LSTMs Exploit Linguistic Attributes of Data
*Paul G. Allen School of Computer Science & Engineering; ○Department of Linguistics, University of Washington;
**Allen Institute for Artificial Intelligence; ○Department of Computer Science, University of Colorado Boulder

TL;DR
- Data with linguistic attributes helps LSTMs learn a non-linguistic memorization task.
- To solve the task, LSTMs use individual neurons to count timesteps.
- We hypothesize that LSTMs pick up on the patterns and structure in linguistic data and use them as additional noisy training signal.

Testbed Memorization Task
- Given a constant-length sequence of tokens, predict the identity of the middle token seen.
- This task is inherently nonlinguistic.

Training Datasets with Various Linguistic Attributes
1. **Language** setting
   - Directly take sequences
   - [Pierre Vinken, 61, years old, will, join the board as]
2. **n-gram** setting
   - Chunk corpus into pieces of size n (n = 2 in this example)
   - Permute the chunks
   - Split into sequences of desired length (4 in this example)
   - [join the Pierre Vinken, [61], years old, [board as], ...]
3. **Uniform** setting
   - Randomly pick tokens from vocabulary
   - [volume vice, a, pilots or sign corp., amid in <unk> banks]

Experiments
- Test data: uniform distribution over the 100 rarest words in the PTB.
- Ensures that models truly generalize and are not just using training data-specific features.

Models trained on data with linguistic features generalize better

What happens if we add more hidden units?
- Uniform data do poorly, even with more hidden units.

Further Analysis
- We further study an LSTM with 100 hidden units trained on Language, where train and test sequences are of length 300.

To solve the task, RNNs learn to count
- The RNN exploits linguistic features to bootstrap itself early in training and learns to generalize later.