Character-level Convolutional Networks for Text Classification

Xiang Zhang, Junbo Zhao and Yann LeCun
Computer Science Department, Courant Institute of Mathematical Sciences, New York University
NEW YORK UNIVERSITY

Introduction: Character-level Temporal Convolutional Networks

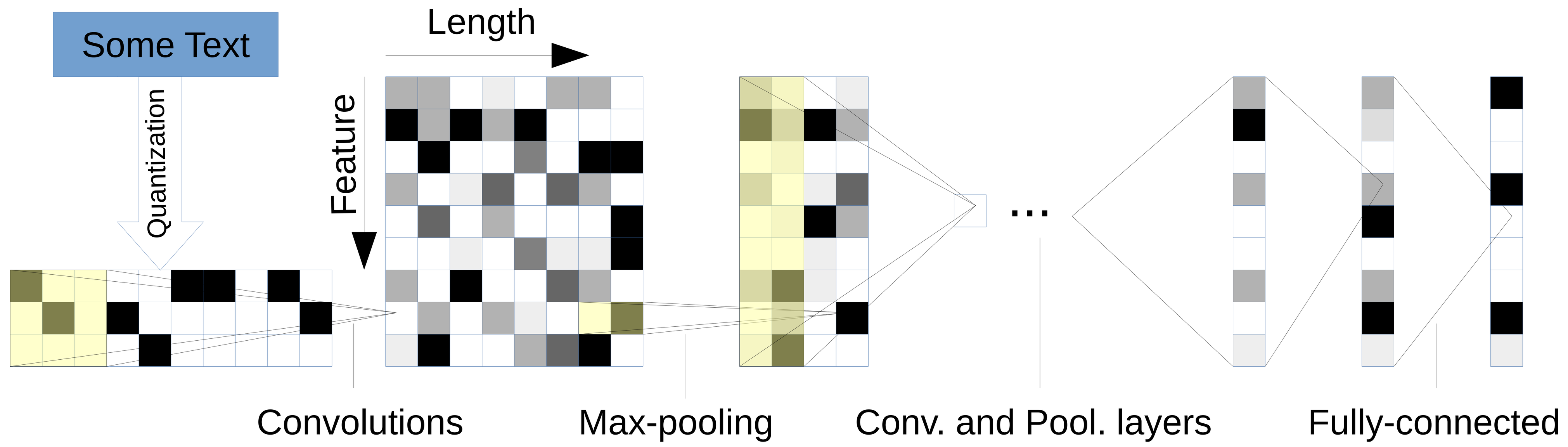
- We constructed several large-scale datasets to show that character-level convolutional networks could achieve state-of-the-art or competitive results for text classification.
- Comparisons are offered against traditional models and deep learning models.

Large-scale Datasets: Veriety of Different Sample Sizes

Dataset	Classes	Train Samples	Test Samples	Epoch Size
AG's News	4	120,000	7,600	5,000
Sogou News	5	450,000	60,000	5,000
DBPedia	14	560,000	70,000	5,000
Yelp Review Polarity	2	560,000	38,000	5,000
Yelp Review Full	5	650,000	50,000	5,000
Yahoo! Answers	10	1,400,000	60,000	10,000
Amazon Review Full	5	3,000,000	650,000	30,000
Amazon Review Polarity	2	3,600,000	400,000	30,000

- Data augmentation using thesaurus: replace words with their synonyms.

The Model: 9-layer Temporal Convolutional Networks



- Large model: A is size of alphabet and C is number of classes.

Layer	1	2	3	4	5	6	7	8	9
Type	Conv.	Conv.	Conv.	Conv.	Conv.	Conv.	Linear	Linear	Linear
Input length	1014	336	110	108	106	104	34	N/A	N/A
Input feature	A	1024	1024	1024	1024	1024	1024	2048	2048
Output feature	1024	1024	1024	1024	1024	1024	2048	2048	C
Kernel size	7	7	3	3	3	3	N/A	N/A	N/A
Pooling	3	3	N/A	N/A	N/A	3	N/A	N/A	N/A

- Small model: Simiar but with 256 features in covolution, 1024 features in linear layers.

Comparison Models: Traditional and Deep Learning

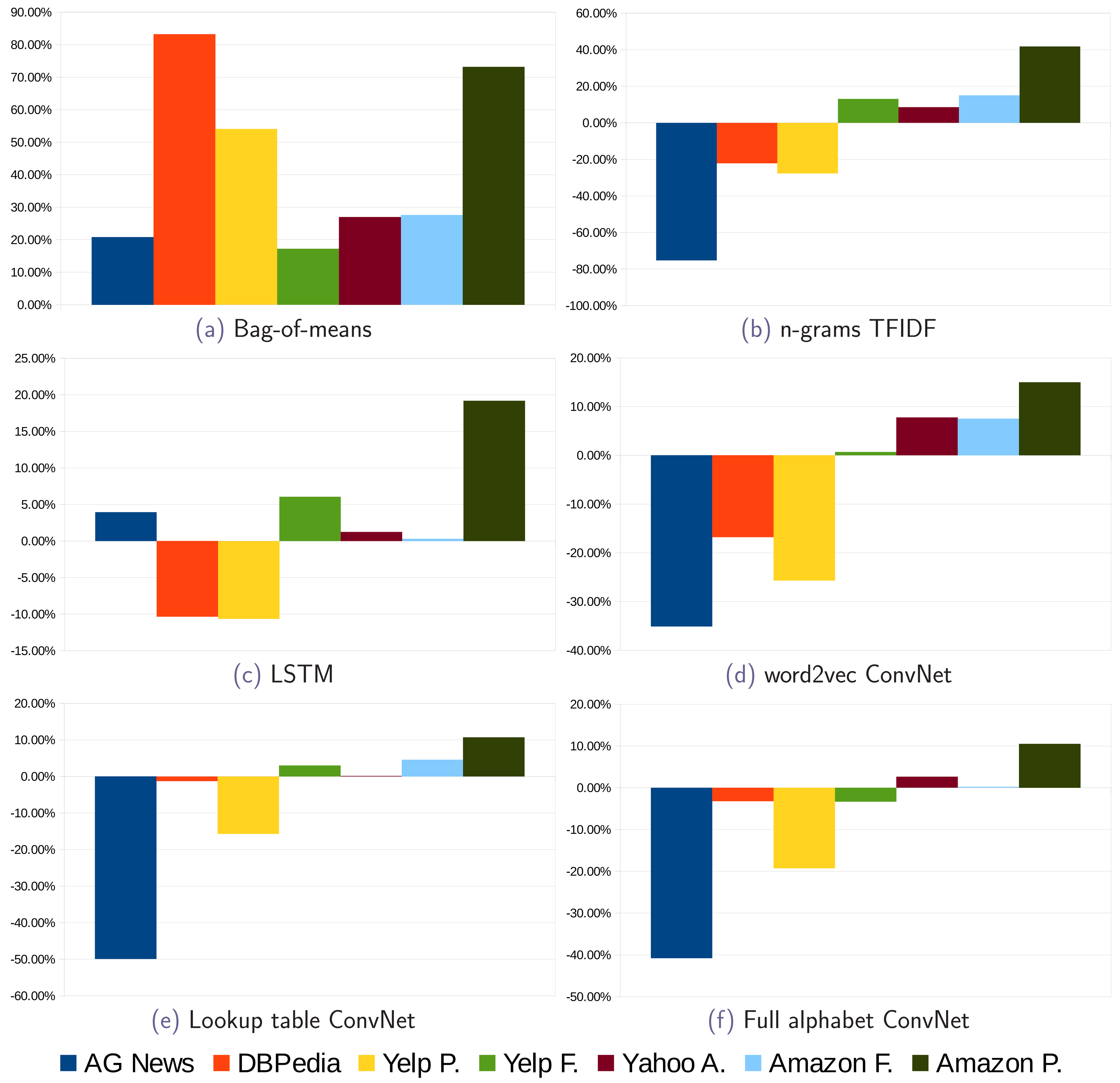
- Traditional models
  - Bag of words and its TFIDF variant: 50,000 most frequent words.
  - Bag of n-grams and its TFIDF variant: 500,000 most frequent n-grams (up to 5-grams).
  - Bag of means on word-embedding: 5000 means on word2vec embeddings.
- Deep learning models
  - Word2vec convolutional networks: using word2vec embeddings.
  - Lookup-table convolutional networks: word embedding trained jointly.
  - Long-short term memory (LSTM): feed the mean of hidden units to classifier.
- Choice of alphabet
  - Distinguish upper-case English letters or not.

To ensure fair comparisons, all these models are of comparable size.

Results and Comparisons: Testing Errors

Model	AG	Sogou	DBP.	Yelp P.	Yelp F.	Yah. A.	Amz. F.	Amz. P.
BoW	11.19	7.15	3.39	7.76	42.01	31.11	45.36	9.60
BoW TFIDF	10.36	6.55	2.63	6.34	40.14	28.96	44.74	9.00
ngrams	7.96	2.92	1.37	**4.36**	43.74	31.53	45.73	7.98
ngrams TFIDF	**7.64**	**2.81**	**1.31**	4.56	45.20	31.49	47.56	8.46
Bag-of-means	**16.91**	**10.79**	**9.55**	**12.67**	**47.46**	**39.45**	**55.87**	**18.39**
LSTM	13.94	4.82	1.45	5.26	41.83	29.16	40.57	6.10
Lg. w2v Conv.	9.92	4.39	1.42	4.60	40.16	31.97	44.40	5.88
Sm. w2v Conv.	11.35	4.54	1.71	5.56	42.13	31.50	42.59	6.00
Lg. w2v Conv. Th.	9.91	-	1.37	4.63	39.58	31.23	43.75	5.80
Sm. w2v Conv. Th.	10.88	-	1.53	5.36	41.09	29.86	42.50	5.63
Lg. Lk. Conv.	8.55	4.95	1.72	4.89	40.52	29.06	45.95	5.84
Sm. Lk. Conv.	10.87	4.93	1.85	5.54	41.41	30.02	43.66	5.85
Lg. Lk. Conv. Th.	8.93	-	1.58	5.03	40.52	28.84	42.39	5.52
Sm. Lk. Conv. Th.	9.12	-	1.77	5.37	41.17	28.92	43.19	5.51
Lg. Full Conv.	9.85	8.80	1.66	5.25	38.40	29.90	40.89	5.78
Sm. Full Conv.	11.59	8.95	1.89	5.67	38.82	30.01	40.88	5.78
Lg. Full Conv. Th.	9.51	-	1.55	4.88	38.04	29.58	40.54	5.51
Sm. Full Conv. Th.	10.89	-	1.69	5.42	**37.95**	29.90	40.53	5.66
Lg. Conv.	12.82	4.88	1.73	5.89	39.62	29.55	41.31	5.51
Sm. Conv.	15.65	8.65	1.98	6.53	40.84	29.84	40.53	5.50
Lg. Conv. Th.	13.39	-	1.60	5.82	39.30	**28.80**	40.45	**4.93**
Sm. Conv. Th.	14.80	-	1.85	6.49	40.16	29.84	**40.43**	5.67

Relative Errors: Comparison against Word-based Methods



(a) Bag-of-means (b) n-grams TFIDF (c) LSTM (d) word2vec ConvNet (e) Lookup table ConvNet (f) Full alphabet ConvNet

Conclusion, Code and Datasets

- Character-level convolutional network is an effective method for text classification.
- How well it performs depends on many factors such as dataset size and choice of alphabet.
- There is not a single model that performs the best in every case.
- Code: http://github.com/zhangxiangxiao/Crepe
- Datasets: http://goo.gl/JyCnZq

Computer Science Department, Courant Institute of Mathematical Sciences, New York University
Email: {xiang, junbo.zhao, yann}@cs.nyu.edu
Website: http://cilvr.cs.nyu.edu