Cross-modal Retrieval with Correspondence Autoencoder

Fangxiang Feng
Beijing University of Posts and Telecommunications
Beijing, China
f.fangxiang@gmail.com

Xiaojie Wang
Beijing University of Posts and Telecommunications
Beijing, China
xjwang@bupt.edu.cn

Ruifan Li
Beijing University of Posts and Telecommunications
Beijing, China
rfli@bupt.edu.cn

ABSTRACT

The problem of cross-modal retrieval, e.g., using a text query to search for images and vice-versa, is considered in this paper. A novel model involving correspondence autoencoder (Corr-AE) is proposed here for solving this problem. The model is constructed by correlating hidden representations of two uni-modal autoencoders. A novel optimal objective, which minimizes a linear combination of representation learning errors for each modality and correlation learning error between hidden representations of two modalities, is used to train the model as a whole. Minimization of correlation learning error forces the model to learn hidden representations with only common information in different modalities, while minimization of representation learning error makes hidden representations are good enough to reconstruct input of each modality. A parameter α is used to balance the representation learning error and the correlation learning error. Based on two different multi-modal autoencoders, Corr-AE is extended to other two correspondence models, here we called Corr-Cross-AE and Corr-Full-AE. The proposed models are evaluated on three publicly available data sets from real scenes. We demonstrate that the three correspondence autoencoders perform significantly better than three canonical correlation analysis based models and two popular multi-modal deep models on cross-modal retrieval tasks.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models; I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms; Design

Keywords

Cross-modal; retrieval; image and text; deep learning; autoencoder

MM'14, November 03–07, 2014, Orlando, FL, USA.
Copyright 2014 ACM 978-1-4503-3063-3/14/11 ...$15.00.
http://dx.doi.org/10.1145/2647868.2654902.

1. INTRODUCTION

The inclusion of multi-modal data in webpages has become a widespread trend. For example, a webpage telling a story often contains some illustrations and other images accompanying the text. A travel photo shared on the web is usually tagged with words. The presence of massive multimodal data on the Internet brings huge cross-modal retrieval requirements, such as using an image query to search for texts and using a text query to search for images. Unlike traditional information retrieval tasks in a single modality, such as using a text query to search for text, cross-modal retrieval focuses on mining the correlations between data from different modalities.

1.1 Previous Work

Many approaches have been proposed to develop solutions to these challenging tasks. There are two main strategies for modeling cross-modal correlations among previous work.

One strategy involves modeling the correlation between different modalities with a shared layer.

A portion of them use topic models to achieve it. Correspondence LDA (Corr-LDA) [4] extended LDA [5] to find the topic-level relationships between images and text annotations, where topics' distributions act as middle layer for both images and texts. For images with loosely-coupled text, a mixture of directed and undirected probabilistic graphical model, called MDRF, was proposed in [14]. This model was built using a Markov random field over LDA. These LDA-based models essentially can be understood as a two-layer architecture with only one hidden layer.

Recently, there has been a trend of developing deep architecture for tackling complex AI problems [2], which is inspired by the architectural depth of the brain. Some representative models, such as deep autoencoders (DAE) [11], deep belief networks (DBN) [13], and deep Boltzmann Machine (DBM) [21], and their corresponding learning algorithms have been proposed. These models also have been extended to model multi-modal data. Ngiam et al. [17] used multimodal DAE to learn a shared representation for both speech and visual inputs. Srivastava and Salakhutdinov [25] used multimodal DBM and DBN to learn a unified representation for both images and texts. We also noticed several deep learning methods [29, 9, 23] for learning a joint embedding space of image and text very recently. Those models have shown significant advantages at image annotation, objective classification and zero-shot learning tasks. But, to our best knowledge, neither of them has been used for cross-modal retrieval up to now.

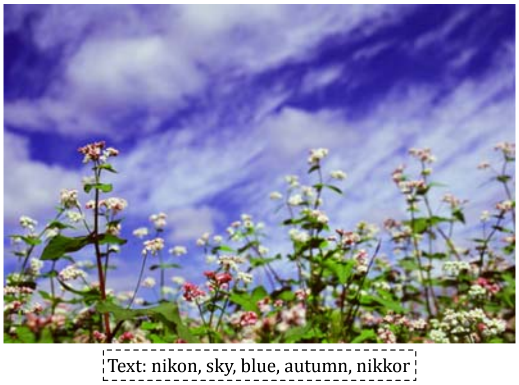


Figure 1: An image and its tags. "sky" and "blue" are common information in both image and text modalities. "nikon" and "nikkor" are text modality-specific information while "flowers" and "clouds" are image modality-specific information. Common information are the key to cross-modal retrieval tasks.

The other strategy involves a two-stage framework. It first learns or extracts features for each modality separately then uses canonical correlation analysis (CCA) [10] to build a lower-dimensional common representation space. This choice is more straightforward for cross-modal retrieval task. CCA is a method of data analysis used to discover a subspace of multiple data spaces. Given the training pairs p and q, CCA finds matrices U and V such that Up and Vq have maximum correlations. The first d canonical components of U and V could be used for projecting new input pairs into a d-dimension space where cross-modal retrieval could be conducted by calculating a simple similarity measurement, such as Euclidean or cosine distance. Rasiwasia et al. [19] proposed correlation matching mapping the features of images and texts extracted by LDA into same representation space using CCA. Ngiam et al. [17] used DAE to extract features and suggested using CCA to form a shared representation of audio and video data. As in Ngiam et al.'s work, Kim et al. [15] learned the shared semantic space of different language with DAE and CCA.

1.2 Motivation

In the first strategy, the shared layer was built by jointly learning from different modalities. We think the shared layer learnt in this way might no fit the need of cross-modal retrieval. Perceived data from different modalities for a same object normally comprise of common information in all modalities and modality-specific information. Figure 1 gives an example. There is a picture accompanying with several text tags in figure 1, where "sky" and "blue" are common information in both image and text modalities. "nikon" and "nikkor" are text modality-specific information which can hardly be acquired from the image. While "flowers" and "clouds" are only image modality-specific information which can not be captured in text tags. Intuitively, common information like "sky" and "blue" are key to cross-modal retrieval. Both text modality-specific and image modality-specific information are not necessary and even harmful to cross-modal retrieval. A shared representation learns both common and modality-specific information. Although it showed its strength on learning complementarity of data from different modalities, it is therefore not a exactly well-fitting representation for cross-modal retrieval. A representation which can learn only common information for different modalities is a better choice. This is the first motivation to build our model in this paper.

The second strategy separates correlation learning from representation learning. This strategy is too weak to exploit the complex correlations of representations from different modalities, especially when autoencoders are used to learn representations. Autoencoders can be used learn different level representation for inputs with different reconstruction errors, one problem for building correlations between data from two different modalities is that it is difficult to decide which level is best for building correlations. In other words, two different modalities may be correlated at different abstract levels of representation. Therefore, correlation learning need be considered with representation learning as a whole. This is the second motivation to build our model in this paper.

1.3 Contribution

In this paper, we propose correspondence autoencoder (Corr-AE) based on two basic uni-modal autoencoders. The difference between two-stage methods and our Corr-AE is illustrated in Figure 2. The two-stage methods ignore the correlation between different modalities when perform representation learning. Corr-AE incorporates representation learning and correlation learning into a single process. A novel loss function is here designed. It includes not only the loss of different autoencoders for all modalities, but also the loss of correlation between different modalities. This model is evaluated using three publicly available data sets. Compared against several multi-modal deep learning models, our Corr-AE demonstrate its effectiveness. Besides, based on two other multi-modal autoencoders, we extend Corr-AE to two correspondence models, Corr-Cross-AE and Corr-Full-AE. Experimental results show that based on different autoencoders, the combination of representation learning and correlation learning is more effective than the two-stage methods.

The remainder of this paper is organized as follows. In the next section, the details of the basic Corr-AE are described. The Corr-AE are then extended to Corr-Cross-AE and a combinational model. Section 3 describes the experimental results and compares the performance of different models. Conclusions are presented in Section 4.

2. LEARNING ARCHITECTURE

In this section, the details of the architecture of the basic Corr-AE are described. Then a loss function suitable for training Corr-AE for learning similar representations of different modalities is proposed. Next, two extensions of the Corr-AE are introduced. Lastly, the deep architecture with the training algorithms is described.

2.1 Correspondence Autoencoder

As illustrated in Figure 3, the Corr-AE architecture consists of two subnetworks, each a basic autoencoder. These two networks are connected by a predefined similarity measure on the code layer. Each subnetwork in the Corr-AE is

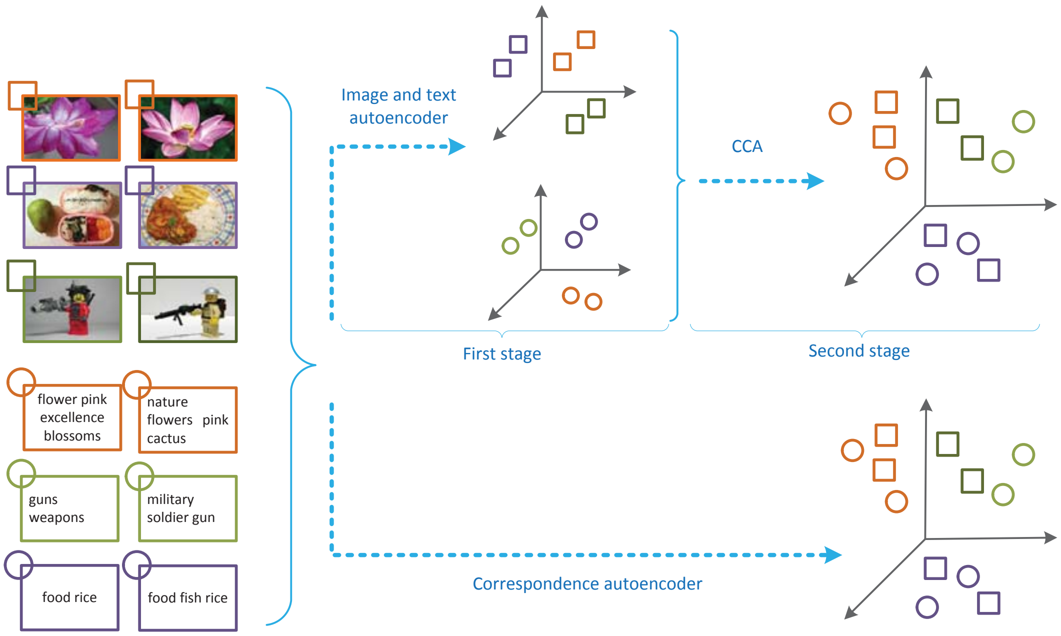

flower pink excellence blossoms
nature flowers pink cactus
guns weapons
military soldier gun
food rice
food fish rice
Image and text autoencoder
CCA
First stage
Second stage
Correspondence autoencoder

Figure 2: Difference between two-stage methods and our Corr-AE: Corr-AE incorporates representation learning and correlation learning into a single process while two-stage methods separate the two processes.

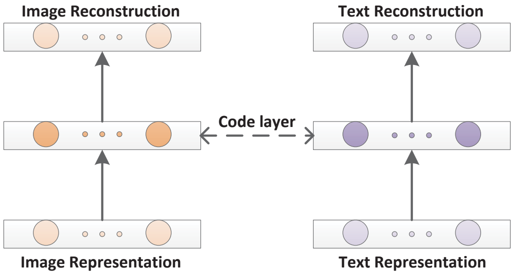


Figure 3: Correspondence autoencoder

responsible for each modality. In this way, the inputs to each subnetwork are features from one modality. During learning, the two subnetworks are coupled at their code layer using a similarity measure. After learning, the two subnetworks in the Corr-AE exhibit different parameters even if they have the same architecture. As a result, the code for new inputs can be obtained using the learned network parameters.

Formally, the mapping from the inputs of these two subnetworks to the code layers is denoted as $f(p; W_f)$ and $g(q; W_g)$, in which, f is image modality and g is text modality; W denotes the weight parameters in these two subnetworks. The subscripts denote the corresponding modalities. f and g are logistic activation function. The similarity measure between ith pair of image representation $p^{(i)}$ and the given text representation $q^{(i)}$ are here defined as follows:

$$C(p^{(i)}, q^{(i)}; W_f, W_g) = \left\| f(p^{(i)}; W_f) - g(q^{(i)}; W_g) \right\|_2^2 \qquad (1)$$

where $\|\cdot\|_2$ is the $\mathcal{L}_2$ norm.

To learn the similar representations of these two modalities for one object, a loss function given input image representation $p^{(i)}$ and its text representation $q^{(i)}$ are established. To simplify the notation, the network parameters W_f, W_g are grouped as Θ. The loss function on any pair of inputs can then be defined as follows:

$$\begin{aligned} L(p^{(i)}, q^{(i)}; \Theta) =& (1-\alpha)\left(L_I(p^{(i)}; \Theta) + L_T(q^{(i)}; \Theta)\right) \\ &+ \alpha L_C(p^{(i)}, q^{(i)}; \Theta) \end{aligned} \qquad (2)$$

where

$$L_I(p^{(i)}, q^{(i)}; \Theta) = \left\| p^{(i)} - \hat{p}_I^{(i)} \right\|_2^2 \qquad (3a)$$

$$L_T(p^{(i)}, q^{(i)}; \Theta) = \left\| q^{(i)} - \hat{q}_T^{(i)} \right\|_2^2 \qquad (3b)$$

$$L_C(p^{(i)}, q^{(i)}; \Theta) = C(p^{(i)}, q^{(i)}; \Theta) \qquad (3c)$$

Here, $\|\cdot\|_2$ is the $\mathcal{L}_2$ norm. L_I and L_T are the losses caused by data reconstruction errors for the given inputs (an image and its text) of two subnetworks, specifically image and text modalities. $\hat{p}_I^{(i)}$ and $\hat{q}_I^{(i)}$ are the reconstruction data from $p^{(i)}$ and $q^{(i)}$ respectively; L_C is the correlation loss; $\alpha(0 < \alpha < 1)$ in the total loss function (2) is a parameter used to trade off between two groups of objectives: correlation losses and reconstruction losses. An appropriate value for α is crucial. If $\alpha = 0$, the loss function degenerates to the loss function of the autoencoders. It cannot then capture any correlations between inputs from different modalities. At the other extreme, considering only correlation loss, α is set to 1. This assumes that any pair of inputs has correlations, regardless of whether the image and text inputs match or not. An intuitive interpretation for this is that the cost function only focuses on the constraints of correlations and ignores the characteristics of the data completely.

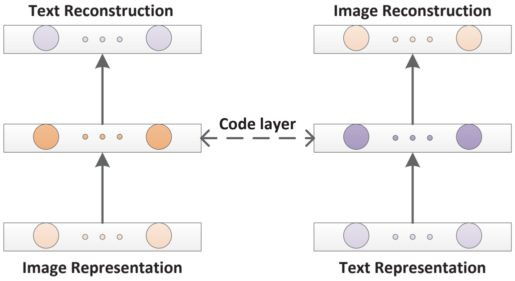


Figure 4: Correspondence cross-modal autoencoder

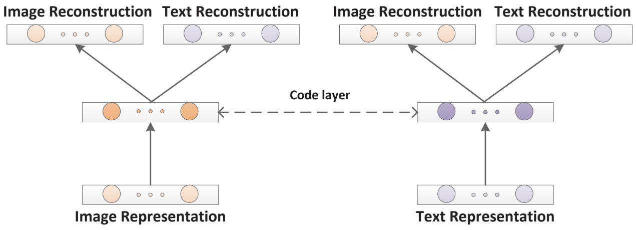


Figure 5: Correspondence full-modal autoencoder

In summary, minimizing the loss function defined in Eq. (2) enables our Corr-AE to learn similar representations from bimodal feature representations.

2.2 Correspondence Cross-modal Autoencoder

As illustrated in Figure 4, we propose the Corr-Cross-AE, which replace the basic autoencoders to cross-modal autoencoders. Unlike the basic autoencoders, which reconstruct the input itself, cross-modal autoencoders reconstruct input from different modalities. The loss function on any pair of inputs of the Corr-Cross-AE is defined as follows:

$$\begin{aligned} L(p^{(i)}, q^{(i)}; \Theta) =& (1-\alpha)\left(L_I(p^{(i)}, q^{(i)}; \Theta) + L_T(p^{(i)}, q^{(i)}; \Theta)\right) \\ &+ \alpha L_C(p^{(i)}, q^{(i)}; \Theta) \end{aligned} \tag{4}$$

where

$$L_I(p^{(i)}, q^{(i)}; \Theta) = \left\| q^{(i)} - \hat{q}_I^{(i)} \right\|_2^2 \tag{5a}$$

$$L_T(p^{(i)}, q^{(i)}; \Theta) = \left\| p^{(i)} - \hat{p}_T^{(i)} \right\|_2^2 \tag{5b}$$

$$L_C(p^{(i)}, q^{(i)}; \Theta) = C(p^{(i)}, q^{(i)}; \Theta) \tag{5c}$$

Here, $\hat{q}_I^{(i)}$ and $\hat{p}_T^{(i)}$ are the reconstruction data from image and text subnet, respectively. The meanings of other symbols are the same as in Eq. (2).

The representation learning of image modality in cross-modal autoencoder considers the information from the text modality and vice-versa. This causes some correlations to be captured in the reconstruction loss.

2.3 Correspondence Full-modal Autoencoder

The full-modal autoencoder can be viewed as a combination of a basic autoencoder and cross-modal autoencoder. This autoencoder is proposed to model the audio and video data in [17]. The basic Corr-AE are also easy to extend to Corr-Full-AE based on full-modal autoencoder. As illustrated in Figure 5, the autoencoder reconstruct not only the input itself but also input from different modalities. This "full" representation space contains information from both modalities. The loss function of any pair of inputs of the Corr-Full-AE is defined as follows:

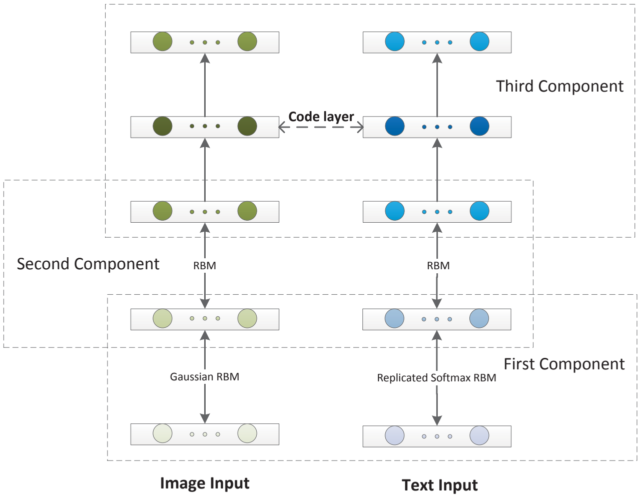


Figure 6: Deep architecture

$$\begin{aligned} L(p^{(i)}, q^{(i)}; \Theta) =& (1-\alpha)\left(L_I(p^{(i)}, q^{(i)}; \Theta) + L_T(p^{(i)}, q^{(i)}; \Theta)\right) \\ &+ \alpha L_C(p^{(i)}, q^{(i)}; \Theta) \end{aligned} \tag{6}$$

where

$$L_I(p^{(i)}, q^{(i)}; \Theta) = \left\| p^{(i)} - \hat{p}_I^{(i)} \right\|_2^2 + \left\| q^{(i)} - \hat{q}_I^{(i)} \right\|_2^2 \tag{7a}$$

$$L_T(p^{(i)}, q^{(i)}; \Theta) = \left\| p^{(i)} - \hat{p}_T^{(i)} \right\|_2^2 + \left\| q^{(i)} - \hat{q}_T^{(i)} \right\|_2^2 \tag{7b}$$

$$L_C(p^{(i)}, q^{(i)}; \Theta) = C(p^{(i)}, q^{(i)}; \Theta) \tag{7c}$$

Here, $\hat{p}_I^{(i)}$ and $\hat{q}_I^{(i)}$ are the reconstruction data from $p^{(i)}$ and $q^{(i)}$ in the image subnet; $\hat{p}_T^{(i)}$ and $\hat{q}_T^{(i)}$ are the reconstruction data from $p^{(i)}$ and $q^{(i)}$ in the text subnet. The meanings of other symbols are the same as in Eq. (2).

2.4 Deep Architecture

Data from different modalities may have very different statistical properties. This makes it difficult to capture correlations across modalities directly. To address this, a deep architecture is here proposed. This first involves using some stacked modality-friendly models to learn higher-level representations that remove such modality-specific properties. Then the Corr-AE are used to learn similar representations at a higher level.

As illustrated in Figure 6, the deep architecture has three stacked components. The first two components are all restricted Boltzmann machines (RBMs). There are two extended RBMs for the first component and two basic RBMs for the second component. To be brief, RBM[22] is an undirected graphical model with stochastic binary units in a visible layer and hidden layer but without connections between the units within these two layers. Given that each unit is distributed by Bernoulli distribution with logistic activation function, a joint probabilistic distribution of visible units and hidden units can be defined. The basic RBM can be extended to exponential family. All the models can be efficiently learned by using the contrastive divergence approximation (CD) [12]. For the first layer, Gaussian RBM [28] and replicated softmax RBM [20] can be used to model the real-valued feature vectors for image and the discrete sparse word count vectors for text, respectively. After learning the RBMs, the hidden layer can then be used as the input for the second component. The second component involves two basic RBMs, which are used to learn higher-level features for image and text. The third component can involve any one of the three correspondence autoencoders given above. The learning for Corr-AE can be performed using standard back-propagation algorithm.

Because different modalities are mapped into the same representation space by the deep architecture, it is straightforward to use the model for cross-modal retrieval tasks. For example, given a text query, we expect the relevant images could be returned[1]. After learning this deep architecture, all test images are mapped into the representation space from the three-layer image subnetwork. A new text query is mapped into the same space from the three-layer text subnetwork. The similarity between the text query and all the candidate images can be calculated by a simple distance metric in the representation space. In this way, in searching for image using text, image ranked list will return by increasing distance for any text query.

3. EXPERIMENTS

We evaluate our models on three public available real-world data sets. In this section, we first give a detailed description of these data sets. Then, the evaluation protocol used in the experiments is introduced. Next, the performance of several models is reported. Finally, an analysis of the impact of α on the models is given.

3.1 Data sets and Feature Extraction

Wikipedia. The data set [19] was collected from "Wikipedia featured articles". It contains 2,866 image/text pairs belonging to 10 semantic categories. The data set was split into three subsets: 2,173 cases as training set, 231 cases as validation set and 462 cases as testing set. For image representation, we extract the following three types of features:

- Pyramid Histogram of Words (PHOW) [6]. For PHOW features, dense SIFT descriptors is first extracted by VLfeat [27] from per training image and then a 1000-dimensional visual word codebook learned with K-means clustering.

- Gist [18]. The gist descriptor is extracted by the public available package[2] with default parameters. This results in a 512-dimensional feature vector.

- MPEG-7 descriptors [16]. We use the public available software [1] to extract four different visual descriptors (CSD, SCD, CLD, EHD) defined in MPEG-7 for image representations. The dimension of obtained MPEG-7 feature vector is 784.

Thus, each image is represented by a 2296-dimensional feature vector. For text representation, we use bag of words model. A dictionary of 3000 high-frequency words is built from all training texts. We use Python Natural Language Toolkit [3] to stem the text. This data set is available at `http://www.svcl.ucsd.edu/projects/crossmodal/`.

Pascal. The data set [8] contains 1,000 image/text pairs from 20 categories, 50 cases per categories. The images are randomly selected from 2008 PASCAL development kit. Each image is labeled with 5 sentences. We split the data into three subsets, 800 for training (40 cases per category), 100 for validation (5 cases per category) and 100 for testing (5 cases per category). The feature extraction for images and texts is the same as for the Wikipedia data set except that the dimension of text is 1,000. This data set is available at `http://vision.cs.uiuc.edu/pascal-sentences/`.

NUS-WIDE-10k. This data set is a subset of NUS-WIDE [7], which contains about 270k images with tag annotations from 81 categories. We only choose 10 categories with the largest quantity and 1,000 image/text pairs per category from NUS-WIDE. Each pair of our NUS-WIDE-10k only belongs to a single category. The 10 categories are animal, clouds, flowers, food, grass, person, sky, toy, water and window. We randomly split the data set into three subsets: 8,000 cases for training (800 cases per category), 1,000 for validation (100 cases per category) and 1,000 for testing (100 cases per category). Each image is represented by six descriptors, including 64-D color histogram, 144-D color correlogram, 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moments and 500-D bag of words based on SIFT descriptions. Each text is represented by 1000-dimensional bag of words. This data set with extracted features is available at `http://lms.comp.nus.edu.sg/research/NUS-WIDE.htm`.

As we can see, these data sets have very distinct properties. For example, the text modality of the three data sets Wikipedia, Pascal and NUS-WIDE-10k is quite different, which are article, sentences and tags, respectively. Besides, the sizes of these data sets range from 1k to 10k and the number of the categories range from 10 to 20.

3.2 Evaluation metric

We consider two cross-modal retrieval tasks: text retrieval from an image query, and image retrieval from a text query. Following [30], retrieval performance is evaluated using two metrics, mean average precision (mAP) and top 20% percentage. The first one represents the ability of learning discriminative cross-modal mapping functions while the later one reveals the ability of learning corresponding latent concepts.

mAP Given one query and first R top-ranked retrieved data, the average precision is defined as

$$\frac{1}{M}\sum_{r=1}^{R} p(r) \cdot rel(r) \tag{8}$$

where M is the number of relevant data in the retrieved result, $p(r)$ is precision of at r, and $rel(r)$ presents the relevance of a given rank (one if relevant and zero otherwise). The retrieved data is considered as relevant if it has the same

[1] Text retrieval using an image query can be done in a similar way.

[2] `http://people.csail.mit.edu/torralba/code/spatialenvelope/`

semantic label as the query. mAP is obtained by averaging AP of all the queries. We report mAP@50 ($R = 50$) in all experiments. Semantic labels are only used for evaluation. Training of our models does not need any semantic label.

Top 20% Jia et al. [14] proposed this evaluation metric for data sets without semantic label. Because there is only one ground-truth for each image-text pair in this case, the position of the ground-truth image/text in the ranked list is used to evaluate performance. Specifically, top 20% percentage is the relative number of images/texts correctly retrieved in the first 20% of the ranked list.

3.3 Baseline

For a fair comparison, the inputs to our models and baseline methods[3] are all features learned by first two components of deep architecture described in section 2.4. That means that the only difference is the third component of the deep architecture. Besides, cosine distance is used to measure the similarity in all experiments. We compare our models with three CCA based models and two multi-modal models:

- CCA-AE[15]. We first use two uni-modal autoencoders to learn higher level image feature and text feature respectively, and then use CCA[4] to learn a common representation space on the learned features.

- CCA-Cross-AE. Instead of using the uni-modal autoencoders with CCA, this method combines immediately cross-modal autoencoders and CCA.

- CCA-Full-AE. Literally, this method combines full-modal autoencoders and CCA.

- Bimodal AE[17]. We train a bimodal autoencoder to perform shared representation learning. Follow the training algorithm described in [17], we add training examples that have zero values for one of the input modalities (e.g., image) and original values for the other input modality (e.g., text), but still require the network to reconstruct both modalities (image and text). Thus, one-third of the training data has only image for input, while another one-third of the data has only text, and the last one-third of the data has both image and text. After learning the bimodal autoencoder, a single modality input(image or text) can be mapped into the shared representation space.

- Bimodal DBN[17, 24]. A bimodal DBN is obtained by connecting image and text features with a joint layer. The model does not give a direct matching function of image and text input. However, it can generate the unknown modality conditioned on a given modality. In other words, queries can be mapped into the feature space of the other modality and a suitable similarity measure can be used to retrieve results that are close to the query.

3.4 Model Architecture

We perform grid search for the number of hidden units of each layer with the setting 32, 64, 128, 256, 512, 1024. In all models, to reduce the search space, the number of units for all hidden layers in the deep architecture is restricted to the same. The validation sets are used to determine the best number of hidden units. The dimensionality of the CCA latent space and the iterations of Bimodal AEs and our three correspondence models are also determined by the validation sets. For our three correspondence models, we need to choose an appropriate value for the parameter α. The parameter α is not sensitive to data sets. So, in all data sets, the value of α for Corr-AE and Corr-Full-AE is set to 0.8. For Corr-Cross-AE, α is set to 0.2. A detailed analysis of α will be given at the end of this section.

For the models involving CCA, to achieve a better performance, CCA is applied to learn the correlation between the image and text hidden layers with different number of units. Three copies of all training data are used in Bimodal AE training. The first copy remains data from both modalities (original data). The second is only image data with text data setting to zeros. And the third is only text data with image data setting to all zeros. We do not weight the reconstruction errors from different modalities. So do our correspondence models. For the Bimodal DBN, variational mean field with ten steps is used to generate the unknown modality conditioned on a given modality. Gibbs sampling does not show improvement in our experiments. The code with parameter specifications of our three models and the baseline methods are available online[5].

3.5 Results

Table 1 summarizes the mAP scores and top 20% of the two cross-modal retrieval tasks for Wikipedia, Pascal and NUS-WIDE-10k data sets respectively. On all data sets, our three correspondence autoencoders significantly outperform other models on both tasks of text and image retrieval. Taking Corr-Full-AE as an example, we compare it with the best results on each task achieved by the five baseline models. It improves mAP scores 12.3% and 16.6% respectively on searching texts by images and searching images by texts on Wikipedia data set, improves 12.4% and 2.2% respectively on the two tasks on Pascal data set, improves 32.4% and 10.2% respectively on the two tasks on NUS-WIDE-10k data set.

Compared with CCA-AE, our Corr-AE improves average mAP value of two tasks 53.6%, 81.5%, 48.3% on Wikipedia, Pascal, NUS-WIDE-10k data sets respectively. Compared with CCA-Cross-AE, our Corr-Cross-AE improves 57.9%, 73.6%, 28.3% on the three data sets. Compared with CCA-Full-AE, our Corr-Full-AE improves 12.8%, 71.2%, 46.7% on the three data sets. Based on three different multi-modal autoencoders, our correspondence models significantly outperform the two-stage methods. The advantage of our correspondence models over two-stage models is that correspondence models combined representation learning and correlation learning into a whole. In other words, compared with our correspondence models, the two-stage methods are sub-optimal. It is also noticeable that the difference within the three correspondence autoencoders are found to be smaller than the difference among the three CCA-AEs. This also demonstrates the effectiveness of the combination of representation learning and correlation learning processes.

Compared with Bimodal AE, our Corr-AE improves mAP scores 15.6% and 10.4% respectively on searching texts by

[3] The code for all baseline models are available online at `https://github.com/nitishsrivastava/deepnet`.

[4] Matlab code of the CCA can be downloaded at `http://www.davidroihardoon.com/Professional/Code.html`.

[5] https://github.com/fangxiangfeng/deepnet

Table 1: Results of two retrieval protocols: *mAP* scores and top 20% on three data sets.

(a) Wikipedia

Model	*mAP*			Top 20%		
	Image Query	Text Query	Average	Image Query	Text Query	Average
CCA-AE[15]	0.213	0.235	0.224	28.35	23.59	25.97
CCA-Cross-AE	0.197	0.230	0.214	25.54	28.14	26.84
CCA-Full-AE	0.293	0.331	0.312	51.08	49.57	50.33
Bimodal AE[17]	0.282	0.327	0.305	44.16	42.42	43.29
Bimodal DBN[17, 24]	0.189	0.222	0.206	26.19	31.6	28.90
Corr-AE	0.326	0.361	0.344	56.06	55.19	55.63
Corr-Cross-AE	**0.336**	0.341	0.338	55.41	**58.66**	57.04
Corr-Full-AE	0.335	**0.368**	**0.352**	**57.36**	57.79	**57.58**

(b) Pascal

Model	*mAP*			Top 20%		
	Image Query	Text Query	Average	Image Query	Text Query	Average
CCA-AE[15]	0.161	0.153	0.157	36	30	33
CCA-Cross-AE	0.137	0.182	0.159	18	16	17
CCA-Full-AE	0.148	0.177	0.163	32	35	33.5
Bimodal AE[17]	0.250	0.270	0.260	68	68	68
Bimodal DBN[17, 24]	0.219	0.219	0.219	55	61	58
Corr-AE	**0.290**	0.279	**0.285**	72	67	69.5
Corr-Cross-AE	0.271	**0.280**	0.276	**78**	**75**	**76.5**
Corr-Full-AE	0.281	0.276	0.279	74	73	73.5

(c) NUS-WIDE-10k

Model	*mAP*			Top 20%		
	Image Query	Text Query	Average	Image Query	Text Query	Average
CCA-AE[15]	0.199	0.268	0.234	37	33.8	35.4
CCA-Cross-AE	0.199	0.344	0.272	29	47.7	38.35
CCA-Full-AE	0.241	0.242	0.242	37.1	38.2	37.65
Bimodal AE[17]	0.250	0.297	0.274	30.2	35.4	32.8
Bimodal DBN[17, 24]	0.173	0.203	0.188	25.3	27	26.15
Corr-AE	0.319	0.375	0.347	47.1	53.5	50.3
Corr-Cross-AE	**0.349**	0.348	0.349	**53.1**	**59.7**	**56.4**
Corr-Full-AE	0.331	**0.379**	**0.355**	49.6	56.5	53.05

images and searching images by texts on Wikipedia data set, improves 16.0% and 3.3% respectively on the two tasks on Pascal data set, improves 27.6% and 26.3% respectively on the two tasks on NUS-WIDE-10k data set. Both Bimodal AE and Bimodal DBN model the multi-modal inputs with a shared hidden layer. Representation learned in shared hidden layer should cover difference between two modalities for achieving good performance in both autoencoders or RBMs. It therefore focuses more on learning complementarity instead of correlation across data from different modalities.

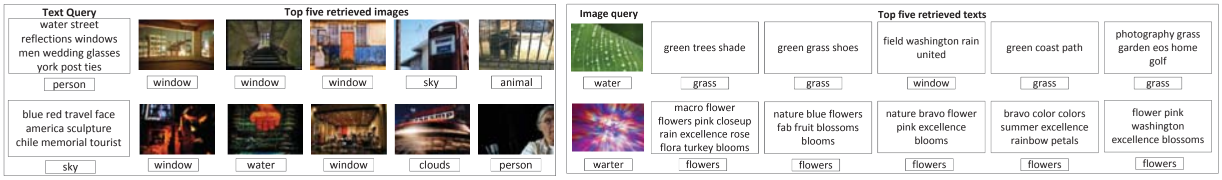
Figure 7 shows three examples of text-based cross-modal using our Corr-Full-AE and the best baseline method. In these examples, the top four retrieved images by Corr-Full-AE are all relevant to correspondence text query. Figure 8 shows several failure cases of Corr-Full-AE on the NUS-WIDE-10k test data set.

As for the top 20% percentage metric, our three correspondence autoencoders also significantly outperform the other baseline models. There is only one ground-truth for each image-text pair under this evaluation metric. Figure 9 shows several top-1-hit retrieval examples by our Corr-Full-AE. In these cases, the first returned results by the query are all ground truth on both image query text and text query image tasks.

3.6 Analysis of α

Here, *mAP* values of three correspondence autoencoders with different values of α in all data sets are given in Figure 10. Both too small values and too large values of α show poor performance in the three data sets. This is consistent with the impact of α in the present models. Too small values of α overemphasize the “individuality” of data and ignore the correlations. Too large values overemphasize the correlations and ignore the “individuality” of the data.

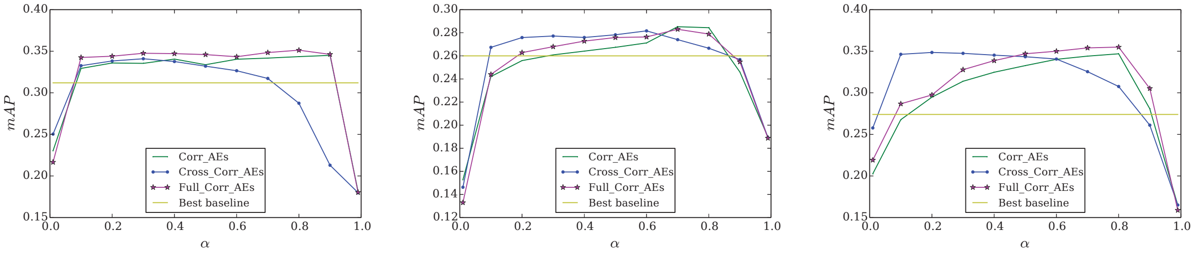
To validate the hypothesis of the effect of α to the “individuality” of data, we use tSNE [26] to visualize the image and text representation learned by our Corr-Full-AE. As shown in Figure 11, when $\alpha = 0.01$, image and text representation space are almost disjoint. In this case, image and text have strong “individuality” so that the learned representation of image and text have no correlations. To the other extreme of the α, when $\alpha = 0.99$, “individuality” of im-


Text Query
Top five retrieved images
water street reflections windows men wedding glasses york post ties
person
window
window
window
sky
animal
blue red travel face america sculpture chile memorial tourist
sky
window
water
window
clouds
person
Image query
Top five retrieved texts
water
green trees shade
grass
green grass shoes
grass
field washington rain united
window
green coast path
grass
photography grass garden eos home golf
grass
warter
macro flower flowers pink closeup rain excellence rose flora turkey blooms
flowers
nature blue flowers fab fruit blossoms blooms
flowers
nature bravo flower pink excellence blooms
flowers
bravo color colors summer excellence rainbow petals
flowers
flower pink washington excellence blossoms
flowers

(a) Failure cases: text query image

(b) Failure cases: image query text

Figure 8: Several failure cases of Corr-Full-AE on the NUS-WIDE-10k test data set. The words under the texts/images are the correspondence semantic labels. In the cases of text query images, the images are related to the subset of the text query. For example, in the first case, the first retrieved image is related to the "windows" and "glasses". In the cases of image query texts, the image queries are too difficult to be recognized. For example, in the second case, the image query is easy to be recognized as the flowers falsely.

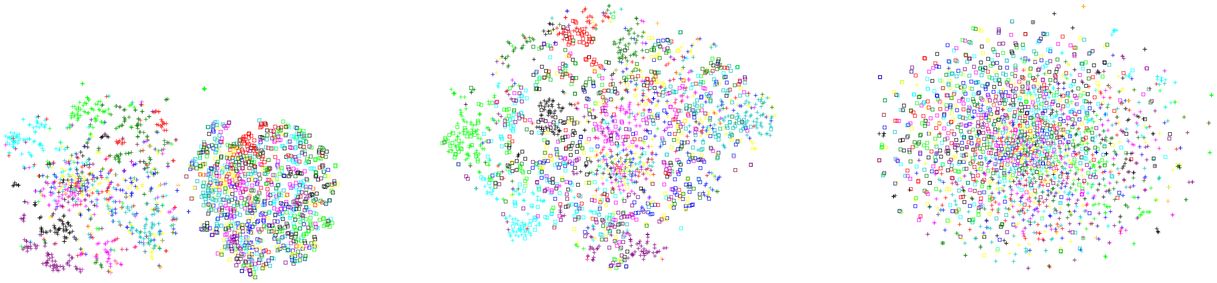


(a) Wikipedia

(b) Pascal

(c) NUS-WIDE-10k

Figure 10: mAP values of three correspondence autoencoders with different values of α in all data sets. The X-axis gives values for α. The Y-axis gives the mAP scores. The yellow line denotes the mAP score of the best baseline model, which does not change with α.



(a) $\alpha = 0.01$

(b) $\alpha = 0.8$

(c) $\alpha = 0.99$

Figure 11: Image and text representation visualization of different value for α on the NUS-WIDE-10k test data set. Different shapes denote different modalities: "square" for image, "plus" for text. Different colors denote different semantic categories.

age and text representation is missing due to the confusion representation space. In this case, any image-text pair has correlations, regardless of whether the image and text inputs match or not.When α is set to 0.8, the representation space is quite effective to the cross-modal retrieval task, since a large number of image-text pairs with same semantic labels are clustered.

As shown in Figure 10, the yellow line denotes the mAP scores of the best baseline model. On all data sets, our three correspondence autoencoders outperform the best baseline when α is in a quite large range.

4. CONCLUSION

In this work, a cross-modal learning model, Corr-AE, is presented. This model incorporates representation learning and correlation learning into a single process, so combining autoencoder cost with correlation cost. Corr-AE is here extended to Corr-Cross-AE by replacing the basic autoencoder with a cross-modal autoencoder. Finally, Corr-Full-AE is built by combining Corr-AE and Corr-Cross-AE. These three models are compared to CCA-based and multimodal deep learning models and found to be effective in cross-modal retrieval of information from three publicly available data sets.

5. ACKNOWLEDGEMENTS

This work was partially supported by National Natural Science Foundation of China(No. 61273365), National High Technology Research and Development Program of China(No. 2012AA011103), discipline building plan in 111 base(No. B08004), the Fundamental Research Funds for the Central Universities (No. 2013RC0304) and Engineering Research Center of Information Networks, Ministry of Education.

6. REFERENCES

[1] M. Bastan, H. Cam, U. GÍźdÍźkbay, and z. Ulusoy. Bilvideo-7: an mpeg-7- compatible video indexing and retrieval system. *IEEE MultiMedia*, 17(3):62–73, 2010.

[2] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009.

[3] E. L. Bird, Steven and E. Klein. In *Natural Language Processing with Python*. O'Reilly Media Inc., 2009.

[4] D. M. Blei and M. I. Jordan. Modeling annotated data. *ACM SIGIR*, pages 127–134, 2003.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[6] A. Bosch, A. Zisserman, and X. Muñoz. Image classification using random forests and ferns. In *ICCV*, pages 1–8. IEEE, 2007.

[7] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng. Nus-wide: A real-world web image database from national university of singapore. In *CIVR*, Santorini, Greece., 2009.

[8] A. Farhadi, S. M. M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. A. Forsyth. Every picture tells a story: Generating sentences from images. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *ECCV (4)*, volume 6314 of *Lecture Notes in Computer Science*, pages 15–29. Springer, 2010.

[9] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[10] D. R. Hardoon, S. SzedmÍćk, and J. Shawe-Taylor. Canonical correlation analysis; an overview with application to learning methods. *Neural Computation*, 16:2639–2664, 2004.

[11] G. Hinton and R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[12] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002.

[13] G. E. Hinton, S. Osindero, and Y. Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

[14] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. *ICCV*, pages 2407–2414, 2011.

[15] J. Kim, J. Nam, and I. Gurevych. Learning semantics with deep belief network for cross-language information retrieval. *COLING*, pages 579–588, 2012.

[16] B. S. Manjunath, J. R. Ohm, V. V. Vinod, , and A. Yamada. Color and texture descriptors. *IEEE Trans. Circuits and Systems for Video Technology, Special Issue on MPEG-7*, 11(6):703–715, June 2001.

[17] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. *ICML*, pages 689–696, 2011.

[18] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[19] 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. *ACM MM*, pages 251–260, 2010.

[20] R. Salakhutdinov and G. Hinton. Replicated softmax: an undirected topic model. *NIPS*, pages 1607–1614, 2009.

[21] R. R. Salakhutdinov and G. G. Hinton. An efficient learning procedure for deep Boltzmann machines. *Neural computation*, 24(8):1967–2006, 2012.

[22] P. Smolensky. Parallel distributed processing: explorations in the microstructure of cognition, vol. 1. chapter Information processing in dynamical systems: foundations of harmony theory, pages 194–281. MIT Press, Cambridge, MA, USA, 1986.

[23] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, pages 935–943, 2013.

[24] N. Srivastava and R. Salakhutdinov. Learning representations for multimodal data with deep belief nets. *ICML Representation Learning Workshop*, 2012.

[25] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep Boltzmann machines. *NIPS*, pages 2231–2239, 2012.

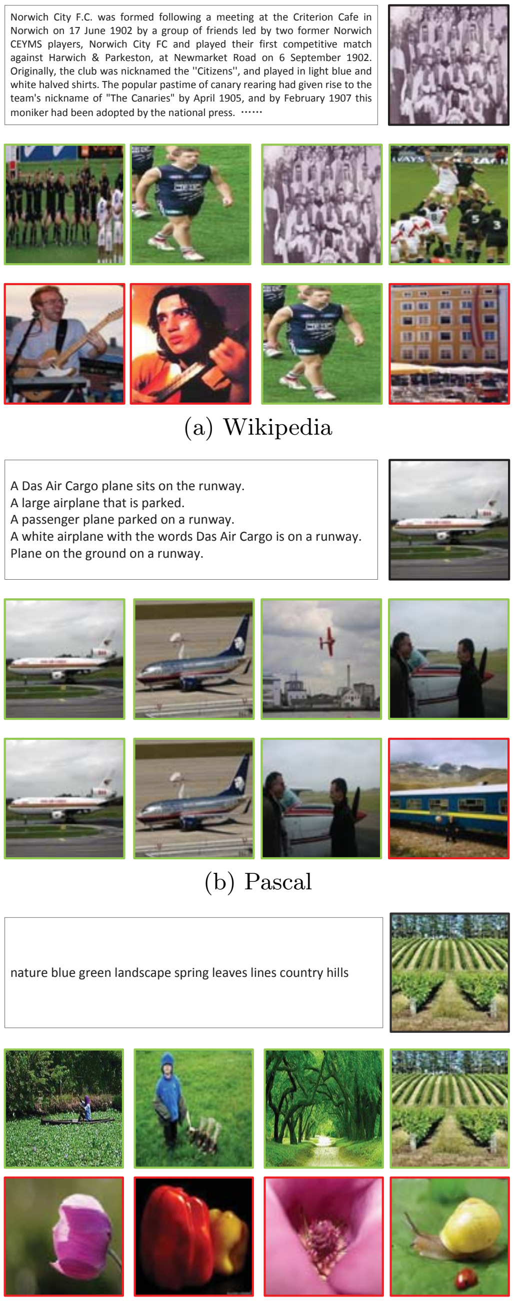
[26] L. van der Maaten and G. Hinton. Visualizing high-dimensional data using t-sne. *JMLR*, 2008.

[27] A. Vedaldi and B. Fulkerson. Vlfeat: an open and portable library of computer vision algorithms. In A. D. Bimbo, S.-F. Chang, and A. W. M. Smeulders, editors, *ACM MM*, pages 1469–1472, 2010.

[28] M. Welling, M. Rosen-Zvi, and G. Hinton. Exponential family harmoniums with an application to information retrieval. In *NIPS*, pages 501–508, Vancouver, 2004. Morgan Kaufmann.

[29] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: Learning to rank with joint word-image embeddings. In *ECML*, pages 21–35, 2010.

[30] Y. Zhuang, Y. F. Wang, F. Wu, Y. Zhang, and W. Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *AAAI*, 2013.



Norwich City F.C. was formed following a meeting at the Criterion Cafe in Norwich on 17 June 1902 by a group of friends led by two former Norwich CEYMS players, Norwich City FC and played their first competitive match against Harwich & Parkeston, at Newmarket Road on 6 September 1902. Originally, the club was nicknamed the ''Citizens'', and played in light blue and white halved shirts. The popular pastime of canary rearing had given rise to the team's nickname of "The Canaries" by April 1905, and by February 1907 this moniker had been adopted by the national press. ······

(a) Wikipedia

A Das Air Cargo plane sits on the runway.
A large airplane that is parked.
A passenger plane parked on a runway.
A white airplane with the words Das Air Cargo is on a runway.
Plane on the ground on a runway.

(b) Pascal

nature blue green landscape spring leaves lines country hills

(c) NUS-WIDE-10k

Figure 7: Three examples of text-based cross-modal using our Corr-Full-AE and best baseline method. In each example, the query text and its correspondence image are shown on the top; retrieved images of our Corr-Full-AE are presented in the middle; retrieved images of the best baseline model are presented at the bottom. Relevant matches are shown with green bounding box. Irrelevant matches are shown with red bounding box. The text queries come from sport, aeroplane and grass category respectively.

War broke out again when Armenia and Iberia revolted against Sassanid rule in 571 AD, following clashes involving Roman and Persian proxies in Yemen and the Syrian desert, and Roman negotiations for an alliance with the Turks against Persia.John of Epiphania, ''History'', , gives an additional for the outbreak of the war: " contentiousness increased even further when Justin did not deem to pay the Medians the five hundred pounds of gold each year previously agreed to under the peace treaties and let the Roman State remain forever a tributary of the Persians."······

It is a commercially important species to the extent that its fins are prized for soup and its meat and oil frequently used. It is used fresh, smoked, dried and salted for human consumption and its hide is used for leather. It is subject to fishing pressure throughout virtually its whole range &mdash; although it is more often taken as by-catch than by design, since it takes bait from longlines intended for other species. Famed oceanographic researcher Jacques Cousteau described the oceanic whitetip as "the most dangerous of all sharks". ······

Andean Peru has been recognized as one of six global areas that saw the indigenous development of civilization, and one of two, along with Mesoamerica, in the Western Hemisphere. Norte Chico has pushed back the horizon for complex societies in the Peruvian region by centuries. The Chavín culture, circa 900 BC, had long been considered the first civilization of the area and is still regularly cited as such in general works. ······

(a) Wikipedia

Two bicyclists and a pedestrian are waiting to cross the road.
Two bicyclists wait to cross a busy road.
Two blonde girls attempt to cross a busy street in town with their bicycles.
Two women watching the street with their bikes to see when they may cross.
Two women with bicycles are trying to cross the street.

A black and white cow in a grassy field stares at the camera.
A black and white cow standing in a grassy field.
A black and white cow stands on grass against a partly cloudy blue sky.
A cow is gazing over the grass he is about to graze.
Black and white cow standing in grassy field.

A group of friends pose with furry hats.
A group of people, four of which are women, two of which are men, pose wearing fuzzy hats with earflaps.
A group of peoples posing in fur hats.
Group of teenagers being silly.
Visitors pose wearing fur hats.

(b) Pascal

nature green quality fly spider

white macro flowers pink rxcellence balcony

food fish mexico

(c) NUS-WIDE-10k

Figure 9: Several image-text pairs of three data sets. In these pairs, the closest images(texts) to the text(image) queries are all ground truth. In figure (a), three image-text pairs come from warfare, biology, history category respectively. In figure (b), three pairs come from bicycle, cow, person category respectively. In figure (c), three pairs come from animal, flower, food category respectively. Various images(texts) can be retrieved effectively.