Two-Layer Hierarchical Softmax in Penn TreeBank Language Modeling

Lei Mao

“Depth Charge”

University of Chicago

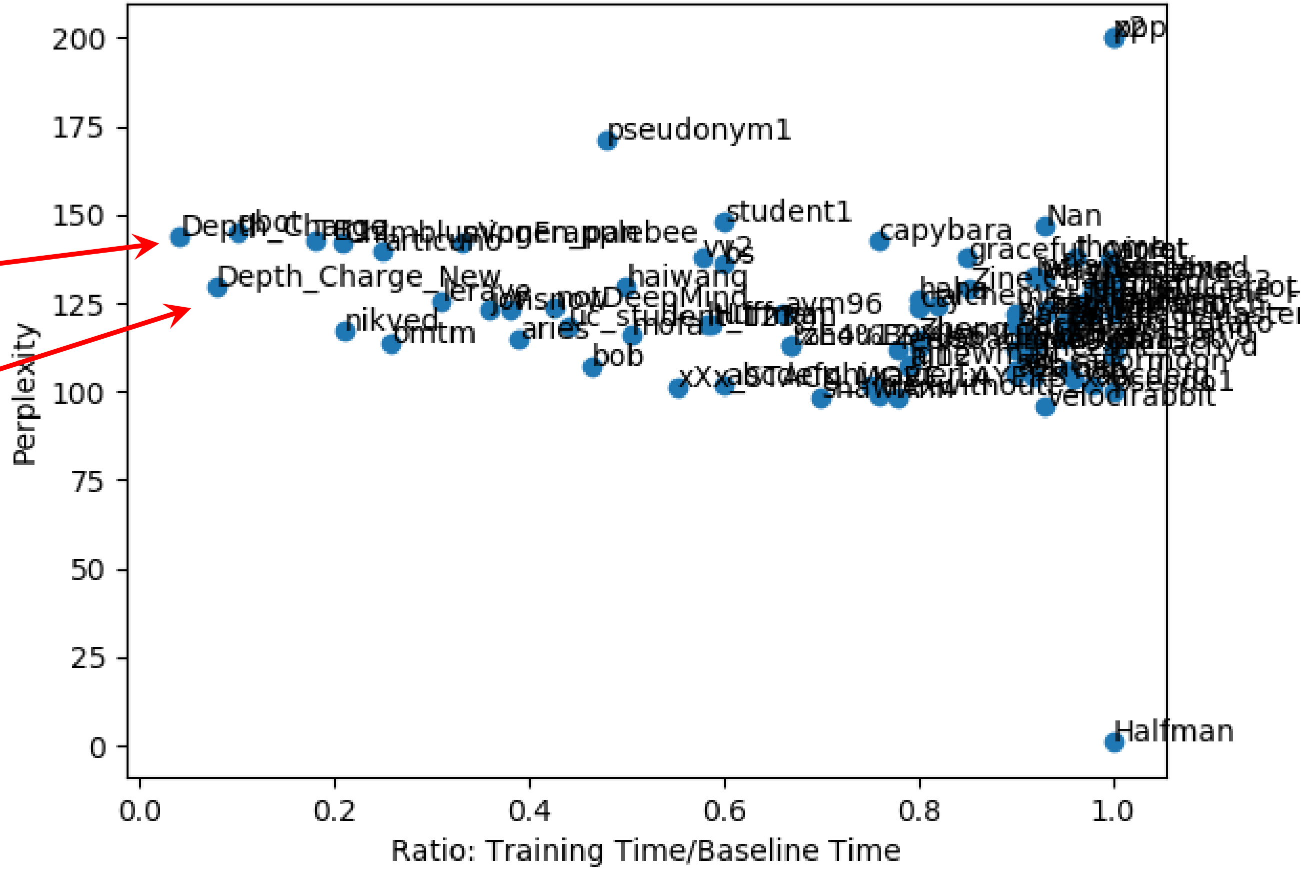
3/8/2018

Pareto Point

“Depth Charge”

(143.79, 0.041)

(129.87, 0.079)



Algorithm Implemented and Tested

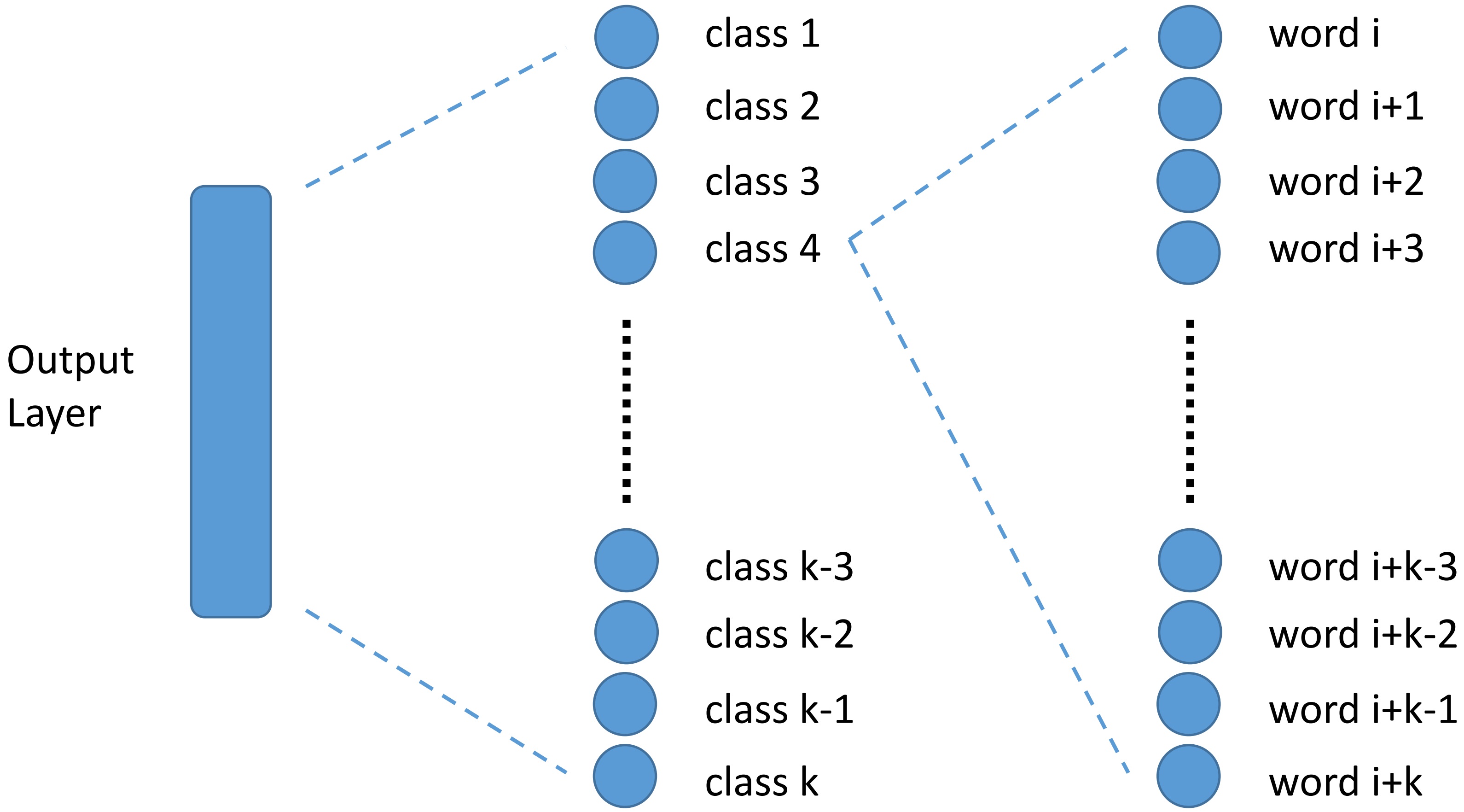
Algorithm	Category	Efficiency (Per Batch)	Validation	Notes	Asymptotic Complexity	Ref	Code
BlackOut	Sampling Softmax	Slow (sampling 200 negative samples worse than full softmax)	Bad ☹	My negative sampling implementation was bad	O(k)	[1]	Not ready for open-source yet
Sampled Softmax	Sampling Softmax	2-6 times faster (sampling 100 – 200 negative samples)	Acceptable Still worse than full softmax	C++ Backended PyTorch code available.	O(k)	[2]	GitHub
Two-Layer Hierarchical Softmax	Hierarchical Softmax	5 times faster	Very good short term.	Official Theano code available.	O(NlogN)	[3]	GitHub

[1] BlackOut: Speeding up Recurrent Neural Network Language Models With Very Large Vocabularies, 2016.

[2] On Using Very Large Target Vocabulary for Neural Machine Translation, 2015.

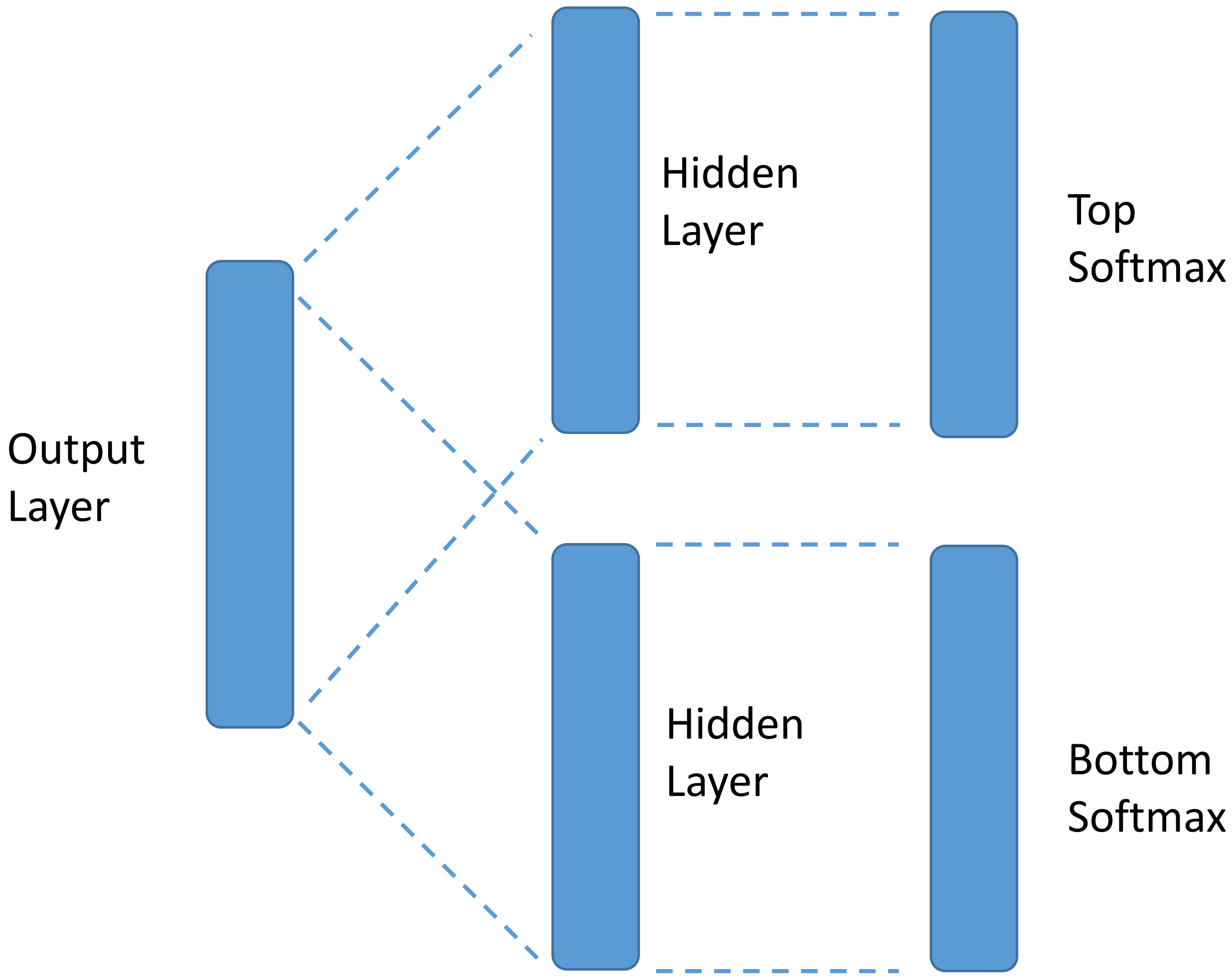
[3] Classes for Fast Maximum Entropy Training, 2001.

Introduction to Two-Layer Hierarchical Softmax



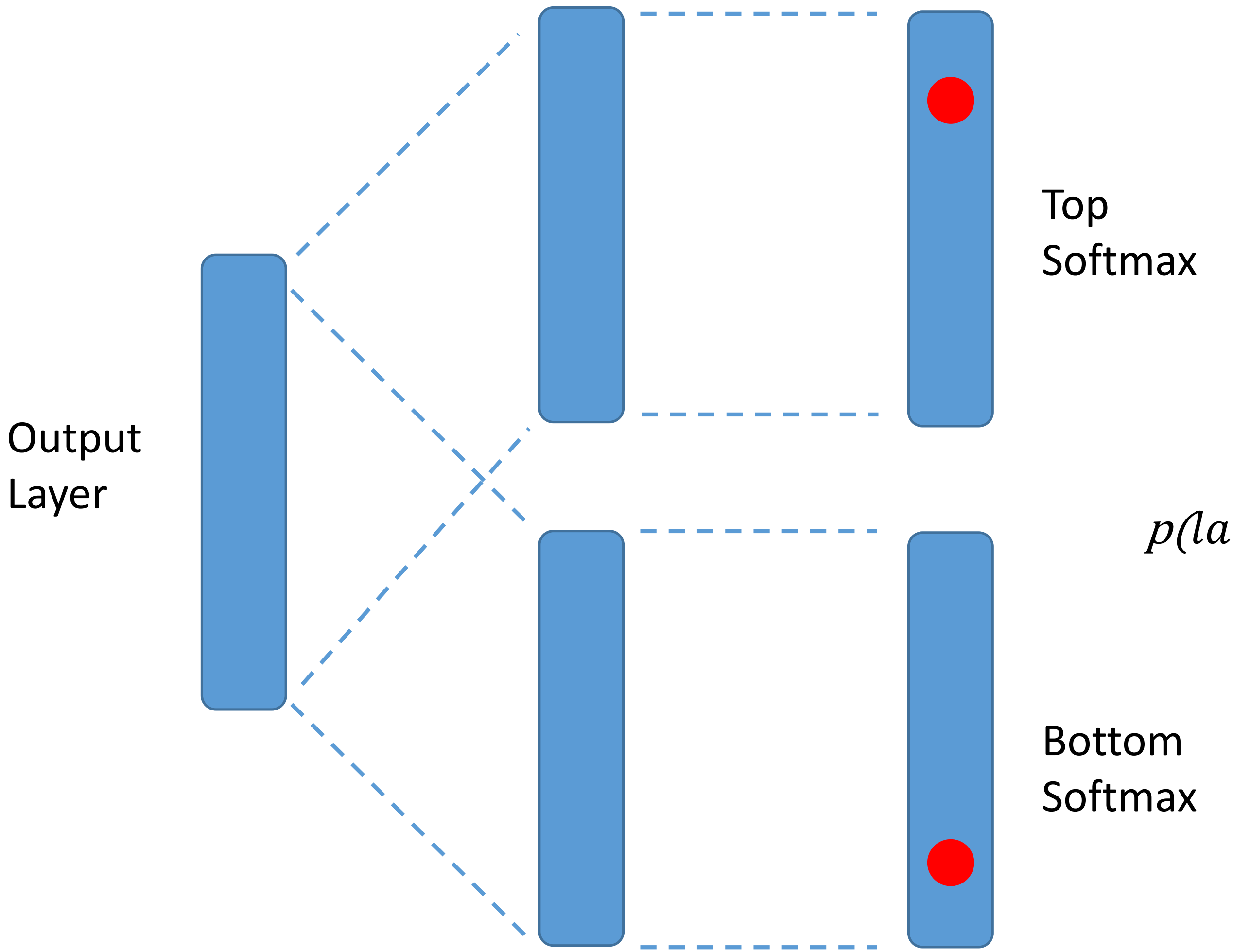
$$p(word_{i+2}) = p(class_4)p(word_{i+2}|class_4)$$

Introduction to Two-Layer Hierarchical Softmax



For Penn TreeBank dataset with corpus size of 10,000, this two-layer hierarchical softmax is probably better than any other complex hierarchical softmax, such as Huffman Tree softmax. Because dataset is small and the model is so **SIMPLE**!

Introduction to Two-Layer Hierarchical Softmax



During training, for any input x, we need to know its target label in both top softmax and bottom softmax.

$$p(word_{i+2}) = \\ p(label\_top\_word_{i+2}, label\_bottom\_word_{i+2}) = \\ p(label\_top\_word_{i+2}) \times \\ p(label\_bottom\_word_{i+2} | label\_top\_word_{i+2})$$

Minimize the loss:

$$loss = -\log p(word\_target | context)$$

Settings and Performance of the Model

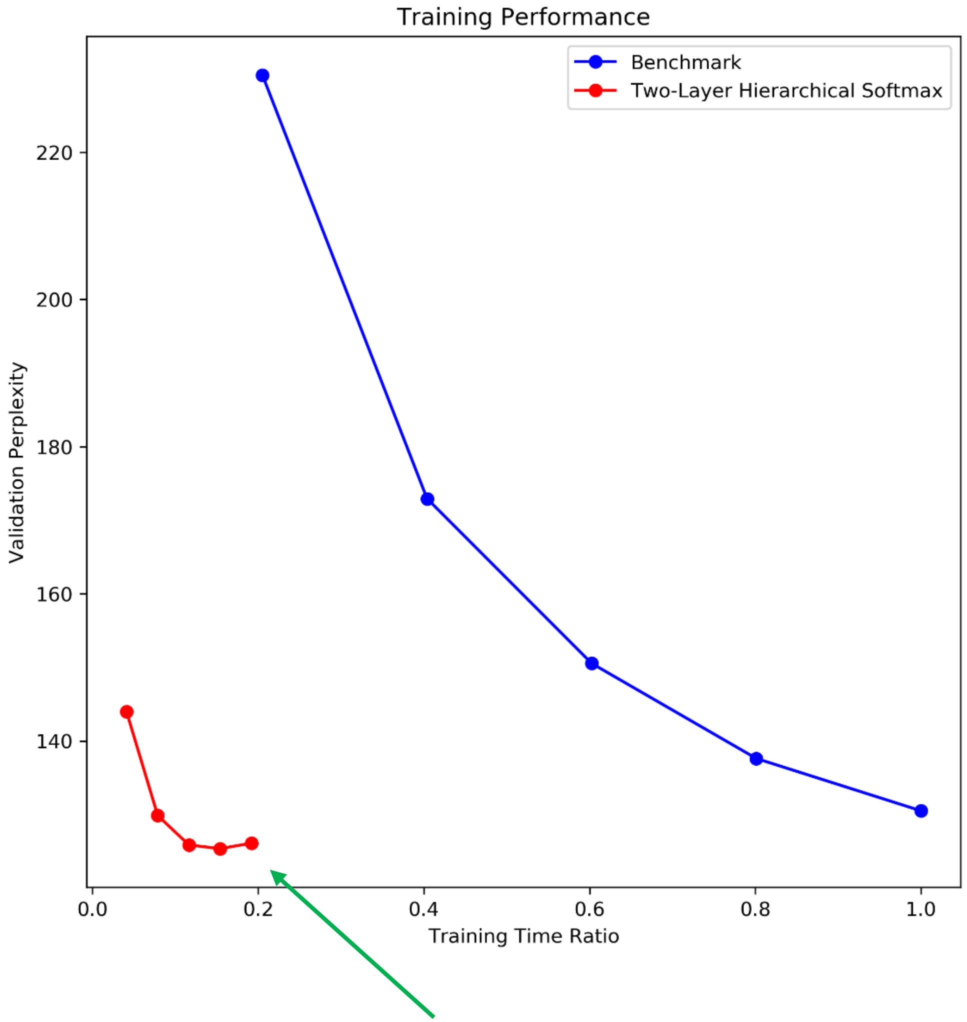
Top Softmax size: 100
Bottom Softmax size: 100
(10,000 = 100 x 100)

Use GRU, Adagrad
Tune lr, bptt, nhid, batch_size

The following settings are bad ☹
But could be implemented easily.

Class 1: Word 1-100
Class 2: Word 101-200
Class 3: Word 201-300
…
Class 100: Word 9901-10000

It's basically randomly grouping the words together…



Random grouping probably limits the model.

Generate a Sample Essay

mae 's british electronics and environmental health and activities law of money <eos> the bonn plan calls proponents in the customer through the strategy of the ongoing agreement to see the <unk> emergency and agreements <eos> the house expected panel use much regrets of negotiations that the u.s. was stopped by <unk> it is n't <unk> <eos> next next October the labor department said it would take an item in <unk> operation to additional N or $ N for the face value of $ N a share for decades <eos> if june N had more than N days such as top police execution vehicle u.s. abortion <eos> <unk> cutler was asked not to be assistant tobacco campaign <eos> in Germany for rome friday hint had <unk> as a labor department of running away from only N <unk> st. louis <unk> cohen and <unk> <unk> a more commuters used to <unk> five of soda lewis <unk> had been told to any congressional relations <eos> for years may reverse the nomination <eos> but mr. b. harrison also referred to the p.m. leader did n't be completed by penn air <eos> and he wrote his office judge greenspan may understand his trip at an antitrust case of the government and regulation <eos> now already as visible but more <unk> supporting markets and whole legislatures <eos> with declaring the foster builds he ca n't be made <eos> wrap a satisfaction is <unk> so in another time when opposition banning song in court <eos> in what attended us allow the board at seats on the won he made a sure to make <unk> and taiwan those technology was that the <unk> that had a $N in N americans and ca n't anxiety about $ N an emergency and drug peak <eos> also followed the plant by april N of the people male child <eos> but any conservative and wo n't help democrat sherwin with

Future Improvements

Dealing with Penn TreeBank dataset, given:

- The training of the model with Two-Layer Hierarchical Softmax is extremely **FAST** per epoch.
- Validation perplexity converges extremely **FAST.**
- Random grouping could even result in acceptable validation perplexity.

I expect using better grouping strategy will result in much better training performance both in the short-term and long-term.

- Cluster the words using fixed pre-trained word embeddings, and assign each word to the corresponding group before the training.
- Internally dynamically reorganize groups during training.

Conclusion

- As a higher level API, PyTorch still has long way to go to catch up TensorFlow and Theano (At least Theano will no longer be maintained in the near future is a good news to PyTorch).
- Coding with tensors in deep learning frameworks always makes people dizzy …

When coding tensor operations,

Thanks

All the codes are open-sourced at https://github.com/leimao.

This is not the “Depth Charge”.

This is the “Depth Charge”.

Character From “Beast War”