Training Autoencoders in Sparse Domain

Biswarup Bhattacharya, Arna Ghosh* & Somnath Basu Roy Chowdhury*
University of Southern California (USA), McGill University (Canada), Indian Institute of Technology Kharagpur (India)

Abstract

Autoencoders (AE) are essential in learning representation of large data (like images) for dimensionality reduction. Images are converted to sparse domain using transforms like Fast Fourier Transform (FFT) or Discrete Cosine Transform (DCT) where information that requires encoding is minimal. By optimally selecting the feature-rich frequencies, we are able to learn the latent vectors more robustly. We successfully show enhanced performance of autoencoders in sparse domain for images.

Introduction

Autoencoders are essential in learning representation of large data (like images) for dimensionality reduction. [1] described an effective way of initializing the weights that allows deep autoencoder networks to learn low-dimensional codes that work much better than principal components analysis (PCA). In this paper, we convert the input images to a sparse domain using transforms like Fast Fourier Transform (FFT) or Discrete Cosine Transform (DCT). Fourier analysis converts a signal from its original domain (usually time or space) to a representation in the frequency domain (and vice versa). A Fast Fourier Transform (FFT) rapidly computes such transformations by factorizing the Discrete Fourier Transform (DFT) matrix into a product of sparse factors, thereby, reducing the complexity of computing the Discrete Fourier Transform.

Our goal: To develop a method to improve information encoding of large data points in a deep learning framework using sparse domain representation.

Key Idea: We propose an encoding scheme of the deep learning framework autoencoder by representing the input in Fourier domain which, being a sparse domain, is easier to encode without significant information loss.

Methodology

We use an autoencoder framework to map the Fourier domain representations of closely linked images - alike images from same distributions to map to similar distributions. This is inspired from the fact that motion cues or temporal cues act as weak supervisory signals for object detection [2] [3].

Architecture

We use a simple two-layered autoencoder network, although the proposed model is applicable to other stacked autoencoder models as well. The network reduces the dimension of MNIST dataset from 784 to 20. Table 1 depicts the entire network architecture. Layers 1 & 2 in Table 1 denote the encoder module and Layers 3 & 4 constitute the decoder module.

Layer	Type	Maps and Neurons
1	Linear	784×400
2	Linear	400×20
3	Linear	20×400
4	Linear	400×784

Table 1: Autoencoder Network architecture

Experiments

We use the MNIST handwritten digits dataset for experimentation. The training set has 60000 images and the test set has 10000 images. The network is trained using Adam Optimizer with a learning rate of 0.001. The batch size is kept at 256 images and the network is trained for 20 epochs.

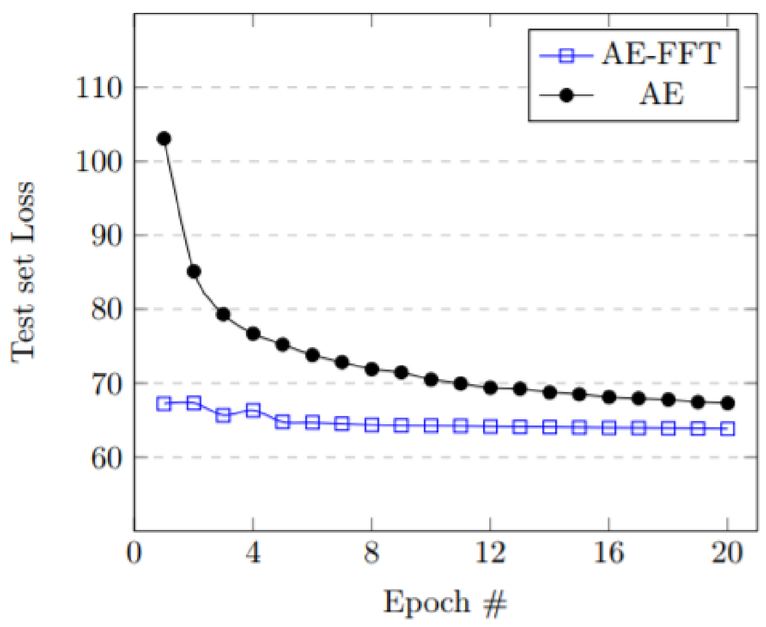

Figure 1: Test-set Loss for 2-layered AE-FFT and AE

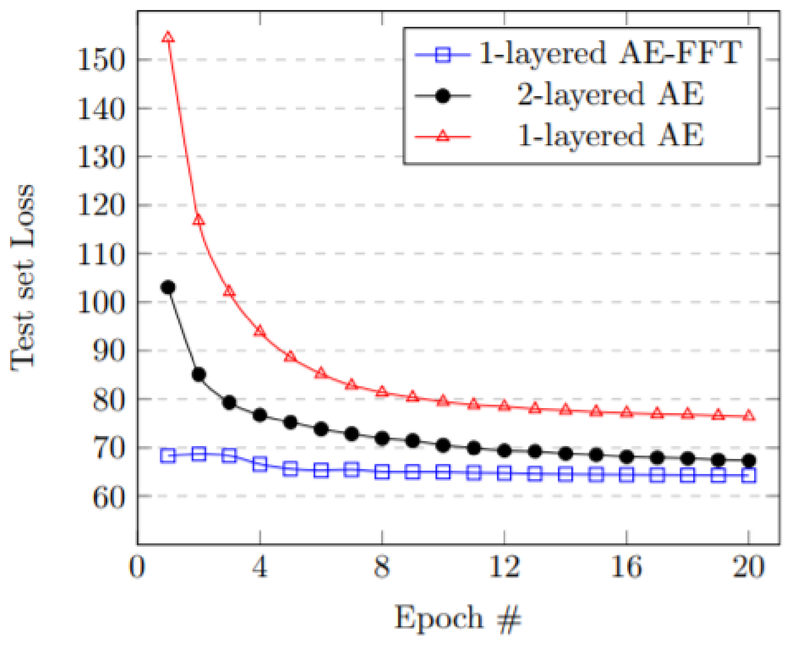

Figure 2: Test-set Loss for AE-FFT, 1-layered vanilla AE & 2-layered vanilla AE

We show that by optimally selecting feature-rich frequencies, we are able to enhance performance of prevalent autoencoder architectures. This reduces the dependency of autoencoder architectures on huge datasets.

References

[1] Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. Science 313(5786):504– 507.

[2] Agrawal, P.; Carreira, J.; and Malik, J. 2015. Learning to see by moving. In Proceedings of the IEEE International Conference on Computer Vision, 37–45.

[3] Goroshin, R.; Bruna, J.; Tompson, J.; Eigen, D.; and LeCun, Y. 2015. Unsupervised feature learning from temporal data. arXiv preprint arXiv:1504.02518.

* Equal contribution