Project Proposal - EE367
From Code to Image - Inverting Neural Networks

Timon Ruban, Jayant Thatte, Vincent Sitzmann

February 8, 2016

1 Introduction

Convolutional Neural Networks (CNNs) can be used to represent images as a vector embedding (code) in a higher abstraction feature space. Gaining a better understanding of such a lower-dimensional representation of images can be helpful in developing new approaches to old problems like 3D reconstruction and novel view synthesis (e.g. choosing the desired features of a scene and then inverting the code to reconstruct a real image).

2 Related Work

An exemplary application is demonstrated in [1], who use a vector embedding corresponding to a 3D representation of computer generated models and then train a CNN to synthesize novel views of 3D objects.

Recently, there has been a lot of effort in inverting neural networks to visualize and understand what a network learns. [3] solves an optimization problem to find an image whose code looks similar to a given code and in addition is imposed with a natural image prior. [5] and [4] use a similar technique to find an image that maximizes a certain class score. Older approaches involve deconvolutional neural networks that manipulate backpropagation to visualize the maximal gradients of the network with respect to the input image ([6]). All of this work focused on understanding CNNs trained to do image classification, like the famous AlexNet ([2]).

3 Project Overview

In this project, we will attempt to better understand what information can be recovered from a code extracted from a network trained to understand 3D object representations. We plan to use a CNN model developed by the Computer Vision and Geometry Laboratory at Stanford and a corresponding dataset consisting of several million images annotated with camera parameters. The network was trained to understand the 3D pose of objects from the training imagery.

The first part of the project is to go from code to image by using inversion techniques similar to the ones used in [3] and [5]. An interesting aspect could be to explore which image priors (e.g. isotropic or anisotropic TV) leads to the

most interpretable images. The second part could involve training a CNN that takes a code as input and outputs a reconstructed image.

We hope that this project will be able to give unique insight and shed light into what features of an image are relevant to recognize the 3D orientation of an object. Such visualization would be be useful in building deeper understanding into the use of CNNs in problems of novel view synthesis and 3D reconstruction.

4 Disclaimer

This project is done in cooperation with Jayant Thatte and Vincent Sitzmann and as a joint project for CS231n.

References

[1] Alexey Dosovitskiy, Jost Tobias Springenberg, and Thomas Brox. Learning to generate chairs with convolutional neural networks. *CoRR*, abs/1411.5928, 2014.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, page 2012.

[3] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. *CoRR*, abs/1412.0035, 2014.

[4] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *CoRR*, abs/1312.6034, 2013.

[5] Jason Yosinski, Jeff Clune, Anh Mai Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *CoRR*, abs/1506.06579, 2015.

[6] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. *CoRR*, abs/1311.2901, 2013.