Explainability in neural networks

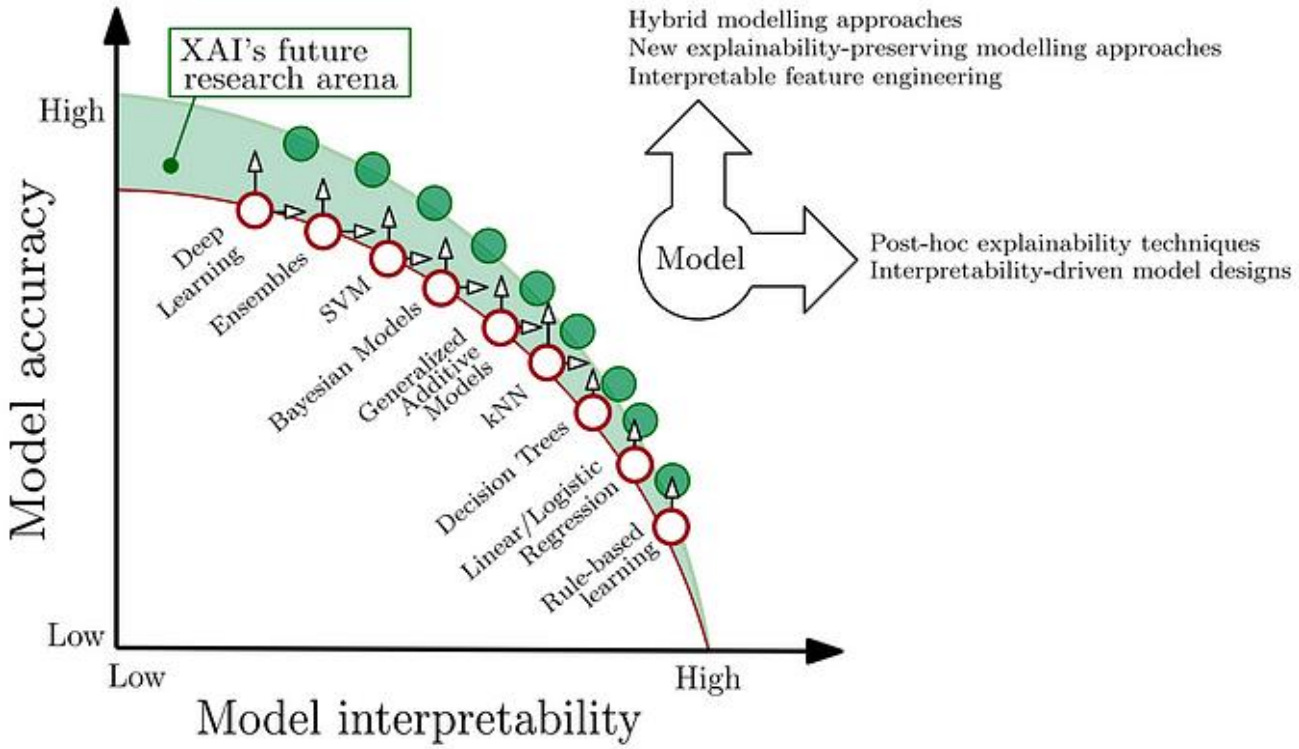
Szymon Bobek

Jagiellonian University
2023

https://geist.re

DNN are blackboxes, but excellent blackboxes

- DNN can be treated as any other Balckbox model and analyzed with all sorts of model-agnostic approaches (SHPA, PDP, LIME)
- There are plenty of methods that were crafted for DNN to understand or debug their operation
- There is also a new trend in building self-explainable DNN



Post-hoc methods for DNN

Class activation maps (CAM)

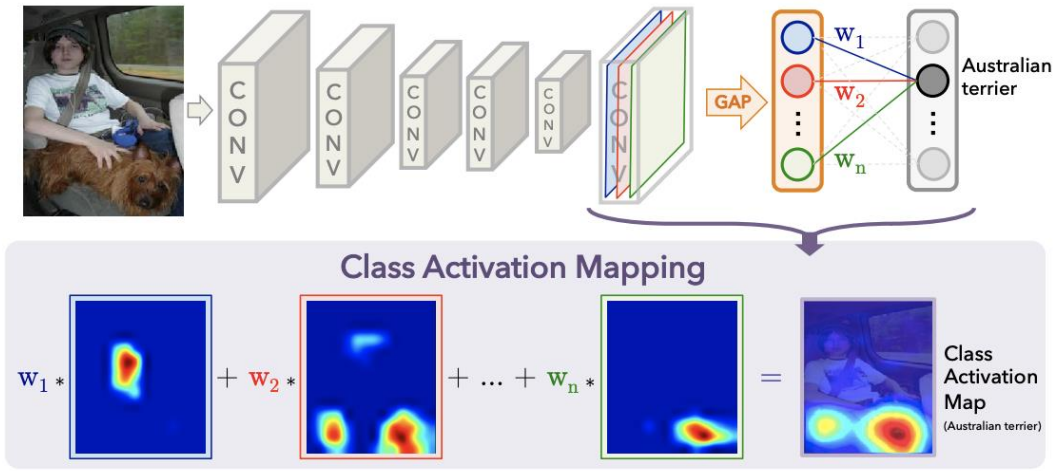


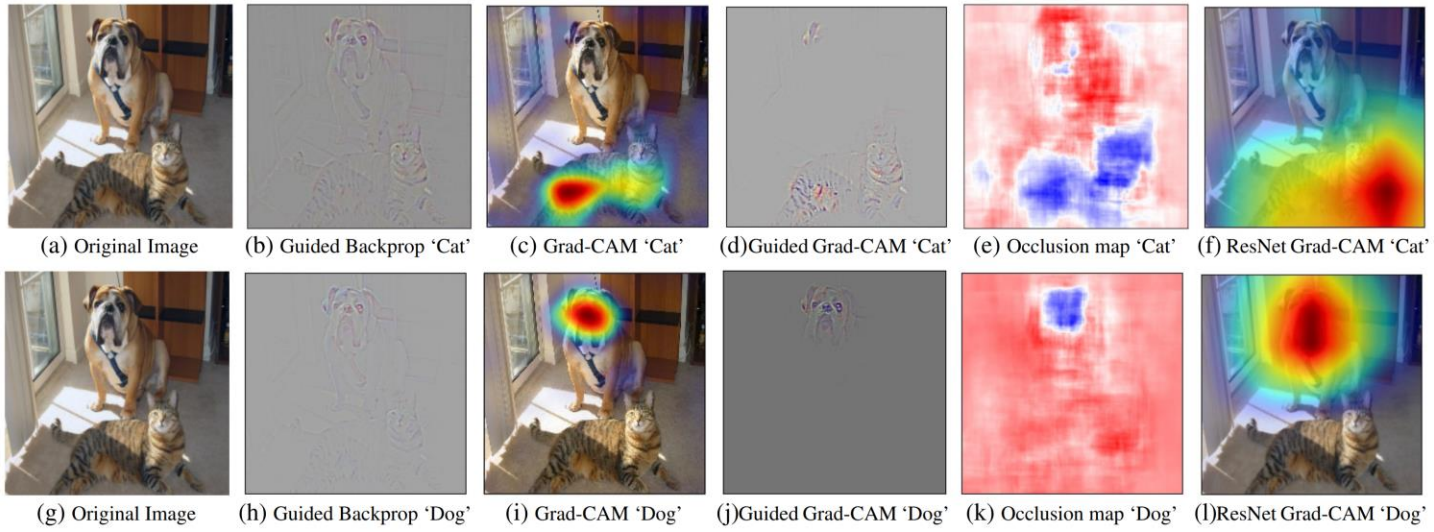
Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

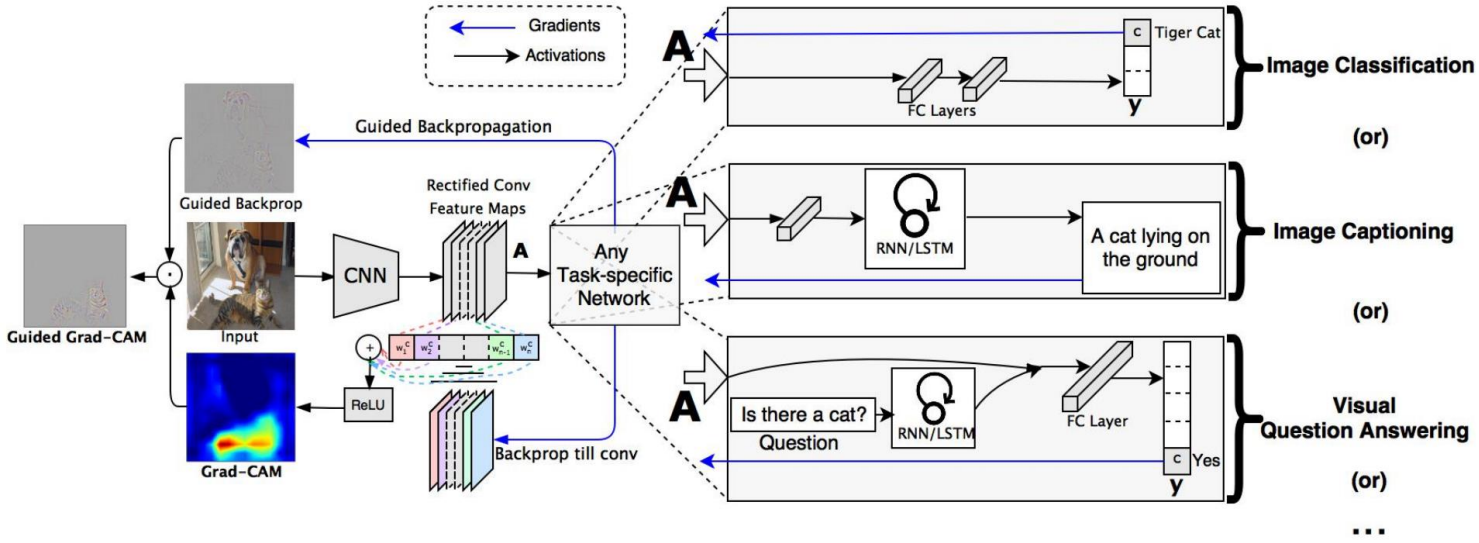
- It exploits locality of CNN arhcitecture
- It uses latent representation from last CNN layer as class activation maps
- Requires modifications to the architecture, i.e. injecting Global Average Pooling (GAP) at the end
- Requires re-training of GAP (with frozen weights of CNN layers)
- It might be to restrictive for more complex tasks than simple classification
- In fact, we are explaining only the last feature map

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva and A. Torralba, "Learning Deep Features for Discriminative Localization," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016 pp. 2921-2929.

GradCAM

- Motivation: come up with CAM-like asrchitecture that not restricts the architecture
- It also uses features maps produced by the last CNN layer of the model
- IN GradCAM we base on gradients, not weights
- Instead of GAP, we use backpropagation to obtain partial derivatives
- Positive gradient contribute to the given class
- We apply average pooling to obtain the weights for feature map
- The scores are passed to ReLU to cut all negative values

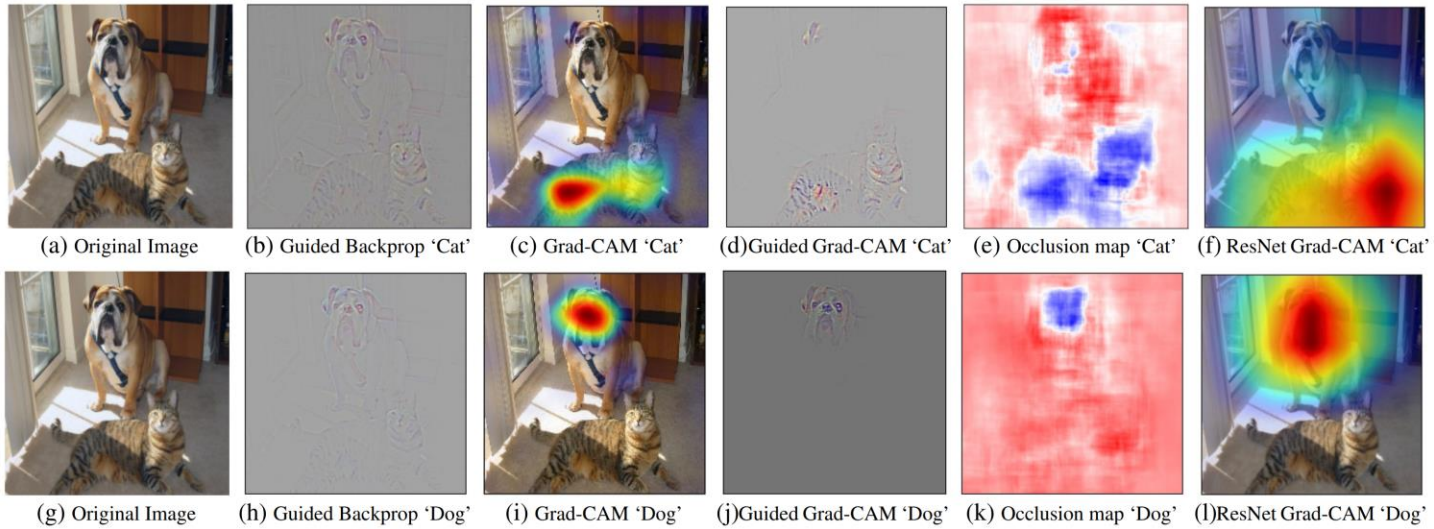




R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

GradCAM

- Motivation: come up with CAM-like asrchitecture that not restricts the architecture
- It also uses features maps produced by the last CNN layer of the model
- IN GradCAM we base on gradients, not weights
- Instead of GAP, we use backpropagation to obtain partial derivatives
- Positive gradient contribute to the given class
- We apply average pooling to obtain the weights for feature map
- The scores are passed to ReLU to cut all negative values



$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

K-th feature map

K stands for k feature maps

Total number of pixels in feature map

R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

Ante-hoc methods for DNN

DNN are blackboxes, but excellent blackboxes

Perspective | Published: 13 May 2019

Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead

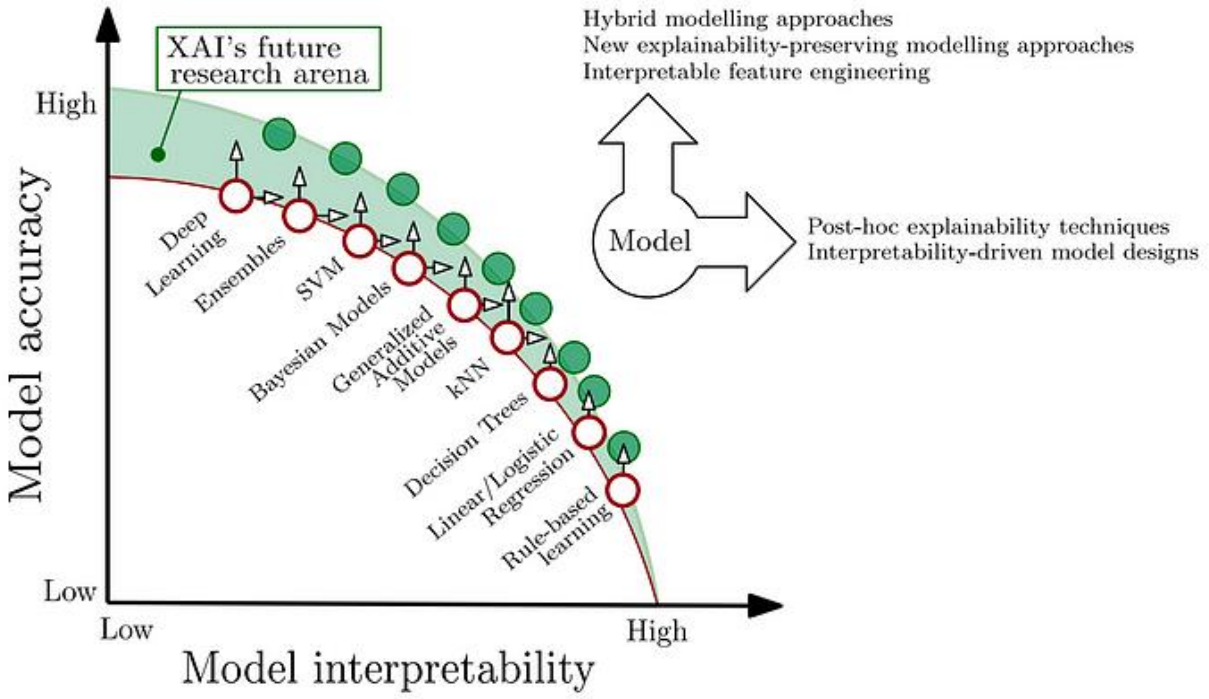
Cynthia Rudin

Nature Machine Intelligence **1**, 206–215 (2019) | Cite this article

A preprint version of the article is available at arXiv.

Abstract

Black box machine learning models are currently being used for high-stakes decision making throughout society, causing problems in healthcare, criminal justice and other domains. Some people hope that creating methods for explaining these black box models will alleviate some of the problems, but trying to explain black box models, rather than creating models that are interpretable in the first place, is likely to perpetuate bad practice and can potentially cause great harm to society. The way forward is to design models that are inherently interpretable. This Perspective clarifies the chasm between explaining black boxes and using inherently interpretable models, outlines several key reasons why explainable black boxes should be avoided in high-stakes decisions, identifies challenges to interpretable machine learning, and provides several example applications where interpretable models could potentially replace black box models in criminal justice, healthcare and computer vision.



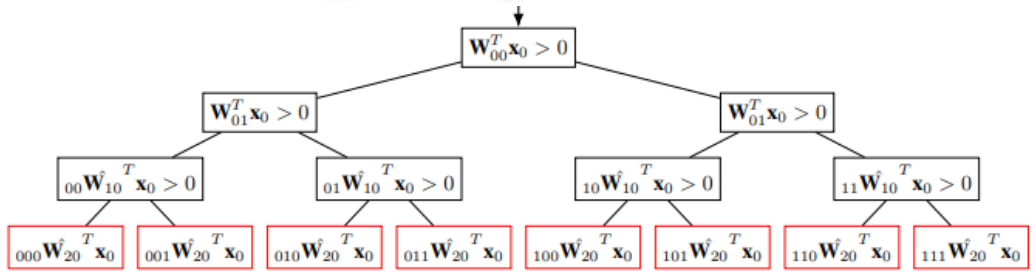
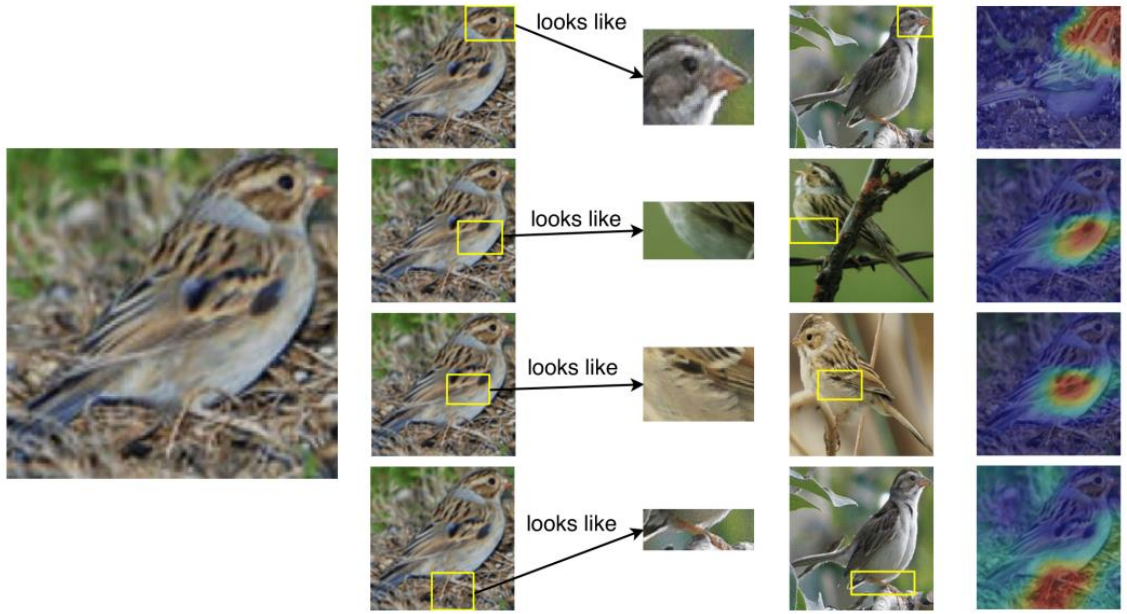


Figure 1. Decision Tree for a 2-layer ReLU Neural Network

ProtoPNet

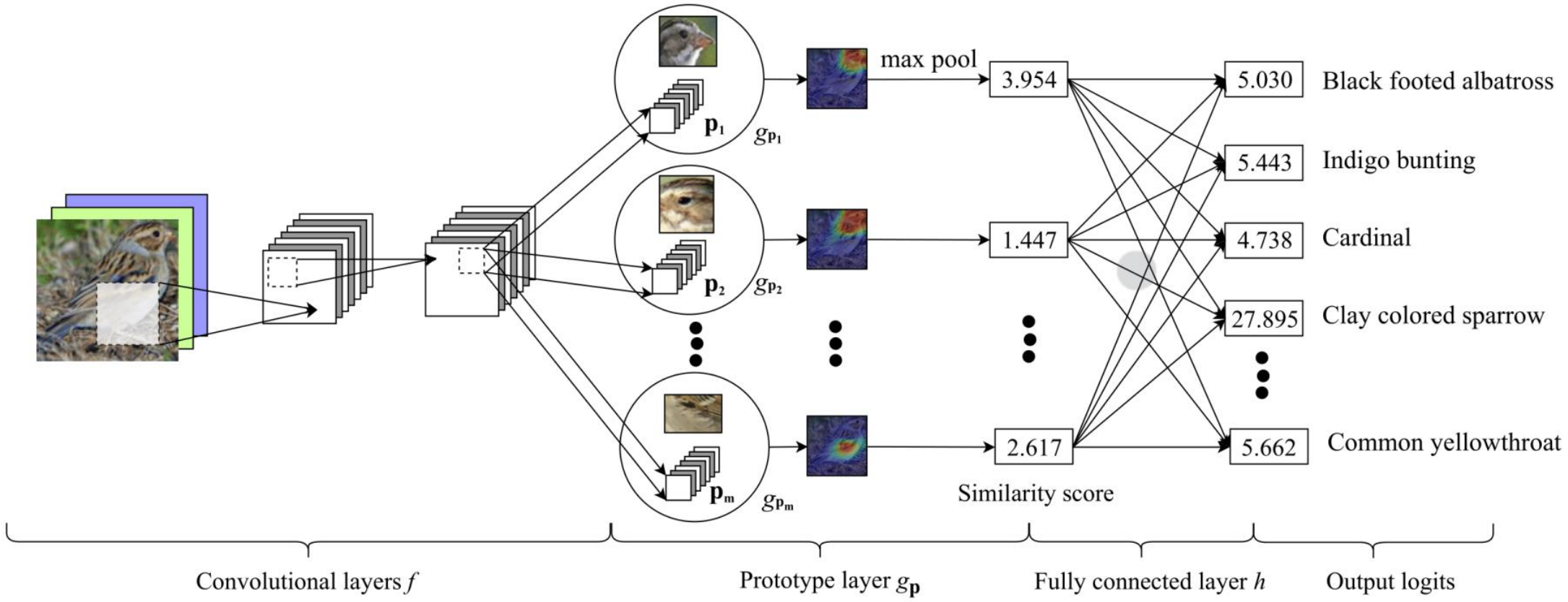


Leftmost: a test image of a clay-colored sparrow
Second column: same test image, each with a bounding box generated by our model -- the content within the bounding box is considered by our model to look similar to the prototypical part (same row, third column) learned by our algorithm
Third column: prototypical parts learned by our algorithm
Fourth column: source images of the prototypical parts in the third column
Rightmost column: activation maps indicating how similar each prototypical part resembles part of the test bird

C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, 'This looks like that: deep learning for interpretable image recognition', in Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA: Curran Associates Inc., 2019, pp. 8930–8941.

ProtoPNet

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

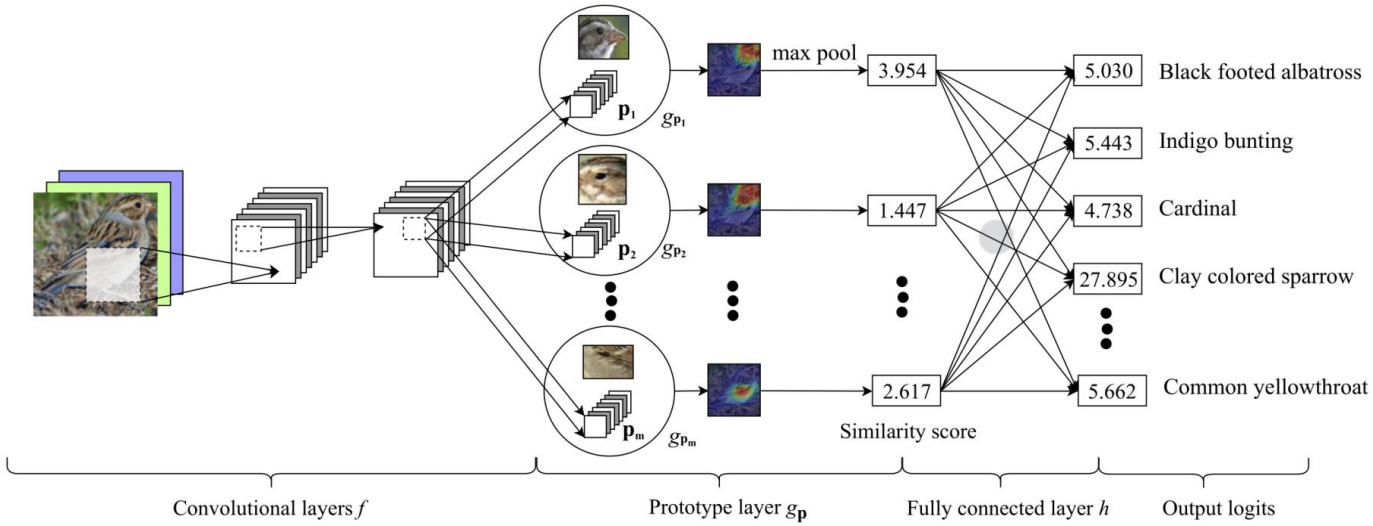


C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, 'This looks like that: deep learning for interpretable image recognition', in Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA: Curran Associates Inc., 2019, pp. 8930–8941.

ProtoPNet

- Main -- Train the prototypes, with CNN fixed, and custom loss function
- Push – to visualize, replace prototype with patches closest to prototype
- Fine-tune the last layer

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



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

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

C. Chen, O. Li, C. Tao, A. J. Barnett, J. Su, and C. Rudin, 'This looks like that: deep learning for interpretable image recognition', in Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA: Curran Associates Inc., 2019, pp. 8930–8941.

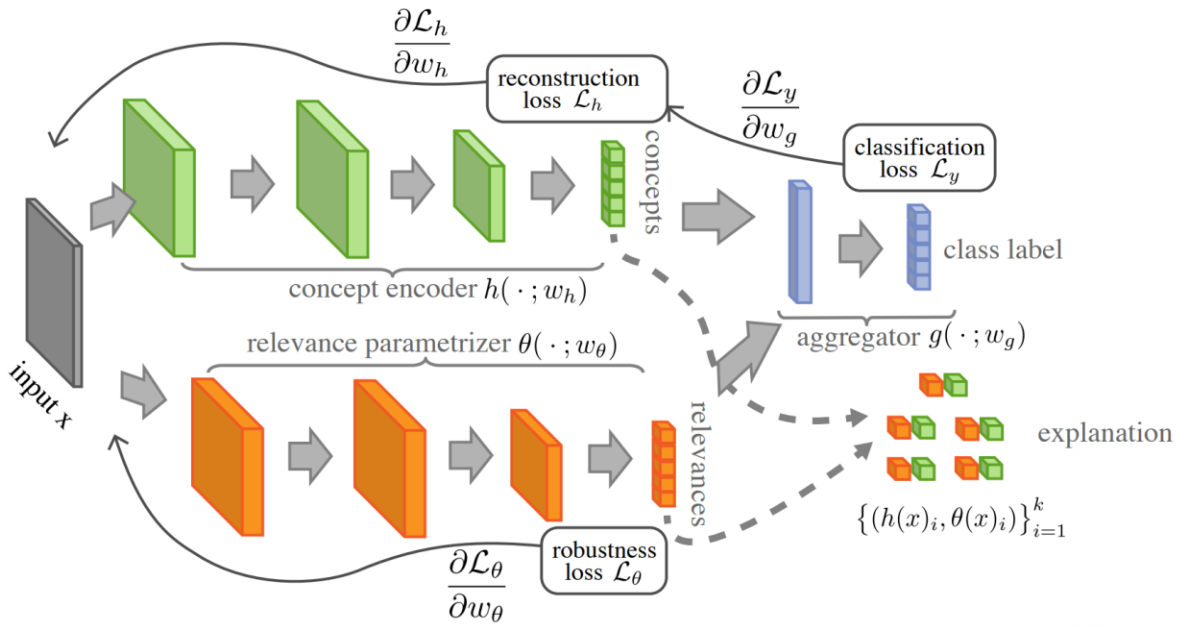
Prost and cons of ProtoPNet

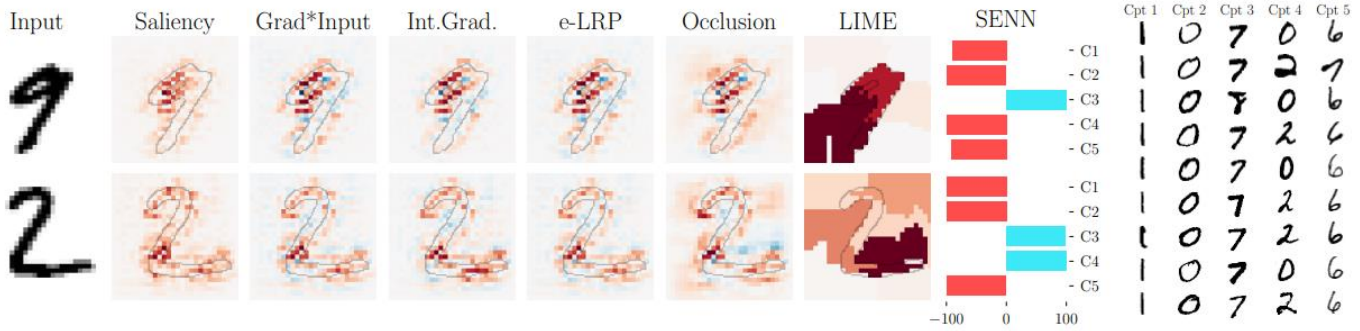
- Pros
  - It is explainable mdel (no post-hoc operations)
  - It is proven to work on fine-grained datsets (all categories are similar, e.g. birds spieces, car models, etc.)
- Cons
  - Training is complex (warm-up, main, push fine-tuning)
  - Not suitable for not fine-grained datasets (ProtoPNet still does not work on ImageNet)
  - Large number of prototypes
  - Classes cannot share prototypes

Self-explainable Neural Netowrks (SENN)

- The authors use an idea of a concept (e.g. prototype)
- They introduce relevance component that ensures that the input can be interpreted through concepts and relevance scores
- The aggregater is an additive, interpretable funcion (e.g. linear)

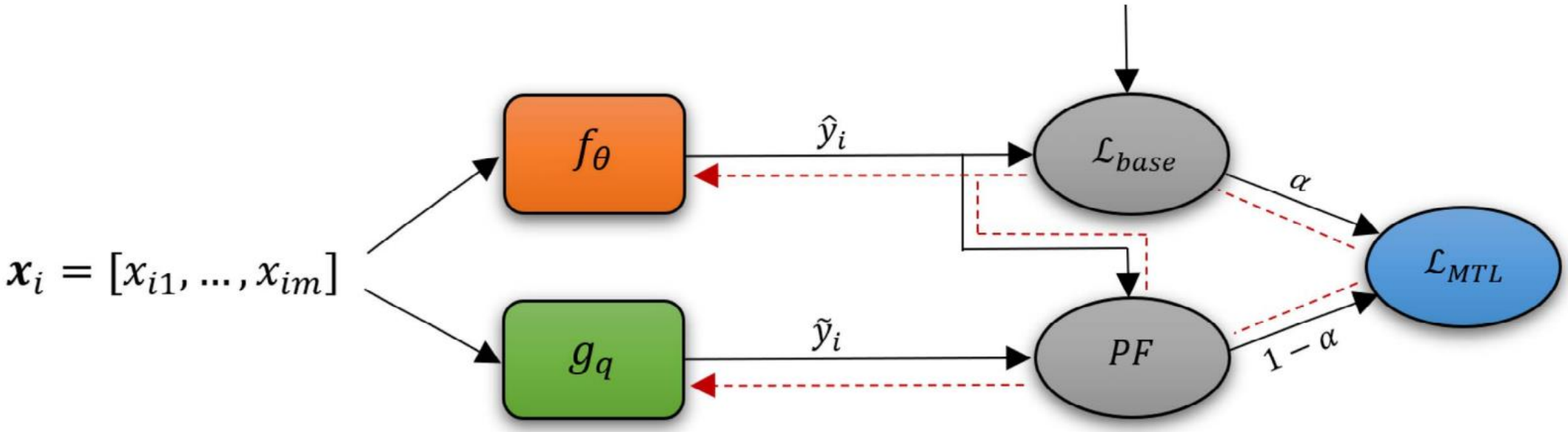
$$\mathcal{L}_y(f(x), y) + \lambda \mathcal{L}_\theta(f) + \xi \mathcal{L}_h(x, \hat{x})$$



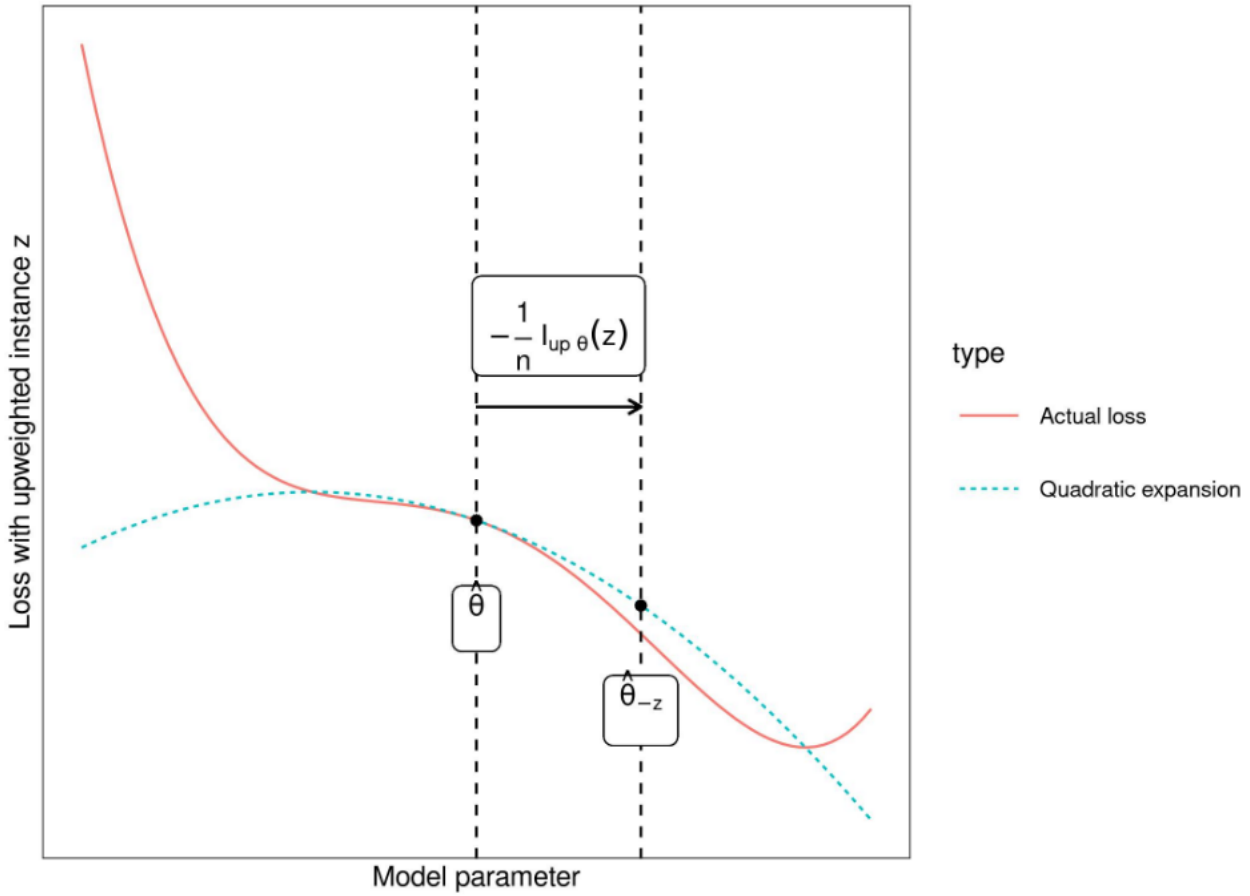


D. Alvarez-Melis and T. S. Jaakkola, 'Towards robust interpretability with self-explaining neural networks', in Proceedings of the 32nd International Conference on Neural Information Processing Systems, in NIPS'18. Red Hook, NY, USA: Curran Associates Inc., Dec. 2018, pp. 7786–7795.

Multi-Task Learning for Explainability



Influential instances



- Updating the model parameter (x-axis) by forming a quadratic expansion of the loss around the current model parameter, and moving 1/n into the direction in which the loss with upweighted instance z (y-axis) improves most.
- This upweighting of instance z in the loss approximates the parameter changes if we delete z and train the model on the reduced data.

AIRA Seminar (this Thursday)

2024-01-11

Speaker: Arkadiusz Tomczyk, Assistant Professor @ Lodz University of Technology

Title: Interpretable components and graph neural networks

Abstract: In the presentation graph neural networks will be discussed. They will be compared both to classic and deep learning techniques (including convolutional neural networks and transformers). Their possible areas of applications will be illustrated by prediction of chemical molecules' properties and structured image analysis. In both cases the explainability aspects will be emphasized. In particular, when it comes to images, it will be argued that proper representation of their content with interpretable components may lead to additional benefits (better communication with domain experts).

Biogram: Arkadiusz Tomczyk received the MSc degree in computer science in 2002 and the PhD with honours in computer science in 2011 from the Faculty of Technical Physics, Information Technology and Applied Mathematics of the Lodz University of Technology, Poland. Since 2002 he has been employed in the Institute of Information Technology of the Lodz University of Technology. His research experience covers image processing and analysis, especially active contour methods, as well as pattern recognition and machine learning techniques. From 2013 to 2017 he was a principal investigator in research grant focused on Cognitive Hierarchical Active Partitions, a method combining active contour approach with structural representation of image content. This project was supported by National Science Centre, project no. 2012/05/D/ST6/03091. Currently he actively participates in projects supported by National Centre for Research and Development and his scientific interests focus on machine learning techniques (convolution neural networks, transformers, graph neural networks) applied to analysis of images and graphs. He is an author and co-author of around 50 journal papers, book chapters and conference contributions.

Thank you for your attention!

JAGIELLONIAN UNIVERSITY
IN KRAKÓW

https://geist.re