Seminar Report On

Multimodal Deep Learning

Submitted in Partial Fulfilment of the
Requirments for the award of
Bachelor of Technology in
Computer Science and Engineering
(2013-2017)

Submitted By:
Sangeetha Mathew
Roll No. 13028102

Guide:
Ms.Divya Madhu
Department of CSE

Muthoot Institute of Technology and Science (MITS)
Varikoli P.O, Puthencruz- 682308

MUTHOOT INSTITUTE OF TECHNOLOGY & SCIENCE

Varikoli P.O, Puthencruz- 682308

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the Seminar report entitled "Multimodal Deep Learning" submitted by **Sangeetha Mathew (13028102)** of Semester VII is a bonafide account of the work done by her under our supervision.

Guide
Ms.Divya Madhu
Dept. of CSE

Head of the Department
Dr. Sanju V.
Dept. of CSE

Acknowledgements

I respect and thank **Dr.RamKumar.S**, Principal of MITS for giving me the opportunity to do this seminar.

I would like to sincerely thank my guide **Ms.Divya Madhu**, Asst. Prof, CSE, for her support and valuable guidance. Her timely advice, meticulous scrutiny, scholarly and scientific approach has helped me complete this project on time.

I am thankful to our seminar coordinator **Asst. Prof.Jency Thomas**, for their insight, supervision and encouragement which has helped me to complete this project on time. Also I would like to thank our college, MITS, for providing the best facility and support.

I thank each and every staff of the Computer Science Department for lending us help and support.I express my heartfelt veneration to all who had been helpful and inspiring throughout this endeavour. Last but not the least, I thank the almighty for blessing me for completing the seminar.

Sangeetha Mathew

Abstract

Deep learning is a new area of machine learning research that imitates the way the human brain works. It has a great number of successful applications in speech recognition, image classification, and natural language processing. It is a particular approach to build and train neural networks. A deep neural network consists of a hierarchy of layers, whereby each layer transforms the input data into more abstract representations. Deep networks have been successfully applied to unsupervised and supervised feature learning for single modalities like text, images or audio. As the developments in technology, an application of deep networks to learn features over multiple modalities has surfaced. It involves relating information from multiple sources. The relevance of multi-modality has enhanced tremendously due to extensive use of social media and online advertising. Social media has been a convenient platform for voicing opinions from posting messages to uploading a media fille, or any combination of messages.There are a number of methods that can be used for multimodal deep learning, but the most efficient one is Deep Boltzmann Machine (DBM). The DBM is a fully generative model which can be utilized for extracting features from data with certain missing modalities. DBM is constructed by stacking one Gaussian RBM and one standard binary RBM. An RBM has three components: visible layer, hidden layer, and a weight matrix containing the weights of the connections between visible and hidden units. There are no connections between the visible units or between the hidden units. That is the reason why this model is called restricted.

Contents

1 Introduction **v**

2 Literature Survey **2**

3 Traditional Models **5**

4 Architectures **7**

5 Algorithms **11**

6 Methodology **14**

7 Conclusion **26**

List of Figures

3.1 Audio-Visual User Recognition Systems proposed 6

4.1 Comparison of model structures 8

6.1 RBM 15
6.2 Left:A three-layer Deeep Belief Network and a three-layer Deep Boltzmann Machine.Right:Pretraining consist of learning a stack of modified RBM's 16
6.3 Multimodal DBM 17
6.4 Multimodal learning setting 21

Chapter 1

Introduction

Multimodal sensing and processing have shown promising results in detection, recognition and identifcation in various applications, such as human-computer interaction, surveillance, medical diagnosis, biometrics, etc. There are many ways to generate multiple modalities; one is via sensor diversity (specially in our everyday life tasks) and the other is via feature diversity (using engineered and/or learned features).

In the last few decades, many machine learning models have been proposed to deal with multimodal data. Here I mostly focus on deep learning models for multimodal deep learning. The group of multimodal deep learning approaches that was discussed is all based on Restricted Boltzmann Machines (RBMs). These methods include Deep Belief Networks (DBNs), Deep Boltzmann Machines (DBNs) and Deep Autoencoders. Not only the building blocks of all these models are RBMs, their training algorithms are also very similar. RBM, the model considered here are a group of non-directed probabilistic energy-based graphical models that assign a scalar energy value to each variable configuration. These models are trained in a way that the plausible configurations are associated with lower energies (higher probabilities). An RBM has three components: visible layer, hidden layer, and a weight matrix containing the weights of the connections between visible and hidden units. There are no connections between the visible units or between the hidden units. That is the reason why this model is called restricted".

In the seminar I have studied, these methods and have been compared to more traditional classification approaches such as SVMs and LDA. For that

reason, before getting into deep learning models, we first briefly introduce SVM and LDA and mention a few of their applications in processing and classifying multimodal data.

Chapter 2

Literature Survey

Machine Learning: Take data, train model on data and use the model to make predictions. Feature Engineering: Art of extracting useful patterns from data that will make it easier for Machine Learning models to distinguish between classes. Feature Learning: Feature learning algorithm and the common patterns that are important to distinguish between classes and extract them automatically to be used in a classification or regression process. Deep Learning: The term deep learning originated from new methods and strategies designed to generate deep hierarchies of non-linear features by overcoming the problems with vanishing gradients so that we can train architectures with dozens of layers of non-linear hierarchical features. Affective computation has been extensively studied in the last decades, and many methods are proposed for handling various media types including textual documents, images , music and movies. Two widely investigated tasks are emotion detection and sentiment analysis. Both of them are standard classification problems with different state spaces. Usually emotion detection is defined on several discrete emotions, such as anger, sadness, joy etc., while sentiment analysis aims at categorizing data into positive or negative.

Affective computation has been extensively studied in the last decades, and many methods are proposed for handling various media types including textual documents , images , music and movies . Two widely investigated tasks are emotion detection and sentiment analysis. Both of them are standard classification problems with different state spaces. Usually emotion detection is defined on several discrete emotions, such as anger, sadness, joy etc., while sentiment analysis aims at categorizing data into positive or neg-

ative. Since the adopted techniques of these two tasks are quite similar, we will not differentiate them in this section. Previous efforts are summarized mainly based on the modality of the data they are working on.

For textual data, lexicon-based approach using a set of pre-defined emotional words or icons has been proved to be an effective way.Researches propose to predict the sentiment of tweets by using the emoticons (e.g., positive emoticon ":)" and negative one ": (") and acronyms [e.g., lol (laugh out loudly), gr8 (great) and rotf (rolling on the floor)]. A partial tree kernel is adopted to combine the emoticons, acronyms and Part-of-Speech (POS) tags. Three lexicon emotion dictionaries and POS tags are leveraged to extract linguistic features from the textual documents. A semantic feature is proposed to address the sparsity of microbloggings. The non-appeared entities are inferred using a pre-defined hierarchical entity structure. For example, "iPad" and "iPhone" indicate the appearance of "Product/Apple". Furthermore, the latent sentiment topics are extracted and the associated sentiment tweets are used to augment the original feature space.A set of sentimental aspects, such as opinion strength, emotion and polarity indicators, are combined as meta-level features for boosting the sentiment classification on Twitter messages.

Affective analysis of images adopts a similar framework with general concept detection. In SentiBank, a set of visual concept classifiers, which are strongly related to emotions and sentiments, are trained based on unlabeled Web images. Then, a SVM classifier is built upon the output scores of these concept classifiers. The performance of SentiBank is recently improved by using deep convolution neural network (CNN). Nevertheless, the utility of SentiBank is limited by the number and kind of concepts (or ANPs). Due to the fact that ANPs are visually emotional concepts, selection of right samples for classifier training could be subjective. In addition to the semantic level features, a set of low-level features, such as color-histogram and visual aesthetics, are also adopted. The combined features are then fed into a multi-task regression model for emotion prediction. Hand-crafted features derived from principles-of-art such as balance and harmony are proposed for recognition of image emotion. The deep CNN is directly used for training sentiment classifiers rather than using a mid-level consisting of some general concepts. Since Web images are weakly labeled, the system progressively select a subset of the training instances with relatively distinct sentiment

labels to reduce the impact of noisy training instances.

For emotional analysis of music, various hand-crafted features corresponding to different aspects (e.g., melody, timbre and rhythm) of music are proposed. In [19], the early fused features are characterized by cosine radial basis function A ListNet layer is added on top of the RBF layer for ranking the music in valence and arousal in Cartesian coordinates. Besides hand-crafted features, the authors adopt deep belief networks (DBN) on the Discrete Fourier Transforms (DFTs) of music signals. Then, SVM classifiers are trained on the latent features from hidden layers.

In the video domain, most research efforts are dedicated to movies. A large emotional dataset, which contains about 9,800 movie clips, is constructed. SVM classifiers are trained on different low-level features, such as audio features, complexity and color harmony. Then, late fusion is employed to combine the classifiers. A set of features are proposed based on psychology and cinematography for affective understanding in movies. Early fusion is adopted to combine the extracted features. Other fusion strategies on auditory and visual modalities are studied, a hierarchical architecture is proposed for predicting both emotion intensity and emotion types.
CRF is adopted to model the temporal information in the video sequence. In addition to movies, a large-scale Web video dataset for emotion analysis is recently proposed , where a simplified multi-kernel SVM is adopted to combine the features from different modalities.
Different from those works, the approach proposed in this paper is a fully generative model, which defines a joint representation for various features extracted in different modalities. More importantly, the joint representation conveying information from multiple modalities can still be generated when some modalities are missing, which means that our model does not restrict to the media types of user generated contents.

Chapter 3

Traditional Models

SVM and LDA for Multimodal Data
Several groups of researchers have proposed multimodal classification and data fusion SVM and LDA-based approached. The authors of [???] claim that existing multi-biometric fusion techniques face a number of limitations since they are based on the assumptions that each biometric modality is local, complete, and static. These limitations are particularly pronounced when considered in the context of biometric identification, as opposed to verification. Key limitations include:
1. Each registered person must be entered into every modality. This may not be plausible and is very restrictive. Moreover, this makes adding additional modalities to an existing system difficult or impossible.
2. All of the classifiers must always be available. This will not be the case if the modalities are part of a distributed system, such as when a multi-biometric fusion may degrade as individuals are later added to or removed from the system.
3. Limited to verification. Due to the other limitations listed above, most existing fusion techniques are explicitly designed for verification only – identification is not supported.
They propose a novel multi-biometric fusion technique that addresses the issues listed above and is suitable for both identification and verification. A mediator agent controls the fusion of the individual biometric match scores, using a "bank" of SVMs that cover all possible subsets of the biometric modalities being considered. This agent selects an appropriate SVM for fusion, based on which modality classifiers are currently available and have sensor data for the identity in question. This fusion technique differs from a

traditional SVM ensemble – rather than combining the output of all of the SVMs, we apply only the SVM that best corresponds to the available modalities. The mediator agent also controls the learning of new SVMs when modalities are added to the system or sufficient changes have been made to the data in existing modalities. The experiments utilize the following biometric modalities: face, fingerprint, and DNA profile data. We empirically show that our multiple SVM technique produces more accurate results than the traditional single SVM approach.The pipeline of this approach is shown in figure below

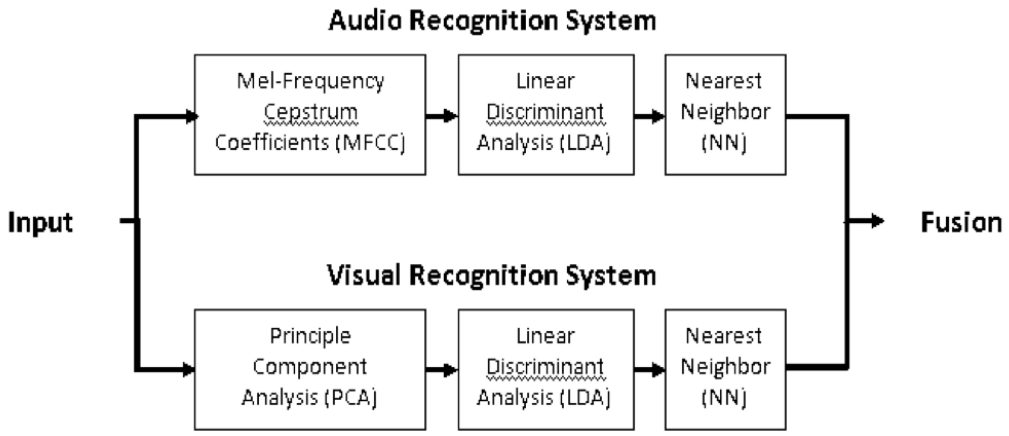

Figure 3.1: Audio-Visual User Recognition Systems proposed

Comparisons and Discussions

Both Linear Discriminant Analysis and Support Vector Machines compute hyperplanes that are optimal with respect to their individual objectives. However, there can be vast differences in performance between the two techniques depending on the extent to which their respective assumptions agree with problems at hand.

It's true that LDA and linear SVM share much in common, both draw a line, the technical differences between these two is significant. As a very informal explanation, LDA draws lines, while SVM can be nonlinear and draw curves instead. Also, as the linear SVM is a super-class (or generalization of) LDA, it is generally the better or more sophisticated approach.

Chapter 4

Architectures

There are a number of deep learning models for multimodal sensing and processing. The first group of multimodal deep learning approaches that we study are all based on Restricted Boltzmann Machines (RBMs). These methods include Deep Belief Networks (DBNs), Deep Boltzmann Machines (DBNs) and Deep Autoencoders. Not only the building blocks of all these models are RBMs, their training algorithms are also very similar.

4.1 Boltzmann Machine

A Boltzmann Machine is a network of symmetrically connected neuron like units that take stochastic decisions about whether to turn on or off. These neurons include both hidden and visible units. An energy function is used for their activation. They are one of the first examples of a neural network capable of learning internal representations, and are able to represent and solve difficult combinatorial problems. However due to a number of issues such as the machine seems to stop learning as it is scaled up owing to exponential time requirements with increase in machine size and number of connections between neurons and noise causing the connection strength to randomize. So Boltzmann Machines with unrestricted connectivity are not much used in machine learning. However Boltzmann Machines with restrictions on connectivity between neurons are useful in field of ML. This is since they are free of these issues we discussed

4.2 Restricted Boltzmann Machine(RBM)

RBM is a kind of Boltzmann Machine whose connections are restricted with the simple rule that it's neurons must form a bipartite graph i.e. its neurons should be dividable into 2 disjoint sets. RBMs perform a kind of factor analysis on input data,extracting a smaller set of hidden variables, that can be used as data representation. It is different from other representation algorithms in Machine Learning(ML) due to 2 things.
Stochastic
Generative
Being stochastic means it's neuron values are calculated based on probability distribution. Since it is generative, it can generate data on its own after learning.

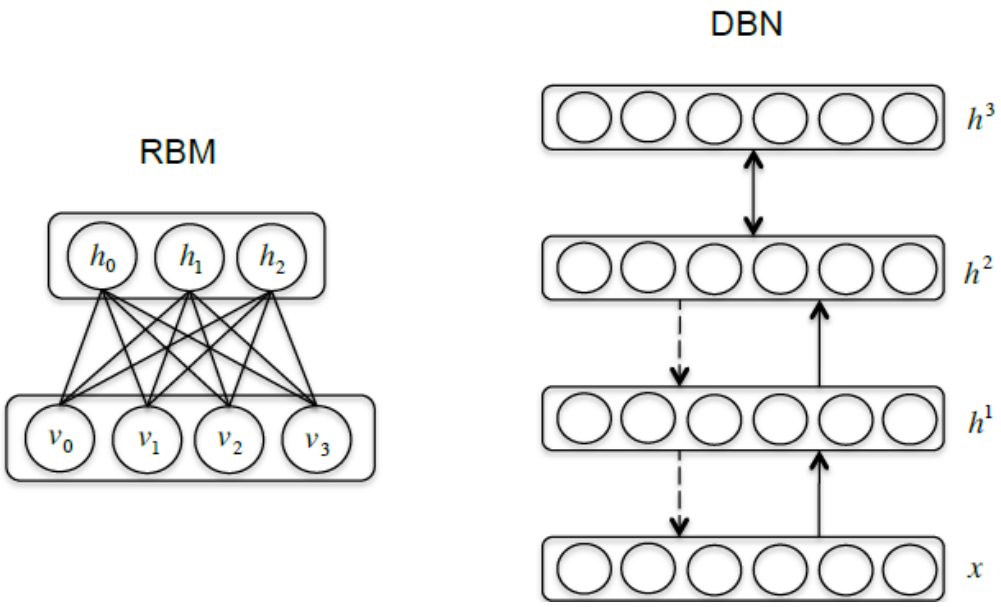

Figure 4.1: Comparison of model structures

4.3 Deep Boltzmann Machine(DBM)

A DBM is a deep multimodal Boltzmann Machine with restrictions. DBM's have the potential of learning internal representations that become increasingly complex, which is considered to be a promising way of solving object

and speech recognition problems. Their training can be done from a large supply of unlabeled sensory inputs and very limited labeled data can then be used to only slightly fine tune the model for a specific task at hand. DBMs also handle ambiguous inputs more robustly. This is since they incorporate top-down feedback for the training procedure.As we can see from figure a DBM may consist of several RBMs connected together.

4.4 Comparison

A naive approach for multimodal deep learning is to concatenate the data descriptors from different input sources to construct a single high-dimensional feature vector and use it to solve a unimodal representation learning problem.However, the correlation between features in each data modality is much stronger than that between data modalities. As a result, the learning algorithms are easily tempted to learn dominant patterns in each data modality separately while giving up learning patterns that occur simultaneously in multiple data modalities.
To resolve this issue, deep learning methods, such as deep autoencoders or deep Boltzmann machines (DBM), have been adapted, where the common strategy is to learn joint representations that are shared across multiple modalities at the higher layer of the deep network, after learning layers of modality-specific networks. The rationale is that the learned features may have less within-modality correlation than raw features, and this makes it easier to capture patterns across data modalities. This has shown promise, but there still remains the challenging question of how to learn associations between multiple heterogeneous data modalities so that we can effectively deal with missing data modalities at testing time.

One necessary condition for a good generative model of multimodal data is the ability to predictor reason about missing data modalities given partial observation. Honglak Lee's research group at the University of Michigan have proposed a new approach to satisfy this condition and improve multimodal deep learning. There emphasis is on efficiently learning associations between heterogeneous data modalities. According to their study, the data from multiple sources are semantically correlated and provide complementary information about each other and a good multimodal model must be able to generate a missing data modality given the rest of the modalities.

They propose a novel leaning framework that explicitly aims at this goal by training the model to minimize the Variation of Information (VI) instead of maximizing the likelihood.

Chapter 5

Algorithms

5.1 Contrastive Divergence

In my paper, the learning process of proposed model is split into two phases. In the first phase, each RBM component of the proposed multimodal DBM is pre-trained by using the greedy layerwise pretraining strategy. Under this scheme, we will train each layer, with a set of different parameters and choose the best performing parameter set for the model. For this a contrastive divergence(CD) algorithm is utilized, since the time complexity for computation increases with number of neurons. The 1-step contrastive divergence(CD1) algorithm is widely used for RBM training, to perform approximate learning for learning parameters. CD allows us to approximate the gradient of energy function. The approximation of the gradient is based on a Markov chain.
In CD1 algorithm a Markov chain is run for one full step and them the parameters are modified to reduce the likelihood of chain wandering of from the initial distribution. This reduces the time and computational effort since we are not waiting the chain to run to equilibrium state and comparing initial and final distribution. The distribution generated by Markov Chain can be thought approximately as the distribution generated by RBMs since they both alter the energy function. The one step running of the Markov Chain is since the results of that single step itself would give us the direction of change of the parameters(gradient).
The CD1 actually performs poorly in approximating the size of the change in parameters. However, it is accurate enough for learning a RBM to provide hidden features for a high-level RBM training. This is since CD1 retains

most of the information about inputs, as it involves single step calculations. The greedy layer-by-layer pretraining algorithm relies on learning a stack of RBM's with a small modification. The key intuition is that for the lower-level RBM to compensate for the lack of top-down input into h1, the input must be doubled, with the copies of the visible-to-hidden connections tied. Conversely, for the top-level RBM to compensate for the lack of bottom-up input into h2, the number of hidden units is doubled. For the intermediate layers, the RBM weights are simply doubled. The stack of RBMs can then be trained in a greedy layer-by-layer fashion using the CD algorithm.

5.2 Greedy layerwise pretraining strategy

Greedy layer-wise supervised training A reasonable question to ask is whether the fact that each layer is trained in an unsupervised way is critical or not. An alternative algorithm is supervised, greedy and layer-wise: train each new hidden layer as the hidden layer of a one-hidden layer supervised neural network NN (taking as input the output of the last of previously trained layers), and then throw away the output layer of NN and use the parameters of the hidden layer of NN as pre-training initialization of the new top layer of the deep net, to map the output of the previous layers to a hopefully better representation. Pseudo-code for a deep network obtained by training each layer as the hidden layer of a supervised one-hidden-layer neural network
During each phase of the greedy unsupervised training strategy, layers are trained to represent the dominant factors of variation extant in the data. This has the effect of leveraging knowledge of X to form, at each layer, a representation of X consisting of statistically reliable features of X that can then be used to predict the output (usually a class label) Y. This perspective places unsupervised pre-training well within the family of learning strategies collectively know as semisupervised methods. As with other recent work demonstrating the effectiveness of semi-supervised methods in regularizing model parameters, we claim that the effectiveness of the unsupervised pre-training strategy is limited to the extent that learning P(X) is helpful in learning P(Y—X). Here, we find transformations of X—learned features—that are predictive of the main factors of variation in P(X), and when the pre-training strategy is effective,2 some of these learned features of X are also predictive of Y. In the context of deep learning, the greedy unsupervised strategy may also have a special function. To some degree it resolves the problem of si-

multaneously learning the parameters at all layers by introducing a proxy criterion. This proxy criterion encourages significant factors of variation, present in the input data, to be represented in intermediate layers.

Chapter 6

Methodology

The learning of our proposed model is not trivial due to multiple layers of hidden units and multiple modalities.Here the methodology is to split the learning process into two stages. First, each RBM component of the proposed multimodal DBM is pretrained by using the greedy layerwise pre- training strategy. In this stage, the time cost for exactly computing the derivatives of the probability distributions with respect to parameters increases exponentially with the number of units in the network. Thus, we adopt 1-step contrastive divergence, an approximate learning method.The second way is to infer the missing modalities by alternating Gibbs sampling. Meanwhile, the joint representation is updated with the generated data of missing modalities.

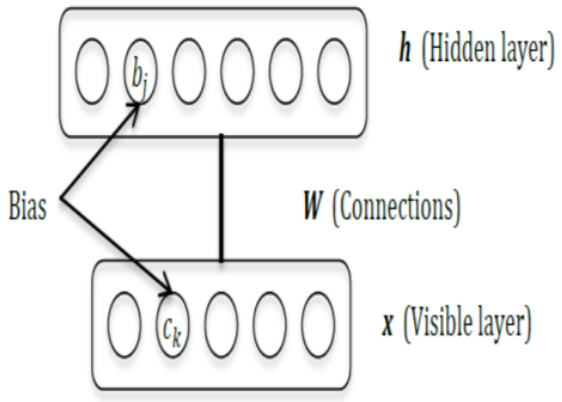

Figure 6.1: RBM

The proposed network architecture, which is shown above is com- posed of three different pathways respectively for visual, auditory and textual modalities. Each pathway is formed by stacking multiple Restricted Boltzmann Machines (RBM), aiming to learn several layers of increasingly complex representations of individual modality.We adopt Deep Boltzmann Machine (DBM) in multimodal learning framework. Different from other deep networks for extracting feature, such as Deep Belief Networks (DBN) and denoising Autoencoders (dA), DBM is a fully generative model which can be utilized for extracting features from data with certain missing modalities.
Additionally, besides the bottom-up information propagation in DBN and dA, a top-down feedback is also incorporated in DBM, which makes the DBM more stable on missing or noisy inputs such as weakly labeled data on the Web. The pathways eventually meet and the sophisticated non-linear relationships among three modalities are jointly learned. The final joint represented in an unified way.Every RBM tries to optimize its energy function in order to maximize the probability of the training data.
DBNs can be trained using the CD algorithm to extract a deep hierarchical representation of the training data. During the learning process, the DBN is first trained one layer at a time, in a greedy unsupervised manner, by treating the values of hidden units in each layer as the training data for the next layer (except for the first layer, which is fed with the raw input data). This learning procedure, called pre-training, finds a set of weights that determine how the variables in one layer depend on the variables in the layer above.

These parameters capture the structural properties of the training data. If the network is to be used for a classification task, then a supervised discriminative fine-tuning is performed by adding an extra layer of output units and back-propagating the error derivatives (using some form of stochastic gradient descent, or SGD).

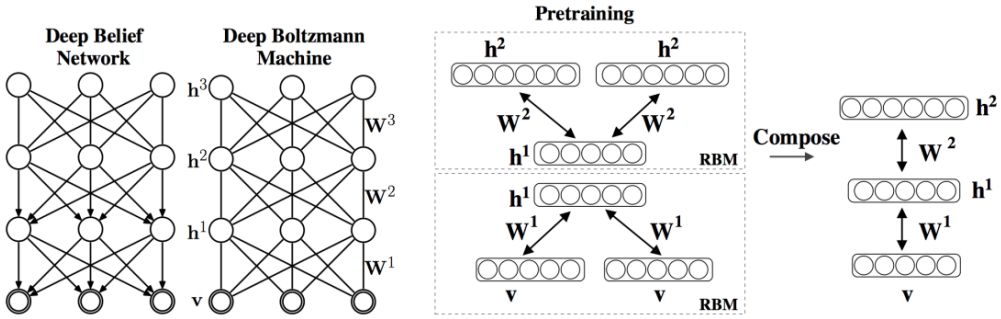

Figure 6.2: Left:A three-layer Deeep Belief Network and a three-layer Deep Boltzmann Machine.Right:Pretraining consist of learning a stack of modified RBM's

To generate a sample from the DBN, we need to perform Gibbs sampling for a long time between the top two layers $h1$ and $h2$ until we converge to a sample of the $h2$ layer, then traverse the rest of the DBN in a top-down manner using the conditional probability distributions to generate the desired sample at the visible layer.
Erhan et al. (2009) studies the reasons why pre-trained deep networks work much better than traditional neural networks and proposes several possible explanations. One possible explanation is that pre-training initializes the parameters of the network in an area of parameter space where optimization is easier and better local optima is found. This is equivalent to penalizing solutions that are outside a particular region of the solution space. Another explanation is that pre-training acts as a kind of regularizer that minimizes the variance and introduces a bias towards configurations of the parameters that stochastic gradient descent can explore during the supervised learning phase, by defining a data-dependent prior on the parameters obtained through the unsupervised learning. In other words, pre-training implicitly imposes constraints on the parameters of the network to specify which minimum out of all local minima of the objective function is desired. The effect

of pre-training relies on the assumption that the true target conditional distribution are structure with the input distribution .

6.1 Multimodal DBM

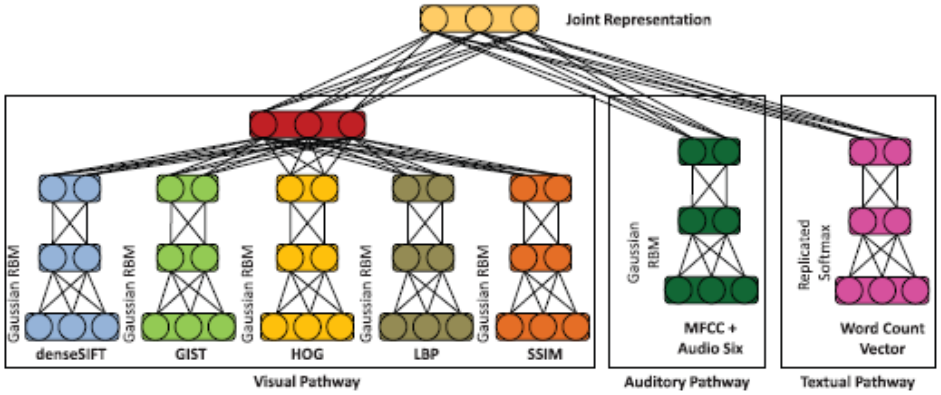

Figure 6.3: Multimodal DBM

Figure above shows the proposed network architecture, which is composed of three different pathways respectively for visual, auditory and textual modalities. Each pathway is formed by stacking multiple Restricted Boltzmann Machines (RBM), aiming to learn several layers of increasingly complex representations of individual modality. Similar to [23], we adopt Deep Boltzmann Machine (DBM) in our multimodal learning framework. Different from other deep networks for extracting feature, such as Deep Belief Networks (DBN) [24] and denoising Autoencoders (dA) [25], DBM is a fully generative model which can be utilized for extracting features from data with certain missing modalities. Additionally, besides the bottom-up information propagation in DBN and dA, a top-down feedback is also incorporated in DBM, which makes

the DBM more stable on missing or noisy inputs such as weakly labeled data on the Web. The pathways eventually meet and the sophisticated non-linear relationships among three modalities are jointly learned. The final joint representation can be viewed as a shared embedded space, where the features with very different statistical properties from different modalities can be represented in an unified way.

Visual Pathway The visual input consists of five complementary low-level features widely used in previous works. As shown in Figure, each feature is modeled with a separate two-layer DBM.Pathway denote the set of five features, respectively as DenseSIFT, GIST, HOG, LBP and SSIM.

Auditory Pathway The input features adopted in auditory pathway are MFCC and Audio-Six (i.e., Energy Entropy, Signal Energy, Zero Crossing Rate, Spectral Rolloff, Spectral Centroid, and Spectral Flux). The Audio-Six descriptor, which can capture different aspects of an audio signal, is expected to be complementary to the MFCC. Since the dimension of Audio-Six is only six, we directly concatenate the MFCC feature with Audio-Six rather than separating them into two sub-pathways as the design in visual pathway. The correlation between these two features can be learned by the deep architecture of DBM. Let denote the real-valued auditory features and and represent h1 and h2(hidden layers) the first and second hidden layers respectively.The DBM is constructed by stacking one Gaussian RBM and one standard binary RBM.

Textual Pathway Different from the visual and auditory modalities, the inputs of the textual pathway are discrete values (i.e., count of words). Thus, we use Replicated Softmax to model the distribution over the word count vectors. Let one visible unit denoting the associated metadata (i.e., title and description) of a video , and denotes the count of the th word in a pre-defined dictionary containing words.

6.2 Modeling Tasks

Generating Missing Modalities: As argued in the introduction, many real-world applications will often have one or more modalities missing. The Multimodal DBM can be used to generate such missing data modalities by clamping the observed modalities at the inputs and sampling the hidden modalities from the conditional distribution by running the standard alternating Gibbs sampler

Inferring Joint Representations: The model can also be used to generate

a fused representation that multiple data modalities. This fused representation is inferred by clamping the observed modalities and doing alternating Gibbs sampling to sample fromtwo layers (if both modalities are present) or from other (if text is missing).
This representation can then be used to do information retrieval for multimodal or unimodal queries. Each data point in the database (whether missing some modalities or not) can be mapped to this latent space. Queries can also be mapped to this space and an appropriate distance metric can be used to retrieve results that are close to the query.
Discriminative Tasks: Classifiers such as SVMs can be trained with these fused representations as inputs. Alternatively, the model can be used to initialize a feed forward network which can then be finetuned. In our experiments, logistic regression was used to classify the fused representations. Unlike finetuning, this ensures that all learned representations that we compare (DBNs, DBMs and Deep Autoencoders) use the same discriminative model.

6.3 Classification Tasks

Multimodal Inputs: Our first set of experiments, evaluate the DBM as a discriminative model for multimodal data. For each model that we trained, the fused representation of the data was extracted and feed to a separate logistic regression for each of the 38 topics. The text input layer in the DBM was left unclamped when the text was missing. Fig. 4 summarizes the Mean Average Precision (MAP) and precision@50 (precision at top 50 predictions) obtained by different models. Linear Discriminant Analysis (LDA) and Support Vector Machines (SVMs) [2] were trained using the labeled data on concatenated image and text features that did not include SIFT-based features. Hence, to make a fair comparison, our model was first trained using only labeled data with a similar set of features (i.e., excluding our SIFT-based features). We call this model DBM-Lab. Fig. 4 shows that the DBM-Lab model already outperforms its competitor SVM and LDA models. DBMLab achieves a MAP of 0.526, compared to 0.475 and 0.492, achieved by SVM and LDA models.To measure the effect of using unlabeled data, a DBM was trained using all the unlabeled examples that had both modalities present. We call this model DBM-Unlab. The only difference between the DBM-Unlab and DBM-Lab models is that DBM-Unlab used unlabeled data during its pretraining stage. The input features for both models remained the

same. Not surprisingly, the DBM-Unlab model significantly improved upon DBM-Lab achieving a MAP of 0.585. Our third model, DBM, was trained using additional SIFT-based features. Adding these features improves the MAP to 0.609. We compared our model to two other deep learning models: Multimodal Deep Belief Network (DBN) and a deep Autoencoder model. These models were trained with the same number of layers and hidden units as the DBM. The DBN achieves a MAP of 0.599 and the autoencoder gets 0.600. Their performance was comparable but slightly worse than that of the DBM. In terms of precision@50, the autoencoder performs marginally better than the rest. We also note that the Multiple Kernel Learning approach proposed in Guillaumin et. al.achieves a MAP of 0.623 on the same dataset. However, they used a much larger set of image features (37,152 dimensions).
Unimodal Inputs: Next, we evaluate the ability of the model to improve classification of unimodal inputs by filling in other modalities. For multimodal models, the text input was only used during training. At test time, all models were given only image inputs.

6.4 Retrieval Tasks

Multimodal Queries: The next set of experiments was designed to evaluate the quality of the learned joint representations. A database of images was created by randomly selecting 5000 imagetext pairs from the test set. We also randomly selected a disjoint set of 1000 images to be used as queries. Each query contained both image and text modalities. Binary relevance labels were created by assuming that if any of the 38 class labels overlapped between a query and a data point, then that data point is relevant to the query. Fig. 5a shows the precision-recall curves for the DBM, DBN, and Autoencoder models (averaged over all queries). For each model, all queries and all points in the database were mapped to the joint hidden representation under that model. Cosine similarity function was used to match queries to data points. The DBM model performs the best among the compared models achieving a MAP of 0.622. The autoencoder and DBN models perform worse with a MAP of 0.612 and 0.609 respectively.Note that even though there is little overlap in terms of text, the model is able to perform well.
Unimodal Queries: The DBM model can also be used to query for unimodal inputs by filling in the missing modality. Fig. 5b shows the precision-recall curves for the DBM model along with other unimodal models, where

each model received the same image queries as input. By effectively inferring the missing text, the DBM model was able to achieve far better results than any unimodal method (MAP of 0.614 as compared to 0.587 for an Image-DBM and 0.578 for an Image-DBN).

6.5 Multimodal learning setting

We will consider the learning settings shown in Figure. The overall task can be divided into three phases – feature learning, supervised training, and testing. We keep the supervised training and testing phases fixed and examine different feature learning models with multimodal data. In detail, we consider three learning settings – multimodal fusion, cross modality learning, and shared representation learning.

	Feature Learning	Supervised Training	Testing
Classic Deep Learning	Audio	Audio	Audio
	Video	Video	Video
Multimodal Fusion	Audio + Video	Audio + Video	Audio + Video
Cross Modality Learning	Audio + Video	Audio	Audio
	Audio + Video	Video	Video
Shared Representation Learning	Audio + Video	Audio	Video
	Audio + Video	Video	Audio

Figure 6.4: Multimodal learning setting

For the multimodal fusion setting, data from all modalities is available at all phases; this represents the typical setting considered in most prior work in audio-visual speech recognition [3]. In cross modality learning, one has access to data from multiple modalities only during feature learning. During the supervised training and testing phase, only data from a single modality is provided. In this setting, the aim is to learn better single modality representations given unlabeled data from multiple modalities. Last, we consider

a shared representation learning setting, which is unique in that different modalities are presented for supervised training and testing. This setting allows us to evaluate if the feature representations can capture correlations across different modalities. Specifically, studying this setting allows us to assess whether the learned representations are modality-invariant.

6.6 Datasets and Task

Since only unlabeled data was required for unsupervised feature learning, we combined diverse datasets to learn features. We used all the datasets for feature learning. AVLetters and CUAVE were further used for supervised classification. We ensured that no test data was used for unsupervised feature learning.

CUAVE 36 individuals saying the digits 0 to 9. We used the normal portion of the dataset where each speaker was frontal facing and spoke each digit 5 times. We evaluated digit classification on the CUAVE dataset in a speaker independent setting. As there has not been a fixed protocol for evaluation on this dataset, we chose to use odd-numbered speakers for the test set and evennumbered ones for the training set.

AVLetters" 10 speakers saying the letters A to Z, three times each. The dataset provided preextracted lip regions at 60x80 pixels. As we were not able to obtain the raw audio information for this dataset, we used it for evaluation on a visual-only lipreading task. We report results on the third-test settings used for comparisons.

AVLetters2: 5 speakers saying the letters A to Z, seven times each. This is a new high definition version of the AVLetters dataset. We used this dataset for unsupervised training only.

Stanford Dataset: 23 volunteers spoke the digits 0 to 9, letters A to Z and selected sentences from the TIMIT dataset. We collected this data in a similar fashion to the CUAVE dataset and used for unsupervised training only.

TIMIT: We used the TIMIT dataset for unsupervised audio feature pretraining.
We note that in all datasets there is variability in the lips in terms of appearance, orientation and size.

Our features were evaluated on speech classification of isolated letters and digits. We extracted features from overlapping windows. Since examples had varying durations, we divided each example into S equal slices and performed average-pooling over each slice. The features from all slices were subsequently concatenated together. We combined features using S = 1 and S = 3 to form our final feature representation for classification using a linear SVM.

6.7 Cross Modality Learning

We first evaluate the learned features in a setting where unlabeled data for both modalities are available during feature learning, while during supervised training and testing phases only a single modality is presented. In these experiments, we evaluate cross modality learning where one learns better representations for one modality (e.g., video) when given multiple modalities (e.g., audio and video) during feature learning. For the bimodal deep autoencoder, we set the value of the other modality to zero when computing the shared representation which is consistent with the feature learning phase. All deep autoencoder models are trained with all available unlabeled audio and video data.

On the AVLetters dataset, there is an improvement over hand-engineered features from prior work. The deep autoencoder models performed the best on the dataset, obtaining a classification score of 65.8%, outperforming the best previous published results.

On the CUAVE dataset (Table 1b), there is an improvement by learning video features with both video audio compared to learning features with only video data. The deep autoencoder models ultimately performs the best, obtaining a classification score of 69.7%. In our model, we chose to use a very simple front-end that only extracts bounding boxes (without any correction for orientation or perspective changes). A more sophisticated visual

front-end in conjunction with our models has the potential to do even better.

The video classification results show that the deep autoencoder model achieves cross modality learning by discovering better video representations when given additional audio data. In particular, even though the AVLetters dataset did not have any audio data, we were able to obtain better performance by learning better video features using other unlabeled data sources which had both audio and video data.

However, we also note that cross modality learning did not help to learn better audio features; since our feature learning mechanism is unsupervised, we find that our model learns features that adapt to the video modality but are not useful for speech classification.

6.8 Cross-modal retrieval

Nowadays, mobile devices and emerging social websites (e.g., Facebook, Flickr, YouTube, and Twitter) are changing the ways people interact with the world and search information of interest. It is convenient if users can submit any media content at hand as the query. Suppose we are on a visit to the Great Wall, by taking a photo, we may expect to use the photo to retrieve the relevant textual materials as visual guides for us. Therefore, cross-modal retrieval, as a natural searching way, becomes increasingly important.

Cross-modal retrieval aims to take one type of data as the query to retrieve relevant data of another type. For example, the text is used as the query to retrieve images. Furthermore, when users search information by submitting a query of any media type, they can obtain search results across various modalities, which is more comprehensive given that different modalities of data can provide complementary information to each other. More recently, cross-modal retrieval has attracted considerable research attention.

The challenge of cross-modal retrieval is how to measure the content similarity between different modalities of data, which is referred as the heterogeneity gap. Hence, compared with traditional retrieval methods, cross-modal retrieval requires cross-modal relationship modeling, so that users can retrieve what they want by submitting what they have. Now, the main research effort is to design the effective ways to make the cross-modal retrieval

more accurate and more scalable.

In the cross-modal retrieval procedure, users can search various modalities of data including texts, images and videos, starting with any modality of data as a query. The general framework of cross-modal retrieval, in which, feature extraction for multimodal data is considered as the first step to represent various modalities of data. Based on these representations of multimodal data, cross-modal correlation modeling is performed to learn common representations for various modalities of data. At last, the common representations enable the cross-modal retrieval by suitable solutions of search result ranking and summarization.Shorly we can say multi-modal retrieval is to use both the image and the text to find similar content (maybe also multi-modal content, i.e. image+text, but maybe just images or just text). That is, I search for content that matches somehow both the image and the text in the query. cross-modal retrieval is to use one modality to find similar content in the other modality. E.g. to use the text to find images matching that text (which would then also match the original image, if the association between the original image and the original text holds).

Chapter 7

Conclusion

This seminar presented a deep model for learning multimodal signals coupled with emotions and semantics. Particularly, we propose a multi-pathway DBM architecture dealing with low-level features of various types and more twenty-thousand dimensions, which is not previously attempted to the best of our knowledge. The major advantage of this model is on capturing the non-linear and complex correlations among different modalities in a joint space. The model enjoys peculiarities such as learning is unsupervised and can cope with samples of missing modalities. Compared with hand-crafted features, our model generates much more compact features and allows natural cross-modal matching beyond late or early fusion. As demonstrated on ImageTweets datasets, the features generated by mapping single modality samples (text or visual) into the joint space consistently outperform hand-crafted features in sentiment classification. In addition, we show the complementary between deep and hand-crafted features for emotion prediction on Video Emotion dataset. Among the eight categories of emotion, nevertheless, the categories 'anticipation' and 'surprise' remain difficult either with learnt or hand-tuned features. For video retrieval, our model shows favorable performances, convincingly outperforms hand-crafted features over different types of queries. Encouraging results are also obtained when applying the deep features for cross-modal retrieval, which is not possible for hand-crafted features.Hence, the learning is fully generative and the model is more expressive.

Bibliography

[1] Lei Pang, Shiai Zhu, and Chong-Wah Ngo *Deep Multimodal Learning for Affective Analysis and Retrieval*IEEE Transactions on Multimedia, Volume 17,No.11,November 2015

[2] Y.-G. Jiang, B. Xu, and X. Xue,*Predicting emotions in user-generated videos,* in Proc. AAAI, pp. 73–79,in 2014.

[3] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau,*Sentiment analysis of Twitter data,* in Proc. Workshop Languages Social, pp. 30–38 Media, 2011.

[4] Dalal, N., and Triggs,, *Histograms of oriented gradients for human detection,* In Proc. of IEEE Conference on Computer Vision and Pattern Recognition 2005.

[5] T. Chen, D. Borth, T. Darrell, and S. Chang,*DeepSentiBank:Visual sentiment concept classification with deep convolutional neuralnetworks,*CoRR, 2014

[6] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, *Large-scale visual sentiment ontology and detectors using adjective noun pairs,* in Proc. ACM MM, pp. 223–232,ACM 2013.

[7] Sivic, J., and Zisserman, *A text retrieval approach to object matching in videos*, in Proceedings of the Ninth IEEE International Conference on Computer Vision (ICCV 2003) 2-Volume Set,2003

[8] Y. Bengio: *Learning deep architectures for ai*, Foundations and Trends in Machine Learning, vol. 2, no. 1, pp. 1–127, 2009.