Model Interpretability

Todo List

1. Smilkov, Daniel, et al. "Smoothgrad: removing noise by adding noise." *arXiv preprint arXiv:1706.03825* (2017).
2. Bach S, Binder A, Montavon G, et al. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation [J]. PloS one, 2015, 10(7): e0130140.
3. Guo, Wenbo, et al. "Lemna: Explaining deep learning based security applications." *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018.
4. Tao Guanhong, Ma Shiqing, Liu Yingqi, et al. Attacks meet interpretability: Attribute-steered detection of adversarial samples [C] //Proc of the 32st Int Conf on Neural Information Processing Systems. USA: Curran Associates Inc., 2018: 7717-7728
5. Liu Ninghao, Yang Hongxia, Hu Xia. Adversarial detection with model interpretation [C] //Proc of the 24th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining. New York: ACM, 2018: 1803-1811
6. Carlini N, Wagner D. Towards evaluating the robustness of neural networks [C] //Proc of the 38th IEEE Symposium on Security and Privacy. Piscataway, NJ: IEEE, 2017: 39-57
7. Papernot N, Mcdaniel P, Jha S, et al. The limitations of deep learning in adversarial settings [C] //Proc of the 1st IEEE European Symp on Security and Privacy. Piscataway, NJ: IEEE, 2016: 372-387
8. Papernot N, Mcdaniel P, Goodfellow I, et al. Practical blackbox attacks against machine learning [C] //Proc of the 12th ACM Asia Conf on Computer and Communications Security. New York: ACM, 2017: 506-519
9. Li Jinfeng, Ji Shouling, Du Tianyu, et al. TextBugger: Generating Adversarial Text Against Real-world Applications [C] //Proc of 26th Annual Network and Distributed Systems Security Symp. Reston, VA: ISOC, 2019
10. Ghorbani A, Abid A, Zou J. Interpretation of neural networks is fragile [J]. arXiv preprint arXiv:1710.10547, 2017
11. Zhang Xinyang, Wang Ningfei, Ji Shouling, et al. Interpretable Deep Learning under Fire [C] //Proc of the 29th USENIX Security Symp. Berkele, CA: USENIX Association, 2020

机器学习模型可解释性方法、应用与安全研究综述

Notes

> 这篇是纪老师关于模型可解释性文章的综述，以下均为个人对文章做的笔记，建议自己阅读原文。另外纪老师在这方面有很多不错的工作，值得关注。

机器学习可解释性问题

1. **模型可解释性问题**：可解释性旨在帮助人们理解机器学习模型是如何学习的，它从数据中学到了什么，针对每一个输入它为什么做出如此决策以及它所做的决策是否可靠；
2. **模型的复杂度与模型准确性相关联，又与模型的可解释性相对立。**
3. 根据**选择结构简单易于解释的模型然后训练它**，还是**训练复杂的最优模型然后开发可解释性技术解释它**，将机器学习模型可解释性总体上分为：**ante-hoc 可解释性**和 **post-hoc 可解释性**；

* Ante-hoc 可解释性

1. Ante-hoc 可解释性指**模型本身内置可解释性**，即对于一个已训练好的学习模型，我们无需额外的信息就可以理解模型的决策过程或决策依据；
2. 自解释模型：

   (1) 可模拟性：在一定时间可以预测模型的每一步计算；

   (2) 可分解性：模型的每个部分都可以得到一个直观的解释；

   (3) 结构简单：由于人类认知的局限性，自解释模型的内置可解释性受模型的复杂度制约，这要求自解释模型结构一定不能过于复杂；
3. 广义加性模型：在简单模型和复杂问题之间的一个折中；

$$g(y) = f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n)$$

4. 注意力机制：通过 Attention 的权重来分析模型关注的重点是什么；

Post-hoc 可解释性

1. Post-hoc 可解释性的重点在于设计高保真的解释方法或构建高精度的解释模型，根据解释目的和解释对象的不同，可分为全局可解释性和局部可解释性；
2. 经典的解释方法如下：

Table 1 Summary of classic post-hoc interpretation methods.
表 1 经典的 post-hoc 解释方法总结

Method	G/L	MA/MS	TML	FCN	CNN	RNN	Fidelity	Security	Domain
inTree [23]	G	MS	✓	×	×	×	○	-	n/a
SGL [47]	G	MS	✓	×	×	×	○	-	n/a
GIRP [53]	G	MA	✓	✓	✓	✓	○	×	CV/NLP
MAGIX [58]	G	MA	✓	✓	✓	✓	○	-	n/a
DeepVID [70]	G	MA	×	×	✓	×	○	×	CV
AM [75]	G	MS	×	✓	✓	×	◐	×	CV
Nguyen et al. [79]	G	MS	×	✓	✓	×	◐	×	CV
Yuan et al. [82]	G	MS	×	×	×	✓	◐	×	NLP
Saliency Mask [93]	L	MA	×	✓	✓	×	○	×	CV
RSRS [94]	L	MA	×	✓	✓	×	○	×	CV
LIME [13]	L	MA	✓	✓	✓	✓	◐	×	CV/NLP
LORE [96]	L	MA	✓	✓	✓	✓	○	×	n/a
Anchor [98]	L	MA	✓	✓	✓	✓	◐	×	CV/NLP
LEMNA [99]	L	MS	×	×	×	✓	◐	×	NLP/Malware
Grad [73]	L	MS	×	✓	✓	✓	○	×	CV/NLP
DeconvNet [80]	L	MS	×	×	✓	×	◐	×	CV
GuidedBP [100]	L	MS	×	×	✓	×	◐	×	CV
Integrated [101]	L	MS	×	✓	✓	✓	○	×	CV/NLP
SmoothGrad [102]	L	MS	×	✓	✓	✓	◐	×	CV/NLP
LRP [105]	L	MS	×	✓	✓	✓	◐	×	CV/NLP
DeepLIFT [106]	L	MS	×	✓	✓	✓	◐	×	CV/Genomics
Guided Inversion [103]	L	MS	×	×	✓	×	◐	✓	CV
CAM [112]	L	MS	×	×	✓	×	◐	×	CV
Grad-CAM [113]	L	MS	×	×	✓	×	◐	×	CV
AI^2 [115]	L	MS	×	×	✓	×	○	✓	CV
OpenBox [116]	G, L	MS	×	✓	×	×	●	✓	CV

Note：G = global, L = local, MA = model-agnostic, MS = model-specific, TML = traditional machine learning, ○ = low, ◐ = middle, ● = high, - = unknown, CV = computer vision, NLP = natural language processing, and n/a = not mentioned in the literature.

3. 全局解释：全局可解释性旨在帮助人们从整体上理解模型背后的复杂逻辑以及内部的工作机制，如模型是如何学习的、模型从训练数据中学到了什么、模型是如何进行决策的 等；

   (1) * 规则提取：通过受训模型中提取解释规则的方式，提供对复杂模型尤其是黑盒模型整体决策逻辑的理解（比较早期的做法）；

   (2) 模型蒸馏：

   - 定义：通过**降低模型复杂度**，解决理解受训模型比较困难的问题，是一种经典的**模型压缩方法**；
   - 核心思想：利用**结构紧凑的学生模型来模拟结构复杂的教师模型**，从而完成从教师模型到学生模型的知识迁移过程，实现对复杂教师模型的知识的 "蒸馏"；
   - 训练损失函数定义：

     $$L_{student} = \alpha L^{(soft)} + (1-\alpha) L^{(hard)}$$

     其中，$L^{(soft)}$ 为**软目标损失**，期望学生模型能够学到教师模型相似的概率分布输出；$L^{(hard)}$ 为**硬目标损失**，要求学生模型能够保留教师模型决策的类别；
   - 模型蒸馏解释方法实现简单，易于理解，且不依赖待解释模型的具体结构信息，因而作为一种模型无关的解释方法，常被用于解释黑盒机器学习模型；
   - 蒸馏模型只是对原始复杂模型的一种全局近似，基于蒸馏模型所做出的解释不一定能够反映待解释模型的真实形为；

> 我的想法：1. 能否通过模型蒸馏的方法去生成一个黑盒模型的替代模型，从而去辅助生成对抗样本？（这个关键在于生成的替代模型能否较好地逼近黑盒模型）2. 模型蒸馏的方法，最终能够得到怎样的模型解释？能否通过模型蒸馏的方法来解释现有的语音识别模型？（这个关键在于如何去分析蒸馏以后的模型）

(3) 激活最大化 (Activation Maximization)：

- 定义：通过在特定的层上**找到神经元的首选输入来最大化神经元激活**，来理解 DNN 中每一层隐含层神经元所捕获的表征；
- 核心思想：通过寻找有界范数的输入模式，最大限度地激活给定地隐藏神经元，而一个单元最大限度地响应的输入模式可能是 ”一个单元正在做什么的“ 良好的一阶表示；
- 形式化定义：

  $$x^* = \arg\max_x f_l(x) - \lambda\|x\|^2$$

  其中左边项期望 x 能够使得当前神经元的激活值最大；右边项期望 x 与原样本尽可能接近（右边应该改用 Δx 来表示）；
- 激活最大化解释方法是一种模型相关的解释方法，相比规则提取解释和模型蒸馏解释，其解释结果更准确，更能反映待解释模型的真实形为；
- 激活最大化本身是一个优化问题，在通过激活最大化寻找原型样本的过程中，优化过程中的噪音和不确定性可能导致产生的原型样本难以解释；
- 激活最大化解释方法**难以用于解释自然语言处理模型和图神经网络模型**；

4. 局部解释：模型的局部可解释性以输入样本为导向，通常可以通过分析输入样本的每一维特征对模型最终决策结果的贡献来实现。

(1) 敏感性分析 (Sensitivity Analysis)：

- 核心思想：通过**逐一改变自变量的值来解释因变量受自变量变化影响大小**的规律；
- 根据**是否需要利用模型的梯度信息**，敏感性分析方法可分为模型相关方法和模型无关方法；
- **模型相关方法**：利用模型的局部梯度信息评估特征与决策结果的相关性，常见的相关性定义如下：

  $$R_i(x) = (\frac{\partial f(x)}{\partial x_i})^2$$

  即为模型梯度的 l_2 范数分解；
- 模型无关方法：无需利用模型的梯度信息，只关注**待解释样本特征值变化对模型最终决策结果的影响。**具体地，该方法通过观察去掉某一特定属性前后模型预测结果的变化来确定该属性对预测结果的重要性，即：

  $$R_i(x) = f(x) - f(x \setminus x_i)$$
- 敏感性分析方法提供的解释结果通常**相对粗糙且难以理解**，且无法分析**多个特征之间的相关关系**；

(2) 局部近似：

- 核心思想：利用**结构简单的可解释模型拟合待解释模型针对某一输入实例的决策结果**，然后基于解释模型对该决策结果进行解释；
- 基于局部近似的解释方法实现简单，易于理解且不依赖待解释模型的具体结构，适于**解释黑盒机器学习模型**；
- 对于每个输入样本都需要训练一个解释模型，**效率不高**，并且**该方法基于特征相互独立的假设**；

(3) ☆ 反向传播：

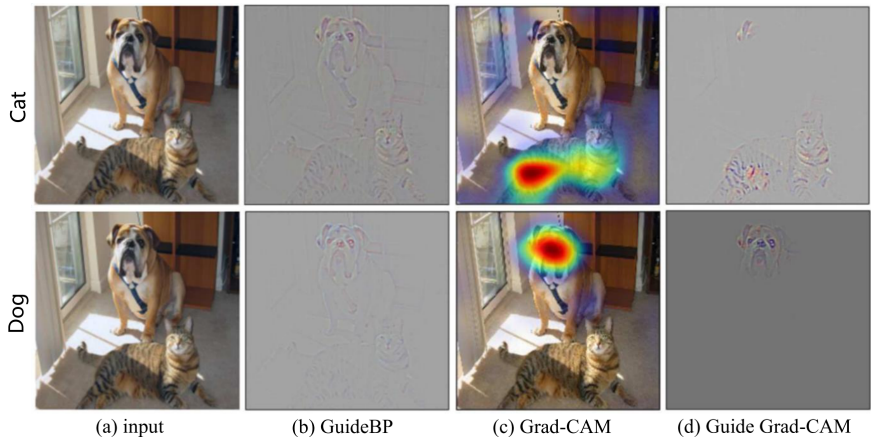
- 核心思想：利用 **DNN 的反向传播机制**将模型的决策重要性信号从模型的输出层神经元逐层传播到模型的输入以推导输入样本的特征重要性；
- 其中，**SmoothGrad 方法**的核心思想：通过向待解释样本中添加噪声对相似的样本进行采样，然后利用反向传播方法求解每个采样样本的决策显著图，最后将所有求解得到的显著图进行平均并将其作为对模型针对该样本的决策结果的解释；
- **分层相关性传播(LRP)方法**的核心思想：利用反向传播将高层的相关性分值递归地传播到底层直至传播到输入层；
- 基于反向传播地解释方法通常实现**简单、计算效率高且充分利用了模型的结构特性**；
- 如果预测函数在输入附近变得平坦，那么预测函数相对于输入的梯度在该输入附近将变得很小，进而导致无法利用梯度信息定位样本的决策特征；

> 反向传播的想法应该可以在语音领域进行尝试。

(4) * 特征反演 (Feature Inversion)：

- 定义：特征反演作为一种可视化和理解 DNN 中间特征表征的技术，可以充分利用模型的中间层信息，以提供对模型整体行为及模型决策结果的解释；
- 特征反演解释方法分为：模型级解释方法和实例级解释方法；

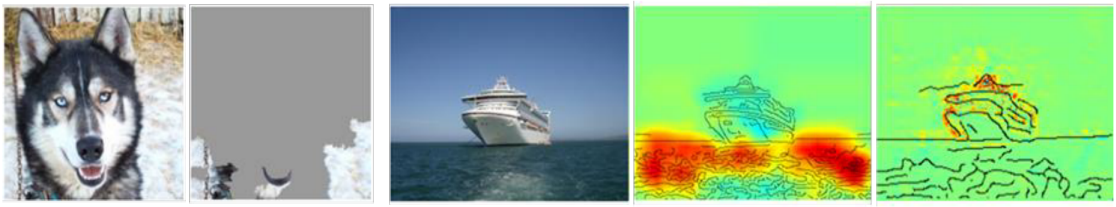
(5) 类激活映射：



- ☆ 前提：**CNN 不同层次的卷积单元包含大量的位置信息，使其具有良好的定位能力。**然而，传统 CNN 模型通常在卷积和池化之后采用全连接层对卷积层提取的特征图进行组合用于最终的决策，因而**导致网络的定位能力丧失**，这个问题需要在这些工作中解决。
- **类激活映射 (Class Activation Mapping, CAM) 解释方法**：利用全局平均池化 (Global Average Pooling)层来替代传统 CNN 模型中除 softmax 层以外的所有全连接层，并通过将输出层的权重投影到卷积特征图来识别图像中的重要区域 => (需要**用全局平均池化层替换模型中的全连接层并重新训练**)。
- **梯度加权类激活映射 (Grad-CAM) 解释方法**：给定一个输入样本，Grad-CAM 首先计算目标类别相对于最后一个卷积层中每一个特征图的梯度并对梯度进行全局平均池化，以获得每个特征图的重要性权重；然后，基于重要性权重计算特征图的加权激活，以获得一个粗粒度的梯度加权类激活图，用于定位输入样本中具有类判别性的重要区域 => (**不需要修改网络后进行重训练**)；
- **导向梯度加权类激活 (Guided Grad-CAM) 解释方法**：即将 GuidedBP 方法和 Grad-CAM 方法进行结合；

可解释性应用

1. 模型验证：由于数据集中可能存在偏差，并且验证集也可能与训练集同分布，我们很难简单地通过评估模型在验证集上的泛化能力来验证模型的可靠性，也很难验证模型是否从训练数据中学到了真正的决策只是。这里**通过可解释性方法，去分析模型在做决策时到底更加关注哪些特征，这些特征是否是合理的。**如下：

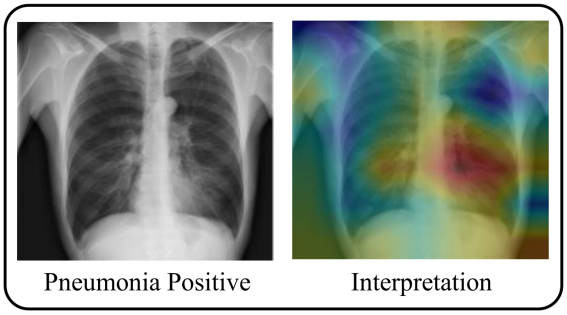


(a) validation of husky classifier[13] (b)validation of boat classifier[117]

2. * 模型诊断：诊断模型中的缺陷；

> 这个东西感觉不太靠谱，模型可解释性的方法很大程度上并不能得到很直观的一个解释，又如何依靠模型可解释性来做模型的诊断。

3. * 辅助分析：如医疗领域，使用可解释性方法来辅助医护人员进行检查；



4. 知识发现：辅助人学习基于大量数据训练的模型中的知识；
    - **LEMNA** 解释方法可以挖掘出检测模型从数据中学到的新知识；

可解释性与安全性分析

> 应该仔细思考：**可解释性方法和安全领域的相关关系，从而去发掘新的应用场景、攻击场景、防御手段等。**

1. 安全隐患消除：模型可解释性方法用于检测对抗样本，并增强模型的鲁棒性；
2. 安全隐患：**攻击者也同样可以利用模型可解释性的方法来生成更好的对抗样本**；
3. ☆ 自身安全问题：由于采用了近似处理或是基于优化手段，大多数解释方法只能提供近似的解释，因而解释结果与模型的真是形为之间存在一定的不一致性
    - 不改变模型的决策结果的前提下，使解释方法解释出错：

$$\arg\max_{\delta} D(I(x_t;N), I(x_t+\delta;N))$$
$$s.t. \|\delta\|_\infty \leq \epsilon, f(x_t+\delta) = f(x_t)$$

      其中，$I(x_t;N)$ 为解释系统对神经网络 N 针对样本 x_t 决策结果 $f(x_t)$ 的解释。
    - 模型结果出错，单不改变解释方法解释出错；

未来方向

1. ☆如何设计更精确、更友好的解释方法，消除解释结果与模型真实形为之间的不一致；
2. ☆如何设计更科学、更统一的可解释性评估指标，以评估可解释方法解释性能和安全性；

Links

- 论文链接：纪守领, et al. "机器学习模型可解释性方法, 应用与安全研究综述." *计算机研究与发展* 56.10 (2019).