Final Group Project Proposal

Group: #6

Team members: KJ, Nayan, Neelanshi and Zane

Problem Statement:

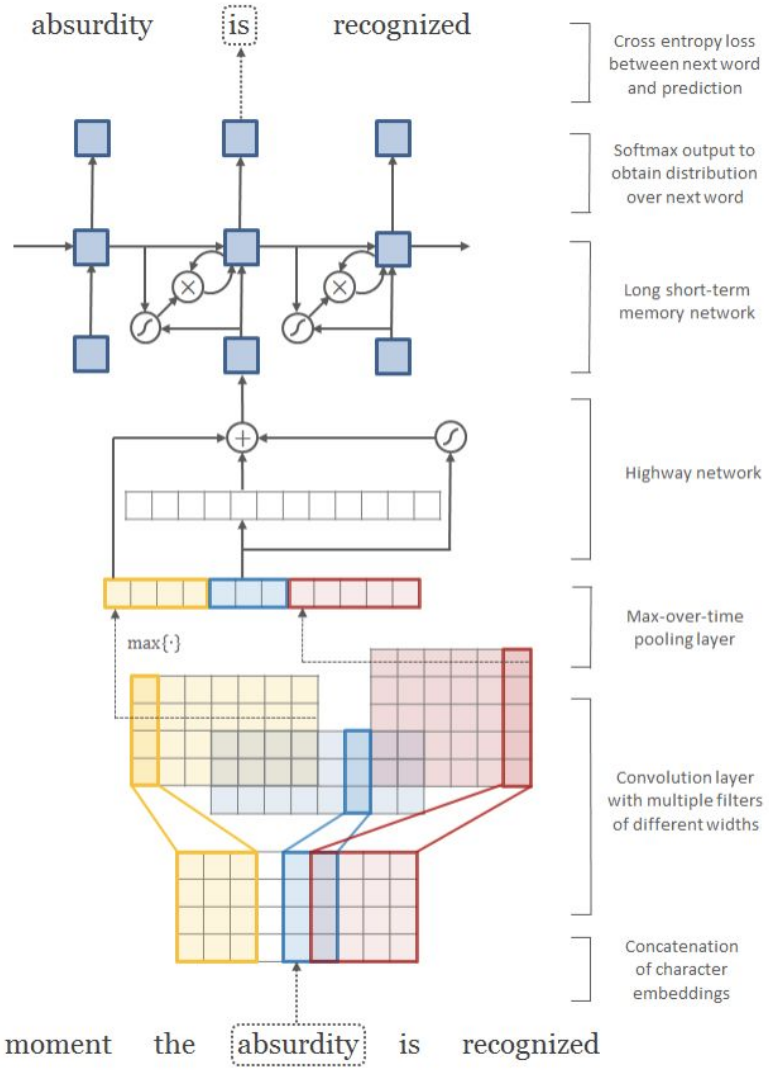
We propose using a Character-level Convolutional Neural Network and it's variants on large versus small datasets and comparing performance. Particularly, we would like to see how highway layers affect the model's performance on large datasets. Yoon Kim et al.'s paper Character-Aware Neural Language Models compares word-level CNNs to character-level CNNs using the Penn Treebank dataset. We would like to compare their char-CNN results on this dataset to our results of a char-CNN on a larger dataset. The paper also lightly investigates the use of highway layers. The paper found that having one to two highway layers was important, but more highway layers generally resulted in similar performance; and having more convolutional layers before max-pooling did not help. However, they note that this may depend on the size of the dataset. We will change the architecture of the model using various amounts and variants of highway layers, dropouts and CNNs. And observe how this affects the model's performance on a larger dataset.

Motivation / Rationale:

In the different papers presented over the past two weeks, we saw different patterns of how models performed on language modeling tasks and wanted to learn and investigate more on the topic. Character level models are particularly efficient because they reduce the embeddings size a large factor since we consider 512 characters at most in the English language and maybe a few more for other languages, whereas for word level embeddings the size is in tens of thousands - requiring a lot more compute power and time. But at the same time there's also the problem of learning context, since for the same context length the amount of information for a word-level embedding is much more than a character level embedding. So we want to know what factors - size of dataset, dropouts, highway layers, etc affect the performance for learning the context and why. One can then take advantage of that particularly important component to drive the future research.

This paper, particularly with its highway layers, is able to have a better perplexity than SOA but it does not quite explain why and how. The paper found that having one to two highway layers was important, but more highway layers generally resulted in similar performance; and having more convolutional layers before max-pooling did not help. However, they note that this may depend on the size of the dataset. The paper does not talk a lot about how dropouts and their selection of CNN affects the performance, which we would further like to dig deep into.

Model Description:



The paper uses a combination of character level LSTM, CNN and Highway layers which altogether results in a better perplexity. It would be interesting to see which component is the most effective for modeling the dataset and how much percentage effect they have. We will be using different variants of this model architecture - modifications to the Highway Network, dropouts, CNN layers and see how the model performs.

Task Description:

Proposed dataset: One of the following:

- A larger Treebank dataset from http://www.nltk.org/nltk_data/,

- Brown Corpus
- 1B word dataset*

Our task has two parts:

1. Adding more highway layers and tweaking the transform gates on them
2. Experimenting with dropouts, CNN Layers (mainly adding and removing them) to see their effect on context learning
3. Comparing the char-CNN's performance on a larger dataset to its performance on a small dataset

The paper uses the Penn Treebank dataset with the char-CNN model. We would like to use the larger Brown Corpus dataset with a char-CNN model and compare our model's performance to that of the paper's.

Evaluation Metric:

Our evaluation metric is Perplexity. Perplexity is a commonly used evaluation in NLP models and is calculated by:

$$perplexity(LM) = \exp\left(-\frac{1}{N}\sum_{n=0}^{N}\ln(p_n)\right)$$

Benchmark Description:

Our benchmarks for comparison are from Character-Aware Neural Language Models. We will be specifically comparing our results to the LSTM-Char models in this table.

	PPL	Size
LSTM-Word-Small	97.6	5 m
LSTM-Char-Small	92.3	5 m
LSTM-Word-Large	85.4	20 m
LSTM-Char-Large	78.9	19 m
KN-5 (Mikolov et al. 2012)	141.2	2 m
RNN† (Mikolov et al. 2012)	124.7	6 m
RNN-LDA† (Mikolov et al. 2012)	113.7	7 m
genCNN† (Wang et al. 2015)	116.4	8 m
FOFE-FNNLM† (Zhang et al. 2015)	108.0	6 m
Deep RNN (Pascanu et al. 2013)	107.5	6 m
Sum-Prod Net† (Cheng et al. 2014)	100.0	5 m
LSTM-1† (Zaremba et al. 2014)	82.7	20 m
LSTM-2† (Zaremba et al. 2014)	78.4	52 m

We will also be comparing our highway network experiments to this table:

	LSTM-Char	
	Small	Large
No Highway Layers	100.3	84.6
One Highway Layer	92.3	79.7
Two Highway Layers	90.1	78.9
One MLP Layer	111.2	92.6

Table 7: Perplexity on the Penn Treebank for small/large models trained with/without highway layers.

We would love to hear your inputs and suggestions that you may have on this topic. Thank you!

References:

[1] Kim, Yoon, et al. "Character-aware neural language models." Thirtieth AAAI Conference on Artificial Intelligence. 2016.