深度学习导论

论文阅读报告

黄瑞轩 PB20111686

论文标题 Take 5: Interpretable Image Classification with a Handful of Features

论文来源 Progress and Challenges in Building Trustworthy Embodied AI @NeurIPS

发表时间 December 2022

Keywords

Interpretable AI Low-dimensional Models Image Classification

1 选阅理由与概述

机器学习可解释性是当前人工智能领域中一个非常重要的话题，涉及模型的可信度、公平性、安全性等多个方面。特别是在智能医疗、无人驾驶等领域，机器学习可解释性相关的研究变得越来越重要。随着机器学习应用场景的不断扩大和深入，越来越多的人开始关注机器学习模型的可解释性问题。我本人近期也在 MSRA 参加了模型可解释性分析相关的课题，对此领域的研究比较感兴趣。

本次阅读报告选读的论文 *Take 5: Interpretable Image Classification with a Handful of Features* 来自 2022 年 NeruIPS，这篇文章提出了一种在深度神经网络中使用可解释、稀疏和低维度决策层的方法，以实现对细粒度图像分类的可解释分析。

2 论文关注的问题

当前针对模型解释性分析的方法主要存在特征数量多、特征难理解（特征是经过深度神经网络计算出来的 deep feature）的问题。在细粒度图像分类任务中，由于图像的特征一般是各像素值，维度非常高，即便是经典的机器学习模型解释起来效果也不好。并且，即使应用于此任务经典机器学习模型已经有一些可解释的分析，但是在深度神经网络模型上并没有相关的分析。

论文从上面提出的特征数量多、特征难理解的现实问题出发，提出了解决问题的方向。因为人在同一时间可考虑的特征数量约为 5 至 9 个，论文希望将分类的维度平均减少至 5 个左右（稀疏性，利用 *glm-saga* 计算选定特征的稀疏线性分类器），这样更加有可能与人类理解对齐。并且，论文希望通过某种特征选择算法选择保留的特征——而不是通过计算来获得新的特征——来保留特征的原始语义，最大程度地保证特征的易理解性。除此之外，针对减少特征可导致可用于决策的总信息受限的问题（冗余特征），论文还提出了一种新的能够确保特征信息多样性的损失函数。

3 论文提出的方法

论文提出了解决问题的一整条 pipeline，即特征学习、特征选择、稀疏化和特征微调。在简单介绍各步骤的前置知识之后，再来观察整个的 pipeline 过程。

3.1 前置：*glm-saga*

论文利用 *glm-saga* 方法计算选定特征的稀疏线性分类器，这里简单介绍下 *glm-saga* 方法。假设预先计算（如正则化）了特征，*glm-saga* 利用之计算一系列稀疏线性分类器（被称为 regularization path）作为 fit 的过程。

$$P = \left[\left(\boldsymbol{W}_1^{\text{sparse}}, \boldsymbol{b}_1\right), \left(\boldsymbol{W}_2^{\text{sparse}}, \boldsymbol{b}_2\right), \dots, \left(\boldsymbol{W}_n^{\text{sparse}}, \boldsymbol{b}_n\right)\right]$$

这里 $\boldsymbol{W}_i$ 的稀疏程度随 i 递减。论文利用 *glm-saga* 在选定的具有更高多样性的特征上创建一个更可解释的模型，然后对特征微调以获得最终的特征表示。

3.2 前置：$\mathcal{L}_{div}$损失函数

目的：让不同的特征捕获不同的语义（可解释）性质。

计算定义：

$$\hat{s}_{i,j}^{l}=\frac{\exp\left(\boldsymbol{M}_{l,i,j}\right)}{\sum_{i'=1}^{h_M}\sum_{j'=1}^{w_M}\exp\left(\boldsymbol{M}_{l,i',j'}\right)}\frac{f_l}{\max \boldsymbol{f}}\frac{|w_{\hat{c}l}|}{\|\boldsymbol{w}_{\hat{c}}\|_2} \qquad \mathcal{L}_{div}=-\sum_{i=1}^{h_M}\sum_{j=1}^{w_M}\max_{1\le k\le n_f}\hat{s}_{i,j}^{k}$$

思路解释：损失函数设计是在整个特征图 $\boldsymbol{M}$ 上使用了 Cross-Channel-Max-Pooling。并使用 softmax 在空间维度上进行归一化。然后缩放这些图，使得它们通过保持 $\boldsymbol{M}$ 的相对均值来强调可见和重要的特征，同时根据预测的类别对它们进行加权。

3.3 前置：特征选择方法

使用改进的 *glm-saga* 方法来选择特征。如果某个$\left(\boldsymbol{W}_j^{\text{sparse}},\boldsymbol{b}_j\right)$用到的特征不在已经选出的特征中，就选择它（加入已选择的特征集合），然后重新运行 *glm-saga* 算法。论文从重要性计算算子方面改进了 *glm-saga*，原始版本的计算原理这里并不打算详细讨论。

3.4 pipeline：创建 SLDD-Model

任务假设：细粒度图像分类任务，经过特征学习后的特征向量为 $\boldsymbol{f}\in\mathbb{R}^{n_f}$，分类输出是 $\boldsymbol{y}=Model(\boldsymbol{f})\in\mathbb{R}^{n_c}$。根据要求，仅使用 $n_f^*\ll n_f$ 个特征，并且使用的 Model 是可解释的（线性层 $\boldsymbol{y}=\boldsymbol{W}\boldsymbol{f}+\boldsymbol{b},\boldsymbol{W}\in\mathbb{R}^{n_c\times n_f^*},\boldsymbol{b}\in\mathbb{R}^{n_c}$），这里的 $\boldsymbol{W}$ 要求非常稀疏。

SLDD-Model 训练的过程分为如下几步：

- 使用损失函数 $\mathcal{L}_{div}$ 训练一个神经网络
- 使用 *glm-saga* 和特征选择算法选择并描述特征（计算 regularization path）
- 微调其他层，使之和输出线性层的稀疏性相适应

4 论文实验的分析

论文在细粒度图像分类领域的四个常见基准数据集上验证了其方法，并且使用可视化方法展示了可解释性和精度的 tradeoff 结果。

为了展示选出的特征多样性，论文开发了衡量特征多样性的指标 diversity@k，其定义为 $\text{diversity@k}=\sum_{i=1}^{h_M}\sum_{j=1}^{w_M}\max_{1\le l\le k}s_{ij}^{l}$，因为论文提出要将每个类别相关的特征数$\left(n_{wc}=\frac{n_w}{n_c}\right)$减少到 5 个以内，所以设定 $k=5$。

4.1 实验概述与结果分析

论文进行了在 n_f 尺度上的非稀疏模型和用论文方法构建的 $n_{wc}\le 5$ 的稀疏模型（SLDD-Model）应用于细粒度图像分类（Resnet50 数据集）任务实验。

实验结果表明，最终层中可以获得很好的稀疏率 $\frac{5}{2048}$（特征数量减少了 97.6%，仅

产生了 50 个特征），相应精度仅降低 0.1～0.4%。论文提出的$\mathcal{L}_{div}$提高了所有稀疏模型的精度（Table 3 所示）。并且实验结果还展示了论文方法的泛用性。不过，论文方法应用于某些数据集上存在一定的不稳定性（Table 4 所示）。

	CUB-2011					FGVC-Aircraft					NABirds					Stanford Cars				
	$n_f^*=2048$			$n_f^*=50$		$n_f^*=2048$			$n_f^*=50$		$n_f^*=2048$			$n_f^*=50$		$n_f^*=2048$			$n_f^*=50$	
$\mathcal{L}_{\text{div}}$	Dense	Sparse	Finet.	Sparse	Finet.	Dense	Sparse	Finet.	Sparse	Finet.	Dense	Sparse	Finet.	Sparse	Finet.	Dense	Sparse	Finet.	Sparse	Finet.
✗	86.6	81.8	85.3	79.5	83.4	90.0	88.4	89.4	87.3	88.1	84.2	79.5	83.3	77.3	80.7	93.2	90.9	92.6	89.3	91.1
✓	86.6	**84.0**	**86.5**	**81.7**	**84.0**	**91.4**	**90.7**	**91.1**	**89.8**	**90.1**	**84.4**	**81.0**	**84.0**	**79.8**	**81.7**	**93.6**	**92.1**	**93.3**	**91.1**	**92.0**
MCL Chang et al. (2020)	86.1	81.9	85.1	79.4	82.8	90.1	88.4	89.0	87.2	88.1	-	-	-	-	-	93.1	91.0	92.5	89.0	90.7
FRL Zheng et al. (2020)	86.4	81.5	85.3	78.9	82.6	90.0	88.5	89.4	87.5	88.2	-	-	-	-	-	93.3	90.8	92.6	89.4	90.9

Table 3: Impact of the loss function on accuracy in percent for Resnet50. Best results are in bold.

	CUB-2011					FGVC-Aircraft					NABirds					Stanford Cars				
	$n_f^*=n_f$			$n_f^*=50$		$n_f^*=n_f$			$n_f^*=50$		$n_f^*=n_f$			$n_f^*=50$		$n_f^*=n_f$			$n_f^*=50$	
Backbone	Dense	Sparse	Finet.	Sparse	Finet.	Dense	Sparse	Finet.	Sparse	Finet.	Dense	Sparse	Finet.	Sparse	Finet.	Dense	Sparse	Finet.	Sparse	Finet.
DenseNet121	86.3	76.2	82.9	75.7	83.1	**91.5**	88.2	89.8	88.1	90.0	84.1	72.8	64.6	71.0	80.5	93.3	87.3	91.7	85.8	91.4
Inception-v3	82.3	78.0	80.3	74.0	78.3	88.9	87.5	88.1	85.9	87.4	79.0	75.8	77.3	73.1	76.5	91.5	88.9	90.3	86.3	89.4
Resnet50	**86.6**	**84.0**	**86.5**	**81.7**	**84.0**	91.4	**90.7**	**91.1**	**89.8**	**90.1**	**84.4**	**81.0**	**84.0**	**79.8**	**81.7**	**93.6**	**92.1**	**93.3**	**91.1**	**92.0**

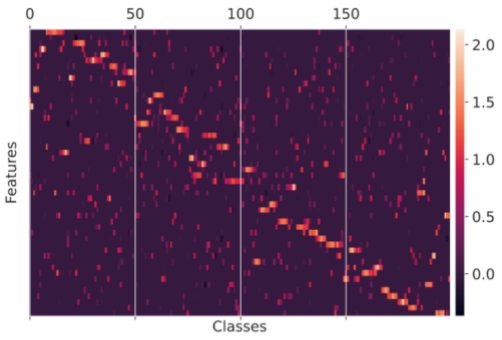
Table 4: Accuracy in percent dependent on backbone. Best results are in bold.

4.2 可解释性方面分析

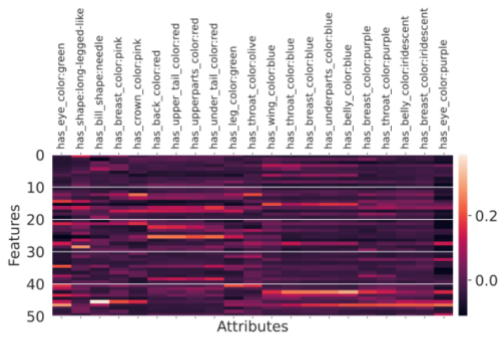
SLDD-Model 的可解释性主要来自于很少的特征数量，因为稀疏线性层很容易解释，因此具有足够好理解特征的完整模型既可以局部解释，也可以全局解释。

对于全局可解释性，SLDD-Model 的最后一层可以完全可视化和分析，这有利于相关从业者验证模型的全局行为。

论文在验证全局解释性时做了可视化分析，这对我来说也是一个启发。前段时间在听相关报告和复现 LIME 时都有相应的可视化分析，包括这篇文章的三种可视化方法各不相同，但是都很清楚的展示了结论的性质。如下面是文章的一个 case，其中名为 has_bill_shape:needle 的特征在所有类别超过 30%的具有该属性的样本上都有非零权重。



(a) Exemplary $\boldsymbol{W}^{\text{sparse}}$. The alignment of the features with attributes in CUB-2011 is displayed in Figure 3b.



(b) Relationship between chosen features and attributes ($C > 20\%$) of CUB-2011 for the exemplary model. Higher values indicate that the feature describes the attribute.

局部可解释性方面学界已经有诸如 LIME 等很完善的分析方法，这篇文章因为对全局解释性做了可视化分析，所以局部解释性也是水到渠成。不过论文并没有继续就给出的可解释性本身的性质好坏做衡量。全局解释性和局部解释性的 tradeoff 也依然是模型解释领域一个比较困难的问题。

4.3 可解释性与精度的权衡

正如我所预料到的那样，大量牺牲特征维度会导致精度出现一定的下降。论文尝试了不同的 n_f^* 数量对任务精度的影响，这些精度下降基本上都在 2%这个可控比例内。这个结果看起来还是比较诱人的。

5 论文优缺点讨论

虽然论文在某些任务上给出了非常诱人的成果，我们还是有必要从原理和实验结果入手谈谈论文方法的优缺点。

＋ 论文提出的方法补充了当前模型解释领域对深度神经网络解释方面的研究，为提高了神经网络模型的理解和可信度做了一定的工作。
＋ 论文的方法能够有效减少特征数量并保留特征的原始语义，这个特征的减少幅度和对应的精度保持相当具有竞争力。
＋ 论文的方法可以提升深度学习模型在需要高度可解释性和可信度的领域——如医学和自动驾驶等——的可信度。

× 论文的方法并不能保证在任何情况下都能获得最佳的稀疏度和维度，仍需进行预实验来验证方法的效果。
× 论文的方法需要进行特征学习、特征选择、稀疏化和特征微调等多个步骤，虽然论文并没有在文章中给出训练用时或者实验用时，但想必时间复杂度不会低。
× 论文并没有解释所用方法在所获特征与人类概念的对齐问题方面的突破，需要进一步改进和优化。

虽然论文的方法并不是万能的，但是该方法可以揭示机器学习模型中的问题，为构建公平和可信任的模型提供机会。在深度学习模型解释领域，稀疏度和维度或许可以成为评估模型可信度的指标之一。

6 论文阅读的收获

6.1 阅读收获

当前的模型解释领域已经将问题细分成模型有关/无关、全局/局部解释等，每个细分领域都有专业的方法。之前没有特别关注过模型解释的对象模型的生成来源的影响，因为文章提到了神经网络模型解释方面的问题，将模型解释继续按模型生成来源分类也许是一个好的方案。因为基于线性模型或者决策树的经典模型都具有比较好的可解释性，或许若我们能可视化模型内的特征流动，深度学习模型的可解释性也能获得好的提升。

论文的另一大贡献就是使我们看到将特征数量降到很低仍然有可能在精度上取得一个比较好的表现（当然也有可能只是在图像分类领域，因为特征维度实在是太高了，所以特征降低率的表现非常明显）。或许可以将论文方法选出的特征和图像超像素聚类得到的特征进行比对分析，并且进一步对双方选择特征的性能进行比对分析。

通过阅读这篇论文，我在模型可解释领域又有了一些新的视野和研究意向（比如说上面提到的一些改进方向），今后有时间会尝试落地。这次阅读忽视了一部分数学原理，还可以审视论文提出的一些复杂的公式相比于一些已有的简单方法来说是否是必要的。

6.2 课程关联

在深度学习导论课程中，我们学习了很多经典的模型，如 RNN、GNN 和 Transformer 等一些生成式模型。这些模型在实际应用中取得了很好的效果，但它们也存在一些问题，主要是在应用领域，其生成结果存在可解释性不足（黑箱模型）等局限性。该论文提出的方法可以帮助我们提高深度神经网络模型的可解释性，从而提高模型的可信度和可靠性，并且其特征减少的幅度让其看起来具有较好的可行性。因此，该论文可以作为深度学习导论课程的一个补充，加深我们对深度学习模型的理解和认识。

对于课程的建议，我认为可以增加一些关于深度学习模型的可解释性和可信度的内容。对于模型可解释性的分析与人的主观意识结合的非常紧密，探索起来非常有意思，并且还可以帮助降低同学们对模型调参「炼丹」的刻板印象。如果学时允许，可以增加一些使用比如 LIME 库来分析文本模型的表现。由于解释领域和人的主观想法比较相关，我之前做的一些调研也有很多借鉴了心理学、生物学上一些实验的方法。课堂上也可以就此问题组成额外的专题讨论，让大家各抒己见，可以帮助我们更好地理解深度学习模型的局限性和可信性。