A New Approach to Cross-Modal Retrieval

Tingting Gou, Libo Liu, Qian Liu and Zhen Deng
College of Information Engineering, Ningxia University, Ningxia, China
E-mail address: gttnxdx@163.com, Liulib@163.com, 578335442@qq.com, dengzhen1025@163.com

Abstract. The semantic correlation matching algorithm, which attempts to integrate semantic information based on the unified space obtained by CCA learning, achieves an improvement in the effect of cross-modal retrieval. However, there is still a lot of room for the algorithm to make further progress in the generation of shared subspace. This paper proposes a semantic correlation matching algorithm based on the single-layer feed forward network. It generates features through neural networks to represent weight and bias. Meanwhile, it maximizes the correlation between different modalities by means of cross-coupling. The experimental results show that the algorithm proposed in this paper is superior to the compared method in terms of cross-modal retrieval task of image-text and has better cross-modal retrieval ability.

1. Introduction
With the rapid development of information technology, the digital data generated in various industries has been increasing rapidly, and has clear correlation between different representations of the same object. Information retrieval problems for single modalities, such as image retrieval and text retrieval, have been extensively studied. However, the researches cannot support a retrieval that occurs between one modality and any other modality. The explore of new technologies to mine the effective information of data, learning the similarity between heterogeneous attributes, and accurately retrieving related content required by users has become one of the challenges in the current information retrieval field. The cross-modal retrieval technology provides a new way to solve the above problems. It aims to look for the relationship among different modalities of same data. The core of cross-modal retrieval is to learn the multimodal representation space, so that the data from different modalities can be compared.

This paper focuses on the mutual retrieval between the images and texts. Many researchers have proposed various methods for it. In general, cross-modal retrieval generally focuses on two aspects: feature representation and similarity measure. As shown in Figure 1, the texts and images can be mapped to the shared subspace after analyzing the correlation between the data and obtaining the corresponding projection matrix. Therefore, cross-modal retrieval can be transformed into traditional single-modal retrieval.
Figure 1. Subspace mapping

In the cross-modal retrieval based on machine learning, Canonical Correlation Analysis (CCA) has been widely used and studied as one of mainstream methods. It uses the idea of Principal Component Analysis (PCA) to study the same modal data as a whole, rather than analyzing the components inside the data. Then, CCA extracts representative new comprehensive variables in each set of variables, and maximizes the correlation of integrated variables. Finally, it uses the method of correlation joint dimension reduction to reduce the data to the associated subspace, and the data correlation of different modalities is also preserved. Since then, many researchers have extended the CCA method. Rasiwasia et al. proposed a Semantic Correlation Matching (SCM) algorithm based on CCA, which uses the posterior probability to compute similarity of the data in the semantic space. Although, this method has a higher improvement in retrieval performance, it can be improved in terms of the generation of shared subspace.

According to the above analysis, this paper proposes a semantic correlation matching algorithm based on the single-layer feed forward network version (SLFN). First, the algorithm uses the CCA to learn the maximum shared subspace between the image and the text, and takes the obtained related features as the underlying features. Then, the similarity measurement is obtained by calculating the distance between the semantics through the single-layer neural network. The algorithm not only has the merits of the original algorithm, but also (1) needs only to set the number of hidden layer nodes of the network, (2) does not need to adjust the input weights and bias of the network during the execution process, and generates a unique solution, (3) achieves superior performance in terms of generalization.

2. Related Work

In the research of cross-modal, some studies have solved the obstacles of different modal feature sizes by constructing a low-dimensional public space, which is the commonly mentioned subspace learning method. The method projects high-dimensional features extracted from different modalities into the low-dimensional common space by learning a set of mapping functions. Then, the similarity between different modalities can be calculated at the same feature dimension.

As one of the most popular unsupervised subspace learning methods, Canonical correlation...
analysis (CCA) has been widely used in various retrieval methods. It was proposed by Hotelling and it measures linear relationship between two multi-dimensional variables. We mainly use it to establish intermodal relationships between data of different modalities by learning common subspaces. Specifically, CCA learns the two directions and makes the data most relevant along the (x, y) data modalities. It is different from the standard dimensionality reduction techniques, such as principal component analysis (PCA) or linear discriminant analysis (LDA), as we mentioned above, dimensionality reduction is performed simultaneously across two (or more) modalities. In recent years, cross-modal retrieval based on deep learning has been extensively and deeply studied, but cross-modal retrieval based on CCA still has a long way to go, because it can provide ideas and supplements for the follow-up related research.

The following methods have been proposed for joint dimensionality reduction. Semantic Matching (SM) was proposed in [1], in this approach two mappings are implemented using classifiers of the two modalities. Then each modality is represented as vector of posterior probabilities, which serves as the common feature representation. Weakly-Paired Maximum Covariance Analysis (WMCA) proposed in [2], learns a correlated discriminative feature space without the need for pair-wise correspondences. However, WMCA is based on maximum correlation analysis while the proposed work extends CCA for a similar problem setting. By directly separating the classes in the joint feature space, Multi-view Discriminant Analysis (MvDA) proposed forgoes the free parameters [3]. While it is not clear how correlated the samples from different modalities are. Weakly-Paired Maximum Covariance Analysis (WMCA) learns a correlated discriminative feature space without the need for pair-wise correspondences [2], but it is based on maximum correlation analysis. Cluster Canonical Correlation Analysis Extends CCA for a similar problem setting [4]. Except for the methods mentioned above, still some other theories are proposed for improvement. A sparse canonical correlation analysis is proposed in [5], it adds $I_p$ constraints on the canonical vectors and shows how to solve the efficiently using linearized alternating direction method of multipliers (ADMM) and using TFOCS as a black box.

With the rise of deep learning technology, some classic methods based on deep learning for cross-modal retrieval are proposed in [6-8]. A deep and two-way representation learning model is proposed in [6], which uses two convolution-based networks to complete the representation of learning images and text through which images and text are mapped into a common space. By associating the hidden representations of two single-mode automatic encoders, a new model involving a corresponding automatic encoder (Corr-AE) is modeled in [7]. Besides, it is recommended to use the restricted communication Boltzmann machine (Corr-RBM) to map the original features of the bimodal data (such as images and text in our settings) to the low-dimensional public space [8]. Then the two RBMs constructed for the image and text respectively pass the correlation loss functions are connected at their respective hidden presentation layers.

There are many kinds of cross-modal retrieval methods, beyond the approaches mentioned above, some other methods are based on machine learning neural network. By employing two independent models of deep neural networks, the approach can map the low-level features of images and texts to their semantic subspace is proposed in [9]. It uses the low-level features of training images and texts with their labels to train two independent models of deep neural networks. Then it inputs the low-level features of testing images and texts to the trained models of neural networks, and regards the top-level outputs of networks as the semantic subspace of testing images and texts. A new way based on a novel combination of Convolutional Neural Networks over images and Recurrent Neural Networks over sentences is proposed [10], by effectively unifying joint multimodal embedding and cross-modal hashing, it can improving cross-modal retrieval efficiency.

3. Canonical Correlation Analysis
Canonical correlation analysis, a method for studying the correlativity between two sets of variables, is aimed to identify and quantify their internal relationship.

Given two sets of vectors, $A = [a_1, a_2, ..., a_p], A \in R^{N \times p}$ and $B = [b_1, b_2, ..., b_p], B \in R^{N \times p}$, the $X = u_i^T A$ and $Y = v_i^T B$, which are linear combinations respectively, are used to study the
correlativity between the primitive variables A and B.

CCA seeks a pair of vectors, named as u and v, to maximize the correlation coefficient $\rho(X, Y)$,

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}} = \frac{u^T \Sigma_{12} v}{\sqrt{u^T \Sigma_{11} u} \sqrt{v^T \Sigma_{22} v}}$$

where $\Sigma_{12}$ is a sample covariance matrix between A and B, $\Sigma_{21}$ between B and A, and $\Sigma_{12} = \Sigma_{21}$. $\Sigma_{11}$ and $\Sigma_{22}$ are the covariance matrices of A and B, respectively.

The correlation coefficients of these random variables do not change if they are multiplied by a constant; some constraints (Eq. (2)) are included in Eq. (1) in order to prevent computational duplication.

$$\text{Var}(X) = u^T \Sigma_{11} u = 1, \text{Var}(Y) = v^T \Sigma_{11} v = 1$$

Let $M = \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}$ and $N = \Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}$. Then, the solution of Eq. (1) can be obtained by solving either of the following two eigenvalue problems

$$Mu = \lambda^2 u, \quad Nu = \lambda^2 v$$

where the square roots of the eigenvalues $\lambda^2$, obtained from Eq. (3), are called canonical correlations, and the vectors $a$ and $b$ are the eigenvectors corresponding to $A$ and $B$, respectively. Consequently, we acquire the $i$th set of canonical variables

$$X_i = u^{(i)} A, Y_i = v^{(i)} B, i = 1, 2, \ldots, p$$

as well as the $i$th canonical correlation coefficient $\lambda$.

4. Semantic Correlation Matching Algorithm Based on Single-Layer Feed Forward Neural Network

The diagram of the proposed the semantic correlation matching algorithm based on single-layer feed forward neural network is shown in Figure 2. The process of our method making the data of different modalities into a common semantic space by introducing a single-layer feedforward neural network is supervised. The method ensures that the data of different modalities maintain correlation in the semantic space, so that each modal can learn its own discriminative semantic features, while maintaining the semantic correlation between them.
Feed forward neural network is a kind of artificial neural network. The algorithm principle uses the sum of the Error Sum of Squares (SSE) of the simulated output value and the expected output value as the error of the neural network according to the maximum likelihood estimation idea. The network weights are adjusted by means of forward propagation of errors from input layer to output layer.

Similar to the Semantic Correlation Matching (SCM) algorithm, this paper also uses CCA to project the underlying features of the data from different modalities into their isomorphic subspace. Suppose $X = [x_1, x_2, ..., x_n]^T \in \mathbb{R}^{m \times n}$ and $Y = [y_1, y_2, ..., y_n]^T \in \mathbb{R}^{m \times n}$ are sample sets of image and text data in isomorphic subspace respectively. $M$ represents the dimension of features and $n$ is the number of samples. The network model with $m$ hidden layer nodes projected into their respective semantic spaces is

$$
\sum_{i=1}^{m} P_{i}^{IT} g(w_i \cdot x_j + b_i) = S_{j}^{IT}, \quad \sum_{i=1}^{m} P_{i}^{IT} g(w_i \cdot y_j + b_i) = S_{j}^{IT}, \quad j = 1, ..., n
$$

(5)

Among them, $w_i$ and $b_i$ are the input weight and bias of the $i$-th hidden layer node respectively, and $g(\cdot)$ is the activation function. $P_{i}^{IT}$ and $P_{i}^{IT}$ are the projection matrices, and $S_{j}^{IT}$ and $S_{j}^{IT}$ are the corresponding projective results in the two modalities.

The semantics of the different modalities of the same sample are correlated, so $S_{j}^{IT}$ and $S_{j}^{IT}$ should also have good semantic correlation. Therefore, cross-coupling between different modalities in the semantic space gives

$$
\sum_{i=1}^{m} P_{i}^{IT} g(w_i \cdot x_j + b_i) = S_{j}^{IT}
$$

$$
\sum_{i=1}^{m} P_{i}^{IT} g(w_i \cdot y_j + b_i) = S_{j}^{IT}
$$

(6)

In order to maintain the semantic correlation between the modalities, the objective function is obtained.

$$
F = \min_{P_{IT}, P_{IT}} \alpha \left\| S_{j}^{IT} - P_{IT} H(I) \right\|^2 + \beta \left\| S_{j}^{IT} - P_{IT} H(T) \right\|^2 + \gamma \left\| P_{IT} \right\|^2 + \left\| P_{IT} \right\|^2
$$

(7)
Among them, $\alpha$ and $\beta$ are equilibrium parameters and $\gamma$ is a regularization parameter.

In order to calculate $\frac{\partial F}{\partial P_{IT}}$ and $\frac{\partial F}{\partial P_{TI}}$, this paper uses the method of cross-iteration to optimize.

For example, when calculating $\frac{\partial F}{\partial P_{IT}}$, fix the value of $P_{TI}$ firstly and get

$$\frac{\partial F}{\partial P_{IT}} = 2\alpha(S^{IT} - P_{IT}H(I)) = 0$$  \hspace{1cm} (8)

$$P_{IT} = P_{IT} H(T)(H(I) - \frac{\gamma}{\alpha} I)^{-1}$$  \hspace{1cm} (9)

$I$ is the identity matrix.

In a similar way,

$$P_{IT} = P_{IT} H(T)(H(I) - \frac{\gamma}{\alpha} I)^{-1}$$  \hspace{1cm} (10)

Therefore, each parameter of the optimization objective function can be obtained by continuously iteratively updating.

5. Experiment

The configuration of the experiment is introduced in this section. And the experimental results are shown and analyzed. Data sets collected from "Wikipedia Feature Articles" to verify the effectiveness of the experiment.

Wikipedia editors have been updating and collecting 2,700 articles since 2009 and commented on them. In addition, Wikipedia classifies each feature article into one of 29 categories, which are assigned to the text and image components of each article. It contains 2866 pairs of images/text, which belongs to 10 semantic categories. Random segmentation is used to generate 2173 training sets and 693 test sets.

A 2296-dimensional feature vectors are extracted for image representation and the word bag model is constructed for text representation. Each image is represented by 500 visual word vectors based on SIFT description, and each text is represented by a 1,000-word words word package.

5.1. W Parameter Tuning Experiment

In our method, the experiment starts with parameter experiment. Figure 1 shows the MAP scores of the image retrieval text and the text retrieval image corresponding to the different number of hidden nodes in the proposed algorithm, and the average of the two MAPs. The number of hidden nodes is set from 1 to 200. It is evidence that the MAP score is the highest when the number of hidden nodes is 19. With the increase of the number of hidden nodes, the three curves in Figure 1 decrease rapidly. Therefore, the hidden nodes of network in our method is set to 19.
5.2. Activation Function Parameter Tuning Experiment

After determining the number of best hidden nodes, the experimental results of a comparative experiment for different activation functions are shown in Figure 2. It is shown that the MAP score of the activation function "sig" is higher than the other four activation functions, whether it is the process of image retrieval or text retrieval. Therefore, "sig" is selected as the activation function in the subsequent experiments.

5.3. Distance Function Parameter Tuning Experiment

Fig. 3 shows the MAP scores corresponding to different distance calculation functions in the retrieval process. According to [11], the distance measurement available for use are: (a) L1-Norm: 
\[ d(x, y) = \sum_{i=1}^{n}|x_i - y_i| \]; (b) L2-Norm: 
\[ \cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \]; (c) KL: 
\[ D(P||Q) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)} \]; (d) NC: 
\[ NC(i, j) = \frac{\sum_{m,n} T(m,n) S^2(m,n)}{\sqrt{\sum_{m,n} T^2(m,n) \sum_{m,n} S^2(m,n)}} \] and (e) CC: 
\[ \cos(\theta) = \frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \times \sqrt{\sum_{i=1}^{n}(y_i)^2}} \]. Although the effect of each function is not much different in general, "L2" still wins by a small margin in the two retrieval processes. Thus, in the subsequent experiments, "L2" function is utilized as the distance function in the retrieval.
5.4. MAP Experiment

The optimal settings of the algorithm are determined through the above experiments. Therefore, the proposed method is compared with three methods including CM, SM, SCM and RBM. Two tasks, which include image retrieval text and text retrieval image, are mainly considered. In the first case, each image is used for a query, and then all the text after the experiment is sorted. In the second case, the roles of text and images are reversed.

Table 1 demonstrates the MAP score of the different methods. The proposed method is based on the unified space obtained by CCA, and the semantic category information is improved. The retrieval efficiency is also 13% higher than the CCA method. It is obvious that several of the above-mentioned methods are effective. However, the average MAP score obtained by this method is 34%, which is at least 10% higher than the SVM method, the SCM method combined with CCA and SVM, and the RBM method. It is shown that the algorithm can greatly improve the efficiency of cross-modal retrieval. Moreover, the method proposed in this paper is more efficient in the process of image retrieval of text than the process of text retrieval, however, the efficiency of text retrieval image needs to be improved.

| Experiment | Image Query | Text Query | Average |
|------------|-------------|------------|---------|
| CM         | 0.2449      | 0.1929     | 0.2189  |
| SM         | 0.2264      | 0.2184     | 0.2224  |
| SCM        | 0.2684      | 0.2276     | 0.248   |
| RBM        | 0.1179      | 0.1179     | 0.1179  |
| Our method | 0.5333      | 0.1662     | 0.34975 |

5.5. Classification Prediction Experiment

Taking "cat" category in Wikipedia data set as an example, four methods are experimented to predict the probability of "cat" category. Figure 4 shows that the proposed method is superior to other methods in both text retrieval and text retrieval and is more accurate in judging categories. However, the probability of the second process is lower than that the first process, which may be due to the fact that category tags are more directly applied to text than images. Consider the example article in Figure 1, the text description clearly belongs to the "history" category, but the image of the castle can also.

Classified as "Art and Architecture" or "Geography and Place" category. It needs to be clear that misclassifications such as these are often justified. This is especially true for text queries. For example, the non-visual art of "movie," "drama," and "music" are often confused with each other, rather than an unrelated category like "sports." Because the above three categories share similar words and images.
6. Conclusion
This paper proposes a semantic correlation matching algorithm based on the single-layer feed forward network for cross-modal retrieval. The algorithm introduces the neural network into the semantic correlation matching method and uses the single-layer neural network to calculate the distance between the semantics to measure the similarity. Comparing with the original algorithm on the "Wikipedia Featured Article" database, the experimental results show that the algorithm proposed in this paper has good advantage in both algorithmic efficiency and the performance of retrieval. Next step we will consider introducing nonlinear feature mapping relationships into the model. It makes the model obtained by learning not confined to linear projection and improve the ability of cross-modal retrieval while maintaining high computational efficiency.

7. Competing interests
The authors declare that they have no competing interests.

8. Acknowledgements
This work is supported by the National Nature Science Foundation of China (No. 61751215, 61862050), Scientific Research and Innovation Project of China (No. ZKZD2017005) and Natural Science Foundation of Ningxia (No. NZ17010).

9. References
[1] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos, "A new approach to cross-modal multimedia retrieval," In Proceedings 18th ACM International Conference on Multimedia, 2010. pp. 251-260.
[2] C. H. Lampert, and O. Krömer, "Weakly-Paired Maximum Covariance Analysis for Multimodal Dimensionality Reduction and Transfer Learning," Computer Vision – ECCV 2010. pp. 566-579.
[3] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-View Discriminant Analysis," In Computer Vision–ECCV 2012, pages 808–821. Springer, 2012. pp. 808-821.
[4] N. Rasiwasia, V. Mahadevan, G. Aggarwal, and D. Mahajan, “Cluster Canonical Correlation Analysis,” Aistats, 2014.

[5] J. A. Ramey, Ii, “Selected Topics in High-Dimensional Statistical Learning,” Dissertations & Theses - Gradworks, 2012.

[6] F. Feng, X. Wang, and R. Li, “Cross-modal Retrieval with Correspondence Autoencoder,” pp. 7-16, 2014.

[7] F. Feng, R. Li, and X. Wang, Deep correspondence restricted Boltzmann machine for cross-modal retrieval: Elsevier Science Publishers B. V., 2015.

[8] N. Srivastava, and R. Salakhutdinov, "Multimodal Learning with Deep Boltzmann Machines," International Conference on Neural Information Processing Systems, 2012. pp. 2222-2230.

[9] Z. bin, “Research on cross-modal multimedia retrieval method based on neural network,” master, shandong normal university, 2018.

[10] Y. Cao, M. Long, J. Wang, Q. Yang, and P. S. Yu, "Deep Visual-Semantic Hashing for Cross-Modal Retrieval," ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016. pp. 1445-1454.

[11] L. Lin, G. Wang, W. Zuo, F. Xiangchu, and L. Zhang, “Cross-Domain Visual Matching via Generalized Similarity Measure and Feature Learning,” IEEE Transactions on Pattern Analysis & Machine Intelligence, vol. 39, no. 6, pp. 1089-1102, 2017.