Understanding Recurrent Neural Architectures by Analyzing and Synthesizing Long Distance Dependencies in Benchmark Sequential Datasets

Abhijit Mahalunkar
Technological University Dublin
Dublin, Ireland
abhijit.mahalunkar@mydit.ie

John D. Kelleher
Technological University Dublin
Dublin, Ireland
john.d.kelleher@dit.ie

ABSTRACT
In order to build efficient deep recurrent neural architectures, it is essential to analyze the complexity of long distance dependencies (LDDs) of the dataset being modeled. In this context, in this paper, we present detailed analysis of the complexity and the degree of LDDs (or LDD characteristics) exhibited by various sequential benchmark datasets. We observe that the datasets sampled from a similar process or task (e.g. natural language, or sequential MNIST, etc) display similar LDD characteristics. Upon analysing the LDD characteristics, we were able to analyze the factors influencing them; such as (i) number of unique symbols in a dataset, (ii) size of the dataset, (iii) number of interacting symbols within a given LDD, and (iv) the distance between the interacting symbols. We demonstrate that analysing LDD characteristics can inform the selection of optimal hyper-parameters for SOTA deep recurrent neural architectures. This analysis can directly contribute to the development of more accurate and efficient sequential models. We also introduce the use of Strictly $k$-Piecewise languages as a process to generate synthesized datasets for language modelling. The advantage of these synthesized datasets is that they enable targeted testing of deep recurrent neural architectures in terms of their ability to model LDDs with different characteristics. Moreover, using a variety of Strictly $k$-Piecewise languages we generate a number of new benchmarking datasets, and analyse the performance of a number of SOTA recurrent architectures on these new benchmarks.

CCS CONCEPTS
• Mathematics of computing → Information theory; • Computing methodologies → Information extraction; Neural networks; Supervised learning by classification; • Theory of computation → Regular languages.

KEYWORDS
Sequential Models, Recurrent Neural Networks

1 INTRODUCTION
Recurrent Neural Networks (RNN) laid the foundation of sequential data modeling [13]. However, recurrent neural architectures trained using backpropagation through time (BPTT) suffer from exploding or vanishing gradients [3, 21, 22]. This problem presents a specific challenge in modeling sequential datasets which exhibit long distance dependencies (LDDs). LDDs describe an interaction between two (or more) elements in a sequence that are separated by an arbitrary number of positions. LDDs are related to the rate of decay of statistical dependence of two points with increasing time interval or spatial distance between them. For example, in English there is a requirement for subjects and verbs to agree, compare: "The dog in that house is aggressive" with "The dogs in that house are aggressive". This dependence can be computed using information theoretic measure i.e. Mutual Information [4, 10, 28, 35].

One of the early attempts at addressing this issue was by El Hihi and Bengio [12] who proposed a Hierarchical Recurrent Neural Network which introduced several levels of state variables, working at different time scales. Various other architectures were developed based on these principles [5, 7]. LSTM introduced by Hochreiter and Schmidhuber [23] attempted to bridge minimal time lags in excess of 1000 discrete time steps by enforcing constant error flow through constant error carousels within special units. More recently, attention and memory augmented networks have delivered good performance in modeling LDDs [19, 33, 40]. The issue of vanishing gradients can also be alleviated by maintaining spectral norm of weight matrix close to unity, thus enforcing orthogonality [44].

A fundamental task for modelling sequential data is Language Modeling. A language model accepts a sequence of symbols and predicts the next symbol in the sequence. The accuracy of a language model is dependent on the capacity of the model to capture the LDDs in the data on which it is evaluated because an inability to model LDDs in the input sequence will result in erroneous predictions. In this paper, we will use the language modeling task, on a range of datasets, to evaluate the ability of RNNs to model LDDs. The standard evaluation metric for language models is perplexity. Perplexity is the measurement of how well a language model predicts the next symbol, and the lower the perplexity of a model the better the performance of the model.

There are a number of benchmark datasets used to train and evaluate language models: PennTree Banks (PTB) [30], WikiText 2 (Wiki-2), WikiText 103 (Wiki-103) [33] and Hutter-Text (Text8 and Enwik8). We reviewed the SOTA language models to check their performance on these datasets. Table 1 lists the perplexity scores for test and valid sets for PTB, Wiki-2 and Wiki-103. There is a general trend that model evaluations on Wiki-103 tend to result in lower perplexity scores followed by Wiki-2 and then PTB.

The similarity of scores for different models on these different benchmark datasets indicate that word-based dataset exhibit similar dependency structures; e.g., they exhibit similar LDD characteristics. Furthermore, our review of the language model SOTA revealed that most research on developing language models fails to explicitly analyze the characteristics of the LDDs within the datasets used to
First, we argue that a key step in modeling sequential data is to understand the characteristics of the LDDs within the data. Second, we present a method to compute and analyze the LDD characteristics on DilatedRNNs. We then argue that understanding LDD characteristics can inform the selection of appropriate hyperparameters for training and evaluating the models. Motivated by these two observations, this paper makes a number of research contributions.

**2 PRELIMINARIES**

### 2.1 LDD Characteristics

The experiments in Section 3.1 and 3.2 analyze the LDD characteristics of sequential datasets. This section describes the algorithm we have developed to calculate the LDD characteristic of a dataset.

Mutual information measures dependence between random variables $X$ and $Y$. These random variables have marginal distributions $p(x)$ and $p(y)$ and are jointly distributed as $p(x, y)$ [10]. Mutual information, $I(X; Y)$, is defined as:

$$I(X; Y) = \sum_{x, y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$  \hspace{1cm} (1)

If $X$ and $Y$ are not correlated, in other words if they are independent to each other, then $p(x)p(y) = p(x, y)$ and $I(X; Y) = 0$. However, if $X$ and $Y$ are fully dependent on each other, then $p(x) = p(y) = p(x, y)$ which results in the maximum value of $I(X; Y)$.

Mutual information can also be expressed using the entropy of $X$ and $Y$ i.e. $H(X)$, $H(Y)$ and their joint entropy, $H(X, Y)$ as given in the equations below:

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$  \hspace{1cm} (2)

$$H(X) = -\sum_x p(x) \log p(x)$$  \hspace{1cm} (3)

Shannon’s Entropy in Eq. 3 is known to be biased, generally underestimating true entropy from finite samples, thus, in this work, we choose the following equation to compensate for insufficient samplings [18]:

$$H(X) = \log N - 1/N \sum_{i=1}^K N_i \psi(N_i)$$  \hspace{1cm} (4)

where $N_i$ is the frequency of unique symbol $i$, $N = \sum N_i$, $K$ is the number of unique symbols, and $\psi(N_i)$ is the logarithmic derivative of the gamma function of $N_i$.

In order to measure dependence between any two symbols at a distance $D$ in a sequence, we design random variables $X$ and $Y$ so that $X$ holds the subsequence of the original sequence from index 0 till $|dataset| - 1 - D$, and $Y$ holds the subsequence from index $D$.
Computing LDD Characteristics

The experiments in Section 3.2 and 3.3 are based on synthetic datasets generated using SP languages. Here we introduce SP languages, following [1, 15, 39].

SP languages form a subclass of regular languages. Subregular languages can be identified by mechanisms much less complicated than Finite-State Automata. Many aspects of human language such as local and non local dependencies are similar to subregular languages [25]. More importantly, there are certain types of long distance (non local) dependencies in human language which allow finite-state characterization [20]. These type of LDDs can easily be characterizable by SPk languages and can be easily extended to other processes.

A language L, is described by a finite set of unique symbols Σ and Σ* (free monoid) is a set of finite sequences or strings of zero or more elements from Σ.

Example 2.1. Consider, Σ = {σ1, σ2, σ3, σ4} where σ1, σ2, σ3, σ4 are the unique symbols. A free monoid over Σ contains all concatenations of these unique symbols. Thus, Σ* = {λ, σ1, σ1σ2, σ1σ3, σ1σ4, σ3σ2, σ3σ3σ1σ3, σ2σ1σ3σ3, ...}.

Definition 2.1. Let, u denote a string, where u=σ1σ3. Length of a string u is denoted by |u| which 2. A string with length zero is denoted by λ.

Definition 2.2. A string v is a subsequence of string w, iff v = σ1σ2 ... σn and w ∈ Σσ1Σ2Σ3 ... ΣnΣn* , where σ ∈ Σ. A subsequence of length k is called a k-subsequence. Let subseqk(w) denote the set of subsequences of w up to length k.

Example 2.2. Consider, Σ = {a, b, c, d}, w = [acbd], u = [bd], v = [ac] and x = [db]. String x is a subsequence of length k = 2 or 2-subsequence of w. String v is a 3-subsequence of w. However, string x is not a subsequence of w as it does not contain [db] subsequence.

SPk languages are defined by grammar GSPk as a set of permissible k-subsequences. Here, k indicates the number of elements in a dependency. Datasets generated to simulate 2 elements in a dependency will be generated using SP2. This is the simplest dependency structure. There are more complex chained-dependency structures which require higher k grammars.

Example 2.3. Consider, L, where Σ = {a, b, c, d}. Let GSP2 be SPk grammar which is comprised of permissible 2-subsequences. Thus, GSP2 = {aa, ac, ad, ba, bb, bc, bd, ca, cb, cc, cd, da, db, dc, dd}. GSP2 grammar is employed to generate SP2 language.

Definition 2.3. Subsequences which are not in the grammar G are called forbidden strings1.

Example 2.4. Consider Example 2.3, although |ab| is a possible 2-subsequence, it is not part of the grammar GSP2. Hence, |ab| is a forbidden substring.

Example 2.5. Consider strings u, v, w u = [bbbcdd], v = [bbbbbcddada] and w = [bbabbbbcdd], where |u| = 6, |v| = 12 and |w| = 10. Strings u and v are valid SP2 strings because they are composed of subsequences that are in GSP2. However, w is invalid SP2 string because w contains |ab| a subsequence which is a forbidden string. These constraints apply for any string x where |x| ∈ Σ.

Example 2.6. Let GSP3 = {aaaa, aab, abb, ba, bab, bba, bb, bbb, ...} and forbidden string = {aba} be SP3 grammar which is comprised of

1Refer section 3.2. Finding the shortest forbidden subsequences in [15] for method to compute forbidden sequences for SPk language.
When applied to natural language datasets, language modelling we analysed the attributes of the standard datasets at both the

while simulating a real world dataset where the probability of oc-

natural language benchmark datasets can be used to train both

are characters). The perplexity scores reported in Section 1 were

are words) or character-based language modeling (unique symbols

WikiText 103 (Wiki-103)

[33] and Hutter-Text (Text8 and Enwik8).

In figure 2b, we fit broken power laws to word-based datasets to

character and word level. Table 2 lists aggregate statistics for these
datasets at both these levels. In order to examine the characteristics
of the LDDs within each of these datasets we plotted a curve of
the mutual information within the dataset at different distances. For
each dataset two of these curves were created, one at the word
level and one at the character level.

To plot a curve for a dataset we first applied the algorithm from
Section 2.1 in an iterative manner for different sizes of D (ranging
from 1 to the length of the dataset). Then these results were plotted
on a log-log axis, with the x-axis denoting the distance between
two symbols (either characters or words, and with a range from
1 to the length of the dataset) and the y-axis denotes the mutual
information (in nats) between these two random variables. We refer
to these plots as the LDD characteristics of a dataset.

LDD characteristics at a character level were computed for PTB,
Wiki-2, Wiki-103, Text8 and Enwik8 and are displayed in figure 3a.
Word-level LDD characteristics were computed for PTB, Wiki-2,
Wiki-103 and Text8 and are displayed in figure 2a.

LDD characteristics of character-based and word-based tasks
follow expected trends [11, 34]. It is seen that mutual information
decay follows a power law [28]. For character-based datasets, strong
dependence (higher power law decay) is observed between char-
acters at a distance up to 30; beyond which the curve exhibits a
long flat tail indicating lower dependence. This point of inflection
is of much interest. For word-based datasets, strong dependence
is observed between words at a distance up to 10 across various
datasets. This inflection point indicates the presence of a broken
power law. A Broken power law is a piecewise function, consisting
of two or more power laws, combined with a threshold (inflection
point) [26]. For e.g. with two power laws:

\[ f(t) = \begin{cases} 
  t^{-a_1} & t \leq t_{\text{thresh}} \\
  t^{-a_2} & t > t_{\text{thresh}} 
\end{cases} \]

(5)

In figure 2b, we fit broken power laws to word-based datasets to

study the features of the LDD characteristics. For a given dataset,
we observe that \( a_1 > a_2 \). The higher value of \( a_1 \) is due to a faster
rate of reduction in the frequency of contextually correlated words
in a sequence, as the spacing between them increases. This signifies
the presence of a strong grammar. Beyond the point of inflection, it
is understood that the pairs are not contextually correlated which
results in a flatter curve or lower value of \( a_2 \). This analysis enables
us to approximate the contextual boundary of the natural language
data. Also, the absolute of value of mutual information is an indi-
cator of the degree of the short and long distance dependencies
present in a dataset.

### Table 2: Features of natural language datasets

| Dataset   | Words Length | Characters Length |
|-----------|--------------|-------------------|
| Enwik8    | NA           | 6062              | 98620454         |
| Text8     | 253855       | 17005208          | 27               | 1000000000 |
| PTB       | 10000        | 1085779           | 48               | 5639751     |
| Wiki2     | 33278        | 2551843           | 282              | 12755448    |
| Wiki103   | 267735       | 103690236         | 1249             | 536016456   |

| Dataset   | Word-based | Character-based |
|-----------|------------|-----------------|
| Wiki103   | 267735     | 103690236       |
The fact that our above analysis of the English datasets found a very large value for $\alpha_1$ indicates that a dataset with good distribution of English text will exhibit a high value of mutual information at lower values of $D$ followed by a steep decay of mutual information. Recall from Section 1 that we noted a trend in the results reported across the standard benchmark datasets where Wiki-103 tended to deliver the best perplexity score followed by Wiki-2 and PTB. Our analysis of the LDD characteristics provides an explanation for this trend. Language models have very good performance on Wiki-2 due to the fact that they can take advantage of large $\alpha_1$ and very low mutual information in the flat region. Furthermore, language models marginally outperform on Wiki-2 as compared to the PTB due to higher mutual information at lower values of $D$.

The natural language datasets analysed above are not the only datasets used to evaluate language models. The sequential MNIST dataset is also widely used as a benchmarking dataset. The dataset contains 240,000 training images and 40,000 test images which are 28x28 pixel wide. In order to use them in a sequential task, the images are converted into a single vector of 784 pixels by concatenating all the rows of a single image. There are 256 unique values and total length of the data is 5488000. We generated an LDD characteristic plot for the entire dataset by concatenating all the sequential data of the images and then applying algorithm 1 on this data. We also designed permuted sequential MNIST datasets with various seeds and computed plotted the LDD characteristics of these datasets. The LDD characteristics are plotted in figure 3b. The MNIST LDD curve shows us that the unpermuted Sequential MNIST (blue line) exhibits a peculiar decay. High mutual information is observed between pixels spaced at small distances. We also observe mutual information peaks at multiples of 28, where 28 is the width of the MNIST images. This is in accordance with the properties of the image, i.e. nearby pixels (along the rows and columns) tend to be similar due to spatial dependencies. And when we convert the 2D image into a 1D sequential vector, the dependent pixels get spaced out by a factor of 28 (width of the image in this case), hence we observe peaks at multiples of 28. Also, when we analyze the decay trend, we observe exponential decay indicating lack of LDDs.
for $D > 300$ [28]. By introducing permutations in the sequential MNIST data, we loose the spatial dependencies within the image. Permutations result in chaotic data, thereby increasing the joint entropy of the two random variables, resulting in significantly lower mutual information. The LDD curve stays almost flat for $D < 300$. Beyond which the mutual information decay is exponential also indicating lack of LDDs for $D > 300$.

RNN models always perform better on unpermutated sequential MNIST compared with permuted sequential MNIST. Our analysis of the LDD characteristics of these datasets provides an explanation for why this is the case. Unpermuted sequential MNIST has LDDs of less than 300, due to the pixel dependencies and exponential decay. Datasets possessing such short-range dependency can be easily modeled using simple models e.g. Hidden Markov Models (HMMs) as they don’t require long memory. In the case of permuted sequential MNIST, we again observe LDDs of not more than 300 (due to exponential decay beyond that). However, the flat curve with very low mutual information is usually a result of noisy data. It is this noisy data (rather than complex LDDs) that is responsible for reducing the performance of the SOTA sequential models on permuted sequential MNIST. Overall, however we would argue that both of these datasets are inadequate at benchmarking sequential models due to their limitation in generating complex LDDs.

We also plotted the LDD characteristics of GPS trajectory dataset collected in Geolife project (Microsoft Research Asia) by 178 users in a period of over four years (from April 2007 to October 2011), see figure 3c. A GPS trajectory of this dataset is represented by a sequence of time-stamped points, each of which contains the information of latitude, longitude and altitude which was converted to a unique location number. These trajectories were recorded by different GPS loggers and GPS phone [47]. Upon analyzing the plot of the LDD characteristics in this data, it’s evident that human mobility also has power law decay suggesting the presence of LDDs.

### 3.2 LDD characteristics of SP$k$ datasets

Natural datasets present little to no control over the factors the affect LDDs. This, limits our ability to understand LDDs in more detail. SP$k$ languages exhibit some types of LDDs occurring in natural datasets. Moreover, by modifying the SP$k$ grammar we can control the LDD characteristics within a dataset generated by the grammar. To understand and validate the interaction between an SP$k$ grammar and the characteristics of the data it generates we used a number of SP$k$ grammars to generate datasets and analysed the properties of these datasets.

We used SP2, SP4 and SP16 grammars to generate these datasets. Using grammars with $k = \{2, 4, 16\}$, enabled us to generate datasets with different dependency structures (2, 4 and 16) and, hence, to analyze the impact of dependency structure on LDD characteristics.

In order to analyse the impact of vocabulary size on LDD characteristics, we generated SP2 grammars where $\Sigma_1 = \{a, b, c, d\}$ (size of vocabulary $= 4$) and $\Sigma_2 = \{a, b, c, d, \ldots, x, y, z\}$ (size of vocabulary $= 26$). We generated strings of maximum length of 20, 100, 200 and 500 using SP2 grammar. As explained in Example 2.6, by increasing the length of the generated strings, the distance between dependent elements is also increased, resulting in longer LDDs. We can then simulate LDD lengths as 20, 100, 200 and 500. We also choose two sets of forbidden strings for SP2 grammar, $\{ab, bc\}$ and $\{ab, bc, cd, dc\}$. We also generate two sizes of the same SP2 grammar to study the impact of the size of the data on the LDD characteristics, where one dataset was twice the size of the other. The datasets were generated using foma [24] and python [29]. Figure 4 shows plots of the LDD characteristics of these datasets.

Figure 4a plots LDD characteristics of SP2 languages with maximum string length of 20, 100, 200, 500. The point where mutual information decay is faster, the inflection point, lies around the same point on x-axis as the maximum length of the LDD. This confirms that SP$k$ can generate datasets with varying lengths of LDDs.

Figure 4b plots the LDD characteristics of SP2, SP4 and SP16 grammars. The strings in all the grammars are up to 100. Hence, we can observe the mutual information decay beyond $D > 100$. $k$ defines the number of correlated or dependent elements in a dependency rule. As $k$ increases the grammar becomes more complex and there is an overall reduction in frequency of the dependent elements in a given sequence (due to lower probability of these elements occurring in a given sequence). Hence, the mutual information is lower. This can be seen with dataset of SP16 vs SP2 and SP4. It is worth noting that datasets with lower mutual information curves tend to present more difficulty during modeling [29].

The impact of vocabulary size can be seen in figure 4c where the LDD characteristics of SP2 datasets with vocabulary size $V = 4$ and 26 are plotted. Both these datasets contain strings of maximum length 20. Hence the mutual information decays at 20. Both curves have identical decay indicating a similar grammar. However the overall mutual information of the dataset with $V = 26$ is much lower then the mutual information of the dataset with $V = 4$. This is because a smaller vocabulary results in an increase in the probability of the occurrence of each elements.

Figure 4d plots the LDD characteristics of SP2 grammar with two set of forbidden strings as $\{ab, bc\}$ and $\{ab, bc, cd, dc\}$. It is seen that the dataset with more forbidden strings exhibited less steep mutual information decay than the one with less number of forbidden strings. This can be attributed to the fact that datasets with more complex forbidden strings tend to exhibit more complex grammar as explained in section 2.2. By introducing more number of forbidden strings, it is possible to synthesize more complex LDDs as seen in the plot. In figure 4e we can observe the impact of the size of the dataset sampled from the same grammar. It can be seen that datasets sampled from the same grammar are less likely to be affected by the size of the dataset.

These grammars allow for the generation of rich datasets by setting the parameter $k$, the maximum length of the strings generated, size of vocabulary and by choosing appropriate forbidden substrings. This presents a compelling case to use these grammars to benchmark state-of-the-art sequential models.

### 3.3 Experiments with Dilated Recurrent Neural Networks

DilatedRNNs use multi-resolution dilated recurrent skip connections to extend the range of temporal dependencies in every layer and upon stacking multiple such layers are able to learn temporal dependencies at different scales [7]. This stacking of multi-resolution layers helps in passing contextual information over long distances.
which otherwise would have vanished via a single layer. Thus, the size of the dilations should, ideally, be tailored to match the LDD characteristics present in the dataset, and, in particular, the max dilation should match the max significant LDDs present in the dataset being modeled. The dilations per layer, and the number of layers, within a DilatedRNN are controlled by hyper-parameters [7].

In Chang et al. [7] it is reported that for PTB and sequential MNIST (unpermuted and permuted) the best performance is achieved with max dilations of 64 and 256 respectively. However, Chang et al. [7] provide no explanation for why these dilations are the optimal settings for these datasets. Given this context, it is interesting that our analysis of the LDD characteristics of PTB and sequential MNIST (unpermuted and permuted) in figures 3a (red line) and 3b respectively, found the mutual information inflection point for each dataset has similar value as the max dilations reported in Chang et al. [7]. For the PTB dataset, the inflection point is between 40 to 60 (x-axis) and for permuted sequential MNIST datasets, the inflection point is between 200-300 (x-axis). Based on this, we formulated the following hypothesis:

**Hypothesis 1.** For DilatedRNNs, the hyperparameter value of maximum dilation which lies in the same region as the inflection point of the LDD characteristics, delivers the best performance.

We trained DilatedRNNs on the permuted sequential MNIST dataset [7] for three sets of hyper-parameters to confirm that the max dilation and the inflection point have similar value. The settings are mentioned in the table 3. The first 3 settings were used for this experiment. To prevent any influence of other hyper-parameters on training, the original code was kept unchanged except for the setting of dilations. The testing accuracy for each task is plotted in figure 5 for all the 3 tasks. As expected, the set of hyper-parameters with max dilation of 256, delivered the best performance. After analyzing the LDD characteristics, the choice of dilations (powers of 2) make sense as they provide maximum LDD coverage with the least number of stacked layers and hyper-parameters.

Our analysis of the LDD characteristics of the sequential MNIST task, section 3.1, indicated that this dataset does not exactly follow power law decay. This presents an interesting opportunity to deviate the set of dilations from the powers of 2 in-order to deliver better performance. We trained a DilatedRNN network with max dilation of 280, as we observed 280 as the point of inflection for permuted sequential MNIST. The settings for this experiment is mentioned task no. 4 in table 3. The performance curve is plotted in figure 5 (purple line). This new set of hyper-parameters delivered better performance as compared to max dilation of 256. Both these experiments and results recorded in [7] confirmed our hypothesis 1.

Hence, by observing the LDD characteristics, we ensured faster and optimal training of DilatedRNNs by preventing the need for a grid search for this optimal hyper-parameter. We suspect max dilations less than the 280 (such as, up-to 128) inhibit the networks ability to fully capture all LDDs in the dataset, whereas dilations greater than the inflection point (such as, 512) results in the network learning contextually uncorrelated pairs from the data, which leads to degraded testing accuracy. After analyzing the performance of DilatedRNNs it is evident that exponential decay of mutual information in sequential MNIST achieved perplexity of 99.2 due to lack of LDDs.

Figure 4: LDD characteristics of datasets of (a) SP\textsuperscript{2} grammar exhibiting LDDs of length 20, 100, 200 and 500. (b) SP\textsuperscript{2}, SP\textsuperscript{4} and SP\textsuperscript{16} grammar exhibiting LDD of length 100. (c) SP\textsuperscript{2} grammar with vocabulary of 4 and 26. (d) SP\textsuperscript{2} grammar with varying forbidden strings (e) SP\textsuperscript{2} grammar with varying size of the datasets (f) All the SP\textsuperscript{k} grammars in a plot.
Table 3: Tasks for training DilatedRNNs

| Task No | Task Name                  | # of layers | Size of Dilations       |
|---------|----------------------------|-------------|-------------------------|
| 1       | Dilations upto 128        | 8           | 1, 2, 4, 8, 16, 32, 64, 128 |
| 2       | Dilations upto 256        | 9           | 1, 2, 4, 8, 16, 32, 64, 128, 256 |
| 3       | Dilations upto 512        | 10          | 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 |
| 4       | Custom Dilations upto 280 | 10          | 1, 2, 4, 8, 16, 32, 64, 128, 256, 512 |

Figure 5: Testing accuracy plot of DilatedRNNs with max dilations of 128, 256, 512 and 280.

4 DISCUSSION

The LDD characteristics of a dataset is indicative of the presence of a certain type of grammar in the dataset. For example, our analysis of word-based and character-based datasets in section 3.1, indicate that the word-based grammar is very different from character-based grammar. Understanding the properties of the underlying grammar that produces a language (data sequence) can aid in choosing optimal sequential model to learn on a given dataset of that language. For example, the maximum length of LDDs is much smaller in word-based datasets as compared to character-based datasets. But at the same time word-based LDD characteristics exhibit higher value of overall mutual information. This is why the sequential model that performs best on the word-based language modeling task will not necessarily be the best for the character based language modeling task.

It can also be noted that even though a specific grammar does induce similar LDD characteristics, there are subtle variations. These variations depend on a number of factors such as size of the vocabulary, size of the dataset, dependency structure (for e.g. "k" and "forbidden strings") and presence of any other noisy data (or presence of another grammar as seen with Enwik8 dataset). Thus, if a sequential model such as recurrent neural architecture intends to model a dataset, knowing these factors would greatly benefit in selecting the best hyper-parameters of the sequential model.

As seen in LDD characteristics of sequential MNIST, it is evident that the use of standard sequential MNIST in benchmarking tasks is out of place due to the absence of long-range correlations. This presents a compelling case to analyze LDD characteristics of benchmark datasets before they are selected for this job. Also, permutations lead to more complex dependency structure. Thus by altering existing datasets (with short-range dependency) in a way to introduce long-range correlations or LDDs and then analyzing the LDD characteristics presents a more systematic way of building more rich datasets. Even in SP languages, the choice of forbidden strings allows for the introduction of more complex dependency structure, hence introducing stronger long-range correlations in the generated dataset. This results in a systematic dependency of the design of benchmarking tasks. This can also be verified by computing LDD characteristics of the generated datasets.

One implication of these experiments is that having multiple benchmark datasets from a single domain does not necessarily improve the experimental testing of a models capacity to model LDDs: essentially, LDDs are fixed within a domain and sampling more datasets from that domain simply results in testing the model on LDDs with similar characteristics. Consequently, the relatively limited set of domains and tasks covered by benchmark datasets indicates that current benchmarks do not provide enough LDD variety to extensively test the capacity of state-of-the-art architectures to model LDDs.

5 RELATED WORK

5.1 Mutual Information and LDDs

Mutual information has previously been used to compute LDDs. In [11], two literary texts, Moby Dick by H. Melville and Grimm’s tales were used to analyze maximum length of LDDs present in English text. Correlations were found between few hundred letters. More specifically, strong dependence was observed (large α1) upto 30 characters indicating strong grammar, beyond which point the curve exhibited a long tail indicating weak dependence.

[28] analyzed LDD characteristics of enwik8. It was observed that LDDs with power-law correlations tend to be more difficult to model. They argued that LSTMs are capable of modeling sequential datasets exhibiting LDDs with power law correlations such as
natural languages far more effectively than Markov models; due to power-law decay of hidden state of the LSTM network controlled by the forget gate.

In another experiment, it was observed that DNA nucleotides exhibited long-range power law correlations [31, 36].

5.2 Neural Networks and Artificial Grammars

Formal Language Theory, primarily developed to study the computational basis of human language is now being used extensively to analyze any rule-governed system [8, 9, 14]. Formal languages have previously been used to train RNNs and investigate their inner workings. The Reber grammar [38] was used to train various 1st order RNNs [6, 41]. The Reber grammar was also used as a benchmarking dataset for LSTM models [23]. Regular languages, studied by Tomita [43], were used to train 2nd order RNNs to learn grammatical structures of the strings [16, 45].

Regular languages are the simplest grammars (type-3 grammars) within the Chomsky hierarchy which are driven by regular expressions. Strictly k-Piecewise languages are natural and can express some of the kinds of LDDs found in natural languages [20, 25]. This presents an opportunity of using SPk grammar to generate benchmarking datasets [1, 29]. In Avcuv et al. [1], LSTM networks were trained to recognize valid strings generated using SP2, SP4, SP8 grammar. LSTM could recognize valid strings generated using SP2 and SP4 grammar but struggled to recognize strings generated using SP8 grammar, exposing the performance bottleneck of LSTM networks. It was also observed that by increasing the maximum length of the generated strings of SP2 language thereby increasing the length of LDDs, the performance of LSTM degraded [29].

6 CONCLUSION

The foundational contribution of this paper represent a synthesis of distinct types of research on LDDs from multiple fields, including information theory, artificial neural networks for sequential data modeling, and formal language theory. The potential impact of this synthesis for neural networks research include: an appreciation of the multifaceted nature of LDDs; a procedure for measuring LDD characteristics within a dataset; an evaluation and critique of current benchmark datasets and tasks for LDDs; an analysis of how LDD characteristics within a dataset; an evaluation and critique of the use of these standard benchmarks and tasks can be misleading in terms of evaluating the capacity of a neural architectures to generalize to datasets with different forms of LDDs; and, a deeper understanding of the relationship between hyper-parameters and LDDs within language model architectures which can directly contribute to the development of more accurate and efficient sequential models.

REFERENCES

[1] Enes Avcuv, Chihiro Shibata, and Jeffrey Heinz. 2017. Subregular Complexity and Deep Learning. In Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML).
[2] Alexei Barvski and Michael Auli. 2019. Adaptive Input Representations for Neural Language Modeling. In International Conference on Learning Representations. https://openreview.net/forum?id=ByzZX20gFQ
[3] Y. Bengio, P. Simard, and P. Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks 5, 2 (March 1994), 157–166. https://doi.org/10.1109/72.279181
[4] Gerlof Bouma. 2009. Normalized (Pointwise) Mutual Information in Collocation Extraction. (01 2009).
[5] Victor Campos, Brendon Jou, Xavier Giró-i Nieto, Jordi Torres, and Shih-Fu Chang. 2018. Skip RNN: Learning to Skip State Updates in Recurrent Neural Networks. In International Conference on Learning Representations.
[6] M. Casey. 1996. The Dynamics of Discrete-Time Computation, with Application to Recurrent Neural Networks and Finite State Machine Extraction. Neural Computation 8, 6 (Aug 1996), 1135–1178. https://doi.org/10.1162/neco.1996.8.6.1135.
[7] Shiyu Chang, Yang Zhang, Wei Han, Mo Yu, Xiaoxiao Guo, Wei Tan, Xiaodong Cui, Michael Witbrock, Mark Hasegawa-Johnson, and Thomas Huang. 2017. Dilated Recurrent Neural Networks. arXiv:1710.02224 (2017).
[8] N. Chomsky. 1956. Three models for the description of language. IBE Transactions on Information Theory 2, 3 (September 1956), 113–124. https://doi.org/10.1109/TTT.1956.1056813
[9] Noam Chomsky. 1959. On certain formal properties of grammars. Information and Control 2, 2 (1959), 157–167. https://doi.org/10.1016/S0019-9958(59)90362-6
[10] Thomas M. Cover and Joy A. Thomas. 1991. Elements of Information Theory. Wiley-Interscience, New York, NY, USA.
[11] Werner Ebeling and Thorsten Poescheil. 2002. Entropy and Long range correlations in literary English. 26. 2 (2002), 241–246. https://doi.org/10.1209/0295-5075/26/4/001.arXiv:cond-mat/0204108
[12] Salah El Hiti and Yoshua Bengio. 1995. Hierarchical Recurrent Neural Networks for Long-term Dependencies. In Proceedings of the 8th International Conference on Neural Information Processing Systems (NIPS ’95). MIT Press, Cambridge, MA, USA, 493–499. http://dl.acm.org/citation.cfm?id=2998828.2998898
[13] Jeffrey L. Elman. 1990. Finding structure in time. COGNITIVE SCIENCE 14, 2 (1990), 179–211.
[14] W. Tecumseh Fitch and Angelina D Fiedereici. 2012. Artificial grammar learning meets formal language theory: an overview. Philosophical Transactions of the Royal Society B: Biological Sciences 367, 1598 (jul 2012), 1933–1955. https://doi.org/10.1098/rstb.2012.0108
[15] Jie Fu, Jeffrey Heinz, and Herbert G. Tanner. 2011. An algebraic characterization of strictly piecewise languages. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 6648 LNCS (2011), 252–263. https://doi.org/10.1007/978-3-642-20875-7_26
[16] C. L. Giles, C. B. Miller, D. Chen, H. H. Chen, G. Z. Sun, and Y. C. Lee. 1992. Learning and Extracting Finite State Automata with Second-Order Recurrent Neural Networks. Neural Computation 4, 3 (May 1992), 393–405. https://doi.org/10.1162/neco.1992.4.3.393
[17] ChengXue Cong, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2018. FRAGE: Frequency-Agnostic Word Representation. CoRR abs/1809.06658 (2018).
[18] P. Grassberger. 2003. Entropy Estimates from Insufficient Samplings. ArXiv e-prints (July 2003). arXiv:physics/0307138
[19] A. Graves, G. Wayne, and I. Danihelka. 2014. Neural Turing Machines. ArXiv e-prints (Oct. 2014). arXiv:1410.5401
[20] Jeffrey Heinz and James Rogers. 2010. Estimating Strictly Piecewise Distributions. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL ’10). Association for Computational Linguistics, Stroudsburg, PA, USA, 886–896. http://aclweb.org/anthology/P10-1086
[21] Sepp Hochreiter. 1991. Untersuchungen zu dynamischen neuronalen Netzen. Master’s thesis. TU Munich.
[22] Sepp Hochreiter, Yoshua Bengio, and Paolo Frasconi. 2001. Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies. In Field Guide to Dynamical Recurrent Networks. J. Kolen and S. Kremer (Eds.). IEEE Press.
[23] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural Computation 9, 8 (1997), 1735–1780.
[24] Mans Hulden. 2009. Foma: a finite-state compiler and library. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 29–32.
[25] G. Jager and J. Rogers. 2012. Formal language theory: refining the Chomsky hierarchy. Philosophical Transactions of the Royal Society B: Biological Sciences 367, 1598 (2012), 1956–1970. https://doi.org/10.1098/rstb.2012.0077
[26] Guðlaugur Jóhannesson, Guðlaugur Björnsson, and Einar H. Gudmundsson. 2006. Afterglow Light Curves and Broken Power Laws: A Statistical Study. The Astrophysical Journal Letters 640, 1 (2006), L5. http://stacks.iop.org/1538-4357/640/i=1/a=L5
[27] Ben Krause, Emmanuel Kemehmher, Iain Murray, and Steve Renals. 2018. Evaluation of Neural Sequence Models. In Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research), Jennifer Dy and Andreas Krause (Eds.). Vol. 80. PMLR, Stockholm, Sweden, 2766–2775. http://proceedings.mlr.press/v80/krause18a.html
[28] Henry W. Lin and Max Tegmark. 2017. Critical Behavior in Physics and Probabilistic Formal Languages. Entropy 19, 7 (2017). https://doi.org/10.3390/e19070299
[29] Abhijit Mahalanobis and John D. Kelleher. 2018. Using Regular Languages to Explore the Representational Capacity of Recurrent Neural Architectures. In Artificial Neural Networks and Machine Learning – ICANN 2018, Věra Kůrková, Yannis Manolopoulos, Barbara Hammer, Lazaros Iliadis, and Ilias Maglogiannis (Eds.). Springer International Publishing, Cham, 189–198.
[30] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The Penn Treebank: Annotating Predicate Argument Structure. In Proceedings of the Workshop on Human Language Technology (HLT ’94). Association for Computational Linguistics, Stroudsburg, PA, USA, 114–119.

[31] S. S. Melnik and O. V. Usatenko. 2014. Entropy and long-range correlations in DNA sequences. Computational biology and chemistry 53 Pt. A (2014), 26–31.

[32] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2017. Regularizing and Optimizing LSTM Language Models. CoRR abs/1708.02182 (2017).

[33] Stephen Merity. Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer Sentinel Mixture Models. CoRR abs/1609.07843 (2016). arXiv:1609.07843 http://arxiv.org/abs/1609.07843

[34] Marcelo A. Montemurro and Pedro A. Pury. 2002. Long-Range Fractal Correlations in Literary Corpora. Fractals 10, 04 (2002), 451–461. https://doi.org/10.1142/S0218348X02001257

[35] Liam Paninski. 2003. Estimation of Entropy and Mutual Information. Neural Computation 15, 6 (2003), 1191–1253. https://doi.org/10.1162/089976603321780272 arXiv:https://doi.org/10.1162/089976603321780272

[36] C. K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley. 1992. Long-range correlations in nucleotide sequences. Nature 356 (12 Mar 1992), 168. http://dx.doi.org/10.1038/356168a0

[37] Jack W. Rae, Chris Dyer, Peter Dayan, and Timothy P. Lillicrap. 2018. Fast Parametric Learning with Activation Memorization. CoRR abs/1803.10049 (2018). arXiv:1803.10049 http://arxiv.org/abs/1803.10049

[38] Arthur S. Reber. 1967. Implicit learning of artificial grammars. Journal of Verbal Learning and Verbal Behavior 6, 6 (1967), 855–863. https://doi.org/10.1016/S0022-5371(67)80149-X

[39] James Rogers, Jeffrey Heinz, Gil Bailey, Matt Edlefsen, Molly Visscher, David Wellcome, and Sean Wibel. 2010. On Languages Piecewise Testable in the Strict Sense. In The Mathematics of Language, Christian Ebert, Gerhard Jager, and Jens Michaelis (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 255–265.

[40] Giancarlo D. Salton, Robert J. Ross, and John D Kelleher. 2017. Attentive Language Models. In IJCNLP.

[41] A. W. Smith and D. Zipser. 1989. Encoding sequential structure: experience with the real-time recurrent learning algorithm. In International 1989 Joint Conference on Neural Networks. 645–648 vol.1. https://doi.org/10.1109/JCNN.1989.118646

[42] Sho Takase, Jun Suzuki, and Masaaki Nagata. 2018. Direct Output Connection for a High-Rank Language Model. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 4599–4609. http://aclweb.org/anthology/D18-1489

[43] Masaru Tomita. 1982. Learning of construction of finite automata from examples using hill-climbing: RR : Regular set Recognizer. Proceedings of Fourth International Cognitive Science Conference (1982), 105–108.

[44] Eugene Vorontsov, Chiheb Trabelsi, Samuel Kadoury, and Christopher Pal. 2017. On orthogonality and learning RNN with long term dependencies. In Proceedings of the 34th International Conference on Machine Learning (ICML’17). https://arxiv.org/abs/1702.00071 arxiv:1702.00071.

[45] Raymond L. Watrous and Gary M. Kuhn. 1991. Induction of Finite-State Automata Using Second-Order Recurrent Networks. In NIPS.

[46] Zhulin Yang, Zizhang Dai, Ruslan Salahutdinov, and William W. Cohen. 2017. Breaking the Softmax Bottleneck: A High-Rank RNN Language Model. CoRR abs/1711.03953 (2017).

[47] Yu Zheng, Hao Fu, Xing Xie, Wei-Ying Ma, and Quanann Li. 2011. Geolife GPS trajectory dataset - User Guide. https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/