Exploring the Limits of Language Modelling

9.2.17

Alex Gamble

Context

- Language Modelling
    - Count-based approaches.
    - Continuous space models.
- Argument that previous research focused on PTB too heavily.

Aims

- Present current research on language models.
- Extend several NN approaches to better tackle issues of:
  - Corpora and vocabulary sizes.
  - Long term and complex language structures.
- Apply and evaluate these approaches to ‘One Billion Word’ benchmark.

Softmax Optimisations

$$P(y=j|\mathbf{x}) = \frac{e^{\mathbf{x}^\mathsf{T}\mathbf{w}_j}}{\sum_{k=1}^{K} e^{\mathbf{x}^\mathsf{T}\mathbf{w}_k}}$$

- |V| * |h|, where *V* is the vocabulary set and *h* is the set of contexts.

- Computationally expensive during training when vocabulary is large.

CNN Softmax

- Calculate embedding for Softmax logit as e_w = *CNN* (chars$_w$)

- Argument made that vector e_w can be precomputed, so no additional computational complexity compared to regular Softmax.

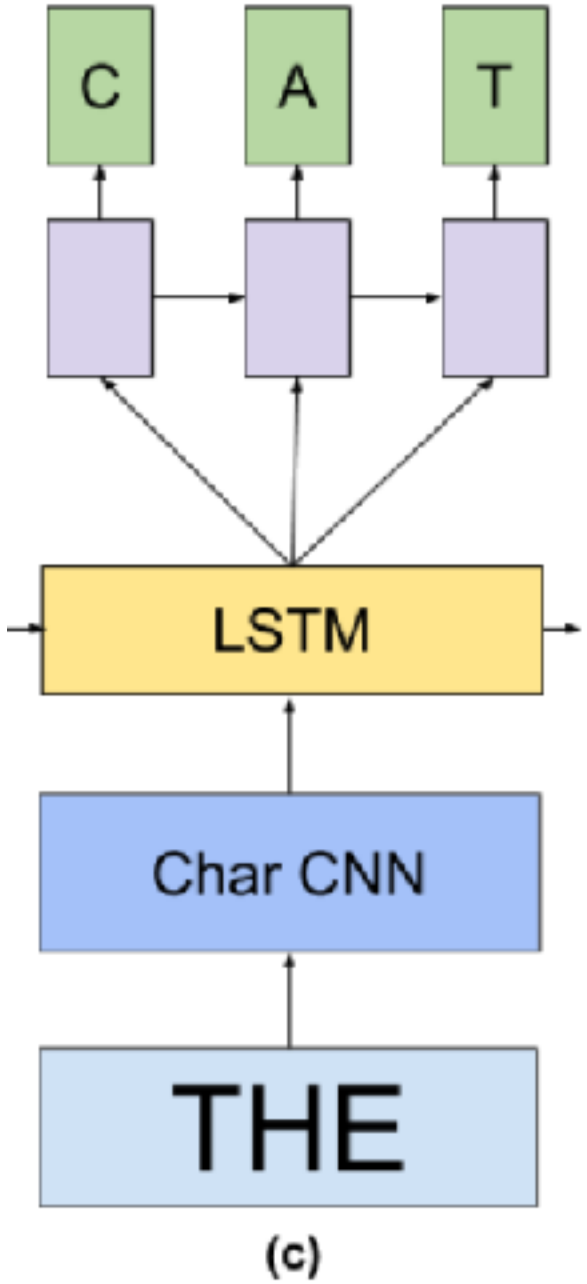
CNN Softmax

$$z_w = h^T CNN(chars_w) + h^T M corr_w$$

- Many logit ties, authors found the mapping from character sequence to embedding is smooth.
- Lower learning rate meaning increased training time.

Char LSTM Predictions

- CNN Softmax layer still slow.
- Instead, provide hidden state of LSTM to C-LSTM, to predict word one character at a time.



Experiments

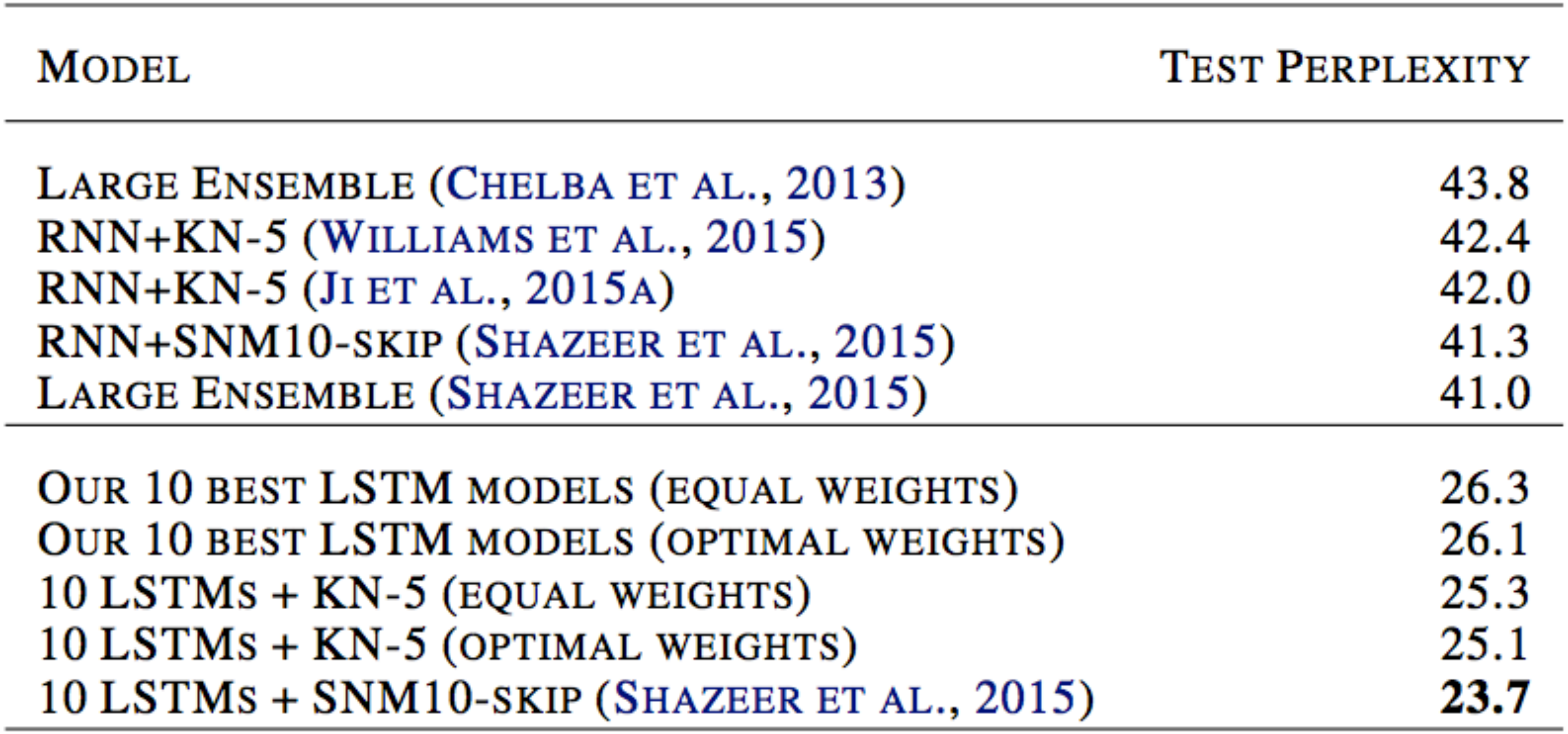
$$PP(w_1 \ldots w_N) \quad = \quad \sqrt[N]{\frac{1}{P(w_1 \ldots w_N)}}$$

- Trained and evaluated on 1B Word Benchmark data.
  - 0.8B words
  - Vocabulary of 793,471
- Perplexity as evaluation metric.
- Evaluation of several literature models.

Table 1. Best results of single models on the 1B word benchmark. Our results are shown below previous work.

Model	Test Perplexity	Number of Params [billions]
Sigmoid-RNN-2048 (Ji et al., 2015a)	68.3	4.1
Interpolated KN 5-gram, 1.1B n-grams (Chelba et al., 2013)	67.6	1.76
Sparse Non-Negative Matrix LM (Shazeer et al., 2015)	52.9	33
RNN-1024 + MaxEnt 9-gram features (Chelba et al., 2013)	51.3	20
LSTM-512-512	54.1	0.82
LSTM-1024-512	48.2	0.82
LSTM-2048-512	43.7	0.83
LSTM-8192-2048 (No Dropout)	37.9	3.3
LSTM-8192-2048 (50% Dropout)	32.2	3.3
2-Layer LSTM-8192-1024 (BIG LSTM)	30.6	1.8
BIG LSTM+CNN Inputs	**30.0**	**1.04**
BIG LSTM+CNN Inputs + CNN Softmax	39.8	**0.29**
BIG LSTM+CNN Inputs + CNN Softmax + 128-dim correction	35.8	**0.39**
BIG LSTM+CNN Inputs + Char LSTM predictions	47.9	**0.23**

Table 2. Best results of ensembles on the 1B Word Benchmark.



Model	Test Perplexity
Large Ensemble (Chelba et al., 2013)	43.8
RNN+KN-5 (Williams et al., 2015)	42.4
RNN+KN-5 (Ji et al., 2015a)	42.0
RNN+SNM10-skip (Shazeer et al., 2015)	41.3
Large Ensemble (Shazeer et al., 2015)	41.0
Our 10 best LSTM models (equal weights)	26.3
Our 10 best LSTM models (optimal weights)	26.1
10 LSTMs + KN-5 (equal weights)	25.3
10 LSTMs + KN-5 (optimal weights)	25.1
10 LSTMs + SNM10-skip (Shazeer et al., 2015)	**23.7**

$< S >$ With even more new technologies coming onto the market quickly during the past three years , an increasing number of companies now must tackle the ever-changing and ever-changing environmental challenges online . $< S >$ Check back for updates on this breaking news story . $< S >$ About 800 people gathered at Hever Castle on Long Beach from noon to 2pm , three to four times that of the funeral cortège . $< S >$ We are aware of written instructions from the copyright holder not to , in any way , mention Rosenberg 's negative comments if they are relevant as indicated in the documents , " eBay said in a statement . $< S >$ It is now known that coffee and cacao products can do no harm on the body . $< S >$ Yuri Zhirkov was in attendance at the Stamford Bridge at the start of the second half but neither Drogba nor Malouda was able to push on through the Barcelona defence .

Advocate

- Good contextualisation
  - Evaluation of a number of models as baselines.
  - Two metrics - parameters and perplexity.

- Experiment
  - Novel ideas attempted in several areas, with varying degrees of success.

Criticism

- Evaluation
  - Motivation to reduce parameterisation of models could be further expanded.

- CNN Softmax
  - Work here seems overly brief and without justification for some methods tried.
  - Would be useful to see function mappings for similarly spelt words.