# Deep Learning Workshop
# EEL6935 Course Project – Sentiment Analysis

Caleb Bryant, Jixin Feng

University of Florida

## Objectives

- Study machine learning tools for natural language processing (NLP)
- Apply text pre-processing and encoding scheme
- Implement sentiment analysis system using both Logistic Regression and LSTM
- Build a operational web app for real time sentiment analysis

## Introduction

The volume of text on the Internet – unstructured text especially – is increasing with drastic speed everyday. Unlike human brains, traditional computer programs have a much more limited ability to extract useful information from unstructured text with satisfactory precision. While traditional machine learning methods have had some success tackling NLP problems with "bag of words" models and feature engineering, deep learning and the subsequent development of robust word embeddings have shown promising results and have made substantial ground towards replacing older methods. In this course project for EEL6935 Big Data Ecosystems, we implemented a sentence classification program for sentiment analysis based on Long-Short Term Memory Network (LSTM) and compare the performance with a logistic regression based baseline model. In addition, we also created a web application able to do real-time sentiment analysis based on the model we created.

## Dataset

The dataset we used to train the system is Stanford Large Movie Review Dataset

- Consists of 50,000 polarized movie reviews and corresponding scores from IMDb
- 25,000 reviews are designated for training
- 25,000 reviews are designated for testing
- Each movie ratin between 1 and 10
- Reviews with a rating of 5 or 6 were excluded

## Developing Environment

The whole project was coded in Python 3 with

- Ubuntu 14.04 LTS
- Scikit-Learn & Tensorflow
- Flask

The computer used for training has:

- 3.1GHz Intel Core i7 CPU with 8MB cache
- 16GB RAM
- nVidia GTX 1050 GPU with 4GB RAM

## Model: LSTM (Cont'd)

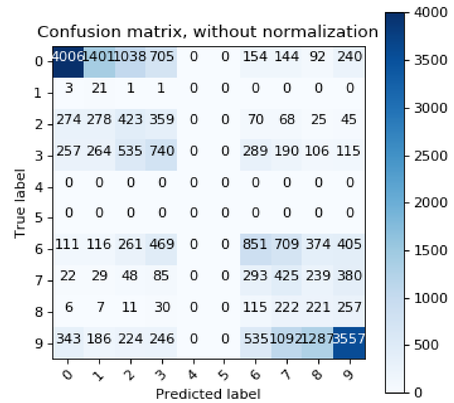
Given a word sequence $S = \{v_0, v_1, \ldots, v_{l-1}\}$ with length $l$, the states of LSTM are updated as:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ c_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} S[h_{t-1}, x_t]$$

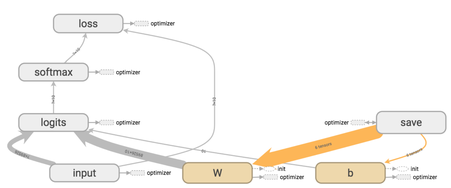
## Sentiment Analysis Result

| Method | Epochs | Binomial Training | Binomial Testing | Multinomial Training | Multinomial Testing |
|---|---|---|---|---|---|
| scikit-learn LR | N/A | 0.9981 | **0.8697** | N/A | N/A |
| tensorflow LR | 20 | 0.8670 | 0.8583 | 0.9982 | 0.3734 |
| larger LSTM | 8 | N/A | N/A | 0.6693 | 0.3657 |
| smaller LSTM | 2 | N/A | 0.8507 | 0.5622 | **0.4098** |

## Result

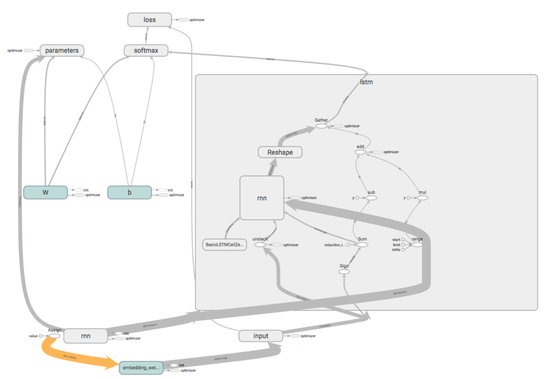



## Model: Logistic Regression




The softmax function maps each dimension of its input to a value between 0 and 1, and the transformation also normalizes the predicted probabilities such that they sum to 1 for multi-class prediction.

## Model: LSTM




## Learn More

Project repository hosted on GitHub (`https://github.com/ufjfeng/EEL6935-Course-Project`). Web app hosted on `http://t21.ecegator.com`. Video demo is hosted on YouTube (`https://youtu.be/MRpfsmDhVII`) All documents including this poster are written in LaTeX.

Code Repo  Web App  Video Demo