This Looks Like *That*: Deep Learning for Interpretable Image Recognition

NeurIPS 2019

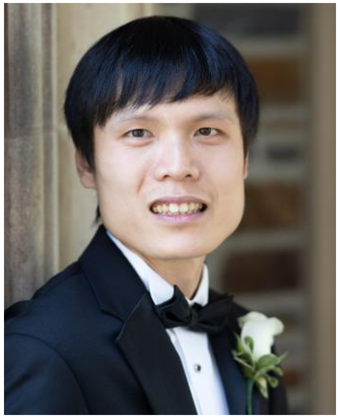
Chuanbin Liu

- 盲人摸象
  - 其触牙者即言象形如芦菔根
  - 其触鼻者言象如杵，
  - 其触腹者言象如瓮
  - ，，，
- 指出典型样
  - prototype
- This looks like that

- 作者介绍
- 研究背景
- 当前方法
- 本文方法
- 实验效果
- 总结&反思

- 作者介绍
- 研究背景
- 当前方法
- 本文方法
- 实验效果
- 总结&反思

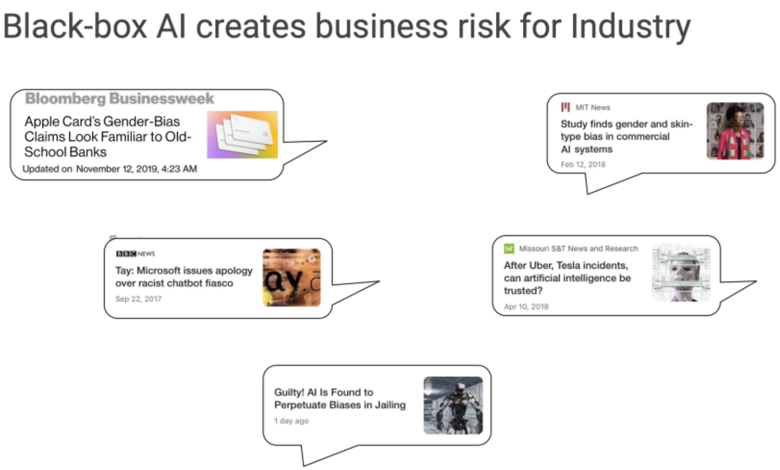
- Chaofan Chen
  - Duke University
  - interpretable machine learning
  - NeurIPS2019+NeruIPS2018 Challenge 1st

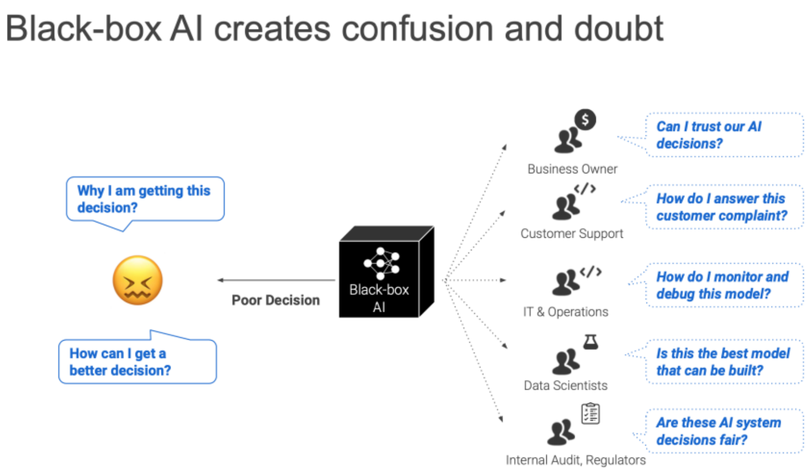


- Cynthia Rudin
  - Duke University
  - Professor, associate director SAMSI
  - interpretable machine learning
    - Stop Explaining Black Box Machine Learning for High Stakes Decisions and Use Interpretable Models Instead

- 作者介绍
- 研究背景
- 当前方法
- 本文方法
- 实验效果
- 总结&反思

- 深度学习黑盒现象
  - 对于深度学习，推理结果为相关而非因果。而且准确度越高的模型（比如深度神经网络），其推理结果越没法解释。
  - AAAI 2020 Tutorial --- Explainable AI

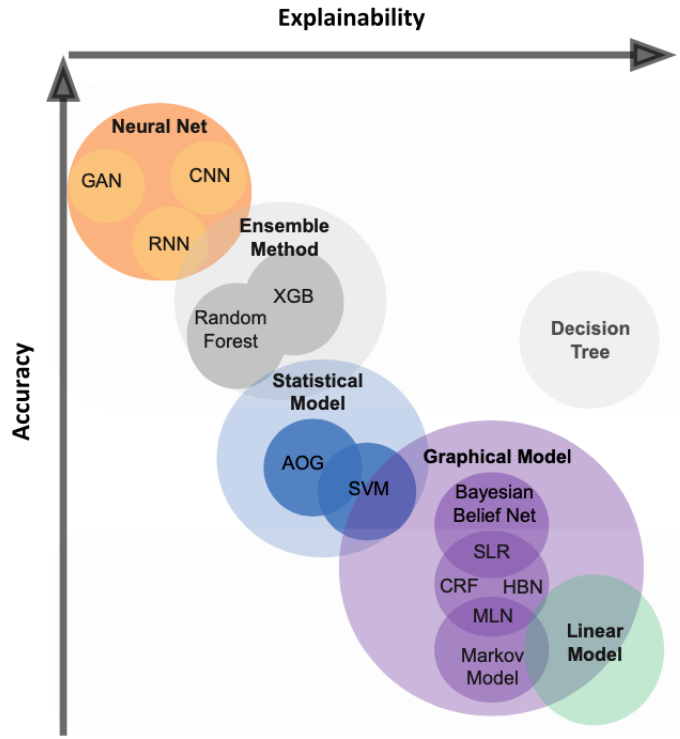




How to Explain? Accuracy vs. Explainability

Learning

- Challenges:
  - Supervised
  - Unsupervised learning
- Approach:
  - Representation Learning
  - Stochastic selection
- Output:
  - **Correlation**
  - **No causation**



Interpretability

Non-Linear functions

Polynomial functions

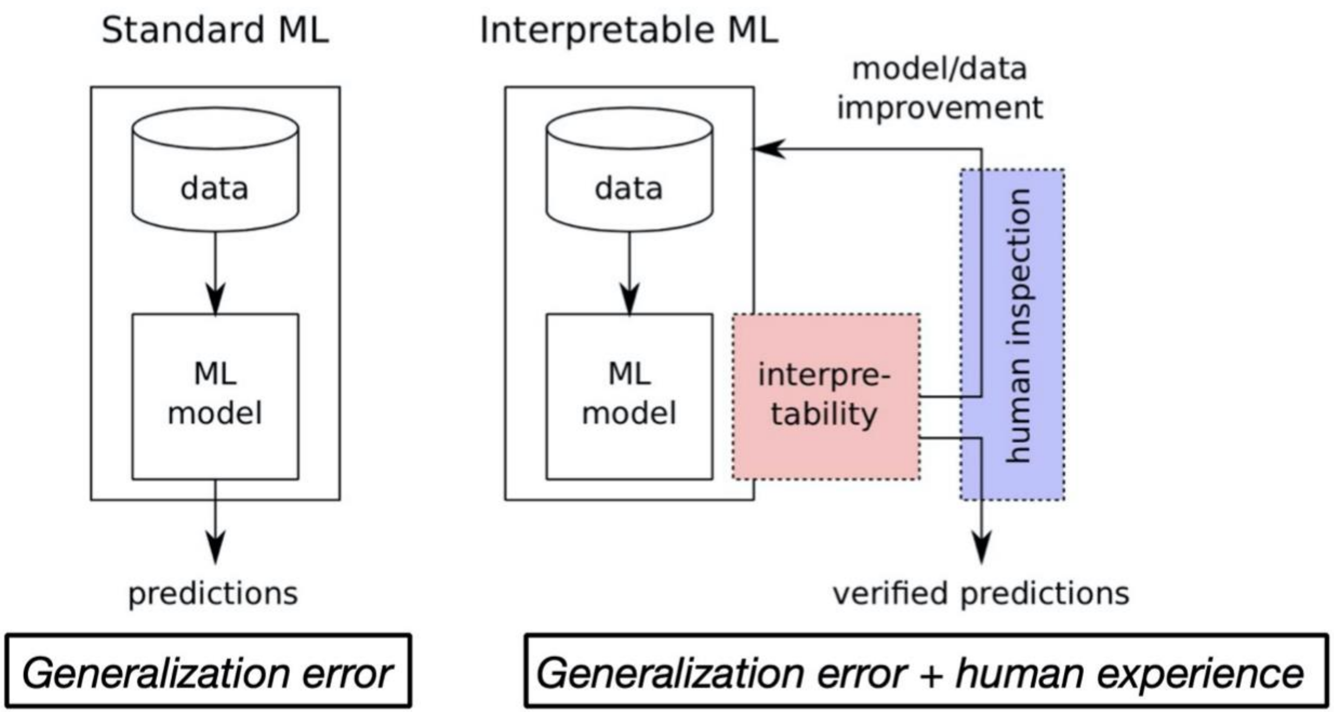
Quasi-Linear functions

Why Explainability: Debug (Mis-)Predictions

Top label: **"clog"**

Why did the network label this image as **"clog"**?

Why Explainability: Improve ML Model



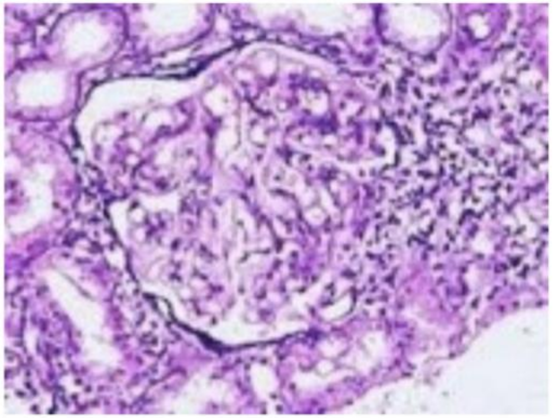
Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

Why Explainability: Verify the ML Model / System

Wrong decisions can be costly and dangerous

"Autonomous car crashes, because it wrongly recognizes …"

"AI medical diagnosis system misclassifies patient's disease …"

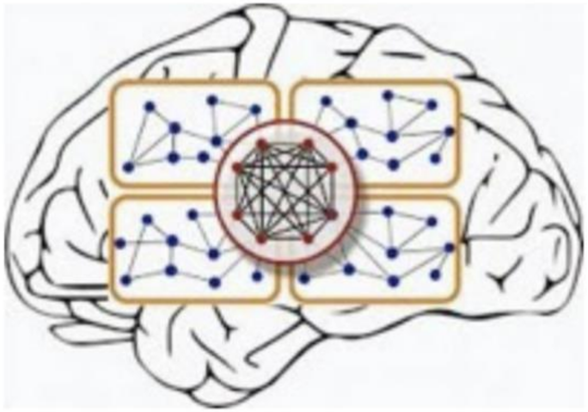


Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18

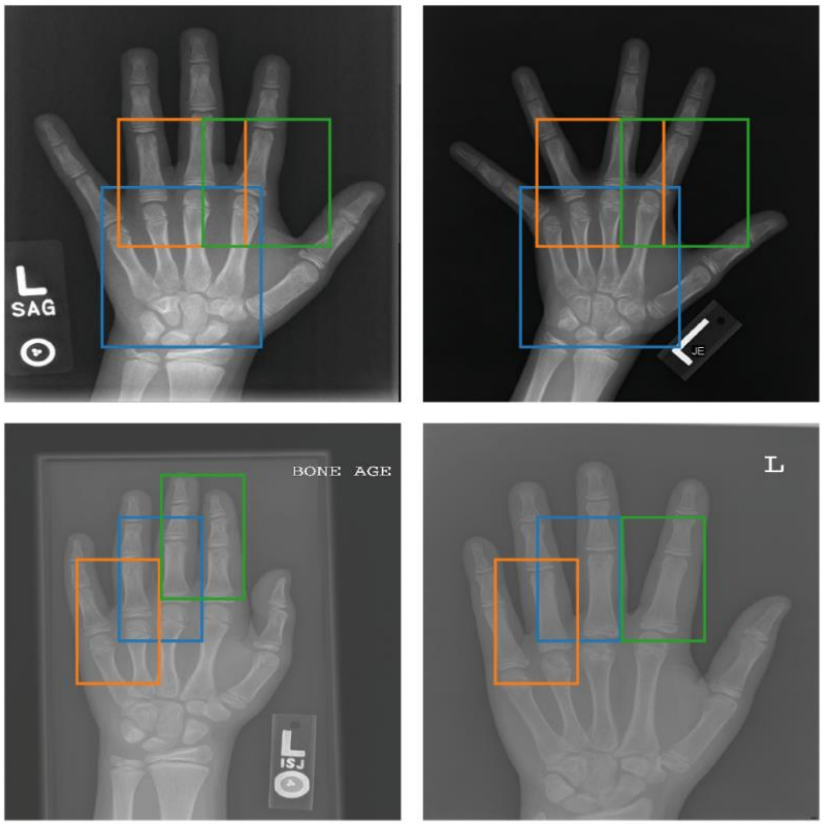
Why Explainability: Learn New Insights

"It's not a human move. I've never seen a human play this move." (Fan Hui)

Old promise:
"Learn about the human brain."



Credit: Samek, Binder, Tutorial on Interpretable ML, MICCAI'18



- The carpal and the proximal pha- langes are extracted for female
- The proximal phalanges of the index finger, middle finger and ring finger are usually separately extracted for male.

Why Explainability: Laws against Discrimination

Citizenship
Immigration Reform and Control Act

Sex
Equal Pay Act of 1963;
Civil Rights Act of 1964

Age
Age Discrimination in Employment Act of 1967

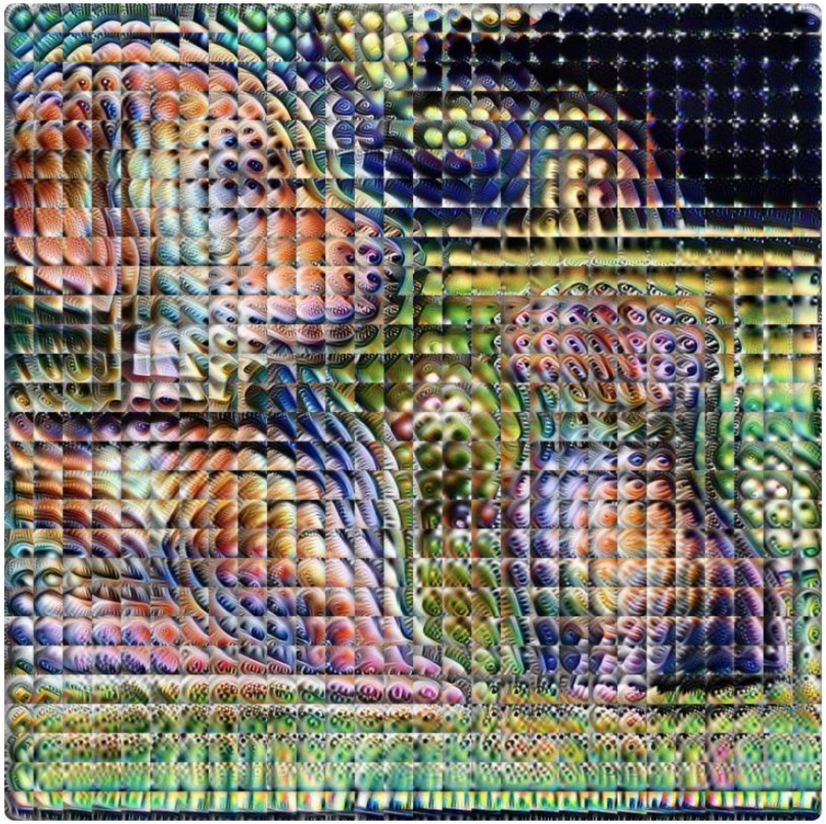
Race
Civil Rights Act of 1964

Disability status
Rehabilitation Act of 1973;
Americans with Disabilities Act of 1990

And more...

- 作者介绍
- 研究背景
- **当前方法**
- 本文方法
- 实验效果
- 总结&反思

- 隐藏层可视化
  - DeepDream
    - https://github.com/google/deepdream
  - Building_blocks
    - https://distill.pub/2018/building-blocks/



□ 语义生产

Western Grebe

Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Western Grebe* is a waterbird with a yellow pointy beak, white neck and belly, and black back.
Explanation: This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

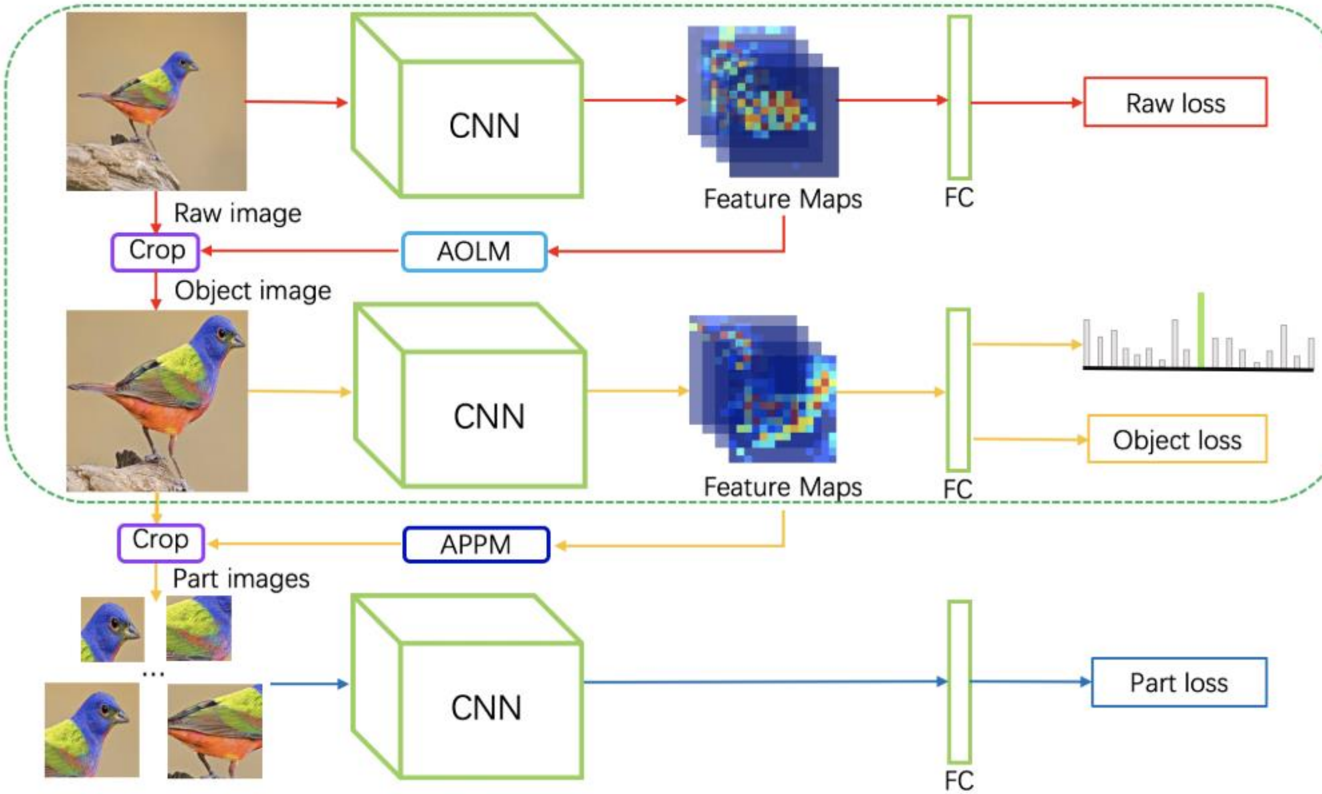
Laysan Albatross

Description: This is a large flying bird with black wings and a white belly.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Laysan Albatross

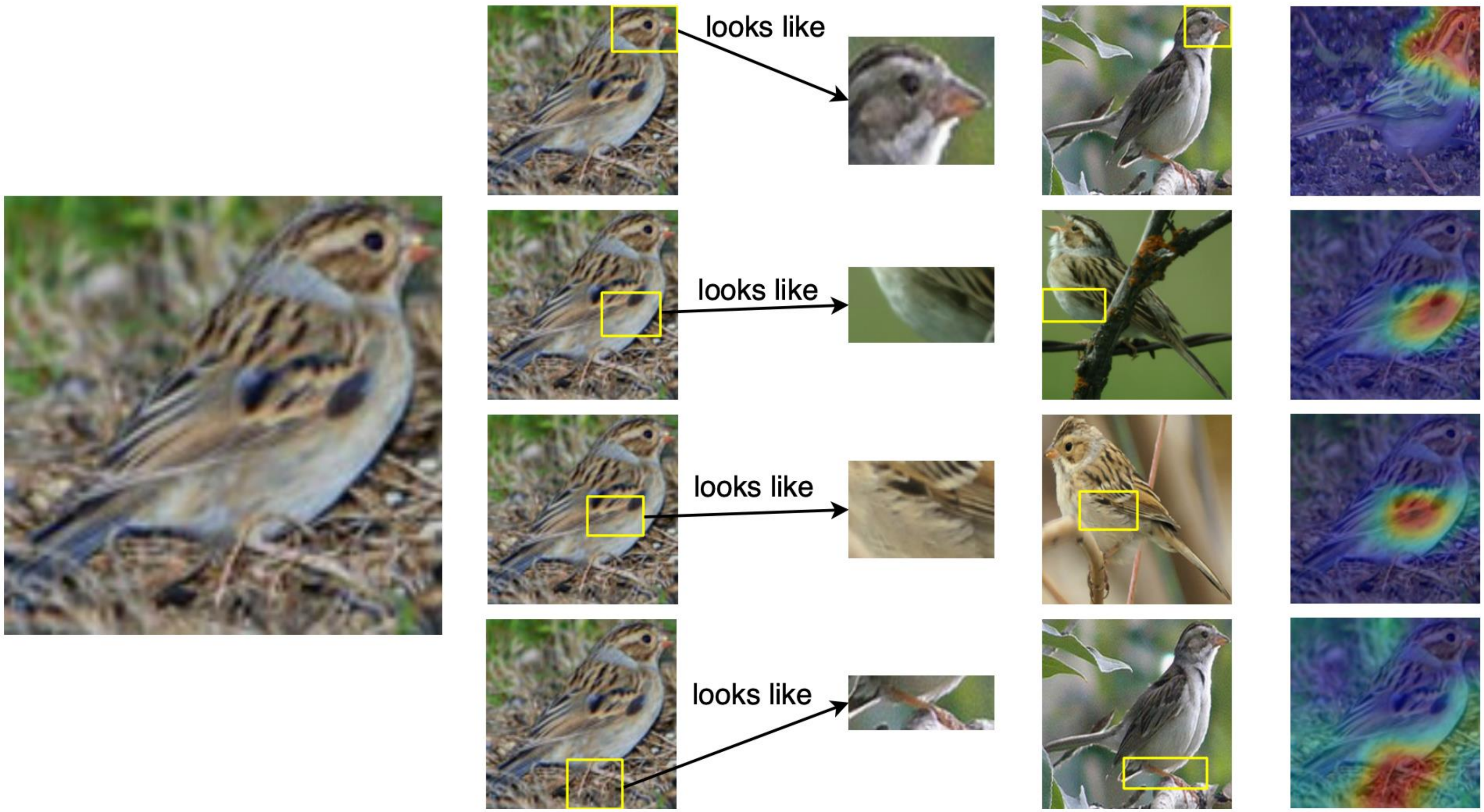
Description: This is a large bird with a white neck and a black back in the water.
Class Definition: The *Laysan Albatross* is a large seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

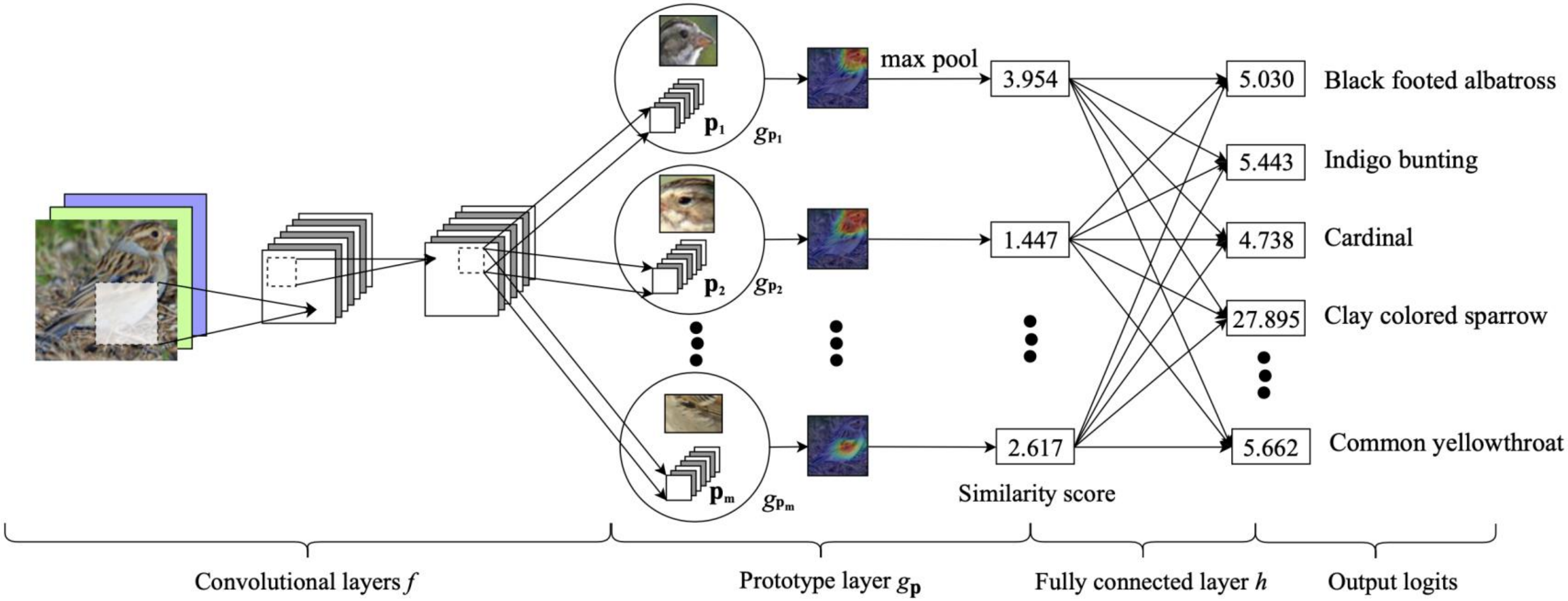
□ 注意力机制



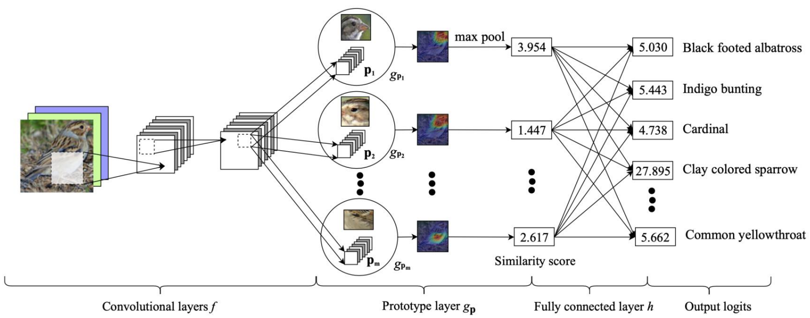
CNN
Raw loss
Feature Maps
FC
Raw image
Crop
AOLM
Object image
CNN
Object loss
Feature Maps
FC
Crop
APPM
Part images
...
CNN
Part loss
FC

- 作者介绍
- 研究背景
- 当前方法
- **本文方法**
- 实验效果
- 总结&反思


looks like
looks like
looks like
looks like



- 卷积网络 f：输入图像的大小为 224*224，通过卷积网络输出的大小为 HWD（e.g H=W=7）这一部分也就是常见的特征提取作用。
- 原型层g_p（prototype layer）：网络学习了 m 个原型 P:这些原型 P 以卷积层的特征图为输入，经过 m 组的卷积网络得到不同 patch 的原型激活值，该原型激活图的大小在本文中为 h=w=1。计算 pj 和 z 之间的 L2 距离，并将这个距离转换为相似度分数。
- 全连接层 h：经过前面的提取特征并聚类到原型得到相似度分数后，m 个相似度分数通过全连接层 h，得到最终的输出单元，经过 softmax 之后得到预测概率，分类图片结果。



- 每个类别预先选取10个训练集图片作为prototype
- 经特征提取，学习输入影像特征图和prototype的特征向量
- 相互计算特征图每个patch特征向量与prptotype特征向量L2距离
- L2距离转换为相似度分数

$$g_{\mathbf{p}_j}(\mathbf{z}) = \max_{\tilde{\mathbf{z}} \in \text{patches}(\mathbf{z})} \log\left((\|\tilde{\mathbf{z}} - \mathbf{p}_j\|_2^2 + 1) / (\|\tilde{\mathbf{z}} - \mathbf{p}_j\|_2^2 + \epsilon) \right)$$

- 可视化
  - 上采样至原图大小
  - 提取95%阈值的bbox，可视化
- M个相似度分数经全连接+softmax得到分类结果

- Stochastic gradient descent (SGD)

$$\min_{\mathbf{P}, w_{\text{conv}}} \frac{1}{n} \sum_{i=1}^{n} \text{CrsEnt}(h \circ g_{\mathbf{p}} \circ f(\mathbf{x_i}), \mathbf{y_i}) + \lambda_1 \text{Clst} + \lambda_2 \text{Sep}, \quad \text{where Clst and Sep are defined by}$$

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

  - The cross entropy loss (CrsEnt)
    - penalizes misclassification on the training data.
  - The minimization of the cluster cost (Clst)
    - encourages each training image to have some latent patch that is close to at least one prototype of its own class
  - The minimization of the separation cost (Sep)
    - encourages every latent patch of a training image to stay away from the prototypes not of its own class.

- Prototype visualization
  - smallest rectangular patch at least as large as the 95th-percentile of all activation values in that same map

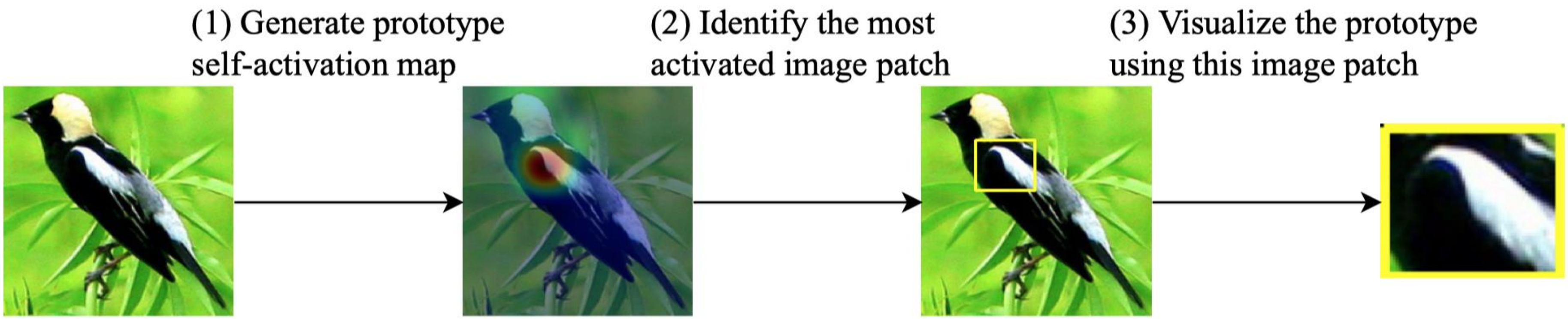


Figure 8: How to visualize a prototype.

- 作者介绍
- 研究背景
- 当前方法
- 本文方法
- **实验效果**
- 总结&反思

Table 1: Top: Accuracy comparison on cropped bird images of CUB-200-2011
Bottom: Comparison of our model with other deep models

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

Interpretability	Model: accuracy
None	**B-CNN**[26]: 85.1 (bb), 84.1 (full)
Object-level attn.	**CAM**[53]: 70.5 (bb), 63.0 (full)
Part-level attention	**Part R-CNN**[50]: 76.4 (bb+anno.); **PS-CNN** [16]: 76.2 (bb+anno.); **PN-CNN** [3]: 85.4 (bb+anno.); **DeepLAC**[25]: 80.3 (anno.); **SPDA-CNN**[49]: 85.1 (bb+anno.); **PA-CNN**[20]: 82.8 (bb); **MG-CNN**[45]: 83.0 (bb), 81.7 (full); **ST-CNN**[17]: 84.1 (full); **2-level attn.**[46]: 77.9 (full); **FCAN**[27]: 82.0 (full); **Neural const.**[36]: 81.0 (full); **MA-CNN**[52]: 86.5 (full); **RA-CNN**[8]: 85.3 (full)
Part-level attn. + prototypical cases	**ProtoPNet** (ours): 80.8 (full, VGG19+Dense121+Dense161-based) 84.8 (bb, VGG19+ResNet34+DenseNet121-based)

Table 1: Accuracy comparison on Stanford Cars

Baseline architecture	Accuracy of ProtoPNet	Accuracy of baseline
VGG19	87.4 ± 0.3	85.9 ± 0.2
ResNet34	86.1 ± 0.1	85.4 ± 0.1
DenseNet121	86.8 ± 0.1	89.7 ± 0.1

- 作者介绍
- 研究背景
- 当前方法
- 本文方法
- 实验效果
- **总结&反思**

- Inspiration
  - 可解释模型
  - 对传统CNN亲和
  - 可拓展性
    - MICCAI 2020
- To improve
  - L2 distance -> better metric
  - Pre-defined prototype
  - bbox anno.
  - Less diversity in selected image patch

Thanks for your listening