FastGRNN: A Fast, Accurate, Stable and Tiny Kilobyte Sized Recurrent Neural Network

Aditya Kusupati †
Manish Singh †
Kush Bhatia †
Ashish Kumar †
Prateek Jain †
Manik Varma †
† Microsoft Research India
† Indian Institute of Technology Delhi
† University of California Berkeley

Recurrent Neural Networks (RNNs)
State-of-the-art for analysing sequences & time series

Limitations of Existing RNNs

- Traditional RNNs: Training is unstable
- Unitary RNNs: Expensive to train and lack accuracy
- Gated RNNs: Large model size and prediction costs

Our Solutions: FastRNN for provably stable training & FastGRNN for state-of-the-art performance in 1-KB size models

Gradients & Theorems

- Setting α = 1/\(T\), β = 1 − α stabilizes the gradients
- FastRNN has convergence rate and generalization error upper bounds independent of \(T\)

FastRNN

- Provably stable training with a residual connection having 2 additional scalars
- Accuracy: RNN < Unitary RNNs < FastRNN < Gated RNNs

FastGRNN

- Extend α & β from scalars to vector gates
- Make \(U\) and \(W\) lower rank (L), sparse (S) and quantized (Q)
- Accuracy: RNN < Unitary RNNs < Gated RNNs < FastGRNN

Architectures' Equations

Simple RNN

\[ h_t = \sigma(W_x + U h_{t-1} + b) \]

FastGRNN

\[ h_t = \sigma(W x + U h_{t-1} + b) \]

\[ 0 \leq x, \beta, \nu \leq 1, \]

and are trainable scalars, parameterized by the sigmoid function

\( \sigma(\cdot) \) can be any non-linearity

\( \odot \) is Hadamard product

Results

- Accuracy (%)
- Model Size (KB)
- F1 Score
- Model Size (KB)

Code: https://github.com/Microsoft/EdgeML

Billions of these resource-constrained devices form the IoT ecosystem

<table>
<thead>
<tr>
<th>Arduino MKR1000 - Time (ms)</th>
<th>Google-12</th>
<th>Google-30</th>
<th>Yelp-5</th>
<th>ISA-19</th>
<th>Pixel MNIST-10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google-12</td>
<td>10000</td>
<td>7500</td>
<td>5000</td>
<td>2500</td>
<td>5000</td>
</tr>
<tr>
<td>Google-30</td>
<td>7500</td>
<td>5000</td>
<td>2500</td>
<td>1250</td>
<td>2500</td>
</tr>
<tr>
<td>Yelp-5</td>
<td>5000</td>
<td>2500</td>
<td>1250</td>
<td>625</td>
<td>1250</td>
</tr>
<tr>
<td>ISA-19</td>
<td>2500</td>
<td>1250</td>
<td>625</td>
<td>312</td>
<td>625</td>
</tr>
<tr>
<td>Pixel MNIST-10</td>
<td>1250</td>
<td>625</td>
<td>312</td>
<td>156</td>
<td>312</td>
</tr>
</tbody>
</table>