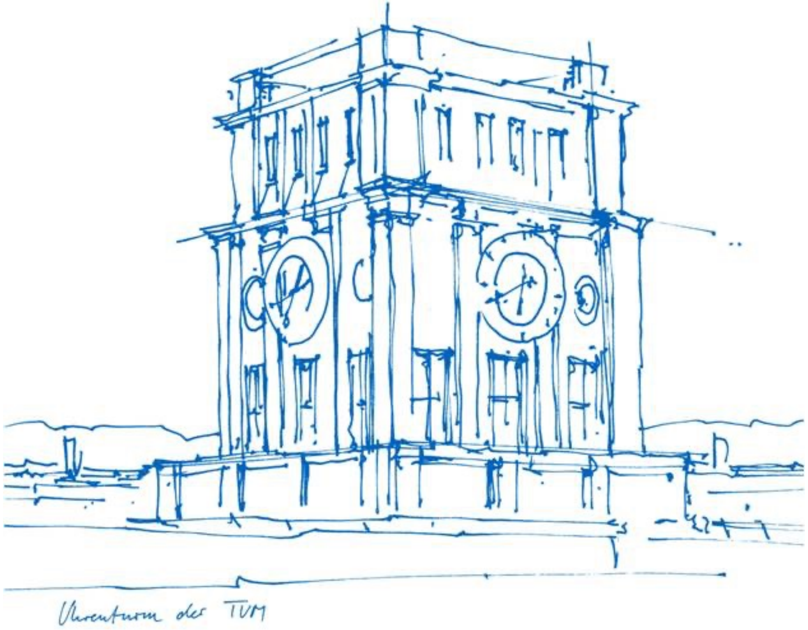
Towards Interpretable Neural Networks for Differential Dementia Diagnosis

Tom Nuno Wolf

TUM Klinikum - Technische Universität München

London, 4 April 2024



Overview

1. Alzheimer‘s Disease
2. Explainability and Related Work
3. PANIC
4. ProtoPFaith
5. Summary

Overview

1. Alzheimer‘s Disease
2. Explainability and Related Work
3. PANIC
4. ProtoPFaith
5. Summary

1. Alzheimer‘s Disease

- Neurodegenerative Disease and most common form of dementia (60-80%)
- Symptoms:
  - Loss of memory
  - Disorientation
  - Mood and behavior changes
  - Difficulty to speak, swallow and walk

[1] Jack, C.R. et al.: The Alzheimer‘s Disease Neuroimaging Initiative (ADNI): MRI Methods. J Magn Reson Imaging, 27(4) (2008)

1. Alzheimer‘s Disease

- Neurodegenerative Disease and most common form of dementia (60-80%)
- Symptoms:
  - Loss of memory
  - Disorientation
  - Mood and behavior changes
  - Difficulty to speak, swallow and walk

- Two major suspected proteins:
  - Plages: between nerve cells (beta-amyloid)
  - Tangles: twisted fibers within cells (tau)

[1] Jack, C.R. et al.: The Alzheimer‘s Disease Neuroimaging Initiative (ADNI): MRI Methods. J Magn Reson Imaging, 27(4) (2008)

1. Alzheimer‘s Disease

- ~150 million affected in 2050
- Disease progression relatively unknown
- Studies like ADNI [1] collect data:

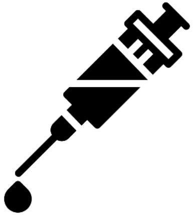
[1] Jack, C.R. et al.: The Alzheimer‘s Disease Neuroimaging Initiative (ADNI): MRI Methods. J Magn Reson Imaging, 27(4) (2008)

1. Alzheimer‘s Disease

- ~150 million affected in 2050
- Disease progression relatively unknown
- Studies like ADNI [1] collect data:

PET Imaging

APOE4

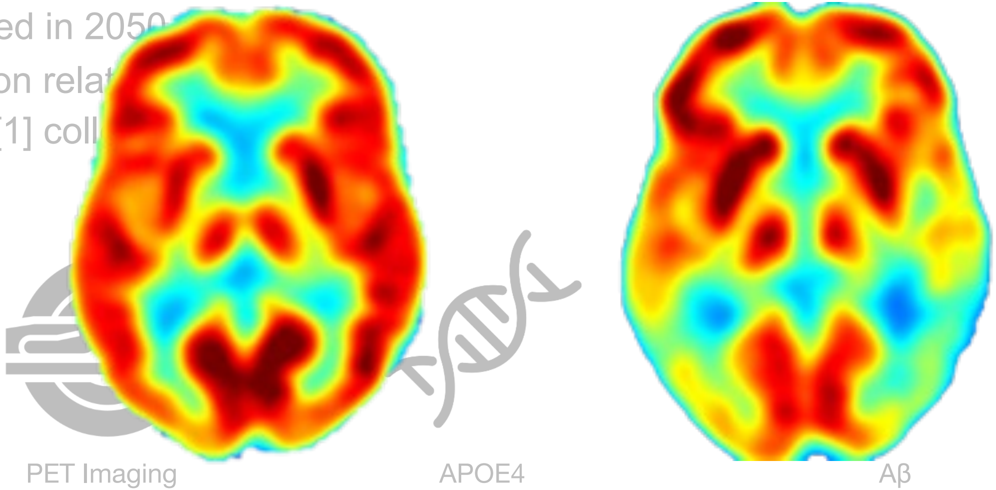


Aβ

[1] Jack, C.R. et al.: The Alzheimer‘s Disease Neuroimaging Initiative (ADNI): MRI Methods. J Magn Reson Imaging, 27(4) (2008)
[DNA by Stock Image Folio | Laboratory Sample by Ben Davis] from the Noun Project

1. Alzheimer‘s Disease

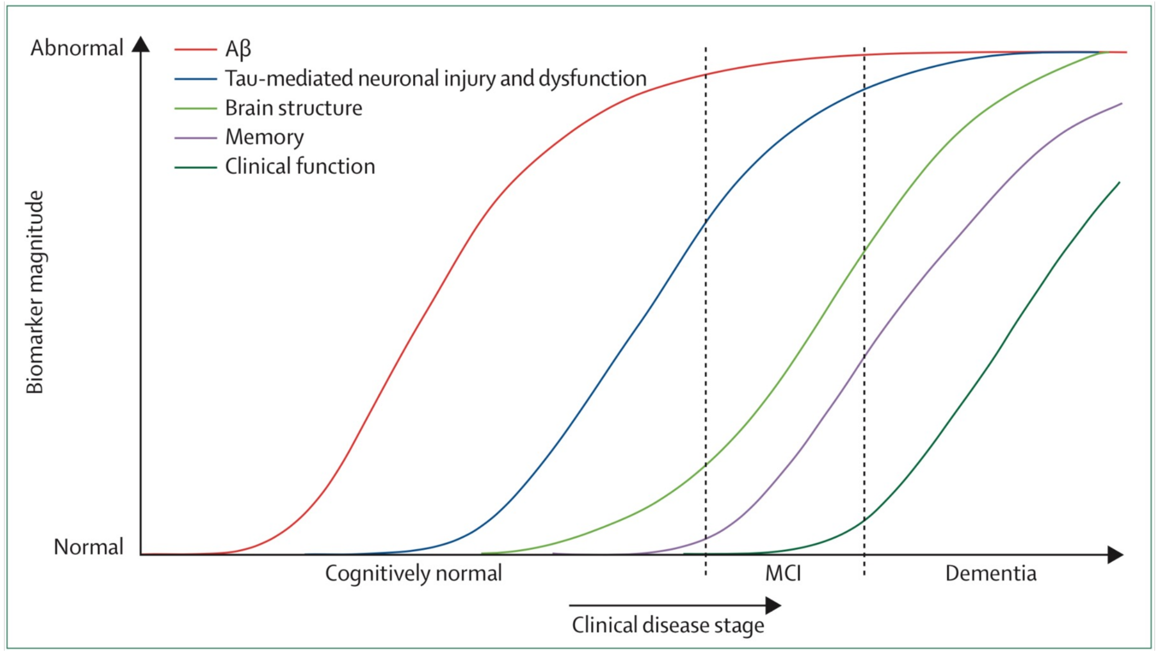
- ~150 million affected in 2050
- Disease progression relat
- Studies like ADNI [1] coll



[1] Jack, C.R. et al.: The Alzheimer‘s Disease Neuroimaging Initiative (ADNI): MRI Methods. J Magn Reson Imaging, 27(4) (2008)

[DNA by Stock Image Folio | Laboratory Sample by Ben Davis] from the Noun Project

1. Alzheimer‘s Disease



Jack Jr, C. et al. Tracking pathophysiological processes in AD an updated hypothetical model of dynamic biomarkers. The Lancet Neurology, 12(2), 207–216 (2013)

Overview

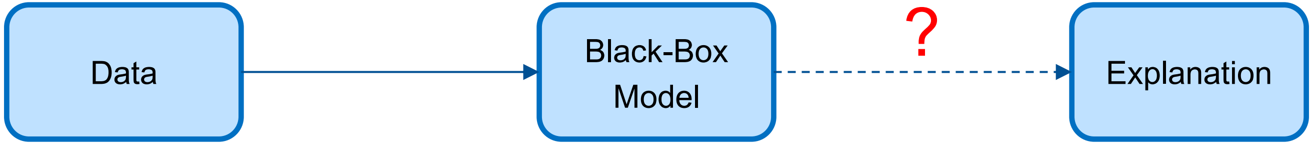
1. Alzheimer‘s Disease
2. Explainability and Related Work
3. PANIC
4. ProtoPFaith
5. Summary

2. Explainability – Introduction

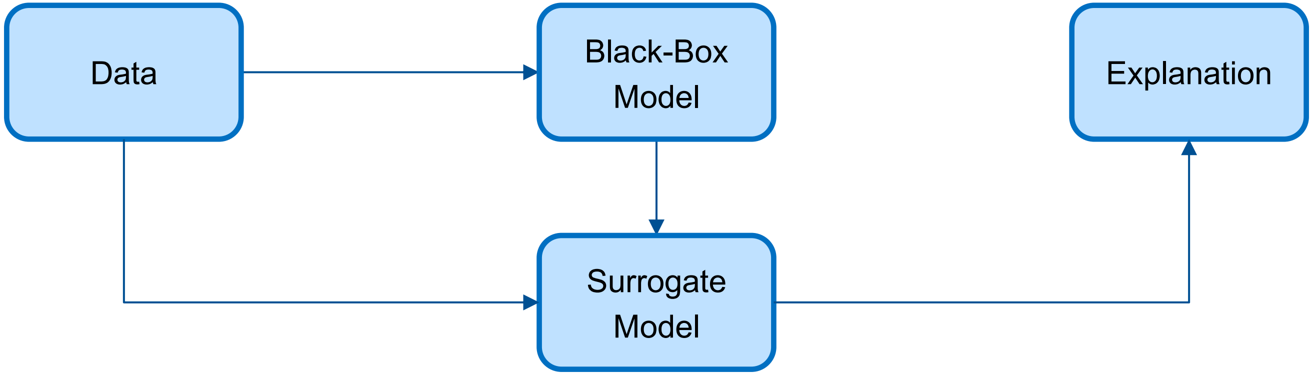
Gradients

Perturbations

2. Explainability – Introduction

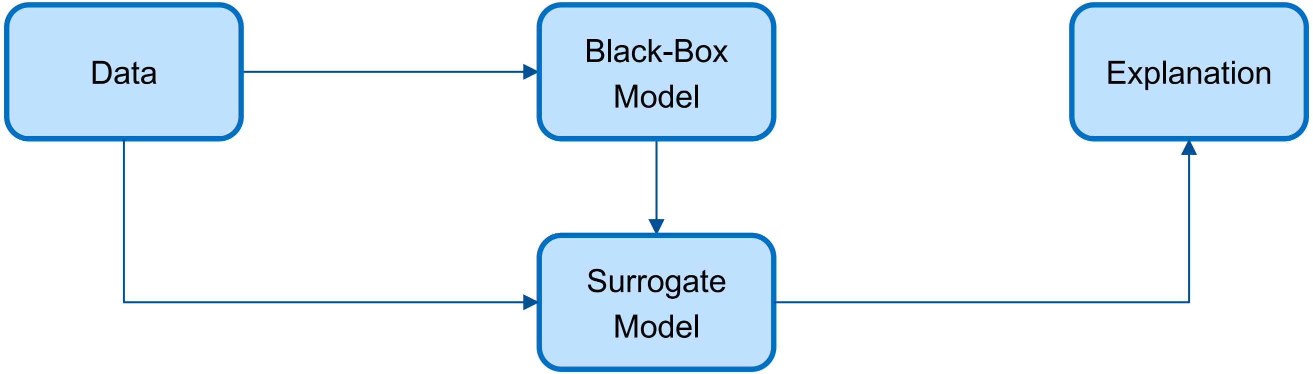


2. Explainability – Introduction



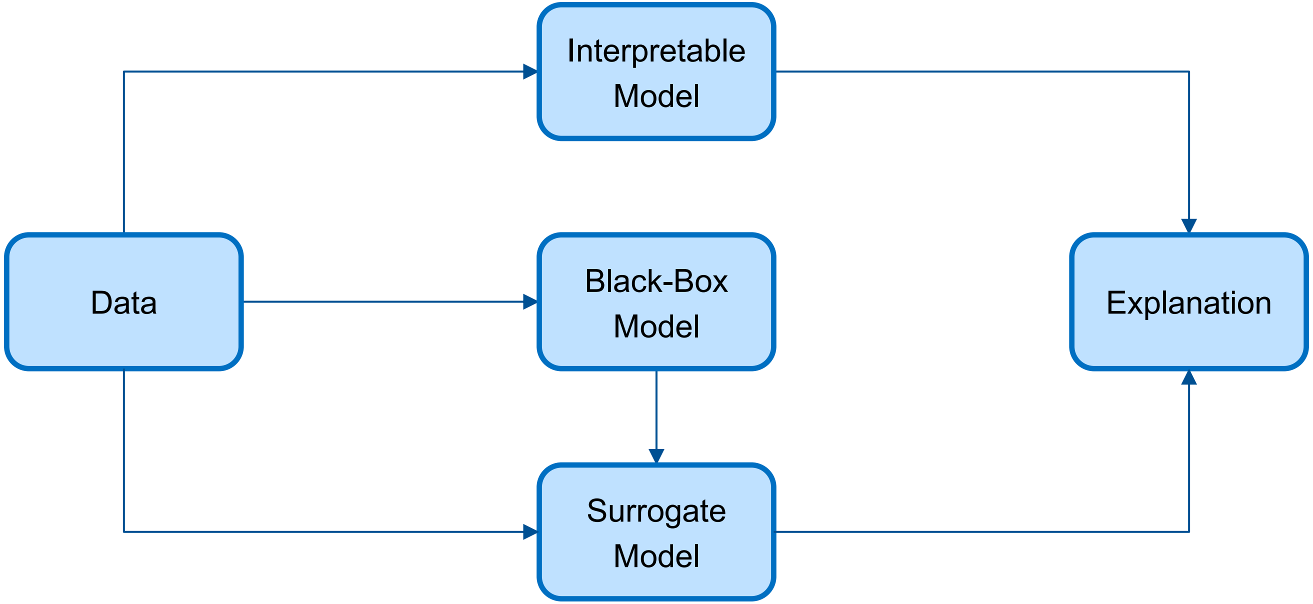
2. Explainability – Introduction

„[Post-Hoc] Explanations must be wrong“ [2]

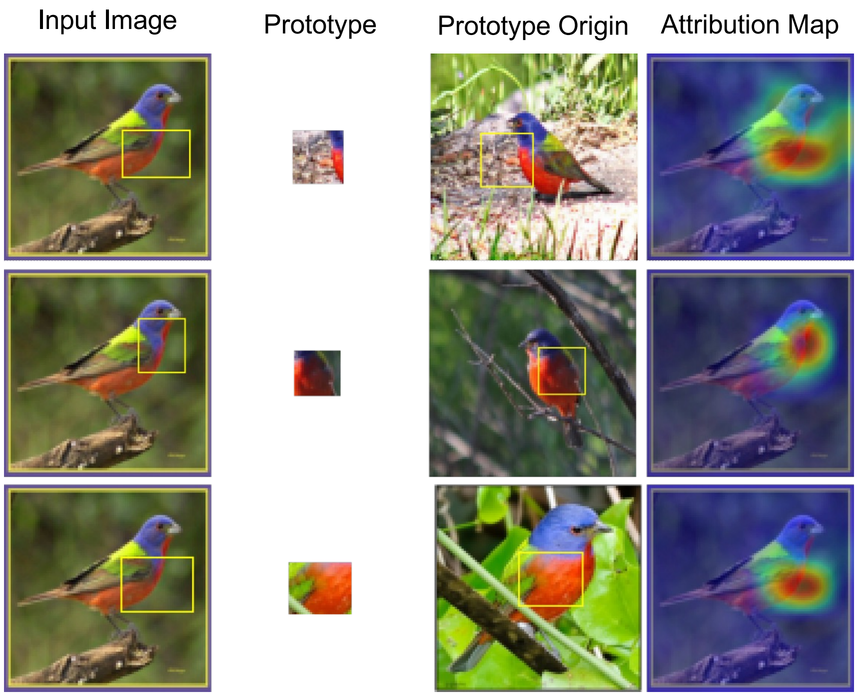


[2] Rudin, C.: Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature Machine Intelligence 1(5), 206-215 (2019)

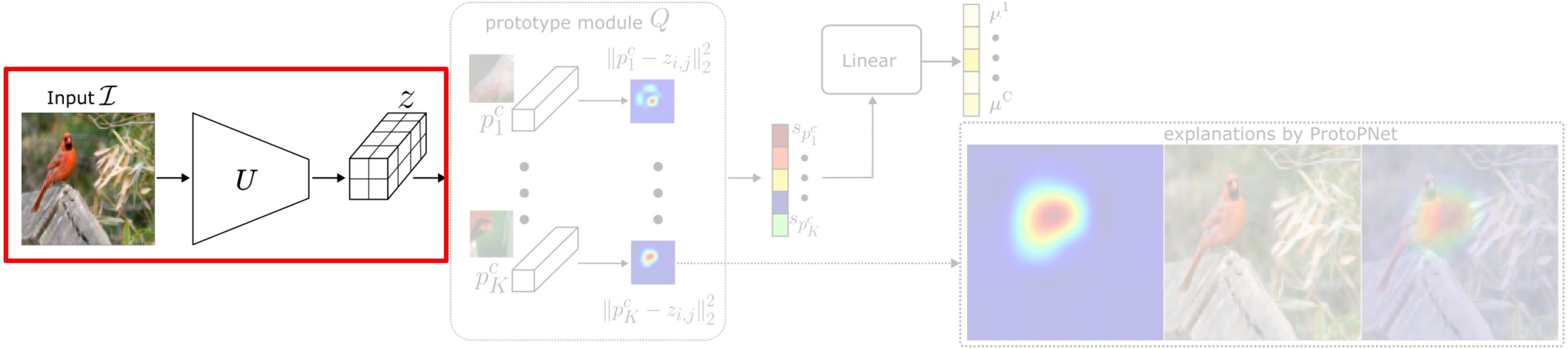
2. Explainability – Introduction



2. Explainability – Case-Based Reasoning

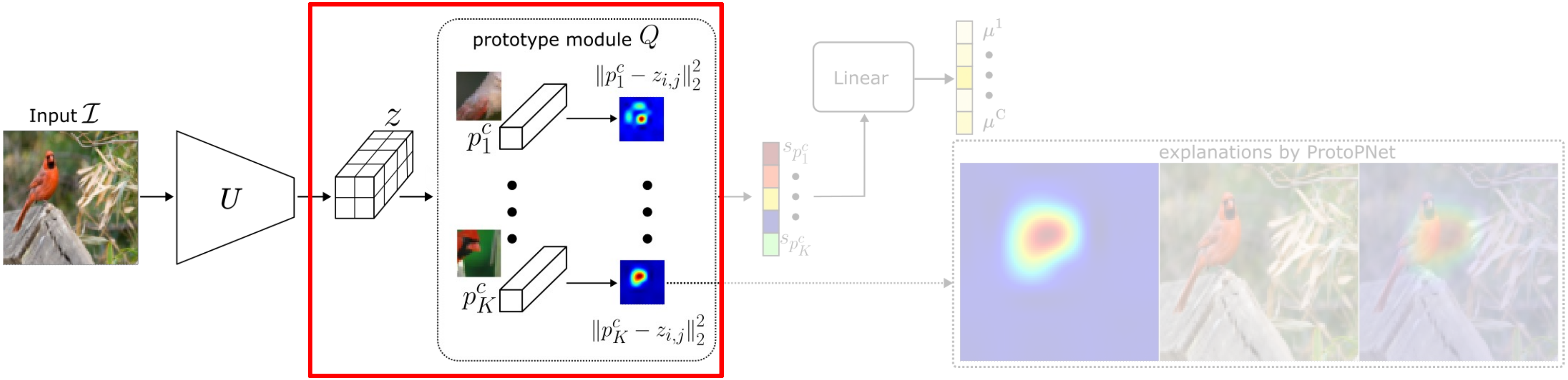


2. Related Work – ProtoPNet [3]



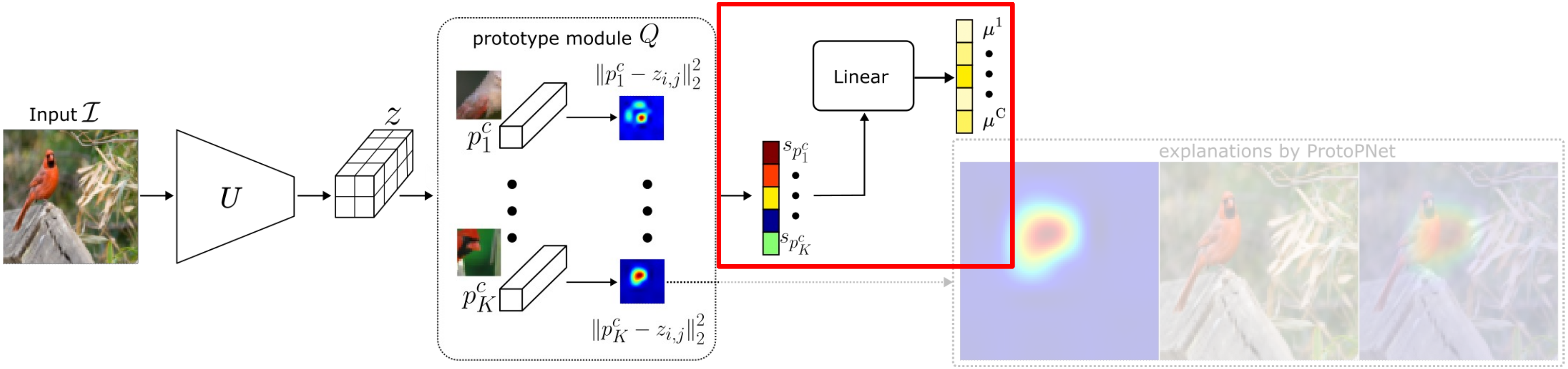
[3] Chen, C. et al.: This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS, vol. 32 (2019)

2. Related Work – ProtoPNet [3]



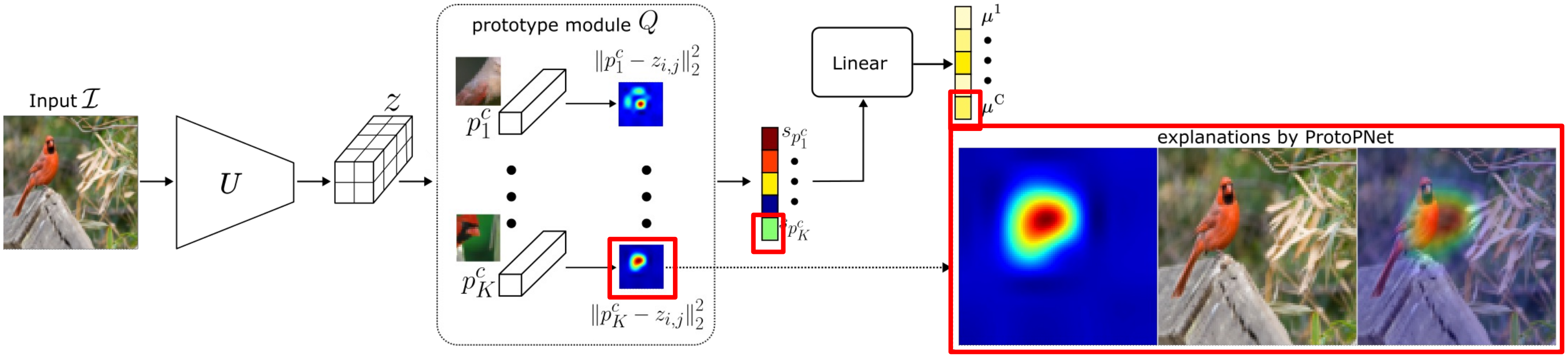
[3] Chen, C. et al.: This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS, vol. 32 (2019)

2. Related Work – ProtoPNet [3]



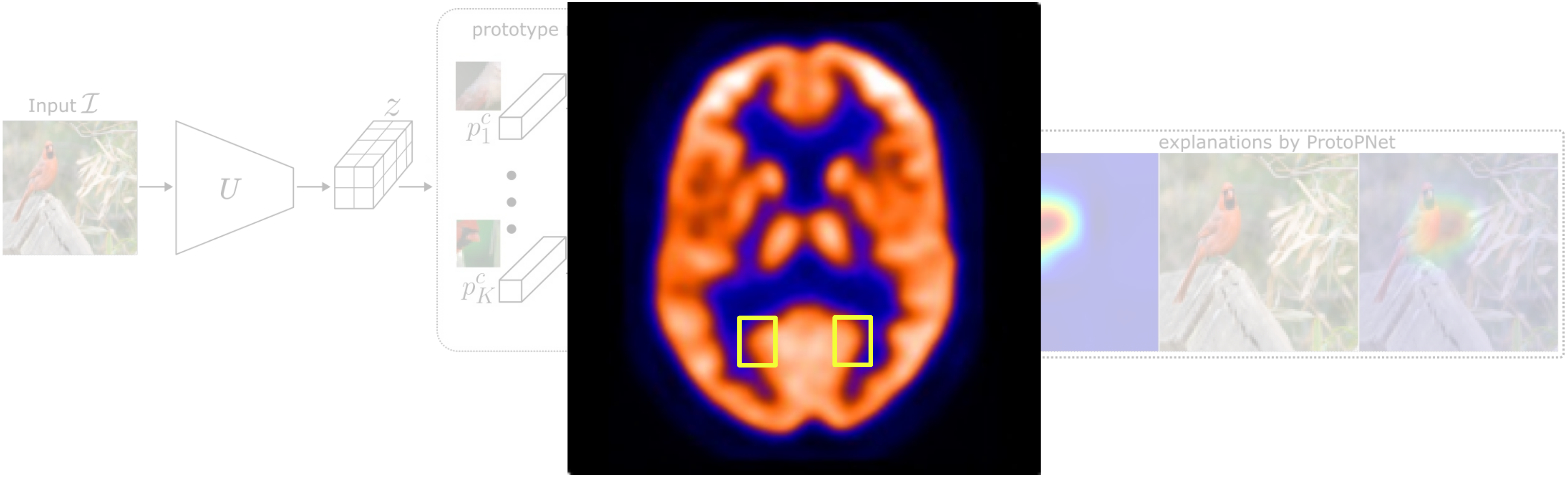
[3] Chen, C. et al.: This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS, vol. 32 (2019)

2. Related Work – ProtoPNet [3]



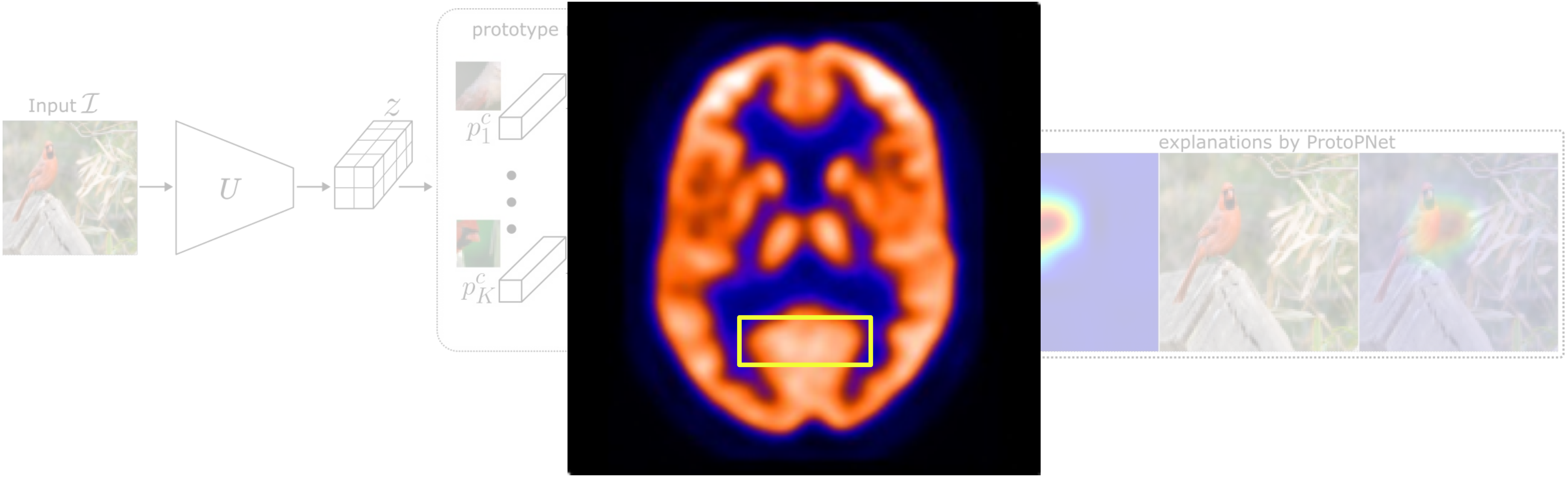
[3] Chen, C. et al.: This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS, vol. 32 (2019)

2. Related Work – ProtoPNet [3]



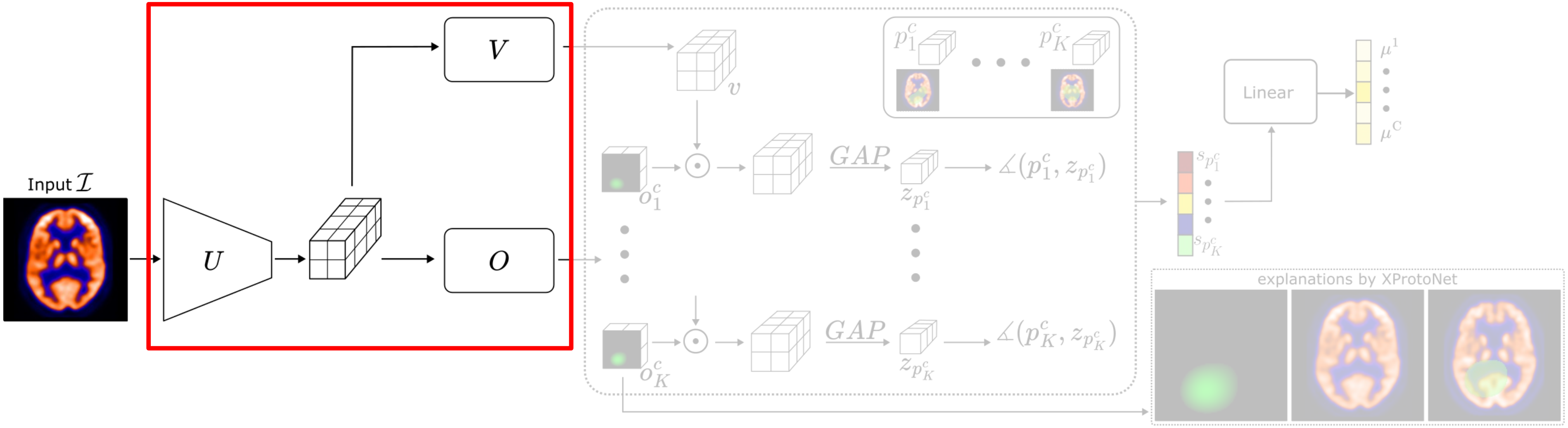
[3] Chen, C. et al.: This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS, vol. 32 (2019)

2. Related Work – ProtoPNet [3]



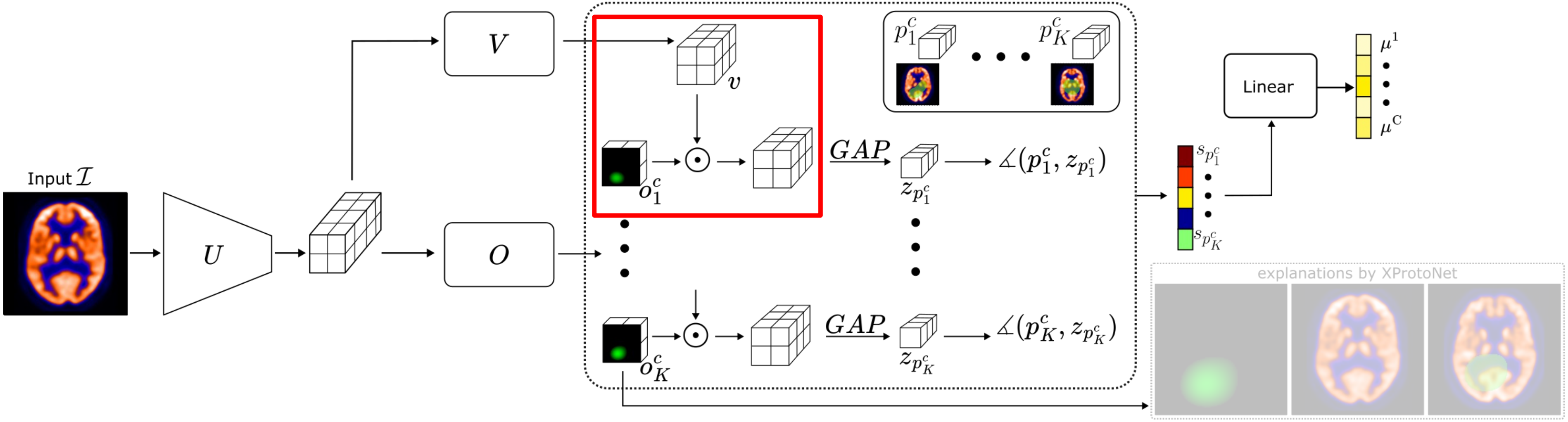
[3] Chen, C. et al.: This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS, vol. 32 (2019)

2. Related Work – XProtoNet [4]



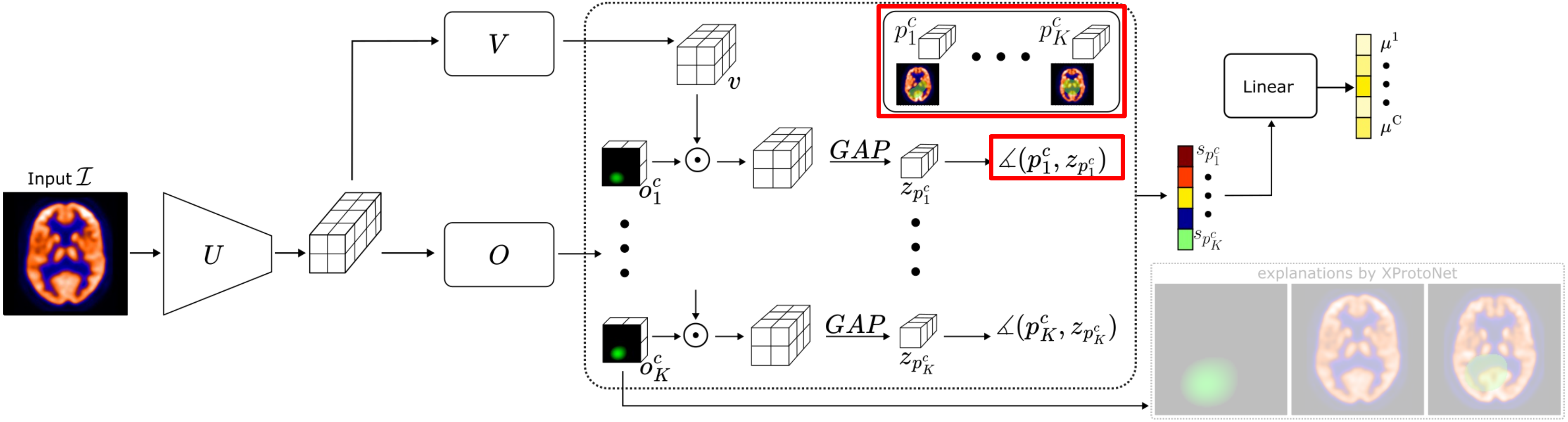
[4] Kim, E. et al.: XProtoNet: Diagnosis in Chest Radiography With Global and Local Explanations. CVPR, pp. 15719-15728 (2021)

2. Related Work – XProtoNet [4]



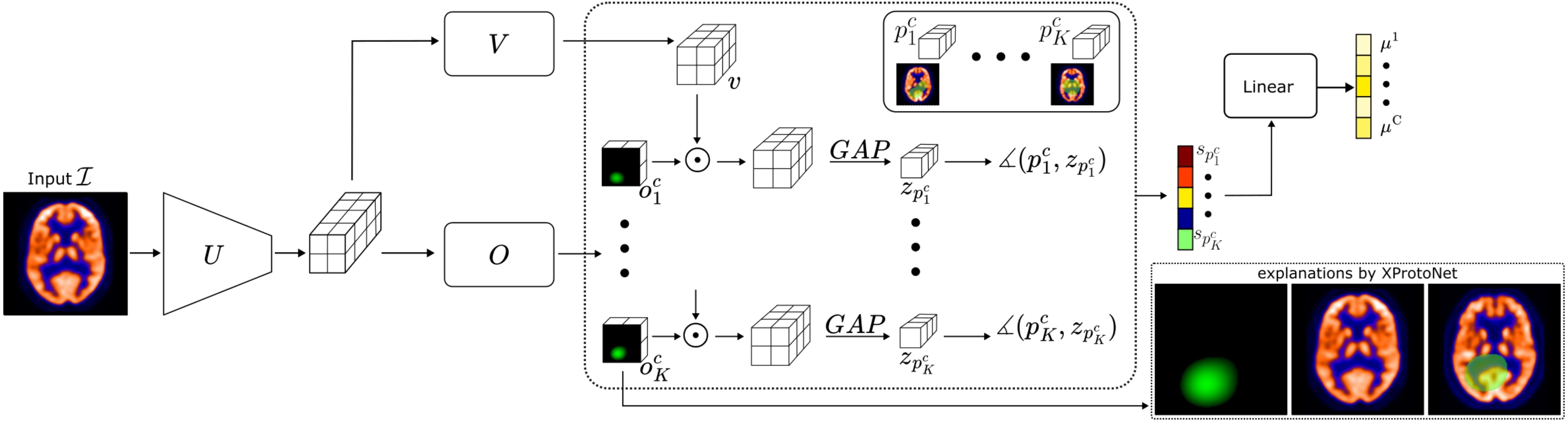
[4] Kim, E. et al.: XProtoNet: Diagnosis in Chest Radiography With Global and Local Explanations. CVPR, pp. 15719-15728 (2021)

2. Related Work – XProtoNet [4]



[4] Kim, E. et al.: XProtoNet: Diagnosis in Chest Radiography With Global and Local Explanations. CVPR, pp. 15719-15728 (2021)

2. Related Work – XProtoNet [4]



[4] Kim, E. et al.: XProtoNet: Diagnosis in Chest Radiography With Global and Local Explanations. CVPR, pp. 15719-15728 (2021)

Overview

3. PANIC

PET Imaging

APOE4

Aβ

Inherently Interpretable Neural Network for Heterogeneous Data

3. PANIC


PET Imaging


APOE4

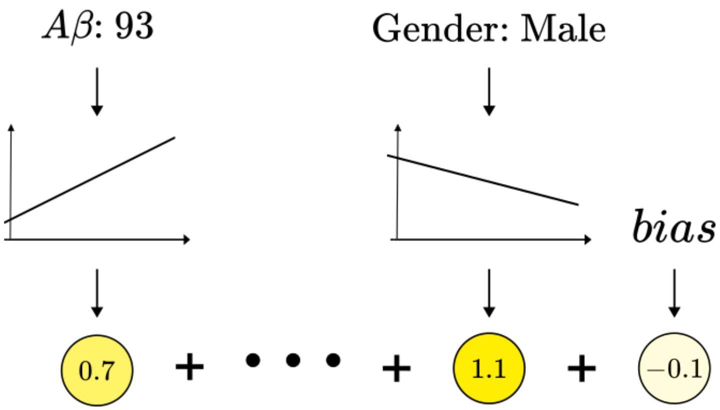

Aβ

Inherently Interpretable Neural Network for Heterogeneous Data

Does not exist!

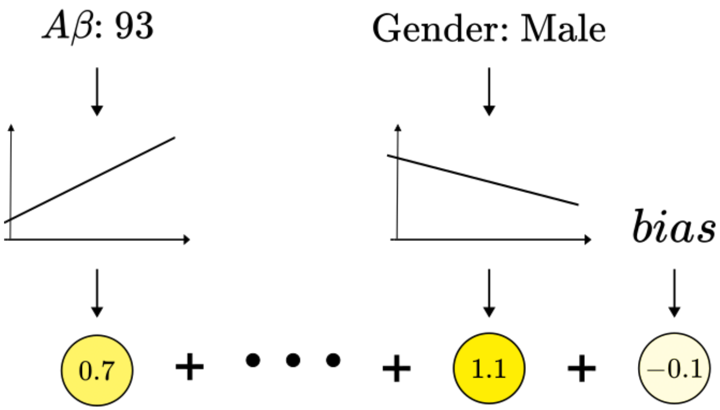
3. PANIC

Generalized Additive Model (GAM):

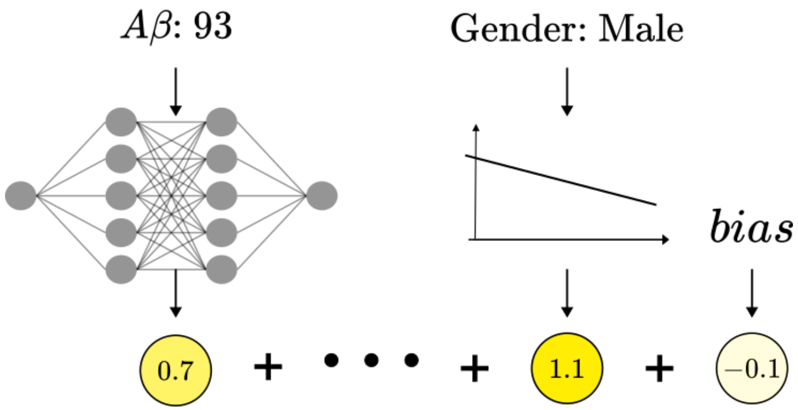


3. PANIC

Generalized Additive Model (GAM):

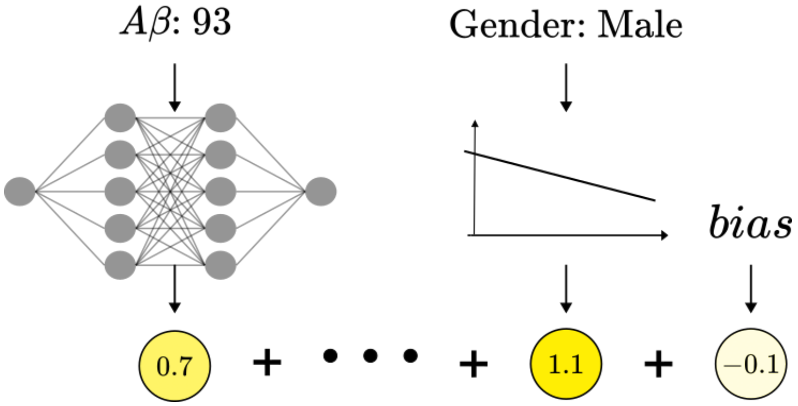


Neural Additive Model (NAM) [5]:



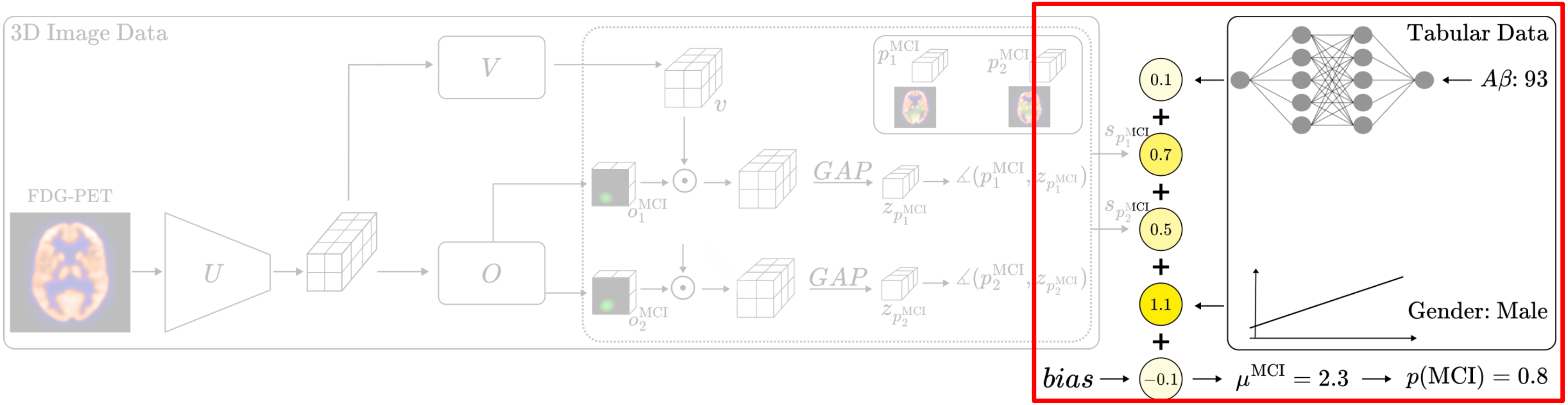
[5] Agerwal, R. et al.: Neural Additive Models: Interpretable Machine Learning with Neural Nets. NeurIPS, vol. 34 (2021)

3. PANIC



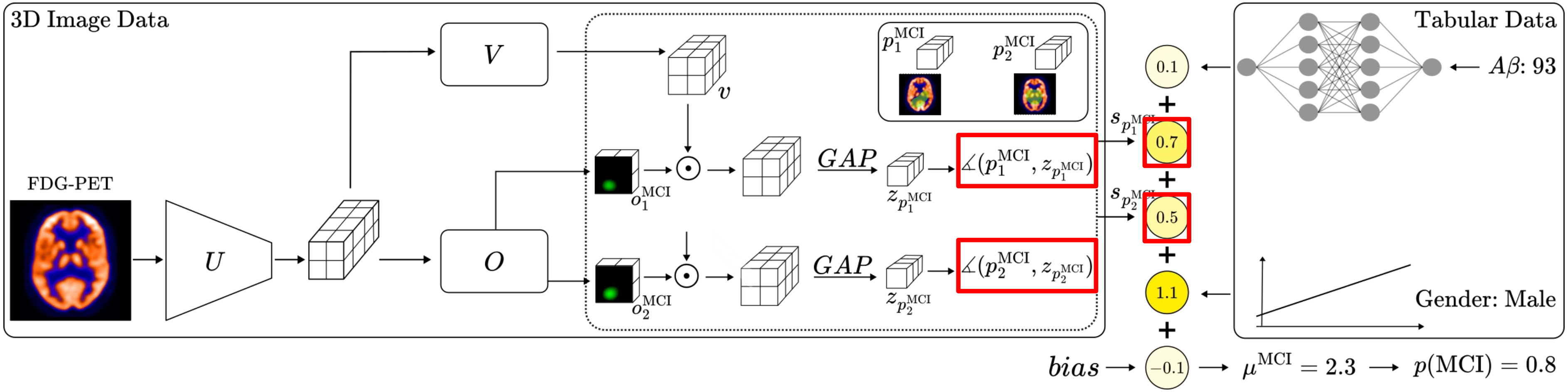
$$f_n^c(x_n) = \begin{cases} s_n^c, & \text{if } x_n \text{ is missing,} \\ \beta_n^c x_n, \ \text{with } \beta_n^c \in \mathbb{R}, & \text{if } x_n \text{ is categorical} \\ \mathrm{MLP}_n^c(x_n), & \text{otherwise.} \end{cases}$$

3. PANIC



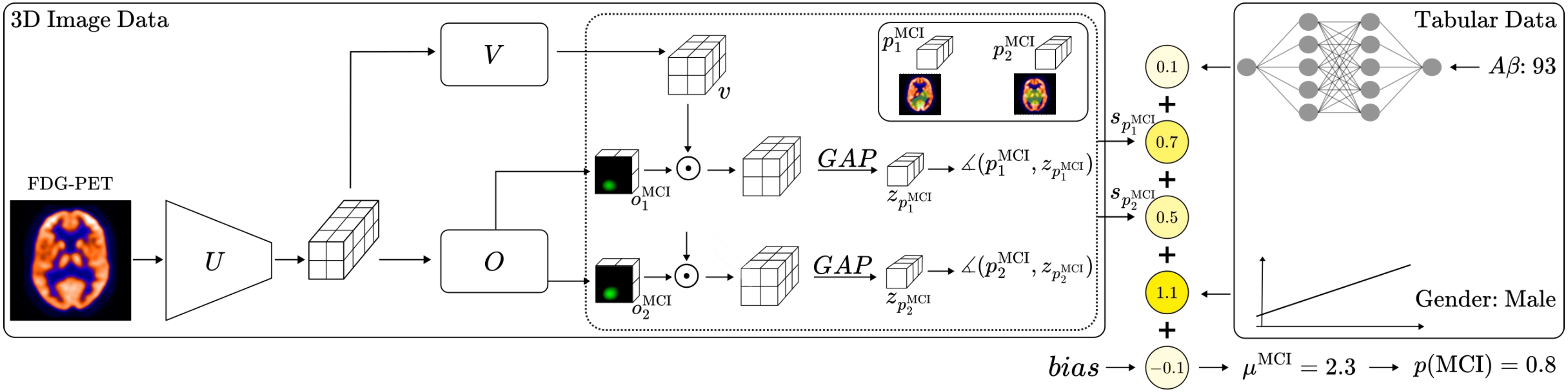
3D Image Data
FDG-PET
U
V
O
v
o_1^{MCI}
o_2^{MCI}
GAP
GAP
$z_{p_1^{\mathrm{MCI}}}$
$z_{p_2^{\mathrm{MCI}}}$
$\measuredangle(p_1^{\mathrm{MCI}}, z_{p_1^{\mathrm{MCI}}})$
$\measuredangle(p_2^{\mathrm{MCI}}, z_{p_2^{\mathrm{MCI}}})$
p_1^{MCI}
p_2^{MCI}
$s_{p_1^{\mathrm{MCI}}}$
$s_{p_2^{\mathrm{MCI}}}$
Tabular Data
$A\beta$: 93
Gender: Male
0.1
0.7
0.5
1.1
−0.1
$bias$
$\mu^{\mathrm{MCI}} = 2.3$
$p(\mathrm{MCI}) = 0.8$

3. PANIC



3D Image Data
FDG-PET
U
V
O
v
p_1^{MCI}
p_2^{MCI}
o_1^{MCI}
o_2^{MCI}
GAP
GAP
$z_{p_1^{MCI}}$
$z_{p_2^{MCI}}$
$\measuredangle(p_1^{MCI}, z_{p_1^{MCI}})$
$\measuredangle(p_2^{MCI}, z_{p_2^{MCI}})$
$s_{p_1^{MCI}}$
$s_{p_2^{MCI}}$
0.1
0.7
0.5
1.1
−0.1
Tabular Data
$A\beta$: 93
Gender: Male
$bias$
$\mu^{MCI} = 2.3 \longrightarrow p(\mathrm{MCI}) = 0.8$

3. PANIC



$$\mathcal{L}(y, x_1, \ldots, x_n, \mathcal{I}) = \mathcal{L}_{\text{CE}}(y, \hat{y}) + \lambda_1 \mathcal{L}_{\text{Tab}}(x_1, \ldots, x_n) + \lambda_2 \mathcal{L}_{\text{clst}}(\mathcal{I}) + \lambda_3 \mathcal{L}_{\text{sep}}(\mathcal{I}) + \lambda_4 \mathcal{L}_{\text{occ}}(\mathcal{I}) + \lambda_5 \mathcal{L}_{\text{affine}}(\mathcal{I})$$

3. PANIC: Results – Data and Performance

Evaluation:

- 1245 baseline samples of ADNI [1]
- 5-fold Cross-Validation, stratified by age, sex, and labels
- Evaluated on Balanced Accuracy (BAcc) mean and standard deviation (SD)

[1] Jack, C.R. et al.: The Alzheimer‘s Disease Neuroimaging Initiative (ADNI): MRI Methods. J Magn Reson Imaging, 27(4) (2008)

3. PANIC: Results – Data and Performance

Evaluation:

- 1245 baseline samples of ADNI [1]
- 5-fold Cross-Validation, stratified by age, sex, and labels
- Evaluated on Balanced Accuracy (BAcc) mean and standard deviation (SD)

Dataset

Labels	CN	379 (30.4%)
	MCI	610 (49.0%)
	AD	256 (20.6%)

Performance

PANIC	BAcc (SD)	60.7% (4.4%)
DAFT [6]	BAcc (SD)	56.2% (4.5%)

[1] Jack, C.R. et al.: The Alzheimer‘s Disease Neuroimaging Initiative (ADNI): MRI Methods. J Magn Reson Imaging, 27(4) (2008)

[6] Wolf, T.N. et al.: DAFT: A Universal Module to Interweave Tabular und 3D Images in CNNs. NeuroImage, p. 119505 (2022)

3. PANIC: Tabular Data

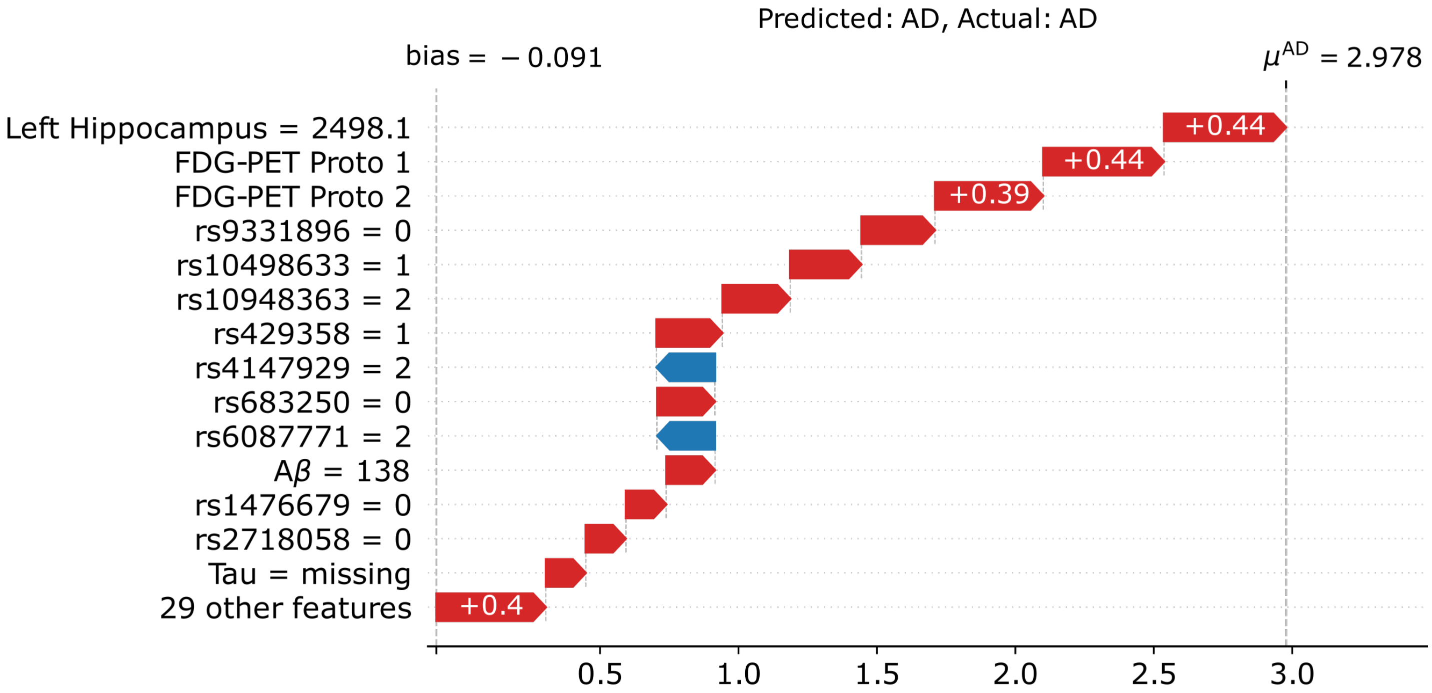
Continuous features:

- Age
- Education
- cerobrospinal fluid markers Aβ, Tau, p- Tau
- MRI-derived volumes of left/right hippocampus and thickness of left/right entorhinal cortex

Categorical features:

- Gender
- 31 AD-related genetic variants

3. PANIC: Results - Local Interpretability



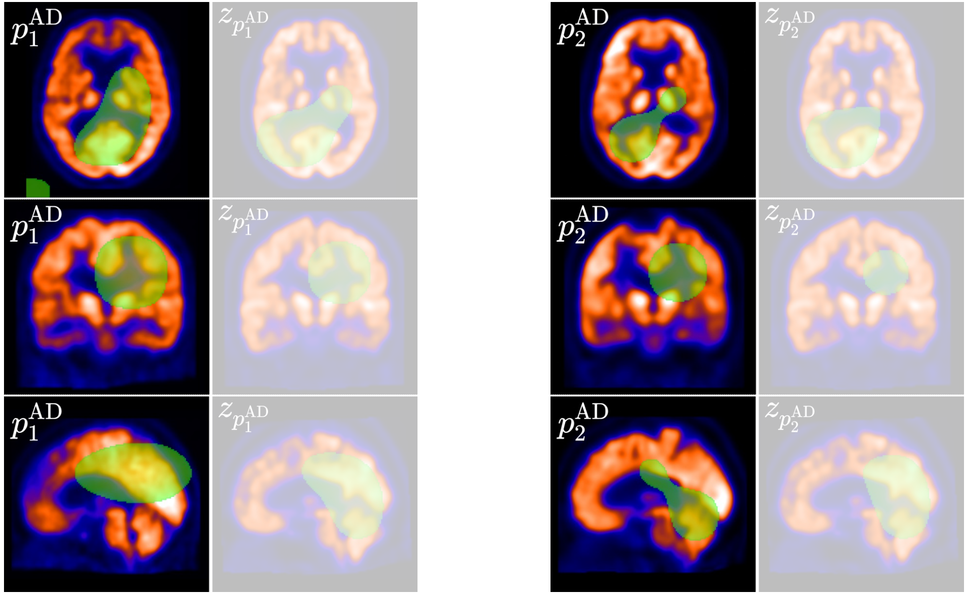
3. PANIC: Results - Local Interpretability

Self Occurrence Prototype 1 | Predicted Occurrence Prototype 1 | Self Occurrence Prototype 2 | Predicted Occurrence Prototype 2

Axial

Coronal

Sagittal



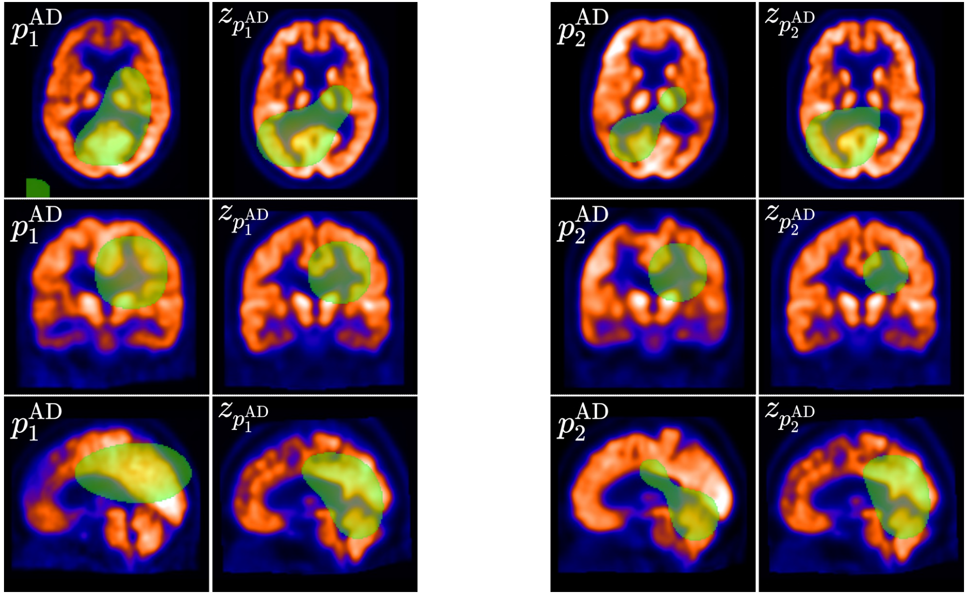
3. PANIC: Results - Local Interpretability

Self Occurrence Prototype 1 | Predicted Occurrence Prototype 1 | Self Occurrence Prototype 2 | Predicted Occurrence Prototype 2

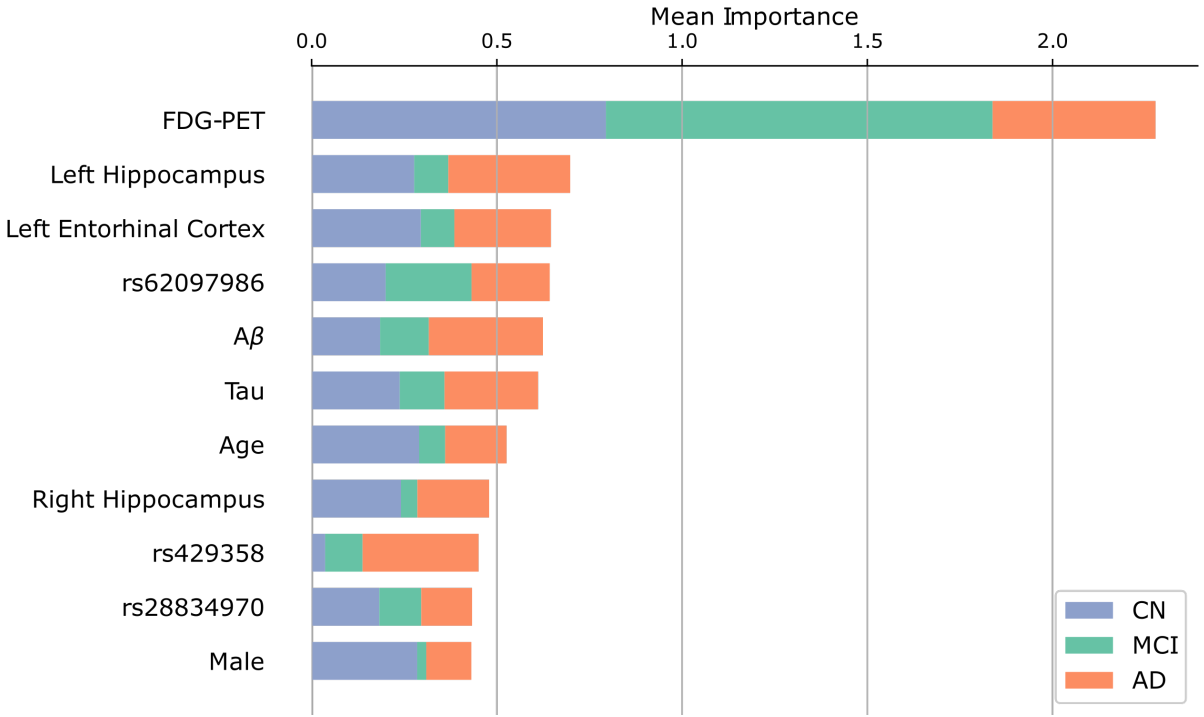
Axial

Coronal

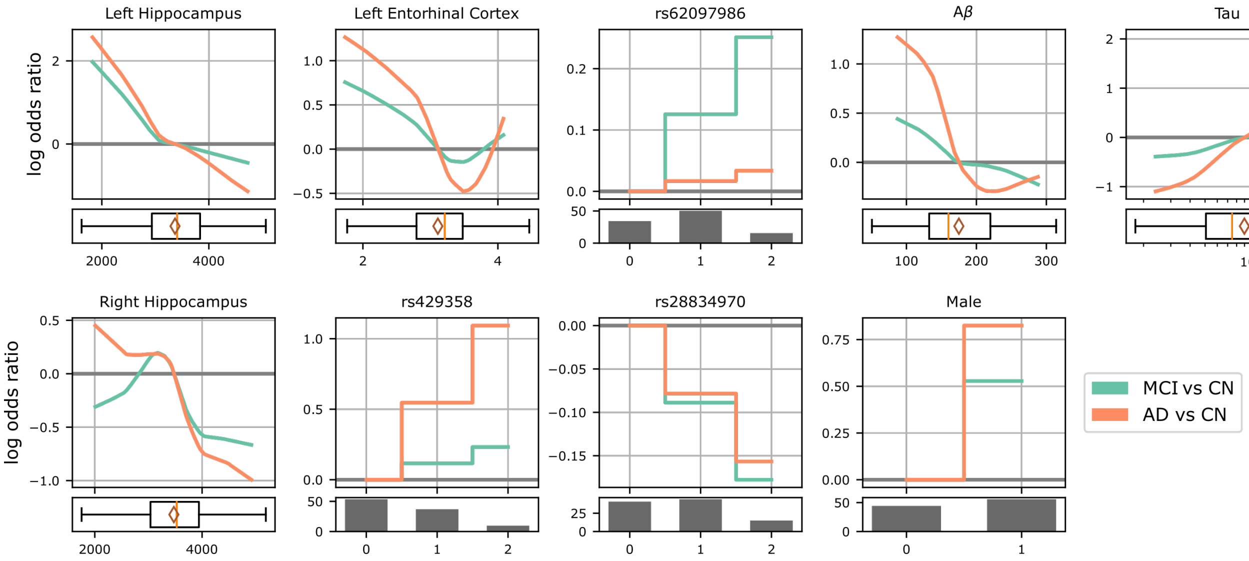
Sagittal



3. PANIC: Results - Global Interpretability

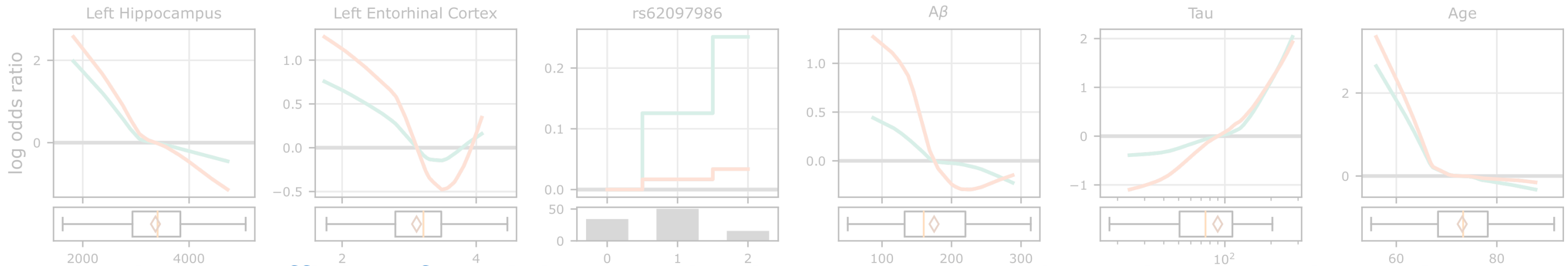


3. PANIC: Results - Global Interpretability

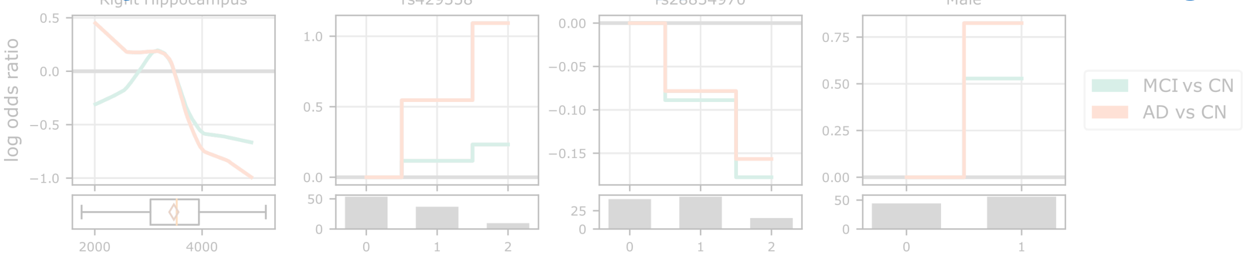


$$\log\left[\frac{p(c\,|\,x_1,\ldots,x_n,\ldots,x_N,\mathcal{I})}{p(\mathrm{CN}\,|\,x_1,\ldots,x_n,\ldots,x_N,\mathcal{I})}\Big/\frac{p(c\,|,x_1,\ldots,x'_n,\ldots,x_N,\mathcal{I})}{p(\mathrm{CN}\,|\,x_1,\ldots,x'_n,\ldots,x_N,\mathcal{I})}\right]$$

3. PANIC: Results - Global Interpretability



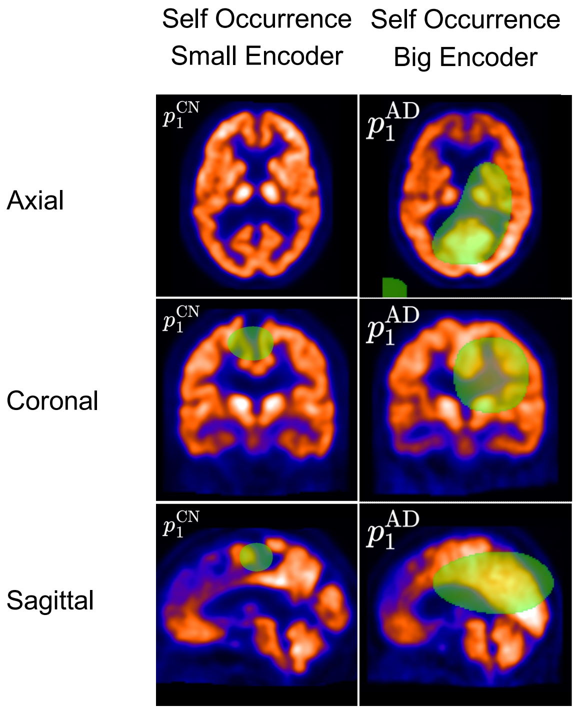
Accuracy is not sufficient for clinical application. Instead, clinical alignment determines use and trust [7]



[7] Tonekaboni, S. et al: What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. Machine Learning for Healthcare Conference, PMLR p.359-380 (2019).

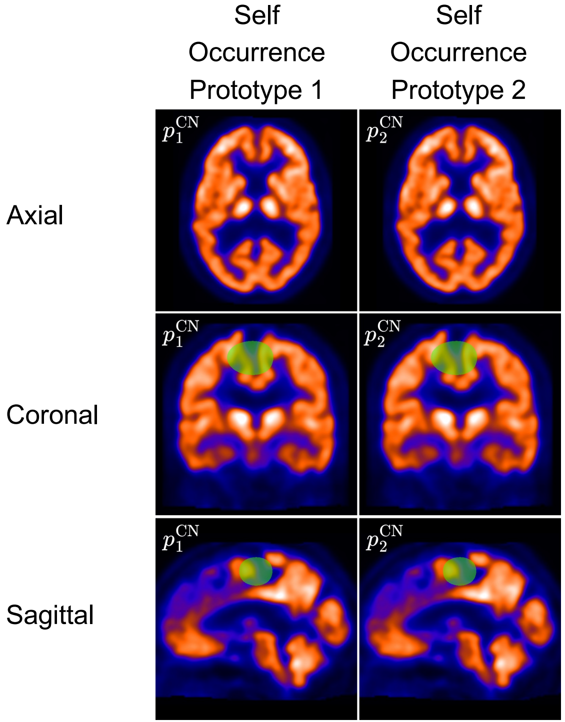
3. PANIC: Discussion - Limitiations

- Granularity of explanations dependent on encoder



3. PANIC: Discussion - Limitiations

- Granularity of explanations dependent on encoder
- Collapse of prototypes of a single class



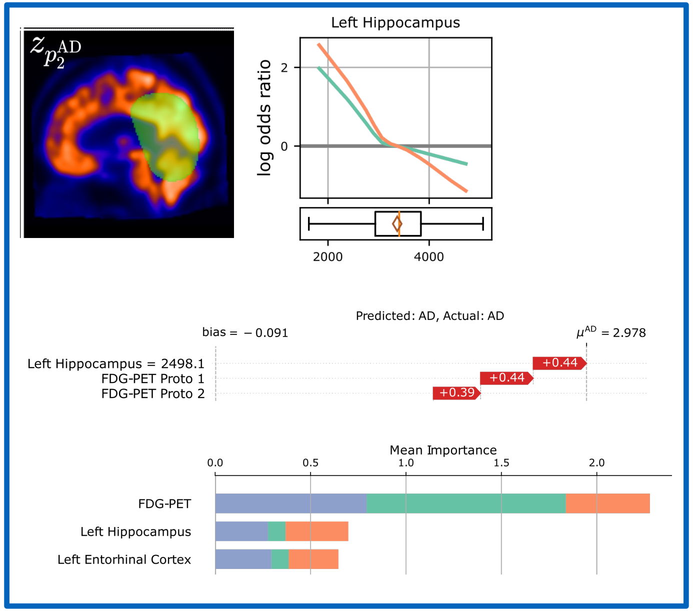
4. Discussion - Limitiations

- Granularity of explanations dependent on encoder
- Collapse of prototypes of a single class
- Convergence

$$\mathcal{L}(y, x_1, \ldots, x_n, \mathcal{I}) = \mathcal{L}_{\mathrm{CE}}(y, \hat{y}) + \lambda_1 \mathcal{L}_{\mathrm{Tab}}(x_1, \ldots, x_n) + \lambda_2 \mathcal{L}_{\mathrm{clst}}(\mathcal{I}) + \lambda_3 \mathcal{L}_{\mathrm{sep}}(\mathcal{I}) + \lambda_4 \mathcal{L}_{\mathrm{occ}}(\mathcal{I}) + \lambda_5 \mathcal{L}_{\mathrm{affine}}(\mathcal{I})$$

3. PANIC: Summary

- Classifying AD still challenging
- First interpretable model for heterogeneous data
- PANIC allows easy toubleshooting of model
- PANIC interpretable both locally and globally
- PANIC closes gap for clinical application [7]



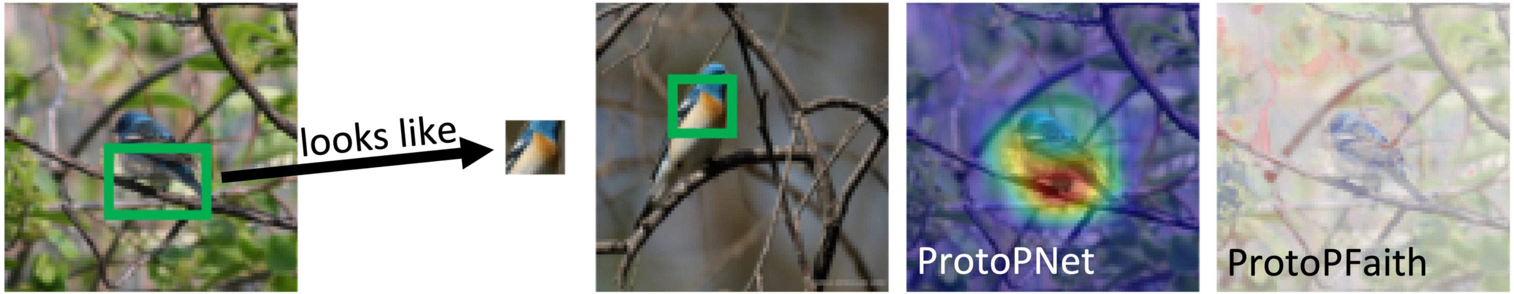
[7] Tonekaboni, S. et al: What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. Machine Learning for Healthcare Conference, PMLR p.359-380 (2019).

Overview

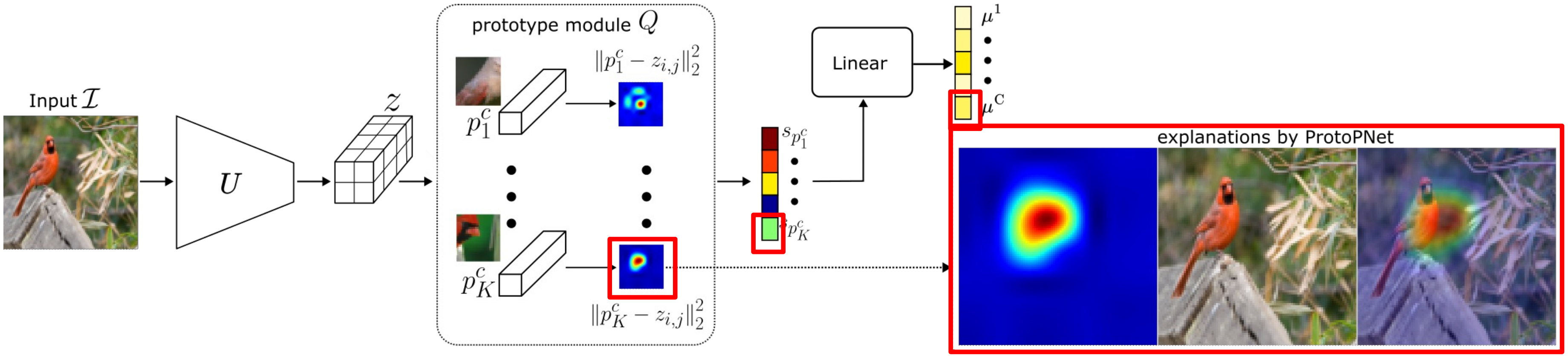
4. ProtoPFaith: Axioms

1. Sensitivity
2. Implementation Invariance
3. Completeness
4. Dummy
5. Linearity
6. Symmetry-Preserving

4. ProtoPFaith: Case-Based Reasoning



4. ProtoPFaith: ProtoPNet [3]



[3] Chen, C. et al.: This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS, vol. 32 (2019)

4. ProtoPFaith: Motivation

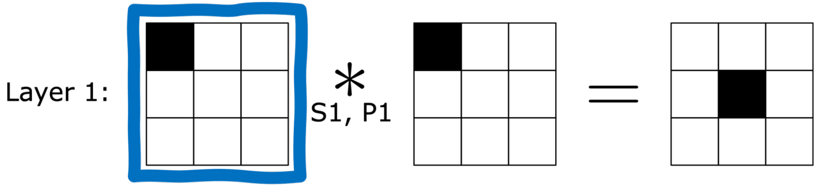
Assumption in ProtoPNet-like [3] architectures:

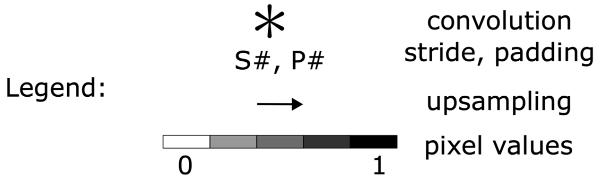
Spatial dependency between
latent features and input domain

4. ProtoPFaith: Motivation

Assumption in ProtoPNet-like [3] architectures:

Spatial dependency between
latent features and input domain



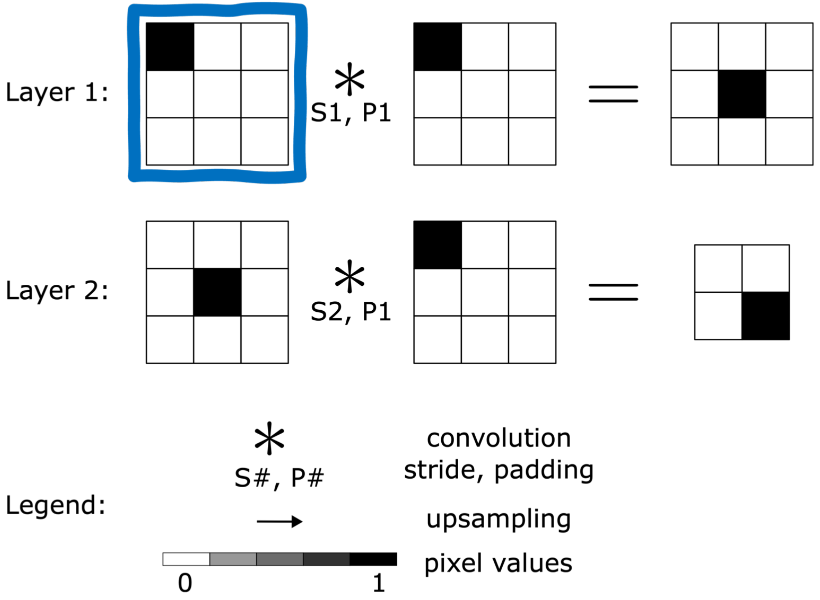


[3] Chen, C. et al.: This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS, vol. 32 (2019)

4. ProtoPFaith: Motivation

Assumption in ProtoPNet-like [3] architectures:

Spatial dependency between
latent features and input domain

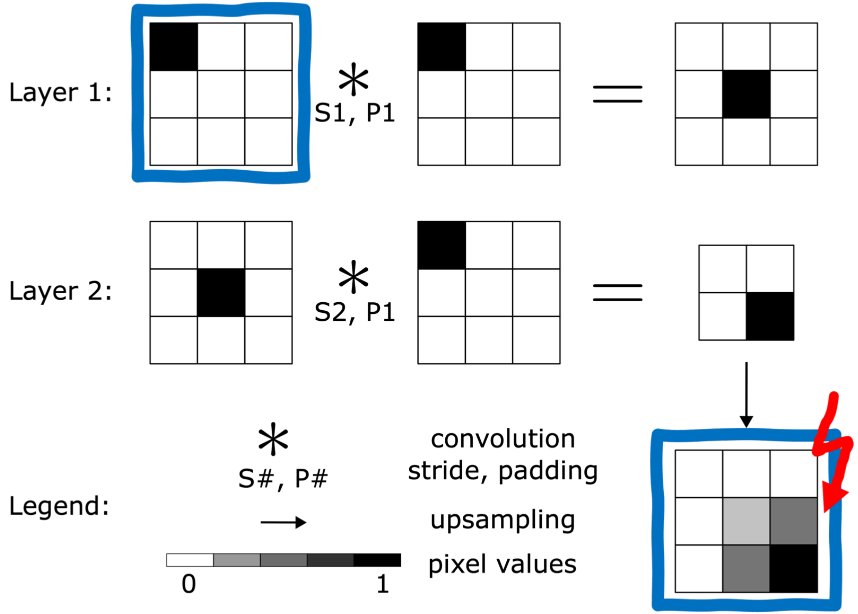


[3] Chen, C. et al.: This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS, vol. 32 (2019)

4. ProtoPFaith: Motivation

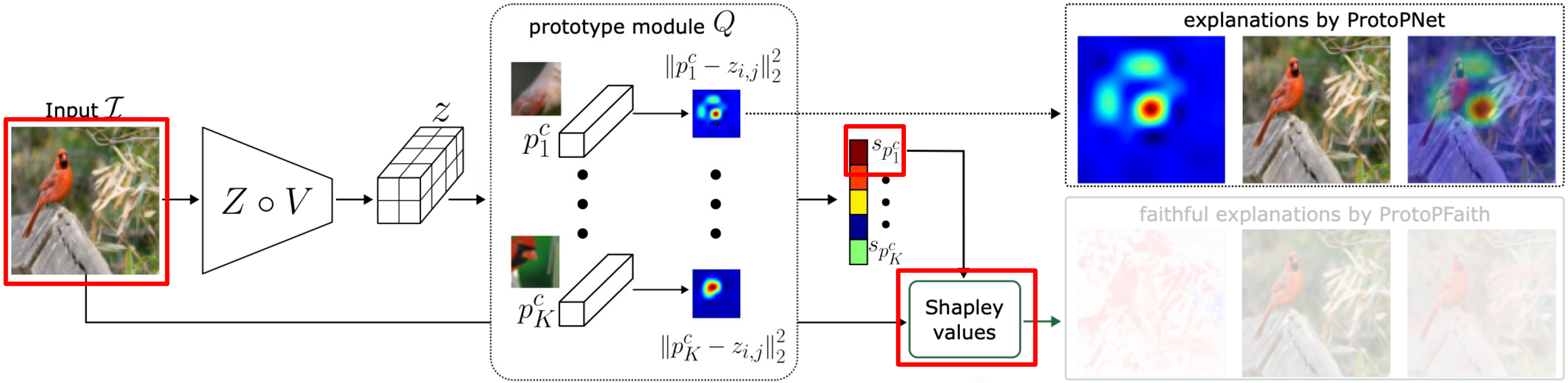
Assumption in ProtoPNet-like [3] architectures:

Spatial dependency between
latent features and input domain



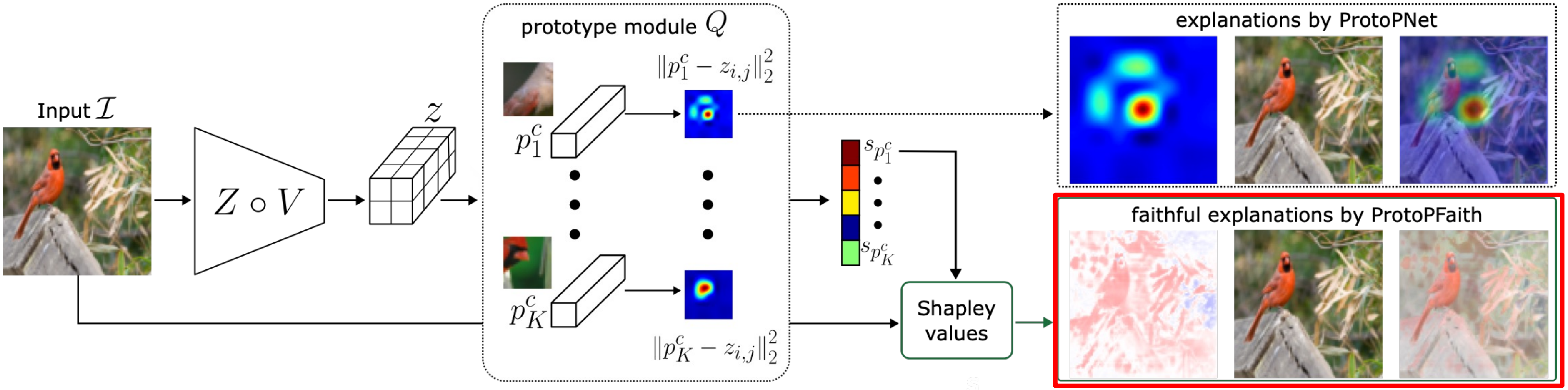
[3] Chen, C. et al.: This Looks Like That: Deep Learning for Interpretable Image Recognition. NeurIPS, vol. 32 (2019)

4. ProtoPFaith: Implementation of Case-Based Reasoning



[8] Wolf, TN et al.: Keep the Faith: Faithful Explanations in Convolutional Neural Networks for Case-Based Reasoning, AAAI (2024)

4. ProtoPFaith: Implementation of Case-Based Reasoning



[8] Wolf, TN et al.: Keep the Faith: Faithful Explanations in Convolutional Neural Networks for Case-Based Reasoning, AAAI (2024)

4. ProtoPFaith: Method

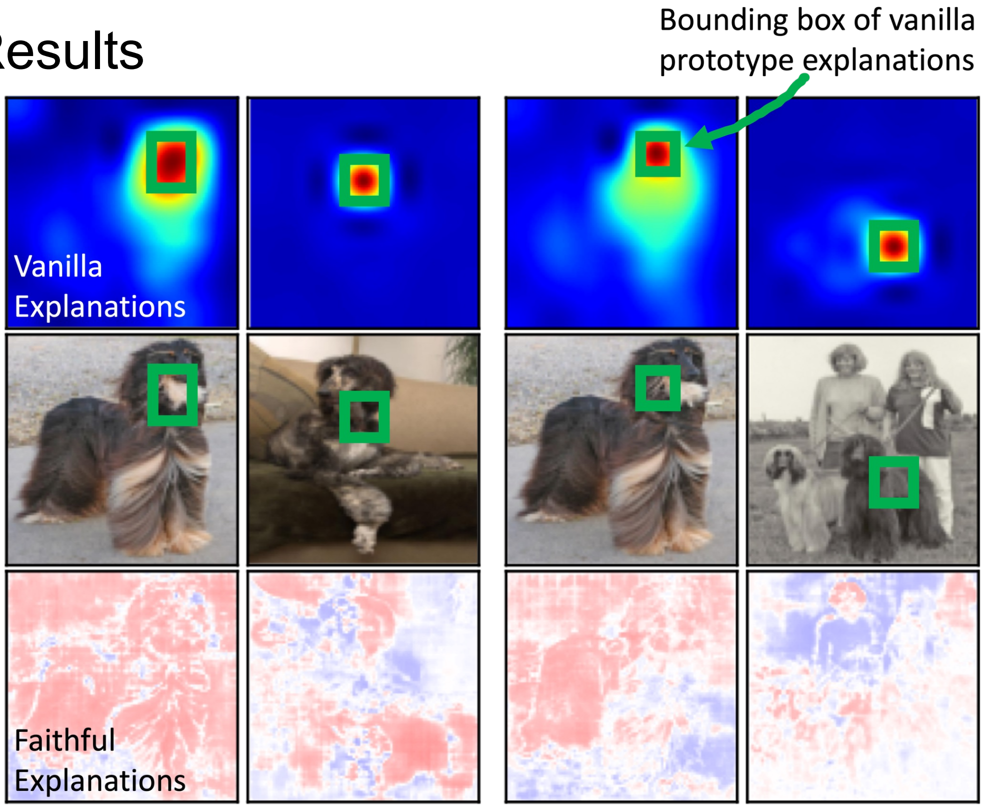
- Convert *trained* ProtoPNet into Lightweight Probabilistic Neural Network [9]
- Extract explanations following DASP [10] over similarity scores *s*
- Explanations are based on Shapley values,
  which satisfy all axioms that we define to be required for *faithfulness*
- Extraction of vanilla explanations still possible for the same model

Requirements

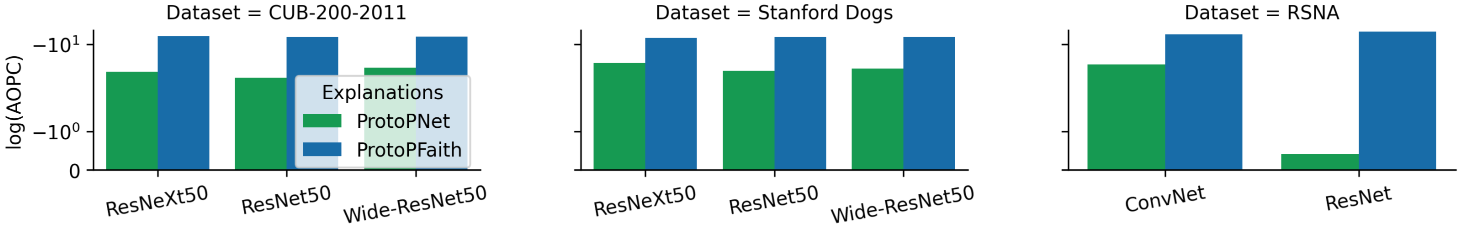
- Closed-Form solution for propagation of normal distributions through all layers

[9] J. Gast, S. Roth: “Lightweight probabilistic deep networks”, CVPR 2018
[10] M. Ancona, C. Oztireli, M. Gross: “Explaining deep neural networks with a polynomial time algorithm for shapley value approximation”, ICML 2019

4. ProtoPFaith: Results



4. ProtoPFaith: Results



4. ProtoPFaith: Discussion

- Theoretical violations manifest in experimental results
- Findings generalize to other implementations of case-based reasoning, e.g. ProtoTrees [11] and XProtoNet [12]
- Faithful explanations difficult to interpret

[11] M. Nauta, R. Van Bree, C. Seifert: “Neural prototype trees for interpretable fine-grained image recognition”, CVPR 2021
[12] E. Kim, S. Kim, M. Seo, S. Yoon: “XProtoNet: Diagnosis in Chest Radiography With Global and Local Explanations”, CVPR 2021

Overview

1. Alzheimer‘s Disease
2. Explainability and Related Work
3. PANIC
4. ProtoPFaith
5. Summary

5. Summary

- Proposed one inherently interpretable neural network for image and tabular data
- Found that implementations of case-based reasoning are not as faithful as anticipated
- Restoring faithful explanations in case-based reasoning ongoing work (reach out for collaboration ☺)

Thank You!

Group:

Don’t PANIC:

Keep the Faith:

Contact via mail: tom_nuno.wolf@tum.de