Kernel Sanders: T1-10: Text Understanding from Scratch - Executive Report

1st Kyle Levy
260604024
Montreal, Canada
kyle.levy@mail.mcgill.ca

2nd Nick Nasirpour
260561658
Montreal, Canada
nick.nasirpour@mail.mcgill.ca

3rd Phil Ryjanovsky
260612028
Montreal, Canada
philip.ryjanovsky@mail.mcgill.ca

The purpose of this project was to test the accuracy performances of our hyperparameter tuned simple classifiers versus the authors' complex convolutional neural networks. We attempted to make these classifications by using methods such as bag-of-words and bag-of-centroids to construct sparse matrices that contained information about word usage in certain datasets. We made no assumptions about the grammar or syntax of the underlying language, such as word placement. One of the main strategies that further increased performance by a range of 1%-2% was the exclusion of certain top 'common' words that have no descriptive meaning. While it served as a slight dimensionality reduction, this elimination was primarily used to disallow these uninformative words from affecting the classification decision.

In our pursuit to beat the authors' convolutional neural networks' performances, we implemented three simple classifiers: naive Bayes, logisitic regression, and linear SVM. Linear SVM and logistic regression used two variations of gradient descent: batch gradient descent and stochastic gradient descent. Stochastic gradient descent proved to be extremely important in executing these classification algorithms in a reasonable amount of time. Random forests which is an ensemble method (essentially bagging with decision trees), however, was used extensively in order to greatly reduce the variance of each learned model. We found that random forests were effective in learning the models of the data sets presented in *Text Understanding from Scratch* by Zhang and LeCun, but still beat by linear SVM. While in most cases we could not beat the authors' small and large CNNs, our simple classifiers made classifications in substantially less time than their complex models, with accuracy performances near, but slightly less than what they had achieved. The following tables show our best simple classifer as well as the results obtained from Zhang's and LeCun's CNNs as well as their baseline model which was logistic regression.

Accuracy Received for Optimum Simple Classifiers AG News Corpus		
Classifier	Train Accuracy	Test Accuracy
Linear SVM	92.1%	**90.4%**
Large CNN	99.4%	**87.2%**
Small CNN	99.2%	84.4%
Baseline	88.02%	86.69%

Accuracy Received for Optimum Simple Classifiers DBPedia		
Classifier	Train Accuracy	Test Accuracy
Linear SVM	98.2%	**97.8%**
Large CNN	99.9%	**98.3%**
Small CNN	99.4%	98.0%
Baseline	96.3%	96.2%
Paper Baseline(w2v)	89.3%	89.1%
Random Forest(w2v)	99.9%	94.5%

Accuracy Received for Optimum Simple Classifiers Yahoo! Answers		
Classifier	Train Accuracy	Test Accuracy
Linear SVM	68.6%	**68.0%**
Large CNN	73.4%	**70.4%**
Small CNN	72.8%	70.2%
Baseline	66.8%	66.6%
Paper Baseline(w2v)	56.37%	56.47%
Random Forest(w2v)	99.9%	61.7%

Test Accuracy Received for Optimum Simple Classifiers Amazon		
Classifier	Full data set	Polarity data set
Logistic Regression	**54.0%**	90.0%
Linear SVM	53.6%	**90.2%**
Large CNN	58.7%	94.5%
Small CNN	**59.5%**	**94.5%**
Baseline	54.17%	89.9%

Accuracy Received for Optimum Simple Classifiers Sogou News		
Classifier	Train Accuracy	Test Accuracy
Logistic Regression(SGD)	89.8%	**89.5%**
Large CNN	**99.1%**	**95.1%**
Small CNN	93.1%	91.4%
Baseline	93.0%	92.8%

It is essential to consider the task at hand when deciding between complex models and simple classifiers. Considerations include time constraints, computation resources, and size of the set. As we saw with our experiments, smaller data sets worked better with tuned simple classifiers. As the size of the data set increased, more complex models like the authors' CNNs outperformed the simpler classifiers. With regards to time constraints simple classifiers proved to obtain accuracy within a few percent of the complex model's accuracy, but ran extremely fast in comparison.