COMPARING REPRESENTATION LEARNING OVER WORD-LEVEL, CHARACTER-LEVEL AND COMBINATION OF BOTH IN NLP TASKS

Anton Serjogin
Centre for Information and Speech Processing, LMU
anton.serjogin@gmail.com

22.05.2017

Objective:

- How different input styles influence the accuracy of Convolutional Neural Networks?
- What is the best set of parameter for CNN?
- How fast the accuracy will be calculated?
- What will change the combination of both word- and character-embeddings?

Representation:

- The commonly used word-embeddings
- Character-level NLP
- Combination of both

The main task is to compare such NLP tasks like sentiment classification and POS tagging. CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network) - two main types of deep neural networks that are widely explored to handle vaious NLP tasks.

Differences:

- RNN can handle arbitrary input/output lengths, can use their internal memory to process arbitrary sequences of inputs, use time-series information, are ideal for text and speech analysis
- CNN take a fixed size input and generate fixed-size outputs, are designed to use minimal amounts of preprocessing, use connectivity pattern between its neurons is inspired by the organization of the animal visual cortex, are ideal for images and videos processing

Changing the following parameters for CNN:

- Embedding size (dimensionality of embeddings)
- Hidden size (number of neurons)
- Batch size (training examples)

Data and implementation:
Stanfort Sentiment Treebank:

- Standfort Sentiment Treebank
  - Contains annotated data from movie reviews

  - Circa 215k unique phrases

- Python Framework Theano
  - Consists of the input layet, hidden layer, output layer
  - While input and output layers stay the same, hidden layers vary in neural networks