Interrogating, understanding, and improving deep learning

Deep learning -- a machine learning paradigm that involves training highly-parameterized, multi-layer neural networks -- has deservedly garnered a lot of attention for its powerful, predictive abilities when given access to large compute and sizable datasets. However, as deep learning begins to be applied to high-impact yet high-risk domains like autonomous driving and medical imaging, there is a need to develop theory and tools to critically interrogate and understand what these networks are learning, as their highly-parameterized design lends themselves to be opaque in nature and thus able to hide unwanted or even malicious biases.

Consider the confirming and transferring qualities of human biases. Today, the majority of U.S. Republicans and Democrats view members of the other party as deeply immoral and find it difficult to comprehend their views (Fingerhut, 2016). This inability to understand another's perspective makes fertile ground for biases to develop and "other" another. Unfortunately, we have a tendency to turn to like-minded people to confirm our own personal values rather than truly seek to understand our own biases and another's viewpoint. Human biases can also transfer to algorithmic ones, which often occurs when datasets reflect human biases. For instance, blink and face detection software frequently fail on Asian and Black faces because they have been trained on datasets with predominately Caucasian faces (Rose, 2010; Tucker, 2017; also see Furl et al., 2002).

With the transfer of human biases from dataset to deep learning practitioner in mind, I propose research that 1. critically interrogates what convolutional neural networks (CNNs) are learning, from each example (i.e., an image), to each labelled concept (i.e., a "cat"), to whole networks themselves, and 2. seeks to understand how concepts are encoded in CNNs and then transferred across tasks, models, and modalities, in order to 3. improve CNNs to be "de-bias-able" and robust against bias in the first instance.

After the revival of CNNs in 2012 (Krizhevsky et al., 2012), visualization techniques that explained "where" a classification network "looked" in a given image soon followed (Simonyan et al., 2013; Zeiler and Fergus, 2014). However, these initial methods primarily tried to produce nice qualitative visualizations to the human eye, and were often used to illustrate that the concepts learned were in line with what a researcher intended. Furthermore, research on explanatory visualizations also began to optimize for performance on other computer vision tasks, such as object localization and segmentation. Similar to the U.S. political spectrum, these latent objectives can make researchers susceptible to a mode of confirmation bias in which we do not easily recognize our own model's inconsistencies. This motivated the development of a more principled visualization metric that could also expose network failure cases, which operates by learning the minimal spatial regions in an image that, when perturbed, maximally deleted evidence for a given classification (Fong and Vedaldi, 2017). We also introduced a novel metric that evaluates how well any visualization technique can identify such minimal, indispensable image regions. Lastly, we also demonstrate how our method can be used to highlight when a network has learned a highly predictive yet spurious correlation that reflects dataset bias rather than a desired causal relationship.

Beyond interrogating what CNNs are "looking" at in a given example, we are also interested in understanding the internal representations of concepts that CNNs are learning. Several works on explaining CNNs focus on describing what individual CNN features encode (Zhou et al., 2015; Bau et al., 2017). While easily human interpretable, this approach over-simplifies CNNs by ignoring how they leverage multiple filters to encode concepts. We propose developing empirically-supported theory and techniques to describe how the co-activation of different CNN filters are used to encode information. Borrowing inspiration from neuroscience, we plan to probe how CNNs respond to internal "lesions", i.e., perturbations to CNN features,

and develop tools for analyzing their effects. Extending our prior work on spatial perturbations (Fong and Vedaldi, 2017), we have begun such research by learning feature perturbation vectors that describe which combinations of CNN filters -- and to what degrees -- are indispensable for encoding a variety of human-interpretable concepts, such as texture, objects, and scenes. We also plan to conduct our network probing research with multi-modal datasets, i.e., ones containing captions and images, so that we can explain CNN encodings linguistically as well as visually.

Furthermore, we plan to investigate how learned representations are transformed when transferred from one model to another, i.e., distilled (Hinton et al., 2015) or compressed networks like SqueezeNet (Iandola et al., 2016), or from one task to another, i.e., fine tuned or multi-task learning. We also are interested in understanding how different models -- from different CNN architectures, such as AlexNet and ResNet, to comparisons with human neural networks as captured in fMRI or EEG recordings -- encode information differently and are affected by architecture design and constraints, i.e., information bottlenecks such as that induced by the number of filters in a layer. In Fong et al., 2017, we demonstrate how a computer vision model can benefit from a form of transfer learning in which it is encouraged to better match the internal visual representation that the human brain has learned, as recorded in fMRI scans. In future work, we plan to focus on understanding what differences in learned representations between human and artificial neural networks confer such performance improvements.

Finally, we propose leveraging the knowledge learned about the internal representations of CNNs to make them more robust against dataset bias (Torralba and Efros, 2011) and malicious, adversarial attacks, which add a pattern of noise to an input image that causes it to be misclassified easily by CNNs (Goodfellow et al., 2014). To do this, we plan to develop a CNN debugging tool to understand how adversarial examples and failure cases deviate statistically from natural images in feature space. We have already demonstrated that our visualization method can be used for unsupervised defense against adversarial examples (Fong and Vedaldi, 2017). We also plan to develop techniques to quantify and attribute sources of network bias using probe datasets with annotations along common dimensions of human biases, i.e., gender, age, race.

In conclusion, given how pervasive and implicit human biases can be, we suspect such biases have been unknowingly transferred to the representations of state-of-the-art machine learning models. Similar to how unity across political divides might be more feasible if individuals sought to understand the beliefs of others rather than confirm their own, we suggest developing theory and tools for explaining CNNs that seek to rigorously interrogate what they have learned and not just affirm what we hope they have learned.

References

Bau, D., et al., CVPR 2017. Network Dissection.

Fingerhut, H., Pew Research 2016. Partisanship and Political Animosity in 2016.

Fong, R., et al., Scientific Reports 2017. Using Human Brain Activity to Guide Machine Learning (in press).

Fong, R. and Vedaldi, A., ICCV 2017. Interpretable Explanations of Black Boxes by Meaningful Perturbation.

Furl, N., et al., Cognitive Science 2002. Face recognition algorithms and the other-race effect.

Goodfellow, I.J., et al., arXiv 2014. Explaining and harnessing adversarial examples.

Hinton, G., et al., arXiv 2015. Distilling the knowledge in a neural network.

Iandola, F.N., et al., arXiv 2016. SqueezeNet.

Krizhevsky, A., et al., NIPS 2012. Imagenet classification with deep convolutional neural networks.

Rose, A., Time 2010. Are Face-Detection Cameras Racist?

Simonyan, K., et al., arxiv 2013. Deep inside convolutional networks.

Torralba, A. and Efros, A.A., CVPR 2011. Unbiased look at dataset bias.

Tucker, I., The Observer 2017. 'A white mask worked better': why algorithms are not colour blind.

Zeiler, M.D. and Fergus, R., ECCV 2014. Visualizing and understanding convolutional networks.

Zhou, B., et al., ICLR 2015. Object Detectors Emerge in Deep Scene CNNs.