Exploring the impact of dependency length on learnability

Ayla Karakaş

Yale University

ayla.karakas@yale.edu

Background The *less is more* hypothesis posits that greater processing restrictions privilege children over adults as language learners [1]. Historically, this hypothesis has been both supported by simulating language learners as limited-capacity neural networks via staged input and memory interference [2], and challenged by conflicting results from replication attempts [3]. Decades later, the resources available for conducting such experiments have improved dramatically: an abundance of naturalistic data, and neural network architectures better suited to natural language tasks (e.g., LSTMs) than those of the past (i.e., SRNs). Mechanistic studies have also found the presence of specific neurons in LSTMs which appear to track grammatical functions, and error patterns that partially mimic human behavior [4], further motivating LSTMs as a reasonable model to test hypotheses about human language learning. This study thus revisits the *less is more* controversy with modern techniques. **Experiment I** extends foundational work showing LSTM ability to correctly classify English subject-verb number agreement over a variety of distances (as number of words between subject and verb) [5], which reported increasing error almost monotonically with distance, with no effect on overall error rate. Short distances were overrepresented in the data, so I replicated with (Fig. 1) and without (Fig. 2) controlling for equal distribution of distances (1-10; 19,200 training sentences in total), and found that the error curve was much flatter when distribution was controlled. The effect of learning from specific distances was then isolated by training on four contiguous subsets of three distances: `short` (1-3), `med1` (4-6), `med2` (7-9), and `long` (10-12). 1,600 examples per distance were included in training (4,800 total). 200 examples per distance included in training were used for validation (600 total), and 200 examples per distance from the full range (2,400 total) were used for testing. Sentences containing agreement attractors were excluded. If short strings play a distinguished role in learning, then generalization from short to long should be better than long to short. But the error rates of the models (Fig. 3) suggest that *less is* not *more* for LSTMs. **Experiment II** addresses why generalization from short to long distances was worse than from long to short distances. [6] showed that finite-precision LSTMs can learn to effectively implement counter automata, using the formal languages $a^n b^n$ and $a^n b^n c^n$, but not infinitely, because LSTMs implement an inherently imprecise counter. To evaluate counting imprecision in a context that better resembles our number classification task, I implemented an LSTM with five hidden units, and generated a dataset of 10,000/1,000/10,000 strings (t/v/t) of the language $a^n b^m$ for the objective: "does the string contain equal a's and b's?" The dataset was evenly split between positive ($n = m$) and negative (n or m is greater by 2) examples. One model was trained only on short strings (2-30 char.), the other only on long (32-60 char.). Both achieved 100% accuracy on classifying strings of the same type they were trained on, but the model trained on short strings was worse at classifying long strings (79% accuracy) than vice versa (99%). Fig. 4-7 display visualizations of the cell state of the LSTM. Unit 3 appears to be the counter unit for `train-short`, analogous to Unit 2 in `train-long`. Counting precision of `train-short` on the $a^{100}b^{100}$ strings is notably worse than that of `train-long` on the $a^{10}b^{10}$ strings. **Conclusion** LSTMs do not seem to show a benefit from learning from just "simple" data. Inclusion of short data in training may still be necessary, but not sufficient, for learning. This may be a fundamental limitation of the LSTM architecture posing a challenge for comparison with human learning.

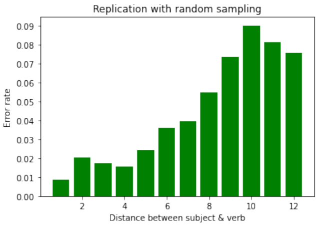


Figure 1: Replication with random sampling

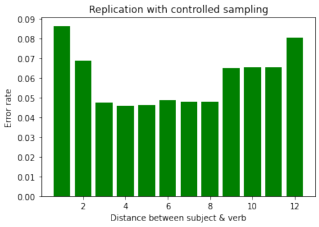


Figure 2: Replication with equal distribution of distances

Figure 3: Extrapolated distances

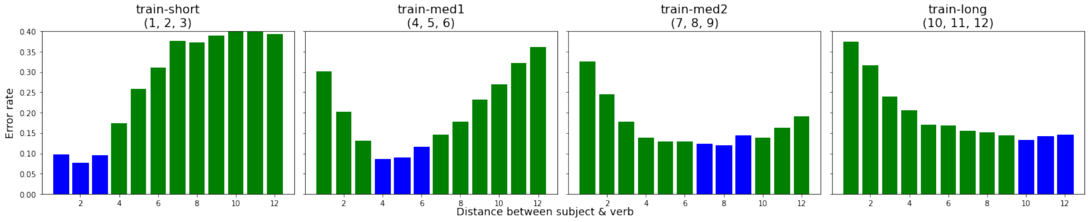


Figure 4: `train-short` $a^{10}b^{10}$

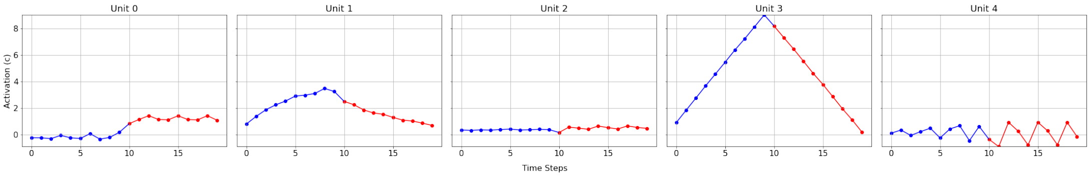


Figure 5: `train-short` $a^{100}b^{100}$

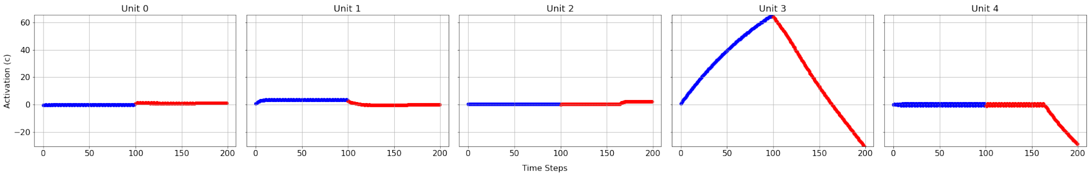


Figure 6: `train-long` $a^{10}b^{10}$

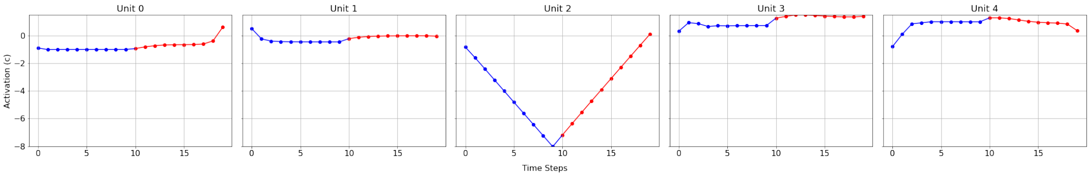
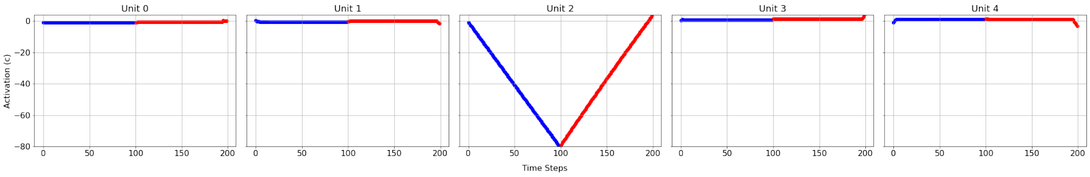


Figure 7: `train-long` $a^{100}b^{100}$



References

[1] Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive science*, *14*(1), 11–28.

[2] Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

[3] Rohde, D. L., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*(1), 67–109.

[4] Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., & Dehaene, S. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, *213*, 104699.

[5] Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, *4*(0), 521–535. https://transacl.org/ojs/index.php/tacl/article/view/972

[6] Weiss, G., Goldberg, Y., & Yahav, E. (2018). On the practical computational power of finite precision rnns for language recognition. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 740–745.