LSTM and CNN Models for News Classification Using Word2Vec Embedding

Eren Aldis, Elif Bilgin, Juhi Malani

Department of Computer Science, Johns Hopkins University

600.682 Machine Learning: Deep Learning
Fall 2018

Intuition and Hypotheses

Classifying news headlines into four categories: Business, Health, Science and Technology and Entertainment with a Kaggle dataset 400,000 entries

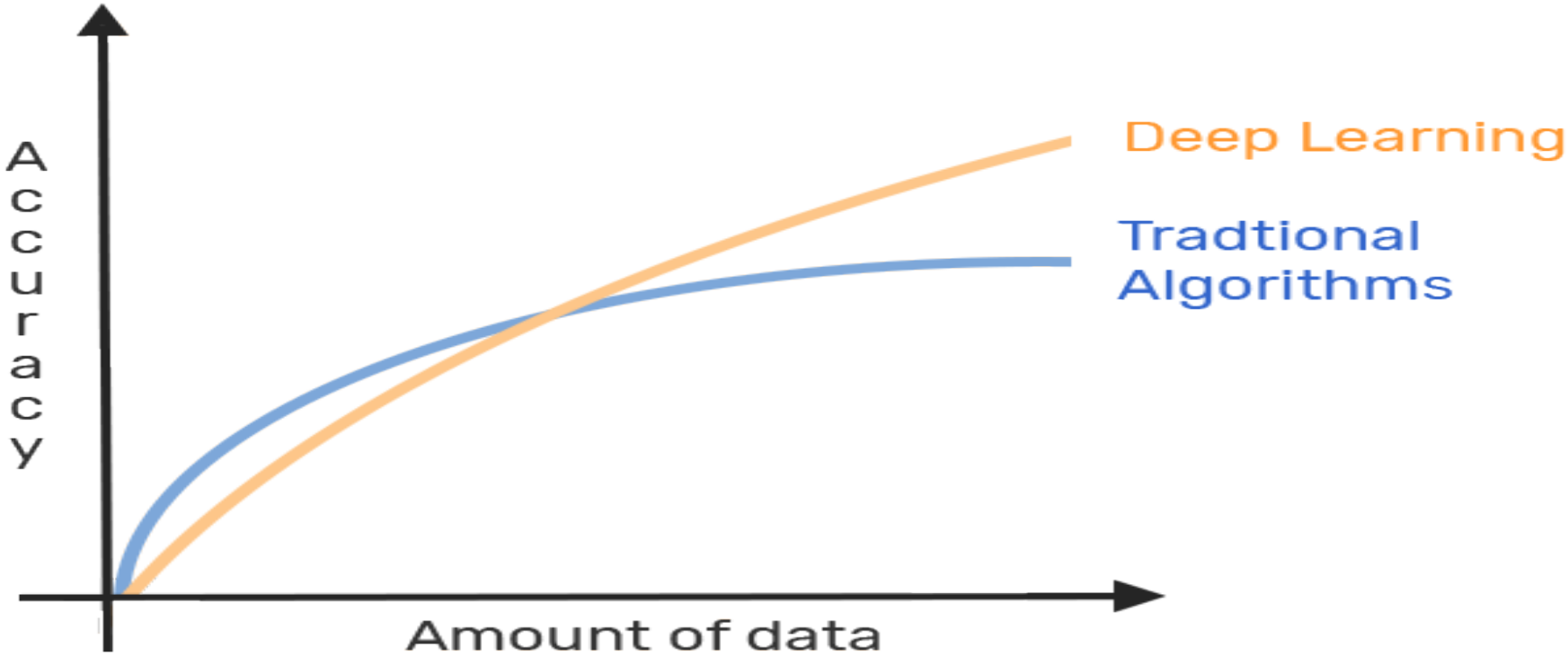

Figure: DL vs ML [**1**]

Deep Learning Methods Used

- Logistic Regression
  - Maximum accuracy; baseline
- CNN
  - Single layer CNN followed by a softmax activation
  - Triple layer CNN
- LSTM
  - Single layer LSTM with a softmax activation
- Hybrid
  - CNN layer followed by a LSTM layer with softmax activation
  - CNN using Word2Vec pre-processing
  - LSTM using Word2Vec pre-processing

S

Result

Cross validation was used for a test set to give the following result

Model	Acc	p-value
Logistic Regression	94.83	-
CNN(1 layer)	94.98	< 0.10
CNN(3 layers)	87.19	$< 10^6$
LSTM	95.02	< 0.02
CNN+LSTM	90.18	$< 10^{-6}$
CNN + Word2Vec	95.01	< 0.02
LSTM + Word2Vec	95.24	< 0.01

Limitations

- Multiple categories
- Categories are subjective
- Artificial limit

Scope

- Tag and label for multiple categories [2]
- Translation and NLP
- Deeper networks have better acurracy

References

[1] Text Classification
https://monkeylearn.com/text-classification/

[2] M. I. Rana and S. Khalid and M. U. Akbar
"News classification based on their headlines: A review"