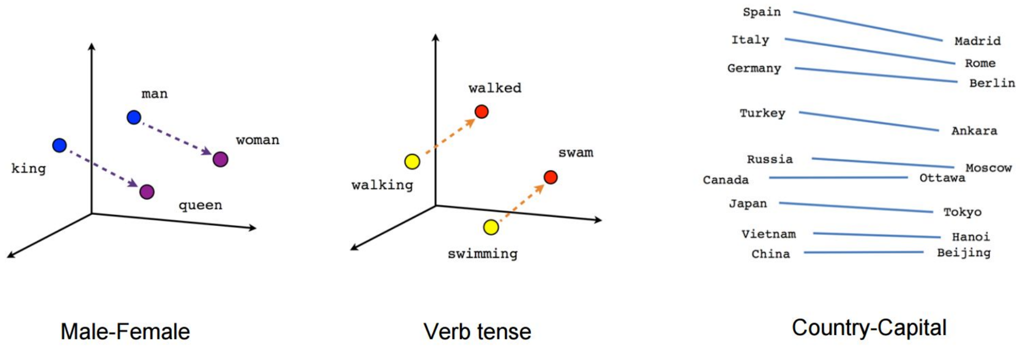
Christopher Rico
MSDS 422
Assignment 8
11/17/19

Introduction

For companies who want to automatically classify customer reviews and call and complaint logs, natural language processing (NLP) is an immeasurably beneficial application of machine learning techniques. The ability to automatically assess the criticality and sentiment of customer interactions can assist in the identification of cases that require rapid response by support personnel, allowing for increased customer satisfaction.

Recurrent neural networks (RNN), considered to be the state of the art in NLP, are a cutting-edge machine learning approach modeled after biological brain architecture that allow for a neural network to have a certain amount of self-feedback and even long-term memory. Often, these models must be coupled with pre-trained word embeddings that map a plain text word to a vector, allowing for an RNN to manipulate and do work on it. There are many such embeddings, each trained with a different number of parameters per word, and each being trained on a different source text.

In furtherance of exploring the most accurate model configuration for NLP modeling , several RNN modeling topologies were trained with two different Google-built word embeddings, and tested to predict sentiment on the IMDB review dataset.



Male-Female

Verb tense

Country-Capital

Example of word embeddings showing the way language relationships can be computed.

Research Design

The IMDB review dataset is a well-known image dataset that consists of 50,000 plain-text movie reviews, broken into a 25,000 review training and 25,000 review test set. Each review is given a label of either 'positive' or ‘negative’. Both test accuracy and processing time were considered when evaluating models, as speed and accuracy are both important factors to be considered when choosing a model.

Programming Work

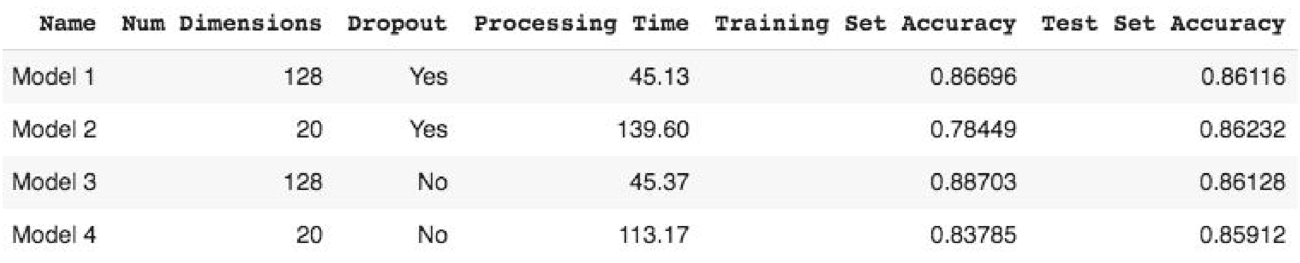
Model construction and evaluation was completed entirely in the cloud, using Google Colaboratory as a development environment and TensorFlow as a RNN modeling environment. TensorFlow is an industry standard neural net library built by Google, with the capacity to model recurrent neural nets.

To begin, four different RNN models were instantiaed. The configuration combinations tested were combinations of two different Google text embeddings – 20 dimensions or 128 dimensions – and a dropout rate of 0.2 or 0. Each RNN was trained on the 25,000 review training set and tested on the 25,000 review test set. Then, training and test set accuracy were reported.

Results and Recommendation

Both training and test set accuracy results are fairly consistent across embedding dimensions and model hyperparameter combinations. Training time increases greatly as the dimensionality of the embedding layer decreases.

It is recommended to use an embedding layer with 128 dimensions, and a layer of dropout at a rate of 0.5 for NLP applications. That said, it would be prudent to explore further model configurations and methods of training set augmentations, as there may be a more optimal configuration not examined.



Name	Num Dimensions	Dropout	Processing Time	Training Set Accuracy	Test Set Accuracy
Model 1	128	Yes	45.13	0.86696	0.86116
Model 2	20	Yes	139.60	0.78449	0.86232
Model 3	128	No	45.37	0.88703	0.86128
Model 4	20	No	113.17	0.83785	0.85912

A shared version of the interactive Jupyter notebook used to run this analysis can be found at:
https://colab.research.google.com/drive/1QMfquzmgL3BF8Tipj4qNF_PRZKlhdZpI