# Interpreting Influence in Convolutional Neural Networks

**Richard Dong**
**11/1/2018**

### Project Web Page

Our project updates will be hosted at https://github.com/rdong97/15-400. This webpage belongs to me and will be maintained through the next semester. Milestone progress reports and project early results will be published as research progresses. This project report will be published in the next website update.

### Project Description

For the following spring semester, I will be working with my project mentor, Matt Fredrikson, who is an Assistant Professor in the School of Computer Science and Institute of Software Research at Carnegie Mellon University. Alongside, I will also be working alongside Klas Leino, a PhD student advised under Prof. Fredrikson who is currently involved in research into transparent deep learning.

Our research topic will cover Convolutional Neural Networks (CNNs), which are a class of neural networks that focus on image based data. Using the underlying image pixels as the input layer, CNNs utilize several layers of convolution and pooling to extract features, which then serve as input to further fully connected layers, finally resulting in an output layer. This output layer allows us to form classification predictions of objects present in the original image.

A major issue of CNNs is their enormous size, comprising of potentially tens of thousands of neurons. As a result, when we back trace the model output, the task of finding an explanation for the output is a major challenge. Additionally, when the model makes incorrect decisions, debugging and correcting the issue is extremely difficult. Neural networks are especially prone to issues such as underfitting and overfitting which lead to high error rates on real world data.

Our research works on bringing the notion of transparency to CNNs by exploring ways to add explanability to existing CNN models. Inside a neural network, neuron activation is passed along via connections between layers. When a intermediate neuron is activated, it was the result of the activation of neurons connected in the network below it.  This influence function can take various forms, but the underlying principle is the same. As a result, we can trace a resulting output to the influential neurons of an intermediate layer recursively all the way to the input layer.

By creating a notion of influence, CNN designers will be able to locate regions of neural networks and even representations on an underlying image which led up to the final network classification. Using this tool, the process of debugging and correcting issues with CNNs will become more rigorous and lead to better model results. One major concern that can be addressed for real world applications of CNNs dealing with discrimination and fairness is the issue of counterfactual reasoning, where model output that are clearly incorrect.

## Project Goals

A major goal of the research project is the creation of an domain specific application case study of the principles of influence in CNNs. Implementation of the application will allow us to experiment around with various models and functions for influence. If everything goes well, we will discover a meaningful model that increases network transparency and will create a publication or presentation to show our results.

If progress moves more slowly than expected we may find that our results are not as meaningful as we originally hoped for. The tasks towards the end will be crunched for time and we will likely have a limited time in exploring and tuning various means of representation.

If progress moves more quickly than expected we expect the opposite to happen. We will have more time to explore and search for a model that leads to good transparency in the given application domain. If time permits we may even have enough time to explore an application case study in another domain altogether.

## Milestones

For the remaining portion of the Fall 2018 semester and following winter, I plan to familiarize myself with the current CNN framework our research group is currently working upon which simulates the behavior of a trained CNN on open source image datasets. I will be creating a case study application utilizing the underlying framework. A decision will be made in the upcoming weeks about which domain the application will be built upon and possible representations of influence could we experiment with.

Found below are projected milestone dates for the upcoming Spring 2019 semester.

| | |
|---|---|
| ● February 1st | Familiar with concepts, framework, and notebook environment |
| ● February 15th | Case study outline with skeleton code |
| ● March 1st | Working case study with preliminary model results |
| ● March 22nd | Exploration of 2-3 representations of influence |
| ● April 5th | Testing model corrections and tuning using influence models |
| ● April 19th | Preliminary draft of research findings |
| ● May 3nd | Presentation or publication of research findings |

## Literature Search

Transparency and fairness have been recent hot topics in the realm of machine learning and lots of literature has been published in subfields including computer vision and convolutional neural networks. Researchers at numerous universities and major technology companies are presently working to discover new techniques and create new tools.

One major publication our work will be based off was co-authored by my project mentor, Prof. Fredrikson is *Influence-Directed Explanations for Deep Convolutional Networks, 2018*. Additional papers have been explored which go into detail of other techniques used to analyze influence, such as influential neurons in network slices and gradient maps. Some of these papers are a bit dense in content and will require additional readings and discussions with the project team.

**Resources Needed**

A major hardware need for this research project is graphics cards which are used to simulate and perform the iterative calculations of large scale neural networks on large datasets of images. A major software our group is currently utilizing is the Jupyter Notebook environment, which combines programming, document creation, and data visualization. The hardware has already been purchased and is being provided by the research group and the software is free to download.