LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision

Mateusz Pach, Koryna Lewandowska, Jacek Tabor, Bartosz Zieliński, **Dawid Rymarczyk**

Post-doc Researcher at Group of Machine Learning Research @ **Jagiellonian University**
Director of Data Science and Artificial Intelligence Center of Excellence @ **Ardigen**

Agenda

1. Interpretability introduction
2. Introduction to inherently interpretable neural networks and prototypical parts
   a. PIPNet (Nauta@CVPR2023)
3. LucidPPN (Pach2024@arxiv)

Interpretability

Introduction

Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature machine intelligence 1.5 (2019): 206-215.

Rudin, Cynthia, et al. "Interpretable machine learning: Fundamental principles and 10 grand challenges." Statistic Surveys 16 (2022): 1-85.

Kodratoff, Y. (1994). The comprehensibility manifesto. KDD Nugget Newsletter.

Li, Xuhong, et al. "Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond." Knowledge and Information Systems 64.12 (2022): 3197-3234.

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. Advances in neural information processing systems, 31.

Interpretability - definition

Model is interpretable when its behaviour is predictable and understandable for the user

Interpretability - definition

Model is interpretable when its behaviour is predictable and understandable for the user

So, the user knows:

- reasons behind predictions
- is able to predict the decision of the model
- is able to predict the explanation of the model

Interpretability vs. XAI

There has been a recent explosion of work on ‘explainable ML’

Interpretability vs. XAI

There has been a recent explosion of work on ‘explainable ML’

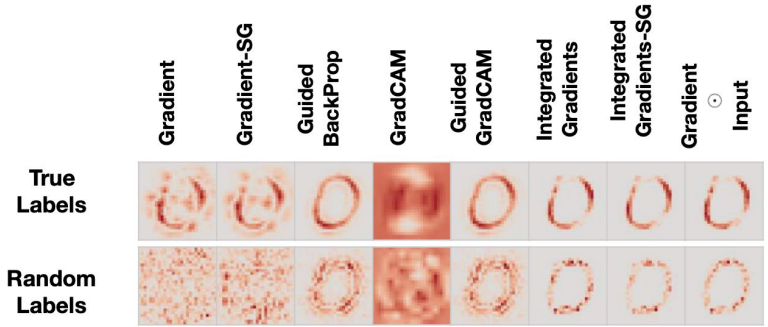
explainable ML -> second (post hoc) model is created to explain the first black box model.
This is problematic.

Interpretability vs. XAI

There has been a recent explosion of work on ‘explainable ML’

explainable ML -> second (post hoc) model is created to explain the first black box model.
This is problematic.

Explanations are often not reliable, and can be misleading.



Interpretability vs. XAI

There has been a recent explosion of work on ‘explainable ML’

explainable ML -> second (post hoc) model is created to explain the first black box model.
This is problematic.

Explanations are often not reliable, and can be misleading.

If we instead use models that are inherently interpretable, they provide their own explanations, which are faithful to what the model actually computes.

Interpretable Machine Learning

XAI or not XAI

Interpretable ML is not a subset of XAI.

The term XAI dates from ~2016, and grew out of work on function approximation; i.e., explaining a black box model by approximating its predictions by a simpler model, or explaining a black box using local approximations.

Interpretable Machine Learning

XAI or not XAI

Interpretable ML is not a subset of XAI.

The term XAI dates from ~2016, and grew out of work on function approximation; i.e., explaining a black box model by approximating its predictions by a simpler model, or explaining a black box using local approximations.

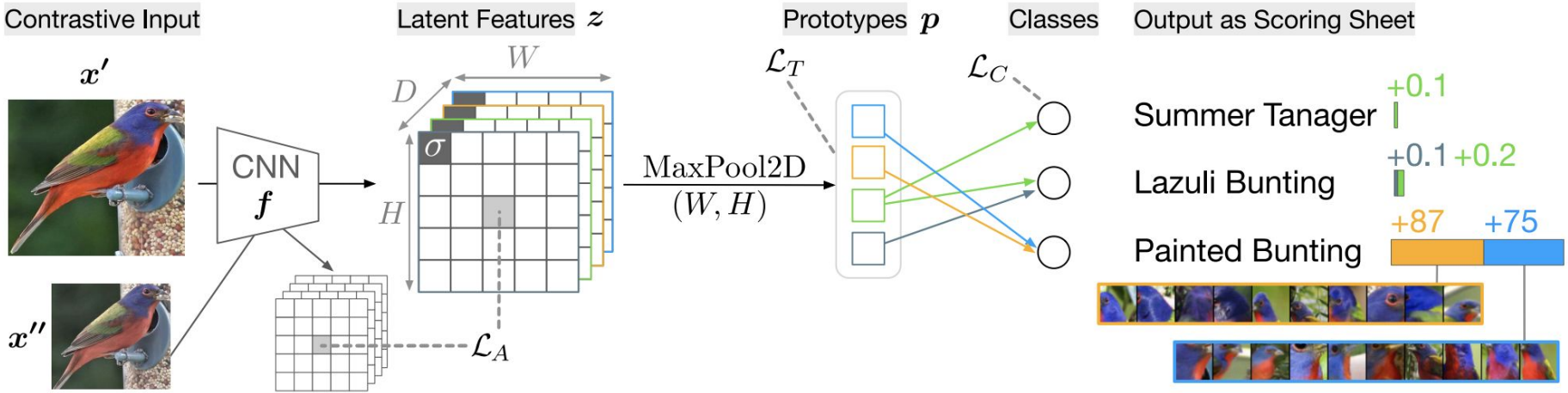
Interpretable ML also has a (separate) long and rich history, dating back to the days of expert systems in the 1950's, and the early days of decision trees.

Introduction to inherently interpretable neural networks and prototypical parts

Nauta, Meike, et al. "Pip-net: Patch-based intuitive prototypes for interpretable image classification." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

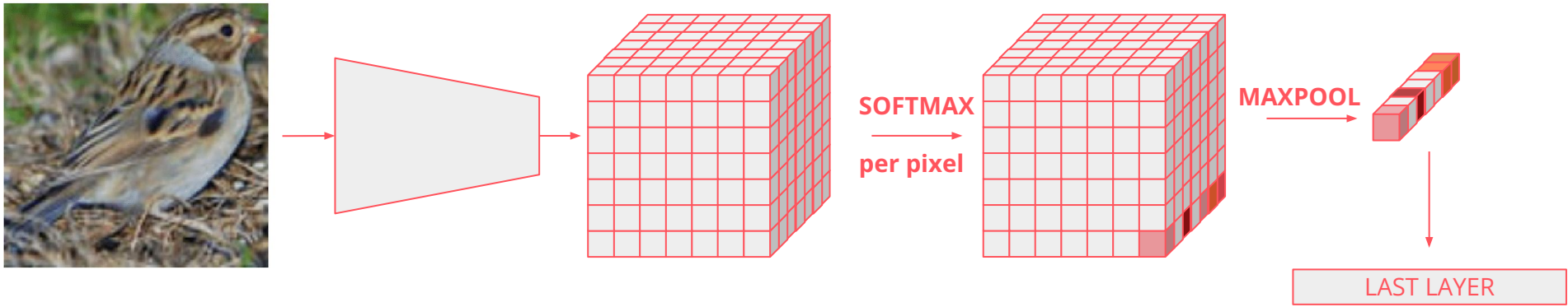
PIPNet

Architecture



PIPNet

How it works?



Limitations of prototypical parts from a user perspective:

group of machine gmum learning research

Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

LucidPPN

What is really important on the image?

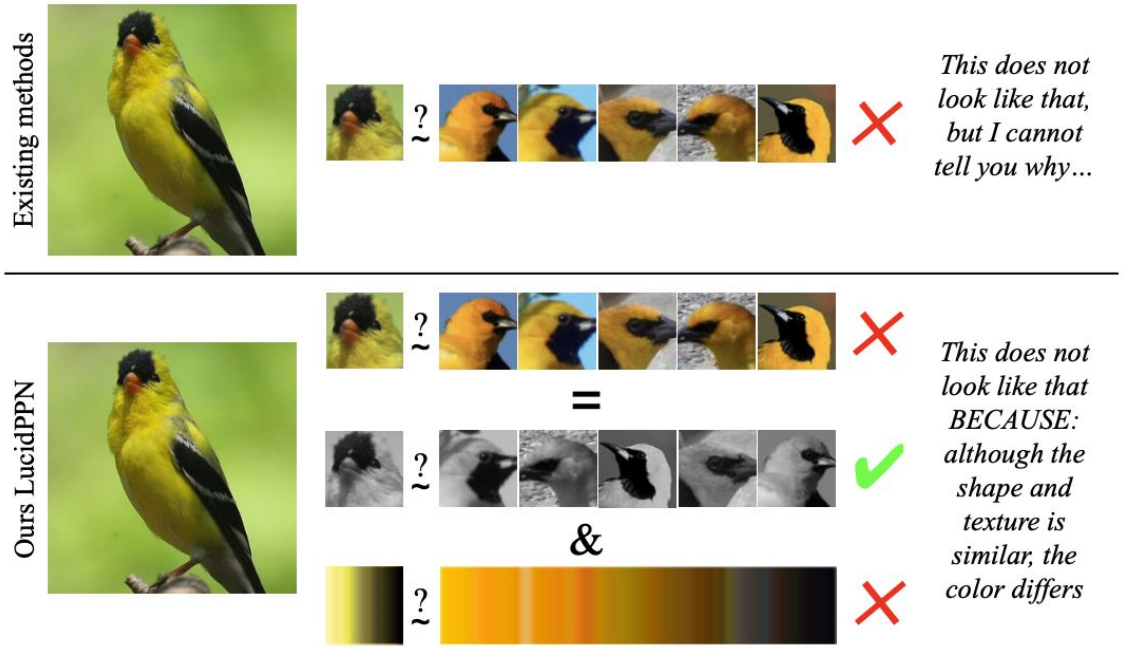


Image source: Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

LucidPPN

What are our contributions?

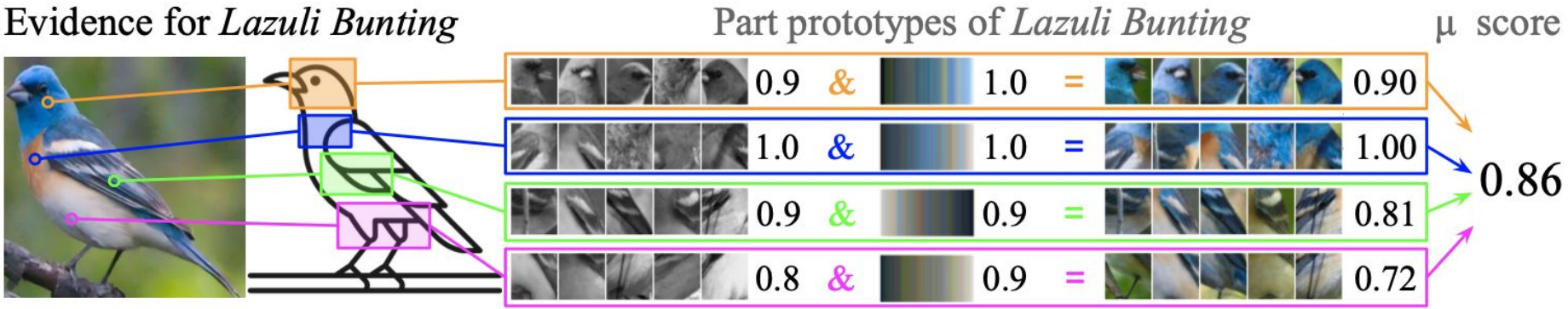


Image source: Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

LucidPPN

Architecture

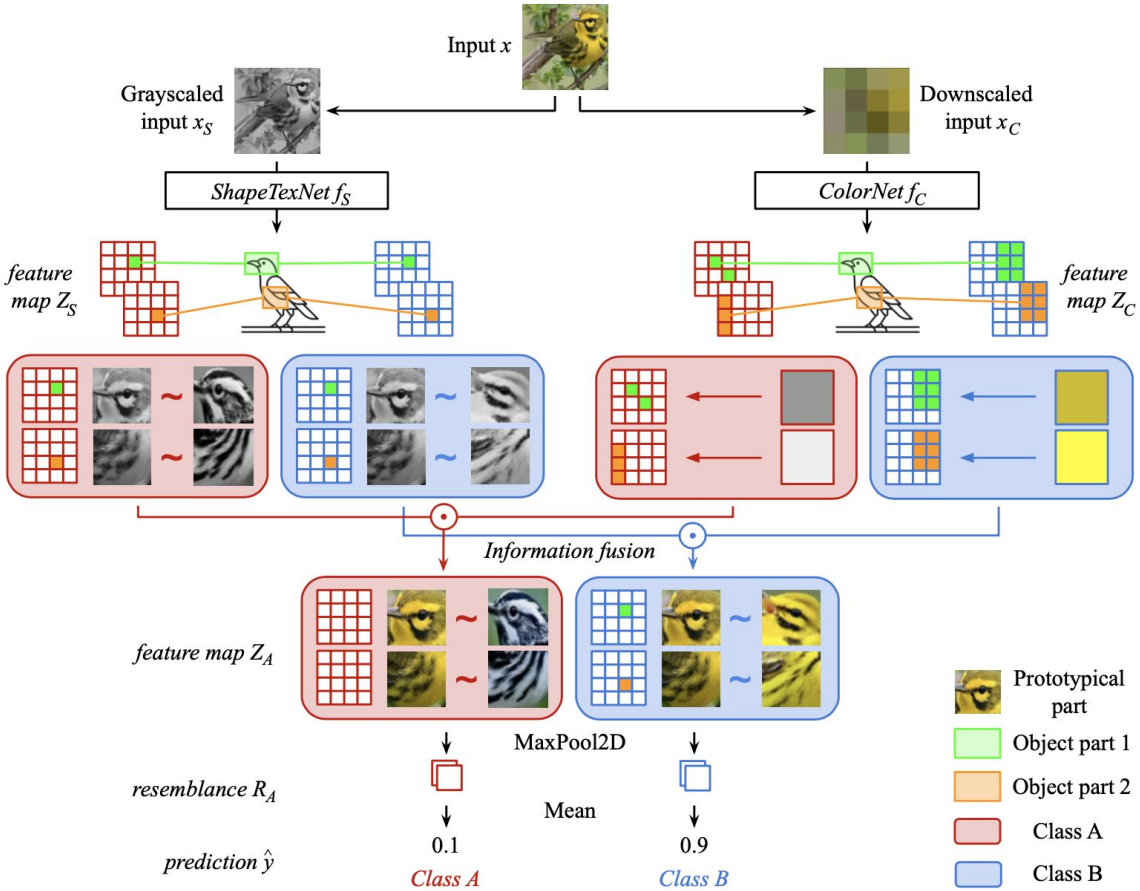


Image source: Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

LucidPPN

Training

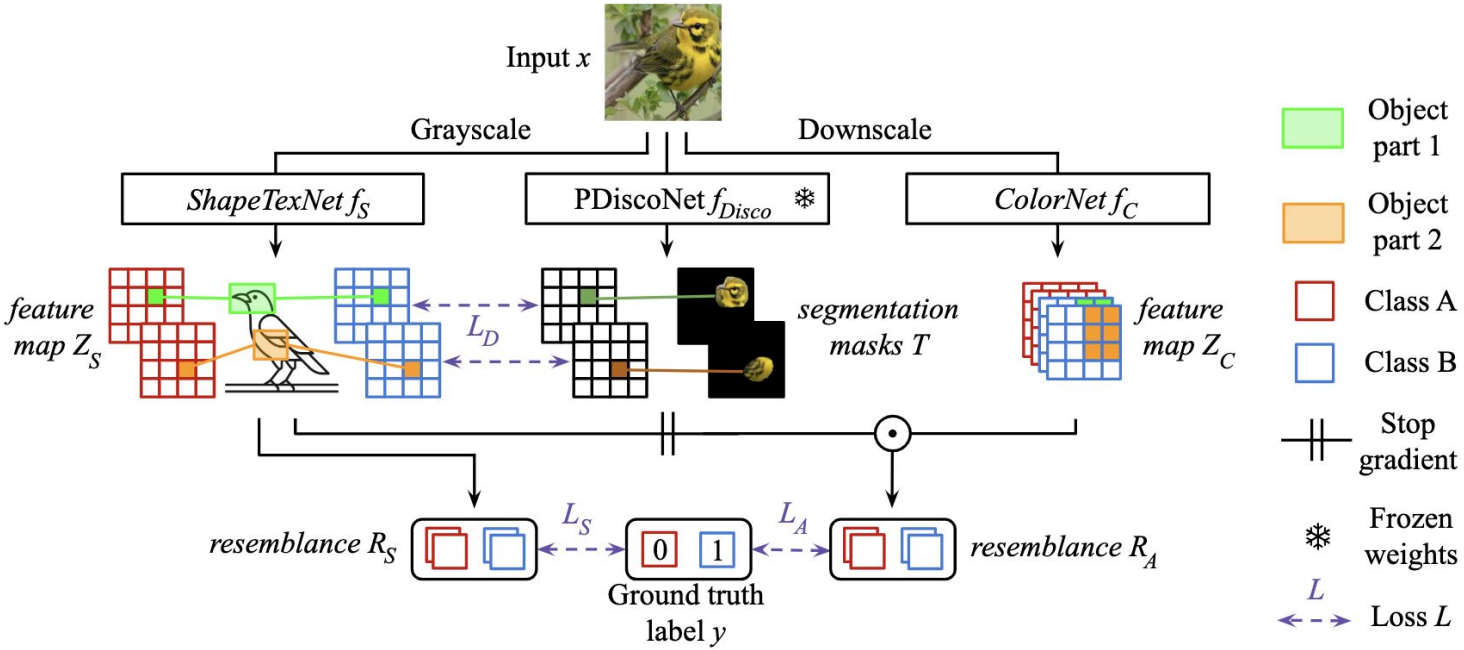


Image source: Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

LucidPPN

Accuracy

	CUB	CARS	DOGS	FLOWER
ProtoPNet (Chen et al., 2019)	79.2	86.1	77.4±0.2	92.1±0.3
ProtoTree (Nauta et al., 2021b)	82.2 ± 0.7	86.6 ± 0.2	–	–
ProtoPShare (Rymarczyk et al., 2021)	74.7	86.4	74.1±0.3	90.3±0.2
ProtoPool (Rymarczyk et al., 2022c)	**85.5 ± 0.1**	88.9 ± 0.1	71.7±0.2	92.7±0.1
PIP-Net (Nauta et al., 2023)	84.3 ± 0.2	88.2 ± 0.5	**80.8 ± 0.4**	91.8 ± 0.5
LucidPPN	81.5 ± 0.4	**91.6 ± 0.2**	79.4 ± 0.4	**95.0 ± 0.3**

Image source: Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

LucidPPN

Influence of color

	CUB	CARS	DOGS	FLOWER
ShapeTexNet	80.4	**91.7**	78.6	93.6
LucidPPN	**81.8**	**91.7**	**78.9**	**95.3**

Image source: Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

LucidPPN

Influence of part correspondence

	CUB	CARS	DOGS	FLOWER
PIP-Net	84.3	88.2	80.8	91.8
LucidPPN	81.5	91.6	79.4	95.0
single branch	**86.6**	**91.9**	**82.7**	**95.6**

Image source: Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

LucidPPN

User study

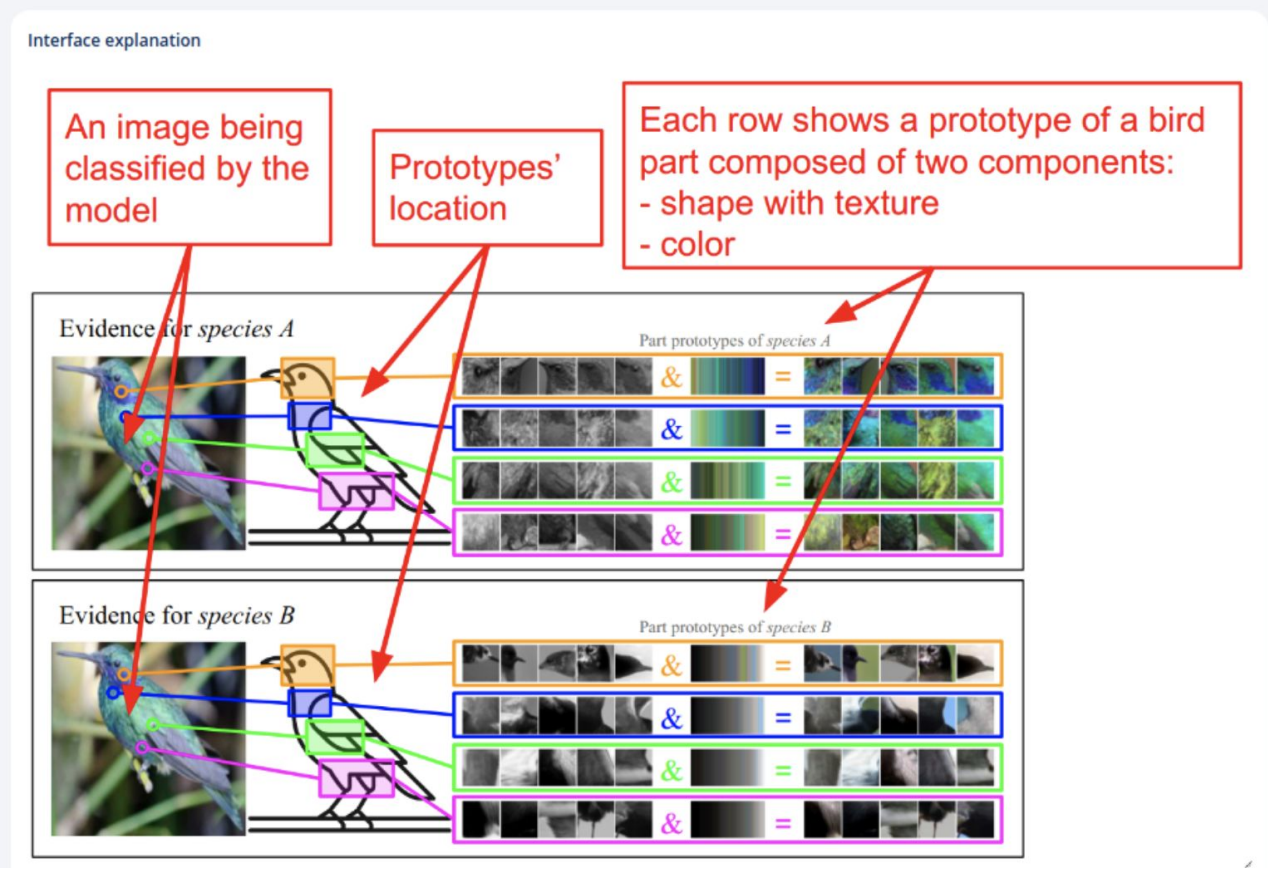


Image source: Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

LucidPPN

Reducing ambiguity of explanations

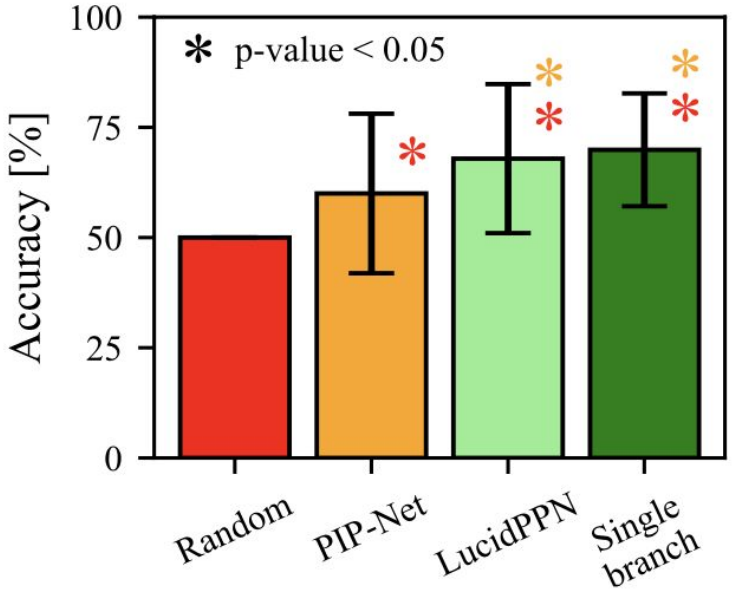


Image source: Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

Thank you!

Q&A?