基于张量计算的卷积神经网络语义表示学习

杨礼吉[1),2)] 王家祺[1),2)] 景丽萍[1),2)] 于剑[1),2)]

[1)](北京交通大学计算机与信息技术学院北京 100044)

[2)](北京交通大学交通数据分析与挖掘北京市重点实验室北京 100044)

摘 要 卷积神经网络已在多个领域取得了优异的性能表现，然而由于其不透明的内部状态，其可解释性依然面临很大的挑战。其中一个原因是卷积神经网络以像素级特征为输入，逐层地抽取高级别特征，然而这些高层特征依然十分抽象，人类不能直观理解。为了解决这一问题，我们需要表征出网络中隐藏的人类可理解的语义概念。本文通过预先定义语义概念数据集(例如红色，条纹，斑点，狗)，得到这些语义在网络某一层的特征图，将这些特征图作为数据，训练一个张量分类器。我们将与分界面正交的张量称为语义激活张量(Semantic Activation Tensors，SATs)，每个 SAT 都指向对应的语义概念。相对于向量分类器，张量分类器可以保留张量数据的原始结构。在卷积网络中，每个特征图中都包含了位置信息和通道信息，如果将其简单地展开成向量形式，这会破坏其结构信息，导致最终分类精度的降低。本文使用 SAT 与网络梯度的内积来量化语义对分类结果的重要程度，此方法称为 TSAT(Testing with SATs)。例如，条纹对斑马的预测结果有多大影响。本文以图像分类网络作为解释对象，数据集选取 ImageNet，在 ResNet50 和 Inceptionv3 两种网络架构上进行实验验证。最终实验结果表明，本文所采用的张量分类方法相较于传统的向量分类方法，在数据维度较大或数据不易区分的情况下，分类精度有显著的提高，且分类的稳定性也更加优秀。这从而保证了本文所推导出的语义激活张量更加准确，进一步确保了后续语义概念重要性量化的准确性。

关键词 深度学习；卷积神经网络；语义建模；张量表示；支持张量机；张量分类
中图法分类号 TP18

Semantic Representation Learning of Convolutional Neural Network Based on Tensor Computation

YANGLi-ji[1),2)] WANG Jia-qi[1),2)] JING Li-ping[1),2)] YU Jian[1),2)]

[1)](School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China)

[2)](Beijing Key Laboratory of Traffic Data Analysis and Mining, Beijing Jiaotong University, Beijing 100044, China)

Abstract Convolutional neural networks have achieved excellent performance in several areas, but their interpretability still faces significant challenges due to their opaque internal state. One reason for this is that convolutional neural networks take pixel-level features as input and extract high-level features layer by layer, but these high-level features are still very abstract and cannot be understood intuitively by humans. To solve this problem, we need to characterize the human-understandable semantic concepts hidden in the network. In this paper, we obtain feature maps of these semantic concepts at one layer of the network by pre-defining a dataset of semantic concepts (e.g. red, stripes, spots, dogs) and use these feature maps as data to train a tensor classifier. We refer to the tensor orthogonal to the partitioned interface as Semantic Activation Tensors (SATs), and each SAT points to a corresponding

本研究得到国家科技研发计划（2020AAA0106800）、国家自然科学基金项目（62176020）、北京市自然科学基金资助项目（Z180006）、中国人工智能学会-华为MindSpore学术奖励基金、中国科学院光电信息处理重点实验室开放课题基金（OEIP-O-202004)资助.**杨礼吉**，硕士研究生，主要研究领域为机器学习、深度学习、可解释学习.E-mail: 19120422.bjtu.edu.cn. **王家祺**，博士研究生，主要研究领域为机器学习、深度学习、可解释学习.E-mail: 19112028@bjtu.edu.cn.**景丽萍**（**通信作者**），博士，教授，博士生导师,主要研究领域为机器学习理论与方法、高维数据挖掘及在智能推荐、多媒体数据处理、社会计算等领域中的应用.E-mail: lpjing@bjtu.edu.cn.**于剑**，博士，教授，博士生导师，主要研究领域为机器学习及应用.E-mail: jianyu@bjtu.edu.cn

semantic concept. In contrast to vector classifiers, tensor classifiers can preserve the original structure of the tensor data. In a convolutional network, each feature map contains location and channel information, and if it is simply expanded into vector form, this would destroy its structural information and lead to a reduction in the final classification accuracy. In this paper, we use the inner product of SATs and network gradients to quantify the importance of semantics on classification results, a method known as TSAT (Testing with SATs). For example, how much streaks affect the prediction results of zebras. In this paper, the image classification network is used as the object of interpretation, and ImageNet is selected as the dataset for experimental validation on both ResNet50 and Inceptionv3 network architectures. The final experimental results show that the tensor classification method used in this paper has a significant improvement in classification accuracy and better classification stability compared to the traditional vector classification method in the case of larger data dimensions or data that are not easily distinguishable. This ensures that the semantic activation tensor derived in this paper is more accurate and further ensures the accuracy of the subsequent quantification of the importance of semantic concepts.

Key words deep learning; convolutional neural network; semantic modeling; tensor representation; support tensor machine; tensor-basedclassification

1 引言

卷积神经网络[1]在物体识别[2~4]、目标检测[5,6]、图像分割[7,8]等计算机视觉任务中取得了前所未有的成功。然而深度学习的“黑盒”效应仍饱受诟病[9][10][11]。虽然这些模型具有出色的性能，但它们缺乏可分解性，无法分解为直观的组件，因此难以解释。为了解决这一问题，网络模型的解释算法应运而生。

一种普遍的解释方法是根据模型的输入特征来描述其预测。例如在逻辑回归中，系数权重通常被解释为每个特征的重要性。然而，一个关键的问题是，在卷积神经网络中，其特征是像素，这并非是人类可以理解的高级语义概念。此外，卷积神经网络的内部值，例如神经元激活值（即特征图），也是不可解释的。

针对上述问题，我们希望可以通过某种方法学习网络内部语义概念的表征，这些表征往往是人类可理解的。学习过程可以分为有监督的学习方法和无监督的学习方法，而无监督方法本身缺乏可解释性，因此我们采用有监督方法。

基于有监督的学习方法，Testing with Concept Activation Vectors（TCAV）[12]这项工作通过使用预先定义的语义概念图片集来学习语义概念在网络中的表示。本文也借鉴此方法，例如，要定义“条纹”这一语义，可以使用一组条纹图像。值得注意的是，这些语义概念不限于训练数据，用户可以使用任何新的数据来定义它们。

但是 TCAV 的缺陷在于学到的是向量形式的语义表示，与网络中原本的张量结构不匹配，而本文所采用的方法可以保留卷积神经网络中张量数据的原始结构。我们使用张量分类方法，从而避免了将网络特征图的原始结构破坏。张量分类方法可以更好地体现数据的结构信息和内在相关性，可以得到更紧凑、更有意义的张量表示，尤其是在高阶张量的情况下，可以提高内积计算的效率，节省存储空间和计算时间，提高分类精度[13][14][15]。

通过在不同语义概念特征图之间训练一个张量分类器（支持张量机 STM），然后取与决策面正交的张量（T_C^l），每个T_C^l都指向其对应的语义概念，我们将这些张量称为语义激活张量（Semantic Activation Tensors，SATs）。

利用 SATs，我们可以在网络的不同卷积层考察模型预测对朝向语义概念方向的输入变化的敏感性，具体可推导为 SAT 与网络梯度的张量积。于是我们可以用网络的梯度与 SAT 之间的张量积来量化模型预测对高层语义概念的敏感性，该方法称之为 TSAT（Testing with SATs）。例如，给定一个可以识别斑马类别的分类模型和一组新的用户定义的“条纹”样本集，TSAT 可以将条纹语义对“斑马”预测的影响量化为一个数值。值得注意的是，我们并不是只解释某一张斑马样本，而是度量语义概念对数据集中所有斑马样本的敏感性，以达到全局解释的目的。具体流程图如图 1 所示。

本项工作主要有以下贡献：

1) 我们提出了一个新的事后可解释算法，该方法可以全局解释网络的决策依据;
2) 我们引入了张量分类方法，保留了卷积神经网络特征图的结构，相比较向量分类方法获得了更高的精度和稳定性，从而学到了更准确的语义表示；
3) 我们可以量化地解释语义概念对网络决策的贡献度，且语义概念是人类可理解的。

2 相关工作

本节将介绍现有的事后可解释方法，主要可以分为显著性解释方法和基于高级语义概念的定量解释方法。

显著性解释方法是图像分类中主要的局部解释方法之一。该方法通常会生成一个热力图，显示某张图片的每个像素对其分类结果的重要性。

CAM[16]是此类方法的开山之作，该方法通过线性地组合特征图来可视化 CNN。每个通道的特征图的权重由对应于目标类别的最后一层的全连接的权重确定。但是，CAM 仅限于 GAP-CNN，也就是说，全连接层之前必须是全局平均池化层。

Grad-CAM[17]是目前适用范围最广，且实现简单的方法。它的灵感来源于 CAM，通过使用梯度加权每个通道的特征图来可视化用于分类任务的任意 CNN。但是此方法尚未说明其机理，即为什么要使用梯度的平均值来加权每个特征图。在此基础上，后续又出现了很多改进版本，如 Grad-CAM++[18]，Ablation-CAM[19]，XGrad-CAM[20]，主要的改进点在于特征图加权值的计算方法。

DeepLIFT[21]提出将每个神经元的激活与其“参考激活”进行比较，并根据差异分配贡献分数。此方法不需要借助梯度信息，从而避免了在反向传播中梯度的消失问题。

尽管显著性解释方法会通过热力图标识相关区域并提供量化评价，但是仍有两个限制：1)由于每次只能解释一张图片（即局部解释），因此用户需要手

动评估每张图片才能得出全局解释。2)热力图关注的区域依然是像素级别的，用户无法获得语义层面的重要度解释。

基于高级语义概念[22][23]的解释方法旨在量化用户定义的高级语义概念对于模型预测的重要性。

LIME[24]用一个简单的线性模型，搭配可解释的语义概念进行适配，并且这个线性模型在局部的表现能够逼近原始 CNN 的效果。其中语义概念（超像素）是通过分割原始图片来获得的，并不能保证用户可以理解每一种特征。

CoCoX[25]可以识别出最小的语义级别特征，即可解释的概念。这些特征需要添加到图片中或从图片中删除才能改变模型对其预测的结果。

Been Kima[12]等人提出了 TCAV，用简单的线性分类器来学出概念的表示，并使用方向导数来量化概念对分类结果的重要程度。但是此方法将张量数据简单地展开成向量形式，并使用向量形式的分类器学到概念表示，丢失了原始数据之间的结构及位置信息。故在较复杂的数据上，分类精度表现一般。

本文的方法，在 TCAV 的基础上进行改进，保留了张量数据的原始结构，采用支持张量机（STM）作为分类方法。最终实验证明，本文所采用的方法在复杂数据上的表现优于 TCAV。

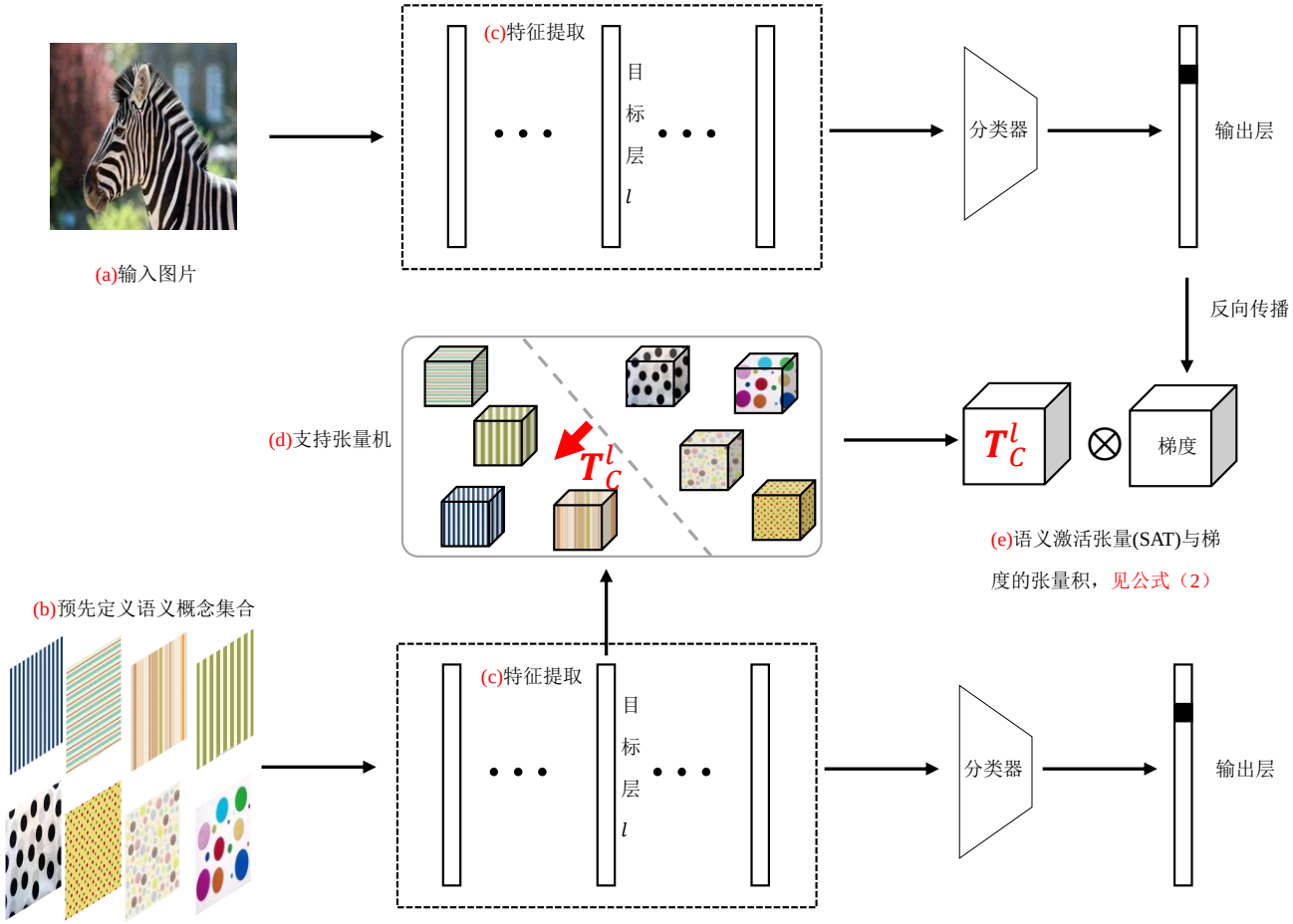

图 1　语义激活张量定量测试(TSAT)流程图。(a)需要解释的训练样本类别，例如斑马。(b)用户预先定义的语义概念集合。(c)已经训练好的卷积神经网络模型的卷积层。(d)将张量分类方法应用于语义概念集合的特征图，导出语义激活张量 SAT，$\boldsymbol{T}_C^l$，红色箭头。(e)TSAT 使用 SAT 与模型梯度的张量积来量化语义的重要性。

3 模型算法

本节主要介绍针对语义概念表示的解释方法，即如何定量解释每个语义概念对模型预测类的重要性，而无需任何模型再训练或修改。

不失一般性，我们考虑一个简单的卷积神经网络。输入图片$\boldsymbol{x} \in R^{C_i \times H_i \times W_i}$，第$l$层卷积层的激活空间为$f \in R^{C_l \times H_l \times W_l}$，最终的输出层为$\boldsymbol{y} \in R^n$。其中，$C, H, W, n$分别为通道数，高度，宽度，类别数。则输入空间到$l$层的激活空间映射可以表示为$f_l: R^{C_i \times H_i \times W_i} \longrightarrow R^{C_l \times H_l \times W_l}$。

3.1 语义概念集

我们方法的第一步是定义一个感兴趣的语义概念。我们可以通过选择一组代表该语义概念的示例或

找到带有该概念标记的独立数据集来简单地做到这一点。此策略的主要好处是，它不会将模型解释局限于预先存在的特征、标签或训练数据。另外，即使是非专家级的模型分析人员，也可以使用相关图片集来定义语义概念，因此本方法具有极大的灵活性。

3.2 语义驱动的激活张量（SATs）

给定一组人类可以理解的语义概念集合，我们在层l的激活空间中寻找一个代表这个语义的张量。为了找到这样一个张量，我们考虑由语义概念集合中的样本与随机负样本产生的l层中的激活。然后我们把“语义激活张量”（SATs）定义为张量分类框架中的权重，以分离模型的激活空间中的语义概念与随机负样本。

当用户对语义概念 S 感兴趣时（比如，条纹，斑点），他们可以收集一组语义概念图片作为正样本P_S，和一组随机负样本 N。例如，当以条纹图片为正样本的时候，可以选择斑点图片作为负样本，这样的好处在于可以同时导出这两个语义激活张量。然后我们可以训练一个二元张量分类器来区分两个样本集合在l层的特征图：$\{f_l(x): x \in P_S\}$和$\{f_l(x): x \in N\}$。最终得到的张量分类器的权重就是我们所需要的 SAT。

3.2.1 张量基础知识

我们首先介绍本算法涉及的张量基本运算。由于本文只涉及到三阶张量，所以以下只讨论在三阶张量下的张量运算法则[26]。

记三阶张量$\boldsymbol{X} \in R^{I_1 \times I_2 \times I_3}$，给出以下定义：

定义 1.张量内积，$\boldsymbol{X} \in R^{I_1 \times I_2 \times I_3}$和$\boldsymbol{Y} \in R^{I_1 \times I_2 \times I_3}$的内积$\langle \boldsymbol{X}, \boldsymbol{Y} \rangle = \sum_{i=1}^{I_1} \sum_{j=1}^{I_2} \sum_{k=1}^{I_3} x_{i,j,k} y_{i,j,k}$。

定义 2.向量外积，$\boldsymbol{X} = \boldsymbol{a}^{(1)} \circ \boldsymbol{a}^{(2)} \circ \boldsymbol{a}^{(3)}$，其中$x_{i_1 i_2 i_3} = a_{i_1}^{(1)} a_{i_2}^{(2)} a_{i_3}^{(3)}$。

定义 3.三阶张量的模式向量积，张量$\boldsymbol{X} \in R^{I_1 \times I_2 \times I_3}$与向量$\boldsymbol{v} \in R^{I_1}$的模式向量积$(\boldsymbol{X} \times_1 \boldsymbol{v})_{i_2 i_3} = \sum_{i_1=1}^{I_1} x_{i_1 i_2 i_3} v_{i_1}$。有以下性质：

1)三阶张量的每次模式向量积都使张量的阶数减一，最后变成零阶张量即标量。

2)三阶张量的模式向量积的顺序可以交换。

图 2 给出了三阶张量$\boldsymbol{X} \in R^{8 \times 7 \times 4}$与向量$\boldsymbol{v_1} \in R^{8 \times 1 \times 1}$，$\boldsymbol{v_2} \in R^{1 \times 7 \times 1}$，$\boldsymbol{v_3} \in R^{1 \times 1 \times 4}$的 3-模式向量积的运算原理图。

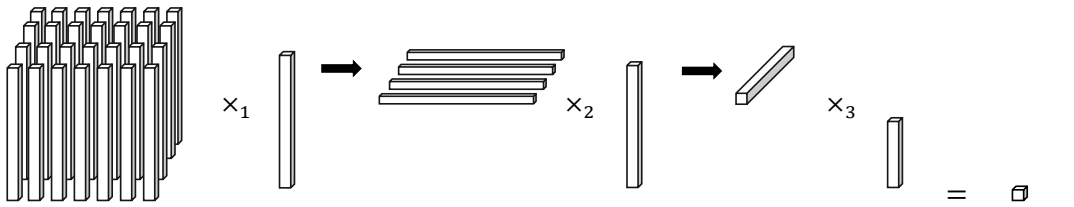

图 2 三阶张量的 3-模式向量积原理图

3.2.2 语义概念表示的学习方法

我们使用张量分类算法来学习语义概念的表示。张量分类算法[27][28][29]由向量分类算法推广而来，以下过程皆可类比向量分类算法。

给定张量数据$\boldsymbol{X}_i|_{i=1}^N \in R^{I_1 \times I_2 \times I_3}$以及标签$y_i = \{+1, -1\}$。我们想要找到一个张量平面分开正负样本。类比支持向量机，张量分类方法的优化方程如下：

$$\min_{\vec{\boldsymbol{w}}_k|_{k=1}^3, b, \vec{\boldsymbol{\xi}}} f\left(\vec{\boldsymbol{w}}_k|_{k=1}^3, \vec{\boldsymbol{\xi}}\right)$$

$$\text{s.t.} \quad y_i\left(\boldsymbol{X}_i \prod_{k=1}^{3} \times_k \vec{\boldsymbol{w}}_k + b\right) \geq 1 - \boldsymbol{\xi}_i, 1 \leq i \leq N \#(1)$$

其中，$f\left(\vec{\boldsymbol{w}}_k|_{k=1}^3, \vec{\boldsymbol{\xi}}\right) = \frac{1}{2}||\boldsymbol{W}||^2 + C\sum_{i=1}^{m} \boldsymbol{\xi}_i$，$\boldsymbol{W} = \boldsymbol{I} \prod_{k=1}^{3} \times_k \vec{\boldsymbol{w}}_k$，$\boldsymbol{I}$是三阶单位张量，$\vec{\boldsymbol{w}}_k$是张量数据$\boldsymbol{X}_i|_{i=1}^N \in R^{I_1 \times I_2 \times I_3}$在每个维度上的权重，$b$为偏置项，$\vec{\boldsymbol{\xi}}$是对应的松弛变量，$\boldsymbol{\xi}_i$用于衡量第$i$个训练样本被分错的成度，$C$为惩罚因子,表示对错误的惩罚程度。

张量分类与向量分类有两个不同点：

1) 向量分类的数据都是向量形式的，而张量分类的数据是张量形式。
2) 向量分类的决策方程定义为$y(\vec{\boldsymbol{x}}) = \text{sign}(\vec{\boldsymbol{w}}^T \vec{\boldsymbol{x}} + b)$，而张量分类的决策方程定义为$y(\boldsymbol{X}) = \text{sign}(\boldsymbol{X} \prod_{k=1}^{3} \times_k \vec{\boldsymbol{w}}_k + b)$。

算法 1 展示了张量分类算法的具体流程。其中 Step4 是传统的向量分类方法，此处采用 SVM。

3.3 语义概念的重要性度量方法（TSAT）

显著性映射（saliency maps）等可解释方法使用 logit 值相对于各个输入特征(例如像素)的梯度，并计算$\frac{\partial h_k(\boldsymbol{x})}{\partial \boldsymbol{x}_{a,b}}$,其中$h_k(x)$是第 k 类样本点的 logit 值，$\boldsymbol{x}_{a,b}$是

$\boldsymbol{x}$在位置(a,b)处的像素值。因此，显著性方法使用导数来衡量输出类别 k 对像素(a,b)大小变化的敏感度。

类比上述方法，我们也引入导数来计算重要性。通过计算模型对 SATs 方向的导数，我们在卷积层l测量模型预测对朝向语义概念方向的输入变化的敏感性。假设$T_C^l \in R^{C_l \times H_l \times W_l}$是$l$层语义概念 C 的一个 SAT 张量，$f_l(\boldsymbol{x})$是输入$\boldsymbol{x}$在$l$层的激活，则第 k 类对语义概念 C 的敏感性可以计算为方向导数$S_{C,k,l}(\boldsymbol{x})$：

$$S_{C,k,l}(\boldsymbol{x}) = lim_{\epsilon \to 0} \frac{h_{l,k}\left(f_l(\boldsymbol{x}) + \epsilon \boldsymbol{T}_C^l\right) - h_{l,k}\left(f_l(\boldsymbol{x})\right)}{\epsilon} = \langle \nabla h_{l,k}\left(f_l(\boldsymbol{x})\right), \boldsymbol{T}_C^l \rangle$$

对上式最直观的理解是，当模型的输入往语义概念的方向变化时，模型的预测到底产生了多大变化。从而进一步导出为 SAT 和模型梯度的张量积。

$S_{C,k,l}(\boldsymbol{x})$可以定量地测量模型预测相对于任何模型层语义概念的敏感性。它不是每个特征的度量指标（例如，与每个像素的显著性映射不同），而是在整个输入或一组输入上计算的每个概念的度量。我们使用上述方法对需要解释的类别的所有样本计算敏感性分数。

假设 k 是类标签，$\boldsymbol{X}_k$表示 k 类的所有输入样本。我们定义 TSAT 分数如下：

$$\mathrm{TSAT}_{C,k,l} = \frac{\left|\boldsymbol{x} \in \boldsymbol{X}_k : S_{C,k,l}(\boldsymbol{x}) > 0\right|}{\left|\boldsymbol{X}_k\right|} \#(3)$$

#(2) 其中|■|表示集合中元素的个数，$\mathrm{TSAT}_{C,k,l} \in [0,1]$。注意，$\mathrm{TSAT}_{C,k,l}$仅取决于$S_{C,k,l}(\boldsymbol{x})$的符号，当然也可以考虑其他度量方式。TAST 度量标准可以轻松地全局解释具体语义概念对某一类别的重要性贡献，而不是仅对单个样本进行局部解释。

算法 1 张量分类问题的交替投影法

输入：训练数据$X_i|_{i=1}^N \in R^{I_1 \times I_2 \times I_3}$, 类标$y_i = \{+1, -1\}$。

输出：支持张量积参数$\vec{w}_k|_{k=1}^3 \in R^{L_k}$和$b \in R$。

第一步：随机初始化$\vec{w}_k|_{k=1}^3$为单位向量。

第二步：重复步骤 3-5，直到满足收敛条件。

第三步：For j= 1 to M：

第四步：求解以下优化问题，得到$\vec{w}_j \in R^{L_j}$。

$$\min_{\vec{w}_j, b, \vec{\xi}} \quad f\left(\vec{w}_j, \vec{\xi}\right)$$

$$\text{s.t.} \quad y_i\left[\vec{w}_j^T (X_i \,\bar{\times}_j\, \vec{w}_j) + b\right] \geq 1 - \xi_i, \quad 1 \leq i \leq N$$

此处$\bar{\times}_j \vec{w}_j$表示对除j维以外的所有维度做模式向量积。

第五步：结束 For 循环。

第六步：检验是否收敛。

如果$\sum_{k=1}^3 [1 - \langle \vec{w}_{k,t}, \vec{w}_{k,t-1} \rangle / (\|\vec{w}_{k,t}\|_2 \cdot \|\vec{w}_{k,t-1}\|_2)] \leq \varepsilon$，收敛。

此处$\vec{w}_{k,t}$是当前的投影向量，$\vec{w}_{k,t-1}$是前一次的投影向量。

第七步：结束。

4 实验与结果

为了证明本文所提出算法的有效性，本文采取与 TCAV 一文中类似的实验方法。我们将本算法分别应用在在不同的网络架构中，包括 ResNet50[4] 和 Inceptionv3[30]。其中，ResNet50 包含 49 层卷积层和 1 层全连接层，而在 PyTorch 的官方模型中，总共分为 4 大层，layer1，layer2，layer3，layer4，每个大层中包含多个卷积层。同样的，Inceptionv3 的 Pytorch 官方模型中包含 1a，2a，2b，3b，4a，5b，5c，5d，6a，6b，6c，6d，6e，7a，7b，7c 这些大层，而 1a，2a，2b，3b，4a 这些层结构简单，并没有复合许多卷积层。因此，为了实验的简明，我们在 ResNet50 的 4 个大层和 Inceptionv3 的 5b~7c 大层上进行实验。本文利用 PyTorch 框架实现文中所涉及的实验，两种网络架构都采用官方提供的预训练模型。

数据来源使用公开数据集 ImageNet，选取斑马，消防车，狗拉雪橇这三种类别进行实验，测试图片均来源于 ImageNet 中这三种类别的测试集。

4.1 TSAT与TCAV在导出语义概念时的分类性能比较

本小节主要介绍 TSAT 与 TCAV 所采用分类方法的性能比较。

我们选取了 3 种类型的语义概念，分别为颜色

（红色 vs 蓝色），纹理（条纹 vs 斑点），物体（斑马 vs 狗），并预先定义好这些语义的数据集。每种语义有 120 张图片，训练集与测试集的比例为 8:2。为了避免分类的偶然性，我们分别将张量分类方法与向量分类方法各测试 50 次。

表 1 与表 2 展示了每种层面的语义在 Inceptionv3 和 ResNet50 各个卷积层的分类精度。可以看到，在网络的低层，低级别语义概念（如颜色）已经可以达到较高的分类精度，且由于颜色分类难度不高，所以无论是 TCAV 还是 TSAT 都能达到很高的精度。而高级特征在网络低层的分类精度表现较差。这证实了较低层充当较低级别的特征检测器,而较高层使用较低级别的特征来进一步检测高级语义特征。从表中我们也可以看出，在某些 TCAV 表现不好的情况下，TSAT 所采用的分类方法依然可以达到一个比较高的分类精度。

由图 3 可知纹理级别的语义在网络的中层精度达到最高，而物体级别的语义则是在最顶层时精度达到最高。这也证明了网络的较低层抽取颜色，纹理等低级特征，而网络高层抽取物体等高级语义特征。

图 4 展示了两种分类方法在 Inceptionv3 和 ResNet50 各个卷积层的分类精度的标准差。可以看到，TCAV 中的分类精度标准差普遍低于 TSAT，这说明了张量分类方法在各个卷积层的分类稳定性都要由于 TCAV 的方法。

同时，通过表 1 和表 2 我们也可以看到，在较低层时，由于某些高级概念并没有被很好的抽取，且低层特征图的维度较大，如果使用 TCAV 的方法，分类精度表现差强人意。而使用 TSAT 的分类方法后，由于考虑到了张量数据的结构、空间信息，分类精度有了较大的提升。由此可见，通过张量学习方法导出的语义概念更具有代表性，更加有利于下一步的解释。

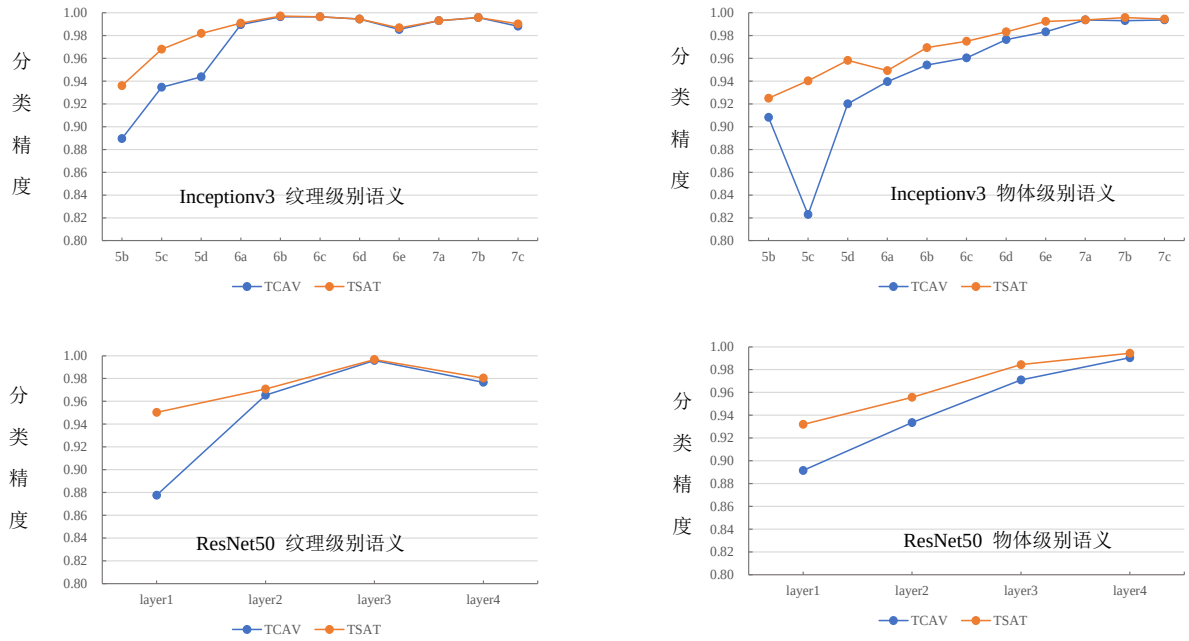

图 3 TSAT 与 TCAV 在不同模型上的分类精度对比图

分类精度标准差

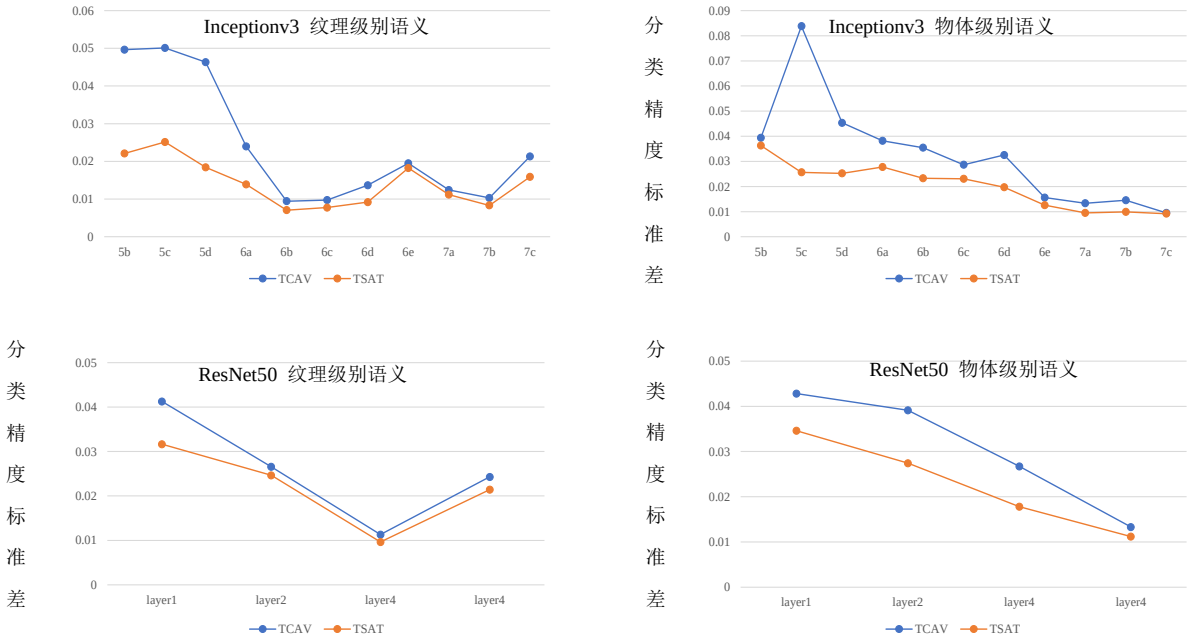

图 4TSAT 与 TCAV 在不同模型上的 50 次分类精度标准差对比图

表 1TSAT 分类方法与 TCAV 分类方法在 Inceptionv3 上的实验结果

	颜色级别的语义		纹理级别的语义		物体级别的语义	
	TCAV	TSAT	TCAV	TSAT	TCAV	TSAT
Mixed_5b	1.0000	1.0000	0.8896	0.9361	0.9083	0.9250
Mixed_5c	1.0000	1.0000	0.9347	0.9680	0.8229	0.9403
Mixed_5d	1.0000	1.0000	0.9438	0.9819	0.9201	0.9583
Mixed_6a	1.0000	1.0000	0.9896	0.9910	0.9396	0.9493
Mixed_6b	1.0000	1.0000	0.9965	0.9972	0.9542	0.9695
Mixed_6c	1.0000	1.0000	0.9965	0.9965	0.9604	0.9750
Mixed_6d	1.0000	1.0000	0.9945	0.9945	0.9765	0.9833
Mixed_6e	1.0000	1.0000	0.9845	0.9868	0.9833	0.9924
Mixed_7a	1.0000	1.0000	0.9931	0.9931	0.9938	0.9938
Mixed_7b	1.0000	1.0000	0.9958	0.9958	0.9931	0.9958
Mixed_7c	1.0000	1.0000	0.9882	0.9903	0.9938	0.9945

表 2TSAT 分类方法与 TCAV 分类方法在 Resnet50 上的实验结果

	颜色级别的语义		纹理级别的语义		物体级别的语义	
	TCAV	TSAT	TCAV	TSAT	TCAV	TSAT
Layer1	1.0000	1.0000	0.8775	0.9503	0.8915	0.9319
Layer2	1.0000	1.0000	0.9654	0.9708	0.9335	0.9557
Layer3	1.0000	1.0000	0.9958	0.9967	0.9709	0.9844
Layer4	1.0000	1.0000	0.9763	0.9804	0.9905	0.9944

4.2 利用TSAT方法量化语义重要性

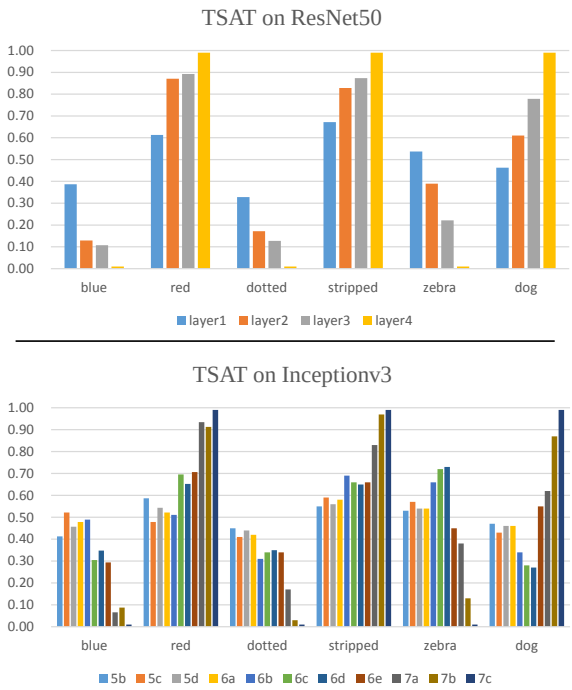

图 5　TSAT 在 ResNet50 和 Inceptionv3 上的实验结果

我们利用 TSAT 方法量化每种语义对相关类别的重要性。在颜色层面，量化红色和蓝色语义对消防车的重要性；在纹理层面，量化斑点和条纹语义对斑马的重要性；在物体层面，量化斑马和狗语义对狗拉雪橇的重要性。

图 5 展示了 TSAT 在 ResNet50 和 Inceptionv3 各个层上的测试结果。TSAT 的定量解释结果证实了我们的常识认知。例如红色语义对消防车的重要性，条纹语义对斑马的重要性以及狗语义对狗拉雪橇的重要性都比它们相对语义的重要性要大。

值得注意的是，该重要度评分的绝对大小并没有实际意义，用户只需关注哪个语义的重要度大，就说明了该语义相对于其负样本语义更具重要性。

从图 5 中我们可以看到，颜色和纹理级别的语义虽然在较低层就已经被很好地区分开来，但是它们对分类的影响仍然需要在较高层才能发挥出来。而物体级别的语义在低层时，TAST 的结果并不符合我们的预期。比如在测试斑马和狗对狗拉雪橇的影响时，可以看到在较低层时斑马反而比狗的影响大。我们推测这是因为在底层时，网络探测的依然是颜色和纹理级别的语义，而斑马的颜色与雪较接近，从而导致此结果。在网络的高层，物体级别语义的 TAST 结果与人类的判断标准也是相符合的。

4.3 TSAT与显著性映射方法对比

我们将 TAST 与传统的显著性映射方法进行比较，具体结果如图 6 所示，其中显著性映射采用的是经典的 Grad-CAM。

Grad-CAM 能够比较精准的突显出斑马所在部位，但是用户依然无法知道模型所依赖的语义概念，以及相关语义概念的重要性。当模型的输入是一些非正常的斑马图片时（例如具有彩色条纹），Grad-CAM 给出的解释则没有什么意义，抗干扰能力欠缺。反观 TSAT，在这两种情况下，都能够根据用户预先定义的语义概念进行解释，且解释结果依然符合人类直觉，条纹概念比斑点概念更加重要。

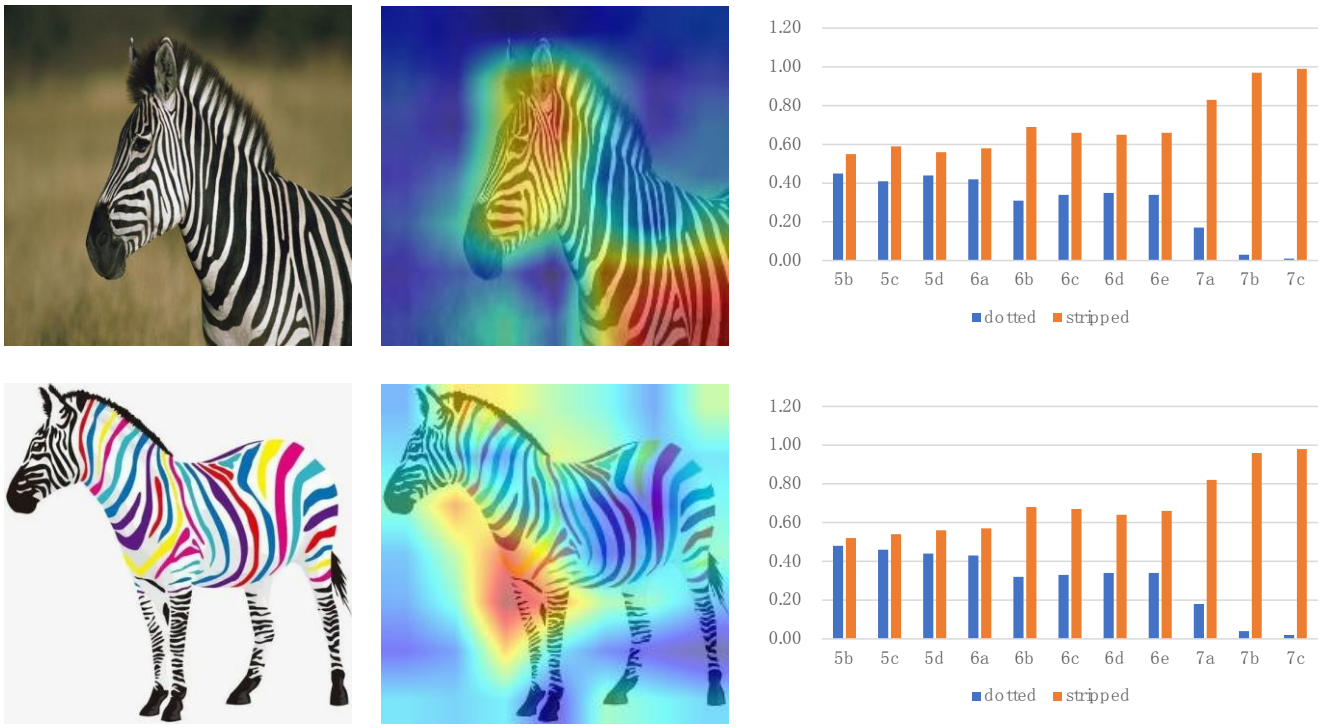

图 6TSAT 与显著性映射的解释结果

4.4 小结

通过以上两个实验结果我们可以看出，在大部分情况下，张量分类方法的精度要比向量分类方法高 1~5 个百分点，且 50 次的分类精度标准差也要高 1~3 个百分点。这充分说明了张量分类方法相较于向量分类方法，精度更高，且稳定性也更好，可以应对复杂的张量形式数据。

TSAT 的解释效果也符合人类的认知常识，且相较于显著性等可视化方法，基于语义概念的解释更具直观性，用户可以很方便地知道网络决策的依据。

另外，从实验结果我们也可以看出网络每个层所探测的特征类型。比如，在 Inceptionv3 中，5b、5c、5d 探测的是颜色层面的语义特征，6a-6e 探测的是纹理层面的语义特征，而 7a、7b、7c 探测的则是物体级别的特征。这对模型分析人员及模型架构师也有一定的帮助。

5 总结

本文提出了一种新的事后可解释方法（TSAT）。该方法以人类可理解的高级语义概念来解释神经网络的内部状态，并量化各个语义概念对分类结果的影响程度。用户通过预先定义语义概念数据集，并使用张量分类算法导出语义激活张量（SATs），继而可以用 SATs 解释任何网络模型。

在导出 SATs 时，我们对比了向量分类器与张量分类器的精度表现。事实证明，当张量数据比较复杂且不易区分时，张量分类器的分类精度相比较普通的向量分类器有较大提升。

TSAT 的定量解释结果证实了人类的普遍识认知，比如条纹对斑马的影响程度比较大。这说明了此方法的有效性。我们分别对不同类型的语义概念做测试，发现低级语义概念在中低层上就会产生一定影响，而高级语义概念在高层上才会产生较大影响，这也符合卷积神经网络的特性，同时也进一步证明了本方法的合理性。

本方法的局限性在于需要人为定义语义概念，未来我们将对此方法进行优化，以期待自动地提取模型中的语义概念。

致 谢 在此，我们向对本文的工作给予支持和建议的老师同学们表示衷心的感谢！

参 考 文 献

[1] LeCun Y. Generalization and network design strategies. Connectionism Perspective, 1989, 19:143-155

[2] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012, 25: 1097-1105.

[3] SzegedyC L W. Going deeper with convolutions//Proceedings of the IEEE Conferenceon Computer Vision and Pattern Recognition. Boston, USA, 2015, 1: 9.

[4] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Las Vegas,USA,2016:770-778.

[5] Geng Z, Sun K, Xiao B, et al. Bottom-Up Human Pose Estimation Via Disentangled Keypoint Regression//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online, 2021: 14676-14686.

[6] Wang X, Shu X, Zhang Z, et al. Towards More Flexible and Accurate Object Tracking with Natural Language: Algorithms and Benchmark//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online, 2021: 13763-13773.

[7] Zhu Z, Xu M, Bai S, et al. Asymmetric non-local neural networks for semantic segmentation//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Korea, 2019: 593-602.

[8] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.

[9] Dong Y, Su H, Wu B, et al. Efficient decision-based black-box adversarial attacks on face recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7714-7722.

[10] Li J, Schmidt F, Kolter Z. Adversarial camera stickers: A physical camera-based attack on deep learning systems//International Conference on Machine Learning. Long Beach, USA,2019: 3896-3904.

[11] Su J, Vargas D V, Sakurai K. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841.

[12] Kim B, Wattenberg M, Gilmer J, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)//International conference on machine learning.Vienna, Austria. PMLR, 2018: 2668-2677.

[13] Makantasis K, Doulamis A D, Doulamis N D, et al. Tensor-based classification models for hyperspectral data analysis. IEEE Transactions on Geoscience and Remote Sensing, 2018, 56(12): 6884-6898.

[14] Liu F, Wang Q. A sparse tensor-based classification method of hyperspectral image. Signal Processing, 2020, 168: 107361.

[15] Klus S, Gelß P. Tensor-based algorithms for image classification. Algorithms, 2019, 12(11): 240.

[16] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas,USA,2016: 2921-2929.

[17] Selvaraju R R, Cogswell M, Das A, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization//Proceedings of the IEEE international conference on computer vision.Venice, Italy, 2017: 618-626.

[18] Chattopadhay A, Sarkar A, Howlader P, et al. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks//2018 IEEE winter conference on applications of computer vision (WACV). Lake Tahoe, USA,2018: 839-847..

[19] Ramaswamy H G. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.Snowmass Village, USA, 2020: 983-991.

[20] Fu R, Hu Q, Dong X, et al. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. arXiv preprint arXiv:2008.02312, 2020.

[21] Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences//International Conference on Machine Learning. Sydney, Australia,PMLR, 2017: 3145-3153.

[22] Ghorbani A, Wexler J, Zou J, et al. Towards automatic concept-based explanations. arXiv preprint arXiv:1902.03129, 2019.

[23] Zhang R, Madumal P, Miller T, et al. Invertible Concept-based Explanations for CNN Models with Non-negative Concept Activation Vectors//Proceedings of the AAAI Conference on Artificial Intelligence.Vancouver, Canada, 2021, 35(13): 11682-11690.

[24] Ribeiro M T, Singh S, Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier//Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco, USA, 2016: 1135-1144.

[25] Akula A, Wang S, Zhu S C. Cocox: Generating conceptual and counterfactual explanations via fault-lines//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(03): 2594-2601.

[26] Kolda T G, Bader B W. Tensor Decompositions and Applications. Siam Review, 2009, 51(3):455-500

[27] Tao D, Li X, Wu X, et al. Supervised tensor learning. Knowledge and Information Systems, 2005, 13(1):450-457

[28] Tan X, Zhang Y, Tang S, et al. Logistic tensor regression for classification//International Conference on Intelligent Science and Intelligent Data Engineering. Springer, Berlin, Heidelberg, 2012: 573-581.

[29] Zhou H, Li L, Zhu H. Tensor regression with applications in neuroimaging data analysis. Journal of the American Statistical Association, 2013, 108(502): 540-552.

[30] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision//Proceedings of the IEEE conference on computer vision and pattern recognition. Las Vegas, USA, 2016: 2818-2826.

YANG Li-ji, M.S. candidate. His research interests include machine learning, deep learningandinterpretable learning.

WANG Jia-qi, Ph.D. His research interests include machine learning, deep learningandinterpretable learning.

JING Li-ping, Ph.D. , professor, Ph. D. superviser. Her research interests include machine learning theory and methods, high-dimensional data mining and applications in intelligent recommendation, multimedia data processing, social computing and other fields.

YU Jian, Ph. D. , professor, Ph. D. superviser. His research interest ismachine learning and application.

Background

Convolutional neural networkshave achieved unprecedented success in computer vision tasks such as object recognition, target detection, and image segmentation. However, the “black box” effect of deep learning is still criticized. Although these models have excellent performance, they lack decomposability and cannot be decomposed into intuitive components and are therefore difficult to interpret. To address this problem, interpretation algorithms for network models have emerged. A common interpretation approach is to characterize the predictions of a model based on its input characteristics. In logistic regression, for example, the coefficient weights are usually interpreted as the importance of each feature. However, a key problem is that in convolutional neural networks, their features are pixels, which are not high-level semantic concepts that humans can understand. Moreover, the internal values of convolutional neural networks, such as neuronal activation values (i.e., feature maps), are not interpretable. To address the above problems, we hope that some method can be used to learn representations of semantic concepts inside the network that are often human-understandable.

In this paper, we obtain feature maps of these semantics at one layer of the network by predefining a dataset of semantic concepts (such as red, stripes, spots, dogs), and use these feature maps as data to train a tensor classifier. We refer to the tensor orthogonal to the partitioned interface as Semantic Activation Tensors (SATs), and each SAT points to the corresponding semantic concept. In this paper, we use the inner product of SATs and network gradients to quantify the importance of semantics on classification results, and this method is called TSAT (Testing with SATs). The quantitative interpretation of TSAT results confirms the general human perception, such as the greater degree of influence of stripes on zebras. This shows the effectiveness of this method. In this paper, we do tests on different types of semantic concepts separately, and find that low-level semantic concepts will have some influence on the low and middle levels, while high-level semantic concepts will have a greater influence only on the high level, which is also in line with the characteristics of convolutional neural networks, and further proves the rationality of this method.

This work was partly supported by the National Key Research and Development Program (2020AAA0106800); the National Natural Science Foundation of China under Grant 62176020; the Beijing Natural Science Foundation under Grant Z180006; CAAI-Huawei MindSpore Open Fund; the Open Project Program Foundation of the Key Laboratory of Opto-Electronics Information Processing, Chinese Academy of Sciences(OEIP-O-202004)