The following papers are important to me because they have been influential for my past research. They are relatively narrow in scope (i.e., mostly cover adversarial examples), mostly because this is the focus of my previous research. I am currently interested in a much broader set of topics than adversarial examples, although I especially enjoy how these papers aim to deeply understand and explain machine learning systems.

Adversarial Examples Are Not Bugs, They Are Features. Ilyas et al. (2019) https://arxiv.org/abs/1905.02175

- Paper demonstrates that the non-semantic patterns observed in adversarial examples are "real" in the sense that they are predictive patterns that exist in the datasets (both training and testing)
- Counters previous thought that adversarial examples are artifacts of how neural networks draw decision boundaries
- Explains why adversarial examples are often transferable
- Proposes methodology to train a neural network classifier that only relies on adversarial patterns and a dataset in which all predictive features are robust (used in my own research)
- Leaves open the question, which I try to answer in my research: are non-robust features patterns that we as humans don't pick up on but are real in nature/photography, or are they something else, such as an artifact of looking for patterns that we as humans do find useful?

Feature Purification: How Adversarial Training Performs Robust Deep Learning. Allen-Zhu et al. (2020) https://arxiv.org/abs/2005.10190

- Argues (theoretically) that two-layer neural networks trained with the non-robust (standard) objective on linear sparsely generated data will learn a dense mixture of the generating features. This means that each neuron in the first layer of the neural network will have weights that are primarily a single basis vector from the generator, plus a small component of every other basis vector. This leads to adversarial vulnerability.
- Argues that when trained with the adversarial-training objective, each neuron in the first layer will be a basis vector from the generator with no added component (i.e., will be a "pure" basis vector or feature), and this leads to adversarial robustness.
- Contains empirical (mostly qualitative) analysis to suggest that this might also be true for deep neural networks.
- Suggests, in the sparse-linear case, that the adversarially vulnerable features are a linear combination of the real generators. This would explain (again, in the

simplified sparse-linear case) why adversarial examples contain predictive generalizing features, as shown empirically by Ilyas et al. (above paper).
- Also suggests to me that adversarial examples (in the sparse-linear case) are related to the robust (pure) features, which is related to my work.

Are perceptually-aligned gradients a general property of robust classifiers? Kaur et al. (2019) https://arxiv.org/abs/1910.08640
- This is a short workshop paper but highlighted a fascinating property of neural networks that was unexpected to me.
- The authors find that computing a smoothed gradient of a single non-robust (standard) neural network would yield analogous semantic-looking gradients to adversarially-trained neural networks. They phrase this in the paper as computing the gradient of a smoothed classifier.
- Perhaps contrary to prior thought, this suggests that non-robust neural networks are looking for, at least in part, semantic looking features. This is some evidence that non-robust features are in some way related to semantic features in the same way as robust features.

The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. Frankle et al. (2018) https://arxiv.org/abs/1803.03635
- Somewhat different from the other papers.
- This paper is fascinating to me because it conflicts with my intuitive notion of how neural networks should learn. The transferability of adversarial examples, and more broadly the universality hypothesis, suggests that neural networks should learn at least overlapping features.
- This paper suggests that what a neural network learns might be to a large extent determined by its initialization.
- Even with a large number of parameters (with many "lottery tickets") how should we expect that neural networks would arrive at a similar enough initialization to learn similar features?
- More generally, demonstrates that gradient descent may be less important than we had previously thought, if gradient descent depends on many lottery tickets (i.e., a large network).

(honorable mention, speculative and more of a blog post but very influential for me) Zoom In: An Introduction to Circuits. Olah et al. (2020) https://distill.pub/2020/circuits/zoom-in/
- The authors visualize neural network circuits and find interesting patterns that appear to be encoded by neural networks but I think the most interesting proposal was their so-called "universality hypothesis."-

- This hypothesizes that different neural networks trained on a highly similar or identical task would learn analogous features.
- If true, it would suggest that exact neural architecture may not matter as much as we previously might have thought. This may limit the amount of necessary training, since we can train a single huge neural network and then re-encode these universal features according to our goal.
- If neural networks learn universal features, then there may be reason to look for these features encoded by the human brain.
- If neural networks learn universal features, then we can empirically analyze a single neural network, and then make claims about the features of every neural network.
- This would explain the transferability of adversarial examples even across highly different architectures.
- This has inspired my own work, where I am investigating the universality hypothesis.