group of machine gmum learning research

Current trends in intrinsically interpretable deep learning

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. ProtoPNet (Chen@NeurIPS2019)
   b. PIPNet (Nauta@CVPR2023)
3. Limitations of prototypical parts from a user perspective:
   a. spatial misalignment (Sacha@AAAI2024)
   b. overconfidence (Kim@ECCV2022)
   c. disambiguation of prototypical parts (Ma@NeurIPS2023, Pach@arxiv2024)
4. Interaction with a user (Bontempelli@ICLR2023)
5. ICICLE - Interpretable Continual Learning (Rymarczyk@ICCV2023)

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

group of machine learning research

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

Interpretability - definition

group of machine
gmum
learning research

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’

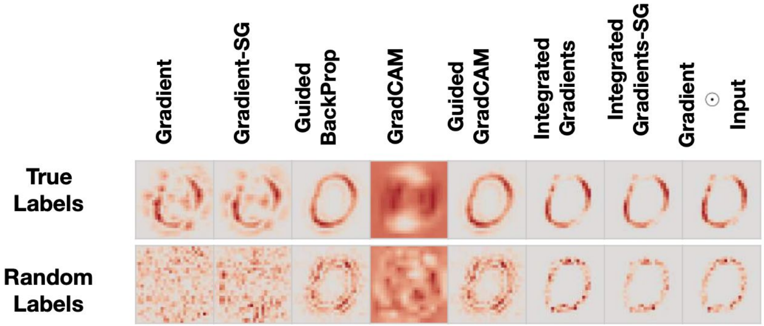
explainable ML -> second (post hoc) model is created to explain the first black box model.
This is problematic.

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

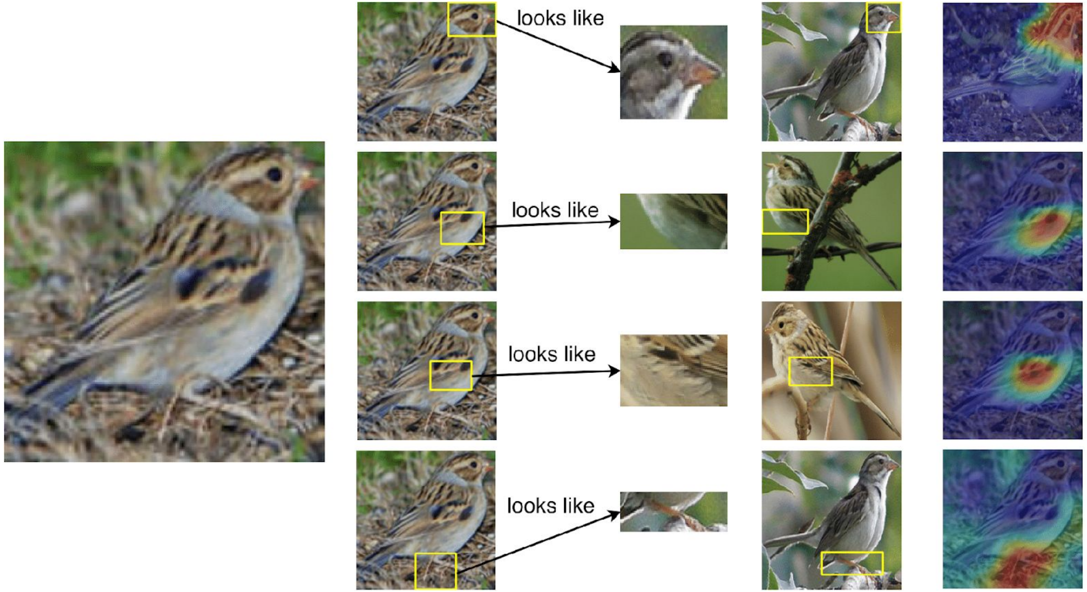
group of machine gmum learning research

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

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.

ProtoPNet

This looks like that



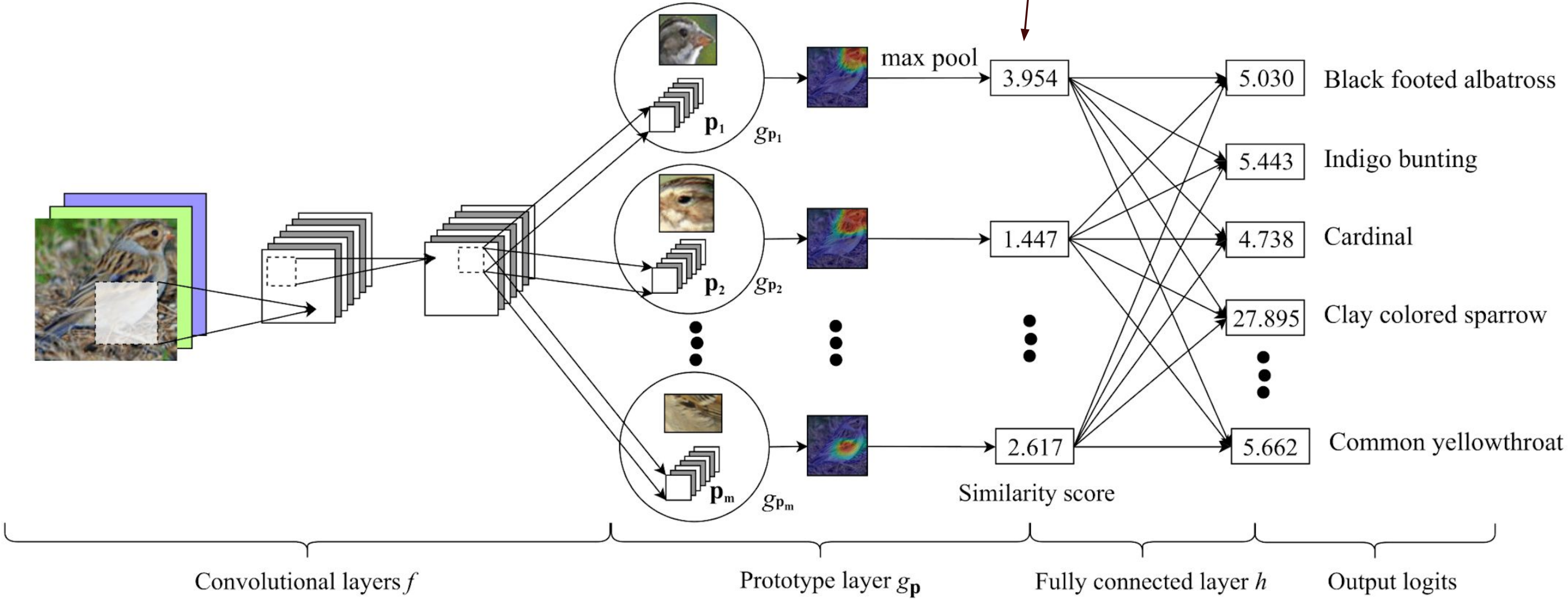
Chen et al. This looks like that: deep learning for interpretable image recognition, NeurIPS 2019

ProtoPNet

Architecture

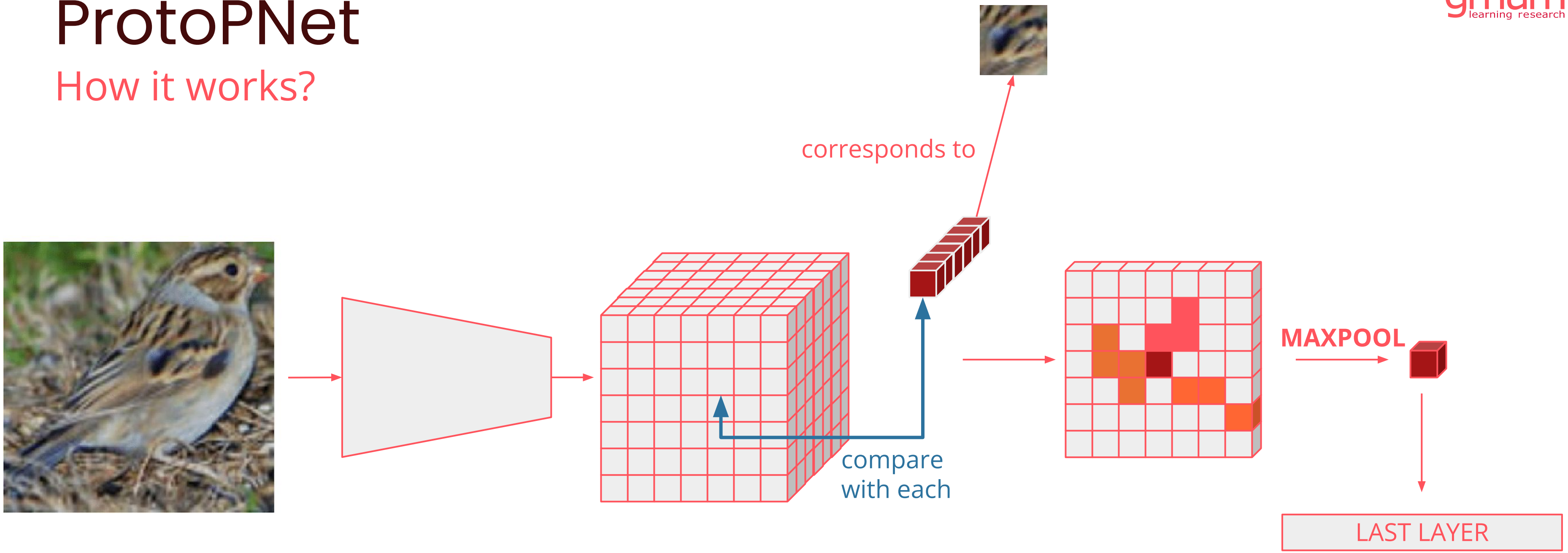
$$g(Z_x, p) = \max_{z \in Z_x} \log\left(\frac{\|z-p\|^2+1}{\|z-p\|^2+\varepsilon}\right) \quad \text{for} \quad \varepsilon > 0.$$

Monotonically decreasing with respect to $\|z-p\|^2$



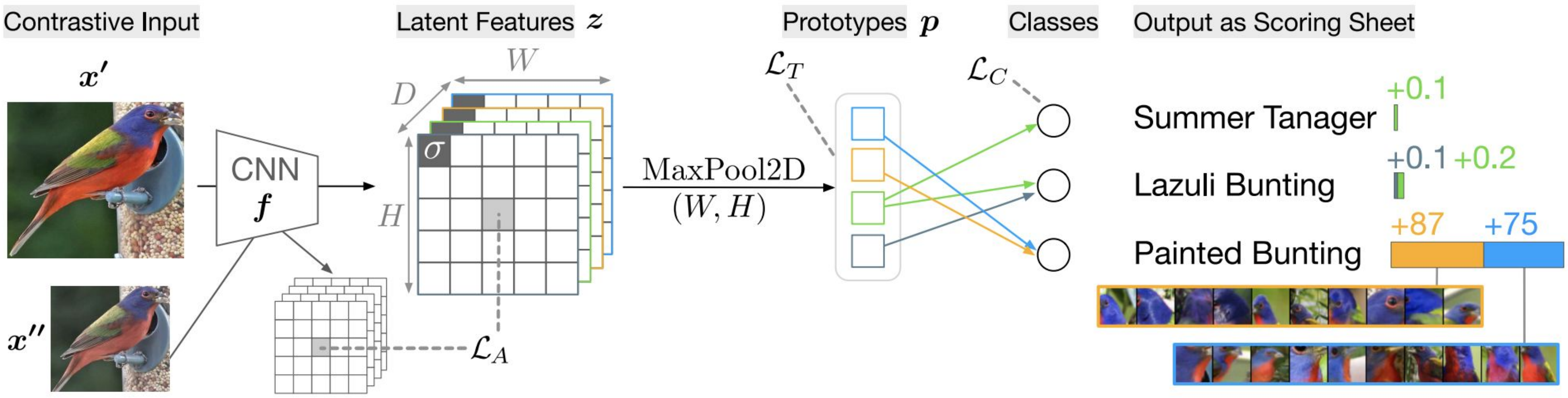
ProtoPNet

How it works?



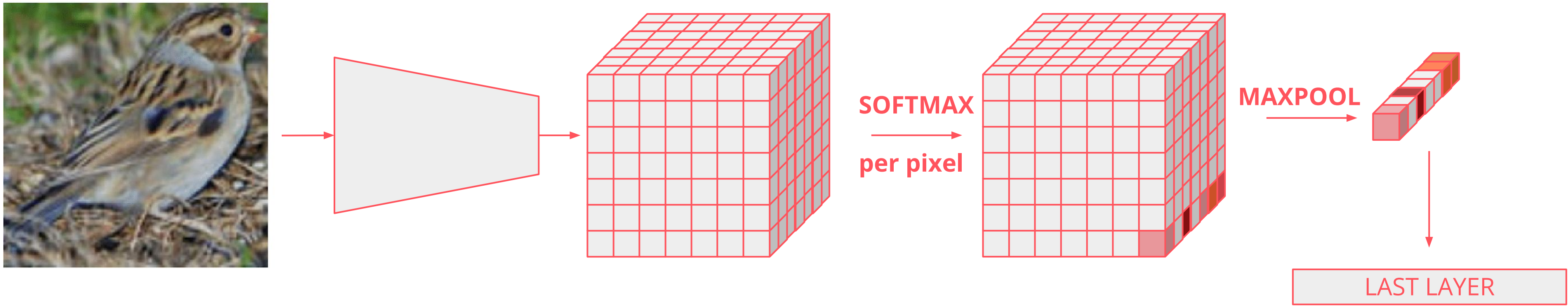
PIPNet

Architecture



PIPNet

How it works?



Limitations of prototypical parts from a user perspective:

Sacha, M., Jura, B., Rymarczyk, D., Struski, Ł., Tabor, J., & Zieliński, B. (2024, March). Interpretability benchmark for evaluating spatial misalignment of prototypical parts explanations. AAAI.

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.

Kim, Sunnie SY, et al. "HIVE: Evaluating the human interpretability of visual explanations." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.

Ma, Chiyu, et al. "This looks like those: Illuminating prototypical concepts using multiple visualizations." Advances in Neural Information Processing Systems 36 (2024).

Spatial Misalignment

Are highlighted pixels really important?

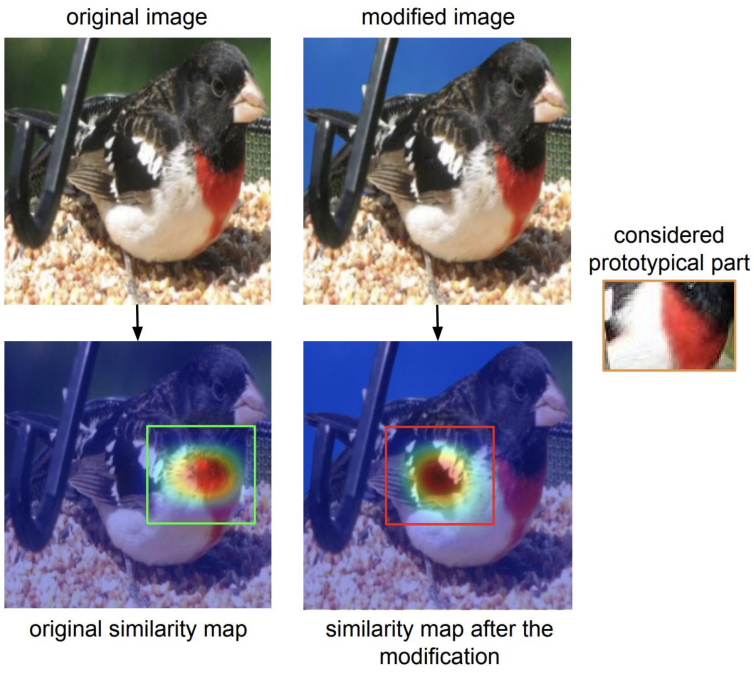


Image source: Sacha, M., Jura, B., Rymarczyk, D., Struski, Ł., Tabor, J., & Zieliński, B. (2024, March). Interpretability benchmark for evaluating spatial misalignment of prototypical parts explanations. AAAI.

Spatial Misalignment

Are highlighted pixels really important?

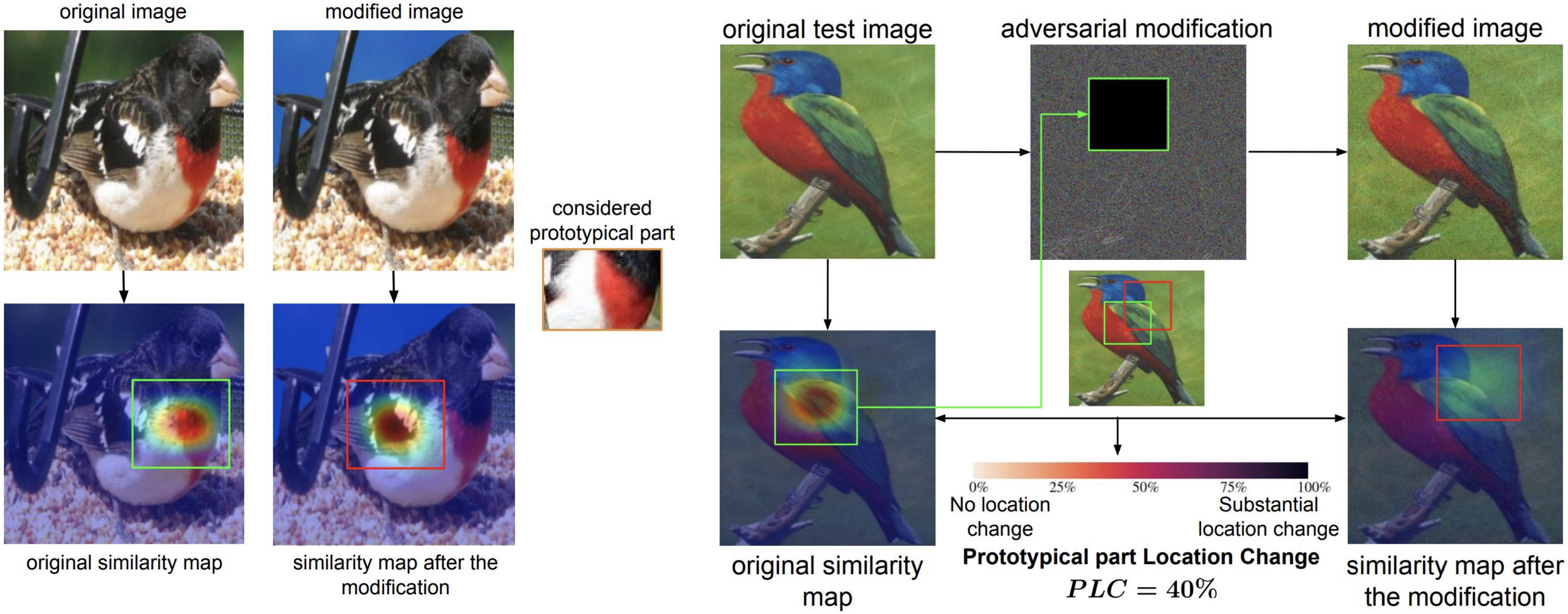
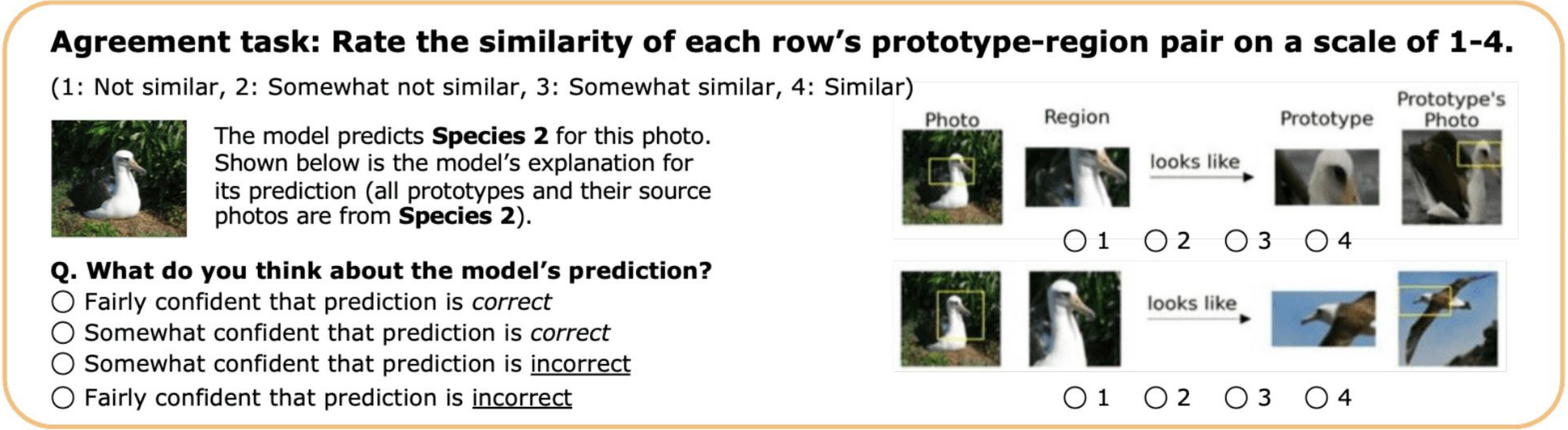


Image source: Sacha, M., Jura, B., Rymarczyk, D., Struski, Ł., Tabor, J., & Zieliński, B. (2024, March). Interpretability benchmark for evaluating spatial misalignment of prototypical parts explanations. AAAI.

Explanations make the user overconfident



Kim, Sunnie SY, et al. "HIVE: Evaluating the human interpretability of visual explanations." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022

Explanations make the user overconfident

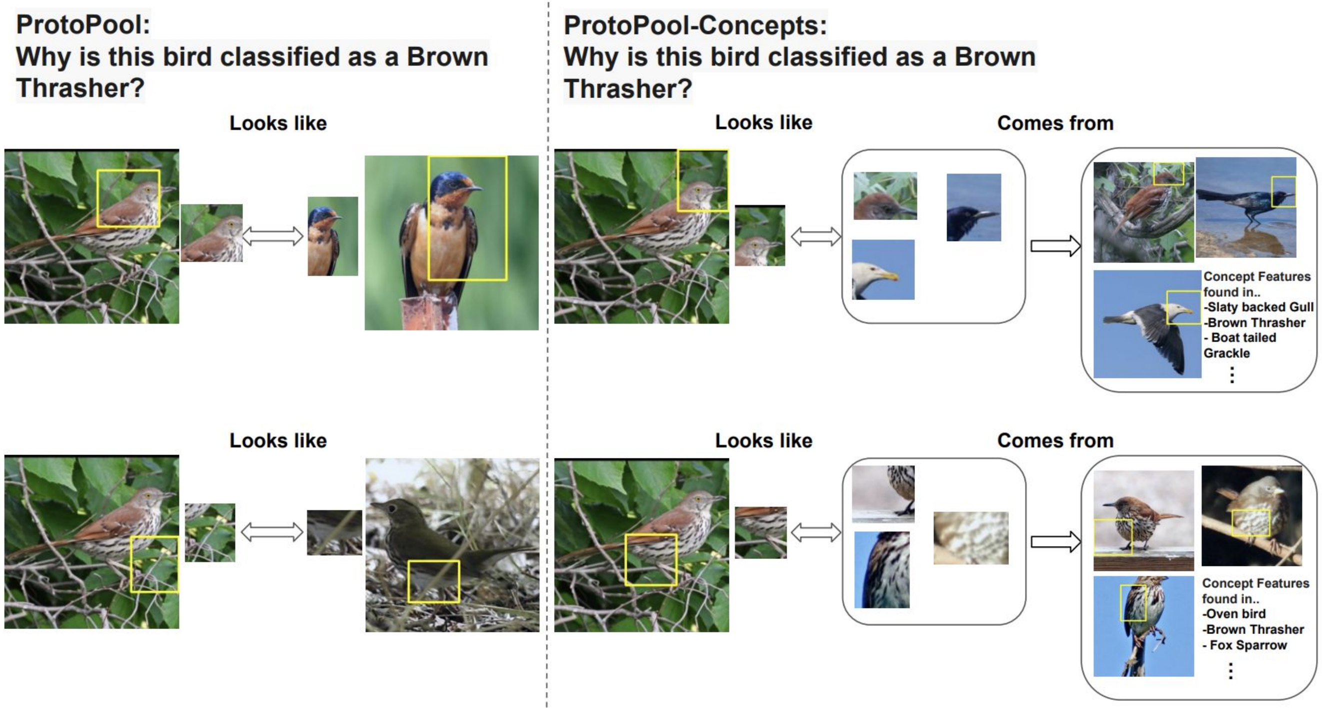
In all studies, participants leaned towards believing that model predictions are correct when provided explanations, regardless of if they are actually correct.

CUB	GradCAM [61]	BagNet [10]	ProtoPNet [15]	ProtoTree [48]
Correct	72.4% ± 21.5 (2.9)	**75.6% ± 23.4 (3.0)**	73.2% ± 24.9 (3.0)	66.0% ± 33.8 (2.8)
Incorrect	32.8% ± 24.3 (2.8)	*42.4% ± 28.7 (2.7)*	***46.4% ± 35.9 (2.4)***	37.2% ± 34.4 (2.7)

Kim, Sunnie SY, et al. "HIVE: Evaluating the human interpretability of visual explanations." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022

Reducing overconfidence

or reducing disambiguation



Ma, Chiyu, et al. "This looks like those: Illuminating prototypical concepts using multiple visualizations." Advances in Neural Information Processing Systems 36 (2024).

LucidPPN

What is really important on the image?

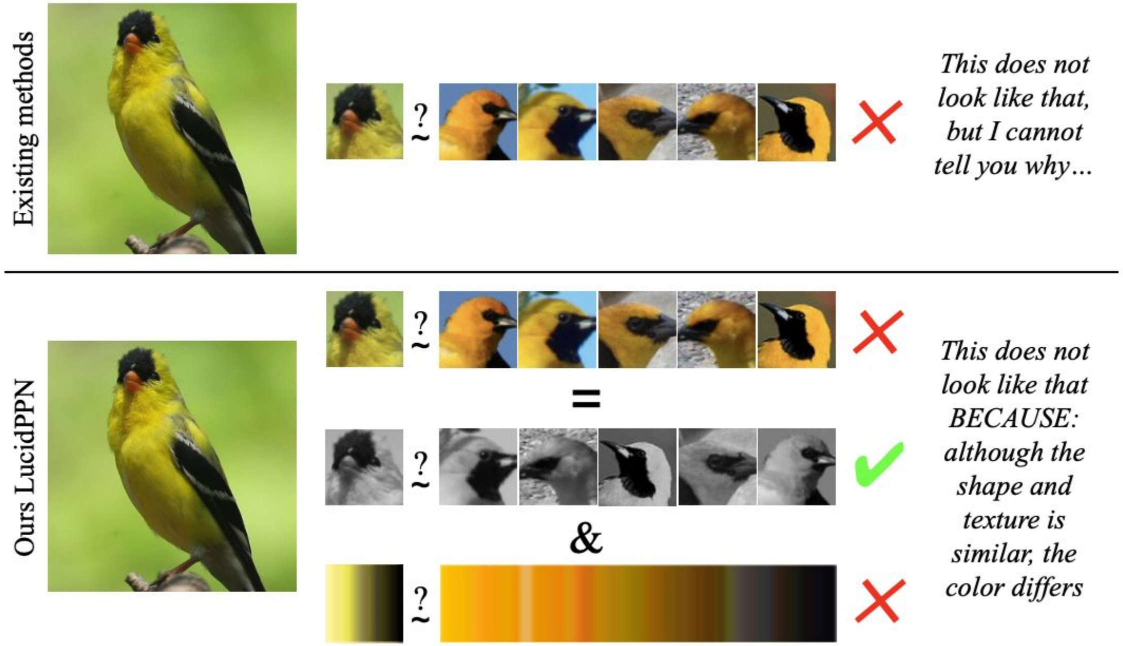


Image source: Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

LucidPPN

What are our contributions?

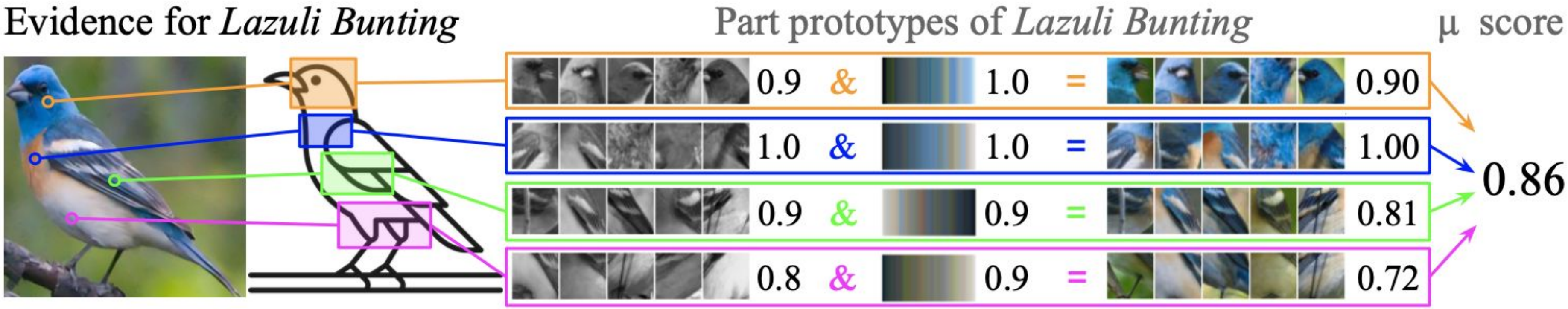


Image source: Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

LucidPPN

Architecture

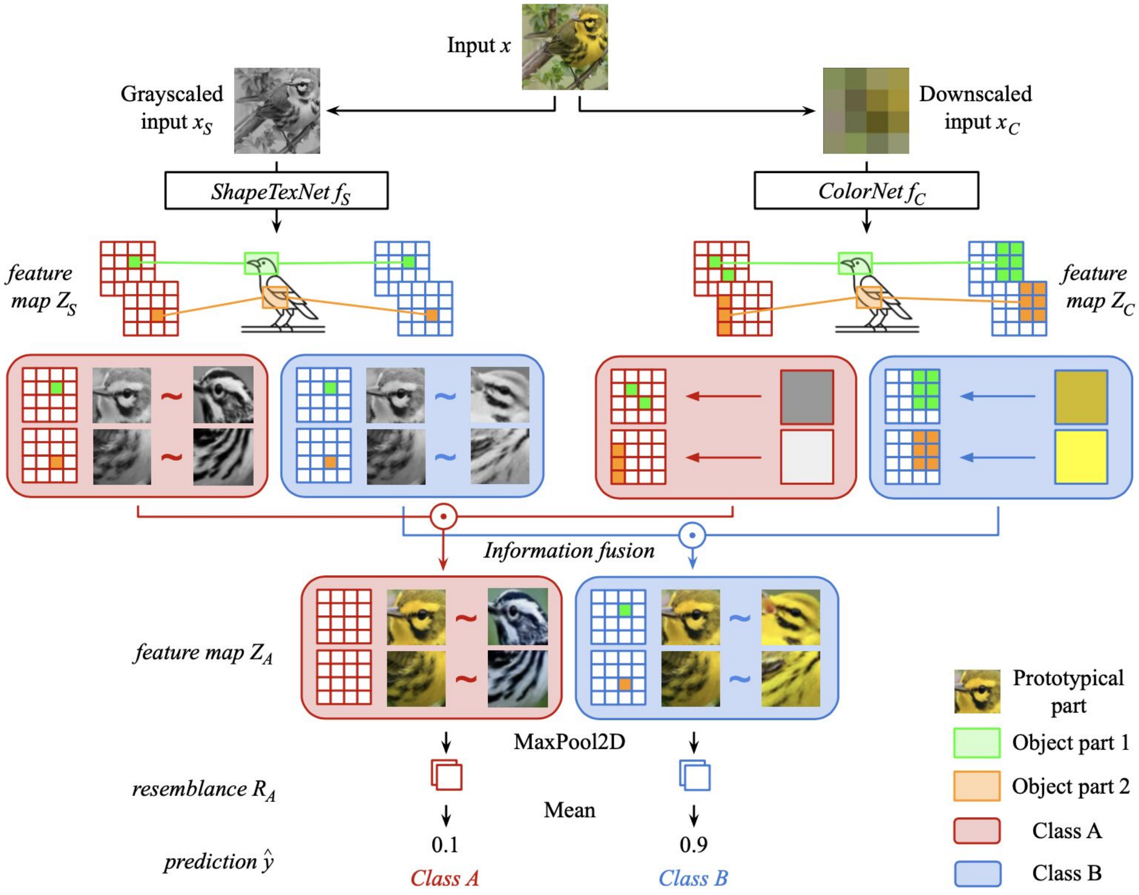
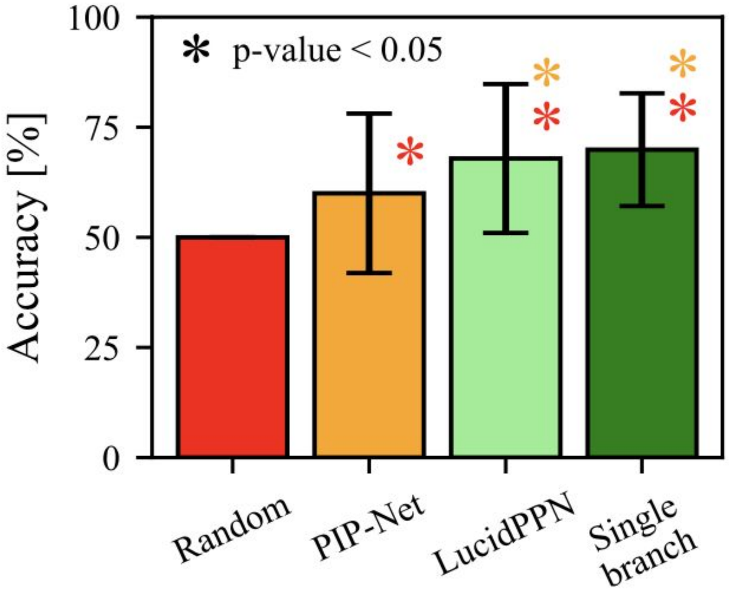


Image source: Pach, M., Rymarczyk, D., Lewandowska, K., Tabor, J., & Zieliński, B. (2024). LucidPPN: Unambiguous Prototypical Parts Network for User-centric Interpretable Computer Vision. arXiv preprint arXiv:2405.14331.

LucidPPN

Reducing ambiguity of explanations



	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.

Interaction with a user

Bontempelli, A., Teso, S., Tentori, K., Giunchiglia, F., & Passerini, A. (2023). Concept-level debugging of part-prototype networks. ICLR.

Interaction with a user

Not only forget but learn a useful thing

ProtoPDebug method allows to forget a concept, but this may harm the model's performance.

Can we redirect model's attention to other part of the image to learn a new concept from human feedback?

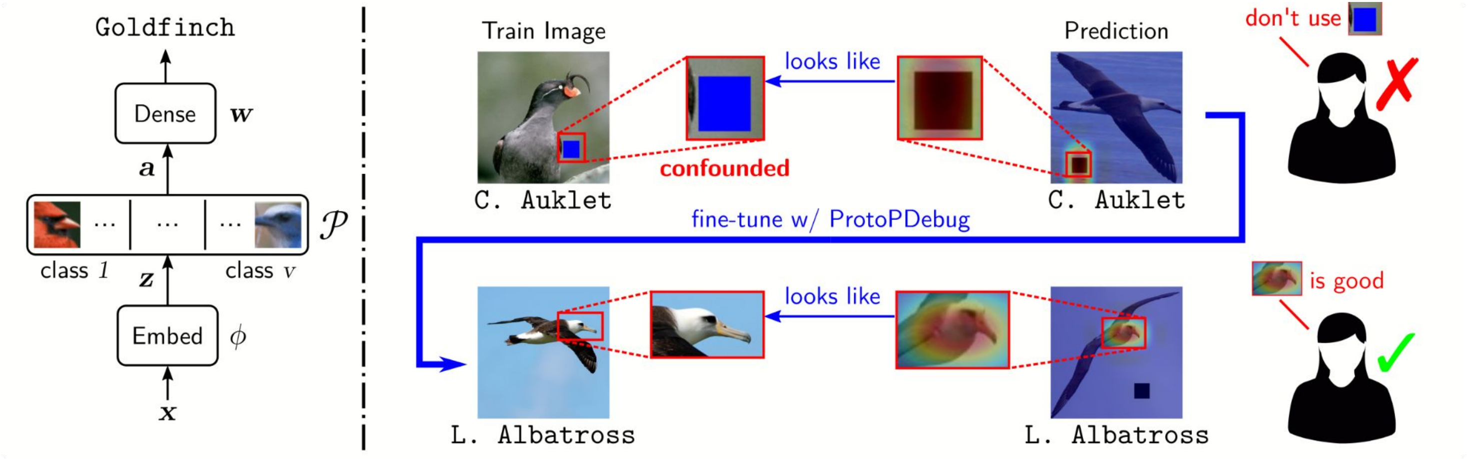


Image source: Bontempelli, A., Teso, S., Tentori, K., Giunchiglia, F., & Passerini, A. (2023). Concept-level debugging of part-prototype networks. ICLR

ICICLE - Interpretable CL

Rymarczyk, Dawid, et al. "Icicle: Interpretable class incremental continual learning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.

ICICLE

Motivation

Interpretable continual learning

Preserving knowledge about the interpretable concepts within the data

Robustness to Interpretability Concept Drift

$$ICD = \mathbb{E}_{i,j=1}^{H,W} \left| sim(p^{t-1}, z_{i,j}^{t}) - sim(p^{t}, z_{i,j}^{t}) \right|$$

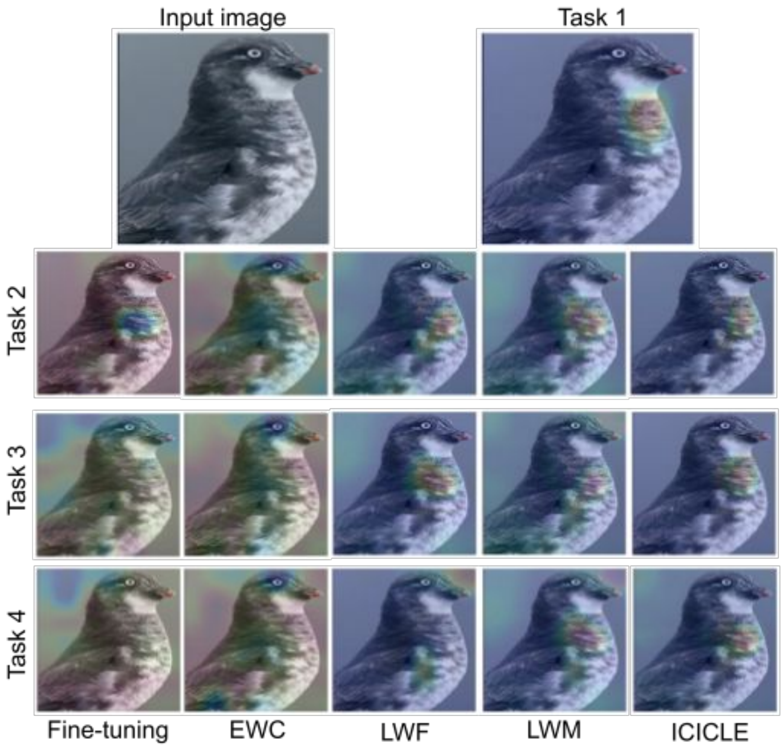


Image source: Rymarczyk, D., van de Weijer, J., Zieliński, B., & Twardowski, B. (2023). Icicle: Interpretable class incremental continual learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1887-1898).

ICICLE

Interpretability regularization

Preserving the knowledge about the concepts

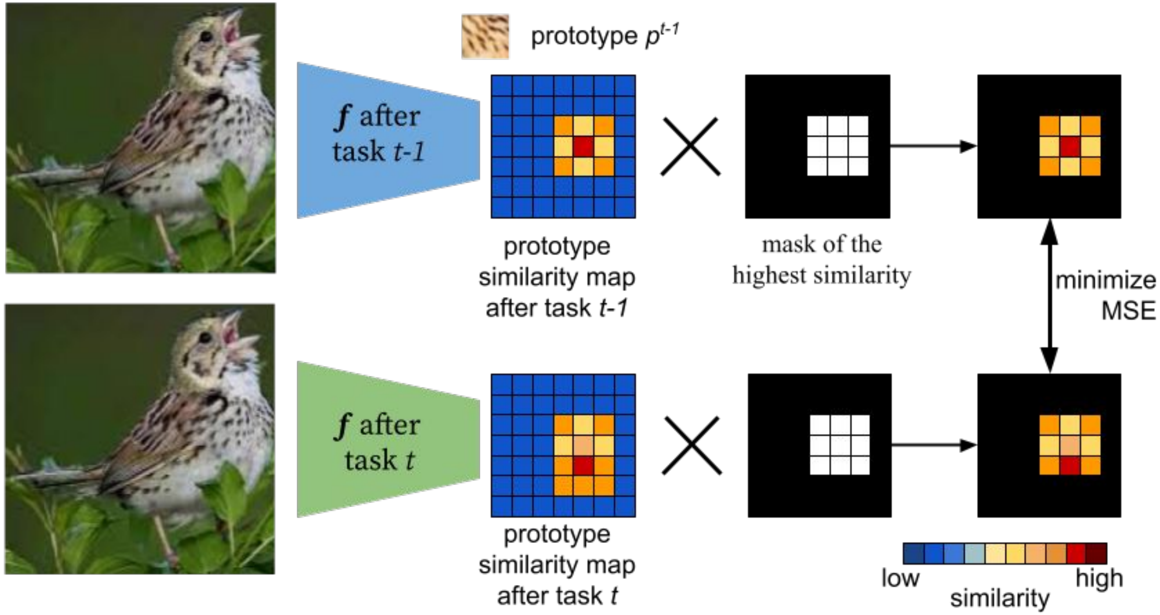


Image source: Rymarczyk, D., van de Weijer, J., Zieliński, B., & Twardowski, B. (2023). Icicle: Interpretable class incremental continual learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1887-1898).

ICICLE

Interpretability regularization

What is distilled defines what kind of plasticity model have when learning new tasks.

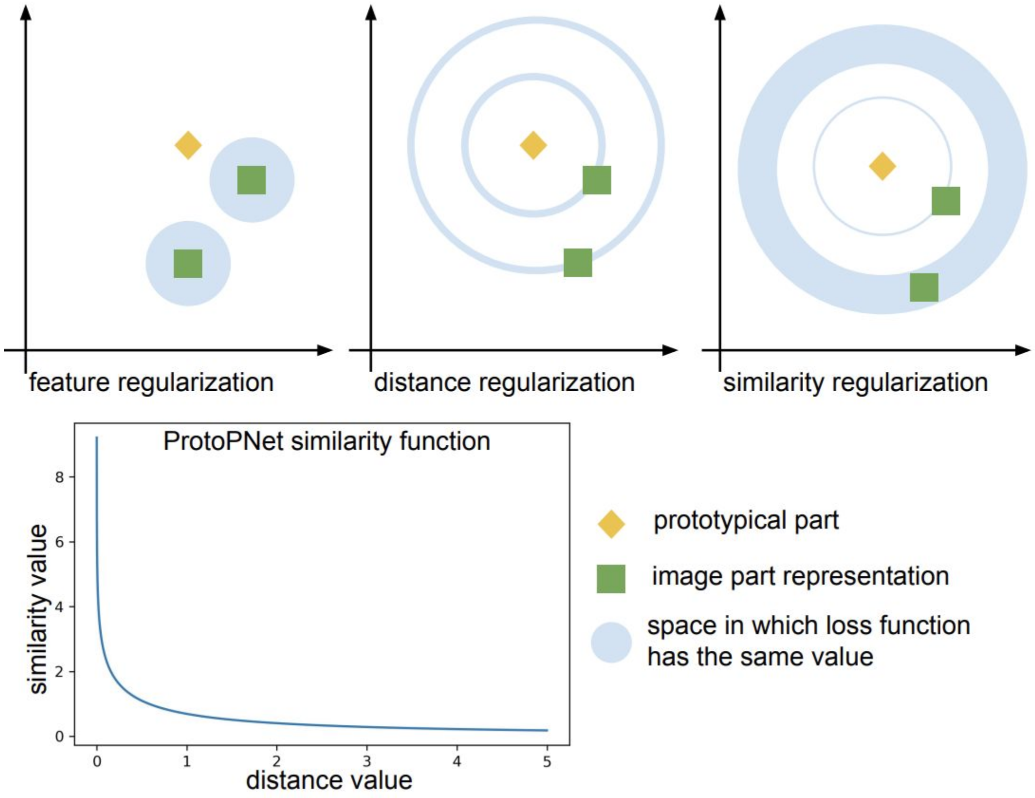


Image source: Rymarczyk, D., van de Weijer, J., Zieliński, B., & Twardowski, B. (2023). Icicle: Interpretable class incremental continual learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1887-1898).

Results

Interpretability concept drift

METHOD	IoU			
	TASK 1	TASK 2	TASK 3	MEAN
FINETUNING	0.115	0.149	0.260	0.151
EWC	0.192	0.481	0.467	0.334
LWF	0.221	0.193	0.077	0.188
LWM	0.332	0.312	0.322	0.325
ICICLE	**0.705**	**0.753**	**0.742**	**0.728**

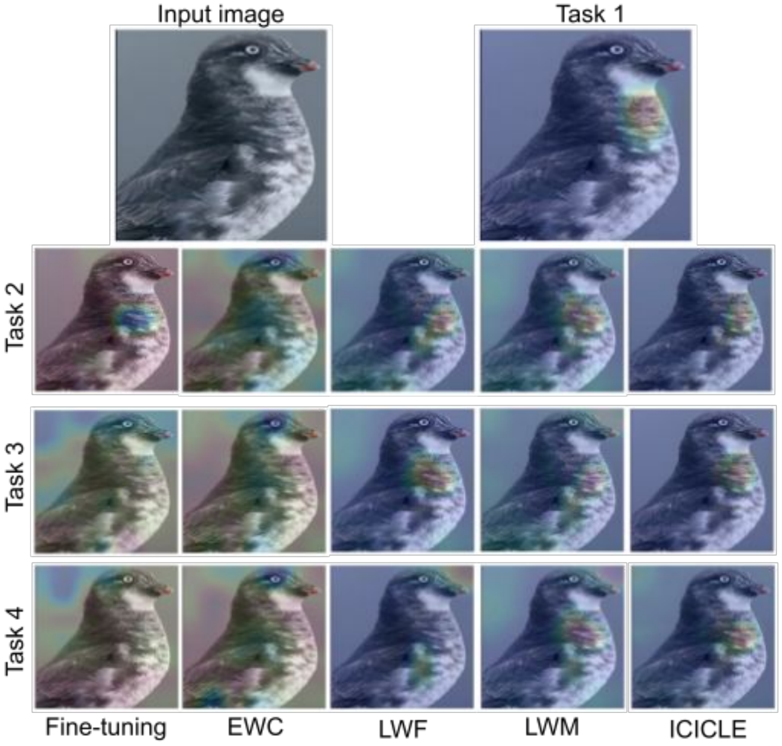


Image source: Rymarczyk, D., van de Weijer, J., Zieliński, B., & Twardowski, B. (2023). Icicle: Interpretable class incremental continual learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1887-1898).

group of machine
gmum
learning research

Thank you!

Q&A?