Fine-Grained Label Learning via Siamese Network for Cross-modal Information Retrieval

Yiming Xu[1,3], Jing Yu[1,2(✉)], Jingjing Guo[1,2], Yue Hu[1(✉)], and Jianlong Tan[1]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
yiming.tsui@gmail.com,{guojingjing,huyue,tanjianlong}@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
yujing02@iie.ac.cn
[3] School of Computer and Information Engineering, Henan University, Kaifeng, China

Abstract. Cross-modal information retrieval aims to search for semantically relevant data from various modalities when given a query from one modality. For text-image retrieval, a common solution is to map texts and images into a common semantic space and measure their similarity directly. Both the positive and negative examples are used for common semantic space learning. Existing work treats the positive/negative text-image pairs as equally positive/negative. However, we observe that many positive examples resemble the negative ones in some degrees and vice versa. These "hard examples" are challenging for existing models. In this paper, we aim to assign fine-grained labels for the examples to capture the degrees of "hardness", thus enhancing cross-modal correlation learning. Specifically, we propose a siamese network on both the positive and negative examples to obtain their semantic similarities. For each positive/negative example, we use the text description of the image in the example to calculate its similarity with the text in the example. Based on these similarities, we assign fine-grained labels to both the positives and negatives and introduce these labels to a pairwise similarity loss function. The loss function benefits from the labels to increase the influence of hard examples on the similarity learning while maximizing the similarity of relevant text-image pairs and minimizing the similarity of irrelevant pairs. We conduct extensive experiments on the English Wikipedia, Chinese Wikipedia, and TVGraz datasets. Compared with state-of-the-art models, our model achieves significant improvement on the retrieval performance by incorporating with fine-grained labels.

Keywords: Fine-grained labeling · Siamese network · Graph convolutional network · Hard examples · Cross-modal information retrieval

This work is supported by the National Key Research and Development Program (Grant No. 2017YFB0803301) and the Innovation Program for College Students by Chinese Academy of Sciences (Grant No. Y8YY261101).

J. M. F. Rodrigues et al. (Eds.): ICCS 2019, LNCS 11537, pp. 304–317, 2019.
https://doi.org/10.1007/978-3-030-22741-8_22

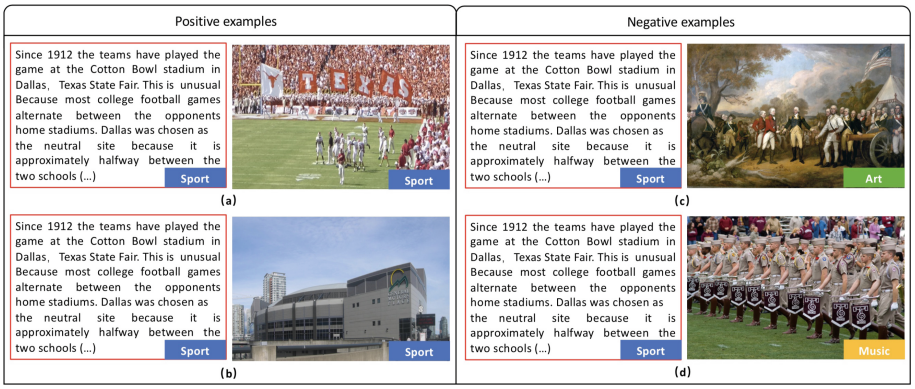

Fig. 1. Illustration of the examples of the positive/negative text-image pairs for cross-modal correlation learning. They are not equally positive/negative.

1 Introduction

In recent decades, online heterogeneous data of different modalities, such as text, image, video and audio, have been accumulated in huge quantity. There is a great need to find semantically relevant information in various modalities. Traditional single-modal search engines retrieve data of the same modality as the query, such as text-based retrieval or content-based image retrieval. With the development of natural language processing (NLP) and computer vision (CV), cross-modal information retrieval (CMIR), which supports retrieval across multi-modal data, brings increasing attention to break the boundary of natural language and vision. The main challenge of CMIR is to bridge the "heterogenous gap" and measure the semantic similarity between different modalities.

The typical solution for CMIR is to learn a common semantic space and project the features of each modality to this space for similarity measurement. The basic paradigm is based on statistical correlation analysis, which learns linear projection matrices by maximizing the statistical correlations of the positive cross-modal pairs. With the advance of deep learning, deep neural network is used for common space learning, which shows great ability of learning the non-linear correlations across modalities. Typically, two subnetworks model positive and negative text-image pairs simultaneously and optimize their common representations by matched and unmatched constrains. Though the progress in the cross-modal retrieval performance, most of existing algorithms treat positive/negative examples as equally positive/negative and ignore their difference in the degrees of "positivity/negativity".

In this paper, we focus on cross-modal information retrieval by exploring fine-grained labels. We observe that many positive examples resemble the negative ones in some degrees and vice versa. To exemplify the above issue, we give an example using the text-image pairs about *Sport* in Fig. 1. The positive text-image pair in Fig. 1(a), which contains an image and a text describing *football*

match, is quite clear to be positive. However, it is very difficult to judge the example in Fig. 1(b) as a positive one, since the *gym* in the image is not definitely belonging to the category of *Sport* according to the visual features. The similar observation exists in the negative examples. The example in Fig. 1(c) consists a text about *football match* and an image of painting, which is quite negative. In contrast, it's hard to distinguish the negative example in Fig. 1(d) from the positive one in Fig. 1(a), since they share many attributes in the images, such as grasses, flags, and people, while having the same content in the texts. As illustrated in Fig. 1, the positive/negative examples are not always equally positive/negative. Assigning examples with labels in different degrees, named as fine-grained labels, will capture more informative characteristics for cross-modal information retrieval.

Our main contribution is to propose a **F**ine-**G**rained **Lab**el learning approach (**FGLab** for short) for cross-modal information retrieval. Our approach first leverage the text description of the image in each text-image example to represent the semantics of the image. Then we propose a siamese network on both the positive and negative examples to obtain their semantic similarities and assign the fine-grained labels accordingly. Finally, we incorporate these labels to a pairwise similarity loss function, which enables the model to pay more attention on the hard examples while maximizing the similarity of positive examples and minimizing the similarity of negative examples. Our proposed approach could be easily applied to other existing models and provide more informative cues for cross-modal correlation learning.

2 Related Work

Cross-modal Information Retrieval. The mainstream solution for CMIR is to project data from different modalities into a common semantic space where the similarity of different modalities data can be measured directly. Traditional statistical correlation analysis methods like Canonical correlation analysis (CCA) [4,5] learn a common semantic space to maximize the pairwise correlation between two sets of heterogeneous data. With the development of deep learning, the DNN-based CMIR methods have attracted much attention due to its strong learning ability. This method usually constructs two subnets to extract the features of different modal data, and the inputs of different media types learn the common semantic space through the shared layer [12,19,23]. In this work, we follow this two-path deep model to learn the common semantic space of the text and image.

Study on Hard Examples. Hard examples analysis can help classification tasks and retrieval tasks obtain better results [1,3,11,24]. Dalal *et al.* [1] construct a fixed set of an initial negative set, train the preliminary detector by using the original dataset, then use the preliminary detector to search the hard examples in the initial negative set. After that, add the hard examples to the original dataset, which was used to re-train the detector. Although the Dalal

et al. method improves the accuracy of the detector, the expanded data set creates more memory and time overhead, which makes the training efficiency of the model declining. In order to solve this problem, Felzenszwalb *et al.* [3] proposed a method of deleting simple examples with high classification accuracy while adding hard examples, and effectively controlling the size of the data set while further improving the accuracy. However, both of the above methods require repeated training of the entire data set to update the model. In order to improve this method, Shrivastava *et al.* [17] proposed an online hard examples mining method. In the training, the negative examples in each batch are sorted, and the k hardest examples are selected for the back propagation to update the model. The above works focus on the mining of hard negative examples. On this basis, Jin *et al.* [6] proposed the view of hard positive, and mining the hard negative and hard positive examples in the video target detection task, which further improved the training efficiency.

These methods focus on updating the model with hard examples, which effectively improves the accuracy of the model. However, different hard examples are not equally "hard", in order to capture more informative cues from the hard examples, Ma *et al.* [10] assign fine-grained labels to hard negative videos based on the hard degree of negative examples, and obtain better detection results in the complex event detection task in the video. Recently, hard examples have begun to receive attention in cross-modal information retrieval. Faghri *et al.* [2] use the hard negative examples in cross-modal information retrieval to converge the model to better results, which significantly improves the retrieval performance. Different for their work, in this paper, we extend the idea of hard negatives to positive examples and assign fine-grained labels to the different degree of hard examples. For text-image pairs in cross-modal information retrieval, we extract the original text information of the image to calculate the similarity between the original text and the text in the examples. Then we assign fine-grained labels to the examples. We modify the loss function of the distance metric learning and add the fine-tuning effect of the fine-grained label, which increases the influence of the hard positive and negative examples.

3 Methodology

Our model consists of two stages, as shown in Fig. 2. The first stage is a fine-grained scoring model based on the similarity degree of text, and the second stage is a cross-modal information retrieval model. In the first stage, the main objective is measuring the correlation between the text in text-image pairs and the text description of images. Compared with the image, the text description often contains richer and more specific information. Therefore, we select the corresponding text description of the image in the example. We use three different feature extraction methods to modeling text, including GCN, Multi-head attention, and Text-CNN and assigning a fine-grained label to the example according to the similarity degree. In the second stage, we use the fine-grained label to adjust the cross-modal information retrieval model.

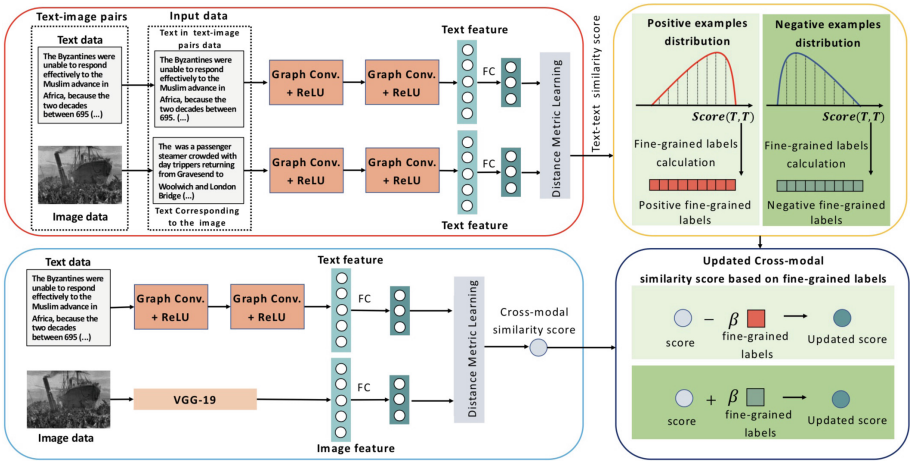

Fig. 2. The proposed model overview is divided into two parts. The top part is the fine-grained label learning based on text similarity evaluation, which includes text feature extraction (red box) and fine-grained labels assign (yellow box). The bottom part is the cross-modal information retrieval model, fine-grained labels play an important role in the model update. (Color figure online)

3.1 Fine-Grained Label Learning

In the first stage, the main objective is to measure the difficulty degree of the text-image pairs and generate fine-grained labels for them. We evaluate the difficulty degree of text-image pairs by the similarity degree of text-image pairs. For positive examples, the smaller the similarity score of text-image pairs, the greater the difficulty degree. For negative examples, the greater the similarity score of text-image pairs, the greater the difficulty degree. Compared with the image, the text usually contains richer and more specific information. Therefore, we calculate the similarity of text-image pairs by selecting the original text description of the image and the text in the example. We use three different feature extraction methods to modeling text and assigning a fine-grained label to the example.

We design a dual-path neural network to extract the text features and text features and learn the potential common semantic space. Define the original dataset as $D = \{(T_i^D, I_i^D)\}_{i=1}^N$, which contains of C classes. Where T_i^D represents the i_{th} text in the original dataset, I_i^D represents the i_{th} image, and (T_i^D, I_i^D) represents the i_{th} text-image pair in the original dataset D. T_i^D is the original text description of the image I_i^D, which have the same semantic and belong to the same class.

Follow the previous work [25], we construct a positive examples dataset $P = \{(T_j^P, I_j^P)\}_{j=1}^M$ and a negative examples $E = \{(T_k^E, I_k^E)\}_{k=1}^K$ dataset based on the original training data D. Specifically, we randomly select T_i^D and I_j^D of the same class from D to constitute the positive examples dataset P, where M represents

the number of text-image pairs in P. Similarly, for E, we randomly select T_x^D and I_k^D which do not belong to the same class from D, where K represents the number of text-image pairs in N. To ensure the same number of positive and negative examples for the model training, we set $M = K$.

We focus on the learning difficulty of different text-image pairs. For positive examples, the closer the semantics of text and image, the lower the learning difficulty, and the larger the semantic difference, the higher the learning difficulty. For negative examples, the closer the semantics of text and image, the higher the learning difficulty, and the larger the semantic difference, the lower the learning difficulty. Compared with the image, the original text description contains richer and more specific information, which can express high-level semantics so that we will learn the difficulty of positive and negative examples through the semantic similarity between texts. Specifically, for all image I_j^P or I_k^E, we extract the original text descriptions T_j^D or T_k^D corresponding to the image in D, forming a positive text-text pair (T_j^P, T_i^D) or the negative text-text pairs (T_k^E, T_i^D). We use an end-to-end dual-path neural network to learn the similarity between the texts. The output of the network is text-text similarity score S, where $S^{Pos}(T_j^P, T_j^D)$ indicates the similarity score of the positive example, $S^{Neg}(T_k^E, T_k^D)$ indicates the similarity score of the negative example. The loss function is formal as:

$$Loss = (\sigma_{T-T}^{2+} + \sigma_{T-T}^{2-}) + \lambda max(0, m - (\mu_{T-T}^{+} - \mu_{T-T}^{-})) \tag{1}$$

Where μ_{T-T}^{+} and σ_{T-T}^{2+} as the mean and variance of the associated text pairs, and μ_{T-T}^{-} and σ_{T-T}^{2-} denote the mean and variance of the unrelated text pairs.

Given positive examples dataset $P = \{(T_j^P, I_j^P)\}_{j=1}^{M}$, we can obtain the positive examples similarity score dataset C^{Pos}. Similarly, given negative examples dataset $E = \{(T_k^E, I_k^E)\}_{k=1}^{K}$, we can obtain the negative examples similarity score dataset C^{Neg}. We assign fine-grained labels L for positive examples and negative examples. The fine-grained labels are allocated from 0 and 1. Given a text-text similarity score S_i, the fine-grained label L_i is defined as follows:

$$L_i^{Pos}(T_i^P, I_i^P) = 1 - \frac{S_i(T_i^P, T_j^D) - S_{min}^{Pos}}{S_{max}^{Pos} - S_{min}^{Pos}} \tag{2}$$

$$L_i^{Neg}(T_i^E, I_i^E) = \frac{S_i(T_i^E, T_j^D) - S_{min}^{Neg}}{S_{max}^{Neg} - S_{min}^{Neg}} \tag{3}$$

where L_i denotes the fine-grained label of i_{th} text-image pair, S_{max}^{Pos} and S_{min}^{Pos} denotes the maximum and the minimum similarity score in positive examples similarity score dataset C^{Pos}, S_{max}^{Neg} and S_{min}^{Neg} denotes the maximum and the minimum similarity score in C^{Neg}.

3.2 Cross-modal Information Retrieval Model

In this stage, we build a dual-path neural network to extract text and image features and learn the potential common semantic space. Then we attain the

similarity of text and image by metric learning. Fine-grained labels are used in the model training process to update the similarity score and to further adjust the loss function. Our model makes hard examples have a greater impact in the model training.

Text and Image Feature Extraction. The GCN model has a strong ability to learn the local and fixed features of the graph and has been successfully used for text classification [9]. In recent research [25], GCN has shown to have a strong capability for text semantic modeling and text categorization. In our model, the text GCN contains two convolution layers each followed by a ReLU. Then, we set up a fully connected layer to map the text features to the common latent semantic space. Given a text T, the text feature v_T can be extracted by the text GCN model $H_T(\cdot)$, which is defined as: $v_T = H_T(T)$.

For image modeling, we use the pre-trained VGG-19 [18] as the basic model to obtain image features. Given a 224×224 image, a 4096-dimensional feature vector is generated from the FC7 layer, which is the penultimate fully connected layer in VGG-19. Next, a fully connected layer map the image to the common semantic space. Given a image I, image vector v_I is extracted by the VGG-19 model $H_I(\cdot)$, which is defined as: $v_I = H_I(I)$.

Object Function. In our model, we set up two paths to obtain the text features v_T and images features v_I. Then we use element-wise product to attain the correlation between the text features and the image features, and a fully connected layer is followed to obtain the similarity score. We use the same loss function with Sect. 3.1, which aims to reduce the proportion of false positives and false negatives, as shown in Fig. 3(a). The left curve represents the distribution of the matched text-image pairs where μ^- denotes the mean and σ^{2-} demotes the variance. The right curve represents the distribution of non-matching text-image pairs where μ^+ denotes the mean and σ^{2+} demotes the variance. The objective function is to maximize μ^+ and minimize μ^-, σ^{2+} and σ^{2-}. In our work, our goal is further reducing the proportion of false positives and false negatives by enhancing the impact of the hard examples in the shadows to make the model converge to better results, which is shown in Fig. 3(b).

In the training process, the distribution of a few of false positives and false negatives in the shaded portion is updated to a more erroneous degree by the fine-grained label, so that the influence of these hard examples is increased, and the model obtains a better descending gradient. Given text features v_T and images features v_I, the similarity of the matched and non-matched text-image pairs is defined as $Y(T, I)$, we use fine-grained labels to update the similarity score of text-image pairs, which is formulated as follow:

$$\widetilde{Y}^{Pos}(T^P, I^P) = Y(T^P, I^P) - \beta L^{Pos}(T^P, I^P) \tag{4}$$

$$\widetilde{Y}^{Neg}(T^E, I^E) = Y(T^E, I^E) - \beta L^{Neg}(T^E, I^E) \tag{5}$$

Where $\widetilde{Y}^{Pos}$ and $\widetilde{Y}^{Neg}$ represents the similarity score after the fine-grained label update for positive and negative examples. β is a hyperparameter that adjusts

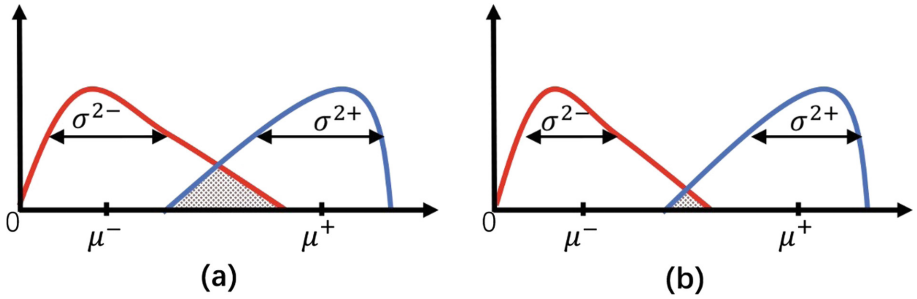

Fig. 3. (a) The original loss function and (b) the upgraded loss function by our fine-grained labels.

the effect of fine-grained labels on similarity scores. The value of β is related to the range of Y. The loss function is defined as follows:

$$Loss = (\sigma^{2+}_{T-I}\sigma^{2-}_{T-I}) + \lambda max(0, m - (\mu^{+}_{T-I} - \mu^{-}_{T-I})) \tag{6}$$

$$\mu^{+}_{T-I} = \sum_{n=1}^{Q_1} \frac{\widetilde{Y}^{Pos}}{Q_1}, \qquad \sigma^{2+}_{T-I} = \sum_{n=1}^{Q_1} \frac{\widetilde{Y}^{Pos} - \mu^{+}_{T-I}}{Q_1}$$
$$\mu^{-}_{T-I} = \sum_{n=1}^{Q_2} \frac{\widetilde{Y}^{Neg}}{Q_2}, \qquad \sigma^{2-}_{T-I} = \sum_{n=1}^{Q_2} \frac{\widetilde{Y}^{Neg} - \mu^{-}_{T-I}}{Q_2}$$

Where μ^{+}_{T-I} and σ^{2+}_{T-I} denote the mean and variance of the matched text and image, and μ^{-}_{T-I} and σ^{2-}_{T-I} are the mean and variance of the non-matched text and image. λ can adjust the ratio of the mean, and m controls the upper limit between the average of the matching and non-matching similarities.

4 Experiments

4.1 Datasets and Evaluation

We accomplish the general cross-modal information retrieval tasks: image-query-texts and text-query-images. We evaluate our model on three benchmark datasets, i.e. English Wikipedia, TVGraz, and Chinese Wikipedia. Each dataset contains a set of text-image pairs, where the texts are long descriptions with rich content instead of tags or captions.

English Wikipedia. The English Wikipedia dataset was divided into 10 categories, containing 2866 image-text pairs. We selected 2173 pairs for training and 693 pairs for testing. Each text was represented by a graph containing 10055 vertices and each image by a 4096-dimensional vector representation of the last layer of the fully connected layer of the VGG-19 model [18].

TVGraz. The TVGraz dataset contained 10 categories from the Caltech-256 [13] dataset, stored in an URL format. We selected more than 10 words of text from 2592 web pages, which comprised 2360 image-text pairs; they were randomly divided into 1885 for training and 475 pairs for testing. Each text was represented by a graph containing 8172 vertices, and each image by a 4096-dimensional VGG-19 feature.

Chinese Wikipedia. Chinese Wikipedia dataset (Ch-Wiki for short) The Chinese Wikipedia dataset [14] was divided into 9 categories, containing 3103 image-text pairs. We randomly selected 2482 pairs for training and 621 pairs for testing. Each text was represented by a graph with 9613 vertices, and each image by a 4096-dimensional VGG-19 feature.

Mean Average Precision (MAP) is used to evaluate our model. MAP is the mean of average precision (AP) for all queries. AP is defined as:

$$AP = \frac{1}{R} \sum_{k=1}^{n} \frac{R_k}{k} \times rel_k \tag{7}$$

where R represents the number of relevant retrieved results. R_k represents the top k results. n represents the number of all the retrieved results. $rel_k = 1$ indicates that the k_{th} result is related, and otherwise 0.

4.2 Implementation Details

Our model consists of two stages, i.e. fine-grained labeling model and cross-modal information retrieval model. We follow the strategy in [25] to randomly select 40,000 positive examples and 40,000 negative examples from the training set in all the datasets. We set the learning rate 0.001 with an Adam optimization, and 50 epochs for training. The regularization is 0.005. m and λ in the loss function are set as 0.8 and 0.35, respectively. In the fine-grained labeling model, the dimension of text features from the two paths is reduced to 1024 after the fully connected layer. Similarly, in the cross-modal information retrieval model, the dimension of both image features and the text features are reduced to 1024 before feeding to the feature fusion module.

4.3 Comparison with State-of-the-Art Methods

We compare the performance of our proposed FGLab with state-of-the-art models, including CCA [16], LCFS [22], ml-CCA [15], LGCFL [7], AUSL [26], JFSSL [21], GIN [25], TTI [14], CM [13], SM [13], SCM, TCM, w-TCM, c-TCM. All the models are well cited in the literature. GIN also serves as the baseline model, which has the same architecture as our cross-modal retrieval model without fine-grained labels in the loss function. We compare the MAP scores with their publicly reported results in Table 1.

From the performance in Table 1, we observe that our model is superior to all the other models for the text queries on the three datasets. Compared with the second best method GIN (baseline), the performance of our method in MAP is

Table 1. MAP score comparison of text-image retrieval on three datasets.

Dataset	Model	Retrieval performance		
		Text query	Image query	Average
Eng-Wiki	CCA	0.187	0.216	0.201
	ml-CCA	0.287	0.352	0.312
	LCFS	0.231	0.297	0.264
	LGCFL	0.316	0.377	0.312
	AUSL	0.332	0.396	0.364
	JFSSL	0.410	**0.467**	0.438
	GIN	0.767	0.452	0.609
	FGLab (ours)	**0.837**	0.457	**0.647**
TVGraz	TTI	0.153	0.216	0.184
	CM	0.450	0.460	0.4550
	SM	0.585	0.619	0.602
	SCM	0.696	0.693	0.694
	TCM	0.706	0.694	0.695
	GIN	0.719	0.818	0.769
	FGLab (ours)	**0.763**	**0.833**	**0.798**
Ch-Wiki	w-TCM	0.298	0.241	0.269
	c-TCM	0.317	0.310	0.313
	GIN	0.384	0.334	0.359
	FGLab (ours)	**0.517**	**0.390**	**0.457**

increased by about 7%, 5%, and 13% on the Eng-Wiki, TVGraz, and Ch-Wiki, respectively. For the image query, FGLab outperforms state-of-the-art models by 1.5% and 5.6% on the TVGraz and Ch-Wiki, respectively. The results on the Eng-Wiki dataset is slightly inferior to the second best model JFSSL. Compare with the image query, the improvement on the text query is more remarkable. It's because that when the fine-grained labels enable the model to pay more attention on the "hard" examples, the parameter tuning of the model has bias on the text modeling path. Specifically, the parameters in the text modeling path are tuned more greater than these parameters in the image modeling path, which enhances the generalization ability of the text representations obviously compared to the image. For the average performance, our model is superior to all the other models. Compared with GIN, the MAP scores is increased by about 4%, 3%, and 10% on the Eng-Wiki, TVGraz, and Ch-Wiki, respectively. It proves that, by incorporated with the "hardness" information of the training examples by fine-grained labels, existing model (GIN) can achieve great improvements on the cross-modal retrieval performance. Meanwhile, the remarkable improvements indicate that our proposed FGLab is able to capture the informative clues of

Table 2. MAP score comparison of FGLab models with different kinds of fine-grained labels on the Eng-wiki dataset.

Model	Text feature	Text query	Image query	Average
FGLab (main model)	GCN	**0.837**	**0.457**	**0.647**
FGLab-Att	Multi-head attention	0.806	0.438	0.622
FGLab-CNN	Text-CNN	0.7501	0.437	0.5931

fine-grained labels and effectively affect the cross-modal correlation learning to focus on "hard" examples.

4.4 The Influence of Text Features on Fine-Grained Labeling

Besides our proposed fine-grained labeling model based on GCN, we also implement another two baseline models based on different text features to evaluate their influence on the fine-grained labeling as well as the cross-modal retrieval performance. The other two models are respectively utilize the multi-head attention [20] (i.e. the encoder in transformer) and Text-CNN [8] instead of GCN in the fine-grained labeling model, with other structures unchanged. The two models are respectively named as FGLab-Att and FGLab-CNN for short.

Figure 4 shows the samples of the fine-grained labels obtained by the three models on the Eng-Wiki dataset. The bar graphs on the top or bottom of each image show the fine-grained labels of the three models using different text features for each positive or negative example. The text is identical for all the examples while the images are different. Because of the space limitation, we only show a part of the whole text, which introduces a war happened on the sea. For clear semantic comparisons, we show the corresponding image of the text on the middle left. For the three positive examples, the images in the text-image pairs from left to right contains *map*, *bird's-eye view of ships*, and the *close shot of a warship*, respectively. Intuitively, the semantic correspondence of the three images with the text increases from left to right while the "hardness" of the three positive examples decrease accordingly. The fine-grained labels obtained by GCN features are identical to human judgement. The results of multi-head attention and Text-CNN are not satisfied to some extent. Similar observations exist for the negative examples and other examples not shown in the paper. Therefore, from the qualitative analysis of the fine-grained labels, the labeling model based on GCN could obtain more accurate labels by human evaluation.

To further prove the effect of fine-grained labels by different text features, we train the retrieval models based on the aforementioned three kinds of labels on the Eng-Wiki dataset. The retrieval results is given in Table 2. It's obvious that FGLab with GCN achieves the highest MAP scores compared the other two models. The performance of FGLab-Att is slightly lower than that of FGLab while FGLab-CNN has the worst performance. The retrieval results of the three models are identical with their performance in fine-grained labeling.

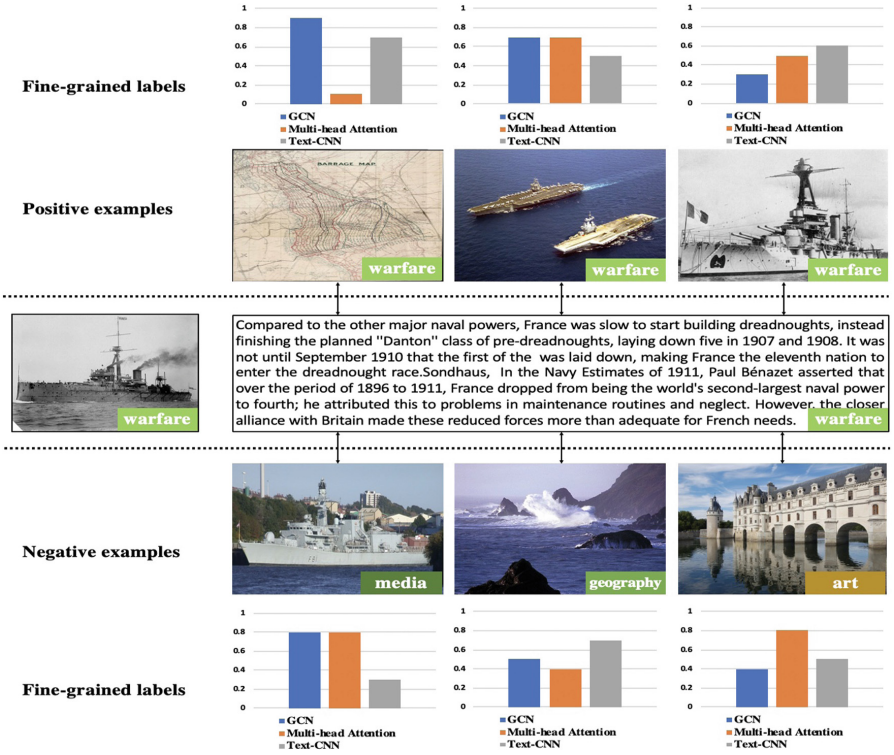

Fig. 4. Samples of the fine-grained labels obtained by three models on the Eng-Wiki dataset. For easy evaluation of the label quality, we show the positive examples and negative examples containing the same text (middle) but different images (top and down, respectively). For clear semantic comparisons, we show the corresponding image of the text on the middle left. The bar graphs on the top/bottom of each image show the fine-grained labels of the three models using different text features.

Both Multi-head attention approach and Text-CNN approach obtains some unreasonable labels, which degrades the retrieval performance.

5 Conclusion

In the paper, we propose a Fine-Grained Label learning approach for cross-modal information retrieval. We design a siamese network to learn fine-grained labels for both the positive and negative examples to capture the degrees of hardness, thus enhancing cross-modal correlation learning. We introduce these labels to a rank-based pairwise similarity loss function. The loss function benefits from the labels to increase the influence of hard examples on the similarity learning while maximizing the similarity of relevant text-image pairs and minimizing the similarity of irrelevant pairs. The experimental results on three widely

used datasets indicate that, comparing with state-of-the-art models, our model achieves significant improvements on the retrieval performance by incorporating with fine-grained labels.

References

1. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
2. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: VSE++: improving visual-semantic embeddings with hard negatives (2017)
3. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
4. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: an overview with application to learning methods. Neural Comput. **16**(12), 2639–2664 (2004)
5. Hotelling, H.: Relations between two sets of variates. Biometrika **28**(3/4), 321–377 (1936)
6. Jin, S.Y., et al.: Unsupervised hard example mining from videos for improved object detection. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11217, pp. 316–333. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01261-8_19
7. Kang, C., Xiang, S., Liao, S., Xu, C., Pan, C.: Learning consistent feature representation for cross-modal multimedia retrieval. IEEE Trans. Multimedia **17**(3), 370–381 (2015)
8. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
9. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
10. Ma, Z., Chang, X., Yang, Y., Sebe, N., Hauptmann, A.G.: The many shades of negativity. IEEE Trans. Multimedia **19**(7), 1558–1568 (2017)
11. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-SVMs for object detection and beyond. In: 2011 IEEE International Conference on Computer Vision (ICCV), pp. 89–96. IEEE (2011)
12. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11), pp. 689–696 (2011)
13. Pereira, J.C., et al.: On the role of correlation and abstraction in cross-modal multimedia retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **36**(3), 521–535 (2014)
14. Qin, Z., Yu, J., Cong, Y., Wan, T.: Topic correlation model for cross-modal multimedia information retrieval. Pattern Anal. Appl. **19**(4), 1007–1022 (2016)
15. Ranjan, V., Rasiwasia, N., Jawahar, C.: Multi-label cross-modal retrieval. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4094–4102 (2015)
16. Rasiwasia, N., et al.: A new approach to cross-modal multimedia retrieval. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 251–260. ACM (2010)

17. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 761–769 (2016)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
19. Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: Advances in Neural Information Processing Systems, pp. 2222–2230 (2012)
20. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)
21. Wang, K., He, R., Wang, L., Wang, W., Tan, T.: Joint feature selection and subspace learning for cross-modal retrieval. IEEE Trans. Pattern Anal. Mach. Intell. **38**(10), 2010–2023 (2016)
22. Wang, K., He, R., Wang, W., Wang, L., Tan, T.: Learning coupled feature spaces for cross-modal matching. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2088–2095 (2013)
23. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5005–5013 (2016)
24. Wu, C.Y., Manmatha, R., Smola, A.J., Krahenbuhl, P.: Sampling matters in deep embedding learning. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2840–2848 (2017)
25. Yu, J., et al.: Modeling text with graph convolutional network for cross-modal information retrieval. In: Hong, R., Cheng, W.-H., Yamasaki, T., Wang, M., Ngo, C.-W. (eds.) PCM 2018. LNCS, vol. 11164, pp. 223–234. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00776-8_21
26. Zhang, L., Ma, B., He, J., Li, G., Huang, Q., Tian, Q.: Adaptively unified semi-supervised learning for cross-modal retrieval. In: International Conference on Artificial Intelligence, pp. 3406–3412 (2017)