Language Models Learn POS First

Naomi Saphra and **Adam Lopez**
n.saphra@ed.ac.uk alopez@ed.ac.uk
Institute for Language, Cognition, and Computation
University of Edinburgh

1 Introduction

A glut of recent research shows that language models capture linguistic structure. Linzen et al. (2016) found that LSTM-based language models may encode syntactic information sufficient to favor verbs which match the number of their subject nouns. Liu et al. (2018) suggested that the high performance of LSTMs may depend on the linguistic structure of the input data, as performance on several artificial tasks was higher with natural language data than with artificial sequential data.

Such work answers the question of *whether* a model represents linguistic structure. But how and when are these structures acquired? Rather than treating the training process itself as a black box, we investigate how representations of linguistic structure are learned over time. In particular, we demonstrate that different aspects of linguistic structure are learned at different rates, with part of speech tagging acquired early and global topic information learned continuously.

2 Methods

2.1 Concentration

We measure the degree to which a neural network has "structured" its representation x of a particular word in a sequence through *concentration*.

$$c(x) = \frac{\|x\|_2}{\|x\|_1} \quad (1)$$

The more similar in value the cells of x are, the smaller its l2/l1 ratio is. Thus if a neural network relies heavily on a small number of cells in an activation pattern, the activation is very concentrated. Likewise, a concentrated gradient is mainly modifying a few specific pathways. For example, it might modify a neuron associated with particular inputs like parentheses (Karpathy et al., 2015), or properties like sentiment (Radford et al., 2017).

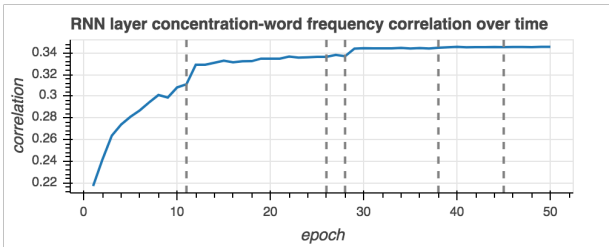

Figure 1: Correlation between mean concentration of a word gradient and word frequency. Vertical dashes mark when the optimizer rescales step size.

2.2 SVCCA

Existing work investigates how language model layers encode tags by training taggers on the activations produced by each layer (Belinkov et al., 2018). We use an alternative technique, SVCCA (Raghu et al., 2017), which interprets an arbitrary selection of neurons in terms of how they relate to another selection of neurons from any network run on the same input data. This method treats a selection of neurons as a subspace, spanned by their activations. Given any 2 sets of neurons, SVCCA projects the 2 distinct views of the same data onto a shared subspace which maximizes correlation between the 2 views.

Intuitively, if both views encode the same semantic information, the correlation in the shared subspace will be high. If the 2 views are encoding disjoint properties, the correlation will be low.

3 Experiments

All experiments are conducted on 1.6GB of English Wikipedia (70/10/20 train/dev/test split) with a 2-layer LSTM language model featuring tied weights in the softmax and embedding layers.

3.1 Gradient Concentration During Training

Over time, the model learns to shape weight structure around familiar words, with more frequent

Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 328–330
Brussels, Belgium, November 1, 2018.

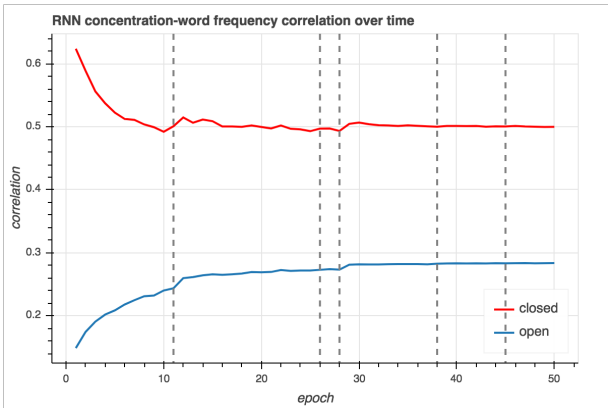

Figure 2: Correlation between mean concentration of a word gradient and word frequency.

words being more concentrated in their gradient. We can inspect this correlation between word frequency and concentration over time in the gradients passed backwards from the decoder layer to the RNN layer in Figure 1. It is clear that frequent words are more concentrated in their representation, and further that generally words become more concentrated in their representation over time. These observations support the idea that gradient concentration can measure the degree to which a word is relied on in shaping specialized structures within the representation.

However, Figure 2 shows that this correlation follows dramatically different trends for open POS classes (e.g., nouns and verbs) and closed classes (e.g., pronouns and prepositions). Initially, frequent words from closed classes are highly concentrated, but soon stabilize, while frequent words from open classes continue to become more concentrated. Why might this pattern emerge?

Closed classes offer clear signals about the current part of speech in a sequence. Open classes, however, contain words which are often ambiguous, such as "report", which may be a noun or verb. Open classes may also offer murkier syntactic signals because there are far more words that may occur in a particular open class POS role. We posit that early in training, closed classes are therefore essential for learning how to prototype syntactic structure, and are essential for shaping network structure. However, open classes are essential for modeling global sentence topic, so their importance in training continues to increase after part of speech tags are effectively modeled.

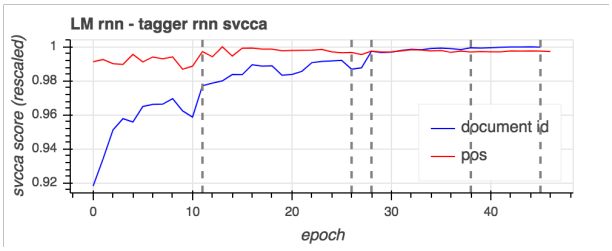

Figure 3: SVCCA correlation scores between LM and taggers. Values are rescaled so maximum score is 1.

3.2 Structure Encoding Over Time

Concentration experiments imply that a network first learns syntax, but topic significance continues to rise later. We test this claim directly.

As a proxy for syntactic representation, we use the task of POS tagging, as in (Belinkov et al., 2017). For document-global topic information, we classify the sequence by which Wikipedia article it came from. Both taggers are single layer LSTMs.

We applied SVCCA to the RNN layers of our language model and each tagger in order to find the correlation between the language model representation and the tagger representation. Indeed, Figure 3 illustrates that the POS structure is effectively represented immediately, and continues to be learned in the early stages of training before the first optimizer step size rescale. After that point, POS structure actually slightly declines and stabilizes below its peak value. Meanwhile, topic structure continues to increase over the course of training.

4 Conclusions

The SVCCA results imply that early in training, representing syntax and POS is the natural way to get initial high performance. However, as training progresses, these low-level aspects of linguistic structure sees diminishing returns from committing more parameters to their representation. Instead, later training realizes more gains from refining representations of global topic.

The concentration experiments tell the same story through a different lens. Early in training, structure is dictated by the closed POS classes, which give clear signals about syntax. However, small collections of directions within the network are increasingly responsive to words from open classes, which are more useful for modeling topic.

Our next step in this work is to develop ways of interpreting syntactic structures during training.

References

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do Neural Machine Translation Models Learn about Morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

Yonatan Belinkov, Llus Mrquez, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James R. Glass. 2018. Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks. *CoRR*, abs/1801.07772.

Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and Understanding Recurrent Networks. *arXiv:1506.02078 [cs]*. ArXiv: 1506.02078.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *arXiv:1611.01368 [cs]*. ArXiv: 1611.01368.

Nelson F. Liu, Omer Levy, Roy Schwartz, Chenhao Tan, and Noah A. Smith. 2018. LSTMs Exploit Linguistic Attributes of Data. *arXiv:1805.11653 [cs]*. ArXiv: 1805.11653.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to Generate Reviews and Discovering Sentiment. *arXiv:1704.01444 [cs]*. ArXiv: 1704.01444.

Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. 2017. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. *arXiv:1706.05806 [cs, stat]*. ArXiv: 1706.05806.