, each test image has six Chinese sentences as ground truth.

Captioning model. We adopt the Show and Tell network [38], originally developed for English caption generation and later found to be effective for Chinese caption generation [14], [18], [19]. The network computes the posterior probability of a sentence S given an input image x by combining a CNN based image encoder and an LSTM based sentence decoder. Consequently, the image is annotated with the sentence yielding the maximal probability. The size of LSTM, *i.e.,* the dimensionality of its hidden vector, is empirically set to 512.

Training strategies. As mentioned above, Flickr8k-cn [18] and AIC-ICC [19] already exist for image captioning in Chinese. One more alternate is COCO-MT, a machine translation version of MS-COCO. We compare COCO-CN with these three datasets. For a fair comparison the Show and Tell model is separately trained using each dataset. The model is trained using standard supervised learning that minimizes the cross entropy loss. The initial learning rate is set to be 0.0005, which decays every three epochs with a decaying factor of 0.8.

While the machine-translated sentences in COCO-MT tend to contain grammatical errors and do not read naturally, this dataset has a much larger vocabulary (with 10,144 words occurring at least 5 times) for training. In order to effectively exploit the large-scale property of COCO-MT and the high-quality property of COCO-CN, we introduce a *Sequential Learning* strategy. Using the joint vocabulary of the two datasets, we first train the captioning model using COCO-MT. After 30 epochs, we pick up the model that has the highest CIDEr score on the COCO-CN validation set. We continue to train this model using COCO-CN, with a learning rate of 0.00005. Sequential Learning is in some way similar to fine tuning, with its focus on appropriate utilization of machine-translated and manually written training sentences. From this viewpoint, Sequential Learning is not new by itself. Rather, it provides a novel and more effective way to exploit cross-lingual resources, when compared to existing works [8], [10].

Baselines. To the best of our knowledge, the *Transfer learning* model [10] and the *Artificial token* model [8] are the only two existing alternatives that target at learning from cross-lingual resources, both originally proposed for generating Japanese captions. In our implementation, the former initializes the visual embedding matrix of a Chinese captioning model using its counterpart in a trained English captioning model, while the latter learns a bilingual model from both manually written Chinese captions and original MS-COCO English captions. In addition, one might consider having COCO-MT and COCO-CN simply blended and used for training the image captioning model in one stage. We term this strategy COCO-Mixed, a necessary baseline to investigate whether the sequential learning strategy really matters.

Evaluation criteria. We use standard automatic evaluation metrics to evaluate caption quality, namely BLEU, METEOR, ROUGE-L and CIDEr. While being convenient, the automated metrics have weak correlations with human judgments of caption quality. We perform a user study, in terms of relevance and fluency, on COCO-CN test100 and NUS-WIDE100. Fluency reflects the extent to which a sentence reads naturally. The same eight subjects participated in the user study. Each sentence was independently rated by two subjects on a Likert scale of 1 to 5 (higher is better) for relevance and fluency. To help a subject give more comparable scores, each test image and sentences generated by different models were presented together to the subject. To avoid bias, the sentences were randomly shuffled in advance.

2) Results: **Automated evaluation**. Table VIII shows automated evaluation results of different captioning models. Models trained on Flickr8k-CN and AIC-ICC have fairly low scores. The relatively limited size of Flickr8k-CN makes it lack the ability for training a good captioning model. As for AIC-ICC, although it is the largest Chinese captioning dataset, it is strongly biased that all the images are about human beings. Consequently, the model trained on this dataset is unsuitable for describing general images. The model trained on COCO-CN is on par with the model trained on machine translated sentences of the full MS-COCO set. As for COCO-Mixed, since manually written sentences from COCO-CN are overwhelmed by machine translated sentences from COCO-MT during training, the benefit of COCO-CN appears to be marginal. By contrast, Sequential learning is quite effective, performing best under all metrics.

Human evaluation. Table IX shows human evaluation of distinct models on the two test sets. The COCO-MT model receives the lowest fluency score of 4.50 on COCO-CN test100 and 4.36 on NUS-WIDE100. Errors in machine translation impair the accuracy and grammar of the sentences generated by the COCO-MT model. With sentences carefully written by native Chinese speakers and covering a wide variety of visual content, COCO-CN is suitable for generic Chinese

TABLE VII
SOME TAGGING AND CAPTIONING RESULTS BY DIFFERENT MODELS. TEXTS IN PARENTHESES ARE ENGLISH TRANSLATIONS, PROVIDED FOR NON-CHINESE READERS.

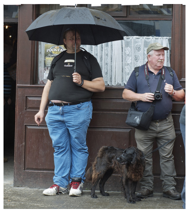
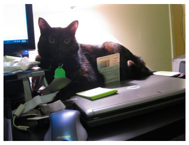
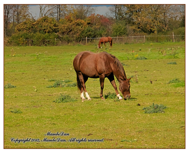
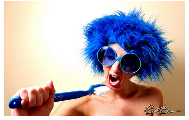
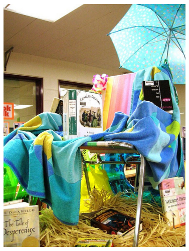
Test image	**Results**
	***Image Tagging***: Manual: 雨伞(umbrella), 帽子(hat), 男人(man) 狗(dog), 相机(camera), 黑色(black) COCO-MT: 狗(dog), 男人(man), 伞(umbrella), 雨伞(umbrella), 女人(woman) COCO-CN: 狗(dog), 男人(man), 黑色(black), 女人(woman), 街道(street) Clarifai: 人(person), 巡逻(patrol), 女人(woman), 男性(male), 狗(dog) Multi-task MLP: 狗(dog), 男人(man), 街道(street), 女人(woman), 黑色(black) Cascading MLP: 狗(dog), 男人(man), 雨伞(umbrella), 黑色(black), 女人(woman) ***Image Captioning***: Manual: 一个挎着相机的男人旁站着一位撑伞的男子和一条黑狗 (A man with a camera stood by a man holding a umbrella and a black dog) AIC-ICC: 一个戴着帽子的男人和一个右手拿着伞的女人走在道路上 (A man in a hat and a woman with an umbrella in her right hand walk on the road) COCO-MT: 一个男人和一个女人拿着一把伞 (A man and a woman hold an umbrella) COCO-CN: 两个男人和一个女人坐在长椅上 (Two men and a woman sitting on a bench) COCO-Mixed: 一个男人和一个女人站在一把伞下 Sequential Learning: 两个男人和一只狗站在一把伞下 (Two men and a dog stand under an umbrella)
	***Image Tagging***: Manual: 猫(cat), 笔记本电脑(laptop) COCO-MT: 猫(cat), 电脑(computer), 笔记本(notebook), 桌子(table), 键盘(keyboard) COCO-CN: 电脑(computer), 猫(cat), 笔记本(notebook), 笔记本电脑(laptop), 书桌(desk) Clarifai: 计算机(computer), 笔记本电脑(laptop), 商业(business), 技术(technology), 房间(room) Multi-task MLP: 猫(cat), 电脑(computer), 笔记本(notebook), 笔记本电脑(laptop), 书桌(desk) Cascading MLP: 电脑(computer), 猫(cat), 笔记本(notebook), 笔记本电脑(laptop), 书桌(desk) ***Image Captioning***: Manual: 一只黑猫趴在一台笔记本电脑上 (A black cat is lying on a laptop) AIC-ICC: 房间里有一个坐在椅子上的女人在看电脑 (There is a woman sitting on the chair in the room watching the computer) COCO-MT: 躺在笔记本电脑上的猫 (The cat lying on the laptop) COCO-CN: 一只猫趴在笔记本电脑上 (A cat is lying on the laptop) COCO-Mixed: 一只黑猫坐在笔记本电脑旁边的桌子上 (A black cat sits on the table next to the laptop) Sequential Learning: 一只黑猫趴在笔记本电脑上 (A black cat is lying on the laptop)
	***Image Tagging***: Manual: 马(horse), 棕色(brown), 草场(meadow) COCO-MT: 马(horse), 草地(grassland), 田野(field), 棕色(brown), 草(grass) COCO-CN: 马(horse), 草地(grassland), 草(grass), 棕色(brown), 草场(meadow) Clarifai: 放牧(graze), 草坪(lawn) 哺乳动物(mammal), 有人(no person), 乡村的(countryside) Multi-task MLP: 马(horse), 草地(grassland), 草(grass), 棕色(brown), 草场(meadow) Cascading MLP: 马(horse), 草(grass), 草地(grassland), 棕色(brown), 草场(meadow) ***Image Captioning***: Manual: 两只棕色的马在草地上吃草 (Two brown horses graze on the grassland) AIC-ICC: 绿油油的草地上有一个穿着白色上衣的女人在骑马 (On the green grass there is a woman in a white coat riding a horse) COCO-MT: 在草地上吃草的马 (A horse that grazes on the grassland) COCO-CN: 一匹马站在草地上 (A horse stands on the grassland) COCO-Mixed: 一匹棕色的马站在一片绿油油的田野上 (A brown horse stands on a green field) Sequential Learning: 一匹棕色的马在草地上吃草 (A brown horse grazes on the grassland)
	***Image Tagging***: Manual: 墨镜(sunglass), 蓝色(blue), 头发(hair), 牙刷(toothbrush), 女孩(girl) COCO-MT: 牙刷(toothbrush), 刷牙(brush one's teeth), 蓝色(blue), 女人(woman), 眼镜(glass) COCO-CN: 牙刷(toothbrush), 刷牙(brush one's teeth), 蓝色(blue), 剪刀(scissor), 墨镜(sunglass) Clarifai: 时尚(fashion), 太阳镜(sunglass), 乐趣(fun), 女人(woman), 肖像(portrait) Multi-task MLP: 牙刷(toothbrush), 刷牙(brush one's teeth), 蓝色(blue), 女人(woman), 剪刀(scissor) Cascading MLP: 牙刷(toothbrush), 蓝色(blue), 刷牙(brush one's teeth), 女人(womaon), 头发(hair) ***Image Captioning***: Manual: 一位蓝色头发带着蓝色墨镜的女孩想用巨大的蓝色牙刷刷牙 (A girl with blue hair and blue sunglasses wants to brush her teeth with a huge blue toothbrush) AIC-ICC: 房间里有一个戴着墨镜的女人在喝水 (There is a woman wearing sunglasses in the room drinking water) COCO-MT: 一个戴着牙刷的人 (A person wearing a toothbrush) COCO-CN: 一只手拿着一把牙刷 (A toothbrush in one hand) COCO-Mixed: 一个戴着眼镜的女人在刷牙 (A woman with glasses is brushing her teeth) Sequential Learning: 一个穿着蓝色衣服的女人正在刷牙 (A woman in blue is brushing her teeth)
	***Image Tagging***: Manual: 雨伞(umbrella), 布料(cloth), 房子(house), 草(grass) COCO-MT: 毛巾(towel), 雨伞(umbrella), 桌子(table), 黄色(yellow), 蓝色(blue) COCO-CN: 蓝色(blue), 椅子(chair), 玩具(toy), 雨伞(umbrella), 伞(umbrella) Clarifai: 有人(no person), 购物(shopping), 市场(market), 存储(storage), 廉价出售(undersell) Multi-task MLP: 椅子(chair), 玩具(toy), 蓝色(blue), 伞(umbrella), 玩偶(doll) Cascading MLP: 雨伞(umbrella), 蓝色(blue), 椅子(chair), 桌子(table), 摆放(display) ***Image Captioning***: Manual: 一间房子里有一把伞，椅子上有一些布料，地上有一些草 (There is an umbrella in a house, with some cloth on the chair and some grass on the ground.) AIC-ICC: 房间里有一个穿着蓝色衣服的女人在看书 (There is a woman in blue in the room reading) COCO-MT: 有许多不同颜色的雨伞的房间 (Rooms with umbrellas of many colors) COCO-CN: 房间里放着许多毛绒玩具 (There are many stuffed toys in the room) COCO-Mixed: 有许多不同颜色的雨伞的房间 (Rooms with umbrellas of many colors) Sequential Learning: 房间里有一把蓝色的雨伞和一把椅子 (There is a blue umbrella and a chair in the room)

TABLE VIII
AUTOMATED EVALUATION OF IMAGE CAPTIONING MODELS TRAINED ON DIFFERENT DATASETS. THE PROPOSED CROSS-LINGUAL TRANSFER PERFORMS THE BEST.

Training	BLEU-4	METEOR	ROUGE-L	CIDEr
Flickr8k-CN	10.1	14.9	33.8	22.9
AIC-ICC	7.4	21.3	34.2	24.6
COCO-MT	30.2	27.1	50.0	86.2
COCO-CN	31.7	27.2	52.0	84.6
COCO-Mixed	29.8	28.6	50.3	86.8
Transfer learning [10]	33.7	28.2	52.9	89.2
Artificial token [8]	31.8	26.9	51.5	85.4
Sequential Learning	**36.7**	**29.5**	**55.0**	**98.4**

TABLE IX
HUMAN EVALUATION OF IMAGE CAPTIONING MODELS ON TWO TEST SETS. THE NUMBERS AFTER ± ARE STANDARD DEVIATIONS. SEQUENTIAL LEARNING OBTAINS THE BEST OVERALL SCORE.

	COCO-CN test100		NUS-WIDE100	
Model	Relevance	Fluency	Relevance	Fluency
Flickr8k-CN	2.34 ±1.10	4.62 ±0.62	1.96 ±1.03	4.71 ±0.48
AIC-ICC	2.63 ±1.24	4.75 ±0.44	2.17 ±1.10	**4.73** ±0.55
COCO-MT	3.89 ±0.77	4.50 ±0.77	2.92 ±1.18	4.36 ±0.86
COCO-CN	3.95 ±0.90	**4.83** ±0.51	2.69 ±1.08	4.59 ±0.72
Transfer Learning [10]	3.89 ±0.83	4.78 ±0.49	2.86 ±1.10	4.62 ±0.61
Artificial token [8]	3.78 ±0.99	4.78 ±0.69	2.61 ±1.10	4.46 ±0.90
Sequential Learning	**4.13** ±0.72	4.76 ±0.59	**3.02** ±1.15	4.67 ±0.60

image captioning. Sequential learning allows us to effectively leverage the manual annotations. While its fluency score of 4.76 is lower than the COCO-CN counterpart on COCO-CN test100 and that of AIC-ICC on NUS-WIDE100, sequential learning gives the best relevance score in terms of both automatic and human evaluation, suggesting that the generated captions are the most descriptives.

Examples of captions generated by different models are shown in Table VII. Consider the second row for instance. The manually written sentence is 一只黑猫趴在一台笔记本电脑上 (A black cat is lying on a laptop). The AIC-ICC model predicts 房间里有一个坐在椅子上的女人在看电脑 (There is a woman sitting on the chair in the room watching the computer), showing a strong bias of the model on describing human actions. Compared to the COCO-CN model which describes the image as 一只猫趴在笔记本电脑上 (A cat is lying on the laptop), both COCO-Mixed and Sequential Learning is more descriptive by predicting 黑猫 (black cat). The result shows the benefit of learning from the cross-lingual resources. Moreover, compared to COCO-Mixed which uses the verb 坐 (sit), Sequential Learning uses the verb 趴 (lying), which is more precise and vivid in the Chinese language.

D. Task III: Cross-Lingual Image Retrieval

1) Setup: The bilingual property of COCO-CN supports cross-lingual image retrieval. We consider in this paper the following setting. Given a Chinese sentence as a query, the goal is to find amidst a set of images the one best matches the query. Each of the images is described by an English sentence. The proposed cross-lingual image retrieval differs from a monolingual task as the former has to look into not only similarities between the visual and textual modalities but also similarities between texts presented in two distinct languages. To accomplish the proposed task, one needs to effectively represent Chinese queries, images and their English descriptions in a common space by cross-modal, cross-lingual representation learning. The task is thus of more research interest than its monolingual counterpart. While cross-modal matching is well known to be difficult, image retrieval by cross-lingual matching is also challenging. Even if the Chinese and English sentences are meant for describing the same image, they were independently written by distinct subjects, presumably with varied educational and cultural background. So they are not necessarily semantically aligned. A desirable method shall exploit cross-modal matching between the query and a given image and cross-lingual matching between the query and the English description of the given image.

Test set. We again use the test set of COCO-CN. For each image, its manually written Chinese sentence is taken as a query, resulting in a query set of size 1,000. Recall that each image has five English sentences from MS-COCO. We take the first sentence as the English description of the image.

Retrieval models. We see three approaches. One, cross-lingual (CL) matching between the Chinese query and the English description, with the visual content ignored. Two, cross-modal (CM) matching between the query and the image, with the English description ignored. Third, cross-lingual and -modal (CLM) matching that combines the previous two approaches. W2VV previously used for sentence recommendation can project a given sentence into a visual feature space. Having sentences of distinct languages projected into the same visual feature space enables both cross-lingual and cross-modal matching. Hence, we adopt W2VV as a unified solution.

We start with the best configuration of W2VV [28]. An input sentence is vectorized in parallel by three strategies: bag-of-words (BoW), Word2Vec (w2v) and a Gated Recurrent Units (GRU) network, as Fig. 6 shows. For GRU based vectorization, W2VV uses the hidden vector at the last time step, which is now known to be over compact to represent the entire sentence [15], [39]. Consider a query sentence "三个塑料饭盒中装着面条，蔬菜和水果" (Three plastic lunch boxes containing noodles, vegetables and fruits) for instance. While the query requires co-presence of multiple objects such as 塑料饭盒 (plastic lunch boxes), 面条 (noodles), 蔬菜 (vegetables) and 水果 (fruits), W2VV tends to over-emphasize the ending part of the query, *i.e.,* fruits. Moreover, W2VV is trained using the Mean Squared Error (MSE), which does not consider negative pairs, *i.e.,* images and sentences that are irrelevant. Based on these two observations, we enhance the encoding module and the loss of W2VV as follows.

First, we improve its GRU part by introducing a soft attention mechanism [39]. As shown in Fig 6, the attention layer is placed right after GRU, summing up the hidden vectors over all time steps with an adaptive weight assigned to each vector. More formally, let h_i be the hidden vector at time step i, with $i = 1, \ldots, n$ and n is the maximum time step. The corresponding weight α_i is computed as

$$\begin{aligned} m_i &= W^T h_i + b, \\ [\alpha_1, \alpha_2, \ldots, \alpha_n] &= softmax([m_1, m_2, \ldots, m_n]), \end{aligned} \tag{1}$$

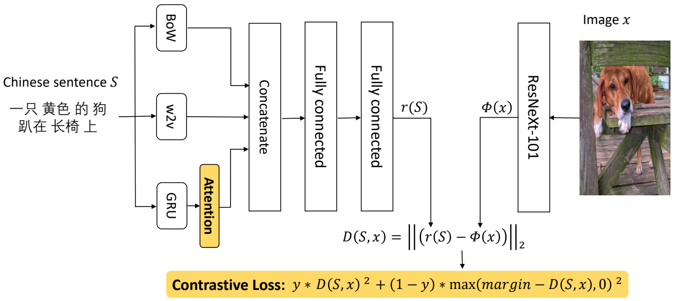


Fig. 6. **W2VV and our improvement**. We enhance W2VV [28] by 1) adding an attention layer after GRU to take into account all the hidden vectors, and 2) substituting the original mean square error for the contrastive loss. The enhanced W2VV is used for cross-lingual image retrieval.

where W and b parameterize the affine transformation within the attention layer. Accordingly we substitute h_n in W2VV for $\sum_{i=1}^{n} \alpha_i h_i$.

Second, in order to exploit the many negative pairs, we substitute the MSE loss for the Contrastive Loss [40]. Given an image x and a sentence S, we use a binary variable y to indicate their relevance, i.e., $y = 1$ if they are relevant and 0 otherwise. The new loss is

$$y \cdot D(S,x)^2 + (1-y) \cdot \max(margin - D(S,x), 0)^2, \quad (2)$$

where $D(S,x)$ indicate the Euclidean distance between the image and sentence vectors and $margin$ is a hyper-parameter, empirically set to be 2. Minimizing Eq. 2 means to minimize (maximize) the distance between relevant (negative) pairs.

We use COCO-CN to train our Chinese model that transforms a given Chinese sentence into a ResNeXt-101 feature. In a similar manner, an English version of our model is trained using the images from the same training set but with MS-COCO sentences. The cross-lingual (CL) similarity between a Chinese sentence and an English sentence is computed as the cosine similarity between their ResNeXt-101 features. Similarly, the cross-modal (CM) similarity between a Chinese sentence and an image is computed.

Given a Chinese query, image retrieval by cross-lingual matching is achieved by sorting the images in descending order according to the CL similarities of their English sentences to the query. Then, image retrieval by cross-modal matching is to sort images in descending order according to their CM similarities to the query. Finally, for image retrieval by cross-lingual and -modal (CLM) matching, the similarity between the query and a given image is computed as a convex combination of the CL and CM similarities. The combination weight is optimized on the validation set.

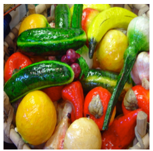
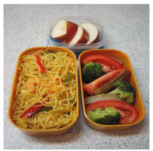
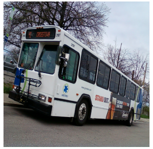
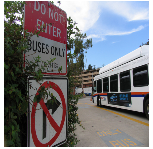
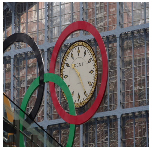
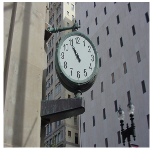
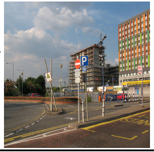
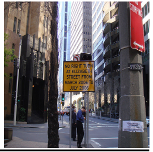
Evaluation criterion. Per query there is only one image known to be relevant with respect to the query. Hence, we report Mean Reciprocal Rank (MRR) at which the relevant item is found. Higher MRR means better performance.

2) Results: Table X shows performance of the different approaches. The Enhanced W2VV outperforms the original W2VV, justifying the effectiveness of the joint use of the attention mechanism and the contrastive loss in the new context. Image retrieval by CM is less effective than its CL counterpart. The best performance is achieved by CLM matching, scoring the highest MRR value of 0.511. This is a relative improvement of 6.2% over the original W2VV. Some qualitative results are given in Table XI. The Enhanced W2VV better models relatively complex queries that require co-occurrence of multiple visual concepts. Notice the exception in the second row, where the ending part of the query, *i.e.,* 红色环 (red ring), is more important. In this case, the shortcoming of W2VV of over-emphasizing the ending part helps.

TABLE X
PERFORMANCE OF DIFFERENT APPROACHES TO CROSS-LINGUAL IMAGE RETRIEVAL, MEASURED BY MRR.

	Matching approaches		
Model	*CL*	*CM*	*CLM*
W2VV [28]	0.450	0.378	0.481
Enhanced W2VV (*attention*)	0.466	0.412	0.499
Enhanced W2VV (*attention + contrastive loss*)	0.468	0.421	**0.511**

TABLE XI
ENHANCED W2VV VERSUS ORIGINAL W2VV FOR IMAGE RETRIEVAL BY CROSS-LINGUAL AND -MODAL (CLM) MATCHING. TEXTS IN PARENTHESES ARE ENGLISH TRANSLATIONS, PROVIDED FOR NON-CHINESE READERS AND NOT USED FOR IMAGE RETRIEVAL. WHILE W2VV TENDS TO OVER-EMPHASIZE THE ENDING WORDS OF A QUERY SENTENCE, ENHANCED W2VV WITH THE ATTENTION MECHANISM BETTER MODELS THE ENTIRE QUERY.

Chinese sentence as query	**Top hit**	
	by W2VV	*by Enhanced W2VV*
三个塑料饭盒中装着面条，蔬菜和水果 (Three plastic lunch boxes containing noodles, vegetables and fruits)		
停车场标志牌旁停着一辆公交车 (Near the sign of the parking lot parked a bus)		
一个时钟安装在一个红色环中 (A clock is installed in a red ring)		
城市建筑中间的一条街道上，竖立着一个指示牌，指示牌下站着等着过马路的男人 (On a street in the middle of city buildings, a sign is erected, under which stood men waiting to cross the road)		

E. Discussion

To justify the viability of COCO-CN for cross-lingual image tagging, captioning and retrieval, we have developed baseline methods per task. While being conceptually simple, much engineering efforts are required to properly implement these methods. For reproducible research, we have released our source code at the same github URL as the dataset.

There is much room for future exploration from an algorithmic perspective. We discuss a few recent methods related to image tagging, captioning and retrieval, albeit in a *monolingual* setting. Wang *et al.* [41] propose a novel generative model, called Bilateral Correspondence Latent Dirichlet Allocation (BC-LDA), to find conditional relationships between images and texts based on good low-dimensional representations. BC-LDA is shown to be effective for microblog image tagging. In order to leverage partially labeled data for image and video annotation, Song *et al.* [42] propose to learn an optimized graph from multiple cues. Relationships between data points are embedded into the graph more precisely. For action recognition in videos, Wang *et al.* [43] introduce two-stream 3-D ConvNet fusion that handles video clips with arbitrary size and length by spatial temporal pyramid techniques. To attack the data selection bias between the training and testing stages, Shen *et al.* [44] introduce causally regularized learning, which results in image classifiers with better generalization ability. In the context of video captioning, Gao *et al.* [2] propose a novel attention-based LSTM model with semantic consistency to explore the correlation between multi-modal representations for generating sentences with rich semantic content. A more recent work by Song *et al.* [45] improves video captioning by implementing the decoding module using multi-modal stochastic RNNs networks, modeling the uncertainty observed in the data using latent stochastic variables. For image retrieval, Cui *et al.* [46] present an interesting study on how to transfer social knowledge sensed from social media platforms for reranking and consequently personalizing image search results from generic image search engines. For content-based video retrieval at large-scale, Song *et al.* [47] propose self-supervised video hashing that simultaneously encodes video temporal and visual information using an end-to-end hierarchical binary auto-encoder and a neighborhood structure. For cross-modal retrieval, Peng *et al.* [48] propose to jointly employ coarse-grained instances and fine-grained patches within a multi-grained fusion network. Based on the observation that different modalities often contain an unequal amount of information when describing the same semantics, a modality-specific cross-modal similarity measurement approach is developed in [49], where an independent semantic space is constructed for each modality. The aforementioned algorithms have shown encouraging results for their own tasks, but all in a monolingual scenario. Repurposing and redesigning these algorithms to let them fit for the cross-ligual setting are beyond the scope of this paper.

Concerning the dataset size, we could have chosen crowd sourcing, which is actually more cheaper and more scalable, to make the numbers more impressive. However, as we noted in the paper and supported by our observations about MS-COCO annotations, crowd sourcing lacks quality control. We thus progressed with the proposed annotation approach, which produced high-quality annotations as demonstrated by the data analytics in Section III-C and the multi-task experiments in this section. Compared to the existing datasets, COCO-CN results in noticeably better models for multiple tasks.

V. Conclusions and Perspectives

The development of COCO-CN indicates that our recommendation-assisted annotation system helps reduce annotators' workload. In particular, 54 percent of the recommended tags were accepted by the annotators. Sentence recommendation is also helpful, albeit less effective than tag recommendation.

COCO-CN has enriched MS-COCO, not just in terms of the language. Compared to the 80 MS-COCO object classes, there are 585 novel concepts. On average a COCO-CN sentence brings in 45.3% keywords not covered by machine translation of MS-COCO sentences. While the number of its images seems to be small, its diverse visual content associated with high-quality annotations makes the dataset superior to Flickr8k-CN and AIC-ICC. Compared to NUS-WIDE, a popular multi-label dataset that has 81 labels and 2.4 labels per image, COCO-CN provides 655 labels with 4.4 labels per image. With the rich and (unaligned) bilingual tags and sentences, COCO-CN is unique to the community, providing a unified and challenging platform for advancing image annotation and retrieval in the cross-lingual direction.

Even though the ground truth of the test set is incomplete, it is adequate for drawing qualitative conclusions about which model is better for cross-lingual image tagging. Cascading MLP outperforms MLP trained on the monolingual resources, multi-task MLP trained on bi-lingual resources, and Clarifai, a commercial image tagging service. Our sequential learning strategy, while being a common practice and thus not new in the monolingual setting, is the first solution that has successfully combined machine-translated and manual annotations for cross-lingual image captioning. The enhanced W2VV establishes a new baseline for cross-lingual image retrieval. COCO-CN, provided with these good baseline methods, opens up interesting avenues for future research on cross-lingual multimedia tasks.

Acknowledgments. The authors thank all six anonymous reviewers for their insightful comments and Miss Xinru Chen for performing typos check on COCO-CN sentences.

Appendix

Choice of Image Features. As shown in Table XII, for all the three tasks models using the ResNeXt-101 feature perform the best, when compared to their counterparts using either the GoogLeNet or ResNet-152 feature.

Image Captioning with Tags. As we have summarized in Table II, a novel property of COCO-CN is that its images are annotated with tags, in addition to sentence descriptions. We conduct an extra experiment, termed image captioning with tags, to demonstrate that the tags can complement the image captioning task. The image captioning model previously trained on COCO-CN is re-used as a baseline. To construct a multi-modal input for a given image, we represent tags associated with this image by a bag-of-words (BoW) feature vector. Each dimension corresponds a specific tag, with its value set to one if the tag is present and zero otherwise. A multi-modal input is formed by concatenating the ResNeXt-101 feature and the BoW feature. As Table XIII shows, image captioning with tags performs better.

TABLE XII

Comparing three CNN features. For a fair comparison, all the features are evaluated in the same setting, *i.e.,* MLP trained on COCO-CN for image tagging, the Show and Tell model trained on COCO-CN for image captioning, and the enhanced W2VV for image retrieval. Models using the ResNeXt-101 feature are the best.

Image Feature	**Image Tagging**			**Image Captioning**				**Image Retrieval**		
	Precision	*Recall*	*F-measure*	*Bleu-4*	*METEOR*	*ROUGE-L*	*CIDEr*	*CL*	*CM*	*CLM*
GoogLeNet	0.394	0.473	0.414	26.1	27.3	49.6	73.5	0.458	0.336	0.463
ResNet-152	0.407	0.496	0.430	26.5	27.3	49.8	74.4	0.450	0.362	0.479
ResNeXt-101	**0.477**	**0.576**	**0.503**	**31.7**	**27.2**	**52.0**	**84.6**	**0.468**	**0.421**	**0.511**

TABLE XIII

Automated evaluation of image captioning with and without tags. COCO-CN is used as training data. Including tags as input gives better performance.

Input	**BLEU-4**	**METEOR**	**ROUGE-L**	**CIDEr**
Image	**31.7**	27.2	52.0	84.6
Image + tags	31.3	**30.1**	**53.2**	**90.0**

References

[1] P. Cui, S. Liu, and W. Zhu, "General knowledge embedded image representation learning," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 198–207, 2018.

[2] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Trans. Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *CVPR*, 2009.

[4] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *CIVR*, 2009.

[5] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *CoRR*, vol. abs/1504.00325, 2015.

[6] Y. Hu, L. Zheng, Y. Yang, and Y. Huang, "Twitter100k: A real-world dataset for weakly supervised cross-media retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 4, pp. 927–938, 2018.

[7] M. Grubinger, P. Clough, H. Müller, and T. Deselaers, "The IAPR benchmark: A new evaluation resource for visual information systems," in *ICLRE*, 2006.

[8] S. Tsutsui and D. Crandall, "Using artificial tokens to control languages for multilingual image caption generation," in *CVPR Language and Vision Workshop*, 2017.

[9] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual English-German image descriptions," in *ACL Workshop on Vision and Language*, 2016.

[10] T. Miyazaki and N. Shimizu, "Cross-lingual image caption generation," in *ACL*, 2016.

[11] J. Hitschler, S. Schamoni, and S. Riezler, "Multimodal pivots for image caption translation," in *ACL*, 2016.

[12] J. Rajendran, M. Khapra, S. Chandar, and B. Ravindran, "Bridge correlational neural networks for multilingual multimodal representation learning," in *NAACL-HLT*, 2016.

[13] Q. Wei, X. Wang, and X. Li, "Harvesting deep models for cross-lingual image annotation," in *CBMI*, 2017, pp. 36:1–36:5.

[14] W. Lan, X. Li, and J. Dong, "Fluency-guided cross-lingual image captioning," in *ACMMM*, 2017.

[15] D. Elliott and Á. Kádár, "Imagination improves multimodal translation," in *IJCNLP*, 2017.

[16] R. Funaki and H. Nakayama, "Image-mediated learning for zero-shot cross-lingual document retrieval." in *EMNLP*, 2015.

[17] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "STAIR captions: Constructing a large-scale japanese image caption dataset," in *ACL*, 2017.

[18] X. Li, W. Lan, J. Dong, and H. Liu, "Adding Chinese captions to images," in *ICMR*, 2016.

[19] J. Wu, H. Zheng, B. Zhao, Y. Li, B. Yan, R. Liang, W. Wang, S. Zhou, G. Lin, Y. Fu, Y. Wang, and W. Y., "AI challenger : A large-scale dataset for going deeper in image understanding," *CoRR*, vol. abs/1711.06475, 2017.

[20] B. Caputo, H. Müller, B. Thomee, M. Villegas, R. Paredes, D. Zellhofer, H. Goeau, A. Joly, P. Bonnet, J. Gomez, I. Varea, and M. Cazorla, "ImageCLEF 2013: The vision, the data and the open challenges," in *CLEF*, 2013.

[21] Clarifai, "Clarifai image and video recognition API," https://clarifai.com/demo, 2018.

[22] S. Gella, R. Sennrich, F. Keller, and M. Lapata, "Image pivoting for learning multilingual multimodal representations," in *EMNLP*, 2017.

[23] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. Snoek, and A. Del Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement and retrieval," *CSUR*, vol. 49, no. 1, pp. 14:1–14:39, 2016.

[24] J. Fu and Y. Rui, "Advances in deep learning approaches for image tagging," *APSIPA Transactions on Signal and Information Processing*, vol. 6, 2017.

[25] X. Li, B. Shen, B.-D. Liu, and Y.-J. Zhang, "Ranking-preserving low-rank factorization for image annotation with missing labels," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1169–1178, 2018.

[26] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks and challenges," *TCSVT*, 2017, in press.

[27] L. Zhang, B. Ma, G. Li, Q. Huang, and Q. Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 128–141, 2018.

[28] J. Dong, X. Li, and C. Snoek, "Predicting visual features from text for image and video caption retrieval," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3377–3388, 2018.

[29] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's Mechanical Turk," in *NAACL-HLT Workshop*, 2010.

[30] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *TACL*, vol. 2, pp. 67–78, 2014.

[31] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *JAIR*, vol. 47, pp. 853–899, 2013.

[32] A. Koochali, S. Kalkowski, A. Dengel, D. Borth, and C. Schulze, "Which languages do people speak on Flickr? a language and geo-location study of the YFCC100m dataset," in *MMCommons*, 2016.

[33] J. Yao, Y. Wang, Y. Zhang, J. Sun, and J. Zhou, "Joint latent dirichlet allocation for social tags," *IEEE Trans. Multimedia*, vol. 20, no. 1, pp. 224–237, 2018.

[34] P. Mettes, D. Koelma, and C. Snoek, "The ImageNet shuffle: Reorganized pre-training for video event detection," in *ICMR*, 2016.

[35] K. Min, C. Ma, T. Zhao, and H. Li, "BosonNLP: An ensemble approach for word segmentation and POS tagging," in *NLPCC*, 2015.

[36] V. I. Levenshtein, "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviet Physics Doklady*, vol. 10, p. 707, Feb. 1966.

[37] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017.

[38] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015.

[39] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *ICLR*, 2015.

[40] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, 2006.

[41] Z. Wang, P. Cui, L. Xie, W. Zhu, Y. Rui, and S. Yang, "Bilateral correspondence model for words-and-pictures association in multimedia-rich microblogs," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 10, no. 4, pp. 34:1–34:21, 2014.

[42] J. Song, L. Gao, F. Nie, H. T. Shen, Y. Yan, and N. Sebe, "Optimized graph learning using partial tags and multiple features for image and

video annotation," *IEEE Trans. Image Processing*, vol. 25, no. 11, pp. 4999–5011, 2016.

[43] X. Wang, L. Gao, P. Wang, X. Sun, and X. Liu, "Two-stream 3-d convnet fusion for action recognition in videos with arbitrary size and length," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 634–644, 2018.

[44] Z. Shen, P. Cui, K. Kuang, B. Li, and P. Chen, "Causally regularized learning with agnostic data selection bias," in *ACMMM*, 2018, pp. 441–419.

[45] J. Song, H. Zhang, X. Li, L. Gao, M. Wang, and R. Hong, "From deterministic to generative: Multi-modal stochastic rnns for video captioning," *IEEE Trans. Neural Networks and Learning Systems*, 2018.

[46] P. Cui, S.-W. Liu, W.-W. Zhu, H.-B. Luan, T.-S. Chua, and S.-Q. Yang, "Social-sensed image search," *ACM Trans. Inf. Syst.*, vol. 32, no. 2, pp. 8:1–8:23, 2014.

[47] J. Song, H. Zhang, X. Li, L. Gao, M. Wang, and R. Hong, "Self-supervised video hashing with hierarchical binary auto-encoder," *IEEE Trans. Image Processing*, vol. 27, no. 7, pp. 3210–3221, 2018.

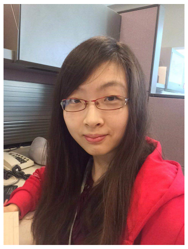
[48] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 405–420, 2018.

[49] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Trans. Image Processing*, vol. 27, no. 11, pp. 5585–5599, 2018.

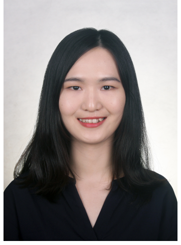
Xirong Li received the B.S. and M.E. degrees from Tsinghua University, Beijing, China, in 2005 and 2007, respectively, and the Ph.D. degree from the University of Amsterdam, Amsterdam, The Netherlands, in 2012, all in computer science. He is currently an Associate Professor with the Key Lab of Data Engineering and Knowledge Engineering, Renmin University of China, Beijing, China. His research includes image and video retrieval.

Prof. Li was recipient of the ACMMM 2016 Grand Challenge Award, the ACM SIGMM Best Ph.D. Thesis Award 2013, the IEEE TRANSACTIONS ON MULTIMEDIA Prize Paper Award 2012, and the Best Paper Award of ACM CIVR 2010. He was area chair of ACMMM 2018 and ICPR 2016.

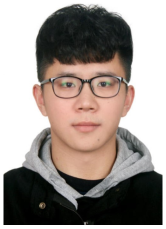
Chaoxi Xu received his B.S. degree in Computer Science from Renmin University of China, Beijing, China in 2017. He is currently a graduate student at School of Information, Renmin University of China, pursuing his master degree on cross-lingual multimedia retrieval.



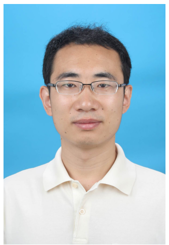
Xiaoxu Wang received her B.S. and M.E. degrees in Computer Science from Renmin University of China, Beijing, China, in 2015 and 2018, respectively. She is currently a software engineer at the Bank of China, Beijing, China.



Weiyu Lan received her B.S. and M.E. degrees in Computer Science from Renmin University of China, Beijing, China, in 2015 and 2018, respectively. She is currently a researcher at Tencent, Shenzhen, China.



Zhengxiong Jia received his B.S. degree in Computer Science from Renmin University of China, Beijing, China, in 2018. He is currently a graduate student at School of Information, Renmin University of China, pursuing his master degree on image captioning.



Gang Yang received his Ph.D. degree in Innovative Life Science from University of Toyama, Toyama, Japan in 2009. He is currently an Associate Professor at School of Information, Renmin University of China, Beijing, China. His current research interests include computational intelligence, multimedia computing and machine learning.

Jieping Xu received her Ph.D. degree in Computer Science from the Institute of Acoustics, Chinese Academy of Sciences in 1999. She is currently an Associate Professor at School of Information, Renmin University of China, Beijing, China. Her research interests include multimedia computing and retrieval.