MAPPING THE TIMESCALE ORGANIZATION OF NEURAL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

In the human brain, sequences of language input are processed within a distributed and hierarchical architecture, in which higher stages of processing encode contextual information over longer timescales. In contrast, in recurrent neural networks which perform natural language processing, we know little about how the multiple timescales of contextual information are functionally organized. Therefore, we applied tools developed in neuroscience to map the "processing timescales" of individual units within a word-level LSTM language model. This timescale-mapping method recovered a small set of units previously known to track long-range sequence dependencies, such as noun number or syntactic features, and revealed a new cluster of previously unreported long-timescale units. Next, we explored the functional role of units by examining the relationship between their processing timescales and network connectivity. We identified two classes of long-timescale units. "Controller" units composed a densely interconnected subnetwork and strongly projected to the forget and input gates of the rest of the network. "Integrator" units expressed connection profiles closer to the mean connection profile and produced few strong projections. Finally, we tested the generalization of these results from word-level models to character-level models. In summary, we demonstrated a model-free technique for mapping the timescale organization in neural network models, and we applied this method to reveal a hierarchical timescale organization within LSTM language models, including a "core" of units with strong gating projections and long processing timescales.

1 INTRODUCTION

Language processing requires tracking information over multiple timescales: to be able to predict the final word "timescales" in the previous sentence, one must consider both the short-range context (e.g. the adjective "multiple") and the long-range context (e.g. the subject "language processing") . How do humans and neural language models encode such multi-scale context information? Neuroscientists have developed methods to study how the human brain encodes information over multiple timescales during sequence processing. By parametrically varying the timescale of intact context, and measuring changes in the neural response, a series of studies Lerner et al. (2011); Xu et al. (2005); Honey et al. (2012) showed that higher-order regions are more sensitive to long-range context change than lower-order sensory regions. These studies indicate the existence of a "hierarchy of processing timescales" in the human brain. More recently, Chien & Honey (2020) used a time-resolved method to investigate how the brain builds shared context, when two groups of people are processing the same sentence preceded by different contexts. By directly mapping the time required for individual brain regions to converge on a shared representation in response to shared input, they again found that higher-order regions take longer time to construct this "shared context". Altogether, these and other lines of investigation suggest that sequence processing in the brain is supported by a distributed and hierarchical structure: sensory regions have short processing timescales and are primarily influenced by the current input and its short-range context, while higher-order cortical regions have longer timescales and track longer-range dependencies (Hasson et al., 2015; Honey et al., 2012; Chien & Honey, 2020; Lerner et al., 2011; Baldassano et al., 2017; Runyan et al., 2017; Fuster, 1997).

How are processing timescales organized within recurrent neural networks (RNNs) trained to perform natural language processing? Long short-term memory networks (LSTMs) (Hochreiter &

Schmidhuber, 1997) have been widely investigated in terms of their ability to successfully solve sequential prediction tasks. However, long-range dependencies have usually been studied with respect to a particular linguistic function (e.g. subject-verb number agreement (Linzen et al., 2016; Gulordava et al., 2018; Lakretz et al., 2019), and there has been less attention on the broader question of how sensitivity to prior context – broadly construed – is functionally organized within these recurrent networks. Therefore, drawing on prior work in the neuroscience literature, we demonstrate a model-free approach to mapping processing timescale in recurrent neural networks. We focus on existing language models that were trained to predict upcoming tokens at the word level Gulordava et al. (2018) and at the character level Hahn & Baroni (2019).

We found that the higher layers of LSTM language models contained a small subset of units which exhibit long-range sequence dependencies; this subset includes previously reported units (e.g. a "syntax" unit (Lakretz et al., 2019)) as well as previously unreported units. After mapping the timescale organization, we asked: how does processing timescale relate to the functional role of each unit in the network? Using a network-theoretic approach, we distinguished two types of long-timescale units: (1) "controller" units which have distinct and strong projections to affect the gating behavior of other units and (2) "integrator" units which have few strong projections and whose connection patterns are closer to the mean pattern. These findings advance our understanding of the timescale distribution and functional organization of LSTM language models, and provide a method for identifying the subset of important nodes for representing and manipulating long-range contextual information.

2 Related Work

Linguistic Context in LSTMs How do LSTMs encode linguistic context at multiple timescales? Although LSTMs are usually not designed with an explicit multi-scale architecture as observed in the human brain, it is possible that long-range and short-range information may be encoded in distinct subsets of LSTM cell state units. Prior work, however, suggested that the units sensitive to long-range dependencies are sparse. By ablating one unit at a time, Lakretz et al. (2019) found two units that encode information required for processing long-range subject-verb number agreement (one for singular and one for plural, respectively), and several other units for short-range number agreement. They further identified a "syntax unit", whose activation tracks the subject and its corresponding verb, regardless of the intervening distance; this "syntax" unit was significantly connected to both of the long-range number units, suggesting that information from the "syntax" unit may support the number units in their functions.

Overall, prior work suggests that a small subset of units track long-range dependencies. If there are very few such dependency-tracking nodes, this may limit the capacity of the models to process long sentences with high complexity, for reasons similar to those thought to limit human sentence processing (Lakretz et al., 2020).

Whole-network context dependence. Previous work by Khandelwal et al. (2018) investigated the duration of prior context that LSTM language models use to achieve word prediction. Context-dependence was measured by permuting the order of words preceding the preserved context, and observing the increase in model perplexity when the preserved context gets shorter. Khandelwal et al. (2018) found that up to 200 word-tokens of prior context were relevant to the model perplexity, but that the precise ordering of words only mattered within the most recent 50 tokens. The context-permutation method employed in this study was analogous to the approach used to measure context-dependence in human brain responses to visual movies (Hasson et al., 2008) and to auditory language Lerner et al. (2011). In the present study, we take a slightly different approach: Instead of entirely permuting the prior context (and preserve a variety of scales of context), we substitute alternative contexts (which is, by itself, a meaningful sentence), and we measure the time-taken for the effect of this prior context to "decay" (see Methods). In addition, we set out to measure the context-dependence at the level of single units in the model, rather than at the level of overall model perplexity.

Here, we build on these prior investigations of context-dependence by: (i) using a model-free approach to map multiple types of short- and long-range context-dependence (Chien & Honey, 2020); (ii) examining the processing timescales of individual units and (iii) by examining the relationship between processing timescales and the network connectivity of each unit. These investigations aim

toward the ultimate goal of developing model-free tools for mapping the flow of information from prior context to present processing within neural language models .

3 METHODS

3.1 LANGUAGE MODELS AND CORPUS

We evaluated the internal representations generated by a pre-trained word-level LSTM language model (WLSTM Gulordava et al. (2018)) as well as a pre-trained character-level LSTM model (CLSTM, Hahn & Baroni (2019)) as they processed sentences sampled from the 427804-word/1965719-character novel corpus: *Anna Karenina* by Leo Tolstoy [1] [2].

For WLSTM, we use the model made available by Gulordava et al. (2018). The model has an 650-dimensional embedding layer, two 650-dimensional hidden layers and an output layer with vocabulary size 5,0000. The model was trained on Wikipedia data, without fine-tuning to the novel. Considering that the model was trained and tested on Wikipedia sentences and may not be familiar with some sentences in the novel, we selected sentences with low perplexity comparable to sentences with similar structure sampled in the testing data in Gulordava et al. (2018).

For CLSTM, we use the model made available by Hahn & Baroni (2019). The model has an 200-dimensional embedding layer, three 1024-dimensional hidden layers and an output layer with vocabulary size 63. The model was trained on Wikipedia data with all character lower-cased and whitespace removed. We tested the model with sentences sampled from *Anna Karenina* as the WLSTM model, and we got bits-per-character (BPC) close to Hahn & Baroni (2019) reported in the original paper.

3.2 TEMPORAL CONTEXT CONSTRUCTION PARADIGM

In order to determine the processing timescales of cell state vectors and individual units, we modified the "temporal context construction" method developed by Chien & Honey (2020) . Thus, the internal representations of the model were compared across two conditions: (1) the intact context condition and (2) the random context condition. In both conditions, the model was processing the same shared sequence of words (for example, segment B), but the preceding sentence differed across the two conditions (Figure 1A, Appendix A.1). In the Intact Context condition, the model processed segment B (the shared segment) preceded by segment A, which was the actual preceding context from the original text. In the Random Context condition, however, the model processed the same shared input (segment B), but it was preceded by a randomly sampled context, segment X, which was sampled from the rest of the corpus. Segment X was therefore not usually coherently related to segment B.

In these two conditions, the model is processing the same input (the shared segment) with different context (the intact vs. random context). We can then examine how the cell state activation differ between the two conditions, while the network is processing the shared segments with identical input. Any difference in internal representations (i.e. hidden state activations) must arise from the context manipulation, since the current input is the same (i.e. Sentence B). Decrease in activation difference over time means that the cell state activation is less and less affected by the context. For a long-timescale unit, whose current state is dependent on information in the far-preceding context, we will see an activation difference is preserved across contexts (Figure 1B, green curve), even while the unit is processing the shared input. On the other hand, for a short-timescale unit, which is driven largely by the current input, we will see that the activation difference drop quickly (Figure 1B, red curve) as the unit processes the shared input.

4 HIERARCHICAL ORGANIZATION OF TIMESCALES ACROSS LAYERS

Do higher levels of the LSTM model exhibit greater context-dependence? Lakretz et al. (2019) observed that long-range functional units were more common in higher layers, and in general higher-levels of hierarchical language model exhibit longer range context-dependence (Jain et al., 2019;

[1](Tolstoy, 2016) translated from Russian to English by Constance Garnett.

[2]retrieved from http://www.gutenberg.org/ebooks/1399

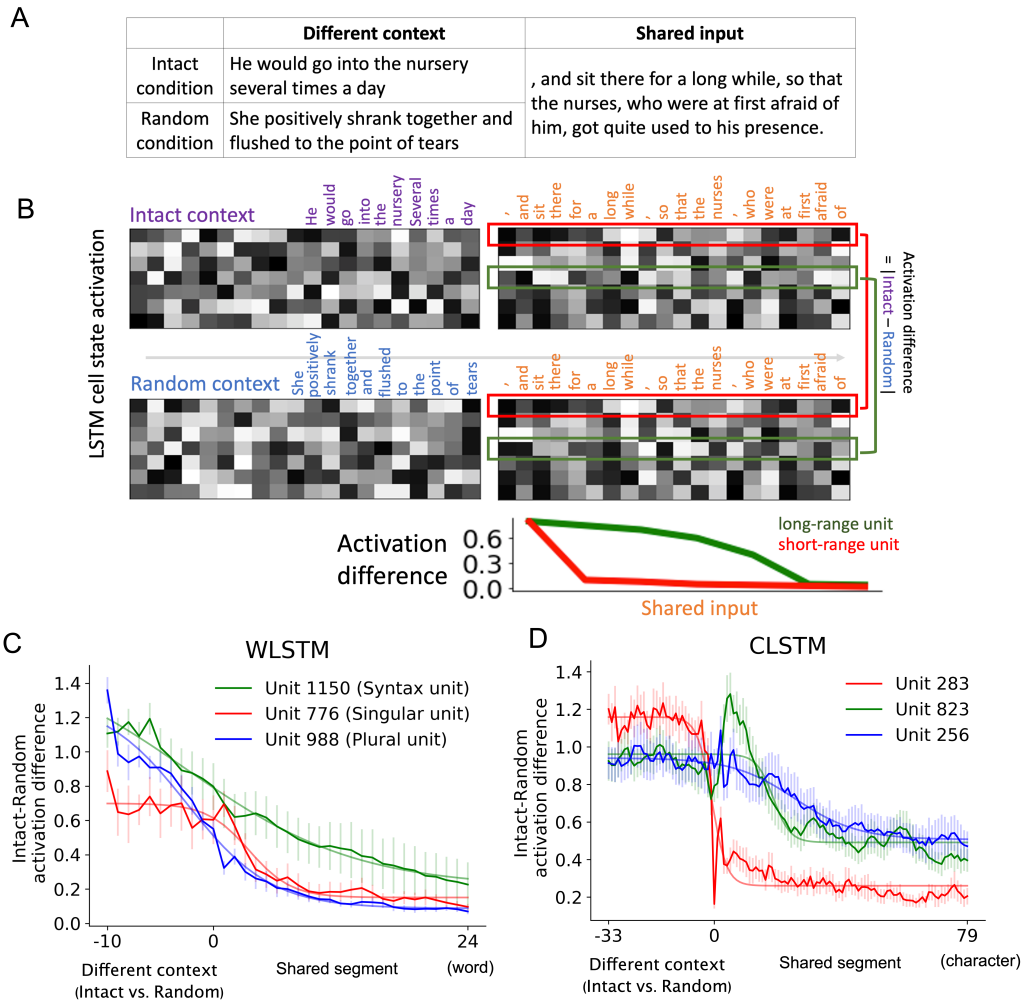

Figure 1: Context construction paradigm for mapping processing timescales of individual units and results from word-level LSTM (WLSTM) and character-level LSTM (CLSTM). **A.** Example sentences for the model to process in the intact and random condition. In the intact condition, the context - shared segment is an intact segment sampled from the novel, while in the random condition, the context sentence is replaced by randomly sampled sentence. **B.** Schematic hidden state representation of the neural network. When the model is processing different sentences, the hidden unit activation is different; when the model starts to process the shared segment, the hidden unit activation difference decreases over time with different rates. We got the activation difference by subtracting unit activation over time between the two conditions (e.g. rows showed in green and red). The expected decreasing pattern of activation difference of a long-timescale unit and a short-timescale unit are shown in green and red, respectively. **C.** Single unit activation differences across tokens for the WLSTM, with logistic curve fits overlaid. The error bars indicate 95% confidence interval across trials. The example units include the functional units reported by (Lakretz et al., 2019) and a long-timescale unit. **D.** Example unit activation difference results in CLSTM model.

Jain & Huth, 2018). Therefore, to validate our stimuli and the sensitivity of our methods, we first compared the processing timescales of different hidden layers in both of the LSTMs.

The higher layers of cell state activation patterns exhibited longer context-dependence scales, relative to the lower layers, in both the CLSTM and WLSTM. We correlated the cell state vectors, column by column, between the Intact condition and Random condition. We found that both layers showed near-zero correlation when processing the different context, and the correlation increase

as they began to process the shared input. In the WLSTM, the correlation increased more slowly for second-level cell state vectors than for first-level cell state vectors. Thus, the representation of second-level cell state is more sensitive to the different context than the first level. Similarly, for the CLSTM model, the third-level cell state exhibited longer-lasting context sensitivity than lower levels (Figure A.1).

This observation of longer context-dependence in higher stages of processing is consistent with prior machine learning analyses (Lakretz et al., 2019; Jain & Huth, 2018) and is also analogous to what is seen in the human brain (Hasson et al., 2015; Chien & Honey, 2020; Lerner et al., 2011; Jain et al., 2019). However, our goal in this study was to examine the functional organization at a finer scale, testing the temporal dependence of individual units within individual layers of the model. Based on the finding of longer processing timescales in higher layers, we examined single units in the highest level hidden units, i.e. the second level of WLSTM (n=650) and the third level of CLSTM (n=1024).

5 Processing Timescales of Individual Units Within LSTM Layers

5.1 Quantifying Timescales via the Rate of Cross-Context Activation Alignment

We examined the absolute single unit activation difference when processing the shared segments preceded by different context. As expected, most of the hidden units showed different activation when the input tokens were different (i.e. while processing the non-shared context in the Intact Context or Random Context conditions). However, once the shared input tokens begin (at $t = 0$) the Intact-Random activation differences drop (Figure 1C, 1D).

We use the rate at which the curves drop to quantify the processing timescale, as this is a measure of how quickly the responses align across differing contexts (Figure 1C, 1D). To quantify the timescale of individual units, we fit the activation difference curves with a logistic function:

$$Y(x) = \frac{L}{1 + e^{-k(x-x_0)}} + d \tag{1}$$

As shown in Figure 1C and Figure 1D, the logistic function fit the raw activation difference curves. We then computed the "timescale" of each unit as the time-to-half-maximum of the logistic decay. In particular, for the WLSTM we used the activation difference $Y(0)$ at the beginning of the shared segment, and at the end of the shared segment $Y(24)$ ($Y(79)$ for the CSLTM) to calculate the time-to-half-maximum of unit i as:

$$timescale_i = \lceil Y^{-1}(\frac{Y_i(0) - Y_i(24)}{2}) \rceil \tag{2}$$

where the inverse function $Y^{-1}(y)$ identifies the largest integer t, for which $Y(t) < y$. We included 635 units in WLSTM and 1012 units in CLSTM for further analysis after excluding the units which cannot be properly fitted by logistic function to quantify their timescales (See Appendix A.2).

5.2 Identifying Long-timescale Functional Units

As we were testing the same word-level LSTM previously studied by Lakretz et al. (2019), we began by examining the timescales of hidden-state units that were already known to be involved in processing context-dependence language information: a "singular number unit" 776, a "plural number unit" 988, and a "syntax unit" 1150. Previously, the "number" units (singular/plural) were shown to increase their activation when encountering the singular/plural nouns, ignoring in-congruent attractors and reacting to the corresponding singular or plural verbs (Lakretz et al., 2020). Most importantly, both "number" units received information from the "syntax" unit, which tracked long-range subject-verb correspondence. We found that compared to other units, the 2 "number" units have similar, medium timescales (3 words), while "syntax" unit has a longer timescale (7 words). This is consistent with the findings by Lakretz et al. (2019) that "syntax" unit, which requires a long timescale to track subject-verb dependency, provides information to "number" units (via strong hidden to input and forget gate projections), which do not necessarily have long timescales by themselves to keep the context.

5.3 Distribution of Processing Timescales in a Word-level LSTM

Most units were insensitive to long-range context, but we identified a substantial subset with long processing timescales. Of the 635 WLSTM units whose processing timescale we mapped, approximately 70% of the units were insensitive to long-range context (processing timescale < 3 words): their activation difference dropped immediately at onset of the shared segment. However, approximately 13% of the units had a timescales equal to or longer than the syntax units (processing timescale > 7 words)(Figure A.2A).

Figure 2A shows the absolute activation difference of all units in WLSTM sorted by timescale (long to short). Some of the longer-timescale units continued to exhibit a large activation difference even when processing the shared segments for more than 20 tokens. This suggests that some of the units have timescales longer than we measured here, consistent with previous findings that LSTM overall perplexity can be affected by up to 50 ordered context tokens (Khandelwal et al., 2018).

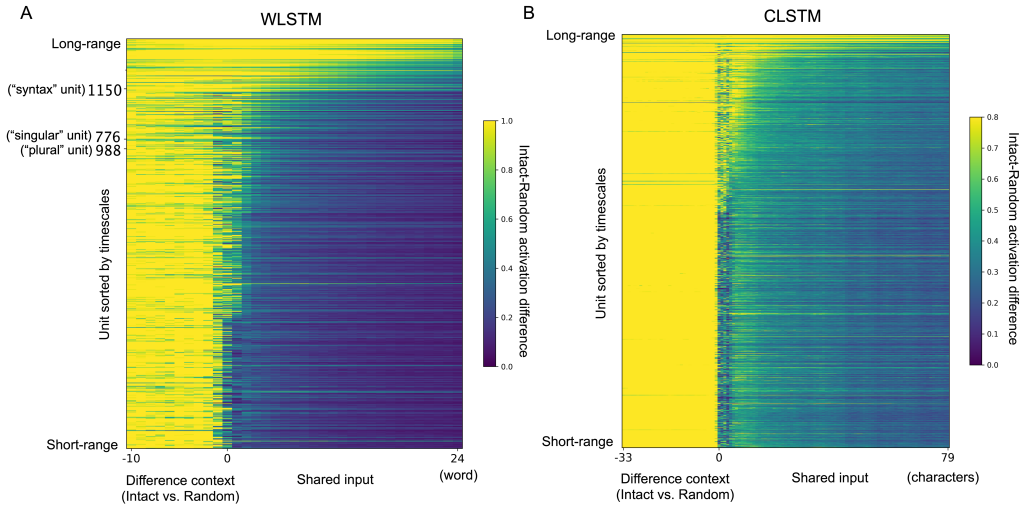

Figure 2: Timescale organization in word-level LSTM (WLSTM) and character-level LSTM (CLSTM) language model. **A.** Absolute activation difference for each WLSTM hidden unit over time, with units (rows) sorted by timescales. A small set of long-timescale units (top) sustain an activation difference during shared segment processing, but most (bottom) are context-insensitive short-timescale units. **B.** Absolute activation difference for each CLSTM unit over time, with units sorted by timescales. Similar to the WLSTM, a small set of long-timescale CLSTM hidden units maintain long-range contextual information.

5.4 Distribution of Processing Timescales in a Character-level LSTM

We repeated the timescale mapping in the CLSTM model, and again identified a small subset of long-timescale units (Figure 2B, Figure A.2B). Although there are overall more units in CLSTM, over 63% of the units are insensitive to the context (timescale < 3 characters). Fewer than 15% of the units exhibited timescale > 10 characters, and the unit with the longest timescale only dropped to its half-maximum activation-difference after 50 characters of shared input.

6 Connectivity of Units with Long and Medium Processing Timescales

Having mapped the timescales of each processing units, we next asked: do units with longer timescales exhibit different connection patterns from units with shorter processing timescales? To answer this question, we analyzed the direct weight projections from one hidden unit at time t to the input and forget gate of other hidden units at time $t + 1$. We set out to measure both (i) the

number of strong projections made by each unit, and (ii) the similarity of the overall connection profile (hidden-to-gates) across different hidden units.

In LSTMs, the amount of contextual (c_{t-1}) and input ($\tilde{c}_t$) information stored in the cell state (c_t) is determined by the forget gate (f_t) and input gate (i_t) activation (Eq. 3); and the activation of the gates i_t and f_t are determined by the current input at time t and the hidden units at time $t+1$ through weight matrices W_i and W_h (Eq. 4, 5).

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{3}$$

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \tag{4}$$

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \tag{5}$$

Thus, to understand how the contextual information over different timescales is projected from the hidden units to the input and forget gates of other units, and how this influences subsequent cell states, we next analyzed the network connectivity, focusing on the weight matrices W_{hi} and W_{hf} within the highest layer of the WLSTM or CLSTM.

6.1 Strong Hidden-Gate Projections from Long-Timescale Units

Units with longer processing timescales made a larger number of strong projections (|z-score|$>$ 5, Appendix A.3) to the input and forget gates of other units in both WLSTM (r=0.31, p$<$0.001, Figure 3A) and CLSTM models (r=0.24, p$<$0.001, Figure A.3A)

In WLSTM, the "syntax" unit (Unit 1150) reported by Lakretz et al. (2019) possessed the most strong projections to input and forget gates of all other units. Furthermore, the target units that Unit 1150 strongly project to are units with medium- to long-timescale units. Figure 3B illustrates all the z-scored weight projections from Unit 1150 to the forget gate and input gate, and the projected units are sorted by timescales.

6.1.1 Controller and Integrator Units in a Word-level LSTM

The presence of strong projections from the "syntax" unit (Unit 1150) to other long-timescale units implies that the long-timescale units with strong projections may also densely interconnect. If so, this set of units may have an especially important role in determining how prior context is updated and how it is used to gate current processing. To identify such "controller units", we binarized the network by identifying a set of 258 "strong" projections (see Appendix A.3), which provided the edges for a network analysis. We then used k-core analysis (Batagelj & Zaversnik, 2003) to identify the "main network core" (the core with the largest degree) amongst these nodes (Figure 3C). At the maximal $k = 5$, the k-core analysis yielded a set of densely interconnected nodes, composed of many long-timescale and medium-timescale units (labeled in red in Figure 3A). We refer to this set as the "controller" set of the network.

Finally, we used multi-dimensional scaling (MDS) to visualize the similarity of connection pattern among units in LSTM. Figure 3D shows the 2-dimensional MDS space of the unit connection patterns of all units, color-coded by timescales. The size of the nodes indicate the number of strong projections from the node, and the lines indicate their input gate/forget gate projections. We observed that the "controller units" (labeled on Figure 3D) are positioned at different locations around the edge of the MDS space, suggesting that these units have distinct connection pattern from other "controller" units and compared to the rest of the network. We also found that several other long-timescale units which have few strong projections, are positioned in the center of the MDS space, suggesting that these "integrator" units have connection profiles that are close to the mean connection pattern.

6.1.2 Controller Units in a Character-level LSTM

Performing the same network analyses on the CLSTM yielded similar results, insofar as (1) the main core network was composed of disproportionately many long-timescale and medium-timescale units (Figure A.3A, A.3B) and (2) the connection pattern of these controller units was distinct from other units in the network, according to the MDS results (Figure A.3C). However, we did not observe the "integrator units" positioned in the center of the MDS space of the CLSTM as was observed in the WLSTM.

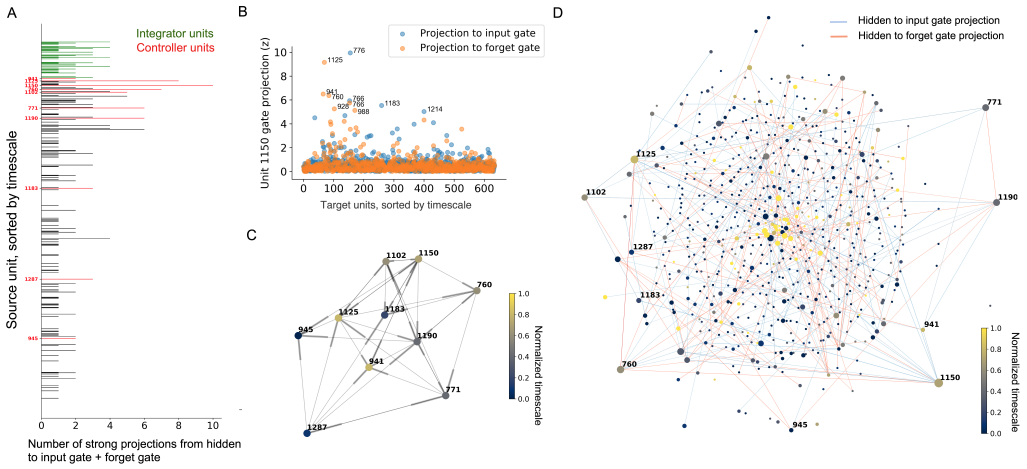

Figure 3: Timescale and connectivity organization in a word-level LSTM. **A.** Long-timescale units show stronger projections from hidden state at time t to the forget gate and input gate at time $t+1$. **B.** Strength of hidden-to-forget gate projections and hidden-input gate projections from high-degree Unit 1150 (a "syntax" unit), to all other units. The units receiving strong projections ($|$z-score$|> 5$) are labeled. **C.** Directed graph of the main core network formed by "controller" units in the WLSTM. **D.** Multi-dimensional scaling representation of network connectivity, with overlaid timescale and strong-connection information. The distance between two nodes indicates the similarity of their hidden-to-gate connection patterns. The size of each node indicates its degree (the number of strong projections from that node to the gate units). An edge between nodes indicates a significant projection between them (i.e. from one node to the gate controlling the other node).

7 Discussion

We demonstrated a new method for mapping the timescale organization in recurrent neural language models. We quantified processing timescales by measuring the number of shared tokens (words or characters) that are required for an individual unit to "align" its activation patterns, when the shared tokens were preceded by distinct contexts. Using this method, we mapped the timescale distributions of units within word-level and character-level neural language models, and identified subsets of 15% of units with long timescales. Furthermore, in both models, we identified a "controller" subnetwork, in which the units (i) possess unusually significantly strong projections to the forget and input gates of one another, and the rest of the network; and (ii) have unusually long processing timescale units on average. Altogether, we proposed a method for discovering timescale organization in RNN and applied it to reveal a hierarchical timescale organization within LSTM language models.

The set of units with long processing timescales included units whose role in long-range language dependencies was already established (Lakretz et al., 2019), but almost all the long timescale units are of unknown function. The timescale mapping procedure described here provides a model-free method for identifying nodes necessary for long-range linguistic and discursive processes (e.g. tracking whether a series of words constitutes an assertion or a question). Future studies of these neural language models could focus on the functions of long-timescale units, and especially the controller units, to understand how and why each unit is tracking long-range dependencies.

Processing timescales may exhibit an analogous hierarchical organization in LSTMs and in the human cerebral cortex: in both cases, a subset of nodes with high degree and high inter-connectivity express unusually long timescales. More detailed testing of this correspondence is required, however, because units within an LSTM layer are not spatially embedded, and thus may not show a spatially graded timescale topography.

Finally, we note that our analyses relied on a relatively small set of long sentences from a single novel; future applications using larger and more diverse text corpora will be required to generalize these findings.

REFERENCES

Christopher Baldassano, Janice Chen, Asieh Zadbood, Jonathan W Pillow, Uri Hasson, and Kenneth A Norman. Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721, 2017.

Vladimir Batagelj and Matjaz Zaversnik. An o (m) algorithm for cores decomposition of networks. *arXiv preprint cs/0310049*, 2003.

Hsiang-Yun Sherry Chien and Christopher J Honey. Constructing and forgetting temporal context in the human cerebral cortex. *Neuron*, 2020.

Joaquın M Fuster. Network memory. *Trends in neurosciences*, 20(10):451–459, 1997.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. *arXiv preprint arXiv:1803.11138*, 2018.

Michael Hahn and Marco Baroni. Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text. *Transactions of the Association for Computational Linguistics*, 7:467–484, 2019.

U. Hasson, E. Yang, I. Vallines, D. J. Heeger, and N. Rubin. A Hierarchy of Temporal Receptive Windows in Human Cortex. *Journal of Neuroscience*, 28(10):2539–2550, March 2008. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.5487-07.2008. URL `http://www.jneurosci.org/cgi/doi/10.1523/JNEUROSCI.5487-07.2008`.

Uri Hasson, Janice Chen, and Christopher J Honey. Hierarchical process memory: memory as an integral component of information processing. *Trends in cognitive sciences*, 19(6):304–313, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Christopher J Honey, Thomas Thesen, Tobias H Donner, Lauren J Silbert, Chad E Carlson, Orrin Devinsky, Werner K Doyle, Nava Rubin, David J Heeger, and Uri Hasson. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, 76(2):423–434, 2012.

Shailee Jain and Alexander Huth. Incorporating Context into Language Encoding Models for fMRI. *Advances in Neural Information Processing Systems*, pp. 6628–6637, 2018.

Shailee Jain, Amanda LeBel, and Alexander Huth. Improving language encoding of fMRI responses with transformers. *Annual Meeting of the Society for Neuroscience.*, 2019.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. Sharp nearby, fuzzy far away: How neural language models use context. *arXiv preprint arXiv:1805.04623*, 2018.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. The emergence of number and syntax units in lstm language models. *arXiv preprint arXiv:1903.07435*, 2019.

Yair Lakretz, Stanislas Dehaene, and Jean-Rémi King. What limits our capacity to process nested long-range dependencies in sentence comprehension? *Entropy*, 22(4):446, 2020.

Yulia Lerner, Christopher J Honey, Lauren J Silbert, and Uri Hasson. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8): 2906–2915, 2011.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Caroline A Runyan, Eugenio Piasini, Stefano Panzeri, and Christopher D Harvey. Distinct timescales of population coding across cortex. *Nature*, 548(7665):92–96, 2017.

Leo Tolstoy. *Anna karenina*. Lulu. com, 2016.

Jiang Xu, Stefan Kemeny, Grace Park, Carol Frattali, and Allen Braun. Language in context: emergent features of word, sentence, and narrative comprehension. *NeuroImage*, 25(3):1002–1015, April 2005. ISSN 10538119. doi: 10.1016/j.neuroimage.2004.12.013. URL http://linkinghub.elsevier.com/retrieve/pii/S1053811904007748.

A APPENDIX

A.1 DETAILS OF TRAINING AND TESTING

The context and shared segments in the Intact Context condition are connected by ", and" in long sentences from the novel. In this way, we ensured that the context segments and the shared segments have temporal dependencies which can be tracked by long-timescale units.

For the WLSTM timescale analysis, we chose long sentences in the Intact Context condition that satisfied the following constraints: (1) mean perplexity across all words in the sentence < 200, (2) the shared segment was longer than 25 words, and (3) the context segment was longer than 10 words. 77 sentences are included as trials in our analyses. To generate a paired "sentence" for the Random Context condition, we preserved the same shared segments and randomly sampled 30 context segments (each longer than 10 words) from other parts of the novel.

To compute the absolute activation difference at a given token position, (i) for each Intact Context sentence we computed the mean absolute activation-difference at that position with each of its 30 associated Random Context segments; and then (ii) we computed the mean of these absolute difference values across all 77 of the Intact Context sentences.

For the CLSTM timescale analysis, we used the same 77 long sentences in the Intact Context condition, and randomly sampled 25 context segments (with length ¿ 33 characters) for the Random Context condition. The mean bits-per-character (BPC) = 1.49, was close to the BPC reported in Hahn & Baroni (2019).

A.2 UNITS EXCLUDED FROM TIMESCALE ANALYSIS

We excluded 1 unit in WLSTM model and 5 units in CLSTM model which failed to be properly fitted using the logistic function; we further excluded 14 units in the WLSTM mode and 7 units in the CLSTM model which either did not show a near-zero activation difference before the shared segment started, or show an increasing activation difference when started to process the shared segment. After these exclusions, there were 635 units remaining in the WLSTM and 1012 units remaining in the CLSTM for further analysis.

A.3 IDENTIFYING STRONG HIDDEN-TO-GATE PROJECTIONS

First, for each unit, we concatenated its corresponding rows in the W_{hi} and W_{hf} matrices, to generate a single "hidden-to-gate" projection vector for that unit. Next we we z-scored the vector to get standardized projection values from that unit to all other units in the network. Using |z-score|> 5 as criterion, we identified a total of 258 "strong projections" from all hidden units to the input gate and forget gate in the WLSTM. The projection strength of each unit was then calculated based on its number of "strong projections" (Figure 3A). Next, we identified the edges corresponding to the top 258 magnitude weight values within the combined W_{hi} and W_{hf} matrices. Together, these edges formed a "strong-projection network". Finally, we used k-core analysis to identify the main core of the strong-projection network. This main core composed our "controller units" (Figure 3C).

Using the same criteria and method, we identified a total of 390 "strong projections" from all hidden units to the input gate and forget gate in the CLSTM. We then extracted the top 390 weight values from the weight matrices to construct a "strong-projection network" and again identified the main core network, composed the "controller units" for the CLSTM model (Figure A.3A, A.3B)

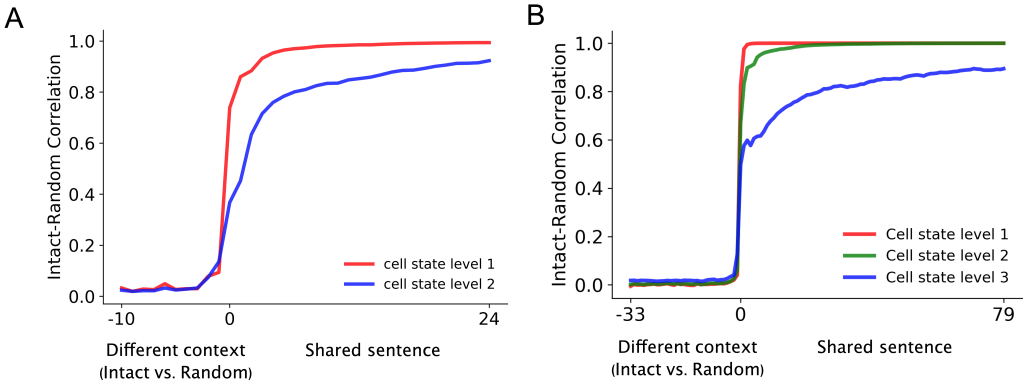

Figure A.1: Context construction across layers in word-level LSTM (WLSTM) and character-level LSTM (CLSTM) language models. **A.** Curve of the correlation of the WLSTM cell-state vectors across the Intact Context condition and Random Context condition as a function of input token. The tokens are ordered from left to right, most distant context (non-shared) to most recent context (shared); and token 0 marks the boundary between the shared and non-shared input. The correlation is near 0 when the model is processing different context segments, and increase towards 1 when the model starts to process the shared segment. Higher-level cell state showed a more gradual increase in correlation, compared to lower-level cell states, indicating that the higher-levels are more sensitive to the prior context. **B.** As for A, but applied to the three levels of CLSTM. Similar to the WLSTM, higher-level cell state of the CLSTM showed more context sensitivity than the lower-level cell state.

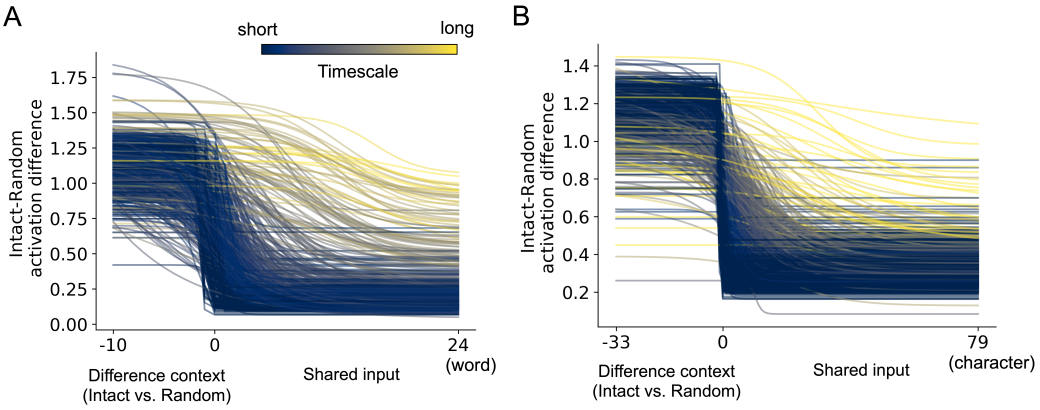

Figure A.2: Logistic fitted curves of activation difference in individual units in word-level and character-level LSTM language model, colored by timescale. **A.** Logistic fitted curves of activation difference over time of all units in word-level LSTM. The color indicates the timescale measured by full-width half-maximum (FWHM) of the curve. **B.** Logistic fitted curves of activation difference over time of all units in character-level LSTM.

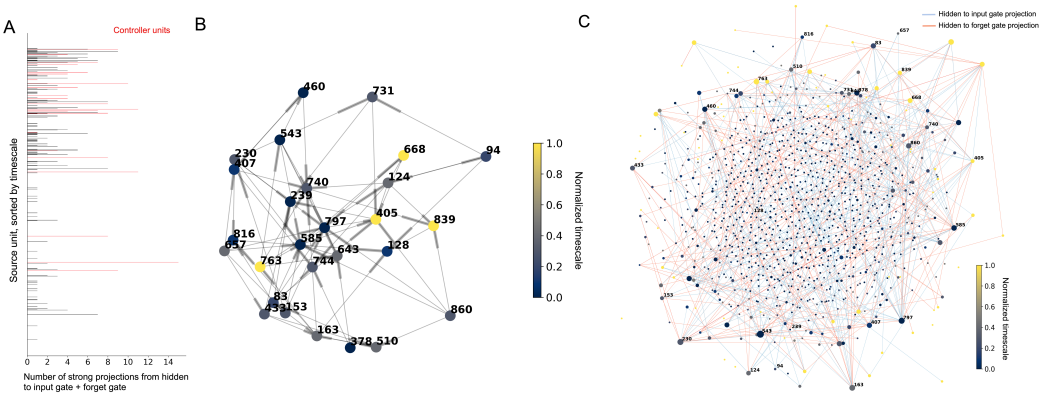

Figure A.3: Timescale and connectivity organization in a character-level LSTM (CLSTM). **A.** Longer-timescale units have overall stronger projections from the hidden units at time t to the forget gates and input gates at time $t+1$. **B.** The main core network ($k = 4$) formed by the controller units in the CLSTM. **C.** The multidimensional scaling space of the hidden-to-gate connection pattern of all units. The distance between nodes indicates their hidden-to-gate connection similarity; the size of the node indicates the number of strong projections from the node; and the line between two nodes indicates a significant projections between them.