Exploring the Long-Term Generalization of Counting Behavior in RNNs

Nadine El-Naggar **Pranava Madhyastha** **Tillman Weyde**
City, University of London
United Kingdom
{nadine.el-naggar,pranava.madhyastha,t.e.weyde}@city.ac.uk

Abstract

In this study, we investigate the generalization of LSTM, ReLU and GRU models on counting tasks over long sequences. Previous theoretical work has established that RNNs with ReLU activation and LSTMs have the capacity for counting with suitable configuration, while GRUs have limitations that prevent correct counting over longer sequences. Despite this and some positive empirical results for LSTMs on Dyck-1 languages, our experimental results show that LSTMs fail to learn correct counting behavior for sequences that are significantly longer than in the training data. ReLUs show much larger variance in behavior and in most cases worse generalization. The long sequence generalization is empirically related to validation loss, but reliable long sequence generalization seems not practically achievable through backpropagation with current techniques. We demonstrate different failure modes for LSTMs, GRUs and ReLUs. In particular, we observe that the saturation of activation functions in LSTMs and the correct weight setting for ReLUs to generalize counting behavior are not achieved in standard training regimens. In summary, learning generalizable counting behavior is still an open problem and we discuss potential approaches for further research.

1 Introduction

Recurrent Neural Networks (RNNs) have been the go-to neural model for sequential tasks since the 1980s and have been popular in recent years for a variety of applications, (see e.g., Karpathy, 2015), including Natural Language Processing (NLP). RNNs are highly expressive, as they are theoretically Turing complete (Siegelmann and Sontag, 1995) and inherit universal approximation capacity (Funahashi and Nakamura, 1992; Leshno et al., 1993) from multi-layer perceptrons. Although RNNs have largely been superseded by Transformers (Vaswani et al., 2017) in NLP, there is recently a renewed interest in their theoretical properties and specifically the counting behavior (Merrill et al., 2020; Srivastava et al., 2022).

Recent work by Suzgun et al. (2019) and Weiss et al. (2018) follows in the footsteps of Gers and Schmidhuber (2001) to evaluate the ability of RNN models to learn formal languages in simplified scenarios. Weiss et al. (2018) show that single-cell Long Short-Term Memory (LSTM) and Rectified Linear Unit (ReLU) networks have the capacity for unbounded counting and modeling a set of formal languages (using saturation of some activation functions in LSTMs), while Gated Recurrent Units (GRUs) and Elman RNNs do not. Weiss et al. (2018) also report that ReLUs are difficult to train, which is consistent with the findings by Hochreiter (1991) and Bengio et al. (1994) that RNNs suffer from vanishing and exploding gradients when trained on tasks with long term dependencies. LSTM, introduced by Hochreiter and Schmidhuber (1997), remedies some of these problems.

Suzgun et al. (2019) show empirically that single-cell LSTMs can learn counting behavior to recognize Dyck-1 languages with 100% accuracy in their experiments. However, several open questions remain,

I Can't Believe It's Not Better Workshop at NeurIPS 2022.

including whether the counting exhibited by LSTMs is sufficiently precise for long sequences and to what extent ReLUs and GRUs can learn counting on long sequences. In this work, we investigate the limits of trained ReLU, LSTM and GRU models in generalizing Dyck-1 recognition to sequences that are substantially longer than those in the training set and analyze when, how and why they fail. The specific contributions of this study are:

1. We observe that single-cell LSTM, ReLU and GRU models generally fail to accept Dyck-1 sequences of length 1000 when trained on lengths up to 50.
2. We further observe that validation loss is a good predictor for long sequence generalization, but reducing loss through longer training is not a practical method to achieve better generalization.
3. We illustrate and analyze the different failure modes of ReLU, LSTM, and GRU cells.

Our code is publicly available at `https://github.com/nadineelnaggar/Long_Term_Counting_RNNs`.

2 Modeling Dyck-1 Sequences

We evaluate the limit of three different single-cell RNN models trained on Dyck-1 sequences in generalizing to longer sequences. Dyck-1 sequences have a) one type of brackets, b) an equal number of opening and closing brackets, and c) no closing brackets before their corresponding opening brackets. We use the same task setup as Suzgun et al. (2019) to evaluate our models, which is to classify which tokens are valid to follow in a Dyck-1 sequence after each token. This is equivalent to classifying complete vs incomplete Dyck-1 sequences. This is limited as a Dyck-1 language acceptance task. While the sequence is classified at every time step, sequences that violate condition c) do not occur. Therefore, this setup does not allow for testing full language acceptance, as defined, e.g., in Merrill (2020).

For training and initial testing, we use the same types of datasets as reported in Suzgun et al. (2019). Our *Training Set* consists of 10000 sequences with lengths of 2 to 50 tokens. We use a *Validation Set* with 5000 sequences of length 2 to 50 tokens. We use a *Long Test Set* with 5000 sequences of lengths 52 to 100 tokens. The disjoint Training, Validation and Long Test Sets are generated in the same way as in Suzgun et al. (2019).

We use single-layer single-cell LSTMs, ReLUs and GRUs in our experiments. We include GRUs in these experiments, although they cannot perform unbounded counting, to compare how this limitation manifests itself empirically. We train online with the Adam optimizer (Kingma and Ba, 2014), and learning rates of 0.01 for the ReLU and LSTM models, and 0.001 for the GRU models. The training was run using PyTorch in a Linux environment.

We train 10 models of each type for 30 epochs and select the best model (out of the epochs) per run by validation loss. We observe convergence typically within 20 epochs of training, often less. We experimented with training for 100 epochs on a subset without observing any additional benefit. For ReLUs we train 30 models and select the 10 best runs, as many ReLUs do not learn to recognize Dyck-1 sequences at all.

3 Very Long Sequence Generalization

We present the prediction perfomance results of our experiments in Table 1. We observe that all three models perform well on both Training and Validation Sets, but there is more variation on the Long Test Set. We also observe in Table 1 that ReLUs tend to show very variable performance. LSTMs perform best on average, and GRUs are on average between LSTMs and ReLUs.

We evaluate the generalization limits of our models that train well on short sequences and generalize well to some longer sequences up to 100 tokens. We use a *Very Long Test Set* with 100 sequences of length 1000 tokens, which are generated using same methods as presented in Suzgun et al. (2019).

The results on the Very Long Test Set are presented in the last column of Table 1 and are shown as First Point of Failure because all models but one fail on these sequences. Only one ReLU model has an accuracy of 4.0% while all the other ReLUs, the LSTMs and the GRUs do not recognize a single

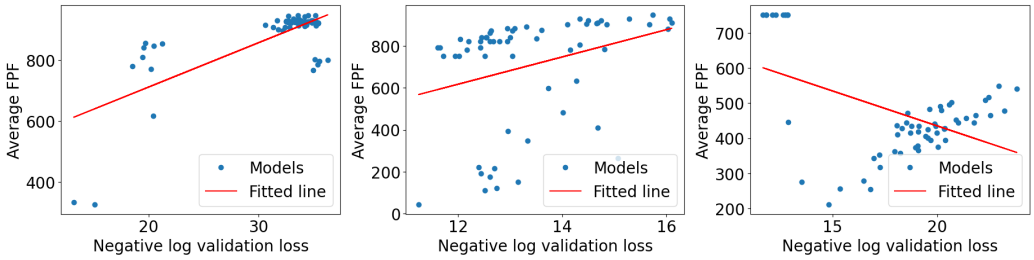
Table 1: For Training, Validation, and Long Test Sets we report the accuracy in percentage of sequences, where a sequence counts as correctly recognized when all tokens are correctly classified. The numbers in the table are the averages and (minimum / maximum) over 10 models. For the Very Long sequences, we report the First Point of Failure (FPF). A maximum FPF value *none* means that the best model(s) did not fail. The best results per column are highlighted in bold font.

Model	Training	Validation	Long	Very Long (FPF)
LSTM	**100** (100 / 100)	**100** (100 / 100)	**100** (99.8 / 100)	**916.8** (802.4 / 946.4)
ReLU	97.9 (90.9 / 100)	97.5 (89.3 / 100)	74.2 (33.5 / 100)	871.7 (543.1 / ***none***)
GRU	100 (100 / 100)	100 (100 / 100)	92.5 (90.72 / 95.6)	472.5 (434.7 / 545.2)

Very Long sequence correctly. Previous work from Weiss et al. (2018) shows that it is not possible for GRUs to perform unbounded counting, so it is not surprising that their performance deteriorates sharply for the longer sequences. ReLUs and LSTMs do have the capacity for unbounded counting, but we observe that training dynamics does not lead to models that count precisely enough in the long run. Interestingly, the only model that recognizes at least some of the Very Long sequences correctly is a ReLU model, despite their lower average performance. We will discuss the reasons for this behavior below.

This result does qualify the statement by Suzgun et al. (2019) that LSTMs can perform dynamic counting, insofar as the counting does not generalize indefinitely. This is actually not surprising given the results by Weiss et al. (2018), where LSTMs fail to generalize to very long sequences on a related task, which can be explained by counting not being precise enough, so that the errors accumulate.

4 Loss and Generalization



(a) LSTM (b) ReLU (c) GRU

Figure 1: Scatter plot with regression lines of negative log of the average validation loss versus the average FPF on the Very Long Test Set. We use a set of 60 models from different stages of the training process of for each of ReLU, LSTM and GRU (see text for details).

In practice, finite counting abilities may be sufficient, hence it would be useful if it is possible during training (through either the training or validation dynamics) to gather sufficient signals in order to understand how well a model will generalize. This is because repeatedly testing on very long sequences can be expensive. In this section, we study the utility of training dynamics towards answering the above question. We use the observation that lower loss values achieved by the models

Table 2: Linear regression between the loss values on the different datasets (Training, Validation and Long Test Sets) and the FPF on the Very Long Test Set with R^2 and p values.

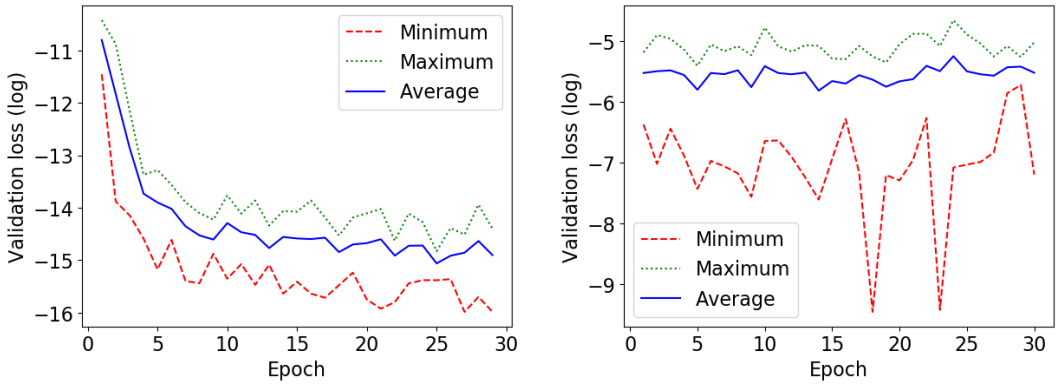
Model	Training R^2 / p	Validation R^2 / p	Long R^2 / p
LSTM	0.434 / 1.06×10^{-8}	0.511 / 1.412×10^{-10}	0.593 / 6.36×10^{-13}
ReLU	0.002 / 0.75	0.094 / 0.017	0.201 / 0.0003
GRU	0.363 / 3.55×10^{-7}	0.218 / 0.00017	0.075 / 0.035

seems to be related to higher FPF values. We quantify this in a linear regression between the negative log loss on Training, Validation and Long Test Sets and the FPF on the on the Very Long Test Set. We use here a different set of models, to include different stages of training with different levels of validation loss. We take from each of the 10 runs the models after epoch 1, 5, 10, 15, 20, and 25, so that we have 60 models for every type. This is done in order to have a range of models including some where the loss values are relatively high because the training was still at the beginning. The R^2 and p values are reported in Table 2.

We observe that the losses and FPF are well correlated for the LSTMs. For the ReLUs, we see increased correlation for the losses on datasets with long sequences, but overall the correlation is low. The GRUs show good correlation on the Training and Validation Sets, but lower correlation values for the Long Test Set, which is plausible given the lower performance on the Long Test Set.

Figure 1 shows scatter plots with the regression lines for the Validation Set loss. For the LSTMs we see that the lower loss (higher negative log loss) does lead to higher FPF, but there are still many values that are 20-25% lower than the length of the sequences. The ReLUs are much more variable, so that a low validation loss does not guarantee a high FPF. The GRUs, are interesting, where we observe that there are a few models with good FPF but relatively high loss. These models lead to a negative correlation overall and even for very low losses the FPF values are relatively low.

4.1 Effect of Training Duration on Loss



(a) LSTM (b) ReLU

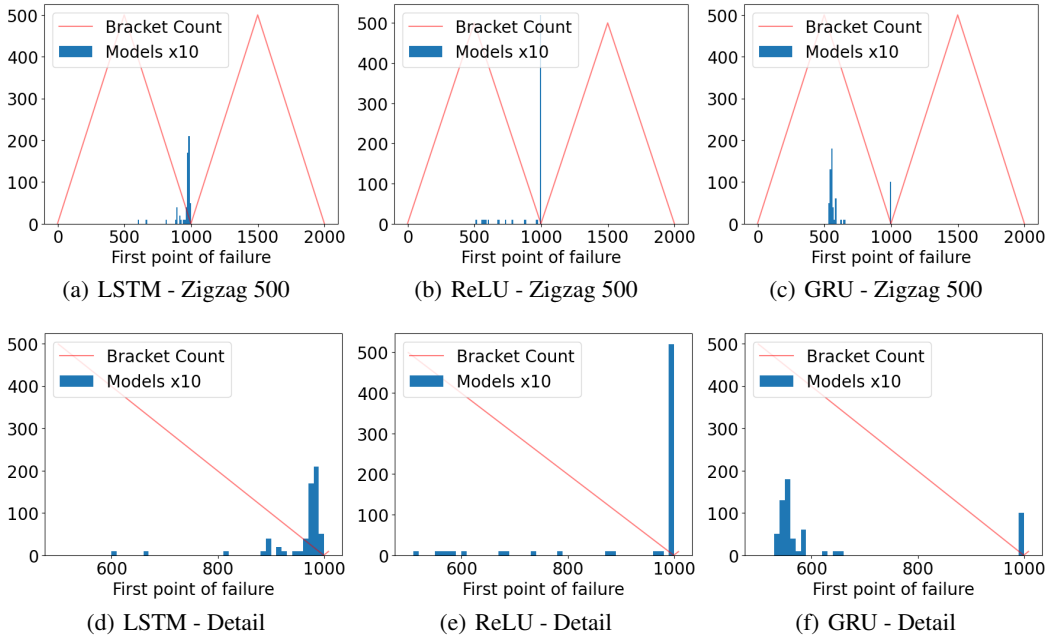
Figure 2: Progression of the validation loss during training for LSTM and ReLU models. We show average, minimum and maximum loss. The LSTMs show effective training while the ReLU results mostly stagnate.

Given that low validation losses correlate with high FPF, a possible strategy for improving long term counting behavior could be to train for more epochs to reduce the loss and thus extend the duration of correct counting. We show the progression of the validation loss during training for LSTM and ReLU models in Figure 2. We do not include GRUs as they do not have the capacity for precise counting, so that this strategy does not apply in any case.

The LSTM training behaves normally, i.e. the loss value tends to decrease throughout training for the LSTM, but the reduction becomes much slower as the training progresses. For ReLUs, there is no clear trend, but strong variations throughout.

In any straightforward solution of exact counting with LSTMs, like the one proposed by Weiss et al. (2018), it is necessary to saturate some of the activation functions. In a cursory inspection we find only a few networks where any saturation occurs. In all the experimentation we do in the context of this paper, we do not encounter a single LSTM model that actually recognizes any Very Long sequence, even with longer training. Thus it seems unlikely that saturation is achievable when training within common ranges. Overall, extensive training is not a practical approach to achieve long term counting behavior as it computationally expensive and not reliable.

5 Failure Modes



(a) LSTM - Zigzag 500 (b) ReLU - Zigzag 500 (c) GRU - Zigzag 500

(d) LSTM - Detail (e) ReLU - Detail (f) GRU - Detail

Figure 3: Histogram of the FPF distribution for the Zigzag sequence with $j = 500$. The red line represents the depth of the sequence (number of opening brackets without a closing bracket), and the histogram represents the number of models failing at different positions. The second row shows details of the distribution from position 500 to 1000, illustrating different failure modes of the different RNN types (see text for a detailed discussion).

When trying to understand why and how the RNN models fail to count correctly for long sequences, it is worth considering where the counting is represented within the RNN. In the single-cell ReLU network, the activation h_t of the ReLU is the only suitable variable. For the LSTM, the variable c_t (in the notation used by Weiss et al. (2018)) is the only one that is not the result of a squashing function, and it has empirically been shown to perform counting in their experiments.

The only actual test needed of the counter is the comparison to zero to test whether the sequence is a complete Dyck-1 sequence (or $a^n b^n$ in the experiments by Weiss et al. (2018)). In order to achieve zero activation at the appropriate points in the sequence, the increment when processing an opening bracket needs to be of the same absolute value but opposite sign as the decrement when processing a closing bracket. The datasets used in these experiments contain only valid Dyck-1 sequences, so that the counter values should not become negative therefore the ReLU should behave linearly. For a linear model, the order of the tokens plays no role. In the solution provided by Weiss et al. (2018) for the LSTM it is necessary that several variables saturate the activation functions to reach 1 or -1 for exact counting. In particular, if f_t (in their notation) does not saturate, the counter state will be reduced over time, independent of the opening or closing brackets. For GRUs, it is not possible to consistently have the same increment and decrement as discussed in Weiss et al. (2018).

We expect these differences in the operation to result in different behaviors that will mainly occur around the points where counter should be 0, because the Dyck-1 sequence is complete. For this purpose we create a dataset that consists of repetitions of j opening brackets followed by j closing brackets for $j = \{10, 20, 25, 50, 100, 125, 200, 250, 500, 1000\}$. All sequences in this dataset are 2000 tokens in length, and distinct. We refer to this dataset as the *Zigzag Set*. We test our 60 models for each RNN type at various stages of training, as in the previous section, on the Zigzag Set and record the FPF achieved for each model on each element of the dataset. The results on the Zigzag Set are that LSTM and GRU models perform better on the sequences with smaller j, showing evidence of a non-linear influence of the counter value (which is proportional to the depth, i.e. the number of open brackets without a corresponding closing bracket).

To illustrate the different behaviors, we plot in Figure 3 the distribution of the FPF values for the Zigzag sequence with $j = 500$ and show in detail graphs positions 500-1000. At position 1000 we have a complete Dyck-1 sequence, so that the counter should be 0 and the RNN should predict the corresponding class. The LSTM models fail mostly at a short distance before position 1000 (Figure 3 (a) and (d)), indicating that the decrements are consistently greater by absolute than the increments. ReLUs fail mostly at position 1000, indicating that the decrement is less by absolute than the increment, but they also fail before (Figure 3 (b) and (e)), indicating that the position of failures is not dependent on the sequence depth. The GRUs' increment and decrement values vary throughout the sequence with the decrements relatively large when the counter value is greater. This leads to a failure soon after closing bracket sequence starts (Figure 3 (c) and (e)).

The GRU behavior confirms the known limitations. The observations of the ReLUs confirm that they can have decrements greater or smaller than increments and independent of the sequence depth, but that finding correct weights by backpropagation with gradient descent is difficult. LSTMs confirm that exact counting does not occur in a normal training setup and saturation of the activation functions is rarely observed, preventing straightforward solutions.

This is a potential explanation for the fact that the best model in our experiments was a ReLU model. In the ReLU, saturation of an activation function is not possible and not needed, so that it is possible for the learning process to occasionally reach a good weight configuration even if the training process is not very effective on average.

6 Conclusions and Future Work

In this study, we investigate the generalization of RNNs trained to recognize Dyck-1 languages. The counting behavior of the trained models is effective on sequences up to double the length of the training sequences, but almost always fails on sequences of 20 times the length (1000 tokens in this case). While Training, Validation and Long Test Set losses are predictive for very long term generalization, the correlation is not perfect and training for longer to achieve better generalization is not a practical solution as long training is not efficient on LSTMs and not reliable on ReLUs. The different failure patterns of LSTMs, ReLUs and GRUs correspond to our hypotheses of their operation and shows that despite the theoretical capacity of LSTMs and ReLUs, the networks do not reach exact solutions even in the simple noiseless test cases we tried.

This study empirically extends and explains the theoretical findings by Weiss et al. (2018) and the experiments by Suzgun et al. (2019). In the future, there are many interesting research topics stemming form these results. Understanding the training dynamics of LSTMs and ReLUs better is important, as it will help determine how to initialize and train LSTMs to reach saturation by backpropagation and for ReLUs how to learn the correct weight configuration in a stable way. Other interesting aspects are the behavior of more commonly used architectures (multi-cell and multi-layer networks), batch training and regularization, and different sequence lengths for training.

Many approaches have been developed to improve the learning of RNNs in general, and they may be helpful in this context. E.g., for ReLUs, gradient clipping has been proposed by Pascanu et al. (2013) and widely applied. Specific weight initialization methods may help especially the ReLU networks to learn more effectively (Le et al., 2015). Targeted discretization of model weights or saturating the activation functions may be effective with LSTMs. Overall, the learning of discrete counting behavior for the long term with continuous neural networks is not yet solved and more work is needed to understand the theoretical foundations and design effective improvements.

References

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166.

Ken-Ichi Funahashi and Yuichi Nakamura. 1992. Neural networks, approximation theory, and dynamical systems. *Notes on the Institute of Mathematical Analysis*, 804:18–37.

Felix A. Gers and Jürgen Schmidhuber. 2001. LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Trans. Neural Networks*, 12(6):1333–1340.

Sepp Hochreiter. 1991. Untersuchungen zu dynamischen neuronalen netzen. *Diploma Thesis, Technische Universität München*.

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

Andrej Karpathy. 2015. The unreasonable effectiveness of recurrent neural networks. *Andrej Karpathy blog*, 21:23.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Quoc V. Le, Navdeep Jaitly, and Geoffrey E. Hinton. 2015. A simple way to initialize recurrent networks of rectified linear units. *arXiv preprint arXiv:1504.00941*.

Moshe Leshno, Vladimir Ya Lin, Allan Pinkus, and Shimon Schocken. 1993. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867.

William Merrill. 2020. On the linguistic capacity of real-time counter automata. *CoRR*, abs/2004.06866.

William Merrill, Gail Weiss, Yoav Goldberg, Roy Schwartz, Noah A. Smith, and Eran Yahav. 2020. A formal hierarchy of RNN architectures. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 443–459. Association for Computational Linguistics.

Razvan Pascanu, Tomás Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1310–1318. JMLR.org.

Hava T. Siegelmann and Eduardo D. Sontag. 1995. On the computational power of neural nets. *J. Comput. Syst. Sci.*, 50(1):132–150.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Mirac Suzgun, Sebastian Gehrmann, Yonatan Belinkov, and Stuart M Shieber. 2019. Lstm networks can perform dynamic counting. *ACL 2019*, page 44.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Gail Weiss, Yoav Goldberg, and Eran Yahav. 2018. On the practical computational power of finite precision rnns for language recognition. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pages 740–745. Association for Computational Linguistics.