MDL-CW: A Multimodal Deep Learning Framework with Cross Weights

Sarah Rastegar, Mahdieh Soleymani Baghshah, Hamid R. Rabiee, Seyed Mohsen Shojaee
AICT Innovation Center, Department of Computer Engineering, Sharif University of Technology
Tehran, Iran
rabiee@sharif.edu

Abstract

Deep learning has received much attention as of the most powerful approaches for multimodal representation learning in recent years. An ideal model for multimodal data can reason about missing modalities using the available ones, and usually provides more information when multiple modalities are being considered. All the previous deep models contain separate modality-specific networks and find a shared representation on top of those networks. Therefore, they only consider high level interactions between modalities to find a joint representation for them. In this paper, we propose a multimodal deep learning framework (MDL-CW) that exploits the cross weights between representation of modalities, and try to gradually learn interactions of the modalities in a deep network manner (from low to high level interactions). Moreover, we theoretically show that considering these interactions provide more intra-modality information, and introduce a multi-stage pre-training method that is based on the properties of multi-modal data. In the proposed framework, as opposed to the existing deep methods for multi-modal data, we try to reconstruct the representation of each modality at a given level, with representation of other modalities in the previous layer. Extensive experimental results show that the proposed model outperforms state-of-the-art information retrieval methods for both image and text queries on the PASCAL-sentence and SUN-Attribute databases.

1. Introduction

In real-world applications, we usually encounter data consisting of different modalities. Images annotated with tags and videos containing audio signal along with visual signal are examples of multimodal data. Although each modality has its own information and statistical properties, different modalities usually share high level concepts. The rationale for using different modalities to learn a shared representation is that those modalities may provide complementary information about a common concept. Moreover, if we could model and learn the cross-modal relations, we can reason about missing modalities using the available modalities or even improve the unimodal models. In recent years, there has been a growing interest in using deep networks for multi-modal learning. Ngiam *et al.* [13] used deep auto-encoder to extract high level features from speech and video signals, and then aggregated these features to find a shared representation. Srivastava *et al.* [20] introduced the multimodal deep Boltzman machine that learns a deep generative model over joint space of image and text inputs. Recently, Sohn *et al.* [19] proposed a training approach for multimodal deep learning that is based on minimizing variation of information instead of maximizing the likelihood. All of these methods for multimodal data have a common strategy in which they first learn layers of modality-specific representations and then learn a shared representation across multiple modalities at the top layer of the deep network.

Besides deep multimodal learning approaches, some probabilistic methods have also been recently introduced for multimodal data. Xing *et al.* [24] proposed dual-wing harmoniums to learn a joint representation of the image and text modalities. Zhen *et al.* [26] introduced a probabilistic generative approach called multimodal latent binary embedding. Ozdemir *et al.* [14] presented a model based on Bayesian nonparametric framework to learn the underlying semantically meaningful and abstract features of multimodal data. However, since different data modalities have different statistical properties, shallow models are not usually able to extract high-level concepts from multimodal data.

One of the other recent challenges about multimodal datasets is their huge size. In order to achieve real time retrieval, some approaches such as multimodal hashing has been introduced which encodes the high-dimensional input vectors into compact binary strings while trying to increase similarity between different modalities that having the same concept in the resulted space. For example, Bronstein *et al.* [2] applied a boosting procedure for cross-modality similarity learning. Zhen *et al.* [25] proposed a co-regularized hashing based on a boosted co-regularization framework.

Rastegari *et al.* [16] presented a predictable dual-view hashing. The main problem of those methods is that they are all discriminative and therefore unable to use large amount of unlabeled data as opposed to the deep learning methods that can use these unlabeled data to learn a better representation. As mentioned above, in all of the previous models, a joint representation for both modalities is considered. This high level joint representation only shows a common representation for both modalities which has information about the common concept behind modalities. The problem with this approach is that since the more powerful modality (e.g. text) has more information about a common concept, it always contributes more to this joint representation. However, in most of the multimodal applications, we look for a representation with more information about the weaker modality (e.g. image).

As mentioned above, in all of the previous deep models, a joint layer is constructed on the top of the modality-specific deep networks to find the shared representation for multimodal data. However, in this paper, we show that considering interactions between modalities may lead to a better representation. The rationale for this approach is that high level concepts may not contain all the useful information about a modality.

In the proposed method, we utilize the cross-weights (between modalities) that enable us to learn a better representation for each modality and to have a more powerful cross modality learning. Specially, the modality that contains the higher level information can help us to find a better representation for other modalities. For example, in the bi-modal data consisting of text and image modalities, the text modality can help to find better representation of the image modality. Moreover, in top layer of the proposed network, we consider a proportion for the dedicated hidden units to each of the modalities. Although our base model is not supervised, we can also consider supervision to achieve a higher performance. The experimental results show that the performance of the proposed supervised model outperforms the state-of-the-art retrieval methods on PASCAL-sentence and SUN-Attribute databases. In addition, the performance of our unsupervised model is comparable to that of the state-of-the-art supervised models.

The rest of this paper is organized as follows. The related works are introduced in Section 2. The motivation of the proposed deep model for multi-modal data is presented in Section 3. The main ideas of the proposed deep architecture and the pre-training method are discussed in Section 4. In Section 5, we introduce the supervised and unsupervised fine-tuning methods of the proposed deep network. Experimental results are reported in Section 7, and finally we conclude the paper in Section 8.

2. Related works

We will use the upper case letters $\mathbf{X}$ and $\mathbf{Z}$ to show random variables corresponding to the modalities and the label random variable is shown as $\mathbf{T}$. Moreover, $\mathbf{h}(\mathbf{X})$ denotes the differential entropy of the random variable $\mathbf{X}$.

In modeling unimodal data, we seek for a representation that has as high as possible information about data and also removes the noise-related information in the input data. For example, if $\mathbf{X}$ shows the input, we usually have its corrupted version, say $\tilde{\mathbf{X}}$, and would like to learn a representation that is a function of $\tilde{\mathbf{X}}$ and has the highest mutual information with the clean input. Let $\mathbf{f}_\theta$ be the mapping on the corrupted input that results in the new representation. We need to solve the following optimization problem for a family of $\mathbf{f}$ functions:

$$\mathbf{f}_{\theta^*} = \max_{\theta} \mathbf{I}(\mathbf{f}_\theta(\tilde{\mathbf{X}}); \mathbf{X}) \tag{1}$$

In [1], it has been shown that using the stacked denoising auto-encoders leads to a good approximation of the mapping $\mathbf{f}_\theta$ that maximizes the mutual information between the obtained representation and the uncorrupted input while also keeping the dimensionality of the representation as low as possible. However, we cannot use this representation learning approach directly on multimodal data (by only considering concatenated data modalities as the input). Indeed, since the modalities have very distinct statistical properties and inter-modality correlations are much more strong, it is usually hard to learn intra-modality correlations by the standard stacked auto-encoders.

To this end, recent methods focus on finding a higher level representation for each modality and then find a joint representation from them. Methods like [13, 20] learn a joint representation on top of two deep networks. However, for cross modality applications, we actually want to retrieve a modality from the other modality. Therefore, it is more rational to use their conditional probability instead of their joint probability. The authors in [19] try to minimize the variation of information instead of maximizing the log likelihood. But, we usually desire to find a more informative representation using a weaker modality (e.g, image). Another approach is introduced in [18, 17, 8] which tries to match different building blocks in one modality to the other modality. However, the problem is that information in modalities are often complementary and not the same.

3. Motivation

The previous methods try to find a joint representation for both modalities or to find an exact match between parts in different modalities. Sometimes, modalities have complementary information that help us to find a better representation. For example, we may have the following pictures; Black dog, Black cat, white cat, and white dog which may be categorized as Cat and Dog but their tags contain the word black or white. Exploiting these complementary information would help us to find a representation which focuses on differences between a Black dog and a Black cat. Thus, while in the previous methods, we may find a representation at the top of the image-specific network that may not consider a difference between a black and a white dog, in the proposed model, the word black will affect the next layer of the image network and leads to a more proper representation.

In fact, an ideal mapping would preserve sufficient information about clean data while inferring the missing modality from the available modalities. To this end, in the proposed method, we use deep networks both for learning modality-specific representation and for learning representation of one modality from another one. We use the property of multi-modal data (stronger inter-modal correlations) and propose a new multistage pre-training and fine-tuning method for learning the weights in the network.

4. Pre-training Deep Multimodal Network with Cross Weights

In this section, we introduce a deep network for multimodal data and propose a suitable pre-training method for this network. Assume that we have found the optimal mappings in Eq. 1 for the modalities as $\mathbf{f}_{\theta_{\mathbf{X}}}(\tilde{\mathbf{X}})$ and $\mathbf{f}_{\theta_{\mathbf{Z}}}(\tilde{\mathbf{Z}})$. For simplicity we define two random variables $\mathbf{Y_1} = \mathbf{f}_{\theta_{\mathbf{X}}}(\tilde{\mathbf{X}})$ and $\mathbf{Y_2} = \mathbf{f}_{\theta_{\mathbf{Z}}}(\tilde{\mathbf{Z}})$. Consider the two generalized stacked denoising autoencoders [21] that learn representation of the two input modalities as shown in Figure 1. The achieved representations (for the modalities) in the second layer can be good approximations for $\mathbf{Y_1}$ and $\mathbf{Y_2}$.

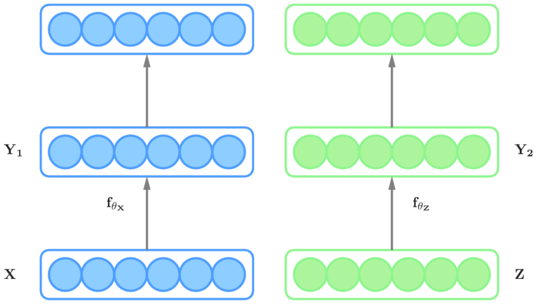

$\mathrm{Y_1}$
$\mathrm{Y_2}$
$\mathrm{f}_{\theta_\mathrm{X}}$
$\mathrm{f}_{\theta_\mathrm{Z}}$
X
Z

Figure 1. Two modality-specific stacked autoencoders.

Then, we try to find $\mathbf{g}_{\theta_{\mathbf{Z}\to\mathbf{Y_1}}}$ and $\mathbf{g}_{\theta_{\mathbf{X}\to\mathbf{Y_2}}}$ such that the two random variables $\mathbf{U_1} = \mathbf{g}_{\theta_{\mathbf{Z}\to\mathbf{X}}}(\tilde{\mathbf{Z}})$ and $\mathbf{U_2} = \mathbf{g}_{\theta_{\mathbf{X}\to\mathbf{Z}}}(\tilde{\mathbf{X}})$ have the same marginal density as $\mathbf{Y_1}$ and $\mathbf{Y_2}$, respectively. The easiest way to find these functions is to learn a mapping from the density of $\mathbf{Z}$ to that of $\mathbf{Y_1}$. In the proposed model, we define cross weights from $\mathbf{Z}$ to $\mathbf{Y_1}$ and pre-train them to minimize the square error of constructing $\mathbf{Y_1}$ from $\mathbf{Z}$. In fact, we consider a single layer neural network with $\mathbf{Z}$ as the input and $\mathbf{Y_1}$ as its output, and try to learn these cross weights as in Figure 2 (cross weights from $\mathbf{X}$ to $\mathbf{Y_2}$ are also found similarly).

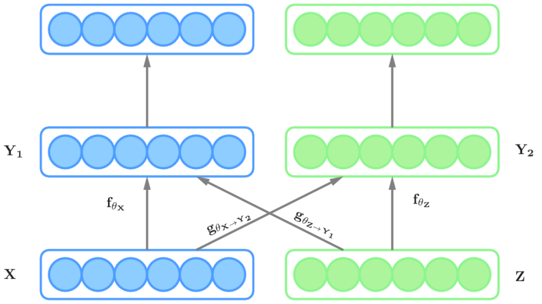

$\mathrm{Y_1}$
$\mathrm{Y_2}$
$\mathrm{f}_{\theta_\mathrm{X}}$
$\mathrm{g}_{\theta_{\mathrm{X}\to\mathrm{Y_2}}}$
$\mathrm{g}_{\theta_{\mathrm{Z}\to\mathrm{Y_1}}}$
$\mathrm{f}_{\theta_\mathrm{Z}}$
X
Z

Figure 2. The network after training the first layer cross weights.

We may continue this method of pre-training cross weights in a deep manner (when we have the higher level representation of each modality) to find all the cross weights. The proposed network is shown in Figure 3.

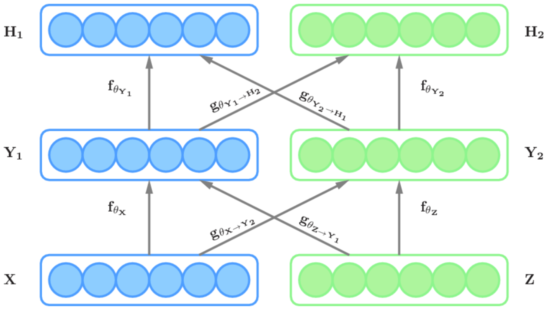

$\mathrm{H_1}$
$\mathrm{H_2}$
$\mathrm{f}_{\theta_{\mathrm{Y_1}}}$
$\mathrm{g}_{\theta_{\mathrm{Y_1}\to\mathrm{H_2}}}$
$\mathrm{g}_{\theta_{\mathrm{Y_2}\to\mathrm{H_1}}}$
$\mathrm{f}_{\theta_{\mathrm{Y_2}}}$
$\mathrm{Y_1}$
$\mathrm{Y_2}$
$\mathrm{f}_{\theta_\mathrm{X}}$
$\mathrm{g}_{\theta_{\mathrm{X}\to\mathrm{Y_2}}}$
$\mathrm{g}_{\theta_{\mathrm{Z}\to\mathrm{Y_1}}}$
$\mathrm{f}_{\theta_\mathrm{Z}}$
X
Z

Figure 3. The network after training all of the cross weights.

Consider different possibilities for the bi-modal input data consisting of text and image, i.e., missing text, missing image, and bi-modal input. For these cases, we intend to discuss the representation obtained in the second layer for the text modality (the resulted representation for the image modality can also be discussed, similarly):

Image is missing: Because $\mathbf{U_1}$ becomes zero, we only have $\mathbf{Y_1}$. Therefore, this situation is similar to the unimodal text-specific stacked auto-encoder before adding the cross weights. Thus, the obtained representation in the second layer of the text-way stacked auto-encoder in the multimodal network, which is shown with the blue color in Figure 3, is the same as the representation obtained in the second layer of the text-specific stacked auto-encoder, i.e., $\mathbf{Y_1}$.

Text is missing: Here, $\mathbf{X}$ contains zeros and we don't have $\mathbf{Y_1}$. However, $\mathbf{Z}$ is available and thus we can calculate $\mathbf{U_1}$ which we have tried to make as close as possible to $\mathbf{Y_1}$. Thus, we can approximately reconstruct $\mathbf{Y_1}$ by only using $\mathbf{Z}$ as input.
Both are present: If both of the modalities are available, we would have $\mathbf{Y_1} + \mathbf{U_1}$ in the second layer of the text stacked autoencoder. We can simply divide this amount into two representations that corresponds to approximation of $\mathbf{Y_1}$ in the second layer.

Summarizing the above cases, we can simply divide the input of the second layer into the number of the available modalities and thus we approximately reach the same representation for unimodal and multimodal inputs.
For the above proposed deep network containing cross-weights, we use a multi-stage learning algorithm. We first pre-train the modality specific autoencoders. Then, the cross weights are pre-trained to minimize the squared error of constructing the modalities from each other. In the next section, we present the theory underlying the proposed network structure and the pre-training method.

4.1. Improved multimodal representation using unimodal representation

In this section, we provide a theorem to show why the proposed method leads to a better representation than the previous deep methods. The graphical model in Figure 4 shows the probabilistic model of our problem. As shown in this figure, only $\mathbf{T}$ causes the two modals $\mathbf{X}$ and $\mathbf{Z}$ to be dependent. Although this is a restricting condition, if $\mathbf{T}$ is a high level concept that is explanatory enough, this model would be the proper generative model for most problems (i.e., if the label random variable is explanatory enough, we can consider it as $\mathbf{T}$).

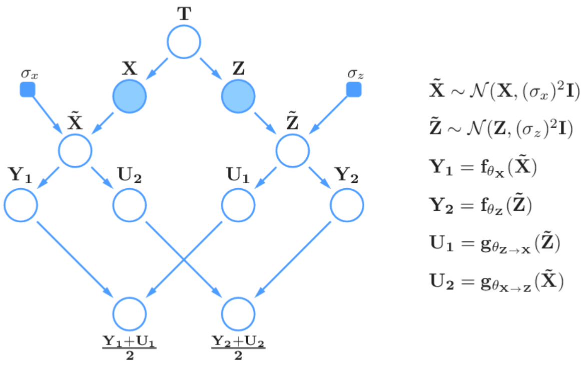

T
X
Z
σ_x
σ_z
$\tilde{\mathbf{X}}$
$\tilde{\mathbf{Z}}$
$\mathbf{Y_1}$
$\mathbf{U_2}$
$\mathbf{U_1}$
$\mathbf{Y_2}$
$\frac{\mathbf{Y_1}+\mathbf{U_1}}{2}$
$\frac{\mathbf{Y_2}+\mathbf{U_2}}{2}$
$\tilde{\mathbf{X}} \sim \mathcal{N}(\mathbf{X}, (\sigma_x)^2\mathbf{I})$
$\tilde{\mathbf{Z}} \sim \mathcal{N}(\mathbf{Z}, (\sigma_z)^2\mathbf{I})$
$\mathbf{Y_1} = \mathbf{f}_{\theta_\mathbf{X}}(\tilde{\mathbf{X}})$
$\mathbf{Y_2} = \mathbf{f}_{\theta_\mathbf{Z}}(\tilde{\mathbf{Z}})$
$\mathbf{U_1} = \mathbf{g}_{\theta_{\mathbf{Z}\to\mathbf{X}}}(\tilde{\mathbf{Z}})$
$\mathbf{U_2} = \mathbf{g}_{\theta_{\mathbf{X}\to\mathbf{Z}}}(\tilde{\mathbf{X}})$

Figure 4. Graphical model for the multimodal input problem. The shaded nodes have been observed.

Definition 4.1 (Variation of Information) *A distance metric in information theory given by* $d(X, Y) = h(X|Y) + h(Y|X)$.

Definition 4.2 (Normalized Variation of Information) $D(X, Y) = 1 - I(X; Y)/h(X, Y)$ *is a set-theoretic metric that is universal. Therefore, if any other distance measure distinguishes that* X *and* Y *are close to each other, then* D *will also judge them close [9].*

Theorem 4.1 *For any two random variables* X *and* Z*, we define four random variables* Y_1 *,* U_1 *,* Y_2 *and* U_2 *as mentioned in section 4. If these random variables have the following properties:*

1. Y_1 *and* U_1 *have the same marginal density conditioned on* X *and* Z *respectively and this marginal density is log-concave.*

2. Y_2 *and* U_2 *have the same marginal density conditioned on* Z *and* X *respectively and this marginal density is log-concave.*

3. Y_1 *conditioned on* X *is independent of* Y_2 *conditioned on* Z.

then the following inequalities hold:

1. $h((Y_1 + U_1)/2, (Y_2 + U_2)/2|X, Z) < h(Y_1, Y_2|X, Z)$
2. $I((Y_1 + U_1)/2; (Y_2 + U_2)/2|X, Z) > I(Y_1; Y_2|X, Z)$
3. $d((Y_1 + U_1)/2, (Y_2 + U_2)/2|X, Z) < d(Y_1, Y_2|X, Z)$
4. $D((Y_1+U_1)/2, (Y_2+U_2)/2|X, Z) < D(Y_1, Y_2|X, Z)$
5. $I((Y_1 + U_1)/2; X, Z) > I(Y_1; X, Z)$

The proof is provided in the supplementary materials. Note that although density functions for $\mathbf{Y_1}$, $\mathbf{U_1}$, $\mathbf{Y_2}$, and $\mathbf{U_2}$ may be complex distributions that are not log-concave, the conditional densities of these variables given the clean input modalities can be usually considered as Gaussian or other exponential family densities that are log-concave[1]. The theorem ensures that if we find suitable $\mathbf{U_1}$ and $\mathbf{U_2}$, i.e., if $\mathbf{U_1}$ and $\mathbf{U_2}$ have enough information about $\mathbf{Z}$ and $\mathbf{X}$ respectively, we can achieve better representation for each modality with more information about other modality and more information about the previous layer of representation compared to the unimodal representation.

5. Fine-tuning Deep Multimodal Network with Cross Weights

The proposed pre-training method in Section 4 is not the only way through which we can consider the same concept

[1] It is the consequence of the chosen model in which we usually assume Gaussian noise and sigmoid activation function that lead to log-concave densities

behind different modalities; we can go further and train the network to produce the same target for uni-modal and multimodal inputs with a common concept. According to the intended application, this target could be reaching the same representation or the same label for these inputs. For example, we intend to find the correct label in classification applications while we prefer to reach the same representation for uni-modal and multimodal inputs with the same concept in cross-modality retrieval applications. Based on the aforementioned applications, we propose both unsupervised and supervised fine-tuning methods for our deep multimodal network.

5.1. Unsupervised Fine-tuning

To achieve the same representation for different modalities, after pre-training, the whole network is fine-tuned with the representation obtained for multimodal data at the last layer. Indeed, the output of the pre-trained network when both modalities are present in the input is considered which is shown inside a red rectangle in Figure 5.

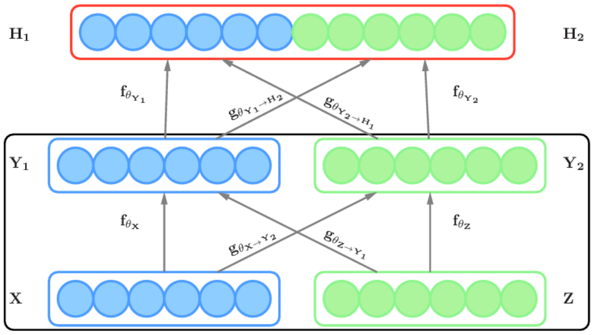


Figure 5. Using last layer representation of the multimodal input as target for unsupervised fine-tuning of Deep Multimodal Network with Cross Weights.

Now this representation is used as the target for both the unimodal and multimodal input cases. Thus, we intend to fine-tune weights such that the obtained representation is close to the representation of multimodal data even when only one of the modalities is present as the input. Therefore, for all the cases of missing text input, missing image input, and bi-modal input, we consider the representation obtained for multimodal data (at the output layer) as the target and use the corresponding unimodal and multimodal samples for the fine-tuning process. This procedure leads to a network which produces approximately the same representation for unimodal and multimodal inputs.

5.2. Supervised Fine-tuning

Since labels can be a good approximation of high level concepts, we can use them to guide our model to learn a better representation. Specially for the text input, labels can improve the learned representation significantly. Our model has the flexablity of using different amount of supervision (i.e., it can be changed from an unsupervised model to a supervised one). Adding supervision usually improves obtained representations and also makes representations of modalities closer to each other.
To incorporate supervisory information, we first learn two modality-specific stacked autoencoders as before. Then, these pre-trained networks are separately fine-tuned with the available labels. After that we learn the cross weights as before. Finally, the whole network is fine-tuned again with available labels (for unimodal and multimodal samples) as opposed to the above unsupervised learning that uses the multimodal representation as the target.

6. Experiments

In this section, we first present results of an experiment conducted on a toy example. Then, we introduce the PASCAL-Sentence and SUN-Attribute datasets on which we run methods and evaluate results. After that, the experimental setup is presented and finally the results of our experiments are reported and discussed[2]

6.1. Toy Example on MNIST

We first evaluate our method on MNIST Dataset [11] composed of hand written digits. As in [19], we halve every image to the left and the right parts and use these parts as the input data modalities. Then, we perform recognition task on the complete input data and halved input data. Table 1 shows recognition errors with different types of input for our method and methods which have been used for comparison in [19].
Our network contains [392 1000 500 300] variables for the left pathway and [392 1000 500 300] for the right one and thus the whole network is composed of [784 2000 1000 600] neurons. Compared to the other methods, the proposed method has higher accuracy for both unimodal and multimodal queries. This in fact shows that our proposed method, uses the information between modalities to improve its knowledge about each of modalities and even leads to a better multimodal representation.

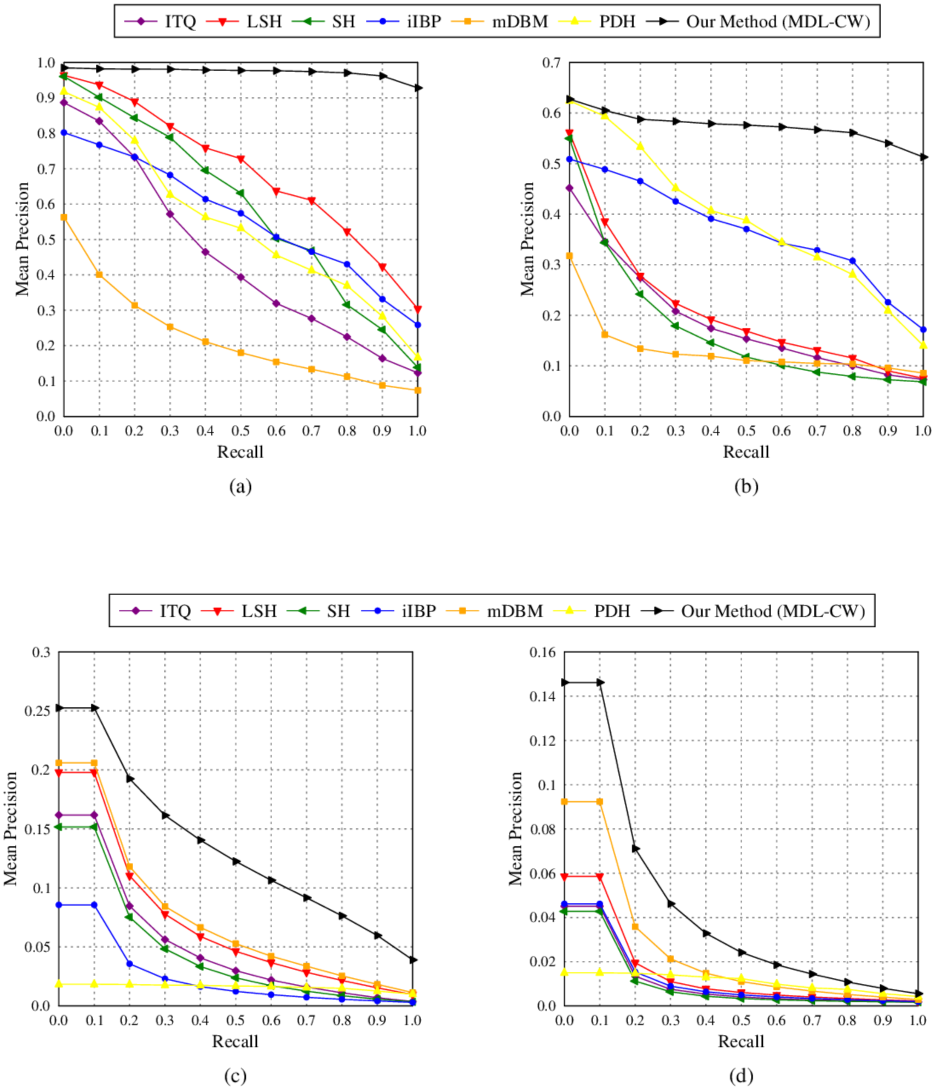
6.2. PASCAL-Sentence

PASCAL-Sentece 2008 database[3] is a collection of images from PASCAL 2008 images along with annotating senteces by Amazon Mechanical Turk workers. There are

[2] The codes are available at http://ml.dml.ir/mdl-cw.

Input modalities at the test time	**Left**	**Right**	**Left+Right**
ML (PCD) [19]	14.98%	18.88%	1.57%
Min VI (CD-Percloss) [19]	9.42%	11.02%	1.71%
Min VI (MP) [19]	6.58%	7.27%	1.73%
Our method (MDL-CW)	4.23%	5.99%	1.39%

Table 1. Test error on MNIST dataset for the methods reported in [19] and our method.



(a)

(b)

(c)

(d)

Figure 6. The result of the category retrieval for all query types (a) text-to-image and (b) image-to-image queries on PASCAL-Sentence dataset and (c) text-to-image and (d) image-to-image queries on SUN-Attribute dataset. Our method using supervised fine-tuning is compared with state-of-the-art supervised methods.

50 images for each of the 20 categories in this dataset. Each image is annotated by five sentences. We used the same visual and textual features as in [3]. Indeed, for images, several object detectors are run and the most confident one is found. For each detector, the coordination of detection along with the confidence value is considered. Moreover, the response of several SVMs trained for each category using GIST descriptor is computed as in [3].

For textual features, a dictionary of discriminative and frequent words in the database sentences is found. For each image, a triplet of <*object, action, scene*> is extracted and the semantic similarity between each word in the triplet and

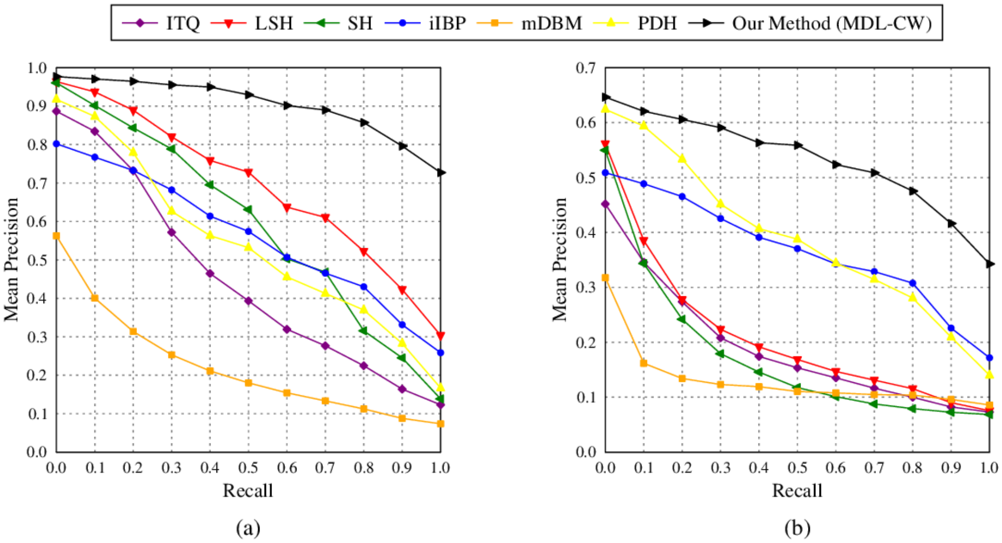


Figure 7. Results of category retrieval for all the query types (a) text-to-image and (b) image-to-image queries on PASCAL-Sentence dataset. Our method using unsupervised fine-tuning is compared with state-of-the-art supervised methods.

all dictionary words is computed by Lin similarity measure [12] on the WordNet hierarchy. Finally, the feature vector is computed as the sum of all the similarity vectors for the words in the triplet [3].

This results in identical features for all of five sentences for an image, to make textual features distinct for every image, bag of words representation of each sentence was concatenated to features described above.

6.3. SUN-Attribute

The SUN-Attribute dataset is a large-scale dataset with 14340 images and 717 categories. There are only 20 images for each category in this dataset and some categories are very close to each other. Each image is annotated with 102 binary attribute labels from three Amazon Mechanical Turk workers. For the final attribute vector, the mean of three annotated vectors is used. The precomputed image features include GIST, 2 × 2 histogram of oriented gradient, self-similarity measure, and geometric context color histograms [15, 23] with dimension 19080. As in [14], we reduce the dimensionality from 19080 to 1000 by randomly selecting features.

6.4. Experimental setup

For the PASCAL-Sentence dataset, we use a network composed of [1408 800 300 128] neurons in the text pathway and [260 400 200 128] neurons in the image pathway for the unsupervised model. Activation function in all encoder layers is set to rectified linear unit (ReLU).

For supervised fine-tuning, a softmax layer with 20 nodes is added on top of each stacked auto-encoder. We also used four dropout [7] modules in between layers to prevent overfitting. Dropout probability is set to 0.35.

For the SUN-Attribute dataset, we use a network composed of [1000 500 200 128] neurons for the image pathway and [102 500 200 128] neurons for the attribute pathway. Rest of network design and learning procedure is similar to what was described for PASCAL-Sentence dataset. Activation used in all layers is ReLU and for supervised fine-tuning using category labels, a softmax layer with size 717 was added on top of each stacked auto-encoder. Also dropout modules with dropout probability of 0.35 was used in between the layers.

6.5. Experimental results

Our method is compared with several methods applied to multimodal data including Locality Sensitivity Hashing(LSH)[5], Spectral Hashing(SH)[22], multimodal Deep Boltzman Machines (mDBM) [20], iterative quantization (ITQ) [6], predictable dual-view hashing [16], and integrative Indian Buffet Process (iIBP) [14]. We applied LSH and SH for each of the modalities separately to show that representation learning using multimodal data will lead to a better performance even on unimodal queries in both modalities.

As in [14], we split the datasets to the same number of test and train images for each category. We used 256 bits for the final representation[3] for all the methods except to (ITQ and PDH) since they don't support this number of bits. For these two methods, we use the maximum number of bits

[3] This was done by threshholding the representation in the last layer for our method and mDBM

Model	Recall@1	Recall@5	Recall@10	Mean rank
Random Ranking	4.0	9.0	12.0	71.0
Socher [18]	23.0	45.0	63.0	16.9
kCCA [10]	21.0	47.0	61.0	18.0
DeViSE [4]	17.0	57.0	68.0	11.9
SDT-RNN [17]	25.0	56.0	70.0	13.4
Deep Fragment [8]	39.0	68.0	79.0	10.5
Our method (MDL-CW)	34.0	70.0	89.0	9.2

Table 2. Pascal1K ranking experiments in image annotation.

Model	Recall@1	Recall@5	Recall@10	Mean rank
Random Ranking	1.6	5.2	10.6	50.0
Socher [18]	16.4	46.6	65.6	12.5
kCCA [10]	16.4	41.4	58.0	15.9
DeViSE [4]	21.6	54.6	72.4	9.5
SDT-RNN [17]	25.4	65.2	84.4	7.0
Deep Fragment [8]	23.6	65.2	79.8	7.6
Our method (MDL-CW)	35.2	72.6	90.6	6.8

Table 3. Pascal1K ranking experiments in image search.

that they support. For each query, we find its representation using each of the above methods and according to the hamming distance between the representation of that query with those of the training images, we order the training images[4].

6.5.1 Supervised model

We consider supevision information for both datasets. However, we didn't fine-tune image related networks in the SUN-Attribute dataset as mentioned in Section 6.4. Mean precision curves for the both datasets are presented in Figure 6. As mentioned above, in the SUN-Attribute dataset, there would be only 10 samples for each class (from 7170 training samples) which leads to a poor performance. As it can be seen in this figure, our method strongly outperforms all the other methods. The qualititative results which is provided in the supplementary materials, show remarkable semantic similarity of both sets of results to the queries.

6.5.2 Unsupervised model

In addition to the supervised model, we have also conducted experiments to evaluate results of our model without any supervision. In Figure 7 , we compared our method with other methods on the PASCAL-Sentence dataset. The proposed method that uses unsupervised fine-tuning outperforms all the previous methods even the supervised ones. These results show that the proposed method can extract high level semantics between modalities even without use of any labels. Therefore, we can take advantage of any amount of supervision which makes our model more flexible than other methods.

In addition to these sets of experiments, in Tables 2 and 3, we compared our method with other state-of-the-art deep learning models. However, we used continuous representation (without thresholding) and cosine similarity in these experiments. The compared models include Deep fragment embeddings for bidirectional image sentence mapping [8], grounded compositional semantics for finding and describing images with sentences [17], parsing natural scenes and natural language with recursive neural networks [18], kernel and nonlinear canonical correlation analysis [10], and DeViSE that is a deep visual-semantic embedding model [4]. Settings for this sort of experiments are as in [8].

7. Conclusion

In this paper, we proposed a novel representation learning framework for multimodal data using deep networks, called MDL-CW. We tried to maximize the mutual information between representations of modalities in a deep manner while the information about individual modalities was also preserved. This leads to a representation that is better than representation obtained for each of the modalities, separately. In the proposed framework, a multi-stage learning method consisting of some pre-training and fine-tuning steps that are useful for multi-modal data was presented. Experimental results on challenging real world datasets demonstrated that MDL-CW outperforms the existing state-of-the-art multimodal methods.

[4] iIBP uses its own method for finding similarity between binary representations.

References

[1] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, pages 899–907, 2013. 2

[2] M. M. Bronstein, A. M. Bronstein, F. Michel, and N. Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3594–3601. IEEE, 2010. 1

[3] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*, pages 15–29. Springer, 2010. 5, 6, 7

[4] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013. 8

[5] A. Gionis, P. Indyk, R. Motwani, et al. Similarity search in high dimensions via hashing. In *VLDB*, volume 99, pages 518–529, 1999. 7

[6] Y. Gong and S. Lazebnik. Iterative quantization: A procrustean approach to learning binary codes. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 817–824. IEEE, 2011. 7

[7] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 7

[8] A. Karpathy, A. Joulin, and F. F. F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014. 2, 8

[9] A. Kraskov, H. Stögbauer, R. G. Andrzejak, and P. Grassberger. Hierarchical clustering using mutual information. *EPL (Europhysics Letters)*, 70(2):278, 2005. 4

[10] P. L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(05):365–377, 2000. 8

[11] Y. LeCun and C. Cortes. The mnist database of handwritten digits, 1998. 5

[12] D. Lin. An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304, 1998. 7

[13] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 689–696, 2011. 1, 2

[14] B. Ozdemir and L. S. Davis. A probabilistic framework for multimodal retrieval using integrative indian buffet process. In *Advances in Neural Information Processing Systems*, pages 2384–2392, 2014. 1, 7

[15] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2751–2758. IEEE, 2012. 7

[16] M. Rastegari, J. Choi, S. Fakhraei, D. Hal, and L. Davis. Predictable dual-view hashing. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1328–1336, 2013. 2, 7

[17] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics*, 2:207–218, 2014. 2, 8

[18] R. Socher, C. C. Lin, C. Manning, and A. Y. Ng. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 129–136, 2011. 2, 8

[19] K. Sohn, W. Shang, and H. Lee. Improved multimodal deep learning with variation of information. In *Advances in Neural Information Processing Systems*, pages 2141–2149, 2014. 1, 2, 5, 6

[20] N. Srivastava and R. R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230, 2012. 1, 2, 7

[21] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *The Journal of Machine Learning Research*, 11:3371–3408, 2010. 3

[22] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *Advances in neural information processing systems*, pages 1753–1760, 2009. 7

[23] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010. 7

[24] E. P. Xing, R. Yan, and A. G. Hauptmann. Mining associated text and images with dual-wing harmoniums. *arXiv preprint arXiv:1207.1423*, 2012. 1

[25] Y. Zhen and D.-Y. Yeung. Co-regularized hashing for multimodal data. In *Advances in Neural Information Processing Systems*, pages 1376–1384, 2012. 1

[26] Y. Zhen and D.-Y. Yeung. A probabilistic model for multimodal hash function learning. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 940–948. ACM, 2012. 1