Understand Deep Nets 1

- The lectures present three ways to understand Deep Networks. These are based on three studies.
- The first is work by Yosinski et al. It attempts to generate the stimuli that best activate neurons (features) in the Deep Network.
- The second is work by B. Zhou et al. who estimate the receptive fields by using human observers (Mechanical Turk) to classify the neurons by classifying the types of image patches that they respond to.
- The third method by J. Wang et al. uses clustering algorithms to determine which frequent patterns of feature activity and evaluates them as key-point and semantic part detectors.

Understand Deep Nets 2: Yosinski

- This approach exploits the fact that the activity of a feature/neuron is a differentiable function of the input x $f_{a,l}(x, \omega)$. (Suffices a, are the channel and layer of the network).
- They initiative an input image x_0 and update it by $x_t \mapsto x_t + \epsilon_t \frac{\partial f_{a,l}(x,\omega)}{\partial x}$. This is steepest ascent on the feature response and will generate an image x which activates the feature strongly.
- This converges to images (see Yosinski handout) which are fairly unrealistic, particularly if the images are initialized with random noise (like static on a TV screen). But if we apply a regularization prior which encourages the images to be weakly smooth (recall the weak membrane models from earlier in the course) then the images start looking like roughly similar to objects/parts.
- Note that this is similar to the algorithm for generating adversarial examples (but the goals are very different). This study also shows that the set of all possible images is much bigger than the set of naturally occurring images.

Understand Deep Nets 3: Zhou

- This study argues that Deep Network features represent object parts, if the Deep Networks are trained for object classification, and objects (object parts) if the Deep Networks are trained for scene classification.
- Their method is to first determine the receptive field of each neuron. This means segmenting the image to give the subregion that causes the neuron to fire strongly.
- Then, for each neuron, they take the top-ranked segmented images patches which activate the neuron. Human observers (Mechanical Turk) are asked to annotate the semantics of the neurons by assigning them labels and types of each neuron (from a pre-specified set).
- This shows that neurons in Deep Networks respond to objects and object parts. But it is possible that the same neurons also respond to other images.

Understand Deep Nets 4: J. Wang

- This work studies the feature activity patterns for Deep Networks where the input are vehicles of fixed size. It analyzes population coding (studying all neurons channels, but individual neurons). It uses clustering to identify the most commonly occurring activity patterns, which are called visual concepts.
- Visual concepts are perceptually tight, in the sense that image patches which activate them are visually extremely similar. Visual concepts have good coverage of the object, meaning that they respond to most parts of the object.
- Visual concepts can be tested as detectors for key-points or semantic-parts. They perform fairly well, as an unsupervised method, but are limited because they typically respond to several semantic-parts.
- Visual concepts perform better if several of them are combined to detect the semantic-part (i.e. each visual concept corresponds to a subpart of the semantic-part).