Sparse Neural Networks:
Improving Feedforward Neural Network Efficiency with Sparse Matrix Constructions

Dr. Krzysztof Choromanski
Google Research, NY

EMAIL@YOURDOMAIN.EDU

Alfred Bouérly
Columbia University

EMAIL@COAUTHORDOMAIN.EDU

Patrick Boueri
Columbia University

EMAIL@COAUTHORDOMAIN.EDU

Abstract

We show empirical evidence that sparse connectivity in between layers in a Feed Forward Neural Network does not impact accuracy significantly, as compared to a fully connected layer. Using the canonical MNIST data set, we compute accuracy measures for many Feed Forward Neural Nets with different connection schemes and topologies, showing there is no significant drop off as low as 10% of connections in the fully connected setting. With sparse connection paradigms, neural networks can be stored with less memory and predictions can be made with faster or with less computation resources, suggesting suitable applications in IoT.

1. Introduction

- Feed Forward Networks, construction, theory and successes (LeCun et al., 1998)

- Problems with Neural Networks, overspecification?

2. Experimental Setup

2.1. Sparse Construction

- Intro to sparsification (What we mean by Sparse)

- How we impose sparsification in training

- Classes of graphs:

  - Random
  - Circulant
  - Pseudo-Random TBD

- Big O Notation of sparse vs. FC for all

2.2. methods

Keraspatal, training params, and hardware

3. Results

Here we discuss the actual measured results in terms of accuracy, and come up with really good metrics of accuracy and a full picture of what the differences are between the Fully Connected and Sparse Paridigm

3.1. Example Image

4. Analysis

Fully Connected Networks do not add as much accuracy, over parameterized, with quick wins coming with very little connections, and additional ones, adding very little. Something like that.

5. Future Work

- Implementing in different Neural Nets (Convolutional neural nets, are in a sense sparse)

- Training analysis (does it take longer, shorter, more or less local minima?)

- Datasets where it succeeds, where it fails
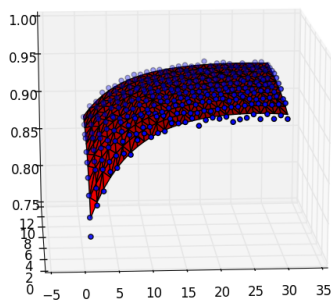
- More information on topologies vs. sparsity

Figure 1. Example Result from Structured.

- Computational resources were a big limitation

- Maybe sparsifying intentionally during training (Drop out and Stay out)

6. Conclusion

Bring home major point : most of NN accuracy, comes with few connections
Link to IoT if possible (it's 'hot')

References

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998.