# A Discriminative Latent Model of Object Classes and Attributes

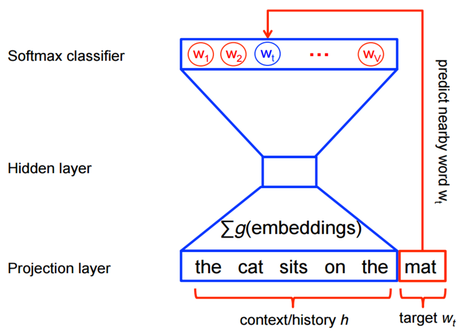
Yang Wang and Greg Mori
Vision and Media Lab, Simon Fraser University, Burnaby, BC, Canada

SFU Vision and Media Lab

## Overview

- Joint modeling of object classes and (correlated) attributes
- Use of word embedding to train Keras layers
- A general learning framework for classification with auxiliary labels

## Model Formulation




**Training data**: IMDB movie review. Labeled as negative and positive

**Training algorithm for embedding** : Hierarchical softmax

**Training algorithm for Keras layers** : softmax with stochastic gradient descent

## Learning and Inference

**Scoring**: score(w,h) computes the compatibility of word w with the context h

$$\begin{aligned}\mathbf{P}(w_t|h) &= softmax(score(w_t,h)) \\ &= \frac{exp(socre(w_t,h)}{\sum_{w' in Vocab} exp(score(w',h))}\end{aligned}$$

**Learning with latent attributes**:

$$\begin{aligned}&\min_{\mathbf{w},\xi} \beta||\mathbf{w}||^2 + \sum_{n=1}^{N} \xi^{(n)} \\ &\text{s.t.} \max_{\mathbf{h}} \mathbf{w}^\top \Phi(\mathbf{x}^{(n)}, \mathbf{h}, y^{(n)}) - \max_{\mathbf{h}} \mathbf{w}^\top \Phi(\mathbf{x}^{(n)}, \mathbf{h}, y) \\ &\quad \geq \Delta(y, y^{(n)}) - \xi^{(n)}, \forall n, \forall y\end{aligned}$$

Another choice is to use the ground-truth attribute labels $\mathbf{h}^n$ (i.e. learning with observed attributes).

### Attribute Relation Graph

Running minimum spanning tree with $\mathrm{NormMI}(j,k)$ as the weight on the edge $(j,k)$.

### Other Loss Functions

A simple modification of $\Delta$ will optimize different (training) errors.

**Overall accuracy:**

$$\Delta_{0/1}(y, y^{(n)}) = \begin{cases} 1 \text{ if } y \neq y^{(n)} \\ 0 \text{ otherwise} \end{cases}$$

**Mean per-class accuracy:**

$$\Delta_{\text{new}}(y, y^{(n)}) = \begin{cases} \frac{1}{m_p} \text{ if } y \neq y^{(n)} \text{ and } y^{(n)} = p \\ 0 \quad \text{otherwise} \end{cases}$$

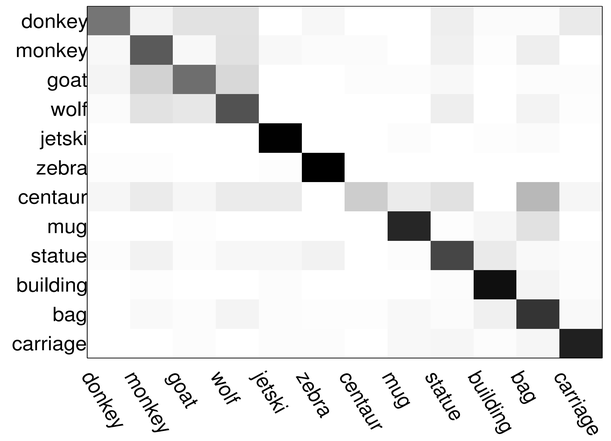
where $m_p$ is the number of training examples with class $p$.

## Experiments

**a-Pascal dataset:**

| method | overall | mean per-class |
|---|---|---|
| Our approach with $\Delta_{0/1}$ | **62.16** | **46.25** |
| Our approach with $\Delta_{\text{new}}$ | **59.15** | **50.84** |
| SVM with $\Delta_{0/1}$ | 58.77 | 38.52 |
| SVM with $\Delta_{\text{new}}$ | 53.74 | 44.04 |
| Farhadi et al. CVPR09 (base features+SVM) | 58.5 | 34.3 |
| Farhadi et al. CVPR09 (best result) | 59.4 | 37.7 |

**a-Yahoo dataset:**




| method | overall | mean per-class |
|---|---|---|
| Our approach with $\Delta_{0/1}$ | **78.67** | **71.45** |
| Our approach with $\Delta_{\text{new}}$ | **79.88** | **73.31** |
| SVM with $\Delta_{0/1}$ | 74.43 | 65.96 |
| SVM with $\Delta_{\text{new}}$ | 74.51 | 66.74 |

The 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 2010