Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2024:7**
http://www.math.su.se

Toward Decoding The Abstract Image Representations in Neural Networks

Martin Björklund*

June 2024

Abstract

In this thesis, we present a decoder that reconstructs images from high-dimensional abstract representations inside a neural network. As our model of interest we use a ResNet-18 model trained on the CIFAR-10 image data and investigate 6 checkpoints in the model, which we use as input to the decoder. Drawing inspiration from autoencoders, our decoder mirrors the architecture of ResNet-18, aiming to reverse the sequence of operations that it has applied to the original images in the dataset.

We find that the decoder works well for checkpoints placed early in the network but that the quality of the reconstructed images deteriorates as one moves toward the output layer in the neural network, resulting in a higher mean squared error for the reconstructions. Moreover, we examine how the decoder reconstructs images based on artificially generated abstract representations.

As a further assessment of the decoder's performance, we let ResNet-18 classify each reconstructed image. Based on its accuracy on the reconstructions, we find that the decoder does not generally preserve the features that are necessary for accurate classification. While this could be due to a loss of information in each checkpoint, we hypothesize that it is because of the loss function used for the decoder. We propose introducing suitable regularization terms to the loss function to ensure that the decoder preserves features in the representations that are relevant for classification.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: kmartinbjorklund@gmail.com. Supervisor: Chun-Biu Li.