Cross-Modality Feature Learning via Convolutional Autoencoder

XUELIANG LIU and MENG WANG, Hefei University of Technology, China
ZHENG-JUN ZHA, University of Science and Technology of China, China
RICHANG HONG, Hefei University of Technology, China

Learning robust and representative features across multiple modalities has been a fundamental problem in machine learning and multimedia fields. In this article, we propose a novel MUltimodal Convolutional AutoEncoder (MUCAE) approach to learn representative features from visual and textual modalities. For each modality, we integrate the convolutional operation into an autoencoder framework to learn a joint representation from the original image and text content. We optimize the convolutional autoencoders of different modalities jointly by exploiting the correlation between the hidden representations from the convolutional autoencoders, in particular by minimizing both the reconstructing error of each modality and the correlation divergence between the hidden feature of different modalities. Compared to the conventional solutions relying on hand-crafted features, the proposed MUCAE approach encodes features from image pixels and text characters directly and produces more representative and robust features. We evaluate MUCAE on cross-media retrieval as well as unimodal classification tasks over real-world large-scale multimedia databases. Experimental results have shown that MUCAE performs better than the state-of-the-arts methods.

CCS Concepts: • **Information systems → Content analysis and feature selection**; • **Computing methodologies → Neural networks**;

Additional Key Words and Phrases: Cross modality, feature learning, convolutional autoencoder

ACM Reference format:
Xueliang Liu, Meng Wang, Zheng-Jun Zha, and Richang Hong. 2019. Cross-Modality Feature Learning via Convolutional Autoencoder. *ACM Trans. Multimedia Comput. Commun. Appl.* 15, 1s, Article 7 (January 2019), 20 pages.
https://doi.org/10.1145/3231740

This work was supported in part by the National 973 Program of China under grant 2014CB347600; in part by the National Natural Science Foundation of China (NSFC) under grants 61432019, 61732008, 61725203, 61632007, 61502139, 61622211, 61472392, and 61620106009; in part by the Natural Science Foundation of Anhui Province under grant 1608085MF128; and in part by the Fundamental Research Funds for the Central Universities under grant WK2100100030.
Authors' addresses: X. Liu and M. Wang, Hefei University of Technology, 193 Tuixi Rd., Hefei, Anhui, 230009, China; emails: liuxueliang@hfut.edu.cn, eric.mengwang@gmail.com; Z.-J. Zha, University of Science and Technology of China, 96 Jinzhai Rd., Hefei, Anhui, 230000, China; email: zhazj@ustc.edu.cn; R. Hong, Hefei University of Technology, 193 Tuixi Rd, Hefei, Anhui, 230009, China; email: hongrc.hfut@gmail.com.

1551-6857/2019/01-ART7 $15.00
https://doi.org/10.1145/3231740

1 INTRODUCTION

Recent years have witnessed the rapid growth of social media websites, such as Flickr, YouTube, and Facebook, on each of which the information is presented in multiple media types (Imran et al. 2015; Qian et al. 2014). For example, images on Flickr are associated with text, such as titles, tags, and comments, while videos on YouTube contain both visual and audio signals. On the other hand, users usually have cross-modal information needs, for example, seeking images or video clips to illustrate text descriptions. The prevalence of multimodal data and the increase of users' cross-modal information needs has led to a pressing demand of cross-modal information processing technology (Lew et al. 2006; Mei et al. 2014; Wang et al. 2015a). However, most existing approaches are only able to handle unimodal data and are not applicable to process multimodal data since the data of different modalities are represented in different feature spaces (Zheng et al. 2018).

To address the above issue, some approaches have been proposed to bridge different modalities (Feng et al. 2015; Kim et al. 2012; Liu et al. 2015; Rasiwasia et al. 2010). Since the features extracted from different modalities cannot be matched directly, a group of mapping functions are learned to model the correlation among different modalities, and then they are used to project data from different modalities into a common latent space. Among these approaches, the Canonical Correlation Analysis (CCA) algorithm has been employed to solve the cross-modality representation learning problem. It learns a shared feature space by maximizing the correlation between projected vectors of two modalities (Costa Pereira et al. 2014; Rasiwasia et al. 2010). Besides CCA, there are some other learning methods, such as graph-based modeling (Wang et al. 2009; Zhang et al. 2014) and kernel methods (Guillaumin et al. 2010; Yang et al. 2009), that are also used for cross-modality modeling. These approaches rely on handcrafted features, which are extracted to represent samples and then used in subsequent cross-modality learning procedures. The resultant features might be suboptimal for cross-modality analysis.

Recently, deep learning has made great achievements in many multimedia and computer vision tasks (LeCun et al. 2015; Zhang et al. 2016). For example, the autoencoder has been developed to learn effective features for multimedia content representation (Feng et al. 2015; Krizhevsky and Hinton 2011; Ngiam et al. 2011). It exploits a neural network to learn representations for given samples to minimize the reconstruction error. Moreover, the convolutional neural network (CNN) is the ideal neural network to learn the universal representation from original content (Donahue et al. 2013; Krizhevsky et al. 2012). Compared with the handcrafted features, the deep representation learned by CNN conveys rich semantic information and has obtained state-of-the-art performance.

Motivated by the above observations, we investigate combining the autoencoder and convolutional neural network to learn representative features from original content automatically. We propose a novel MUltimodal Convolutional AutoEncoder (MUCAE) network, which correlates hidden representations of two convolutional autoencoders. Compared with the conventional autoencoder network, the proposed convolutional autoencoder utilizes the convolutional operation between layers to learn a comprehensive hidden representation automatically. We design a new objective function, which minimizes both the error of representation learning for each modality and the correlation divergence between hidden representations corresponding to various modalities. By optimizing the objective function, the two autoencoders are learned jointly. The framework of our proposed approach is illustrated in Figure 1. Compared with previous works (Feng et al. 2015; Ngiam et al. 2011) that employ deep neural networks to learn a latent representation from handcrafted features, such as LLC (Yu et al. 2009) and bag of visual words for visual content, or tf-idf for textual content, the proposed MUCAE network could not only learn an effective representation from original content automatically but also learn the latent representation of different modalities jointly in the cross-modal modeling task. We evaluate MUCAE on two important multimedia

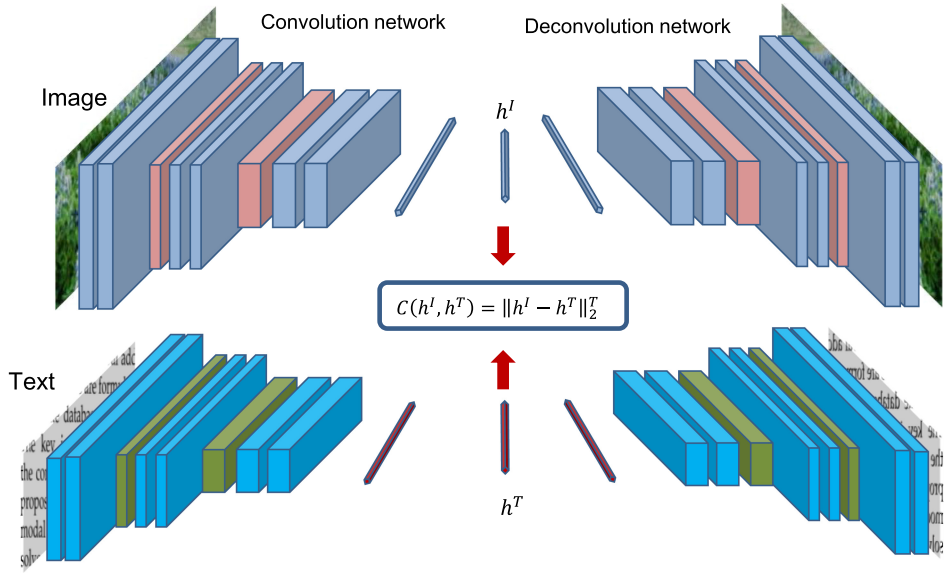


Fig. 1. The framework of the proposed multimodal convolutional autoencoder, which learns the hidden representation from different modalities and minimizes their divergence jointly.

tasks, i.e., cross-media retrieval and unimodal image classification. The multiple modalities are encoded jointly by MUCAE to learn the hidden representation and the resultant features are used for retrieval and classification. We conduct extensive evaluation on two real-world large-scale multimedia datasets, and the experiment results have demonstrated the effectiveness of the proposed MUCAE approach. The main contributions of this article can be summarized as follows:

- We propose to integrate CNN into the autoencoder framework to learn representative and effective representation from original multimodal content, such as image pixels and text characters.
- To address the cross-modal feature learning problem, we design a novel objective function to minimize the error of representation learning for each modality and maximize the correlation between hidden representations of different modalities jointly.
- The learned features are evaluated on cross-media retrieval and unimodal classification tasks with two evaluation benchmarks, and the results show the efficacy of the proposed method.

The reminder of this article is organized as follows. Section 2 reviews related works, and Section 3 presents the proposed MUCAE approach in detail. Section 5 reports experimental results and analysis, followed by conclusions in Section 6.

2 RELATED WORKS

In this section, we review related works on multimodal feature learning and deep feature learning.

2.1 Multimodal Feature Learning

Multimodal feature learning plays a fundamental role in many multimedia and computer vision tasks (Hong et al. 2014). It aims at learning a joint representation from various modalities. Many multimodal feature learning methods have been proposed in recent years. One of the most widely

used learning method is the Canonical Correlation Analysis (CCA) (Hardoon et al. 2004). CCA could optimize linear combinations of two feature matrices to maximize their correlation and identify the basis vectors for feature matrices from different modalities, while the correlation between the projected representations of original features is mutually maximized. In Costa Pereira et al. (2014), CCA was employed to learn joint representation from image and text data, and three solutions to cross-modal retrieval were then developed based on the learned representation. CCA was also used to exploit the correlations between web images and text captions in Hardoon et al. (2004) as well as that between images and audio signals in Li et al. (2003).

Some other works used the manifold learning algorithm for multimodal representation learning. These methods learn a manifold from a distance matrix of multimodal data by joining the distances of each individual modality. For example, Yang et al. (2009) presented a cross-media retrieval framework, where the query example and the retrieved results can be of different media types. The multimodal correlation space was constructed by exploring the semantic correlation of different media modalities. Zhuang et al. (2008) proposed a uniform cross-media correlation graph to mine the semantic correlations among media objects of different modalities uniformly. Hong et al. (2017) constructed a joint semantic-visual space by modeling visual descriptors and semantic attributes jointly to bridge the semantic gap into a single framework. Wang et al. (2016a) developed a method for learning joint embeddings of images and text using a two-branch neural network with multiple layers of linear projections followed by nonlinearities. Mahadevan et al. (2011) proposed a maximum covariance unfolding algorithm to compute a common low-dimensional embedding of two given sources. It maximized the cross-modal correlations while preserving the local distances. In Yan et al. (2016), a cross-media active learning algorithm is proposed to reduce the effort on labeling images for training the image classifiers. In this proposal, the text description is taken as privileged information since text is missing in the testing data. In our work, we use both textual and visual data equivalently to train a cross-modality neural network.

There are also some other learning techniques used in cross-modality modeling, such as graph/hypergraph modeling (Wang et al. 2009, 2015b; Zhang et al. 2014) and the kernel method (Guillaumin et al. 2010). These approaches achieved success on a particular task with predefined rules but failed to generalize as a uniform cross-modality modeling framework.

2.2 Deep Feature Learning

In recent years, deep learning has obtained impressive performance on various multimedia and computer vision tasks (Bengio et al. 2013; Schmidhuber 2015). Given the efficacy of deep learning, several deep neural-network-based feature learning approaches have been proposed recently. According to the learning algorithm used in them, the approaches could be categorized into two groups, i.e., unsupervised and supervised deep feature learning methods.

On one hand, feature learning with unsupervised learning algorithms aims at reconstructing input samples based on predefined rules. The autoencoders (Tan 2008) could learn representative features to reconstruct the input samples with minimal reconstruction error. In Xue et al. (2015), autoencoders were employed to fuse the audio and lyrics for music mood classification. Autoencoders and their variants were also applied to multimodal representation learning (Alain and Bengio 2014; Chen et al. 2014). In Ngiam et al. (2011), the autoencoder network was used to extract intermediate-level representation between visual and textual modalities. The combination of these two modalities boosted the performance of recognition. Feng et al. (2015) proposed a correspondence autoencoder network to learn the hidden representations between visual and textual content, minimizing the correlation learning error between the hidden representations of the two modalities. In Han et al. (2016), a denoising autoencoder was employed to learn robust and

representative features in an unsupervised manner and applied to train salience detection models from raw image data.

Similar to the autoencoder, the deep Boltzmann machine (Srivastava and Salakhutdinov 2012) was also used to extract a meaningful representation of multimodal data. In order to measure the similarity between image and text, Fang et al. (2015) developed a deep multimodal similarity model that learns two neural networks to map images and text fragments to a common representation. The similarity between image and text could be measured by cosine similarity between their corresponding representation. Zhao et al. (2016) proposed a probabilistic graphical model based on the restricted Boltzmann machine to learn a low-dimensional latent semantic representation for recognizing complex activities and events in uncontrolled web videos.

On the other hand, the supervised deep network was also used for feature learning (Li et al. 2013; Zhong et al. 2015). The convolutional neural network had success in image classification recently. This demonstrates its ability in learning effective and rich representations. Some recent works (Oquab et al. 2014; Simonyan and Zisserman 2014) demonstrated that the CNN models trained on large-scale datasets with data diversity could be directly applied to extract deep visual features and obtain promising results on visual recognition tasks. Sharif Razavian et al. (2014) demonstrated that features extracted from the pretrained CNN could be utilized as a generic image representation for diverse visual recognition tasks, such as scene recognition, fine-grained recognition, attribute detection, and image retrieval. Oquab et al. (2014) transferred the image representations from CNN trained on large-scale datasets to other visual recognition tasks with a limited amount of training samples. In Babenko et al. (2014), the responses from the top layer of a pretrained CNN were used as the high-level representation for image retrieval and achieved promising performance. Besides the CNN, the recurrent neural network (RNN) is also employed on representation learning. Lev et al. (2016) combined RNN and Fisher vectors method to encode sequences and provide effective representations.

While the convolutional networks could learn effective features from unimodal and raw content, the autoencoder could learn a joint representation from the features of different modalities. In this article, we integrate the CNN and autoencoder in a unified framework to learn representative features from multimodal raw content jointly.

3 MULTIMODAL CONVOLUTIONAL AUTOENCODER

In this section, we elaborate the proposed MUCAE approach. For the sake of clarity, we introduce MUCAE with visual and textual modalities, though it can deal with various modalities.

3.1 Problem Formulation

Given a multimedia dataset $\mathcal{D}= \{\mathbf{X}^T, \mathbf{X}^I\}$, with $\mathbf{X}^T$ and $\mathbf{X}^I$ being the text and image samples, respectively, where $\mathbf{X}^T \in \mathbb{R}^{N\times M^T}$ and $\mathbf{X}^I \in \mathbb{R}^{N\times M^I}$. N is the size of the dataset and M^T and M^I are the feature dimensionalities of $\mathbf{X}^T$ and $\mathbf{X}^I$, respectively. The objective is to learn a joint feature representation across textual and visual modalities. The important notations and definitions are listed in Table 1.

3.2 The Autoencoder Neural Network

An autoencoder is a neural network that learns representative features for given samples automatically. The model could reconstruct the samples with minimal reconstruction errors and capture their essential representation. Hence, it is a promising solution to learn the latent representation of a particular modality.

Table 1. Notations and Definitions

Notation	Definition
$\mathbf{X^T}$	$X^T \in \mathbb{R}^{N \times M^T}$ denotes the textual feature for the given dataset, where N is the size of the dataset and M^T is the feature dimension
$\mathbf{X^I}$	$X^I \in \mathbb{R}^{N \times M^I}$ denotes the visual feature for the given dataset, where M^I is the feature dimension
$\mathbf{W}_l$	The convolutional filter parameter of the lth layer
$\mathbf{b}_l$	The bias parameter of the lth layer
E	The reconstruction error
C	The correlation divergence of different modalities
Ω	The objective loss of the MUCAE model

For a feature vector $\mathbf{x}$, the autoencoder transforms it to a latent representation vector $\mathbf{z} = s(\mathbf{W}\mathbf{x} + \mathbf{b})$ by an activation function s and the parameters $\mathbf{W}$ and $\mathbf{b}$. The latent representation $\mathbf{z}$ is subsequently used to reconstruct the feature vector $\hat{\mathbf{x}} = s(\mathbf{W'}\mathbf{z} + \mathbf{b'})$. The objective is to make $\hat{\mathbf{x}}$ as close to $\mathbf{x}$ as possible. Hence, the parameters are optimized to minimize the least square reconstruction error as follows:

$$\mathbf{W}^\star, \mathbf{b}^\star, \mathbf{W'}^\star, \mathbf{b'}^\star = \underset{\mathbf{W}, \mathbf{b}, \mathbf{W'}, \mathbf{b'}}{\arg\min} \|x - \hat{x}\|_2^2 . \tag{1}$$

By the autoencoder, the latent representation h could preserve the main information of the original feature x (Bengio et al. 2007). A stacked autoencoder network could be constructed by stacking multiple autoencoders by way of feeding the output from the ith hidden layer as the input to the $(i+1)$-th hidden layer.

3.3 Multimodal Convolutional Autoencoder

As aforementioned, conventional autoencoders rely on handcrafted features and their performances are influenced by the efficacy of the features. Recently, the CNN has achieved impressive performance on many computer vision and text analysis tasks (Ji et al. 2013; Krizhevsky et al. 2012; Zhang and LeCun 2015). The significant achievements demonstrate its effectiveness in abstracting the semantics from raw features, such as image pixels and text characters. Here, we propose a novel MUCAE network that integrates the CNN and autoencoder and exploits the advantages of both the CNN and autoencoder. MUCAE utilizes the CNN to extract the latent representation from textual and visual modalities jointly and exploits the correlation between the two modalities.

As shown in Figure 1, the network architecture consists of two stacked autoencoder subnetworks, which are connected by a predefined similarity measure in the encoding layer. Each subnetwork is similar to the conventional autoencoder network. The output from each layer is the convolutional responses of the corresponding input with a group of filters. Let $\mathbf{x}_l^I$ and $\mathbf{x}_l^T$ denote the output of layer l in image and text pathways, respectively, each of which has L layers in total. $\mathbf{x}_l^I$ and $\mathbf{x}_l^T$ could be obtained by the following equation in layer-wise processing:

$$\begin{aligned} \mathbf{x}_l^I &= f_l^I \left(\mathbf{x}_{(l-1)}^I * \mathbf{W}_{(l)}^I + \mathbf{b}_{(l)}^I \right), \\ \mathbf{x}_l^T &= f_l^T \left(\mathbf{x}_{(l-1)}^T * \mathbf{W}_{(l)}^T + \mathbf{b}_{(l)}^T \right), \end{aligned} \tag{2}$$

where f_l^I and f_l^T are the "ReLU" nonlinear activation functions of layer l in the convolutional autoencoders for visual and text modalities, respectively. $\{\mathbf{W}_{(i)}^I, \mathbf{b}_{(i)}^I\}_{l=1}^L$ and $\{\mathbf{W}_{(i)}^T, \mathbf{b}_{(i)}^T\}_{l=1}^L$ are the corresponding parameters. $*$ denotes the 2D convolutional operation. The first layer takes visual and textual features as input.

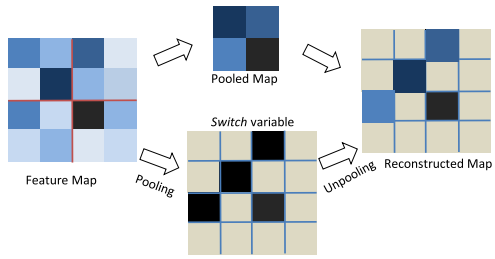


Fig. 2. Pooling and unpooling with switch variables.

In this work, multiple single convolutional autoencoder layers are stacked into deep neural networks to learn a semantic representation from data. In the stacked convolutional autoencoders, we need the reconstructed output to have the same spatial dimensions and the same number of channels as the input, which could be implemented by deconvolution networks.

There are two types of layers in deconvolution networks, the unpooling layers and deconvolutional layers. As we know, the max-pooling layer is usually incorporated into the convolutional neural network to reduce the feature dimensionality and avoid "overfitting." It filters noises from upper layers, leading to a robust representation. In the proposed convolutional autoencoder, to decode the latent representation, an unpooling layer is used to reverse the pooling operation and reconstruct the original intermediate-level representation. As max pooling is noninvertible, we build the unpooling layer in the same way as Zeiler and Fergus (2014), where the location of maximum is recorded by a switch variable for each pooling region. The variable ensures that each activation could be placed back to its original pooled location during unpooling. The remaining elements are set to zeros. Figure 2 illustrates the pooling and unpooling operations.

However, the output of an unpooling layer is a sparse feature map. To obtain dense features, the deconvolutional layer is performed on the sparse activations obtained from unpooling (Zeiler and Fergus 2014). Deconvolution is the opposite of convolution, which uses nonzero padding convolution operation to associate a single input with multiple outputs (Dumoulin and Visin 2016). The output of the deconvolutional layer is an enlarged and dense feature map. To make its output the same size as its input, the boundary of the enlarged feature map is cropped.

In addition, similar to the convolution network, a hierarchical structure of unpooling/deconvolutional layers is stacked to capture details in a different level. With the unpooling layers and deconvolutional layers, the convolutional autoencoder finally generates a recovering feature map $\hat{\mathbf{x}}$ with the same size as the input. And the reconstruction error is defined as the mean squared error between the stacked convolutional autoencoder input $\mathbf{x}$ and reconstructed feature map $\hat{x}$:

$$\begin{aligned} E^T &= \left\| \hat{\mathbf{x}}^T - \mathbf{x}^T \right\|_2^2, \\ E^I &= \left\| \hat{\mathbf{x}}^I - \mathbf{x}^I \right\|_2^2, \end{aligned} \tag{3}$$

where E^T and E^I are the reconstruction errors of visual and textual samples, respectively. The network configuration will be detailed in Section 4.2.

As aforementioned, the multimodal convolutional autoencoder not only learns a semantic representation from each modality but also exploits the correlations among different modalities. Accordingly, we define the loss function of the multimodal convolutional autoencoder as follows:

$$\Omega = \lambda(E^T + E^I) + (1 - \lambda)C + \frac{\alpha}{2}\Psi, \tag{4}$$

where C is the correlation divergence between the latent visual and textual representation. Many types of metric could be used to measure correlation divergence. For the sake of simplicity, we adopt Euclidean distance. As shown in Figure 1, we denote the latent feature learned by the two stacked convolutional autoencoders by $\mathbf{h}^I$ and $\mathbf{h}^T$, respectively, and the correlation divergence between different modalities could be defined as

$$C(\mathbf{h}^I, \mathbf{h}^T) = \left\| \mathbf{h}^I - \mathbf{h}^T \right\|_2^2. \tag{5}$$

Ψ is a regularization term defined as follows:

$$\Psi = \sum_{l=1}^{L} \left(\left\| \mathbf{W}_{(l)}^T \right\|_F^2 + \left\| \mathbf{W}_{(l)}^I \right\|_F^2 + \left\| \mathbf{b}_{(l)}^T \right\|_2^2 + \left\| \mathbf{b}_{(l)}^I \right\|_2^2 \right). \tag{6}$$

Now the objective function in Equation (4) can be rewritten as

$$\begin{aligned} \Omega = {} & \lambda \left(\left\| \hat{\mathbf{x}}^T - \mathbf{x}^T \right\|_2^2 + \left\| \hat{\mathbf{x}}^I - \mathbf{x}^I \right\|_2^2 \right) + (1-\lambda) \left(\left\| \mathbf{h}^I - \mathbf{h}^T \right\|_2^2 \right) \\ & + \frac{\alpha}{2} \sum_{l=1}^{L} \left(\left\| \mathbf{W}_{(l)}^T \right\|_F^2 + \left\| \mathbf{W}_{(l)}^I \right\|_F^2 + \left\| \mathbf{b}_{(l)}^T \right\|_2^2 + \left\| \mathbf{b}_{(l)}^I \right\|_2^2 \right). \end{aligned} \tag{7}$$

There are three terms in the objective function defined by Equation (7). The first term seeks to minimize the reconstruction error for the given visual and text samples in the deep convolutional autoencoders. The second term aims to minimize the correlation divergence between the latent visual and textual representation. The last term imposes an l_2 regularization constraint to avoid overfitting.

Different from recent multimodal feature learning approaches (Ngiam et al. 2011; Srivastava and Salakhutdinov 2012) that exploit deep neural networks to learn latent representation from handcrafted features, the proposed multimodal convolutional autoencoder learns the latent representation from original data directly.

3.4 Model Optimization

The back-propagation algorithm is used to optimize network parameters. The parameters are updated from top layers down through the entire network based on the gradient of loss function as the following formula:

$$\mathbf{W}_{(l)}^T = \mathbf{W}_{(l)}^T - \eta \frac{\partial \Omega}{\partial \mathbf{W}_{(l)}^T}, \tag{8}$$

$$\mathbf{W}_{(l)}^I = \mathbf{W}_{(l)}^I - \eta \frac{\partial \Omega}{\partial \mathbf{W}_{(l)}^I}, \tag{9}$$

$$\mathbf{b}_{(l)}^I = \mathbf{b}_{(l)}^I - \eta \frac{\partial \Omega}{\partial \mathbf{b}_{(l)}^I}, \tag{10}$$

$$\mathbf{b}_{(l)}^T = \mathbf{b}_{(l)}^T - \eta \frac{\partial \Omega}{\partial \mathbf{b}_{(l)}^T}, \tag{11}$$

where η denotes the learning rate. The gradient of Ω with respect to each parameter can be derived as follows:

$$\begin{aligned}\frac{\partial \Omega}{\partial \mathbf{W}_{(l)}^T} &= \lambda\left(\mathbf{x}^T * \frac{\partial h^T}{\mathbf{W}_{(l)}^T} + h * \frac{\partial \hat{\mathbf{x}}}{\mathbf{W}_{(l)}^T}\right) \\ &\quad + 2(1-\lambda)(h^T - h^I)\frac{\partial h^T}{\mathbf{W}_{(l)}^T} + \alpha \mathbf{W}_{(l)}^T,\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{\partial \Omega}{\partial \mathbf{W}_{(l)}^I} &= \lambda\left(\mathbf{x}^I * \frac{\partial h}{\mathbf{W}_{(l)}^I} + h * \frac{\partial \hat{\mathbf{x}}}{\mathbf{W}_{(l)}^I}\right) \\ &\quad + 2(1-\lambda)(h^I - h^T)\frac{\partial h^I}{\mathbf{W}_{(l)}^I} + \alpha \mathbf{W}_{(l)}^I,\end{aligned} \tag{13}$$

$$\begin{aligned}\frac{\partial \Omega}{\partial \mathbf{b}_{(l)}^T} &= 2\lambda(\hat{\mathbf{x}}^T - \mathbf{x}^T)\frac{\partial \hat{\mathbf{x}}^T}{\mathbf{b}_{(l)}^T} \\ &\quad + 2(1-\lambda)(h^T - h^I)\frac{\partial h^T}{\mathbf{b}_{(l)}^T} + \alpha \mathbf{b}_{(l)}^T,\end{aligned} \tag{14}$$

$$\begin{aligned}\frac{\partial \Omega}{\partial \mathbf{b}_{(l)}^I} &= 2\lambda(\hat{\mathbf{x}}^I - \mathbf{x}^I)\frac{\partial \hat{\mathbf{x}}^I}{\mathbf{b}_{(l)}^I} \\ &\quad + 2(1-\lambda)(h^T - h^I)\frac{\partial h^T}{\mathbf{b}_{(l)}^I} + \alpha \mathbf{b}_{(l)}^I.\end{aligned} \tag{15}$$

Based on the above formula, the network parameters could be optimized by the stochastic gradient descent algorithm.

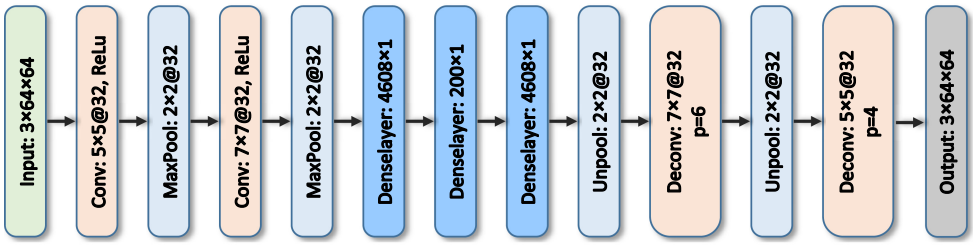
4 IMPLEMENTATION

4.1 Visual and Textual Samples

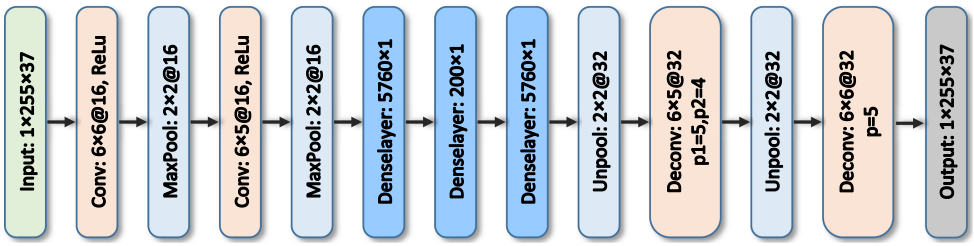
The proposed MUCAE performs basic convolution operations on samples that are required in the form of a matrix. Here, we scale images into the same size and use their pixel matrices as the input to MUCAE. In conventional text processing methods, a text sample is usually represented as a vector of words with a fixed length by BoW or n-gram. With the rapid development of deep learning, methods of understanding text by deep learning techniques have gradually attracted attention in the academic community. Most of these techniques use word2vec (Mikolov et al. 2013) or similar techniques to produce word embeddings. Specifically, these methods employ two-layer neural networks to project words into an embedding vector space such that words sharing common contexts in the corpus are closed to one another in the embedding space (dos Santos and Gatti 2014; Kim 2014). More recently, inspired by the success of ConvNet on computer vision, there is also a new trend in formulating text at the character level with the convolutional neural network (Kim et al. 2016; Zhang and LeCun 2015). These works show that the character-level convolutional neural network is an effective method for text modeling without the need for words. We think this could also be explained by the recent discovery about the cognitive process when reading written text, which demonstrates that readers can understand the meaning of words in a sentence even when the interior letters of each word are interchanged (Rawlinson 1976). And the convolutional network could capture the local structure in a word. Similar to Zhang and LeCun (2015), MUCAE takes a sequence of encoded characters as input. The encoding is done by projecting the input

	a	d	d	e	b	d	a
a	1	0	0	0	0	0	1
b	0	0	0	0	1	0	0
c	0	0	0	0	0	0	0
d	0	1	1	0	0	1	0
e	0	0	0	1	0	0	0

Fig. 3. Illustration of the text quantization method, with text "addebda" and dictionary "abcde."



(a) The convolutional autoencoder for visual content



(b) The convolutional autoencoder for textual content

Fig. 4. The convolutional autoencoder structure. There are 11 layers in the neural networks, which include convolutional, maxpool, dense, deconvolutional, and unpooling layers.

sample with a predefined character dictionary of size m. Each character is quantized using 1-of-m encoding. Then, the sequence of characters is transformed to a sequence of such m-sized vectors with a fixed length l. Any character exceeding length l is ignored, and any characters that are not in the dictionary are quantized as all-zero vectors. The vectors in the sequence are then concatenated and a text sequence is finally encoded as an $m \times l$ matrix, as illustrated in Figure 3. All text samples are converted into lowercase. The dictionary used in the experiments is of size 37 including 26 English characters *a-z*, 10 Arabic numerals *0-9*, and the blank character.

4.2 Deep Network Architecture

The multimodal convolutional autoencoders are stacked into a deep network to learn more representative features from textual and visual content. The layer parameters for each modality are detailed in Figure 4. In particular, there are 11 layers in the model to learn the latent representation. First, the image and text pathways follow the same architecture in the network, which consists of two groups of convolutional components. In each component, the convolutional layer with the ReLU activation function is followed by the maxpool layer. Following, in the middle of the networks, the outputs are used to feed the dense layer jointly, leading to a final 200D latent presentation. Finally, the deconvolutional part with deconvolutional layers and unpooling layers

inverses the function of convolutional/pool layers, as described in Section 3.3. The convolutional network used in our method is based on LeNet (LeCun et al. 1998), which has been popularly used in many recognition tasks. We use this lightweight network to show the possibility of designing an end-to-end learning framework using deep neural networks for cross-modality feature learning from scratch.

5 EXPERIMENTS

In this section, we evaluate the proposed MUCAE model on two important multimedia tasks, i.e., cross-media retrieval and unimodal object classification. We compare MUCAE to the following multimodal feature learning methods:

- **CCA** (Rasiwasia et al. 2010). It learns a common feature subspace shared by two modalities, leading to joint representation in the subspace.
- **CCA-AE** (Kim et al. 2012). It first learns two unimodal autoencoders to generate higher-level visual and textual features, respectively, based on which CCA is then used to learn joint visual-textual representation.
- **Corr-AE** (Feng et al. 2015). Similar to **CCA-AE**, **Corr-AE** also trains two unimodal autoencoders to learn latent visual and textual features and maximizes the correspondence between the two modalities jointly. Compared with the proposed method, the dense layers rather than the convolutional layers are used in Corr-AE and cannot learn features from raw input directly.
- **MSAE** (Wang et al. 2016b). MSAE utilizes the stacked autoencoder framework to learn a joint latent semantic space from visual and textual content. In the model, a stacked autoencoder is trained for each modality, and then the two modalities are joined at the top layer. We use the output of each single SAE as the latent features in the following evaluation.
- **MGCMH** (Xie et al. 2016). MGCMH is an unsupervised method that integrates multigraph learning to learn hash functions for cross-media retrieval. In the implementation, we use text and visual features as Xie et al. (2016) did to train the model, and the size of the embedding code is set to 200.

To train the CCA-AE, Corr-AE, MSAE, and MUCAE models, we use the Glorot initialization with uniform distribution, leading to the same gradient variance in each layer. We update model parameters using 10,000 iterations of stochastic gradient descent with a momentum of 0.9, weight decay of 0.0005, base learning rate of 0.005, and batch size of 256.

5.1 Datasets

We use two real-world image datasets, i.e., MIRFlickr and NUS-WIDE, to evaluate the above methods. It is worth noting that both the datasets were collected from the Internet and widely used in the literature (Feng et al. 2015; Jiang and Li 2016). The MIRFlickr dataset (Huiskes and Lew 2008) consists of 1 million images retrieved from the photo-sharing website Flickr. User-provided tags associated with the images are also included in the dataset. Twenty-five thousand images were labeled with 24 topics such as bird, sky, night, and people. We use the remaining 975,000 unlabeled images as training data and the labeled images for testing. Moreover, the MIRFlickr dataset provides a set of thumbnails of size 64×64 pixels, which are used instead of the original images in order to save computation cost in the experiments. The NUS-WIDE (Chua et al. 2009) dataset contains 269,648 images with their associated tags collected from Flickr. The images were annotated with 81 concepts. After removing the images without any tags, we have 267,465 images left. In the experiments, all the images are rescaled to 64×64 pixels. We use a subset of NUS-WIDE, called NUS-WIDE-OBJECT, for testing (Gao et al. 2011; Li et al. 2012), and the rest of the data as training

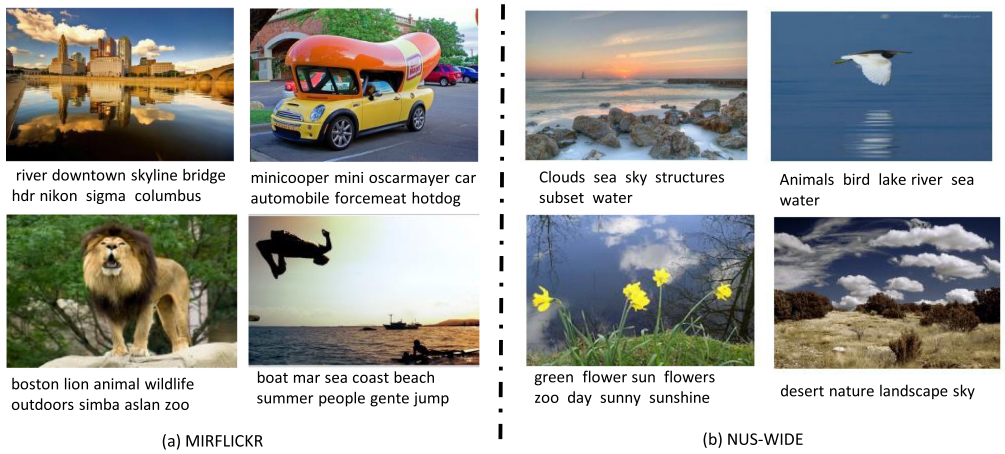


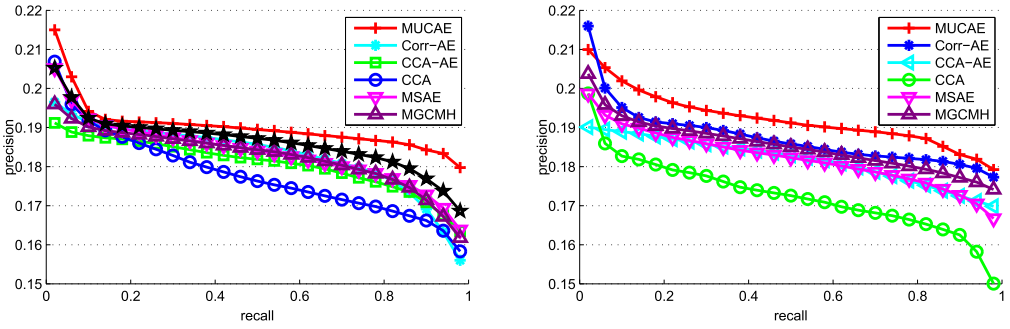
Fig. 5. The data samples of MIRFlickr and NUS-WIDE datasets. Both of the datasets are collected from Flickr. Besides the visual content, each image is associated with tags, which are shared by the uploaders.

data. The subset contains 30,000 images of 31 classes of objects, such as "bear," "boat," and "computer." In the two datasets, the tags of each sample are concatenated as the textual description. Some samples of the two datasets are shown in Figure 5. For MUCAE, all the images and tags are quantized as matrices. For CCA, CCA-AE, MGCMH, and Corr-AE, the conventional 1,024D bag of visual words and 500D tf-idf features are extracted from images and tags, respectively, and used as the input.

5.2 Cross-Media Retrieval

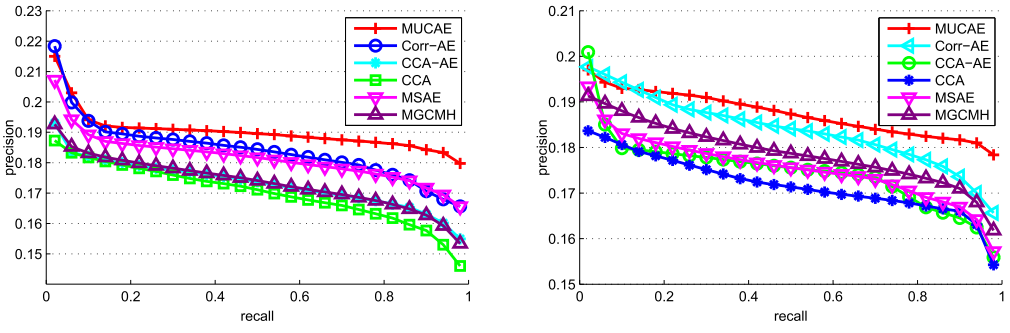
Cross-media retrieval aims at retrieving content of one media type given a query of another media type (Wu et al. 2015; Zheng et al. 2014). The key is to exploit the correlation between the modalities corresponding to the two media types. The retrieval is performed based on the normalized correlation between the representations of different modalities. Here, we evaluate two retrieval tasks, i.e., using text to search image (T2I) and using image to search text (I2T). We repeat the evaluation for each task 10 times. At each time, we randomly select 500 text or image queries, and the remaining data are used as the retrieval database. The average performance over the queries in terms of precision-recall (PR) curves and mean average precision (MAP) are reported. The PR curves of both of the two cross-media retrieval tasks on the two datasets are illustrated in Figure 6 and Figure 7, respectively, while the MAP scores are reported in Table 2. From the results, we can obtain the following observations:

(1) The proposed MUCAE approach obtains the best performance in terms of both MAP and PR curves as compared to other methods.
(2) MUCAE has consistent and robust performance improvements on both the T2I and I2T retrieval tasks over the two datasets.
(3) MUCAE obtains about 2.45%, 5.67%, 11.30%, 5.49%, and 3.64% relative improvements on MAP compared to other methods, respectively, for T2I on the MIRFlircker dataset. The performance improvements for I2T are 4.59%, 5.87%, 7.53%, 4.19%, and 4.75%, respectively. On the NUS-WIDE dataset, MUCAE outperforms Corr-AE, CCA-AE, CCA, MSAE, and MGCMH by 1.93%, 7.23%, 9.04%, 6.84%, and 4.80% MAP for the T2I task and 3.48%, 9.93%, 11.93%, 4.85%, and 9.92% MAP for the I2T task, respectively.



(a) Using text to search image (T2I) (b) Using image to search text (I2T)

Fig. 6. The Precision recall curves of cross-modality query on MIRFlickr dataset.



(a) Using text to search image (T2I) (b) Using image to search text (I2T)

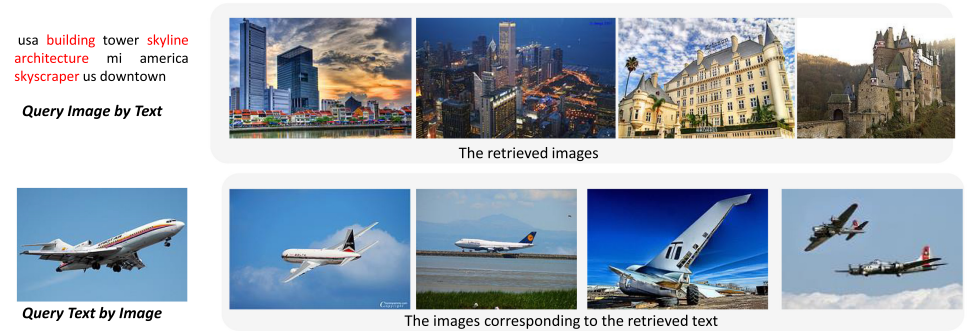
Fig. 7. The Precision Recall curves of cross-modality query on NUS-OBJECT dataset.

Table 2. MAP of Various Methods on the Two Retrieval Tasks

	T2I		I2T	
	MIRFLICKR	NUSWIDE	MIRFLICKR	NUSWIDE
MUCAE	0.1913	0.1904	0.1921	0.1874
Corr-AE	0.1829	0.1840	0.1875	0.1838
CCA-AE	0.1807	0.1732	0.1818	0.1747
CCA	0.1779	0.1701	0.1726	0.1718
MSAE	0.1836	0.1816	0.1821	0.1754
MGCMH	0.1827	0.1735	0.1854	0.1788

(4) As shown in the experimental results, the proposed MUCAE achieves much better performance than conventional graph- and autoencoder-based methods. It is because the deep features extracted from the character and pixel level directly by MUCAE are more representative than handcrafted features used in these methods. In addition, MUCAE also outperforms CCA and its variants. The reason could be that the nonlinear projection used in MUCAE is able to better model the semantic correlation than the linear function in CCA-based methods.

(a) MIRFLICKR



(b) MIRFLICKR

Fig. 8. Examples of querying image by text and querying text by image over MIRFLICKR and NUS-WIDE datasets. For querying images by text task, the corresponding images of retrieved texts are provided to demonstrate the results.

Some examples of querying image by text and querying text by image over the NUS-WIDE and MIRFLICKR datasets are illustrated in Figure 8.

5.3 Unimodal Object Classification

Besides cross-media retrieval, we also evaluate the efficacy of the features of each modality learned by MUCAE. We evaluate the textual and visual features individually with object classification tasks on the NUS-WIDE-OBJECT dataset. Specifically, we use the textual and visual features learned by CCA, CCA-AE, Corr-AE, and MUCAE to train a group of SVM classifiers for classification. For each category, we use 80% of data for classifier training and the remaining 20% for testing. The experimental performance of the textual and visual modalities in terms of accuracy is shown in Tables 3 and 4, respectively. From the results, we can obtain the following observations:

(1) MUCAE obtains the best overall performance in terms of average accuracy compared to other methods.
(2) MUCAE performs the best on "zebra" and "fish" out of all 31 categories on both textual and visual classification tasks. Some of the improvements are significant. For example, in the visual classification task, in comparison with Corr-AE, MUCAE achieves 4.92%, 4.96%, and 8.74% improvement on "coral," "plane," and "zebra" categories, respectively, in terms

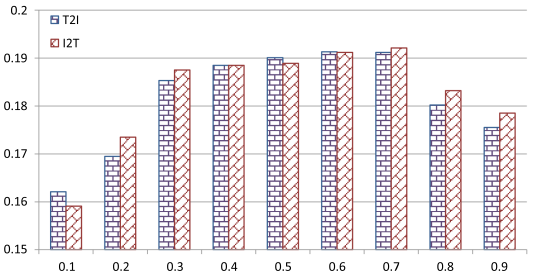


Fig. 9. The retrieval performance in terms of MAP with respect to different λ, which is used to balance the reconstruction errors of two modalities and the correlation divergence between the latent representations from the modalities.

of accuracy. When digging the dataset, it could be found that these categories are with rich with related textual descriptions. In the text classification task, the improvement on "boats," "coral," "train," and "plane" is 4.58%, 4.38%, 3.62%, and 2.82%, respectively, compared with Corr-AE. In the dataset, it could be found that the images of these categories are more discriminant semantically. And this observation verifies that MUCAE could exploit the semantics from external content with high quality to learn representative features.

(3) MUCAE deteriorates on some categories, such as "books," "fox," and "cars" in the visual classification task and "sand," "sun," and "leaf" in the text classification task. In these categories, Corr-AE also achieves worse results than CCA and CCA-AE. The reason may be that both methods learn the hidden representation with external content. For example, in the visual classification task, the feature is learned from not only visual content but also text content. But for these categories, too much external content is irrelevant to the topic and brings noise and hence deteriorates the final classification result.

5.4 Parameter Analysis

MUCAE has several parameters, i.e., λ, α, and size of h in the loss function in Equation (7). α is used to weight the regularization term and is set as 1e-4 empirically in the experiments. Here, we investigate the effects of λ and the size of h and use the cross-media retrieval as the experimental task.

5.4.1 Evaluation of Parameter λ. The parameter λ in Equation (4) is used to balance the reconstruction errors of two modalities and the correlation divergence between the latent representations from the modalities. When λ is 0, MUCAE only optimizes the correlation between modalities but does not persist with the information in original features, while when λ is 1, it degenerates to reconstruct the image and text samples separately but cannot bridge the two modalities. Here, we investigate the retrieval performance with respect to different values of λ from 0.1 to 0.9 on the MIRFlickr dataset. Figure 9 illustrates the experimental results, from which we can see that the retrieval performance has a "$\wedge$" distribution of λ on both T2I and I2T retrieval tasks. That is, the performance first increases with λ until a peak, then decreases when λ increases. The best performance is obtained when λ is 0.6 and 0.7 on the T2I and I2T retrieval tasks, respectively.

5.4.2 Evaluation of the Dimensionality of $\mathbf{h}$. MUCAE encodes features as the hidden representation $\mathbf{h}$. The dimensionality of $\mathbf{h}$, i.e., $|\mathbf{h}|$, impacts the amount of information loss in feature encoding. A low-dimensional $\mathbf{h}$ leads to significant information loss. Thus, the hidden representation preserves a limited amount of information from the original data. A high-dimensional $\mathbf{h}$ decreases information loss but increases computational cost in subsequent processing. We investigate $|\mathbf{h}|$

Table 3. Accuracy of Classification with Textual Modality on NUSWIDE-OBJECT Dataset

	bear	birds	boats	book	cars	cat	comput.	coral	cow	dog	elk	fish	flags	flower	fox	horse	leaf	plane	rocks	sand	sign	statue	sun	tiger	tower	toy	train	tree	vehicle	whale	zebra	Avg
CCA	62.10	60.76	65.50	55.07	64.42	60.05	64.34	70.63	75.69	60.16	71.59	71.98	74.68	59.34	72.92	67.82	59.20	63.62	61.20	62.21	61.64	63.29	60.96	64.02	62.50	60.91	55.38	59.08	59.97	78.79	80.56	64.85
CCA-AE	65.73	61.18	60.54	60.87	60.74	66.28	65.89	58.74	65.97	59.76	69.89	70.15	59.49	59.76	62.50	70.12	60.72	61.70	59.72	59.71	56.39	58.94	61.72	60.37	58.78	57.51	56.99	60.17	59.57	74.75	81.94	62.79
Corr-AE	66.90	67.64	66.36	53.77	58.71	66.44	65.19	71.82	69.86	69.68	80.91	79.19	71.01	68.14	65.69	73.05	58.76	71.70	58.23	57.02	58.85	58.60	58.36	74.15	63.82	60.44	66.88	57.73	63.57	79.90	70.56	66.22
MUCAE	68.32	66.79	72.43	53.77	59.33	60.44	62.09	78.11	77.50	65.90	74.09	80.48	65.29	70.12	70.56	75.92	56.60	75.75	59.79	54.51	59.84	62.46	55.53	74.76	63.48	57.04	71.72	58.82	66.04	78.89	78.89	66.94
MSAE	67.04	62.40	61.75	62.09	61.95	67.61	67.21	59.92	67.29	60.96	71.28	71.55	60.68	60.96	63.75	71.52	61.94	62.94	60.91	60.90	57.52	60.12	62.96	61.57	59.96	58.66	58.13	61.37	60.77	76.24	83.58	64.05
MGCMH	66.31	64.41	63.45	57.32	59.72	66.36	65.54	65.28	67.92	64.72	75.40	74.67	65.25	63.95	64.10	71.58	59.74	66.70	58.97	58.36	57.62	58.77	60.04	67.26	61.30	58.97	61.94	58.95	61.57	77.32	76.25	64.51
dCorr-AE	67.61	67.22	69.39	53.77	59.02	63.44	63.64	74.97	73.68	67.79	77.50	79.84	68.15	69.13	68.13	74.48	57.68	73.72	59.01	55.76	59.34	60.53	56.94	74.45	63.65	58.74	69.30	58.28	64.80	79.39	74.72	66.58

We can find that MUCAE obtains the best average accuracy compared to other methods and has good performance on the "boats," "coral," "train," and "plane" concepts.

Table 4. Accuracy of Classification with Visual Modality on NUSWIDE-OBJECT Dataset

	bear	birds	boats	book	cars	cat	comput.	coral	cow	dog	elk	fish	flags	flower	fox	horse	leaf	plane	rocks	sand	sign	statue	sun	tiger	tower	toy	train	tree	vehicle	whale	zebra	Avg
CCA	55.24	59.07	59.27	62.32	60.12	59.82	62.02	58.74	65.28	59.96	61.36	60.99	62.03	60.48	62.50	60.06	57.69	61.70	58.39	58.66	58.36	60.39	63.50	56.71	60.30	62.04	64.52	60.25	59.46	65.66	58.33	60.49
CCA-AE	60.08	56.26	61.66	65.22	52.76	61.20	58.14	57.34	59.72	61.75	61.36	63.92	59.49	59.16	57.64	62.93	62.24	62.13	60.73	63.88	59.34	58.94	60.33	63.42	59.80	63.17	57.53	60.02	58.34	54.55	72.22	60.49
Corr-AE	62.90	58.65	60.54	59.42	60.12	63.97	60.47	60.84	58.33	61.55	67.05	60.26	62.03	62.28	53.47	56.61	59.20	63.19	60.73	60.13	60.00	61.84	59.06	53.66	58.78	60.58	63.44	59.47	58.68	63.64	65.28	60.52
MUCAE	64.92	60.06	58.95	53.62	57.06	60.74	61.24	67.13	62.50	61.36	67.61	63.74	62.03	63.71	58.33	59.77	63.38	69.79	60.97	61.17	60.33	60.87	60.96	63.42	61.49	61.23	62.37	60.25	62.77	68.69	77.78	62.52
MSAE	61.28	57.38	62.89	66.52	53.82	62.43	59.30	58.49	60.92	62.99	62.59	65.20	60.68	60.35	58.79	64.19	63.48	63.37	61.95	65.16	60.53	60.12	61.54	64.68	60.99	64.43	58.68	61.22	59.51	55.64	73.67	61.70
MGCMH	61.49	57.45	61.10	62.32	56.44	62.59	59.30	59.09	59.03	61.65	64.20	62.09	60.76	60.72	55.56	59.77	60.72	62.66	60.73	62.00	59.67	60.39	59.70	58.54	59.29	61.87	60.48	59.74	58.51	59.09	68.75	60.51
dCorr-AE	63.91	59.35	59.74	56.52	58.59	62.36	60.85	63.99	60.42	61.45	67.33	62.00	62.03	62.99	55.90	58.19	61.29	66.49	60.85	60.65	60.16	61.35	60.01	58.54	60.14	60.91	62.90	59.86	60.72	66.16	71.53	61.52

We can find that MUCAE obtains the best average accuracy compared to other methods and has good performance on the "zebra," "fish," "coral," "plane," and "zebra" concepts.

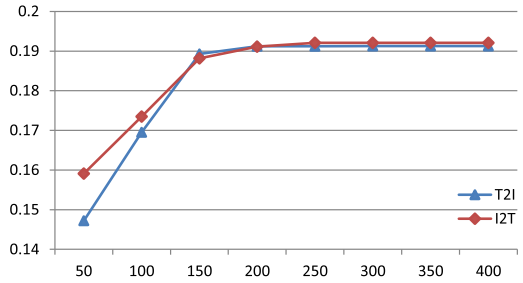


Fig. 10. The retrieval performance in terms of MAP with respect to the dimensionality of $\mathbf{h}$.

with a range from 50 to 400 with an interval of 50. The evaluation results on the retrieval task are reported in Figure 10. It can be observed that the retrieval performance first increases when increasing $|\mathbf{h}|$ and then becomes stable on both T2I and I2T retrieval tasks. Specifically, the MAP scores of T2I and I2T retrieval are 0.147 and 0.159, respectively, when $|\mathbf{h}|$ is 50. The scores increase to 0.189 and 0.188 when $|\mathbf{h}|$ is 150 and become stable until $|\mathbf{h}|$ is 200. Hence, the optimal value for $|\mathbf{h}|$ is 200 in the experiments.

6 CONCLUSION

In this article, we have proposed a novel cross-modal feature learning approach that learns an effective joint representation from multimodal data. In particular, the image and text modalities are modeled by the convolutional neural network to learn representative features from original image pixels and text characters directly, while the convolutional neural network is stacked into an autoencoder framework. We also design a novel objective function to exploit the correlation between the hidden representations from each modality, so that both the representation learning error of each modality and the correlation divergence between different modalities are minimized jointly. We conducted extensive experiments to evaluate the approach on the tasks of cross-modal retrieval and unimodal classification on two real-world large-scale datasets consisting of both image and text data. Experimental results have shown the effectiveness of the proposed approach and its advantages over the state-of-the-art methods.

REFERENCES

Guillaume Alain and Yoshua Bengio. 2014. What regularized auto-encoders learn from the data-generating distribution. *Journal of Machine Learning Research* 15, 1 (Jan. 2014), 3563–3593.

Artem Babenko, Anton Slesarev, Alexandr Chigorin, and Victor Lempitsky. 2014. Neural codes for image retrieval. In *Proceedings of European Conference on Computer Vision*. Vol. 8689. Springer International Publishing, 584–599.

Yoshua Bengio, Aaron Courville, and Pierre Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 8 (2013), 1798–1828.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, Hugo Larochelle. 2007. Greedy layer-wise training of deep networks. In *Proceedings of Advances in Neural Information Processing Systems*, Vol. 19. 153.

Minmin Chen, Kilian Q. Weinberger, Fei Sha, and Yoshua Bengio. 2014. Marginalized denoising auto-encoders for nonlinear representations. In *Proceedings of International Conference on Machine Learning*. 1476–1484.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao. Zheng. 2009. NUS-WIDE: A real-world web image database from national university of Singapore. In *Proceedings of ACM Conference on Image and Video Retrieval*.

Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert Lanckriet, Roger Levy, and Nuno Vasconcelos. 2014. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36, 3 (March 2014), 521–535.

Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. 2013. Decaf: A deep convolutional activation feature for generic visual recognition. *arXiv preprint arXiv:1310.1531* (2013).

Cicero dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of the International Conference on Computational Linguistics: Technical Papers*. 69–78.

Vincent Dumoulin and Francesco Visin. 2016. A guide to convolution arithmetic for deep learning. *arXiv preprint arXiv:1603.07285* (2016).

Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollar, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*.

Fangxiang Feng, Xiaojie Wang, Ruifan Li, and Ibrar Ahmad. 2015. Correspondence autoencoders for cross-modal retrieval. *ACM Transactions on Multimedia Computing, Communications, and Applications* 12, 1s, Article 26 (Oct. 2015), 22 pages.

Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. 2011. Multi-layer group sparse coding for concurrent image classification and annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2809–2816.

Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. 2010. Multimodal semi-supervised learning for image classification. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 902–909.

Junwei Han, Dingwen Zhang, Shifeng Wen, Lei Guo, Tianming Liu, and Xuelong Li. 2016. Two-stage learning to predict human eye fixations via SDAEs. *IEEE Transactions on Cybernetics* 46, 2 (Feb. 2016), 487–498.

David R. Hardoon, Sandor R. Szedmak, and John R. Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Computing* 16, 12 (Dec. 2004), 2639–2664.

Richang Hong, Lei Li, Junjie Cai, Dapeng Tao, Meng Wang, and Qi Tian. 2017. Coherent semantic-visual indexing for large-scale image retrieval in the cloud. *IEEE Transactions on Image Processing* 26, 9 (2017), 4128–4138.

Richang Hong, Meng Wang, Yue Gao, Dacheng Tao, Xuelong Li, and Xindong Wu. 2014. Image annotation by multiple-instance learning with discriminative feature mapping and selection. *IEEE Transactions on Cybernetics* 44, 5 (2014), 669–680.

Mark J. Huiskes and Michael S. Lew. 2008. The MIR Flickr retrieval evaluation. In *Proceedings of ACM Conference on Multimedia Information Retrieval*. ACM, New York.

Muhammad Imran, Carlos Castillo, Fernando Diaz, and Sarah Vieweg. 2015. Processing social media messages in mass emergency: A survey. *Computing Surveys* 47, 4, Article 67 (June 2015), 67:1–67:38 pages.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (Jan. 2013), 221–231. DOI: https://doi.org/10.1109/TPAMI.2012.59

Qing-Yuan Jiang and Wu-Jun Li. 2016. Deep cross-modal hashing. *arXiv preprint arXiv:1602.02255* (2016).

Jungi Kim, Jinseok Nam, and Iryna Gurevych. 2012. Learning semantics with deep belief network for cross-language information retrieval. In *Proceedings of International Conference on Computational Linguistics*. 579–588.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. 1746–1751.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI'16)*. 2741–2749.

Alex Krizhevsky and Geoffrey E. Hinton. 2011. Using very deep autoencoders for content-based image retrieval. In *Proceedings of European Symposium on Artificial Neural Networks*.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of Advances in Neural Information Processing Systems*. 1097–1105.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 512 (2015), 436–444.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86, 11 (1998), 2278–2324.

Guy Lev, Gil Sadeh, Benjamin Klein, and Lior Wolf. 2016. RNN Fisher vectors for action recognition and image annotation. In *European Conference on Computer Vision*. Springer, 833–850.

Michael S. Lew, Nicu Sebe, Chabane Djeraba, and Ramesh Jain. 2006. Content-based multimedia information retrieval. *ACM Transactions on Multimedia Computing Communications and Applications* 2, 1 (2006), 1–19.

Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K. Sethi. 2003. Multimedia content processing through cross-modal association. In *Proceedings of ACM Conference on Multimedia*. 604–611.

Hong Li, Yantao Wei, Luoqing Li, and C. L. Philip Chen. 2013. Hierarchical feature extraction with local neural response for image recognition. *IEEE Transactions on Cybernetics* 43, 2 (April 2013), 412–424. DOI: https://doi.org/10.1109/TSMCB.2012.2208743

Yangxi Li, Bo Geng, Dacheng Tao, Zheng-Jun Zha, Linjun Yang, and Chao Xu. 2012. Difficulty guided image retrieval using linear multiple feature embedding. *IEEE Transactions on Multimedia* 14, 6 (2012), 1618–1630.

Xianglong Liu, Yadong Mu, Danchen Zhang, Bo Lang, and Xuelong Li. 2015. Large-scale unsupervised hashing with shared structure learning. *IEEE Transactions on Cybernetics* 45, 9 (Sept. 2015), 1811–1822. DOI: https://doi.org/10.1109/TCYB.2014.2360856

Vijay Mahadevan, Chi Wah Wong, Jose Costa Pereira, Thomas T. Liu, Nuno Vasconcelos, and Lawrence K. Saul. 2011. Maximum covariance unfolding: Manifold learning for bimodal data. In *Proceedings of Advances in Neural Information Processing Systems*. 918–926.

Tao Mei, Yong Rui, Shipeng Li, and Qi Tian. 2014. Multimedia search reranking: A literature survey. *Computing Surveys* 46, 3 (2014), 57–76.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In *Proceedings of International Conference on Machine Learning*.

Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. 1717–1724. DOI: https://doi.org/10.1109/CVPR.2014.222

Xueming Qian, Xian-Sheng Hua, Yuan Yan Tang, and Tao Mei. 2014. Social image tagging with diverse semantics. *IEEE Transactions on Cybernetics* 44, 12 (Dec. 2014), 2493–2508.

Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of ACM Conference on Multimedia*. 251–260.

Graham Rawlinson. 1976. *The Significance of Letter Position in Word Recognition*. Ph.D. Dissertation. University of Nottingham.

Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61 (2015), 85–117.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN features off-the-shelf: An astounding baseline for recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Nitish Srivastava and Ruslan R. Salakhutdinov. 2012. Multimodal learning with deep Boltzmann machines. In *Proceedings of Advances in Neural Information Processing Systems*. 2222–2230.

Chun Chet Tan. 2008. *Autoencoder Neural Networks: A Performance Study Based on Image Recognition, Reconstruction and Compression*. Ph.D. Dissertation. Multimedia University.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016a. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5005–5013.

Meng Wang, Xian-Sheng Hua, Richang Hong, Jinhui Tang, Go-Jun Qi, and Yan Song. 2009. Unified video annotation via multigraph learning. *IEEE Transactions on Circuits and Systems for Video Technology* 19, 5 (2009), 733–746.

Meng Wang, Weisheng Li, Dong Liu, Bingbing Ni, Jialie Shen, and Shuicheng Yan. 2015a. Facilitating image search with a scalable and compact semantic mapping. *IEEE Transactions on Cybernetics* 45, 8 (2015), 1561–1574.

Meng Wang, Xueliang Liu, and Xindong Wu. 2015b. Visual classification by l_1-hypergraph modeling. *IEEE Transactions on Knowledge and Data Engineering* 27, 9 (2015), 2564–2574.

Wei Wang, Xiaoyan Yang, Beng Chin Ooi, Dongxiang Zhang, and Yueting Zhuang. 2016b. Effective deep learning-based multi-modal retrieval. *VLDB Journal* 25, 1 (2016), 79–101.

Fei Wu, Xinyang Jiang, Xi Li, Siliang Tang, Weiming Lu, Zhongfei Zhang, and Yueting Zhuang. 2015. Cross-modal learning to rank via latent joint representation. *IEEE Transactions on Image Processing* 24, 5 (2015), 1497–1509.

Liang Xie, Lei Zhu, and Guoqi Chen. 2016. Unsupervised multi-graph cross-modal hashing for large-scale multimedia retrieval. *Multimedia Tools and Applications* 75, 15 (2016), 9185–9204.

Hao Xue, Like Xue, and Feng Su. 2015. Multimodal music mood classification by fusion of audio and lyrics. In *Proceedings of International Conference on MultiMedia Modeling*. Springer, 26–37.

Yan Yan, Feiping Nie, Wen Li, Chenqiang Gao, Yi Yang, and Dong Xu. 2016. Image classification by cross-media active learning with privileged information. *IEEE Transactions on Multimedia* 18, 12 (2016), 2494–2502.

Yi Yang, Dong Xu, Feiping Nie, Jiebo Luo, and Yueting Zhuang. 2009. Ranking with local regression and global alignment for cross media retrieval. In *Proceedings of ACM Conference on Multimedia*. 175–184.

Kai Yu, Tong Zhang, and Yihong Gong. 2009. Nonlinear learning using local coordinate coding. In *Proceedings of Advances in Neural Information Processing Systems*, Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta (Eds.). 2223–2231.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Proceedings of European Conference on Computer Vision*. Springer, 818–833.

Hanwang Zhang, Xindi Shang, Huanbo Luan, Meng Wang, and Tat-Seng Chua. 2016. Learning from collective intelligence: Feature learning using social images and tags. *ACM Transactions on Multimedia Computing, Communications and Applications* 13, 1, Article 1 (Nov. 2016), 23 pages.

Luming Zhang, Yue Gao, Chaoqun Hong, Yinfu Feng, Jianke Zhu, and Deng Cai. 2014. Feature correlation hypergraph: Exploiting high-order potentials for multimodal recognition. *IEEE Transactions on Cybernetics* 44, 8 (2014), 1408–1419.

Xiang Zhang and Yann LeCun. 2015. Text understanding from scratch. *CoRR abs/1502.01710* (2015).

Fang Zhao, Yongzhen Huang, Liang Wang, Tao Xiang, and Tieniu Tan. 2016. Learning relevance restricted Boltzmann machine for unstructured group activity and event understanding. *International Journal of Computer Vision* 3 (2016), 1–17.

Liang Zheng, Shengjin Wang, and Qi Tian. 2014. Coupled binary embedding for large-scale image retrieval. *IEEE Transactions on Image Processing* 23, 8 (2014), 3368–3380.

Liang Zheng, Yi Yang, and Qi Tian. 2018. SIFT meets CNN: A decade survey of instance retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 5 (2018), 1224–1244.

Lin Zhong, Qingshan Liu, Peng Yang, Junzhou Huang, and Dimitris N. Metaxas. 2015. Learning multiscale active facial patches for expression analysis. *IEEE Transactions on Cybernetics* 45, 8 (Aug. 2015), 1499–1510.

Yue-Ting Zhuang, Yi Yang, and Fei Wu. 2008. Mining semantic correlation of heterogeneous multimedia data for cross-media retrieval. *IEEE Transactions on Multimedia* 10, 2 (Feb. 2008), 221–229.

Received October 2017; revised April 2018; accepted June 2018