

Scaling Hidden Markov Models

Justin T Chiu and Alexander Rush
Cornell Tech

March 11, 2021

Hidden Markov Models in NLP

- ▶ Historically significant latent variable models
- ▶ Are thought to be very poor language models
- ▶ We show they are not!

Lessons from Large Neural Language Models

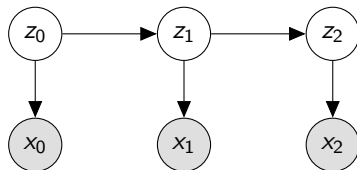
Large models perform better but are . . .

1. Slow to train
2. Prone to overfitting

We must overcome these issues when scaling HMMs

HMMs

For times t , model states $z_t \in \mathcal{Z}$, and tokens $x_t \in \mathcal{X}$,



We wish optimize

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z})$$

3 Techniques for Training Large HMMs

- ▶ Block-sparse emission constraints

↑ Speed

- ▶ Compact neural parameterization

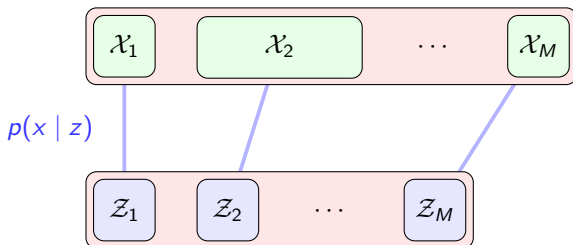
↑ Generalization

- ▶ State dropout

↑ Speed ↑ Generalization

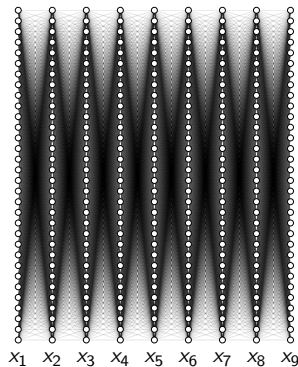
Technique 1: Block-Sparse Emission Constraints

- ▶ Reduce cost of marginalization by enforcing structure
- ▶ Partition words and states jointly
- ▶ Words can only be emit by states in same group

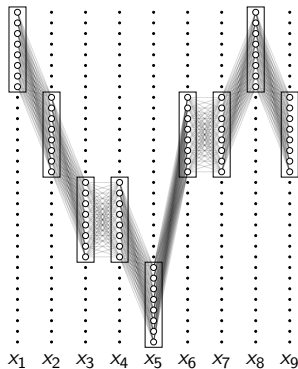


Block-Sparse Emissions: Effect on Inference

Given each word x_t , only the states in the correct group can occur



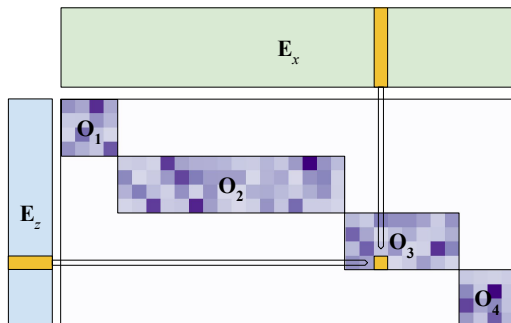
(a) No constraints



(b) Block-sparse emission

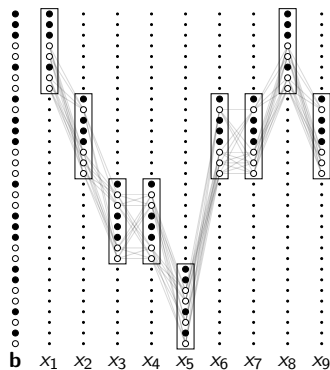
Technique 2: Neural Parameterization

- ▶ A neural parameterization allows for parameter sharing
- ▶ Generate conditional distributions from state \mathbf{E}_z and token representations \mathbf{E}_x



Technique 3: State Dropout

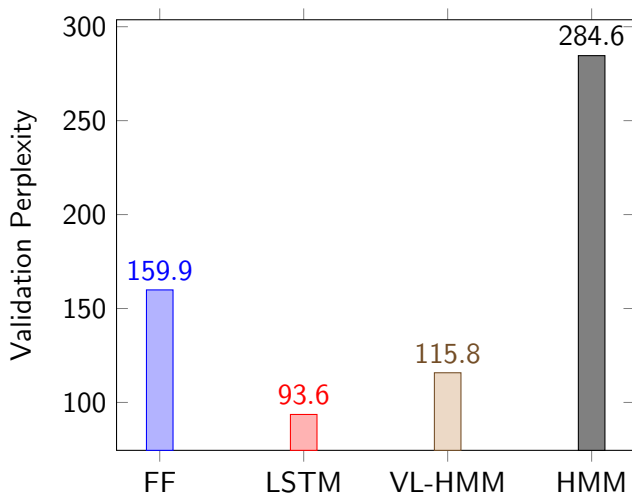
- ▶ State dropout encourages broad state usage
- ▶ At each batch, sample dropout mask $\mathbf{b} \in \{0, 1\}^{|Z|}$



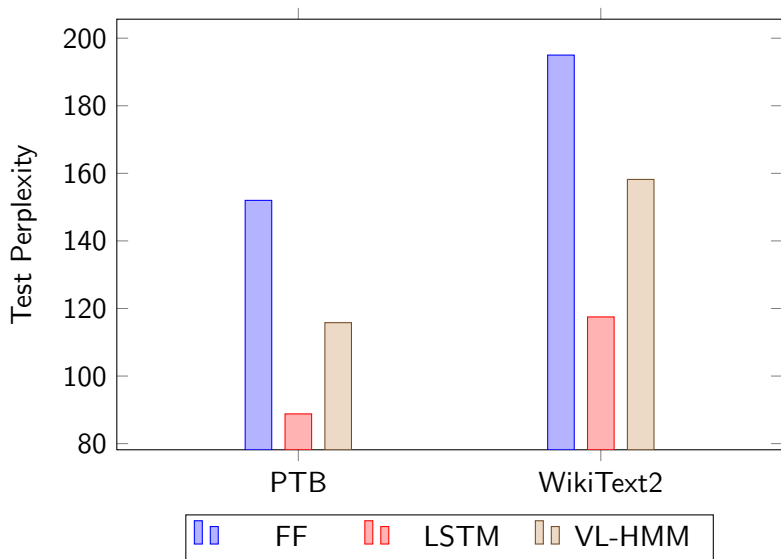
Experiments

- ▶ Language modeling on Penn Treebank and Wikitext-2
- ▶ Baselines
 - ▶ Feedforward 5-gram model
 - ▶ 2-layer LSTM
 - ▶ A 900 state HMM (Buys et al 2018)
- ▶ Model
 - ▶ 2^{15} (32k) state very large HMM (VL-HMM)
 - ▶ $M = 128$ groups (256 states each), obtained via Brown Clustering
 - ▶ Dropout rate of 0.5 during training

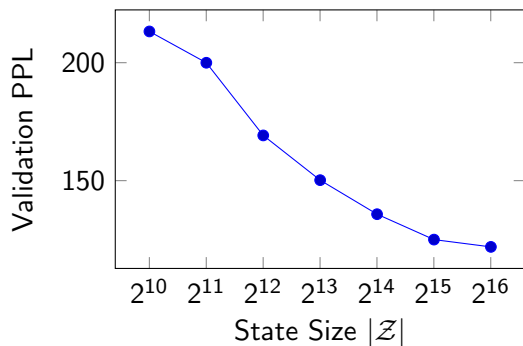
Results on PTB Validation Data



Results on PTB and WT2 Test Data



State Size Ablation



Validation perplexity on PTB by state size ($\lambda = 0.5$ and $M = 128$)

Other Ablations

Model	Param	Train	Val
VL-HMM (2^{14})	7.2M	115	134
- neural param	423M	119	169
- state dropout	7.2M	88	157

Embedded Structure Prediction



Generalized Softmax



Embedded Inference



Results

Conclusion

- ▶ HMMs are competitive language models
- ▶ Introduced 3 techniques for tackling speed and overfitting
- ▶ A great time to revisit other discrete latent variable models

EOS

Citations