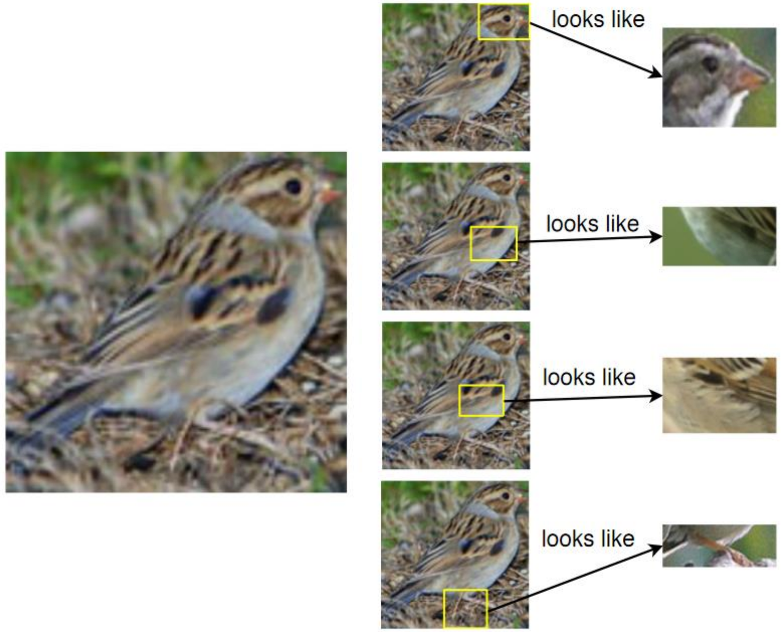
This Looks Like That... Does it?
Shortcomings of Latent Space Prototype Interpretability in Deep Networks

Are Deep Interpretable Networks Always Interpretable?

Prototypical Part Networks



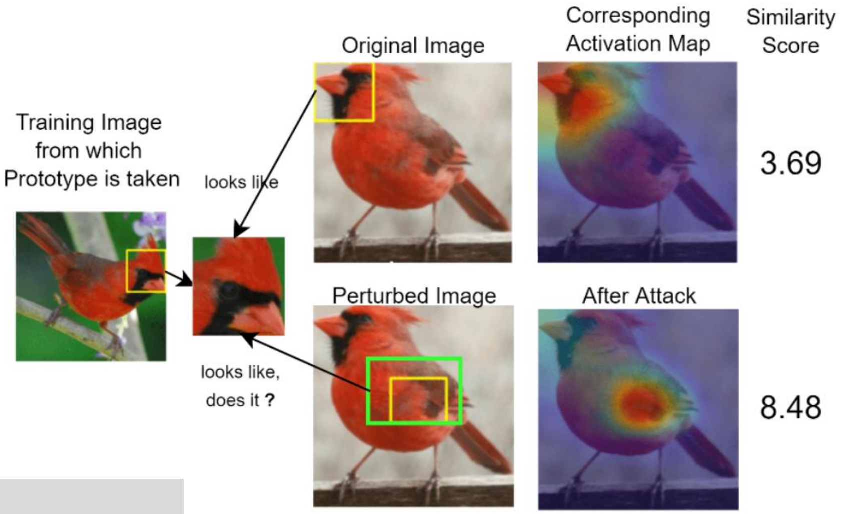
Defining Interpretability for ProtoPNets

Interpretablility holds i.f.f.

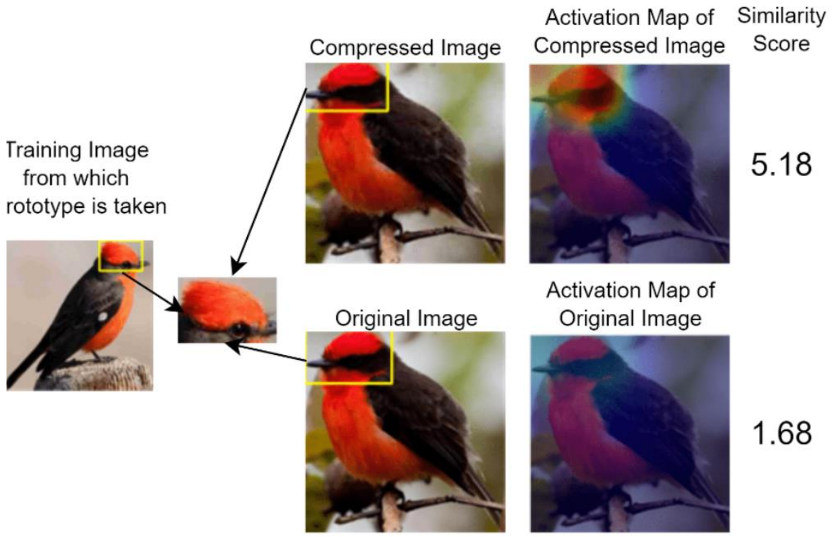
Two image patches look similar to a ProtoPNet ⇔ *Two image patches look similar to a human*

not always true

Head-on-Stomach Experiment



JPEG Experiment


Training Image from which rototype is taken
Compressed Image
Activation Map of Compressed Image
Similarity Score
5.18
Original Image
Activation Map of Original Image
1.68

Key Takeaways:

- Use interpretable architectures like ProtoPNets, **with caution**
  - Interpretations might not faithfully convey the network's underlying reasoning process
  - Be aware of different types of noise in dataset when collecting data from different sources

Adrian Hoffmann*, Claudio Fanconi*, Rahul Rade*, Jonas Kohler, ETH Zurich, Switzerland