Interpreting Influence in Convolutional Neural Network Decisions

Richard Dong
10/31/2018

Team

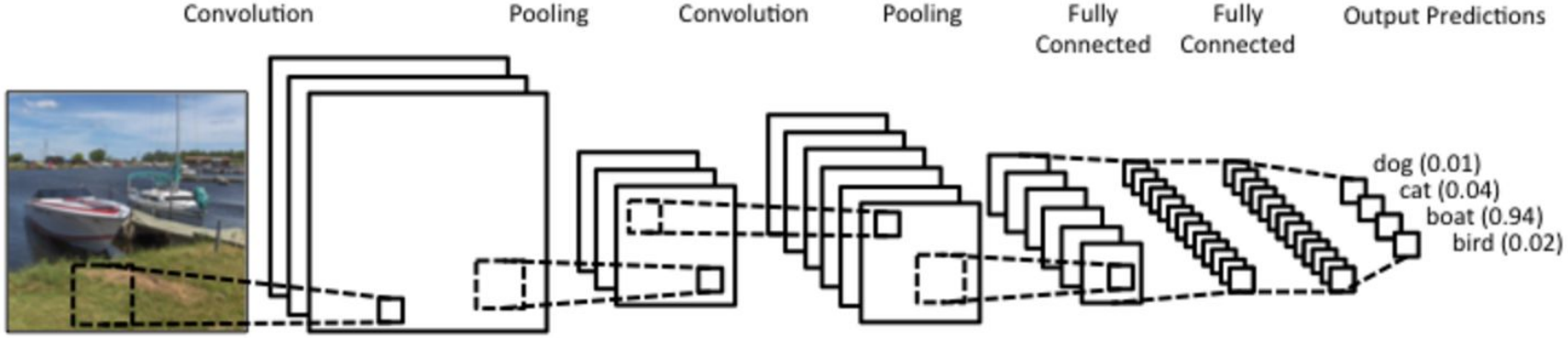
Project Mentor

- Matt Fredrikson
  - Assistant Professor
  - SCS/ISR

Graduate Student

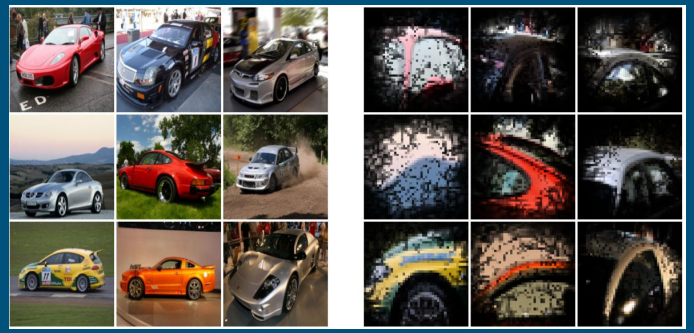
- Klas Leino
  - PhD advised under Prof. Fredrikson
  - Concentration in transparent deep learning

Project Background - Convolutional Neural Network (CNN)



Influence and Explainability

- Input layer contains **features**: clusters of pixels that allow us to conclude whether or not an object is present in a given image



Explaining Influence

1. Underlying pixels lead to intermediate layer neuron activation
2. Internal neuron activation causes change on output layer result

Application: Case Study and New Interpretation Techniques

- Find interpretations for influential components in the instance of counterfactual reasoning
  - Explain why a model produced incorrect results
- Utilize existing framework to develop domain-specific application
  - Exploring different representations (percentiles of most influential pixels, regions, etc.)
  - Additional techniques for measuring influence such as utilizing numerous layers and combinations of neurons

Project Impact - Power of Explainability

- Recent Deep Learning and Convolutional Neural Networks designs are very large (thousands and thousands of neurons), leading to opacity and black-box behavior
- Debugging and fixing models becomes exceedingly difficult because we don’t know which feature caused the behavior
  - Ex. Models need to be trained, and black-box behavior leads to underfitting and overfitting on test data
- Useful in real world applications where fairness and discrimination play into factors

Citations

Leino, Klas, et al. “Influence-Directed Explanations for Deep Convolutional Networks.”

arxiv.org/pdf/1802.03788.pdf

Matt Fredrikson. www.cs.cmu.edu/~mfredrik/research.html