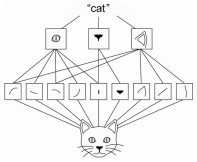
How to Read the "Mind" of a Neural Network

Thomas Walton, Advisor: Mohammad R. Hasan

Introduction & Motivation

- Deep Learning (DL) is a powerful tool for solving computer vision problems, but has shortcomings that are not well understood
- Algorithmic bias plagues the decision-making process of DL models, leading to nongeneralizable results and limited capability



A CNN analyzing an image of a cat

- A Convolutional Neural Network (CNN) is a DL model used to learn the semantic identity (category) of data
- CNNs are intelligent if they can generalize knowledge learned from data onto novel, unseen data
- Algorithmic bias limits the capability of CNNs to learn

Objective

- Use Class Activation Maps (CAMs) to read the "mind" of a CNN
- Create ScoreCAM maps to make these CAMs human readable
- Leverage this analysis to discover algorithmic bias

A ScoreCAM map of a dandelion

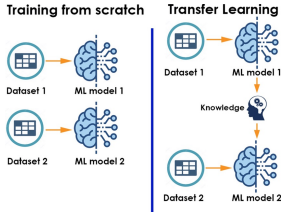
Scientific Research Questions

- SRQ1: What are the reasons for a vision model to fail in its predictions?
- SRQ2: When a vision model identifies an object accurately, does it necessarily mean that it recognizes the object in the image? If not, then why?

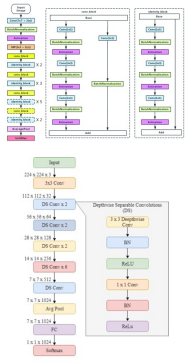
The decision-making process is filtered through layers of neurons, each contributing to the final prediction

Methodology

- Analyze ScoreCAM maps of a scratch trained model and a transfer learning model
- Identify biases and shortcuts in the learning process

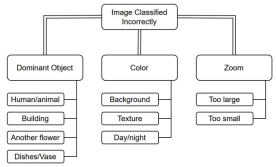


- Analyze the behaviors of ResNet-50 (scratch) and MobileNetV3 (transfer)



Results

SRQ1: Both ResNet-50 and MobileNetV3 had similar downfalls



Flowchart of common inaccuracy reasons

CNNs can make mistakes for various reasons, often due to confusion with how to process images outside the norm

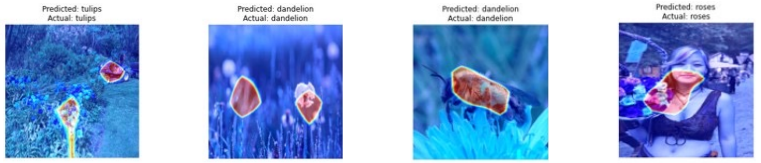
Conclusions

- On their own, SRQ1 and SRQ2 tell only part of the story of why algorithmic bias occurs
- When failures from SRQ1 are combined with suspicions from SRQ2, bias is exposed

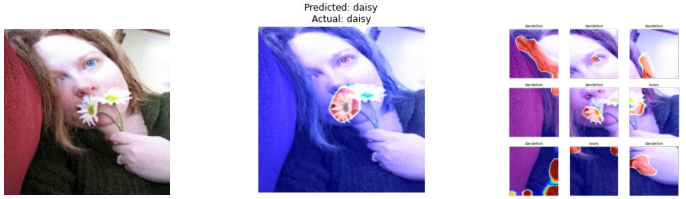
Adding unexpected objects lead to misclassification on images where the model was once confident

- By adding objects that lead to failure onto images where the vision model was confident, weaknesses in CNNs become apparent
- The lack of adaptability from this vision model implies that while it may be good at classifying flowers, it is not truly learning intelligent representations

SRQ2: While a model may seem to have learned intelligence, further analysis shows that this may not be true



Despite being correct in their predictions, both models used shortcuts or other objects to draw conclusions



While the CNN correctly classifies the image, when broken up, trouble arises

Questions? Email tomwalton262@gmail.com