Lab: Character-Level LMs

David Vilar
david.vilar@nuance.com
MT Marathon 2016
Goals

- Implement a simple FF LM using theano
- Experiment with the hyperparameters and get a feeling of their influence
- Extend the model to increase accuracy and performance
The task:

- Predict missing items in a text (see also https://www.kaggle.com/c/billion-word-imputation)

Practical considerations: Work on character level

- No OOVs
- Small vocabulary
alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought alice, 'without pictures or conversations?'
Test data:

the garden of live flowers
but how curiously it twi @sts!
this goes straight back to the
The Code

- Skeleton code and data available under
  /net/mtm/mtm006/lmExercise
- Heavily based on
  http://deeplearning.net/tutorial/mlp.html and
  http://deeplearning.net/tutorial/logreg.html
  - Most of the documentation on these pages is valid for this code, check it!
- Gaps to be filled are marked as <SNIP>
- Potentially problematic code marked as # <ADAPT?>
Example Run

```
$ ./mlp.py -v data/voc -t data/alice-letters.txt \  
    -V data/validation-letters.txt \  
    -T data/test-letters-gaps.txt \  
    -l 2 -c 3 -n 100
```

Data sizes:
  train: 142645 valid: 17190 test: 1665

... building the model

... training

epoch 1 (patience 10), validation PPL 11.51660, test error 71.71%
epoch 2 (patience 10), validation PPL 9.77725, test error 67.21%
epoch 3 (patience 10), validation PPL 9.03788, test error 65.29%
epoch 4 (patience 10), validation PPL 8.58992, test error 64.02%
epoch 5 (patience 10), validation PPL 8.23185, test error 62.88%
epoch 6 (patience 10), validation PPL 7.93009, test error 60.72%
epoch 7 (patience 12), validation PPL 7.68329, test error 59.10%
epoch 8 (patience 14), validation PPL 7.47755, test error 58.56%
epoch 9 (patience 16), validation PPL 7.29955, test error 58.14%
epoch 10 (patience 18), validation PPL 7.14263, test error 58.02%

...
Possible Enhancements

• Bidirectional context
• Efficient data representation
  • Data storage is extremely inefficient
  • Implement embedding layer as lookup table
• Go recurrent!
• ...
Lab: Character-Level LMs

David Vilar
david.vilar@nuance.com
MT Marathon 2016