Implementation of Character-level Convolutional Networks for Text Classification

Paper ID: 113
Ateeksha Mittal: 2017A8PS0431P
Dishita Malav: 2017A7PS0164P
Shefali Tripathi: 2017A7PS0139P

About the Paper

- Offers an empirical exploration on the use of **character-level convolutional networks** (ConvNets) for text classification.
- Author constructed several large-scale datasets to show that character-level convolutional networks could achieve **state-of-the-art** or **competitive results.**
- Offered **comparisons** against traditional models such as bag of words, n-grams and their TFIDF variants, and deep learning models such as **word-based ConvNets** and recurrent neural networks.

Our Problem Statement

- Implement **Character-level Convolutional Network** for Text Classification on the AG News dataset using PyTorch.
- Compare the results obtained with that of **Word-Based Convolutional Network** for Text Classification on the same dataset.

Deeper into the paper

- To date, almost all techniques of text classification are based on **words**, in which simple statistics of some ordered word combinations (such as n-grams) usually perform the best
- **This work is the first to apply ConvNets only on "Characters".**

Why characters?

- ConvNets do not require the knowledge about the syntactic or semantic structure
  of a language
- **Abnormal character combinations such as misspellings and emoticons may be naturally learnt.**

Dataset Used

We obtained the **AG's corpus** of news article from http://groups.di.unipi.it/~gulli/AG_corpus_of_news_ar .

- It contains 496,835 categorized news articles from more than 2000 news sources. We choose the 4 largest classes from this corpus to construct our dataset, using only the title and description fields.

- The number of training samples for each class is 30,000 and testing 1900.

AG News Raw Dataset

We obtained the dataset from the link (as provided to us) :

http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html

```
Yahoo Business  http://us.rd.yahoo.com/dailynews/rss/business/*http://story.news.yahoo.com/news?tmpl=story2&u=/nm/20040814/bs_nm/column_stocks_week_dc
Wall St. Pullback Reflects Tech Blowout (Reuters)   none    Business    Reuters - Wall Street's long-playing drama,\
"Waiting for Google," is about to reach its final act, but its\
stock market debut is ending up as more of a nostalgia event\
than the catalyst for a new era.    5   0000-00-00 00:00:00 \N
Yahoo Business  http://us.rd.yahoo.com/dailynews/rss/business/*http://story.news.yahoo.com/news?tmpl=story2&u=/nm/20040814/bs_nm/markets_bears_dc   Wall
St. Bears Claw Back Into the Black (Reuters)   none    Business    Reuters - Short-sellers, Wall Street's dwindling\
band of ultra-cynics, are seeing green again.   5   0000-00-00 00:00:00 \N
Yahoo Business  http://us.rd.yahoo.com/dailynews/rss/business/*http://story.news.yahoo.com/news?tmpl=story2&u=/nm/20040814/bs_nm/column_mergers_dc
Carlyle Looks Toward Commercial Aerospace (Reuters) none    Business    Reuters - Private investment firm Carlyle Group,\
which has a reputation for making well-timed and occasionally\
controversial plays in the defense industry, has quietly placed\
its bets on another part of the market. 5   0000-00-00 00:00:00 \N
```

Data Cleaning

- In order to make the dataset usable for the CNN, we cleaned the dataset and populated two files, **train.csv and test.csv**, which contain 120,000 and 7600 entries respectively. (**train-validation sets: 1,00,000 - 20,000**)
- Both files contain three columns: **Class, Title and Description**.
- There are four classes for the news articles assigned a number from 1 to 4:
  - World
  - Sports
  - Business
  - Sci/tech
- ***We also obtained the training and testing data by using the clean dataset provided by one of the authors of the paper, for the purpose of comparison. This dataset has the same size, columns and classes as the datasets prepared by us.***

Data Processing

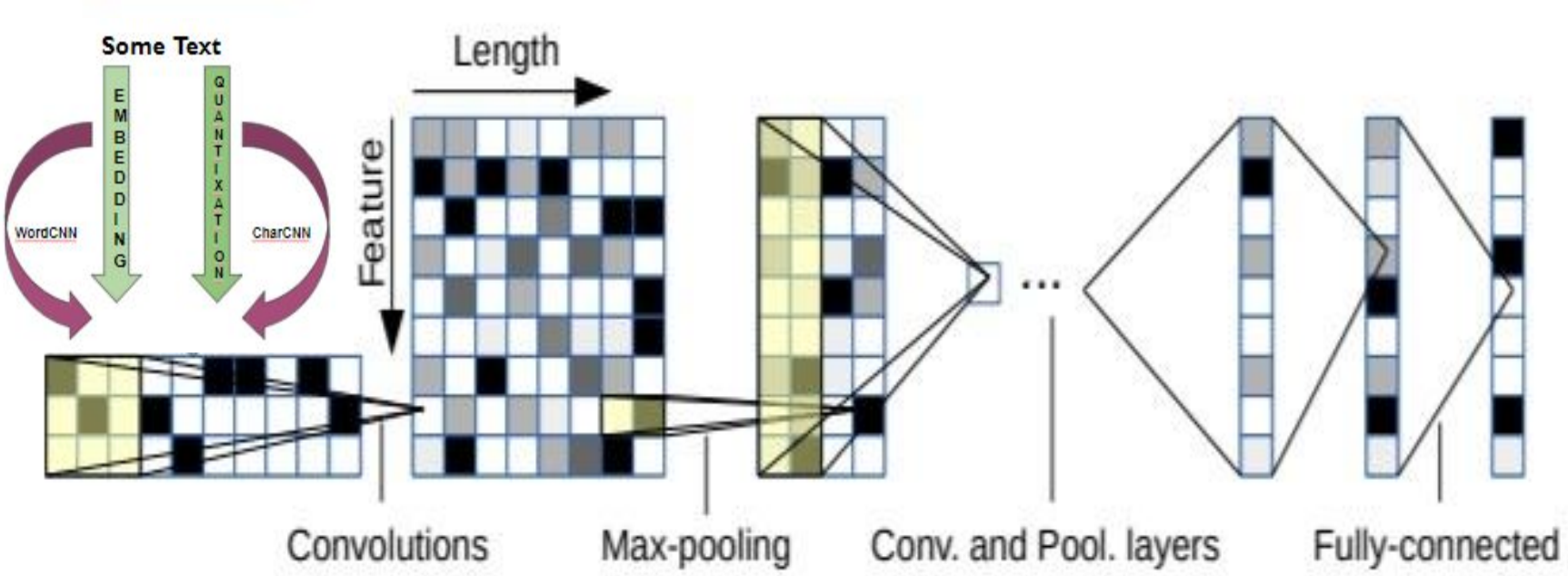
The characters are quantized using "**one-hot**" encoding. Every character from the given vocabulary list is given a unique vector of length 70.

"abcdefghijklmnopqrstuvwxyz0123456789-,;.!?:'"/\|_@#$%^&*~
'+-=<>()[]{}"

Every sentence is truncated or zero-padded to a fixed length of **1014** characters

```
[0., 0., 1., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.,
 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0., 0.])
```

Network


Some Text
EMBEDDING
QUANTIXATION
WordCNN
CharCNN
Length
Feature
Convolutions
Max-pooling
Conv. and Pool. layers
Fully-connected

Layer Description

Layer	Large Feature	Small Feature	Kernel	Pool
1	1024	256	7	3
2	1024	256	7	3
3	1024	256	3	N/A
4	1024	256	3	N/A
5	1024	256	3	N/A
6	1024	256	3	3

Layer	Output Units Large	Output Units Small
7	2048	1024
8	2048	1024
9	Depends on the problem	

Along with 2 drop-out modules in between fully connected layers with a drop-out probability of 0.5.

Test Runs:-

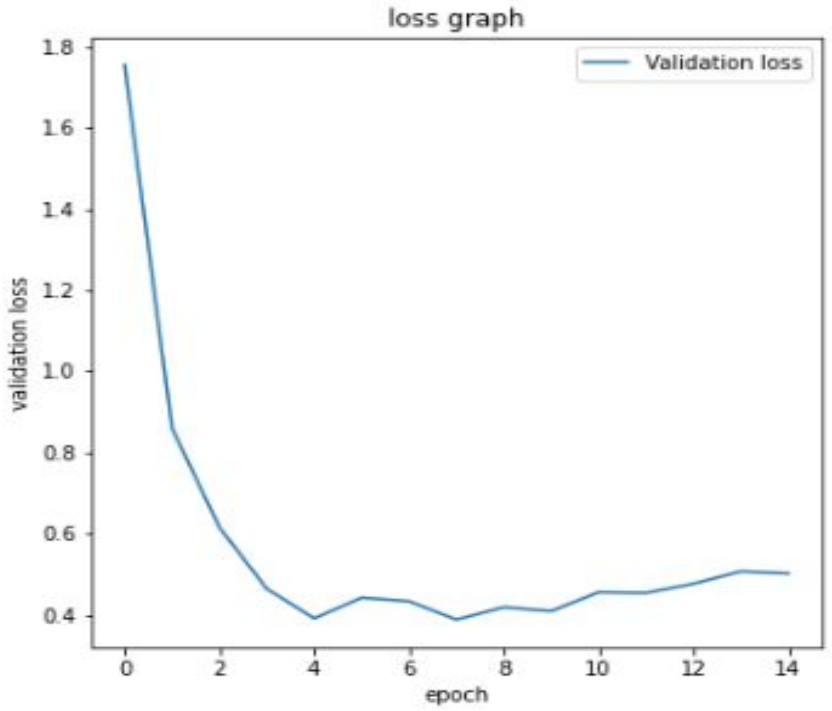
Networks designed for CharCNN and WordCNN were trained with and on the following:

1. **Small feature with author's dataset**
2. **Large feature with author's dataset**
3. **Small feature with self-cleaned dataset**
4. **Large feature with self-cleaned dataset**

Results - Validation Loss Curve

We observed the validation loss of the models and concluded that:

- **The loss is subsequently low for 10 epochs and hence we trained the model(s) for 15 epochs to keep a safe margin for comparison of results.**



Results - Test Accuracy

We compared the test accuracy of the models and concluded that:

- **Author's dataset gave better results with the cleaning techniques used.**
- **Using large size convolution layers contributed to better accuracy due to accomodation of more information.**
- **WordCNN performed slightly better than CharCNN due to pre-trained word-embeddings**

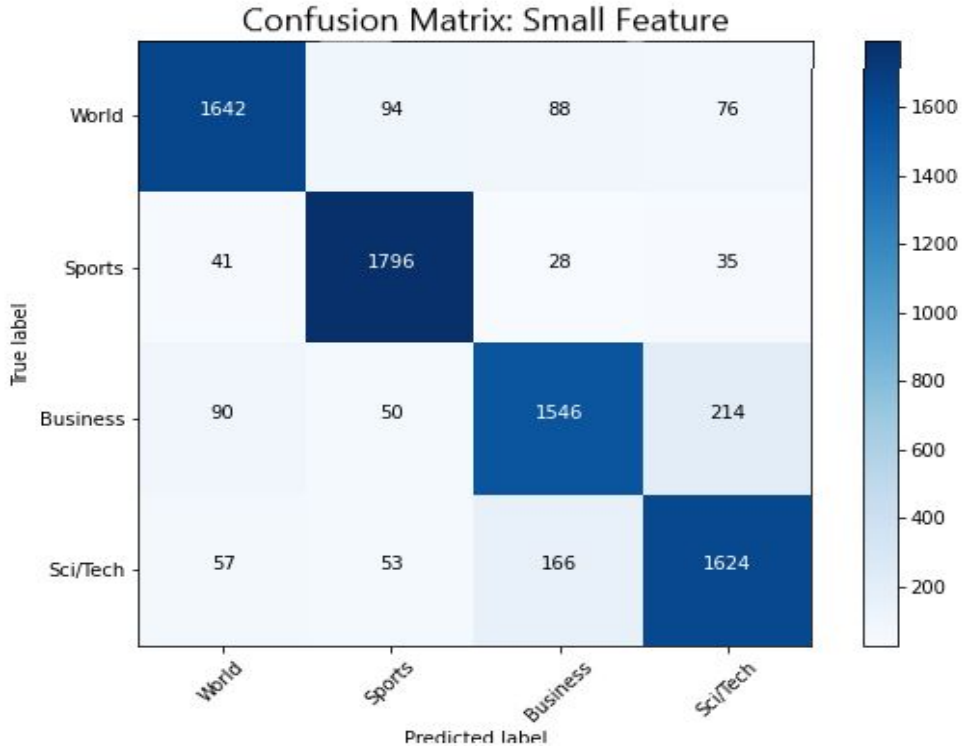
CharCNN			
Our Dataset		Author's Dataset	
Small Features	Large Features	Small Features	Large Features
0.81	0.84	0.87	0.89

WordCNN			
Our Dataset		Author's Dataset	
Small Features	Large Features	Small Features	Large Features
0.81	0.82	0.88	0.89

Results - Confusion Matrix

We compared the confusion matrix from all the 8 test runs of the models and concluded that:

- **The model was repeatedly successful in classifying Sports News.**
- **Similar trend was observed in WordCNN.**
- **The predicted reason behind this is the presence of more numeral characters in sports news as compared to other news items.**



Results - Other Comparison Measures

	CharCNN															
	Our Dataset								Author's Dataset							
	Small Feature				Large Feauture				Small Feauture				Large Feauture			
	Sci/Tech	Sports	Business	World	Sci/Tech	Sports	Business	World	Sci/Tech	Sports	Business	World	Sci/Tech	Sports	Business	World
Precision	0.75	0.89	0.72	0.89	0.84	0.87	0.77	0.9	0.83	0.9	0.85	0.9	0.87	0.94	0.85	0.89
Recall	0.76	0.89	0.78	0.81	0.8	0.94	0.82	0.81	0.85	0.95	0.81	0.86	0.85	0.95	0.85	0.89
f1-Score	0.76	0.89	0.75	0.85	0.82	0.9	0.79	0.86	0.84	0.92	0.83	0.88	0.86	0.95	0.85	0.89

	WordCNN															
	Our Dataset								Author's Dataset							
	Small Feature				Large Feauture				Small Feauture				Large Feauture			
	Sci/Tech	Sports	Business	World	Sci/Tech	Sports	Business	World	Sci/Tech	Sports	Business	World	Sci/Tech	Sports	Business	World
Precision	0.69	0.93	0.82	0.86	0.71	0.93	0.83	0.88	0.84	0.94	0.85	0.9	0.85	0.95	0.86	0.92
Recall	0.86	0.88	0.65	0.86	0.87	0.91	0.69	0.85	0.87	0.96	0.82	0.88	0.88	0.97	0.84	0.89
f1-Score	0.76	0.91	0.73	0.86	0.78	0.92	0.75	0.87	0.86	0.95	0.83	0.89	0.86	0.96	0.85	0.9

Comparison with the Paper

We compared our results with those mentioned in the paper. The errors stated by the author were found to be almost similar (As can be seen by out Test Accuracy) to those obtained by our trial tests.

Hence, the results of the project were reliable.

Model	AG
BoW	11.19
BoW TFIDF	10.36
ngrams	7.96
ngrams TFIDF	7.64
Bag-of-means	16.91
LSTM	13.94
Lg. w2v Conv.	9.92
Sm. w2v Conv.	11.35
Lg. w2v Conv. Th.	9.91
Sm. w2v Conv. Th.	10.88
Lg. Lk. Conv.	8.55
Sm. Lk. Conv.	10.87
Lg. Lk. Conv. Th.	8.93
Sm. Lk. Conv. Th.	9.12
Lg. Full Conv.	9.85
Sm. Full Conv.	11.59
Lg. Full Conv. Th.	9.51
Sm. Full Conv. Th.	10.89
Lg. Conv.	12.82
Sm. Conv.	15.65
Lg. Conv. Th.	13.39
Sm. Conv. Th.	14.80

Testing Errors of the models

Random News Sample Predictions

We use four random News Samples, in order to check our model's Qualitative Performance.

Sample 1 (Sci/Tech): 'LOS ANGELES (Reuters) - A group of technology companies including Texas Instruments Inc. <TXN.N>, STMicroelectronics <STM.PA> and Broadcom Corp. <BRCM.O>, on Thursday said they will propose a new wireless networking standard up to 10 times the speed of the current generation.'

Sample 2 (Sports): 'The Cleveland Indians pulled within one game of the AL Central lead by beating the Minnesota Twins, 7-1, Saturday night with home runs by Travis Hafner and Victor Martinez.'

Sample 3 (World): BEIJING (Reuters) - Beijing on Monday accused a Chinese-American arrested for spying for Taiwan of building an espionage network in the United States, and said he could go on trial very soon.'

Sample 4 (Business): 'HONG KONG (Dow Jones)--China Mobile (Hong Kong) Ltd. (CHL), the listed unit of China #39;s biggest cellular phone operator, posted Wednesday a 7.8 rise in first-half net profit on a 23 increase in its subscriber base. '

Random News Sample Predictions

	Predicted Class							
	CharCNN				WordCNN			
	Our Dataset		Author's Dataset		Our Dataset		Author's Dataset	
Actual Class	Small Features	Large Features	Small Features	Large Features	Small Features	Large Features	Small Features	Large Features
Business	Sci/Tech	Sci/Tech	Business	Business	Sci/Tech	Sci/Tech	Business	Business
Sci/Tech	Sci/Tech	Sci/Tech	Sci/Tech	Sci/Tech	Sci/Tech	Sci/Tech	Sci/Tech	Sci/Tech
World	Business	Business	World	World	World	World	World	World
Sports	Sports	Sports	Sports	Sports	Sports	Sports	Sports	Sports

The models when run on the Author's Dataset, predicts the classes of the random news samples correctly, on both large as well as small feature. On our Dataset, the models don't perform similarly well, which comply with the test accuracy achieved on both the datasets.

Observations and Final Remarks

- **Character Level CNN can directly work for text classification without the need for words because of comparable accuracy of both the models. This is a strong indication that language can be thought of as a signal no different from any other kind.**
- **The differences in the results for Author's and our dataset can be traced to differences in the cleaning process.**
- **Large Features always provide better results.**
- **The same networks can also be tested with a different vocabulary set to bring out further analysis.**
- **In future, Character-level ConvNets can be applied for a broader range of language processing tasks especially when structured outputs are needed.**

Thank You