Toxic Comments Classification in The Cyber Community

Wenyi Jones
wenyi503@stanford.edu
https://youtu.be/SnmOHr_KIHg

Data

The data sets are from a Kaggle competition "Toxic Comment Classification Challenge: Identify and classify toxic online comments"

There are about 160,000 examples for both training and testing. Each column is a binary classification of comment ranging from "toxic" to "identity hate", of six possible categories.

Models

Deep Neural Network (DNN)

- Input features: Converted comments to integers and constrained the maximum sentence length to 435.
- 3 fully connected layers each with different number (e.g. 8, 8, 6 and 50, 50, 6) of fully connected nodes; softmax as activation for the last layer.
- Running rate: 0.01; optimizer: adam; loss: categorical_crossentropy.

LSTM Recurrent Neural Network (RNN)

- Input features: same as that for DNN.
- 1 layer of LSTM followed by 1 layer of fully connected network.

Accuracy

	toxic	severe	obscene	threat	insult	Identity_hate
DNN (8, 20 epochs)	63.51	99.76	97.59	99.86	97.48	99.53
DNN (8, 50 epochs)	63.63	99.76	97.21	99.84	97.76	99.53
DNN (50, 20 epochs)	96.01	99.76	61.74	99.85	97.74	99.54
RNN (20 epochs)	96.02	99.76	97.59	99.86	97.76	99.53

Results

For the two DNN models with the same size of hidden layers and different epochs, there is virtually no difference in terms of accuracy results. For the third DNN model with 50 fully connected nodes, toxic's accuracy improved dramatically just as much as obscene's accuracy dropped.

The LSTM RNN model has the best performance with great accuracy across all six classifications. This comes as no surprise as LSTM allows us to compute the hidden state and thus great for predicting sequential words.

Future Steps

1. The results show that the LSTM RNN model we employed outperforms DNN models. In the future, we could zoom in on various LSTM RNN models and techniques.
2. We constrained the maximum sentence length to 435, with more resources, we can try improving accuracy without such constraint.
3. There are comments in foreign languages. Also, we could discard unknown characters (of invalid language) before training.

References: see paper report