Unsupervised Learning of Multimodal Features: Images and Text

Maurizio Calo Caligaris
Advised by Andrew L. Maas and Andrew Y. Ng
CS229 Final Project
{maurizio,amaas,ang}@cs.stanford.edu

1 Introduction

Multimodal learning involves relating information from disparate sources. For example, Wikipedia contains text, audio and images; YouTube contains audio, video and text; and Flickr contains images and text. Our goal is to find meaningful representations of mutimodal data so as to capture as much information as possible.

Hand-engineering task-specific features for single modalities (e.g. audio or vision) is by itself a difficult task and is often very time-consuming. The challenge gets significantly pronounced when the data comes from various different modalities (e.g. images and text).

Thus, we propose an unsupervised learning model which uses images and tags from Flickr to learn joint features that model image and text correlations. Furthermore, we demonstrate cross-modality feature learning, in which better features for one modality (e.g. images) can be learned if multiple modalities (e.g. images and text) are present during feature learning time.

In the following sections, we present the network architectures we use to learn bi-modal and cross-modal features. We describe an experimental setting which demonstrates that we are indeed able to learn features that effectively capture information from different modalities and that we can further improve on computer vision features if we have other modalities (e.g text) available during feature learning time. We then conclude and offer suggestions for further work.

2 Dataset

The NUS-WIDE dataset provided by the University of Singapore[3] contains links to 269,648 images from image sharing site Flickr.com, together with their corresponding tags. [1] The NUS-WIDE dataset also provides a list of 5018 unique tags which appear more than 100 times in the dataset and also appear in the WordNet. By considering only tags present in this list, we essentially get rid of the problem that many tags in Flickr are noisy (e.g. misspelled tags or in tags in another language) or are irrelevant for the task of feature learning (e.g. proper names, model of the camera used to take picture).

3 Methodology

A key challenge in this work is to figure out a way to combine both visual and linguistic aspects in a way which allows for an autoencoder to learn meaningful representations of the data. Since the correlations between image and text data are highly non-linear, it is hard for an autoencoder or a Restricted Boltzmann Machine (RBM) to form multimodal representations of the data when fed in unprocessed text and images as input.

[1] These images were obtained by randomly crawling more than 300,000 images from Flickr's publically available API, and after removing duplicates as well as those images that contain inappropriate length-width ratios or whose sizes are too small.

Subsequently, we describe the deep learning approach we used to learn features that jointly model image and text correlations, taking as input images of variable size along with their corresponding tags.

3.1 Visual Aspect

To learn the visual features, we have used the "visual words" model often used in computer vision. In this model, we dense-sample each image to extract low-level SIFT descriptors. We then apply the k-means clustering algorithm to find cluster centroids, which forms a "codebook" or "visual words" of canonical descriptors. Finally, we use the codebook to map input patches into 1-of-K code vector ("hard assignment"), and ultimately represent each image as a "histogram" of visual words- a 1000-dimensional vector in which the *k*-th entry indicates how many times the *k*-th canonical descriptor appears in a given image. That histogram is length-normalized to take into account the variability in size of the different images.

3.2 Linguistic Aspect

We use a bag-of-words model to represent the tags associated with an image. Using the dictionary provided by the NUS-WIDE dataset, we represent text data corresponding to a each image as a 5018-dimensional binary vector whose *i*-th entry is either 0 or 1 depending on whether the *i*-th word from the dictionary belongs to the list of tags for the given image.

Since the tags are so sparse (~8-9 tags per image on average) it is difficult for an autoencoder or a Restricted Boltzmann Machine(RBM) to learn meaningful representations of the data. Thus, we map the binary valued vector of words into a more compact vector space as in [2]. More specifically, we form a vector space model which learns semantically sensitive word representations via a probabilistic model of tag co-occurrences and represents each word in the dictionary as a 20-dimensional vector. For a given document, we use as features the mean representation vector, an average of the word representations for all words appearing in the document.

3.3 Joint Feature Learning

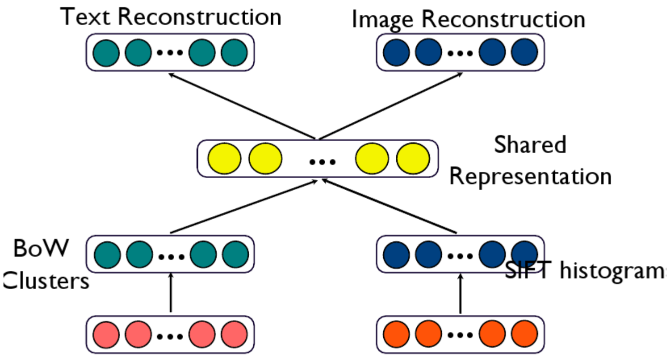


Figure 1 : Joint Feature Learning Architecture

Having learned a vector space model for the linguistic data, we concatenate the mean tag vectors representing each image to the corresponding visual histogram, and feed that as input to an autoencoder, which attempts to reconstruct both modalities. The autoencoder consists of one input layer, one over-complete hidden layer which captures cross-modal correlations between image and text (this layer is what constitutes our "joint" features), and a linear output layer which we set to be equal to the input. Given that the text features are very low-dimensional compared to the visual features, we modify the objective function of the autoencoder to account for different weighting between image and text features. The final objective function for the model is given by

$$J = \frac{1}{2}\left\|W^{(2)\mathrm{T}}h - \upsilon\right\|_2^2 + \frac{1}{2}\mu\left\|W^{(2)\mathrm{T}}h - \tau\right\|_2^2 - \frac{\lambda}{2}\left\|W^{(1)}\right\|_F^2 - \frac{\lambda}{2}\left\|W^{(2)}\right\|_F^2$$

Where h is the hidden layer representation, $W^{(1)}$ and $W^{(2)}$ are the weights going from input to hidden layer, and from hidden to the (linear) output layer, υ and τ are the visual and textual feature vectors respectively and μ is a free parameter of the model which controls the relative importance of the visual and linguistic features. The first two terms thus correspond to the reconstruction errors for the different modalities, whereas the last two terms are regularization terms which tend to decrease the Frobenius norm of the weight matrices and prevent overfitting. We run stochastic gradient descent to find the optimal weight parameters, which are used to compute the hidden activations for each example (our desired “joint” feature representation).

3.4 Cross-Modality Learning

The joint feature learning model is not very robust to missing modalities, so it can't be used in settings in which multiple modalities are available during training time but not in testing time. Thus, we propose two alternate models which improve existing computer vision features if textual information is available during feature learning time.

In the first model ("Cross-Modal I", Figure 2a), we train a network which learns to reconstruct the text features given only visual features as input. Therefore, if we only have visual input available at test time, we use the learnt weights to compute the corresponding hidden unit activations and hence obtain a multi-modal representation of the data. We also propose a similar network ("Cross-Modal II", Figure 2b) which is trained to reconstruct both modalities when given only visual data. We hypothesize that in the process of reconstructing both modalities, the network will learn a better hidden representation of the data.

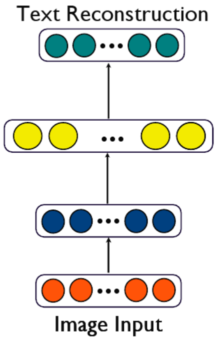


Figure 2a)

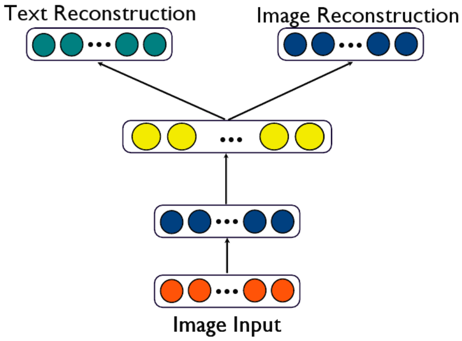


Figure 2b)

4 Experiments

We evaluate the performance of the individual components of the system on a tag-suggestion task. The NUS-WIDE dataset provides ground-truth for 81 "concepts" or "tags", which mainly correspond to the most frequently occurring tags on Flickr and are consistent with concepts commonly used in image categorization (see Table 1). For each of the 269,648 images, the dataset provides a binary-valued 81 dimensional vector, in which its *i*-th entry indicates whether the *i*-th concept corresponds to the image (which does not necessarily mean that the *i*-th concept appears in Flickr as a tag for the given image). The fact that the ground-truths for the 81 concepts were manually annotated circumvents the problem that tags in Flickr are generally incomplete and thus allows us to execute supervised learning experiments. To evaluate our system in reasonable time, we only consider the 10 most frequently occurring concepts for our experiments.

Concept	**Occurrence (%)**
Sky	7.5
Water	7.1
Clouds	5.8
Sunset	4.3
Beach	3.2
Tree	2.9
Reflection	2.7
Animal	2.7
Street	2.6
Sun	2.6

For each of the 10 most frequently occurring concepts in the dataset, we remove the corresponding tag from the dictionary (otherwise the task is trivial for components which take inputs relating to text) and train our models (joint and cross-modal) in an unsupervised fashion to learn an appropriate feature representation for the data. We use the ground-truths for these labels to separately train L2-regularized logistic regression classifiers. Each of these tag detectors learns whether a specific concept corresponds to a given image. We use a training set size of 50000 examples and test on another 50000 examples, using the Area Under the (ROC) Curve (AUC) metric to evaluate performance of each component of the system on each category.

We PCA whiten the visual input (histogram) to 200 dimensions (which we have found is enough) and normalize the tag vector representations so as to have unit variance and zero mean. We train the networks using 1.4x over-complete hidden layer representation and using a weighting parameter μ=10 to control the relative importance given to the textual features (for joint network and for cross-modal II). The parameter μ was chosen over a small grid search of parameters to find the one with best performance on the training set.

4 Results

Feature Representation	**Mean AUC**
Raw Tags (Bag of Words)	0.850
SIFT	0.755
Raw Tags+SIFT	0.824
Semantic Word Vectors	0.821
LSA	0.803
Joint Features	**0.861**
Cross-Modal I	0.762
Cross-Modal II	**0.773**

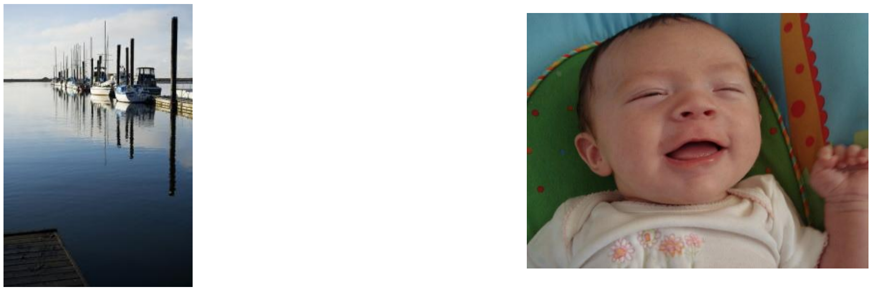
Table 2: Performance of individual components and combination of components of our system on the tag-suggestion task.

Table 2 shows the AUC scores of the different components averaged over the 10 concepts. It turns out that linguistic information is generally more useful for this task than the visual features (although SIFT does better than text on suggesting 'clouds' for instance). When simply concatenating SIFT histograms to the textual features, the performance decreases because of the poorer visual features.

The non-linearity introduced by the transformation of the tags into a more compact feature space makes it harder for a linear classifier to perform well in the task of tag suggestion. However, our semantic word vector representation still does better than other methods such as Latent Semantic Analysis (LSA) and is what ultimately allows us to train autoencoder models that outperform any other part of the system. In particular, the joint feature learning model achieves the best results on this task by taking into account all of the information available-visual and semantic-and combining them to form meaningful representations of the data. Moreover, both cross-modal networks outperform SIFT, which shows that better features for computer vision can be learned if semantic information is available during training.

4.3 Visualizations of Learned Features

To get an idea of how well the cross-modal system is doing at reconstructing the input when given only image features, we looked at a few examples in the test set (for which the system was only given access to the SIFT features but not to its corresponding tags) and computed the mean tag vector reconstructions by the cross-modal II system. By the way the system is built, it is not possible to recover the actual tags, but by looking at the tags closest in the vector space model to the reconstruction vector we can qualitatively see that the system is indeed doing a reasonable job at reconstructing a semantic representation of the original input.



Flickr Tags: January, sailboats, surrey.
Reconstructed: Water, rocks, sailboat, sea, agua, pier, fishing, canoe, creek, seascape.

Flickr Tags: Child, baby, infant, newborn
Reconstructed: Adorable, sweater, hair, expression, smiling, fingers, look, playful, fluffy, cute.

Figure 3: Visualizations of Cross-Modal II reconstructions.

4 Conclusions and Further Work

We have shown that our system effectively relates information from disparate data sources by learning meaningful representations that capture correlations across different modalities. We envision this work may have practical applications, such as in image retrieval or in the organization of large personal photo albums, given that the system can be used to automatically suggest tags, categorize images and find visually and semantically similar images (and as a matter of fact we already have).

Moreover, this work can have a significant impact in the area of computer vision research. As a next step, we would like to evaluate performance on a standardized dataset. In particular, we may use the same approach used to improve SIFT features to improve the currently best performing image features on a dataset such as ImageNet, and use those learnt features to beat the current state-of-the-art in ImageNet.

Acknowledgments

I would like to thank the University of Singapore for providing the NUS-WIDE Flickr dataset. I would further like to thank Andrew Maas and Prof. Andrew Ng for their invaluable guidance on this project, as well as Jiquan Ngiam for his very helpful collaboration.

References

[1] K. E. A. van de Sande, T. Gevers & C. G. M. Snoek.(2010). Evaluating color descriptors for object and scenerecognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (in press).

[2]A. L. Maas & A. Y. Ng. (2010). A Probabilistic Model for Semantic Word Vectors. *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning.*

[3] T.S Chua, J. Tang, R. Hong, H. Li, Z. Luo, & Yan-Tao Zheng.(2009). NUS-WIDE: A Real-World Web Image Database from National University of Singapore. *ACM International Conference on Image and Video Retrieval*.

[4]J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee & A. Y. Ng. (2010). Multimodal Deep Learning. *NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning.*

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