DL2023 Major Project:SparseRNN- Temporal skip connections in sequential models

Ishaan Shrivastava, Vikash Yadav, Jahnab Dutta

Contributing authors: B20AI013 (shrivastava.9@iitj.ac.in); B20AI061 (yadav.41@iitj.ac.in); B20CS091 (dutta.4@iitj.ac.in);

1 Abstract

The study of useful inductive biases in sequence to sequence modelling has taken a backseat since the advent of transformers. The relaxation of inductive biases puts no additional constraints on a model, as a result they can find a better optimum if more data is given. However, this is not always possible in data and resource-constrained settings. Inspired by skip connections in large vision models, we propose a low-cost alternative to gating mechanisms in Recurrent Neural Networks with exponentially dilated recurrent connections called SparseRNN. We show that our approach generalises better than LSTMs, and is competitive with other recent models, and set the grounds for further investigations.

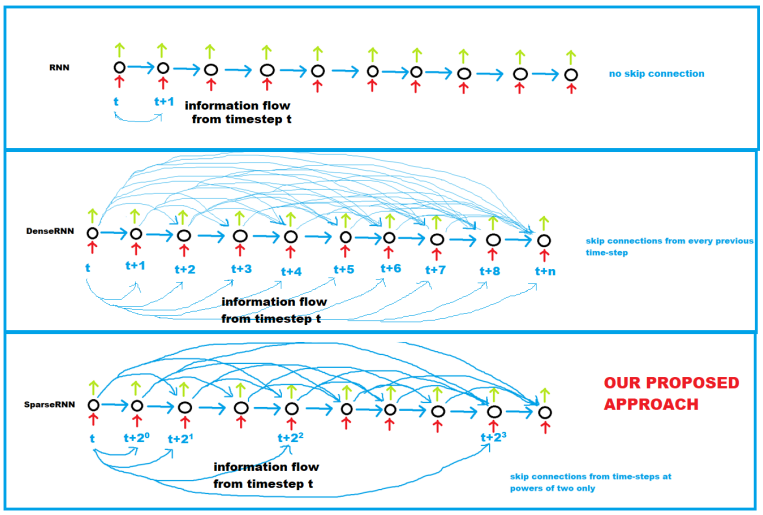
2 Motivatioin

In deep convolutional networks, skip connections have been known to mitigate vanishing gradients and hence improve performance metrics. These also allow the network to take a more varied input at each node and get knowledge from the previous part of the network. However the concept of these skip connections seem to unexplored when it comes to RNNs. Techniques like LSTMs and GRUs seem to dominate the world of seq2seq models. We attempt to explore the possiblity of usage of skip connections in RNNs. We believe these skip connections through time dimension will firstly solve the vanishing gradient problem prevelant in RNNs, secondly we think this will make the model better where use case involves long range temporal dependencies, as temportal skip connections through time give a better input of the previous nodes.

3 Methodology

3.1 Architecture

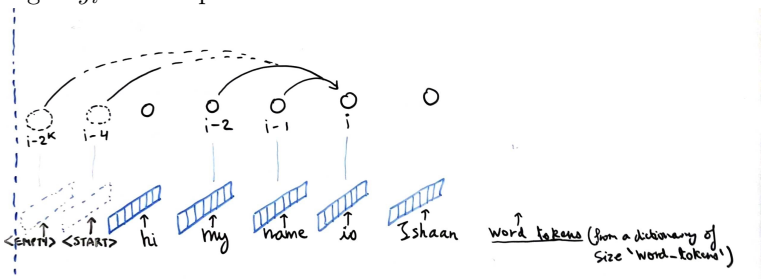
We introduce additional skip connections (also referred to as dilated recurrent skip connections in this survey) according to the diagram given below this section. It can be advantageous to utilize dilated connections between every timestep, however this suffers from an $O(n^2)$ complexity, thus we propose to introduce skip connections between timesteps characterized by gaps of powers of two, which pushed the time complexity of inference down to the subquadratic domain, i.e., $O(nlogn)$ while retaining much of the advantages a dense skip-connection scheme would provide.



3.2 Implementation

We have implemented our model to take in batched sequences of tokens of size (N, K) where N is the batch size and K is the fixed input sequence length, typically set as $2 + \underset{x \in \mathcal{D}}{max}|x|$, that is, the maximum sequence length over the dataset, plus two more spaces for the start and stop tokens. The model has a stacked RNN structure where the hidden layer output h_i^l at a given stack level l and position i are computed by concatenating the hidden layer outputs $\forall h_{i-2^j}^l$ and the hidden layer outputs from the lower level h_i^{l-1}. j is indexed $\in 1, ...K$, even for $i < 2^j$ in which case a zero vector of the appropriate shape is passed.

We use a stacked RNN structure for the SparseRNN with hidden layers of sizes [512, 128] and a fully connected linear layer for mapping the highest level hidden states h_i^L to the logits $\hat{y}_i$ at that position.



4 Experiments

4.1 Tasks & Datasets

We tested our architecture on Sentiment Analysis and Part of Speech Tagging. Our intuition is that our architecture would prove to give good results where there are long-range dependencies among the inputs, so tried to incorporate such tasks in our experiments.

For the **Sentiment analysis** task, we use the IMDB, a dataset for binary sentiment classification which contains input sentences upto 4000 words. The dataset contains 50K movie reviews.

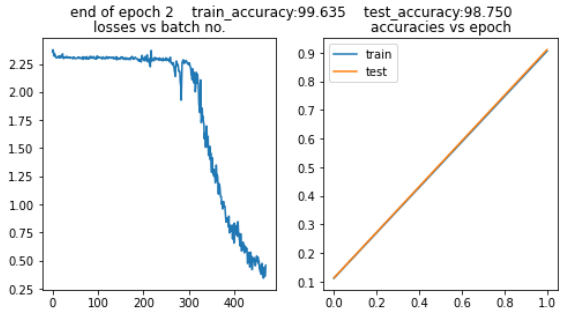
For **Part of Speech tagging**, we have used 3 datasets. First is the Penn Treebank dataset which contains articles of Wall Street Journal (WSJ), and is one of the most known and used corpus for the evaluation of models for sequence labeling. Second, we used Brown Corpus which contains one million words of American English texts printed in 1961. And third one is the Conll-2000 Dataset which is a dataset for dividing text into syntactically related non-overlapping groups of words, so-called text chunking.

4.2 Comparison

We compare our architecture with the Vanilla LSTM model and a simple RNN model on the 2 above tasks. The metric for evaluation is just accuracy. We have used WandB tool we learnt about in our labs to monitor and track our experiments and to plot our results for comparison. Here's the link to our WandB Project Page page containing all results.

4.3 Results & Findings

Preliminary results on MNIST dataset classification using our SparseRNN architecture. We flatten MNIST images to form the Sequential MNIST dataset (the height, width and input channels are flattened to a single dimension) and achieved a 98.75% accuracy on the test set without any significant tuning of the hyper-parameters. These results on Sequential MNIST are competitive to other previous baselines set on the Sequential MNIST task. **In particular, we beat LSTMs and Transformers on this dataset** which hints at the ability of our approach to excel in settings where correlations between different positions in the sequence follow different patterns from text-based tasks. The results of other approaches are tabulated against ours in figure **Fig.1**.
Further, the accuracy of the the three trained models for part of speech tagging over 20 epochs is displayed in **Fig.2** while the same for sentiment classification is in **Fig.3**. Finally, the tabulated form of the results are displayed in **Fig.4**.



Approach	SequentialMNIST accuracy (%)
IRNN	95.0
uRNN	95.1
RNN-path	96.9
LSTM	98.2
Transformer	98.2
R-Transformer	98.75
SparseRNN	**98.75**
dense-IndRNN	99.48

Fig. 1 Results of SparseRNN on SequentialMNIST & comparison with previous literature

5 Conclusion

In this study, we propose to use exponentially dilated recurrent skip connections. The proposed SparseRNN takes a subset of previous time-steps as input for a current timestep while unrolling the recurrent network, to incorporate the flow of information between far-apart time-steps. Compared with LSTMs, our model is slower to converge but has a distinctly better generalisation capability on the test set. The rationale for

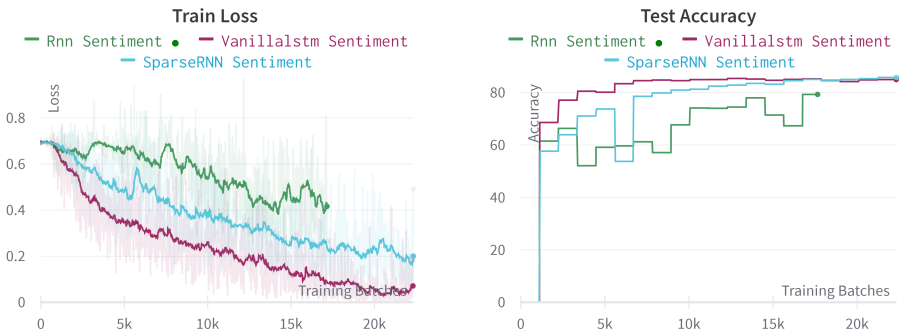


Fig. 2 Sentiment-Classification

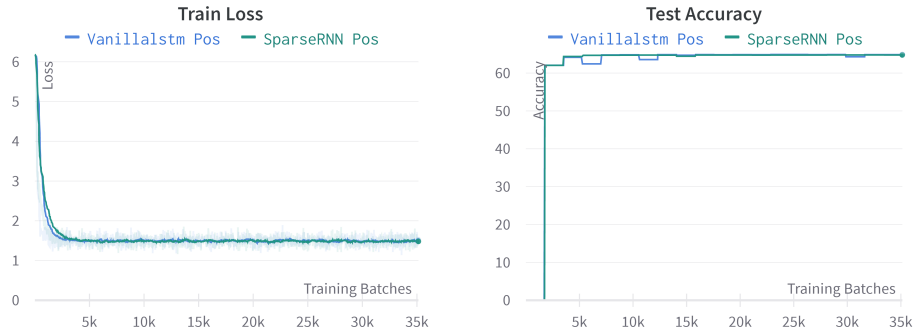


Fig. 3 Part of Speech Tagging

Approach	**POS tagging accuracy (%)**
RNN	-
Vanilla-LSTM	64.879
SparseRNN	**64.77**

Approach	**Sentiment analysis accuracy (%)**
RNN	80.62
Vanilla-LSTM	84.96
SparseRNN	**85.79**

Fig. 4 Results of RNN,LSTM and SparseRNN on POS-tagging and sentiment analysis

the specific architectural choices we used is simple and motivated by considering the computational complexity in a serialized setting, and thus it has desirable properties that make our approach a suitable candidate for low-cost resource constrained settings.