

This Looks Like That: Deep Learning for Interpretable Image Recognition

Chaofan Chen^{*†}, Oscar Li^{*†}, Chaofan Tao[†],
Alina Jade Barnett[†], Jonathan Su[^], Cynthia Rudin[†]

* Contributed equally

† Duke University

^ MIT Lincoln Laboratory

The authors (Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin) reserve the right to make reproductions for use in teaching, scholarship, and research, the right to borrow portions of the work for use in other works, the right to make derivative works, the right to alter the work, add to the work, or update the content of the work, the right to be identified as the authors of the work, the right to perform or display the work, the right to include all or part of this material in the thesis or dissertation, the right to make oral presentation of the material in any forum, the right to archive and preserve the work as part of either a personal or institutional initiative (including, but not limited to, archiving the work on personal websites or institutional repositories), and the copyright in every draft and pre-print version of the work.

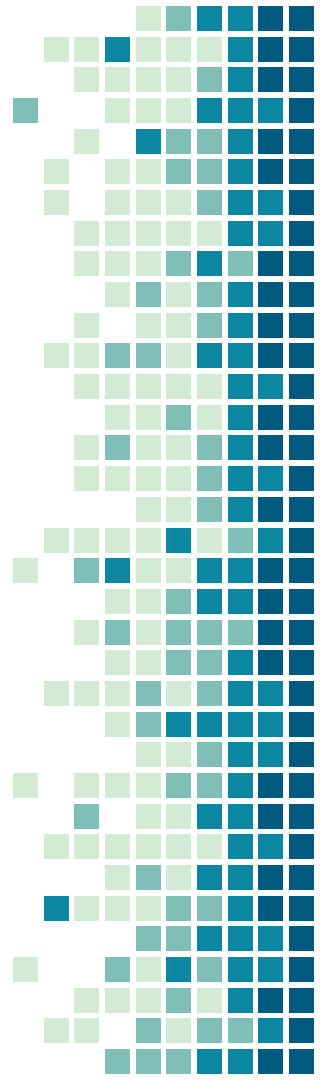
DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

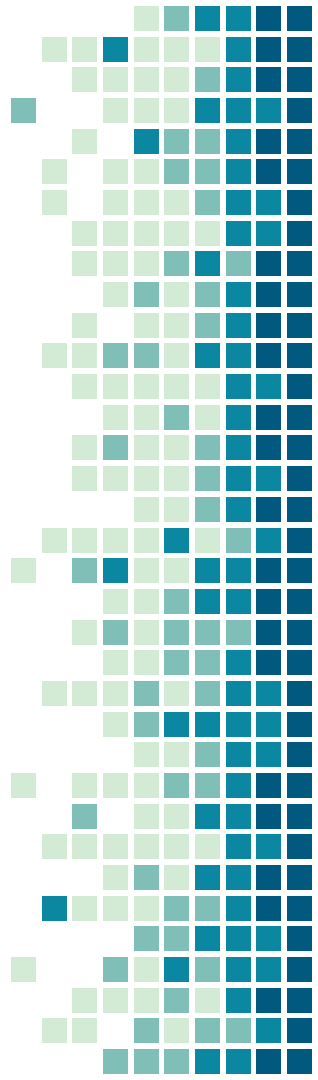
© 2019 Massachusetts Institute of Technology.

Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.

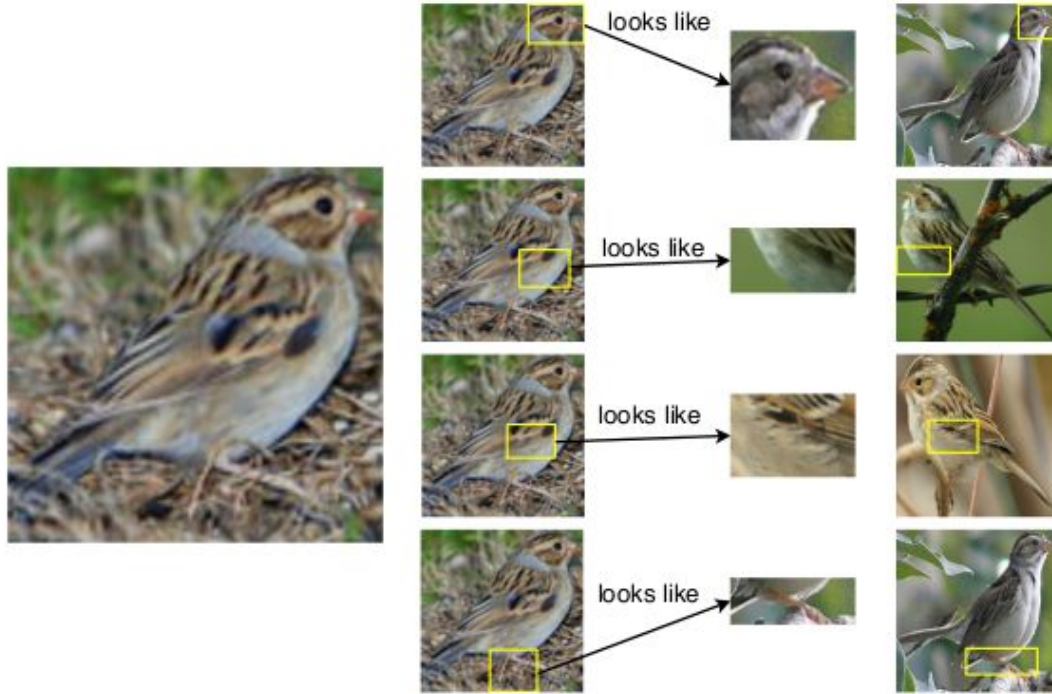
A new form of interpretability...



A new form of interpretability...



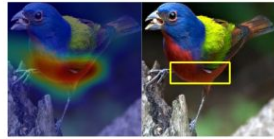
A new form of interpretability...



...with richer explanations

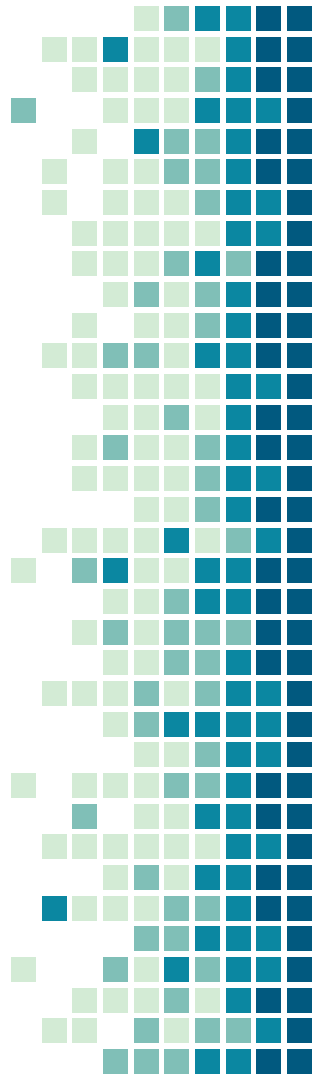


(a) Object attention
(class activation map)



(b) Part attention
(attention-based models)

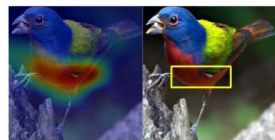
Previous methods



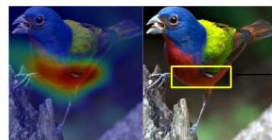
...with richer explanations



(a) Object attention
(class activation map)



(b) Part attention
(attention-based models)



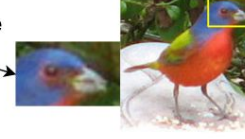
looks like



looks like

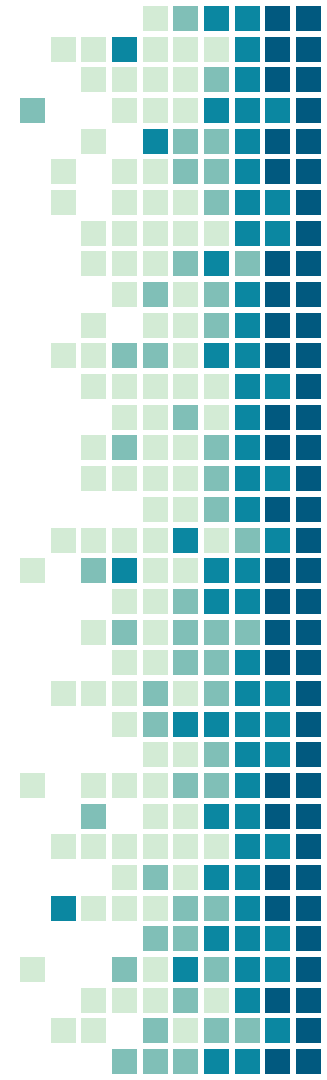


looks like

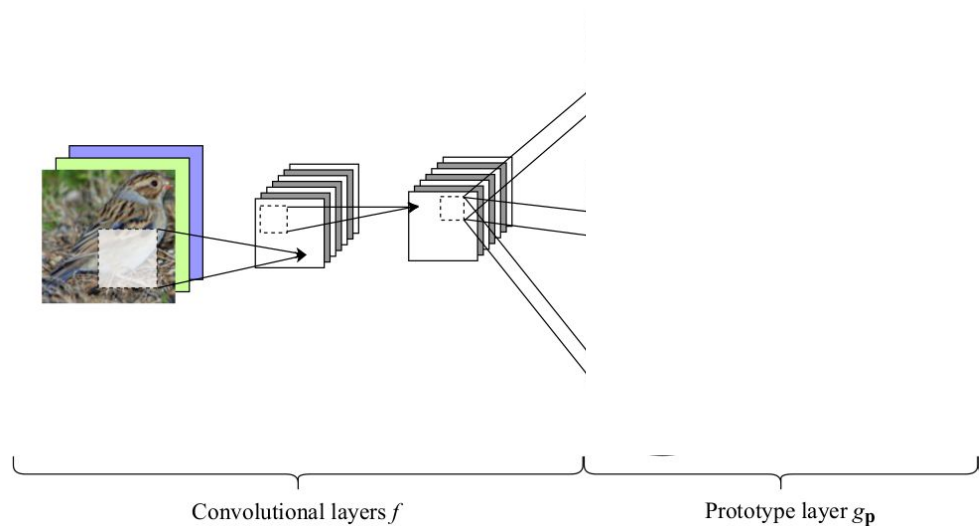


(c) Part attention + comparison with learned
prototypical parts (our model)

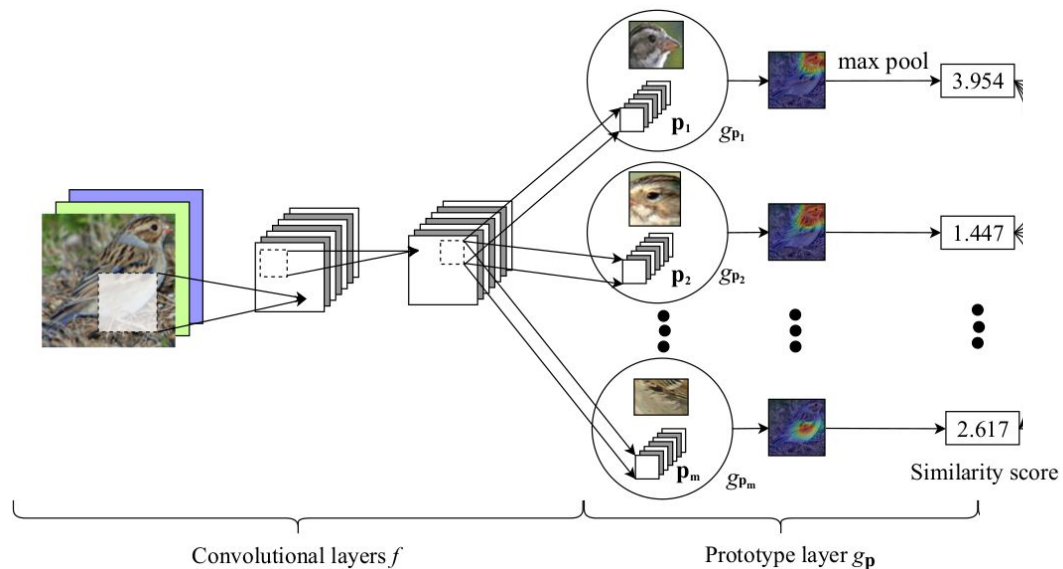
Previous methods



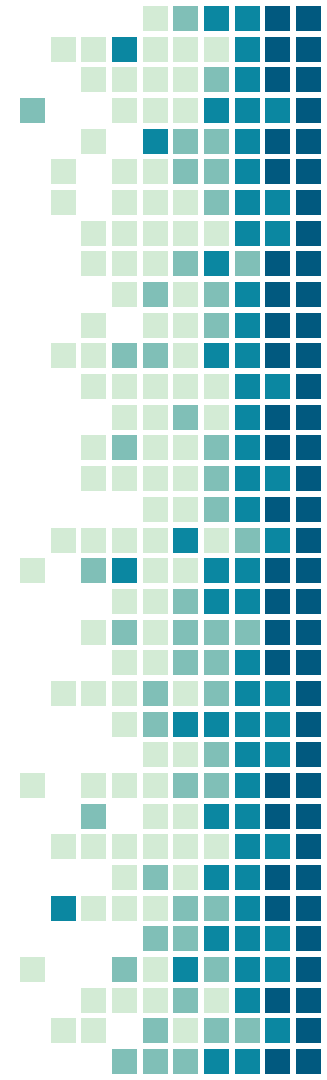
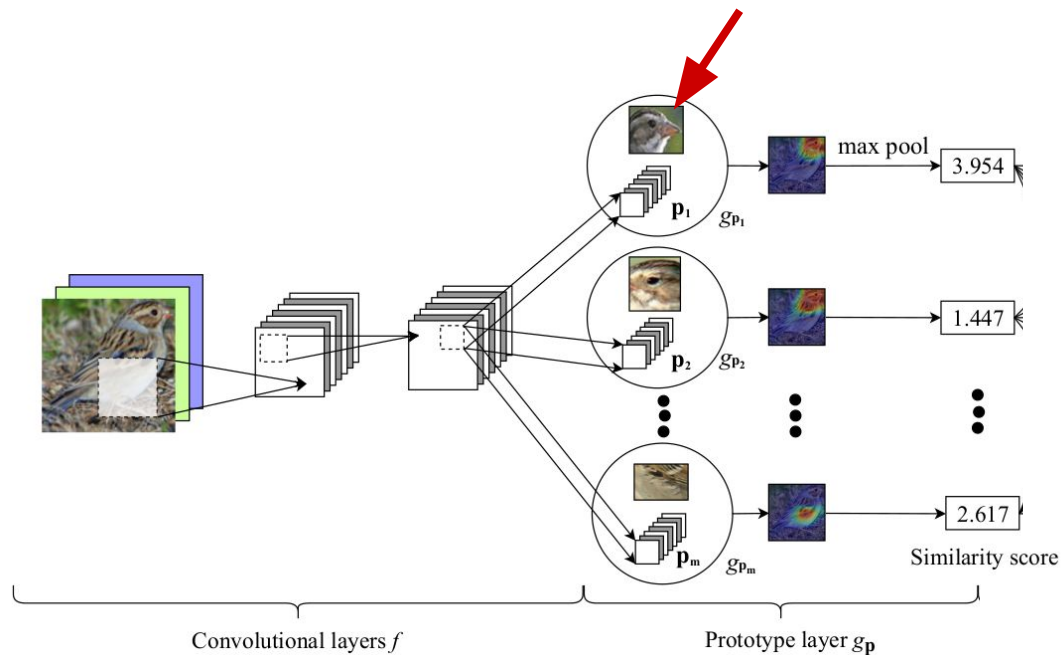
ProtoNet Architecture



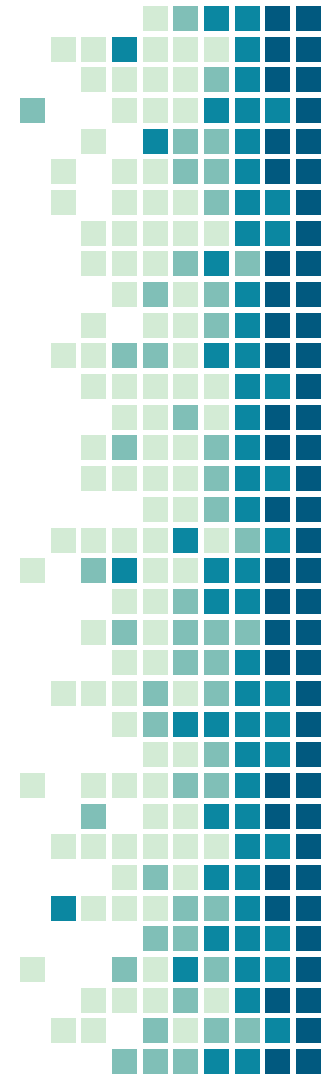
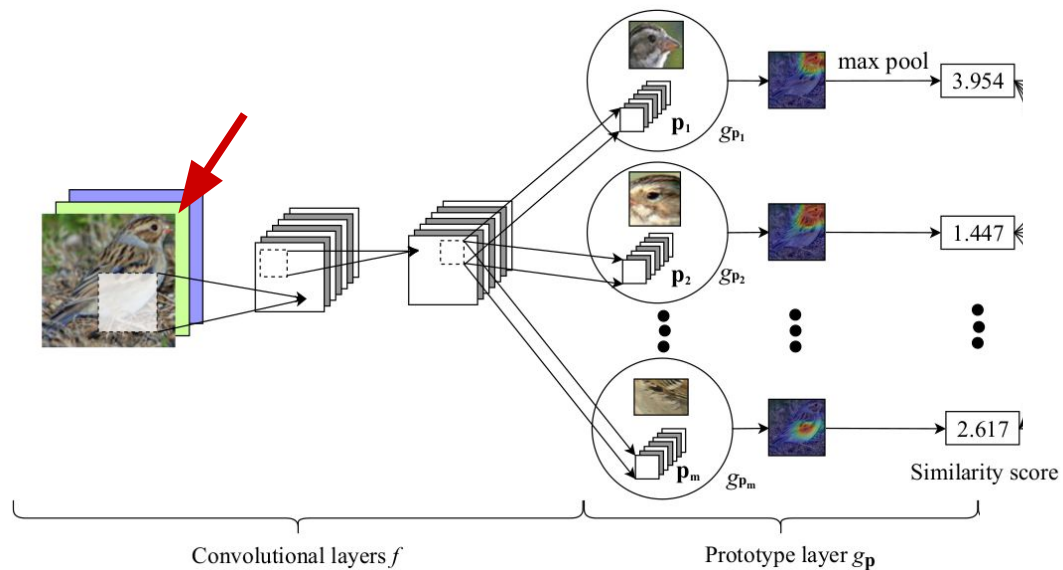
ProtoNet Architecture



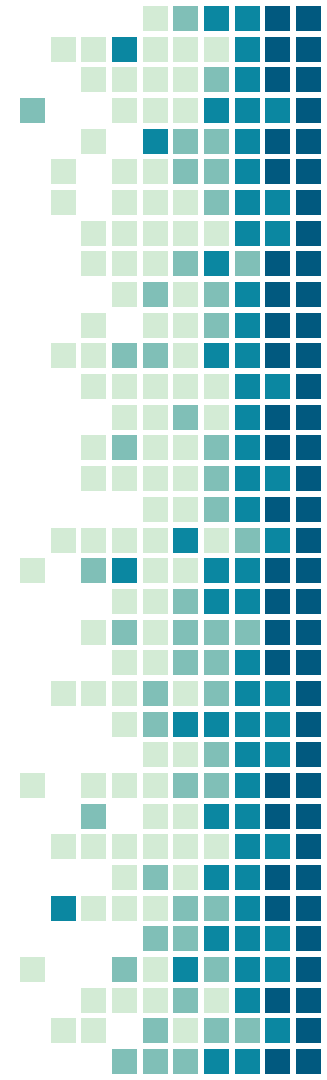
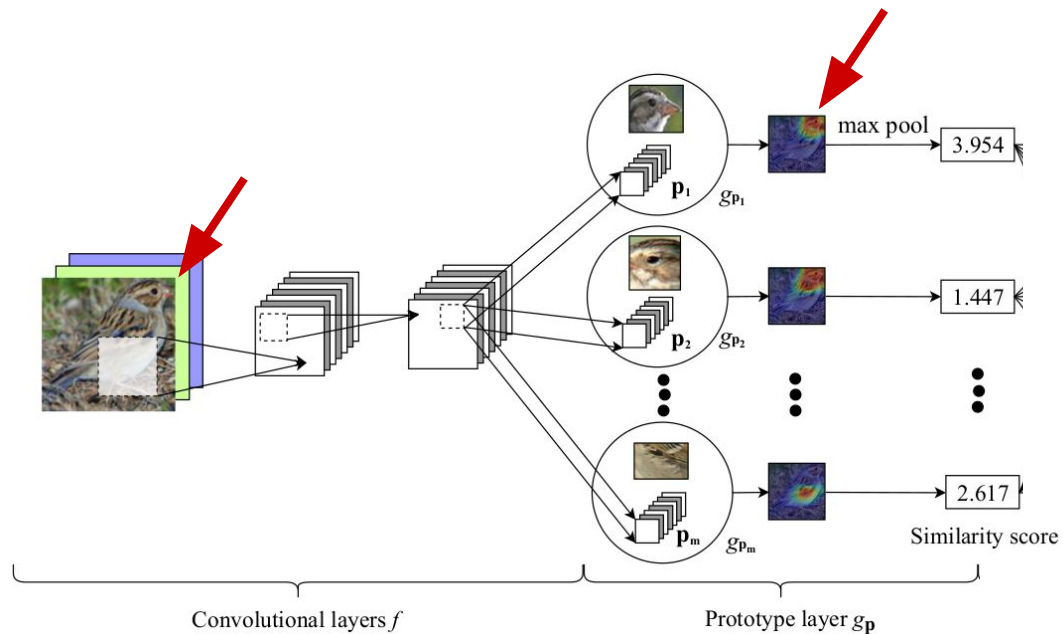
ProtoNet Architecture



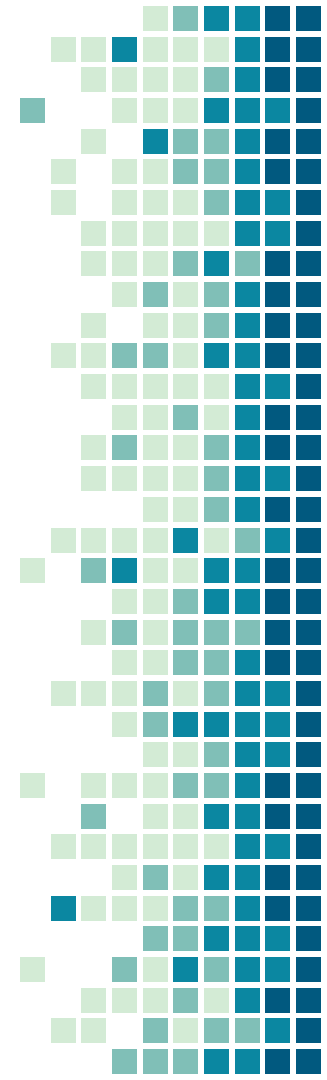
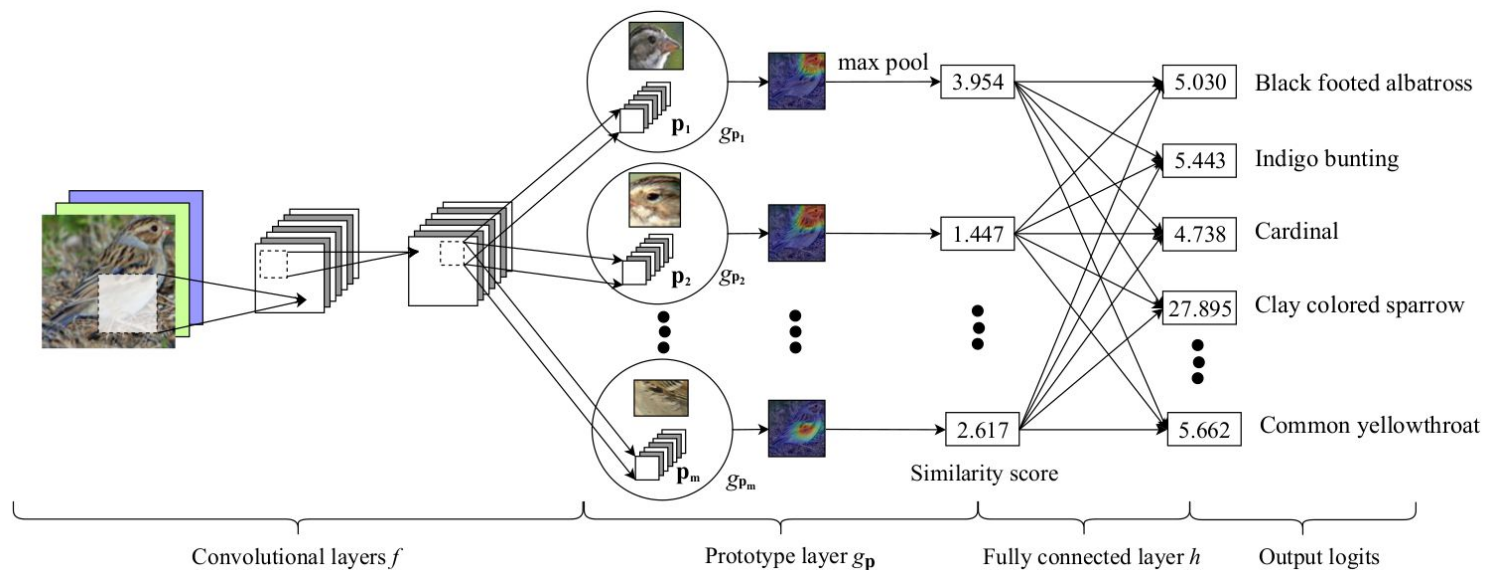
ProtoNet Architecture



ProtoNet Architecture



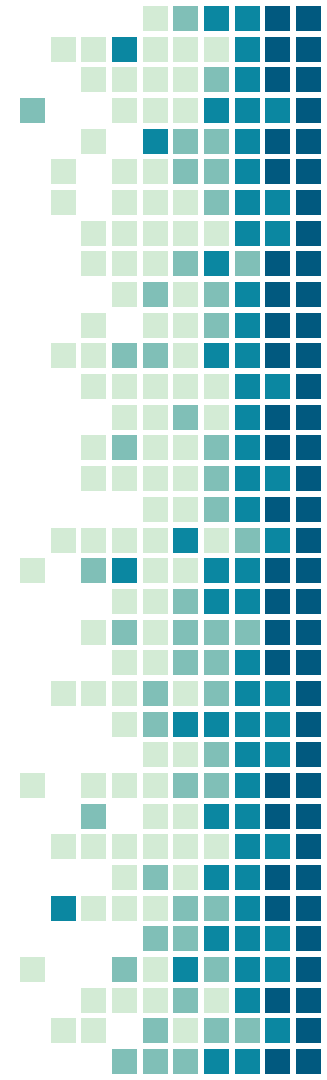
ProtoNet Architecture



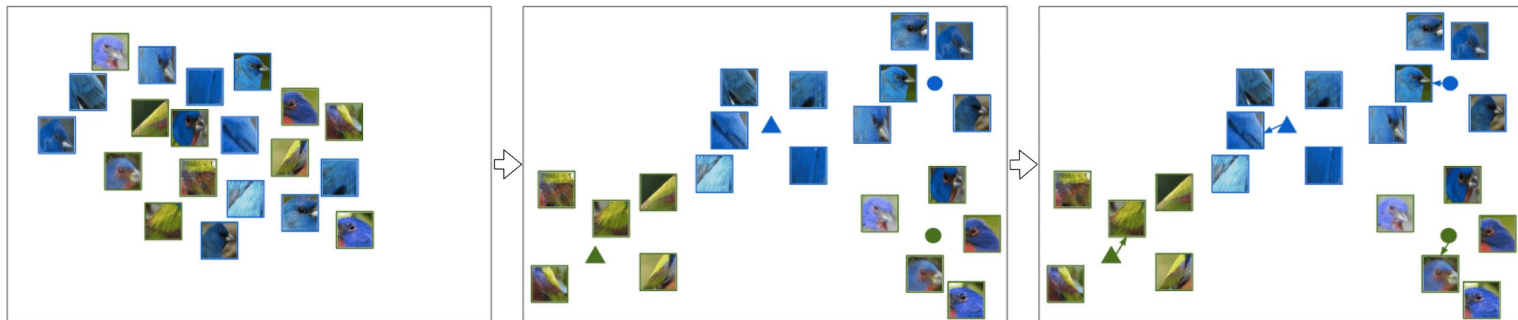
ProtoPNet as Scoring Sheets

	Original image (box showing part that looks like prototype)	Prototype	Training image where prototype comes from	Activation map	Similarity score	Class connection	Points contributed
					6.499	1.180	$= 7.669$
					4.392	1.127	$= 4.950$
					3.890	1.108	$= 4.310$
	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Total points to red-bellied woodpecker: 32.736



Training Algorithm



Stage 1: stochastic gradient descent (SGD) of layers before last layer

$$\min_{\mathbf{P}, w_{\text{conv}}} \frac{1}{n} \sum_{i=1}^n \text{CrsEnt}(h \circ g_{\mathbf{p}} \circ f(\mathbf{x}_i), \mathbf{y}_i) + \lambda_1 \text{Clst} + \lambda_2 \text{Sep}, \quad \text{where}$$

$$\text{Clst} = \frac{1}{n} \sum_{i=1}^n \min_{j: \mathbf{p}_j \in \mathbf{P}_{y_i}} \min_{\mathbf{z} \in \text{patches}(f(\mathbf{x}_i))} \|\mathbf{z} - \mathbf{p}_j\|_2^2, \text{Sep} = -\frac{1}{n} \sum_{i=1}^n \min_{j: \mathbf{p}_j \notin \mathbf{P}_{y_i}} \min_{\mathbf{z} \in \text{patches}(f(\mathbf{x}_i))} \|\mathbf{z} - \mathbf{p}_j\|_2^2.$$

Stage 2: projection of prototypes

$$\mathbf{p}_j \leftarrow \arg \min_{\mathbf{z} \in \mathcal{Z}_j} \|\mathbf{z} - \mathbf{p}_j\|_2, \text{ where } \mathcal{Z}_j = \{\tilde{\mathbf{z}} : \tilde{\mathbf{z}} \in \text{patches}(f(\mathbf{x}_i)) \forall i \text{ s.t. } y_i = k\}.$$

Stage 3: Convex optimization of last layer

$$\min_{w_h} \frac{1}{n} \sum_{i=1}^n \text{CrsEnt}(h \circ g_{\mathbf{p}} \circ f(\mathbf{x}_i), \mathbf{y}_i) + \lambda \sum_{k=1}^K \sum_{j: \mathbf{p}_j \notin \mathbf{P}_k} |w_h^{(k,j)}|.$$

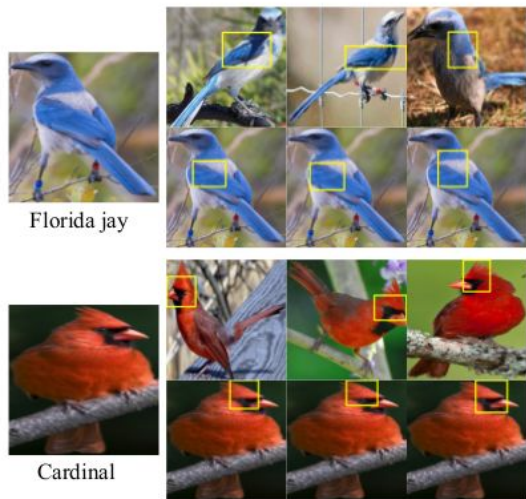
Accuracy Comparison

Base	ProtoPNet	Baseline	Base	ProtoPNet	Baseline
VGG16	76.1 \pm 0.2	74.6 \pm 0.2	VGG19	78.0 \pm 0.2	75.1 \pm 0.4
Res34	79.2 \pm 0.1	82.3 \pm 0.3	Res152	78.0 \pm 0.3	81.5 \pm 0.4
Dense121	80.2 \pm 0.2	80.5 \pm 0.1	Dense161	80.1 \pm 0.3	82.2 \pm 0.2

Interpretability	Model: accuracy
None	B-CNN : 85.1 (bb), 84.1 (full)
Object-level attn.	CAM : 70.5 (bb), 63.0 (full)
Part-level attention	Part R-CNN : 76.4 (bb+anno.); PS-CNN : 76.2 (bb+anno.); PN-CNN : 85.4 (bb+anno.); DeepLAC : 80.3 (anno.); SPDA-CNN : 85.1 (bb+anno.); PA-CNN : 82.8 (bb); MG-CNN : 83.0 (bb), 81.7 (full); ST-CNN : 84.1 (full); 2-level attn. : 77.9 (full); FCAN : 82.0 (full); Neural const. : 81.0 (full); MA-CNN : 86.5 (full); RA-CNN : 85.3 (full)
Part-level attn. + prototypical cases	ProtoPNet (ours): 80.8 (full, VGG19+Dense121+Dense161-based) 84.8 (bb, VGG19+ResNet34+DenseNet121-based)



Analysis of Latent Space



(a) nearest prototypes of two test images

left: original test image

right: top: three nearest prototypes of the image, with prototypical parts shown in box

below: test image with patch closest to each prototype shown in box

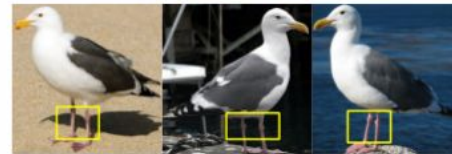
Prototype
(in bounding box)



Nearest training patches
(in bounding box)



Nearest test patches
(in bounding box)



(b) nearest image patches to prototypes

left: prototype, with prototypical parts in box

middle: nearest training images to prototype, with patch closest to prototype in box

right: nearest test images to prototype, with patch closest to prototype in box

This Looks Like That: Deep Learning for Interpretable Image Recognition

Chaofan Chen^{*†}, Oscar Li^{*†}, Chaofan Tao[†],
Alina Jade Barnett[†], Jonathan Su[^], Cynthia Rudin[†]

* Contributed equally

† Duke University

^ MIT Lincoln Laboratory

The authors (Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin) reserve the right to make reproductions for use in teaching, scholarship, and research, the right to borrow portions of the work for use in other works, the right to make derivative works, the right to alter the work, add to the work, or update the content of the work, the right to be identified as the authors of the work, the right to perform or display the work, the right to include all or part of this material in the thesis or dissertation, the right to make oral presentation of the material in any forum, the right to archive and preserve the work as part of either a personal or institutional initiative (including, but not limited to, archiving the work on personal websites or institutional repositories), and the copyright in every draft and pre-print version of the work.

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

© 2019 Massachusetts Institute of Technology.

Delivered to the U.S. Government with Unlimited Rights, as defined in DFARS Part 252.227-7013 or 7014 (Feb 2014). Notwithstanding any copyright notice, U.S. Government rights in this work are defined by DFARS 252.227-7013 or DFARS 252.227-7014 as detailed above. Use of this work other than as specifically authorized by the U.S. Government may violate any copyrights that exist in this work.