Handwriting to text conversion for mathematical texts

Submitted by Erik Beerepoot for CSCI E-89: Deep Learning for Natural Language Processing, Summer 20119

Problem statement: Researchers in the sciences spend significant time transcribing pre-existing mathematics (handwritten and typeset) into Latex code.

Approach:

- Explore CNN encoder architecture for single character detection.
- Use a sequence2sequence model similar to Google's "Show, Attend and Tell" neural image captioning paper, applied to the im2latex dataset from Harvard NLP, applying architecture from exploration.
- Used distributed Tensorflow to train model on 4 Tesla V100 accelerators.

Benefits:

- Save time by generating Latex for mathematics in the scientific literature
- Transcribe handwritten mathematics into high-quality rendering for notes & publishing.

Challenges:

- Full end-to-end model changing is slow — training with full dataset is time/cost prohibitive.
- Long formula lengths are challenging for RNNs because of the vanishing gradient problem.
- Compiling the predicted latex into an image requires correctness — even a high degree of similarity can still be syntactically invalid.

Data sources:

- **ICFHR 2016 Competition on Recognition of On-line Handwritten Mathematical Expressions (CROHME) Dataset**
  - http://tc11.cvc.uab.es/datasets/ICFHR-CROHME-2016_1
- **Im2latex dataset**
  - https://zenodo.org/record/56198#.V2p0KTXT6eA

Results: When training the final model for 60 epochs on ~32k examples, the training loss is 20 (down from 120) and validation loss is 18. Subjectively, the model makes accurate predictions on short formulas, but often struggles to produce syntactically valid Latex.

Notebooks:

1. **Symbol Recognition using CNN - Data Preprocessing:** Pre-processing data for single symbols recognition.
2. **Symbol Recognition using CNN - Training:** Training a simple CNN to recognize Latex symbols.
3. **Formula recognition using seq2seq - Data Preprocessing:** Pre-processing the im2latex dataset for the complex, whole-formula case.
4. **Formula recognition using seq2seq - Training:** Training a seq2seq model on the data, and visualizing the results.
5. **Formula recognition using seq2seq - Distributed Training:** Adapting the model to run on more GPUs.

Video: https://youtu.be/Fwr0lHJuzDI **Github:** https://github.com/erikbeerepoot/img-to-latex