Improving Prototypical Visual Explanations via Reward Reweighing, Reselection, and Retraining

Aaron J. Li[1], Robin Netzorg[2], Zhihan Cheng[2], Zhuoqin Zhang[2], Bin Yu[2]

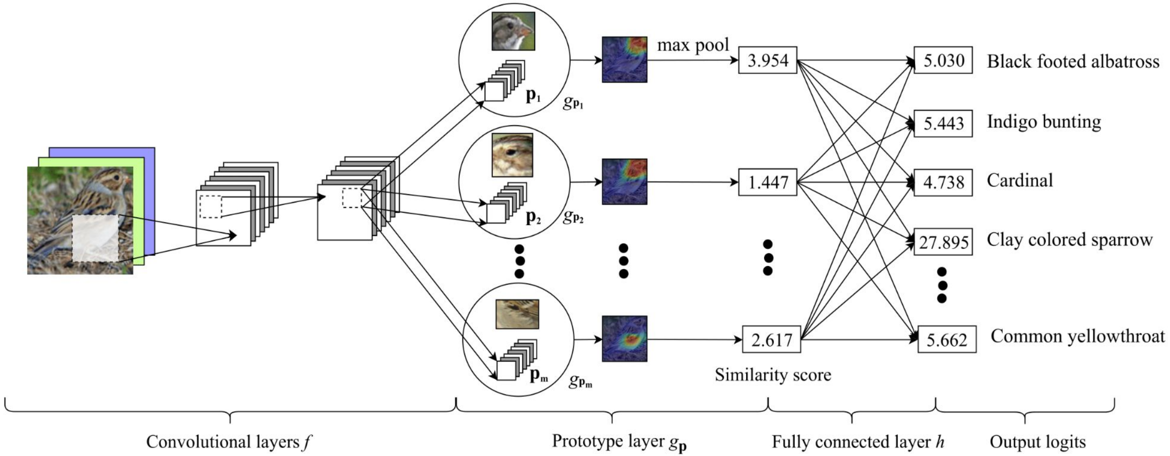
[1]Harvard University, [2]University of California, Berkeley

ICML
International Conference
On Machine Learning

Background & Motivation

Prototypical Part Network[1] (ProtoPNet):

- A prototype layer on top of a CNN base architecture
- Maintains an intuitive reasoning structure by enforcing each prototype to be similar to a particular training image patch



A key limitation of prototype based neural networks:

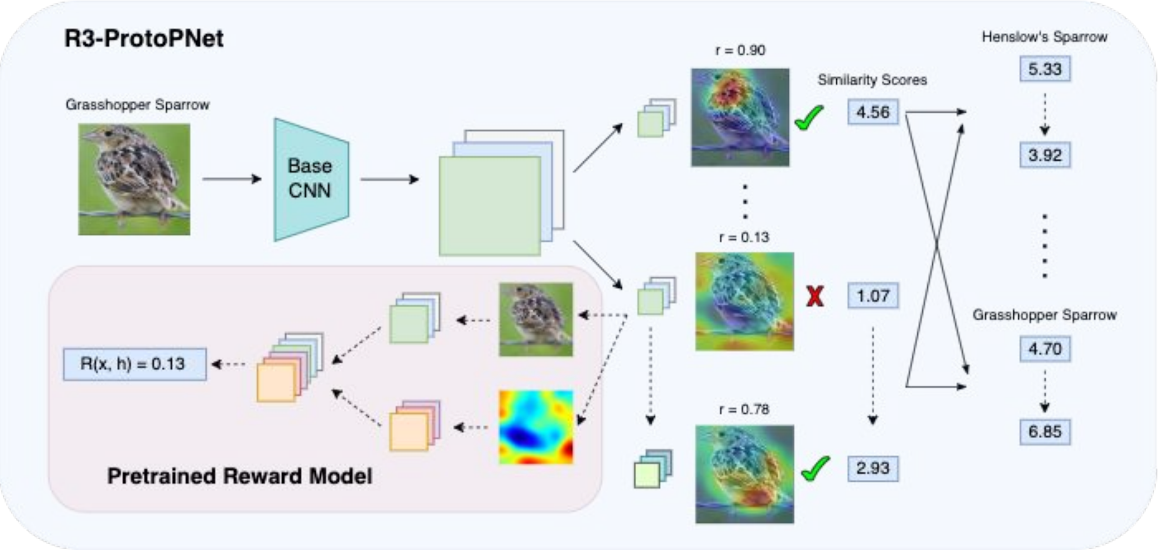
- Learned prototypes are counter-intuitive and not semantically meaningful
- E.g. Background image patches are highly activated; multiple body parts are highlighted by a single prototype

[1] Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C., and Su, J. K. This looks like that: deep learning for interpretable image recognition. Advances in neural information processing systems, 32, 2019.

Reward-Reweighing, Reselecting, and Retraining (R3) Debugging Framework

An offline and efficient concept-level debugging[2, 3] framework with the following steps:

- Train a reward model to predict human user preferences given a prototype visualization
- Use the learned reward model to update the prototypes for the pretrained ProtoPNet
- Retrain the ProtoPNet to align the rest of model with the updated prototypes



Contributions:

- Shows the effectiveness of using learned reward model as a quantified metric of prototypical visual explanation quality and model interpretability
- The proposed R3 framework empirically improves both the prototype meaningfulness and model predictive performance

[2] Bontempelli, A., Teso, S., Tentori, K., Giunchiglia, F., and Passerini, A. Concept-level debugging of part-prototype networks, 2023
[3] Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., and Rudin, C. Iaia-bl: A case-based interpretable deep learning model for classification of mass lesions in digital mammography, 2021

Human Data Collection & Reward Modeling

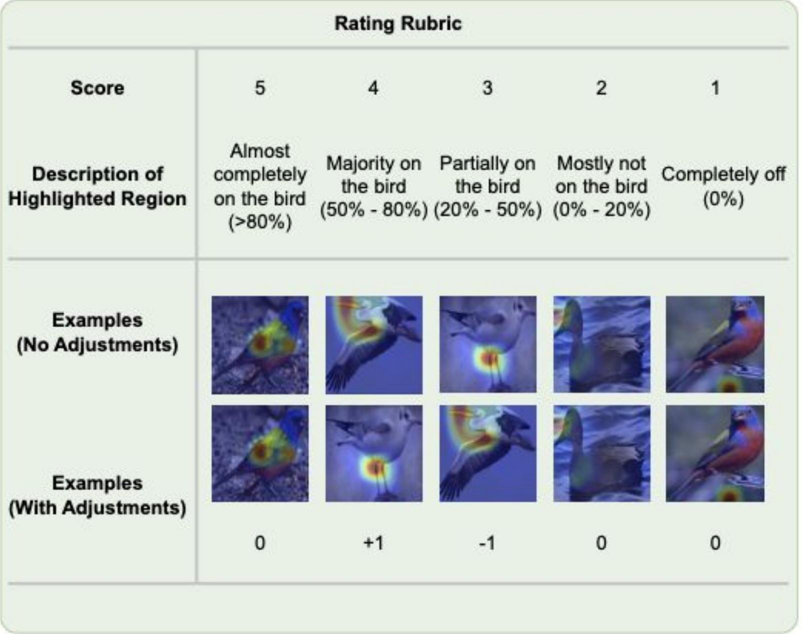
ICML
International Conference On Machine Learning

- Human preference model:

$r(x_i, h_{ij}) \in [0,1]$, where h_{ij} is the activation pattern of p_j on x_i such that $Class(p_j) = y_i$

- Data Collection:
  - ~500 (x, h, r) tuples are sampled from a pretrained ProtoPNet
  - Collected labels r are on a discrete rating scale of 1 to 5, given by human raters
  - Generate a comparison dataset by pairing each collected sample with one another[4, 5]
- The reward model is trained with the Bradley-Terry Model[6], using the paired human preference dataset

$$\mathcal{L}_{\text{reward}} = -\sum_{i \neq i' \text{ or } j \neq j'} \left[\mathbf{1}_{c_{iji'j'}=-1} \log \left(\frac{\exp(r(x_i, h_{ij}))}{\exp(r(x_i, h_{ij})) + \exp(r(x_{i'}, h_{i'j'}))} \right) + \mathbf{1}_{c_{iji'j'}=1} \log \left(\frac{\exp(r(x_{i'}, h_{i'j'}))}{\exp(r(x_i, h_{ij})) + \exp(r(x_{i'}, h_{i'j'}))} \right) \right]$$



[4] Bontempelli, A., Teso, S., Tentori, K., Giunchiglia, F., and Passerini, A. Concept-level debugging of part-prototype networks, 2023
[5] Barnett, A. J., Schwartz, F. R., Tao, C., Chen, C., Ren, Y., Lo, J. Y., and Rudin, C. Iaia-bl: A case-based interpretable deep learning model for classification of mass lesions in digital mammography, 2021
[6] Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. Biometrika, 39(3/4):324–345, 1952

R3 Debugging Steps

ICML
International Conference On Machine Learning

- Reward Reweighing:
  - Used to locally “move” the focus of a prototype according to human preference

$$\max_{p_j} \mathcal{L}_{reweigh}(z_i^*, p_j) = \max_{p_j} \sum_{i \in I(p_j)}^{n} \frac{r(x_i, p_j)}{\lambda_{dist} \|z_i^* - p_j\|_2^2 + 1} \quad (1)$$

where $z_i^* = \text{argmin}_{z \in \text{patches}(f(x_i))} \|z - p_j\|_2^2$,

- Prototype Reselection:
  - Used to completely discard the original prototype and reselect a new one (e.g. when the old prototype completely focuses on the background)
  - For each suboptimal prototype, the choice between reward reweighing and reselection is based on its predicted reward value (a reward threshold is empirically determined)
- Retraining
  - Same as the original ProtoPNet training
  - used to align the rest of the model with the updated prototypes

Algorithm 1 Reward Reweighed, Reselected, and Retrained Prototypical Part Network (R3-ProtoPNet)

1: **Initialize:** Collect high-quality human feedback data and train a reward model.
2: **Reward Reweighing:** Perform the reward-reweighed update for the ProtoPNet, defined in Equation 1. Optimize the loss function, which leads to locally maximal solutions, improving the prototypes.
3: **Prototype Reselection:** Run the reselection procedure based on a reward threshold.
If $\frac{1}{n_k} \sum_{i \in I(p_j)} r(x_i, p_j) < \alpha$, reselect the prototype by sampling from patch candidates and temporarily setting the prototype to a new candidate that passes the acceptance threshold and is unique from other current prototypes.
4: **Retraining:** Retrain the model with the same loss function used in the original ProtoPNet update, to realign the prototypes and the rest of the model.

Experiment Results (CUB-200-2011)

ICML
International Conference
On Machine Learning

Table 1. Predictive Accuracy

BASE (m_k)	PROTOPNET	R2-PROTOPNET	R3-PROTOPNET
VGG-19 (5)	76.33 ± 0.12	62.76 ± 1.18	$\mathbf{77.80} \pm 0.18$
VGG-19 (10)	77.58 ± 0.22	50.41 ± 1.36	$\mathbf{79.60} \pm 0.25$
RESNET-34 (10)	78.73 ± 0.13	58.11 ± 2.71	$\mathbf{80.21} \pm 0.22$
RESNET-50 (10)	78.52 ± 0.17	56.36 ± 2.40	$\mathbf{80.25} \pm 0.22$
DENSENET-121 (10)	79.64 ± 0.23	54.67 ± 2.29	$\mathbf{80.42} \pm 0.26$
DENSENET161 (10)	$\mathbf{79.75} \pm 0.27$	62.75 ± 2.43	79.48 ± 0.36
ENSEMBLE OF ABOVE	82.92 ± 0.09	70.46 ± 0.82	$\mathbf{84.37} \pm 0.20$

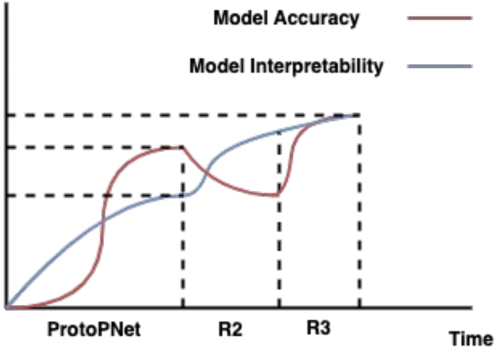
Table 3. Average Activation Precision (AP)

BASE (m_k)	PROTOPNET	RESELECTED	REWEIGHED	R3-PROTOPNET
VGG19 (5)	70.31	79.81	85.64	86.61
VGG19 (10)	63.12	75.95	82.72	81.62
RESNET-34 (10)	85.63	88.81	90.33	92.23
RESNET-50 (10)	71.45	79.29	83.69	83.52
DENSENET-121 (10)	66.22	81.64	86.73	89.38
DENSENET-161 (10)	82.56	85.24	87.55	87.60
AVERAGE	73.22	81.79	86.11	86.83

* The AP metric has been used in Bontempelli et. al., 2023 [2] and Barnett et. al., 2021 [3]

Table 2. Estimated reward values

BASE (m_k)	PROTOPNET	RESELECTED	REWEIGHED	R3-PROTOPNET
VGG19 (5)	0.61	0.66	0.70	0.71
VGG19 (10)	0.46	0.55	0.64	0.67
RESNET-34 (10)	0.40	0.47	0.51	0.54
RESNET-50 (10)	0.36	0.45	0.50	0.54
DENSENET-121 (10)	0.48	0.53	0.58	0.58
DENSENET-161 (10)	0.48	0.51	0.57	0.56
AVERAGE	0.47	0.53	0.58	0.60



Trade-off Between Accuracy and Interpretability (qualitative)

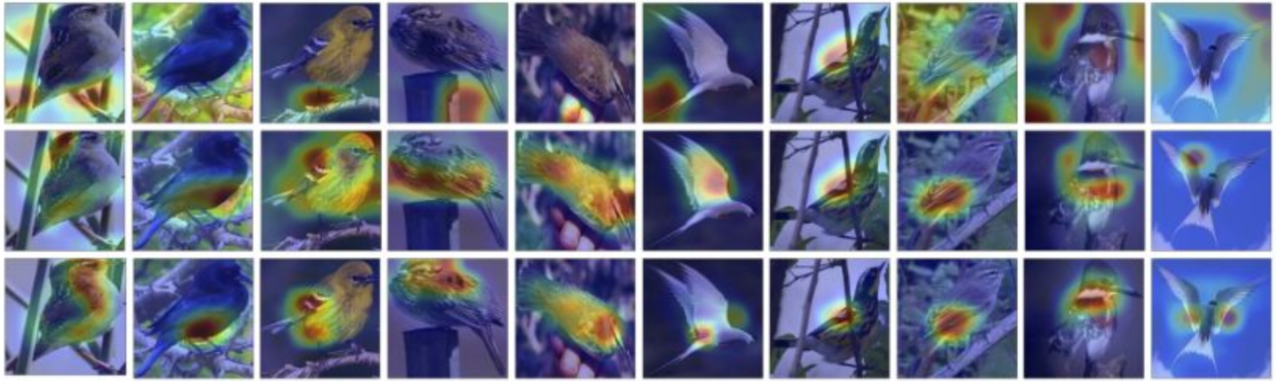
ICML
International Conference
On Machine Learning

Visualized Examples

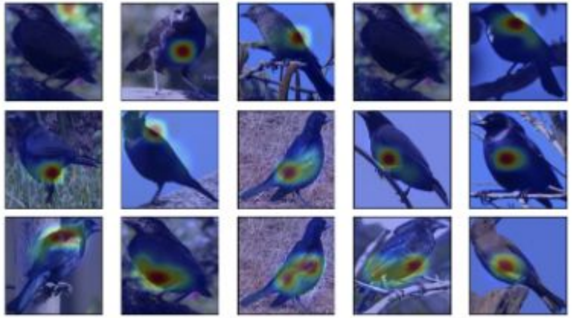
ProtoPNet

After Reweighing and Reselection
(R2-ProtoPNet)

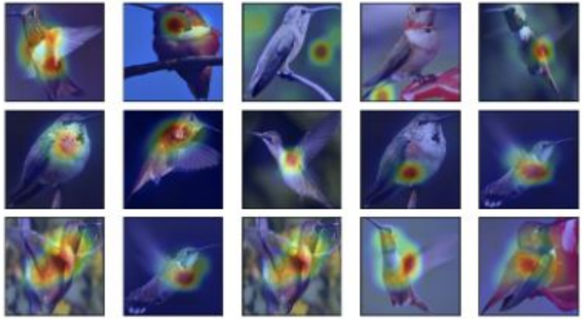
After Retraining (R3-ProtoPNet)



P1 P2 P3 P4 P5



P1 P2 P3 P4 P5



ProtoPNet

R2-ProtoPNet

R3-ProtoPNet