Pixel-Grounded Prototypical Part Networks

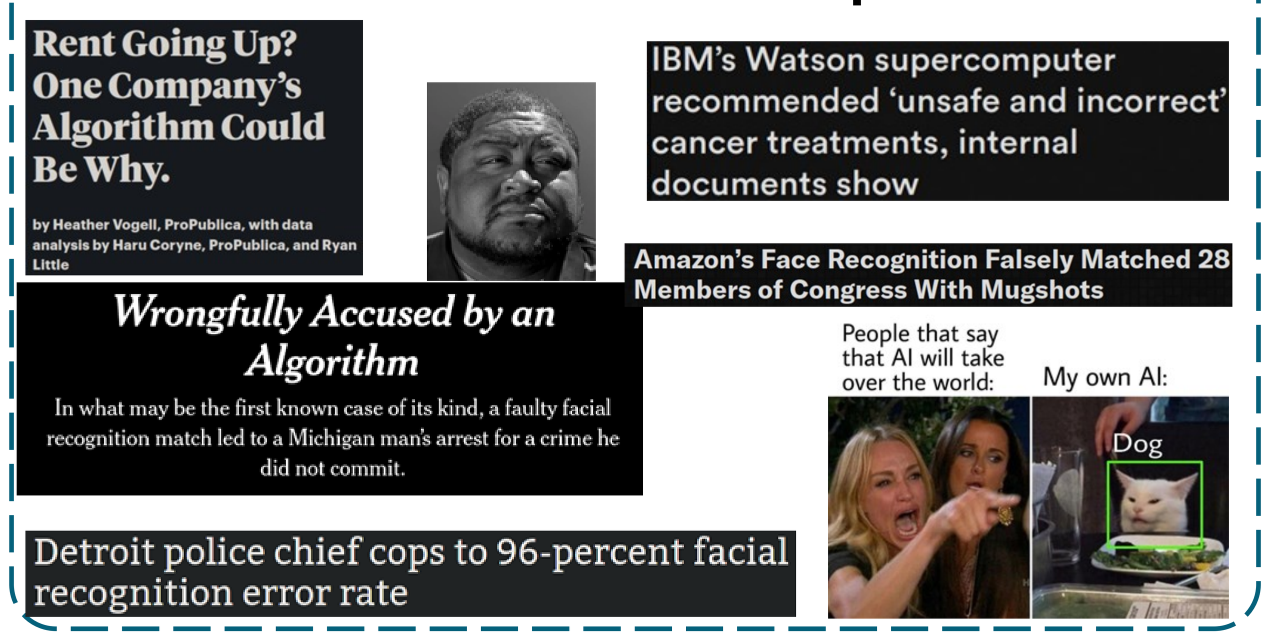
Zachariah Carmichael*[1,2], Suhas Lohit[2], Anoop Cherian[2], Michael Jones[2], Walter J Scheirer[1]

[1]*University of Notre Dame*, [2]*Mitsubishi Electric Research Labs*

**zcarmich@nd.edu*

Background

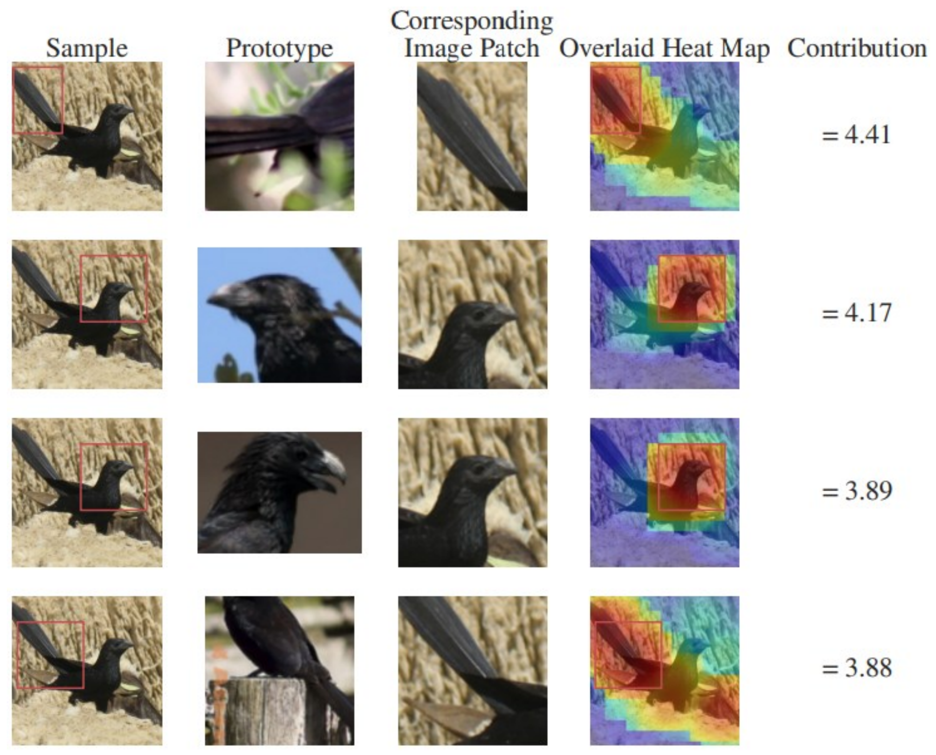
AI has a trustworthiness problem.



Post hoc explanation has a trustworthiness problem.

- Post hoc explainers disagree
- Post hoc explanation fidelity is unverifiable
- Post hoc explainers can be fooled
- Humans can be fooled by explanations
- "Researcher degrees of freedom"

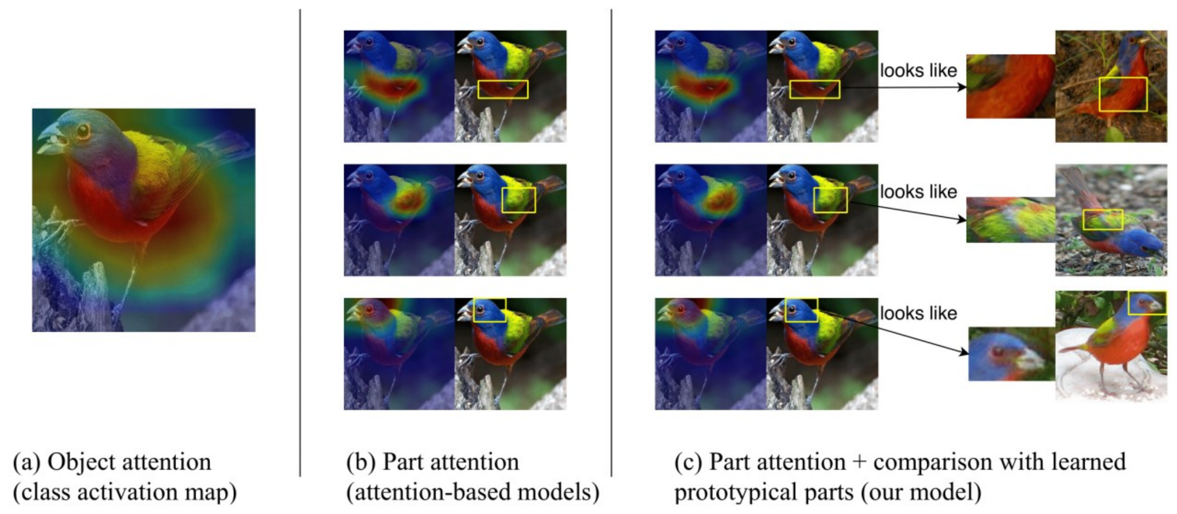
Example Explanation



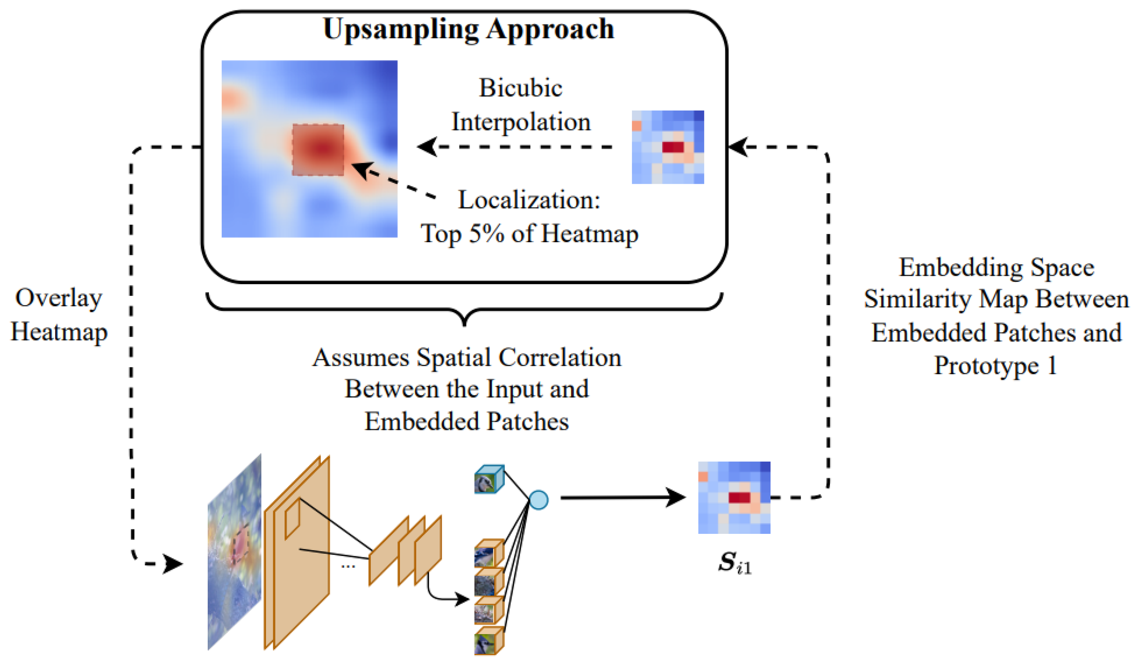
Problem

Prototypical Part Neural Networks

- Goal: Produce explanations of the form, *This* looks like *that*

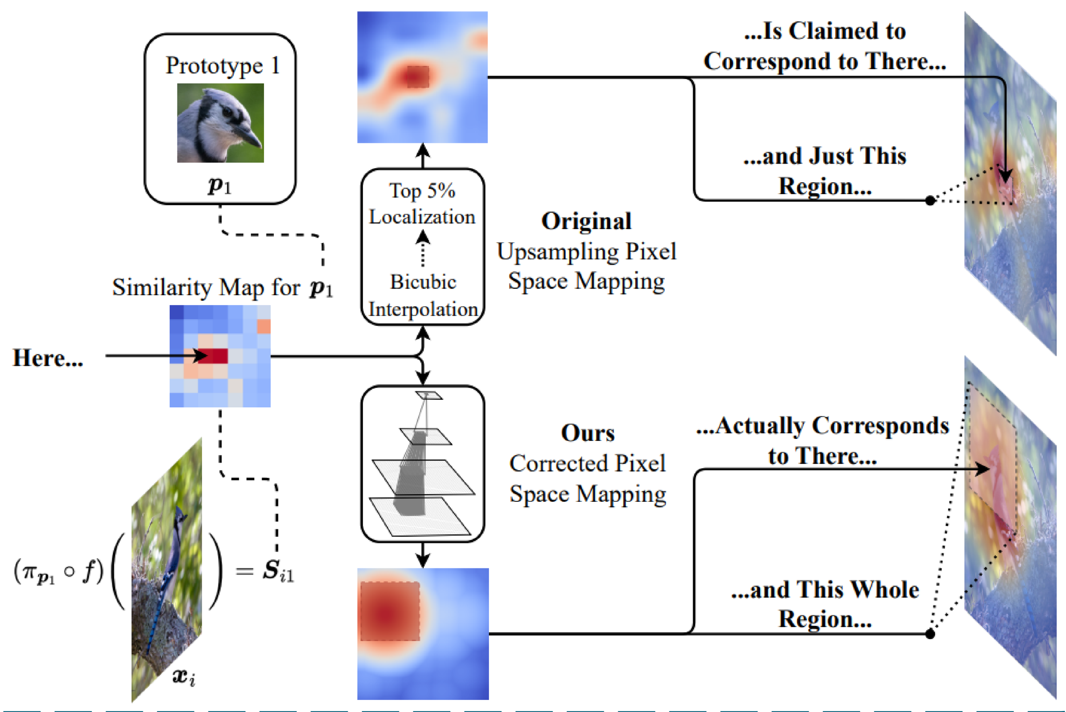


- *Inference*
  - Image embedding
  - Similarity pooling
  - Linear combination of scores
- *Prototype Training*
  - Learnable prototype vectors
  - Project closest training patches onto prototype vectors
- *Prototype Visualization & Localization*



Glaring Problem

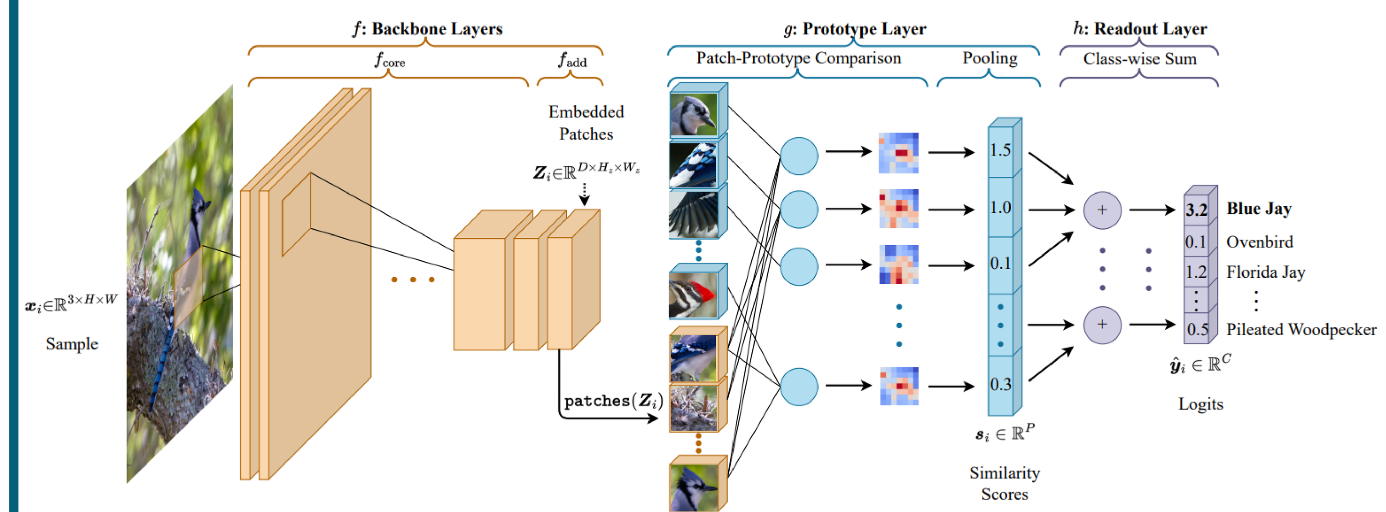
- Pixel Space Mapping is ill-formed
- Is it fair to say just 5% of the input contributed to the similarity score?
- Is it fair to say positions in latent feature maps correspond to parts of the input?



Solution

PixPNet: Pixel-Grounded Prototype Network

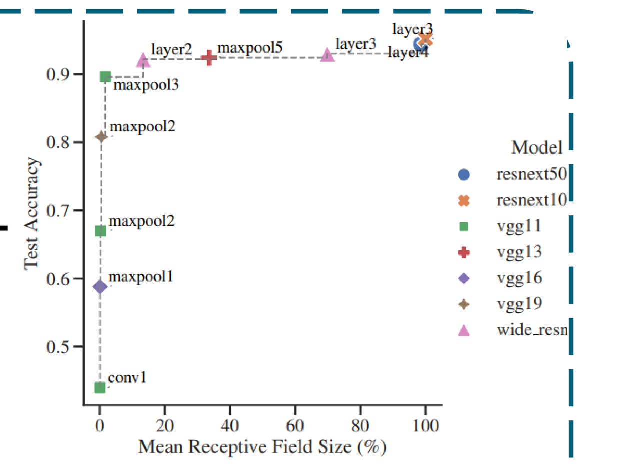
- Guarantee faithful part localization by design
- Key idea: constrain backbone receptive field (accuracy-localization precision trade-off)

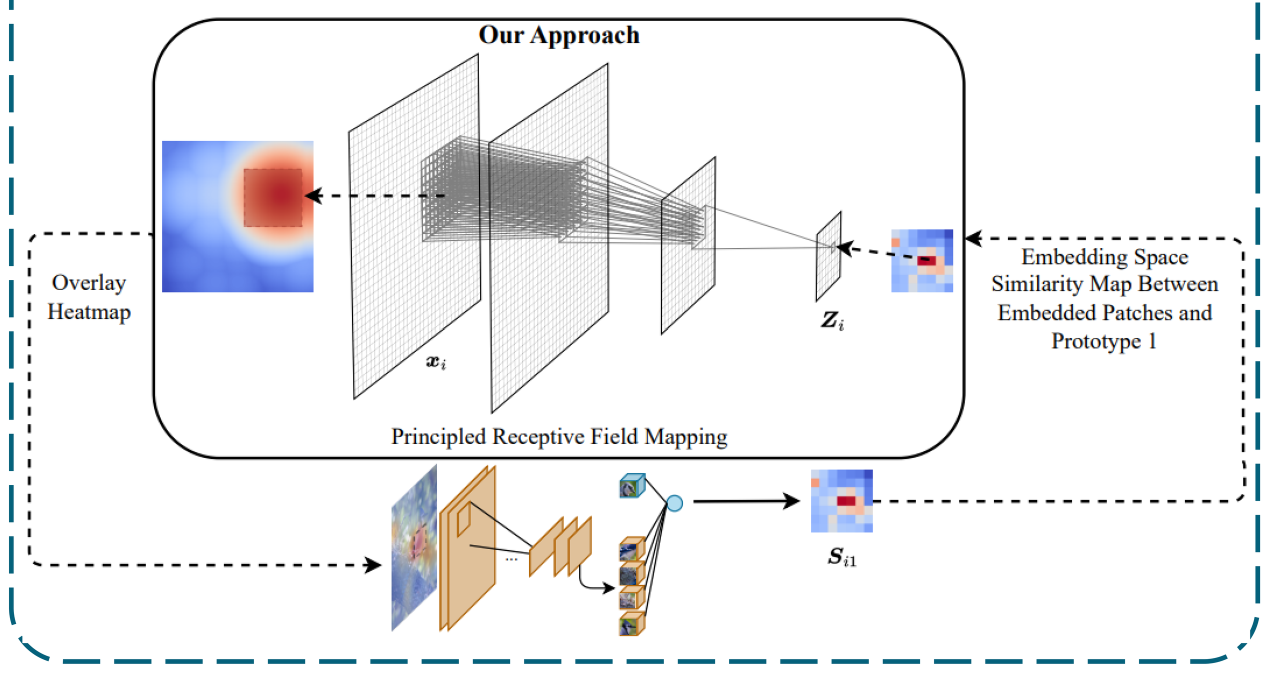


Proposed Pixel Space Mapping

- Select input corresponding to receptive field
- Assign Value:

```
Max(Current, Gaussian(
  Similarity))
```





Results

- Outperforms ProtoPNet accuracy on CUB-200-2011 and Stanford Cars datasets
- Does not rely on bounding box annotations
- *Interpretability Metrics*:
- Semantic consistency
- Semantic Stability
- Relevance Ordering Test

Backbone	MRF	Acc. ↑	PSM	S_{con} ↑	S_{sta} ↑	AUSC ↑	%2R ↓
VGG11 @maxpool4	8.31	72.9	Ours	**65.3**	**48.3**	**0.99**	**11.2**
			Orig.	45.8	44.0	0.90	30.5
VGG13 @maxpool4	9.69	75.3	Ours	**66.9**	**45.0**	**0.97**	**13.0**
			Orig.	48.1	41.8	0.88	84.1
VGG16 @maxpool4	15.7	76.4	Ours	**62.0**	**46.4**	**1.02**	**6.98**
			Orig.	46.8	42.2	0.89	35.5
VGG19 @maxpool4	22.8	77.1	Ours	**60.1**	**42.5**	**0.94**	**21.4**
			Orig.	48.4	41.3	0.80	99.9
VGG13 @maxpool5	33.5	78.1	Ours	**67.0**	**42.5**	**0.90**	**29.5**
			Orig.	43.7	39.9	0.81	99.2
VGG16 @maxpool5	52.5	79.8	Ours	**69.5**	**51.6**	**0.90**	**32.0**
			Orig.	44.1	42.4	0.82	55.5
WRN50 @layer3	69.9	80.1	Ours	**56.4**	**64.7**	**0.93**	**13.0**
			Orig.	56.4	47.6	0.85	39.6
VGG19 @maxpool5	70.4	80.1	Ours	**47.6**	**64.2**	**0.92**	**43.4**
			Orig.	45.8	46.0	0.85	92.9
ResNet18 @layer2	15.4	57.2	Ours	**59.2**	**46.6**	**0.98**	**4.10**
			Orig.	25.2	45.6	0.88	96.8
			PRP	–	–	0.95	25.4
ResNet50 @layer3	69.8	76.6	Ours	47.9	**62.0**	**0.58**	**72.8**
			Orig.	**53.5**	42.7	0.42	97.8
			PRP	–	–	0.34	100.0

Dombrowski, Ann-Kathrin, et al. "Explanations can be manipulated and geometry is to blame." Advances in neural information processing systems 32 (2019).
Slack, Dylan, et al. "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. 2020.

Hedström, Anna, et al. "The Meta-Evaluation Problem in Explainable AI: Identifying Reliable Estimators with MetaQuantus." arXiv preprint arXiv:2302.07265 (2023).

Krishna, Satyapriya, et al. "The disagreement problem in explainable machine learning: A practitioner's perspective." arXiv preprint arXiv:2202.01602 (2022).

Li, Oscar, et al. "Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions." *Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 32. No. 1. 2018.*

Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." *Advances in neural information processing systems 32 (2019).*