

# Efficient Cross-modal Retrieval via Discriminative Deep Correspondence Model

Zhikai Hu<sup>1,2</sup>, Xin Liu<sup>1,2</sup>(✉), An Li<sup>1,2</sup>, Bineng Zhong<sup>1,2</sup>, Wentao Fan<sup>1,2</sup>,  
and Jixiang Du<sup>1,2</sup>

<sup>1</sup> Department of Computer Science, Huaqiao University, Xiamen 361021, China  
xliu@hqu.edu.cn

<sup>2</sup> Xiamen Key Laboratory of Computer Vision and Pattern Recognition,  
Huaqiao University, Xiamen 361021, China

**Abstract.** Cross-modal retrieval has recently drawn much attention due to the widespread existence of multi-modal data, and it generally involves two challenges: how to model the correlations and how to utilize the class label information to eliminate the heterogeneity between different modalities. Most previous works mainly focus on solving the first challenge and often ignore the second one. In this paper, we propose a discriminative deep correspondence model to deal with both problems. By taking the class label information into consideration, our proposed model attempts to seamlessly combine the correspondence autoencoder (Corr-AE) and supervised correspondence neural networks (Super-Corr-NN) for cross-modal matching. The former model can learn the correspondence representations of data from different modalities, while the latter model is designed to discriminatively reduce the semantic gap between the low-level features and high-level descriptions. The extensive experiments tested on three public datasets demonstrate the effectiveness of the proposed approach in comparison with the state-of-the-art competing methods.

**Keywords:** Cross-modal retrieval  
Discriminative deep correspondence model  
Correspondence autocoder · Semantic gap

## 1 Introduction

In recent years, cross-modal retrieval has attracted considerable attention due to the rapid growth of abundant multi-modal data. Unlike uni-modal retrieval, cross-modal retrieval searches the information of one modality by given information of another different modality, and this type of matching problem has been existing in many fields, e.g., multimedia retrieval. For instance, one often attempts to seek the picture that best illustrates a given text, or find the text that best describes a given picture. Since the samples from different modalities are always heterogeneous, it is still a non-trivial task to bridge the semantic gap, measure the similarity and establish the connections among different modalities.

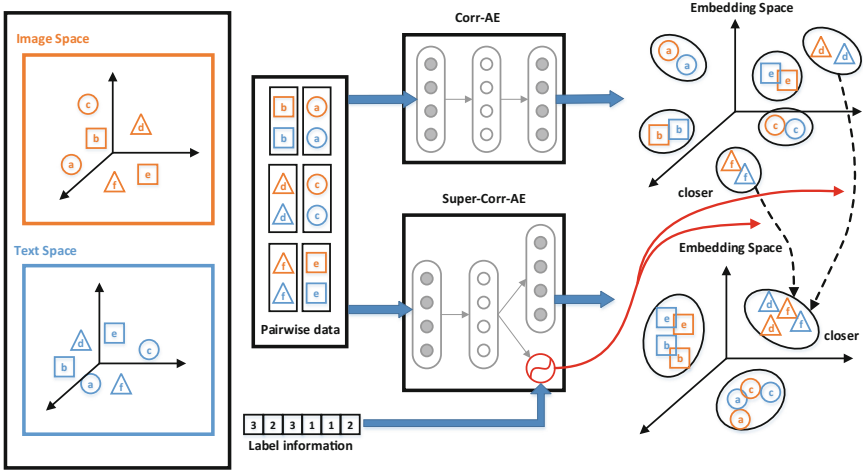


Fig. 1. The difference between the Corr-AE [5] and our proposed model.

In the past, different kinds of cross-modal retrieval approaches have been exploited, and most of them attempted to learn a common latent subspace to make all data comparable. That is, these approaches project the representations of multiple modalities into a common space with same dimension, whereby the similarity can be easily measured by counting Euclidean or other distances [13–15]. Along this line, one of the most common subspace learning methods is Canonical Correlation Analyse (CCA), which learns a pair of linear transformations to maximize the correlations between two different modalities. For instance, Rasiwasia et al. [13] proposed a semantic correlation matching scheme for cross-modal retrieval, where the CCA is applied to the maximally correlated feature representations. Later, some improved versions [12] and extensions [11] were also exploited to produce an isomorphic semantic space for cross-modal retrieval. In addition, bilinear model (BLM) [15] and partial least squares (PLS) [14], were also popularly utilized in cross-modal retrieval. Despite these contributions have been made to the solution of cross-modal retrieval, most of their performances are still far from satisfactory. A plausible reason is that these traditional approaches did not fully consider the label information or high-level featuring mapping for discriminative analysis, thereby the connections of different modalities were not well built in practice.

Recently, significant progress has been made in deep learning, and some deep models have been developed to tackle cross-modal retrieval problem [9]. For instance, Srivastava and Salakhutdinov [17] first utilized two separated Restricted Boltzmann Machines (RBM) to learn the low-level representations of different modalities, and then fused them into a joint representation by a high-level full-connection layer. Similarly, Feng et al. [5] employed a correspondence autoencoder (Corr-AE) to represent different modalities, in which an efficient loss function was designed for minimization of correlation learning error

and representation learning error synchronously. Although these typical methods were able to produce considerable performance, their cross-modal retrieval performance were still a bit poor. The main reason lies that these learning models can only preserve the inter-modal similarity, but which often ignore intra-modal similarity. As a result, the corresponding retrieval performances were far from satisfactory.

In general, the more information embedded in a model for discriminative analysis, the greater performance it reaches. Evidently, label information often provides discriminative cues for the classification task, and some deep learning methods [2, 19, 20] thus attempted to utilize the label information in their models. For instance, Wang et al. [19] first utilized the label information to learn the discriminative representation of data from each modality individually, then investigated the highly non-linear semantic correlation between different modalities. This strategy separates the representation learning and correlation learning, which can not utilize label information sufficiently.

In this paper, we propose a discriminative deep correspondence model for cross-modal retrieval, which can learn the representation and correlation of different modalities in an integral model. As shown in Fig. 1, our proposed model inherently differs from Corr-AE model [5], and improves this method by providing the following three contributions:

- By using the label information, a supervised correspondence neural networks (Super-Corr-NN), is proposed to learn the representation and correlation of different modalities simultaneously, while preserving more intra-modal similarity.
- We propose a supervised Corr-AE (Super-Corr-AE) model and extend it into another two forms, whereby the data from different modalities can be well retrieved.
- The proposed discriminative deep correspondence model seamlessly combines the correspondence autoencoder (Corr-AE) and supervised correspondence neural networks (Super-Corr-NN) for data representation learning, which can preserve both inter-modal and intra-modal similarity for efficient cross-modal matching. The experimental results have shown its outstanding performance.

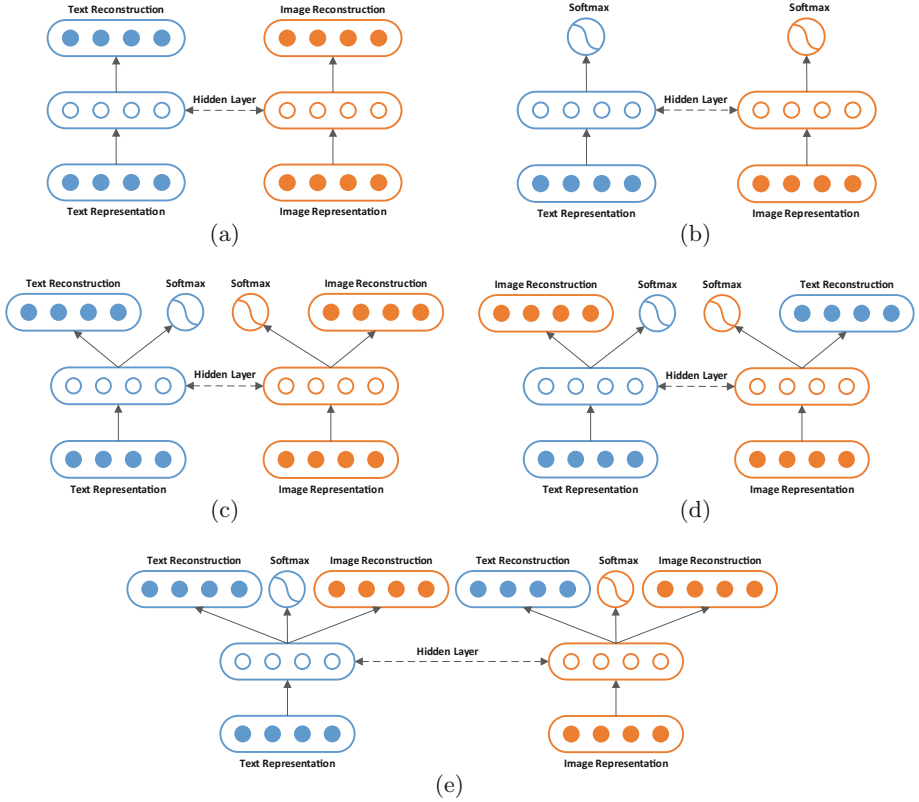
The rest of this paper is organized as follow. Section 2 presents the detail of our proposed model. In Sect. 3, we present the experimental results and related comparisons. Finally, we draw a conclusion in Sect. 4.

## 2 The Proposed Deep Learning Model

In this section, Corr-AE [5] is simply introduced for illustration, and then the proposed deep correspondence models are presented in details.

### 2.1 Correspondence AutoEncoder

As shown in Fig. 2(a), Corr-AE consists of two interdependent autoencoders, whose hidden layers are connected with each other by defining a similarity metric. Two autoencoders correspond to image and text modalities respectively,



**Fig. 2.** (a) Correspondence autoencoder, (b) Supervised correspondence neural networks, (c) Supervised correspondence autoencoder, (d) Supervised correspondence cross-modal autoencoder, (e) Supervised correspondence full-modal autoencoder.

receive data at input layers, reconstruct them at output layers, and extract the information of hidden layers as final learned representations. Let  $X_1 = \{x_i^{(i)}\}_{i=1}^n$  and  $X_2 = \{x_2^{(i)}\}_{i=1}^n$  denote two different modalities, e.g., image and text, and share the same label  $Y = \{y^{(i)}\}_{i=1}^n$ , where  $n$  is the number of objects,  $y^{(i)} \in \{1, 2, \dots, c\}$  and  $c$  is the number of categories. The hidden layers of two autoencoders are denoted as  $H_1 = f(X_1; W_f)$  and  $H_2 = g(X_2; W_g)$ , where  $W_f$  and  $W_g$  are parameters of two autoencoders respectively,  $f$  and  $g$  are the corresponding activation functions. The similarity metric between two modalities are defined as follow:

$$L_H(X_1, X_2; W) = \|H_1 - H_2\|_F^2 = \|f(X_1; W_f) - g(X_2; W_g)\|_F^2 \quad (1)$$

where  $W = [W_f, W_g]$  is the whole parameters. Accordingly, the reconstruction losses of two modalities can be described as follows:

$$L_I(X_1, X_2; W) = \|X_1 - \hat{X}_1^{(I)}\|_F^2 \quad (2)$$

$$L_T(X_1, X_2; W) = \|X_2 - \hat{X}_2^{(T)}\|_F^2 \tag{3}$$

where  $\hat{X}_1^{(I)}$  and  $\hat{X}_2^{(T)}$  are outputs of the image and text autoencoder respectively. Consequently, the total loss can be derived as follow:

$$L_{total} = (1 - \alpha)(L_I + L_T) + \alpha L_H \tag{4}$$

where  $\alpha \in [0, 1]$  is balance parameter.

### 2.2 Supervised Correspondence Neural Networks

In general, label information often provides discriminative cues for the vision task [18]. Inspired by the theories of Corr-AE, we proposed a Supervised Correspondence Neural Networks (Super-Corr-NN) by taking the label information into the consideration. As shown in Fig. 2(b), two autoencoders are replaced by two 3-layers multilayer perceptron in Super-Corr-NN. Given data set  $X = \{x^{(i)}\}_{i=1}^n$  and its label set  $L = \{l^{(i)}\}_{i=1}^n$ ,  $l^{(i)} \in \{1, 2, \dots, c\}$  and  $c$  is the number of categories, the softmax loss is presented as follow:

$$J(X, L; \Theta) = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^c I\{l^{(i)} = j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{k=1}^c e^{\theta_k^T x^{(i)}}} \quad j = 1, 2, \dots, c \tag{5}$$

where  $\Theta$  is the parameter of network,  $\theta_i$  is the  $i$ -th column of  $\Theta$ , and  $I\{z\}$  equals 1 if  $z$  is true, otherwise equals 0. Differing from the autoencoder model, the output of multilayer perceptron is a probability vector of a sample belonging to different classes, which can utilize label information to preserve intra-modal similarities and make results more discriminative. Accordingly, the overall softmax loss function in Super-Corr-NN can be described as:

$$L_S = J(X_1, Y; \Theta_1) + J(X_2, Y; \Theta_2) \tag{6}$$

where  $\Theta_1$  and  $\Theta_2$  are the parameters of two multilayer perceptrons respectively. Specifically, the same similarity metric is adopted to learn representation and correlation simultaneously. By taking the correspondence loss defined in Eq. (1), the total loss function can be derived by:

$$L_{total} = (1 - \alpha)L_S + \alpha L_H \tag{7}$$

where  $\alpha \in [0, 1]$  is balance parameter.

### 2.3 Supervised Correspondence AutoEncoder

In essence, Corr-AE, as an unsupervised learning method, can learn a correspondence representation and correlation between different modalities. In contrast to this, Super-Corr-NN could learn more discriminative representation by using the label information, but which performs weak in correlation mining between different modalities. Meanwhile, Corr-AE can preserve more inter-modal similarity

while Super-Corr-NN can preserve more intra-modal similarity. Inspired by this finding, As shown in Fig. 2(c), we propose a supervised correspondence autoencoder (Super-Corr-AE) by combing the Corr-AE and Super-Corr-NN, which are complementary to each other. In Super-Corr-AE, Super-Corr-NN and Corr-AE share their input layer and hidden layer, with their output layers separated independently. As a result, this model could simultaneously preserve both inter-modal and intra-modal similarities and make the learned representation more discriminative. In this model, the total loss consists of three parts: reconstruction loss  $L_I + L_T$ , correspondence loss  $L_H$  and softmax loss  $L_S$ :

$$L_{total} = (1 - \alpha)(L_I + L_T) + \alpha L_H + \beta L_S \quad (8)$$

where  $\alpha \in [0, 1]$  and  $\beta \geq 0$  are balance parameters.

## 2.4 Supervised Correspondence Cross-modal AutoEncoder

As shown in Fig. 2(d), we further replace the Corr-AE in Super-Corr-AE to Corr-Cross-AE. Differing from the Corr-AE model, Corr-Cross-AE restricts the outputs of each autoencoder by the inputs of different modality. Accordingly, the reconstruction loss is given as:

$$L_{Icross}(X_1, X_2; W) = \|X_2 - \hat{X}_2^{(I)}\|_F^2 \quad (9)$$

$$L_{Tcross}(X_1, X_2; W) = \|X_1 - \hat{X}_1^{(T)}\|_F^2 \quad (10)$$

where  $\hat{X}_2^{(I)}$  and  $\hat{X}_1^{(T)}$  are outputs of image autoencoder and text autoencoder, respectively. Then, the total loss can be derived as follow:

$$L_{total} = (1 - \alpha)(L_{Icross} + L_{Tcross}) + \alpha L_H + \beta L_S \quad (11)$$

where  $\alpha \in [0, 1]$  and  $\beta \geq 0$  are balance parameters.

## 2.5 Supervised Correspondence Full-modal AutoEncoder

As shown in Fig. 2(e), we replace the Corr-AE in Super-Corr-AE to Corr-Full-AE. As a combination of Corr-AE and Corr-Cross-AE, Corr-Full-AE utilizes the inputs of all modalities to restrict the outputs of each autoencoder. Consequently, the reconstruction loss can be derived as follows:

$$\begin{aligned} L_{Ifull}(X_1, X_2; W) &= L_I(X_1, X_2; W) + L_{Icross}(X_1, X_2; W) \\ &= \|X_1 - \hat{X}_1^{(I)}\|_F^2 + \|X_2 - \hat{X}_2^{(I)}\|_F^2 \end{aligned} \quad (12)$$

$$\begin{aligned} L_{Tfull}(X_1, X_2; W) &= L_T(X_1, X_2; W) + L_{Tcross}(X_1, X_2; W) \\ &= \|X_2 - \hat{X}_2^{(T)}\|_F^2 + \|X_1 - \hat{X}_1^{(T)}\|_F^2 \end{aligned} \quad (13)$$

where  $\hat{X}_1^{(I)}$  and  $\hat{X}_2^{(I)}$  are outputs of image autoencoder,  $\hat{X}_1^{(T)}$  and  $\hat{X}_2^{(T)}$  are outputs of text autoencoder, respectively. Then, the total loss can be derived as follows:

$$L_{total} = (1 - \alpha)(L_{Ifull} + L_{Tfull}) + \alpha L_H + \beta L_S \quad (14)$$

where  $\alpha \in [0, 1]$  and  $\beta \geq 0$  are balance parameters.

### 3 Experimental Results

#### 3.1 Datasets

For experimental evaluation, three publicly available datasets are selected:

**Wikipedia**<sup>1</sup>. This dataset [13] is collected from “Wikipedia features articles”, which contains 10 categories and 2,866 image&text pairs. Images are represented by 2,296-dimensions features, which consists of three type of low-level features extracted from images, including 1,000-dimensions pyramid histogram of words (PHOW) [1], 512-dimensions gist descriptor [10], 784-dimensions MPEG-7 descriptors [8]. And texts are represented by 3,000-dimensions high-frequency words vector. These features are all extracted by [5]. We use 2,173 pairs as training set, 231 pairs as validation set, 462 pairs as test set.

**Pascal**<sup>2</sup>. This dataset [4] contains 10 categories and 1,000 image&text pairs. Image are represented by 2,296-dimensions features, shares the same extraction with wikipedia dataset. And texts are represented by 1,000-dimensions high-frequency words vector. We use 800 pairs as training set, 100 pairs as validation set, 100 pairs as test set.

**NUS-WIDE-10k**. This dataset is selected from the real-world web image dataset NUS-WIDE<sup>3</sup> [3], which contains 81 concepts and 269,648 images with tags. We only choose 10 largest concepts and the corresponding 10,000 image&text pairs, each concept contains 1,000 pairs. Images are represented by 1134-dimensions features, which consists of six types of low-level features extracted from images, including 64-dimensions color histogram, 144-dimensions color correlogram, 73-dimensions edge direction histogram, 128-dimensions wavelet texture, 225-dimensions block-wise color moments and 500-dimensions bag of words based on SIFT descriptions. And texts are represented by 1,000-dimensions bag-of-words. We use 8,000 pairs as training set, 1,000 pairs as validation set, and 1,000 pairs as test set.

#### 3.2 Implementation Details

In the experiments, two RBMs were selected to pre-process the data. In the first RBM, Guassian RBM and replicated softmax RBM were adopted to process image and text data, respectively. The second RBM was a basic RBM. As a result, the pre-processed data was utilized as the input of the proposed deep correspondence models.

In particular, the settings of parameters  $(\alpha, \beta)$  in different datasets were shown in Table 1, although parameters  $(\alpha, \beta)$  are not sensitive to datasets (more empirical analysis will be given in the coming sections), to reach best performance, we set different  $(\alpha, \beta)$  on three different datasets and different model and we fixed these values in all related experiments.

<sup>1</sup> <http://www.svcl.ucsd.edu/projects/crossmodal/>.

<sup>2</sup> <http://vision.cs.uiuc.edu/pascal-sentences/>.

<sup>3</sup> <http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm>.

**Table 1.** Parameters setting on different methods and datasets

Method	Wikipedia		Pascal		NUS-WIDE-10k	
	$\alpha$	$\beta$	$\alpha$	$\beta$	$\alpha$	$\beta$
Super-Corr-NN	0.8	-	0.9	-	0.7	-
Super-Corr -AE	0.9	11	0.9	19	0.9	1
Super-Corr-Cross-AE	0.9	50	0.8	30	0.7	3
Super-Corr-Full-AE	0.9	10	0.9	12	0.9	2

### 3.3 Baseline Methods

Meanwhile, three CCA basic methods: CCA-AE [6], CCA-Cross-AE [5], CCA-Full-AE [5], two multi-modal methods: Bimodal AE [9], Bimodal DBN [9, 16], and three correspondence autoencoder methods [5]: Corr-AE, Corr-Cross-AE, Corr-Full-AE, were selected for comparison.

### 3.4 Evaluation Metric

We perform two kinds of cross-modal retrieval: retrieve text by given image query and vice versa. Mean average precision (mAP) of  $R$  top-rank retrieved data is employed in our experiments to evaluate the performance of the results:

$$mAP = \frac{1}{M} \sum_{i=1}^M \left( \frac{1}{L} \sum_{r=1}^R P(r) \times \delta(r) \right) \quad (15)$$

where  $M$  is the number of queries,  $L$  is the number of relevant data in retrieved dataset,  $P(r)$  is precision of top  $r$  retrieved data, and  $\delta(r)$  equals 1 if the retrieved data shares the same label with the query, otherwise equals 0. We employ  $mAP@50$  ( $R = 50$ ) in our experiments.

### 3.5 Analyse of Results

Typical mAP scores obtained by different methods and respectively tested on Wikipedia, Pascal and NUS-WIDE-10k datasets, were reported in Table 2. Comparing with the baseline methods on image retrieval and text retrieval, it can be found that our proposed models have significantly improved the performances in both image and text retrieval. In particular, the retrieval performance obtained by our proposed Super-Corr-NN was competitive with the results obtained by Corr-AE. For instance, Super-Corr-NN has produced a better retrieval performance than Corr-AE on Wikipedia and NUS-WIDE-10k datasets. Meanwhile, Super-Corr-AE has achieved the better cross-modal retrieval performances than Corr-AE models on three datasets, and also produced the better result than Super-Corr-NN on Wikipedia and Pascal datasets. For instance, Super-Corr-AE improved the mAP scores, respectively 3.1% and 13.9% on image query and text



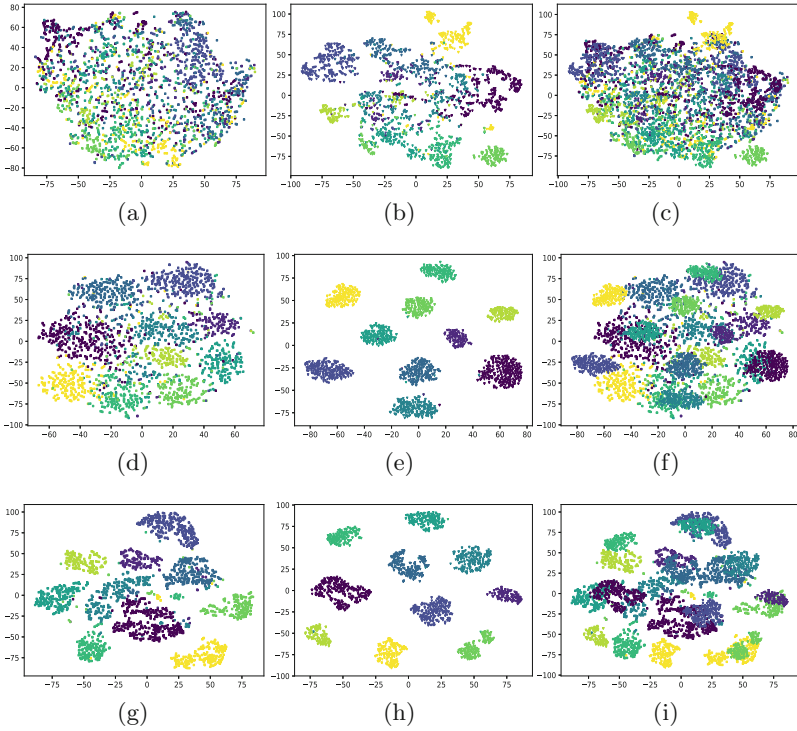
**Table 2.** mAP scores tested on three different datasets (I: image; T: text).

Method	Wikipedia			Pascal			NUS-WIDE-10k		
	I→T	T→I	Avg	I→T	T→I	Avg	I→T	T→I	Avg
CCA-AE [6]	0.213	0.235	0.224	0.161	0.153	0.157	0.199	0.268	0.234
CCA-Cross-AE [5]	0.197	0.230	0.214	0.137	0.182	0.159	0.199	0.344	0.272
CCA-Full-AE [5]	0.293	0.331	0.312	0.148	0.177	0.163	0.241	0.242	0.242
Bimodal AE [9]	0.295	0.307	0.302	0.253	0.278	0.266	0.251	0.295	0.273
Bimodal DBN [9,16]	0.197	0.216	0.207	0.208	0.211	0.210	0.163	0.204	0.184
Corr-AE [5]	0.326	0.361	0.344	0.290	0.279	0.285	0.319	0.375	0.347
<b>Super-Corr-NN</b>	0.330	0.383	0.356	0.202	0.206	0.204	0.337	<b>0.435</b>	<b>0.386</b>
<b>Super-Corr-AE</b>	<b>0.336</b>	<b>0.411</b>	<b>0.379</b>	<b>0.301</b>	<b>0.320</b>	<b>0.310</b>	<b>0.353</b>	0.415	0.384
Corr-Cross-AE [5]	0.336	0.341	0.338	0.271	0.280	0.276	<b>0.349</b>	0.348	0.349
<b>Super-Corr-Cross-AE</b>	<b>0.340</b>	<b>0.446</b>	<b>0.393</b>	<b>0.292</b>	<b>0.295</b>	<b>0.293</b>	0.319	<b>0.395</b>	<b>0.357</b>
Corr-Full-AE [5]	0.335	0.368	0.352	0.281	0.276	0.279	0.331	0.379	0.355
<b>Super-Corr-Full-AE</b>	<b>0.352</b>	<b>0.423</b>	<b>0.387</b>	<b>0.300</b>	<b>0.302</b>	<b>0.301</b>	<b>0.358</b>	<b>0.415</b>	<b>0.386</b>

query than Corr-AE tested on Wikipedia dataset, 3.8% and 14.7% on Pascal dataset, 10.7% and 10.7% on NUS-WIDE-10k. Further, Super-Corr-Cross-AE has improved the mAP scores, respectively 1.2% and 30.8% on image query and text query than Corr-Cross-AE on Wikipedia dataset, 7.7% and 5.4% on Pascal dataset. This superiority can be attributed to the physical meaning of the model illustrated in Sect. 2.3, which seamlessly combined the Corr-AE and Super-Corr-NN by using the class label information. Remarkably, Super-Corr-Full-AE has achieved the satisfactory performance, and the experimental results have shown its outstanding performance.

### 3.6 Visualization

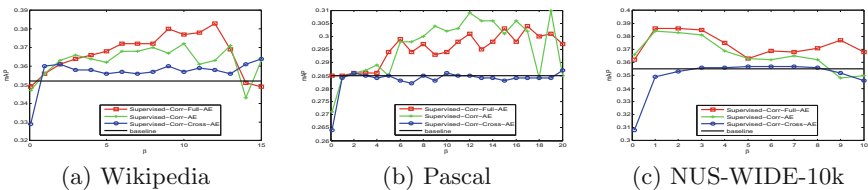
Further, we utilized the t-SNE [7] to visualize the image and text representation for discrimination illustration intuitively. As shown in Fig. 3, it can be found that Super-Corr-NN was able to learn discriminative representations, and the data points of text modality can be well divided into their corresponding classes. In contrast to this, Corr-Full-AE has failed to separate the text modalities of different classes. The main reason lies that the Corr-Full-AE model focuses more on pairwise data and often fails to preserve any global information in single modality. As a result, the data points from different classes are distributed disorderly. Further, the proposed Super-Corr-Full-AE could produce discriminative separations on both image and text modalities, the data points can be well divided into different classes visually. Therefore, it can be concluded that our discriminative deep correspondence model could learn discriminative representations of different modalities for cross-modal matching and retrieval.



**Fig. 3.** Visualization of image, text and image&text pair in Wikipedia dataset, respectively shown in (a)–(c) by Corr-Full-AE, (d)–(f) by Super-Corr-NN, and (g)–(i) by Super-Corr-Full-AE.

### 3.7 Analyse of $\beta$

In addition, there are two parameters  $\alpha$  and  $\beta$  in our models, and the sensitivity of parameter  $\alpha$  is well analysed in work [5]. By fixing  $\alpha = 0.9$ , we conduct various experiments to analyse the effect of different parameter  $\beta$  values. The mAP values obtained by different models and tested on different datasets were shown in Fig. 4, it can be easily found that parameters  $\beta$  is not sensitive in our models and could produce satisfactory performance in a wide range of the values. Therefore, it is adequate to fix the  $\beta$  with an appropriate value in practice.



**Fig. 4.** mAP values obtained by the proposed deep models with different  $\beta$  values.

## 4 Conclusion

In this paper, we have presented a discriminative deep correspondence model by seamlessly combining the correspondence autoencoder and supervised correspondence neural networks for cross-modal matching. The proposed deep model not only can learn the representation and correlation of different modalities simultaneously, but also could preserve both inter-modal and intra-modal similarity for efficient cross-modal matching. The experimental results have shown its outstanding performance in comparison with existing counterparts.

**Acknowledgment.** The work described in this paper was supported by the National Science Foundation of China (Nos. 61673185, 61502183, 61673186), National Science Foundation of Fujian Province (No. 2017J01112), Promotion Program for Young and Middle-aged Teacher in Science and Technology Research (No. ZQN-PY309), the Promotion Program for graduate student in Scientific research and innovation ability of Huaqiao University (No. 1611314006).

## References

1. Bosch, A., Zisserman, A., Munoz, X.: Image classification using random forests and ferns. In: Proceedings of IEEE ICCV, pp. 1–8 (2007)
2. Castrejn, L., Aytar, Y., Vondrick, C., Pirsivash, H., Torralba, A.: Learning aligned cross-modal representations from weakly aligned data. In: Proceedings of IEEE CVPR, pp. 2940–2949 (2016)
3. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from National University of Singapore. In: Proceedings of ACM International Conference on Image and Video Retrieval, pp. 48:1–48:9 (2009)
4. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: generating sentences from images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6314, pp. 15–29. Springer, Heidelberg (2010). [https://doi.org/10.1007/978-3-642-15561-1\\_2](https://doi.org/10.1007/978-3-642-15561-1_2)
5. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: Proceedings of ACM Multimedia, pp. 7–16 (2014)
6. Kim, J., Nam, J., Gurevych, I.: Learning semantics with deep belief network for cross-language information retrieval. In: Proceedings of IEEE International Conference on Computational Linguistics, pp. 579–588 (2012)
7. van der Maaten, L., Hinton, G.E.: Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**(2), 2579–2605 (2008)
8. Manjunath, B.S., Ohm, J.R., Vasudevan, V.V., Yamada, A.: Color and texture descriptors. *IEEE Trans. Circuits Syst. Video Technol.* **11**(6), 703–715 (2002)
9. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of IEEE ICML, pp. 689–696 (2011)
10. Oliva, A., Torralba, A.: Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **42**(3), 145–175 (2001)
11. Pereira, J.C., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **36**(3), 521–535 (2013)
12. Quadrianto, N., Lampert, C.H.: Learning multi-view neighborhood preserving projections. In: Proceedings of IEEE ICML, pp. 425–432 (2011)

13. Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R.G., Levy, R., Vasconcelos, N.: A new approach to cross-modal multimedia retrieval. In: Proceedings of IEEE International Conference on Multimedia. pp. 251–260 (2010)
14. Rosipal, R., Krmer, N.: Overview and recent advances in partial least squares. In: Proceedings of IEEE International Conference on Subspace, Latent Structure and Feature Selection, pp. 34–51 (2005)
15. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: a discriminative latent space. In: Proceedings of IEEE CVPR, pp. 2160–2167 (2012)
16. Srivastava, N., Salakhutdinov, R.: Learning representations for multimodal data with deep belief nets. In: Proceedings of IEEE ICML workshop (2012)
17. Srivastava, N., Salakhutdinov, R.: Multimodal learning with deep Boltzmann machines. *J. Mach. Learn. Res.* **15**(8), 1967–2006 (2012)
18. Tang, J., Wang, K., Shao, L.: Supervised matrix factorization hashing for cross-modal retrieval. *IEEE Trans. Image Process.* **25**(7), 3157–3166 (2016)
19. Wang, C., Yang, H., Meinel, C.: Deep semantic mapping for cross-modal retrieval. In: Proceedings of IEEE International Conference on Tools Artificial Intelligence pp. 234–241 (2016)
20. Wang, W., Yang, X., Ooi, B.C., Zhang, D., Zhuang, Y.: Effective deep learning-based multi-modal retrieval. *VLDB J.* **25**(1), 79–101 (2016)