Learning a Semantic Space by Deep Network for Cross-media Retrieval

Zhao Li[†][⋆][∗], Wei Lu[†], Egude Bao[†], Weiwei Xing[†]
[†] School of Software Engineering, Beijing Jiaotong University, Bejing 100044, China
[⋆] Shandong Computer Science Center(National Supercomputing Center in Jinan), Jinan Shandong 250014, China
[∗] Shandong Provincial Key Laboratory of Computer Network, Jinan Shandong 250014, China
liz@sdas.org {luwei, baoe, wwxing}@bjtu.edu.cn

***Abstract*—With the growth of multimedia data, the problem of cross-media (or cross-modal) retrieval has attracted considerable interest in the cross-media retrieval community. One of the solutions is to learn a common representation for multimedia data. In this paper, we propose a simple but effective deep learning method to address the cross-media retrieval problem between images and text documents for samples either with single or multiple labels. Specifically, two independent deep networks are learned to project the input feature vectors of images and text into an common (isomorphic) semantic space with high level abstraction (semantics). With the same dimensional feature representation in the learned common semantic space, the similarity between images and text documents can be directly measured. The correlation between two modalities is built according to their shared ground truth probability vector. To better bridge the gap between the images and the corresponding semantic concepts, an open-source CNN implementation called Deep Convolutional Activation Feature (DeCAF) is employed to extract input visual features for the proposed deep network. Extensive experiments on two publicly available multi-label datasets, NUS-WIDE and PASCAL VOC 2007, show that the proposed method achieves better results in cross-media retrieval compared with other state of the art methods.**

***Keywords*-cross-media retrieval; cross-modal retrieval; deep learning.**

I. Introduction

Nowadays, with the development of Internet, an enormous amount of multimedia data, e.g., image, text documents and videos, have been generated. These data with various modalities usually co-occur to describe the same objects or events. For example, images are usually accompanied with a textual description to represent the same meaning. Learning the relationships among different modalities is becoming an interesting research topic which can benefit many important applications, such as multimedia retrieval and content creation. In this work, we address the cross-media retrieval problem between images and text documents, i.e., using an image to retrieve text and using text to retrieve images, as illustrated in Fig. 1. Although here we only focus on two modalities, i.e., image and text, our method can be easily adapted to other modalities.

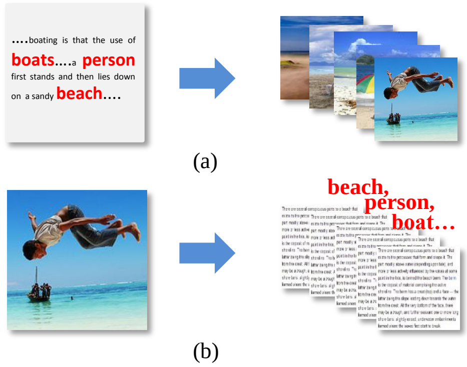
....boating is that the use of boats....a person first stands and then lies down on a sandy beach....
(a)
beach,
person,
boat…
(b)

Figure 1: The cross-media retrieval task considered in this paper. (a) Using text as a query to retrieve relevant images. (b) Using an image as a query to retrieve relevant text.

During the past few years, many cross-media retrieval methods have been proposed [1], [2], [3]. As two typical methods, Canonical Correlation Analysis (CCA) [4] and Partial Least Squares [5], [6] are usually adopted to learn a couple of projections to maximize the correlations between two variables. Some methods have been proposed based on CCA. One of them is a Semantic Correlation Matching method [1], which leverages multi-class logistic regression to produce an isomorphic semantic space for cross-media retrieval. In [2], a generic framework called Generalized Multiview Analysis was presented to address multimedia problems. More recently, [3] proposed a multi-view CCA model by introducing a semantic view to achieve a better separation for multimedia data of different classes in the learned isomorphic space.

Although these methods have made contributions to the solution of cross-media retrieval tasks, their performance is still far from satisfactory. This is because the performance of cross-media retrieval between images and text is highly dependent on visual feature representation, but traditional feature extraction techniques have been undergoing a bottleneck period for image understanding in the past few years. Recently, significant progress has been made in image classification due to the development of convolutional neural networks (CNN) [7], [8], [9]. Especially, [9] has demonstrated promising results for image classification in the 2012 ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [10]. More recently, [11] has released an

DOI reference number: 10.18293/DMS2015-005

open-source implementation of an eight-layer network called DeCAF, which is trained on ImageNet with 1,000 classes, and has demonstrated that CNN features are suitable as general features for various tasks.

In this paper, we propose a deep learning method to address the cross-media retrieval problem with multi-label. As shown in Fig. 2, two independent fully-connected deep networks are trained to project input feature vectors of images and text into an isomorphic semantic space with higher level abstraction. Specifically, we employ DeCAF to extract the visual feature of each image as the input feature vector of Image_Net. For Text_Net, the input feature vector can be extracted by many traditional feature extraction techniques, such as the bag-of-words. Since the proposed method is supervised, the correlation between two modalities can be built according to their shared ground truth probability vector. Both two networks have two hidden layers and one output layer, and the squared loss is employed as the loss function. Experiments on NUS-WIDE and PASCAL VOC 2007 demonstrate a significant performance improvement over other methods.

The remainder of this paper is organized as follows. We briefly review the related work of cross-media retrieval in Section II. In Section III, we present the proposed method in details. After that, experimental results and analysis are reported in Section IV. Finally, Section V presents the conclusions.

II. Related Work

Based on Canonical Correlation Analysis (CCA) [4], some methods [1], [2], [3], [12] are proposed to lean a common space for multimedia data of different modality, in which the distance between two media objects with similar semantics could be minimized while those with different semantics could be maximized. Specifically, Gong *et al.* [3] presented a multi-view CCA method via introducing a third view, i.e., semantic view, to better separate the multimedia data with different semantics in the learned latent common space. The semantic view representation can be obtained through supervised information as well as clustering analysis. Similarly, a cluster CCA method, which also focuses on learning discriminant common space to maximize the correlation of different kinds of multimedia data, was proposed by [12]. In this work, the separation for multimedia data with of different semantics was achieved via an unsupervised way.

Besides, with the ever-growing large-scale multimedia data on the Internet, much attention has been devoted to nearest neighbor search. To address this time-consuming problem, some hashing-based methods [13], [14], [15] have attracted a lot of interest. In [13], a cross view hashing (CVH) method was proposed to generate hash codes by minimizing the distance of hash codes for similar data while maximizing the distance for dissimilar data. Wu *et al.* [14] presented a sparse hashing method to obtain sparse code sets for the data of different modalities through joint multimedia dictionary learning.

In addition, with the development of deep learning, some deep models [16], [17], [18] have been proposed to address multimedia problem. Specifically, Andrew *et al.* [17] adapt the CCA into the deep model to learn complex nonlinear transformations of different multimedia data. Based on Restricted Boltzmann Machine, Ngiam *et al* [16] proposed to learn a shared representation between different modalities of multimedia data.

III. The Proposed Method

In this section, we will detail the proposed deep learning method for cross-media retrieval. We will first describe the architectures of the two deep networks as well as the training parameters, and then introduce the Euclidean loss function used in the training process.

A. Network Architecture

As shown in Fig. 2, we build two independent networks, i.e., Image_Net and Text_Net, to map images and text from their input feature spaces into a common semantic space respectively. Each network consists of two hidden layers and one output layer. For Image_Net, we employ DeCAF for image feature extraction. Specifically, each image is firstly resized to 256×256 and fed into DeCAF, which is pre-trained on the ImageNet dataset with 1,000 classes. Different from the previous work [11], which used CNN features (DeCAF$_5$, DeCAF$_6$ or DeCAF$_7$; refer to [11] for more details) to represent a given image, we utilize the 1000 dimensional predictive scores of each image as the input visual feature of the proposed deep network. The reason for this choice is that the predictive scores provide a probability distribution over 1,000 classes from the ImageNet dataset and the relationship between this kind of visual features and ground truth can be easily built. For Text_Net, since textual features usually have greater discriminative power than traditional visual features (e.g., SIFT and HOG), the relationship between textual features and ground truth can be more easily built. Therefore, many feature extraction techniques, such as bag-of-words, can be employed to extract the input textual features for Text_Net.

Denote $\boldsymbol{h}^{(0)} \in R^{d_0}$ as the input feature vector of Image_Net (or Text_Net). d_t is regarded as the output dimension of the tth layer (the input can be considered as the 0th layer for convenience). The outputs of the subsequent three layers (two hidden layers and one output layer) can be defined as

$$\boldsymbol{h}^{(t)} = \sigma\left(W_t \boldsymbol{h}^{(t-1)} + \boldsymbol{b}_t\right),\ t = 1, 2, 3, \tag{1}$$

where $\boldsymbol{h}^{(t)}$ is the output vector, $W_t \in R^{d_t \times d_{t-1}}$ is the matrix of weights and $\boldsymbol{b}_t \in R^{d_t}$ is the vector of biases. $\sigma(\cdot)$ is the activation function. In our work, we use the rectified linear units (ReLU) as the nonlinear activation function.

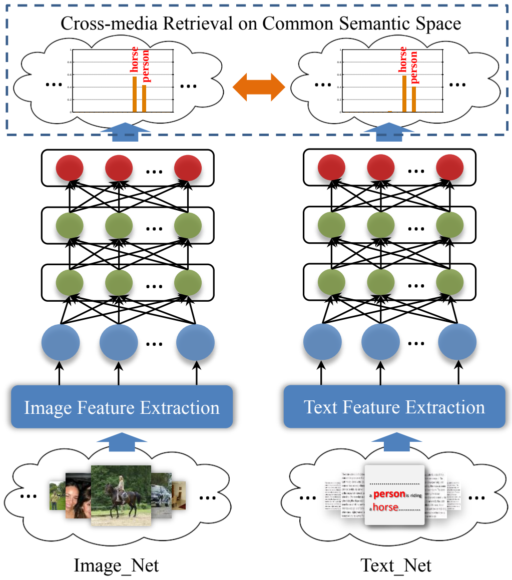

Figure 2: Two independent deep networks, Image_Net and Text_Net, are trained to project images and text from their original feature spaces into a common semantic space for the cross-media retrieval. Both networks have three layers, including two hidden layers and one output layer. Blue, green and red nodes represent input features, hidden units and output units, respectively.

The two networks are trained by stochastic gradient descent with momentum of 0.9. Besides, each hidden layer is followed by a dropout operation with a dropout ratio of 0.5 to combat overfitting. The global learning rate of these two networks is set as 0.01 at the beginning and dynamically changed according to the Euclidean loss(see below).

B. Euclidean Loss

To achieve the target that pairs of image and text can retain similar feature representation in the common semantic space, we utilize Euclidean loss as the cost function to optimize both Image_Net and Text_Net. The output of the last layer is fed into a c-way ($c = d_3$) softmax, which generates predictive scores over the c class labels. Suppose there are n *image-text* pairs in the training set. The predicted probability for the jth class of the ith input vector $\boldsymbol{h}_{\boldsymbol{i}}^{(0)}$ (image or text) can be defined as

$$P(y = j|\boldsymbol{h}_{\boldsymbol{i}}^{(0)}) = \frac{\exp(f_j(\boldsymbol{h}_{\boldsymbol{i}}^{(0)}))}{\sum_{k=1}^{c} \exp(f_k(\boldsymbol{h}_{\boldsymbol{i}}^{(0)}))}, \tag{2}$$

where $f(\cdot)$ can be considered as the mapping from the input layer to the output layer and $f_j(\boldsymbol{h}_{\boldsymbol{i}}^{(0)})$ is the activation value of $\boldsymbol{h}_{\boldsymbol{i}}^{(0)}$ on class j. y indicates the class label. Since the proposed method is targeted at multi-label problems, we can form a label vector $\boldsymbol{y}_i = [y_{i1}, y_{i2}, ...y_{ic}]$ for each *image-text* pair. $y_{ij} = 1$ ($j = 1...c$) if the given sample is annotated with class j, and otherwise $y_{ij} = 0$. We define the ground truth probability vector of $\boldsymbol{h}_{\boldsymbol{i}}^{(0)}$ as $\hat{\boldsymbol{p}}_i = \boldsymbol{y}_i/||\boldsymbol{y}_i||_1$ and the predictive probability vector as $\boldsymbol{p}_{\boldsymbol{i}} = [p_{i1}, p_{i2}, ...p_{ic}]$, where $p_{ij} = P(y = j|\boldsymbol{h}_{\boldsymbol{i}}^{(0)})$ ($j = 1...c$). Then, the cost function to be minimized can be defined as

$$J = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{c} (p_{ik} - \hat{p}_{ik})^2. \tag{3}$$

IV. Experiments

A. Dataset and Settings

We evaluate the performance of the proposed method compared with other methods on two publicly available datasets.

NUS-WIDE [19]: The dataset contains 269,648 images. Each image is accompanied with 81 ground truth labels (classes) and 1,000 text annotations. We select those pairs belonging to one of the 20 most frequent classes and ignore those pairs containing images without any ground truth label or text annotation. Then, a subset of 52,829 pairs for training and 35,216 pairs for testing can be obtained for evaluation.

PASCAL VOC 2007 [20]: There are 9,963 images of 20 classes in this dataset. A set of 399 words provided by [21] is employed as the textual description for each image. We conduct experiments on *trainval* and *test* splits, which contain 5,011 and 4,952 pairs respectively.

We experimentally set $d_1 = 512$, $d_2 = 256$, $d_3 = 20$ for both Image_Net and Text_Net. Euclidean distance is used to measure the similarity between features in the semantic space. We compare the proposed method with three popular methods, including two unsupervised methods, Canonical Correlation Analysis (CCA) [4] and Partial Least Squares (PLS) [5] which only utilize the pair-wise information, and the linear version of one supervised method, Multi-view CCA (Multi-CCA)[1] [3]. We vary the embedding dimensionality for these methods, i.e., 20, 128 and 512, and report the best performance. All the methods project images and text from their input feature spaces into a 20 dimensional common semantic space.

To get effective feature representation of visual representations for both NUS-WIDE and PASCAL VOC 2007, DeCAF [2] is employed in our method to extract the 1,000 dimensional predictive scores as the visual feature for each image. We use the 1,000 dimensional bag-of-words features provided by [19] as the textual features for NUS-WIDE and use the 798 dimensional tag ranking features (relative and absolute) provided by [21] as the textual features for PASCAL VOC 2007. Besides, to validate the effectiveness of the visual features used in this work, we do a comparison with other visual features based on CCA, PLS and multi-view CCA. For NUS-WIDE, we use the 500 dimensional Bag-of-SIFT-Words (SIFT-BoW) [22] features provided by [19]

[1] http://www.unc.edu/ yunchao/crossmodal.htm
[2] https://github.com/UCB-ICSI-Vision-Group/decaf-release

Table I: Cross-media retrieval performance on the NUS-WIDE dataset (mAP scores).

Method	I2T	T2I	Average
CCA (SIFT-BoW)	0.226	0.205	0.216
CCA (DeCAF)	0.277	0.262	0.270
PLS (SIFT-BoW)	0.316	0.181	0.249
PLS (DeCAF)	0.203	0.314	0.259
Multi-CCA (SIFT-BoW)	0.353	0.280	0.317
Multi-CCA (DeCAF)	0.439	0.315	0.377
Proposed	**0.486**	**0.409**	**0.448**

Table II: Cross-media retrieval performance on the PASCAL VOC 2007 dataset (mAP scores).

Method	I2T	T2I	Average
CCA (GIST+HSV+SIFT-BoW)	0.368	0.345	0.357
CCA (DeCAF)	0.638	0.618	0.628
PLS (GIST+HSV+SIFT-BoW)	0.380	0.348	0.364
PLS (DeCAF)	0.351	0.574	0.463
Multi-CCA (GIST+HSV+SIFT-BoW)	0.475	0.436	0.456
Multi-CCA (DeCAF)	0.709	0.577	0.643
Proposed	**0.781**	**0.689**	**0.735**

as the visual representations. For PASCAL VOC 2007, the 776 dimensional visual features, each of which contains a 512 dimensional GIST [23] feature, a 64 dimensional color feature (i.e., HSV) and a 200 dimensional SIFT-BoW feature, provided by [21] are employed as the visual representations.

B. Evaluation Metrics

In this paper, we consider two retrieval tasks, i.e., using image to retrieve text documents (I2T) and using text document to retrieve images (T2I). Retrieval performance is evaluated by mean average precision (mAP), which is one of the standard information retrieval metrics. In particular, given a set of queries, the average precision (AP) of each query is defined as:

$$AP = \frac{\sum_{k=1}^{R} P(k)rel(k)}{\sum_{k=1}^{R} rel(k)},$$

where R denotes the number of retrieved results. $rel(k) = 1$ if the item at rank k is relevant, $rel(k) = 0$ otherwise. $P(k)$ is the precision of retrieved results ranked at k. We can get the mAP score by averaging AP for all queries. Since NUS-WIDE and PASCAL VOC 2007 are two multi-label datasets, it is regarded as a relevant result if the retrieved result shares at least one class label with the query.

C. Results

Table I and Table II report our experimental results on NUS-WIDE and PASCAL V0C 2007, respectively. We can observe that the proposed method makes a significant improvement over any compared method on NUS-WIDE and PASCAL VOC 2007 (44.8% and 73.5%). This is because the proposed two deep networks can effectively build the relationship between the input feature vectors and the shared ground truth probability vectors with Euclidean loss function. In addition, the improvement of our method may also depend on the effective visual features. It can be observed that DeCAF works better than other image feature extraction techniques. This observation is reasonable, since DeCAF is pre-trained with about 100 million labeled images (i.e., ImageNet), which can make the learned visual features have sufficient representational power. Fig. 3 shows some examples accompanied with their visual features and textual features in the learned common semantic space on PASCAL VOC 2007. From Fig. 3, it can be seen that each *image-text* pair usually has a similar probability distribution in the common semantic space, which can further validate the effectiveness of our method.

V. Conclusions

In this paper, we proposed a deep learning method for cross-media retrieval. We trained two independent deep networks to map input feature vectors of images and text documents into an isomorphic semantic space, respectively. Especially, we took 1,000 dimensional predictive scores produced by an open-source CNN implementation called DeCAF, which is pre-trained on the ImageNet dataset with 1,000 classes, as the input visual features of Image_Net. Extensive experimental results on NUS-WIDE and PASCAL VOC 2007 show that the proposed method can achieve a better performance in cross-media retrieval task compared with other methods.

References

[1] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos, “A new approach to cross-modal multimedia retrieval,” in *MM*, 2010, pp. 251–260.

[2] A. Sharma, A. Kumar, H. Daume, and D. W. Jacobs, “Generalized multiview analysis: A discriminative latent space,” in *CVPR*, 2012, pp. 2160–2167.

[3] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, “A multi-view embedding space for modeling internet images, tags, and their semantics,” *IJCV*, vol. 106, no. 2, pp. 210–233, 2014.

[4] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods,” *Neural Computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[5] R. Rosipal and N. Krämer, “Overview and recent advances in partial least squares,” in *Subspace, Latent Structure and Feature Selection*, 2006, pp. 34–51.

[6] A. Sharma and D. W. Jacobs, “Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch,” in *CVPR*, 2011, pp. 593–600.

[7] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, “What is the best multi-stage architecture for object recognition?” in *ICCV*, 2009, pp. 2146–2153.

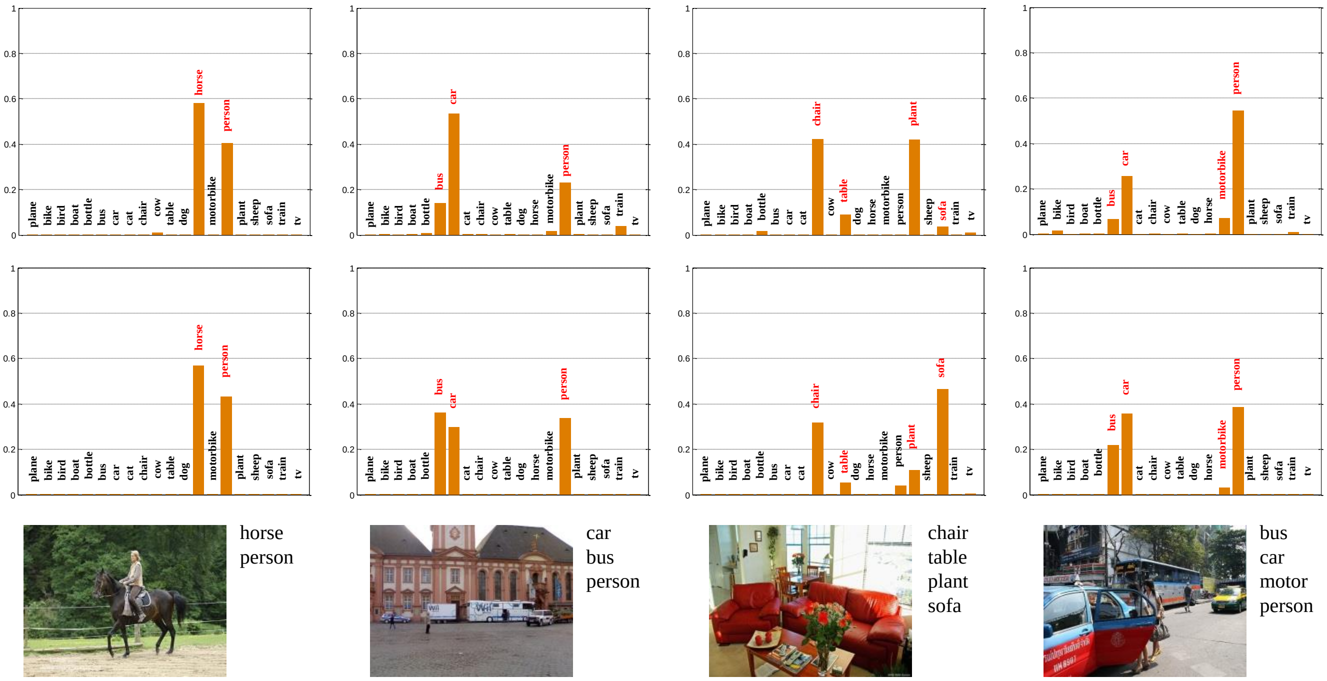

Figure 3: Some examples on PASCAL VOC 2007. The first and second rows plot the 20 dimensional semantic feature vectors of image and text corresponding to examples given in the third row. Ground truth labels are highlighted with red color.

[8] Y. LeCun, K. Kavukcuoglu, and C. Farabet, “Convolutional networks and applications in vision,” in *ISCAS*, 2010, pp. 253–256.

[9] A. Krizhevsky, I. Sutskever, and G. Hinton, “Imagenet classification with deep convolutional neural networks,” in *NIPS*, 2012, pp. 1106–1114.

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *CVPR*, 2009, pp. 248–255.

[11] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, “Decaf: A deep convolutional activation feature for generic visual recognition,” *arXiv preprint arXiv:1310.1531*, 2013.

[12] N. Rasiwasia, D. Mahajan, V. Mahadevan, and G. Aggarwal, “Cluster canonical correlation analysis,” in *AISTATS*, 2014, pp. 823–831.

[13] S. Kumar and R. Udupa, “Learning hash functions for cross-view similarity search,” in *IJCAI*, vol. 22, no. 1, 2011, p. 1360.

[14] F. Wu, Z. Yu, Y. Yang, S. Tang, Y. Zhang, and Y. Zhuang, “Sparse multi-modal hashing,” *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 427–439, 2014.

[15] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios, “Data fusion through cross-modality metric learning using similarity-sensitive hashing,” in *CVPR*, 2010, pp. 3594–3601.

[16] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal deep learning,” in *ICML*, 2011, pp. 689–696.

[17] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, “Deep canonical correlation analysis,” in *ICML*, 2013, pp. 1247–1255.

[18] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov *et al.*, “Devise: A deep visual-semantic embedding model,” in *NIPS*, 2013, pp. 2121–2129.

[19] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, “Nus-wide: a real-world web image database from national university of singapore,” in *CIVR*, 2009, p. 48.

[20] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes (voc) challenge,” *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.

[21] S. J. Hwang and K. Grauman, “Accounting for the relative importance of objects in image retrieval.” in *BMVC*, 2010, pp. 1–12.

[22] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[23] A. Oliva and A. Torralba, “Building the gist of a scene: The role of global image features in recognition,” *Progress in brain research*, vol. 155, pp. 23–36, 2006.