FastGRNN: A Fast, Accurate, Stable and Tiny Kilobyte Sized Gated 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

\[ \mathbf{U} = f(\ldots, \mathbf{W}^T \mathbf{Q} = \mathbf{Q} \left[ \begin{array}{c} \mathbf{Q}^T \end{array} \right] \mathbf{Q}^T, \ldots) \]

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 1KB size models

Gradients & Theorems
- Setting \( \alpha = O(1/T), \beta = 1 - \alpha \) 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
- \( \alpha, \beta \) from scalars to vector gains
- Make \( \mathbf{U} \) and \( \mathbf{W} \) lower rank (L), sparse (S) and quantized (Q)
- Accuracy: RNN ≤ Unitary RNNs < Gated RNNs = FastGRNN

Architectures’ Equations
- Simple RNN
  \[ \mathbf{h}_t = \sigma (\mathbf{W}_t \mathbf{x}_t + \mathbf{U}_t \mathbf{h}_{t-1} + \mathbf{b}) \]
- FastGRNN
  \[ \mathbf{h}_t = \sigma (\mathbf{W}_t \mathbf{x}_t + \mathbf{U}_t \mathbf{h}_{t-1} + \mathbf{b}) \]
  \( 0 \leq \alpha, \beta, \gamma \leq 1 \), and are trainable scalars, parameterized by the sigmoid function \( \sigma(\cdot) \) can be any non-linearity

Results
- Accuracy (%)

Model Size (KB)

Arduino MKR1000 - Time (ms)

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

Billions of these resource-constrained devices form the IoT ecosystem.