CS 182 Lecture 9: Visualization and Style Transfer 6/29

Hubel & Wiesel: "thought experiments" - visualized stimuli of what cat saw.

By understanding what CNNs "see", we can tap into its creativity.

Visualizing neuron responses → look for certain images that "excite" specific units (filters)
→ see which pixels maximally "influence" value @ filter.

For idea 1: use gradients $\left(\frac{\partial \text{unit}}{\partial \text{pixel of image}}\right)$. Higher = more influence. Use backprop.

"guided backprop" - a hack to drastically improve quality of "vision" by "zeroing out" negative gradients at each layer (~ReLU).

Idea 2: "Optimize" image to maximally excite filter. $x \leftarrow \arg\max_x S(x) + R(x)$
($S(x)$: activations; $R(x)$: regularizer to prevent "crazy" images, e.g. $d(\|x\|_2^2)$)

More nuanced $R(x)$? → "blur" image after updating image and zero out small values.

Using backprop to modify features - "transport" feature distributions to other images.

DeepDream - visualize classes thru "forced hallucination"
↳ literally set $dx = x$ (gradient to activations) to accentuate further

Style Transfer: extracting style, updating spatial positions of content

We can quantify style as a co-occurrence of features → $Cov_{lkm} = \sum [f_{lk} f_{lm}]$
(↑ average over all positions in image.)

new image: $x \leftarrow \arg\min_x \mathcal{L}_{style}(x) + \mathcal{L}_{content}(x)$ Gram matrix: $G_{lkm} = Cov_{lkm}$ ("style")

$$\mathcal{L}_{style}(x) = \sum_l \sum_{km} \left(G^l_{km} - A^l_{km}(x)\right)^2 w_l$$

(G^l_{km}: source image activation; $A^l_{km}(x)$: new image activation)

How can we quantify content? "match" the features.
↑ for a specific layer (choosing said layer is also important)

$$\mathcal{L}_{content}(x) = \sum_{ij} \sum_k \left(f^l_{ijk}(x_{content}) - f^l_{ijk}(x)\right)^2$$