Learning Multimodal Neural Network with Ranking Examples

Xinyan Lu, Fei Wu, Xi Li, Yin Zhang, Weiming Lu, Donghui Wang, Yueting Zhuang
College of Computer Science
Zhejiang University, China
{xinyanlu,wufei,xilizju,zhangyin98,luwm,dhwang,yzhuang}@zju.edu.cn

ABSTRACT

To support cross-modal information retrieval, cross-modal *learning to rank* approaches utilize ranking examples (e.g., an example may be a text query and its corresponding ranked images) to learn appropriate ranking (similarity) function. However, the fact that each modality is represented with intrinsically different low-level features hinders these approaches from better reducing the heterogeneity-gap between the modalities and thus giving satisfactory retrieval results. In this paper, we consider learning with neural networks, from the perspective of optimizing the listwise ranking loss of the cross-modal ranking examples. The proposed model, named Cross-Modal Ranking Neural Network (CMRNN), benefits from the advance of both neural networks on learning high-level semantics and *learning to rank* techniques on learning ranking function, such that the learned cross-modal ranking function is implicitly embedded in the learned high-level representation for data objects with different modalities (e.g., text and imagery) to perform cross-modal retrieval directly. We compare CMRNN to existing state-of-the-art cross-modal ranking methods on two datasets and show that it achieves a better performance.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models

Keywords

Cross-modal Ranking; Learning to rank; Representation Learning

INTRODUCTION

With the rapid development of multimedia technology, today many real-world applications involve multimodal data. For conducting cross-modal retrieval, one category of approaches are based on the rankings of the data related to the queries (e.g., an example may be a text query and its corresponding ranked images) to learn ranking functions which optimize for certain ranking loss, such as PAMIR [5], SSI [1] and LSCMR [6].

It is obvious that the intra-modality feature representation plays an important role in learning cross-modal ranking functions for these approaches: imagine how much easier when dealing with multimodal data that are all with high-level semantic features. However, each modality usually has a different kind of low-level representation and intrinsic structure. For example, text is usually represented as discrete sparse word-count vectors, whereas an image is represented using pixel intensities or low-level visual feature vectors. This makes it hard for these approaches to discover the relationships across modalities.

On the other hand, deep neural networks (DNNs) that learn a transformation of a low-level representation to a high-level representation have shown their powerful ability to the tasks of learning multimodal representation. One straightforward consideration is that train individual neural network separately for different modality and then apply existing approaches based on the features previously learned by the DNNs, which will be used as baselines in the experiments. However in this scheme, the final ranking performance is limited since it is hard for the ranking loss to be back-propagated to the pre-trained neural network which can boost the discriminative representation for cross-modal ranking.

In this work, we seek to bridge the gap between neural networks and cross-modal ranking. By adapting techniques of *learning to rank*, we propose a new multimodal neural network named Cross-Modal Ranking Neural Networks (CMRNN) to support cross-modal ranking. Specifically, CMRNN outputs the relevance scores of the retrieved documents given a query in another modality, and then the documents are ranked by their relevance scores (see Figure 1).

It is worthwhile to highlight the main contribution of the proposed model. Not only the model learns high-level feature representation for data objects with different modalities, but more importantly the ranking loss is back-propagated to train the modality-specific neural networks and thus the learned cross-modal ranking function is implicitly embedded in the learned high-level feature representation that are optimized for cross-modal retrieval performance (in other words, the learned high-level features have the discriminative power for cross-modal ranking). Therefore, the proposed model benefits from both the most recent advances in neural networks and learning to rank techniques.

Related work The authors of [8] propose a multimodal Deep Boltzmann Machine for learning a generative model of data that consists of multiple input modalities. The model works by learning a joint representation over the space of multimodal inputs, which is useful for cross-modal retrieval. Note that the multimodal DBM is trained with paired multimodal training data, and it is not optimized directly for the final retrieval performance. Methods like [7] based on autoencoders and [4] based on CCA have similar incentive.

A model termed DeViSE, proposed in [3], leverages textual data to learn semantic relationships between labels and explicitly maps images into a rich semantic embedding space via a linear transfor-

MM'14, November 3–7, 2014, Orlando, Florida, USA.

http://dx.doi.org/10.1145/2647868.2655001.

mation. The model is trained to produce a high dot-product similarity between the transformed visual output and the vector representation of the correct label. Given an image, pairwise ranking loss is first calculated between the correct label and the other labels, and then back-propagated into the core visual model to fine-tune the visual representation. In the proposed model, the explicit linear transformation alignment is absorbed into the process of learning the high-level representation; Moreover, listwise ranking strategy is used in the proposed model.

THE CMRNN MODEL

For an explicit articulation in the rest of this section, the model is only described in the case of text-query-image retrieval. We report the experiments in both scenarios of image-query-text retrieval and text-query-image retrieval.

Notation

Given a training set of N samples, each contains a text query $q^{(i)}$ $(i = 1, \ldots, N)$ as well as a list of corresponding retrieved images $D^{(i)} = (d_1^{(i)}, d_2^{(i)}, \ldots, d_{n^{(i)}}^{(i)})$, where $n^{(i)}$ denotes the size of $D^{(i)}$. Furthermore, each list of retrieved images $D^{(i)}$ is associated with their true judgments (ranking scores) $Y^{(i)} = (y_1^{(i)}, y_2^{(i)}, \ldots, y_{n^{(i)}}^{(i)})$ where $y_j^{(i)}$ denotes the judgment on image $d_j^{(i)}$ with respect to text query $q^{(i)}$. The judgment $y_j^{(i)}$ represents the relevance degree of $d_j^{(i)}$ to $q^{(i)}$, and can be a score explicitly or implicitly given by humans. For example, $y_j^{(i)}$ can be the number of clicks on $d_j^{(i)}$ given query $q^{(i)}$, or an expert judgment on the relevance level of $d_j^{(i)}$ to query $q^{(i)}$. The higher $y_j^{(i)}$, the stronger relevance.

The ranking function

A ranking function $f(q, d)$ between a text query q and an image d is to be learned according to a pre-defined ranking loss. The learned function f maps each text-image pair to a ranking score based on their semantic relevance. Given a text query q and an image d, we tend to learn the scoring function as a dot product of two vectors $f_u(q)$ and $f_v(d)$ in the K-dimensional latent space:

$$f(q, d) = f_u(q)^T f_v(d) \tag{1}$$

where $f_u(q) \in \mathbb{R}^K$ and $f_v(d) \in \mathbb{R}^K$.

To better discover modality-intrinsic structure, neural networks are used to represent the text queries and images in this work: $f_u(q)$ refers to map the query text q from its original text space to the K-dimensional latent space by a neural network, and $f_v(d)$ refers to map the retrieved image d from its original image space to the K-dimensional latent space by another neural network. Therefore, the text query and the retrieved image are mapped to a common K-dimensional latent space, and then their similarity is measured by a dot product of the two vectors in the K-dimensional space, which is commonly used to measure the matching between textual vectors.

Given a text query $q^{(i)}$, for each image $d_j^{(i)}$ in $D^{(i)}$, the ranking function outputs a relevance score $z_j^{(i)} = f(q^{(i)}, d_j^{(i)})$. For the list of images $D^{(i)}$, a list of scores $Z^{(i)} = (z_1^{(i)}, z_2^{(i)}, \ldots, z_{n^{(i)}}^{(i)})$ is obtained. The objective of learning is formalized as minimizing the empirical ranking loss with respect to the training data:

$$E = \sum_{i=1}^{N} L(Y^{(i)}, Z^{(i)}),$$

where L is a listwise loss function that we will define shortly.

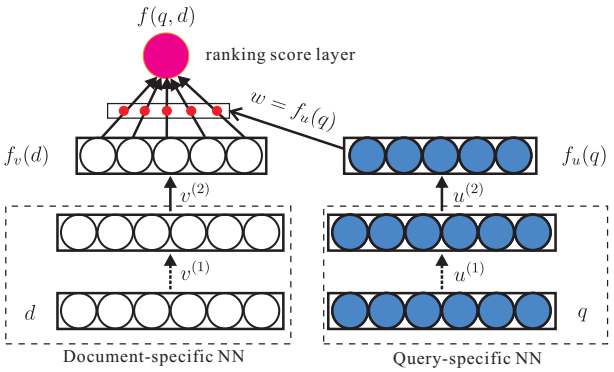


Figure 1: The architecture of the proposed cross-modal ranking neural network. If the output layers of the two modality-specific NN have unequal number of units, one can extend both the architectures by one *deeper* layer with an equal number of output units.

The architecture of the proposed cross-modal ranking neural network is depicted in Figure 1. Consider modeling each data modality (i.e., text or imagery) using separate neural networks where the text and images are represented as the outputs of the modality-specific NN, respectively. Then the document-specific NN (i.e., the image-specific NN) is extended by an additional ranking score layer with only one output unit, where the weights that connect the scoring unit are specified by the outputs of the query-specific NN (i.e., the text-specific NN). In this way, the outputs of the query-specific NN act as part of parameters of the ranking function which ranks the retrieved images by their relevance to the query text. It is shown that the output of the scoring unit is exactly $f(q, d)$ in Eq. (1) when we set the activation function of the unit to be linear, and then the learned multimodal representation is stored as the outputs of the modality-specific NN.

For conducting text-query-image retrieval, the neural network is first fed by the text query and the output of the text-specific NN is taken as the weights that connect the score unit. The images are then fed into the neural network to output their ranking scores. Finally, we rank the images in descending order of their scores.

Motivated by [2], the top one probability of an image being ranked on the top for a given text query, given the scores of all the target retrieved images, is defined as

$$P_s(d_j^{(i)}) = \frac{\exp(s_j^{(i)})}{\sum_{k=1}^{n^{(i)}} \exp(s_k^{(i)})}^{1} \tag{2}$$

where $s_j^{(i)}$ is the ranking score of image d_j with respect to query $q^{(i)}$. With the use of top one probability, given two lists of scores we use Cross Entropy as metric (note that $\sum_{j=1}^{n^{(i)}} P_s(d_j^{(i)}) = 1$), the listwise loss function becomes

$$L(Y^{(i)}, Z^{(i)}) = -\sum_{j=1}^{n^{(i)}} P_{y^{(i)}}(d_j^{(i)}) \log P_{z^{(i)}}(d_j^{(i)}) \tag{3}$$

Learning method

Denote the ranking function based on the neural network model as f, the query-specific neural network as f_u (parameterized by

[1]More precisely, $P_s(d_j^{(i)})$ should be $P_s(d_j^{(i)}, q^{(i)})$ where we have omitted the query for simplicity.

u) and the document-specific neural network as f_v (parameterized by v). The weights that connecting to the ranking score unit are denoted as w (i.e., $w = f_u(q)$). The rest is to learn the parameters u and v. Note that for simplicity we have omitted the superscript i denoting the i-th training example.

Given the parameters of the neural networks, a forward propagation is first performed for one or a mini-batch of ranking example(s), then the errors propagate backwards from the output nodes to the input nodes regarding the network's modifiable weights and finally the weights are updated by a gradient descent step. Taking derivatives of $L(Y, Z)$ with respect to w, we have

$$\nabla w = \frac{\partial L(Y,Z)}{\partial w} = -\sum_{j=1}^{n} P_y(d_j) \frac{\partial \log(P_z(d_j))}{\partial w}$$

Note that

$$\log P_z(d_j) = f(q,d_j) - \log \sum_{i=1}^{n} \exp(f(q,d_i))$$

$$\frac{\partial \log P_z(d_j)}{\partial w} = \frac{\partial f(q,d_j)}{\partial w} - \sum_{i=1}^{n} P_z(d_i) \frac{\partial f(q,d_i)}{\partial w}$$

and with some derivation we get

$$\nabla w = \sum_{j=1}^{n} (P_z(d_j) - P_y(d_j)) f_v(d_j) \tag{4}$$

The derivatives of $L(Y, Z)$ with respect to u and v are deduced in a similar manner as follows:

$$\nabla u = \sum_{j=1}^{n} (P_z(d_j) - P_y(d_j)) f_v^T(d_j) \frac{\partial f_u(q)}{\partial u}$$

$$= (\nabla w)^T \frac{\partial f_u(q)}{\partial u} \tag{5}$$

$$\nabla v = \sum_{j=1}^{n} (P_z(d_j) - P_y(d_j)) f_u^T(q) \frac{\partial f_v(d_j)}{\partial v}$$

$$= w^T \left(\sum_{j=1}^{n} (P_z(d_j) - P_y(d_j)) \frac{\partial f_v(d_j)}{\partial v} \right) \tag{6}$$

Note that we can deduce $\partial f_u(q)/\partial u$ and $\partial f_v(d_j)/\partial v$ via classical back-propagation algorithm which depend on the structure of the query-specific NN and the document-specific NN, respectively. Therefore, we can evaluate the required derivatives ∇u and ∇v, and apply the gradient descent step.

EXPERIMENTS AND RESULTS

Datesets Two public real-world datasets are used in the comparative experiments. Both the datasets are bi-modal with the image and the associated text modalities. The statistics of the two datasets are summarized in Table 1.

The Wikipedia feature articles[2] dataset consists of 2,866 images, each with a short paragraph describing the image. The images are labeled with exactly one of the 10 different semantic classes, such as art and geography. For text, we extract 5,000D bag-of-words feature vectors with the TF-IDF weighting scheme. For images, 1,000D bag-of-visual-words feature vectors are extracted by clustering SIFT points with k-means. A target document is relevant to a query if they belong to the same semantic class. The NUS-WIDE[3] dataset contains 133,208 images with 1,000 tags and 81 concepts, which are pruned from the NUS dataset by keeping the images that have at least one tag and one concept. For the feature representation, we use the publicly available 1,000D text feature vector (namely tags) and 500D image feature vector based on SIFT BoVW kindly provided by the authors. A target document is relevant if it shares at least one concept with the query. However, we also note that the relevant judgement can be obtained from the abundance of users' clickthrough data with little overhead.

[2] http://www.svcl.ucsd.edu/projects/crossmodal/

[3] http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm

Table 1: The statistics of the datasets used.

	Wikipedia	NUS-WIDE
Avg. # of words/image	117.5	7.73
# of neurons in Image NN	1000-50	500-50
# of neurons in Text NN	5000-50	1000-50
Documents	1,500/500	13,320/23,977
Partition[a]	866	95,911
Queries	1,500/500	13,320/2,000
Partition[a]	866	2,000

[a] Partitions are ordered by training/validation/testing.

Experimental setup The training examples are generated as follows: for each text (image) query, we randomly selected 40 images (text documents) in the other modality in the training set as candidates and then the selected target documents are automatically labeled as relevant or irrelevant. We observed little difference when different sampling strategies were applied. While for validation and testing, we randomly select text (image) queries and all the images (text documents) in the other modality in the validation/test set are regarded as retrieval candidates.

The structures of modality-specific neural networks are listed in Table 1. The proposed model is trained with sigmoid activation, momentum of 0.3 and weight decay of 0.0001. We adjust the learning rate manually from 0.01 to 0.0001. The size of each mini-batch is set to 100. Autoencoder with one hidden layer (50 units) is used to pre-train the modality-specific neural networks. The learned features by autoencoders are also served as the input features to all the comparative methods.

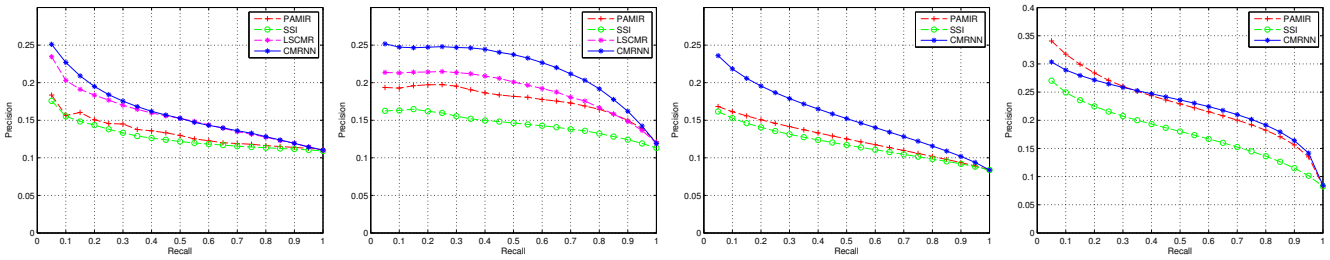
Evaluation metric The standard ranking performance metric *Mean Average Precision* (MAP) is used for comparison. Let $p^* = \text{rank}(y)$ (true ranking with two rank value $+1$ and -1) and $p = \text{rank}(z)$ (predicted ranking with a total order). Given a query and a set of R retrieved target documents, the *Average Precision* (AP) is defined as

$$\text{AP}(p^*, p) = \frac{1}{M} \sum_{j=1}^{R} Prec(j) \cdot Rel(j) \tag{7}$$

where M is the number of the relevant documents in the retrieved set, $Prec(j)$ the percentage of the relevant documents in the top j documents in predicted ranking p and $Rel(j)$ an indicator function equaling 1 if the item at rank j in predicted ranking p is a relevant document, zero otherwise. We then average the AP values from all the queries in the query set to obtain the MAP score. In the experiments, R is the number of the retrieved documents to be examined, where we set $R = 50$ or $R = all$ for all the retrieved documents.

Performance comparison We compare CMRNN with other state-of-the-art models. All the comparative models (PAMIR[5], SSI[1] and LSCMR[6]) are elaborately chosen to be trained with ranking examples for fair comparison.

Table 2 reports the performance of CMRNN and the other comparative models on the test set of the Wiki dataset, showing that CMRNN outperforms all the comparing methods on both direc-



(a) Wiki Text Query (b) Wiki Image Query (c) NUS Text Query (d) NUS Image Query

Figure 2: Precision-Recall curves on the two datasets.

Table 2: The performance comparison in terms of MAP@R scores on the Wiki dataset.

	Text Query		Image Query	
	$R = 50$	$R = all$	$R = 50$	$R = all$
PAMIR	0.2344	0.1361	0.1763	0.1786
SSI	0.2058	0.1309	0.1834	0.1470
LSCMR	0.2520	0.1597	0.2010	0.1919
CMRNN	**0.2712**	**0.1649**	**0.2563**	**0.2216**

Table 3: The performance comparison in terms of MAP@R scores on the NUS dataset.

	Text Query		Image Query	
	$R = 50$	$R = all$	$R = 50$	$R = all$
PAMIR	0.2026	0.1265	**0.4682**	**0.2331**
SSI	0.2163	0.1200	0.4259	0.1830
LSCMR	-	-	-	-
CMRNN	**0.3193**	**0.1569**	0.4402	0.2319

tions of the retrieval. The Precision-Recall curves on both directions are reported in Figure 2(a) and 2(b).

The improvement of CMRNN on the NUS dataset is not as significant as that on the Wiki dataset. The MAP scores of all the methods over NUS dataset are shown in Table 3 and the Precision-Recall curves are reported in Figure 2(c) and 2(d). For text-query-image retrieval, CMRNN outperforms the other comparative methods again, while for the other direction, PAMIR have a slightly better overall performance than CMRNN. The reason why PAMIR even beat CMRNN in the case of image-query-text retrieval may be as follows. As shown in Table 1, the average number of words per image in the NUS dataset is much smaller than that in the Wiki dataset. The short text doesn't provide sufficient word occurrences and thus is fuzzy. Given an image query, fine-tuning the modality-specific NN by the ranking loss with respect to fuzzy short text may play a counteractive effect and degrade the performance.

It is observed that LSCMR performs most closely to CMRNN on the Wiki dataset, however LSCMR fails to train a ranking model on the NUS dataset. For NUS dataset, we sample 10% of the documents to form the training set, and thus for each direction of retrieval about 13k queries (with 530k corresponding documents) are used for training. Note that the weight updating step of LSCMR requires the full batch of training examples, which makes it hard for LSCMR to deal with such large set of ranking examples. By contrast, the weight updating step of the other three comparative methods (PAMIR, SSI and CMRNN) requires only a mini-batch of ranking examples, and consequently the three methods (including the proposed CMRNN) can benefit from the large amount of ranking examples.

CONCLUSION

In this paper, a new multimodal model is presented to solving the problem of cross-modal ranking. Benefiting from both neural networks and learning to rank techniques, the proposed model can learn high-level feature representation which has the discriminative power for cross-modal ranking and the learning procedure is shown efficient for practice. We have also demonstrated the effectiveness of the proposed method CMRNN and have shown significant improvements over the comparative methods on two real-world datasets.

ACKNOWLEDGEMENTS

This work is supported in part by National Basic Research Program of China (2010CB327900), NSFC (61103099), 863 program (2012AA012505), the Fundamental Research Funds for the Central Universities, Chinese Knowledge Center of Engineering Science and Technology (CKCEST), Program for New Century Excellent Talents in University, and Zhejiang Provincial Natural Science Foundation of China (LQ13F020001, LQ14F010004).

REFERENCES

[1] B. Bai, J. Weston, D. Grangier, R. Collobert, K. Sadamasa, Y. Qi, O. Chapelle, and K. Weinberger. Learning to rank with (a lot of) word features. *Information Retrieval*, 13(3):291–314, 2010.

[2] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136. ACM, 2007.

[3] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.

[4] A. Galen, A. Raman, B. Jeff, and L. Karen. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[5] D. Grangier and S. Bengio. A discriminative kernel-based approach to rank images from text queries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(8):1371–1384, 2008.

[6] X. Lu, F. Wu, S. Tang, Z. Zhang, X. He, and Y. Zhuang. A low rank structural large margin method for cross-modal ranking. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, pages 433–442. ACM, 2013.

[7] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 689–696, 2011.

[8] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems 25*, pages 2231–2239, 2012.