Multimodal Stacked Denoising Autoencoders

Patrick Poirson
School of Computing
University of South Alabama
plp1003@jagmail.southalabama.edu

Haroon Idrees
Center for Research in Computer Vision
University of Central Florida
haroon@eecs.ucf.edu

Abstract

We propose a Multimodal Stacked Denoising Autoencoder for learning a joint model of data that consists of multiple modalities. The model is used to extract a joint representation that fuses modalities together. We have found that this representation is useful for classification tasks. Our model is made up of layers of denoising autoencoders which are trained locally to denoise corrupted versions of their inputs. This allows the model to learn useful features from unlabeled data in an unsupervised fashion. Our results on multimodal data consisting of images and text show that our model significantly outperforms SVMs and LDA on discriminative tasks. We also present our work towards generating tags from a image.

1. Introduction

In this paper, we consider the problem of image classification using the image along with its associated keywords or tags, such as those found on the photo sharing website Flickr [6]. The goal of image classification is to decide whether or not an image belongs to a certain category. Utilization of existing image tags along with features within an image should improve image classification results since the tags provide additional information about the image that may not otherwise be derived.

Our brains are able to process various types of information through multiple input channels and fuse this information to form associations between the separate but correlated modalities. Images are associated with captions or tags while videos contain visual and audio signals [8]. Because these distinct modalities each have different statistical properties, it is impossible to ignore that they come from separate input channels. For example, images are usually represented as real-valued outputs of feature extractors, whereas text is represented by discrete sparse word count vectors. Therefore, it is much harder to discover relationships across separate modalities than it is among features from the same modality. We propose a multimodal stacked denoising autoencoder (SDA) to learn a joint representation between images and their tags.

In the following sections, we first review the relevant literature in §2. We then present the building blocks of our model along with detailed methods in §3. Finally, we report experimental results §4 and conclude §5.

2. Previous Works

There have been several approaches to learning from multimodal data. In particular, Huiskes *et al*. [6] showed that using captions or tags in addition to standard low-level image features significantly improves classification accuracy of Support Vector Machines (SVM) and Linear Discriminant Analysis (LDA) models. A similar approach of Guillaumin *et al*. [5], based on multiple kernel learning framework, further demonstrated that an additional text modality can improve the accuracy of SVMs on various object recognition tasks. However, all of these approaches cannot make use of large amounts of unlabeled data, which is one benefit of deep learning models [3][4].

On the deep learning side, the recent approach of Ngiam *et al*. [7] uses a deep sparse autoencoder for speech and vision fusion. Most similar to our work is Srivastava *et al*. [8] which proposes a multimodal Deep Boltzmann Machine (DBM) model to learn the joint representation between images and tags. There are, however, several crucial differences. First, in this work we propose a stacked denoising autoencoder, whereas the aforementioned use a sparse autoencoder. Second, we will be performing unimodal experiments on autoencoders.

3. Methods

Autoencoders have been used effectively in modeling distributions over binary-valued data. Recent work has shown that denoising autoencoders often learn better features than regular autoencoders. In this section, we begin with a brief review of these models because they serve as the foundation of our multimodal model. Then, we propose our multimodal model.

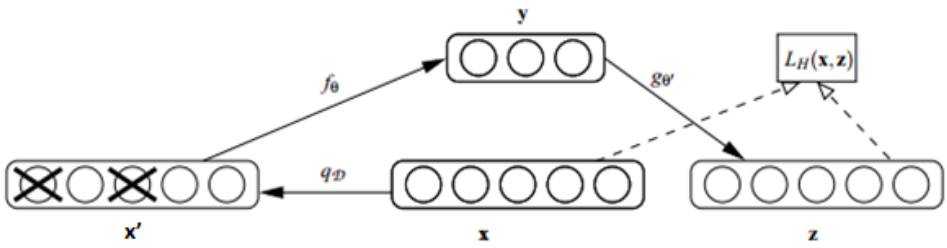

Figure 1. The denoising autoencoder architecture. An example x is corrupted via q_D to x'. The autoencoder then maps it to y (via encoder f_θ and attempts to reconstruct x via decoder $g_{\theta'}$, producing reconstruction z. Reconstruction error is measured by $L_H(x, z)$ [9]

3.1. Autoencoders

Autoencoders take an input $x \in [0, 1]^d$ and maps it using an encoder to a hidden representation $y \in [0, 1]^d$ through a mapping eq. 1

$$y = f_\theta(x) = s(Wx + b) \tag{1}$$

Where s is a non-linear function such as the sigmoid function. Y is then mapped back through a decoder into a reconstruction z eq. 2

$$z = f_{\theta'}(x) = s(W^{'} x + b) \tag{2}$$

through a similar mapping as the previous mapping. The weights, W and W', are trained to minimize the mean reconstruction error eq. 3.

$$L(x, z) = \|x - z\|^2 \tag{3}$$

Intuitively, if a representation attains a good reconstruction of its input, it has retained much of the information that was present in the original input [1].

3.2. Denoising Autoencoders

Simply retaining information about the input X is not enough to learn a useful representation. For example, in an autoencoder where the input X is the same dimensions as the output Y, the autoencoder can achieve perfect reconstruction by simply learning the identity function. Therefore, further constraints must be applied to autoencoders to force them to learn better features. This is the motivation for sparse autoencoders and denoising autoencoders (DAE).

What separates DAE from sparse autoencoders is that rather than constraining the representation, the reconstruction criterion is changed to cleaning partially corrupted input, in other words denoising [1][9]. DAE extract more robust features than autoencoders by changing the criterion to denoising. These features are also more invariant to noise in the input which makes them even more useful. However, the training of DAE is not much different than training autoencoders.

First, the initial input x is corrupted into x' by the mapping 4.

$$x' \sim q_D(x'|x) \tag{4}$$

The corrupted input x' is then mapped, as with the basic autoencoder. The corrupted input x' is mapped to the hidden representation 5

$$y = f_\theta(x') = s(Wx' + b) \tag{5}$$

from which we then reconstruct 6.

$$z = f_{\theta'} = s(W'x' + b) \tag{6}$$

The parametars θ and θ' are trained so that the average reconstruction error eq. 3 is minimized. The complete process is shown in Figure 1.

3.3. Stacked Denoising Autoencoders

Like autoencoders and RBM multiple layers of DAE can be stacked upon each other to form a SDA. Input corruption is used for the initial denoising-training of each individual layer so that it may learn useful feature extractors. However, once the mapping f_θ has been learned, it is used on uncorrupted inputs. No corruption is applied to produce the representation that will serve as input for training the next layer [1]. The next layer will then add noise and proceed as a normal denoising autoencoder. The complete procedure is shown in Figure 2.

3.4. Proposed Multimodal Model

Our proposed model Figure 3 features a separate layer of denoising autoencoders for each modality. This allows the model to learn useful features for the separate modalities before they are combined, and the joint representation is learned. By representing the data through learned first layer representations, it can be easier for the model to learn higher-order correlations across modalities [9].

Because our model has output layers that are the same dimensions as the image and text inputs, we are able to measure the reconstruction error between the input and the output of our model for each modality. We use this in several of our algoritms.

3.5. Generating Text

One initial goal of the project was being able to perform unimodal experiments on our model by filling in the missing text modality. For example, given an image we want our model to generate the missing text modality. We tested several different methods for generating tags with no success.

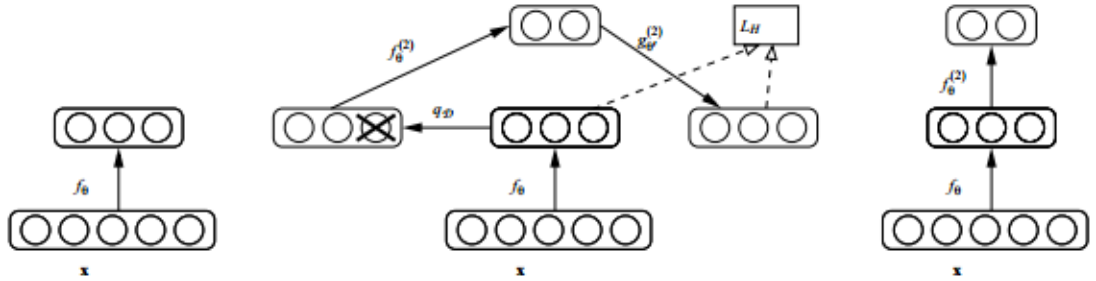

Figure 2. Stacking denoising autoencoders. After training a first level denoising autoencoder (see Figure 1) its learnt encoding function f_θ is used on clean input (left). The resulting representation is used to train a second level denoising autoencoder (middle) to learn a second level encoding function $f_\theta^{(2)}$. From there, the procedure can be repeated (right). [9]

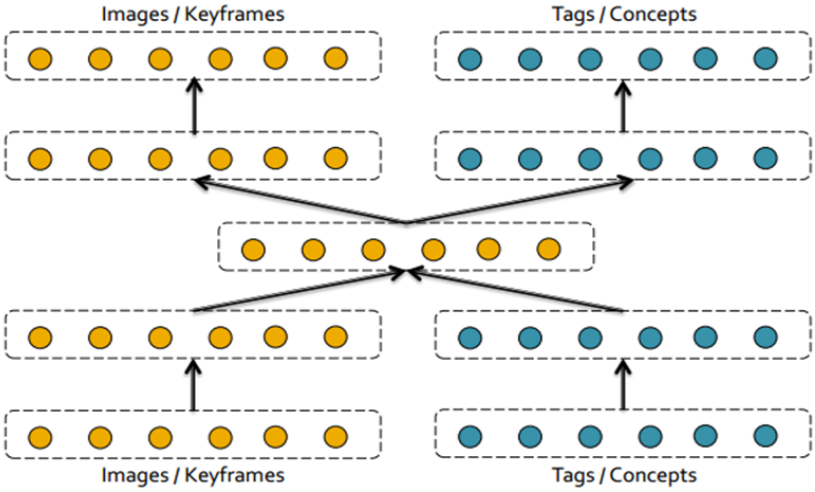

Figure 3. Multimodal Stacked Denoising Autoencoders

We will briefly explain the methods that we have tested and propose a possible solution to this problem.

Given an image, we pass the image along with the tags from the test set to our model. We measure the reconstruction error eq. 3 where x is the image input vector and z is the image output vector. We then take the k lowest scores $error$ and their tags. We then use eq. 7

$$\sum_{i=1}^{k} w_i = \frac{1}{\exp^{-error_i}} \tag{7}$$

to map the reconsturction errors to weights. Then, to get one tag we multiply the weights by their tags then sum them together to get one tag vector for the image 8.

$$tag = \sum_{i=1}^{k} w_i * tag_i \tag{8}$$

In another approach we tried, we found the k lowest reconstruction error scores tags and then took the 10 tags that overlapped the most to get one tag vector for the image. There has been recent work [2] on generative autoencoders, which could be useful for generating tags. Given more time, I think this could prove to be successful for generating tags.

4. Experiments

4.1. Dataset and Feature Extraction

The MIR Flickr Data set was used in our experiments. The data set consists of 1 million images retrieved from the photography website Flickr along with their user assigned tags; 25,000 of the images have been annotated for 24 topics including object categories such as: dog, bird, car, food and scene categories including sunset, night, and indoor. Fourteen of these classes were annotated with a category only if that category was salient in the image [6]. Together, there are a total of 38 classes where an image may belong to multiple classes. The unlabeled 975,000 images and tags were used only for pretraining. We use 15,000 images for training and 10,000 for testing, following Huiskes *et al.* [6]. Mean Average Precision (MAP) is used as the performance metric.

Tags are represented as vocabulary of the 2000 most frequent tags. The average number of tags is 5.15 [6] with a standard deviation of 5.13. Images are represented by a 3857-dimensional features vector, the feature vector consists of the following concatenated features: Pyramid Histogram of Words (PHOW) features, Gist, and MPEG-7 descriptors (EHD, HTD, CSD, CLD, SCD). PHOW features are bags of image words obtained by extracting dense SIFT

Method	MAP	Prec@50
Random	.124	.124
LDA	.492	.754
SVM	.475	.758
DBM	.609	.873
SDA (Proposed)	.533	.858

Figure 4. Multimodal Classification Results

features over multiple scales and clustering them [6].

4.2. Model Architecture and Learning

The image pathway consists of a DAE with 3857 visible units and 1024 hidden units. Similarly, the text pathway consists of 2000 visible units and 1024 hidden units. The joint layer contains 2048 hidden units. Both modalities also have output units of the same size as their inputs. The joint layer is mapped from 2048 units to 1024 hidden units for each modalities output pathway. The hidden representation is then mapped to the text and image outputs, respectively. Each layer of weights was pretrained in a greedy layer-wise training.

4.3. Multimodal Classification

We evaluate our model as a discriminative model for multimodal data. We evaluated the model by training the joint representation of the data which was then fed to a separate regression SVM for each of the 38 topics. Figure 5 summarizes the Mean Average Precision (MAP) and precision@50 (precision at top 50 predictions) obtained by the different models. The multimodal DBM does outperform our model MAP of .609, compared to .533, but our model does outperform LDA and SVM. If you compare the prec@50 of our model to the multimodal DBM the results, .858 and .873 respectively, are much closer.

One difference between our multimodal classification experiments and the ones conducted in Srivastava *et al*. [8] is that we use a regression SVM with no tuning of the paramaters to evaluate the learned joint representation discriminatively. Whereas, Srivastava *et al*. [8] performs logistic regression with finetuned parameters.

5. Conclusion

We proposed a Stacked Denoising Autoencoder for learning multimodal data representations. Large amounts of unlabeled data can be effectively utilized by the model. Pathways for each modality can be pretrained independently and concatenated for doing joint training. We initially set out to perform unimodal experiments using images without tags. For these experiments, we wanted to generate the missing modality. We tried several different methods for generating the missing modality with no success. The recent work on generative autoencoders in Bengio *et al*. [2] would be a good thing to explore applying to solving this problem.

References

[1] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, 2(1):1–127, 2009. Also published as a book. Now Publishers, 2009.

[2] Y. Bengio, L. Yao, G. Alain, and P. Vincent. Generalized denoising auto-encoders as generative models. *CoRR*, abs/1305.6663, 2013.

[3] D. Erhan, A. Courville, Y. Bengio, and P. Vincent. Why does unsupervised pre-training help deep learning? In *Proceedings of AISTATS 2010*, volume 9, pages 201–208, May 2010.

[4] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, and P. Vincent. The difficulty of training deep architectures and the effect of unsupervised pre-training. pages 153–160, Apr. 2009.

[5] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[6] M. J. Huiskes, B. Thomee, and M. S. Lew. New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative. In *Proceedings of the international conference on Multimedia information retrieval*, MIR '10, pages 527–536, New York, NY, USA, 2010. ACM.

[7] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *International Conference on Machine Learning (ICML)*, Bellevue, USA, June 2011.

[8] N. Srivastava and R. Salakhutdinov. Multimodal learning with deep boltzmann machines. In P. Bartlett, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2231–2239. 2012.

[9] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, ICML '08, pages 1096–1103, New York, NY, USA, 2008. ACM.