Concept-Based Methods for Neural Network Interpretation

Riccardo Massidda

https://pages.di.unipi.it/massidda/

Università di Pisa

Abstract

Concept-based methods attempt to interpret existing neural networks or to design inherently interpretable models by exploiting human-comprehensible concepts. In the current talk, I will present few significant examples of such methods, discussing their commonalities, their underlying assumptions, and their applications. More in detail, I will focus on the semantic alignment of neural directions and visual concepts in CNNs for computer vision. In this context, different existing approaches might be understood in terms of a unified general framework. Furthermore, I will show the impact of acknowledging semantic relations on such framework. Finally, the talk discusses the main issues affecting concept-based methods and hints to possible research strategies to tackle them.

Table of Contents

Motivation

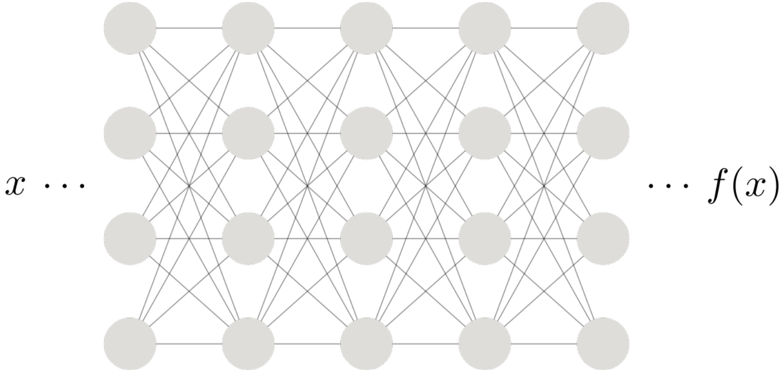


Figure 1: An Artificial Neural Network f is composed of a set of neural units U which are displaced into consecutive and interconnected layers.

Function and Concept

Frege (1891)

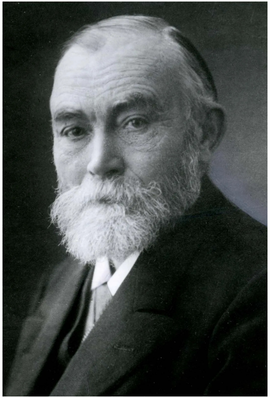


Figure 2: Gottlob Frege

"We thus see how closely that which is called a concept in logic is connected with what we call a function. Indeed, we may say at once: a *concept* is a *function* whose value is always a *truth-value*."

Function and Concept

Frege (1891)

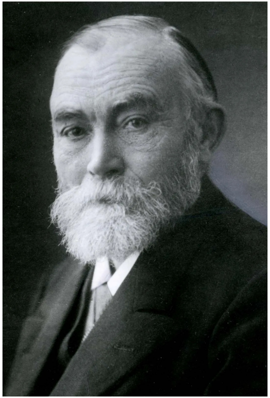


Figure 2: Gottlob Frege

"We thus see how closely that which is called a concept in logic is connected with what we call a function. Indeed, we may say at once: a *concept* is a *function* whose value is always a *truth-value*."

$$I_c : W \to \{T, F\} \tag{1}$$

Function and Concept

Frege (1891)

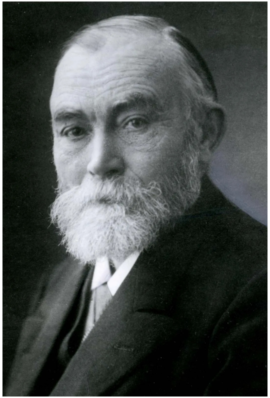


Figure 2: Gottlob Frege

"We thus see how closely that which is called a concept in logic is connected with what we call a function. Indeed, we may say at once: a *concept* is a *function* whose value is always a *truth-value*."

$$I_c : W \to \{T, F\} \tag{1}$$

$$E_c = \{o \in W : I_c(o)\} \tag{2}$$

Concept-Based Methods

Posthoc analysis:

- Latent Object Detectors (Zhou, Khosla, et al. 2014)
- Feature Visualization (Olah, Mordvintsev, and Schubert 2017)
- Network Dissection (Zhou, Bau, et al. 2019)
- TCAV (Kim et al. 2018)
- CaCe (Goyal et al. 2020)
- Interpretable Basis Decomposition (Zhou, Sun, et al. 2018)
- Net2Vec (Fong and Vedaldi 2018)
- ConceptSHAP (Yeh et al. 2020)

Inherently interpretable:

- Concept Bottleneck Models (Koh et al. 2020)
- Debiased CBMs (Bahadori and Heckerman 2021)
- Graph CBMs for algorithmic reasoning (Georgiev et al. 2021)
- ProtoPNet (C. Chen et al. 2019)
- Concept Whitening (Z. Chen, Bei, and Rudin 2020)

Concept-Based Methods

Posthoc analysis:

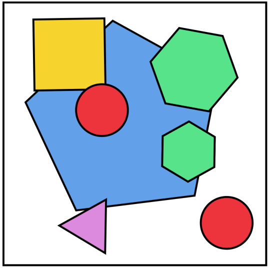
- Latent Object Detectors (Zhou, Khosla, et al. 2014)
- Feature Visualization (Olah, Mordvintsev, and Schubert 2017)
- **Network Dissection** (Zhou, Bau, et al. 2019)
- **TCAV** (Kim et al. 2018)
- CaCe (Goyal et al. 2020)
- Interpretable Basis Decomposition (Zhou, Sun, et al. 2018)
- Net2Vec (Fong and Vedaldi 2018)
- ConceptSHAP (Yeh et al. 2020)

Inherently interpretable:

- Concept Bottleneck Models (Koh et al. 2020)
- Debiased CBMs (Bahadori and Heckerman 2021)
- Graph CBMs for algorithmic reasoning (Georgiev et al. 2021)
- ProtoPNet (C. Chen et al. 2019)
- Concept Whitening (Z. Chen, Bei, and Rudin 2020)

Network Dissection

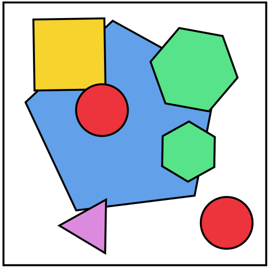
Bau et al. (2020)



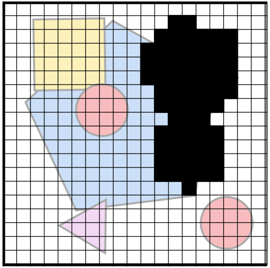
(a) $x \in X$

Network Dissection

Bau et al. (2020)



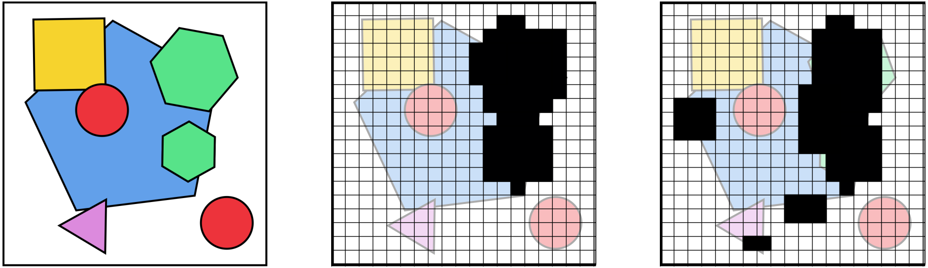
(a) $x \in X$



(b) $L_c(x)$

Network Dissection

Bau et al. (2020)



(a) $x \in X$ (b) $L_c(x)$ (c) $M_u(x)$

Figure 3: Decomposition of an image example x for a specific concept c and a given unit u according to the Network Dissection approach.

Network Dissection

Bau et al. (2020)

$$\mathrm{IoU}(u, c) = \frac{\sum_{x \in X} |M_u(x) \wedge L_c(x)|}{\sum_{x \in X} |M_u(x) \vee L_c(x)|} \tag{3}$$

Concept Activation Vectors (CAVs)

Kim et al. (2018)

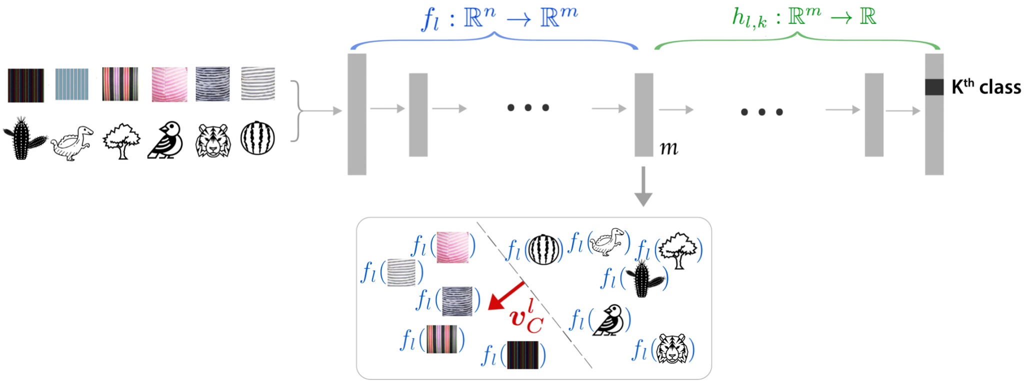


Figure 4: Schema of the learning procedure for a Concept Activation Vector.

Testing with CAVs (TCAV)

Kim et al. (2018)

$$
\begin{aligned}
S_{C,k,l}(x) &= \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(x) + \epsilon v_C^l) - h_{l,k}(f_l(x))}{\epsilon} \\
&= \nabla h_{l,k}(f_l(x)) \cdot v_C^l
\end{aligned} \tag{4}
$$

Testing with CAVs (TCAV)

Kim et al. (2018)

$$
\begin{aligned}
S_{C,k,l}(x) &= \lim_{\epsilon \to 0} \frac{h_{l,k}(f_l(x) + \epsilon v_C^l) - h_{l,k}(f_l(x))}{\epsilon} \\
&= \nabla h_{l,k}(f_l(x)) \cdot v_C^l
\end{aligned} \tag{4}
$$

$$
\mathrm{TCAV}_{C,k,l} = \frac{|\{x \in X_k : S_{C,k,l}(x) > 0\}|}{|X_k|} \tag{5}
$$

Table of Contents

Schematic representation

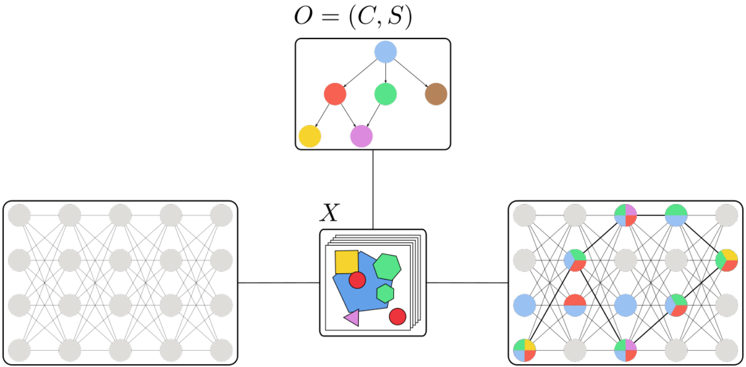


Figure 5: Overview of the proposed methodology. A set of neural directions D is semantically aligned with an ontology O through a pixel-level annotated dataset X, whose labels are in a two-way relationship with the ontology concepts C. Semantic relations S enable the retrieval of subgraphs composed of architecturally connected and semantically related directions.

Higher-Level Concepts

(a) Induced taxonomy (b) $x \in X$ (c) $L_c(x)$

Figure 6: Given the taxonomy induced by the specialisation relation, it is possible to analyze concept masks not directly annotated in the input.

Higher-Level Concepts

(a) Induced taxonomy (b) $x \in X$ (d) $L_c(x)$

Figure 6: Given the taxonomy induced by the specialisation relation, it is possible to analyze concept masks not directly annotated in the input.

Higher-Level Concepts

(a) Induced taxonomy (b) $x \in X$ (e) $L_c(x)$

Figure 6: Given the taxonomy induced by the specialisation relation, it is possible to analyze concept masks not directly annotated in the input.

Neural Directions

A neural direction

$$d = (l, v) \tag{6}$$

identifies a specific vector direction v in the output space of the l-th layer of the network.

Neural Directions

A neural direction

$$d = (l, v) \tag{6}$$

identifies a specific vector direction v in the output space of the l-th layer of the network. For a given input x, its ouput is given by the dot-product

$$A_d(x) = f^l(x) \cdot v. \tag{7}$$

Neural Directions

A neural direction

$$d = (l, v) \tag{6}$$

identifies a specific vector direction v in the output space of the l-th layer of the network. For a given input x, its ouput is given by the dot-product

$$A_d(x) = f^l(x) \cdot v. \tag{7}$$

$\approx$ essentially a CAV!

Neural Directions

A neural direction

$$d = (l, v) \tag{6}$$

identifies a specific vector direction v in the output space of the l-th layer of the network. For a given input x, its ouput is given by the dot-product

$$A_d(x) = f^l(x) \cdot v. \tag{7}$$

$\approx$ essentially a CAV!

When $\exists i. v = e^{(i)}$, the direction refers to the i-th unit of the l-th layer.

Neural Directions

A neural direction

$$d = (l, v) \tag{6}$$

identifies a specific vector direction v in the output space of the l-th layer of the network. For a given input x, its ouput is given by the dot-product

$$A_d(x) = f^l(x) \cdot v. \tag{7}$$

$\approx$ essentially a CAV!

When $\exists i. v = e^{(i)}$, the direction refers to the i-th unit of the l-th layer.

$\approx$ as in NetDissect!

Semantic Alignment

Given a set of neural directions D and an ontology $O = (C, S)$, the semantic alignment is estimated by an arbitrary performance metric

$$\sigma : D \times C \to [0, 1] \tag{8}$$

over the classification boundary defined by the direction.

Candidate σ

The Jaccard similarity, also known as Intersection over Union (IoU),

$$\sigma_{\mathrm{IoU}}(d, c) = \frac{\sum_{x \in X} |M_d(x) \wedge L_c(x)|}{\sum_{x \in X} |M_d(x) \vee L_c(x)|}, \tag{9}$$

Candidate σ

The Jaccard similarity, also known as Intersection over Union (IoU),

$$\sigma_{\mathrm{IoU}}(d, c) = \frac{\sum_{x \in X} |M_d(x) \wedge L_c(x)|}{\sum_{x \in X} |M_d(x) \vee L_c(x)|}, \tag{9}$$

or the Sørensen–Dice coefficient, also known as F1 score,

$$\sigma_{\mathrm{F1}}(d, c) = \frac{\sum_{x \in X} 2|M_u(x) \wedge L_c(x)|}{\sum_{x \in X} |M_u(x)| + |L_c(x)|}, \tag{10}$$

consitute insightful measures of semantic alignment.

Acknowledging Polysemanticity

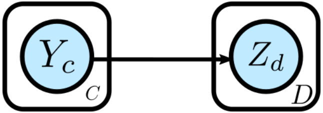


Figure 7: Ideal interaction between visual concepts C and neural directions D.

$$\begin{aligned}\sigma_{\mathcal{L}}(d, c) &= \mathcal{L}(Y_c = 1 \mid Z_d = 1) \\ &= \frac{\sum_x |L_c(x) \wedge M_d(x)|}{\sum_x |L_c(x)|}\end{aligned} \tag{11}$$

Acknowledging Polysemanticity

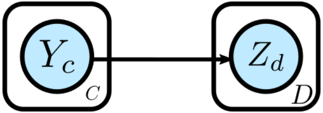


Figure 7: Ideal interaction between visual concepts C and neural directions D.

$$\begin{aligned}\sigma_{\mathcal{L}}(d, c) &= \mathcal{L}(Y_c = 1 \mid Z_d = 1) \\ &= \frac{\sum_x |L_c(x) \wedge M_d(x)|}{\sum_x |L_c(x)|} \\ &= \text{recall}\end{aligned} \tag{11}$$

Unit-alignment comparison

(a) Examples with maximal activations

Synset	$\sigma(u, c)$
hovel.n.01	0.021
roof.n.03	0.025
building.n.01	0.031
shelter.n.01	0.035
house.n.01	0.098

(b) $\mathrm{IoU}(u, c)$

Synset	$\sigma(u, c)$
circus_tent.n.01	0.401
greenhouse.n.01	0.403
shed.n.01	0.469
pavilion.n.01	0.568
bandstand.n.01	0.631

(c) $\mathcal{L}(c \mid u)$

Figure 8: Semantic alignment of unit 196 in the last residual block of ResNet-18.

Network Alignment Ψ

We define the set of τ-sufficiently aligned direction-concept pairs as

$$\Psi_\tau = \{(d, c) \mid \sigma(d, c) \geq \tau\} \subseteq D \times C. \tag{12}$$

Network Alignment Ψ

We define the set of τ-sufficiently aligned direction-concept pairs as

$$\Psi_\tau = \{(d, c) \mid \sigma(d, c) \geq \tau\} \subseteq D \times C. \tag{12}$$

Alignment-pairs can be connected in a directed graph

$$G = (\Psi, E) \tag{13}$$

where an edge between two pairs exists if and only if directions are architecturally dependent and concepts are semantically related.

Network Alignment Ψ

We define the set of τ-sufficiently aligned direction-concept pairs as

$$\Psi_\tau = \{(d, c) \mid \sigma(d, c) \geq \tau\} \subseteq D \times C. \tag{12}$$

Alignment-pairs can be connected in a directed graph

$$G = (\Psi, E) \tag{13}$$

where an edge between two pairs exists if and only if directions are architecturally dependent and concepts are semantically related.
By extracting each non-trivial connected component, we obtain a set

$$T = \{t \mid t \subseteq \Psi, |t| > 1, G[t] \text{ is connected}\}, \tag{14}$$

where each $t \in T$ is a semantically related and architecturally connected neural circuit.

Circuits analysis

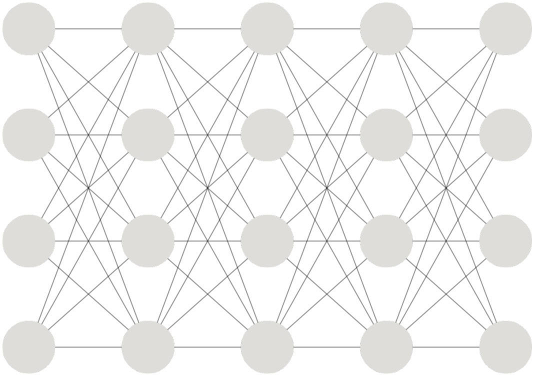


Figure 9: The ontological structure of visual concepts enables the retrieval of architecturally connected and semantically related directions.

Circuits analysis

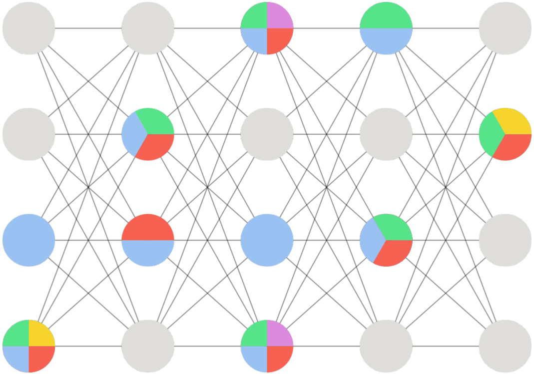


Figure 9: The ontological structure of visual concepts enables the retrieval of architecturally connected and semantically related directions.

Circuits analysis

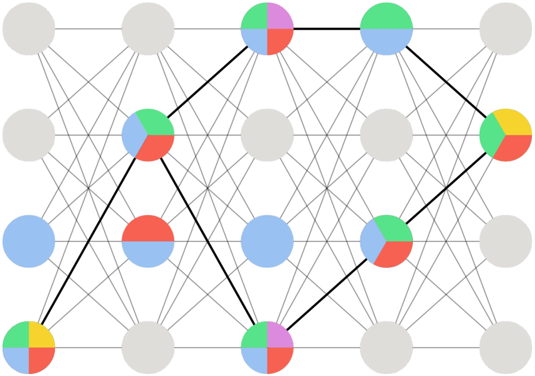


Figure 9: The ontological structure of visual concepts enables the retrieval of architecturally connected and semantically related directions.

Circuit 16

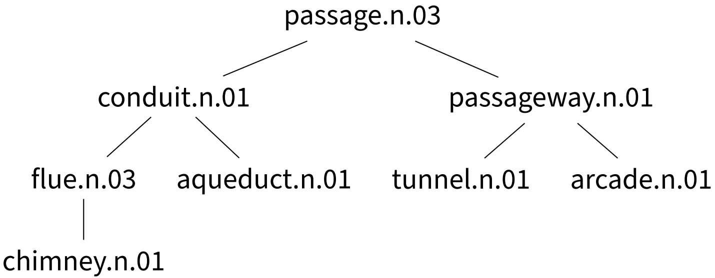


Figure 10: Hierarchy of WordNet synsets within Circuits 16 from AlexNet/Broden pretrained on Places365.

Circuit 16

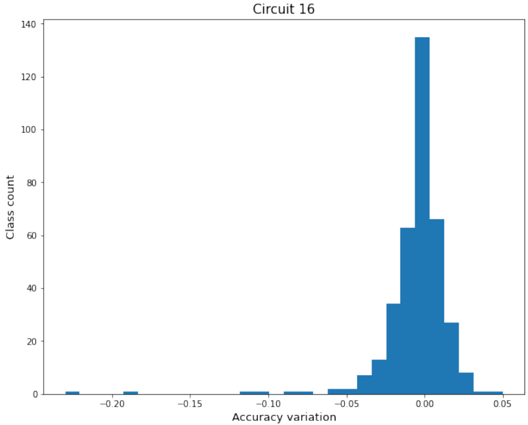


Figure 11: Accuracy drop histogram for Circuit 16 from AlexNet/Broden pretrained on Places365.

k	Description	Drop
96	/c/clothing_store	-0.05
121	/d/dining_room	-0.05
159	/g/gazebo/exterior	-0.06
91	/c/church/outdoor	-0.06
12	/a/arch	-0.08
260	/p/pavilion	-0.09
288	/r/river	-0.1
66	/b/bridge	-0.11
347	/v/viaduct	-0.19
10	/a/aqueduct	-0.23

Table 1: Class accuracy drop for Circuit 16 from AlexNet/Broden pretrained on Places365.

Implementation

Bisturi

`https://github.com/rmassidda/bisturi`

Table of Contents

The network has learned concept c.

The network ~~has learned~~ is able to represent concept c.

The network ~~has learned~~ is able to represent concept c.

...well, so?

Issues

Discussion points

Issues

Discussion points

- Intuitive approach, without formal statements.

Issues

Discussion points

- Intuitive approach, without formal statements.
- Do we really want symbols to get in through the backdoor?
  (Fodor and Pylyshyn 1988; Chalmers 1990)

Issues

Discussion points

- Intuitive approach, without formal statements.
- Do we really want symbols to get in through the backdoor?
  (Fodor and Pylyshyn 1988; Chalmers 1990)
- How much transformations to justify the previous "conceptual statement"?

Issues

Discussion points

- Intuitive approach, without formal statements.
- Do we really want symbols to get in through the backdoor?
  (Fodor and Pylyshyn 1988; Chalmers 1990)
- How much transformations to justify the previous "conceptual statement"?
- Massive need of label data w/o self-supervision.

Issues

Discussion points

- Intuitive approach, without formal statements.
- Do we really want symbols to get in through the backdoor?
  (Fodor and Pylyshyn 1988; Chalmers 1990)
- How much transformations to justify the previous "conceptual statement"?
- Massive need of label data w/o self-supervision.
- “Explanations must be wrong.” (Rudin 2019)

Issues

Discussion points

- Intuitive approach, without formal statements.
- Do we really want symbols to get in through the backdoor?
  (Fodor and Pylyshyn 1988; Chalmers 1990)
- How much transformations to justify the previous "conceptual statement"?
- Massive need of label data w/o self-supervision.
- “Explanations must be wrong.” (Rudin 2019)

  > *If the explanation was completely faithful to what the original model computes, the explanation would equal the original model, and one would not need the original model in the first place, only the explanation.*

References I

Bahadori, Mohammad Taha and David Heckerman (2021). "Debiasing Concept-based Explanations with Causal Analysis". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=6puUoArESGp.

Bau, David et al. (2020). "Understanding the role of individual units in a deep neural network". In: *Proceedings of the National Academy of Sciences*. ISSN: 0027-8424. DOI: 10.1073/pnas.1907375117. URL: https://www.pnas.org/content/early/2020/08/31/1907375117.

Chalmers, David (1990). "Why Fodor and Pylyshyn were wrong: The simplest refutation". In: *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society, Cambridge, Mass*, pp. 340–347.

Chen, Chaofan et al. (2019). "This Looks Like That: Deep Learning for Interpretable Image Recognition". In: *Proceedings of Neural Information Processing Systems (NeurIPS)*.

References II

- Chen, Zhi, Yijie Bei, and Cynthia Rudin (Dec. 2020). “Concept whitening for interpretable image recognition”. en. In: *Nature Machine Intelligence* 2.12, pp. 772–782. ISSN: 2522-5839. DOI: 10.1038/s42256-020-00265-z. URL: https://www.nature.com/articles/s42256-020-00265-z (visited on 01/24/2022).
- Fodor, Jerry A and Zenon W Pylyshyn (1988). “Connectionism and cognitive architecture: A critical analysis”. In: *Cognition* 28.1-2, pp. 3–71.
- Fong, Ruth and Andrea Vedaldi (Mar. 2018). “Net2Vec: Quantifying and Explaining how Concepts are Encoded by Filters in Deep Neural Networks”. In: *arXiv:1801.03454 [cs, stat]*. arXiv: 1801.03454. URL: http://arxiv.org/abs/1801.03454 (visited on 01/24/2022).
- Frege, Gottlob (1891). *Function und Begriff*. Jena: Hermann Pohle. Trans. by Peter Geach and Max Black. As “Function and Concept”. (Philosophical Library, 1952).

References III

Georgiev, Dobrik et al. (July 2021). "Algorithmic Concept-based Explainable Reasoning". In: *arXiv:2107.07493 [cs]*. arXiv: 2107.07493. URL: http://arxiv.org/abs/2107.07493 (visited on 01/24/2022).

Goyal, Yash et al. (Feb. 2020). "Explaining Classifiers with Causal Concept Effect (CaCE)". In: *arXiv:1907.07165 [cs, stat]*. arXiv: 1907.07165. URL: http://arxiv.org/abs/1907.07165 (visited on 01/25/2021).

Kim, Been et al. (June 2018). "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)". In: *arXiv:1711.11279 [stat]*. arXiv: 1711.11279. URL: http://arxiv.org/abs/1711.11279 (visited on 01/25/2021).

Koh, Pang Wei et al. (Dec. 2020). "Concept Bottleneck Models". In: *arXiv:2007.04612 [cs, stat]*. arXiv: 2007.04612. URL: http://arxiv.org/abs/2007.04612 (visited on 01/25/2021).

Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert (2017). "Feature Visualization". In: *Distill*. DOI: 10.23915/distill.00007. URL: https://distill.pub/2017/feature-visualization.

References IV

- Rudin, Cynthia (May 2019). "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: *Nature Machine Intelligence* 1, pp. 206–215.
- Yeh, Chih-Kuan et al. (June 2020). "On Completeness-aware Concept-Based Explanations in Deep Neural Networks". In: *arXiv:1910.07969 [cs, stat]*. arXiv: 1910.07969. URL: http://arxiv.org/abs/1910.07969 (visited on 01/25/2021).
- Zhou, Bolei, David Bau, et al. (Sept. 2019). "Interpreting Deep Visual Representations via Network Dissection". eng. In: *IEEE transactions on pattern analysis and machine intelligence* 41.9, pp. 2131–2145. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2018.2858759.
- Zhou, Bolei, Aditya Khosla, et al. (2014). "Object detectors emerge in deep scene cnns". In: *arXiv preprint arXiv:1412.6856*.

References V

Zhou, Bolei, Yiyou Sun, et al. (2018). "Interpretable Basis Decomposition for Visual Explanation". en. In: *Computer Vision – ECCV 2018*. Ed. by Vittorio Ferrari et al. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 122–138. ISBN: 9783030012373. DOI: 10.1007/978-3-030-01237-3_8.