Advanced Machine Learning: Final Project

Aman Gulati
ag3743

Joaquim Lyrio
jc4637

Ricardo Pommer
rap2194

Henrique Saboya
hs2923

December 8, 2017

Abstract

Neural networks, although empirically validated, remain a theoretically opaque tool. Once we have trained a particular architecture, what within its layers, neurons, or channels, drive its classification performance? How do we interpret the results beyond out-of-sample loss measures? Feature visualization attempts to elucidate this question by, generally speaking, projecting intermediate layers to pixel-space. In this report, we review and consolidate different methods of feature visualization, their definitions, theoretical approaches and performance on three convolutional neural networks (CNN).

0.1 Definitions

As any young field, neural networks and their interpretability has yet to agree on definitions for a vast number of methods and parameters. Here, we provide our own that may at times conflict with other authors'.

0.1.1 Feature Visualization

We refer to **feature visualization** as any projection that relays information about a subset of the network's architecture. The distinction will become increasingly important as we delve into the details of each technique. Furthermore, some authors (Olah et al., 2017) distinguish between **feature visualization** and **attribution**. We draw no such distinction.

1 A Simple Example

To motivate our survey, we begin with a toy model and dataset of the letters "E", "F" and "L". The network and dataset are based on (Orbanz, 2017) as presented in lecture notes. We generated 2,000 images of each category and added noise distributed as $\prime, \prime . \in$, where all negative noise samples were brought back to 0.