



SAPIENZA  
UNIVERSITÀ DI ROMA

Sapienza University of Rome

Department of Computer, Control and Management Engineering  
PhD in Computer Engineering

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

# Explaining Deep Neural Networks by Leveraging Intrinsic Methods

Thesis Advisor

**Prof. Roberto Capobianco**

Candidate

**Biagio La Rosa**

Co-Advisor

**Prof. Daniele Nardi**

Academic Year 2023-2024 (XXXVI cycle)



SAPIENZA  
UNIVERSITÀ DI ROMA

Sapienza University of Rome

Department of Computer, Control and Management Engineering  
PhD in Computer Engineering

THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

# Explaining Deep Neural Networks by Leveraging Intrinsic Methods

Thesis Advisor

**Prof. Roberto Capobianco**

Candidate

**Biagio La Rosa**

Co-Advisor

**Prof. Daniele Nardi**

Academic Year 2023-2024 (XXXVI cycle)

## Abstract

Deep neural networks have been pivotal in driving AI advancements over the past decade, revolutionizing domains such as gaming, biology, autonomous systems, and voice and text assistants. Despite their impact, these networks are often regarded as black-box models due to their intricate structures and the absence of explanations for their decisions. This opacity poses a significant challenge to AI systems' wider adoption and trustworthiness. This thesis addresses this issue by contributing to the field of eXplainable AI, focusing on enhancing the interpretability of deep neural networks.

The core contributions lie in introducing novel techniques aimed at making these networks more interpretable by leveraging an analysis of their inner workings. Specifically, the contributions are threefold. Firstly, the thesis introduces designs for self-explanatory deep neural networks, such as the integration of external memory for interpretability purposes and the usage of prototype and constraint-based layers across several domains. These proposed architectures are specifically designed to preserve most of the black-box networks, thereby maintaining or improving their performance. Secondly, this research delves into novel investigations on neurons within trained deep neural networks, shedding light on overlooked phenomena related to their activation values. Lastly, the thesis conducts an analysis of the application of explanatory techniques in the field of visual analytics, exploring the maturity of their adoption and the potential of these systems to convey explanations to users effectively.

In summary, this thesis contributes to the growing field of Explainable AI by proposing intrinsic techniques to enhance the interpretability of deep neural networks. By mitigating the opacity issue of deep neural networks and applying them to several different applications, the research aims to foster trust in AI systems and facilitate their wider adoption across several applications.



## Acknowledgements

I strongly believe that the goals achieved in life need to be shared with all the people you interact with. My life has been full of sliding doors, chaos, dreams, failures, and discoveries, and all of them involved other persons in the process. Therefore, thank you to all the people who met me in their lives, no matter if you like or hate me: each of you pushed me to become who I am.

Special thanks for their contribution to my research path must be given to:

- Prof. Roberto Capobianco, for supporting my freedom, my research, and guiding my studies. We embarked on this mad and risky journey of explainability together, and I really enjoyed the time under your supervision. Thank you for pushing me to go outside my comfort zone and for all your precious suggestions. The calls during the evening to discuss novel ideas and the long night for the IJCAI paper are examples of memories that I will remember forever! arigatō Roberto!
- Prof. Leilani Gilpin for welcoming me as a visitor to your lab, for granting me freedom during the months there, and for being so supportive even in tasks where you were not supposed to be. You are a rare case in academia: your way of supervising and the environment that you built around me allowed me to exploit the best of my skills and reach the goals that I wanted to reach. Therefore, thank you for everything!
- KRL group for our dinners and casual chats! My first year alone as a student and under the pandemic was hard, but thanks to your presence, I discovered what it means to be in a lab group of smart people.
- XAI-KRL group for becoming a house where discussing the field, ideas, and opinions and grow together.
- the AWARE group (Marco, Graziano, and Prof. Santucci) for introducing me to a novel field and opening the doors to so many future directions and collaborations.
- Prof. Daniele Nardi, for the chance to be in your lab, for your support whenever I needed it, and for being kind and nice to me! I really appreciate everything that you have done in these last three years!
- Prof. Roberto Navigli, for introducing me to the field of AI and research when I was an undergraduate without any knowledge of AI. Your suggestions and your supervision definitely shaped my future, transforming me from a future software engineer to a future researcher. Your compliments and advice pushed me to surpass myself at each step of my career until today.

Thank you, Alessio Ragno, for becoming my French partner in crime in XAI. Your presence and our battles helped me to grow as a supervisor and as a reviewer 2. Because of you, I learned how to choose the best paper title and how to sell ice to a polar bear. Thank you, Michela Proietti, for being patient and kind in this last period. I'm sure you will become a great scientist; just relax and enjoy your journey!

Additionally, I would like to thank all the students that I supervised both for theses and projects: interacting with you allowed me to experiment and improve my skills as a supervisor and become a better researcher. Thank you to all the reviewers<sup>2</sup> who rejected my papers and proposals.

During this path, I had the chance to collaborate with great colleagues whose contribution and help to this thesis are valuable in multiple aspects. Other than my French partners Alessio and Michela, I thank Romain Giot, Graziano Basili, Sayo M. Makinwa, Enrico Bertini, Marco Angelini, Romain Bourqui, Rino Ragno, David Auber, and Giuseppe Santucci for contributing to the research presented in this thesis. I hope to collaborate again with all of you in the future.

Thank you to all my family for being supportive during my long academic travel. I hope that today you will be proud of me and you can finally relax. Thank you also to my friends for pushing me to prove to them that doing research is “a job”. Finally, thank you to Roberta for being so patient during the writing process of this thesis and for postponing your drama. Thank you for encouraging me to make difficult decisions even if these decisions will impact your life and future. I really appreciate your help, and I hope we will walk together for a long time.

# Contents

<b>List of Figures</b>	vii
<b>List of Tables</b>	x
<b>Nomenclature</b>	xiii
<b>Acronyms</b>	xiv
<b>1 Introduction</b>	1
1.1 Introduction . . . . .	1
1.2 Contributions . . . . .	3
1.3 Thesis Organization . . . . .	4
1.4 Related Publications . . . . .	5
<b>I Preliminaries</b>	7
<b>2 Background</b>	8
2.1 Fundamentals of Deep Neural Networks . . . . .	8
2.1.1 Architectures . . . . .	11
2.2 Explainable AI . . . . .	15
2.2.1 Categorization . . . . .	16
2.2.2 Evaluation of XAI methods . . . . .	17
2.3 Visual Analytics . . . . .	20
<b>3 Related Work</b>	22
3.1 Post-Hoc Methods . . . . .	22
3.1.1 Feature Attribution . . . . .	22
3.1.2 Learned Features . . . . .	23
3.1.3 Explanations by examples . . . . .	24
3.1.4 Counterfactuals . . . . .	25
3.2 Self-Explainable Deep Neural Networks . . . . .	25
3.3 Relations with Thesis Contribution . . . . .	27
<b>II Self-Explainable Neural Networks</b>	29
<b>4 Prototype-based Interpretable Graph Neural Networks</b>	30
4.1 Design . . . . .	30
4.1.1 Training Process. . . . .	31
4.1.2 Explanations . . . . .	33
4.2 Experiments . . . . .	34
4.2.1 Performance . . . . .	34

4.2.2	Explanations . . . . .	35
4.3	Design Choices . . . . .	39
4.4	Contributions . . . . .	42
<b>5</b>	<b>Self-Explainable Memory-Augmented Deep Neural Networks</b>	<b>43</b>
5.1	Simplified Differentiable Neural Computers . . . . .	44
5.1.1	Design . . . . .	44
5.1.2	Experiments . . . . .	48
5.1.3	Design Choices . . . . .	53
5.2	Memory Wrap . . . . .	54
5.2.1	Design . . . . .	55
5.2.2	Experiments . . . . .	59
5.2.3	Design choices . . . . .	68
5.3	Contributions . . . . .	71
<b>III</b>	<b>Explaining Latent Representations</b>	<b>72</b>
<b>6</b>	<b>Graph Concept Whitening</b>	<b>73</b>
6.1	Design . . . . .	73
6.1.1	Alignment Process . . . . .	74
6.1.2	Explanations . . . . .	75
6.2	Experiments . . . . .	75
6.2.1	Concept Selection . . . . .	75
6.2.2	Performance . . . . .	76
6.2.3	Explanations . . . . .	77
6.3	Design Choices . . . . .	79
6.4	Contributions . . . . .	81
<b>7</b>	<b>Clustered Compositional Explanations</b>	<b>82</b>
7.1	Algorithm Design . . . . .	82
7.1.1	Proof of Admissibility of MMESH . . . . .	86
7.2	Metrics . . . . .	88
7.3	Experiments . . . . .	89
7.3.1	Motivation . . . . .	89
7.3.2	Performance . . . . .	90
7.3.3	Explanations . . . . .	90
7.3.4	Metrics . . . . .	94
7.4	Design Choices . . . . .	95
7.5	Contributions . . . . .	97
<b>IV</b>	<b>Bringing Explanations to the User: Visual Analytics</b>	<b>98</b>
<b>8</b>	<b>Explaining Deep Learning in Visual Analytics Systems</b>	<b>99</b>
8.1	Overview . . . . .	99
8.1.1	Feature Attribution . . . . .	100
8.1.2	Learned Features . . . . .	101
8.1.3	Explanations by Examples and Counterfactuals . . . . .	102
8.1.4	Model Behavior . . . . .	103
8.2	Analysis . . . . .	104
8.3	Towards a Better Integration . . . . .	108
8.4	Contributions . . . . .	109

<b>V Conclusions</b>	<b>110</b>
<b>9 Conclusions</b>	<b>111</b>
9.1 Summary . . . . .	111
9.2 Limitations . . . . .	112
9.3 Future Research Directions . . . . .	113
<b>Bibliography</b>	<b>115</b>

# List of Figures

2.1	An example of a 3-layer neural network. Each node in the hidden layers is connected with all the nodes of the previous layer and all the nodes of the next layer. . . . .	9
2.2	A simplified version of an artificial neuron. . . . .	9
2.3	An unfolded view of an LSTM cell. . . . .	11
4.1	The architecture of PIGNN when the task is graph classification. . . . .	31
4.2	Avg. Fidelity and standard deviation achieved by the intrinsic method, DeepLIFT, and GNNExplainer on GCN (a) and GAT(b) models across several datasets. . . . .	36
4.3	Avg. Fidelity and standard deviation reached by the intrinsic method, DeepLIFT, and GNNExplainer on GIN models across several datasets. . . . .	37
4.4	Sub-graphs associated with the prototypes learned by PIGNN based on GCN on the Ba-2Motifs dataset. The most activated node is marked with a red borderline. The blue shade encodes the importance of the node for the subgraph. . . . .	38
4.5	Visualization of explanation for a prediction. The red border indicates the analyzed node; the Input Activations column reports the similarity between the graph’s nodes and the prototype displayed in the same row in the Prototype Activation column. . . . .	38
4.6	Accuracy of PIGNN when varying the number of prototypes. . . . .	39
4.7	Visualization comparison between projecting the similarities towards all the nodes in the graph and only towards the k-hop most important subgraph. . . . .	40
4.8	Avg. Accuracy and standard deviation of PIGNN when jointly trained with the feature extractor (Classic), trained after a pre-training on the same dataset (Pre-training), and fine-tuning a model trained in a different dataset via transfer learning (Transfer) . . . . .	41
5.1	Design of SDNC. The blue boxes represent the read-and-write history. $\langle C_i, x_j \rangle$ indicates that feature $x_j$ is stored in cell $C_i$ . The ranking sorts the subsequence of the input based on their importance. . . . .	47
5.2	Example of a 5-step modified T-maze, where the starting position (in white) indicates the relevant cell (in yellow). . . . .	49
5.3	Cumulative reward and standard deviation of ten agents trained on the T-Maze problem. . . . .	50
5.4	Matching accuracy score and standard deviation of ten agents trained on the T-Maze problem. . . . .	50
5.5	Example outputs on the Story Cloze Test. A relevance score is associated with each premise. The blue color indicates the best premise, while the orange indicates the worst one. . . . .	51
5.6	Output of LIME [197] applied to a couple of samples from the Story Cloze Test dataset. . . . .	52
5.7	Architecture of a deep neural network augmented with Memory Wrap. . . . .	55
5.8	Illustration of a snapshot of the memory of Memory Wrap. Each row can be associated with one of the three sets $M_0$ , $M_=$ , and $M_{\neq}$ based on its associated content weight $w_c$ and prediction. . . . .	57
5.9	Performance comparison between memory-based modules. . . . .	60

5.10	Performance comparison between k-NN variants and Memory Wrap variants. . . . .	61
5.11	Avg. accuracy and standard deviation over 15 runs of the black-box model, the full Memory Wrap, and its ablated version on subsets of CIFAR10. . . . .	62
5.12	Avg. accuracy and standard deviation over 15 runs of the black-box model, the full Memory Wrap, and its ablated version on subsets of CINIC10. . . . .	63
5.13	Avg. Input representativeness scores and standard deviation reached by CHP (gray), KNN* (pink), the random baseline (green), and Memory Wrap (blue). Lower is better. . . . .	65
5.14	Avg. Prediction representativeness scores and standard deviation reached by CHP (gray), KNN* (pink), the random baseline (green), and Memory Wrap (blue). Lower is better. . . . .	65
5.15	Inputs (first rows) from SVHN (a) and CIFAR10 (b), their associated predictions, and an overview of the samples in the memory set that have an active influence on the decision process (second row). . . . .	67
5.16	An example of a model biased towards the background. The figure shows how the inspection of the memory samples used by the network makes clear the reasons behind the bad performance of the network. . . . .	68
5.17	Comparison between different encoders trained using Cosine Similarity (solid lines) and $L_2$ distance (dashed lines) on a subset SVHN dataset. . . . .	70
6.1	Separability score of the latent space of a black-box model and its interpretable version. Lower is better. . . . .	78
6.2	Avg. importance scores assigned to five concepts in the layers of a GAT model trained on BBBP. . . . .	78
6.3	Feature attribution for the concept #Heteroatoms in the BBBP dataset. . . . .	79
7.1	Explanations in terms of recognized concepts and sample examples of different activation ranges for unit #1. . . . .	83
7.2	Visualization of information used by MMESH: the size of the concept mask (white), size of the intersection (yellow), minimum extension (red bounding box), maximum extension (blue bounding box). . . . .	85
7.3	Percentage of labels' category associated with the labels returned by CoEx over different activation ranges. Left: ranges from quantile to infinite. Right: range from 0 to quantile. Results are computed over 100 randomly extracted units. . . . .	89
7.4	Examples of specialization (left) and polysemy (right). Neuron #368 (right) recognizes unrelated concepts among different activation ranges (windmill, closet, amusement park). . . . .	94
8.1	Example of complex visualization in VA systems of attention flow. . . . .	100
8.2	Example of abstraction in VA system. . . . .	101
8.3	Example of interface to steer a model to add, delete, and revise prototypes. All edits are revertable and traceable in the editing history. Image from Ming et al. [157] © 2019 IEEE . . . . .	103
8.4	View to interact with a latent representation. The user can assign values to dimensions of the latent space in the parallel coordinate plot, and the system will adjust the other views accordingly. Image from Wang et al. [258] ©2020 IEEE . . . . .	104
8.5	Distribution of target users in the analyzed VA systems . . . . .	104
8.6	Summary of user support for different XAI categories. The edge width is proportional to the number of systems connecting users and categories. The opacity of XAI categories is proportional to the total number of contributions. . . . .	105
8.7	Target domain of VA systems. Research refers to research papers targeting general applications. . . . .	106
8.8	Distribution of interactions supported for the user. . . . .	107

8.9	Distribution of evaluation procedures used in the analyzed systems. . . . .	107
8.10	(a) Explanation categories and (b) models coverage over the years. . . . .	108

# List of Tables

4.1	Avg. Accuracy and standard deviation over 15 runs of GNNs with and without prototype layers on synthetic datasets. . . . .	34
4.2	Avg. Accuracy and standard deviation over 15 runs of GNNs with and without prototype layers on molecular datasets. . . . .	34
4.3	Avg. Accuracy and standard deviation over 15 runs of GNNs with and without prototype layers on citation datasets. . . . .	35
4.4	Explanations performances on big datasets. $Fidelity^{+prob}$ and $Fidelity^{+ROC-AUC}$ (Equation (2.24)) are reported for the self-interpretable models and the extrinsic methods on the black-box GCN model. . . . .	37
4.5	Fidelity comparison between projecting the similarities towards all the nodes in the graph and only towards the k-hop most important subgraph. . . . .	39
4.6	Performance comparison between the full and the reduced loss applied to PIGNN. .	40
4.7	Examples of specific cases when the reduced loss can work better. . . . .	40
4.8	Interpretability scores achieved by PIGNN when using the full or reduced loss. . .	40
5.1	Average test Accuracy reached by the models in the testing Story Cloze Test dataset.	50
5.2	Avg. fidelity over ten runs on the Story Cloze Test when feeding the model with only a random, the best and the worst premise. . . . .	51
5.3	Accuracy and $Fidelity^{+acc}$ when the model is untrained and then labels are shuffled.	52
5.4	Avg. fidelity obtained on the development Story Cloze Test dataset using different $topk^r$ values. . . . .	54
5.5	Avg. accuracy and standard deviation of the baselines and Memory Wrap when the training datasets are the whole SVHN, CIFAR10 and CINIC10 datasets. . . . .	63
5.6	Avg. and standard deviation of the matching accuracy reached over 15 runs by the sample in the memory set with the highest sparse content-based attention weight (Top), the example with the lowest weight but greater than zero (Bottom) and a random sample (Random) . . . . .	64
5.7	Avg. IM1, IM2, and IIM1 scores on the SVHN dataset reached by Memory Wrap and a prototype-based method. Lower is better. . . . .	66
5.8	Avg. and standard deviation of the Pearson product-moment correlation coefficients between the correctness of the predictions, the number of counterfactuals (number), and the position of the first counterfactual (position). . . . .	67
5.9	Accuracy reached by the model on SVHN when the sample with the highest weight in the memory set is a counterfactual (Top Counter) and when there are no counterfactuals at all (No Counter). . . . .	67
5.10	Avg. accuracy and standard deviation over five runs Memory Wrap using different selection mechanisms to select samples in the memory set. . . . .	68
5.11	Avg. accuracy and standard deviation over five runs of Memory Wrap trained using a variable number of samples in memory, when the training dataset is a subset of SVHN. . . . .	70

6.1	Avg. Accuracy and ROC-AUC over 15 runs on ClinTox when adding one concept at a time for the GCW layers. We highlight the best results of combined accuracy and ROC-AUC.	76
6.2	Avg. ROC-AUC and standard deviation over 15 runs before and after adding GCW to the models.	77
6.3	Avg. and standard deviation of the Fidelity+ score of feature attributions computed on both black-boxes and models employing GCW layers.	78
6.4	Avg. ROC-AUC and standard deviation reached by models including GCW using different activation modes.	80
6.5	Avg. ROC-AUC and standard deviation of models when GCW replaces several types of normalization layers.	80
7.1	Avg. and standard deviation of steps required to reach the optimal solution within the beam.	90
7.2	Avg. and Std Dev. across metrics reached by NetDissect, CoEx, and Clustered Compositional Explanations using Ade20K as a concept dataset.	91
7.3	Avg. and Std Dev. across metrics reached by NetDissect, CoEx, and Clustered Compositional Explanations on ResNet. Concepts are retrieved from the Pascal dataset.	91
7.4	Avg. and Std Dev. of the desired qualities over the labels returned by NetDissect, CoEx, and Clustered Compositional Explanations (our) applied to ResNet pre-trained on ImageNet. Results are computed for 50 randomly extracted units.	92
7.5	Percentage of predictions changed when the activations of a given cluster are masked.	93
7.6	Percentage of unspecialized and weakly specialized activation ranges in ResNet18 (ReLU Yes) and DenseNet161(ReLU No).	93
7.7	Avg. and Std Dev. of the scores obtained by optimizing the algorithm using Detection Accuracy. Underlined results are those that reach lower scores compared to when using the IoU score. Bold results are those that reach higher scores.	95
7.8	Avg. and Std Dev. of the proposed metrics when using a variable number of clusters. Results are computed for 50 randomly extracted units.	96
7.9	Avg. and standard deviation of visited states per unit. Results are computed for 100 randomly extracted units.	97

# Nomenclature

- Element-wise Multiplication
- $\|\cdot\|^n$   $n$ -Norm Distance
- $\|\cdot\|$  Square Distance
- $\mathfrak{D}$  Dataset
- $\mathfrak{L}^l$  Set of logical connections of arity  $l$
- $\oplus$  Concatenation
- $\sigma_{sig}(x)$  Sigmoid Function
- $\sigma_{soft+}(x)$  Softplus Function
- $\sigma_{soft}(x)$  Softmax Function
- $\sigma_{tanh}(x)$  Tanh Function
- $abs()$  Absolute Value
- $f_{CL}$  Classifier
- $f_{FE}$  Feature Extractor
- a**  $n$ -Norm Distance

# Acronyms

**AI** Artificial Intelligence

**CBM** Concept Bottleneck Models

**CNN** Convolutional Neural Network

**CoEx** Compositional Explanations

**DL** Deep Learning

**DNC** Differentiable Neural Computer

**DNN** Deep Neural Network

**GAT** Graph Attention Network

**GCN** Graph Convolutional Network

**GCW** Graph Concept Whitening

**GIN** Graph Isomorphism Network

**GNN** Graph Neural Network

**MANN** Memory Augmented Neural Network

**ML** Machine Learning

**MLP** Multi-Layer Perceptron

**NetDissect** Network Dissection

**NN** Neural Network

**PIGNN** Prototype-based Interpretable Graph Neural Network

**SDNC** Simplified Differentiable Neural Computer

**VA** Visual Analytics

**XAI** eXplainable Artificial Intelligence

# Chapter 1

## Introduction

### 1.1 Introduction

Artificial Intelligence (AI) is a field of computer science aimed at developing machines capable of solving tasks that typically require human intelligence. The first successful AI approaches were rooted in expert systems relying on rules and symbolic reasoning. However, despite early optimism, symbolic AI has revealed limited adaptation capabilities. These systems often assume perfect knowledge of tasks and disregard uncertainty or ambiguity in data. Consequently, pure symbolic AI struggles to manage complex tasks for which humans cannot describe the rules governing the phenomenon. For instance, tasks like translating text, recognizing images, and exploring unknown environments were deemed impossible to solve.

Machine Learning (ML) emerged as a paradigm shift in AI research to mitigate these problems. ML provides algorithms capable of learning from data and improving performance over time without explicit knowledge of a particular phenomenon's rules. Classical ML algorithms use statistical theory for pattern recognition on a set of data. Examples of this category include decision trees, logistic regression, and Support Vector Machines [82], which were the state-of-the-art tools to deal with complex tasks until recently.

In recent years, the internet's expansion and the availability of cheaper hardware, open-source platforms, and big data have enabled the collection of vast raw datasets. However, classical ML struggles to fully exploit these datasets due to their size, complexity, and lack of explicit semantics associated with data. Conversely, **Deep Neural Networks (DNNs)** are explicitly designed to process raw data and memorize large quantities of information in the interconnections between network layers. Therefore, the field of Deep Learning (DL), which leverages DNNs, has become the primary subfield of ML.

A DNN consists of thousands or millions of interconnected neurons. Since the work of Rosenblatt [199], the design of neurons and DNNs has become increasingly complex, involving functions such as non-linear operations, convolution, memorization, attention, and skip connections. The complexity of these designs enables DNNs to achieve impressive performance in various tasks, often surpassing human performance. Games [159, 215], vision [196, 233], robotics [128, 130], and natural language processing [24, 245] have all undergone a revolution in their respective field. Nowadays, mainstream applications for speech recognition, machine translation, and text generation are all powered by DL systems.

However, the gain in performance comes at the cost of transparency. While symbolic systems are easy to understand in terms of encoded knowledge and decision process, classical ML tends to be more opaque in both aspects. Indeed, keeping track of the learning process is challenging, and the explainability of these systems is often limited to extracting the learned decision process. But, even in these cases, there is a trade-off between complexity and transparency. For instance, when using large ML models (e.g., a wide and deep decision tree), extracting explanations about the behavior that are easy to understand could be challenging since explanations can be extremely long. The challenge is exacerbated for DNNs, where there is no semantic associated with data, and tracking how inputs are transformed into outputs is a nightmare due to the complexity of the interconnections. Therefore, **DNNs are commonly referred to as black-boxes, where one feeds an input and receives an output without understanding the motivations behind the results.**

The field of eXplainable Artificial Intelligence (XAI) aims to address the need for transparency and interpretability in machine learning and AI systems. XAI aims to provide insights into the inner workings of AI models, enabling users to understand and interpret their outputs effectively. XAI encompasses a broad range of techniques, spanning from methods that highlight the most important parts of the input to those that extract the knowledge learned by machine learning models.

In classical ML, XAI methods mainly focus on providing concise explanations summarizing the already known rationale behind decisions. Examples of these methods are feature importance analysis [292], which identifies the most relevant features contributing to model predictions, and decision tree visualization [237], which provides a graphical representation of the decision-making process. In the context of DL, XAI methods deal with an unknown decision process and an unknown learned knowledge and aim at approximating, guessing, or probing the real model behavior.

The first and most popular techniques in the area are the so-called *extrinsic methods*, which approximate the behavior and generate explanations for DL models by exploiting external means. For example, several techniques employ surrogate models [140, 197], generative models [16], or perturbation-based analysis [59] to approximate the decision process of the networks around a given point. Despite the flexibility and high compatibility with existing models, recent works [4, 200] argue against relying solely on extrinsic methods. Indeed, extrinsic methods often struggle to capture the complexity of model behavior of DNNs, are biased by the selection of the external means, and require too much time to return a reliable approximation due to the complexity of the tasks managed by DNNs.

To address this issue, researchers are starting to explore *intrinsic methods*, which aim to enhance the interpretability of deep models by leveraging the inner workings of the models. This objective can be achieved by modifying the design of DNNs to make it more explainable, adjusting their training process to produce explainable representations, or analyzing and connecting the working mechanisms of their components. These methods include techniques such as attention mechanisms [13], which highlight relevant input features, activations analysis [17, 163], and self-explainable DNNs [33, 35]. Intrinsic methods offer the advantage of directly linking interpretability to the model’s design. They are generally faster than extrinsic methods and are more faithful to the model behavior. Nonetheless, they are usually tailored to specific settings (architectures, training processes, etc.) and, in the case of self-explainable DNNs, can incur performance trade-offs, thereby limiting the adoption from the general audience.

This thesis contributes to the ongoing research effort on XAI intrinsic methods by proposing methods that leverage DNN inner workings for explaining DL. In this context, the thesis proposes multiple designs of self-explainable DNNs and a post-training method to investigate the neurons' recognition capabilities. **The goal of the proposed self-explainable DNNs is to reduce the performance trade-off and broaden the applicability of such methods.** To achieve the first goal, the proposed layers can be inserted into black-box models without disrupting their structure and preserving most of their representation power. The second goal is achieved either by extending approaches to novel domains [33, 256], treating the preceding layers as black-boxes (i.e., not exploiting specific shapes or structures), and evaluating the proposed techniques across several architectures. As a side product, the thesis also increases the diversity of approaches in self-explainable DNNs since it introduces a new family of architectures: memory-based self-explainable DNNs.

Similarly, **the proposed post-training method** shares the underlying principle of compatibility and reliance on DNNs inner workings (i.e., activations) and **advances our knowledge of the semantics encoded in neurons.** Namely, it enables the investigation of broader settings than those explored in literature [17, 163], shedding light on novel phenomena related to the neurons' activation spectrum.

Finally, the thesis delves into the ongoing discourse on rendering the explanations usable and beneficial for users. Recently, several interactive systems have emerged to connect users, DL systems, and explanations by exploiting interactive interfaces [236, 263] and dialogue systems powered by large language models [221]. This thesis contributes to this research by surveying the combination of XAI techniques and Visual Analytics (VA) systems. **By surveying existing methodologies and advocating for the integration of XAI techniques into VA systems, the thesis aims to increase the awareness of the XAI and VA communities of each other and foster a new alternative direction for interactive explanations.**

## 1.2 Contributions

The main contributions of this thesis to the field of XAI are the following:

- the introduction of a **novel prototype-based layer for Graph Neural Networks (GNNs)**, which generalizes prototype-based self-explainable DNNs [33] in terms of domains, architectures, and assumptions. Differently from the current research, the layer represents prototypes as node embeddings, allowing its application to both graph and node classification tasks. Furthermore, our contribution includes a sparse explanation visualization for graphs that preserves the faithfulness of explanations while enhancing their understandability compared to commonly used visualization methods for prototype-based networks.
- the proposal of **novel memory modules designed to enhance the interpretability of existing neural networks while preserving their performance.** The utilization of memory modules for interpretability purposes represents a novelty in current literature. By operating within the same training settings as black-box models, these memory modules address some of the issues associated with the use of self-explainable DNNs. Additionally, these designs are flexible and can facilitate the retrieval of various types of explanations.

- the **application and adaptation of concept-whitening** [35] to the chemical domain and drug discovery tasks to ensure that latent representations encode semantics related to molecular properties. Our contribution includes the generalization of the method [35] to the graph and chemical domain, thereby expanding its applicability beyond the vision domain. The generalization includes the type of data, the covered architectural families, and the type of normalization layers, an area not explored in the literature.
- the **design of a novel algorithm for computing compositional explanations of neurons behavior**. This algorithm enables explanations for a broader spectrum of neurons behavior than the one investigated in the literature and overcomes issues related to the computational complexity of previous approaches [163]. Our contributions include the discovery and discussion of novel phenomena related to neuron activations, as well as the design of novel metrics for measuring the properties of explanation methods for latent representation.
- a **review and analysis of VA systems as a means to connect users and explanations** in an interactive environment without requiring XAI experts during the usage. By focusing on systems dealing with DNNs and employing XAI methods, the thesis contributes to identifying strengths and weaknesses of the proposed VA solutions, laying the foundation for future enhancements in the integration and collaborations between VA and XAI research.

### 1.3 Thesis Organization

The thesis is organized into five parts: Preliminaries, Self-Explainable Neural Networks, Explaining Latent Representations, Bringing Explanations to the User: Visual Analytics, and Conclusions. Each part is organized as follows:

- **Part I Preliminaries** provides the necessary background knowledge and terminology to comprehend the chapters of this thesis and reviews the related work on explainable AI applied to DNNs.
  - **Chapter 2** introduces the terminology and key concepts related to DNNs, XAI, and VA; describes the families of deep architectures examined in this thesis; introduces the categorization of explanation methods; discusses the evaluation challenges in XAI; and introduces the metrics used to evaluate the proposed methods.
  - **Chapter 3** reviews the literature on post-hoc XAI methods for explaining DL, focusing on techniques similar to those proposed in this thesis; reviews self-explainable DNNs; analyzes the limitations of current methods; and discusses the relationships between existing approaches and the thesis techniques.
- **Part II Self-Explainable Neural Networks** presents modules that can be added to existing neural networks to improve their explainability; presents a prototype-based design for self-explainable DNNs; introduces the novel family of self-explainable DNNs based on memory;
  - **Chapter 4** presents the design of a prototype-based layer for enhancing the interpretability of GNNs; discusses the benefits in terms of interpretability; evaluates the performance and the explanations over several architectures; and discusses alternative design choices.

- **Chapter 5** introduces a design to improve the interpretability of recurrent models on sequential data; introduces a memory module to enhance the interpretability of black-box models on image data; introduces the unified mechanism of memory tracking to compute different types of explanations; evaluates the performance and the explanations of the proposed modules; discusses alternative design choices.
- **Part III Explaining Latent Representations** presents methods to inspect the knowledge learned by DNNs; explore a method to enforce semantics in the latent representation of GNNs; presents a technique to extract semantics from neurons activations.
- **Chapter 6** presents a normalization for the graph neural networks and graph data domain to enforce latent representation to be aligned to molecular properties in the chemical domain and drug discovery; discusses the benefits in terms of interpretability; evaluates the performance and the explanations over several architectures, datasets, and layers; and discusses alternative design choices.
- **Chapter 7** introduces a heuristic-guided algorithm to compute the semantics encoded in a wide spectrum of neurons' activations; proposes a set of metrics to describe the properties of explanations returned by algorithms that compute explanations for neurons' activations; discusses the computational efficiency of the proposed algorithm; evaluates the explanations computed by the algorithm; discusses novel phenomena related to neurons activations; discusses alternative designs.
- **Part IV Bringing Explanations to the User: Visual Analytics** reviews and analyzes VA systems that incorporate XAI methods to help users understand DNNs.
- **Chapter 8** collects, reviews, and analyzes, from a XAI perspective, the VA systems that use XAI methods to improve the explainability of DNNs; analyzes the potential and benefits of interactive systems; discusses limitations and future direction of the field.
- **Part V Conclusions** summarizes the contributions of the thesis and discusses limitations and future research directions.
- **Chapter 9** summarizes the thesis's contribution; highlights the advantages and limitations of the proposed approaches; and discusses short and long-term future research directions that this thesis opened.

## 1.4 Related Publications

Part of the thesis has been previously published in the following journal and conference articles:

- *Towards a fuller understanding of neurons with Clustered Compositional Explanations.*  
**Biagio La Rosa**, Leilani Gilpin, and Roberto Capobianco  
In Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS 2023)
- *Explainable AI in Drug Discovery: Self-interpretable Graph Neural Network for molecular property prediction using Concept Whitenening.*

Michela Proietti, Alessio Ragno, **Biagio La Rosa**, Rino Ragno, and Roberto Capobianco.  
Machine Learning (Journal), pp. 1–32 (2023)

- *The State of The Art of Visual Analytics for eXplainable Deep Learning.*  
**Biagio La Rosa**, Graziano Blasilli, Romain Bourqui, David Auber, Giuseppe Santucci, Roberto Capobianco, Enrico Bertini, Romain Giot, and Marco Angelini  
Computer Graphic Forum (Journal)  
Presented also at 25th EG Conference on Visualization (EuroVIS 2023)
- *A self-interpretable module for deep image classification on small data.*  
**Biagio La Rosa**, Roberto Capobianco and Daniele Nardi.  
Applied Intelligence (Journal) (2023)
- *Prototype-based Interpretable Graph Neural Networks.*  
Alessio Ragno, **Biagio La Rosa**, and Roberto Capobianco  
IEEE Transactions on Artificial Intelligence (Journal) (2022)
- Detection Accuracy for Evaluating Compositional Explanations of Units.  
Sayo M. Makinwa, **Biagio La Rosa**, Roberto Capobianco.  
In Proceedings of AIxIA 2021 - Advances in Artificial Intelligence, pages 550–563. Springer International Publishing, (2021)
- *A Discussion about Explainable Inference on Sequential Data via Memory-Tracking.*  
**Biagio La Rosa**, Roberto Capobianco and Daniele Nardi.  
Discussion Papers of AIxIA 2021
- *Explainable Inference on Sequential Data via Memory-Tracking.*  
**Biagio La Rosa**, Roberto Capobianco and Daniele Nardi.  
In Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20) (2020)

# Part I

## Preliminaries

# Chapter 2

## Background

This chapter serves as a foundational introduction necessary for comprehending the subsequent content of the thesis. It introduces the terminology and concepts used to describe the methods proposed in this thesis.

The chapter is organized as follows: Section 2.1 introduces terminology and concepts related to Deep Learning (DL) and the architectures used in this thesis; Section 2.2 introduces terminology, metrics, evaluation, and categorizations related to eXplainable Artificial Intelligence (XAI); finally, Section 2.3 introduces terminology, concepts, and categorizations related to Visual Analytics (VA).

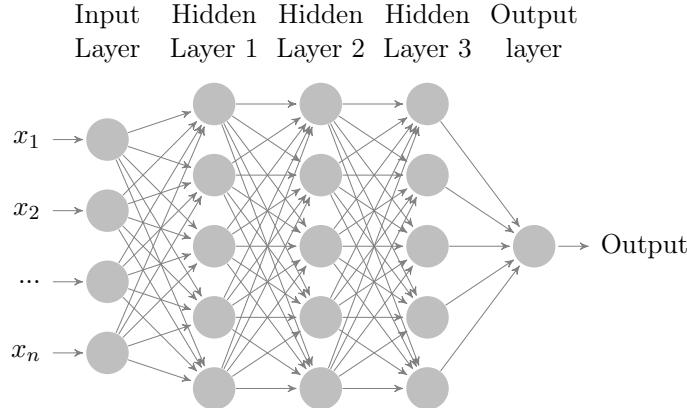
### 2.1 Fundamentals of Deep Neural Networks

This section describes the terminology used for the building blocks of a Neural Network (NN) and DL and introduces the families of architectures examined in this thesis.

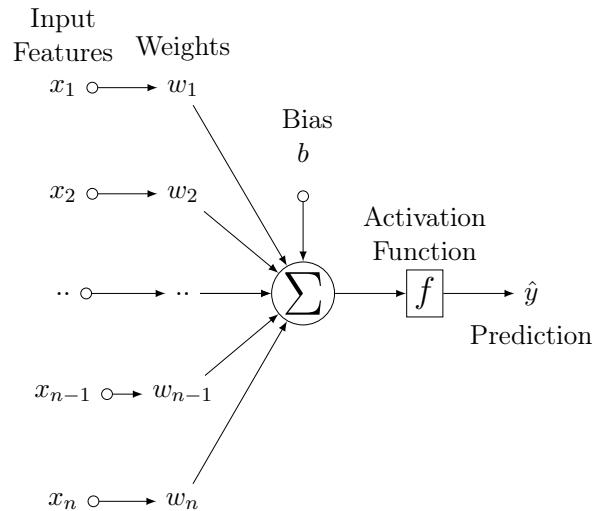
We begin the description by highlighting the objective of a NN: learn how to solve a task based on data observations. A data observation comprises several elements termed *features*. The collection of observations forms the *training dataset*, which is then used to train the NN in a process known as *training process*. In supervised learning and classification tasks, the primary paradigms explored in this thesis, each observation is associated with a *ground truth* label. The ground truth denotes the desired output and is used as feedback during the learning process.

A NN constitutes a hierarchical mathematical model, composed of multiple layers of interconnected artificial neurons, which receives an observation and yields an output referred to as a *prediction*. The layers of a NN can be categorized based on their position within the architecture: input, hidden, and output layers (Figure 2.1). The *input layer*, constituting the lowest layer in the hierarchy, feeds input features to the subsequent layers. A *hidden layer* receives the output of neurons from the previous layers and outputs to the next layer. The outputs of a hidden layer collectively form the *latent representation* of the sample. The manifold encompassing all possible latent representations of a given hidden layer is termed the *latent space*. The *output layer*, positioned at the apex of the hierarchy, accepts the output of the preceding layers as input and delivers the prediction. A network featuring at least two hidden layers is termed a Deep Neural Network (DNN).

The neurons across layers are interconnected through edges termed *weights*. Each weight is associated with a value. In its simplest form (Figure 2.2), a neuron processes the outputs of



**Figure 2.1:** An example of a 3-layer neural network. Each node in the hidden layers is connected with all the nodes of the previous layer and all the nodes of the next layer.



**Figure 2.2:** A simplified version of an artificial neuron.

neurons from the preceding layer, each multiplied by the corresponding weight connecting one of the previous neurons to the current one. Subsequently, an aggregation function (e.g., summation) combines all inputs. Finally, an activation function is applied to the aggregated value to compute the neuron output. The weight values are adjusted throughout the training process to achieve the desired output [199]. The mechanism for updating the weights, known as the *error-correction learning rule*, can be expressed as:

$$w_{i,t+1} = w_{i,t} + \eta[y_{j,d} - y_{j,t}]x_{i,t} \quad (2.1)$$

where  $w_{i,t}$  denotes the weight  $i$  at iteration  $t$ ,  $y_{j,d}$  represents the desired output for the input  $j$  and  $y_{j,t}$  denotes the current neuron output. If the prediction is correct, then  $y_{d,j} = y_{j,t}$  and the weights remain unaltered. The  $\eta \in [0, 1]$  parameter denotes the *learning rate* controlling the magnitude of the update: a small value leads to gradual adjustments, averaging the past inputs; large value facilitates rapid adaptation, albeit with lesser consideration for past errors. In the case

of neurons in hidden layers, where no pre-defined desired output exists (i.e., the ground truth), the backpropagation algorithm is employed to update their weights. This algorithm computes the error for each neuron  $j$  as follows:

$$\delta_j^L = \phi'(a_j) \sum_k w_{jk} \delta_k^{L+1} \quad (2.2)$$

where the superscript indicates the layer of the neuron,  $\delta_k^{L+1}$  represents the error of neuron  $k$  in the layer  $L + 1$ , connected to the current neuron via weight  $w_{jk}$ , and  $\phi'$  denotes the derivative of the activation function of the node  $j$ . The error is then multiplied by the input received by the neuron to obtain the gradient of the error with respect to the weight  $w_{ji}$ :

$$\frac{\partial E_n}{\partial w_{ji}} = \delta_j^L z_i^{L-1} \quad (2.3)$$

The set of the partial derivatives for all the weights in the network forms the gradient vector  $\nabla E_n$ , which is used to update the weights of the network:

$$w^{t+1} = w^t - \eta \nabla E_n \quad (2.4)$$

These equations serve as a general framework extended by the optimization algorithm called *optimizers* such as Adam [112], RMSProp, or AdaGrad [50]. Each optimizer presents its own learning schedule, differing in how weights are updated at each step. For instance, Adam and RMSProp employ distinct learning rates for different parameters and adjust them during the training. The highly interconnected structure and the training process enable each neuron to specialize in detecting particular feature correlations, enabling the entire network to represent any function.

The training process is divided into *epochs*, where the entire training dataset is fed to the network to update weights. As the dataset may be too large to be stored in memory, it is usually divided into *batches* of  $m$  samples progressively fed during the epoch. At each step, the optimizer computes the *loss* (i.e., error) associated with the prediction and updates the weights. Various loss functions can be employed based on the task.

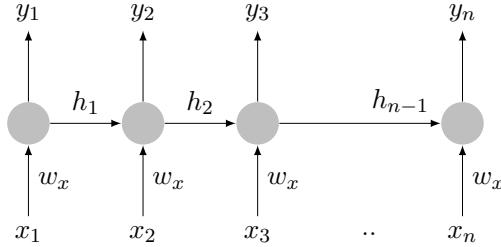
In recent years, numerous pre-trained DNNs have been available publicly. These models are trained on large corpora to learn fundamental features shared across several tasks. They can then serve as initial configurations for training models on smaller downstream tasks. The idea of training pre-trained models on downstream tasks starts from the observation that **most DNNs can be represented as the composition of two functions, the feature extractor  $f_{FE}$  and the classifier  $f_{CL}$** :

$$f_{FE} : \mathbf{x} \rightarrow \mathbf{h} \quad (2.5)$$

$$f_{CL} : \mathbf{h} \rightarrow \mathbf{y} \quad (2.6)$$

$$f : f_{CL}(f_{FE}(\mathbf{x})) \quad (2.7)$$

The feature extractor  $f_{FE}$  transforms input  $\mathbf{x}$  from features representation to a latent representation  $\mathbf{h}$  capturing feature relations. The latent representation denotes the output of the last layer of the feature extractor. Then, this representation is fed to the classifier  $f_{CL}$ , which transforms this representation into a prediction. Therefore, since the common features shared across tasks are likely



**Figure 2.3:** An unfolded view of an LSTM cell.

encoded into the feature extractor, the idea is to preserve the feature extractor’s learned knowledge and change the classifier for the downstream task. We can distinguish between two forms of training for novel tasks: *fine-tuning* and *transfer learning*. In the former, the weights of the feature extractor remain *frozen* and are not updated during training, with adjustments confined to the classifier’s weights. Conversely, in transfer learning, all the parameters are updated during the new training phase and the feature extractor’s weights are used to initialize the network’s parameters.

At the end of the training process, the network’s capability is assessed by providing a set of samples unseen during training, referred to as the *testing dataset*. The quality of the learning process is evaluated using a set of metrics that summarizes the network’s performance on the testing dataset. This thesis uses two metrics to evaluate the performance of DNNs: accuracy and ROC-AUC. *Accuracy* is the ratio between the number of correctly predicted samples and the total number of samples in the dataset. It ranges from 0 (there are no correct predictions) to 1 (all predictions are correct). *ROC-AUC* additionally takes into account the ratio between true positive (e.g., prediction is 1 when the ground truth is 1) and false positive (e.g., prediction is 0 when the ground truth is 1) and it is useful when the dataset is unbalanced in the number of samples per class.

### 2.1.1 Architectures

As mentioned in the previous section, neurons and layers can be interconnected in several ways, and multiple activation functions can be employed to compute the output of neurons. Based on the configuration schema of neurons and layers, we can distinguish different families of architectures. This section briefly describes the main layers and architectures used throughout the thesis: LSTM, attention, memory-augmented neural networks, convolutional neural networks, and graph neural networks.

#### LSTM

Long-Short Term Memory (LSTM) [86] networks are recurrent neural networks that deal with sequential data. Sequential data is a type of data where data points have a temporal dependency on other data points. A dataset of sequential data is a dataset where points chained by a temporal dependency are collected into a structure called a *sequence*. Each item in the sequence is a *timestep*.

In an LSTM, each neuron takes one timestep at a time as input and the previous neuron’s output, yielding a prediction. Therefore, if the sequence includes  $n$  timesteps, the LSTM yields  $n$  predictions (Figure 2.3).

In LSTMs, each neuron is a *memory cell* with its state  $s$ . For each timestep, the neuron takes as input the current element  $x_t$  in the sequence, the previous output  $h_{t-1}$ , and the previous cell state  $s_{t-1}$ . The purpose of the cell state is to store, update, and carry information through the timesteps. The content of the cell state is managed by two gates: the forget and the input gate. The *forget gate*  $g_f$  controls which information to delete from the state; the *input gate*  $g_i$  controls which information to write into the state. Finally, a tanh function  $\sigma_{\text{tanh}}$  computes the information to be stored  $\hat{s}_t$  as:

$$\hat{s}_t = \sigma_{\text{tanh}}(\mathbf{x}_t, h_{t-1}) \quad (2.8)$$

The next cell state is computed as the weighted sum between the current state and the information to be stored, weighted by the gates:

$$s_t = g_f s_{t-1} + g_i \hat{s}_t \quad (2.9)$$

Finally, the output is obtained by multiplying the *output gate*  $g_o$ , which controls the contribution of the cell state, and the candidate output, which is computed by applying the tanh function to the next cell state:

$$y_t = g_o \sigma_{\text{tanh}}(h_{t-1}, \mathbf{x}_t) \quad (2.10)$$

The architecture and the memory cell allow LSTM to store information longer than classical recurrent networks, achieving better performance.

## Attention Layers

Attention [12] is a mechanism used to model and exploit long-range correlations between features. This mechanism combines three components: the queries  $\mathbf{Q}$ , the keys  $\mathbf{K}$ , and the values  $\mathbf{V}$ . These components can be either distinct vectors or linear projections of the same input [245]. The process begins by comparing each query  $\mathbf{q}$  to the keys and computing a score value for each pair:

$$e_{\mathbf{q}, \mathbf{k}_i} = f_{\text{scoring}}(\mathbf{q}, \mathbf{k}_i) \quad (2.11)$$

The score value is computed by a scoring function  $f_{\text{scoring}}$  that changes based on the type of attention implemented. Then, a softmax function computes the attention weights based on the scores:

$$w_{\mathbf{q}, \mathbf{k}_i} = \sigma_{\text{soft}}(e_{\mathbf{q}, \mathbf{k}_i}) \quad (2.12)$$

Finally, the layer outputs the weighted sum of the values, weighted by the attention weights:

$$\mathbf{y} = \text{attention}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \sum_i w_{\mathbf{q}, \mathbf{k}_i} \mathbf{v}_{k_i} \quad (2.13)$$

Several attention variants have been proposed by changing the scoring functions [142, 270], the projections [19], or the applied steps. The attention layers are the fundamental blocks of the Transformer [245] architecture, which is the backbone of several state-of-the-art models.

### Memory Augmented Neural Networks

Memory Augmented Neural Networks (MANNs) are a class of neural networks designed to overcome the limitation of classical recurrent architectures like LSTM. Indeed, LSTMs suffer from the vanishing gradient problem. When sequences are particularly long, gradients are multiplied repeatedly during backpropagation through the recurrent connections. This multiplication causes the gradients to get smaller and smaller until they disappear. This problem leads to a limited capability of memorization and exploitation of information from the earlier steps of the sequence.

MANNs mitigate this problem by employing an external memory for storing information for longer periods of time. A MANNs is characterized by five elements: the controller, the memory, the write heads, the read heads, and the classifier. The *controller* is a DNN that takes an input and returns an output, usually in the form of a latent representation. If the input is a sequence, then the output of the controller for each step is given by the following equation:

$$\mathbf{h}_t = f(\mathbf{x}_t) \quad (2.14)$$

The *memory* is a matrix  $\mathbf{M}$  of dimensions  $N \times M$  that can be written or read by the network. Writing and reading operations are performed by functions called *write heads* and *read heads*. Most of the time, heads are based on attention mechanisms. Both of them usually depend on the output of the controller. In this way, the controller can learn how to use the memory. The output of writing operations is the updated memory:

$$\mathbf{M}_{t+1} = wh(\mathbf{M}_t, \mathbf{h}_t) \quad (2.15)$$

while the output of reading operations are readings  $\mathbf{r}$  whose shape depends on the specific design:

$$\mathbf{r}_t = rw(\mathbf{M}_t, \mathbf{h}_t) \quad (2.16)$$

Finally, the *classifier* yields the prediction by exploiting the readings from the memory and the controller output.

$$\mathbf{y}_t = f_{CL}(\mathbf{h}_t, \mathbf{r}_t) \quad (2.17)$$

MANN designs can include all or parts of the described elements. Only the controller and the memory are common to all the MANNs.

### Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are networks originally designed to deal with image data characterized by the presence of *convolutional layers* and *pooling layers*. The output of each neuron in a convolutional layer depends on three components: the filter, the receptive field, and the inputs it receives. The receptive field is the set of neurons connected to the current neuron. While in most DNNs, each neuron receives information from all the neurons of the previous layers, in the case of CNNs, each neurons receive the outputs of only a subset of neurons in the previous layer. The filters are the weights that connect the neuron to its receptive field. In CNNs, all the neurons in a layer share the same filters. Convolutional layers are designed to exploit spatial information. Indeed, neurons are organized in sequence, grids, or 3d volumes based on the dimension of the input. For

2d data, filters and receptive fields are grids and each neuron can be identified by the indices on the grid. In mathematical terms, the output of a neuron placed in position  $[i, j]$  is associated with a  $d_r \times d_r$  receptive field and a filter  $d_f \times d_f$

$$h_{i,j}^l = \sigma \left( \sum_{a=0}^{d_f-1} \sum_{b=0}^{d_f-1} w_{ab} h_{(i+k)(j+b)^{l-1}} \right) \quad (2.18)$$

where  $\sigma$  is a non-linear activation.

Generally, convolutional are interleaved with pooling layers, which reduce the dimensionality of the input they receive. The reduction consists of clustering a portion of the input and then combining the elements of each cluster into a single value. For example, a Max-Pooling layer yields the maximum values of each cluster as outputs. The clusters are computed using spatial information, for example, extracting multiple  $3 \times 3$  squares of adjacent values in the matrix the layer receives. Several designs of CNN architectures that interleave pooling and convolutional layers have been proposed in the literature. Among them, the thesis include experiments on ResNet [80], EfficientNet [234], MobileNet [207], GoogLeNet [232], DenseNet [92], ShuffleNet [281], and WideResNet [277].

### Graph Neural Networks

Graph Neural Networks (GNNs) are networks built to work with graph data. Their central element is the message-passing paradigm used in their graph layers, where each node recursively updates its representations by aggregating the ones of its neighbors [267]. How representations are updated and aggregated varies based on the layer's design. We report the aggregation and update functions used by the three architectures we use in this thesis: *Graph Convolutional Networks (GCNs)*, *Graph Attention Networks (GATs)*, and *Graph Isomorphism Networks (GINs)*.

GCNs aggregate latent representation of a node  $v$  by averaging the latent representations of its neighbor nodes:

$$\mathbf{h}^l = W \cdot (\{\mathbf{h}_u^{l-1} : u \in N(v)\}) \quad (2.19)$$

where  $W$  is the weight matrix of the convolution,  $N(v)$  is the set of neighbors of the node  $v$  and  $\mathbf{h}_u^{l-1}$  is the representation of the neighbor node  $u$  in the previous layer.

GATs [246] weight the convolutional aggregation via attention scores,

$$\mathbf{h}^l = \sum (\{\mathbf{a}(u, v) \cdot W \cdot \mathbf{h}_u^{l-1} : u \in N(v)\}) \quad (2.20)$$

where  $\mathbf{a}(u, v)$  is the attention weight between the nodes  $u$  and  $v$  and the other parameters are the same of GCN.

GINs [268] use learned parameters  $\epsilon$  to weight the aggregation performed by a Multi-Layer Perceptron (MLP).

$$\mathbf{h}^l = MLP((1 + \epsilon)h_v^{l-1} + \sum_{u \in N(v)} h_u^{l-1}) \quad (2.21)$$

Node and graph classification are two popular tasks in GNNs. In both cases, the input is usually a single graph. In *node classification*, the task is to predict multiple labels, each associated with a node in the graph (e.g., predicting properties of the nodes). In *graph classification*, the task is to predict a single output, usually the class or a general property of the whole graph. In the first case,

the output of the layers described above is directly fed to a classifier. In the latter, the output of the layers is first fed to a readout layer, such as global pooling layers, and then to the classifier.

## 2.2 Explainable AI

By analyzing the architectures presented in the previous section, we can note that the introduced components and advancements between one architecture and another are designed to exploit novel and more complex relations encoded in the data, thereby improving the capability of the networks. However, the combination of these advancements in complexity, the opacity of the training process, and the usage of raw data make the behavior of these architectures opaque. To address this issue, the XAI field aims to develop methods that can improve the explainability of artificial intelligence systems.

As in the case of the definition of intelligence, there is no universally accepted definition of explainability [66, 155, 200]. In this thesis, we use explainability as a general term to refer to all the methods that “*enable human users to understand, appropriately trust, and effectively manage the emerging generation of AI systems*” [73]. **We use the terms explainability and interpretability interchangeably throughout the text.**

There are several ways to classify XAI methods proposed in the field. Here, we focus on categories tailored to DNNs and this thesis. A first distinction separates *post-hoc* approaches and *self-interpretable DNNs* [10, 282]:

- **post-hoc** “*methods target models that are not readily interpretable by design by resorting to diverse means to enhance their interpretability*” Arrieta et al. [10];
- **self-explainable DNNs** are architectures that include components that can be used directly to ease the explanation of the results [12, 33, 132].

A second distinction, introduced in the previous chapter, separates between *intrinsic* and *extrinsic* methods:

- **intrinsic methods** aim to enhance the interpretability of deep models by looking at and leveraging the inner workings of the models;
- **extrinsic methods** focus on generating explanations for DL models by exploiting external means (e.g., gradients, surrogate models).

While often self-explainable DNNs are referred to as intrinsic methods and post-hoc as extrinsic ones, we prefer to keep the categories separated, given the proliferation of methods between these categories. For example, methods that look at attention patterns [2] in Transformers exploit only the components of the models, and thus, they are intrinsic. However, Transformers cannot be considered self-explainable architectures. Therefore, a method of this kind could be just referred to as a post-hoc intrinsic method.

A further distinction [3] can be done by separating *local* and *global* methods:

- **local** methods provide explanations that are valid for a limited set of data (usually a single prediction), and thus, their explanations do not generalize to other components;

- **global** methods explain a model’s whole logic (usually approximating average outcomes).

Local methods are more faithful and precise in explaining single predictions but do not offer a guarantee of generalizability. On the other end, global methods are effective in providing an overview of the main learned relations but tend to be less precise and are less reliable in explaining individual components or predictions.

### 2.2.1 Categorization

XAI methods can also be split based on the type of explanations they return. While several XAI taxonomies have been proposed in the literature to perform such a split, this section reports a simplified categorization [121] by grouping the types of methods discussed in the thesis. Each category is described in terms of the goals of its methods and explanation objects.

#### Feature attribution

Feature attributions are scores assigned to each input feature or group of features representing the impact of that item on the network’s decision process. Methods focusing on feature attributions provide answers to questions like “Why does the model return this specific output?” and “What are the most relevant features exploited by the model to recognize objects in this class?”. Attributions can be either global or local. Global feature attribution measures the importance of features on average. Local feature attribution measures the importance of features exclusively for the current prediction. Scores are usually normalized and visualized as heat maps or numbers.

#### Learned Features

Learned features are features or concepts neurons, groups of neurons, or layers recognize. In the context of this thesis, a **concept** is intended as a set of semantically connected features annotated in a dataset (e.g., the concept of a wheel in a car). They address the question “What has this component learned during the training process?” and thus provide global post-hoc explanations. They can be visualized as samples exhibiting the learned knowledge or by exposing groups of features connected to such knowledge.

#### Explanations by Example

Explanations by examples are samples similar to the item that the user wants to explain and associated with the same prediction. Typically, the samples are extracted from the training dataset and share most features with the object to explain. Inspecting explanations by example helps users extract patterns exploited by the models. Explanations by examples can be applied to extract both global and local information. For example, when applied to the input of the model, explanations by example provide local information, highlighting features exploited by the model for the specific input. Conversely, these explanations provide (global) insights into the learned dynamics when applied to latent representations (e.g., as shown in Chapter 4).

## Counterfactuals

Counterfactuals are samples representing alternative configurations of input features. A counterfactual is a sample as close as possible to the input but associated with a different prediction. Samples can be either generated or extracted from a dataset. Methods of this category aim to find the minimum magnitude of meaningful edits needed on the current input sample to obtain a different prediction. These explanations are particularly useful for recourse and, most of the time, are local methods.

### 2.2.2 Evaluation of XAI methods

One of the open questions of the XAI field regards the evaluation of explanation methods. Indeed, there is no ground truth as the real inner workings of the model are unknown. We can distinguish between three categories of evaluation: user-study, datasets, and proxy metrics.

*User studies* involve users testing explanations in a real-world task and providing feedback. This form of evaluation was among the first to be explored. We identify three common workflows supporting user studies. The first consists of presenting users with multiple explanations and asking them to rank their preferences, which are then used to measure the quality of the tested methods. The second one consists of providing explanations to the users to assist them while solving a task, followed by collecting their feedback as a qualitative evaluation. Lastly, the third workflow permits users to interact with explanations and tasks at the same time. This task can be facilitated through the usage of interactive interfaces like dashboards, dialogue systems, and visual analytics systems. At the end of the interactive session, both quantitative and qualitative metrics and feedback can be collected.

Evaluations based on user studies, and particularly the interactive ones, can be used to evaluate the usefulness of the explanations, their impact on specific tasks, and the preference regarding explanation visualization [48]. However, these evaluations are challenging to reproduce, causing problems in the comparison between methods proposed at different times. Moreover, user satisfaction is more influenced by their belief and mental models than the precision of the explanations in capturing the true behavior of the DNN. Therefore, involving users should occur only after explanation methods have demonstrated a connection to the real decision process of the network. Otherwise, explanations can easily cause overtrust and mislead the user [99].

In an effort to establish more reproducible environments and link the evaluations to the decision process, *Dataset evaluations* propose to use datasets with available ground truth explanations [137] to evaluate explanations. These evaluations draw inspiration from the common evaluation practices applied in machine learning and discussed in Section 2.1. However, this form of evaluation is feasible only for a subset of methods and simplified tasks where the rules governing the task are known and there is only one unique way to solve the task.

The third choice involves using a *mathematical proxy* to compare explanation methods quantitatively. They are not intended as replacements for user evaluations but as a necessary prerequisite for selecting the appropriate set of explanation methods. Although several proxy metrics have been proposed to measure the quality of the explanations quantitatively in recent years, there is no standard global metric yet, and each explanation type and context has its own set of metrics.

**This thesis mainly uses proxy metrics to evaluate the proposed methods.** The metrics

are chosen among each category's most recent and popular ones. When available (Chapter 5), we support the evaluation on toy datasets where ground truth can be retrieved and the models satisfy the appropriate requirements (e.g., perfect accuracy). We also discuss the interactive systems supporting user studies in Chapter 8. The following paragraphs list the metrics used to evaluate the quality of the explanations for each category presented in the previous section.

### Feature Attributions.

**Definition 2.2.1.** The ***fidelity+prob*** [275] metric measures the difference in probability of the original prediction and the prediction generated when the most relevant features are masked.

$$Fidelity^{+prob} = \frac{1}{N} \sum_{i=1}^N (f(x_i) - f(\theta_1(x_i))) \quad (2.22)$$

where  $\theta_1(x)$  is a function that mask  $x$  by removing the most important features. Higher is better. This metric is equivalent to the deletion score proposed in [185].

**Definition 2.2.2.** The ***fidelity+acc*** [275] metric measures the difference in the accuracy of the original prediction and the prediction generated when the most relevant features are masked.

$$Fidelity^{+acc} = \frac{1}{N} \sum_{i=1}^N (\arg \max f(x_i) = \arg \max f(\theta_1(x_i))) \quad (2.23)$$

where  $\theta_1(x)$  is a function that mask  $x$  by removing the most important features. Higher is better. Fidelity can be also computed by using the ROC-AUC instead of the accuracy: in this case, the metric is denoted as *Fidelity+ROC-AUC*.

**Definition 2.2.3.** The ***fidelity-acc*** [275] metric measures the difference in accuracy between the original prediction and the prediction generated when the least relevant features are masked.

$$Fidelity^{-acc} = \frac{1}{N} \sum_{i=1}^N (\arg \max f(x_i) = \arg \max f(\theta_2(x_i))) \quad (2.24)$$

where  $\theta_2(x)$  is a function that mask  $x$  by removing the least important features. Higher is better.

### Learned Features.

**Definition 2.2.4.** The **intra-concept similarity** [35] metric measures the mean pairwise cosine similarity between samples of the same concept.

$$d_{ii} = \frac{1}{n^2} \left( \sum_{j=1}^n \sum_{k=1}^n \frac{\mathbf{x}_{ij} \cdot \mathbf{x}_{ik}}{\|\mathbf{x}_{ij}\| \|\mathbf{x}_{ik}\|} \right) \quad (2.25)$$

where  $j$  is the index of the sample,  $i$  is the index of the concept,  $\mathbf{x}_{ij}$  is the latent representation for the sample  $j$  of the concept  $i$ , and  $n$  is the number of samples including the concept  $i$ .

**Definition 2.2.5.** The **inter-concept similarity** [35] metric measures the mean pairwise cosine similarity between samples of two different concepts.

$$d_{pq} = \frac{1}{nm} \left( \sum_{j=1}^n \sum_{k=1}^m \frac{\mathbf{x}_{pj} \cdot \mathbf{x}_{qk}}{\|\mathbf{x}_{pj}\| \|\mathbf{x}_{qk}\|} \right) \quad (2.26)$$

where  $p$  and  $q$  are the index of the two concepts, and  $m$  and  $n$  are the number of samples for concept  $p$  and  $q$ , respectively.

**Definition 2.2.6.** The **separability** [35] metric measures the separability of two concepts in a latent space. It is expressed as the ratio between intra and inter-concept similarities.

$$Sep_{ij} = \frac{d_{ij}}{\sqrt{d_{ii}d_{jj}}} \quad (2.27)$$

A lower score is considered better.

**Definition 2.2.7.** The **intersection over union (IoU)** [17] metric measures the alignment between a concept annotation and the firing rate of a latent representation.

$$IoU(L, \mathfrak{D}) = \frac{\sum_{x \in \mathfrak{D}} |M^k(x, i) \cap C(x, c)|}{\sum_{x \in \mathfrak{D}} |M^k(x) \cup C(x, c)|} \quad (2.28)$$

where  $k$  represents the index of a neuron,  $c$  represents a concept,  $M^k(x, i)$  is a function that returns the binary mask that indicates the parts of the input on which the neuron fires, and  $C(x, c)$  corresponds to a function that returns a binary mask that represents which parts of the input are associated with the concept  $c$ .

### Explanations by Examples.

**Definition 2.2.8.** The **input non-representativeness** [106] metric measures the  $L_1$  distance in the logits obtained by feeding the current input and the explanation by example to the model.

$$Inr = L^1(f(x), f(x_e)) \quad (2.29)$$

A lower input non-representativeness means the explanation is of better quality.

**Definition 2.2.9.** The **prediction non-representativeness** [188] metric measures the cross entropy loss between the current predicted class and the class predicted by using the explanation by example as input.

$$Pnr = CrossEntropy(y_x, y_{x_e}) \quad (2.30)$$

A lower prediction non-representativeness means the explanation is of better quality.

### Counterfactuals.

**Definition 2.2.10.** Given a predicted class  $y_x$ , a counterfactual class  $y_c$ , an autoencoder  $f_{y_x}$  trained on reconstructing samples of the predicted class, and an autoencoder  $f_{y_c}$  trained on reconstructing samples of the counterfactual class, the **IM1** [138] metric measures the ratio between errors of the two autoencoders to reconstruct the counterfactual.

$$\frac{\|x_{cf} - AE_{y_c}(x_{cf})\|}{\|x_{cf} - AE_{y_x}(x_{cf})\|} \quad (2.31)$$

A low IM1 score means the counterfactual is closer to the counterfactual class data distribution than the input one and it is of a better quality. This is useful for methods that start from the input and perturb it to generate a counterfactual.

**Definition 2.2.11.** Given a predicted class  $y_x$ , a counterfactual class  $y_c$ , an autoencoder  $f_{y_x}$  trained on reconstructing samples of the predicted class, and an autoencoder  $f_{y_c}$  trained on reconstructing samples of the counterfactual class, **IIM1 score** [138] measures the ratio between errors of the two autoencoders to reconstruct the counterfactual.

$$\frac{\|x_{cf} - AE_{y_x}(x_{cf})\|}{\|x_{cf} - AE_{y_c}(x_{cf})\|} \quad (2.32)$$

IIM1 metric [124] is the inverted version of the IM1 score. A low score means that the counterfactual is closer to the input's class data distribution than the counterfactual's. This is useful for methods that select counterfactuals from a pool of samples associated with different predictions.

**Definition 2.2.12.** Given a predicted class  $y_x$ , an autoencoder  $f_{y_x}$  trained on reconstructing samples of the predicted class, and an autoencoder  $f_{all}$  trained on reconstructing samples of all the classes in the dataset, the **IM2** [138] metric measures the difference in the reconstruction errors of the counterfactual between the autoencoder trained on the input classes and the one trained on all the classes.

$$\frac{\|AE_{y_x}(x_{cf}) - AE_{all}(x_{cf})\|}{\|x_{cf}\|_1} \quad (2.33)$$

A low IM2 score means the data distribution of the counterfactual class describes the counterfactuals as good as the distribution over all classes and the counterfactual can be considered interpretable. Historically, it is a score proposed to evaluate the usefulness of a generated counterfactual.

## 2.3 Visual Analytics

The XAI field is not the sole field that is actively working on supporting users to understand DNNs. Indeed, according to Choo and Liu [37], VA systems are playing “*a critical role in enhancing the interpretability of deep learning models, and it is emerging as a promising research field*”. This section lists the main components used in the Chapter 8 to analyze the VA systems aiming to help users understand DNNs through XAI methods.

We begin by enunciating the definition of VA: it is the science of analytical reasoning supported by interactive visual interfaces [238]. A VA system helps users synthesize and derive insights from data by detecting the expected relationships, discovering the unexpected ones, and communicating these findings to the human user for further actions [116].

The VA process [104, 203] can be divided into four components:

- **data** are the starting point of the system and, in the context of this thesis, correspond to inputs, outputs, activations, explanations, etc.;

- **model** applies transformations to the data and, in the context of this thesis, corresponds to the DNN;
- **visualization** is the interface between users and the system that allows the detection and discovery of relationships and insights on the data and model;
- **knowledge** is the user-driven component and “*consists in finding evidence for existing assumptions or learning new knowledge about the problem domain*” La Rosa et al. [121]

The user interacts with the VA system through interfaces and interactions [240]. Interactions allow the user to analyze data, generate new visualizations, and steer the analyzed model [53, 87]. Interactions represent the means by which the users can achieve the desired goal and help to create a better mental model of the investigated problem. Given a trained model loaded in a VA system, we can distinguish between three types of interactions [121]:

- **passive observations** involve only the navigation across data (samples, layers, explanations, etc.) in terms of filtering and selection;
- **interactive input observations** involve the modification or creation of novel inputs and feeding them to the model to evaluate changes in the models’ response;
- **interactive model observations** involve the modification of the deep learning model components to evaluate changes in the models’ response;

Finally, VA systems are tailored to specific target users and tasks. Strobelt et al. [227] propose the following classification for target users:

- **architects** are advanced DL experts whose goal is developing new architectures or modifying the design of the current ones;
- **trainers** are users (with sufficient background knowledge in DL) whose goal is to apply existing architectures to new domains, tasks, and datasets;
- **end-users** have limited or no DL knowledge and use already trained models in their specific application domain.

Note that end-users cannot use a system built for the other categories, while all the other categories of users can use systems designed for end-users.

# Chapter 3

## Related Work

This chapter reviews existing methods for explaining Deep Neural Networks (DNNs). By utilizing the categorization introduced in Section 2.2, this chapter discusses post-hoc methods and self-explainable DNNs. Since the chapter cannot provide an exhaustive overview of the state-of-the-art methods for the whole XAI field, the review of post-hoc methods focuses attention on methods that have already been applied to DNNs, are data agnostic, and related to the approaches proposed in the next chapters. The chapter is organized as follows: Section 3.1 reviews post-hoc methods for deep learning, Section 3.2 reviews current self-explainable DNNs, and section 3.3 discusses the relation between the reviewed approaches and the methods proposed throughout the thesis.

### 3.1 Post-Hoc Methods

This section briefly describes the state of the art of post-hoc methods for the categories described in Section 2.2. It focuses only on methods tailored for DNNs or explicitly tested on them. For a deeper description of eXplainable Artificial Intelligence (XAI) methods in machine learning, please refer to the recent surveys on the topic [3, 66, 72].

#### 3.1.1 Feature Attribution

Feature attribution methods assign scores to every input feature based on their relevance to a set of predictions. It is one of the oldest and most extensively explored research areas in the XAI field.

Gilpin et al. [66] distinguish between methods based on perturbations, backpropagations, gradients, and surrogate models. The concept behind perturbation-based methods is straightforward: modify the value of a feature (or a set of features), input the perturbed instance into the model, and collect the results. Scores are computed by iteratively repeating this process and measuring the deviation between the original prediction and the predictions of perturbed instances [1, 279, 291]. An optimization process can be used to guide the generation of meaningful perturbations [59, 273].

Approaches utilizing surrogate models, such as LIME [197] and SHAP [140], also follow the idea of using perturbations. However, in this case, perturbations are not directly used to estimate attributions but serve as samples for training a local interpretable surrogate model. The surrogate model learns to mimic the black-box model’s decision process in the samples’ neighborhood to be explained. The weights learned by the surrogate model are then used as feature attribute scores. However, these approaches are slow, especially with a large number of features, as in the case of

DNNs. Additionally, due to the nonlinear nature of DNNs, the result is influenced by the number of features removed altogether at each iteration and the number of permutations.

Gradient-based methods offer a faster alternative to perturbation-based methods by considering the partial derivative of the target output as scores for feature attributions [216]. However, vanilla gradients produce noisy explanations. To address this problem, several works propose enhancements such as multiplicative terms to the gradients [30, 209, 214, 224], the use of integrals and baselines [103, 231, 269], or smoothing functions [25, 222]. While popular for their simplicity, these methods tend to be loosely linked to the decision process [4] and suffer from the gradient-shattering effect.

Backpropagation methods [11, 214, 279] and forward propagation methods [68] work similarly to gradient-based methods but employ custom rules and quantities instead of gradients. For instance, LRP [11] computes relevance for each neuron of the network during input parsing and then back-propagates the relevance and prediction to compute attribution scores. DeepLIFT [214] extends this idea by considering baseline inputs as proposed in gradient-based methods.

In parallel with these approaches, there has been recent interest in computing scores for sets of semantically connected features (i.e., *concepts*) rather than individual features. Particularly, TCAV [111] and its derivatives [57, 63] investigate the impact of concepts on the decision process. These methods collect samples with and without the target concept, build a hyperplane to separate these samples, and use directional derivatives with respect to this hyperplane as attribution scores. The sensitivity to the collection of samples in terms of diversity and number of samples and the computational time required to probe for several concepts represent the major limitations of these approaches.

Finally, all the methods described thus far explain individual predictions. To obtain global explanations, researchers propose various methods that aggregate attribution scores of individual samples, either by averaging or clustering them [126, 165, 206].

### 3.1.2 Learned Features

This category of methods focuses on extracting information regarding what individual components (e.g., layers, neurons) of neural networks have learned during the training process. Research in this area has explored two main directions [27, 66, 204]: feature synthesis and feature probing using external datasets.

Features synthesis involves generating synthetic explanations through iterative processes [54, 175] or using external models [169, 171]. These processes are typically tailored to maximize (or minimize) an objective function related to the component to be explained. For instance, in the case of neurons, several approaches aim to generate synthetic inputs that maximize the activation of a specific neuron [54, 144, 175, 217]. The general process consists of generating this synthetic input by iteratively altering the features of a starting random input. However, vanilla maximization of the activation can yield abstract explanations that are difficult to interpret. To address this issue, several regularizers have been proposed to increase variance [144] or impose constraints on the generation process [58, 171, 175, 254, 274].

Nonetheless, these methods encounter several challenges. For example, the stochastic nature of the process [161] may produce different explanations for the same neuron. Additionally, while neurons can recognize multiple features due to superposition [23, 52], these methods tend to converge

towards one or a few concepts, overlooking the multifaceted nature of these components [170]. Furthermore, despite their popularity and the advancements to reduce abstractness, humans often struggle to comprehend these explanations and instead prefer feature probing explanations [22, 290].

Feature probing methods employ real samples to represent the learned features. In this category, the natural counterpart of activation maximization methods is to select samples from the dataset that maximize the neuron’s activation [27]. In contrast to the activation maximization method, the selected samples are not abstract and naturally exhibit diversity. However, the connection between high activations and the explanation is weaker. Indeed, it is often unclear if the cause of the high activation is the whole sample or specific elements within it. While providing more samples can mitigate this issue, an excessive number of samples can diminish the usefulness of the explanation.

Bau et al. [17] propose addressing this issue using a concept-annotated dataset. In this case, the algorithm selects the concept that maximally activates a neuron, removing factors unrelated to the high activation. Similar works reverse the process by fixing a concept and searching for the neurons associated with the highest activations when the concept is included in the samples [9, 41, 51, 83, 164]. Like the generative approaches, associating a single concept with a neuron is inadequate due to superposition. Therefore, successive works [26, 76, 150, 163] propose associating logical formulas of concepts with neurons by finding the formula whose annotations are the most aligned with the highest activations through search algorithms.

One weakness of feature probing approaches is their dependency on concept-annotated datasets. Indeed, the same neuron can be associated with two different explanations simply by changing the annotated dataset [192]. Moreover, the analysis of learned features is constrained by the annotated concepts and the quality of annotations. Recent works suggest replacing the concept-annotated dataset with captioning or multimodal models [84, 173] to mitigate this dependency. However, these models require human annotations during training and are susceptible to the domain adaptation problem, thus shifting the dependency from dataset to model aspects.

### 3.1.3 Explanations by examples

Explanations by examples are samples similar to the current input and associated with the same prediction. Therefore, one of the first explored research directions was providing samples that approximate data prediction distribution as explanations. These approaches [74, 110] rely on clustering the data, identifying prototypes for each cluster, and using them as explanations. While effective for small tabular datasets, this approach falls short in providing descriptive prototypes for clusters in large datasets.

Conversely, post hoc methods for explanations by examples dealing with DNNs and large datasets typically employ case-based reasoning approaches. These approaches utilize a proxy model, such as a K-NN classifier [39], to explain the black-box by learning a mapping between them. Approaches of this category differ in the information provided to the proxy. The simplest approach is to directly use the training samples to train the proxy. However, since typically that not all features are equally important for a given model, several approaches propose weighting the features to yield more informative explanations [172, 183]. For example, a popular technique involves weighting each feature in the sample using feature attribution scores, thereby linking the explanations to the important part of the input [5, 105, 106]. Lastly, another viable approach is to use the activations of the last layers of a DNN as samples to train the proxy [179] or a meaningful subset of them [108].

### 3.1.4 Counterfactuals

Counterfactuals are samples similar to the current input but associated with a different prediction. Finding counterfactuals for DNNs dealing with non-tabular data is challenging. Indeed, DNNs work with raw data, where there are no formal constraints on the position and values of the features, and the number of features involved is large. Therefore, it is challenging to generate plausible counterfactuals. Moreover, the decision process of these networks is extremely sensitive, and it is often possible to obtain a different prediction using adversarial attacks that change the value of a few features in a way that is not discernible to a human. While these modifications align with the definition of counterfactuals (Section 2.2.1), they are often meaningless and cannot be considered explanations. Conversely, counterfactual explanation methods aim to provide counterfactuals that are both plausible and meaningful.

The pioneer work of this category is the method proposed by Wachter et al. [251]. This method generates counterfactuals by an iterative perturbation process guided by a loss function that minimizes the difference between the predictions on the perturbed instance, the desired outcome, and the L1 norm of the perturbations. Successive works propose alternative losses that take into account additional factors like the closeness of features [46, 127], plausibility [271], or the distance from a set of prototypes [138]. One of the drawbacks of these methods is their latency due to the iterative process. One possible solution to mitigate such an issue is to use generative models [16, 107, 135] or genetic algorithms [212]. However, since these procedures are black-boxes themselves, it is difficult to understand why a particular counterfactual has been selected as a good candidate.

Another possibility is to remove features from the current input until the prediction is flipped [193, 247] by guiding the process with feature attribution methods. However, feature attribution methods focus on the feature important for a specific instance and do not detect discriminative features that the model uses to discriminate among classes, limiting the applicability of these approaches. Connected to this line of research, some works aim at extracting the contrastive features [67, 101, 260], which are highly discriminant features for a class and uninformative for the others [72]. These methods can be considered complementary to methods that select counterfactuals from a dataset based on user-specified properties and classes [189].

## 3.2 Self-Explainable Deep Neural Networks

While post-hoc methods represent the most popular tool for explaining DNNs, recently, the field has observed a rising interest in the so-called self-explainable DNNs. These networks return explanations alongside their predictions or provide designs that can be easily inspected to provide explanations. The advantage of these methods is that explanations are fast to compute and are directly linked to the decision process of the networks. Moreover, they represent the natural next step for the progress in machine learning. Self-explainable DNNs can be divided into three categories: prototype-based, constraints-based, and attention-based.

**Prototype-based networks** have been introduced by Snell et al. [223] to deal with few-shot classification. In their networks, prototypes are computed as the average of the learned embedding of a set of data points. However, since the prototypes are the average of multiple embeddings, they are not interpretable. The concept of prototypes has been merged with the concept of self-interpretable neural networks [8] in ProtoPNet [33]. In this case, the network learns a set of prototypes repre-

senting a *part* of the input (i.e., a set of features semantically linked). Given an input, the network compares the input’s latent representation against the learned prototypes and computes the prediction based on the similarity between the prototypes and the input representation. While in the networks proposed by Snell et al. [223], prototypes are not associated with semantic meaning, ProtoPNet enforces semantics by using losses and a projection phase that projects the prototypes to real training points. ProtoPNet has been recently extended to enforce diversity [256], a better prototype organization [77, 166, 201, 202], and semantic correspondence [167] of prototypes. Networks inspired by the ProtoPNet design have also been proposed in reinforcement learning [109, 191], sequence classification [156], healthcare [160], and graph classification [283].

The interpretability is ensured by the fact that the prototypes are real samples and can be visualized. Moreover, the classification layer is easily interpreted since the predictions are based on a weighted average of the prototypes’ activations. By comparing the input and the most activated prototypes, a user can extract insights about important features in the input, similar to feature attribution. Conversely, by extracting samples close in the latent space to the prototypes, a user can extract global explanations by example, both positive and negative [218] (i.e., this does not look like that prototype).

**Constraint-based architectures** encourage the network to learn more interpretable representations in the form of disentangled or concept-aligned representations [35, 115, 244]. Concept Bottleneck Modelss (CBMs) [115] is an example of such architecture. Given a latent representation of a sample, CBMs are trained to predict the presence or absence of a set of pre-defined concepts in the sample. The probabilities of all the concepts are then combined to classify the sample. The idea is that a user can look at the probabilities associated with the set of concepts to understand which ones are influencing the most the prediction. Moreover, the user can also intervene by changing the probability of a concept to the desired one. Several works extend CBMs by improving the concept representations [78], the performance [55, 276], and the induced bias [174, 272]. Nonetheless, the induced bottleneck limits the expressiveness of the network, and thus their performance is usually lower than the black-box counterparts.

An alternative approach is to enforce the latent disentanglement without introducing the bottleneck [35, 139, 229, 244]. This goal can be achieved either using additional semantic layers that project the latent representation in another more aligned subspace [139] or using a normalization layer, as proposed in Concept Whitening [35]. Despite the progress and improvements in performance, several open problems are connected to the usage of these networks, such as concept leakage and dependencies related to the concept dataset or models.

**Attention-based architectures** use attention modules to improve the performance and the interpretability of the decision process [61, 129, 285]. The idea is that since attention is a weighted sum of the vector representations, attention weights can be used directly as feature attribution scores as long as the vector representations are meaningful. Attention has been used to discover relations such as the coreference and syntax in natural language processing tasks [38, 235, 249] and modality relations in multimodal models [32]. For architecture employing multiple attention layers, Abnar and Zuidema [2] propose a post-hoc intrinsic method to reconstruct the flow of attention weights along the network and compute feature attribution based on the flow.

Several works in literature explored the conditions under which attention can be considered as a reliable proxy for explanations [265]. For example, Wiegrefe and Pinter [264] and Serrano and Smith

[211] find altering the attention weights does not affect the predictions of some configurations of attention-based architectures, casting doubts on their reliability. In this case, the problem is caused by input dispersion, a phenomenon connected to the accumulation of information from different sources in a single vector representation.

Finally, apart from these general categories, there are several other self-explainable DNNs tailored for specific tasks or domains [40, 71] that combine concepts of the previous categories or propose novel ones [34, 61, 195, 198], lacking however in generalizability.

Overall, since the area of self-explainable DNNs is relatively recent and emerging, several open challenges exist to address. Among them, we can mention the limited generalizability since most of the approaches are tested only on specific families of architectures, the limited diversity among the approaches, the accuracy-explainability trade-off, and the bias induced by their design.

### 3.3 Relations with Thesis Contribution

This thesis contributes to the research efforts in explaining DNNs outlined in the previous section by: (i) proposing novel designs for self-explainable DNNs; (ii) advancing the understanding of features learned by DNNs; and (iii) discussing the integration of visual analytics and explanation methods to deliver explanations to users.

The proposed designs of self-explainable DNNs address issues related to limited generalizability and diversity, to the usage of custom training recipes, and the accuracy-explainability trade-off. While existing self-explainable DNNs are primarily applied to vision tasks and convolutional neural networks (Section 3.2), our contributions extend to novel domains and architectures. Specifically, our proposed designs based on concepts and prototypes (Chapter 4 and Chapter 6) relax common assumptions, like the availability of pre-trained models, and can be applied to various families of graph neural networks, an area often overlooked by state-of-the-art approaches. The generalization contribution is directed not only to the type of data and the families of covered architectures but also to the type of normalization layers, an unexplored area in the current literature.

While these designs contribute to the literature in terms of generalization, they reiterate prototypes and concept-based architectures and use custom training recipes to train the models, following the current literature. However, using recipes different from those used to train the black-box models can cause instability in the training process when applied to settings where the models use non-standard recipes. These issues are addressed by the proposed designs of self-explainable DNNs based on memory modules (Chapter 5). Exploiting memory modules for explainability purposes represents a novelty in the field, increasing the diversity of approaches. These designs can be directly applied to black-box models following their training settings. The combination of memory modules and attention mechanisms mitigates the problem of input dispersion, typical of attention-based models, and is more flexible than current popular designs since it can provide proxies for feature attribution, explanations by example, and, for the first time for a self-explainable DNNs, for local counterfactuals.

The designs mentioned above introduce novel architectures replacing the current black-box models. Conversely, the third and fourth parts of the thesis contribute to the research about explaining black-box DNNs without modifying them.

The Part III advances the state-of-the-art feature probing methods by investigating a wider

spectrum of neuron behavior than the one investigated in past research (i.e., only the highest activations) and proposing metrics to evaluate such methods. Specifically, the proposed method overcomes the issues related to the computational complexity of the previous approaches [163] and provides explanations that better capture the superposition of neuron activations [52]. Additionally, due to the wider analyzed spectrum, the thesis contributes to this area by shedding light on novel overlooked activation phenomena connected to the activations of popular DNNs.

Finally, Part IV contributes to the literature by reviewing visual analytics systems as a means to provide explanations computed by XAI methods to the user. This effort targets exclusively Visual Analytics (VA) systems dealing with DNNs and using XAI methods to enhance their interpretability, and it represents the first effort in this direction. Indeed, current VA literature [7, 89] targets more general settings, and XAI literature has limited awareness of these tools. Therefore, it is unclear whether the current VA solutions are effective and what the strengths and weaknesses of the current literature are. Both these contributions, together with increasing the awareness of the XAI community and paving the steps for increasing the collaborations, are discussed in the Part IV.

## Part II

# Self-Explainable Neural Networks

## Chapter 4

# Prototype-based Interpretable Graph Neural Networks

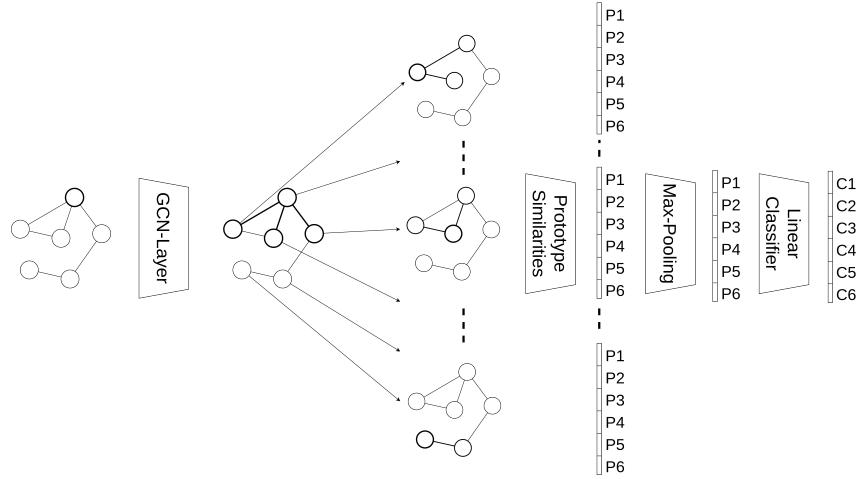
This chapter represents the entry point of the thesis contribution on self-explainable designs for Deep Neural Networks (DNNs). As mentioned in Section 1.1, these designs aim to enhance the interpretability of DNNs and help developers to debug and improve their models. Contextually, these designs aim to reduce the typical performance trade-off of self-explainable DNNs and broaden the applicability of such methods. The idea shared by **all the techniques proposed in this part of the thesis is to insert the novel components into the DNN without disrupting its structure and preserving most of its representation power**. Therefore, the proposed layers are placed in between the feature extractor  $f_{FE}$  and the classifier  $f_{CL}$  (Section 2.1).

As a first step, this chapter proposes a **self-explainable prototype-based layer for Graph Neural Networks (GNNs)** inspired by the design of prototype-based architectures [33] commonly used for image classification. Specifically, the resulting architecture, called Prototype-based Interpretable Graph Neural Network (PIGNN), learns a set of prototypes representing nodes and uses the similarity between the node representations of the current input and the learned prototypes to perform both graph and node classification. This design enhances the interpretability of black-box GNNs, supporting developers in understanding and improving their models. Indeed, this design supports the identification of pitfalls of the dataset or in the training setup as well as error analysis through explanations by examples and feature attributions.

The chapter is organized as follows: Section 4.1 describes the proposed prototype layer and how users can probe its behavior; Section 4.2 evaluates PIGNN both in terms of performance and explanations quality; Section 4.3 discusses alternative design choices for PIGNN; finally, Section 4.4 summarized the contributions of this chapter.

### 4.1 Design

As previously mentioned, the **PIGNN** design introduces a layer after the feature extractor  $f_{FE}$  to enhance interpretability. This layer includes a set of  $m$  learned vectors  $\mathbf{P} \in \mathbb{R}^{m \times emb} = \{p_j\}^m$ , termed *prototypes*. **Each prototype is optimized during training to represent a node embedding** of dimension  $emb$ . Given an input  $\mathbf{x}_i$ , the feature extractor returns the latent representation  $\mathbf{h} = f_{FE}(\mathbf{x}_i) \in \mathbb{R}^{n \times emb}$ , where  $n$  is the number of nodes in the graph input. The latent



**Figure 4.1:** The architecture of PIGNN when the task is graph classification.

representation is then compared against the full set of prototypes, producing a score that estimates their similarity:

$$\mathbf{s} = \text{Sim}(\mathbf{h}, \mathbf{P}) \in \mathbb{R}^{n \times m} \quad (4.1)$$

Each prototype is associated with  $n$  scores, reflecting how strongly the semantic encoded in the prototype is represented in each node embedding. The layer employs global Max Pooling (Section 2.1) to select the maximum score for each prototype:

$$\mathbf{a}_i = \max_{n \in N_p} (\mathbf{s}) \in \mathbb{R}^{n \times m} \quad (4.2)$$

The resulting scores can be regarded as how much the current graph input activates each prototype. In the case of graph classification, the classifier takes the activation scores as input to make predictions (Figure 4.1):

$$\mathbf{y}_i = f_{CL}(\mathbf{a}_i) \quad (4.3)$$

For node classification, the classifier directly utilizes the similarity scores as inputs, as each node is associated with a prediction:

$$\mathbf{y}_i = f_{CL}(\mathbf{s}_i) \quad (4.4)$$

#### 4.1.1 Training Process.

PIGNN, like all ProtoPNet [33] variants, **requires a custom training recipe**. Each training epoch comprises three phases: weight optimization, prototype projection, and classifier optimization.

**Weight Optimization.** This phase adjusts the learned latent space of the features extractor to cluster relevant subgraphs around the prototypes and to separate prototypes from each other. To achieve the goal, the feature extractor and the prototype layer are optimized by minimizing the

following loss [256]:

$$L = L_{pred} + \lambda_1 L_{cluster} + \lambda_2 L_{orth} - \lambda_3 L_{class\_separation} - \lambda_4 L_{separation} \quad (4.5)$$

and  $\lambda_1, \lambda_2, \lambda_3, \lambda_4$  are hyperparameters that balance the weight of the losses. Here,  $L_{pred}$  is the task loss used to optimize the model (e.g., cross-entropy between prediction and ground truth):

$$L_{pred} = \frac{1}{n} \sum_{i=1}^n \text{CrossEntropy}(f(x_i), y) \quad (4.6)$$

$L_{cluster}$  encourages subgraphs important for similar predictions to cluster together:

$$L_{cluster} = \frac{1}{n} \sum_{i=1}^n \min_{j: \mathbf{p}_j \in P_{y_i}} \min_{\mathbf{h} \in \text{khop\_subgraph}(x_i)} d(\mathbf{h}, \mathbf{p}_j) \quad (4.7)$$

where  $\text{khop\_subgraph}(x_i)$  is the set of all the subgraphs of the graph  $x_i$  such that they are centered in one of the nodes in  $x_i$  and include all the neighbors of that node at a maximum distance equal to  $k$  and  $d(\mathbf{h}, \mathbf{p}_j)$  is the distance between the latent representation and the prototype  $\mathbf{p}_j$  computed as:

$$d(\mathbf{h}, \mathbf{p}_j) = \frac{\mathbf{h}^T \mathbf{p}_j}{\|\mathbf{h}\|} \quad (4.8)$$

$L_{orth}$  enforces the orthonormality of prototypes, avoiding prototypes converging toward the same semantics:

$$L_{orth} = \sum_{k=1}^K \|P_k P_k^T - \mathbb{I}\| \quad (4.9)$$

$L_{class\_separation}$  encourages separation between prototypes associated with different classes:

$$L_{class\_separation} = \frac{1}{\sqrt{2}} \sum_{k_1=1}^{K-1} \sum_{k_2=k_1+1}^K \|P_{k_1} P_{k_1}^T - P_{k_2} P_{k_2}^T\| \quad (4.10)$$

Finally,  $L_{separation}$  promotes differentiation between prototypes associated with other classes:

$$L_{separation} = \frac{1}{n} \sum_{i=1}^n \min_{j: \mathbf{p}_j \notin P_{y_i}} \min_{\mathbf{h} \in \text{khop\_subgraph}(x_i)} d(\mathbf{h}, \mathbf{p}_j) \quad (4.11)$$

Note that during this phase, the weights of the classifier  $f_{CL}$  are frozen. The total loss is a generalization of the ones proposed in Chen et al. [33], which can be recovered by setting  $L_{orth}$  and  $L_{class\_separation}$  to zero.

**Prototype Projection.** Since the first phase can produce out-of-distribution or unrealistic prototypes, this phase maps each prototype to the closest subgraph in the training dataset:

$$\mathbf{p}_j = \arg \min_{\mathbf{h} \in FE_j} \|\mathbf{h} - \mathbf{p}_j\| \quad (4.12)$$

where  $FE_j$  is the set of node embeddings returned by the feature extractor for all the samples in the dataset of the same class of the prototype  $\mathbf{p}_j$ . This phase is crucial to keep the prototypes

interpretable (i.e., associated with a clear semantic) and visualize them (Section 4.1.2).

**Classifier Optimization.** This phase optimizes the classifier  $f_{CL}$  while keeping the feature extractor and the prototype layer frozen. It aims to obtain a sparse connection between prototypes and output neurons, which is a desirable property for interpretability. To achieve the goal, we use a loss that penalizes the weights between prototypes associated with a class and the output neurons corresponding to all the other classes:

$$L_{CL} = L_{pred} + \lambda_5 + \sum_{k=1}^{nc} \sum_{j: p_j \notin P_k} \text{abs}(w_{CL}^{j,k}) \quad (4.13)$$

where the weight  $w^{j,k}$  connect the the prototype  $p_j$  to the output neuron corresponding to the class  $k$ .

#### 4.1.2 Explanations

The decision process of the described PIGNN can be probed in two ways. The first consists of understanding the semantics of the learned prototypes, shedding light on what the model deems important during the training process. The second leverages prototype activation to yield node attributions. **To understand the semantics of the learned prototypes, we can use explanations by examples, considering prototypes as inputs.** As described in the previous section, during the second training stage, the prototype is projected into one of the subgraphs included in the dataset, becoming a point in the training distribution of node embeddings. Consequently, we can extract neighbors of each prototype using an algorithm like K-NN [39] over the node embeddings of dataset samples. Then, we can use and visualize the neighbors as explanations by examples, helping users understand the semantics encoded by the prototypes.

Regarding **node attribution**, note that, by design, a node embedding encodes information of the *k-hop subgraph* originating from the node. This subgraph includes all the neighbors within the radius  $k$ , where  $k$  denotes the number of layers before the considered embedding. The k-hop subgraph that excites the prototype the most is identified by selecting the node embedding with the highest similarity to the prototype (determined by Equation (4.2)). Feature attribution for each node  $i$  in the k-hop subgraph is then computed as the sum of similarities between its node embedding and each prototype associated with the prediction. This computation is weighted by the learned weights connecting the prototypes and the predicted class:

$$e(x_i) = \sum_{p_j \in \mathbf{P}_c} w_{jc} s_{ij} \quad (4.14)$$

where  $x_i$  is one of the nodes in the k-hop subgraph,  $c$  is the predicted class for the input  $x$ ,  $w_{jc} \in W_{CL}$  denotes the learned weight of the classifier connecting the prototype  $p_j$  with the output neuron corresponding to the class  $c$ , and  $\mathbf{P}_c$  is the set of prototypes associated with the class  $c$ .

This process can be applied to all nodes in the graph. However, given that graphs are typically compact structures and tasks often involve identifying specific substructures, we propose considering only the **k-hop subgraph of the current input corresponding to the most activated prototype**. This approach yields explanations that are both easy to understand and sparse, aligning

them with the inherent characteristics of graph structures.

## 4.2 Experiments

### 4.2.1 Performance

This section evaluates the performance of popular GNNs augmented with the proposed prototype layer. Specifically, we train Graph Convolutional Network (GCN) [113], Graph Attention Network (GAT) [246], and Graph Isomorphism Network (GIN) [268] on various datasets: BA-shapes [273], BA-Community [273], MUTAG [44], BBBP [149], BACE [230], BA-2Motifs [141], Tree-Grid [273], Cora [151], Citeseer [64] and Pubmed [210]. BA-shapes, BA-Community, Tree-Grid, and BA-2Motifs are syntetic datasets built for node classification (BA-shapes, BA-Community, Tree-Grid) and graph classification (BA-2Motifs). MUTAG, BBBP, and BACE are molecule datasets for graph classification, where the task is to classify molecules as active or inactive. Cora, Citeseer, and Pubmed are datasets for node classification, where nodes represent scientific publications, and the edges represent citations. **Differently from the standard procedures for prototype-based models in vision, we train these models from scratch** since there are no publicly available pre-trained GNNs on large corpora. Additionally, we jointly optimize the feature extractor and the prototype layers during the same training process. We adopt the setup of Ying et al. [273], training 3-layers GCN, GAT, and GIN with 128 units each. Models are trained with and without prototype layers using 15 different seeds. Following [33] and [256], we set the number of prototypes to ten per class, the value of  $\lambda_1$  to 0.8, the value of  $\lambda_2$  to 1e-7, 0.08, the value of  $\lambda_3$  to 0.7, the value of  $\lambda_4$  to 0.8, and the value of  $\lambda_5$  to 1e-3.

**Table 4.1:** Avg. Accuracy and standard deviation over 15 runs of GNNs with and without prototype layers on synthetic datasets.

Model	BA-2Motifs	BA-Community	BA-Shapes	Tree-Grid
GCN	<b>87.20 ± 4.19</b>	<b>97.34 ± 0.14</b>	91.23 ± 1.08	84.79 ± 1.76
+ Proto	59.00 ± 4.12	<b>97.31 ± 0.29</b>	<b>96.30 ± 0.42</b>	<b>97.17 ± 0.59</b>
GAT	<b>51.20 ± 2.32</b>	<b>86.61 ± 1.76</b>	<b>76.54 ± 1.18</b>	<b>58.49 ± 0.00</b>
+ Proto	<b>50.10 ± 3.08</b>	<b>87.94 ± 1.66</b>	42.86 ± 0.00	<b>58.49 ± 0.00</b>
GIN	<b>99.50 ± 0.67</b>	92.34 ± 1.95	89.79 ± 3.89	87.03 ± 1.93
+ Prototypes	<b>99.80 ± 0.40</b>	<b>98.36 ± 0.34</b>	<b>96.96 ± 0.14</b>	<b>98.67 ± 0.28</b>

**Table 4.2:** Avg. Accuracy and standard deviation over 15 runs of GNNs with and without prototype layers on molecular datasets.

Model	BBBP	MUTAG	BACE
GCN	85.75 ± 0.61	<b>77.00 ± 6.78</b>	76.32 ± 1.74
+ Proto	<b>87.61 ± 1.18</b>	<b>80.50 ± 5.68</b>	<b>79.28 ± 1.67</b>
GAT	<b>87.32 ± 1.20</b>	<b>77.50 ± 8.14</b>	<b>78.22 ± 2.66</b>
+ Proto	<b>87.02 ± 0.73</b>	<b>79.00 ± 6.24</b>	<b>77.57 ± 2.44</b>
GIN	<b>85.61 ± 1.56</b>	<b>82.50 ± 7.50</b>	75.40 ± 1.45
+ Proto	<b>87.17 ± 1.05</b>	<b>85.50 ± 3.50</b>	<b>79.34 ± 2.68</b>

We compare the black-box models against the same models augmented with a prototype layer in synthetic (Table 4.1) and molecular datasets (Table 4.2) to assess the validity of the augmentation. The results suggest that **adding the layer to the black-box model and jointly training them does not compromise performance**. In most configurations, models with prototype layers exhibit the same or higher performance than those without prototypes. Exceptions include GAT on BA-Shapes and GCN on BA-2motif, which can be addressed using a simplified loss (Section 4.3). We observe that the performance gain depends on the particular combination of dataset, model, and tasks. The highest gap in accuracy is reached on node classification tasks and synthetic datasets and by the GIN model. These results align with the ones reported by Wang et al. [256] on image data for architectures augmented by prototypes and Xu et al. [268] for GIN superiority. They also suggest that the better the model, the higher the gap between the black-box and the augmented model.

With the approach validated, we compare its performance against an alternative explainable GNN based on prototypes (ProtGNN [283]), another type of explainable network (SEGNN [40]), and the black-box models on Cora, Citeseer, and PubMed (Table 4.3). ProtGNN is a prototype-based model that computes prototypes as graph embedding and represents the alternative configuration of prototype-based layers for GNN. SEGNN is an explainable GNN that classifies nodes based on K-nearest nodes found on the basis of the sub-graphs similarity.

While SEGNN achieves the best performances, PIGNN closely matches its performance on two over of three datasets and outperforms the alternative prototype-based design in all tasks. Lower prototype-based network performance against SEGNN suggests ten prototypes (Section 4.3) may be insufficient to capture enough patterns in citation networks. However, in the next section, we observe that better-quality explanations of these networks somewhat mitigate this issue.

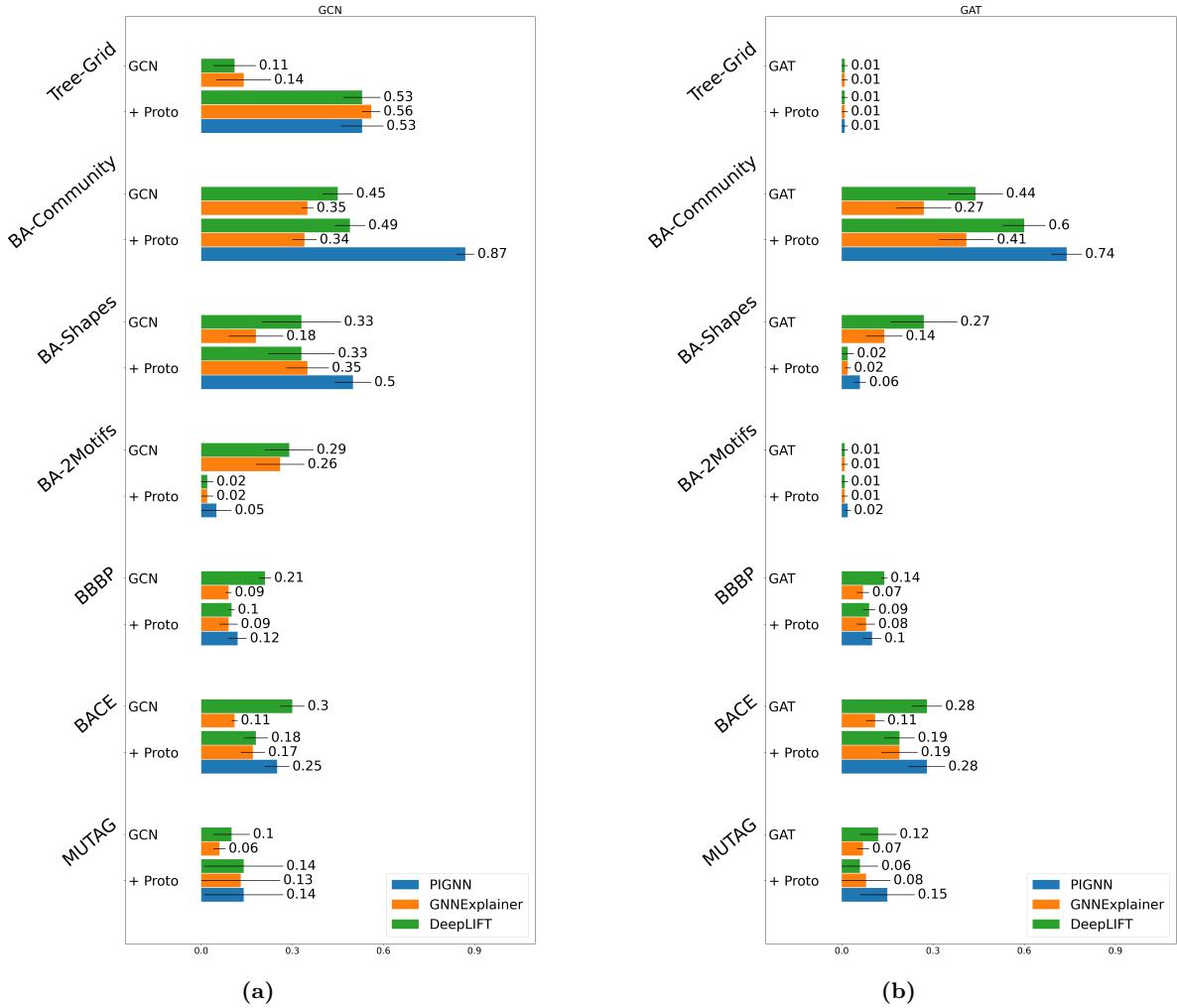
### 4.2.2 Explanations

This section discusses and evaluates the explanations retrievable using PIGNN. We begin by evaluating the quality of feature attributions computed using prototype activations in terms of  $Fidelity + prob$  (Equation (2.22)). We compare its scores against DeepLIFT [214], a popular model-agnostic post-hoc extrinsic method, and GNNExplainer [273], a recent feature attributions method tailored for GNNs. We compare the performance of both models with and without prototypes.

Figure 4.2a, Figure 4.2b, and Figure 4.3 summarize the explanation quality of explanation computed on synthetic and molecular datasets over GAT, GCN, and GIN models. Overall, **PIGNN’s explanations exhibit better quality than extrinsic methods**. In only a few cases, the combination of black-box models and extrinsic methods outperforms PIGNN. When this occurs, the

**Table 4.3:** Avg. Accuracy and standard deviation over 15 runs of GNNs with and without prototype layers on citation datasets.

	Citeseer	Pubmed	Cora
GCN	$73.60 \pm 0.35$	$74.52 \pm 0.54$	$79.09 \pm 0.81$
GCN + Proto	$70.72 \pm 1.01$	<b><math>76.77 \pm 0.86</math></b>	$78.65 \pm 0.80$
SEGNN	<b><math>74.19 \pm 0.51</math></b>	<b><math>76.92 \pm 0.30</math></b>	<b><math>79.66 \pm 0.56</math></b>
ProtGNN	$53.07 \pm 1.98$	$75.93 \pm 1.14$	$59.73 \pm 1.95$

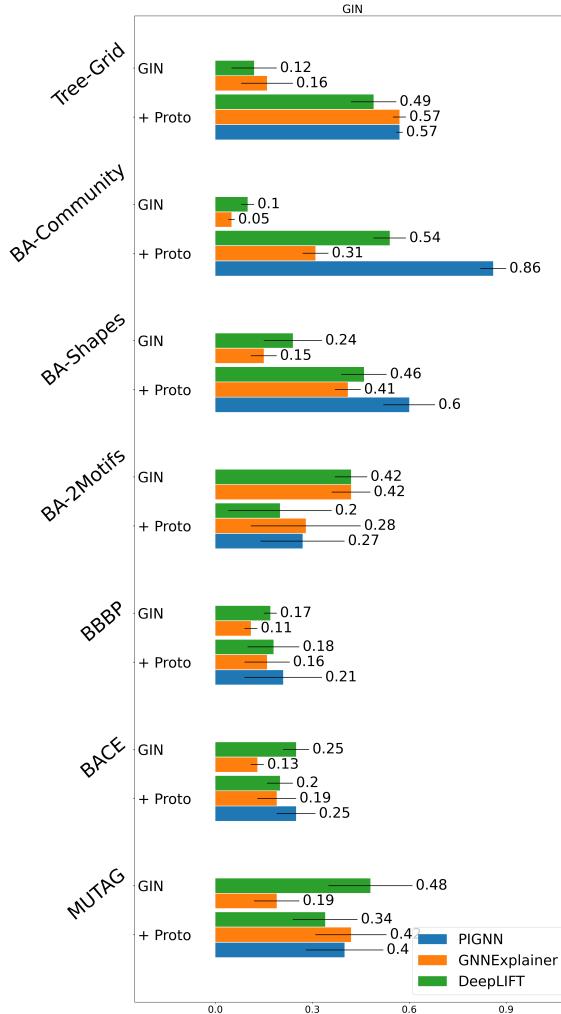


**Figure 4.2:** Avg. Fidelity and standard deviation achieved by the intrinsic method, DeepLIFT, and GNNExplainer on GCN (a) and GAT(b) models across several datasets.

gap is small, and the explanations require more computation time (up to 20x for GNNExplainer). The good quality is also confirmed in Table 4.4, where PIGNN outperforms SEGNN and ProtGNN in terms of both  $Fidelity + prob$  and  $Fidelity + ROC-AUC$  when masking the most important subgraph. Notably, SEGNN is the worst performer in terms of interpretability. We hypothesize that its low performance is due to its dense decision process, reliant on the similarity between k-hop subgraphs, and thus, able to recover the prediction using similarities among other nodes. Conversely, prototype-based networks rely on the bottleneck induced by the prototypes, making their decision process more sparse and sensitive to removing crucial subgraphs.

Now, we describe how developers can **exploit the design and characteristics of PIGNN to understand the model’s behavior**. The examples presented here are drawn from a PIGNN model based on GIN and trained on the Tree-Grid dataset. Nevertheless, the same approach is applicable across different scenarios.

Figure 4.4 shows the globally most influential k-hop graphs used by the model for making predictions. We can note that the first two rows of the image display a diverse array of patterns associated with class 0, including various tree structures. Conversely, the lowest rows show dupli-

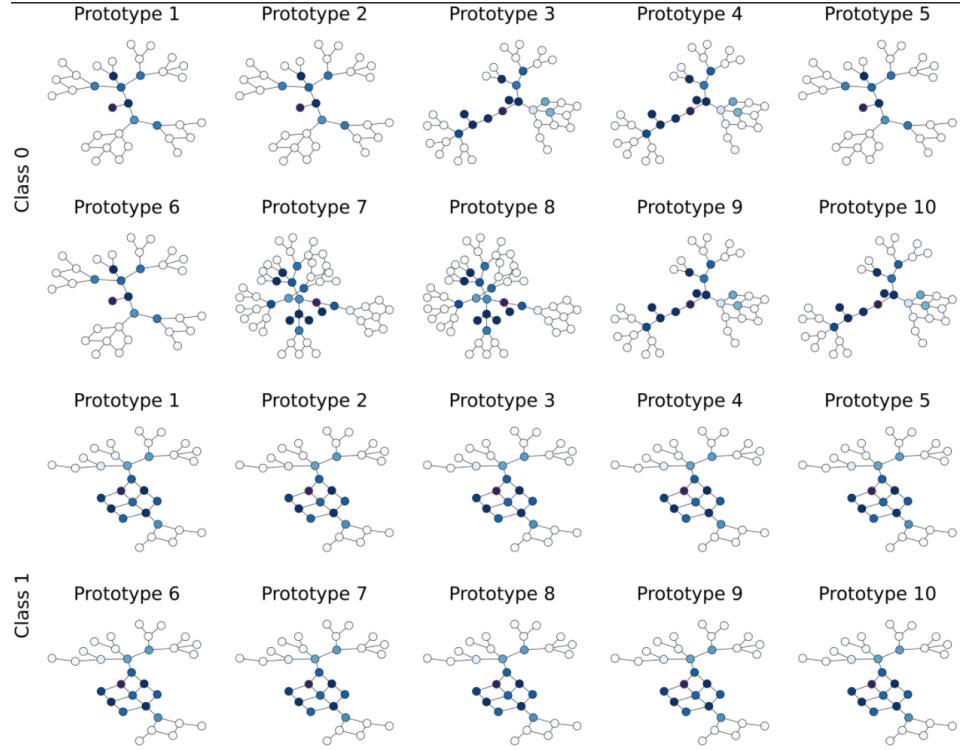


**Figure 4.3:** Avg. Fidelity and standard deviation reached by the intrinsic method, DeepLIFT, and GNNExplainer on GIN models across several datasets.

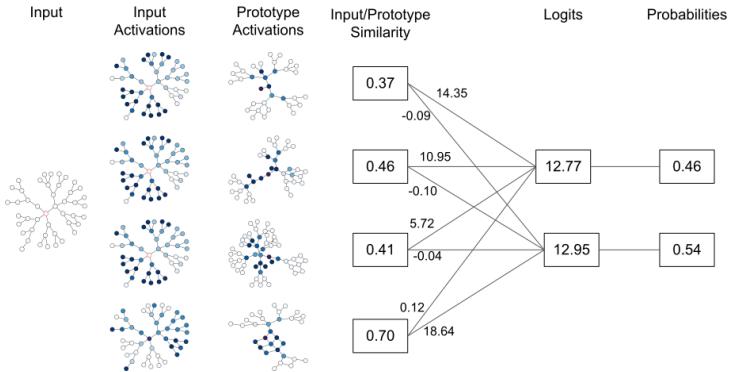
cations in the prototypes for class 1. By inspecting the dataset, a developer could verify that there is low variability in the patterns associated with class 1, most of which are associated with 3x3 grid patterns. These dataset's characteristics cause the convergence of PIGNN towards the recognition of class 1 associated with the recognition of 3x3 grid patterns, captured by all the prototypes. Once that the problem has been identified, the developer can mitigate the issue, for example, by limiting

**Table 4.4:** Explanations performances on big datasets.  $Fidelity+^{prob}$  and  $Fidelity+^{ROC-AUC}$  (Equation (2.24)) are reported for the self-interpretable models and the extrinsic methods on the black-box GCN model.

	$Fidelity+^{prob}$ ( $\uparrow$ )			$Fidelity+^{ROC-AUC}$ ( $\downarrow$ )		
	Citeseer	Pubmed	Cora	Citeseer	Pubmed	Cora
PIGNN	<b><math>0.29 \pm 0.03</math></b>	<b><math>0.24 \pm 0.02</math></b>	<b><math>0.50 \pm 0.06</math></b>	<b><math>0.81 \pm 0.01</math></b>	<b><math>0.87 \pm 0.01</math></b>	$0.88 \pm 0.02$
SEGNN	$0.03 \pm 0.00$	$0.02 \pm 0.01$	$0.10 \pm 0.01$	$0.86 \pm 0.02$	$0.90 \pm 0.03$	$0.96 \pm 0.00$
ProtGNN	$0.02 \pm 0.02$	$0.06 \pm 0.02$	$0.07 \pm 0.07$	<b><math>0.81 \pm 0.03</math></b>	$0.89 \pm 0.02$	<b><math>0.83 \pm 0.01</math></b>



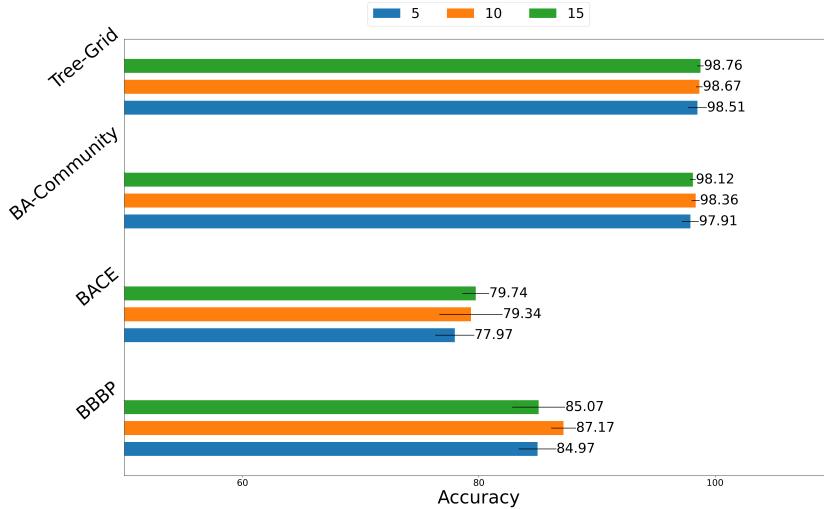
**Figure 4.4:** Sub-graphs associated with the prototypes learned by PIGNN based on GCN on the Ba-2Motifs dataset. The most activated node is marked with a red borderline. The blue shade encodes the importance of the node for the subgraph.



**Figure 4.5:** Visualization of explanation for a prediction. The red border indicates the analyzed node; the Input Activations column reports the similarity between the graph's nodes and the prototype displayed in the same row in the Prototype Activation column.

the number of prototypes or increasing diversity in the dataset.

The visualization of prototypes, together with the similarity associated with them, can also be used to **understand individual predictions and perform error analysis**, as shown in Figure 4.5. In this particular case, the model incorrectly predicts the class of a node as 1. This misclassification is indicated by a high similarity score of 0.7 with the fourth prototype, which represents class 1. The analysis of the prototype's meaning (Prototype Activations column) reveals that the misclassification arises from the model detecting a grid pattern within a portion of the input graph, highlighted in the fourth row of the Input Activation column.



**Figure 4.6:** Accuracy of PIGNN when varying the number of prototypes.

### 4.3 Design Choices

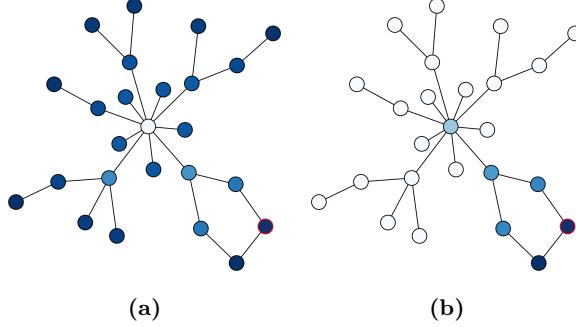
This section discusses alternative design choices for PIGNN and experiments varying the values of hyperparameters.

**Number of Prototypes.** We begin the exploration of alternative design choices for PIGNN by examining the impact of varying the number of prototypes in the model. As observed in the previous section, an excess of prototypes may lead to duplication despite incorporating the orthonormality loss. Figure 4.6 depicts the performance of PIGNN with 5, 10, and 15 prototypes across the BBBP, BACE, BA-Community, and Tree-Grid datasets. We observe that there are no clear winners: while employing a larger number of prototypes generally seems beneficial, this trend does not hold true for BBBP, where using either 5 or 15 prototypes results in equivalent performance deterioration. Therefore, the number of prototypes is a hyperparameter to be tuned before training. Both domain-specific insights and prototype inspection, as shown in the previous section, can be used to guide the selection of the right number of prototypes.

**Similarity Projection.** A design choice about the explanations is whether to project the prototypes' similarities onto all the graph nodes or solely onto the k-hop most important subgraph, as proposed in Section 4.1.2. Table 4.5 compares the fidelity of both approaches, revealing no discernible difference when considering only a subset of nodes. This observation is motivated by the nature of graph-related tasks, where often, identifying a single pattern suffices for graph or

**Table 4.5:** Fidelity comparison between projecting the similarities towards all the nodes in the graph and only towards the k-hop most important subgraph.

	MUTAG	BACE	BBBP	Ba-2Motifs
k-hop	$0.35 \pm 0.10$	$0.25 \pm 0.06$	$0.18 \pm 0.07$	$0.16 \pm 0.15$
Full Nodes	$0.36 \pm 0.13$	$0.24 \pm 0.06$	$0.18 \pm 0.05$	$0.16 \pm 0.15$



**Figure 4.7:** Visualization comparison between projecting the similarities towards all the nodes in the graph and only towards the  $k$ -hop most important subgraph.

node classification. While interpretability scores remain unaffected, the visualization of the two approaches, as depicted in Figure 4.7, differs significantly, with a sparse visualization offering enhanced comprehension and immediacy.

**Loss Function.** Another design choice is the one related to the selection of the loss function to be used during the first training phase. As mentioned earlier, an alternative loss function to the

**Table 4.6:** Performance comparison between the full and the reduced loss applied to PIGNN.

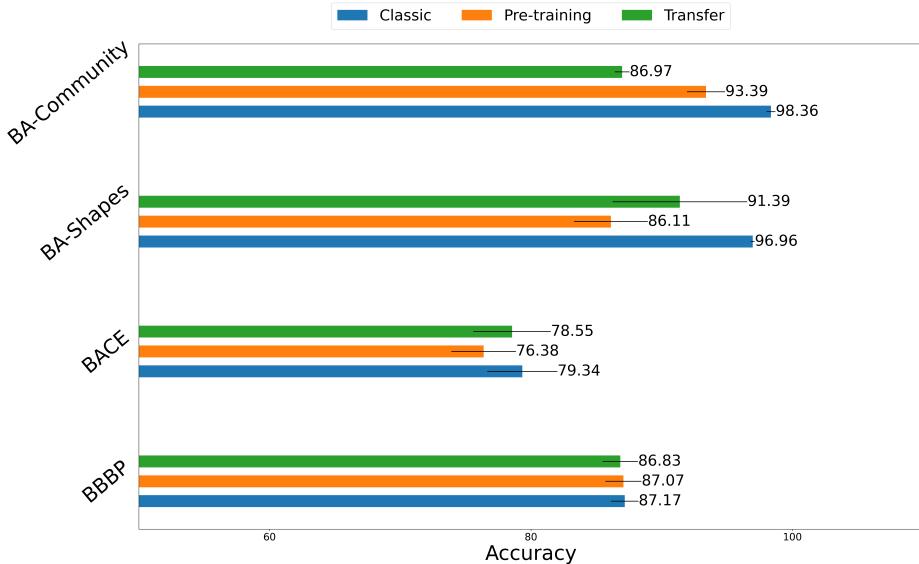
Dataset	Full Loss	Reduced Loss
Citeseer	<b>70.72 ± 1.01</b>	<b>70.12 ± 0.90</b>
Pubmed	<b>76.77 ± 0.86</b>	<b>76.38 ± 1.07</b>
Cora	<b>78.65 ± 0.80</b>	<b>78.62 ± 1.57</b>
BBBP	<b>87.17 ± 1.05</b>	85.66 ± 1.74
MUTAG	<b>85.50 ± 3.50</b>	<b>83.50 ± 5.50</b>
BACE	<b>79.34 ± 2.68</b>	<b>80.86 ± 3.12</b>
BA-2Motifs	<b>99.80 ± 0.40</b>	98.90 ± 1.14
BA-Community	<b>98.36 ± 0.34</b>	96.96 ± 0.14
BA-Shapes	<b>96.65 ± 0.74</b>	85.11 ± 1.41
Tree-Grid	<b>98.67 ± 0.28</b>	82.42 ± 6.82

**Table 4.7:** Examples of specific cases when the reduced loss can work better.

Dataset	Model	Full Loss	Reduced Loss
BA-2Motifs	GCN	59.00 ± 4.12	<b>99.40 ± 0.80</b>
BA-2Motifs	GAT	50.10 ± 3.08	<b>83.30 ± 5.92</b>

**Table 4.8:** Interpretability scores achieved by PIGNN when using the full or reduced loss.

Dataset	$Fidelity + prob$ ( $\uparrow$ )		$Fidelity + ROC-AUC$ ( $\downarrow$ )	
	Full Loss	Reduced Loss	Full Loss	Reduced Loss
Citeseer	<b>0.29 ± 0.03</b>	<b>0.29 ± 0.02</b>	<b>0.81 ± 0.01</b>	0.86 ± 0.01
Pubmed	<b>0.24 ± 0.02</b>	0.20 ± 0.01	<b>0.87 ± 0.01</b>	0.90 ± 0.01
Cora	<b>0.50 ± 0.06</b>	0.33 ± 0.06	<b>0.88 ± 0.02</b>	0.94 ± 0.00



**Figure 4.8:** Avg. Accuracy and standard deviation of PIGNN when jointly trained with the feature extractor (Classic), trained after a pre-training on the same dataset (Pre-training), and fine-tuning a model trained in a different dataset via transfer learning (Transfer)

one outlined in Section 4.1 is proposed by Chen et al. [33], which omits the orthogonal and class separation losses. Table 4.6 illustrates that the full loss remains the optimal choice, as the reduced loss achieves comparable performance on only half of the datasets. However, specific scenarios, detailed in Table 4.7, demonstrate instances where the reduced loss outperforms the full one by a significant margin. In such cases, inherent characteristics of datasets like BA-2Motifs, such as dependency on a single motif, coupled with additional constraints imposed by the added losses, may cause the model to be stuck on a local minimum. A search over hyperparameters tailored to specific cases could mitigate this issue. Furthermore, as evidenced in Table 4.8, in terms of interpretability, the full loss consistently demonstrates superiority in both  $Fidelity + prob$  and  $Fidelity + ROC - AUC$ , thus suggesting that the added losses help the network to better separate the prototypes and thus improving their usefulness for interpretability purposes. Therefore, the full loss represents the best choice.

**Transfer Learning** Finally, unlike other prototype-based architectures [33, 256], PIGNN does not use pre-trained models and jointly trains both the feature extractor and the prototype layer simultaneously. Here, we explore the alternative scenario of transfer learning. It's worth noting that no publicly available pre-trained models on large graph datasets exist due to the disparate meanings of edges and nodes across different tasks. Therefore, we test transfer learning by pre-training the feature extractor in a dataset and then fine-tuning the whole model using the target dataset. Note that we select the first dataset among the ones sharing the same data type (i.e., molecular). We also consider the case where we pre-train the feature extractor and then fine-tune the same model augmented with the proposed layer on the same dataset.

Figure 4.8 demonstrates that simultaneously training from scratch the feature encoder and the prototype layer yields superior results, outperforming the alternative in all considered tasks. The lower performance of transfer learning was expected, given that molecules from different datasets

often belong to separate chemical spaces, thereby leading to the problem of the domain of applicability and lowering the general performance. On the other hand, the better performance and the large gap between simultaneous training and a pre-training phase are surprising and underscores the potential for further exploration of this phenomenon in different domains.

## 4.4 Contributions

This chapter contributed to the research on self-explainable neural networks in the emerging field of GNNs. Specifically, we investigated the application of prototype layers to graph data. We analyzed the differences and specific characteristics of the graph applications, along with the advantages and weaknesses associated with their utilization for enhancing the interpretability of GNNs.

Notably, the proposed approach preserves the performance of the black-box models while improving their interpretability in terms of fidelity with respect to their decision process. Unlike the image domain, prototype layers proved themselves to be more effective when jointly trained with the feature extractor. Moreover, they provide more sparse and easy-to-understand global and local explanations. This chapter showcased how they can be used to highlight the learned subgraphs in terms of prototypes, the most important subgraph for a specific prediction, and, more in general, to debug and improve DNNs models. Averaging the quality of performance and explanations, PIGNN represents an advancement with respect to the state-of-the-art self-explainable graph neural networks and a better choice than using black-box models coupled with extrinsic methods.

Alongside these strengths, we also analyzed some of the drawbacks of this approach. For example, the number of prototypes in prototype layers represents a parameter that is difficult to choose and impacts both the performance and quality of explanations, especially in unbalanced datasets. Additionally, despite the improvements in terms of joint training, the proposed layer still uses the custom training process of prototype-based architectures. Using these training recipes could cause instability in the training process when applied to settings where the black-box models are trained using non-standard procedures. This problem motivated us to investigate designs beyond the current self-explainable DNNs, described in the next chapter.

## Chapter 5

# Self-Explainable Memory-Augmented Deep Neural Networks

The previous chapter contributes to the literature in terms of generalization, thus following most of the established procedures and inheriting some of its limitations. These limitations could impact the applicability of the proposed techniques beyond the standard settings explored in the research. To mitigate these limitations, this chapter takes a further step in exploring self-explainable deep neural networks and presents designs that augment black-box DNNs with memory, thereby transforming them into explainable Memory Augmented Neural Networks (MANNs) (Section 2.1).

While the augmentation of a DNN with a memory module has already been explored to mitigate the issue of catastrophic forgetting [60, 69, 70] or the scarcity of data [223, 250], explicitly exploiting memory modules for explainability purposes represents a novelty in the field, increasing the diversity of approaches.

As in the previous chapter, the idea is to keep most of the DNNs' structure intact. Therefore, we assume that the DNN can be represented as the composition of two functions, the feature extractor  $f_{FE}$  and the classifier  $f_{CL}$ , and place the novel memory modules after the feature extractor. Additionally, in this case, the designs aim to preserve the training recipe as much as possible.

Similar to other DNNs, MANNs are commonly regarded as black-box models. The opacity of a MANN arises from three primary sources within its decision process: the controller, the memory, and the classifier. In a MANN, the controller transforms the input representation of the current sample into a latent representation, which is then used to interact with the memory. Since the controller is often a black-box DNN, the transformations applied to the input often lack interpretability, rendering the resultant latent representation devoid of semantics.

The memory can be challenging to interpret for various reasons. Firstly, the information stored within the memory may lack a clear link to discernible semantics. Secondly, the writing mechanism may merge new and old information, blurring the distinction between different instances. Lastly, the reading mechanism itself may be opaque, particularly when multiple readings are combined through non-linear operations.

Finally, the interpretability of the classifier depends on how the controller output and the readings are combined, in addition to the conventional factors influencing interpretability, such as the inputs and the design of the classifier itself.

Since MANNs are more opaque than a DNN alone, transforming a DNN into a MANN to improve

its interpretability may seem counter-intuitive. Nevertheless, we propose **designs that exploit the advantages of MANN and, at the same time, can effectively mitigate the challenges mentioned above via memory tracking operations**, thereby enhancing the interpretability of DNNs. Specifically, our approaches treat the architectural components of the original black-box DNN until the penultimate layer (feature extractor) as the controller of MANN and place memory between the feature extractor and the classifier. Therefore, the feature extractor remains a black-box, and the proposed designs improve the interpretability of the last part of the network.

These designs can be used to support developers in understanding and improving their models by supporting tasks like hypothesis verifications, bias, uncertainty detection, and more, in general, to better understand the decision process of these models.

This chapter is split into three blocks: the first block provides an analysis of our proposed solution starting from the case of sequential data (Section 5.1); the second block proposes a self-interpretable module for image classification (Section 5.2); the last block summarizes the results and contributions (Section 5.3).

## 5.1 Simplified Differentiable Neural Computers

This section proposes a first design of explainable MANN able to work on sequential data. First, Section 5.1.1 describes the architectural design and how to retrieve explanations. Then, Section 5.1.2 evaluates the proposed architecture both in terms of performance and explanations. Finally, Section 5.1.3 discusses several alternative design choices.

### 5.1.1 Design

Building upon the design of Differentiable Neural Computers (DNCs) [70], **Simplified Differentiable Neural Computers (SDNCs)** [122] are MANNs fully differentiable in all their components. Given a sequential input  $\mathbf{x} = [x_1; x_2; \dots; x_n]$ , the basic components of a SDNC are the controller  $f_{FE}$  and a memory  $M$ . The controller takes the input of the current step  $t$  and returns an output  $\mathbf{h}_t$  for each step. The memory is a matrix  $M \in \mathbb{R}^{NR \times NC}$  of  $NR$  row and  $NC$  columns where the controller can read and write information into. Writings and readings within SDNCs are executed using  $NH_R$  read heads and  $NH_W$  write heads. During the training process, the architecture learns a set of weights to transform the controller output into a set of parameters, subsequently utilized to interact with the memory:

- $\mathbf{W}_e$  controls the information to be erased in memory;
- $\mathbf{W}_w^i$  manages the information to be written in memory by each write heads  $i$ ;
- $\mathbf{W}_r^i$  regulates the information to be read in memory by each read heads  $i$ ;
- $\mathbf{W}_y^i$  controls how to use the memory and controller information to produce the prediction.

First, the architecture normalizes the controller's output for the current step in a sequence:

$$\mathbf{h}_t = h_{i,k}^t = \frac{h_{i,k}^t - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} \quad (5.1)$$

where  $\sigma$  and  $\mu$  are the mean and the variance across the controller's outputs, respectively. Then, the model computes a series of vectors used for controlling readings and writings based on the set of parameters above-mentioned:

- $e \in [0, 1]^{NC}$ , the erase vector representing the information to be erased in memory computed as  $\sigma_{sig}(\mathbf{h}_t \mathbf{W}_e)$ ;
- $v \in \mathbb{R}^{NC}$ , the write vector representing the information to be written into memory, computed as  $\mathbf{h}_t \mathbf{W}_w$ ;
- $k \in \mathbb{R}^{NH_R \times NC}$ , a set of read keys, one for each read head;
- $\beta^i \in [1, \infty]^{NH_R}$ , the read strengths of the read heads.

Readings and writing operations are performed using attention mechanisms, employing dynamic memory allocation for writing and content-based addressing for reading.

Dynamic memory allocation is a differentiable mechanism used to free memory. This mechanism updates the *usage vector*  $\mathbf{u}_t \in [0, 1]^{NR}$ , which tracks the usage of each location in memory. Specifically, the location usage is incremented upon being written in the preceding step:

$$\mathbf{u}_t = \mathbf{u}_{t-1} + \mathbf{w}_{t-1}^w - \mathbf{u}_{t-1} \mathbf{w}_{t-1}^w \quad (5.2)$$

Based on the usage vector and a list  $\phi_t$  containing memory location indices sorted by ascending usage, the architecture computes the *allocation vector*:

$$\mathbf{a}_t[\phi_t[j]] = (1 - \mathbf{u}_t[\phi_t[j]]) \prod_{i=1}^{j-1} \mathbf{u}_t[\phi_t[i]] \quad (5.3)$$

When the location usage is 1, the corresponding index in the allocation vector is set to 0, indicating that the memory location cannot be overwritten. Conversely, when the usage is 0, the location is available for writing operations and the allocation is set to 1. When the usage is between 0 and 1, the allocation is weighted by the position of the location in the sorted list: a higher position results in a lower allocation value. Subsequently, the write heads combine the allocation vector with the *writing gate* to generate the *write weightings*:

$$\mathbf{w}_t^w = \mathbf{g}_t^w \mathbf{a}_t \quad (5.4)$$

The *writing gate* controls the extent to which information from the current step should be written. Therefore, if the gate is 0, no memory writing occurs during the current step. The architecture then updates each memory location by erasing information based on the *erase vector*, writing the information contained in the *write vector*, and controlling the information flow using the *writing weightings*:

$$\mathbf{M}_t = \mathbf{M}_{t-1} \circ (1 - \mathbf{w}^w e^T) + \mathbf{w}^w v^T \quad (5.5)$$

Thus, a location is partially erased when both the erase vector and the write weightings do not contain a 0 in the index corresponding to that location. A location is written when neither the write weightings nor the write vector contain a 0 in the index corresponding to that location.

Read operations within the architecture rely on the *content addressing* mechanism. Content addressing is determined by the cosine similarity between two vectors, as defined by the equation:

$$D(\mathbf{v}_1, \mathbf{v}_2) = \frac{\mathbf{v}_1 \cdot \mathbf{v}_2}{\|\mathbf{v}_1\| \|\mathbf{v}_2\|} \quad (5.6)$$

In this context, the two vectors under consideration are the memory of the previous step  $\mathbf{M}_{t-1}$  and the read keys computed based on the current output of the controller. Thus, the architecture calculated the read weightings as the softmax of the cosine similarity between the memory of the previous step and the read keys, weighted by the read strength:

$$C(\mathbf{M}, \mathbf{k}, \beta)[i] = \frac{e^{D(\mathbf{k}, \mathbf{M}[i])\beta}}{\sum_j e^{D(\mathbf{k}, \mathbf{M}[j])\beta}} \quad (5.7)$$

$$\mathbf{w}_t^{r,i} = C(\mathbf{M}_{t-1}, \mathbf{k}_t^{r,i}, \beta_t^i) \quad (5.8)$$

The read weightings are then used to compute the *read vectors* as the weighted sum of the memory content:

$$\mathbf{r}_t^i = \mathbf{M}_t^T \mathbf{w}_t^{r,i} \quad (5.9)$$

Finally, the architecture concatenates the read vectors and the controller's output and sends them to the classifier to compute the prediction. To encourage the usage of memory during the inference process, a dropout is applied to the controller's output.

$$y_t = W_y((\mathbf{h}_t \mathbf{p}_t) \oplus \mathbf{r}_t^1 \oplus \mathbf{r}_t^2 \oplus \dots \oplus \mathbf{r}_t^{N_r}) \quad (5.10)$$

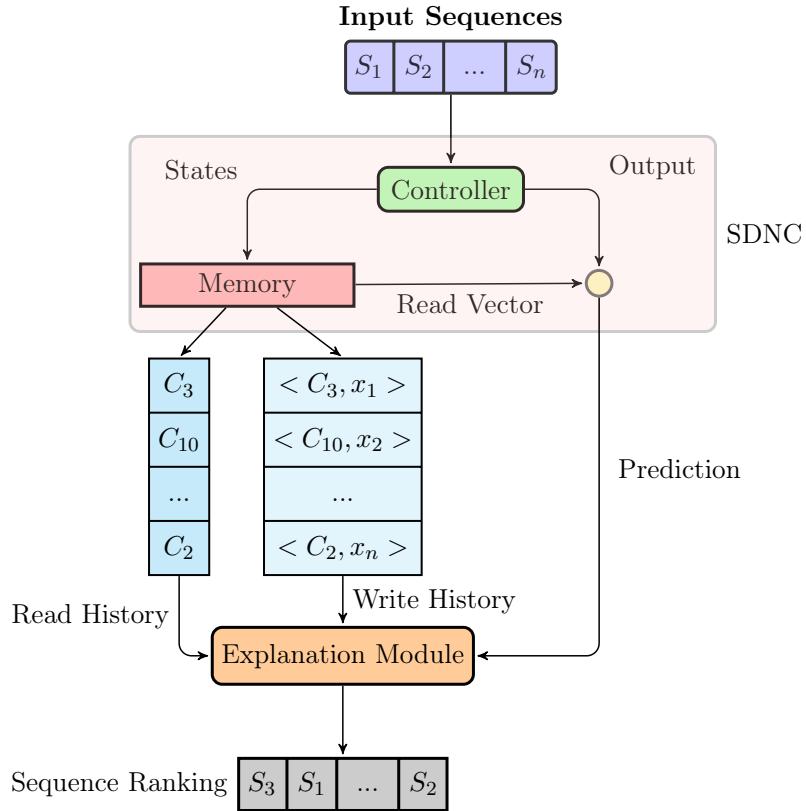
$$\mathbf{p}_t = \sum_{i=0}^n p_i \in [0, 1] \sim Bernoulli(r_p) \quad (5.11)$$

where  $r_p$  is the dropout rate that controls how many elements of the vectors should be set to zero on average.

Regarding the differences with DNC, the latter includes more parameters than the SDNC, a different write mechanism, a different read mechanism, and it includes a double recurrence. Specifically, DNC includes two additional gates, a set of write strengths, and a set of write keys. The writing mechanism also considers the *content weights* to compute the write weightings. Therefore, the architecture can write new information in a location containing similar content even if the usage vector is high. The double recurrence refers to concatenating the input with the read vectors of the previous steps. While these features are important to reach higher performance than the controller alone, they lower the interpretability of memory operations since they lead to overwriting operations or increase the opacity of the semantics encoded in a state. Finally, DNCs also include the temporal linkage matrix for computing read weightings, which we removed following the same analysis of Franke et al. [60].

### Explaining Via Memory Tracking

As introduced in the previous section, the idea is to treat the architectural components of the controller as a black-box and improve the interpretability of the last part of the network by exploiting



**Figure 5.1:** Design of SDNC. The blue boxes represent the read-and-write history.  $\langle C_i, x_j \rangle$  indicates that feature  $x_j$  is stored in cell  $C_i$ . The ranking sorts the subsequence of the input based on their importance.

the memory layer. We can observe that information stored and read from the memory corresponds to the states and outputs of the black-box controller. Therefore, for interpretability purposes, the idea is to extract patterns in the sequence of outputs and connect these patterns to the final prediction. To achieve the goal, the proposed **memory tracking mechanism monitors both the reading and writing operations to derive explanations** (Figure 5.1). During the writing process, the tracking focuses on the writing weightings. The objective is to establish a mapping between steps and memory locations that were written during each step. As previously explained, a location is written if the value in the writing weightings corresponding to that location is zero. However, due to the differentiability of the operators, the weighting values of the are never exactly zero but rather close to it. Therefore, we relax the definition of written locations to introduce more flexibility. Let be  $\mu_t^w$  the average value of the write weightings:

$$\mu_t^w = \frac{\sum \mathbf{w}_t^w}{NR} \quad (5.12)$$

Then the written locations at the step  $t$  are represented by the vector of indices  $\mathbf{ind}_t^w$ , computes as:

$$\mathbf{ind}_t^w = \{j \mid \mathbf{w}_t^w[j] > \mu_t^w\}_{topk^w} \quad (5.13)$$

where  $topk^w$  is a hyperparameter and the notation  $_{topk^w}$  indicates that only the highest  $topk^w$  locations are retained among those satisfying the condition. In other words, a location is considered written if its write weight is above the mean and is among the  $topk^w$  highest values in  $\mathbf{w}^w$ .

Similarly, during the reading process, the tracking focuses on the read weightings. In this case, the goal is to establish a mapping between steps and locations that were read during each step. A location can be considered read if the value in the read weightings is greater than zero. However, a small value has little or no impact on the decision process. Therefore, we propose a more strict definition of read locations. The average value of the read weighting is defined as: Let be  $\mu_t^{r,i}$  the average value of the read weightings of the read head  $i$ :

$$\mu_t^{r,i} = \frac{\sum \mathbf{w}_t^{r,i}}{NR} \quad (5.14)$$

The read locations at the step  $t$  are represented by the vector of indices  $\mathbf{ind}_t^{r,i}$ , computed as:

$$\mathbf{ind}_t^{r,i} = \{j \mid \mathbf{w}_t^{r,i}[j] > \mu_t^{r,i}\}_{topk^r} \quad (5.15)$$

Here,  $topk^r$  and  $topk^w$  are parameters similar to those defined for writing operations but specific for reading operations. In summary, a location is considered read if its read weight is above the mean and is among the  $topk^r$  highest values in  $\mathbf{w}^{r,i}$ .

Subsequently, the module ranks the features based on the frequency of readings. The ranking construction is flexible and can be adapted to tasks and needs related to explainability.

In the case of a multi-step decision process, a common scenario in MANNs, the readings of the steps in the decision process are summed, and **attribution is assigned based on the frequency of readings**:

$$a(t_k : t_m) = \sum_{t_{f_0} < t_f < t_{f_n}} \sum_{k < t < m} \sum_{0 < i < R} |j : j \in \mathbf{ind}_t^w \text{ } \& \text{ } j \in \mathbf{ind}_{t_f}^{r,i}| \quad (5.16)$$

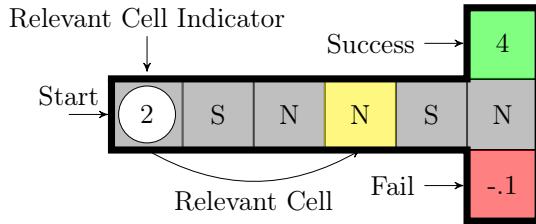
where  $a(t_k : t_m)$  is the score assigned to the subsequence  $[x_k : x_m]$ ,  $t_{f_0}$  is the first step of the decision process, and  $t_{f_n}$  is the last step of the decision process. The equation is the most general one, encompassing more specific cases like the feature-level attribution (setting  $t_f = t_{f_0} = t_{f_n}$ ) and single-step decision processes (setting  $t = k = m$ ). The explanation process is summarized in Figure 5.1.

### 5.1.2 Experiments

#### Performance

SDNC is proposed as a network augmentation for a black-box model to enhance its interpretability. This section compares the performance achieved by the standalone black-box model and our proposed augmentation. While the design draws inspiration from DNCs [70] and Advanced DNCs [60], SDNC is not intended to enhance the controller's performance but rather to improve its interpretability. Specifically, we removed certain components of DNC that might complicate interpretability while potentially boosting performance. Therefore, we expect DNC to outperform SDNC, and we do not consider DNC and its variants as competitors.

We evaluate the performance of SDNC on a modified T-Maze task and the Cloze Story Test. For both tasks, we use a feedforward LSTM layer with 128 units as a black-box and as a controller of SDNC. The choice is motivated by the fact that this model has already been successfully tested



**Figure 5.2:** Example of a 5-step modified T-maze, where the starting position (in white) indicates the relevant cell (in yellow).

in all the scenarios considered in this chapter: on T-Maze [15], on Story-Cloze test [153], and as a controller of DNC variants [60, 70].

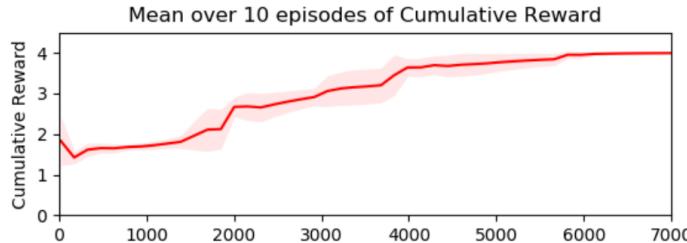
T-maze [15] is a non-Markovian discrete control task where an agent navigates a corridor of variable length from a starting position to a T-junction. The agent can move North, East, South, or West at each step. In our modified T-maze version, at each step, the agent observes a symbol indicating the direction to take at the junction to reach the goal. The starting position specifies the index of the step containing the right suggestion. All the other symbols are assigned randomly. The agent receives a reward of 4 for the correct action and a reward of  $-0.1$  otherwise (see Figure 5.2). We cast the problem as a classification problem and train all the models on a dataset of 2000 samples, split following the rule of 80/20 between the train and test datasets.

The Story Cloze Test [162] is a commonsense reasoning task to predict the correct ending of a story composed of four sentences (*premises*). For this task, we embed each word using word2vec [154] and append three boolean flags to each vector, indicating whether the word is included in the story, the first ending, or the second one. We also include a zeroed query vector to request an answer from the model. The training dataset is split into training and validation data (10% of the training data).

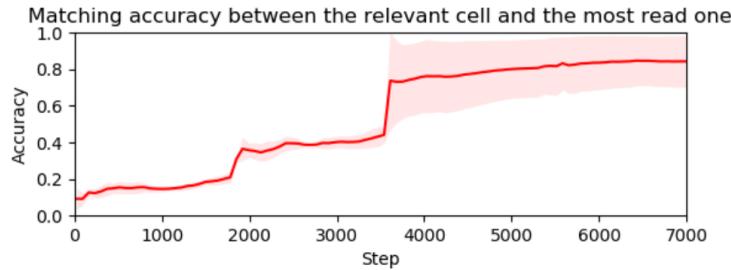
We train the model for ten epochs in T-Maze and use an early stopping with patience equal to ten epochs in the Story Cloze Test. The model associated with the highest accuracy on the validation data is considered the final model. In the case of the Story Cloze Test, we adopt the configuration of Mihaylov and Frank [153] for the controller configuration, using the encoding of the last step of the story as the initial state of the encoding of both the possible endings. Regarding the memory parameters, in the Story Cloze Test, we fix the memory size to 512, the read heads number to 4 read heads, and the write heads number to 1. We apply the Adam optimizer using a starting learning rate of  $1e-3$  and a dropout rate equal to 0.2. Conversely, in T-Maze, we use the RMSprop optimizer using the default parameters of the Keras implementation<sup>1</sup>, and we fix the memory size to 50, and both the read heads and the write heads number to 1.

As depicted in Figure 5.3 and Table 5.1, the network effectively solves the task in the T-Maze case and **attains comparable generalization power** in terms of accuracy in the Story Cloze Test when compared to the controller alone. The black-box performance is retrieved from Mihaylov and Frank [153]. While the achieved performance is promising, it is important to acknowledge some weaknesses. Due to the increased parameters in SDNC, the training process is slower (over 2x). Additionally, note that **SDNC employs a distinct configuration** compared to the one of Mihaylov and Frank [153] with a lower learning rate, a smaller number of units, and a different

<sup>1</sup>[https://www.tensorflow.org/api\\_docs/python/tf/keras/optimizers/experimental/RMSprop](https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/experimental/RMSprop)



**Figure 5.3:** Cumulative reward and standard deviation of ten agents trained on the T-Maze problem.



**Figure 5.4:** Matching accuracy score and standard deviation of ten agents trained on the T-Maze problem.

batch size (1). Indeed, while the black-box benefits from large batch sizes [153], we observe that SDNC performs optimally when trained using small batch sizes. Therefore, we expect that a user transitioning from a black-box to an SDNC should conduct an optimization search over the controller configuration parameters to achieve comparable performance.

### Explanations

The main advantage carried on by the proposed SDNC is on the interpretability side. Indeed, the controller is a black-box model and does not provide any way to inspect its decision process. Conversely, this section showcases and evaluates explanations obtained by applying the memory tracking mechanism (Section 5.1) to SDNC.

In the T-Maze problem, users could be interested in **verifying whether the models learned as intended** and leverage information from the relevant step corridor to solve the task. In this case, since the rules governing the dataset are known, we can retrieve the ground truth explanations (Section 2.2), which correspond to the relevant step corridor. To achieve the goal, memory tracking specializes Equation (5.16) by collecting the two most read cells during the decision step. We consider two locations to also account for the first corridor step, which could be equally significant. To evaluate these explanations, we check the frequency (*matching accuracy*) with which the memory locations containing that information rank among the first two positions in the memory tracking mechanism's output.

**Table 5.1:** Average test Accuracy reached by the models in the testing Story Cloze Test dataset.

Model	Test Acc. % (Avg)
Black-Box	0.72
SDNC	0.72

Earl woke up early to make some coffee. (48.3%) He wanted to be alert for work that day. (47.4%) The aroma woke up all his roommates. (0%) They wanted to make coffee too. (4.2%)
E1. All of his roommates made coffee. (CORRECT) – E2. All of his roommates were sick of coffee.
Samantha had recently purchased a used car. (15.6%) She loved everything about the car except for the color. (30.3%) She took her car to her local paint shop. (31%) She got it painted a bright pink color. (23%)
E1. Samantha likes the color of her car now. (CORRECT) – E2. Samantha thinks her bus looks pretty now.
Tim didn't like school very much. (23.6%) His teacher told him he had a test on Friday. (15%) If he didn't pass this test, he could not go on the class trip. (4.5%) Tim decided to play with his kites instead of study for the test. (56.8%)
E1. Tim was unprepared and failed the test. – E2. Tim aced the test and passed with flying colors. (WRONG)
Neil took a ferry to the island of Sicily. (87.2%) The wind blew his hair as he watched the waves. (0%) Soon it docked, and he stepped onto the island. (0%) It was so breathtakingly beautiful. (12.7%)
E1. Neil enjoyed Sicily. (CORRECT) – E2. Sicily was the worst place neil had ever been.

**Figure 5.5:** Example outputs on the Story Cloze Test. A relevance score is associated with each premise. The blue color indicates the best premise, while the orange indicates the worst one.

Figure 5.4 illustrates the agent’s gradual improvement in the accuracy of its explanations during the learning process. Observing the similar behavior in the curves of Figure 5.4 and Figure 5.3, we infer that, as expected, better detection of the relevant step corridor among the most read locations corresponds to improved task accuracy.

In the Story Cloze Test, we evaluate the memory-tracking mechanism’s precision in **highlighting the story’s most important subsequence** (premise) using the  $Fidelity-acc$  metric. In this case, the memory tracking mechanism, during each step of parsing the predicted ending, collects the ten most read locations. These locations are mapped to the stored words in memory, and a premise ranking is built based on the reading frequency, as explained in Equation (5.16). Three kinds of premises are identified from the ranking: the best, the worst, and a random one. The *best premise* is determined by the most frequently read words, the *worst premise* by the least frequently read words, and the *random premise* is arbitrarily chosen. Figure 5.5 provides some examples of the computed explanations and their associated frequency. Each of these premises is then fed as isolated input to the network, and their predictions are collected. Intuitively, if a premise is important, the network should more consistently achieve the same prediction as when fed the full input, compared to a random premise. Table 5.2 validates the quality of the approach, indicating that the best premise achieves higher fidelity than the random premise on the training, development, and testing datasets.

**Table 5.2:** Avg. fidelity over ten runs on the Story Cloze Test when feeding the model with only a random, the best and the worst premise.

	Avg. $Fidelity-acc$			
	Premise	Train	Dev	Test
Random	0.57	0.60	0.56	
Best	0.66	0.73	0.64	
Worst	0.51	0.53	0.54	

Earl woke up early to make some coffee . He wanted to be alert for work that day . The aroma woke up all his roommates . They wanted to make coffee too . All of his roommates made coffee. All of his roommates were sick of coffee.	Samantha had recently purchased a used car . She loved everything about the car except for the color . She took her car to her local paint shop . She got it painted a bright pink color . Samantha likes the color of her car now. Samantha thinks her bus looks pretty now.
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

**Figure 5.6:** Output of LIME [197] applied to a couple of samples from the Story Cloze Test dataset.

Additionally, the worst premise attains the lowest fidelity, reinforcing the reliability of the ranking returned by the memory tracking mechanism.

To ensure that the elements used for explanations are connected to the decision process, two additional tests are performed: parameter randomization and data randomization tests. These tests evaluate the link between learned parameters, labels, and explanations [4] starting from the observation that **good explanations should be linked to the learned parameters and the input-label mapping**. The parameter randomization test checks the fidelity of the best and worst premises on an untrained network, while the data randomization test shuffles the labels in the training dataset, trains a model by using it, and evaluates the fidelity of explanations on the test dataset.

Table 5.3 shows that the best and the worst fidelity are nearly the same in the parameter randomization test, implying that the method cannot differentiate between a good and a bad premise in both cases. This result indicates that memory tracking correctly depends on the model’s learned parameters. The results of the data randomization test are similar, with the best premise accuracy failing to predict the original outcomes, confirming the link with the learned mapping. Moreover, we observe that the worst premise outperforms the best one. The cause of such behavior is an attempt of the model to exploit the changed labels. Indeed, when a label is altered, the model should select the premise with the minimum similarity with the memory output to predict the changed label. This behavior is rewarded by some of the correctly labeled examples due to the overabundant nature of the task.

The explanations retrieved for the two benchmarks underscore the **flexibility of memory-tracking** to adapt to the user query. For comparison, consider Figure 5.5 and Figure 5.6, which presents the explanations returned by LIME [197] for a couple of samples in the Story Cloze Test. It is easy to see that LIME focuses on words in the endings, indicating which words influence the prediction towards the first ending (blue) or the second one (orange) if removed. A similar output can be achieved by considering the readings of the endings’ words, instead of the premises ones, using the memory tracking mechanism (i.e., the same approach used in the T-Maze task). Conversely, adapting LIME to extract which premises are important is challenging and requires a redesign of LIME or a modification at the model level. Similar to other model-agnostic methods,

**Table 5.3:** Accuracy and  $Fidelity^{acc}$  when the model is untrained and then labels are shuffled.

	Randomization Test	Avg. $Fidelity^{acc}$		
		Accuracy	Best	Worst
Parameters		0.51	0.56	0.55
Data		0.59	0.52	62.8

LIME treats the model as a black-box, and it simply analyzes the relation between modified inputs and the outputs. Consequently, guiding the explanation process to answer specific user questions becomes difficult. While Ribeiro et al. [197] suggest using super-pixels permutations to extract feature attribution for a group of features (such as the premises), finding thousands of permutations of premises that preserve the semantics of the story is not trivial for sequential and text data. The last alternative to assign scores to the premises is to aggregate attributions at the word level into a sentence-level one. However, as mentioned earlier, since there is no way to impose constraints on the input, the highest-scored words are those in the endings, while all the words of the premises receive low associated weights, making the entire process unreliable. Conversely, in the SDNC case, it is sufficient to modify the indices of Equation (5.16) to switch from one case to another, making these adaptations trivial and more preferable than post-hoc extrinsic methods.

While these results are encouraging and represent a big step forward with respect to the black-box controller, we note two weaknesses related to explanations based on SDNC. The first is related to the skip connection that connects the controller to the classifier. While the scores are reliable on average, as highlighted by the fidelity and the additional tests, there could be cases where the classifier ignores the memory output and performs the classification based only on the controller output. These cases represent a current limitation but can be detected by using external means (like the gradients) to inspect the contribution of each element in the read vectors, thus combining the advantages of both approaches. The second limitation regards the input dispersion. Indeed, we observe that applying memory tracking for individual feature attribution (i.e., assigning scores to individual features) is less effective. In these cases, the highest scores can be associated with later steps, like punctuation at the end of the sentence. The nature of the controller itself causes this problem since LSTMs store information on their memory cells, and thus, consecutive steps can have similar information [123]. This problem is also the cause of the variance observable in Figure 5.4. Section 5.2 addresses some of these problems by proposing a different memory-based design.

### 5.1.3 Design Choices

This section delves into the hyperparameters and design choices of SDNCs, highlighting their connection to the interpretability of the results.

**Memory Size.** The first parameter to consider is the size of the memory. The settings used in the previous examples include a large memory size to minimize overwriting the same locations. Indeed, given the design of SDNC, the network is forced to utilize empty locations as long as possible. Once the memory is full, SDNC selects the least used locations as candidates for the writing operations. As the only way to increase the usage of a location is by writing on it, the least used locations are those written during the first steps of the sequence. Future work could explore adjusting the design of SDNC to improve this mechanism. A simple alternative could be to increase the usage of read locations, but this choice is task-dependent since there are several tasks where a location could be useful only at the end (e.g., T-Maze), and, thus, it should be read only at the end of the task. Therefore, we suggest using a memory big enough to avoid overwritings. This mechanism represents one of the directions that future work could investigate to improve the design of SDNC. Concerning the number of read heads and write heads, the reader is encouraged to refer to the DNC paper [70]. There are no differences between single and multiple heads in terms of interpretability.

The only aspect to consider is that using more write heads implies more writing, and thus, a larger memory is needed to avoid overwriting.

**Number of Locations for explanations.** Another important hyperparameter is  $topk^r$ , which controls how many locations to consider for each step. This hyperparameter should be tuned in a validation set or selected based on the task. For example, in the T-Maze problem, the number is set to two based on the desired behavior to probe and the problem's structure. Conversely, in the case of the Story Cloze Test task, we perform a grid search over the values  $\{1, 5, 10, 25 > \mu_t^{r,i}\}$ , as shown in Table 5.4.

When considering a smaller number of read locations for each step, both the best and worst premise increase their fidelity. In contrast, when considering a larger number of read locations, both premises lower their scores. Therefore,  $topk^r$  can be selected based on the objective and the desired explanations. If we are interested only in the precision of the most important premise, then fewer steps are the best choice. Conversely, for the entire ranking, a larger number of steps increases the gap between indices.

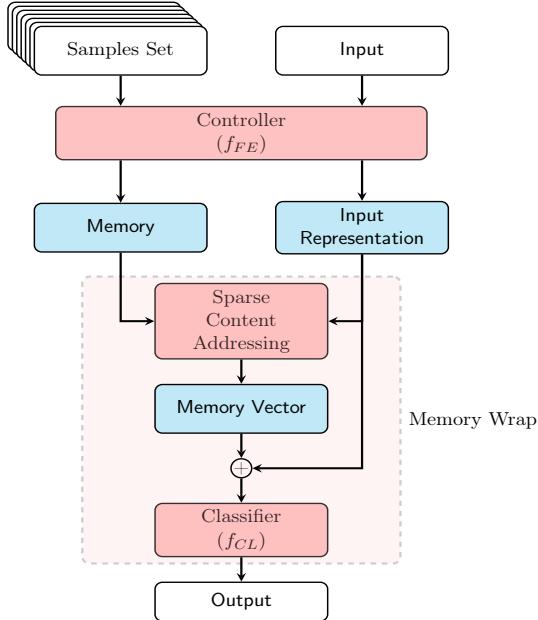
Note that the number of locations changes the type of promoted readings. For example, in the case of Table 5.4, if a few locations are selected for each step, at the end of the ending parsing, it is unlikely that the read words are associated with all the premises. Often, they are associated with only one or two premises, making the choice of the worst premise nearly random. However, these few selected locations are associated with the highest content weights and are highly influential; thus, the identification of the best premise is reliable. Conversely, if a large number of locations are selected for each step, all the premises will be represented, and the choice of the worst premise becomes more reliable. Instead, the choice of the best premise is influenced by the noise of the multiple readings, which reward words read most often rather than words read more constantly. Therefore, a premise read multiple times but in a single step can be scored as higher than one read more constantly during the parsing but using few locations per step.

## 5.2 Memory Wrap

The previous section described a first attempt at using memory modules for enhancing the interpretability of DNNs. SDNC is designed to deal with data sequence since MANNs have been historically applied in this domain. Being able to keep most of the training process intact, the SDNC design represents a step forward with respect to the prototype layers proposed in Chapter 4. However, we observed issues related to the controller dealing with data sequence.

**Table 5.4:** Avg. fidelity obtained on the development Story Cloze Test dataset using different  $topk^r$  values.

Avg. Fidelity- $acc$						
		Number of locations				
Premise		1	5	10	25	$ ind_t^{r,i} $
Best		<b>0.76</b>	<b>0.76</b>	0.73	0.71	0.40
Worst		0.60	0.60	0.53	<b>0.40</b>	0.49



**Figure 5.7:** Architecture of a deep neural network augmented with Memory Wrap.

Starting from this analysis, here we further develop the employment of memory modules for enhancing the interpretability of DNNs. The goal is to take a step forward to mitigate the issues highlighted in the previous section and investigate memory modules and the memory-tracking mechanism beyond the data sequence domain.

In particular, this section investigates the augmentation of popular DNNs with memory layers in a non-traditional domain for MANNs, the vision domain. In this context, the few available memory-based architectures [223, 250] usually target settings where few data are available for training (e.g., few-shot learning), and none of them is designed for interpretability.

To deal with such a domain, we propose Memory Wrap, **a memory-based layer that can be attached to any DNN and that does not require modifications in the training recipe of the augmented DNN**. Memory Wrap is designed to preserve the performance of the black-box model and avoid the problem of input dispersion (Section 3.2) by assigning a precise semantic to each memory slot. Additionally, its structure can be used by users to extract explanations by examples and counterfactuals. These explanations are useful for inspecting the decision process and detecting biased and uncertain predictions.

This section is organized as follows: Section 5.2.1 describes the architectural design of Memory Wrap and how to retrieve explanations; Section 5.2.2 evaluates the proposed module both in terms of performance and explanations; and Section 5.2.3 discusses alternative design choices for the module.

### 5.2.1 Design

Similarly to all the approaches discussed previously, **Memory Wrap** [124] can be placed after the feature extractor  $f_{FE}$ , which acts as a controller (see Figure 5.7).

The module, denoted as  $f_{MW}$ , receives two inputs: the latent representation of the current input  $\mathbf{x}_i$ , and the latent representations of a set of  $n$  sample  $S_j = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ . This set of samples serves as the memory  $\mathcal{M}$  of Memory Wrap.

Therefore, the output of the augmented DNN can be expressed as:

$$\begin{aligned} y &= f(\mathbf{x}_i, \mathbf{S}_j) \\ &= f_{MW}(f_{FE}(\mathbf{x}_i), f_{FE}(\mathbf{S}_j)) \end{aligned} \quad (5.17)$$

In contrast to the dynamic memory discussed in the previous section, where the controller decides at each step which information to store based on later outcomes, Memory Wrap is designed for data independence. Each sample in  $\mathbf{S}_j$  is treated as independent, free from temporal dependencies. Hence, the controller does not learn how to write the memory but can learn which information to read from it.

To fill the memory, an external function  $g$  is responsible for extracting samples from the dataset based on specific objectives and writing their latent representation into  $\mathbf{M}$ . After the writing operation, each memory row contains the latent representation generated by the controller for one of the samples in input. Optionally, a set of linear layers can project data into different manifolds and perform writing based on multi-head attention, similarly to Section 5.1.

Conversely, for reading operations, the module employs a sparse content addressing mechanism to assign zero probability to irrelevant input tokens. The content addressing weights are computed, as in Equation (5.6) (Section 5.1), thus, as the cosine similarity between the memory rows and the latent representation of the current input.

However, in this case, the mechanism applies the sparsemax function as a replacement for the softmax function. This replacement reduces noise in the decision process and improves interpretability, as highlighted by recent literature [146, 148]. Therefore, the content weights are computed as follows:

$$\mathbf{w}_c = C(\mathbf{x}, \mathbf{M}) = \text{Sparsemax}(D(f_{FE}(\mathbf{x}), \mathbf{M})) \quad (5.18)$$

The sparsemax is a sparse differentiable function proposed by Martins and Astudillo [147] that can be computed by using the algorithm proposed by Peters et al. [184]. This sorting-based algorithm finds the probability distribution satisfying the following equation:

$$\phi_2(x_j) = \arg \min_{\mathbf{p} \in \Delta^{n-1}} \mathbf{p}^T \mathbf{x} + \mathbf{H}_\alpha^t(\mathbf{p}) \quad (5.19)$$

where  $\Delta^{n-1}$  is the probability simplex,  $\alpha \in [0, 2]$  is a hyperparameter controlling the smoothness of the function, and  $\mathbf{H}_\alpha^t$  is the Tsallis entropy [241]:

$$\mathbf{H}_\alpha^t(\mathbf{p}) = \begin{cases} \frac{1}{\alpha(\alpha-1)} \sum_j (p_j - p_j^\alpha) & \alpha \neq 1 \\ -\sum_j p_j \log p_j & \alpha = 1 \end{cases} \quad (5.20)$$

For  $\alpha = 2$  (i.e., the value used for sparsemax), the Tsallis entropy reduces to the well-known Gibbs-Boltzmann-Shannon entropy.

To compute weights, we can use the following equation, which returns the solution to the system:

$$\phi_2(x_j) = \text{ReLU}([( \alpha - 1 ) \mathbf{x} - \tau \mathbf{1}]^{\frac{1}{\alpha-1}}) \quad (5.21)$$

where  $\tau$  is the Lagrange multiplier corresponding to the  $\sum_i p_i = 1$  constraint. For further details about the algorithm and the proof of the derivation, please refer to the work of Martins and Astudillo

$w_c$	$f_{FE}(\mathbf{x}_j)$	$f(\mathbf{x}_j)$
0.40	Sample 1	1
0.11	Sample 2	1
0.10	Sample 3	3
0.03	Sample 4	3
0.02	Sample 5	1
...	...	...
0	...	6
0	Sample n-1	4
0	Sample n	5

**Figure 5.8:** Illustration of a snapshot of the memory of Memory Wrap. Each row can be associated with one of the three sets  $M_0$ ,  $M_=$ , and  $M_≠$  based on its associated content weight  $w_c$  and prediction.

[147].

Since the readings are based on the content addressing mechanism, the network learns to read (i.e., select) samples from memory similar to the current input, aiding the classifier in the current prediction. The retrieval of similar memory samples and the combination with the input enables the network to overcome missing or noisy information in the current input. Moreover, the employment of sparse mechanisms and the disjoint encoding mitigates the issue of input dispersion.

After the computation of the content weights, Memory Wrap computes the memory vector  $\mathbf{v}_M$  as the weighted sum of the memory  $\mathbf{M}$ , where the weights are set equal to the sparse content weights computed in the previous step.

$$\mathbf{v}_M = \mathbf{M}^T \mathbf{w}_c. \quad (5.22)$$

Since the module employs the sparsemax, a few rows of the memory are represented inside  $\mathbf{v}_M$ . Finally, the classifier  $f_{CL}$  takes the concatenation of the memory vector and the input latent representation as input and returns the final prediction (see Figure 5.7):

$$y = f_{CL}(\mathbf{v}_M \oplus f_{FE}(\mathbf{x}_i)) \quad (5.23)$$

The idea is that by using the similarity and the sparsemax together, strong features of the target class will be more represented than features of other classes, thus helping the network in the decision process.

### Explaining Via Memory Tracking

In Section 5.1, the memory contains states connected to the parsing of input features. Hence, the memory tracking mechanism results in feature attribution explanations. Conversely, in this case, the memory contains different samples, and the memory tracking mechanism computes example-based explanations: explanations by examples and counterfactuals.

Since Memory Wrap uses the sparsemax, we can identify three disjoint sets of rows in the memory  $\mathbf{M}$  (Figure 5.8). Each row  $j$  in  $\mathbf{M}$  corresponds to the latent representation generated by the controller when fed with the sample  $x_j$ . First, we can isolate the samples that have no impact on the decision process because they are associated with content weights equal to 0. Consequently, these samples are not represented in the read vector and are collected in the set  $\mathbf{M}_0$ . Since these samples are associated with zero weights, the neural network deems them different from the current input, and thus, they do not represent good candidates for explanations. Hence, the set  $\mathbf{M}_0$  is disregarded when searching for potential candidates for explanations. Note that this set exists due to the employment of the sparsemax function: any softer version would produce only the other two sets.

To identify the other two sets, each memory sample associated with a content weight greater than zero is fed to the network, and the corresponding predictions are gathered. Among these samples, a distinction is made between those associated with the same prediction as the current input, collected in the set  $\mathbf{M}_\equiv$ , and those associated with a different prediction, collected in the set  $\mathbf{M}_\neq$ . In mathematical terms,  $\mathbf{M}_\equiv$  and  $\mathbf{M}_\neq$  can be defined as:

$$\mathbf{M}_\equiv = \{f_{FE}(\mathbf{x}_j) \mid f(\mathbf{x}_i, S_j) = f(\mathbf{x}_j, S_k) : \mathbf{x}_j \in S_j \wedge S_j \neq S_k\} \quad (5.24)$$

$$\mathbf{M}_\neq = \{f_{FE}(\mathbf{x}_j) \mid f(\mathbf{x}_i, S_j) \neq f(\mathbf{x}_j, S_k) : \mathbf{x}_j \in S_j \wedge S_j \neq S_k\} \quad (5.25)$$

and clearly,  $\mathbf{M}_\equiv$ ,  $\mathbf{M}_\neq$ , and  $\mathbf{M}_0$  satisfy the following constraints:

$$\mathbf{M}_\equiv \cup \mathbf{M}_\neq \cup \mathbf{M}_0 = \mathbf{M} \quad (5.26)$$

$$\mathbf{M}_\equiv \cap \mathbf{M}_\neq \cap \mathbf{M}_0 = \emptyset \quad (5.27)$$

$\mathbf{M}_\equiv$  and  $\mathbf{M}_\neq$  inherently represent good sets of candidates for explanations by examples and counterfactuals, respectively. More precisely, the best candidate for being used as an explanation by example is the sample in  $\mathbf{M}_\equiv$  associated with the highest weight:

$$\mathbf{X}_e = \{\mathbf{x}_k \mid f_{FE}(\mathbf{x}_k) \in \mathbf{M}_\equiv \wedge \forall \mathbf{x}_j \in \mathbf{M}_\equiv : w_j \leq w_k\} \quad (5.28)$$

Intuitively, the sample selected  $\mathbf{x}_e \in \mathbf{X}_e$  represents a good explanation by example for three reasons. Firstly, it is highly activated, actively contributing to the current prediction. Secondly, a high content weight indicates similarity between the encoding of the input  $\mathbf{x}_i$  and the encoding of  $\mathbf{x}_e$ ; thus, the network treats them similarly. Lastly, it is predicted in the same class as the current input, and thus, the common features between the explanation by example and the current input can inform the user about the features exploited by the network for that prediction.

Similarly, the best candidate for being used as a counterfactual is the sample in  $\mathbf{M}_\neq$  associated with the highest weight:

$$\mathbf{X}_c = \{\mathbf{x}_k \mid f_{FE}(\mathbf{x}_k) \in \mathbf{M}_\neq \wedge \forall \mathbf{x}_j \in \mathbf{M}_\neq : w_j \leq w_k\} \quad (5.29)$$

In this case, the sample selected  $\mathbf{x}_c \in \mathbf{X}_c$  represents a good counterfactual explanation because it is associated with a high content weight and predicted in a different class. A different prediction

indicates that despite the high similarity between the two samples, the network yields different predictions. Therefore, examining the differences between the samples can help users discern the necessary changes to achieve a different prediction. This information aligns with the criteria of a good counterfactual (Section 2.2). Finally, note that  $\mathbf{x}_c \in \mathbf{X}_c$  is partially included in the memory vector, and frequently, the counterfactual class is the second or third predicted class. Therefore, its inspection can help users understand the uncertainties of the neural network.

In the rare case of multiple samples associated with the highest weights (i.e., the cardinality of  $\mathbf{X}_c$  or  $\mathbf{X}_e$  greater than 0), several criteria can be employed to resolve the tie. For instance, the samples closer in the input space can be selected, or multiple explanations can be provided for the same input.

### 5.2.2 Experiments

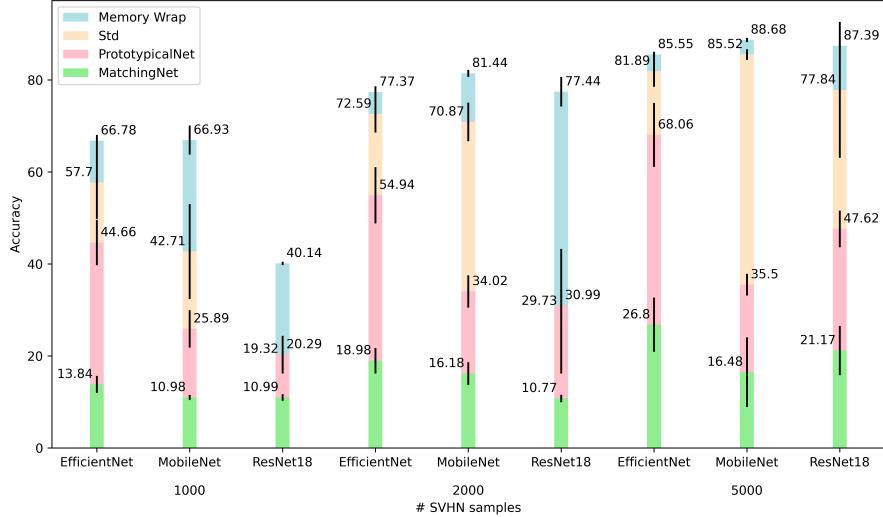
#### Performance

We test Memory Wrap on image classification tasks on several popular DNNs. Namely, we train from scratch and augment ResNet18 [80], EfficientNetB0 [234], MobileNet-v2 [207], GoogLeNet [232], DenseNet [92], ShuffleNet [281], WideResNet 28x10 [277], and ViT [49] with Memory Wrap and compare their performance against the networks trained without the module on the Street View House Number (SVHN) [168], CINIC10 [42] and CIFAR10 [117] datasets. To train the models, we follow the training setup of the respective papers, repositories, or datasets **without any modification**.

Specifically, we follow the setup of Huang et al. [92] for SVHN and CIFAR10 and the suggested procedure for training for CINIC10 for all the models but WideResNet [277] and ViT [49], for which we follow the alternative setup described in their papers. The procedure for all the other models consists of training the models using the Stochastic Gradient Descent (SGD) algorithm and decreasing the learning rate by a factor of 10 after 50% and 75% of epochs. In all the cases, we initialize the learning rate to 1e-1. We apply a data augmentation based on random horizontal flips and train the networks for 40 epochs in the SVHN dataset and 300 epochs in CIFAR10 and CINIC10 (200 when the controller is ViT). Regarding the parameters of Memory Wrap, we use a memory of 100 randomly extracted samples from the dataset (i.e., the  $g$  function is the random selection) and, as a classifier, a multi-layer perception with one hidden layer containing a number of units equal to double the dimension of the input. Note that, ideally, each sample should be associated with a different memory set. However, supporting this type of mapping would require too much space, even for small batch sizes. Indeed, if  $m$  is the size of memory and  $n$  is the dimension of the batch, then the new input would contain  $m \times n$  samples in place of  $n$ . To reduce the memory footprint, we provide a shared memory set for each batch, thus reducing the space required to  $m + n$ .

Since MANNs [223, 250] used in the vision domain usually target settings where few data are available for training, we test four different sizes for the training data: 1000, 2000, 5000 samples, and the entire dataset. The sets correspond to the 1%, 2%, 5%, and 100% of the labeled samples in CINIC10 and 2%, 4%, 10%, and 100% of the labeled samples in the other datasets.

We begin the evaluation by comparing the networks augmented with Memory Wrap against networks without the module (std), Matching Networks [250], Prototypical Networks [223], K-NN [39], sparse K-NN, and an ablated version of Memory Wrap when trained on the above-mentioned



**Figure 5.9:** Performance comparison between memory-based modules.

datasets' subsets. These competitors cover several alternatives regarding memory usage, interpretable decision processes, architectures for small data settings, and module designs.

Specifically, Prototypical Networks are memory-based networks that compute prototypes for each class as the mean of the samples of that class in the memory set, and then they use the distance between the prototypes and the input to compute the prediction.

Matching Networks are memory-based networks that classify inputs based on a weighted linear combination of the labels in the memory set, where the weights are the distances between its samples and the current input computed over an embedding learned by a recurrent network.

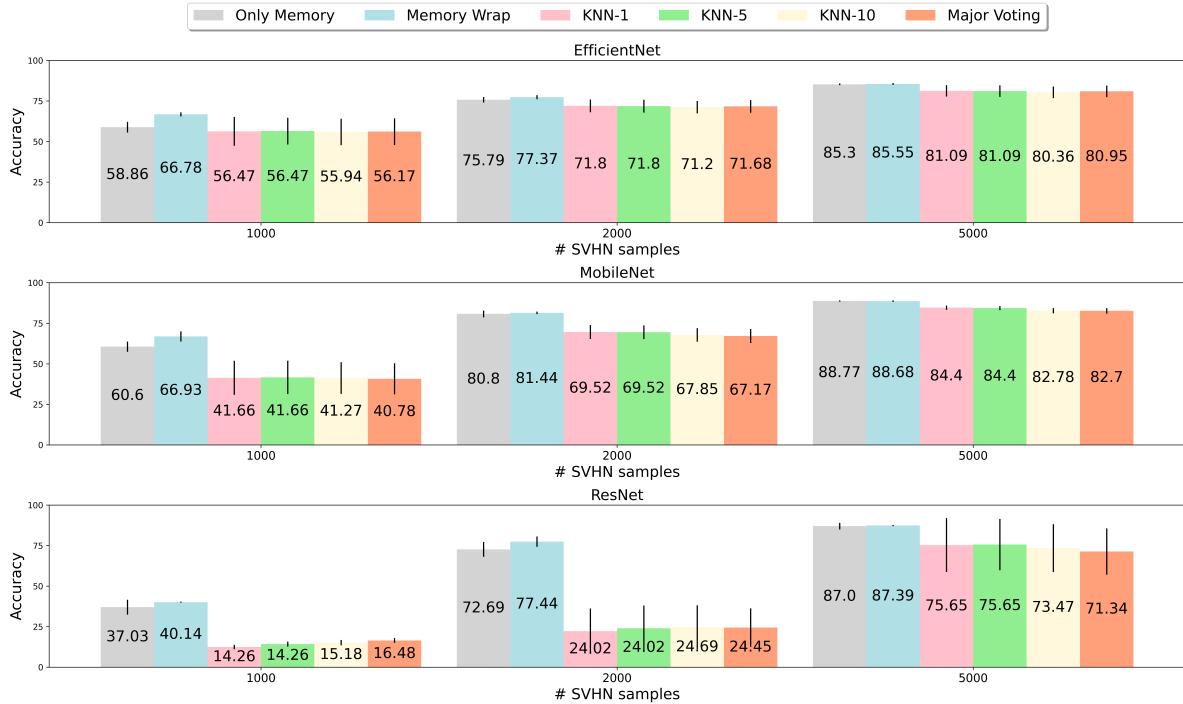
K-NN [39] is a popular baseline in these settings [250], and it is an interpretable model since the prediction of each input is based on the mode of the labels of its  $k$  nearest neighbors (i.e., the closest in the latent space).

Sparse K-NN replaces the selection of the  $k$  parameter with the usage of sparsemax. In this case, the label is chosen based on the mode of the labels of samples in memory associated with a weight greater than zero.

Finally, the ablated version of Memory Wrap (MemOnly) uses only the read vector as input to the classifier, removing the skip connection from the controller. Therefore, this network uses the input only to compute the sparse content-based attention weights and compute the read vectors.

We use MobileNet, ResNet18, and EfficientNet trained on SVHN as the largest testbed to filter the best competitors for the full experiments. In this test, we use three different sizes for the training data: 1000, 2000, and 5000 samples. We collect the competitors' performance on the official testing dataset over 15 different experiments and analyze the average accuracy.

In Figure 5.9 and Figure 5.10, we observe that **Memory Wrap reaches the highest performance across all the configurations against both the memory competitors and the interpretable ones in small data settings**. In these settings, the interpretable models are more performant than the memory ones. Indeed, the K-NN baseline reaches comparable performance with respect to the black-box model alone (std), dropping its performance of just 1-2% but gaining a more interpretable decision process. This result aligns with recent findings about this strong



**Figure 5.10:** Performance comparison between k-NN variants and Memory Wrap variants.

baseline [250]. Note that while the best  $k$  to be used as a hyperparameter of this baseline varies across the configurations, the results across them are similar most of the time, and thus, the choice of  $k$  has no big impact.

Matching Networks and Prototypical Networks reach the lowest performance across all the configurations. We argue that the cause of low performance resides in how they encode and use the memory samples, tailored to the few-shot learning scenarios. For example, Matching Networks embed both the input and memory using recurrent networks (LSTM). When trained on small datasets, the resulting embeddings capture a small portion of the semantics. Thus, they are less effective since the embeddings of the memory set can be misaligned with respect to the input one. Prototypical Networks mitigate this problem by encoding both the input and memory set using the controller in the same way as Memory Wrap. However, the fact that the output of Prototypical Networks is based on all the samples in memory (i.e., by computing the average encoding) can create instability. Indeed, the presence of outliers in memory, a common scenario when samples are not selected a priori, can produce a large shift in the prediction.

Memory Wrap combines the positive qualities of all the competitors. Like Prototypical Networks, Memory Wrap learns how to use the memory during the training process and encodes both the input and the memory samples using the same controller. Like K-NNs, it selects a subset of similar samples from memory, and thus, only a small number of samples in memory can have an impact on the decision process, avoiding the problem of outliers. The difference with K-NN is that the latter is trained independently from the black-box model, and thus, the performance of the black-box model represents an upper bound. Conversely, Memory Wrap jointly learns the controller and how to use its encodings to exploit the memory mechanisms, improving both components at the same time. These characteristics allow Memory Wrap to reach an improvement in performance and data

efficiency between  $2.5\times$  (ResNet) and  $1.5\times$  (EfficientNet and MobileNet) over the black-box models. Regarding the ablated version, we can observe that it performs better than the black-box models. However, the information carried on by the skip connection in the full Memory Wrap is crucial to achieving the best absolute performance, likely due to the access of shortcuts in it (e.g., the presence of rare features).



**Figure 5.11:** Avg. accuracy and standard deviation over 15 runs of the black-box model, the full Memory Wrap, and its ablated version on subsets of CIFAR10.

From the previous results, we select the three best competitors (Memory Wrap, the black-box model, and the ablated Memory Wrap) and evaluate their performance on CIFAR10 (Figure 5.11), CINIC10 (Figure 5.12), and the rest of the selected DNNs. We observe that Memory Wrap achieves the best performance across all the models, confirming the results of the previous case. However, here we can note two differences: the ablated version reaches lower performance than the black-box models in ViT and WideResNet, and the gap between Memory Wrap and the black-box models is shorter on average than the gaps reported in SVHN. Indeed, CIFAR10 and CINIC10 represent a less structured dataset than SVHN. In CIFAR10 and CINIC10, classes share several features, like colors, and there is a high variability among the samples of the same class. For example, a “stealth airplane” has completely different colors and shapes from commercial airplanes. These characteristics lower the benefits of exploiting the similarity in memory since samples retrieved by similarity can be associated with other classes. In these cases, the skip connection of Memory Wrap becomes crucial, allowing the model to tolerate wrong retrieval from memory better than the ablated version.

Finally, Table 5.5 reports the performance of Memory Wrap when it is trained using the full dataset. We observe that Memory Wrap achieves, on average, the same performance as the black-box models. Therefore, **in the full dataset scenario, there is no gain in performance when**



**Figure 5.12:** Avg. accuracy and standard deviation over 15 runs of the black-box model, the full Memory Wrap, and its ablated version on subsets of CINIC10.

**applying Memory Wrap.** However, in contrast with black-box models, Memory Wrap provides ways to interpret its behavior, which we discuss in the next paragraph, thus representing a useful way to augment the black-box models also in these cases.

## Explanations

This paragraph evaluates and showcases the explanations extracted by analyzing the Memory Wrap structure. All the experiments and figures use MobileNet as a controller. However, the findings also hold for all the other black-box models.

Following the evaluation conducted in Section 5.1.2, we begin our evaluation of the explanations linked to the Memory Wrap design by investigating the reliability of the ranking returned by the

**Table 5.5:** Avg. accuracy and standard deviation of the baselines and Memory Wrap when the training datasets are the whole SVHN, CIFAR10 and CINIC10 datasets.

Encoder	Model	Full Datasets Avg. Accuracy %		
		SVHN	CIFAR10	CINIC10
EfficientNet	Black-Box	94.39 ± 0.24	88.13 ± 0.38	77.31 ± 0.35
	Memory Wrap	<b>94.67 ± 0.16</b>	<b>88.05 ± 0.20</b>	<b>77.34 ± 0.27</b>
MobileNet	Black-Box	<b>95.95 ± 0.09</b>	<b>88.78 ± 0.41</b>	<b>78.97 ± 0.31</b>
	Memory Wrap	95.63 ± 0.08	<b>88.49 ± 0.32</b>	<b>79.05 ± 0.15</b>
ResNet18	Black-Box	<b>95.70 ± 0.10</b>	<b>91.94 ± 0.19</b>	<b>82.05 ± 0.25</b>
	Memory Wrap	95.49 ± 0.11	91.49 ± 0.17	<b>82.04 ± 0.16</b>

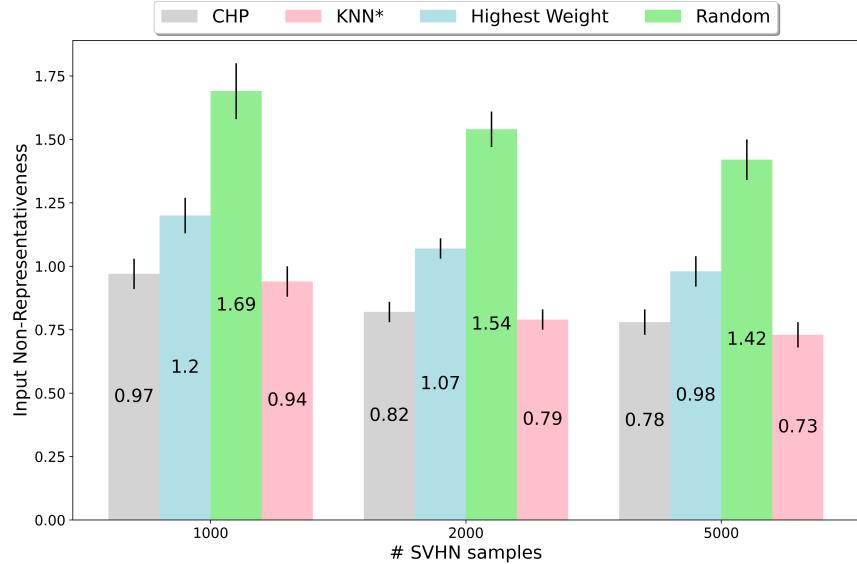
sparse content-based attention weights. To examine the ranking, we select the sample  $\in M_{ce}$  with the highest weight (*Top*), the sample  $\in M_{ce}$  with the lowest weight (*Bottom*), and a random sample from memory. Then, we feed these samples as input to the network and collect their predictions. Finally, we measure the *prediction matching accuracy* by counting how frequently these predictions correspond to the prediction associated with the current input.

Table 5.6 shows that the matching accuracy achieved by the sample with the highest weight is consistently higher than that of the randomly selected sample and the sample with the lowest weight. The random selection is the worst performer because it chooses a random example from samples of all classes. Given that the dataset has ten classes, the probability of selecting the correct class is approximately 10%. Conversely, the sample with the lowest weight actively contributes to the decision-making process, and as it resembles the current input, it may belong to the same class or a different class that could be misclassified into the same class. Furthermore, we note that the higher the model’s accuracy, the higher the bottom sample’s matching accuracy. This observation is consistent with the finding that when Memory Wrap achieves high accuracy, the memory samples associated with a weight greater than zero are typically all samples from the same class, thereby enhancing the matching accuracy. Collectively, these results confirm the **reliability of the sparse content-based attention weights ranking and the quality of the sample with the highest weight as a good proxy for prediction**.

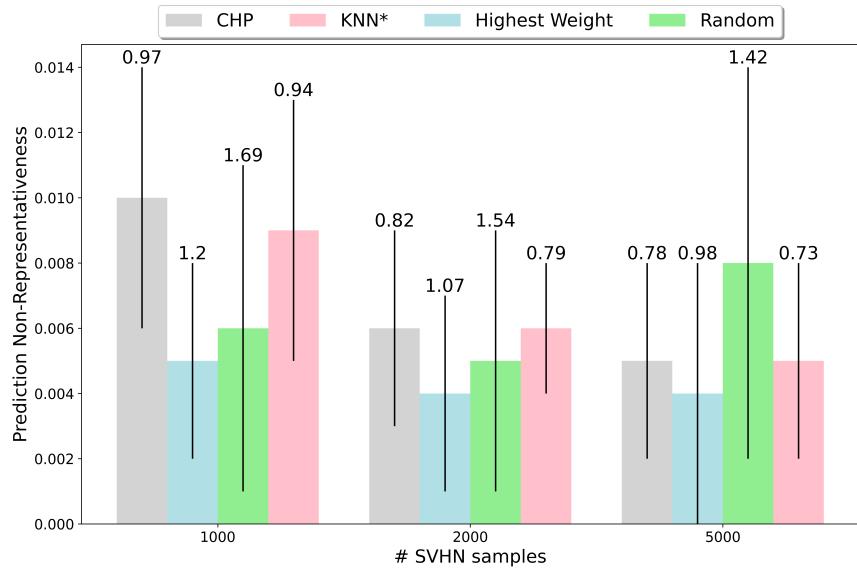
Once the reliability of the ranking has been proved, we can assess the quality of explanations by examples and counterfactuals extracted using the method outlined in Section 5.2.1. The explanations by examples are evaluated using the *input non-representativeness* [106] (Equation (2.29)) and the *prediction non-representativeness* [188] (Equation (2.30)) metrics. The counterfactuals are evaluated using IM1 (Equation (2.31)), IM2 [138](Equation (2.33)), and IIM1 (Equation (2.32)) scores. Competitors include Memory Wrap, the random selection, and the posthoc methods CHP [106] and KNN\* [179] for explanations by examples and a method based on prototypes [138] for counterfactuals. These methods are chosen for their publicly available implementation, recent relevance, and compatibility with Memory Wrap’s assumptions (i.e., they have been tested on image data and do not require additional knowledge). Note that the post-hoc methods require more computation time than Memory Wrap and that Memory Wrap is not optimized during the training process to select

**Table 5.6:** Avg. and standard deviation of the matching accuracy reached over 15 runs by the sample in the memory set with the highest sparse content-based attention weight (Top), the example with the lowest weight but greater than zero (Bottom) and a random sample (Random)

	Prediction matching accuracy %		
Example	1000	2000	5000
SVHN			
Top	<b>84.24 ± 1.22</b>	<b>90.59 ± 0.52</b>	<b>94.47 ± 0.22</b>
Bottom	<b>46.46 ± 1.77</b>	<b>57.39 ± 1.09</b>	<b>69.94 ± 1.37</b>
Random	11.76 ± 0.30	11.66 ± 0.17	11.71 ± 0.13
CIFAR10			
Top	<b>82.04 ± 1.14</b>	<b>87.75 ± 0.72</b>	<b>91.76 ± 0.22</b>
Bottom	<b>46.01 ± 1.92</b>	<b>60.10 ± 1.29</b>	<b>69.94 ± 0.82</b>
Random	10.22 ± 0.28	10.23 ± 0.20	9.80 ± 0.43
CINIC10			
Top	<b>76.31 ± 0.73</b>	<b>78.50 ± 0.50</b>	<b>78.45 ± 0.62</b>
Bottom	<b>37.01 ± 1.12</b>	<b>41.34 ± 0.73</b>	<b>37.55 ± 1.49</b>
Random	10.47 ± 0.19	10.30 ± 0.11	10.16 ± 0.12



**Figure 5.13:** Avg. Input representativeness scores and standard deviation reached by CHP (gray), KNN\* (pink), the random baseline (green), and Memory Wrap (blue). Lower is better.



**Figure 5.14:** Avg. Prediction representativeness scores and standard deviation reached by CHP (gray), KNN\* (pink), the random baseline (green), and Memory Wrap (blue). Lower is better.

the best possible explanations, but the explanations are a side product of its design.

Figure 5.13 and Figure 5.14 show that **explanations by example selected by Memory Wrap are better than the competitors** in terms of prediction non-representativeness and worse than them in terms of input non-representativeness. While the lower performance in input non-representativeness is expected since competitors are designed to optimize this score, the discrepancy in non-representativeness prediction deserves further analysis. This behavior can be explained by analyzing the cases when the network is uncertain, causing the logits of several classes to be close. Since, by design, the considered post-hoc methods select samples with low L<sub>1</sub> distance, they pick other uncertain samples, leading to low average input non-representativeness. As the samples

selected by the post-hoc methods are challenging samples for the network, the network may predict them in different classes, especially if trained on the lowest settings, increasing the prediction non-representativeness, which is based solely on the predicted classes. In Memory Wrap, the uncertainty is usually encoded in the counterfactuals included in the set  $M_{\neq}$ . Therefore, the sample selected as an explanation by example tends to be a strong prototype of the predicted class, resulting in low prediction non-representativeness and higher input non-representativeness.

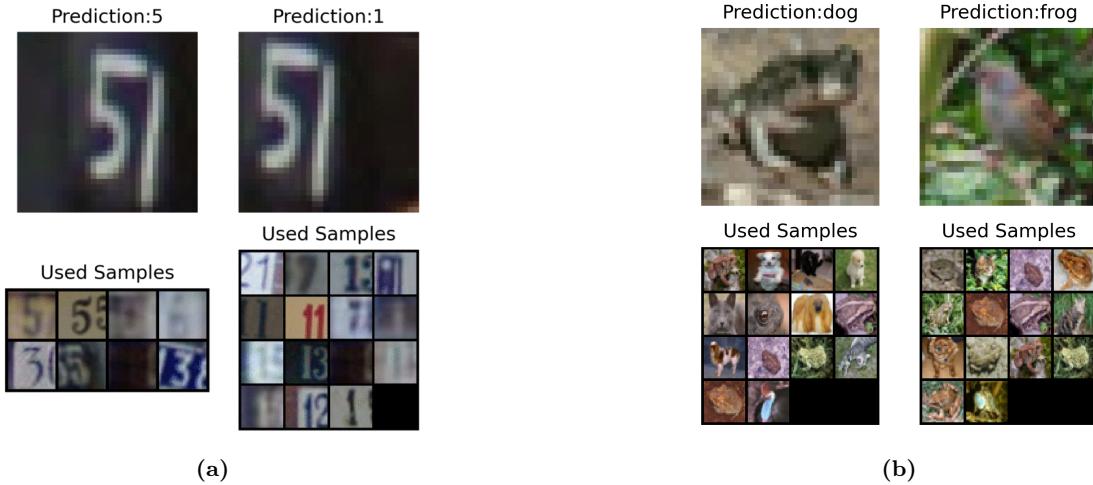
Table 5.7 compares counterfactuals in terms of IM1, IM2 [138], and IIM1 scores [124]. Since the post-hoc method generates counterfactuals, it scores lower in both IM1 and IM2. Indeed, Memory Wrap selects real samples by design, keeping the distribution identical to the training one. Even when it selects edge cases (e.g., confusing images), their distribution is usually closer to one of the counterfactual classes than one of the input classes. More interestingly, Memory Wrap reaches worse scores in IIM1, but the gap is not large, despite the fact Memory Wrap has limited access to the possible counterfactuals due to the random selection of the memory set and the sparsemax, which filter out most of the candidate counterfactuals.

Now, we describe how users can exploit the design and characteristics of Memory Wrap to gain insights into its decision process.

First of all, as described in Section 5.2.1, the **presence of counterfactuals can be used as a means to detect uncertainty** in the decision process. This property is tested on MobileNet and SVHN by measuring the Person correlation [225] between the correctness of the prediction, the number of counterfactuals in  $M_{\neq}$  and the index of the counterfactual in the list of memory samples sorted by their content weights. Table 5.8 confirms this hypothesis, revealing a positive correlation between the position of the highest-rated counterfactual and the prediction correctness and a negative correlation between counterfactual count and correctness. Thus, an increase in both the number and rating of counterfactuals in memory is a marker of increased uncertainty in the network’s prediction. These results align with the findings in Table 5.9, where predictions featuring the highest rated counterfactual in memory reach a substantially lower accuracy than predictions with no counterfactuals in memory. Consequently, the absence of counterfactuals or their index can offer valuable insights into the reliability of the current prediction.

**Table 5.7:** Avg. IM1, IM2, and IIM1 scores on the SVHN dataset reached by Memory Wrap and a prototype-based method. Lower is better.

Counterfactuals Scores Avg.			
	Samples		
	Score	Dataset	MemoryWrap Proto [138]
IM1	1000	0.991	0.994
	2000	0.985	0.997
	5000	0.985	0.990
IM2(x10)	1000	1.703	2.623
	2000	1.674	2.707
	5000	1.698	2.770
IIM1	1000	1.014	-
	2000	1.015	-
	5000	1.013	-



**Figure 5.15:** Inputs (first rows) from SVHN (a) and CIFAR10 (b), their associated predictions, and an overview of the samples in the memory set that have an active influence on the decision process (second row).

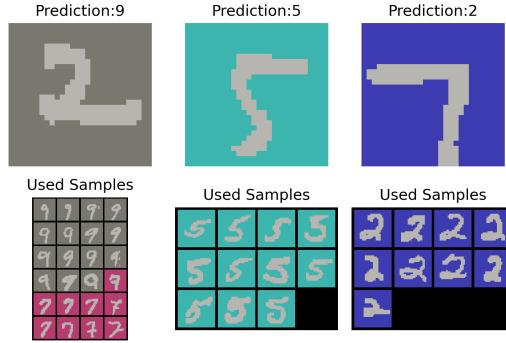
Figure 5.15a and Figure 5.15b show another couple of examples from SVHN and CIFAR10. Two horizontally shifted input images are fed with the same memory samples in the first case (Figure 5.15a). Despite their similarity, Memory Wrap selects different samples from the memory for each input. We can analyze the samples associated with a positive weight to **understand the decision process**. The left image is associated with samples containing images of “5” and “3”. On the other hand, the right image is associated with samples containing images of “7” and “1”. Despite their high similarity, no samples have been associated with both images, indicating the network correctly focuses on the center image while disregarding other objects. Furthermore, the inspection of the memory sample reveals that the shapes have a bigger impact than the background colors in these decisions, which is evident from the low rank of samples with dark backgrounds. Conversely, in Figure 5.15b, colors and background exert a bigger influence, particularly in the right image (i.e.,

**Table 5.8:** Avg. and standard deviation of the Pearson product-moment correlation coefficients between the correctness of the predictions, the number of counterfactuals (number), and the position of the first counterfactual (position).

Avg. Pearson Coefficient			
Model	1000	2000	5000
Position	0.39 ± 0.02	0.40 ± 0.01	0.41 ± 0.01
Number	-0.42 ± 0.01	-0.45 ± 0.01	-0.46 ± 0.01

**Table 5.9:** Accuracy reached by the model on SVHN when the sample with the highest weight in the memory set is a counterfactual (Top Counter) and when there are no counterfactuals at all (No Counter).

Avg. Accuracy % (Coverage%)		
Counter	1000	2000
Top	35.45 (19.32)	38.29 (12.06)
No	95.20 (18.73)	97.00 (36.47)



**Figure 5.16:** An example of a model biased towards the background. The figure shows how the inspection of the memory samples used by the network makes clear the reasons behind the bad performance of the network.

the bird predicted as a frog): the majority of memory samples include a green background or brown animals in the center.

Finally, memory inspection can be used by developers to **detect biased decisions**. For example, Figure 5.16 shows some examples of predictions of a biased EfficientNetB0 model trained on CMNIST [14], a dataset where we introduce a strong bias that correlates the background of images with their labels at training time. In this case, it is easy to see that the inspection and visualization of the memory samples help in detecting the bias. Indeed, the samples associated with a weight greater than zero show how the model uses the background to classify the digits: in all the cases, the highest-rated samples are images with the same color background, although the digit is different, thus exposing the learned bias.

### 5.2.3 Design choices

In the following paragraphs, we explore alternative design choices for Memory Wrap.

**g function.** We begin with the selection of the  $g$  function responsible for choosing the memory samples from the datasets. To be feasible, the  $g$  function should be sufficiently fast to avoid slowing

**Table 5.10:** Avg. accuracy and standard deviation over five runs Memory Wrap using different selection mechanisms to select samples in the memory set.

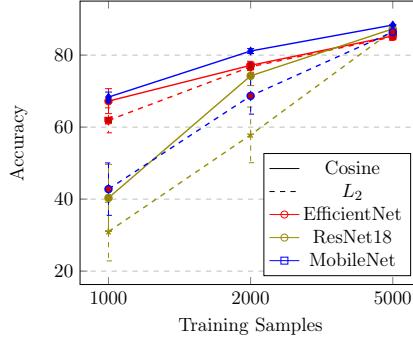
		SVHN Avg. Accuracy %		
Encoder	Distance	1000	2000	5000
EfficientNet	Random	$67.22 \pm 3.47$	$77.16 \pm 0.97$	$85.32 \pm 0.90$
	Balanced	$65.80 \pm 2.74$	$77.19 \pm 1.27$	$85.80 \pm 0.41$
	Replay-Last	$66.53 \pm 1.80$	$76.36 \pm 1.09$	$85.80 \pm 0.49$
	Replay-Last5	$66.32 \pm 1.70$	$77.56 \pm 1.57$	$85.50 \pm 0.54$
MobileNet	Random	<b><math>68.34 \pm 1.40</math></b>	$81.14 \pm 0.69$	<b><math>88.36 \pm 0.09</math></b>
	Balanced	<b><math>67.19 \pm 1.77</math></b>	$81.45 \pm 1.30$	<b><math>88.36 \pm 0.68</math></b>
	Replay-Last	<b><math>66.92 \pm 1.95</math></b>	$81.17 \pm 1.35$	<b><math>88.51 \pm 0.29</math></b>
	Replay-Last5	$65.93 \pm 1.98$	$82.09 \pm 0.93$	$88.19 \pm 0.26$
ResNet18	Random	$40.35 \pm 9.31$	$74.24 \pm 2.70$	$87.29 \pm 0.39$
	Balanced	$40.74 \pm 8.98$	$75.07 \pm 4.20$	<b><math>87.43 \pm 0.25</math></b>
	Replay-Last	$38.75 \pm 11.90$	<b><math>78.56 \pm 1.84</math></b>	$87.35 \pm 0.85$
	Replay-Last5	$38.73 \pm 14.25$	<b><math>79.55 \pm 1.16</math></b>	<b><math>87.70 \pm 0.34</math></b>

down the training process. Consequently, we exclude functions based on the similarity to a specific input (e.g., selecting the top- $k$  most similar samples in the training set). Instead, we test several random selection methods: pure random, balanced random, from the previous batch, and from the previous  $k$  batches. The *pure random* competitor selects  $m$  training samples randomly from the dataset; the *balanced random* competitor works similarly to the pure random one but ensures that each class has  $\frac{m}{\text{num\_classes}}$  samples in memory on average; the third competitor selects  $m$  samples from the last batch, assuming the batch size is larger than  $m$ ; the last competitor selects  $m$  samples from the pool of samples from the last  $k$  batches. Note that the last two competitors apply the selection at training time but not at testing time since, at testing time, no learning occurs, and the benefits of selecting the previous batches diminish.

Table 5.10 presents the performance across the SVHN dataset and three different DNNs. No clear winners emerge, with all competitors achieving almost identical performance, especially when applied to EfficientNet. Consequently, they all converge towards random selection. This phenomenon can be explained by analyzing the potential advantages of the competitors. The balanced random competitor selects a fair amount of samples for each class. However, due to sparsemax, Memory Wrap selects a few samples to be used during the inference process. Thus, providing more samples for each class does not enhance the quality of the memory unless samples similar to the current input are chosen. Since the balanced competitor is random and the difference in the number of samples per class is minimal compared to the pure random competitor, the advantages of 1-2 additional samples per class are nullified by the design of Memory Wrap itself. Moreover, the rare cases where few or no samples are provided for a specific class can be considered random noise useful to regularize the training process. Regarding the competitors using the last batches, the potential advantage lies in feeding samples that the network is able to recognize since the network’s weights have just been updated to recognize them better. However, while this advantage can play a role during the first epochs of the training process, it diminishes as soon as the network improves its performance. Additionally, note that the similarity is computed with the current input, which is novel with respect to recent samples on which the weights have been updated. Therefore, when the current input is not correctly recognized, the advantage becomes a weakness of these selection mechanisms.

Regarding the testing time, the selection of memory samples is flexible and can be adapted to the application context. We argue that the random choice serves a robust baseline and, on average, is more resistant to adversarial and privacy attacks, the memory samples are in-distribution with respect to the training distribution, and the randomness ensures diversity. However, in contexts where the user can access perfect representative prototypes satisfying the requested objective (e.g., fairness, diversities, etc.) for each class, a fixed selection could be employed without any loss of precision.

**Memory Size.** Linked to the  $g$  function, a second design choice concerns the **number of samples** stored in memory. This number is influenced by the number of classes in the dataset, the memory footprint, and the training time. Given fixed hardware and a minimum number of samples per class, the memory size grows according to the number of classes. A dataset with 1000 classes (e.g., ImageNet) requires a memory of 10000 samples to provide ten samples per class, which is a challenging requirement for most common workstations. Conversely, a binary dataset requires



**Figure 5.17:** Comparison between different encoders trained using Cosine Similarity (solid lines) and  $L_2$  distance (dashed lines) on a subset SVHN dataset.

a small memory, and the footprint is negligible. Moreover, the larger the memory, the longer the training time due to the processing time of the samples. Table 5.11 reveals that providing fewer than ten samples per class negatively impacts the performance. Providing more than ten samples is beneficial for some configurations. However, since the gain tends to diminish progressively and the impact on training time tends to increase, the number of samples becomes a hyperparameter to tune for Memory Wrap as a trade-off between them.

**Similarity.** Finally, Figure 5.17 compares  **$L_2$  distance measures** and the cosine similarity when used as a similarity for the content addressing mechanism. We observe that, as noted by previous studies [70, 250], the cosine similarity outperforms  $L_2$  in almost all the configurations, representing the optimal choice for memory operations in neural networks.

**Table 5.11:** Avg. accuracy and standard deviation over five runs of Memory Wrap trained using a variable number of samples in memory, when the training dataset is a subset of SVHN.

Encoder	Memory	Reduced SVHN Avg. Accuracy%		
		Samples		
EfficientNet	20	64.95 ± 2.46	75.80 ± 1.17	84.86 ± 0.99
	100	67.16 ± 1.33	77.02 ± 2.20	85.82 ± 0.45
	300	66.70 ± 1.58	77.97 ± 1.34	85.37 ± 0.68
	500	66.76 ± 0.98	77.67 ± 1.17	85.25 ± 1.02
MobileNet	20	63.42 ± 2.46	80.92 ± 1.42	88.33 ± 0.36
	100	68.31 ± 1.53	81.28 ± 0.69	88.47 ± 0.10
	300	65.08 ± 0.30	82.05 ± 0.75	88.93 ± 0.37
	500	69.88 ± 1.76	80.92 ± 1.74	88.61 ± 0.32
ResNet18	20	39.32 ± 7.21	72.54 ± 3.03	87.30 ± 0.41
	100	40.38 ± 9.32	74.36 ± 2.69	87.39 ± 0.45
	300	44.42 ± 10.97	74.63 ± 3.28	87.75 ± 0.62
	500	40.59 ± 12.27	76.97 ± 2.48	87.55 ± 0.35

### 5.3 Contributions

This chapter delved into the design of self-explainable DNNs based on memory. Specifically, it introduced two distinct designs of MANNs. The first design is fully differentiable and capable of employing memory for both readings and writings. The second is restricted to reading from the memory. We analyzed the advantages and weaknesses associated with their utilization for enhancing the interpretability of DNNs.

One of the strongest advantages lies in the fact that the augmentation based on memory improves or maintains the performance of the black-box model alone. These results align with the one obtained in the previous chapter (Chapter 4). Consequently, the trade-off of opting for a more interpretable model is moved from the performance aspect to considerations related to the training time and the memory footprint. These methods are also model-agnostic, denoting their agnosticism concerning the models they augment. Indeed, we applied their augmentation to three classes of architectures: recurrent, convolutional, and attention-based. Memory Wrap is also data-agnostic since it can be potentially applied to sequential data or any other kind of data as long as they are divided into separate samples. Additionally, we demonstrated their high flexibility in the types of explanations that can be derived, spanning from single and group feature attributions to explanations by example and counterfactuals. Alongside these strengths, we also analyzed some of the drawbacks of these approaches, such as the number of hyperparameters to tune in the case of fully differentiable architectures, their impact on the explanation process, the large footprint required by MANNs, and the reliance on the skip connection of the controller, which remains a source of opacity in the decision process.

This chapter closes the Part II focused on the description of the designs that add novel layers to DNNs to enhance their interpretability. Specifically, the contribution of this part includes a novel prototypes-based layer for GNNs and the introduction of a novel family of self-explainable DNNs: explainable memory-based architectures. The proposed designs preserve or even improve the performance of the black-box models to which they attach these layers, thus nullifying the typical performance trade-off of self-explainable DNNs.

More importantly, we proved and showcased multiple kinds of intrinsic explanations that can be retrieved by inspecting the proposed designs and their components. The extracted explanations reach similar or better scores than the extrinsic methods in most of the considered metrics and appear more flexible than them. We show how retrieved explanations can help developers to understand better the models they are developing in verifying the correct behavior, detecting and understanding errors and biased predictions, or detecting reliable and unreliable predictions.

However, all these designs require a modification of the architecture, introducing a bottleneck, and training from scratch of the model. Moreover, these designs introduce novel hyperparameters that must be tuned. Conversely, the next part of the thesis focuses on techniques that can be applied to trained networks and require small or no modifications at all.

## Part III

# Explaining Latent Representations

## Chapter 6

# Graph Concept Whitening

In the previous part, the outlined techniques aimed to integrate novel modules between the feature extractor and the classifier, with the primary goal of enhancing interpretability while preserving the performance of black-box models to the greatest extent possible. These modules strive to simplify the understanding of the final computation stages and are jointly trained with the black-box feature extractor. Conversely, techniques described in this part focus on pre-trained models and how to improve the interpretation of the learned knowledge (i.e., the latent space).

As a first step, this chapter presents Graph Concept Whitening (GCW) [190]. This normalization technique aims to ease the interpretation of the semantics encoded in the latent representations learned by the feature extractor, subsequently used by the classifier to generate predictions. GCW includes a few modifications of the feature extractor and a small alignment process, aiming as the previous part to preserve most of the black-box structure to not lower the performance. GCW is inspired and adapts the concept whitening technique [33] to Graph Neural Networks (GNNs) and the chemical domain. In particular, it encourages the components of the latent representations to represent molecular properties, providing a useful entry point for network inspection and interpretability. The goal of such alignment is to support users of this domain on the tasks of molecular property prediction and drug discovery.

The chapter is organized as follows: Section 6.1 describes GCW and how users can understand the learned behavior of the network; Section 6.2 evaluates GCW both in terms of performance and explanations quality; Section 6.3 discusses alternative design choices; finally, Section 6.4 summarizes the contributions of this chapter.

### 6.1 Design

The primary objective of **GCW** [190] is to facilitate the interpretation of the latent space of the GNN by normalizing the latent space and ensuring that each dimension in the node embeddings represents a molecular property. The inspection of these dimensions aids the users in understanding the semantics encoded within the latent space.

Let  $C = \{c_1, c_2, \dots, c_k\}$  denote a predefined set of concepts (i.e., in our case, molecular properties). For each concept  $c_j$ , we can define a subset of the training dataset  $X_{c_j}$  comprising only samples that exhibit that concept. Let  $\mathbf{h}_i^l$  represent the latent representation of the sample  $i$  returned by the layer  $l$ . The layer's goal is to align the  $j$ -th dimension of  $\mathbf{h}$  to the concept  $c_j$  and ensure that

samples in  $X_{c_j}$  exhibit larger values of  $h_j$  compared to other samples in the training dataset.

To achieve this goal, the layer applies a normalization that centers around the mean, decorrelates, and rotates the latent space to ensure that samples  $\in X_{c_j}$  related to the concept  $c_j$  are highly activated along the  $j$ -th axis.

To decorrelate and standardize the data, the layer applies the following function to a set of latent representation  $\mathbf{H} \in \mathbb{R}^{n \times d}$  of  $n$  samples:

$$g(\mathbf{H}) = W(\mathbf{H} - \mu\mathbf{1}) \quad (6.1)$$

where  $\mu$  is the sample mean:

$$\mu = \frac{1}{n} \sum_{i=1}^n \mathbf{H}_i \quad (6.2)$$

The matrix  $W$  of the previous equation is the whitening matrix, and it is defined such that:

$$W^T W = \frac{1}{n} (H - \mu\mathbf{1}^T)(H - \mu\mathbf{1}^T)^T \quad (6.3)$$

This matrix can be computed using several algorithms. Here, we follow Chen et al. [35] and use the zero-phase component analysis [93]:

$$\mathbf{H} = \mathbf{D}\Lambda^{-\frac{1}{2}}\mathbf{D}^T \quad (6.4)$$

where  $\Lambda \in \mathbb{R}^{d \times d}$  is the eigenvalue diagonal matrix and  $\mathbf{D} \in \mathbb{R}^{d \times d}$  is the eigenvector matrix. Finally, to rotate the latent representation, the layer multiplies the  $H$  matrix with an orthogonal matrix  $Q \in \mathbb{R}^{d \times d}$  where each column  $q_j$  represents the  $j^{th}$  axis.

$$\arg \max_{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_k} \sum_{j=1}^k \frac{1}{n_j} \mathbf{q}_j^T g(H_{c_j} \mathbf{1}) \quad (6.5)$$

subject to the constraint  $Q^T Q = I_D$ . Here,  $H_{c_j}$  is the matrix containing the latent representations of  $X_{c_j}$

### 6.1.1 Alignment Process

The alignment process is applied as a fine-tuning phase and optimizes both the performance and alignment between the presence of the concepts and the axes. The alignment is achieved by updating the rotation matrix  $Q$  via stochastic gradient descent. Each epoch of the alignment process of a model incorporating GCW consists of two phases: *accuracy optimization* and *concept alignment*. During accuracy optimization, the objective is to minimize the loss associated with classification accuracy (e.g., the cross-entropy loss):

$$L_y = \sum_{i=1}^n \text{CrossEntropy}(f(x_i), y_i) \quad (6.6)$$

In the second phase, the layer focuses on aligning the presence of concepts with the activation of axes. Specifically, it employs the following concept alignment loss:

$$\arg \max_{\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_k} \sum_{j=1}^k \frac{1}{n_j} \sum_{x_i^{c_j} \in X_{c_j}} \mathbf{q}_j^T g(H^l) \quad (6.7)$$

such that  $Q^T Q = I_d$ .

During the first phase,  $\mathbf{Q}$  remains fixed and all the other parameters are updated. Conversely, during the second phase, the alignment process optimizes  $\mathbf{Q}$  while freezing all other parameters. Specifically, the orthogonal matrix  $\mathbf{Q}$  is updated using the Cayley transform:

$$\mathbf{Q}^{(t+1)} = (I + \frac{\eta}{2} \mathbf{A})^{-1} \mathbf{Q}^{(t+1)} = (I + \frac{\eta}{2} \mathbf{A}) \mathbf{Q}^{(t)} \quad (6.8)$$

where  $\mathbf{A} = \mathbf{G}(\mathbf{Q}^{(t)})^T - \mathbf{Q}^{(t)} \mathbf{G}^T$  represents a skew-symmetric matrix,  $\mathbf{G}$  is the gradient of the loss function and  $\eta$  is the learning rate. Since  $\mathbf{G}$  requires a single scalar to compute the gradient, the concept activations must be projected into a scalar. GCW uses the max function, selecting the maximum value over the concept activation. However, various alternatives can be explored during the hyperparameters tuning (Section 6.3).

### 6.1.2 Explanations

GCW is designed to improve the interpretation of the latent space activations. Specifically, the inspection of the activations of a concept  $c_k$  can reveal the influence of that concept in the parsing of the input in the given layer. Conversely, the inspection of the full representation can help users understand the amount of semantics captured by the model until that point. Moreover, since GCW is a normalization layer, we can apply the gradient or post-hoc techniques to project the concept's importance onto the input space. This feature is particularly useful in the case of abstract concepts, where there is no fixed direct correspondence between the concept and a set of features. This is a common scenario in the case of graph data and the chemical domain. Moreover, following the trace of importance across the layers, it is possible to inspect circuits [176] of how the network combines elements during its decision process.

## 6.2 Experiments

### 6.2.1 Concept Selection

As described in the previous section, GCW requires specifying the concepts to which the latent space must align. Since we deal with molecular datasets, in this case, we select concepts expressing known relevant molecular properties for each of the studied tasks. A concept is considered present in the current input when the corresponding property value is above or below a given threshold.

The selected concepts are:

- Quantitative Estimate of Drug-likeness [20] (QED), which expresses the similarity to the properties of oral drugs;

- TPSA [178], which expresses the correlation with passive molecular transport through membranes;
- logarithm of the octanol-water Partition coefficient ( $\log P$ ), which estimates the molecules' hydro-lipophilicity balance [119];
- the number of nitrogen and oxygen atoms (NOCount);
- the number of heteroatoms (# Heteroatoms);
- molecular weight (mol\_weight), which is the sum of the atomic weight values of its atoms;
- number of H-Bond Acceptors (HBA);
- number of H-Bond Donors (HBD).

### 6.2.2 Performance

We start by comparing the black-box models and the same models fine-tuned by replacing Batch Normalization layers with GCW layers. As black-boxes, we evaluate Graph Convolutional Network (GCN), Graph Attention Network (GAT), and Graph Isomorphism Network (GIN). We test these models on BBBP, BACE, ClinTox, and HIV datasets. These datasets are popular chemical datasets for studying molecular properties and drug design. For BBBP, BACE, and HIV datasets, concepts are selected based on established literature [114, 205, 219, 230]. We use QED, TPSA, LogP, NoCount, and n\_heteroatoms for BBBP; QED, TPSA, mol\_weight, HBA, and HBD for BACE; and QED, nDoubleBonds, nO, HBA, HBD, and LogP for HIV. For ClinTox, we choose the best-performing concepts among those used for BBBP and BACE: QED, mol\_weight, HBA, and n\_heteroatoms (Table 6.1). Threshold values are determined based on Lipinski's rule of five [131] for molecular weight, logP, HBA, and HBD, while other thresholds are set as the mean of the property across each dataset. All the models include 128 units per layer. In GAT, we set the number of attention heads to 2. Models are trained on BBBP and HIV using a batch size of 64, a learning rate of  $1e - 3$ , and a weight decay of  $5e - 4$ . Models are trained on BACE and ClinTox using a batch size equal to 128, a learning rate equal to  $1e - 2$ , and without weight decay. The fine-tuning lasts a

**Table 6.1:** Avg. Accuracy and ROC-AUC over 15 runs on ClinTox when adding one concept at a time for the GCW layers. We highlight the best results of combined accuracy and ROC-AUC.

Architecture	Accuracy	ROC-AUC
Baseline	$0.94 \pm 0.01$	$0.86 \pm 0.08$
GCW - QED	<b><math>0.95 \pm 0.02</math></b>	<b><math>0.95 \pm 0.04</math></b>
GCW - TPSA	$0.94 \pm 0.02$	$0.93 \pm 0.04$
GCW - mol_weight	<b><math>0.95 \pm 0.02</math></b>	<b><math>0.93 \pm 0.08</math></b>
GCW - HBA	<b><math>0.95 \pm 0.01</math></b>	<b><math>0.95 \pm 0.02</math></b>
GCW - HBD	$0.94 \pm 0.01$	$0.93 \pm 0.08$
GCW - logP	<b><math>0.95 \pm 0.01</math></b>	<b><math>0.93 \pm 0.04</math></b>
GCW - NOCount	$0.95 \pm 0.02$	$0.91 \pm 0.09$
GCW - # Heteroatoms	<b><math>0.95 \pm 0.02</math></b>	<b><math>0.95 \pm 0.03</math></b>

maximum of 50 epochs with an early stopping with a patience of five epochs. All parameters have been fixed after a grid search over the hyperparameter space. Note that the fine-tuning phase is longer than the commonly used in the vision domain [35]. Black-box models are pre-trained using an early stopping with a patience of 20 epochs. We run the models from 15 different initializations and compute the average ROC-AUC and its standard deviation.

Table 6.2 reports the ROC-AUC of the black-box models, the same models fine-tuned by using GCW, and three commonly used baselines in these datasets: random forest, Multi-Layer Perceptron (MLP), and message-passing neural networks (MPNN) [65]. The results demonstrate the performance benefits of fine-tuning models using GCW. Specifically, **models fine-tuned with GCW outperform the black-box variants** in all the configurations and the baselines in all but in the HIV dataset. In this dataset, MLP reaches the best performance against all the competitors. The improvement of the GCW models can be attributed to the ability of the GCW layers to encourage node embeddings to represent class-specific information, aiding discrimination among classes. These results align with the ones reported by Chen et al. [35]. The only difference is in the number of epochs needed to reach comparable results. While they fine-tune the models for just one epoch, GNNs require up to 50 epochs to align the representations and reach comparable or superior performance.

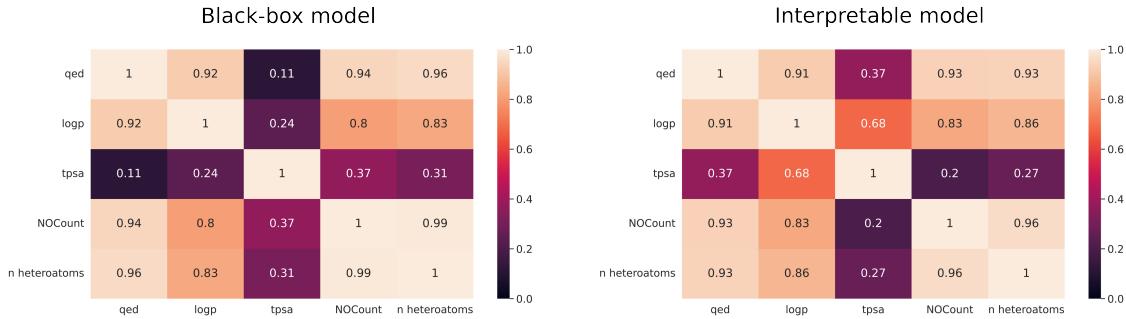
### 6.2.3 Explanations

In this section, we investigate and showcase the interpretability improvements of models employing GCW layers. We use GAT trained on BBBP as the example model. However, similar results can be obtained for other models.

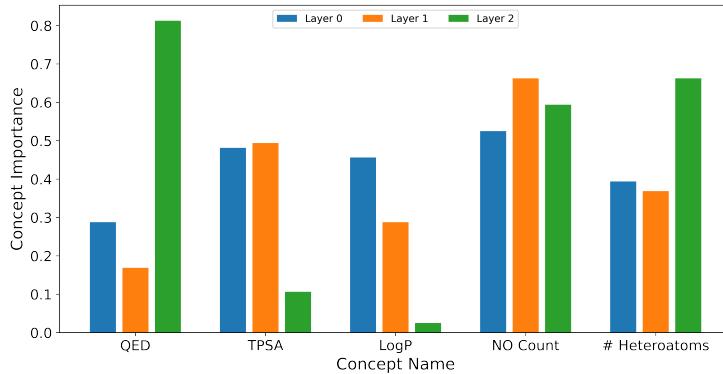
We begin the investigation by comparing the separability (Equation (2.27)) in the latent space exposed by the black-box models and the models incorporating GCW. Figure 6.1 shows that **models with GCW layers achieve superior separability**. In particular, note that molecular properties that are not similar, like NoCount and TPSA, exhibit better separation in in the interpretable model than in the black-box model (score 0.2 vs 0.37). Conversely, chemically related properties (e.g., TPSA and LogP) are similar in the representation (score 0.68). These results confirm a better disentanglement induced by GCW useful for improving the interpretability of the axes.

**Table 6.2:** Avg. ROC-AUC and standard deviation over 15 runs before and after adding GCW to the models.

Model	BBBP	BACE	ClinTox	HIV
GCN	$0.88 \pm 0.03$	$0.86 \pm 0.02$	$0.86 \pm 0.08$	<b><math>0.79 \pm 0.03</math></b>
+ GCW	<b><math>0.91 \pm 0.02</math></b>	<b><math>0.92 \pm 0.02</math></b>	<b><math>0.93 \pm 0.05</math></b>	<b><math>0.79 \pm 0.02</math></b>
GAT	$0.89 \pm 0.03$	$0.86 \pm 0.02$	$0.84 \pm 0.08$	<b><math>0.80 \pm 0.03</math></b>
+ GCW	<b><math>0.97 \pm 0.02</math></b>	<b><math>0.93 \pm 0.01</math></b>	<b><math>0.88 \pm 0.05</math></b>	<b><math>0.80 \pm 0.02</math></b>
GIN	<b><math>0.89 \pm 0.02</math></b>	$0.87 \pm 0.03$	<b><math>0.75 \pm 0.11</math></b>	<b><math>0.78 \pm 0.03</math></b>
+ GCW	<b><math>0.91 \pm 0.02</math></b>	<b><math>0.91 \pm 0.02</math></b>	<b><math>0.74 \pm 0.11</math></b>	<b><math>0.79 \pm 0.02</math></b>
Random forest	$0.79 \pm 0.01$	$0.79 \pm 0.01$	$0.74 \pm 0.01$	$0.65 \pm 0.01$
MLP	$0.87 \pm 0.01$	$0.84 \pm 0.01$	$0.75 \pm 0.05$	$0.84 \pm 0.01$
MPNN	$0.88 \pm 0.02$	$0.70 \pm 0.03$	$0.72 \pm 0.02$	$0.75 \pm 0.01$



**Figure 6.1:** Separability score of the latent space of a black-box model and its interpretable version. Lower is better.



**Figure 6.2:** Avg. importance scores assigned to five concepts in the layers of a GAT model trained on BBBP.

To quantify the improvement in interpretability, we examine the impact of the interpretable design on other types of explanations, such as feature attribution toward the predicted class. To measure the impact, we report in Table 6.3 the *fidelity+<sup>prob</sup>* of the feature attribution computed by GNNExplainer [273] before and after the fine-tuning phase. The results demonstrate that incorporating GCW layers does not compromise the interpretability scores compared to when applied on black-boxes. While performance is comparable, the mean fidelity is usually higher and the standard deviation lower, suggesting **more stable and robust explanations**.

Once the interpretability improvements have been validated, we showcase how the concept-aligned latent space can be used to study molecular properties and help users in **drug discovery**. The first example explores the flow of learned knowledge across the network. In particular, we compute the concepts’ importance across layers by using the positive directional derivative of the gradient cumulated on the concept axis. As shown in Figure 6.2, some properties are more important

**Table 6.3:** Avg. and standard deviation of the Fidelity+ score of feature attributions computed on both black-boxes and models employing GCW layers.

Model	BBBP	BACE	ClinTox	HIV
Black-box	$0.24 \pm 0.41$	$0.45 \pm 0.45$	$-0.12 \pm 0.25$	$0.00 \pm 0.01$
+GCW	$0.31 \pm 0.38$	$0.45 \pm 0.41$	$-0.06 \pm 0.12$	$0.02 \pm 0.19$



**Figure 6.3:** Feature attribution for the concept `#Heteroatoms` in the BBBP dataset.

in lower layers than higher ones (e.g., TPSA and logP), while others exhibit the opposite trend (e.g., QED and `#Heteroatoms`). This phenomenon resembles the behavior observed for CNN on the vision domain, where lower layers detect shapes and colors, while higher layers detect the composition of concepts. Moreover, lower layers exhibit an almost uniform importance across different concepts; thus, the network uses information from all the concepts at this stage. Conversely, higher layers exhibit significant differentiation, and the network uses only a subset of concepts. We also highlight that the most important concepts in the last hidden layer, which are then used by the classifier to perform the prediction, are QED, `#Heteroatoms`, and NOCount. These findings align with the literature on drug discovery and the ones of Sakiyama et al. [205].

The second example shows how it is possible to use the concept axes to **discover atom relationships that can be exploited for the optimization of the molecules’ activities** to design novel drugs. As anticipated in Section 6.1.2, it is possible to project concept activation onto input features and understand the influence of features on input parsing. Figure 6.3 showcase an application of this procedure to the `#Heteroatoms` concept. The figure displays three samples that mostly activate the concept’s axis. Notably, oxygen and nitrogen atoms receive the largest attribution. These results confirm the quality of the learned alignment as oxygen and nitrogen are heteroatoms and decisive in determining the molecule’s class.

### 6.3 Design Choices

This section delves into alternative choices for GCW, specifically focusing on the activation mode and the layers to replace.

**Activation Modes.** In the previous sections, we assumed the maximum value over the node embeddings as the activation score. Here, we explore alternative activation modes: *mean*, *pos\_mean*, *topk\_pool*, and *weighted\_topk\_pool*. The *mean* mode computes the mean over the node embedding. The *pos\_mean* computes the mean over only the positive value of the node embedding. *Topk\_pool* computes the mean over the subgraph node embeddings extracted using the recently proposed *topk\_pool* operator [62]. Finally, the *weighted\_topk\_pool* activation pool computes the mean as in *topk\_pool* but weighs the contribution of each component by the projection score returned by the *topk\_pool* operator [62].

From Table 6.4, we can observe that the max activation mode is the best across several datasets and models. More in detail, in the BBBP and HIV datasets, this activation mode outperforms all the others by large margins, except for the *pos\_mean* mode where the gap is lower. Conversely, in BACE and ClinTox all the activation modes achieve comparable performance, with *weighted\_topk\_pool* even reaching the maximum possible mean in one case (i.e., in the ClinTox dataset).

**Normalization Layers.** Here, we test the efficacy of GCW when used to replace different types of normalization layers. We train from scratch the best-performing architectures of the previous experiments, replacing their Batch Normalization layers with Layer Normalization, Instance Normalization, and Graph Normalization layers. Then, we fine-tune the trained model, replacing the normalization layers with GCW, and collect the ROC-AUC before and after the fine-tuning. Table 6.5 suggests that GCW can be used as a replacement for all the types of normalization layers since, most of the time, it reaches comparable or better performance than the original normalization. Note the significant gap when GCW replaces the Instance Normalization and Graph Normalization layers. We hypothesize that the gap is due to a mismatch between these normalizations and the backbone architectures. In these cases, GCW recovers the performance since it is a variant of batch

**Table 6.4:** Avg. ROC-AUC and standard deviation reached by models including GCW using different activation modes.

Activation Mode	BBBP	BACE	ClinTox	HIV
GCN				
– max	<b>0.91 ± 0.02</b>	<b>0.92 ± 0.02</b>	<b>0.93 ± 0.05</b>	<b>0.79 ± 0.02</b>
– mean	0.88 ± 0.03	<b>0.92 ± 0.03</b>	<b>0.95 ± 0.03</b>	0.72 ± 0.02
– pos_mean	<b>0.89 ± 0.03</b>	<b>0.92 ± 0.02</b>	<b>0.95 ± 0.02</b>	<b>0.77 ± 0.02</b>
– topk_pool	0.87 ± 0.03	<b>0.92 ± 0.01</b>	0.92 ± 0.08	0.70 ± 0.02
– weighted_topk_pool	0.81 ± 0.06	<b>0.93 ± 0.02</b>	0.92 ± 0.10	0.69 ± 0.03
GAT				
– max	<b>0.97 ± 0.02</b>	<b>0.93 ± 0.01</b>	<b>0.88 ± 0.05</b>	<b>0.80 ± 0.02</b>
– mean	0.92 ± 0.05	<b>0.93 ± 0.02</b>	<b>0.90 ± 0.03</b>	0.75 ± 0.02
– pos_mean	0.91 ± 0.02	0.91 ± 0.02	<b>0.88 ± 0.04</b>	<b>0.80 ± 0.02</b>
– topk_pool	0.92 ± 0.03	<b>0.94 ± 0.02</b>	<b>0.90 ± 0.04</b>	0.76 ± 0.02
– weighted_topk_pool	0.89 ± 0.03	<b>0.93 ± 0.02</b>	<b>0.90 ± 0.04</b>	0.74 ± 0.04
GIN				
– max	<b>0.91 ± 0.02</b>	<b>0.91 ± 0.02</b>	<b>0.74 ± 0.11</b>	<b>0.79 ± 0.02</b>
– mean	0.88 ± 0.03	<b>0.92 ± 0.02</b>	<b>0.70 ± 0.15</b>	0.71 ± 0.02
– pos_mean	<b>0.90 ± 0.02</b>	<b>0.92 ± 0.02</b>	<b>0.69 ± 0.12</b>	<b>0.79 ± 0.02</b>
– topk_pool	0.84 ± 0.04	0.82 ± 0.02	<b>0.73 ± 0.13</b>	0.70 ± 0.03
– weighted_topk_pool	0.85 ± 0.03	<b>0.92 ± 0.02</b>	<b>0.77 ± 0.08</b>	0.71 ± 0.03

**Table 6.5:** Avg. ROC-AUC and standard deviation of models when GCW replaces several types of normalization layers.

Normalization type	BBBP	BACE	ClinTox	HIV
BatchNorm	0.89 ± 0.03	0.86 ± 0.02	0.86 ± 0.08	<b>0.80 ± 0.03</b>
– GCW	<b>0.97 ± 0.02</b>	<b>0.93 ± 0.01</b>	<b>0.93 ± 0.05</b>	<b>0.80 ± 0.02</b>
LayerNorm	<b>0.88 ± 0.03</b>	<b>0.88 ± 0.03</b>	<b>0.83 ± 0.03</b>	0.71 ± 0.04
– GCW	<b>0.88 ± 0.03</b>	<b>0.86 ± 0.03</b>	0.71 ± 0.09	<b>0.81 ± 0.02</b>
InstanceNorm	0.60 ± 0.12	0.69 ± 0.07	0.58 ± 0.06	0.61 ± 0.05
– GCW	<b>0.92 ± 0.03</b>	<b>0.94 ± 0.02</b>	<b>0.92 ± 0.05</b>	<b>0.81 ± 0.02</b>
GraphNorm	0.59 ± 0.20	0.76 ± 0.05	0.56 ± 0.09	0.60 ± 0.05
– GCW	<b>0.89 ± 0.03</b>	<b>0.92 ± 0.02</b>	<b>0.82 ± 0.10</b>	<b>0.80 ± 0.02</b>

normalization layers, and thus, it is more similar to the one used in the design of the backbone models.

## 6.4 Contributions

This chapter contributed to the research on making the interpretation of latent space easier by investigating the application of concept whitening to graph data and in the chemical domain. The chapter analyzed the differences and specific characteristics of the graph applications, along with the advantages and weaknesses associated with their utilization for enhancing the interpretability of GNNs.

Notably, the approach preserves the performance of the black-box models while improving their interpretability in terms of the separability of the latent space and explanation fidelity when used in combination with post-hoc methods. With GCW, we take a step forward with respect to the approaches described in the previous chapters. Instead of adding elements to the network, this approach modifies only the normalization layers to improve the interpretability of the network. We discussed how this modification can be performed in the fine-tuning stage without loss of accuracy and how to inspect the decision process of the network via the inspection of the latent space. In particular, we showed how a disentangled latent space can help chemists in molecular activity prediction and drug discovery tasks.

Alongside these strengths, GCW presents some drawbacks. For example, the selection of concepts and the dependence on the prior design of the architecture (i.e., it performs better when replacing BatchNormalization layers) represent limitations to the applicability of the approach compared to all the previous approaches, which are, instead, model-agnostic. Moreover, it still requires a new training phase and a modification of the architecture, even if it is small. These limitations will be addressed in the next chapter, where we will present a post-training algorithm applicable to any DNN.

## Chapter 7

# Clustered Compositional Explanations

This chapter shares the same motivation as the previous chapter, elucidating the semantics encoded in the learned latent representations. While Graph Concept Whitening proposes modifying the architecture to enforce the semantics, the technique described in this chapter aims to extract the semantics encoded in the components of the latent representation of any pre-trained black-box model without any modification.

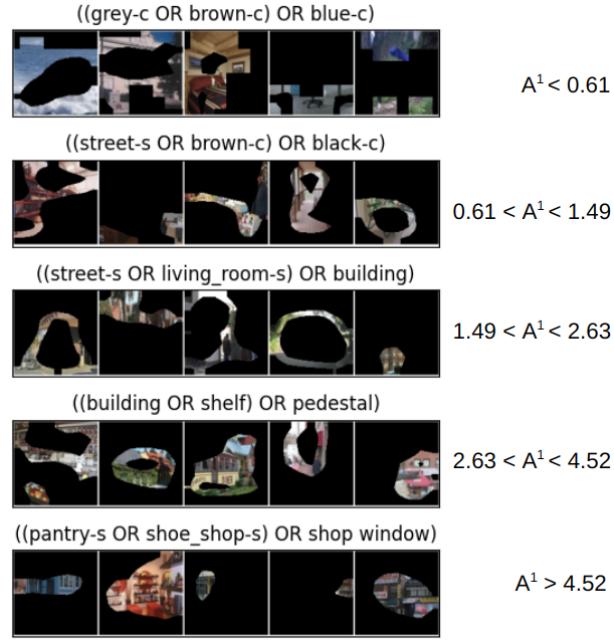
This chapter focuses on the family of methods that investigate what kind of information is recognized by neurons by probing them using a concept dataset [17, 150, 163] (Section 3.1). The idea is to probe each neuron and select, as an explanation, the label whose annotations are aligned the most with its activations. The current literature [17, 84, 150, 163] focuses only on the exceptionally high activations due to the computational challenges associated with considering wider ranges. Conversely, this chapter describes an **algorithm that combines heuristics and clustering to extract the composition of concepts recognized by neurons at several intervals of activations**, mitigating the computational challenges connected to their exploration and the superposition issue (Section 2.2). By using this algorithm, it is possible to extract novel findings about the phenomena related to models' activations, helping us to understand these black-boxes better.

The chapter is organized as follows: Section 7.1 describes the proposed algorithm and its heuristic; Section 7.2 introduces and summarizes metrics to use to evaluate the explanations' quality; Section 7.3 evaluates the algorithm in terms of efficiency and explanations quality and sheds light on phenomena related to neurons activations; Section 7.4 discusses design choices for the proposed algorithm; finally, Section 4.4 summarized the contributions of this chapter.

### 7.1 Algorithm Design

The proposed algorithm, named **Clustered Compositional Explanations algorithm**, aims to describe a wider spectrum of neuron activations than current methods. To achieve the goal, the algorithm combines clustering algorithms with a beam search guided by heuristics. The idea is to cluster the activations of a neuron, compute explanations for each cluster, and provide a summary of the explanations to the user (Figure 7.1).

Specifically, let  $\mathfrak{D}$  be a dataset where each sample can include concept annotations. Each annotation is represented by a function  $S(x, L)$  that returns a binary mask for a sample  $x$  where



**Figure 7.1:** Explanations in terms of recognized concepts and sample examples of different activation ranges for unit #1.

only the features associated with a concept labeled as  $L$  are set to one. Given the matrix  $A^k(x)$ , which represent the activation map of the  $k$  neuron when the input is the sample  $x$ ; the algorithm computes a set of activation thresholds:

$$T = \{[min(Cls), max(Cls)], \forall Cls \in Clustering(A^k(\mathfrak{D}))\} \quad (7.1)$$

where  $Clustering(A^k(\mathfrak{D}))$  returns a set of  $n_{cls}$  clusters including non-zero activations of the  $k$  neuron. For simplicity, we assume the clusters are disjoint so that two values are enough to identify a cluster precisely. Therefore, each element in  $T$  is a pair of threshold  $[\tau_i, \tau_j]$  where  $\tau_i$  is the minimum value of activations included in the given cluster and  $\tau_j$  is the maximum value. However, the approach can also be applied to overlapping clusters as long as they can be identified by a set of thresholds.

For each pair of thresholds (or equivalently for each cluster), the algorithms compute the matrix  $M_{[\tau_i, \tau_j]}^k(\mathfrak{D})$ , which is a binary mask of the activations  $A^k(\mathfrak{D})$  obtained by setting to 0 all the values lower than  $\tau_i$  and greater than  $\tau_j$ .

At this point, the algorithm assigns a label to each cluster by extracting the one whose annotations overlap the most with the matrix  $M_{[\tau_i, \tau_j]}^k(\mathfrak{D})$

$$L^{best} = \{\arg \max_{L \in \mathcal{L}^n} IoU(L, \tau_i, \tau_j, \mathfrak{D}), \forall [\tau_i, \tau_j] \in T\} \quad (7.2)$$

where  $\mathcal{L}^l$  is the set of logical connections arity  $l$  that can be built between concepts  $L$  from the concept dataset  $\mathfrak{D}$ ;

In this equation,  $IoU(L, \tau_i, \tau_j, \mathfrak{D})$  refer to Equation (2.28), adapted to the context so that:

$$IoU(L, \tau_1, \tau_2, \mathfrak{D}) = \frac{\sum_{x \in \mathfrak{D}} |M_{[\tau_1, \tau_2]}(x) \cap S(x, L)|}{\sum_{x \in \mathfrak{D}} |M_{[\tau_1, \tau_2]}(x) \cup S(x, L)|} \quad (7.3)$$

For each cluster, the label can be extracted by using the Compositional Explanations (CoEx) algorithm [163].

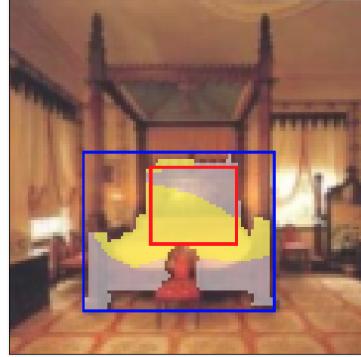
This algorithm assigns logical formulas of arity  $n$  to a range of activations by performing a beam search of size  $b$  over the most promising logical connections of atomic concepts. At each step  $i$ , only the  $b$  best labels of arity  $i$  are used as bases for computing labels of arity  $i + 1$ . The first beam is selected among the labels associated with the best scores computed by the Network Dissection (NetDissect) algorithm [17]. NetDissect optimizes Equation (7.2) by performing an exhaustive search over the space of all the atomic concepts in the dataset.

However, by design, the vanilla beam search implemented by the CoEx algorithm would require  $n_{cls} \times (n - 1) \times b$  times the computation time of NetDissect to compute the labels for all the clusters in a single neuron, where  $b$  and  $(n - 1)$  are the width and deepness of beam search tree, respectively. This computational time is practically unfeasible for most of the current deep neural networks, thus impeding the analysis of a wider spectrum of activations.

To mitigate this issue, we propose the *Min-Max Extension per Sample Heuristic* (**MMESH**) to guide the beam search and cut the search space. The purpose of the heuristic is to estimate the IoU score and sort the search space based on the estimation. If the heuristic is admissible, and thus, it never underestimates the true IoU, then it is possible to cut the search space at each step of the beam search by removing all labels whose estimated IoU score is below the best score found so far. Obviously, the time employed to estimate the scores and to sort the search space must be lower than the time needed to explore the full search space. In the considered scenario, the bottleneck is represented by the repeated multiplication between large matrices (as large as the dataset), and a good heuristic should avoid performing them as much as possible.

To achieve this goal, **MMESH exploits the coordinates that identify the smallest and the largest possible extension of the concepts in each dataset sample**. Given a sample  $x$ , *the smallest possible extension corresponds to the longest contiguous segment that includes only concept features*. For example, in the case of 1D data, it corresponds to the shortest subsequence that includes only concept features, and the coordinates are the sequence's first and last index. In the image data, the smallest possible extension corresponds to the largest rectangle inscribed inside the polygon defined by the segmentation mask associated with the concept, and the coordinates are the ones of the top left and bottom right corners of the rectangle. Conversely, *the largest possible extension corresponds to the shortest contiguous segment that includes all the features associated with the concept*. In this case, for 1D data, it corresponds to the subsequence that starts from the first occurrence of the concept's features and ends in the last occurrence. For image data, the largest possible extension corresponds to the bounding box covering the entire concept's mask, and the coordinates are the ones of the box's top left and bottom right corners.

Given a sample  $x$ , a neuron  $k$ , and a label  $L \in \mathcal{L}^i$ , MMESH estimates the IoU score by combining the two quantities described above, the size of the label mask on the sample  $S(x, L)$ , and the size of the intersection between the label's terms mask and the neuron's activation on the sample  $IMS(x, t) \forall t \in L$ . The last quantity can be computed during the execution of the previous beam.



**Figure 7.2:** Visualization of information used by MMESH: the size of the concept mask (white), size of the intersection (yellow), minimum extension (red bounding box), maximum extension (blue bounding box).

Figure 7.2 visualizes the information used by the heuristic.

MMESH supports the estimation of logical formulas connected by OR, AND, and AND NOT operators, which are the most used operators [76, 145, 150, 163] for compositional explanations. To estimate the IoU score, MMESH must estimate both the intersection  $\hat{I}$  and the union  $\hat{U}$ . Expanding  $\hat{U}$ , we can rewrite Equation (7.3) as:

$$\begin{aligned} IoU(L, \tau_1, \tau_2, \mathfrak{D}) &= \frac{\hat{I}}{\hat{U}} = \frac{\sum_{x \in \mathfrak{D}} \hat{I}_x}{\sum_{x \in \mathfrak{D}} \hat{U}_x} = \\ &= \frac{\hat{I}_x}{\sum_{x \in \mathfrak{D}} |M_{[\tau_1, \tau_2]}(x)| + \sum_{x \in \mathfrak{D}} |\widehat{S(x, L)}| - \hat{I}_x} \end{aligned} \quad (7.4)$$

Since  $|M_{[\tau_1, \tau_2]}(x)|$  is unique, it is shared among all the labels. Therefore, it can be computed one time per cluster, and it does not need to be estimated. As previously mentioned, the heuristic needs to be admissible in order to be used effectively. Therefore, the estimation of the numerator  $\hat{I}$  must be an overestimation, and the estimation of the denominator  $\hat{I}$  must be an underestimation. Hence, the heuristic assumes the best-case scenario for each logical operator for the estimation of the numerator since it must be an overestimation:

$$\hat{I}_x = \min(|IMS_{[\tau_1, \tau_2]}(x, L_{\leftarrow})| + |IMS_{[\tau_1, \tau_2]}(x, L_{\rightarrow})|, |M_{[\tau_1, \tau_2]}(x)|) \quad op = OR \quad (7.5)$$

$$\hat{I}_x = \min(|IMS_{[\tau_1, \tau_2]}(x, L_{\leftarrow})|, |IMS_{[\tau_1, \tau_2]}(x, L_{\rightarrow})|) \quad op = AND \quad (7.6)$$

$$\hat{I}_x = \min(|IMS_{[\tau_1, \tau_2]}(x, L_{\leftarrow})|, |M_{[\tau_1, \tau_2]}(x)| - |IMS_{[\tau_1, \tau_2]}(x, L_{\rightarrow})|) \quad op = AND\ NOT \quad (7.7)$$

Where  $op$  is the logical connector of the formula,  $L_{\leftarrow} \in \mathfrak{L}^{l-1}$  and  $L_{\rightarrow} \in \mathfrak{L}^1$  denote the left side and the right side of a label of arity  $i$  obtained by adding an atomic term to the label at each step, respectively. Equation (7.5) corresponds to assuming disjoint labels (i.e., masks). Conversely, Equation (7.6) and Equation (7.7) assume fully overlapping masks.

Conversely, in the case of the denominator, since the estimation must be an underestimation, the heuristic assumes the worst-case scenario for each operator:

$$\widehat{S(x, L)} = \begin{cases} \max(|S(x, L_{\leftarrow})|, |S(x, L_{\rightarrow})|, S(x, \widehat{L_{\leftarrow} \cup L_{\rightarrow}}), \widehat{I}_x), & op = OR \\ \max(\text{MinOver}(L), \widehat{I}_x) & op = AND \\ \max(|S(x, L_{\leftarrow})| - \text{MaxOver}(L), \widehat{I}_x) & op = AND NOT \end{cases} \quad (7.8)$$

$$(7.9)$$

$$(7.10)$$

where  $\text{MaxOver}(L)$  returns the maximum possible overlap between the largest possible extensions of the formula's terms,  $\text{MinOver}(L)$  returns the minimum possible overlap between the smallest possible extensions of the formula's terms, and  $S(x, \widehat{L_{\leftarrow} \cup L_{\rightarrow}}) = |S(x, L_{\leftarrow})| + |S(x, L_{\rightarrow})| - \text{MaxOver}(L)$ . In this case, Equation (7.8) corresponds to assuming fully overlapping masks between the labels connected by the operator. Conversely, Equation (7.9) and Equation (7.10) assume disjoint masks.

### 7.1.1 Proof of Admissibility of MMESH

This section proves that **MMESH is an admissible heuristic** (i.e., it never underestimates the true IoU score). As previously explained, in order to be admissible,  $\widehat{I}_x$  and  $\widehat{S(x, L)}$  must satisfy the following equations:

$$\left\{ \begin{array}{l} |\widehat{I}_x| \geq |I_x| \\ 0 \leq \widehat{S(x, L)} - |\widehat{I}_x| \leq |S(x, L)| - I \end{array} \right. \quad (7.11)$$

$$(7.12)$$

$\forall x \in \mathfrak{D}$ .

We prove the two equations by proving they hold in the two extreme cases: fully overlapping between the labels connected by the operators and disjoint masks. As a consequence, they will also hold for partially overlapping masks.

#### Proof of Equation (7.11)

**Disjoint Masks.** In this case, the real intersection is given by the following quantities:

$$I_x = \begin{cases} |IMS_{[\tau_1, \tau_2]}(x, L_{\leftarrow})| + |IMS_{[\tau_1, \tau_2]}(x, L_{\rightarrow})| & op = OR \\ 0 & op = AND \end{cases} \quad (7.13)$$

$$(7.14)$$

$$(7.15)$$

We can prove Equation (7.11) by observing that:

- eq. (7.13)  $\leq$  eq. (7.5) because the intersections  $IMS_{[\tau_1, \tau_2]}(x, \cdot)$  are disjoint subsets of the matrix  $M_{[\tau_1, \tau_2]}(x)$  and, thus, eq. (7.13) = eq. (7.5);
- eq. (7.14)  $\leq$  eq. (7.6) due to the non-negativity property of the cardinality;
- eq. (7.15)  $\leq$  eq. (7.7) holds because the presence of the minimum operator and the fact that the mask  $M_{[\tau_1, \tau_2]}(x) - IMS_{[\tau_1, \tau_2]}(x, L_{\rightarrow})$  contains  $IMS_{[\tau_1, \tau_2]}(x, L_{\leftarrow})$  since  $L_{\leftarrow}$  and  $L_{\rightarrow}$  are associated with disjoint masks.

**Fully Overlapping Masks.** In this case, the real intersection is given by the following quantities:

$$I_x = \begin{cases} \max(|IMS_{[\tau_1, \tau_2]}(x, L_{\leftarrow})|, |IMS_{[\tau_1, \tau_2]}(x, L_{\rightarrow})|) & op = OR \\ \min(|IMS_{[\tau_1, \tau_2]}(x, L_{\leftarrow})|, |IMS_{[\tau_1, \tau_2]}(x, L_{\rightarrow})|) & op = AND \end{cases} \quad (7.16)$$

$$I_x = \min(|IMS_{[\tau_1, \tau_2]}(x, L_{\leftarrow})|, |M_{[\tau_1, \tau_2]}(x)| - |IMS_{[\tau_1, \tau_2]}(x, L_{\rightarrow})|) \quad op = AND NOT \quad (7.18)$$

### Proof of Equation (7.12)

To prove Equation (7.12), it is sufficient to prove that the following equation holds for both disjoint and overlapping masks:

$$\widehat{S(x, L)} \leq S(x, L) \quad (7.19)$$

Indeed, once the previous equation is proved, then also the constraint  $\widehat{S(x, L)} - \widehat{I}_x \leq S(x, L) - I_x$  holds because  $\widehat{I}_x \geq I_x$  as shown before. Moreover,  $\widehat{S(x, L)} - \widehat{I}_x \geq 0$  is already satisfied by  $\widehat{S(x, L)}$  because all the cases include the max operator and  $\widehat{I}_x$  as a term.

We prove the equation by proving it holds in the two extreme cases: fully overlapping between the labels connected by the operators and disjoint masks. As a consequence, it will also hold for partially overlapping masks.

**Disjoint Masks.** In this case, the real joint label's mask is given by:

$$S(x, L) = \begin{cases} |S(x, L_{\leftarrow})| + |S(x, L_{\rightarrow})| & op = OR \end{cases} \quad (7.20)$$

$$S(x, L) = 0 \quad op = AND \quad (7.21)$$

$$S(x, L) = |S(x, L_{\leftarrow})| \quad op = AND NOT \quad (7.22)$$

We can prove Equation (7.19) by observing that:

- eq. (7.8)  $\leq$  eq. (7.20) because the maximum between two cardinalities is lower than their sum and  $MaxOver(L) \geq 0$ ;
- eq. (7.9)  $\leq$  eq. (7.21) because the masks are disjoint and thus  $MinOver(L) = 0$  and  $\widehat{S(x, L)} = S(x, L)$ ;
- eq. (7.10)  $\leq$  eq. (7.22) because  $|\widehat{I}_x| \leq |S(x, L_{\leftarrow})|$  since the intersection is a subset of the whole set, and  $MaxOver(L)$  returns a positive number, implying that  $|S(x, L_{\leftarrow})| - MaxOver(L) \leq |S(x, L_{\leftarrow})|$ .

**Fully Overlapping Masks.** In this case, the real joint label's mask is given by:

$$S(x, L) = \begin{cases} \max(|S(x, L_{\leftarrow})|, |S(x, L_{\rightarrow})|) & op = OR \end{cases} \quad (7.23)$$

$$S(x, L) = \min(|S(x, L_{\leftarrow})|, |S(x, L_{\rightarrow})|) \quad op = AND \quad (7.24)$$

$$S(x, L) = \max(|S(x, L_{\leftarrow})| - |S(x, L_{\rightarrow})|, 0) \quad op = AND NOT \quad (7.25)$$

We can prove Equation (7.19) by observing that:

- eq. (7.8)  $\leq$  eq. (7.23) because  $S(x, \widehat{L_{\leftarrow} \cup L_{\rightarrow}})$  cannot be larger than individual masks due to the overlap;

- eq. (7.9)  $\leq$  eq. (7.24) because, in the case of fully overlapping masks,  $MinOver(L)$  returns a subset of the smallest label's mask and, thus,  $MinOver(L) \leq \min(|S(x, L_{\leftarrow})|, |S(x, L_{\rightarrow})|)$ ;
- eq. (7.10)  $\leq$  eq. (7.25) due to the non-negativity of the cardinality.

Therefore, since the heuristic satisfies Equation (7.11) and Equation (7.12), the heuristic is admissible and returns the optimal formula inside the beam.

## 7.2 Metrics

This section introduces our **set of metrics to evaluate the qualities of logical explanations associated with neurons**. While the IoU metric has been only reframed to be compatible with threshold values, all the other metrics are proposed by this thesis.

**Intersection Over Union** measures the alignment between labels' annotations and activation map in the given activation range  $(\tau_1, \tau_2)$ . A higher IoU means the algorithm can better capture the pre-existent alignment [163].

$$IoU(L, \tau_1, \tau_2, \mathfrak{D}) = \frac{\sum_{x \in \mathfrak{D}} |M_{[\tau_1, \tau_2]}(x) \cap S(x, L)|}{\sum_{x \in \mathfrak{D}} |M_{[\tau_1, \tau_2]}(x) \cup S(x, L)|} \quad (7.26)$$

**Detection Accuracy** measures the bijection from the overlapping set between label annotations and neuron activation inside the activation range to the set of label annotations [145]. In other words, it measures the quantity of annotations recognized at the given activation range. Higher is better.

$$DetAcc(L, \tau_1, \tau_2, \mathfrak{D}) = \frac{\sum_{x \in \mathfrak{D}} |M_{[\tau_1, \tau_2]}(x) \cap S(x, L)|}{\sum_{x \in \mathfrak{D}} |S(x, L)|} \quad (7.27)$$

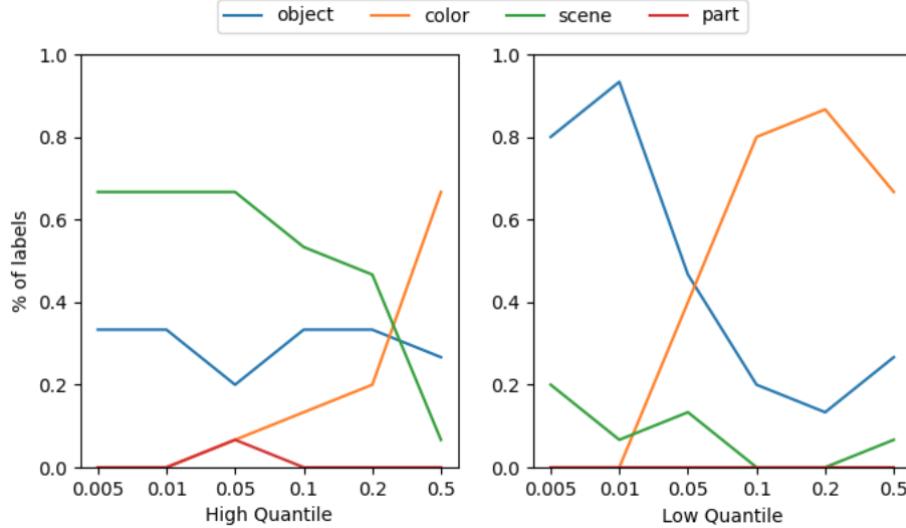
**Samples Coverage** measures the bijection from the set of samples where there is an overlap between label annotations and neuron activation inside the activation range to the set of samples associated with label annotations. In other words, it is the percentage of candidate samples recognized at the given activation range. Higher is better.

$$SampleCov(L, \tau_1, \tau_2, \mathfrak{D}) = \frac{|\{x \in \mathfrak{D} : |M_{[\tau_1, \tau_2]}(x) \cap S(x, L)| > 0\}|}{|\{x \in \mathfrak{D} : |S(x, L)| > 0\}|} \quad (7.28)$$

**Activation Coverage** measures the bijection from the overlapping set between label annotations and neuron activation inside the activation range to the set of neuron activations within the range. In other words, measures the label  $L$  exclusivity for the activation range  $(\tau_1, \tau_2)$ . Higher is better.

$$ActCov(L, \tau_1, \tau_2, \mathfrak{D}) = \frac{\sum_{x \in \mathfrak{D}} |M_{[\tau_1, \tau_2]}(x) \cap S(x, L)|}{\sum_{x \in \mathfrak{D}} |M_{[\tau_1, \tau_2]}(x)|} \quad (7.29)$$

**Explanation Coverage** measures the bijection from the set of samples where there is an overlap between label annotations and neuron activation inside the activation range to the set of samples associated with label annotations. In other words, it measures the correlation's strength between the label's presence and the neuron activation. Higher is better.



**Figure 7.3:** Percentage of labels' category associated with the labels returned by CoEx over different activation ranges. Left: ranges from quantile to infinite. Right: range from 0 to quantile. Results are computed over 100 randomly extracted units.

$$ExplCov(L, \tau_1, \tau_2, \mathcal{D}) = \frac{|\{x \in \mathcal{D} : |M_{[\tau_1, \tau_2]}(x) \cap S(x, L)| > 0\}|}{|\{x \in \mathcal{D} : |M_{[\tau_1, \tau_2]}(x)| > 0\}|} \quad (7.30)$$

**Label Masking** measures the correlation between the explanation label  $L$  and the activation range  $[\tau_1, \tau_2]$ . Given a function  $\theta(x, L)$  that masks all the features in  $x$  but the ones corresponding to the label  $L$ , the score corresponds to the cosine similarity between activation of neuron  $k$  obtained by feeding the full and the masked input. Higher is better.

$$LabMask(L, \tau_1, \tau_2, \mathcal{D}) = \frac{\sum_{x \in \mathcal{D}} CosSim(M_{[\tau_1, \tau_2]}^k(x)A^k(\theta(x, L)), M_{[\tau_1, \tau_2]}^k(x)A^k(x))}{|\{x \in \mathcal{D} : |M_{[\tau_1, \tau_2]}(x)| > 0\}|} \quad (7.31)$$

## 7.3 Experiments

### 7.3.1 Motivation

The section begins by showcasing the limitations of the current state-of-the-art algorithms, which prompted the development of our algorithm. Existing methods, such as NetDissect and CoEx, focus on understanding a neuron's recognition capabilities by examining its highest activations, typically within the 0.005 percentile of activations. The objective function that they optimize is the same as Equation (7.2), but they investigate only the range  $T = [\tau^{top}, \infty]$  and NetDissect limits the explanations to the one in  $L \in \mathcal{L}^1$ .

Figure 7.3 presents the results obtained by varying the percentile to different values. Specifically, we explore values ranging from the highest to the lowest [0.005, 0.01, 0.02, 0.05, 0.1, 0.2, 0.5] percentile of neuron activations from 100 randomly extracted units in a ResNet18 model. For each run, we collect the labels associated with each unit by the CoEx algorithm. Then, we map each label to its category based on the Pascal dataset [56] annotations and record the percentage of labels for each category. The results are shown in Figure 7.3.

It is evident that the **distribution of labels varies with changes in the percentile**. For instance, when considering the highest activations, a larger quantile rewards colors while penalizing scene labels. Conversely, when considering the lowest activations, a larger quantile penalizes object labels. Notably, the distribution differs when considering the highest or lowest activations for the same quantile value. These findings suggest that neurons may recognize different concepts at different *activation levels*, indicating that solely considering exceptionally high activations provides a partial view. This limited perspective can lead to issues when comparing the interpretability of latent space, as they may overlook concepts to which the neurons are aligned (e.g., as in the previous chapter).

### 7.3.2 Performance

We begin the analysis of the proposed heuristic by comparing the computation steps needed to find the most aligned label against those required by CoEx, which does not use heuristics. In this scenario, we use ResNet18 [80] as the model to explain and Ade20k [287, 289] as the concept dataset. The results (Table 7.1) are computed over 100 randomly selected units and follow the settings of Mu and Andreas [163]. The results show that **MMESH is faster than CoEx by two orders of magnitude**. This improvement corresponds to transitioning from a computation time of 60 minutes per unit (CoEx) to less than two minutes (our algorithm)<sup>1</sup>. We expect that this enhancement will be more substantial in more complex settings, supporting larger datasets in terms of samples and concepts. Thus, for a reasonably low number of clusters, our algorithm can compute composition explanations for clusters within the same time frame (or less) taken by CoEx to compute compositional explanations for a single cluster.

### 7.3.3 Explanations

This section evaluates the explanations retrieved by the Clustered Compositional Explanation and analyzes phenomena related to activations emerging after the application of our algorithm to the latent space of popular Deep Neural Networks (DNNs). Following the literature on compositional explanations [145, 150, 163], we set the maximum explanation’s length to three. The first experiment compares our algorithm against NetDissect and CoEx using the metrics described in Section 7.2. This experiment test all the competitors in ResNet18 [80], DenseNet [92] and AlexNet [118] pre-trained on the Place365 dataset [288] and ResNet18 [80] and VGG16 [216] pre-pretrained on ImageNet. As a concept dataset, we use combinations of the models on both Ade20k [287, 289] and the Pascal dataset [56]. In all experiments, we focus on explaining the last layers of the networks following the setup of Mu and Andreas [163]. However, the results can be easily extended to any

**Table 7.1:** Avg. and standard deviation of steps required to reach the optimal solution within the beam.

Algorithm	Steps
NoHeuristics	$39656 \pm 12659$
MMESH	$129 \pm 712$

<sup>1</sup>Timing collected using a workstation powered by an NVIDIA GeForce RTX-3090 graphic card.

layer of a DNN. Finally, we employ K-Means as a clustering algorithm and fix the number of clusters to five.

Table 7.2 and Table 7.3 report the scores for individual clusters and their averages across different models and datasets, where the activation range orders clusters (i.e., Cluster 1 corresponds to the

**Table 7.2:** Avg. and Std Dev. across metrics reached by NetDissect, CoEx, and Clustered Compositional Explanations using Ade20K as a concept dataset.

	Cluster	IoU	ExplCov	SampleCov	ActCov	DetAcc	LabMask
ResNet							
NetDissect		0.05 ± 0.03	0.14 ± 0.20	0.58 ± 0.27	0.14 ± 0.17	0.15 ± 0.15	0.62 ± 0.26
CoEx		0.08 ± 0.03	0.14 ± 0.15	0.53 ± 0.25	0.15 ± 0.10	0.18 ± 0.10	0.57 ± 0.22
Our Avg.		0.15 ± 0.07	0.60 ± 0.36	0.69 ± 0.24	0.37 ± 0.17	0.20 ± 0.09	0.61 ± 0.16
Our	1	0.27 ± 0.02	0.99 ± 0.00	0.96 ± 0.02	0.54 ± 0.02	0.34 ± 0.03	0.64 ± 0.06
Our	2	0.15 ± 0.02	0.91 ± 0.16	0.80 ± 0.06	0.49 ± 0.09	0.18 ± 0.03	0.68 ± 0.07
Our	3	0.12 ± 0.03	0.56 ± 0.27	0.64 ± 0.15	0.33 ± 0.10	0.16 ± 0.05	0.62 ± 0.14
Our	4	0.10 ± 0.04	0.35 ± 0.19	0.52 ± 0.22	0.25 ± 0.12	0.15 ± 0.07	0.53 ± 0.19
Our	5	0.09 ± 0.05	0.21 ± 0.18	0.53 ± 0.26	0.22 ± 0.15	0.18 ± 0.09	0.58 ± 0.23
AlexNet							
NetDissect		0.03 ± 0.02	0.25 ± 0.34	0.32 ± 0.21	0.15 ± 0.19	0.13 ± 0.10	0.27 ± 0.21
CoEx		0.05 ± 0.02	0.24 ± 0.27	0.22 ± 0.16	0.14 ± 0.11	0.11 ± 0.07	0.23 ± 0.18
Our Avg.		0.12 ± 0.07	0.65 ± 0.34	0.67 ± 0.28	0.33 ± 0.18	0.17 ± 0.08	0.48 ± 0.17
Our	1	0.25 ± 0.04	0.99 ± 0.00	0.91 ± 0.09	0.57 ± 0.03	0.30 ± 0.06	0.55 ± 0.11
Our	2	0.12 ± 0.03	0.88 ± 0.20	0.76 ± 0.16	0.43 ± 0.13	0.15 ± 0.03	0.60 ± 0.13
Our	3	0.10 ± 0.03	0.63 ± 0.25	0.56 ± 0.16	0.29 ± 0.09	0.14 ± 0.04	0.54 ± 0.13
Our	4	0.08 ± 0.03	0.47 ± 0.25	0.37 ± 0.15	0.23 ± 0.10	0.12 ± 0.03	0.42 ± 0.13
Our	5	0.06 ± 0.02	0.27 ± 0.25	0.23 ± 0.10	0.14 ± 0.10	0.13 ± 0.06	0.27 ± 0.14
DenseNet							
NetDissect		0.05 ± 0.03	0.05 ± 0.04	0.52 ± 0.30	0.07 ± 0.05	0.15 ± 0.09	0.40 ± 0.26
CoEx		0.06 ± 0.03	0.09 ± 0.10	0.51 ± 0.27	0.11 ± 0.06	0.13 ± 0.07	0.42 ± 0.25
Our Avg.		0.20 ± 0.08	0.77 ± 0.30	0.83 ± 0.20	0.43 ± 0.15	0.27 ± 0.10	0.52 ± 0.27
Our	1	0.14 ± 0.05	0.63 ± 0.27	0.74 ± 0.12	0.35 ± 0.13	0.20 ± 0.07	0.61 ± 0.27
Our	2	0.26 ± 0.06	0.96 ± 0.09	0.96 ± 0.07	0.53 ± 0.06	0.35 ± 0.08	0.48 ± 0.25
Our	3	0.26 ± 0.03	0.99 ± 0.01	0.98 ± 0.04	0.55 ± 0.02	0.33 ± 0.05	0.38 ± 0.20
Our	4	0.20 ± 0.07	0.84 ± 0.23	0.85 ± 0.21	0.44 ± 0.13	0.27 ± 0.10	0.57 ± 0.27
Our	5	0.11 ± 0.05	0.42 ± 0.28	0.62 ± 0.21	0.28 ± 0.14	0.18 ± 0.08	0.56 ± 0.30

**Table 7.3:** Avg. and Std Dev. across metrics reached by NetDissect, CoEx, and Clustered Compositional Explanations on ResNet. Concepts are retrieved from the Pascal dataset.

	Cluster	IoU	ExplCov	SampleCov	ActCov	DetAcc	LabMask
NetDissect		0.05 ± 0.05	0.32 ± 0.24	0.25 ± 0.17	0.23 ± 0.20	0.08 ± 0.07	0.28 ± 0.17
CoEx		0.09 ± 0.05	0.31 ± 0.22	0.22 ± 0.15	0.19 ± 0.17	0.09 ± 0.07	0.26 ± 0.15
Our Avg.		0.15 ± 0.05	0.68 ± 0.31	0.64 ± 0.26	0.39 ± 0.18	0.20 ± 0.10	0.55 ± 0.17
Our	1	0.28 ± 0.02	0.99 ± 0.00	0.97 ± 0.01	0.56 ± 0.02	0.35 ± 0.02	0.63 ± 0.08
Our	2	0.15 ± 0.01	0.99 ± 0.01	0.84 ± 0.03	0.57 ± 0.04	0.17 ± 0.02	0.68 ± 0.08
Our	3	0.11 ± 0.02	0.61 ± 0.27	0.60 ± 0.09	0.28 ± 0.10	0.17 ± 0.05	0.59 ± 0.10
Our	4	0.10 ± 0.06	0.42 ± 0.15	0.49 ± 0.16	0.27 ± 0.09	0.18 ± 0.09	0.53 ± 0.15
Our	5	0.09 ± 0.06	0.40 ± 0.21	0.30 ± 0.16	0.28 ± 0.20	0.12 ± 0.08	0.34 ± 0.15

range including the lowest activations). Inspecting the mere scores, NetDissect exhibit similar performance CoEx. CoEx demonstrates a slight superiority in IoU, while NetDissect is superior in Sample Coverage. Overall, the **Clustered Compositional Explanations algorithm assigns labels of better quality** across all considered clusters, models, and datasets. Therefore, there is no rationale related to explanation quality to justify disregarding the full spectrum of neuron activations when explaining neuron behavior.

However, another reason to avoid considering the full spectrum could be that the network exclusively uses the highest activation to process specific concepts. To verify if this is the case, we masked the activations within a given range of the last layer of the feature extractor. Intuitively, if the network relies solely on the highest activations to recognize a concept, masking these activations should produce a more significant shift in predictions than masking lower activations. Table 7.5 demonstrates that this is not the case and changes in predictions are similar between using the narrow range of CoEx and the ranges used by the clusters of our algorithm. These results align with the essence of DNNs, designed to **leverage all activations of a neuron**, regardless of their magnitude.

Interestingly, Table 7.5 indicates that the labels associated with the lowest cluster have the least impact on the decision process. In Table 7.2, this cluster is associated with Explanation and Sample coverage close to one, suggesting that the neurons fire in almost all samples containing the cluster’s label and rarely if the label is not included. By inspecting the labels associated with these clusters, the labels often refer to combinations of colors and background concepts (e.g., Sky OR Blue), with low variability across neurons, as most neurons are associated with the same labels. The same behavior is observed in the third cluster in DenseNet and, to a lesser extent, in

**Table 7.4:** Avg. and Std Dev. of the desired qualities over the labels returned by NetDissect, CoEx, and Clustered Compositional Explanations (our) applied to ResNet pre-trained on ImageNet. Results are computed for 50 randomly extracted units.

	Cluster	IoU	ExplCov	SampleCov	ActCov	DetAcc	LabMask
ResNet18							
NetDissect		$0.04 \pm 0.02$	$0.13 \pm 0.17$	$0.36 \pm 0.27$	$0.10 \pm 0.12$	$0.12 \pm 0.11$	$0.37 \pm 0.24$
CoEx		$0.05 \pm 0.03$	$0.13 \pm 0.16$	$0.28 \pm 0.20$	$0.12 \pm 0.08$	$0.12 \pm 0.08$	$0.32 \pm 0.19$
Our Avg.		$0.13 \pm 0.08$	$0.65 \pm 0.37$	$0.65 \pm 0.29$	$0.36 \pm 0.19$	$0.18 \pm 0.11$	$0.56 \pm 0.18$
Our	1	$0.28 \pm 0.02$	$0.99 \pm 0.01$	$0.98 \pm 0.02$	$0.54 \pm 0.02$	$0.36 \pm 0.03$	$0.64 \pm 0.06$
Our	2	$0.15 \pm 0.03$	$0.95 \pm 0.12$	$0.89 \pm 0.06$	$0.53 \pm 0.07$	$0.18 \pm 0.04$	$0.73 \pm 0.08$
Our	3	$0.13 \pm 0.05$	$0.79 \pm 0.23$	$0.66 \pm 0.11$	$0.39 \pm 0.12$	$0.10 \pm 0.03$	$0.63 \pm 0.11$
Our	4	$0.07 \pm 0.03$	$0.32 \pm 0.21$	$0.42 \pm 0.18$	$0.18 \pm 0.10$	$0.12 \pm 0.06$	$0.48 \pm 0.15$
Our	5	$0.06 \pm 0.04$	$0.19 \pm 0.16$	$0.29 \pm 0.19$	$0.15 \pm 0.11$	$0.12 \pm 0.08$	$0.34 \pm 0.18$
VGG-16							
NetDissect		$0.03 \pm 0.02$	$0.13 \pm 0.18$	$0.35 \pm 0.26$	$0.11 \pm 0.11$	$0.11 \pm 0.11$	$0.39 \pm 0.25$
CoEx		$0.04 \pm 0.03$	$0.10 \pm 0.10$	$0.28 \pm 0.19$	$0.10 \pm 0.07$	$0.10 \pm 0.09$	$0.33 \pm 0.18$
Our Avg.		$0.08 \pm 0.06$	$0.38 \pm 0.34$	$0.45 \pm 0.26$	$0.22 \pm 0.17$	$0.11 \pm 0.07$	$0.36 \pm 0.16$
Our	1	$0.16 \pm 0.06$	$0.92 \pm 0.16$	$0.78 \pm 0.16$	$0.48 \pm 0.13$	$0.20 \pm 0.07$	$0.38 \pm 0.13$
Our	2	$0.07 \pm 0.03$	$0.43 \pm 0.24$	$0.56 \pm 0.20$	$0.20 \pm 0.10$	$0.10 \pm 0.05$	$0.43 \pm 0.17$
Our	3	$0.06 \pm 0.03$	$0.29 \pm 0.21$	$0.40 \pm 0.18$	$0.18 \pm 0.11$	$0.08 \pm 0.04$	$0.39 \pm 0.15$
Our	4	$0.08 \pm 0.04$	$0.16 \pm 0.15$	$0.28 \pm 0.16$	$0.13 \pm 0.10$	$0.08 \pm 0.04$	$0.33 \pm 0.15$
Our	5	$0.05 \pm 0.04$	$0.10 \pm 0.12$	$0.24 \pm 0.18$	$0.12 \pm 0.10$	$0.08 \pm 0.07$	$0.28 \pm 0.17$

the adjacent clusters. We apply the Clustered Compositional Explanation algorithm to random initialized networks to understand the convergence towards the same labels. Surprisingly, all the clusters in these networks are associated with labels similar to those in the lowest clusters of the trained network. Therefore, these labels represent unspecialized activations, which are activations that cannot be linked to specific behaviors and are less important than the others. We term these labels as *default labels*.

Further investigation into activation-related phenomena reveals insights into unspecialized and weakly specialized activations. By analyzing labels across clusters, we compute the percentage of default labels. We define *unspecialized activations* as those for which the whole compositional explanation is a default label. Conversely, activations where only the last part of the label converges on a different concept are defined as *weakly specialized activations*. Table 7.6 shows that **Cluster 1 in ReLU networks and Cluster 3 in non-ReLU networks include unspecialized activations**. These clusters correspond to activations closer to 0. Indeed, ReLU layers return non-negative activations; thus, the lowest cluster is the one that includes activations close to zero. Conversely, non-ReLU can return negative activations, which are captured by the lowest clusters, and activations close to zero are captured by middle clusters. The percentage of unspecialized activations decreases when moving away from zero. **Weakly specialized units are rare**, especially in ReLU networks. The increment in percentage in adjacent clusters may be due to clustering algorithm artifacts (i.e., spurious outliers in a cluster). This phenomenon can be further investigated in future work by using a cleaning process on the clusters or developing clustering algorithms tailored to the task.

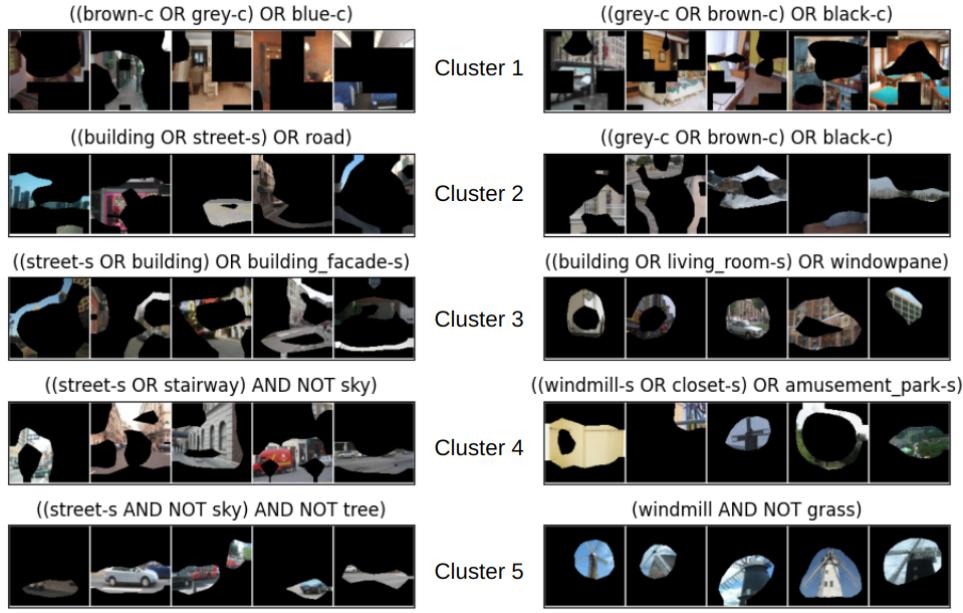
In ReLU layers, **we also observe progressive specialization of activations** with lower activations recognizing general concepts like scenes and buildings, middle activations combining general concepts with more complex ones, and the highest activations focusing on specific and smaller objects (Figure 7.4). This phenomenon also appears spatially aligned, with lower activations detecting background and side objects, while the highest ones focus on smaller foreground objects.

**Table 7.5:** Percentage of predictions changed when the activations of a given cluster are masked.

Prediction Change %	
CoEx	12.9
Cluster 1	9.7
Cluster 2	12.90
Cluster 3	12.29
Cluster 4	11.62
Cluster 5	13.79

**Table 7.6:** Percentage of unspecialized and weakly specialized activation ranges in ResNet18 (ReLU Yes) and DenseNet161(ReLU No).

	ReLU	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Unspecialized	Yes	0.93	0.37	0.01	0.00	0.00
	No	0.05	0.60	0.95	0.30	0.05
Weakly Specialized	Yes	0.00	0.03	0.01	0.01	0.00
	No	0.05	0.12	0.02	0.17	0.02



**Figure 7.4:** Examples of specialization (left) and polysemy (right). Neuron #368 (right) recognizes unrelated concepts among different activation ranges (windmill, closet, amusement park).

The specialization property highlighted by Mu and Andreas [163] can be considered related to progressive specialization. Figure 7.4 (left) shows an example of such specialization, where the neuron recognizes streets in more and more specific contexts.

Regarding individual clusters, we can analyze the performance of Cluster 5, which contains the highest activations. Therefore, this cluster also includes the activation range used by NetDissect and CoEx. Comparing its scores against CoEx, we observe that slightly increasing the range size does not lower the quality of explanations, as Cluster 5 outperforms CoEx or reaches the same performance in all scores.

Lastly, similarly to Mu and Andreas [163], we investigate the polysemy of activations, generalizing the polysemy to the clusters. A label is monosemic if all the concepts involved in the label refer to similar or related concepts. For example, the label ((chair AND table) OR kitchen) is monosemic since its atomic concepts are related. Conversely, a label is polysemic if it involves unrelated concepts. Therefore, we define a neuron as monosemic if all its clusters are associated with monosemic labels, excluding the default ones. Conversely, a neuron is polysemic if at least one of its clusters is associated with polysemic labels. In our manual inspection, **15% of the neurons meet the criterion to be monosemic**. In these cases, neurons are referred to as **highly specialized**. The percentage is much lower than the one found by Mu and Andreas [163], who consider only the highest activations (%85). This difference underscores the necessity of considering a wider spectrum of activations when studying their recognition capability.

### 7.3.4 Metrics

In this section, we investigate the feasibility of exploiting the aforementioned metrics in the optimization process, with a specific focus on non-sample-based metrics.

We begin our discussion with IoU, which has been the metric optimized by all algorithms

discussed in this chapter to identify the most aligned concepts. Hence, the previously reported results correspond to those obtained by solely optimizing this metric. One limitation of IoU, as highlighted in previous literature [145, 163], is that the IoU can be artificially increased by increasing the formula’s length. Common fixes for this issue include presetting the length for all neurons before applying the algorithm or introducing a term to penalize lengthy explanations [163].

Concept masking involves multiple iterations of feeding the entire dataset to the architecture. Consequently, optimizing this metric during the beam process poses computational time challenges.

In the case of optimizing for activation coverage, given the fixed denominator, only the numerator - the intersection between labels and activations - matters. In this case, the intersection disproportionately favors larger concepts covering the entire activation, even if the covered portion is a tiny fraction of the label annotation. Consequently, at first sight, scene concepts covering the whole image are expected to be associated with the neurons. However, scene concepts do not cover the full dataset, resulting in several zeros in the equation. Therefore, the algorithm converges towards default rules composed of colors since they are the most frequent ones and the largest among the most frequent.

Table 7.7 shows the results using Detection Accuracy as an optimization metric. While improvements in terms of Sample Coverage and Detection accuracy are expected, it is interesting to note that Concept Masking also benefits from this metric. Conversely, the IoU score and Activation Coverage drop down to zero, indicating a potential correlation between these metrics and an opposite behavior to the Detection Accuracy.

**These results underscore the absence of a straightforward method in the optimization process to artificially inflate the overall score using novel metrics** and reinforce the notion that multiple metrics should be used at the same time to have a fuller understanding of the quality of these approaches.

## 7.4 Design Choices

This section delves into design choices connected to the implementation of our proposed algorithm.

**Number of Clusters.** One critical design choice is the number of clusters per neuron. Here, we evaluate the quality of explanations when the number of clusters ranges from 5 to 15. Table 7.8

**Table 7.7:** Avg. and Std Dev. of the scores obtained by optimizing the algorithm using Detection Accuracy. Underlined results are those that reach lower scores compared to when using the IoU score. Bold results are those that reach higher scores.

	IoU	ExplCov	SampleCov	ActCov	DetAcc	LabMask
Our	$0.01 \pm 0.03$	$0.01 \pm 0.03$	$0.74 \pm 0.26$	$0.01 \pm 0.04$	<b><math>0.68 \pm 0.23</math></b>	$0.68 \pm 0.21$
Cluster 1	$0.00 \pm 0.00$	$0.00 \pm 0.00$	$0.90 \pm 0.19$	$0.00 \pm 0.00$	<b><math>0.89 \pm 0.12</math></b>	$0.60 \pm 0.17$
Cluster 2	$0.00 \pm 0.00$	$0.00 \pm 0.00$	$0.82 \pm 0.22$	$0.00 \pm 0.00$	<b><math>0.74 \pm 0.21</math></b>	$0.72 \pm 0.23$
Cluster 3	$0.00 \pm 0.00$	$0.00 \pm 0.00$	$0.69 \pm 0.26$	$0.00 \pm 0.00$	<b><math>0.68 \pm 0.22</math></b>	<b><math>0.70 \pm 0.23</math></b>
Cluster 4	$0.00 \pm 0.00$	$0.00 \pm 0.01$	<b><math>0.65 \pm 0.27</math></b>	$0.00 \pm 0.00$	<b><math>0.53 \pm 0.20</math></b>	$0.69 \pm 0.21$
Cluster 5	$0.03 \pm 0.06$	$0.03 \pm 0.07$	<b><math>0.64 \pm 0.24</math></b>	$0.03 \pm 0.08$	<b><math>0.54 \pm 0.19</math></b>	<b><math>0.68 \pm 0.22</math></b>

suggests that increasing the number of clusters does not yield better explanations. Moreover, by inspecting the labels associated with the clusters, most labels (75%) are repeated across adjacent clusters when the number of clusters is high. Note also that computing explanations for 15 clusters requires more computation time ( $\times 3$ ) than computing explanations for five clusters. Therefore, there is a trade-off between computational time and novel coverage. While these results suggest that a lower number of clusters is the best option, we hypothesize that the iterative procedure of K-means can also influence them. Therefore, developing a clustering algorithm tailored to the task could potentially flip these results by finding that a higher number of clusters leads to better explanations.

**Clustering Algorithm.** Connected to the choice of the number of clusters, another design choice regards the clustering algorithm. The algorithm can be chosen based on the dataset's size and the activation vectors' size. In the settings reported in this thesis, the dataset and the activations vectors are so large that the K-Means algorithm is the only viable option among the clustering algorithms publicly available. Indeed, to compute the clusters, more sophisticated algorithms would require either an amount of time greater than the one saved by our heuristics or too much memory space to be feasible in standard workstations.

**Heuristics.** The third choice is about the heuristic used to speed up the beam search. Here, we evaluate two alternative heuristics designed to use less information for estimating the IoU score: the *Areas* and the *Coordinates-Free* heuristics.

The **Areas** heuristic uses only the information about the mask size of the terms composing the current label:

$$\widehat{IoU}(L, \tau_1, \tau_2, \mathfrak{D}) = \frac{\widehat{I}_x}{\sum_{x \in \mathfrak{D}} |M_{[\tau_1, \tau_2]}(x)| - \widehat{I}_x} \quad (7.32)$$

where

$$\widehat{I}_x = \begin{cases} \min(|S_{[\tau_1, \tau_2]}(x, L_{\leftarrow})| + |S_{[\tau_1, \tau_2]}(x, L_{\rightarrow})|, |M_{[\tau_1, \tau_2]}(x)|) & op = OR \\ \min(|S_{[\tau_1, \tau_2]}(x, L_{\leftarrow})|, |S_{[\tau_1, \tau_2]}(x, L_{\rightarrow})|) & op = AND \end{cases} \quad (7.33)$$

$$\widehat{I}_x = \begin{cases} \min(|S_{[\tau_1, \tau_2]}(x, L_{\leftarrow})|, \text{size}(x) - |S_{[\tau_1, \tau_2]}(x, L_{\rightarrow})|) & op = AND NOT \end{cases} \quad (7.35)$$

The **Coordinates-Free Heuristic (CFH)** avoids the estimation of  $\widehat{S(L, x)}$  by setting it to 0, thus saving the computation time needed to compute the maximum and minimum possible extension:

$$IoU(L, \tau_1, \tau_2, \mathfrak{D}) = \frac{\widehat{I}_x}{\sum_{x \in \mathfrak{D}} |M_{[\tau_1, \tau_2]}(x)| - \widehat{I}_x} \quad (7.36)$$

**Table 7.8:** Avg. and Std Dev. of the proposed metrics when using a variable number of clusters. Results are computed for 50 randomly extracted units.

Clusters	IoU	ExplCov	SampleCov	ActCov	DetAcc	LabMask
5	$0.15 \pm 0.37$	$0.60 \pm 0.23$	$0.70 \pm 0.17$	$0.34 \pm 0.11$	$0.21 \pm 0.10$	$0.55 \pm 0.22$
10	$0.09 \pm 0.05$	$0.48 \pm 0.35$	$0.68 \pm 0.23$	$0.30 \pm 0.16$	$0.13 \pm 0.06$	$0.58 \pm 0.17$
15	$0.07 \pm 0.04$	$0.44 \pm 0.36$	$0.66 \pm 0.21$	$0.25 \pm 0.16$	$0.09 \pm 0.05$	$0.56 \pm 0.18$

where  $\widehat{I}_x$  is defined as in MMESH.

Table 7.9 compares the three competitors based on the number of visited labels before finding the optimal label. The reported results are the average across 100 randomly selected units. Labels are computed using the standard settings of CoEx (i.e., by fixing  $T = [\tau^{top}, \infty]$ ). Not surprisingly, MMESH achieves the target label much faster than the other heuristics. However, in scenarios where computing the coordinates of the maximum and minimum extension is prohibitively expensive, CFH serves as a viable alternative, as it is one order of magnitude faster than not using heuristics at all. Regarding the Areas heuristics, while it performs worst among those tested, it offers a key advantage: it does not require running the first step of the algorithm (i.e., NetDissect), as it does not use the information about the minimum possible intersection. Consequently, it can be employed within NetDissect to speed up the search process when the label search space is large.

## 7.5 Contributions

This chapter contributes and end the part of this thesis dedicated to the research on explaining the latent space learned by DNNs. It introduced a method applicable to any DNN for analyzing alignment between latent representations and a predefined set of concepts. Specifically, in contrast with previous literature, it studies the full spectrum of activations by combining clustering, heuristics, and beam search.

The chapter discussed the gain in terms of the efficiency of the algorithm with respect to the state-of-the-art ones and how analyzing a wider spectrum of activations does not lower the quality of explanations and instead provides a fuller picture of the neurons' behavior. The chapter also sheds light on novel phenomena connected to neuron activations, such as the unspecialized activations around zero and the progressive specialization. Additionally, this chapter contributes to the research on metrics for explanation methods by proposing novel metrics for evaluating approaches that explain latent representations. Alongside these strengths, the chapter also analyzed some of the drawbacks of these approaches, like the challenges related to the usage of clustering algorithms.

Finally, this chapter ends the part of the thesis devoted to the introduction of novel methods for explaining deep learning. We began with the introduction of a generalization of a known architecture (Chapter 4) to a specific domain and settings, and subsequently, we progressively discussed more generalized approaches aimed at mitigating previous issues and expanding the potential applications. This latest algorithm, being model-independent and not necessitating modification or additional training, signifies the apex of such efforts. Therefore, the next chapter describes a more general analysis of the XAI field. Then, the final chapter discusses the limitations of the proposed approaches and possible future directions, including the ones discussed in this chapter.

**Table 7.9:** Avg. and standard deviation of visited states per unit. Results are computed for 100 randomly extracted units.

Heuristic	Info from		Visited Labels
	NetDissect	Dataset	
Areas	-	✓	$23602 \pm 3420$
CFH	✓	✓	$5990 \pm 3066$
MMESH	✓	✓	$129 \pm 712$

## Part IV

# Bringing Explanations to the User: Visual Analytics

## Chapter 8

# Explaining Deep Learning in Visual Analytics Systems

In the preceding chapters, this thesis explored various mechanisms to explain the behavior of Deep Neural Networks (DNNs). We showcased several usage scenarios where explanations are employed to assist users. In those instances, the explanations are static, target the model developer, and necessitate the presence of an eXplainable Artificial Intelligence (XAI) expert.

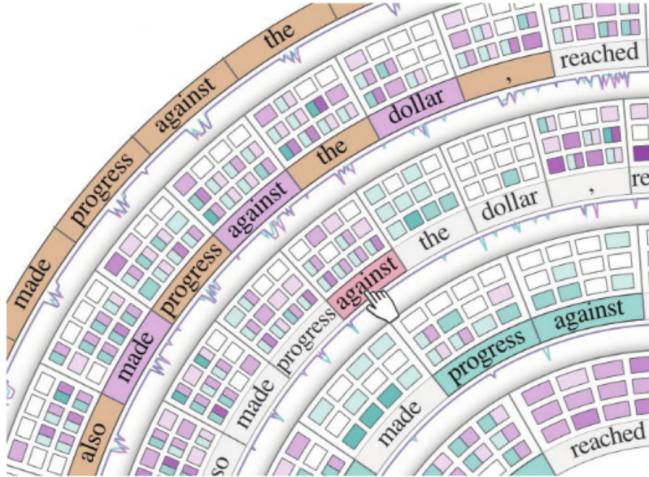
This chapter discusses a promising direction for delivering explanations to users and allowing them to interact with models and explanations. Specifically, it discusses Visual Analytics (VA) systems that employ explanation methods to understand DNNs. The VA community has recently developed intuitive and interactive user interfaces to aid users in understanding machine learning models. Unlike XAI research, VA systems are typically designed for specific tasks and target users. The benefits and novelty of these methods rely more on the functionalities offered to the user and the tasks they can solve rather than the development of specific XAI methods. In particular, VA excels in ease of exploration, hypothesis verification, and discovery when dealing with large datasets.

A natural question arises: can VA and XAI complement each other by combining their strengths and integrating XAI methods into VA systems? This chapter evaluates the current state of integration between the two disciplines by reviewing and analyzing 67 papers proposing VA systems that incorporate XAI methods for understanding DNNs, which is the topic of this thesis.

The chapter is organized as follows: Section 8.1 provides a brief overview of visual analytics systems employing explanation methods; Section 8.2 analyzes strengths and weaknesses of the current integration between XAI and VA; Section 8.3 discusses challenges and next steps needed in the field; and Section 8.4 summarized the contributions of this chapter.

### 8.1 Overview

This section provides an overview of how VA systems use explanation methods to enhance the interpretability of DNNs. Specifically, this section analyzes the support provided for each of the explanation categories presented in Section 2.2. Additionally, it discusses and presents a novel VA-specific category: model behavior. For each category, this section describes both the main visualization and interaction techniques used in the systems and the tasks the users can accomplish thanks to the integration of these XAI methods into the system. While some of these tasks could be



**Figure 8.1:** Example of complex visualization in VA systems of attention flow.

achieved by using XAI methods alone, VA systems expedite the analysis process through broader exploration and provide tools to the users without the need for Deep Learning (DL) or XAI expert involved in the process.

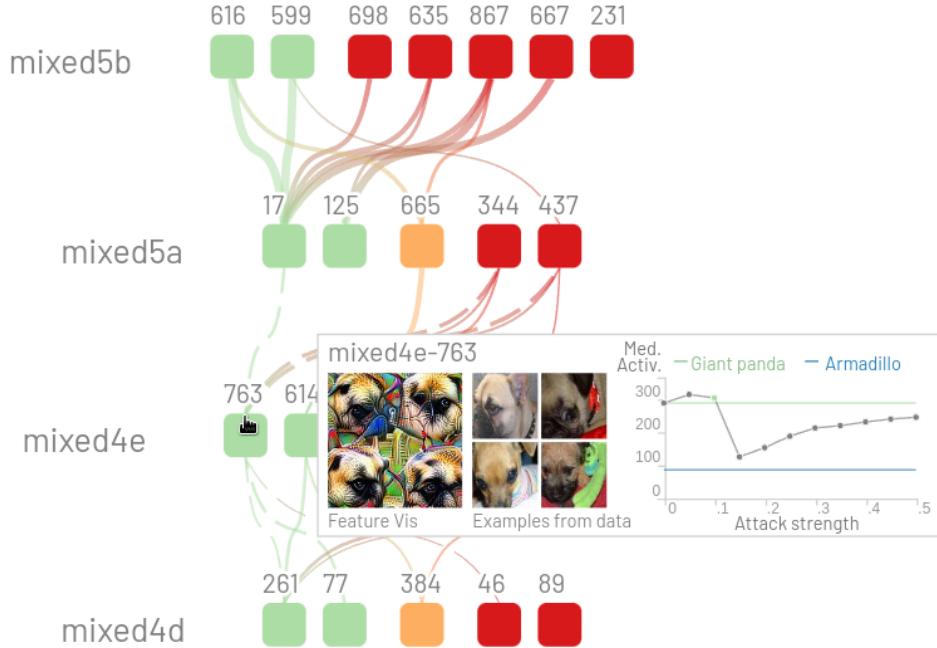
### 8.1.1 Feature Attribution

Unsurprisingly, feature attribution is the most popular category for VA systems to enhance interpretability. For individual explanations and standard settings, VA systems generally employ common visualization techniques also used in XAI papers, such as heatmaps [28, 29, 31, 75, 85, 94, 95, 97, 220, 226, 242, 257, 278] and matrices [45, 96, 136, 180, 253]. However, for more advanced settings, VA exposes its full strength by proposing novel and more complex visualizations. For example, the system proposed by DeRose et al. [45] visualizes Transformers' layers in a radial layout, helping users keep track of the flow of attention weights across multiple layers simultaneously (Figure 8.1). Users can interact with the interface by rotating the radians, filtering layers, heads, and input tokens. This kind of support is common to the systems supporting the analysis of the most complex architectures [45, 220, 262].

Regarding the interactions, sorting [31, 120, 157, 180, 248] and filtering [47, 96, 97, 253] are the most basic and commonly supported for this explanation category. These two interactions are particularly useful for global feature attribution, where filtering can aid users in detecting features relevant to specific subsets of the data [31, 79, 97, 97, 120, 180, 213, 242, 284].

Advanced interactions for this category include real-time what-if analysis and steering the model training by using attributions. In the former case, local explanations can guide features removal (e.g., by brushing over an image) and monitor the changes in predictions [75, 94, 120] or in the decision process when the input is modified [96, 136, 181, 228]. In the latter case, users can impose constraints on the training process by specifying the desired values in feature attributions. For example, some systems support the user in specifying the desired attention values in self-explainable DNNs [120, 157]. Then, the model is optimized to align attention values closer to the specified ones.

Papers proposing these systems demonstrate that combining visualizations and iterative interactions with feature attribution helps users on several tasks, such as assessing the reliability



**Figure 8.2:** Example of abstraction in VA system.

[181, 242] of predictions, discovering bias in the network’s decision process [96, 248], and correct failures [31, 85, 257]. Moreover, in domains where domain experts are involved, these systems lead to the discovery of new knowledge in terms of interactions between factors [242] and pitfalls in the design of the DNNs themselves [75].

### 8.1.2 Learned Features

Several systems use learned features methods for helping users understand the knowledge learned by DNNs. In this case, one of the major benefits of VA systems lies in the extensive supported exploration and the summarization capability.

While the visualization of learned features for a specific component is mostly aligned with the current literature in XAI, using real or generated patches in the vision domain [43, 43, 88, 90, 90, 98, 134, 143, 182, 194, 194, 226, 274, 274, 280, 284] and word clouds in the text data domain [43, 88, 90, 143, 194, 226, 274], the exploration and simultaneous visualization of learned features for entire parts of the network require more complex visualizations and interactions.

VA solutions address this challenge by providing abstractions summarizing the behavior of macro-components and then enabling the user to explore individual components within the macro ones through interactions. Abstractions are provided by aggregating activations, based on their average or similarity [90, 97], using clustering algorithm [133, 182, 286], or based on the importance of the individual components for the task [43].

For example, Figure 8.2 shows an abstraction provided by BLUFF [43], which visualizes, for each layer, the neurons whose activations diverge the most when under an adversarial attack. This abstraction allows users to focus on the most problematic neurons and then explore the rest of the

architecture. Abstractions are visualized in the form of Sankey diagrams [43, 90], graphs [134, 194] or scatter plots using dimensionality reduction techniques [186, 243].

In systems employing learned features techniques, interactions are mainly designed to ease the exploration of the networks. Zoom-in [90, 182] and filtering [90, 182] are the standard interactions available in most of the analyzed systems. More advanced interactions include the definition of custom clusters [133, 286], switching between facets [90, 133], and filtering the granularity of the visualization or the modification of the importance criterion [43].

Papers proposing VA systems employing learned features techniques demonstrate their efficacy in supporting users in the analysis of how low-level features are aggregated into high-level features [43, 90, 97, 133, 134, 182], in the diagnosis of the training process [187, 286], in the comparison between models or configurations [143], and to validate that the learned knowledge is aligned to the expected one [97, 133].

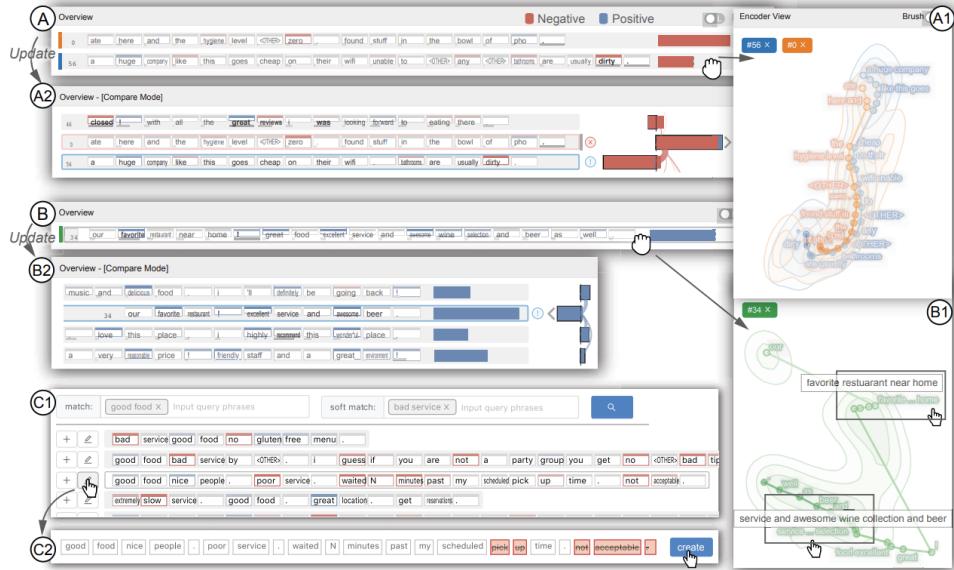
### 8.1.3 Explanations by Examples and Counterfactuals

Explanations by examples and counterfactuals are rarely integrated into VA systems, and when employed, they typically serve as complementary tools for other techniques. Therefore, they represent categories that need further research. This section jointly analyzes these two categories because half of the analyzed papers support both simultaneously, suggesting they are often considered complementary.

The most common visualizations employ a list of inputs enriched by metadata such as the similarity scores, predictions [29], highlights of common features, or highlights of differences between the explanation and the input [29, 29, 91, 227]. Average statistics are usually summarized in bar charts [36] or tables [36, 263]. These visualizations align with the visualizations used in XAI for this category [108]. When the explanations are retrieved using similarity on patterns (e.g., attention [97] or features [81]), these patterns are usually visualized and highlighted along with the explanations by examples.

Regarding interactions, they are mainly based on sorting mechanisms [36, 136], reducing or increasing the number of explanations [36, 97], and selecting input features that should be shared by the explanation by example [227]. A special case is represented by self-explainable DNNs based on prototypes, where explanations by examples are used to extract the semantics encoded in prototypes, similarly to the process presented in Chapter 4. In these cases, Ming et al. [157] propose to use interactions to steer the model training. The system allows the user to specify a set of desired neighbors (i.e., explanations by examples), and these samples are used to generate new prototypes that include the desired samples as neighbors (Figure 8.3).

VA systems utilize both these types of explanations to enable users to understand individual predictions [228] and suggest edits to the user [136, 228, 252, 263, 266]. Explanations by examples are used to estimate the meaning of latent vectors [91, 227, 228], improve the representativeness of a prototype representation [157], and keep track of the decision process layer by layer [91, 228]. Conversely, several systems use counterfactuals to verify hypotheses [36, 227], or explore alternative scenarios in reinforcement learning [158]



**Figure 8.3:** Example of interface to steer a model to add, delete, and revise prototypes. All edits are revertable and traceable in the editing history. Image from Ming et al. [157] © 2019 IEEE

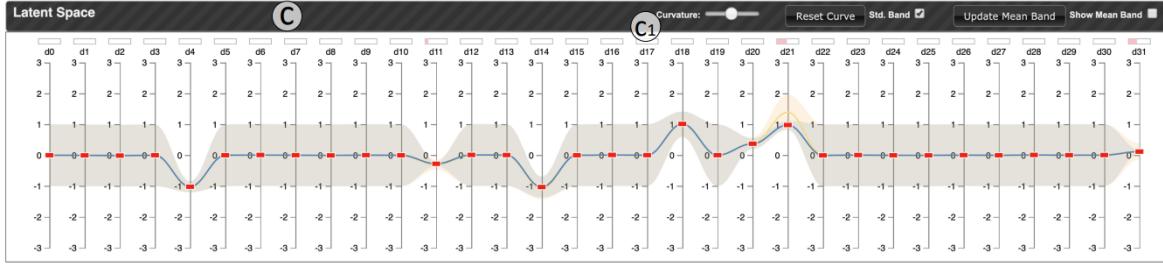
#### 8.1.4 Model Behavior

Model behavior is a category specific to VA systems and aims at explaining the DNN by facilitating the discovery of patterns in terms of predictions or values of the model components and linking them to specific behaviors. By design, these methods are inherently post-hoc and global, combining pattern mining, clustering, interactions, and sometimes explanations of other categories [95, 255]. They are especially useful when other explanation methods fail to provide a comprehensive understanding if used alone, or classical XAI methods are not applicable.

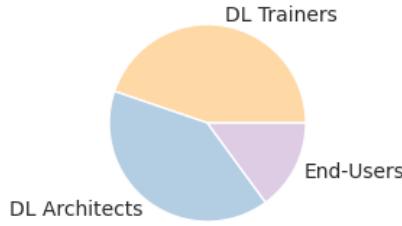
The goal of this category is to improve the simulability of a network by providing answers to questions such as “How does the model react in a given situation?”. This category is strongly linked to intrinsic methods of XAI since the patterns often consist of the values of internals (e.g., neuron activations, attention weights). The methods of this category are also loosely related to the XAI research on circuits [176] and pathways discovery [152].

One of the distinctive characteristics of VA systems supporting this category of explanations is the extensive use of linked views. In this case, several views (i.e., parts of the interface) are connected, and a modification of an element in a view causes a change in the visualization of the other connected views. This feature facilitates the detection of patterns from multiple points of view simultaneously. Patterns can be visualized as lists of neurons [181], tables [261], partial dependency plots [252] or summarized in a more compact visualization, like decision trees [18, 98] and scatter plots [278].

Interactions are the core components of these systems since the explanations are human-driven. The most popular interactions include the selection of subsets of interest [21, 102, 213], zooming functionalities [95, 100, 259], sorting mechanisms, and definition of patterns [255]. Several systems provide support for *interactive input observations* and *interactive model observations*. For example, Wang et al. [258] (Figure 8.4) and Strobelt et al. [227] use Parallel Coordinates Plots to represent input features or layer activations and let the user modify the values selected in each axis in real



**Figure 8.4:** View to interact with a latent representation. The user can assign values to dimensions of the latent space in the parallel coordinate plot, and the system will adjust the other views accordingly. Image from Wang et al. [258] ©2020 IEEE



**Figure 8.5:** Distribution of target users in the analyzed VA systems

time to visualize how the network changes its behavior accordingly.

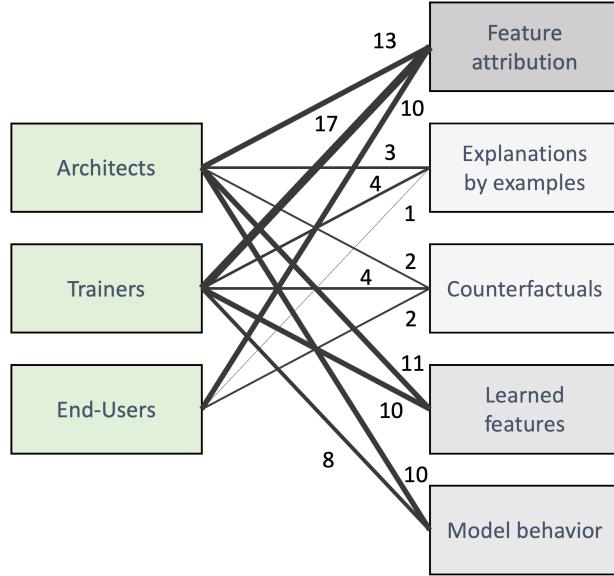
Papers proposing VA systems supporting this category successfully help users to extract policies from deep reinforcement learning [255], approximate the decision process of models, layer, and neurons [18, 95, 98, 181, 255, 261, 278], analyze errors [100, 134], and formulate and refine hypotheses about the semantics associated with the latent spaces [213, 227, 258, 259].

## 8.2 Analysis

This section analyzes the considered VA systems by discussing the target users and domains of these systems, the kind of interactions supported, the evaluation of their usefulness, and their state of integration with DL and XAI fields.

**Target Users.** We begin the analysis by examining the target users of the collected systems. Figure 8.5 shows the distribution of target users in the analyzed VA systems, revealing that DL architects and trainers are the primary focus. This is not surprising since DL experts (i.e., individuals with a background in DL) are also the predominant target users of XAI methods, particularly when used for debugging, which is one of the current main applications of XAI techniques. Conversely, only 14% of systems target end-users without knowledge about DL. This distribution can be considered an indication of the maturity of the XAI field: the field seems still confined mainly in research applications.

The analysis of the categories supported by systems targeting end-users (Figure 8.6) reveals a reliance on feature attribution methods, with limited support for explanations by examples and counterfactuals and no support for model behavior and learned features categories. These findings

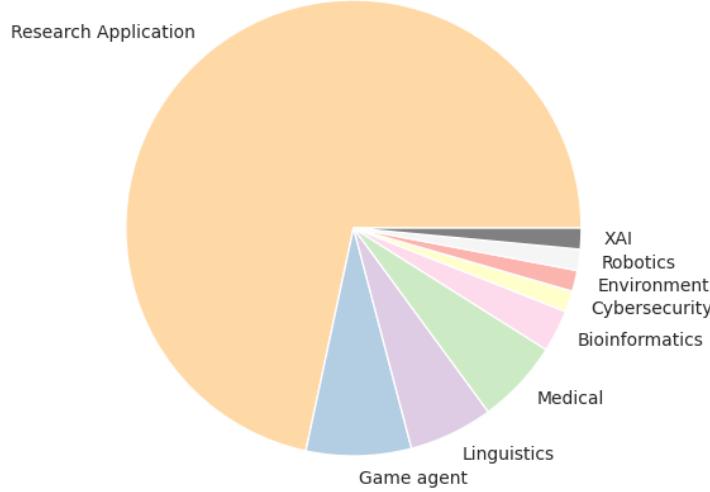


**Figure 8.6:** Summary of user support for different XAI categories. The edge width is proportional to the number of systems connecting users and categories. The opacity of XAI categories is proportional to the total number of contributions.

suggest that the latter explanation types may not yet be sufficiently mature, and further research is needed in this direction. For instance, inspecting and navigating the network’s neurons to understand what the network learned during the training process may require too much effort and expertise for an end-user. In such cases, future directions could involve the employment of summarization techniques and circuits [176] that go beyond the currently supported XAI methods.

Another interesting aspect is that most systems targeting end-users deal with time series and text data, while only one system deals with image data. This finding is surprising given the long-standing maturity of feature attribution methods for image data. The popularity of time series may be attributed to its relevance in healthcare, a primary domain for XAI techniques [6, 177, 239]. Conversely, the low presence of image data suggests that a few real-world problems can benefit from the progress in this research yet.

**Target Domains.** The second analysis discusses the domains investigated in the collected systems. Figure 8.7 corroborates the results regarding the users by showing that the majority of papers target research applications. The other applications target mainly *games*, *linguistics*, and *medical domains*. The inclusion of games and linguistics may be tied to the recent surge in reinforcement learning and large language models, while the medical domain represents a common testing scenario for XAI approaches, as noted before. Regarding the supported categories, in alignment with the analysis for the users, most systems targeting non-research rely on feature attribution methods to provide explainability. However, in this case, these methods are often complemented with model behavior (40%), explanations by examples (10%), or counterfactuals (10%) categories. As for end-users, the learned features category is not supported by any system, suggesting that this type of explanation is not mature enough to be digested by domain experts with limited knowledge of DL.



**Figure 8.7:** Target domain of VA systems. Research refers to research papers targeting general applications.

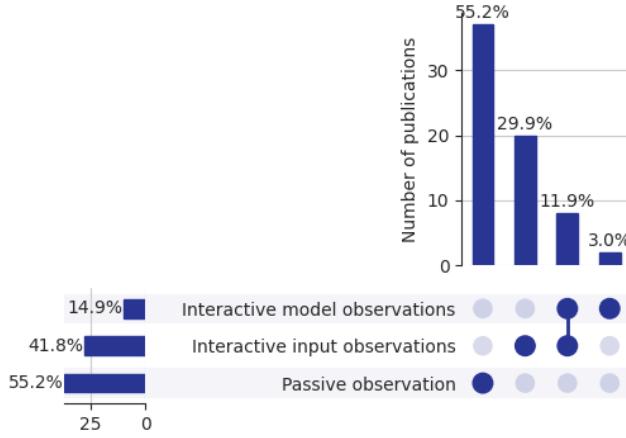
**Interactions.** Figure 8.8 presents the distribution of the interaction degree these systems offer to the target users. Notably, over half of the systems provide only *passive observations*, particularly those relying on feature attribution, learned features, and model behavior. These systems exploit interactivity to support exploration, summarization, and filtering capabilities mainly. One-third of the systems additionally support *interactive input observations*, allowing users to modify the input features. Systems providing this type of interaction support feature attribution, counterfactuals, and model behavior. These interactions are essential for the latter categories to support counterfactual reasoning [263] and verify hypotheses on the model behavior.

Finally, only 15% of systems support *interactive model observations*. Most of these systems employ feature attribution to provide explainability and target attention or prototype-based models. Among them, only a couple of works support steering ability and the possibility of monitoring the training process. Thus, better support for these tasks represents a promising future research direction.

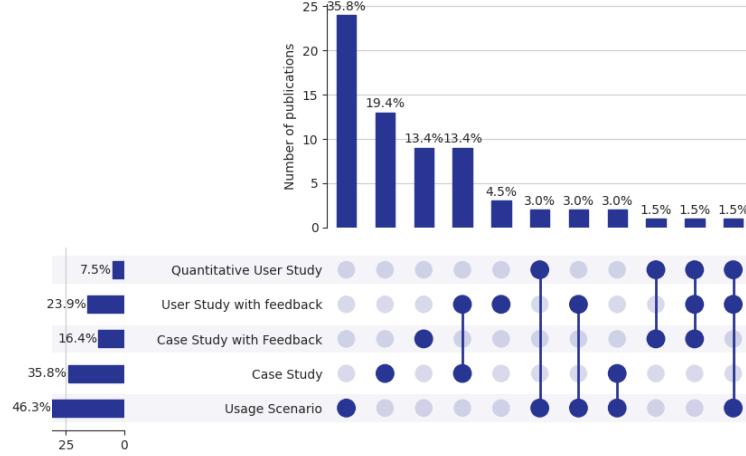
**Evaluation.** The evaluation of the efficacy of the analyzed systems adheres to the established procedures in the VA field. Specifically, these systems are usually evaluated based on user satisfaction.

Nearly half of the papers include a usage scenario, offering potential yet fictitious scenarios where their systems could prove beneficial (Figure 8.9). Since no participants or experts are involved in the process, the usage scenario can be considered proof-of-concept or entry points for further evaluation. Slightly more reliable evaluations are provided through case studies. They include a description of workflows and discoveries of real usage sessions carried out by a user in controlled settings. The case study and the usage scenario lack a quantitative and qualitative evaluation from the users.

The qualitative assessment is provided through case studies with feedback and user studies with feedback. These evaluation procedures are used in 40% of the analyzed papers. The case study with feedback follows the same workflow as the case study without them, but the user provides feedback



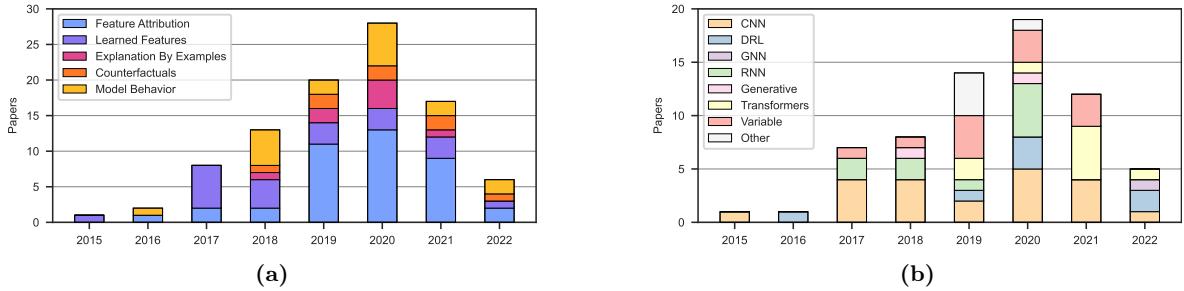
**Figure 8.8:** Distribution of interactions supported for the user.



**Figure 8.9:** Distribution of evaluation procedures used in the analyzed systems.

in the form of an interview at the end of the test session. The goal of this evaluation is to assess the generalization of the results over similar scenarios. Conversely, user studies with feedback relax the controlled settings and involve multiple testing sessions with different users, collecting their feedback. Finally, quantitative user studies provide quantitative evaluations, asking users to fill out questionnaires and collecting quantitative metrics during interactions with the system. Quantitative user studies represent the most general, reliable, and costly evaluation procedure. Unfortunately, less than 10% of the analyzed systems incorporate this evaluation process (Figure 8.9).

Overall, the evaluation practices align with the VA standards, which heavily rely on user satisfaction. In scenarios where these systems employ explanation methods from the XAI literature and apply them to standard settings, these evaluations can further validate the quality of such approaches. However, we observe a lack of a proper quantitative evaluation, in terms of user-independent metrics (Section 2.2), of explanation methods when these are novel and proposed



**Figure 8.10:** (a) Explanation categories and (b) models coverage over the years.

alongside the system and when XAI techniques are applied on settings different from their original ones.

**Support for State-of-the-Art models and methods.** Here, we analyze the diversity and coverage in VA systems of DNNs and explanation methods used to support the different categories. We observe that (Figure 8.10b) convolutional neural networks were the most supported architecture until 2018. Subsequently, other architectures, especially Transformers, gained popularity, following the trend in AI research.

Regarding the supported categories (Figure 8.10a), feature attribution emerged as the preferred method for enhancing interpretability in 2019, leading to a reduction in the percentage of learned features. Looking more in detail on the specific methods used to support the categories, we noted a large gap between the availability of approaches in XAI research and adoption in VA systems. Indeed, the most recent supported methods are SHAP [140], Grad-CAM [208], and LIME [197], which nowadays are considered baselines in the XAI field and have been outperformed by several other methods. Moreover, only five systems include these methods, while the others prefer even older methods like Deconvolution [279] and vanilla Saliency Maps [216]. These choices may limit the widespread of these systems and their adoption from DL experts beyond the specific cases for which these systems are built.

### 8.3 Towards a Better Integration

The previous section analyzed the current state-of-the-art VA systems, discussing their strengths and weaknesses. This section, building upon the previous analysis, outlines research directions and action items necessary to mitigate such issues and advance the field.

We argue that the weaknesses related to models and explanations coverage are connected to the lack of **involvement of DL and XAI** experts in the design process. In fact, analyzing the papers' authors, most of VA systems are solely designed by VA experts, with DL experts considered only as users. Few papers list XAI experts among the authors, and they are rarely involved as users of the systems. This behavior results in limited support for recent and more effective explanation methods and restricted coverage of DL models. Additionally, despite the availability of popular libraries for both XAI and DL, these systems rarely support their integration and typically favor one specific XAI method. Conversely, supporting and displaying explanations returned by different methods of the same class would enhance the reliability of such systems, considering the lack of ground truth and the ongoing debate about evaluation in XAI (Section 2.2).

Several actions could foster collaboration between these fields and mitigate these issues. Firstly, researchers should develop a **greater awareness** of each other’s work. The knowledge of weaknesses and strengths of each field can help developers build more effective VA systems. The awareness can be increased by hosting workshops and events in venues of both fields. While some conferences have hosted workshops discussing VA for explaining DL, these discussions have often been peripheral rather than central. Given the increasing importance of interactivity in XAI with the rise of chatbots and more sophisticated dialogue systems, specialized workshops focused on this topic could significantly enhance the success of the discussed systems.

A closer collaboration would make the support for novel models and explanation methods faster, aligning the state-of-the-art of both fields. As VA systems become more widespread, XAI research stands to benefit from their analytics tools, using them to test and refine their methods. Finally, as the field matures, the development of **common libraries** between VA, DL, and XAI could facilitate research by providing an interface that abstracts access to DL frameworks, XAI methods, and VA components. This would alleviate the difficulty researchers face when adapting or changing components of a system, such as datasets, models, and workflows, speeding up research in the field.

## 8.4 Contributions

This chapter reviewed and analyzed visual analytics systems employing XAI methods to explain DNNs. It presented XAI categories supported by these systems and introduced a novel category specific to these approaches: the model behavior category. The analysis highlighted the current support for different users, domains, models, and XAI methods, along with the evaluation procedures employed in papers proposing such systems. The chapter underscored the potentiality of VA solutions and the areas where future research could yield additional benefits. It advocates for increased collaboration between XAI, VA, and DL fields and the promotion of joint events to align research findings. Additionally, it highlighted drawbacks of current solutions, such as the lack of quantitative evaluation of proposed XAI methods and the limited presence of VA solutions tailored for end-users without DL knowledge.

# Part V

## Conclusions

# Chapter 9

## Conclusions

This final chapter concludes the thesis by summarizing its contributions (Section 9.1), discussing its limitations (Section 9.2), and outlining potential future research directions stemming from this work (Section 9.3).

### 9.1 Summary

This thesis contributed to the field of explainable Artificial Intelligence by focusing on intrinsic techniques to enhance the interpretability of deep neural networks. The overarching goal was to leverage elements already embedded within modern neural networks to probe their behavior, thereby enabling users to understand and trust the decisions made by these models. The underlying idea is to treat the layers preceding the classifier as a black-box feature extractor and focus on the last part to enhance interpretability. In particular, the thesis investigated methods that add elements after the feature extractor or make its representations more interpretable. These approaches address a common limitation of self-explainable deep neural networks: constrained generalization across different architectures and lower performance than black-box models.

The first part of the thesis introduced novel designs for self-explainable neural networks aimed at improving the interpretability of deep neural networks while preserving or enhancing their performance. Chapter 4 introduced Prototype-based Interpretable Graph Neural Networks to address the opacity of graph neural networks by integrating prototype layers between the feature extractor and the classifier. Through extensive experiments, we demonstrated that this design preserves the performance of the black-box models while enhancing their interpretability, particularly in the context of graph data and chemical domains.

Building upon the insights gained from Chapter 4, Chapter 5 introduced Self-Explainable Memory-Augmented Deep Neural Networks, which augmented black-box DNNs with memory modules. The memory facilitates the storage and utilization of crucial information during the decision process. The chapter proposed tracking mechanisms of writing and reading processes to provide interpretability. The proposed designs are model agnostic and, thus, can be applied to any neural network and preserve the performance of the black-box models. Moreover, these designs have been proven flexible in the types of explanations that can be derived, spanning from single and group feature attributions to explanations by example and counterfactuals.

The second part of this thesis aimed at making the interpretation of latent representations

easier. Chapter 6 explored Graph Concept Whitening, which aims to make the interpretation of the semantics encoded in the latent representations of graph neural networks easier. By modifying the normalization layers, this technique encourages the latent space to represent molecular properties on its axes. In this way, it provides a useful entry point for network inspection and interpretability in the chemical domain while preserving model performance.

Chapter 7 pushed forward the interpretation of latent representations by introducing Clustered Compositional Explanations, a method applicable to black-box DNNs without any modification of its structure for analyzing alignment between latent representations and a predefined set of concepts. By studying the full spectrum of neuron activations and employing clustering algorithms, this method provided novel insights into the behavior of neurons and improved the efficiency and quality of explanations for DNNs.

Lastly, the last part and Chapter 8 analyzed a promising direction for delivering explanations to users and allowing them to interact with models and explanations. Namely, Chapter 8 reviewed the emerging field of visual analytics, which is working on the problem of explaining deep learning to the user through interactive interfaces. Specifically, the chapter analyzed visual analytics systems employing XAI methods to explain DNNs, discussing the potentiality and steps needed for full integration and alignment between visual analytics and the XAI field.

In conclusion, this thesis has contributed significantly to the ongoing research on explainable deep learning. By exploring various techniques and methodologies, we have advanced the state-of-the-art of self-explainable neural networks, particularly in graph neural networks, memory-augmented neural networks, and explainable latent space analysis. These contributions enhance the interpretability of deep learning models and pave the way for more transparent and trustworthy AI systems in various domains.

## 9.2 Limitations

Despite the advancements made in this thesis, several open challenges related to the discussed approaches need further research and investigation.

Concerning prototype-based layers, a distinction from the vision domain is the difficulty in projecting the prototypes back to the input space. As the design relies on node embeddings, the interpretability of prototypes depends on the ratio between the number of layers preceding the prototype layers and the size of the relevant subgraphs. In the context of very deep neural networks, which is currently uncommon for graph neural networks, and relatively small graphs, each projection could cover big portions of the graph, thereby undermining the design principles of this architecture type. This aspect, along with the difficulty in choosing the correct number of prototypes and the induced bottleneck of prototype layers, represent the main limitations to be addressed in the future.

Conversely, memory-based architectures avoid the bottleneck problem by employing skip connections. However, these connections also pose a limitation to such methods as they enable the network to ignore the interpretable elements, thereby constraining the interpretability assurances of these designs. Striking a balance between skip connections and the induced bottlenecks remains an open problem to be tackled. Additionally, these architectures require enhancements in terms of memory footprint and computational time needed for training to compete with black-box models in resource-limited scenarios.

Probing processes and concept-based approaches rely on the availability of concept datasets [17, 125, 163] or external multi-modal models [84, 173]. These dependencies introduce bias in the process that can compromise the interpretability of these methods. Probing processes may overlook factors captured by the networks that are not aligned with human knowledge. Concept-based approaches are susceptible to concept leakage and noise from factors correlated with the presence of a concept (e.g., background).

While the thesis contributes to the generalization of eXplainable Artificial Intelligence (XAI) methods across architectures and domains, it primarily focuses on classification problems. Most of the proposed techniques do not make assumptions about the problem's structure. However, it is not guaranteed that their extension to other types of problems is straightforward, thus representing an open question.

### 9.3 Future Research Directions

Moving forward, the insights and methodologies presented in this thesis lay a solid foundation for future research in explainable deep learning. In this section, we discuss potential future research directions and areas for further development of the proposed techniques that can mitigate the limitations of current approaches.

This thesis discussed two designs of self-explainable memory-augmented neural networks and how the memory tracking mechanism is used to compute explanations from the proposed designs. In this direction, we foresee a unified framework to apply the same concept to any memory-augmented neural network, studying the impact of different solutions and laying the foundation for this research area. Other research areas related to memory-based architectures include reduction mechanisms for the large footprint and the adaptation strategy for the training process of the black-box models. In this context, integrating the proposed techniques in a visual analytics system could open the door to a human-centered evaluation. This evaluation could assess the effectiveness and usability of the proposed methods in real-world scenarios, providing valuable insights into their practical utility and informing further improvements based on user feedback.

Regarding neural explanations, most dataset-based techniques are based on the beam search algorithm, which does not guarantee the optimality and completeness of the explanations. A research direction could examine the effectiveness of current approaches and identify the challenges connected to the search for optimality. In particular, by leveraging state-of-the-art clustering algorithms and exploring novel approaches for identifying meaningful clusters in high-dimensional activation spaces, we could improve the granularity and interpretability of explanations provided by this method. Additionally, future research could work on methods that mitigate biases and ensure the reliability and generalization of the probing process, for example, investigating the application of methods rooted in pattern recognition.

Finally, looking at the long-term vision for the research on self-explainable deep neural networks, the eXplainable Artificial Intelligence (XAI) field needs a unified framework that can be applied and modified easily to generate different types of explanations for several problems. Such a design could have an impact similar to that of Transformers for the deep learning field, applied nowadays to multiple different problems and domains, and be able to push research beyond its limits.

By following the research directions described above, we can continue to advance the field of

explainable deep learning and develop more transparent, interpretable, and trustworthy AI systems.

# Bibliography

- [1] Uncovering the hidden significance of activities location in predictive process monitoring. In *Pre-proceedings of the ML4PM 2023 Fourth International Workshop On Leveraging Machine Learning In Process Mining*, pages 1–12, November 2023. 4th International Workshop on Leveraging Machine Learning in Process Mining, ML4PM 2023, ML4PM 2023 ; Conference date: 23-10-2023 Through 27-10-2023.
- [2] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.385.
- [3] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. doi: 10.1109/access.2018.2870052.
- [4] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [5] Ajaya Adhikari, David M. J. Tax, Riccardo Satta, and Matthias Faeth. Leafage: Example-based and feature importance-based explanations for black-box ml models. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, June 2019. doi: 10.1109/fuzz-ieee.2019.8858846.
- [6] A.S. Albahri, Ali M. Duham, Mohammed A. Fadhel, Alhamzah Alnoor, Noor S. Baqer, Laith Alzubaidi, O.S. Albahri, A.H. Alamoodi, Jinshuai Bai, Asma Salhi, Jose Santamaría, Chun Ouyang, Ashish Gupta, Yuantong Gu, and Muhammet Deveci. A systematic review of trustworthy and explainable artificial intelligence in healthcare: Assessment of quality, bias risk, and data fusion. *Information Fusion*, 96:156–191, August 2023. ISSN 1566-2535. doi: 10.1016/j.inffus.2023.03.008.
- [7] Gulsum Alicioglu and Bo Sun. A survey of visual analytics for explainable artificial intelligence methods. *Computers & Graphics*, 2021. ISSN 0097-8493. doi: <https://doi.org/10.1016/j.cag.2021.09.002>.
- [8] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31.

- Curran Associates, Inc., 2018. URL <https://proceedings.neurips.cc/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf>.
- [9] Omer Antverg and Yonatan Belinkov. On the pitfalls of analyzing individual neurons in language models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=8uz0EWPQIMu>.
  - [10] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, jun 2020. doi: 10.1016/j.inffus.2019.12.012.
  - [11] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):e0130140, jul 2015. doi: 10.1371/journal.pone.0130140.
  - [12] Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. January 2015. 3rd International Conference on Learning Representations, ICLR 2015;
  - [13] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL <http://arxiv.org/abs/1409.0473>.
  - [14] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 528–539. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/bahng20a.html>.
  - [15] Bram Bakker. Reinforcement learning with long short-term memory. In T. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001. URL [https://proceedings.neurips.cc/paper\\_files/paper/2001/file/a38b16173474ba8b1a95bc30d3b8a5-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2001/file/a38b16173474ba8b1a95bc30d3b8a5-Paper.pdf).
  - [16] Alejandro Barredo-Arrieta and Javier Del Ser. Plausible counterfactuals: Auditing deep learning classifiers with realistic adversarial examples. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2020. doi: 10.1109/ijcnn48605.2020.9206728.
  - [17] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

- [18] Franziska Becker, Arthur Drichel, Christoph Muller, and Thomas Ertl. Interpretable visualizations of deep neural networks for domain generation algorithm detection. In *2020 IEEE Symposium on Visualization for Cyber Security (VizSec)*. IEEE, oct 2020. doi: 10.1109/vizsec51108.2020.00010.
- [19] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. April 2020. doi: 10.48550/ARXIV.2004.05150.
- [20] Richard Bickerton, Gaia Paolini, Jérémie Besnard, Sorel Muresan, and Andrew Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4:90–8, 02 2012. doi: 10.1038/nchem.1243.
- [21] Alsallakh Bilal, Amin Jourabloo, Mao Ye, Xiaoming Liu, and Liu Ren. Do convolutional neural networks learn class hierarchy? *IEEE Transactions on Visualization and Computer Graphics*, 24(1):152–162, jan 2018. doi: 10.1109/tvcg.2017.2744683.
- [22] Judy Borowski, Roland Simon Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain {cnn} activations better than state-of-the-art feature visualization. In *International Conference on Learning Representations*, 2021. URL <https://openreview.net/forum?id=Q09-y8also->.
- [23] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Connelly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.
- [24] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf).
- [25] Kirill Bykov, Anna Hedström, Shinichi Nakajima, and Marina M.-C. Höhne. Noisegrad — enhancing explanations by introducing stochasticity to model weights. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6):6132–6140, June 2022. ISSN 2159-5399. doi: 10.1609/aaai.v36i6.20561.
- [26] Kirill Bykov, Laura Kopf, Shinichi Nakajima, Marius Kloft, and Marina MC Höhne. Labeling neural representations with inverse recognition. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=gLfgyIWiWW>.

- [27] Stephen Casper, Tilman Rauker, Anson Ho, and Dylan Hadfield-Menell. Sok: Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. In *First IEEE Conference on Secure and Trustworthy Machine Learning*, 2023. URL <https://openreview.net/forum?id=8C5zt-0Utdn>.
- [28] Junghoon Chae, Shang Gao, Arvind Ramanathan, Chad A. Steed, and Georgia Tourassi. Visualization for classification in deep neural networks. Technical report, Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States), 10 2017. URL <https://www.osti.gov/biblio/1407764>.
- [29] Gromit Yeuk-Yin Chan, Enrico Bertini, Luis Gustavo Nonato, Brian Barr, and Claudio T. Silva. Melody: Generating and visualizing machine learning model summary to understand data and classifiers together. *arXiv preprint arXiv:2007.10609*, July 2020.
- [30] Aditya Chattpadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, March 2018. doi: 10.1109/wacv.2018.00097.
- [31] Piyush Chawla, Subhashis Hazarika, and Han-Wei Shen. Token-wise sentiment decomposition for ConvNet: Visualizing a sentiment classifier. *Visual Informatics*, 4(2):132–141, jun 2020. doi: 10.1016/j.visinf.2020.04.006.
- [32] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42(4):1–10, July 2023. ISSN 1557-7368. doi: 10.1145/3592116.
- [33] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: Deep learning for interpretable image recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/adf7ee2dcf142b0e11888e72b43fc75-Paper.pdf>.
- [34] Peipei Chen, Wei Dong, Jinliang Wang, Xudong Lu, Uzay Kaymak, and Zhengxing Huang. Interpretable clinical prediction via attention-based neural network. *BMC Medical Informatics and Decision Making*, 20(S3), July 2020. ISSN 1472-6947. doi: 10.1186/s12911-020-1110-7.
- [35] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, December 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00265-z.
- [36] Furui Cheng, Yao Ming, and Huamin Qu. DECE: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1438–1447, feb 2021. doi: 10.1109/tvcg.2020.3030342.
- [37] Jaegul Choo and Shixia Liu. Visual analytics for explainable deep learning. *IEEE Computer Graphics and Applications*, 38(4):84–92, jul 2018. doi: 10.1109/mcg.2018.042731661.

- [38] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does bert look at? an analysis of bert’s attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/w19-4828.
- [39] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, jan 1967. doi: 10.1109/tit.1967.1053964.
- [40] Enyan Dai and Suhang Wang. Towards self-explainable graph neural network. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM ’21, page 302–311, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482306. URL <https://doi.org/10.1145/3459637.3482306>.
- [41] Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James Glass. What is one grain of sand in the desert? analyzing individual neurons in deep NLP models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6309–6317, jul 2019. doi: 10.1609/aaai.v33i01.33016309.
- [42] Luke N. Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J. Storkey. Cinic-10 is not imagenet or cifar-10. Technical report, University of Edinburgh, 2018. [dataset].
- [43] Nilaksh Das, Haekyu Park, Zijie J. Wang, Fred Hohman, Robert Firstman, Emily Rogers, and Duen Horng Polo Chau. Bluff: Interactively deciphering adversarial attacks on deep neural networks. In *2020 IEEE Visualization Conference (VIS)*. IEEE, oct 2020. doi: 10.1109/vis47514.2020.00061.
- [44] Asim Kumar Debnath, Rosa L. Lopez de Compadre, Gargi Debnath, Alan J. Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of Medicinal Chemistry*, 34(2):786–797, Feb 1991. ISSN 0022-2623. doi: 10.1021/jm00106a046. URL <https://doi.org/10.1021/jm00106a046>.
- [45] Joseph F. DeRose, Jiayao Wang, and Matthew Berger. Attention flows: Analyzing and comparing attention mechanisms in language models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1160–1170, feb 2021. doi: 10.1109/tvcg.2020.3028976.
- [46] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL [https://proceedings.neurips.cc/paper\\_files/paper/2018/file/c5ff2543b53f4cc0ad3819a36752467b-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2018/file/c5ff2543b53f4cc0ad3819a36752467b-Paper.pdf).
- [47] Zhihang Dong, Tongshuang Wu, Sicheng Song, and Mingrui Zhang. Interactive attention model explorer for natural language processing tasks with unbalanced data sizes. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, jun 2020. doi: 10.1109/pacificvis48177.2020.1031.

- [48] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. February 2017. doi: 10.48550/ARXIV.1702.08608.
- [49] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.
- [50] John C. Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. In Adam Tauman Kalai and Mehryar Mohri, editors, *COLT 2010 - The 23rd Conference on Learning Theory, Haifa, Israel, June 27-29, 2010*, pages 257–269. Omnipress, 2010.
- [51] Nadir Durrani, Hassan Sajjad, Fahim Dalvi, and Yonatan Belinkov. Analyzing individual neurons in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.395.
- [52] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022.
- [53] Alex Endert, Patrick Fiaux, and Chris North. Semantic interaction for sensemaking: Inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2879–2888, 2012. doi: 10.1109/TVCG.2012.260.
- [54] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.
- [55] Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 21400–21413. Curran Associates, Inc., 2022. URL [https://proceedings.neurips.cc/paper\\_files/paper/2022/file/867c06823281e506e8059f5c13a57f75-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2022/file/867c06823281e506e8059f5c13a57f75-Paper-Conference.pdf).
- [56] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88(2):303–338, sep 2009. doi: 10.1007/s11263-009-0275-4.
- [57] Zhengqing Fang, Kun Kuang, Yuxiao Lin, Fei Wu, and Yu-Feng Yao. Concept-based explanation for fine-grained images and its application in infectious keratitis classification. In *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, October 2020. doi: 10.1145/3394171.3413557.

- [58] Thomas FEL, Thibaut Boissin, Victor Boutin, Agustin Martin Picard, Paul Novello, Julien Colin, Drew Linsley, Tom ROUSSEAU, Remi Cadene, Lore Goetschalckx, Laurent Gardes, and Thomas Serre. Unlocking feature visualization for deep network with MAgnitude constrained optimization. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=J7VoDuzuKs>.
- [59] Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [60] Jörg Franke, Jan Niehues, and Alex Waibel. Robust and scalable differentiable neural computer for question answering. In *Proceedings of the Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/w18-2606.
- [61] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, July 2017. doi: 10.1109/cvpr.2017.476.
- [62] Hongyang Gao and Shuiwang Ji. Graph u-nets. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2083–2092. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/gao19a.html>.
- [63] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pages 9273–9282, 2019.
- [64] C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. Citeseer: an automatic citation indexing system. In *Proceedings of the third ACM conference on Digital libraries - DL '98*, DL '98. ACM Press, 1998. doi: 10.1145/276675.276685.
- [65] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1263–1272. PMLR, 06–11 Aug 2017. URL <https://proceedings.mlr.press/v70/gilmer17a.html>.
- [66] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, October 2018. doi: 10.1109/dsaa.2018.00018.
- [67] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2376–2384. PMLR, 09–15 Jun 2019. URL <https://proceedings.mlr.press/v97/goyal19a.html>.

- [68] Isel Grau, Gonzalo Nápoles, Marilyn Bello, Yamisleydi Salgueiro, and Agnieszka Jastrzebska. Forward composition propagation for explainable neural reasoning. *IEEE Computational Intelligence Magazine*, 19(1):26–35, February 2024. ISSN 1556-6048. doi: 10.1109/mci.2023.3327834.
- [69] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, May 2013. doi: 10.1109/icassp.2013.6638947.
- [70] Alex Graves, Greg Wayne, Malcolm Reynolds, Tim Harley, Ivo Danihelka, Agnieszka Grabska-Barwińska, Sergio Gómez Colmenarejo, Edward Grefenstette, Tiago Ramalho, John Agapiou, Adrià Puigdomènech Badia, Karl Moritz Hermann, Yori Zwols, Georg Ostrovski, Adam Cain, Helen King, Christopher Summerfield, Phil Blunsom, Koray Kavukcuoglu, and Demis Hassabis. Hybrid computing using a neural network with dynamic external memory. *Nature*, 538(7626):471–476, oct 2016. doi: 10.1038/nature20101.
- [71] Shurui Gui, Hao Yuan, Jie Wang, Qicheng Lao, Kang Li, and Shuiwang Ji. Flowx: Towards explainable graph neural networks via message flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–12, 2023. ISSN 1939-3539. doi: 10.1109/tpami.2023.3347470.
- [72] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, April 2022. ISSN 1573-756X. doi: 10.1007/s10618-022-00831-6.
- [73] David Gunning and David W. Aha. Darpa’s explainable artificial intelligence program. *AI Magazine*, 40(2):44–58, June 2019. ISSN 2371-9621. doi: 10.1609/aimag.v40i2.2850.
- [74] Karthik S. Gurumoorthy, Amit Dhurandhar, Guillermo Cecchi, and Charu Aggarwal. Efficient data representation by selecting prototypes with importance weights. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, nov 2019. doi: 10.1109/icdm.2019.00036.
- [75] Xiaoyang Han, Han-Wei Shen, Guan Li, Xuyi Lu, Guihua Shan, and Yangang Wang. IVDAS: an interactive visual design and analysis system for image data symmetry detection of CNN models. *Journal of Visualization*, 24(3):615–629, jan 2021. doi: 10.1007/s12650-020-00721-3.
- [76] Rafael Harth. Understanding individual neurons of ResNet through improved compositional formulas. In *Pattern Recognition and Artificial Intelligence*, pages 283–294. Springer International Publishing, 2022. doi: 10.1007/978-3-031-09282-4\_24.
- [77] Peter Hase, Chaofan Chen, Oscar Li, and Cynthia Rudin. Interpretable image recognition with hierarchical prototypes. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7:32–40, October 2019. ISSN 2769-1330. doi: 10.1609/hcomp.v7i1.5265.
- [78] Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL [https://openreview.net/forum?id=tglniD\\_fn9](https://openreview.net/forum?id=tglniD_fn9).

- [79] Subhashis Hazarika, Haoyu Li, Ko-Chih Wang, Han-Wei Shen, and Ching-Shan Chou. NNVA: Neural network assisted visual analysis of yeast cell polarization simulation. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2019. doi: 10.1109/tvcg.2019.2934591.
- [80] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. IEEE, 2016.
- [81] Wenbin He, Teng-Yok Lee, Jeroen van Baar, Kent Wittenburg, and Han-Wei Shen. DynamicsExplorer: Visual analytics for robot control tasks involving dynamics and LSTM-based control policies. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, jun 2020. doi: 10.1109/pacificvis48177.2020.7127.
- [82] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, and B. Scholkopf. Support vector machines. *IEEE Intelligent Systems and their Applications*, 13(4):18–28, July 1998. ISSN 1094-7167. doi: 10.1109/5254.708428.
- [83] Lucas Torroba Hennigen, Adina Williams, and Ryan Cotterell. Intrinsic probing through dimension selection. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.15.
- [84] Evan Hernandez, Sarah Schwettmann, David Bau, Teona Bagashvili, Antonio Torralba, and Jacob Andreas. Natural language descriptions of deep features. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=NudBMY-tzDr>.
- [85] Jacob Hilton, Nick Cammarata, Shan Carter, Gabriel Goh, and Chris Olah. Understanding RL vision. *Distill*, 5(11), nov 2020. doi: 10.23915/distill.00029.
- [86] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, November 1997. ISSN 1530-888X. doi: 10.1162/neco.1997.9.8.1735.
- [87] Marius Hogräfer, Marco Angelini, Giuseppe Santucci, and Hans-Jörg Schulz. Steering-by-example for progressive visual analytics. *ACM Trans. Intell. Syst. Technol.*, 13(6), 9 2022. ISSN 2157-6904. doi: 10.1145/3531229. URL <https://doi.org/10.1145/3531229>.
- [88] Fred Hohman, Nathan Hodas, and Duen Horng Chau. ShapeShop. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, may 2017. doi: 10.1145/3027063.3053103.
- [89] Fred Hohman, Minsuk Kahng, Robert Pienta, and Duen Horng Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2674–2693, aug 2019. doi: 10.1109/tvcg.2018.2843369.
- [90] Fred Hohman, Haekyu Park, Caleb Robinson, and Duen Horng Polo Chau. Summit: Scaling deep learning interpretability by visualizing activation and attribution summarizations. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):1096–1106, jan 2020. doi: 10.1109/tvcg.2019.2934659.

- [91] Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. exBERT: A visual analysis tool to explore learned representations in transformer models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-demos.22.
- [92] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4700–4708. IEEE, 2017.
- [93] Lei Huang, Xianglong Liu, Bo Lang, Adams Wei Yu, Yongliang Wang, and Bo Li. Orthogonal weight normalization: Solution to optimization over multiple dependent stiefel manifolds in deep neural networks. In *AAAI*, pages 3271–3278, 2018.
- [94] Xinyi Huang, Suphanut Jamonnak, Ye Zhao, Tsung Heng Wu, and Wei Xu. A visual designer of layer-wise relevance propagation models. *Computer Graphics Forum*, 40(3):227–238, jun 2021. doi: 10.1111/cgf.14302.
- [95] T. Jaunet, R. Vuillemot, and C. Wolf. Drlviz: Understanding decisions and memory in deep reinforcement learning. *Computer Graphics Forum*, 39(3):49–61, 2020. doi: <https://doi.org/10.1111/cgf.13962>.
- [96] Théo Jaunet, Corentin Kervadec, Romain Vuillemot, Grigory Antipov, Moez Baccouche, and Christian Wolf. Visqa: X-raying vision and language reasoning in transformers. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):976–986, 2022. doi: 10.1109/TVCG.2021.3114683.
- [97] Xiaonan Ji, Yamei Tu, Wenbin He, Junpeng Wang, Han-Wei Shen, and Po-Yin Yen. USE-Vis: Visual analytics of attention-based neural embedding in information retrieval. *Visual Informatics*, 5(2):1–12, jun 2021. doi: 10.1016/j.visinf.2021.03.003.
- [98] Shichao Jia, Peiwen Lin, Zeyu Li, Jiawan Zhang, and Shixia Liu. Visualizing surrogate decision trees of convolutional neural networks. *Journal of Visualization*, 23(1):141–156, nov 2019. doi: 10.1007/s12650-019-00607-z.
- [99] Weina Jin, Xiaoxiao Li, and Ghassan Hamarneh. The xai alignment problem: Rethinking how should we evaluate human-centered ai explainability techniques. March 2023. doi: 10.48550/ARXIV.2303.17707.
- [100] Zhihua Jin, Yong Wang, Qianwen Wang, Yao Ming, Tengfei Ma, and Huamin Qu. Gnnlens: A visual analytics approach for prediction error diagnosis of graph neural networks. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2022. doi: 10.1109/TVCG.2022.3148107.
- [101] Hong-Gyu Jung, Sin-Han Kang, Hee-Dong Kim, Dong-Ok Won, and Seong-Whan Lee. Counterfactual explanation based on gradual construction for deep networks. *Pattern Recognition*, 132:108958, December 2022. ISSN 0031-3203. doi: 10.1016/j.patcog.2022.108958.

- [102] Minsuk Kahng, Pierre Y. Andrews, Aditya Kalro, and Duen Horng Chau. ActiVis: Visual exploration of industry-scale deep neural network models. *IEEE transactions on visualization and computer graphics*, 24(1):88–97, jan 2018. doi: 10.1109/tvcg.2017.2744718.
- [103] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: an adaptive path method for removing noise. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021. doi: 10.1109/cvpr46437.2021.00501.
- [104] Daniel Keim, Gennady Andrienko, Jean-Daniel Fekete, Carsten Görg, Jörn Kohlhammer, and Guy Melançon. *Visual Analytics: Definition, Process, and Challenges*, pages 154–175. Springer Berlin Heidelberg. ISBN 9783540709565. doi: 10.1007/978-3-540-70956-5\_7.
- [105] Eoin M. Kenny and Mark T. Keane. Twin-systems to explain artificial neural networks using case-based reasoning: Comparative tests of feature-weighting methods in ANN-CBR twins for XAI. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pages 2708–2715. International Joint Conferences on Artificial Intelligence Organization, 2019.
- [106] Eoin M. Kenny and Mark T. Keane. Explaining deep learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI. *Knowledge-Based Systems*, 233:107530, dec 2021. doi: 10.1016/j.knosys.2021.107530.
- [107] Eoin M. Kenny and Mark T Keane. On generating plausible counterfactual and semi-factual explanations for deep learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11575–11585, May 2021. ISSN 2159-5399. doi: 10.1609/aaai.v35i13.17377.
- [108] Eoin M. Kenny, Eoin Delaney, and Mark T. Keane. Advancing post-hoc case-based explanation with feature highlighting. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, IJCAI-2023. International Joint Conferences on Artificial Intelligence Organization, August 2023. doi: 10.24963/ijcai.2023/48.
- [109] Eoin M. Kenny, Mycal Tucker, and Julie Shah. Towards interpretable deep reinforcement learning with human-friendly prototypes. In *The Eleventh International Conference on Learning Representations*, 2023. URL [https://openreview.net/forum?id=hWwY\\_Jq0xsN](https://openreview.net/forum?id=hWwY_Jq0xsN).
- [110] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
- [111] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viagas, and Rory sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2668–2677. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/kim18d.html>.

- [112] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.
- [113] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations (ICLR)*, 2017.
- [114] Rudolf Kiralj and Márcia M.C. Ferreira. A priori molecular descriptors in qsar: a case of hiv-1 protease inhibitors: I. the chemometric approach. *Journal of Molecular Graphics and Modelling*, 21(5):435–448, 2003. ISSN 1093-3263. doi: [https://doi.org/10.1016/S1093-3263\(02\)00201-2](https://doi.org/10.1016/S1093-3263(02)00201-2). URL <https://www.sciencedirect.com/science/article/pii/S1093326302002012>.
- [115] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/koh20a.html>.
- [116] Jörn Kohlhammer, Daniel Keim, Margit Pohl, Giuseppe Santucci, and Gennady Andrienko. Solving problems with visual analytics. *Procedia Computer Science*, 7:117–120, 2011. ISSN 1877-0509. doi: <https://doi.org/10.1016/j.procs.2011.12.035>. Proceedings of the 2nd European Future Technologies Conference and Exhibition 2011 (FET 11).
- [117] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. [dataset].
- [118] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL [https://proceedings.neurips.cc/paper\\_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf).
- [119] Jacek Kujawski, Hanna Popielarska, Anna Myka, Beata Drabińska, and Marek Bernard. The log p parameter as a molecular descriptor in the computer-aided drug design – an overview. *Computational Methods in Science and Technology*, 18:81–88, 08 2012. doi: 10.12921/cmst.2012.18.02.81-88.
- [120] B. C. Kwon, M. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):299–309, Jan 2019. ISSN 1941-0506. doi: 10.1109/TVCG.2018.2865027.
- [121] B. La Rosa, G. Blasilli, R. Bourqui, D. Auber, G. Santucci, R. Capobianco, E. Bertini, R. Giot, and M. Angelini. State of the art of visual analytics for explainable deep learning. *Computer Graphics Forum*, 42(1):319–355, February 2023. ISSN 1467-8659. doi: 10.1111/cgf.14733.

- [122] Biagio La Rosa, Roberto Capobianco, and Daniele Nardi. Explainable inference on sequential data via memory-tracking. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pages 2006–2013. International Joint Conferences on Artificial Intelligence Organization, 2020.
- [123] Biagio La Rosa, Roberto Capobianco, and Daniele Nardi. A discussion about explainable inference on sequential data via memory-tracking. In *AIxIA 2021 Discussion Papers*, volume 3078, pages 33–44. CEUR Workshop Proceedings, 2021. URL <http://ceur-ws.org/Vol-3078/paper-24.pdf>.
- [124] Biagio La Rosa, Roberto Capobianco, and Daniele Nardi. A self-interpretable module for deep image classification on small data. *Applied Intelligence*, aug 2022. doi: 10.1007/s10489-022-03886-6.
- [125] Biagio La Rosa, Leilani H. Gilpin, and Roberto Capobianco. Towards a fuller understanding of neurons with clustered compositional explanations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=51PLYhMFWz>.
- [126] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1), March 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-08987-4.
- [127] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. Comparison-based inverse classification for interpretability in machine learning. In *Communications in Computer and Information Science*, pages 100–111. Springer International Publishing, 2018. doi: 10.1007/978-3-319-91473-2\_9.
- [128] Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016. URL <http://jmlr.org/papers/v17/15-522.html>.
- [129] Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara. Scouter: Slot attention-based classifier for explainable image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1046–1055, October 2021.
- [130] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.
- [131] Christopher A. Lipinski, Franco Lombardo, Beryl William Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 46 1-3:3–26, 2001.

- [132] Hui Liu, Qingyu Yin, and William Yang Wang. Towards explainable nlp: A generative explanation framework for text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, 2019.
- [133] M. Liu, J. Shi, Z. Li, C. Li, J. Zhu, and S. Liu. Towards better analysis of deep convolutional neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):91–100, Jan 2017. ISSN 1941-0506. doi: 10.1109/TVCG.2016.2598831.
- [134] Mengchen Liu, Shixia Liu, Hang Su, Kelei Cao, and Jun Zhu. Analyzing the noise robustness of deep neural networks. In *2018 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, oct 2018. doi: 10.1109/vast.2018.8802509.
- [135] Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative counterfactual introspection for explainable deep learning. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 1–5, 2019.
- [136] Shusen Liu, Zhimin Li, Tao Li, Vivek Srikumar, Valerio Pascucci, and Peer-Timo Bremer. NLIZE: A perturbation-driven visual interrogation tool for analyzing and interpreting natural language inference models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):651–660, jan 2019. doi: 10.1109/tvcg.2018.2865230.
- [137] Yang Liu, Sujay Khandagale, Colin White, and Willie Neiswanger. Synthetic benchmarks for scientific research in explainable machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL <https://openreview.net/forum?id=R7vr14ffhF9>.
- [138] Arnaud Van Looveren and Janis Klaise. Interpretable counterfactual explanations guided by prototypes. pages 650–665. Springer International Publishing, 2021. doi: 10.1007/978-3-030-86520-7\_40.
- [139] Max Maria Losch, Mario Fritz, and Bernt Schiele. *Semantic Bottlenecks: Quantifying and Improving Inspectability of Deep Representations*, pages 15–29. Springer International Publishing, 2021. ISBN 9783030712785. doi: 10.1007/978-3-030-71278-5\_2.
- [140] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [141] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. *Advances in Neural Information Processing Systems*, 33, 2020.
- [142] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2015. doi: 10.18653/v1/d15-1166.

- [143] Yuxin Ma, Arlen Fan, Jingrui He, Arun Reddy Nelakurthi, and Ross Maciejewski. A visual analytics framework for explaining and diagnosing transfer learning processes. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1385–1395, feb 2021. doi: 10.1109/tvcg.2020.3028888.
- [144] Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.
- [145] Sayo M. Makinwa, Biagio La Rosa, and Roberto Capobianco. Detection accuracy for evaluating compositional explanations of units. In *AIxIA 2021 - Advances in Artificial Intelligence*, pages 550–563. Springer International Publishing, 2022. doi: 10.1007/978-3-031-08421-8\_38.
- [146] Chaitanya Malaviya, Pedro Ferreira, and André F. T. Martins. Sparse and constrained attention for neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2018. doi: 10.18653/v1/p18-2059.
- [147] Andre Martins and Ramon Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1614–1623, New York, New York, USA, 20–22 Jun 2016. PMLR. URL <https://proceedings.mlr.press/v48/martins16.html>.
- [148] André F. T. Martins and Ramón F. Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML’16, page 1614–1623. JMLR.org, 2016.
- [149] Ines Filipa Martins, Ana L. Teixeira, Luis Pinheiro, and Andre O. Falcao. A bayesian approach to in silico blood-brain barrier penetration modeling. *Journal of Chemical Information and Modeling*, 52(6):1686–1697, jun 2012. doi: 10.1021/ci300124c. URL <https://doi.org/10.1021%2Fci300124c>.
- [150] Riccardo Massidda and Davide Bacciu. Knowledge-driven interpretation of convolutional neural networks. In *Machine Learning and Knowledge Discovery in Databases*, pages 356–371. Springer International Publishing, 2023. doi: 10.1007/978-3-031-26387-3\_22.
- [151] Andrew Kachites McCallum, Kamal Nigam, Jason Rennie, and Kristie Seymore. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2):127–163, 2000. ISSN 1386-4564. doi: 10.1023/a:1009953814988.
- [152] Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. Mass-editing memory in a transformer. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=MkbcAHIYgyS>.
- [153] Todor Mihaylov and Anette Frank. Story cloze ending selection baselines and data examination. In Michael Roth, Nasrin Mostafazadeh, Nathanael Chambers, and Annie Louis, editors,

- Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 87–92, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-0913. URL <https://aclanthology.org/W17-0913>.
- [154] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.
  - [155] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, February 2019. ISSN 0004-3702. doi: 10.1016/j.artint.2018.07.007.
  - [156] Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. Interpretable and steerable sequence learning via prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD ’19. ACM, July 2019. doi: 10.1145/3292500.3330908.
  - [157] Yao Ming, Panpan Xu, Furui Cheng, Huamin Qu, and Liu Ren. ProtoSteer: Steering deep sequence model with prototypes. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):238–248, jan 2020. doi: 10.1109/tvcg.2019.2934267.
  - [158] A. Mishra, U. Soni, J. Huang, and C. Bryan. Why? why not? when? visual explanations of agent behaviour in reinforcement learning. In *2022 IEEE 15th Pacific Visualization Symposium (PacificVis)*, pages 111–120, Los Alamitos, CA, USA, apr 2022. IEEE Computer Society. doi: 10.1109/PacificVis53943.2022.00020. URL <https://doi.ieee.org/10.1109/PacificVis53943.2022.00020>.
  - [159] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 1476-4687. doi: 10.1038/nature14236.
  - [160] Sanaz Mohammadjafari, Mucahit Çevik, Mathusan Thanabalasingam, Ayse Basar, and Alzheimer’s Initiative. Using protopnet for interpretable alzheimer’s disease classification. *Proceedings of the Canadian Conference on Artificial Intelligence*, 06 2021. doi: 10.21428/594757db.fb59ce6c.
  - [161] Alexander Mordvintsev, Nicola Pezzotti, Ludwig Schubert, and Chris Olah. Differentiable image parameterizations. *Distill*, 3(7), jul 2018. doi: 10.23915/distill.00012.
  - [162] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and cloze evaluation for deeper understanding of commonsense stories. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego,

- California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1098. URL <https://aclanthology.org/N16-1098>.
- [163] Jesse Mu and Jacob Andreas. Compositional explanations of neurons. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17153–17163. Curran Associates, Inc., 2020. URL <https://proceedings.neurips.cc/paper/2020/file/c74956fffb38ba48ed6ce977af6727275-Paper.pdf>.
- [164] Seil Na, Yo Joong Choe, Dong-Hyun Lee, and Gunhee Kim. Discovery of Natural Language Concepts in Individual Units of CNNs. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=S1EERs09YQ>.
- [165] Karthikeyan Natesan Ramamurthy, Bhanukiran Vinzamuri, Yunfeng Zhang, and Amit Dhurandhar. Model agnostic multilevel explanations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5968–5979. Curran Associates, Inc., 2020. URL [https://proceedings.neurips.cc/paper\\_files/paper/2020/file/426f990b332ef8193a61cc90516c1245-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2020/file/426f990b332ef8193a61cc90516c1245-Paper.pdf).
- [166] Meike Nauta, Ron van Bree, and Christin Seifert. Neural prototype trees for interpretable fine-grained image recognition. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021. doi: 10.1109/cvpr46437.2021.01469.
- [167] Meike Nauta, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2023. doi: 10.1109/cvpr52729.2023.00269.
- [168] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- [169] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL [https://proceedings.neurips.cc/paper\\_files/paper/2016/file/5d79099fcdf499f12b79770834c0164a-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2016/file/5d79099fcdf499f12b79770834c0164a-Paper.pdf).
- [170] Anh Nguyen, Jason Yosinski, and Jeff Clune. Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. February 2016. doi: 10.48550/ARXIV.1602.03616.
- [171] Anh Nguyen, Jeff Clune, Yoshua Bengio, Alexey Dosovitskiy, and Jason Yosinski. Plug & play generative networks: Conditional iterative generation of images in latent space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

- [172] Conor Nugent and Pádraig Cunningham. A case-based explanation system for black-box systems. *Artificial Intelligence Review*, 24(2):163–178, oct 2005. doi: 10.1007/s10462-005-4609-5.
- [173] Tuomas Oikarinen and Tsui-Wei Weng. CLIP-dissect: Automatic description of neuron representations in deep vision networks. In *International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=iPWiWWhc1V>.
- [174] Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=F1Cg47MNvBA>.
- [175] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11), nov 2017.
- [176] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3), March 2020. ISSN 2476-0757. doi: 10.23915/distill.00024.001.
- [177] Mohsen Abbaspour Onari, Isel Grau, Marco S. Nobile, and Yingqian Zhang. Trustworthy artificial intelligence in medical applications: A mini survey. In *2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, August 2023. doi: 10.1109/cibcb56990.2023.10264883.
- [178] Katrin Palm, Patric Stenberg, Kristina Luthman, and Per Artursson1. *Pharmaceutical Research*, 14(5):568–571, 1997. doi: 10.1023/a:1012188625088.
- [179] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. March 2018.
- [180] Cheonbok Park, Jaegul Choo, Inyoup Na, Yongjang Jo, Sungbok Shin, Jaehyo Yoo, Bum Chul Kwon, Jian Zhao, Hyungjong Noh, and Yeonsoo Lee. SANVis: Visual analytics for understanding self-attention networks. In *2019 IEEE Visualization Conference (VIS)*. IEEE, oct 2019. doi: 10.1109/visual.2019.8933677.
- [181] Cheonbok Park, Soyoung Yang, Inyoup Na, Sunghyo Chung, Sungbok Shin, Bum Chul Kwon, Deokgun Park, and Jaegul Choo. Vatun: Visual analytics for testing and understanding convolutional neural networks. *EuroVis 2021 - Short Papers*, 2021. doi: 10.2312/EVS.20211047.
- [182] Haekyu Park, Nilaksh Das, Rahul Duggal, Austin P. Wright, Omar Shaikh, Fred Hohman, and Duen Horng Polo Chau. NeuroCartography: Scalable automatic visual summarization of concepts in deep neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):813–823, jan 2022. doi: 10.1109/tvcg.2021.3114858.
- [183] Jae Heon Park, Kwang Hyuk Im, Chung-Kwan Shin, and Sang Chan Park. Mbnr: Case-based reasoning with local feature weighting by neural network. *Applied Intelligence*, 21(3):265–276, November 2004. ISSN 0924-669X. doi: 10.1023/b:apin.0000043559.83167.3d.
- [184] Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual*

- Meeting of the Association for Computational Linguistics*, pages 1504–1519, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1146. URL <https://aclanthology.org/P19-1146>.
- [185] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.
  - [186] N. Pezzotti, T. Höllt, B. Lelieveldt, E. Eisemann, and A. Vilanova. Hierarchical stochastic neighbor embedding. *Computer Graphics Forum*, 35(3):21–30, jun 2016. doi: 10.1111/cgf.12878.
  - [187] Nicola Pezzotti, Thomas Holtt, Jan Van Gemert, Boudewijn P.F. Lelieveldt, Elmar Eisemann, and Anna Vilanova. DeepEyes: Progressive visual analytics for designing deep neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):98–108, jan 2018. doi: 10.1109/tvcg.2017.2744358.
  - [188] An phi Nguyen and María Rodríguez Martínez. On quantitative aspects of model interpretability. July 2020.
  - [189] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. Face: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’20. ACM, February 2020. doi: 10.1145/3375627.3375850.
  - [190] Michela Proietti, Alessio Ragno, Biagio La Rosa, Rino Ragno, and Roberto Capobianco. Explainable AI in drug discovery: self-interpretable graph neural network for molecular property prediction using concept whitening. *Machine Learning*, oct 2023. doi: 10.1007/s10994-023-06369-y.
  - [191] Ronilo Ragodos, Tong Wang, Qihang Lin, and Xun Zhou. Explaining a reinforcement learning agent via prototyping. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=nyBJcnhjAoy>.
  - [192] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, and Olga Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept salience, and human capability. July 2022.
  - [193] Yanou Ramon, David Martens, Foster Provost, and Theodoros Evgeniou. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: Sedc, lime-c and shap-c. *Advances in Data Analysis and Classification*, 14(4):801–819, September 2020. ISSN 1862-5355. doi: 10.1007/s11634-020-00418-3.
  - [194] Archit Rathore, Nithin Chalapathi, Sourabh Palande, and Bei Wang. Topoact: Visually exploring the shape of activations in deep learning. *Computer Graphics Forum*, 40(1):382–397, 2021. doi: <https://doi.org/10.1111/cgf.14195>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.14195>.

- [195] Yeasir Rayhan and Tanzima Hashem. Aist: An interpretable attention-based deep learning model for crime prediction. *ACM Transactions on Spatial Algorithms and Systems*, 9(2):1–31, April 2023. ISSN 2374-0361. doi: 10.1145/3582274.
- [196] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016. doi: 10.1109/cvpr.2016.91.
- [197] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’16. ACM, August 2016. doi: 10.1145/2939672.2939778.
- [198] Mattia Rigotti, Christoph Mikovic, Ioana Giurgiu, Thomas Gschwind, and Paolo Scotton. Attention-based interpretability with concept transformers. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=kAa9eDS0Rd0>.
- [199] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. ISSN 0033-295X. doi: 10.1037/h0042519.
- [200] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0048-x.
- [201] Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz Zieliński. Protopshare: Prototypical parts sharing for similarity discovery in interpretable image classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD ’21. ACM, August 2021. doi: 10.1145/3447548.3467245.
- [202] Dawid Rymarczyk, Łukasz Struski, Michał Górszczak, Koryna Lewandowska, Jacek Tabor, and Bartosz Zieliński. *Interpretable Image Classification with Differentiable Prototypes Assignment*, pages 351–368. Springer Nature Switzerland, 2022. ISBN 9783031197758. doi: 10.1007/978-3-031-19775-8\_21.
- [203] Dominik Sacha, Andreas Stoffel, Florian Stoffel, Bum Chul Kwon, Geoffrey Ellis, and Daniel A. Keim. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1604–1613, 2014. doi: 10.1109/TVCG.2014.2346481.
- [204] Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. Neuron-level interpretation of deep NLP models: A survey. *Transactions of the Association for Computational Linguistics*, 10:1285–1303, 2022. doi: 10.1162/tacl\_a\_00519.
- [205] Hiroshi Sakiyama, Motohisa Fukuda, and Takashi Okuno. Prediction of blood-brain barrier penetration (bbbP) based on molecular descriptors of the free-form and in-blood-form datasets. *Molecules*, 2021.

- [206] Shaeke Salman, Seyedeh Neelufar Payrovnaziri, Xiuwen Liu, Pablo Rengifo-Moreno, and Zhe He. Deepconsensus: Consensus-based interpretable deep neural networks with application to mortality prediction. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2020. doi: 10.1109/ijcnn48605.2020.9206678.
- [207] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [208] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.
- [209] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vision*, 128(2):336–359, feb 2020. ISSN 0920-5691. doi: 10.1007/s11263-019-01228-7. URL <https://doi.org/10.1007/s11263-019-01228-7>.
- [210] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Gallagher, and Tina Eliassi-Rad. Collective classification in network data. *AI Magazine*, 29(3):93–106, September 2008. ISSN 2371-9621. doi: 10.1609/aimag.v29i3.2157.
- [211] Sofia Serrano and Noah A. Smith. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1282.
- [212] Shubham Sharma, Jette Henderson, and Joydeep Ghosh. Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES ’20. ACM, February 2020. doi: 10.1145/3375627.3375812.
- [213] Qiaomu Shen, Yanhong Wu, Yuzhe Jiang, Wei Zeng, Alexis K H LAU, Anna Vianova, and Huamin Qu. Visual interpretation of recurrent neural network on multi-dimensional time-series forecast. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, jun 2020. doi: 10.1109/pacificvis48177.2020.2785.
- [214] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 3145–3153. JMLR.org, 2017.
- [215] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, October 2017. ISSN 1476-4687. doi: 10.1038/nature24270.

- [216] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. September 2014. doi: 10.48550/ARXIV.1409.1556.
- [217] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014.
- [218] Gurmail Singh and Kin-Choong Yow. These do not look like those: An interpretable deep learning model for image recognition. *IEEE Access*, 9:41482–41493, 2021. doi: 10.1109/access.2021.3064838.
- [219] Suzanne Sirois, Christos Tsoukas, Kuo-Chen Chou, Dong-Qing Wei, Christina Boucher, and G Hatzakis. Selection of molecular descriptors with artificial intelligence for the understanding of hiv-1 protease peptidomimetic inhibitors-activity. *Medicinal chemistry (Shāriqah (United Arab Emirates))*, 1:173–84, 04 2005. doi: 10.2174/1573406053175238.
- [220] Blaž Škrlj, Shane Sheehan, Nika Eržen, Marko Robnik-Šikonja, Saturnino Luz, and Senja Poljak. Exploring neural language models via analysis of local and global self-attention spaces. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 76–83, Online, April 2021. Association for Computational Linguistics. URL <https://aclanthology.org/2021.hackashop-1.11>.
- [221] Dylan Slack, Satyapriya Krishna, Himabindu Lakkaraju, and Sameer Singh. Explaining machine learning models with interactive natural language conversations using talktomodel. *Nature Machine Intelligence*, 5(8):873–883, July 2023. ISSN 2522-5839. doi: 10.1038/s42256-023-00692-8.
- [222] Daniel Smilkov, Nikhil Thorat, Charles Nicholson, Emily Reif, Fernanda B. Viégas, and Martin Wattenberg. Embedding projector: Interactive visualization and interpretation of embeddings. *arXiv preprint arXiv:1611.05469*, November 2016.
- [223] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 4080–4090. Curran Associates Inc., 2017. ISBN 9781510860964.
- [224] J Springenberg, Alexey Dosovitskiy, Thomas Brox, and M Riedmiller. Striving for simplicity: The all convolutional net. In *ICLR (workshop track)*, 2015.
- [225] Stephen M. Stigler. Francis galton's account of the invention of correlation. *Statistical Science*, 4(2), may 1989. doi: 10.1214/ss/1177012580.
- [226] Gjorgji Strezoski and Marcel Worring. Plug-and-play interactive deep network visualization. *VADL: Visual Analytics for Deep Learning*, pages 0100–0106, 2017.
- [227] Hendrik Strobelt, Sebastian Gehrmann, Hanspeter Pfister, and Alexander M. Rush. LST-MVis: A tool for visual analysis of hidden state dynamics in recurrent neural networks. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):667–676, jan 2018. doi: 10.1109/tvcg.2017.2744158.

- [228] Hendrik Strobelt, Sebastian Gehrmann, Michael Behrisch, Adam Perer, Hanspeter Pfister, and Alexander M. Rush. Seq2seq-vis: A visual debugging tool for sequence-to-sequence models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):353–363, jan 2019. doi: 10.1109/tvcg.2018.2865044.
- [229] Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. Spine: Sparse interpretable neural embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2159-5399. doi: 10.1609/aaai.v32i1.11935.
- [230] Govindan Subramanian, Bharath Ramsundar, Vijay Pande, and Rajiah Aldrin Denny. Computational modeling of  $\beta$ -secretase 1 (bace-1) inhibitors using ligand based approaches. *Journal of Chemical Information and Modeling*, 56(10):1936–1949, October 2016. ISSN 1549-960X. doi: 10.1021/acs.jcim.6b00290.
- [231] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 3319–3328. JMLR.org, 2017.
- [232] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, June 2015.
- [233] Yaniv Taigman, Ming Yang, Marc’Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2014. doi: 10.1109/cvpr.2014.220.
- [234] Mingxing Tan and Quoc Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019.
- [235] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1452.
- [236] Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, and Ann Yuan. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-demos.15.
- [237] Soon Tee Teoh and Kwan-Liu Ma. Paintingclass: interactive construction, visualization and exploration of decision trees. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD03. ACM, August 2003. doi: 10.1145/956750.956837.

- [238] J.J. Thomas and K.A. Cook. A visual analytics agenda. *IEEE Computer Graphics and Applications*, 26(1):10–13, 2006. doi: 10.1109/MCG.2006.5.
- [239] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):4793–4813, November 2021. ISSN 2162-2388. doi: 10.1109/tnnls.2020.3027314.
- [240] Christian Tominski and Heidrun Schumann. *Interactive Visual Data Analysis*. AK Peters Visualization Series. CRC Press, 2020. ISBN 9781498753982. doi: 10.1201/9781315152707. URL <https://ivda-book.de>.
- [241] Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *Journal of Statistical Physics*, 52(1-2):479–487, jul 1988. doi: 10.1007/bf01016429.
- [242] Astrid van den Brandt, Mark Christopher, Jasmin Rezapour, Derek S Welsbie, Andrew S Camp, Sally L Baxter, Jiun Do, Sasan Moghimi, Akram Belghith, Christopher Bowd, et al. Glance: A visual analytics approach for opening the black box to explain deep learning predictions of glaucomatous visual field damage from optical coherence tomography scans. *Investigative Ophthalmology & Visual Science*, 61(7):4527–4527, 2020.
- [243] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008. URL <http://jmlr.org/papers/v9/vandermaaten08a.html>.
- [244] Saurabh Varshneya, Antoine Ledent, Robert A. Vandermeulen, Yunwen Lei, Matthias Enders, Damian Borth, and Marius Kloft. Learning interpretable concept groups in CNNs. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, aug 2021. doi: 10.24963/ijcai.2021/147.
- [245] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL [https://proceedings.neurips.cc/paper\\_files/paper/2017/file/3f5ee243547dee91fb0d053c1c4a845aa-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fb0d053c1c4a845aa-Paper.pdf).
- [246] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks. *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=rJXMpikCZ>.
- [247] Tom Vermeire, Dieter Brughmans, Sofie Goethals, Raphael Mazzine Barbossa de Oliveira, and David Martens. Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications*, 25(2):315–335, January 2022. ISSN 1433-755X. doi: 10.1007/s10044-021-01055-y.
- [248] Jesse Vig. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, 2019.

- [249] Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/w19-4808.
- [250] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3637–3645. Curran Associates Inc., 2016. ISBN 9781510838819.
- [251] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law and Technology*, 31(2):841–887, 2018.
- [252] Chuan Wang and Kwan-Liu Ma. Hyppersteer: Hypothetical steering and data perturbation in sequence prediction with deep learning. *arXiv preprint arXiv:2011.02149*, November 2020.
- [253] Chuan Wang, Takeshi Onishi, Keiichi Nemoto, and Kwan-Liu Ma. Visual reasoning of feature attribution with deep recurrent neural networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1661–1668, 2018. doi: 10.1109/BigData.2018.8622502.
- [254] Feng Wang, Haijun Liu, and Jian Cheng. Visualizing deep neural network by alternately image blurring and deblurring. *Neural Networks*, 97:162–172, January 2018. ISSN 0893-6080. doi: 10.1016/j.neunet.2017.09.007.
- [255] J. Wang, L. Gou, H. Shen, and H. Yang. Dqnviz: A visual analytics approach to understand deep q-networks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):288–298, Jan 2019. ISSN 1941-0506. doi: 10.1109/TVCG.2018.2864504.
- [256] Jiaqi Wang, Huafeng Liu, Xinyue Wang, and Liping Jing. Interpretable image recognition by constructing transparent embedding space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 895–904, October 2021.
- [257] Junpeng Wang, Liang Gou, Wei Zhang, Hao Yang, and Han-Wei Shen. DeepVID: Deep visual interpretation and diagnosis for image classifiers via knowledge distillation. *IEEE Transactions on Visualization and Computer Graphics*, 25(6):2168–2180, jun 2019. doi: 10.1109/tvcg.2019.2903943.
- [258] Junpeng Wang, Wei Zhang, and Hao Yang. SCANViz: Interpreting the symbol-concept association captured by deep neural networks through visual analytics. In *2020 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, jun 2020. doi: 10.1109/pacificvis48177.2020.3542.
- [259] Junpeng Wang, Wei Zhang, Hao Yang, Chin-Chia Michael Yeh, and Liang Wang. Visual analytics for RNN-based deep reinforcement learning. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2021. doi: 10.1109/tvcg.2021.3076749.

- [260] Pei Wang and Nuno Vasconcelos. Scout: Self-aware discriminant counterfactual explanations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. doi: 10.1109/cvpr42600.2020.00900.
- [261] Xingbo Wang, Jianben He, Zhihua Jin, Muqiao Yang, Yong Wang, and Huamin Qu. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):802–812, 2022. doi: 10.1109/TVCG.2021.3114794.
- [262] Zijie J. Wang, Robert Turko, and Duen Horng Chau. Dodrio: Exploring transformer models with interactive visualization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 132–141, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-demo.16.
- [263] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viegas, and Jimbo Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, pages 1–1, 2019. doi: 10.1109/tvcg.2019.2934619.
- [264] Sarah Wiegreffe and Yuval Pinter. Attention is not explanation. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1002. URL <https://aclanthology.org/D19-1002>.
- [265] Sarah Wiegreffe and Yuval Pinter. Attention is not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2019. doi: 10.18653/v1/d19-1002.
- [266] Austin P. Wright, Omar Shaikh, Haekyu Park, Will Epperson, Muhammed Ahmed, Stephane Pinel, Duen Horng (Polo) Chau, and Diyi Yang. RECAST: Enabling user recourse and interpretability of toxicity detection models with interactive visualization. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–26, apr 2021. doi: 10.1145/3449280.
- [267] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, January 2021. doi: 10.1109/tnnls.2020.2978386. URL <https://doi.org/10.1109/tnnls.2020.2978386>.
- [268] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=ryGs6iA5Km>.

- [269] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2020. doi: 10.1109/cvpr42600.2020.00970.
- [270] Shuangjie Xu, Yu Cheng, Kang Gu, Yang Yang, Shiyu Chang, and Pan Zhou. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.
- [271] Linyi Yang, Eoin Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. Generating plausible counterfactual explanations for deep transformers in financial text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.541.
- [272] Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19187–19197, June 2023.
- [273] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL <https://proceedings.neurips.cc/paper/2019/file/d80b7040b773199015de6d3b4293c8ff-Paper.pdf>.
- [274] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.
- [275] Hao Yuan, Haiyang Yu, Shurui Gui, and Shuiwang Ji. Explainability in graph neural networks: A taxonomic survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–19, 2022. ISSN 1939-3539. doi: 10.1109/tpami.2022.3204236.
- [276] Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL <https://openreview.net/forum?id=nA5AZ8CEyow>.
- [277] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *Proceedings of the British Machine Vision Conference 2016*. British Machine Vision Association, 2016. doi: 10.5244/c.30.87.
- [278] Tom Zahavy, Nir Ben-Zrihem, and Shie Mannor. Graying the black box: Understanding dqns. In *International Conference on Machine Learning*, pages 1899–1908. PMLR, 2016.
- [279] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833. Springer International Publishing, 2014. doi: 10.1007/978-3-319-10590-1\_53.

- [280] Haipeng Zeng, Hammad Haleem, Xavier Plantaz, Nan Cao, and Huamin Qu. Cnncomparator: Comparative analytics of convolutional neural networks. *arXiv preprint arXiv:1710.05285*, October 2017.
- [281] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6848–6856. IEEE, jun 2018. doi: 10.1109/cvpr.2018.00716.
- [282] Yu Zhang, Peter Tino, Ales Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, oct 2021. doi: 10.1109/tetci.2021.3100641.
- [283] Zaixi Zhang, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. ProtGNN: Towards self-explaining graph neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(8):9127–9135, June 2022. doi: 10.1609/aaai.v36i8.20898. URL <https://doi.org/10.1609/aaai.v36i8.20898>.
- [284] Junhan Zhao, Zeng Dai, Panpan Xu, and Liu Ren. ProtoViewer: Visual interpretation and diagnostics of deep neural networks with factorized prototypes. In *2020 IEEE Visualization Conference (VIS)*. IEEE, oct 2020. doi: 10.1109/vis47514.2020.00064.
- [285] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017. doi: 10.1109/iccv.2017.557.
- [286] Wen Zhong, Cong Xie, Yuan Zhong, Yang Wang, Wei Xu, Shenghui Cheng, and Klaus Mueller. Evolutionary visual analysis of deep neural networks. In *ICML Workshop on Visualization for Deep Learning*, page 9, 2017.
- [287] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [288] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, jun 2018. doi: 10.1109/tpami.2017.2723009.
- [289] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.
- [290] Roland Simon Zimmermann, Judy Borowski, Robert Geirhos, Matthias Bethge, Thomas S. A. Wallis, and Wieland Brendel. How well do feature visualizations support causal understanding of CNN activations? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. URL <https://openreview.net/forum?id=vLPqnPf9k0>.

- [291] Luisa M Zintgraf, Taco S Cohen, Tameem Adel, and Max Welling. Visualizing deep neural network decisions: Prediction difference analysis. In *International Conference on Learning Representations*, 2017. URL <https://openreview.net/forum?id=BJ5UeU9xx>.
- [292] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, August 2013. ISSN 0219-3116. doi: 10.1007/s10115-013-0679-x.