Namdar Kabolinejad COOMP 550 - Reading Assignment 1
260893536

This paper gives an empirical analysis of character-level CNN for text classification. The model consists of a temporal 1-D CNN taking a "one-hot" sequence of character encoding. The main advantage of this approach in text classification is that it doesn't require knowledge of the syntactic or semantic structure of the language and works with characters. The authors offer a comprehensive comparison between their model and traditional (BoW, n-grams) and deep learning (LSTM) word-level approaches. The authors have cultivated several large-scale datasets containing several hundred thousand samples. Apart from these, the experiments contain several different variants of the proposed model, for example, the use of thesaurus augmentation, all lower-case chars, large vs small CNNs. The final results are encouraging as they show that character-based CNN is an effective method for text classification, however, the performance depends on several factors one of which is the scale of the datasets. One problem with the paper is that it doesn't provide any benchmarks from other literature that could help support the results obtained.

There are several steps taken that help with reducing overfitting and handling OOVs:
Firstly, given the character-level approach of the CNN model, the model doesn't require knowledge of the words or the language. This focus on working only at the character level rather than words means that words from different languages, misspellings, typos, emoticons, … that may result in OOV items can be naturally learned.
Furthermore, the authors are using thesaurus data augmentation where they randomly replace a certain number of words with their synonyms. This step will add extra words to the training data that are likely to show up in the test corpus, thus reducing overfitting, and reducing the number of OOVs the model will encounter. These approaches are a bit different from trying to mathematically smooth the probability distributions as we have done in class.
Another detail in the paper that might help with reducing OOVs is the offset constant ($c = k - d + 1$) added to the temporal convolutional module. Although I was not able to make complete sense of the math, it seems that the offset constant is similar to add-δ smoothing introduced in class.

It isn't a surprise that one model doesn't work well on all datasets. This is because there are many different aspects to the approach each model takes in text classification. For example, word vs char level analysis, how the features are selected, etc. All these aspects affect how the model processes the corpus data and how it ultimately performs. This means that depending on the characteristics of the dataset - for example, size, how the data is curated, possible misspellings, … - each model will have a different performance.

It can be seen that in 4 of the 8 data sets the character-based CNN model with data augmentation outperforms word-based models. As expected the proposed model that uses a CNN works better on larger datasets (scale of million of samples) that might contain noisier content (e.g. misspellings, abbreviations, ...), that is the model performs better on crowd-sourced reviews and answers rather than professionally published articles. On the other hand, the better performance of n-grams on smaller and better-structured datasets, e.g. new articles, can be justified. The bad performance on all datasets of Bag-of-means is not entirely logical, the performance might be due to the parameters of the model for example the dimension of the embedding or number of means, and might improve by changing these parameters.
Unfortunately, the paper doesn't provide enough validation to be able to make concise conclusions about the results, the points above are merely observed from the error rates provided.