Computer Vision and Image Understanding 000 (2017) 1–11

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Simple to complex cross-modal learning to rank

Minnan Luo[a], Xiaojun Chang[b,*], Zhihui Li[c], Liqiang Nie[d], Alexander G. Hauptmann[b], Qinghua Zheng[a]

[a] SPKLSTN Lab, Department of Computer Science, Xi'an Jiaotong University, Xi'an, China
[b] School of Computer Science, Carnegie Mellon University, PA, USA
[c] Faculty of Engineering and Information Technology, University of Technology Sydney, Australia
[d] School of Computing, National University of Singapore, Singapore

ARTICLE INFO

Article history:
Received 22 August 2016
Revised 22 June 2017
Accepted 4 July 2017
Available online xxx

Keywords:
Cross-modal retrieval
Learning to rank
Self-paced learning
Diversity regularization

ABSTRACT

The heterogeneity-gap between different modalities brings a significant challenge to multimedia information retrieval. Some studies formalize the cross-modal retrieval tasks as a ranking problem and learn a shared multi-modal embedding space to measure the cross-modality similarity. However, previous methods often establish the shared embedding space based on linear mapping functions which might not be sophisticated enough to reveal more complicated inter-modal correspondences. Additionally, current studies assume that the rankings are of equal importance, and thus all rankings are used simultaneously, or a small number of rankings are selected randomly to train the embedding space at each iteration. Such strategies, however, always suffer from outliers as well as reduced generalization capability due to their lack of insightful understanding of procedure of human cognition. In this paper, we involve the self-paced learning theory with diversity into the cross-modal learning to rank and learn an optimal multi-modal embedding space based on non-linear mapping functions. This strategy enhances the model's robustness to outliers and achieves better generalization via training the model gradually from easy rankings by diverse queries to more complex ones. An efficient alternative algorithm is exploited to solve the proposed challenging problem with fast convergence in practice. Extensive experimental results on several benchmark datasets indicate that the proposed method achieves significant improvements over the state-of-the-arts in this literature.

1. Introduction

In many real-world applications, data related to the same underlying object (content) are often exhibited in diverse modalities for better human cognition (Lux et al., 2004; Zhai et al., 2013). For example, when we want to know what is a dinosaur, we prefer to find the results across various modalities, such as searching images (videos) to figure out what a dinosaur looks like, also searching text description on its size and other biology information for best comprehension. As a result, cross-modal retrieval attracts increasing attention and plays an important role to describe the content of an image with natural language and conversely retrieve image given textual query (Amir et al., 2004; Chang et al., 2017b; Pereira and Vasconcelos, 2014). However, since data in diverse modalities are presented in heterogeneous feature spaces and usually have varying statistical properties, it is a significant challenge to bridge the heterogeneity-gap between multi-modal data (Grangier and Bengio, 2008; Ranjan et al., 2015).

In the past decades, a large number of efforts have been devoted to revealing the inter-modal correspondence via learning a shared embedding space for cross-modal similarity measurement (Chang et al., 2017a; Irie et al., 2015; Jin et al., 2015; Kang et al., 2015a; Menon et al., 2015; Wang et al., 2015). For example, Canonical Correlation Analysis (CCA) and its extensions to kernel version (Hardoon et al., 2004; Hotelling, 1936) aim to learn a common representation by mutually maximizing the correlation between their projections onto the shared basis vectors; Latent Dirichlet Allocation (LDA) based methods (Barnard et al., 2003; Blei and Jordan, 2003; Jia et al., 2011; Wang et al., 2009; Xiaojun Chang and Hauptmann, 2017) establish the shared latent semantic model through the joint distribution of images and the corresponding annotations as well as the conditional relationships between them. However, these methods separate the shared space learning from the ultimate ranking performance, and thus usually suffer from poor generalization capability (Lu et al., 2013). Motivated by the Ordered

* Corresponding author.
E-mail addresses: minnluo@xjtu.edu.cn (M. Luo), cxj273@gmail.com (X. Chang), zhihuilics@gmail.com (Z. Li), nieliqiang@gmail.com (L. Nie), alex@cs.cmu.edu (A.G. Hauptmann), qhzheng@xjtu.edu.cn (Q. Zheng).

http://dx.doi.org/10.1016/j.cviu.2017.07.001
1077-3142/

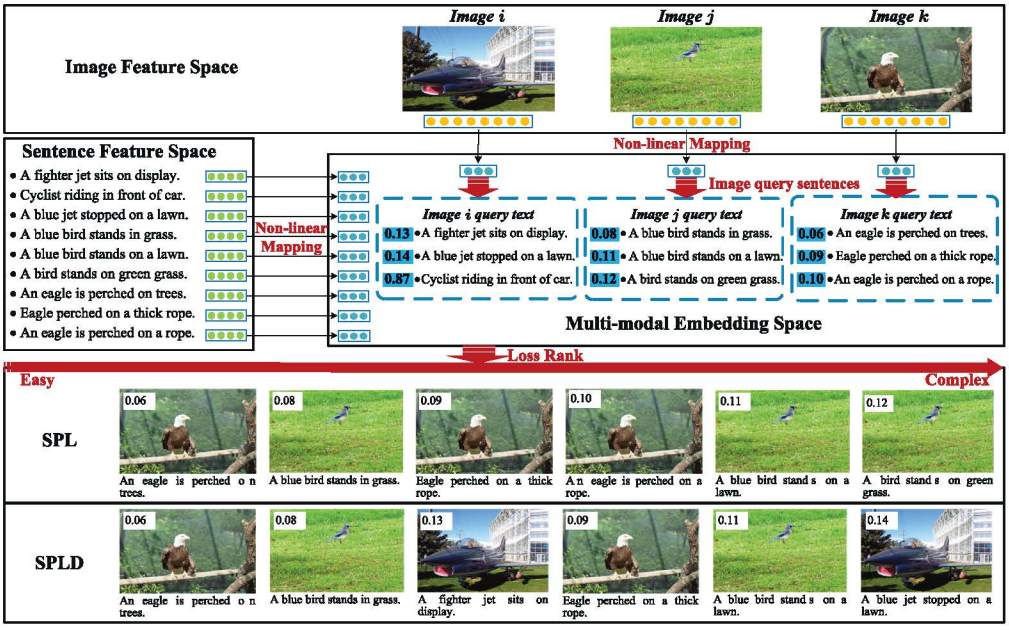


Fig. 1. The framework of the proposed simple to complex cross-modal learning to rank.

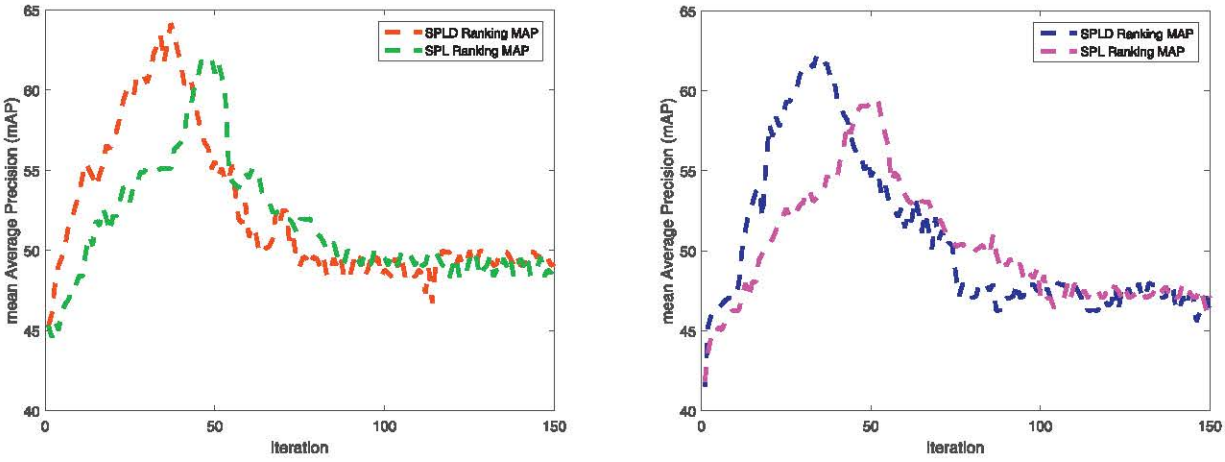
Weighted Pairwise Classification (OWPC) loss (Usunier et al., 2009), Weston et al. (2011) involved a dynamic importance for different ranking and proposed a Weighted Approximate-Rank Pairwise loss (WARP) for multi-label annotation problem. Gong et al. (2013) developed the deep extension of this method. However, WARP is parameterized by a set of decreasing weights which are predefined. Instead, Cross-modal Learning to Rank (CMLR) learns the multi-modal embedding space through minimizing a ranking-based loss function (Deng et al., 2016; Grangier and Bengio, 2008; Kang et al., 2015b; McFee and Lanckriet, 2010; You et al., 2016). Since CMLR is designed orientating to the performance of cross-modal ranking directly, it has become an increasingly important research direction in cross-modal information retrieval. Many approaches have been proposed based on this strategy (Habibian et al., 2015; Jiang et al., 2015; Li et al., 2013; Wang et al., 2016a, 2014; Wu et al., 2015, 2013; Zhu et al., 2013).

However, the task of CMLR remains a significant challenge because it requires an understanding of the content of images, sentences, and their inter-modal correspondence simultaneously (Karpathy et al., 2014). To the best of our knowledge, most methods employ linear mapping functions to translate the image and text feature vectors into a shared embedding space respectively. Although these linear mapping functions are easy to construct, they might not be capable of faithfully reflecting more sophisticated cross-modal correspondence (Jiang et al., 2015). Additionally, given an image query, previous methods suppose that all of the texts in the rank list are of equal importance, and thus either all ranking texts are utilized simultaneously, or a small number of ranking texts are selected randomly at each iteration to train the embedding space. Indeed, the texts ranked higher are more accurate, and thus, should be more important than those ranked lower (Jiang et al., 2014a). As a result, it is significantly necessary to develop more sophisticated mapping functions and discriminate the contributions of each ranking in a theoretically sound manner.

To this end, we incorporate a self-paced learning with diversity (SPLD) theory (Bengio et al., 2009) into CMLR to train an optimal embedding space based on non-linear mapping functions. In such a way, the model is learned gradually from easy rankings with respect to diverse queries to more complex ones. For a better understanding, we take image-query-sentence as an example to illustrate the proposed framework, as shown in Fig. 1. Through non-linear mapping, we translate images and sentences lying in heterogeneous feature spaces into a shared embedding space to facilitate the similarity measurement between image and sentence. Given each image query, the retrieved sentences are ordered according to their ranking loss, as specified by the numbers in Fig. 1. It is reasonable to believe that the sentences ranked higher, i.e., with a smaller loss, are usually more accurate and important. These ranking sentences with the corresponding image query are referred to as easiness in this paper. To learn an optimal embedding space, we follow the self-paced learning (SPL) and select ranking sentences together with the corresponding image queries from easy to more complex (See the row of SPL in Fig. 1 for example). However, SPL only considers about the easiness, not about the diversity of the selected ranking sentences with respect to different image queries. Indeed, studies have suggested that diversity is an important aspect of learning because performance is enhanced significantly through samples that are dissimilar from what has already been learned (Jiang et al., 2014b; Zhao et al., 2015). Ignoring the diversity may lead to over-fitting to a subset of easy rankings by some specific queries. This is significant since the over-fitting becomes increasingly severe as the rankings by some specific queries are kept adding into training while ignoring the easy rankings by other queries. As shown in the row of SPL in Fig. 1, all the rankings with respect to the *i*th image query fails to be selected with SPL. For this issue, we further improve the SPL by considering both easiness and diversity, such that the selected easy rankings are scattered across all image queries as much as possible. As indicated in the last row of Fig. 1, SPL with diversity (SPLD) helps to select from easy ranking sentences with respect to diverse image queries to more complex ones.

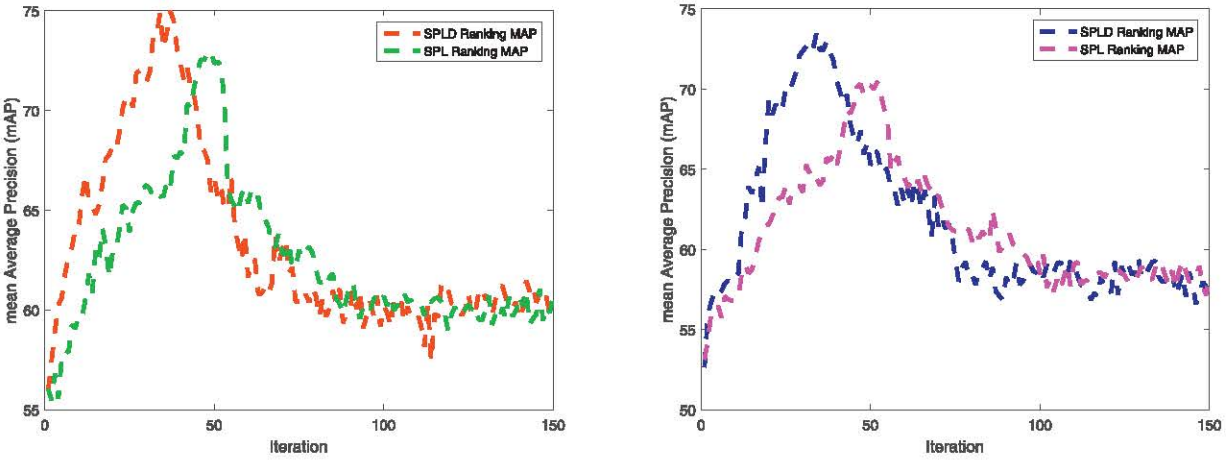
In summary, we describe the contributions of this paper as follows: (1) From a new perspective, we adaptively assign each ranking with an importance weight and learn a more optimal multi-modal embedding space gradually from easy to more complex rankings with respect to diverse image queries. (2) We employ non-linear mapping functions to learn the multi-modal embedding space, such that more sophisticated cross-modal correspondence

Please cite this article as: M. Luo et al., Simple to complex cross-modal learning to rank, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.07.001



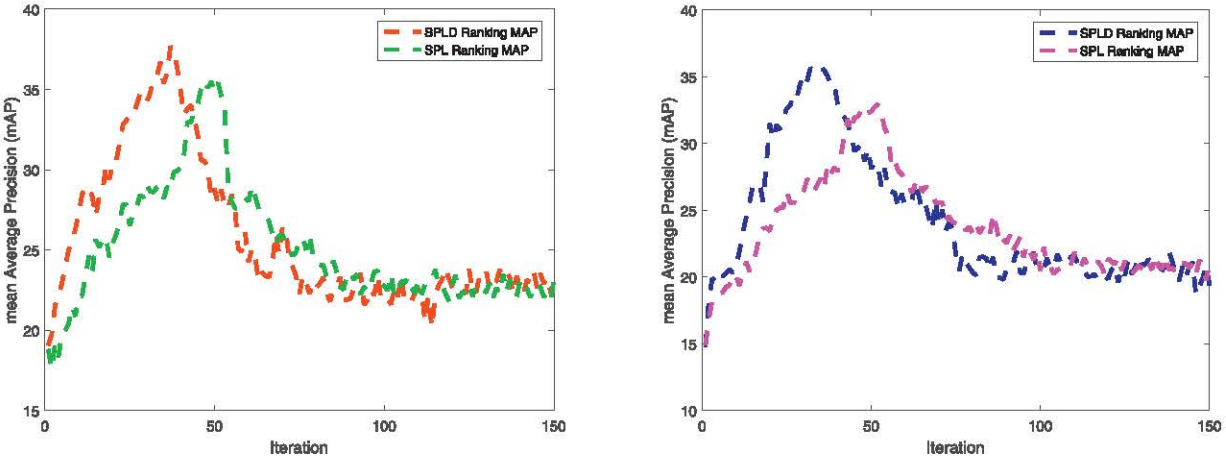
(a) Pascal'07 Image-Query-Text (b) Pascal'07 Text-Query-Image

Fig. 6. Performance comparison over Pascal'07 dataset between SCCM w/t and w/o diversity w.r.t various iterations.



(a) NUS-WIDE Image-Query-Text (b) NUS-WIDE Text-Query-Image

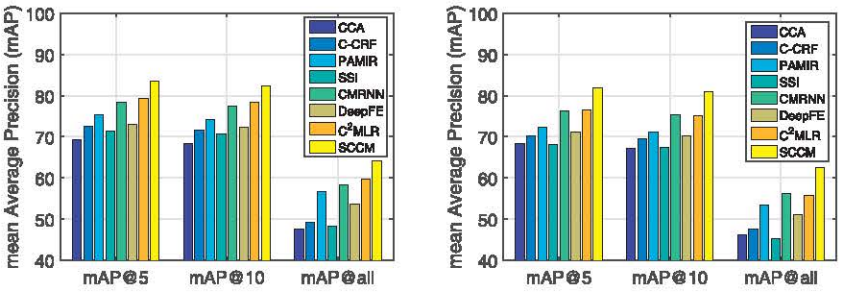
Fig. 7. Performance comparison over NUS-WIDE dataset between SCCM w/t and w/o diversity w.r.t various iterations.



(a) Wiki Image-Query-Text (b) Wiki Text-Query-Image

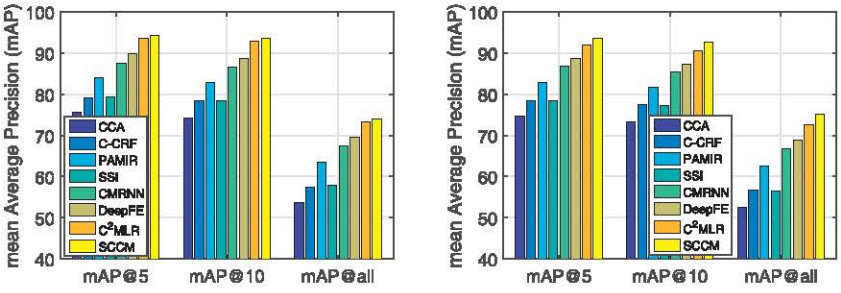
Fig. 8. Performance comparison over Wiki dataset between SCCM w/t and w/o diversity w.r.t various iterations.

Please cite this article as: M. Luo et al., Simple to complex cross-modal learning to rank, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.07.001



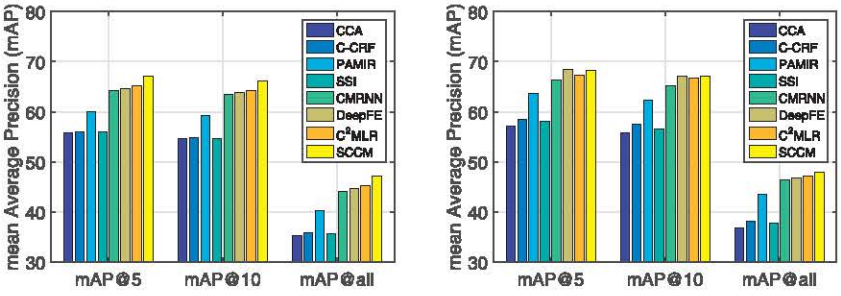
(a) Pascal'07 Image-Query-Text (b) Pascal'07 Text-Query-Image

Fig. 3. The performance comparison in terms of mAP over Pascal'07 dataset.



(a) NUS-WIDE Image-Query-Text (b) NUS-WIDE Text-Query-Image

Fig. 4. The performance comparison in terms of mAP over NUS-WIDE dataset.



(a) Wiki Image-Query-Text (b) Wiki Text-Query-Image

Fig. 5. The performance comparison in terms of mAP over Wiki dataset.

Table 1
Performance comparison of the proposed algorithm w/t and w/o diversity. Mean average precision (mAP) is used as an evaluation metric. Results are shown in percentages. Larger mAP indicates better performance.

Dataset		IQT		TQI	
		w/t div.	w/o div.	w/t div.	w/o div.
Pascal'07	mAP@5	**83.6**	79.7	**81.8**	77.4
	mAP@10	**82.3**	78.9	**80.9**	76.2
	mAP@all	**64.1**	59.6	**62.6**	57.5
NUSWIDE	mAP@5	**94.2**	90.8	**93.5**	89.2
	mAP@10	**93.6**	89.9	**92.6**	88.4
	mAP@all	**74.1**	69.7	**75.3**	70.5
Wiki	mAP@5	**67.1**	62.3	**68.2**	63.6
	mAP@10	**66.2**	61.6	**67.1**	62.8
	mAP@all	**47.1**	44.7	**47.9**	43.6

modified framework without diversity. Regarding to the task of Image-Query-Text (IQT) and Text-Query-Image (TQI), Table 1 shows the performance comparison between with diversity (w/t div.) and without diversity (w/o div.). The results clearly demonstrated that the proposed method with diversity consistently outperforms the framework without diversity in both retrieval directions over the three datasets. For example, for the retrieval task of Image-Query-Text, the framework with diversity outperforms the model without diversity, which is 64.1 *v.s.* 59.6 over the Pascal'07 dataset. We attribute this significant improvement to adaptive distinguishing the contributions of varying rankings to the shared space learning and considering the diversity of rankings by different queries.

In addition, we plot performance comparison between SCCM w/t and w/o diversity w.r.t various iterations in terms of mAP@all on all three datasets in two retrieval directions in Figs. 6–8. A common phenomenon is that the performance improves as the iteration increases. After the performance arrives at its peak, the performance will drop if more iterations are conducted. This is because that some complex image-sentence pairs with large loss values have negative effect on the performance. By comparing SCCM w/t and w/o diversity, we observe that if diversity is considered, SCCM attains a better solution within fewer iterations. For example, in Fig. 6 a, SCCM w/t diversity obtains the best performance by 36 iterations while SCCM w/o diversity arrives at its peak by 58 iterations. This result indicates that SCCM w/t diversity converges much faster than SCCM w/o diversity.

Please cite this article as: M. Luo et al., Simple to complex cross-modal learning to rank, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.07.001

- **PAMIR** (Grangier and Bengio, 2008): PAMIR utilizes global alignment to learn the latent representations of pairs of images and text texts in a pairwise ranking manner. Image and texts are embedded into a global common space using a linear projection.
- **SSI** (Bai et al., 2010): SSI discriminatively trains a class of nonlinear models to map from the word content in a query-document or document-document to a ranking score in a pairwise ranking manner.
- **CMRNN** (Lu et al., 2014): CMRNN is built on top of neural networks and learning to rank techniques, which learns high-level feature representation with discriminative power for cross-modal ranking.
- **DeepFE** (Karpathy et al., 2014): DeepFE learns a multi-modal embedding space for fragments of images and sentences and reasons about their latent, inter-modal alignment. It considers the local alignment of images and sentences.
- **C^{2}MLR** (Jiang et al., 2015): C^{2}MLR considers learning a multi-modal embedding from the perspective of optimizing a pairwise ranking problem while enhancing both local alignment and global alignment. C^{2}MLR learn a ranking manner using the local common space and the global common space jointly, where the local common space is computed by local alignment of visual objects and textual words and the global common space is from the global alignment of images and text.

5.4. Evaluation metric

Mean Average Precision (mAP) is used as an evaluation metric, which is one of the most widespread performance evaluations of information retrieval. Given the Average Precision (AP) of all queries, mAP is the mean of all AP values. And the value AP of a query is calculated according to the formula (30).

$$AP(Y, Y') = \frac{1}{R}\sum_{j=1}^{R} Prec(j)Rel(j) \tag{30}$$

where Y and Y' denotes the true ranking list and the predicted ranking list namely; R is the number of retrieved texts to be examined in the ranking list if R is 5, the mAP is represented mAP@5, and when the value of R is the number of all texts, the mAP is represented mAP@all; $Prec(.)$ is the percentage of the relevant texts in the top j texts in the predicted ranking; $Rel(.)$ is the indicator function equaling to 1 if the document at rank j is relevant. mAP is a more suitable measure than other mentioned metrics in this particular task. Note that mAP indeed considers both precision and recall of the retrieved results, and thus it is a suitable measure for specific task of cross-media learning to rank.

5.5. Performance comparison

We report the performance of cross-modal ranking in terms of mAP on Pascal'07 dataset in Fig. 3 a (Image-Query-Text) and Fig. 3 b (Text-Query-Image). Note that whenever possible we have quoted the numbers directly from the references, while if not available we used code from the respective authors to obtain the results ourselves. From Figs. 3 a and b, it can be seen that the proposed algorithm outperforms the other alternatives by a large margin. We have the following observations on Pascal'07:

- The performances of all the baseline algorithms are much better than those reported in Jiang et al. (2015) and Lu et al. (2014). We attribute this improvement to the adoption of ImageNet Shuffle model for feature extraction. To the best of our knowledge, this is the first work to use ImageNet Shuffle model for learning to rank algorithm.
- CCA obtains very similar performance in both directions of the retrieval. This is because CCA learns the joint representation from paired multi-modal data in which the pair-correspondence of images and text texts ensures an equal contribution to the learned metric in both modalities.
- PAMIR performs much better than CCA and C-CRF in both directions, which confirms that learning a good representation for multi-modal data is crucial for cross-modal ranking.
- We observe that cross-modal with local alignment (i.e., DeepFE) obtains a poor performance on this dataset. This is because many annotated tags of images in Pascal'07 dataset do not explicitly align with visual objects in images.
- The proposed **SCCM** outperforms all baseline methods, which confirms the assumption that learning from simple to complex and considering diversity for learning to rank is instrumental.

Fig. 4a (Image-Query-Text) and Fig. 4b (Text-Query-Image) show the performance comparison of cross-modal ranking in terms of mAP over NUS-WIDE dataset. We have the following observations from the experimental results:

- The proposed **SCCM** outperforms the other alternatives on both search directions by a large margin in terms of all the evaluation metrics.
- SSI generally performs better than CCA. Compared to CCA, SSI introduces a nonlinear projection to map the multi-modal data into a common space. This observation verifies that a nonlinear projection for multi-modal learning achieves better performance than a linear projection.
- The ranking algorithms with local alignment or global alignment generally outperforms the other alternatives. For example, DeepFE and C^{2}MLR achieves significant improvement over CCA, C-CRF, PAMIR and CMRNN. However, these algorithms are built on the assumption that there is an explicit alignment between visual objects and textual words.

We report the experimental results over Wiki dataset in Fig. 5a (Image-Query-Text) and Fig. 5b (Text-Query-Image). By comparing the performance of different alternatives on this dataset, we have the following observations:

- By comparing the performance in the two directions (image-query-text and text-query-image), most of the performance obtain unbalanced performance. In contrast, CCA gets very similar performances in both directions of the retrieval. The reason is that CCA takes strictly paired multi-modal data as the training instances, which makes CCA tend to capture the pair-correspondence between multi-modal data and is unable to capture the discriminative information between multi-modal data.
- PAMIR gets the best performance among the compared non-deep methods. This phenomenon is because PAMIR maps the images into the textual space while the high-level semantics delivered by the textual space is reasonable enough to get good performance on ranking text.

Based on the above observations, we conclude that (1) nonlinear mapping function contributes to the cross-modal learning to rank algorithm because it is able to capture more sophisticated cross-modal correspondence; (2) it is advantageous to learn an optimal multi-modal embedding space gradually from easy to complex rankings by diverse image queries.

5.6. Does diversity help?

In this section, we first conduct an experiment to evaluate whether diversity contribute much to the subsequent performance. By setting $\gamma = 0$ in the optimization problem (10), we obtain a

Please cite this article as: M. Luo et al., Simple to complex cross-modal learning to rank, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.07.001

Algorithm 1 Algorithm of optimizing importance weight $\mathbf{v}$.

Input: Tetrads set $\mathcal{T}_{\mathcal{X}} = \{(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}) \in \mathcal{T}_{\mathbf{x}}^k : j \neq k; k = 1, 2, \cdots, n\}$; Current embedding parameters W; two trade-off parameters λ and γ.

Output: The global optimum $\mathbf{v}^*$ of optimization problem (21).

1: **for** $k = 1, 2, \cdots, n$ **do**

2: Sort the Tetrads in ascending order according to their loss values as $l_{j_1}^k \leq l_{j_2}^k \leq \cdots \leq l_{j_{n-1}}^k$;

3: Compute $U = \{u : l_{j_u}^k < \lambda + \frac{\gamma}{2\sqrt{j_u}}\}$;

4: **if** $U \neq \emptyset$ **then** $v_{j_u}^k = 1, \forall u \in U$;

5: Find $u' = \arg\min_{1,2,\cdots,n-1}\{u : v_{j_u}^k = v_{j_{u+1}}^k = \cdots = v_{j_{u+m-1}}^k, l_{j_u}^k \geq \lambda + \frac{\gamma}{2\sqrt{j_u}}\}$ and set $v_{j_{u'+m}}^k = v_{j_{u'+m+1}}^k = \cdots = v_{j_{n-1}}^k = 0$,

$$v_{j_{u'}}^k = \cdots = v_{j_{u'+m-1}}^k = \frac{\gamma/2(l_{j_{u'}}^k - \lambda)^2 - j_{u'} + 1}{m};$$

6: **end for**

Algorithm 2 Algorithm for optimization probelem (10).

Input: Tetrads set $\mathcal{T}_{\mathcal{X}} = \{(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}) \in \mathcal{T}_{\mathbf{x}}^k : j \neq k; k = 1, 2, \cdots, n\}$; two trade-off parameters λ and γ.

Output: parameters W.

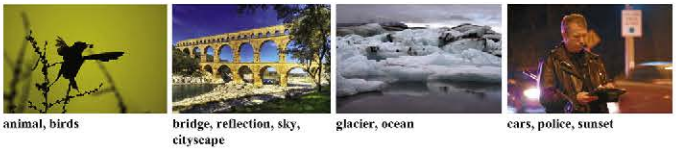
1: **repeat**

2: Updating W according to (19) and (20)

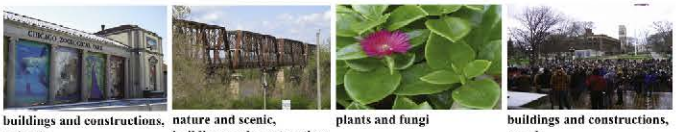
3: Updating $\mathbf{v}$ using Algorithm 1.

4: **until** convergence

(a) Pairs of image-text examples in Pascal'07 dataset



(b) Pairs of image-textexamples in NUS-WIDE dataset



(c) Pairs of image-text examples in Wiki dataset

Fig. 2. The pairs of image-text examples in Pascal'07, NUS-WIDE and Wiki datasets, respectively.

5.1. Dataset description and experimantal setup

The details of the datasets are introduced as follows, together with the feature representation of the texts for each dataset. Some example pairs of image-text are shown in Fig. 2 for a better understanding.

- **Pascal'07:** The Pascal'07 dataset is a widely used benchmark dataset in category recognition and multi-modal classification. It consists of 10,000 images from 20 different categories. 804 corresponding tags are downloaded from Flickr for each image in the dataset and are represented as an 804-dimensional feature vector, each of whose dimension indicates if a tag appears. There are 9587 images in the dataset with the user tags available, which are the image-tag pairs we use in the experiment.
- **NUS-WIDE:** This dataset contains 269,648 images with 1000 associated tags from Flickr. Each image with the corresponding tags has several of the 81 concepts as the ground-truth. We represent the corresponding tags of each image as a 1000-dimensional vector, each dimension of which is a binary indicator to indicate whether a tag appears or not.
- **Wiki image-text:** The dataset contains 2886 articles with the corresponding image in each of the articles. All of the Wikipedia articles are categorized as one of the ten semantic classes. We extract a feature vector from each article with the bag of words model (BoW), resulting a 1000-dimensional representation.

For a fair comparison, we use the same experimental setting in Rasiwasia et al. (2010). Specifically, for Pscal'07 and NUS-WIDE datasets, 2000 images and the associated tags are randomly selected as the training set. 1000 images and the corresponding tags are selected as a validation set used for parameter tuning. The rest are used for testing. For the Wiki dataset, 1200 images and the corresponding text documents are randomly selected as the training set. 500 images and the corresponding text documents are selected as a validation set used for the parameter tuning. Note that the text consisting of multiple tags is used as a query in the experiments.

5.2. Feature extraction

Previous researches have demonstrated that pre-trained model with ImageNet dataset can boost the performance of other important tasks. Different from the leading approaches, who all pre-train CNN models from the 1000 classes defined in the ImageNet Large Scale Visual Recognition Challenge, we leverage the complete ImageNet hierarchy for pre-training deep networks following (Mettes et al., 2016). The key insight in Mettes et al. (2016) is that by utilizing the graph structure of ImageNet to combine and merge classes into balanced and reorganized hierarchies, a significant improvement on the visual recognition task can be achieved. To deal with the problems of over-specific classes and imbalanced classes, we adopt a bottom-up and top-down approach for reorganization of the ImageNet hierarchy. After the training data has been reorganized, we pre-train a CNN model using the same architecture as GooLeNets (Szegedy et al., 2015). The Caffe toolkit (Jia et al., 2014) is used in our experiment. After pre-training, we extract features at the pool5 layer, with a 1,024-dimensional frame representation. We normalize the representation by ℓ_2-normalization.

5.3. Competitors

To evaluate the effectiveness of the proposed method **SCCM**, we compare with the following alternatives. We choose the best parameters using 5-fold cross-validation. All the parameters are tuned in the range of $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$.

- **CCA** (Hardoon et al., 2004): CCA maps the pairs of images and text texts into a latent space, and therefore, the latent representations of the images and text texts are obtained. After the latent representations are individually obtained, CCA performs cross-modal retrieval by measuring the relevance of queries and texts with the cosine similarity regarding their individually latent representation.
- **C-CRF** (Qin et al., 2008): C-CRF first maps the pairs of images and text texts into a latent space, followed by performing the cross-modal retrieval with C-CRF in a list-wise ranking manner.

Please cite this article as: M. Luo et al., Simple to complex cross-modal learning to rank, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.07.001

In summary, we pioneer to associate each tetrad with an adaptive importance weight and employ self-pace regularization $\phi(\mathbf{v}, \lambda, \gamma)$ to guide the learning in a theoretically sound manner. This strategy enhances model's robustness to outliers and improve its generalization capability.

4. Optimization procedure

In this section, we exploit an alternative optimization algorithm to solve the proposed challenging problems via updating embedding parameters W and importance vector $\mathbf{v}$ iteratively with the other one fixed.

4.1. Optimize W

In this step, we seeks to estimate the embedding parameters, i.e., W_1, $\mathbf{b}_1$ and W_2, $\mathbf{b}_2$. With fixed importance weight vector $\mathbf{v}$, the optimization problem (10) degenerates to the following form:

$$\min_{W} \ \frac{1}{2}\|W\|^2 + \sum_k \sum_{j\neq k} v_j^k \ l(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}; W). \tag{15}$$

For this optimization problem, we use the gradient descent method to update embedding parameters W at each iteration. Let the objective in optimization problem (15) be $f(W; \mathbf{v})$. The derivatives of $f(W; \mathbf{v})$ with respect to parameter W_1, denoted by ∇f_{W_1}, can be computed as

$$\begin{aligned}\nabla f_{W_1} &= \frac{\partial f(W;\mathbf{v})}{\partial W_1} \\ &= W_1 + \sum_k \sum_{j\neq k} v_{kj} \frac{\partial l(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}; W)}{\partial W_1}\end{aligned} \tag{16}$$

where according to the definition of loss function in (8), we calculate its gradient with respect to parameter W_1 as follows

$$\frac{\partial l(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}; W)}{\partial W_1} = y_{kj}\left(\frac{\partial S(\mathbf{x}_k, \mathbf{z}_i)}{\partial W_1} - \frac{\partial S(\mathbf{x}_k, \mathbf{z}_j)}{\partial W_1}\right). \tag{17}$$

From the gradient above, we observe that the cost function value is back propagated into the gradient of similarity measurement $S(\mathbf{x}, \mathbf{z})$ with respect to parameter W_1, i.e.

$$\begin{aligned}\frac{\partial S(\mathbf{x},\mathbf{z})}{\partial W_1} &= \frac{\partial\left(h(\mathbf{x})^\top g(\mathbf{z})\right)}{\partial W_1} \\ &= \frac{\partial\left(f(W_1\mathbf{x}+\mathbf{b}_1)^\top g(W_2\mathbf{z}+\mathbf{b}_2)\right)}{\partial W_1} \\ &= \left(g(W_2\mathbf{z}+\mathbf{b}_2)\odot f'(W_1\mathbf{x}+\mathbf{b}_1)\right)\mathbf{x}^\top\end{aligned} \tag{18}$$

where $\odot$ represents the element-wise multiplication and $f'(\cdot)$ refers to the derivative of $f(\cdot)$ with respect to its input. Based on the three equations above, we achieve the gradient of $\Omega(W; \mathbf{v})$ with respect to parameter W_1. According to similar derivations, we can also obtain the gradient of $f(W; \mathbf{v})$ with respect to parameter W_2, denoted by ∇f_{W_2}. In summary, the embedding parameters W_1 and W_2 can be updated as

$$W_1 \leftarrow W_1 + a\nabla f_{W_1} \tag{19}$$

$$W_2 \leftarrow W_2 + b\nabla f_{W_2} \tag{20}$$

where a and b are the step size which can be found by linear search.

4.2. Optimize v

After updating the embedding parameters, we renew the weights $v_{kj}(j\neq k)$ to reflect the adaptive importance of tetrad $(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj})$. Following the algorithm proposed in Zhang et al. (2015), when W is fixed, $\mathbf{v}$ can be updated by solving optimization problem

$$\begin{aligned}&\min_{\mathbf{v}} \ \sum_k \sum_{j\neq k} v_j^k \ l(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}; W) + \phi(\mathbf{v}; \lambda, \gamma) \\ &s.t. \quad v_j^k \in [0,1] \qquad (\forall k, \ j \neq k)\end{aligned} \tag{21}$$

where $\phi(\mathbf{v}; \lambda, \gamma) = -\lambda \sum_k \sum_{j\neq k} v_j^k - \gamma \sum_k \sqrt{\sum_{j\neq k} v_j^k}$. Let $l_j^k = l(\mathbf{x}_i, \mathbf{z}_k, \mathbf{z}_j, y_{kj}^k; W)$. Since the objective function (21) is independent between different k, we can estimate the importance weight $\mathbf{v}^k = [v_1^k, \cdots, v_{k-1}^k, v_{k+1}^k, \cdots, v_n^k]$ individually via solving the following optimization problems

$$\begin{aligned}&\min_{\mathbf{v}^k} \ \psi(\mathbf{v}^k) = \sum_{j\neq k} v_j^k \ l_j^k - \lambda\|\mathbf{v}^k\|_1 - \gamma\sqrt{\sum_{j\neq k} v_j^k} \\ &s.t. \quad v_j^k \in [0,1] \qquad (\ j \neq k)\end{aligned} \tag{22}$$

for each $k = 1, 2, \cdots, n$. In terms of Lagrangian parameters

$$\alpha^k = [\alpha_1^k, \cdots, \alpha_{k-1}^k, \alpha_{k+1}^k, \cdots, \alpha_n]^\top \in \mathbb{R}^{n-1}; \tag{23}$$

$$\beta^k = [\beta_1^k, \cdots, \beta_{k-1}^k, \beta_{k+1}^k, \cdots, \beta_n]^\top \in \mathbb{R}^{n-1}, \tag{24}$$

the Lagrangian function of $\psi(\mathbf{v}^k)$ is formulated as

$$L(\mathbf{v}^k, \alpha^k, \beta^k) = \psi(\mathbf{v}^k) - \sum_j \alpha_j^k v_j^k - \sum_{j\neq k} \beta_j^k (1 - v_j^k). \tag{25}$$

Consequently, we arrive at the corresponding KKT conditions (Boyd and Vandenberghe, 2004) as:

$$\frac{\partial L}{\partial \mathbf{v}_j^k} = l_j^k - \lambda - \frac{\gamma}{2\sqrt{\sum_{j\neq k} v_j^k}} - \alpha_j^k + \beta_j^k = 0 \tag{26}$$

$$\alpha_j^k v_j^k = 0; \tag{27}$$

$$\beta_j^k (1 - v_j^k) = 0; \tag{28}$$

$$\alpha_j^k \geq 0; \quad \beta_j^k \geq 0. \tag{29}$$

Thanks to the convexity of objective function $\psi(\cdot)$, we get a global optimum $\mathbf{v}^*$ satisfying these KKT conditions (26)–(29). In summary, we describe the algorithm of optimizing $\mathbf{v}$ in Algorithm 1. The overall alternative optimization algorithm for the proposed self-paced CMLR with diversity regularization is summarized in Algorithm 2. The proposed algorithm is efficient due to the following analysis. Let z be the average number of tetrads selected by the self-paced function. In each iteration, the complexity mainly lies in Step 2 and 3 of Algorithm 2. In Step 2, the main computational overhead comes from obtaining similarity score for the selected tetrads with complexity $O(z^2)$. In Step 3, the main computational cost of updating $\mathbf{v}$ lies in the calculation of loss of each tetrad with complexity $O(n^2)$.

5. Experiments

To illustrate the effectiveness and superiority of the proposed **S**imple to **C**omplex **C**ross-**M**odel learning to rank framework (**SCCM**), we perform extensive experiments over some benchmark datasets and demonstrate the efficiency of the diversity regularization used in self-paced learning.

Please cite this article as: M. Luo et al., Simple to complex cross-modal learning to rank, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.07.001

Through non-linear mapping h and g, the similarity measurement (relevance score) $S(\mathbf{x}, \mathbf{z})$ between image query $\mathbf{x}$ and the retrieved text $\mathbf{z}$ can be obtained via computing the cosine similarity in the shared embedding space, i.e.,

$$S(\mathbf{x}, \mathbf{z}) = h(\mathbf{x})^\top g(\mathbf{z}). \tag{3}$$

In this case, the underlying correspondence between image and text lies in the embedding parameters W_1, $\mathbf{b}_1$ and W_2, $\mathbf{b}_2$. We add one dimension values 1 in each input feature $\mathbf{x} \in \mathcal{X}$ and $\mathbf{z} \in \mathcal{Z}$ and view the biases $\mathbf{b}_1$ and $\mathbf{b}_2$ as an extra column of the corresponding transformation matrices W_1 and W_2, respectively. Without loss of generality, we still denote the feature vectors of image and text by $\mathbf{x}$ and $\mathbf{z}$ for a better representation.

To verify the effectiveness of the learned similarity measurement based on embedding parameters, we follow the intuitive strategies used in Jiang et al. (2015) and assume the aligned text $\mathbf{z}_k$ should ranks higher than the other text $\mathbf{z}_j \in \mathcal{Z}$ $(j \neq k)$ given an image query $\mathbf{x}^k \in \mathcal{X}$, i.e.,

$$S(\mathbf{x}_k, \mathbf{z}_k) \geq S(\mathbf{x}_k, \mathbf{z}_j) \quad (\forall j \neq k). \tag{4}$$

In such a way, we associate each image query $\mathbf{x}_k$ a tetrad set $\mathcal{T}_{\mathbf{x}}^k = \{(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}) : j = 1, 2, \cdots, k-1, k+1, \cdots, n\}$, where y_{kj} is assigned on the basis of similarity measurement S by

$$y_{kj} = \begin{cases} 1, & S(\mathbf{x}_k, \mathbf{z}_k) \geq S(\mathbf{x}_k, \mathbf{z}_j); \\ -1, & \text{otherwise.} \end{cases} \tag{5}$$

for any $j \neq k$. As a result, the following inequality should be fulfilled for each tetrad $(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}) \in \mathcal{T}_{\mathbf{x}}^k$,

$$y_{kj}\big[S(\mathbf{x}_k, \mathbf{z}_k) - S(\mathbf{x}_k, \mathbf{z}_j)\big] \geq 0. \tag{6}$$

For each ranking text by the kth image query $\mathbf{x}_k$, we define the incurred ranking loss function as

$$\mathcal{L}(\mathcal{T}_{\mathbf{x}}^k; W) = \sum_{j \neq k} l(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}; W), \tag{7}$$

where $W = \{W_1, \mathbf{b}_1, W_2, \mathbf{b}_2\}$ collects the embedding parameters used in functions (1) and (2); $l(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}; W)$ is usually given as a hinge loss by

$$l(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, Y_{kj}; W) = \max(0, y_{kj}[S(\mathbf{x}_k, \mathbf{z}_j) - S(\mathbf{x}_k, \mathbf{z}_k)] + \Delta) \tag{8}$$

with margin $\Delta \geq 0$. This objective loss encourages aligned image-text pairs in $\mathcal{D}$ to have a higher score than misaligned pairs by a margin.

3.2. Self-paced CMLR with diversity regularization

In this part, we incorporate the SPLD into CMLR framework to enhance the embedding space learning. This idea tends to distinguish faithful tetrads $(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj})$ from easy (high-confidence) ones, and then gradually transfer the learning knowledge to recognize more complex ones. To this end, we collect the entire tetrads over all image queries into set $\mathcal{T}_{\mathcal{X}}$,

$$\mathcal{T}_{\mathcal{X}} = \{(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}) \in \mathcal{T}_{\mathbf{x}}^k : j \neq k; k = 1, 2, \cdots, n\} \tag{9}$$

and assign each tetrad $(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}) \in \mathcal{T}_{\mathcal{X}}$ a weight v_j^k to reflect the importance of ranking text $\mathbf{z}_j$ by image query $\mathbf{x}_k$. Specifically, the importance weight vector $\mathbf{v} \in \mathbb{R}^{n(n-1)}$ over set $\mathcal{T}_{\mathcal{X}}$ is defined as

$$\mathbf{v}^\top = [\underbrace{v_2^1, v_3^1, \cdots, v_n^1}_{\mathbf{v}^1 \in \mathbb{R}^{n-1}}, \underbrace{v_1^2, v_3^2, \cdots, v_n^2}_{\mathbf{v}^2 \in \mathbb{R}^{n-1}}, \cdots, \underbrace{v_1^n, v_2^n, \cdots, v_{n-1}^n}_{\mathbf{v}^n \in \mathbb{R}^{n-1}}].$$

In particular, given image query $\mathbf{x}_k$, the loss incurred by ranking text $\mathbf{z}_j$ has no effect on embedded space learning if $v_j^k = 0$, i.e., the tetrad $(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj})$ will not be evolved in the procedure of training.

With the importance weight vector $\mathbf{v}$, the idea of self-paced CMLR with diversity regularization is formalized as solving the following optimization problem:

$$\min_{W, \mathbf{v}} \ \frac{1}{2}\|W\|^2 + \sum_k \sum_{j \neq k} v_j^k \, l(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}; W) + \phi(\mathbf{v}; \lambda, \gamma) \tag{10}$$

$$s.t. \quad v_j^k \in [0, 1] \quad (\forall k, \ j \neq k)$$

where the self-pace capability with diversity is achieved through regularization function

$$\phi(\mathbf{v}; \lambda, \gamma) = -\lambda\|\mathbf{v}\|_1 - \gamma\|\mathbf{v}\|_{2,1}^* \tag{11}$$

where λ and γ are the regularizer penalty parameters which are imposed on the negative l_1-norm term (easiness term) and the $l_{2,1}$-norm-like term $\|\mathbf{v}\|_{2,1}^*$ (diversity term), respectively. Specifically, the easiness term is defined as

$$-\lambda\|\mathbf{v}\|_1 = -\lambda \sum_k \sum_{j \neq k} v_j^k. \tag{12}$$

It favors selecting from easy tetrads in $\mathcal{T}_{\mathcal{X}}$ to more complex ones. Without considering the diversity term, i.e., $\gamma = 0$, the importance weight $v_j^k \in [0, 1]$ is updated for each tetrad $(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj})$ with fixed embedding parameters W, according to

$$v_j^k = \begin{cases} 1, & l(\mathbf{x}_k, \mathbf{z}_k, \mathbf{z}_j, y_{kj}; W) \leq \lambda; \\ 0, & \text{otherwise.} \end{cases} \tag{13}$$

As a result, the tetrad with smaller loss is taken as an easy one and therefore should be learned preferentially by setting $v_j^k = 1$ and vice versa. According to this strategy, the parameter set W is updated iteratively only on the selected preferable tetrads with importance weight $v_j^k = 1$. As the increase of λ, more tetrads with larger loss will be gradually involved to learn a more "nature" model (Jiang et al., 2014a). Thus, by incorporating the estimation of importance weight vector $\mathbf{v}$ with negative l_1-norm regularization into the procedure of CMLR, we indeed achieve to learn the embedding of cross-media data in a self-paced fashion.

However, the easiness regularization might lead to some importance weight vector, for instance $\mathbf{v}^{k_0}$ with respect to the k_0th textual query, becomes zero vector, i.e., $v_j^{k_0} = 0$ for $j = 1, 2, \cdots, k_0 - 1, \cdots, k_0 + 1, \cdots, n$. It leads that all of the tetrads in set $\mathcal{T}_{\mathbf{x}}^{k_0}$ for the k_0th image query are never selected to update embedding parameter W in the next iteration. This is because the updating of importance weight vector $\mathbf{v}$ according to (13) does not consider the diversity of the selected tetrads, where the diversity is an important aspect of learning because performance is usually enhanced significantly through samples that are dissimilar from what has already been learned (Jiang et al., 2014b; Zhao et al., 2015). For this issue, it is necessary to impose the diversity regularization on the importance weights vector, such that the selected tetrads are scattered over different image queries. In this paper, we following the strategy used in Zhang et al. (2015) and define the diversity regularization as a $l_{2,1}$-norm-like term $\|\mathbf{v}\|_{2,1}^*$,

$$\|\mathbf{v}\|_{2,1}^* = \sum_k \sqrt{\sum_{j \neq k} v_j^k}. \tag{14}$$

Intuitively, the diversity regularization evolved into the objective function (10) aims to prevent the non-zero importance weights from concentrating in some image queries and ignoring others, i.e., to select dissimilar tetrads from different image queries as much as possible. Note that the diversity regularization defined in (14) makes $\|\mathbf{v}\|_{2,1}^*$ non-convex, while guaranteeing the convexity of its negative. This strategy preserves the previous axiomatic definition for SPL regularization. Moreover, it makes the solution of the optimization problem (10) more sufficient and simple.

Please cite this article as: M. Luo et al., Simple to complex cross-modal learning to rank, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.07.001

can be captured for cross-media retrieval. (3) An efficient alternative algorithm is exploited to solve the proposed challenging problem with a fast convergence in practice. Extensive experimental results over several benchmark datasets demonstrate the effectiveness and superiority of the proposed algorithm.

The remainder of this paper is organized as follows. We give a brief introduction to the related work on CMLR and SLP in Section 2. In Section 3, we firstly introduce a non-linear mapping to characterize the multi-modalities embedding space, and then we associate each ranking by cross-modal query with an importance weight to train the CMLR model in an SPL fashion with diversity. We exploit an efficient alternating algorithm in Section 4 to address the proposed challenging optimization problem. In Section 5, we conduct extensive experiments over several benchmark data sets to illustrate the effectiveness and superiority of the proposed method. Section 6 concludes this work.

2. Related work

In this section, we briefly review the related work on CMLR and self-paced learning theory and applications.

2.1. Cross-media retrieval

Grangier et al. pioneered to formalize the cross-modal retrieval tasks as a pair-wise ranking problem and maximize the final retrieval performance with a Passive-Aggressive algorithm, namely Passive-Aggressive Model for Image Retrieval (PAMIR). However, since this method verifies the pairwise ranking criterion with mapping from image query space to the document space, its performance may be deteriorated by the skewed multi-modal data. Consequently, some efforts are devoted to formalize cross-media retrieval as a list-wise ranking loss optimization problem. For example, Xu et al. propose to optimize the list-wise ranking loss with a low-rank embedding; Wu et al. (2015) learn the latent joint representation of multi-modal data through a conditional random field. Inspired by dictionary learning together with sparse coding techniques, multi-modal dictionary learning is also studied by associating each modal data with a dictionary (Jia et al., 2010; Monaci et al., 2007). Additionally, hashing technique is also employed to solve the problem of CMLR due to its efficiency for large-scale datasets (Cao et al., 2016; Wang et al., 2016b; Yu et al., 2014; Zhu et al., 2013).

Note that the mentioned approaches above commonly use linear mapping functions to translate multi-modal data into the shared space for its simplicity. However, linear mapping function might not be sophisticated enough to reveal the explicit correspondences between different modalities. For this reason, Feng et al. (2014) leverage correspondence autoencoder with deep architectures to learn the mid-level presentation of multi-modal data; Jiang et al. (2015) assume a deep compositional cross-modal semantic representation is more attractive for CMLR and optimize the pairwise ranking using non-linear mapping. These techniques have shown their effectiveness to learn a more sophisticated embedding space with large scale training collections. However, an expensive computational cost is usually required by these methods due to a large number of parameters. Additionally, the ranking performance is limited when there is not enough training data available for some real world applications.

It is noteworthy that all of the previous methods suppose the rankings of multi-modal data are of equal importance, without distinguishing each ranking's contribution to multi-modal embedding space learning. In this paper, we pioneer to assign each ranking an appropriate importance weight and use non-linear mapping to learn an optimal multi-modal embedding space in a self-paced manner.

2.2. SPL and SPLD

Enlightened by the learning principle underlying the cognitive process of humans and animals, SPL theory is raised lately to learn the model from easy samples to gradually more complex ones (Kumar et al., 2010; Meng et al., 2017). This idea is indeed an improvement of the curriculum learning which specifies a sequence of gradually added training samples (Bengio et al., 2009). Intuitively, since the samples are organized from easiness to hardness instead of using all samples simultaneously or randomly sampling, the SPL (curriculum) can effectively avoid poor local minimum and achieve a better generalization (Bengio et al., 2009). SPL is independent of particular model objectives and has been attracted increasing attention in the field of machine learning and computer vision tasks, such as object localization and tracking (Shi and Ferrari, 2016; Xiao and Lee, 2016), reranking of multimedia search (Jiang et al., 2014a; Liang et al., 2017), image classification (Gong et al., 2016; Tang et al., 2012; Tudor Ionescu et al., 2016), matrix factorization (Zhao et al., 2015) and cotraining of multi-view tasks (Ma et al., 2017). However, traditional SPL theory just considers the easiness while ignoring the diversity of the selected samples. For this issue, Jiang et al. (2014b) enhance SPL with non-convex diversity regularization such that the selected easy samples dissimilar from what has already been learned; Zhang et al. (2015) introduce a convex diversity regularization and use SPLD for co-saliency detection. In this paper, we incorporate the SPLD into CMLR to training the multi-modal embedding space gradually from easy rankings to more complex ones by diverse queries.

3. Self-paced cross-media learning to rank

In this section, we first introduce the framework of CMLR with non-linear mapping functions from image and text feature spaces to a shared embedding space. Then a self-paced CMLR model is proposed with diversity regularization to learn a more optimal embedding space in a theoretically sound manner.

3.1. Problem formulation

We associate each image with a natural language description such that the training dataset consists of n image-text pairs, i.e. $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{z}_i) : i = 1, 2, \cdots, n\}$, where $\mathbf{x}_i \in \bar{\mathcal{X}} \subseteq R^p$ represents a p-dimensional visual feature vector extracted from the ith image and $\mathbf{z}_i \in \bar{\mathcal{Z}} \subseteq R^q$ refers to a q-dimensional feature vector extracted from the ith text (sentence). For a better representation, we collect all sentences and images in $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ and $\mathcal{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n\}$, respectively. Note that the order of components in $\mathcal{X}$ and $\mathcal{Z}$ should correspond with each other such that the ith image $\mathbf{x}_i \in \mathcal{X}$ and the ith text $\mathbf{z}_i \in \mathcal{Z}$ come from the same pair in $\mathcal{D}$.

To explore the underlying correspondence between relevant text and image, a shared multimodal embedding space $\bar{\mathcal{E}} \subseteq R^d$ is learned in CMLR to facilitate the similarity measurement between different modalities. Given an image query $\mathbf{x} \in \mathcal{X}$, we define a non-linear mapping from image feature space into the shared multi-modal embedding space via $h : \bar{\mathcal{X}} \to \bar{\mathcal{E}}$,

$$h(\mathbf{x}) = \sigma(W_1\mathbf{x} + \mathbf{b}_1), \tag{1}$$

where the non-linear mapping $\sigma(\cdot)$ is specified as a Sigmoid function in this paper; W_1 is a $d \times p$ transformation matrix and $\mathbf{b}_1 \in \mathbb{R}^d$ is a bias vector. Similarly, we map each text feature into the shared embedding space by non-linear mapping $g : \bar{\mathcal{Z}} \to \bar{\mathcal{E}}$,

$$g(\mathbf{z}) = \sigma(W_2\mathbf{z} + \mathbf{b}_2), \tag{2}$$

where W_2 is a $d \times q$ transformation matrix and $\mathbf{b}_2 \in \mathbf{R}^d$ is a bias vector.

Please cite this article as: M. Luo et al., Simple to complex cross-modal learning to rank, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.07.001

Xiaojun Chang, M.L.Y.Y., Zhigang Ma, Hauptmann, A.G., 2017. Feature interaction augmented sparse learning for fast kinect motion detection. IEEE Trans. Image Process. 26 (5), 3911–3920.
You, Q., Luo, J., Jin, H., Yang, J., 2016. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In: WSDM, pp. 13–22.
Yu, Z., Wu, F., Zhang, Y., Tang, S., Shao, J., Zhuang, Y., 2014. Hashing with list-wise learning to rank. ACM SIGIR.
Zhai, X., Peng, Y., Xiao, J., 2013. Heterogeneous metric learning with joint graph regularization for cross-media retrieval. AAAI.
Zhang, D., Meng, D., Li, C., Jiang, L., Zhao, Q., Han, J., 2015. A self-paced multiple-instance learning framework for co-saliency detection. ICCV.
Zhao, Q., Meng, D., Jiang, L., Xie, Q., Xu, Z., Hauptmann, A.G., 2015. Self-paced learning for matrix factorization. AAAI.
Zhu, X., Huang, Z., Shen, H.T., Zhao, X., 2013. Linear cross-modal hashing for efficient multimedia search. ACM MM.

Please cite this article as: M. Luo et al., Simple to complex cross-modal learning to rank, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.07.001

6. Conclusion

In this paper, we employ non-linear mapping functions from heterogeneous feature spaces into a shared embedding space and incorporate the SPLD theory into the CMLR to train an optimal multi-modal embedding space gradually from easy rankings by diverse queries to more complex ones. This method adaptively distinguishes the contributions of varying rankings to the shared space learning and explicitly considers the diversity of rankings by different queries at the same time. These strategies effectively enhance model's robustness to outliers in a theoretically sound manner and improve its generalization capability with more sophisticated non-linear mapping. The comprehensive experimental results on three benchmark datasets have demonstrated the effectiveness and superiority of the proposed approach on both tasks of text query image and image queried text. We also experimentally illustrate the significant necessary of diversity regularization imposed on importance weight vector for cross-modal retrieval. A possible direction for future work may lie in studying list-wise self-paced CMRL problem based on weak-supervised learning and exploiting the potentials of the proposed model in other applications, such as attribute detection (Wang et al., 2016d), face aging (Wang et al., 2016c) and action recognition (Wang et al., 2016e).

Acknowledgment

This work was funded by the National Science Foundation (NSF) (No. IIS-1650994, No. IIS-1735591), the National Science Foundation of China (Nos. 61502377, 61532004), and China Postdoctoral Science Foundation (No. 2015M582662).

References

Amir, A., Basu, S., Iyengar, G., Lin, C.-Y., Naphade, M., Smith, J.R., Srinivasan, S., Tseng, B., 2004. A multi-modal system for the retrieval of semantic video events. Comput. Vision Image Understanding 96 (2), 216–236.
Bai, B., Weston, J., Grangier, D., Collobert, R., Sadamasa, K., Qi, Y., Chapelle, O., Weinberger, K., 2010. Learning to rank with (a lot of) word features. Inf. Retr. 13 (3), 291–314.
Barnard, K., Duygulu, P., Forsyth, D., De Freitas, N., Blei, D.M., Jordan, M.I., 2003. Matching words and pictures. J. Mach. Learn. Res. 3, 1107–1135.
Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning. ICML.
Blei, D.M., Jordan, M.I., 2003. Modeling annotated data. ACM SIGIR.
Boyd, S., Vandenberghe, L., 2004. Convex Optimization. Cambridge University Press.
Cao, Y., Long, M., Wang, J., Yang, Q., Yu, P.S., 2016. Deep visual-semantic hashing for cross-modal retrieval. KDD.
Chang, X., Ma, Z., Yang, Y., Zeng, Z., Hauptmann, A.G., 2017. Bi-level semantic representation analysis for multimedia event detection. IEEE Trans. Cybern. 47 (5), 1180–1197.
Chang, X., Yu, Y.-L., Yang, Y., Xing, E.P., 2017. Semantic pooling for complex event analysis in untrimmed videos. IEEE Trans. Pattern Anal. Mach. Intell. 39 (8), 1617–1632.
Deng, C., Tang, X., Yan, J., Liu, W., Gao, X., 2016. Discriminative dictionary learning with common label alignment for cross-modal retrieval. IEEE Trans. Multimedia 18 (2), 208–218.
Feng, F., Wang, X., Li, R., 2014. Cross-modal retrieval with correspondence autoencoder. ACM MM.
Gong, C., Tao, D., Maybank, S.J., Liu, W., Kang, G., Yang, J., 2016. Multi-modal curriculum learning for semi-supervised image classification. IEEE Trans. Image Process. 25 (7), 3249–3260.
Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S., 2013. Deep convolutional ranking for multilabel image annotation. arXiv:1312.4894.
Grangier, D., Bengio, S., 2008. A discriminative kernel-based approach to rank images from text queries. IEEE Trans. Pattern Anal. Mach. Intell. 30 (8), 1371–1384.
Habibian, A., Mensink, T., Snoek, C.G., 2015. Discovering semantic vocabularies for cross-media retrieval. ICMR.
Hardoon, D.R., Szedmak, S., Shawe-Taylor, J., 2004. Canonical correlation analysis: an overview with application to learning methods. Neural Comput. 16 (12), 2639–2664.
Hotelling, H., 1936. Relations between two sets of variates. Biometrika 28 (3/4), 321–377.
Irie, G., Arai, H., Taniguchi, Y., 2015. Alternating co-quantization for cross-modal hashing. ICCV.
Jia, Y., Salzmann, M., Darrell, T., 2010. Factorized latent spaces with structured sparsity. NIPS.
Jia, Y., Salzmann, M., Darrell, T., 2011. Learning cross-modality similarity for multinomial data. ICCV.
Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. ACM MM.
Jiang, L., Meng, D., Mitamura, T., Hauptmann, A.G., 2014. Easy samples first: Self-paced reranking for zero-example multimedia search. ACM MM.
Jiang, L., Meng, D., Yu, S.-I., Lan, Z., Shan, S., Hauptmann, A., 2014. Self-paced learning with diversity. NIPS.
Jiang, X., Wu, F., Li, X., Zhao, Z., Lu, W., Tang, S., Zhuang, Y., 2015. Deep compositional cross-modal learning to rank via local-global alignment. ACM MM.
Jin, C., Mao, W., Zhang, R., Zhang, Y., Xue, X., 2015. Cross-modal image clustering via canonical correlation analysis. AAAI.
Kang, C., Liao, S., He, Y., Wang, J., Niu, W., Xiang, S., Pan, C., 2015. Cross-modal similarity learning: a low rank bilinear formulation. CIKM.
Kang, C., Xiang, S., Liao, S., Xu, C., Pan, C., 2015. Learning consistent feature representation for cross-modal multimedia retrieval. IEEE Trans. Multimedia 17 (3), 370–381.
Karpathy, A., Joulin, A., Li, F.F., 2014. Deep fragment embeddings for bidirectional image sentence mapping. NIPS.
Kumar, M.P., Packer, B., Koller, D., 2010. Self-paced learning for latent variable models. NIPS.
Li, Z., Liu, J., Xu, C., Lu, H., 2013. Mlrank: multi-correlation learning to rank for image annotation. Pattern Recognit. 46 (10), 2700–2710.
Liang, J., Jiang, L., Hauptmann, A.G., 2017. Webly-supervised learning of multimodal video detectors. AAAI.
Lu, X., Wu, F., Li, X., Zhang, Y., Lu, W., Wang, D., Zhuang, Y., 2014. Learning multimodal neural network with ranking examples. ACM MM.
Lu, X., Wu, F., Tang, S., Zhang, Z., He, X., Zhuang, Y., 2013. A low rank structural large margin method for cross-modal ranking. ACM SIGIR.
Lux, M., Granitzer, M., Kienreich, W., Sabol, V., Klieber, W., Sarka, W., 2004. Cross media retrieval in knowledge discovery. In: Practical Aspects of Knowledge Management, pp. 343–352.
Ma, F., Meng, D., Xie, Q., Li, Z., Xuanyi, D., 2017. Self-paced cotraining. ICML.
McFee, B., Lanckriet, G.R., 2010. Metric learning to rank. ICML.
Meng, D., Zhao, Q., Jiang, L., 2017. A theoretical understanding of self-paced learning. Inf. Sci. To appear.
Menon, A.K., Surian, D., Chawla, S., 2015. Cross-modal retrieval: a pairwise classification approach. ICDM.
Mettes, P., Koelma, D.C., Snoek, C.G.M., 2016. The imagenet shuffle: Reorganized pre-training for video event detection. ICMR.
Monaci, G., Jost, P., Vandergheynst, P., Mailhe, B., Lesage, S., Gribonval, R., 2007. Learning multimodal dictionaries. IEEE Trans. Image Process. 16 (9), 2272–2283.
Pereira, J.C., Vasconcelos, N., 2014. Cross-modal domain adaptation for text-based regularization of image semantics in image retrieval systems. Comput. Vision Image Understanding 124, 123–135.
Qin, T., Liu, T., Zhang, X., Wang, D., Li, H., 2008. Global ranking using continuous conditional random fields. NIPS.
Ranjan, V., Rasiwasia, N., Jawahar, C.V., 2015. Multi-label cross-modal retrieval. ICCV.
Rasiwasia, N., Pereira, J.C., Coviello, E., Doyle, G., Lanckriet, G.R.G., Levy, R., Vasconcelos, N., 2010. A new approach to cross-modal multimedia retrieval. ACM MM.
Shi, M., Ferrari, V., 2016. Weakly supervised object localization using size estimates. ECCV.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. CVPR.
Tang, Y., Yang, Y.-B., Gao, Y., 2012. Self-paced dictionary learning for image classification. ACM MM.
Tudor Ionescu, R., Alexe, B., Leordeanu, M., Popescu, M., Papadopoulos, D.P., Ferrari, V., 2016. How hard can it be? estimating the difficulty of visual search in an image. CVPR.
Usunier, N., Buffoni, D., Gallinari, P., 2009. Ranking with ordered weighted pairwise classification. ICML.
Wang, C., Blei, D., Li, F.-F., 2009. Simultaneous image classification and annotation. CVPR.
Wang, J., He, Y., Kang, C., Xiang, S., Pan, C., 2015. Image-text cross-modal retrieval via modality-specific feature learning. ICMR.
Wang, K., He, R., Wang, L., Wang, W., Tan, T., 2016. Learning to rank with (a lot of) word features. IEEE Trans. Pattern Anal. Mach. Intell..
Wang, K., Tang, J., Wang, N., Shao, L., 2016. Semantic boosting cross-modal hashing for efficient multimedia retrieval. Inf. Sci. 330, 199–210.
Wang, W., Cui, Z., Yan, Y., Feng, J., Yan, S., Shu, X., Sebe, N., 2016. Recurrent face aging. CVPR.
Wang, W., Yan, Y., Winkler, S., Sebe, N., 2016. Category specific dictionary learning for attribute specific feature selection. IEEE Trans. Image Process. 25 (3), 1465–1478.
Wang, W., Yan, Y., Zhang, L., Hong, R., Sebe, N., 2016. Collaborative sparse coding for multiview action recognition. IEEE Multimedia 23 (4), 80–87.
Wang, Y., Wu, F., Song, J., Li, X., Zhuang, Y., 2014. Multi-modal mutual topic reinforce modeling for cross-media retrieval. ACM MM.
Weston, J., Bengio, S., Usunier, N., 2011. Wsabie: scaling up to large vocabulary image annotation. IJCAI.
Wu, F., Jiang, X., Li, X., Tang, S., Lu, W., Zhang, Z., Zhuang, Y., 2015. Cross-modal learning to rank via latent joint representation. IEEE Trans. Image Process. 24 (5), 1497–1509.
Wu, F., Lu, X., Zhang, Z., Yan, S., Rui, Y., Zhuang, Y., 2013. Cross-media semantic representation via bi-directional learning to rank. ACM MM.
Xiao, F., Lee, Y.J., 2016. Track and segment: An iterative unsupervised approach for video object proposals. CVPR.

Please cite this article as: M. Luo et al., Simple to complex cross-modal learning to rank, Computer Vision and Image Understanding (2017), http://dx.doi.org/10.1016/j.cviu.2017.07.001