End-To-End Spam Classification With Neural Networks

Christopher Lennan, Bastian Naber, Jan Reher, Leon Weber

Humboldt-Universität zu Berlin, Germany

Motivation

- Until a few years ago the majority of email traffic was due to spam
- Machine learning algorithms must quickly adapt to changing tactics of spammers
- Neural Networks are suited for this task as they do not require feature engineering

Methods

We trained two Convolutional Neural Networks (CNN) and a baseline Support Vector Machine (SVM) on the SpamAssassin and TREC 2007 spam/ham data sets using two encoding methods:

- **Character-level methods** (each character in email is one-hot encoded based on an alphabet)
  - Character-level linear SVM baseline
  - Character-level CNN similar to [1]
- **Word embeddings** (each word in email is transformed to a dense vector)
  - CNN with embedding layer

Results

Accuracy scores on two data sets

	TREC 2007	SpamAssassin
SVM (character level)	**1**	0.96
CNN (character-level)	**1**	0.97
CNN (word embeddings)	**1**	**0.98**

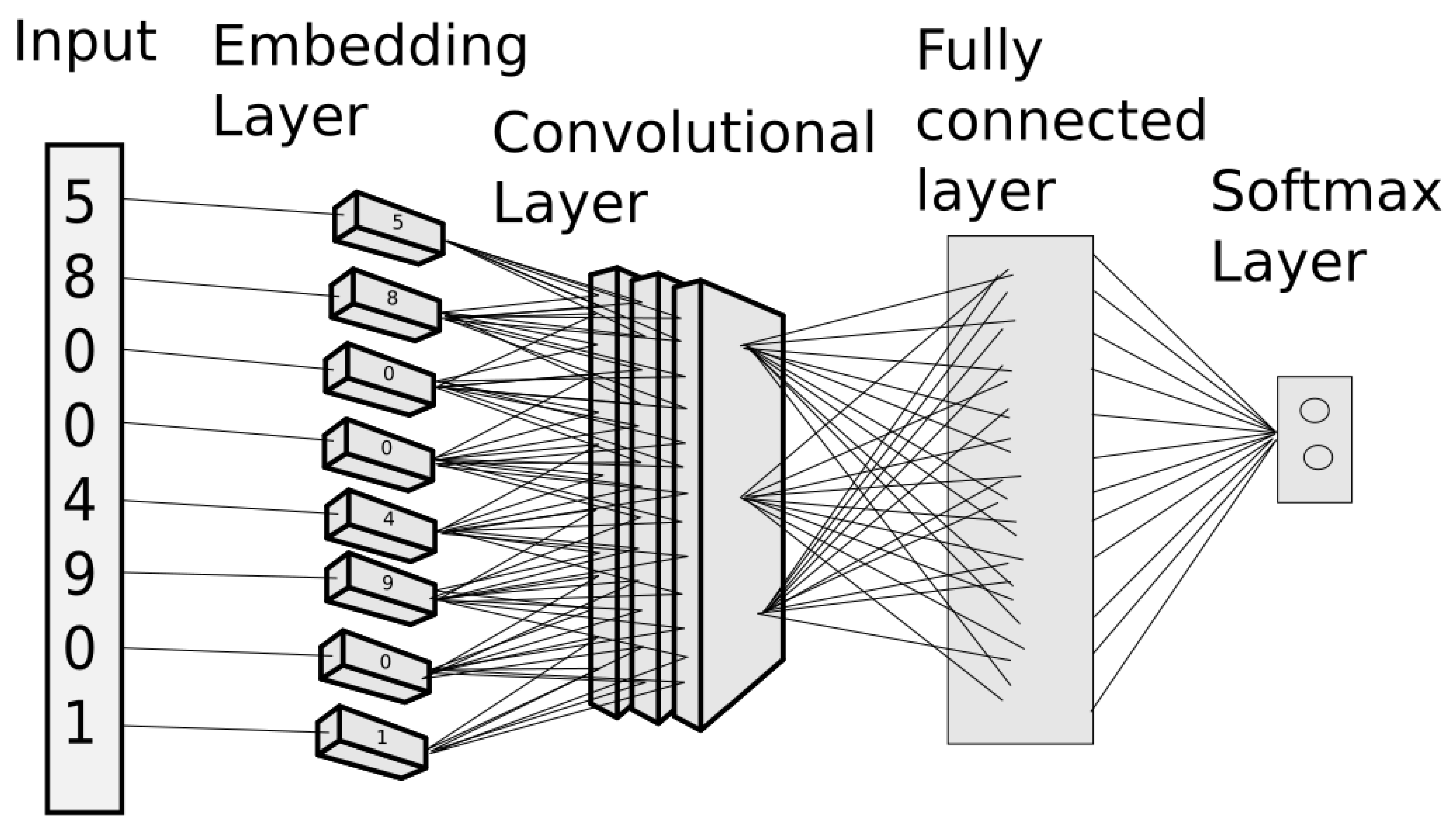
Conclusions

- Character-level SVM performs suprisingly well
- Promising results for CNN-based spam classification with especially good accuracy using word embeddings
- TREC 2007 data set seems unrepresentative

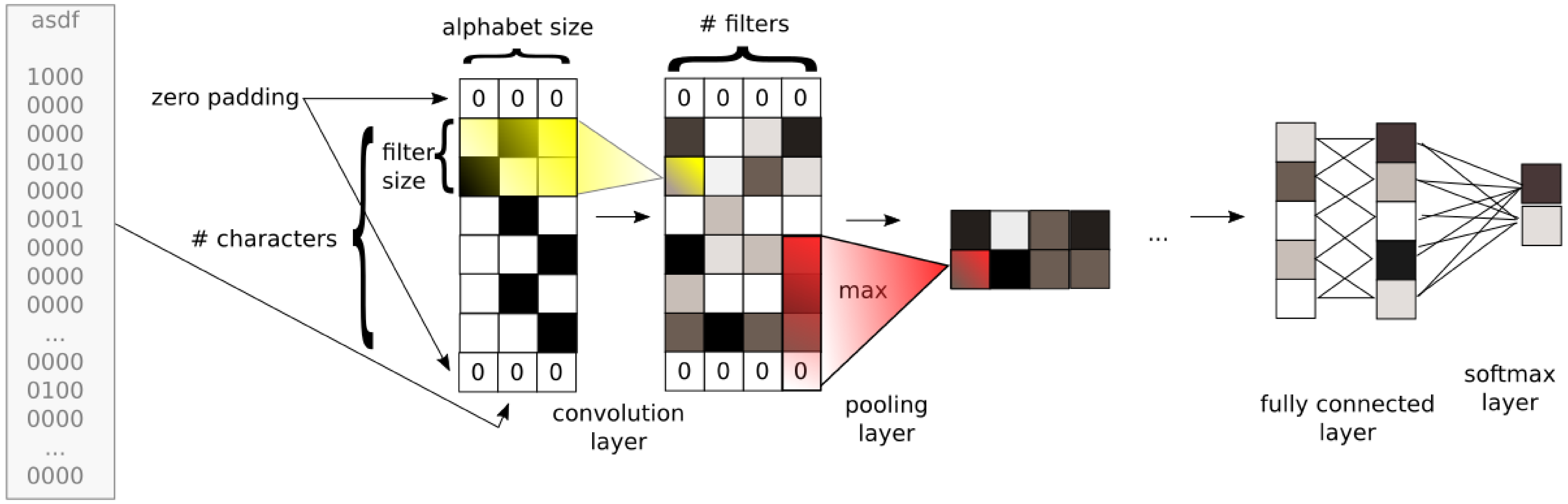
Forthcoming Research

- Create larger data sets and more diverse spam/ham representations by combining multiple data sets and thus achieve better generalization performance of the trained algorithm

Word embeddings CNN



Character-level CNN



References

[1] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. *arXiv preprint arXiv:1509.01626*, 2015.