

神经网络可解释性综述

A Survey on Neural Network Interpretability

arXiv: 2012.14261



张宇
唐珂



Peter Tino
Ales Leonardis

智源社区分享

2021.1.30

目录

- ▶ Interpretability Disambiguation
- ▶ **What.** 什么是可解释性
- ▶ **Why.** 为什么需要可解释性
- ▶ **How.** 如何得到可解释性
 - ▶ A 3D Taxonomy
- ▶ Discussions

Interpretability Disambiguation

Before we really get to our topic, it is important to **distinguish two groups of work** usually sharing the term “interpretability of DNNs”

Interpretability of concrete DNNs

input $\xrightarrow{?}$ output



cat



dog

ATCGC...GAT

special
protein

Interpretability of Deep learning methodology

? $\xrightarrow{?}$ DL is good

convolution,

generalize

pooling,

well/

SGD,

no

dropout, ...

overfitting

e.g.

“Depth can be exponentially more valuable than width for standard feedforward NNs”

R Eldan & O Shamir, 2016

Interpretability Disambiguation

Before we really get to our topic, it is important to **distinguish two groups of work** usually sharing the term “interpretability of DNNs”

Interpretability of concrete DNNs

input $\xrightarrow{?}$ output



cat



dog

ATCGC...GAT

special
protein



Interpretability of Deep learning methodology

? $\xrightarrow{?}$ DL is good

convolution,
pooling,
SGD,
dropout, ...

generalize
well/
no
overfitting

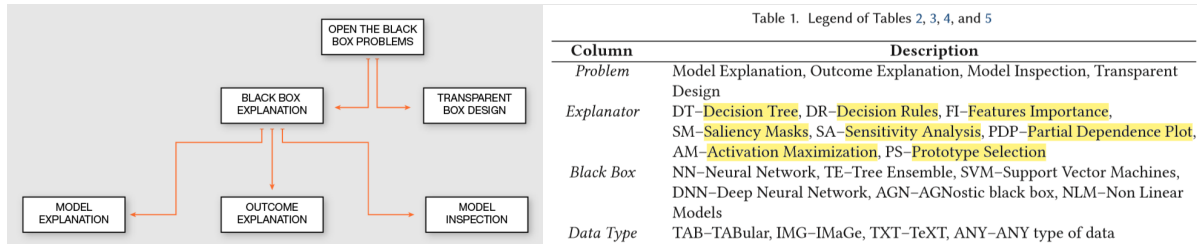
e.g.

“Depth can be exponentially more valuable than width for standard feedforward NNs”

R Eldan & O Shamir, 2016

现有综述的不足之处

e.g. 📄 “A Survey of Methods for Explaining Black Box Models”. *ACM Computing Surveys*. R Guidotti et al., 2018



- ▶ 共性上，主要依赖于一些 pre-recognized **explanators**（解释器）
比如 决策树、feature importance、可视化、代理模型 等等
- ▶ 而这些 explanators 之间的关系比较混乱
有些相互包含，而有些可能完全不在同一个层面
- ▶ （亦或者有些综述仅仅关注某一类方法，不全面）

我们沿用 [F Doshi-Velez & B Kim, 2017](#) 中提到的可解释性的定义，但是额外增加两条解读

Interpretability (of a DNN) is the ability to provide *explanations*¹ in *understandable terms*² to a human.

1. 解释 Explanations, 说到底需要用某种「语言」描述

理想情况下当然使用 逻辑规则 最好 [D Pedreschi et al., 2019](#)，而实践上人们往往不强求「完整的解释」，只需关键信息加脑补

2. 可理解的术语 Understandable terms, 构成解释的基本单元

不同领域的模型的解释需要建立在不同的领域术语之上，比如 CV 中的 image patches，NLP 中的 单词，Bioinformatics 中的 motifs

Explanation-centred!

高可靠性要求

神经网络在实践中经常有**意想不到的错误**（更不用提对抗攻击）
这对于要求高可靠性的系统来说很危险（**不信任**）

可解释性有助于发现潜在的**错误**（比如发现模型逻辑和 domain knowledge 不相符）；也可能可以帮助 debug，改进模型

伦理 / 法规要求

药物设计，医疗器械，需要 FDA 批准
欧盟 **GDPR** (right to explanation)

作为其它科学研究的工具

神经网络已经在众多科学领域比如 生物信息，天文，甚至 社会科学 取得了惊人的效果
科学研究是为了发现**新知识**，可解释性可以用来揭示它

一、事后解释 vs. 主动干预

Passive (post hoc) vs. **Active** (interpretability intervention)

是否在模型的 架构设计 或者 训练过程 中进行干预

二、所产生的解释的表现形式

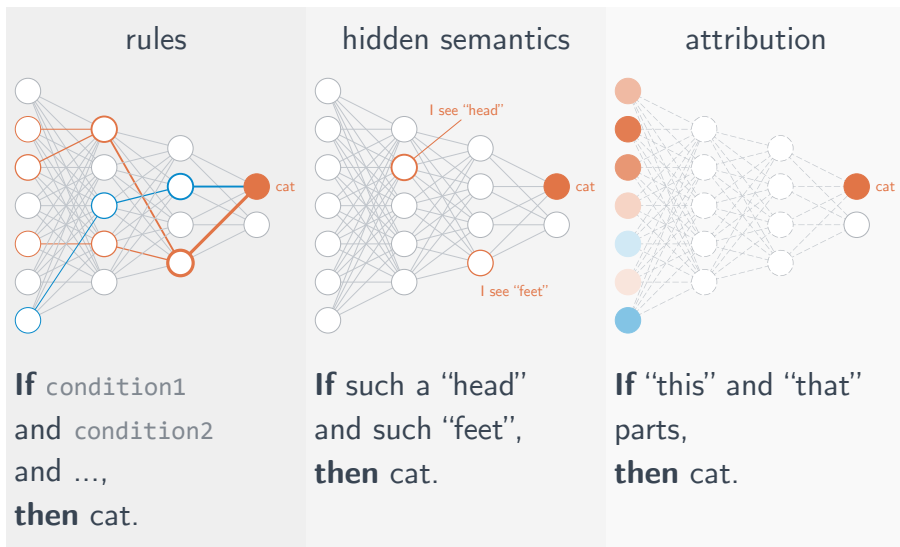
Types/Formats of explanations

比如 逻辑规则、显著性图 等，稍后会详细介绍

三、解释的「覆盖」范围 (w.r.t. *the input space*)

Local/simi-local/global interpretability

比如解释 单独一个输入 或者 整个模型



by examples

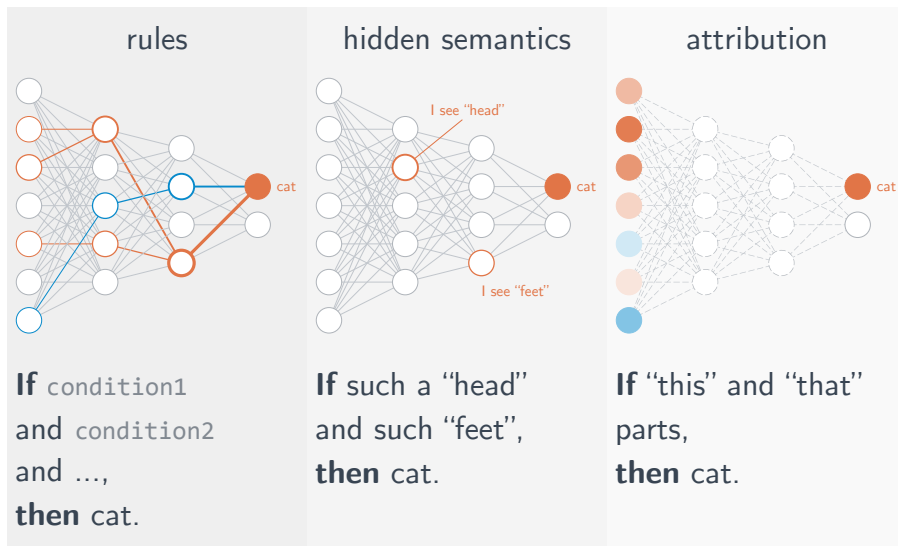
If all the same as "this example", then cat (or not).

解释的表现形式

Dimension 2

Question: Why do you think it is a cat?

Answer:



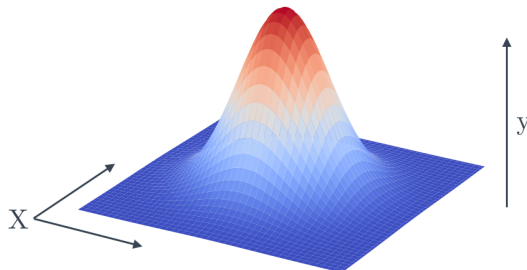
Explicit

Implicit

全局 vs. 局部解释

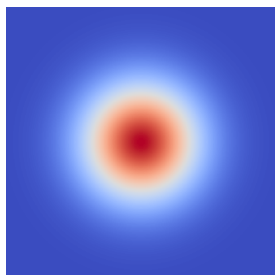
Dimension 3

Truth:

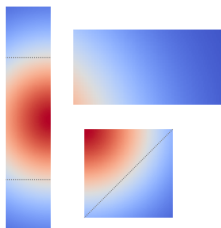


Explanations:

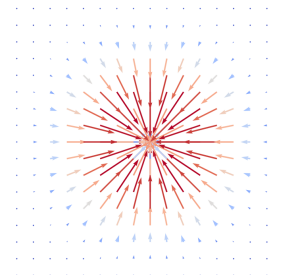
Global



Semi-local



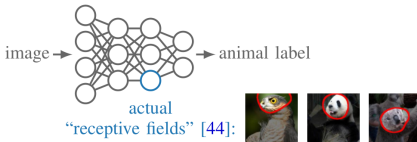




Local



If $x_1^2 + x_2^2 > 1, y = \blacksquare$

$\nabla f|_{x=x^{(i)}}$



	Local (and semi-local) interpretability applies to a certain input $\mathbf{x}^{(i)}$ (and its associated output $\hat{y}^{(i)}$), or a small range of inputs-outputs	Global interpretability w.r.t. the whole input space
Rule as explanation	<ul style="list-style-type: none"> The result "$\mathbf{x}^{(i)}$ is classified as $\hat{y}^{(i)}$" is because $\mathbf{x}_1, \mathbf{x}_4, \dots$ are present and $\mathbf{x}_3, \mathbf{x}_5, \dots$ are absent [38]. (Semi-local) For \mathbf{x} in the neighbourhood of $\mathbf{x}^{(i)}$, if $(\mathbf{x}_1 > \alpha) \wedge (\mathbf{x}_3 < \beta) \wedge \dots$, then $y = \hat{y}^{(i)}$ [34]. 	<p>The neural network can be approximated by</p> $\left\{ \begin{array}{l} \text{If } (\mathbf{x}_2 < \alpha) \wedge (\mathbf{x}_3 > \beta) \wedge \dots, \text{ then } y = 1, \\ \text{If } (\mathbf{x}_1 > \gamma) \wedge (\mathbf{x}_5 < \delta) \wedge \dots, \text{ then } y = 2, \\ \dots \\ \text{If } (\mathbf{x}_4 \dots) \wedge (\mathbf{x}_7 \dots) \wedge \dots, \text{ then } y = M \end{array} \right.$
Explaining hidden semantics (make sense of certain hidden neurons/layers)	<p>*Some local attribution methods (see below) can be easily modified to "explain" a hidden neuron rather than the final output.</p>	<p>Tries to explain a hidden neuron/layer etc.</p> <ul style="list-style-type: none"> An example active method [39] adds a special loss term that encourages filters to learn consistent and exclusive patterns (e.g. head patterns of animals) 
Attribution as explanation	<p>For $\mathbf{x}^{(i)}$:  → neural net → $\hat{y}^{(i)}$: junco bird</p> <p>The "contribution"¹ of each pixel:  [45]</p> <p>a.k.a. saliency map, which can be computed by different methods like gradients [40], sensitivity analysis² [41] etc.</p>	<p>(n.b. For a linear model, the coefficients is the global attribution to its input features.)</p> <ul style="list-style-type: none"> Kim et al. [42] calculate attribution to a target "concept" rather than the input pixels of a certain input. For example, how sensitive is the output (a prediction of zebra) to a concept (the presence of stripes)?
Explanation by showing examples	<p>For $\mathbf{x}^{(i)}$:  → neural net → $\hat{y}^{(i)}$: fish</p> <p>By asking how much the network will change $\hat{y}^{(i)}$ if removing a certain training image, we can find:</p> <p>most helpful² training images:  [43]</p>	<ul style="list-style-type: none"> Adds a (learnable) prototype layer to the network. Every prototype should be similar to at least an encoded input. Every input should be similar to at least a prototype. The trained network explains itself by its prototypes. [46]

Passive. Rules

Global—i.e. Rule extraction

Rule Format

- ▶ Propositional logic rule ... , L Fu, 1991, GG Towell & JW Shavlik, 1993
- ▶ First-order logic rule R Nayak, 2009
- ▶ Fuzzy logic rule ... , S Mitra & Y Hayashi, 2000, JL Castro et al., 2002

Methods

Assume we are interpreting a network f parameterized by Θ

- ▶ **Decompositional** approaches
(extract rules from network weights Θ)
- ▶ **Pedagogical** approaches (e.g. O Boz, 2002)
(extract rules from new training set $\{x^i, \hat{y}^i\}^N$, $\hat{y} = f(x; \Theta)$)
classic rule learning/decision tree learning algorithms can be used.
e.g. CART, C4.5

Passive. Rules

Local

e.g. 📄 “Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives”. *NeurIPS*. [A Dhurandhar et al., 2018](#)

针对某个输入 x ，尝试寻找如下形式的解释

因为特征 x_i, \dots, x_k 存在 (sufficiently present),
并且特征 x_m, \dots, x_p 不存在 (necessarily absent),
所以 y 被预测为某一类

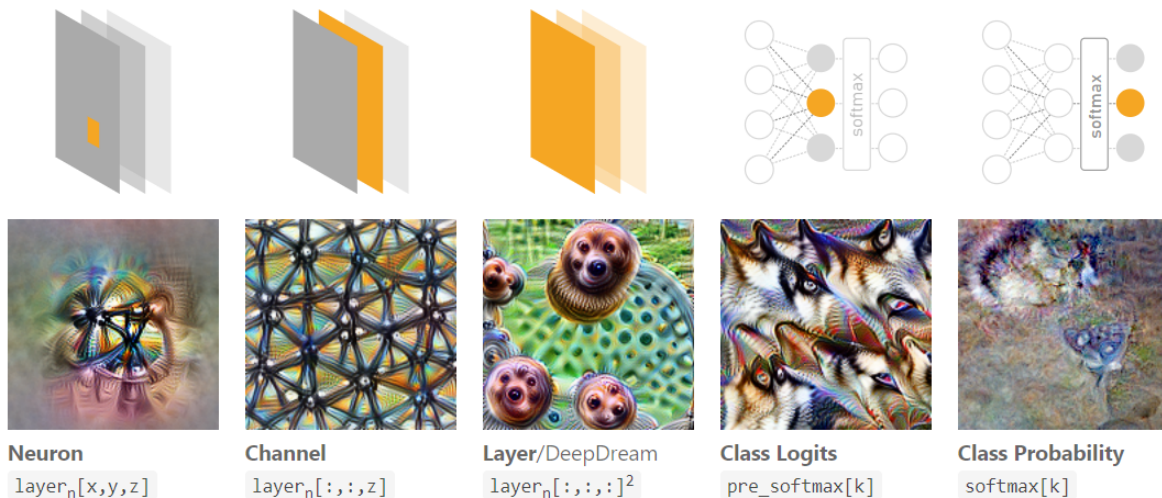
其它方法例如

- ▶ Anchors [MT Ribeiro et al., 2018](#)
- ▶ Interpretable partial substitution [T Wang, 2019](#)

Passive. Hidden Semantics

A Typical Method in CV—Visualization

Make sense of certain hidden neurons/layers



Source: [Feature Visualization - Distill](#)

Passive. Hidden Semantics

Visualization

Idea: find a **representative input** that a neuron/layer is looking for

Method: activation maximization

$$\arg \max_x act(x) - \lambda \Omega(x)$$

Problem: found input/patterns are **unrealistic** and **unrecognizable**

Solution: find a good regularizer/image prior

- ▶ L_2 norm K Simonyan et al., 2013
- ▶ *Total variation* (low-pass filter) A Mahendran & A Vedaldi, 2015
encourage neighbouring pixels to have similar values
- ▶ Clipping pixels with small norm or small contribution
J Yosinski et al., 2015
- ▶ Generative network of a GAN A Nguyen et al., 2016



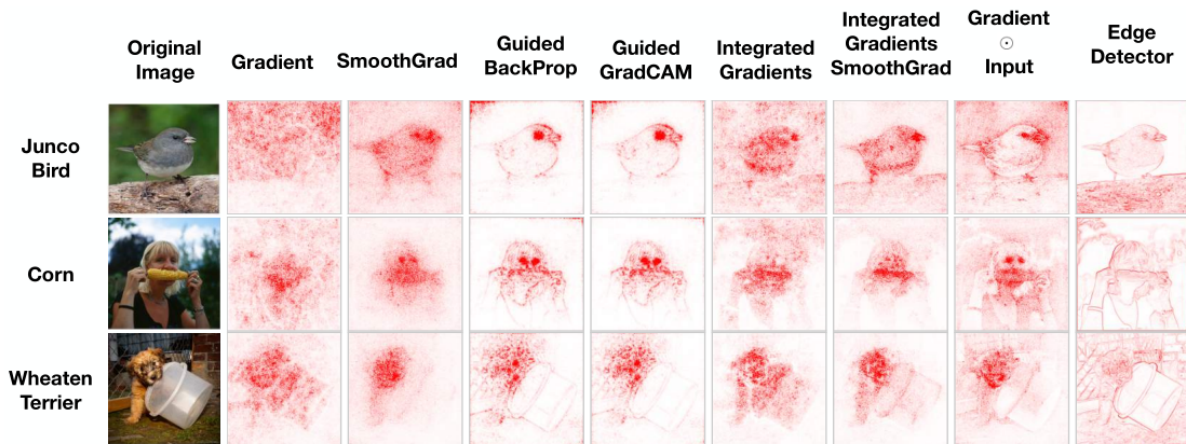
Passive. Hidden Semantics

其它方法例如

- ▶ Network Dissection [D Bau et al., 2017](#)
- ▶ Net2Vec [R Fong & A Vedaldi, 2018](#)
- ▶ “Analyzing individual neurons in deep NLP models”
[F Dalvi et al., 2019](#)

Passive. Attribution

Usually Local



“Sanity Checks for Saliency Maps”. *NeurIPS*. [J Adebayo et al., 2018](#)

注：线性代理模型 (proxy models) 本质上也是提供 attribution 解释

Passive. Attribution

Usually Local

For individual predictions, **attribution methods** try to identify **which attributes** (e.g. pixels) **contribute most** (or least) to *a single prediction*.

Intuitively, instead of interpreting f , it tries to interpret $f|_{x=x^i}$ for each sample x^i

Two representative categories: gradient-related methods and Model agnostic attribution

Gradient-related Methods

- ▶ (Immediate) gradient MT Ribeiro et al., 2016
- ▶ Discrete gradient A Binder et al., 2016, A Shrikumar et al., 2017
between a *reference input* x^{ref} and the target input x^i
- ▶ Integrated gradient M Sundararajan et al., 2017
the *path integral* of all the gradients between x^{ref} and x^i



Passive. Attribution

Local, Model Agnostic

e.g. Shapley Value

Shapley value is a solution to a game theory problem how to fairly assign payoff for each player in a coalition.

If we have a set of players N , $S \subseteq N$, $v(S)$ is the total gain of the set of players S . Then, the payoff for player i is

$$\phi_i(v) = \frac{1}{|N|} \sum_{S \subseteq N \setminus \{i\}} \binom{|N| - 1}{|S|}^{-1} (v(S \cup \{i\}) - v(S))$$

$v(S \cup i) - v(S)$ is the marginal contribution of player i to coalition S .
The rest of the formula is a normalization factor.

The practical problem: the computational complexity (exponential).

Some approximation methods are needed.


..., S Lundberg & S Lee, 2017



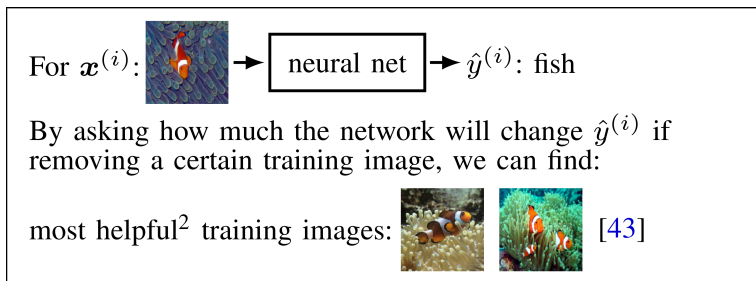
Passive. By Showing Examples

Usually Local

一种直观方法：寻找和待解释的 input 最「相似」的一个训练样本
(在模型的 inner representation 层面上相似) [R Caruana et al., 1999](#)

e.g.  “Understanding black-box predictions via influence functions”.
[ICML. PW Koh & P Liang, 2017](#)

改变一个 训练样本 会影响 模型权值，进而影响对 测试样本 的预测
而用 influence functions 可以找出对某个 target test input 影响最大的
训练样本



Active Interpretability Intervention

区别于前面 passive (post hoc) 方法, active 方法会主动地在模型的架构设计 或者 训练过程 中加入约束, 也即 interpretability loss

Rules as Explanations

e.g. 📄 “Beyond sparsity: Tree regularization of deep models for interpretability”. AAAI. M Wu et al., 2018

想法: 希望模型 $f_{\mathbf{W}}: \mathcal{X} \rightarrow \mathcal{Y}$ 能容易地被浅层的决策树拟合

$$\min_{\mathbf{W}} \left(\sum \text{Loss}(y, f(x; \mathbf{W})) + \lambda \cdot \text{TreeReg}(\mathbf{W}) \right)$$

其中 `TreeReg` 表示 能够基本近似该网络的决策树 的平均**树深度**

之后还有续作 “[Regional Tree Regularization ...](#)”. AAAI. M Wu et al., 2020
提供 semi-local 可解释性

Active. Hidden Semantics

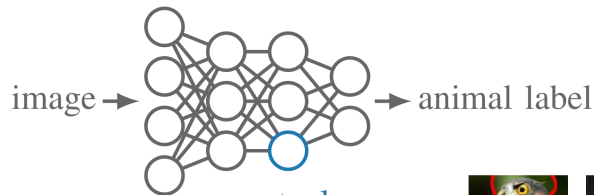
e.g. 📄 “Interpretable convolutional neural networks”. *CVPR*.

Q Zhang et al., 2018

核心思想：使高层的每个 filter 尽量只表示一种概念 (concept)
makes a filter to either have a consistent activation pattern or keep inactivated

Tries to explain a hidden neuron/layer etc.

- An example active method [39] adds a special loss term that encourages filters to learn consistent and exclusive patterns (e.g. head patterns of animals)



actual
“receptive fields” [44]:



Active. Attribution

Local

- ▶ ExpO [G Plumb et al., 2020](#)
在训练时约束模型，希望有助于得到更 *fidelitous*, *stable* 的 attribution
- ▶ DAPr [E Weinberger et al., 2020](#)
在做 attribution 时加入一些（粗略的）领域先验 (priors)



Global

- ▶ Dual-net [M Wojtas & K Chen, 2020](#)
同时训练两个网络，selector 网络用于选择 feature set，operator 网络则使用该 feature set 来完成预测任务

💬 这三篇文章都发表于 *NeurIPS* 2020

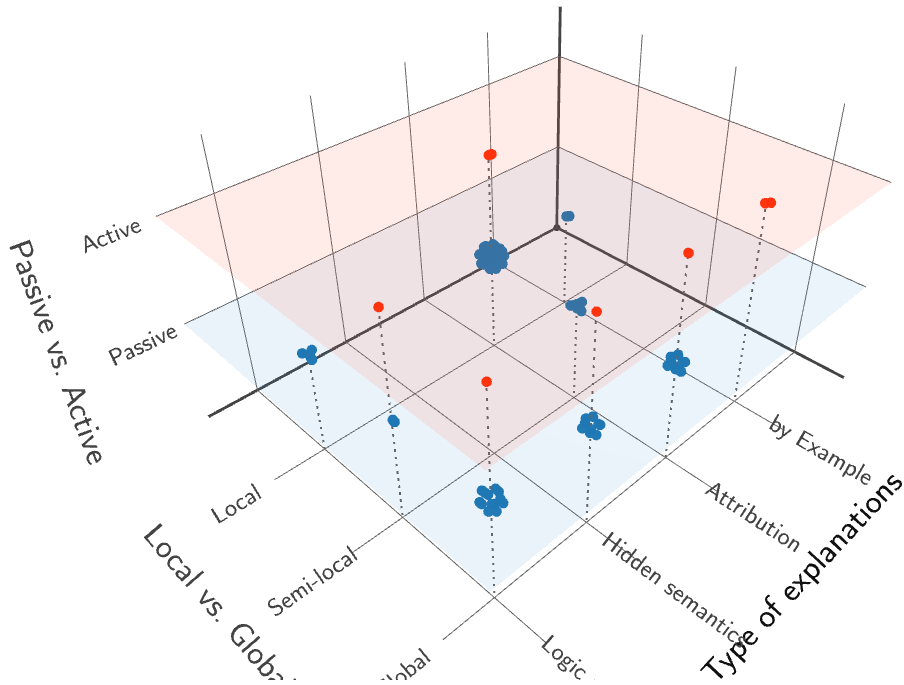


Active. By Showing Examples

- ▶  “Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions”. *AAAI*. [O Li et al., 2018](#)
Adds a **prototype** layer to an autoencoder
- ▶  “This looks like that: deep learning for interpretable image recognition”. *NeurIPS*. [C Chen et al., 2019](#)
Adds a **prototype** layer to a CNN

The Interpretability Paper Space

Hide colored panes (for better hover information)



Q&A

Takeaways

- ▶ Interpretability of a DNN vs. DL theory
- ▶ **What** is interpretability: an explanation-centred definition
- ▶ **Why** interpretability:
 - ▶ High reliability - Ethical/legal requirements - Scientific usages
- ▶ **How** to get interpretability: A 3D Taxonomy
 - ▶ Passive vs. Active - Types/formats of explanations - Local/semi-local/global interpretability

Links

- ▶ [Interpretability Paper Space \(online, interactive\)](#) 
- ▶ [arXiv link \(2012.14261\)](#) 
- ▶ [神经网络可解释性综述 — 知乎](#) 



References

- [1] Aravindh Mahendran, Andrea Vedaldi. Understanding deep image representations by inverting them. *CVPR*. 2015.
- [2] Anh Nguyen, Alexey Dosovitskiy, ..., Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *NIPS*. 2016.
- [3] Alexander Binder, Grégoire Montavon, ..., Wojciech Samek. Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers. *Artificial Neural Networks and Machine Learning – ICANN*. 2016.
- [4] Avanti Shrikumar, Peyton Greenside, Anshul Kundaje. Learning Important Features Through Propagating Activation Differences. *ICML*. 2017.
- [5] Amit Dhurandhar, Pin-Yu Chen, ..., Payel Das. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. *NIPS*. 2018.
- [6] Chaofan Chen, Oscar Li, ..., Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *NeurIPS*. 2019.
- [7] David Bau, Bolei Zhou, ..., Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. *CVPR*. 2017.
- [8] Dino Pedreschi, Fosca Giannotti, ..., Franco Turini. Meaningful explanations of Black Box AI decision systems. *AAAI*. 2019.

References

- [9] Ethan Weinberger, Joseph Janizek, Su-In Lee. Learning Deep Attribution Priors Based On Prior Knowledge. *NeurIPS*. 2020.
- [10] Finale Doshi-Velez, Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv*. 2017.
- [11] Fahim Dalvi, Nadir Durrani, ..., James Glass. What is one grain of sand in the desert? Analyzing individual neurons in deep nlp models. *AAAI*. 2019.
- [12] Geoffrey G. Towell, Jude W. Shavlik. Extracting refined rules from knowledge-based neural networks. *Machine Learning*. 1993.
- [13] Gregory Plumb, Maruan Al-Shedivat, ..., Amee Talwalkar. Regularizing Black-box Models for Improved Interpretability. *NeurIPS*. 2020.
- [14] J. L. Castro, C. J. Mantas, J. M. Benitez. Interpretation of artificial neural networks by means of fuzzy rules. *IEEE Transactions on Neural Networks*. 2002.
- [15] Jason Yosinski, Jeff Clune, ..., Hod Lipson. Understanding Neural Networks Through Deep Visualization. *ICML Deep Learning Workshop*. 2015.
- [16] Julius Adebayo, Justin Gilmer, ..., Been Kim. Sanity Checks for Saliency Maps. *NIPS*. 2018.

References

- [17] Karen Simonyan, Andrea Vedaldi, Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv*. 2013.
- [18] Limin Fu. Rule Learning by Searching on Adapted Nets. *AAAI*. 1991.
- [19] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. Why Should I Trust You?: Explaining the Predictions of Any Classifier. *KDD*. 2016.
- [20] Mukund Sundararajan, Ankur Taly, Qiqi Yan. Axiomatic Attribution for Deep Networks. *ICML*. 2017.
- [21] Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. Anchors: High-precision model-agnostic explanations. *AAAI*. 2018.
- [22] Mike Wu, Michael C Hughes, ..., Finale Doshi-Velez. Beyond sparsity: Tree regularization of deep models for interpretability. *AAAI*. 2018.
- [23] Mike Wu, Sonali Parbhoo, ..., Finale Doshi-Velez. Regional Tree Regularization for Interpretability in Deep Neural Networks.. *AAAI*. 2020.
- [24] Maksymilian Wojtas, Ke Chen. Feature Importance Ranking for Deep Learning. *NeurIPS*. 2020.

References

- [25] Olcay Boz. Extracting Decision Trees from Trained Neural Networks. *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2002.
- [26] Oscar Li, Hao Liu, Chaofan Chen, Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. *AAAI*. 2018.
- [27] Pang Wei Koh, Percy Liang. Understanding black-box predictions via influence functions. *ICML*. 2017.
- [28] Quanshi Zhang, Ying Nian Wu, Song-Chun Zhu. Interpretable convolutional neural networks. *CVPR*. 2018.
- [29] Ronen Eldan, Ohad Shamir. The power of depth for feedforward neural networks. *Conference on learning theory*. 2016.
- [30] Riccardo Guidotti, Anna Monreale, ..., Dino Pedreschi. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*. 2018.
- [31] Richi Nayak. Generating rules with predicates, terms and variables from the pruned neural networks. *Neural Networks*. 2009.
- [32] Ruth Fong, Andrea Vedaldi. Net2Vec: Quantifying and Explaining How Concepts Are Encoded by Filters in Deep Neural Networks. *CVPR*. 2018.

References

- [33] Rich Caruana, Hooshang Kangarloo, ..., David Johnson. Case-based explanation of non-case-based learning methods. *Proceedings of the AMIA Symposium*. 1999.
- [34] S. Mitra, Y. Hayashi. Neuro-fuzzy rule generation: survey in soft computing framework. *IEEE Transactions on Neural Networks*. 2000.
- [35] Scott Lundberg, Su-In Lee. A Unified Approach to Interpreting Model Predictions. *NIPS*. 2017.
- [36] Tong Wang. Gaining Free or Low-Cost Interpretability with Interpretable Partial Substitute. *ICML*. 2019.