Explaining Image Classifiers by Counterfactual Generation

Chun-Hao Chang, Elliot Creager, Anna Goldenberg, David Duvenaud
University of Toronto, Vector Institute
What is an explanation?

• Something that, if it had been different, would have changed the answer.

• Example: This part of the image makes me think it’s cancer. If it had been the usual color, I wouldn’t have a reason to worry.
How to automate explanation?

• Need:

  1. Automatic answer-giver (i.e. image classifier)
  2. Automatic source of plausible counterfactuals

• Can ask: “What part of this image, had it been different, would have changed the classification”
Gradient Saliency Maps

- Original “saliency maps” simply plot gradient: $d \text{class} / d \text{image}$.
- Answers question: Which direction of change in pixels would most change the label?
- A sort of instantaneous counterfactual.

Simonyan et al., 2014
Conditional Counterfactual Generation

- For image classifiers, need to generate plausible alternative in-fills of images.
- Can use variational autoencoders, or GANs.
- Sum over all possible in-fills:  \( p_M(c | \mathbf{x}_{\backslash r}) = \mathbb{E}_{x_r \sim p(x_r | x_{\backslash r})} \left[ p_M(c | x_{\backslash r}, x_r) \right] \)
The converse question

- Can also ask: “Which part of the image, if the rest were obscured, would keep the class the same?”
- I.e. what is the most important part of the image
- Our method (FIDO): Optimize to mask out as much of image as possible while keeping counterfactual answer same.
Qualitative Results
Number of salient pixels required to change normalized classification score.
Limiting Factors

- Quality of generative models
- Speed of inference (necessary for fast infilling)
- Optimization over shape of masked region
- More than one explanation
Takeaways

- Conditional generative models let us automatically reason about counterfactuals
- Scope to experiment with question being asked