CS231A Course Project Proposal
Unsupervised Multi-Modal Feature Learning: Images and Text

Maurizio Calo Caligaris
Stanford University
maurizio@cs.stanford.edu

Mirek Kukla
Stanford University
mkukla@stanford.edu

Abstract

Hand-engineering task-specific features for single modalities (e.g. vision) is a difficult and time-consuming task. Furthermore, the challenge gets significantly more pronounced when the data comes from multiple sources (e.g. images and text). In this work, we seek to leverage freely available images on the web along with nearby text to create meaningful feature representations that capture both visual and sematic information. Our hypothesis is that these learnt features can then be used to improve on many different computer vision features on a wide variety of computer vision tasks.

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. To maximize performance on specific tasks, we would like to use all of the information available to us. It is not clear how to hand-engineer features that capture information from different modalities; thus, we adopt an unsupervised feature learning approach that learns joint correlations between different modalities-in particular, between images and text- using deep neural networks [2].

Specifically, we would like to leverage vast amounts of freely available data on the web - images along corresponding text captions that appears nearby- to create meaningful feature representations of multimodal data and ultimately improve on existing computer vision features.

In this particular project, we learn meaningful feature representations from images and tags crawled from Flickr. Our framework can be summarized as follows:

1. We are given a computer vision feature (e.g. HOG) along with a computer vision task (e.g. scene classification).
2. We train a neural network that learns to predict Flickr tags from the given computer vision features applied to the corresponding Flickr images.
3. The hidden layer constitutes the learned joint features that captures image/text correlations.
4. We then forward propagate the network using as input the images from the dataset to obtain a new set of features for the specified task.

Our hypothesis is that the learnt features (possibly concatenated with the original features) will improve the performance of the original features . Since this approach does not take into account any task-specific knowledge, and the text data is available at no extra cost, we envision this approach to work with a number of computer vision features for a wide variety computer vision tasks.

2. Experimentation

We specifically evaluate the performance of our method in the SUN scene classification task[1]. Torralba *et al*. have put up together a large-database of images containing 397 scene categories to evaluate numerous state-of-the-art algorithms for scene recognition. Our hope is to improve on the best performing algorithm for the task (In this case, HOG 2x2 features).

They provide code to compute a number of different image descriptors, such as color histograms, GIST, SIFT and HOG. We have already been able to reproduce the results they report - that is, scene classification accuracies for different . In particular, we expect to improve on their results when n, the number of training samples per class, is small ($n = 1$ or $n = 5$, say).

References

[1] J. Xiao, J. Hays, K. Ehinger, A. Oliva, A. Torralba SUN Database: Large-scale Scene Recognition from Abbey to Zoo *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[2] G. Hinton and R. Salakhutdinov *Science*, 313(5876):504–507, 2006.

3. Appendix

This project is part of a larger reseasrch project with Prof. Andrew Ng. The plan is to submit a paper to CVPR.