Response Characterization for Auditing Cell Dynamics in Long Short-term Memory Networks

Ramin M. Hasani [*]
CPS, TU Wien
ramin.hasani@tuwien.ac.at

Alexander Amini [*]
CSAIL, MIT
amini@mit.edu

Mathias Lechner
CPS, TU Wien
mathias@cps.tuwien.ac.at

Felix Naser
CSAIL, MIT
fnaser@mit.edu

Radu Grosu
CPS, TU Wien
radu.grosu@tuwien.ac.at

Daniela Rus
CSAIL, MIT
rus@csail.mit.edu

Abstract

In this paper, we introduce a novel method to interpret recurrent neural networks (RNNs), particularly long short-term memory networks (LSTMs) at the cellular level. We propose a systematic pipeline for interpreting individual hidden state dynamics within the network using response characterization methods. The ranked contribution of individual cells to the network's output is computed by analyzing a set of interpretable metrics of their decoupled step and sinusoidal responses. As a result, our method is able to uniquely identify neurons with insightful dynamics, quantify relationships between dynamical properties and test accuracy through ablation analysis, and interpret the impact of network capacity on a network's dynamical distribution. Finally, we demonstrate generalizability and scalability of our method by evaluating a series of different benchmark sequential datasets.

1 Introduction

A key challenge for modern deep learning architectures is that of robust interpretation of its hidden dynamics and how they contribute to the system's decision making ability as a whole. Many safety critical applications of deep neural networks (NNs), such as robotic control and autonomous driving [27, 25, 8], require metrics of explainability before they are deployed into the real world. In particular, interpreting the dynamics of recurrent neural networks (RNNs), which can process sequential data, and are vastly used in such safety critical domains requires careful engineering of network architecture [21]. This is because investigating their behavior enables us to reason about their hidden state-dynamics in action and thus design better models.

The hidden state representations of long short-term memory (LSTM) networks [17], a subset of RNNs with explicit gating mechanisms, have been evaluated by gate-ablation analysis [9, 14] and feature visualization in linguistics [21, 35]. While these studies provide criteria for networks with interpretable cells, they are limited to feature visualization techniques, focus on hidden state dynamics in networks for text analysis, and thus suffer from poor generalizability. A robust, systematic method for assessing RNN dynamics across all sequential data modalities has yet to be developed.

In this paper, we introduce a novel methodology to predict and interpret the hidden dynamics of LSTMs at the individual cell and global network level. We utilize response characterization techniques [29], wherein a dynamical system is exposed to a controlled set of input signals and the associated outputs are systematically characterized. Concretely, we present a systematic testbench to interpret the relative contributions, response speed, and even the phase shifted nature of learned LSTM models. To analyze hidden state dynamics, we isolate individual LSTM cells from trained networks and

[*] Denotes co-first author

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

expose them to defined input signals such as step and sinusoid functions. Through evaluation of output attributes, such as response settling time, phase-shift, and amplitude, we demonstrate that it is possible to predict sub-regions of the network dynamics, rank cells based on their relative contribution to network output, and thus produce reproducible metrics of network interpretability.

For example, step response settling time delineates cells with fast and slow response dynamics. In addition, by considering the steady-state value of the cellular step response and the amplitude of the sinusoid response, we are able to identify cells that significantly contribute to a network's decision. We evaluate our methodology on a range of sequential datasets and demonstrate that our algorithms scale to large LSTM networks with millions of parameters.

The key contributions of this paper can be summarized as follows:

1. Design and implementation of a novel and lightweight algorithm for systematic LSTM interpretation based on response characterization;
2. Evaluation of our interpretation method on four sequential datasets including classification and regression tasks; and
3. Detailed interpretation of our trained LSTMs on the single cell scale via distribution and ablation analysis as well as on the network scale via network capacity analysis.

First, we discuss related work in Sec. 2 and then introduce the notion of RNNs as dynamic systems in Sec. 3. Sec. 4 presents our algorithm for response characterization and defines the extracted interpretable definitions. Finally, we discuss the interpretations enabled by this analysis in Sec. 5 through an series of experiments, and provide final conclusions of this paper in Sec. 6.

2 Related Work

Deep Neural Networks Interpretability - A number of impactful attempts have been proposed for interpretation of deep networks through feature visualization [10, 38, 37, 21, 35, 7]. Feature maps can be empirically interpreted at various scales using neural activation analysis [28], where the activations of hidden neurons or the hidden-state of these neurons is computed and visualized. Additional approaches try to understand feature maps by evaluating attributions [33, 11, 22, 36]. Feature attribution is commonly performed by computing saliency maps (a linear/non-linear heatmap that quantifies the contribution of every input feature to the final output decision). The contributions of hidden neurons, depending on the desired level of interpretability, can be highlighted at various scales ranging from individual cell level, to the channel and spatial filter space, or even to arbitrary groups of specific neurons [28]. A dimensionality reduction method can also be used to abstract from high dimensional feature maps into a low dimensional latent space representation to qualitatively interpret the most important underlying features [26, 6]. However, these methods often come with the cost of decreasing cell-level auditability.

Richer infrastructures have been recently developed to reason about the network's intrinsic kinetics. LSTMVis [35], relates the hidden state dynamics patterns of the LSTM networks to similar patterns observed in larger networks to explain an individual cell's functionality. A systematic framework has also been introduced that combines interpretability methodologies across multiple network scales [28]. This enables exploration over various levels of interpretability for deep NNs; however, there is still space to incorporate more techniques, such as robust statistics [23], information theory approaches [32], gradients in correlation-domain [15] and response characterization methods which we address in this paper.

Recurrent Neural Networks Interpretability - Visualization of the hidden-state of a fixed-structure RNNs on text and linguistic datasets identifies interpretable cells which have learned to detect certain language syntaxes and semantics [21, 35]. RNNs have also been shown to learn input-sensitive grammatical functions when their hidden activation patterns were visualized [19, 20]. Moreover, gradient-based attribution evaluation methods were used to understand the RNN functionality in localizing key words in the text. While these techniques provide rich insight into the dynamics of learned linguistics networks, the interpretation of the network often requires detailed prior knowledge about the data content. Therefore, such methods may face difficulties in terms of generalization to other forms of sequential data such as time-series which we focus on in our study.

Another way to build interpretability for RNNs is using the attention mechanism where the network architecture is constrained to attend to a particular parts of the input space. RNNs equipped with an attention mechanism have been successfully applied in image captioning, the fine-alignments in machine translation, and text extraction from documents [16]. Hidden-state visualization is a frequently shared property of all of these approaches in order to effectively understand the internals of the network. Hudson et al. [18] also introduced Memory, Attention, and Composition (MAC) cells which can be used to design interpretable machine reasoning engines in an end-to-end fashion. MAC is able to perform highly accurate reasoning, iteratively directly from the data. However, application of these modification to arbitrary network architectures is not always possible, and in the case of LSTM specifically, the extension is not possible in the current scope of MAC.

Recurrent Neural Networks Dynamics- Rigorous studies of the dynamical systems properties of RNNs, such as their activation function's independence property (IP) [5], state distinguishability [3], and observability [12, 13] date back to more than two decades. Thorough analyses of how the long term dynamics are learned by the LSTM networks has been conducted in [17]. Gate ablation analysis on the LSTM networks has been performed to understand cell's dynamics [14, 9]. We introduce the response characterization method, as a novel building block to understand and reason about LSTM hidden state dynamics.

3 Dynamics of Recurrent Neural Networks

In this section, we briefly we recap kinetics of RNNs. We denote the global dynamics of the hidden state values as h_t^l, with $t \in \{1..T\}$ denoting the time, and $l \in \{1..L\}$ representing the layers of the neural network. A *vanilla recurrent neural network* (RNN) can be formulated as [30, 21]:

$$h_t^l = \tanh\left(W^l \begin{pmatrix} h_t^l \\ h_{t-1}^l \end{pmatrix}\right), \tag{1}$$

where $W^{l\ [n \times 2n]}$ shows the weight matrix. h_t^0 retains an input vector x_t and h_t^L holds a vector at the last hidden layer, L, that is mapped to an output vector y_t which is ultimately the function of all input sequence $\{x_1, \ldots, x_T\}$.

RNNs are formulated as control dynamical systems in the form of the following differential equation (For the sake of notation simplicity, we omit the time argument, t):

$$\dot{h} = \sigma(Rh + Wx), \quad y = Cx, \tag{2}$$

where h denotes its internal state ($'\ \dot{}\ '$ illustrates time-shift or time derivative for the discrete and continuous-time systems, respectively), x stands for the input, and $R^{[n \times n]}$, $W^{[n \times m]}$ and $C^{[p \times n]}$ are real matrices representing recurrent weights, input weights and the output gains, respectively. $\sigma : \mathbb{R} \to \mathbb{R}$ indicates the activation function. In the continuous setting, σ should be locally Lipschitz (see [4] for a more detailed discussion).

3.1 Long Short-term Memory

Long short term Memory (LSTM) [17], are gated-recurrent neural networks architectures specifically designed to tackle the training challenges of RNNs. In addition to memorizing the state representation, they realize three gating mechanisms to read from input (i), write to output (o) and forget what the cell has stored (f). Activity of the cell can be formulated as follows [14]:

$$c_t^l = z \odot i + f \odot c_{t-1}^l \tag{3}$$

$$y_t^l = o \odot \tanh(c_t^l) \tag{4}$$

$$\begin{pmatrix} z \\ i \\ f \\ o \end{pmatrix} = \begin{pmatrix} \tanh \\ \sigma \\ \sigma \\ \sigma \end{pmatrix} W^l \begin{pmatrix} y_t^{l-1} \\ y_{t-1}^l \end{pmatrix} \tag{5}$$

where c_t^l is layer l's cell state at time t, W^{4n*2n} is the weight matrix, z stands for the input block, and y_t^l denotes the cell's output state.

For analytical interpretation of a dynamical system, the first necessary condition is to check its observability property. A dynamical system is observable if there is some input sequence that gives

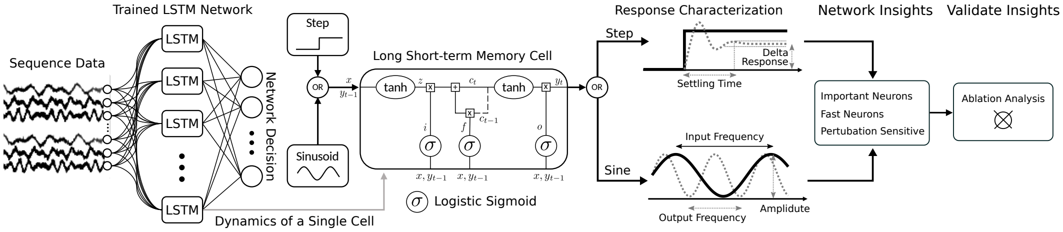

Figure 1: **Response characterization method for LSTM cells.** We take individual LSTM cells from a trained network, and characterize their step and sinusoidal response. These responses predict quantitative and interpretable measures for the dynamics of the single units within the network. We then validate the predictions by performing a neuronal ablation analysis.

rise to distinct outputs for two different initial states at which the system is started [34]. Observable systems realize unique internal parameter settings [5]. One can then reason about that parameter setting to interpret the network for a particular input profile. Information flow in LSTM networks carries on by the composition of static and time-varying dynamical behavior. This interleaving of building blocks makes a complex partially-dependent sets of nonlinear dynamics that are hard to analytically formulate and to verify their observability properties As an alternative, in this paper we propose a technique for finding sub-regions of hidden observable dynamics within the network with a quantitative and systematic approach by using response characterization.

4 Methodology for Response Characterization of LSTM cells

In this section, we explore how response characterization techniques can be utilized to perform systematic, quantitative, and interpretable understanding of LSTM networks on both a macro-network and micro-cell scale. By observing the output of the system when fed various baseline inputs, we enable a computational pipeline for reasoning about the dynamics of these hidden units. Figure 1 provides a schematic for our response characterization pipeline. From a trained LSTM network, comprising of M LSTM units, we isolate individual LSTM cells, and characterize their output responses based on a series of interpretable response metrics. We formalize the method as follows:

Definition 1 *Let G, be a trained LSTM network with M hidden LSTM units. Given the dynamics of the training dataset (number of input/output channels, the main frequency components, the amplitude range of the inputs), we design specific step and sinusoidal inputs to the network, and get the following insights about the dynamics of the network at multi-scale resolutions:*

- *the relative strength or contribution of components within the network;*
- *the reactiveness of components to sudden changes in input; and*
- *the phase alignment of the hidden outputs with respect to the input.*

Specifically, we analyze the responses of (1) the step input and (2) the sinusoidal input. We use the classic formulations for each of these input signals wherein (1) step: $x_t = \left[\left[t > \frac{T}{2}\right]\right]$; and (2) sinusoid: $x_t = \sin\left(2\pi f\, t\right)$; where $[[\cdot]]$ represents the mathematical indicator function.

Across a network of LSTM units we can approximate sub-regions of the dynamics of a single cell, u, by extracting the input and recurrent weights corresponding to that individual cell. We then define a sub-system consisting of just that single cell and subsequently feed one of our baseline input signals, $x_t\ \forall_{t \in \{1..T\}}$ to observe the corresponding output response, y_t. In the following, we define the interpretable response metrics for the given basis input used in this study:

Definition 2 *The **initial** and **final response** of the step response signal is the starting and steady state responses of the system respectively, while the **response output change** represents their relative difference.*

Response output change or the delta response for short determines the strength of the LSTM unit with a particular parameter setting, in terms of output amplitude. This metric can presumably detect significant contributor units to the network's decision.

Definition 3 *The **settling time** of the step response is elapsed time from the instantaneous input change to when the output lies within a 90% threshold window of its final response.*

Computing the *settling time* for individual LSTM units enables us to discover "fast units" and "slow units". This leads to the prediction of active cells when responding to a particular input profile.

Definition 4 *The **amplitude** and **frequency** of a cyclic response signal is the difference in output and rate at which the response output periodically cycles. The response frequency, $\hat{f}$, is computed by evaluating the highest energy component of the power spectral density: $\hat{f} = \arg\max S_{yy}(f)$.*

The *amplitude* metric enables us to rank LSTM cells in terms of significant contributions to the output. This criteria is specifically effective in case of trained RNNs on datasets with a cyclic nature. Given a sinusoidal input, phase-shifts and phase variations expressed at the unit's output, can be captured by evaluating the *frequency* attribute.

Definition 5 *The **correlation** of the output response with respect to the input signal is the dot product between the unbiased signals: $\sum_{t=1}^{T}(x_t - \mathbb{E}[x]) \cdot (y_t - \mathbb{E}[y])$*

The *correlation* metric correspondes to the phase-alignments between input and output of the LSTM unit.

Systematic computation of each of the above responses metrics for a given LSTM dynamics, enables reasoning on the internal kinetics of that system. Specifically, a given LSTM network can be decomposed into its individual cell components, thus creating many smaller dynamical systems, which can be analyzed according to their individual response characterization metrics. Repeating this process for each of the cells in the entire network creates two scales of dynamic interpretability. Firstly, on the individual cell level within the network to identify those which are inherently exhibiting *fast* vs *slow* responses to their input, quantify their relative contribution towards the system as a whole, and even interpret their underlying phase-shift and alignment properties. Secondly, in addition to characterizing responses on the cell level we also analyze the effect of network capacity on the dynamics of the network as a whole. Interpreting hidden model dynamics is not only interesting as a deployment tool but also as a debugging tool to pinpoint possible sources of undesired dynamics within the network.

While one can use these response characterization techniques to interpret individual cell dynamics, this analysis can also be done on the aggregate network scale. After computing our response metrics for all decoupled cells independently we then build full distributions over the set of all individual pieces of the network to gain understanding of the dynamics of the network as a whole. This study of the response metric distributions presents another rich representation for reasoning about the dynamics, no longer at a local cellular scale, but now, on the global network scale.

5 Experimental Results

In the following section, we provide concrete results of our system in practice to interpret the dynamics of trained LSTMs for various sequence modeling tasks. We present our computed metric response characteristics both on the decoupled cellular level as well as the network scale, and provide detailed and interpretable reasoning for these observed dynamics. We chose four benchmark sequential datasets and trained on various sized LSTM networks ranging from 32 to 320 LSTM cell networks. The results and analysis presented in this section demonstrate applicability of our algorithms to a wide range of temporal sequence problems and scalability towards deeper network structures.

We start by reasoning how our response characterization method can explain the hidden-state dynamics of learned LSTM networks for a sequential MNIST dataset and extend our findings to three additional datasets. We perform an ablation analysis and demonstrate how some of our metrics find cells with significant contributions to the network's decision, across all datasets.

5.1 Response characterization metrics predict insightful dynamics for individual cells

We start by training an LSTM network with 64 hidden LSTM cells to classify a sequential MNIST dataset. Inputs to the cells are sequences of length 784 generated by stacking the pixels of the 28×28

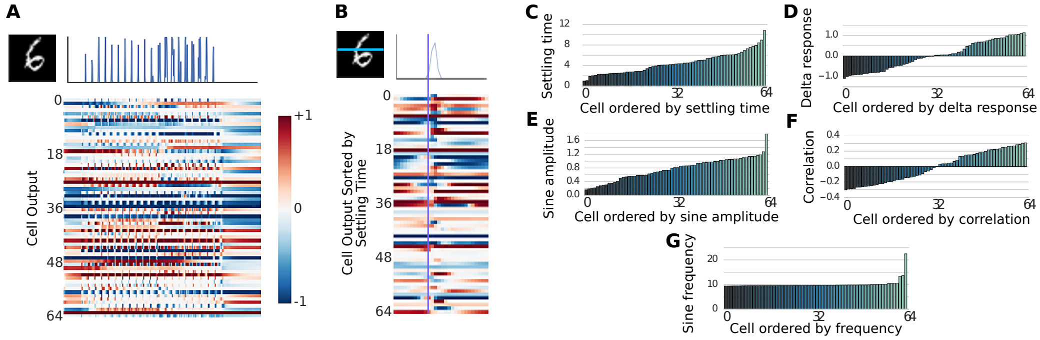

Figure 2: **Cell level interpretation of sequential MNIST.** A) An example sequence for digit 6 together with the network dynamics for a 64-neuron LSTM network. B) One slice sequence from digit 6 and its underlying network dynamics sorted for the settling time attribute. C) Settling time distribution. D) Delta response distribution. E) Sine-wave amplitude distribution. F) Correlation distribution G) Sine-frequency distribution.

Table 1: **Hidden dynamic distributions by dataset.** Systematic interpretation of internal dynamics distributions (mean and variance) of 128 cell LSTMs trained on various different benchmark datasets. The table shows the global speed and amplitude of the activity of network in terms of dynamical properties of the response characterization metrics.

	Step Response		Sinusoidal Response		
Dataset	**Settle Time**	**Output Change**	**Amplitude**	**Correlation**	**Frequency**
Sequential-MNIST [24]	6.96 ± 4.08	-0.04 ± 0.58	0.73 ± 0.30	0.17 ± 0.08	9.83 ± 0.46
S&P 500 Stock [2]	5.62 ± 1.73	0.02 ± 0.16	0.31 ± 0.05	0.03 ± 0.02	2.86 ± 2.19
CO_2 Concentrations [1]	5.65 ± 1.64	0.01 ± 0.12	0.27 ± 0.04	0.03 ± 0.01	9.83 ± 0.08
Protein Sequencing [31]	7.96 ± 6.65	0.08 ± 0.54	0.68 ± 0.22	2.07 ± 1.21	10.36 ± 1.65

hand-writing digits, row-wise (cf. Fig. 2A) and the output is the digit classification. Individual LSTM cells were then isolated and their step and sine-response were computed for the attributes defined formerly (cf. Fig. 4). Fig. 2C-G represent the distribution of cell activities, ranked by the specific metrics. The distribution of the settling time of the individual LSTM cells from a trained network, predicts low time-constant, (fast) cells (the first 20 neurons), and high-time constant (slow) cells (neurons 55-64) (Fig. 2C). This interpretation allows us to indicate fast-activated/deactivated neurons at fast and slow phases of a particular input sequence. This is validated in Fig. 2B, where the output state of individual LSTM cells are visually demonstrated when the network receives a sequence of the digit 6. The figure is sorted in respect to the predicted settling time distribution. We observe that *fast-cells* react to fast-input dynamics almost immediately while *slow-cells* act in a slightly later phase. This effect becomes clear as you move down the heatmap in Fig. 2B and observe the time difference from the original activation.

The distribution of the delta-response, indicates inhibitory and excitatory dynamics expressed by a 50% ratio (see Fig. 2D). This is confirmed by the input-output correlation criteria, where almost half of the neurons express antagonistic behavior to their respective sine-wave input (Fig. 2F). The sine-frequency distribution depicts that almost 90% of the LSTM cells kept the phase, nearly aligned to their respective sine-input, which indicates existence of a linear transformation. A few cells learned to establish a faster frequencies than their inputs, thereby realizing phase-shifting dynamics (Fig. 2G). The sine-amplitude distribution in Fig. 2E demonstrates that the learned LSTM cells realized various amplitudes that are almost linearly increasing. The ones with a high amplitude can be interpreted as those maximally contributing to the network's decision. In the following sections, we investigate the generalization of these effects to other datasets.

5.2 Generalization of response metrics are to other sequential datasets

We trained LSTM networks with 128 hidden cells, for four different temporal datasets: sequential MNIST [24], S&P 500 stock prices [2] and CO_2 concentration for the Mauna Laua volcano [1] forecasting, and classification of protein secondary structure [31]. Learned networks for each dataset are denoted seq-MNIST, Stock-Net, CO_2-Net and Protein-Net. Table 1 summarizes the statistics for

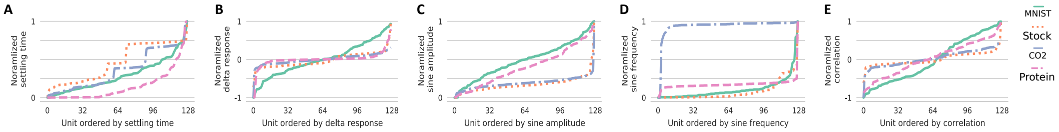

Figure 3: **Cell level response distributions.** (A-E) Response characterization metrics for networks with 128 individually ranked LSTM cells. The analyses predict A) cells with fast-dynamics and slow dynamic, (B and C) cells that are significantly contributing to the network decision, D) cells that realize phase shifting dynamics, and E) cells that are excitatory or inhibitory.

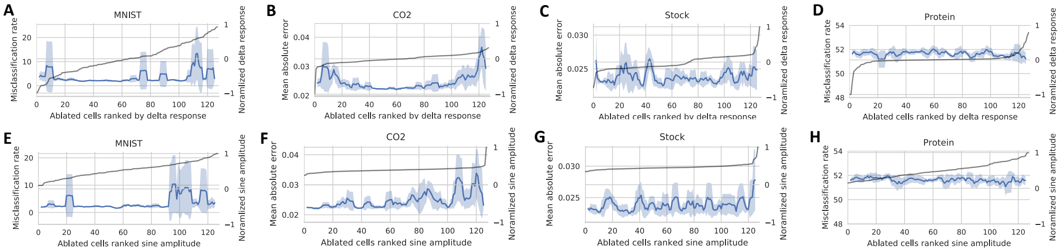

Figure 4: **Cell level ablation analysis**. Ablation of individual cells inside trained 128 cell LSTM networks across all four datasets (left to right). Changes in the predictive error are visualized against the ranked delta response (top) and sine amplitude (bottom) of the ablated cell. The Gray solid line represents the predictions of our method (right side vertical axis) as a function of the particular response metric. The blue solid line shows the mean and the shadows represents the standard deviation of a moving average filter on the 23 ablated impact of individual neurons. This is done to highlight the trend of the ablation impact with respect to the sorted particular metric.

all five metrics with the network size of 128. It represents the average cell response metric attributes for various datasets and demonstrates the global speed and amplitude of the activity of network in terms of dynamical properties of the response characterization metrics.

Fig 3A-E, represents the distributions for the metrics sorted by the value of their specific attribute across all datasets. Cells in Protein-Net realized the fastest dynamics (i.e. smallest settling time) compared to the other networks, while realizing a similar trend to the seq-MNIST (Fig. 3A). The settling time distribution for the LSTM units of CO_2 and Stock-Net depicts cell-groups with similar speed profiles. For instance neurons 52 to 70 in Stock-Net, share the same settling time (Fig. 3A). Sine frequency stays constant for all networks except from some outliers which tend to modify their input-frequency (Fig. 3D). The delta response and the correlation metrics (Fig. 3B and Fig. 3E) both indicate the distribution of the inhibitory and excitatory behavior of individual cells within the network. Except from the Seq-MNIST net, neurons in all networks approximately keep a rate of 44% excitatory and 56% inhibitory dynamics. The high absolute amplitude neurons (at the two tails of Fig. 3C), are foreseen as the significant contributors to the output's decision. We validate this with an ablation analysis subsequently. Moreover, most neurons realize a low absolute delta-response value, for all datasets except for MNIST (Fig. 3B). This is an indication for cells with an equivalent influence on the output accuracy. Sine-amplitude stays invariant for most neurons in Stock and CO_2-Nets (Fig. 3C). For the seq-MNIST net and Protein-net, this distribution has a gradually increasing trend with weak values. This predicts that individual cells are globally equivalently contributing to the output.

5.3 Response metrics predict significant contributing cells to the network's decision

To assess the quality of the predictions and interpretations of the provided response characterization metrics, we performed individual cell-ablation analysis on LSTM networks and evaluated the cell-impact on the output accuracy (misclassification rate), for the classification problems and on the output performance (mean absolute error), for the regression problems. We knocked out neurons from trained LSTM networks with 128 neurons. Fig. 4A-H illustrate the performance of the network for individual cell ablations for all four datasets. The gray solid line in each subplot, stands for the predictions of the response metrics. For CO_2-Net, this confirms that neurons with higher sine amplitude tend to disrupt the output more (Fig 4D). For the same network, the delta response predicted that neurons with high negative or positive value, are more significant in output's prediction. This is

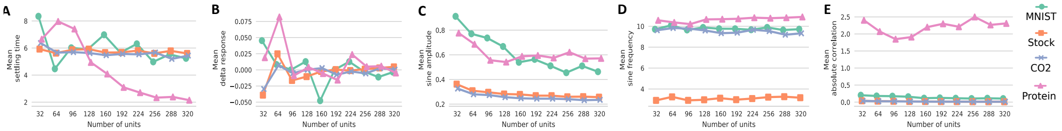

Figure 5: **Network capacity analysis.** (A-E) Response metrics as a function of the network's capacity. The analyses illustrate how response metrics provide insights on the global network scale.

clearly illustrated in Fig. 4C. For seq-MNIST-Net, the same conclusions held true; neurons with high absolute value of delta response or sine-amplitude reduce the accuracy at the output dramatically (Fig. 4A-B). By analyzing the sine-amplitude and delta-response of Protein-Net, we observe that neurons are equivalently valued and tend to contribute equivalently to the output accuracy. This is verified in the ablation analysis, shown in Fig. 4G and 4H, where the mean-misclassification error rate stays constant for all neural ablations. The absolute value for Stock-Net was also weak in terms of these two metrics, though there were some outliers at the tails of their distribution that predicted dominant neurons. This is clearly notable when comparing the neurons 120 to 128 of Fig. 4F to their prediction (gray line) where the amplitude of the response is maximal. In Fig. 4E ablation experiments for neurons 1 to 40 and 100 to 128 impose higher impact on the overall output. This was also observed in the delta response prediction shown in 4B, since neurons with stronger output response were present at the two tails of the distribution.

5.4 Network-level Interpretability for Trained LSTMs

While we analyzed the response characterization distributions on a cellular level above, in this subsection we focus on the effect of network capacity on observed hidden dynamics of the system on a global scale. Reasoning on this scale allows us to draw conclusions on how increasing the expressive capacity of LSTM networks trained on the same dataset can result in vastly different learned dynamics. We experimentally vary the capacity by simply adding hidden LSTM cells to our network and retraining on the respective dataset from scratch. The relationship between each response characteristic metric and the network capacity is visualized in Fig. 5A-E. The trends across datasets are visualized in a single subplot to compare respective trends. One especially interesting result of this analysis is the capacity relationship with response amplitude (cf. Fig. 5C). Here we can see that the amplitude response decays roughly proportionally to $\frac{1}{N}$, for all datasets, where N is the number of LSTM cells. In other words, we get the intuitive finding that as we increase the number of LSTM cells, the magnitude of each cell's relative contribution needed to make a prediction will subsequently decrease. Yet another key finding of this analysis is that the distribution of settling time is relatively constant across network capacity (cf. Fig. 5A). Intuitively, this means that the network is able to learn the underlying time delay constants represented in the dataset irrespective of the network capacity. One particularly interesting point comes for Protein-Net which exhibits vastly different behavior for both settling time (Fig. 5A) and correlation (Fig. 5E) than the remainder of the datasets. Upon closer inspection, we found that Protein-Net was heavily overfitting with increased capacity. This can be seen as an explanation for the rapid decay in its settling time as the addition of LSTM cells would increase specificity of particular cells and exhibit dynamical properties aligning with effectively memorizing pieces of the training set.

6 Conclusion

In this paper, we proposed a method for response characterization for LSTM networks to predict cell-contributions to the overall decision of a learned network on both the cell and network-level resolution. We further verified and validated our predictions by performing an ablation analysis to identify cell's which contribution heavily to the network's output decision with our simple response characterization method. The resulting method establishes a novel building block for interpreting LSTM networks. The LSTM network's dynamic-space is broad and cannot be fully captured by fundamental input sequences. However, our methodology demonstrates that practical sub-regions of dynamics are reachable by response metrics which we use to build a systematic testbench for LSTM interpretability. We have open-sourced our algorithm to encourage other researchers to further explore dynamics of LSTM cells and interpret the kinetics of their sequential models.In the future, we aim to extend our approach to even more data modalities and analyze the training phase of LSTMs to interpret the learning of the converged dynamics presented in this work.

7 Acknowledgment

We gratefully acknowledge the support of NVIDIA Corporation with the donation of GPUs used for this research. R.M.H., M.L. and R.G. are partially supported by Horizon-2020 ECSEL Project grant No. 783163 (iDev40), and the Austrian Research Promotion Agency (FFG), Project No. 860424. A.A. is supported by the National Science Foundation (NSF) Graduate Research Fellowship Program.

References

[1] R. F. Keeling, S. C. Piper, A. F. Bollenbacher, and S. J. Walker, Mauna Loa, Atmospheric Carbon Dioxide Record. `http://cdiac.ess-dive.lbl.gov/ftp/trends/co2/maunaloa.co2`. Accessed: 2018-03-17.

[2] Yahoo Finance Website, S&P 500 Stock. `https://finance.yahoo.com/quote/%5EGSPC/`. Accessed: 2018-04-13.

[3] Francesca Albertini and Paolo Dai Pra. Recurrent neural networks: Identification and other system theoretic properties. *Neural Network Systems Techniques and Applications*, 3:1–41, 1995.

[4] Francesca Albertini and Eduardo D Sontag. For neural networks, function determines form. *Neural Networks*, 6(7):975–990, 1993.

[5] Francesca Albertini and Eduardo D Sontag. Uniqueness of weights for recurrent nets. *MATHEMATICAL RESEARCH*, 79:599–599, 1994.

[6] Alexander Amini, Wilko Schwarting, Guy Rosman, Brandon Araki, Sertac Karaman, and Rus Daniela. Variational autoencoder for end-to-end control of autonomous driving with novelty detection and training de-biasing. 2018.

[7] Alsallakh Bilal, Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Do convolutional neural networks learn class hierarchy? *IEEE transactions on visualization and computer graphics*, 24(1):152–162, 2018.

[8] Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[9] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[10] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

[11] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *arXiv preprint arXiv:1704.03296*, 2017.

[12] Max Garzon and Fernanda Botelho. Observability of neural network behavior. In *Advances in Neural Information Processing Systems*, pages 455–462, 1994.

[13] Max Garzon and Fernanda Botelho. Dynamical approximation by recurrent neural networks. *Neurocomputing*, 29(1-3):25–46, 1999.

[14] Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2017.

[15] Ramin M Hasani, Mathias Lechner, Alexander Amini, Daniela Rus, and Radu Grosu. Re-purposing compact neuronal circuit policies to govern reinforcement learning tasks. *arXiv preprint arXiv:1809.04423*, 2018.

[16] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pages 1693–1701, 2015.

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

[18] Drew A Hudson and Christopher D Manning. Compositional attention networks for machine reasoning. *arXiv preprint arXiv:1803.03067*, 2018.

[19] Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. Lingusitic analysis of multi-modal recurrent neural networks. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 8–9, 2015.

[20] Akos Kádár, Grzegorz Chrupała, and Afra Alishahi. Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*, 43(4):761–780, 2017.

[21] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

[22] Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, and Sven Dähne. Patternnet and patternlrp–improving the interpretability of neural networks. *arXiv preprint arXiv:1705.05598*, 2017.

[23] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. *arXiv preprint arXiv:1703.04730*, 2017.

[24] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[25] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

[26] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[27] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[28] Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

[29] Willsky A. S. Oppenheim, A. V. and I. T. Young. *Signals and systems*. Prentice-Hall, 1983.

[30] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013.

[31] Ning Qian and Terrence J Sejnowski. Predicting the secondary structure of globular proteins using neural network models. *Journal of molecular biology*, 202(4):865–884, 1988.

[32] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[33] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[34] Eduardo D Sontag. *Mathematical control theory: deterministic finite dimensional systems*, volume 6. Springer Science & Business Media, 2013.

[35] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M Rush. Lstmvis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE transactions on visualization and computer graphics*, 24(1):667–676, 2018.

[36] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

[37] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[38] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.