Exercise set 6: Convolutional networks

In this exercise we will try to classify German words as either noun or verb using CNNs.

The data we use for this set of exercises is available in two formats at the following URLs:

- Python pickle http://coltekin.net/cagri/courses/ml/data/e6-data.pickle
- JSON http://coltekin.net/cagri/courses/ml/data/e6-data.json.gz

The data consists of German nouns and verbs with associated class labels. All words are converted to lowercase. It is pre-processed for your convenience. You can load the data (pickle version) using a command similar to:

```
w, c, l, x, y = pickle.load(open('e6-data.pickle','rb'))
```

Where `w` is the words, `c` is the class labels (`noun` or `verb`), `l` contains the list of letters for each word, `x` is a numeric representation of `l` which replaces every letter with an integer index, and `y` is the binary class labels (`noun=1`, `verb=0`). The variables `x` and `y` are particularly suitable for use with Keras' `Embedding` layer.

Exercises

1. Define and train a CNN for predicting the class of words in the data. Use task-specific emebedding vectors for the letters.

   A good example for Keras can be found at https://github.com/fchollet/keras/blob/master/examples/imdb_cnn.py. Make sure to adjust all parameters to more reasonable values.

2. Train the same model with one-of-K representation (without embeddings).

3. Train embeddings for the letters externally using GloVe or word2vec, use these general-purpose embeddings for training a CNN.