**Xi’an Jiaotong University**

Project for Academic English (I)

1st Term of School Year 2018-2019

**Title of the Project**

Name: Zhang Chaofeng(张超峰)

Student ID: 3118311115

School: Xi’an Jiaotong University

Research Field: Computer Science

English Class: Class 38, Xingqing Campus

English Instructor: Fang Wenjuan

Submission Date: 10/01/2019

****

**Abstract**

In this paper, we propose a novel method with comprehensive distance-preserving autoencoders (CDPA) to address the problem of unsupervised cross-modal retrieval. The CDPA consists of four components. First, denoising autoencoders are used to retain the information from the representations and to reduce the negative inﬂuence of redundant noises. Second, a comprehensive distance-preserving common space is proposed to explore the correlations among diﬀerent representations. Third, a novel joint loss function is defined to simultaneously calculate the reconstruction loss of the denoising autoencoders and the correlation loss of the comprehensive distance-preserving common space. Finally, an unsupervised cross-modal similarity measurement is proposed to further improve the retrieval performance. The CDPA is tested on four public datasets with two cross-modal retrieval tasks: “query images by texts” and “query texts by images”.

**Keywords:** cross-modal; retrieval

**Introduction**

Over the last decade, the Internet and social media have developed rapidly, and the volume of multimedia data on the Internet has increased tremendously. Multimedia data on the Internet exists as a range of different media types and comes from heterogeneous data sources, for example, a webpage may contain text, audio, images, and video. Although these data are represented by different modalities they have a strong semantic correlation. Cross-modal retrieval is designed for scenarios where the queries and retrieval results are from different modalities [1], [2].

The key challenge of cross-modal retrieval is how to measure the similarity between representations of diﬀerent media types [3], which is referred as the heterogeneity gap. The current mainstream methods for solving the heterogeneity gap are common space learning methods, which are designed to learn an intermediate common space for features of diﬀerent media types and measure their similarities in the intermediate common space [3]. These methods can be grouped into three categories according to the need for label information: the supervised methods [5], the semi-supervised methods [6], and the unsupervised methods [1].

The supervised and semi-supervised methods require label information. However, collecting label information is usually time-consuming and expensive in practice. The unsupervised methods rely only on cross-modal data without any additional information.

Even though the previous unsupervised cross-modal retrieval methods perform well, there are still two problems to solve: 1) how to reduce the negative inﬂuence of redundant noises in features, and 2) how to directly use the relationships between representations of diﬀerent objects.

In this paper, we propose a novel method called unsupervised cross-modal retrieval with comprehensive distance preserving autoencoders (CDPA) to address the two problems mentioned above.

To alleviate the noise problem, we use denoising autoencoders. We notice that the procedure, which sets part of the input elements as zeroes, simulates the removal of the redundant noises from the inputs.

To explicitly calculate the similarity between features from diﬀerent modalities, CDPA proposes a comprehensive distance-preserving common space to consider all of the relationships between two representations, no matter whether they belong to the same object or not. The preserved distances include pairwise distances, heterogeneous distances, and homogeneous distances

**Denoising Autoencoder**

the CDPA consists of four parallel denoising autoencoders. Two of the denoising autoencoders are related to image features, and the rest are related to text features. The denoising autoencoders, which are responsible for the same modality, share the same parameters, so representations of the same modalities also have the same transforms. In each iteration, four representations between two modalities extracted from two objects are used as inputs. In this way, the relationships of diﬀerent representations can be calculated on the code layer, no matter whether these representations belong to the same object or same modality.

In the denoising autoencoder, a fixed number of input components are randomly set to zero, while the others are left untouched [7]. This procedure simulates the removal of the redundant noises from the inputs; therefore, it reduces the negative inﬂuence of such noises. In addition, the zeroing process can be viewed as a process of data augmentation, and it strengthens the connections between the local structures within the representation that were drawn from diﬀerent modalities

**Distance-Preserving Common Space**

The comprehensive distance-preserving common space is defined on the code layer by preserving three kinds of distances among the input representations. The preserved distances include pairwise distances, heterogeneous distances, and homogeneous distances. All the distances are measured according to the distances between corresponding objects in the image and text modalities.

The CDPA uses the cosine distance to measure the similarity of features in the same media spaces. The distance denote as: D(x , y)

there are three kinds of distances in the comprehensive distance-preserving common space: pairwise distances, heterogeneous distances, and homogeneous distances. The distance is calculated as

The pairwise distance loss is defined as:

L-pair = D(vi, ti) + D(vj, tj)

The heterogeneous distance loss is defined as:

L-heter = |D(vi, tj) - d| + |D(vj, ti) - d|

The homogeneous distance loss is defined as:

L-homo = |D(vi,vj) - d| + |D(ti, tj) - d|

Thus, the correlation loss in the comprehensive distance-preserving common space is defined as follows.

L-corr = L-pair + λ1(L-heter + L-homo)

Here, λ1 is the parameter to trade oﬀ between the pairwise distance loss and other distance losses…

**Evaluation Metric**

Two cross-modal retrieval tasks are used for testing: text retrieval from an image query (Img2Txt) and image retrieval from a text query (Txt2Img). The retrieval performance is evaluated using Mean Average Precision (MAP). Given one query and the first R top-ranked retrieved data, the average precision (AP) is defined as:

AP=1/M \* sum(M(k)/k \* Rel(k)) (k is count, from 1 to R)

Here, M is the number of relevant data in the retrieved results, Mk is the number of relevant items in the top k returns, and Rel(k) represents the relevance of a given rank. Rel(k) = 1 if the item ranked at the kth position is relevant; otherwise, it is zero. The MAP is obtained by averaging the AP of all the queries. We report the MAP@50 (R = 50) in all experiments following [6]. In addition, we also display the precision-scope curves for all of the methods.

**Conclusion**

In this paper, a novel unsupervised cross-modal retrieval method, the Comprehensive Distance-Preserving Autoencoders (CDPA), is proposed to solve the problem of redundant noises and to further explore the correlations among representations from diﬀerent modalities. The CDPA includes four components: denoising autoencoders, a comprehensive distance-preserving common space, a joint loss function, and an unsupervised cross-modal similarity measurement.

**Bibliography**

[1] Galen Andrew, Raman Arora, Jeﬀ Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In International Conference on Machine Learning. 1247–1255

[2] Fangxiang Feng, Ruifan Li, and Xiaojie Wang. 2015. Deep correspondence restricted Boltzmann machine for cross-modal retrieval. Neurocomputing 154 (2015), 50–60.

[3] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In Proceedings of the 22nd ACM international conference on Multimedia. ACM, 7–16

[4] YuxinPeng, XinHuang, and YunzhenZhao.2017. An over view of cross-media retrieval: Concepts, methodologies, benchmarks and challenges. IEEE Transactions on Circuits and Systems for Video Technology (2017).

[5] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. 2018. CCL: Cross-modal Correlation Learning With Multigrained Fusion by Hierarchical Network. IEEE Transactions on Multimedia 20, 2 (2018), 405–420.

[6] Yuxin Peng, Xiaohua Zhai, Yunzhen Zhao, and Xin Huang. 2016. Semi-supervised cross-media feature learning with unified patch graph regularization. IEEE Transactions on Circuits and Systems for Video Technology 26, 3 (2016), 583–596

[7] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th international conference on Machine learning. ACM, 1096–1103.