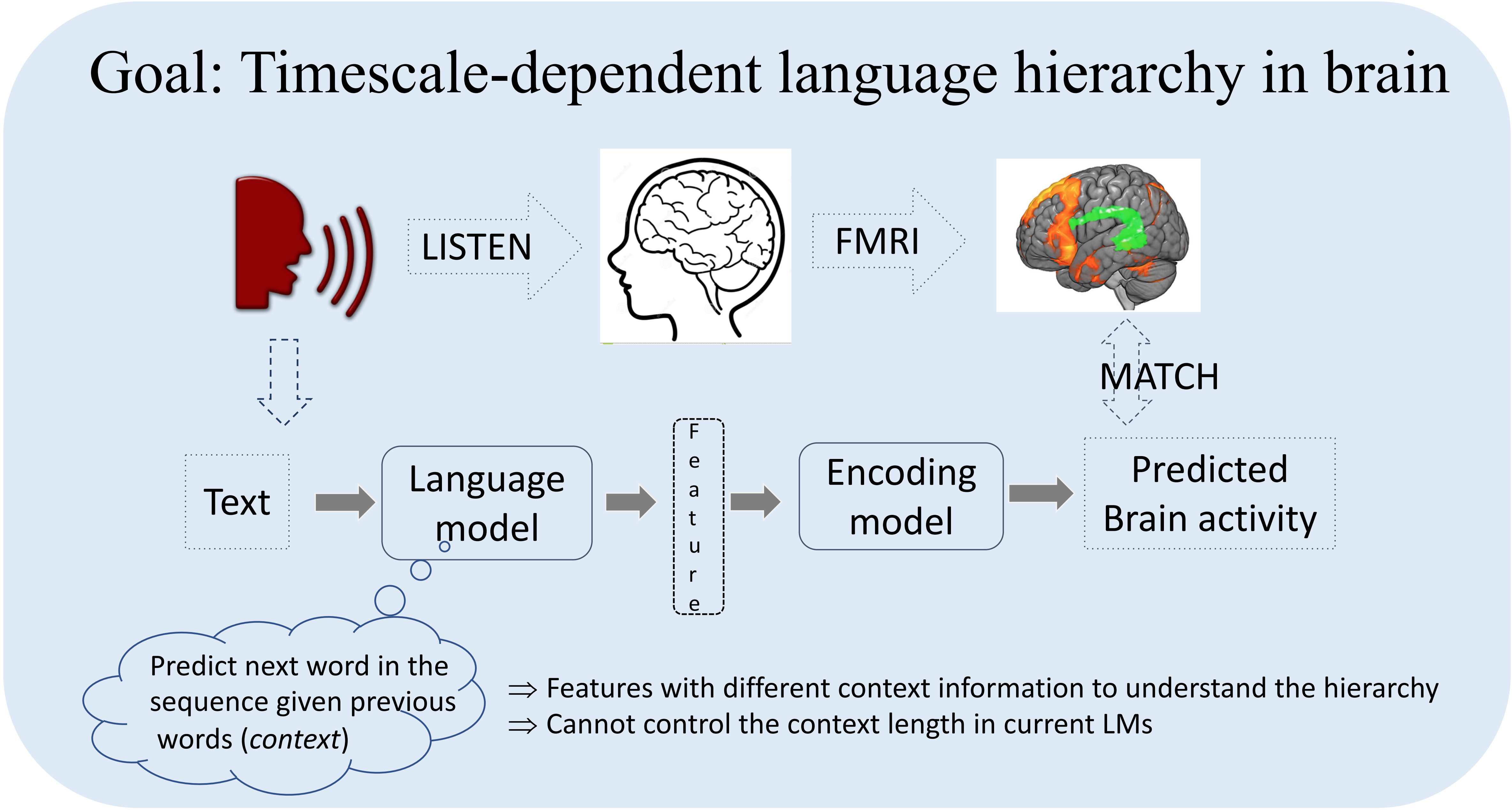
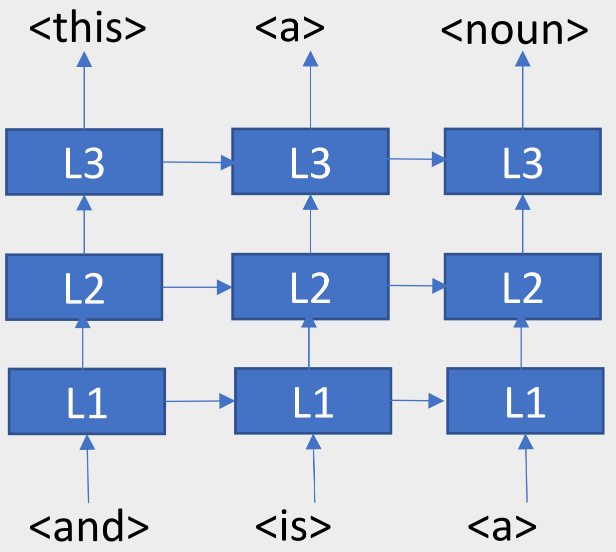
Timescale-dependent representation learning in Language Models

Shivangi Mahto


Goal: Timescale-dependent language hierarchy in brain
LISTEN
FMRI
MATCH
Text
Language model
Feature
Encoding model
Predicted Brain activity
Predict next word in the sequence given previous words (*context*)
⇒ Features with different context information to understand the hierarchy
⇒ Cannot control the context length in current LMs

Objective: LSTM LMs with explicit timescales



- Layers with different contextual memory
  L1: 20 words, L2: 10 words , L3: 5 words

- Layer units with different contextual memory
  L1[1-25]: 20 words; L1[25-50]: 10 words

Approach: Chrono Initialization*

- Forget gate **bias** initialization affects the memory of LSTM

- $b_f \sim \log\left(\mathcal{U}(1, T_{max} - 1)\right)$; $b_i = -b_f$

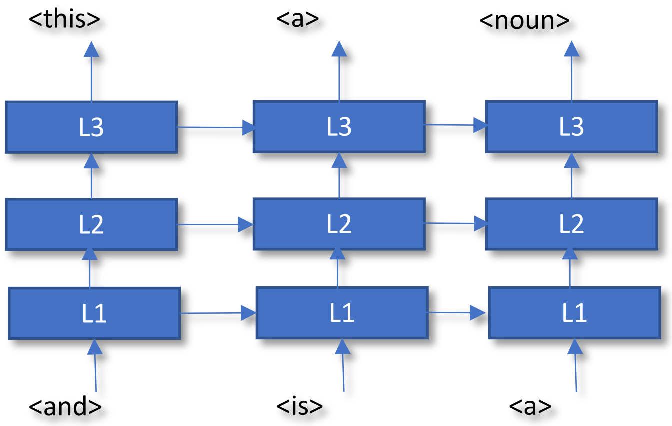
where $\mathcal{U}$ is uniform distribution and T_{max} is expected long term dependency

- Timescale-dependent LSTM layers: CHRONO LSTM LMs

$$\text{L1: } b_f \sim \log\left(\mathcal{U}(1{,}19)\right)$$
$$\text{L2: } b_f \sim \log\left(\mathcal{U}(1{,}9)\right)$$
$$\text{L3: } b_f \sim \log\left(\mathcal{U}(1{,}4)\right)$$

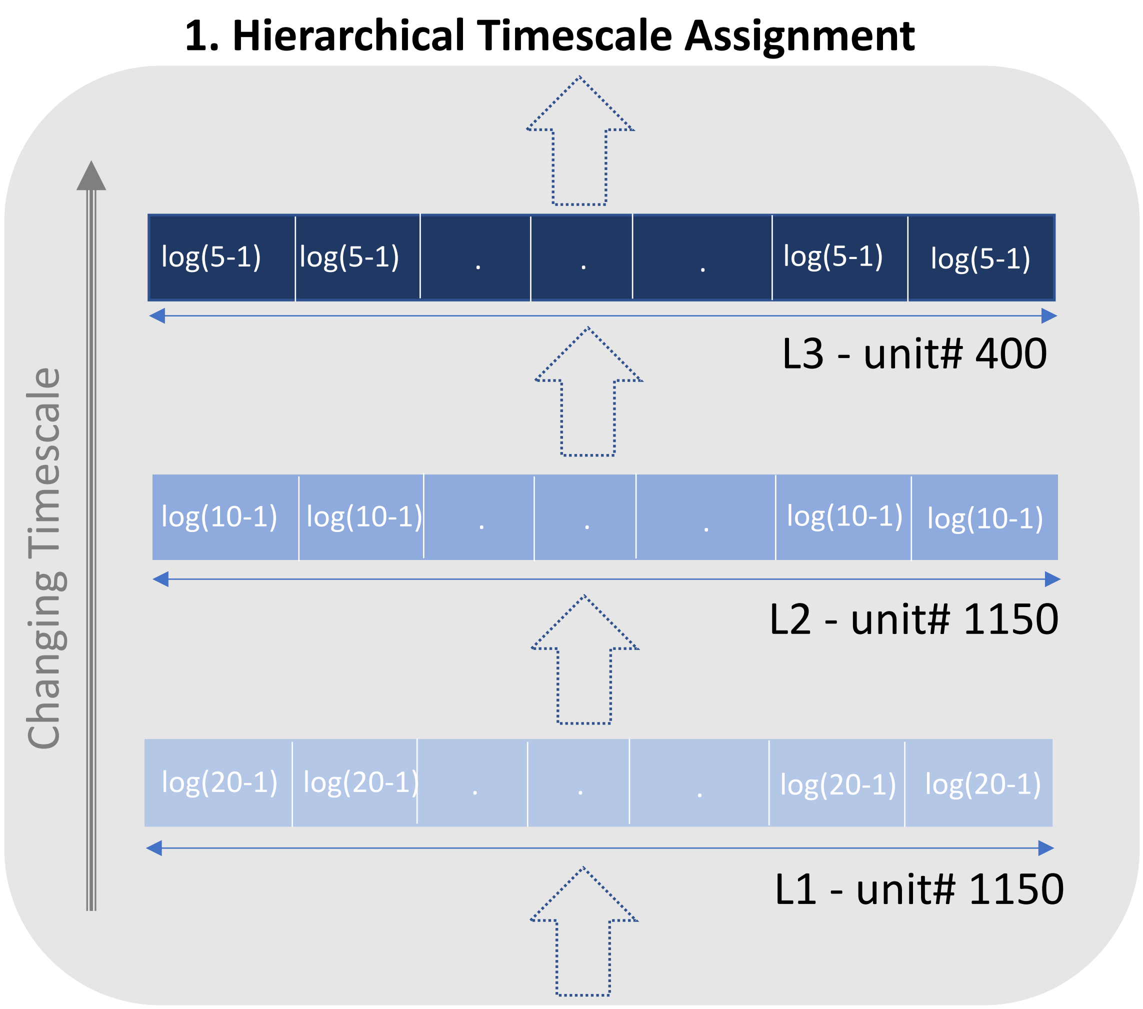
LSTM layer architecture

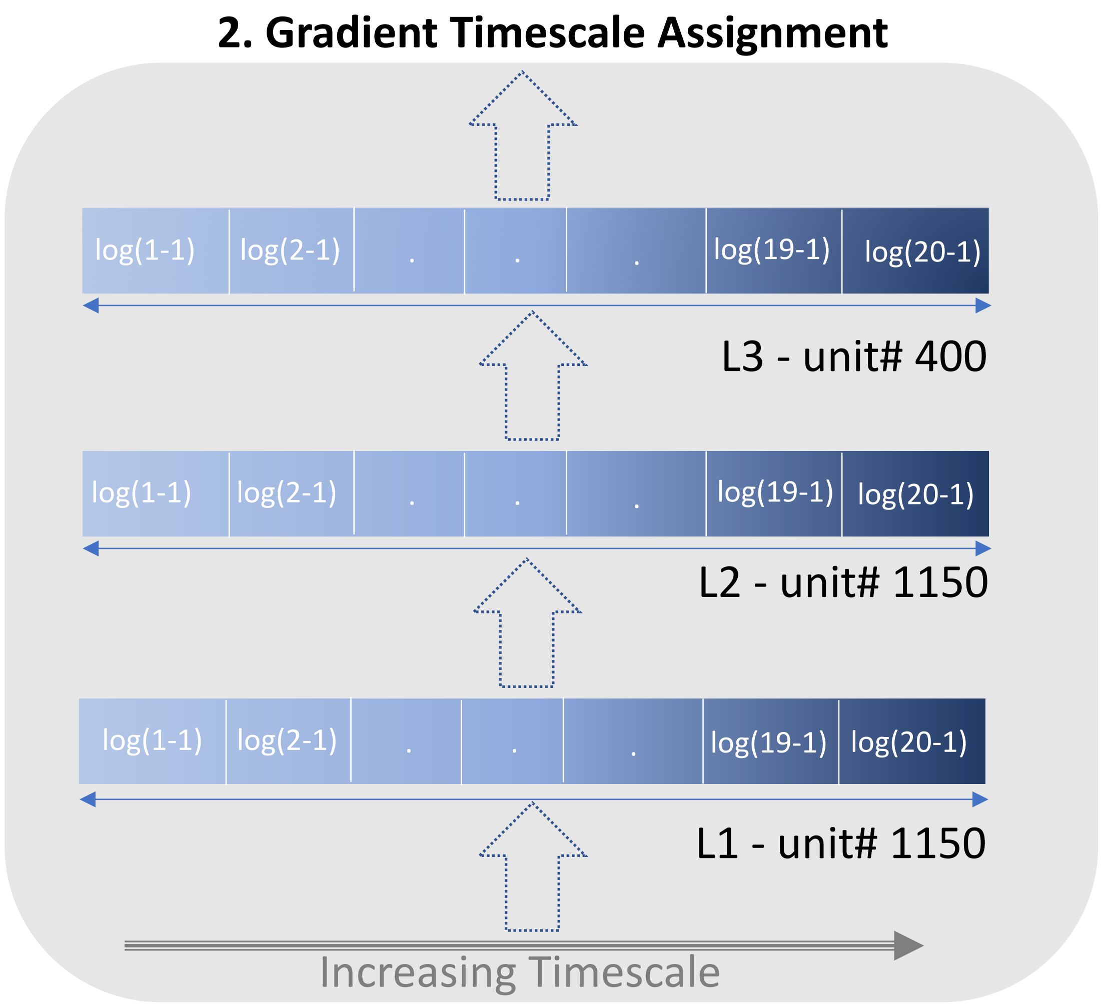
$$\begin{aligned}
i_t &= \sigma(U_i x_t + W_i s_{t-1} + V_i c_{t-1} + b_i),\\
f_t &= \sigma(U_f x_t + W_f s_{t-1} + V_f c_{t-1} + b_f),\\
g_t &= f(U x_t + W s_{t-1} + V c_{t-1} + b),\\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t,\\
o_t &= \sigma(U_o x_t + W_o s_{t-1} + V_o c_t + b_o),\\
s_t &= o_t \cdot f(c_t),\\
y_t &= g(V s_t + M x_t + d),
\end{aligned}$$



*Tallec, Corentin, and Yann Ollivier. "Can recurrent neural networks warp time?." *ICLR* 2018

Variations of Chrono-Initialization for LSTM-LM





Evaluation

- Compare Standard LSTM* and Chrono-LSTM LMs
- Datasets: Penn Tree Bank (PTB)**
  - 10k words in vocabulary, Training/Test (1177*70 words)/Validation
- Metric: Perplexity (PPL) of LMs
  - $2^{\frac{-1}{N}\sum_{t=1}^{N} \mathrm{LM}(w_t \mid w_1, w_2, \ldots, w_{t-1})}$
- Architecture of baseline:
  - 3 LSTM-layer language model – 1150, 1150, 400 (tied encoder-decoder)
  - Sequence length – 70
  - ASGD optimizer

* Merity, Stephen, Nitish Shirish Keskar, and Richard Socher. "Regularizing and optimizing LSTM language models."(2017).

** Mikolov, Tomáš, et al. "Subword language modeling with neural networks." (2012).

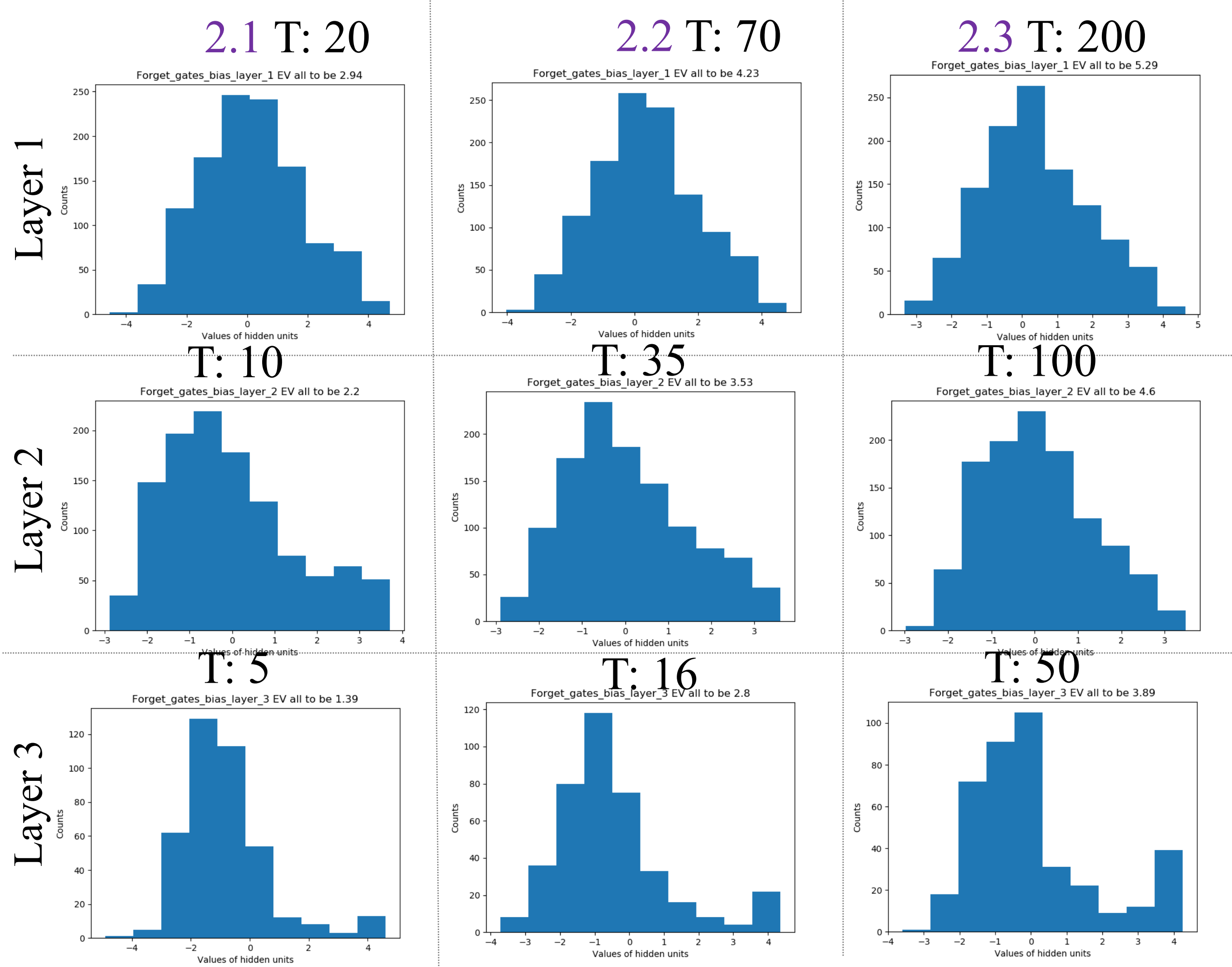
Experiment 1: Chrono-initialize forget gates biases

Model type	Test Perplexity	Description
1. Baseline	61.70	
2. Chrono layers initialization		Fixed layer-wise initialization - L1 log(T-1) - L2 log(T/2 - 1) - L3 log (T/4 -1)
1. T: 20	60.70	
2. T: 70	61.10	
3. T: 200	60.90	
3. Chrono gradient initialization		Gradient init. across all 3 layers For all L1, L2, L3: Init units with log(1----> T)
1. T: 20	60.73	
2. T: 70	61.17	
3. T: 200	61.28	
4. Inverse chrono layers initialization 1. T: 70	61.12	Fixed layer-wise initialization - L1 log(T/4-1) - L2 log(T/2 - 1) - L3 log (T-1)

Performance is not affected by the initialization at all!

What happened to forget gate biases after training?

- Forget gate bias distribution
  - not affected by initialization
  - specific to each layer rather than initial values
- Training overrides initial values
  - robust training
- What if we **fix** the forget gate bias values ?



Experiment 2: Fixed forget gates biases

Model type	Test Perplexity	Description
1. Baseline	61.70	
2. Fixed bias using Chrono layers initialization		Fixed layer-wise bias values - L1 log(T-1) - L2 log(T/2 - 1) - L3 log(T/4 -1)
1. T: 20	61.71	
2. T: 70	65.53	
3. T: 200	68.25	
3. Fixed bias using Chrono Gradient initialization		Gradient bias values across all 3 layers: Fixed units bias with log(1----> T-1)
1. T: 20	61.02	
2. T: 70	64.45	
3. T: 200°	67.23	

A T-dependent pattern among performances!

Lets, look closer into effect of different timescales in different layers…

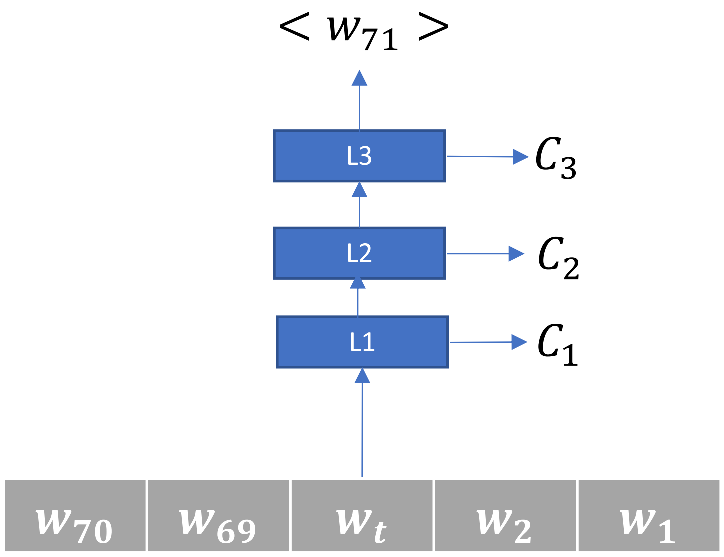
Experiment 3: Variation of timescale across layers

Model type	Test Perplexity	Variation in timescales Chrono Layers
1. Baseline	61.70	
2. First layer – longest timescale		Fixed layer-wise bias values - L1 log(T-1) – longest - L2 log(T/2 - 1) – midrange - L3 log (T/4 -1) – smallest
1. T: 20	61.71	
2. T: 70	65.53	
3. Middle layer – longest timescale		Fixed layer-wise bias values - L1 log(T/4-1) - smallest - L2 log(T - 1) - longest - L3 log (T/2 -1) – mid range
1. T:20	60.61	
2. T:70	63.15	
4. Last layer – longest timescale		Fixed layer-wise bias values - L1 log(T/4-1) – smallest - L2 log(T/2 - 1) – mid range - L3 log (T -1) – longest
1. T: 20	60.54	
2. T: 70	63.48	

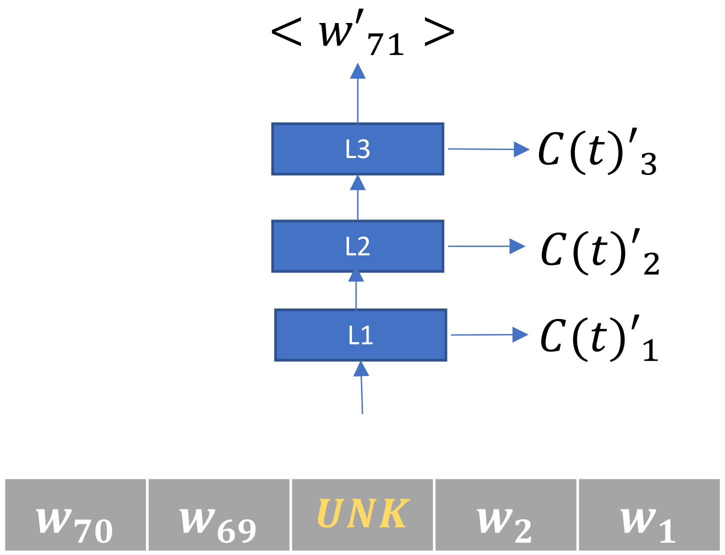
- Small timescale in first layer – most effective

Change in cell state vs. word position

- Ablate words at different positions from context during inference



Original context

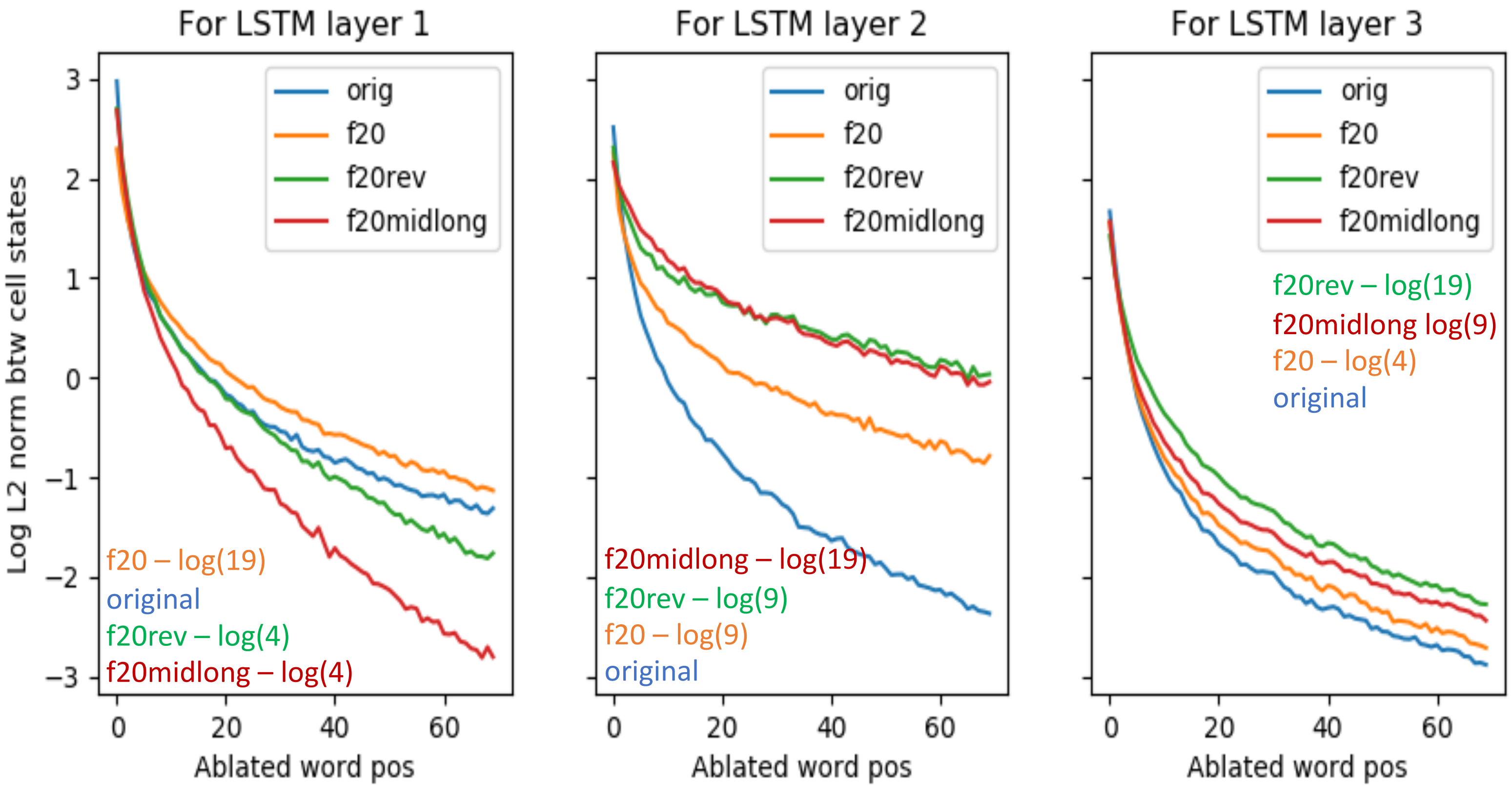


Ablated context

- Memory in layer between $< w_{71} >$ and $< w_t >$ = $\|C_i - C(t)'_i\|_2$

$$
\begin{aligned}
i_t &= \sigma(U_i x_t + W_i s_{t-1} + V_i c_{t-1} + b_i), \\
f_t &= \sigma(U_f x_t + W_f s_{t-1} + V_f c_{t-1} + b_f), \\
g_t &= f(U x_t + W s_{t-1} + V c_{t-1} + b), \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t, \\
o_t &= \sigma(U_o x_t + W_o s_{t-1} + V_o c_t + b_o), \\
s_t &= o_t \cdot f(c_t), \\
y_t &= g(V s_t + M x_t + d),
\end{aligned}
$$

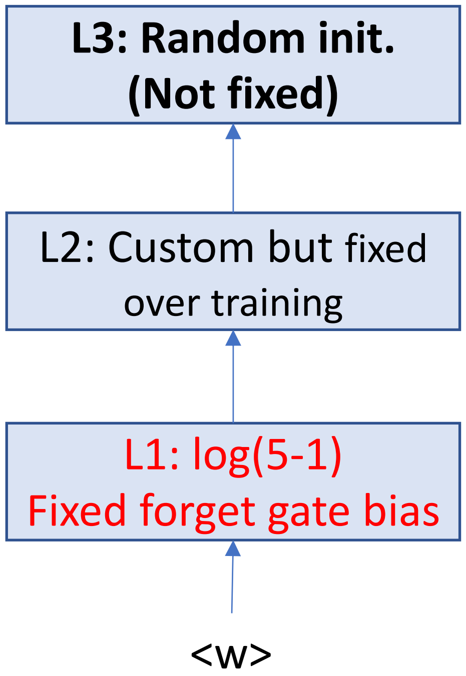
Change in cell memory vs. word position



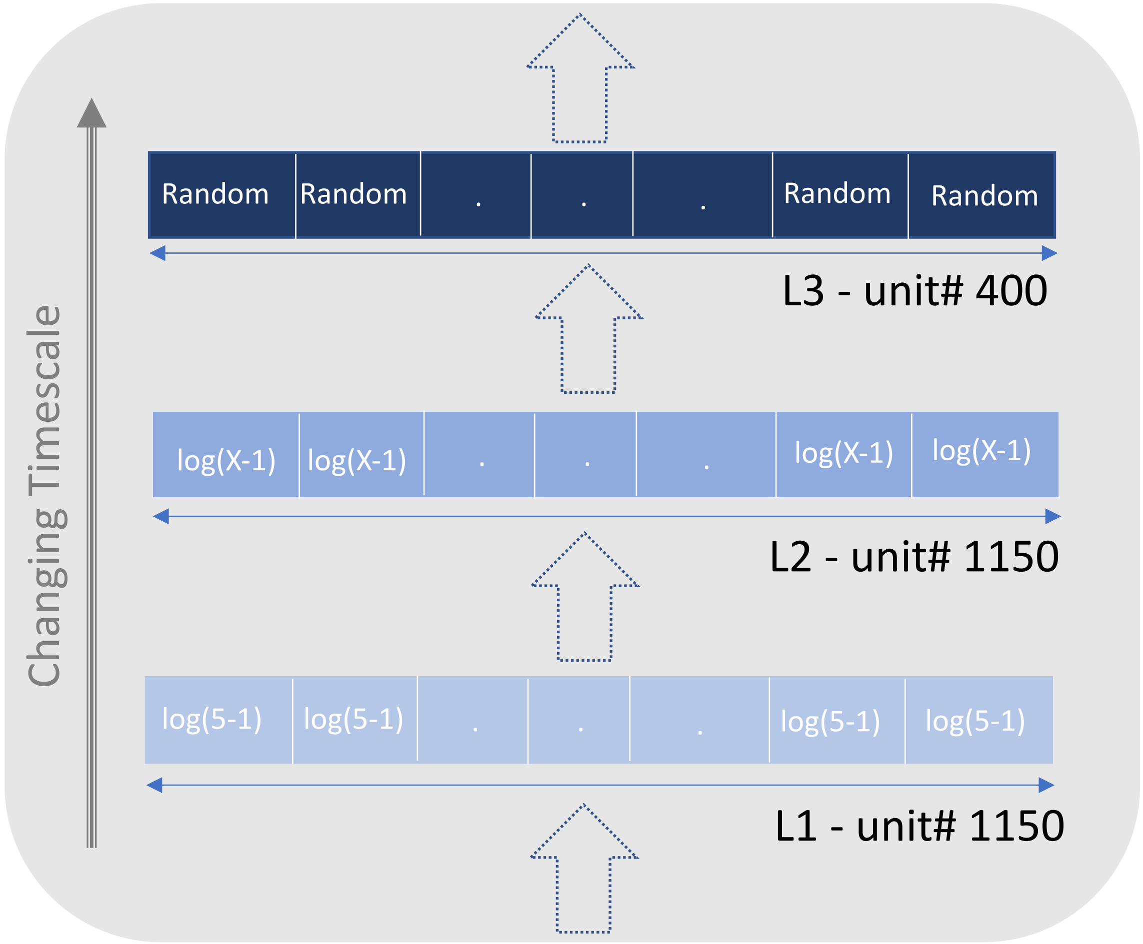
- Slope of graphs – information about timescale
  - Slow decay –> higher information between distant words

- Layer 1 – small timescale most effective – intuitive ..??
- Layer 3 – all model has similar slope – irrespective of fixed bias values – task-dependency ?

Experiment 4: Customize timescale in layers

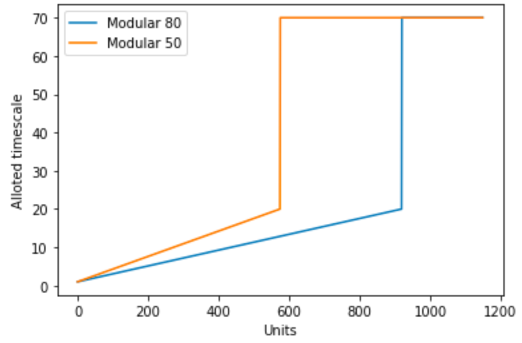


3. Customize Timescale Assignment

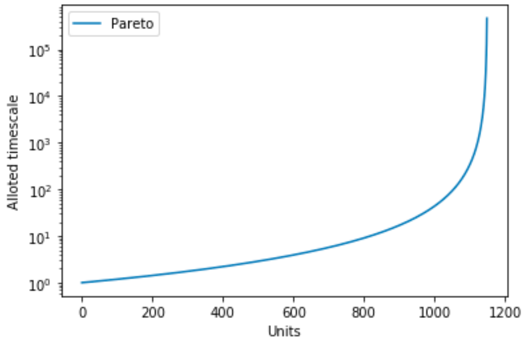


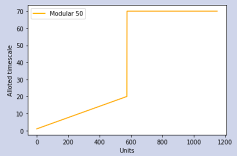
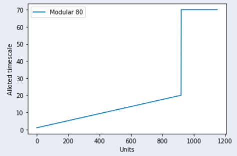
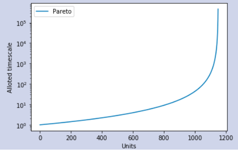
1. Tried different values of X:
   5, 10, 20, 30 ,70

2. Tried mixture of values:
   - 1—20 short memory
   - 70 – long memory

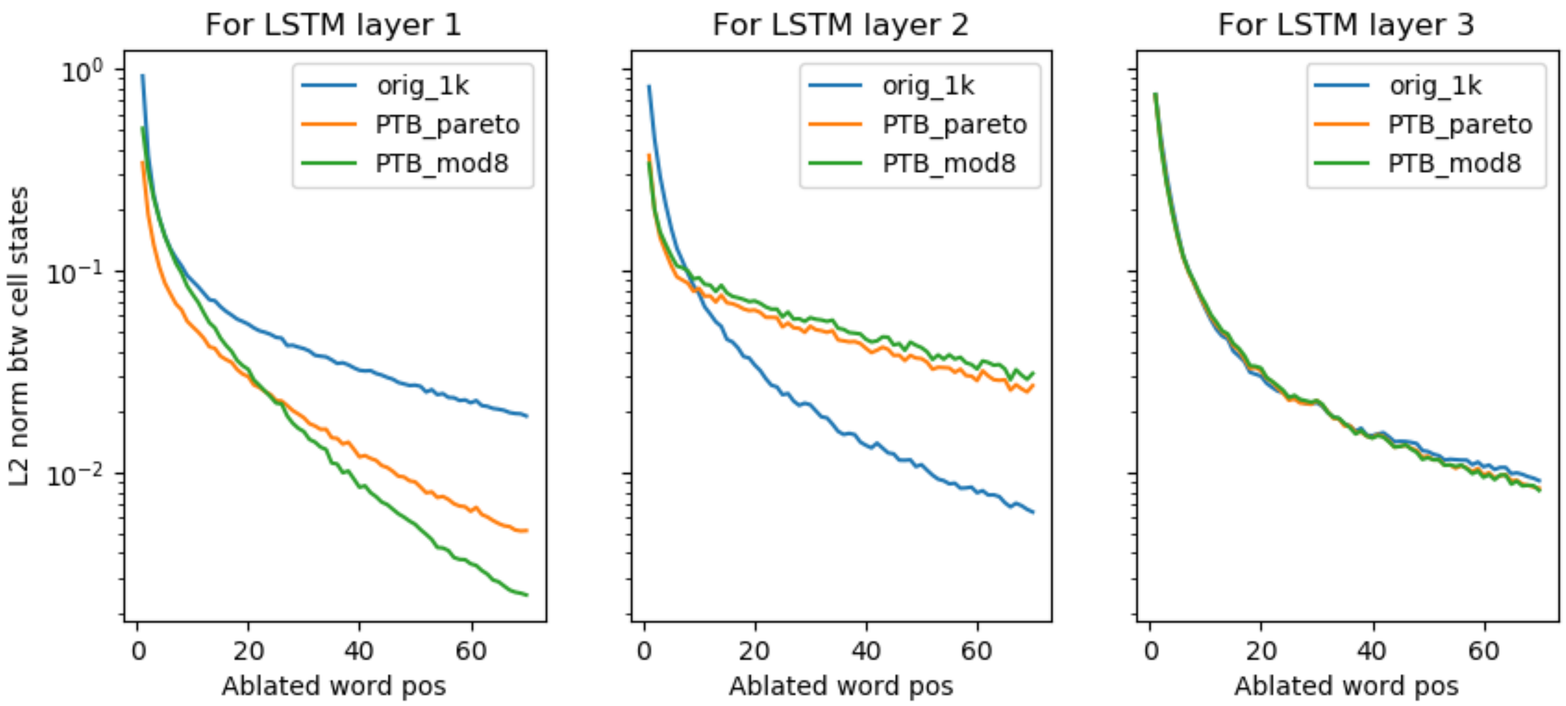


3. Pareto – smooth module



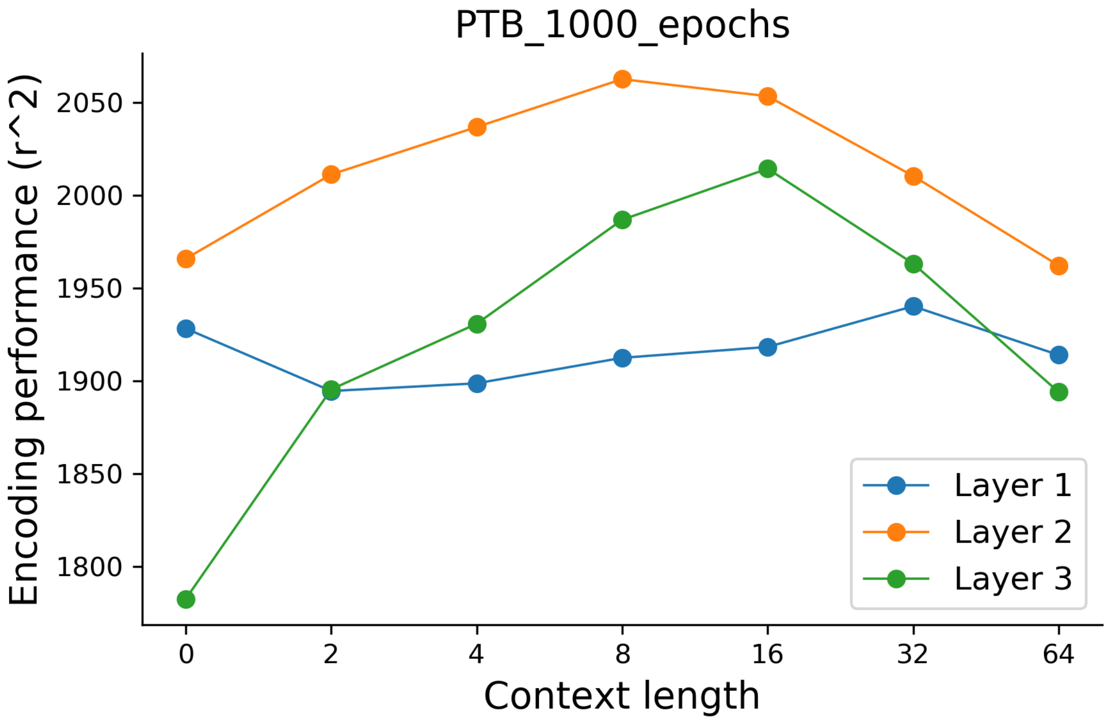
Model type	Test ppl	Bootstrapped samples performance	Averaged performance model trained with different seeds (6)
1. Baseline (1k epochs)	61.40	61.42 +/- 1.52	61.64 +/- 0.28
2. L2: Bias value: log(5-1)	60.17	60.19 +/- 1.50	
3. L2: Bias value: log(10-1)	59.88	59.90 +/- 1.49	
4. L2: Bias value: log(20-1)	59.97	60.00 +/- 1.50	
5. L2: Bias value: log(30-1)	60.09	60.12 +/- 1.49	
6. L2: Bias value: log(70-1)	60.71	60.73 +/- 1.52	
7. L2: Modular I (MOD 5) • Half units: Timescale gradient 1--> 20 • Other half units: Fixed timescale 70 	60.16	60.18 +/- 1.50	60.34 +/- 0.15
8 L2: Modular II (MOD 8) • 80% units: Timescale gradient 1 --> 20 • Other 20% units: Fixed timescale 70 	59.88	59.89 +/- 1.49	59.83 +/- 0.09
9 L2: PARETO : timescales sampled from pareto dist. (1/x**(1.54)) 	59.52	59.55 +/- 1.48	59.66 +/- 0.17

Cell state vs ablated word position



- Can we relate the behavior of cell state change to Mutual Information in Language?
- Can we understand how the LSTM is trying to fit with the mutual information in Language?

Encoding model



Thank you! ☺