arXiv:2301.02437v1 [stat.ML] 6 Jan 2023

深度学习驱动的显著区域的有效 P值

Daiki Miwa* Vo Nguyen LeDuy*名古屋工业大学 理化学研究所 miwa.daiki.mllab.nit@gmail.com
duy.mllab.nit@gmail.com

Ichiro Takeuchi†

名古屋大学和理化学研究所

ichiro.takeuchi@mae.nagoya-u.ac.jp

2023年1月9日

摘要

人们提出了各种显著图方法，用于阐释和解释深度学习模型的预测结果。显著性图使我们能够了解输入信号的哪些部分对预测结果有较大影响。然而，由于显著性图是通过深度学习模型中的复杂计算得到的，因此往往难以判断显著性图本身的可靠性。本研究提出一种方法，以 p-值的形式量化显著区域的可靠性。我们的思路是将显著区域视为训练好的深度学习模型所选定的假设，并采用选择性推断框架。所提方法能够以可证明的方式控制显著区域假阳性检测的概率。我们通过合成数据集和真实数据集上的数值示例，验证了所提方法的有效性。此外，我们开发了一个基于 Keras 的框架，无需额外的实现成本，即可对各类卷积神经网络进行所提出的选择性推断。

*共同贡献
†通讯作者

1 引言

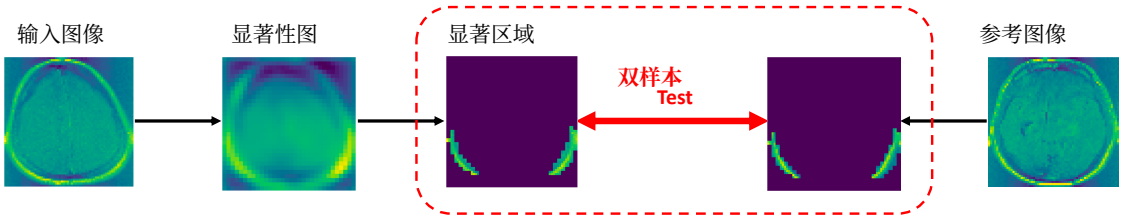
深度神经网络（DNN）凭借其能够从所提供的数据中自动发现预测任务所需表征的能力，在各个领域的众多实际应用中展现出卓越的预测性能。为了确保深度神经网络的决策过程透明且易于理解，有效地解释和阐释深度神经网络的表征至关重要。例如，在图像分类任务中，获取显著区域 使我们能够解释输入图像的哪些部分对分类结果有较大影响。

人们提出了多种显著图方法，用于解释和阐释深度神经网络模型的预测 [Ribeiro 等人，2016，Bach 等人，2015，Doshi-Velez 和 Kim，2017，Lundberg 和 Lee，2017，Zhou 等人，2016，Selvaraju 等人， 2017]。然而，显著性方法得到的结果并不稳健 [Kindermans 等人，2017，Ghorbani 等人，2019，Melis 和 Jaakkola，2018，Zhang 等人，2020，Dombrowski 等人，2019，Heo 等人， 2019]。因此，开发一种量化深度神经网络驱动的显著区域可靠性的方法十分重要。

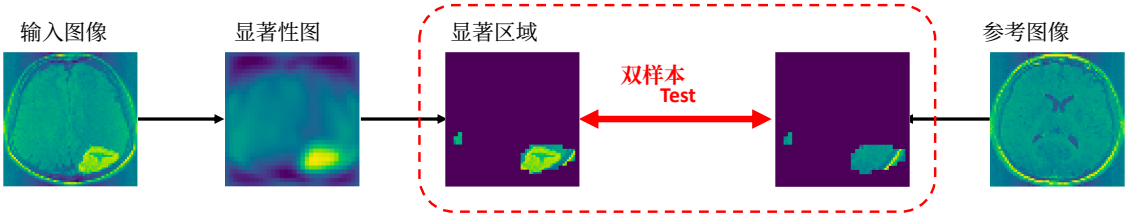
我们的想法是将显著区域视为由训练好的深度神经网络模型驱动的假设，并采用统计假设检验框架。我们使用 p-值作为衡量深度神经网络驱动的假设在统计上是否可靠的标准。遗憾的是，由于存在选择偏差，为深度神经网络驱动的显著区域构建有效的统计检验颇具挑战。换言之，由于训练好的深度神经网络根据所提供的数据选择 显著区域，选择后的重要性评估会产生向上的偏差。

为校正深度神经网络驱动的显著区域所产生的选择偏差并计算有效的 p 值，我们提出了一种条件选择性推断 （SI）方法。条件 SI 通过在假设（显著区域）由训练好的深度神经网络选出的事件条件下进行检验统计量推断，从而校正选择偏差。我们的主要技术贡献是开发了一种计算方法，可明确推导一大类卷积神经网络（CNN）的显著区域的精确（非渐近）条件抽样分布，从而能够进行条件 SI 并计算有效的 p 值。图1展示了问题设定的一个示例。

相关工作。 本研究聚焦于用于事后分析的统计假设检验，即当测试输入实例输入模型时，对训练好的深度神经网络模型所识别的显著区域进行统计显著性量化。人们已开发出多种方法来可视化和理解训练好的深度神经网络。许多此类事后方法 [Mahendran and Vedaldi, 2015, Zeiler and Fergus, 2014, Dosovitskiy and Brox, 2016, Simonyan et al., 2013] 致力于开发可视化工具，以呈现给定训练好的深度神经网络后的显著性图。其他方法则旨在给定训练好的网络，识别输入图像中的判别区域 [Selvaraju et al., 2017, Fong



(a) 不含肿瘤区域的图像。朴素方法-$p = \mathbf{0.00}$（错误检测）和选择性-$p = \mathbf{0.43}$（真阴性）



(b) 图像　与肿瘤区域。朴素方法-$p = \mathbf{0.00}$（真阳性）和选择性推断-$p = \mathbf{0.00}$（真阳性　）

图1：脑肿瘤数据集上的问题设定和所提方法示例。将名为 CAM 的显著性方法应用于查询输入图像 [Zhou et al., 2016] 后，我们得到显著区域。我们的目标是通过比较显著区域与参考图像中对应区域，进行双样本检验，并以p-值的形式给出该显著区域的统计显著性。请注意，由于显著区域是基于数据选出的，所选区域的显著程度会向上偏倚。在上方图像中，由于不存在真脑肿瘤，不考虑选择偏差而得出的朴素方法 p-值接近于零，表明该显著区域是假阳性发现。另一方面，采用所提出的条件选择性推断方法得出的选择性 p-值为 0.43，表明所选显著性区域在统计上并不显著。在下方图像中，由于存在真脑肿瘤，朴素方法 p-值和选择性 p-值都非常小，表明是真阳性发现。这些结果说明，朴素方法 p-值不能用于量化基于深度神经网络的显著区域的可靠性。相反，借助选择性 p-值，我们能够以预期的错误率成功识别假阳性和真阳性检测。

以及 Vedaldi，2017，Zhou 等人，2016，Lundberg 和 Lee， 2017]。此外，近期一些研究表明，许多流行的解释与诠释方法在面对输入数据和模型的扰动或对抗攻击时并不稳定 [Kindermans 等人，2017，Ghorbani 等人，2019，Melis 和 Jaakkola，2018，Zhang 等人，2020，Dombrowski 等人，2019，Heo 等人，2019]。然而，据我们所知，迄今尚无研究在严格的统计推断框架下，对深度神经网络驱动的显著区域进行定量评估并检验其可复现性。

近年来，条件选择性推断已被视为一种有前景的新方法，可用于评估

数据驱动假设的统计显著性。条件选择性推断主要用于对通过 Lasso 等特征选择方法选出的线性模型特征进行推断 [Lee 等人，2016，Liu 等人，2018，Hyun 等人，2018，Le Duy 和 Takeuchi，2021]以及逐步特征选择 [Tibshirani 等人，2016，Sugiyama 等人，2021a]。条件选择性推断研究的核心思想是基于选择事件进行推断，从而得出检验统计量的精确抽样分布。此外，条件选择性推断还应用于各种问题 [Fithian 等人，2015，Tian 和 Taylor，2018，Yang 等人，2016，Hyun 等人，2021，Duy 等人，2020，Sugiyama 等人，2021b，Chen 和 Bien，2019，Panigrahi 等人，2016，Tsukurimichi 等人，2021，Hyun 等人，2018，Tanizaki 等人，2020，Duy 和 Takeuchi，2021，Tibshirani 等人，2016，Sugiyama 等人，2021a，Suzumura 等人，2017，Das 等人，2021，Duy 和 Takeuchi，2022]。

与本研究最相关的现有工作是 Duy 等人 [2022], 的研究，他们提出了一个框架，用于计算基于深度神经网络的图像分割结果的有效 p值。本文对这项工作进行了推广，使得可以考虑由网络任意内部节点表征的假设，从而能够量化显著区域的统计显著性。这与 Duy 等人 [2022]的工作不同，后者仅考虑了分割任务中对深度神经网络输出的推断。此外，我们还引入了一个基于 Keras 的实现框架，使我们能够在无需额外实现成本的情况下，对广泛的卷积神经网络开展选择性推断。这与 Duy 等人 [2022]的工作不同，后者在网络架构发生变化时必须重新实现选择事件。另一方面，Burns 等人 [2020] 将黑箱模型可解释性视为一个多重假设检验问题。他们旨在通过检验模型预测与将特征替换为其反事实后预期结果之间差异的显著性，推断重要特征。这种多重假设检验方法的难点在于需要考虑的假设数量很大（例如，对于一幅包含 n 个像素的图像，可能的显著区域数量为 2^n）。当假设数量很大时，Bonferroni 校正等多重检验校正方法会非常保守。为避开这一难题，他们仅考虑由人类专家或目标检测器选出的、数量可处理的区域；由于这些候选区域是根据数据选出的，因此会导致选择偏差。

贡献。 我们的主要贡献如下：

- 我们基于选择性推断（SI）概念，为显著区域提供了一种精确（非渐近）推断方法。据我们所知，这是首个提出提供有效 p 值的方法，可用于统计量化深度神经网络驱动的显著区域的可靠性。

- 我们提出了一种新颖的算法及其实现。具体而言，我们提出了一种基于 Keras 的实现，使我们能够对广泛类型的卷积神经网络进行条件选择性推断，而无需额外的实现成本。

• 我们在合成数据集和真实世界数据集上进行了实验，结果表明，所提方法能够成功控制假阳性率，在计算效率方面表现良好，并且在实际应用中取得了良好效果。我们在补充文档中详细介绍了实现。我们的代码可在以下位置获取：

https://github.com/takeuchi-lab/选择性推断 深度神经网络 显著区域。

2 问题表述

本文研究这样一个问题：将测试输入实例输入训练好的深度神经网络模型后，如何量化该模型识别出的显著区域的统计显著性。考虑一个n维查询输入向量

$$\boldsymbol{X} = (X_1, ..., X_n)^\top = \boldsymbol{s} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathbb{N}(\mathbf{0}, \sigma^2 I_n)$$

以及一个 n维参考输入向量，

$$\boldsymbol{X}^{\mathrm{ref}} = (X_1^{\mathrm{ref}}, ..., X_n^{\mathrm{ref}})^\top = \boldsymbol{s}^{\mathrm{ref}} + \boldsymbol{\varepsilon}^{\mathrm{ref}}, \quad \boldsymbol{\varepsilon}^{\mathrm{ref}} \sim \mathbb{N}(\mathbf{0}, \sigma^2 I_n),$$

其中， $\boldsymbol{s}, \boldsymbol{s}^{\mathrm{ref}} \in \mathbb{R}^n$ 分别是查询输入向量和参考输入向量的信号， $\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^{\mathrm{ref}} \in \mathbb{R}^n$ 分别是它们的噪声。我们假设信号 $\boldsymbol{s}$ 和 $\boldsymbol{s}^{\mathrm{ref}}$ 未知，而噪声$\boldsymbol{\varepsilon}$ 和 $\boldsymbol{\varepsilon}^{\mathrm{ref}}$ 的分布已知（或可通过外部独立数据估计），且服从 $\mathbb{N}$（$\mathbf{0}, \sigma^2 I_n$），即均值向量为$\mathbf{0}$ 、协方差矩阵为 $\sigma^2 I_n$的 n维正态分布；这些噪声彼此独立。在§1所示的示例中， $\boldsymbol{X}$ 是潜在患者的查询脑部图像（我们不知道此人是否患有脑肿瘤），而 $\boldsymbol{X}^{\mathrm{ref}}$ 是已知没有脑肿瘤的健康人的脑部图像。

考虑一种用于训练好的卷积神经网络的显著性方法。我们将该显著性方法表示为函数$\mathcal{A}$: $\mathbb{R}^n \to \mathbb{R}^n$ ，它接收查询输入向量 $\boldsymbol{X} \in \mathbb{R}^n$ 并返回显著性图 $\mathcal{A}(\boldsymbol{X}) \in \mathbb{R}^n$。对于查询输入向量 $\boldsymbol{X}$ ，我们将显著区域 $\mathcal{M}_{\boldsymbol{X}}$ 定义为显著性图值大于某个阈值的元素集合

$$\mathcal{M}_{\boldsymbol{X}} = \{i \in [n] : \mathcal{A}_i(\boldsymbol{X}) \geq \tau\}, \tag{1}$$

其中， $\tau \in \mathbb{R}$ 表示给定的阈值。本研究以 CAM [Zhou 等人， 2016] 为例，介绍显著性方法以及基于阈值的显著区域定义。我们的方法也可应用于其他显著性方法和其他显著区域定义。

统计推断。 为量化显著性区域 $\mathcal{M}_{\boldsymbol{X}}$ 的统计显著性，我们考虑采用此类双样本检验 来量化查询输入向量中显著

区域与查询输入向量 $\boldsymbol{X}_{\mathcal{M}_{\boldsymbol{X}}}$ 中的显著区域以及参考输入向量 $\boldsymbol{X}^{\mathrm{ref}}_{\mathcal{M}_{\boldsymbol{X}}}$ 中对应区域之间差异的统计显著性。作为双样本检验的具体示例，我们考虑均值零假设检验：

$$\mathrm{H}_0 : \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{X}}} s_i = \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{X}}} s_i^{\mathrm{ref}} \quad \text{v.s.} \quad \mathrm{H}_1 : \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{X}}} s_i \neq \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{X}}} s_i^{\mathrm{ref}}. \tag{2}$$

和全局零假设检验：

$$\mathrm{H}_0 : s_i = s_i^{\mathrm{ref}},\ \forall i \in \mathcal{M}_{\boldsymbol{X}}, \quad \text{v.s.} \quad \mathrm{H}_1 : s_i \neq s_i^{\mathrm{ref}},\ \exists i \in \mathcal{M}_{\boldsymbol{X}}, \tag{3}$$

在式 (2) 所示的均值零假设检验中，我们考虑一个零假设：显著区域 $\mathcal{M}_{\boldsymbol{X}}$ 中的平均信号在 $\boldsymbol{X}$ 和 $\boldsymbol{X}^{\mathrm{ref}}$ 之间相同。相比之下，在式 (3) 的全局零假设检验中，我们考虑一个零假设：显著区域 $\mathcal{M}_{\boldsymbol{X}}$ 中的所有信号元素在 $\boldsymbol{X}$ 和 $\boldsymbol{X}^{\mathrm{ref}}$ 之间都相同。这些双样本检验的 p 值可用于量化显著区域 $\mathcal{M}_{\boldsymbol{X}}$ 的统计显著性。

检验统计量。对于在 $\boldsymbol{X}_{\mathcal{M}_{\boldsymbol{X}}}$ 和 $\boldsymbol{X}^{\mathrm{ref}}_{\mathcal{M}_{\boldsymbol{X}}}$ 之间进行的双样本检验，我们考虑一类称为条件线性检验统计量的检验统计量，其表达式为

$$T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}}) = \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{pmatrix}, \tag{4}$$

和条件 χ 检验统计量，其表达式为

$$T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}}) = \sigma^{-1}\left\|P_{\mathcal{M}_{\boldsymbol{X}}}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{pmatrix}\right\|, \tag{5}$$

其中，$\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}} \in \mathbb{R}^{2n}$ 是一个向量，$P_{\mathcal{M}_{\boldsymbol{X}}} \in \mathbb{R}^{2n\times 2n}$ 是一个依赖于显著性区域 $\mathcal{M}_{\boldsymbol{X}}$ 的投影矩阵。

均值零假设检验和全局零假设检验的检验统计量分别可以写成式 (4) 和式 (5) 的形式。对于式 (2) 中的均值零假设检验，我们考虑以下检验统计量

$$T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}}) = \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{pmatrix} = \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{X}}} X_i - \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{X}}} X_i^{\mathrm{ref}},$$

其中，$\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}} = \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\begin{pmatrix}\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{X}}}\\ -\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{X}}}\end{pmatrix} \in \mathbb{R}^{2n}$. 对于公式（3）中的全局零假设检验，我们考虑以下检验统计量

$$T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}}) = \sigma^{-1}\left\|P_{\mathcal{M}_{\boldsymbol{X}}}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{pmatrix}\right\| = \sqrt{\sum_{i\in\mathcal{M}_{\boldsymbol{X}}}\left(\frac{X_i - X_i^{\mathrm{ref}}}{\sqrt{2}\sigma}\right)^2},$$

其中

$$P_{\mathcal{M}_{\boldsymbol{X}}} = \frac{1}{2}\begin{pmatrix}\mathrm{diag}(\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{X}}}) & -\mathrm{diag}(\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{X}}})\\ -\mathrm{diag}(\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{X}}}) & \mathrm{diag}(\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{X}}})\end{pmatrix}. \tag{6}$$

要获得这些双样本检验的 p 值，我们需要知道检验统计量的抽样分布。遗憾的是，推导检验统计量的抽样分布颇具挑战，因为它们取决于显著区域 $\mathcal{M}_{\boldsymbol{X}}$，而该区域是通过训练好的卷积神经网络中的复杂计算得到的。

3 通过条件选择性推断计算有效的 p-值

本节介绍一种基于条件选择性推断 [Lee 等人， 2016]概念的方法，用于计算查询输入向量 $\boldsymbol{X}$ 与参考输入向量 $\boldsymbol{X}^{\mathrm{ref}}$ 之间显著区域 $\mathcal{M}_{\boldsymbol{X}}$ 的双样本检验的有效 p-值。

3.1 条件分布与选择性 p-值

条件分布。 条件选择性推断的基本思想是，考虑给定选择事件条件下检验统计量的抽样分布。具体而言，我们考虑以下条件分布的抽样性质

$$T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}}) \;\Big|\; \{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{X}_{\mathrm{obs}}}\}, \tag{7}$$

其中， $\boldsymbol{X}_{\mathrm{obs}}$ 是随机向量 $\boldsymbol{X}$的观测值（实现值）。式（7）中的条件表示，在获得与观测到的 $\mathcal{M}_{\boldsymbol{X}^{\mathrm{obs}}}$ 相同的显著区域 $\mathcal{M}_{\boldsymbol{X}}$ 这一事件发生的条件下， $\boldsymbol{X}$ 的随机性。通过以显著区域 $\mathcal{M}_{\boldsymbol{X}}$为条件，条件线性和 χ 检验统计量 T（$\boldsymbol{X}$， $\boldsymbol{X}^{\mathrm{ref}}$）的抽样分布推导，分别可化为对（$\boldsymbol{X}$， $\boldsymbol{X}^{\mathrm{ref}}$）的线性函数和二次函数分布的推导。

选择性 p-值。考虑式（7）中的条件抽样分布后，我们引入以下选择性 p-值:

$$p_{\mathrm{selective}} = \mathbb{P}_{\mathrm{H}_0}\Big(\big|T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}})\big| \geq \big|T(\boldsymbol{X}_{\mathrm{obs}}, \boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}})\big| \;\Big|\; \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{X}_{\mathrm{obs}}},\ \mathcal{Q}_{\boldsymbol{X},\boldsymbol{X}^{\mathrm{ref}}} = \mathcal{Q}_{\mathrm{obs}} \Big), \tag{8}$$

其中

$$\mathcal{Q}_{\boldsymbol{X},\boldsymbol{X}^{\mathrm{ref}}} = \Omega_{\boldsymbol{X},\boldsymbol{X}^{\mathrm{ref}}}, \quad \mathcal{Q}_{\mathrm{obs}} = \mathcal{Q}_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}}$$

with

$$\Omega_{\boldsymbol{X},\boldsymbol{X}^{\mathrm{ref}}} = \left(I_{2n} - \frac{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}} \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top}}{\|\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}\|^2} \right) \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}^{\mathrm{ref}} \end{pmatrix} \in \mathbb{R}^{2n}$$

在均值零假设检验的情况下，以及

$$\mathcal{Q}_{\boldsymbol{X},\boldsymbol{X}^{\mathrm{ref}}} = \{\mathcal{V}_{\boldsymbol{X},\boldsymbol{X}^{\mathrm{ref}}}, \mathcal{U}_{\boldsymbol{X},\boldsymbol{X}^{\mathrm{ref}}}\}, \quad \mathcal{Q}_{\mathrm{obs}} = \mathcal{Q}_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}}$$

with

$$\mathcal{V}_{\boldsymbol{X},\boldsymbol{X}^{\mathrm{ref}}} = \sigma P_{\mathcal{M}_{\boldsymbol{X}}} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}^{\mathrm{ref}} \end{pmatrix} \Big/ \left\| P_{\mathcal{M}_{\boldsymbol{X}}} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}^{\mathrm{ref}} \end{pmatrix} \right\| \in \mathbb{R}^{2n}, \quad \mathcal{U}_{\boldsymbol{X},\boldsymbol{X}^{\mathrm{ref}}} = P^{\perp}_{\mathcal{M}_{\boldsymbol{X}}} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}^{\mathrm{ref}} \end{pmatrix} \in \mathbb{R}^{2n}$$

在全局零假设检验的情况下。$\mathcal{Q}_{\boldsymbol{X},\ \boldsymbol{X}^{\mathrm{ref}}}$ 是干扰参数的充分统计量，为了便于进行推断，需要以其为条件 [1]。

式（8）中的选择性 p 值具有以下所需的抽样性质

$$\mathbb{P}_{\mathrm{H}_0}\left(p_{\text{selective}} \leq \alpha \mid \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{X}_{\mathrm{obs}}}\right) = \alpha, \quad \forall \alpha \in [0,1]. \tag{9}$$

这意味着，选择性 p 值 $p_{\text{selective}}$ 可用作显著区域 $\mathcal{M}_{\boldsymbol{X}}$ 的有效统计显著性度量。

3.2 条件数据空间的刻画

为计算式（8）中的选择性 p 值，我们需要刻画条件数据空间，其刻画将在下一节中介绍。我们将满足式（8）中条件的 $(\boldsymbol{X}\ \boldsymbol{X}^{\mathrm{ref}})^{\top} \in \mathbb{R}^{2n}$ 集合定义为

$$\mathcal{D} = \left\{ (\boldsymbol{X}\ \boldsymbol{X}^{\mathrm{ref}})^{\top} \in \mathbb{R}^{2n} \;\middle|\; \mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{X}_{\mathrm{obs}}}, \mathcal{Q}_{\boldsymbol{X},\boldsymbol{X}^{\mathrm{ref}}} = \mathcal{Q}_{\mathrm{obs}} \right\}. \tag{10}$$

根据第二个条件，$\mathcal{D}$ 中的数据被限制在 $\mathbb{R}^{2n}$ 中的一条直线上，如以下引理所述。

引理 1. 令我们定义，

$$\boldsymbol{a} = \Omega_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}} \quad \textit{and} \quad \boldsymbol{b} = \frac{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}}{\|\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}\|^2} \in \mathbb{R}^{2n}. \tag{11}$$

在均值零假设检验中，以及

$$\boldsymbol{a} = \mathcal{U}_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}} \quad \textit{and} \quad \boldsymbol{b} = \mathcal{V}_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}} \tag{12}$$

在全局零假设检验的情况下。然后，通过使用标量参数 $z \in \mathbb{R}$，式（*10*）中的集合 $\mathcal{D}$ 可改写为 $\mathcal{D} = \left\{ \left(\boldsymbol{X}\ \boldsymbol{X}^{\mathrm{ref}}\right)^{\top} = \boldsymbol{a} + \boldsymbol{b}z \mid z \in \mathcal{Z} \right\}$，其中

$$\mathcal{Z} = \left\{ z \in \mathbb{R} \mid \mathcal{M}_{\boldsymbol{a}_{1:n} + \boldsymbol{b}_{1:n} z} = \mathcal{M}_{\boldsymbol{X}_{\mathrm{obs}}} \right\}. \tag{13}$$

$\boldsymbol{x}_{1:n}$ 表示 $\boldsymbol{x}$ 中从1 到 n 的元素向量。

[1]该干扰参数 $\mathcal{Q}_{\boldsymbol{X},\ \boldsymbol{X}^{\mathrm{ref}}}$ 对应于开创性的条件选择性推断论文 [Lee 等人2016] 中的分量 $\boldsymbol{z}$（见第 5 节、公式 5.2 和定理 5.2），以及 $\boldsymbol{z}$，$\boldsymbol{w}$ Chen 和 Bien 2019]中的分量 [（见第 3 节、定理 3.7）。我们注意到，在条件选择性推断文献中，额外地对 $\mathcal{Q}_{\boldsymbol{X},\ \boldsymbol{X}^{\mathrm{ref}}}$ 进行条件化是一种标准做法，几乎所有条件选择性推断相关研究都采用了这一做法。这里需要指出，选择性 p-值取决于 $\mathcal{Q}_{\boldsymbol{X},\ \boldsymbol{X}^{\mathrm{ref}}}$，但由于我们可以对 $\mathcal{Q}_{\boldsymbol{X},\ \boldsymbol{X}^{\mathrm{ref}}}$ 的所有取值进行边缘化，因此无需这一额外条件，性质 (9) 仍然成立（参见 Lee 等人 [2016] 中定理 5.2 证明的后半部分，以及 Chen 和 Bien [2019] 中定理 3.7 的证明）。

证明。 证明详见附录 A.1 ■

引理 1 表明，我们无需考虑 $2n$ 维数据空间。相反，我们只需考虑 (13) 中的一维投影数据空间 $\mathcal{Z}$ 。现在，考虑一个随机变量 $Z \in \mathbb{R}$ 及其观测值 $Z_{\text{obs}} \in \mathbb{R}$，它们满足 $(\boldsymbol{X} \ \boldsymbol{X}^{\text{ref}})^\top = \boldsymbol{a} + \boldsymbol{b}Z$ 和 $(\boldsymbol{X}_{\text{obs}} \ \boldsymbol{X}^{\text{ref}}_{\text{obs}})^\top = \boldsymbol{a} + \boldsymbol{b}Z_{\text{obs}}$。选择性 p-值 (8) 可改写为

$$p_{\text{selective}} = \mathbb{P}_{\text{H}_0}\left(|Z| \geq |Z_{\text{obs}}| \mid Z \in \mathcal{Z}\right). \tag{14}$$

在零假设下，均值零假设检验中的变量 $Z \sim \mathbb{N}(0, \sigma^2\|\boldsymbol{\eta}\|^2)$ 和全局零假设检验中的变量 $Z \sim \chi$ (Trace(P)) 分别服从截断正态分布和截断 χ 分布。一旦确定截断区域 $\mathcal{Z}$，计算式 (14) 中的选择性 p 值便十分直接。因此，剩下的任务是确定 $\mathcal{Z}$。

一般而言，计算式 (13) 中的 $\mathcal{Z}$ 很困难，因为我们需要针对 $z \in \mathbb{R}$ 的所有取值确定选择事件 $\mathcal{M}_{\boldsymbol{a}_{1:n}+\boldsymbol{b}_{1:n}z}$，这在计算上具有挑战性。在下一节中，我们将说明，对于一大类问题，这一难题可以得到解决。

4 分段线性网络

针对选定显著区域计算选择性 p 值的问题，可以转化为确定一组区间 $\mathcal{Z} = \{z \in \mathbb{R} \mid \mathcal{M}_{\boldsymbol{X}(z)} = \mathcal{M}_{\boldsymbol{X}_{\text{obs}}}\}$ 的问题。考虑到训练好的深度神经网络中的显著性计算十分复杂，似乎很难获得 $\mathcal{Z}$。然而，本节将说明，对于一大类卷积神经网络，这是可行的。

卷积神经网络中的分段线性组件关键在于注意到，训练好的卷积神经网络中使用的大多数基本运算和常见激活函数，都可以表示为以下形式的分段线性函数:

定义 **1.** （分段线性函数）分段线性函数 $f\colon \mathbb{R}^n \mapsto \mathbb{R}^m$ 定义为 *s:*

$$f(\boldsymbol{X}) = \begin{cases} \Psi_1^f \boldsymbol{X} + \boldsymbol{\psi}_1^f, & \text{if } \boldsymbol{X} \in \mathcal{P}_1^f := \{\boldsymbol{X}' \in \mathbb{R}^n \mid \Delta_1^f \boldsymbol{X}' \leq \boldsymbol{\delta}_1^f\}, \\ \Psi_2^f \boldsymbol{X} + \boldsymbol{\psi}_2^f, & \text{if } \boldsymbol{X} \in \mathcal{P}_2^f := \{\boldsymbol{X}' \in \mathbb{R}^n \mid \Delta_2^f \boldsymbol{X}' \leq \boldsymbol{\delta}_2^f\}, \\ \quad \vdots & \\ \Psi_{K(f)}^f \boldsymbol{X} + \boldsymbol{\psi}_{K(f)}^f, & \text{if } \boldsymbol{X} \in \mathcal{P}_{K(f)}^f := \{\boldsymbol{X}' \in \mathbb{R}^n \mid \Delta_{K(f)}^f \boldsymbol{X}' \leq \boldsymbol{\delta}_{K(f)}^f\} \end{cases}$$

其中，Ψ_k^f、 $\boldsymbol{\psi}_k^f$、 Δ_k^f和 $\boldsymbol{\delta}_k^f$对于 $k \in [K(f)]$ 是维度适当的某些矩阵和向量，$\mathcal{P}_k^f$: $= \{\boldsymbol{x} \in \mathbb{R}^n \mid \Delta_k^f \boldsymbol{x} \leq \boldsymbol{\delta}_k^f\}$ 是 $\mathbb{R}^n$ 中的一个多胞体，适用于 $k \in [K(f)],$，而 $K(f)$ 是函数 f对应的多胞体数量。

Ex训练好的卷积神经网络中的分段线性组件示例见附录 A.2。

分段线性网络

定义 2。（分段线性网络）由分段线性函数拼接和复合而成的网络称为分段线性网络。

由于分段线性函数的拼接和复合显然仍是分段线性函数，因此分段线性网络中任意节点的输出都可以表示为输入向量 $\boldsymbol{X}$ 的分段线性函数。显著性图函数 $\mathcal{A}_i(\boldsymbol{X})$ $i \in [n]$ 也具有这一性质。此外，如 §4 所述，我们可以将输入向量限制为 $\boldsymbol{X}(z) = \boldsymbol{a}_{1:n} + \boldsymbol{b}_{1:n} z$ 的形式，其中由标量参数 $z \in \mathbb{R}$ 参数化。因此，每个元素的显著性图值都可以表示为标量参数 z 的分段线性函数，即

$$\mathcal{A}_i(\boldsymbol{X}(z)) = \begin{cases} \kappa_1^{\mathcal{A}_i} z + \rho_1^{\mathcal{A}_i}, & \text{if } \ z \in [L_1^{\mathcal{A}_i}, U_1^{\mathcal{A}_i}], \\ \kappa_2^{\mathcal{A}_i} z + \rho_2^{\mathcal{A}_i}, & \text{if } \ z \in [L_2^{\mathcal{A}_i}, U_2^{\mathcal{A}_i}], \\ \quad \vdots & \\ \kappa_{K(\mathcal{A}_i)}^{\mathcal{A}_i} z + \rho_{K(\mathcal{A}_i)}^{f}, & \text{if } \ z \in [L_{K(\mathcal{A}_i)}^{\mathcal{A}_i}, U_{K(\mathcal{A}_i)}^{\mathcal{A}_i}], \end{cases} \tag{15}$$

其中 $K(\mathcal{A}_i)$ 是该分段线性函数的线性片段数，$\kappa_k^{\mathcal{A}_i}$，$\rho_k^{\mathcal{A}_i}$ 是某些标量参数，$[L_k^{\mathcal{A}_i}, U_k^{\mathcal{A}_i}]$ 是 $k \in [K(\mathcal{A}_i)]$ 的区间（注意，将 $\mathbb{R}^n$ 中的多胞体投影到一维空间时，它会缩减为一个区间）。

这意味着，对于分段线性函数的每一段，我们都可以如下确定 z 的区间，使得 $\mathcal{A}_i(\boldsymbol{X}(z)) \geq \tau$ [2]

$$z \in \begin{cases} \left[\max\left(L_k^{\mathcal{A}_i}, \left(\tau - \rho_k^{\mathcal{A}_i}\right)/\kappa_k^{\mathcal{A}_i}\right),\ U_k^{\mathcal{A}_i}\right] & \text{if } \kappa_k^{\mathcal{A}_i} > 0 \\ \left[L_k^{\mathcal{A}_i}, \min\left(U_k^{\mathcal{A}_i}, \left(\tau - \rho_k^{\mathcal{A}_i}\right)/\kappa_k^{\mathcal{A}_i}\right),\right] & \text{if } \kappa_k^{\mathcal{A}_i} < 0 \end{cases} \quad \Rightarrow \quad \mathcal{A}_i(\boldsymbol{X}(z)) \geq \tau. \tag{16}$$

稍微滥用一下记号，我们将 $z \in \mathbb{R}$ 上由 $L_k^{\mathcal{A}_i}, U_k^{\mathcal{A}_i}, (\tau - \rho_i^{\mathcal{A}_i}/\kappa_k^{\mathcal{A}_i})$ 对所有 $(k, i) \in [K(\mathcal{A}_i)] \times [n]$ 定义的有限个区间统称为

$$[z_0, z_1], [z_1, z_2], \ldots, [z_{t-1}, z_t], [z_t, z_{t+1}], \ldots, [z_{T-1}, z_T],$$

其中，$z_{\min} = z_0$ 和 $z_{\max} = z_T$ 的定义使得 $z < z_{\min}$ 和 $z > z_{\max}$ 的概率质量都可以忽略不计。

[2]为简便起见，我们省略 $\kappa_k^{\mathcal{A}_i} = 0$ 情况的描述。在这种情况下，如果 $\rho_k^{\mathcal{A}_i} \geq \tau$，那么 $z \in [L_k^{\mathcal{A}_i}, U_k^{\mathcal{A}_i}] \Rightarrow i \in \mathcal{M}_{\boldsymbol{X}(z)}$。

算法1 SI DNN 显著性_

输入： $\boldsymbol{X}^{\text{obs}}$，$z_{\min}$，$z_{\max}$，$\mathcal{T} \leftarrow \emptyset$

1: 获取 $\mathcal{E}_{\text{obs}}$，计算 $\boldsymbol{\eta}$ 以及 $\boldsymbol{a}$ 和 $\boldsymbol{b} \leftarrow$ 公式 (12)，并初始化： $t = 1$, $z_t = z_{\min}$

2: **对于** $t \leq T$ **执行**

3: 通过自动条件化计算 z_{t+1} （参见 §5）

4: **如果** $\mathcal{E}_{\boldsymbol{X}(z),\ \boldsymbol{X}^{\text{ref}}(z)} = \mathcal{E}_{\text{obs}}$ 在 $z \in [z_t, z_{t+1}]$ 中（使用公式 (16)）**则**

5: $\mathcal{T} \leftarrow \mathcal{T} + \{t\}$

6: **结束条件判断**

7: $t = t + 1$

8: **结束循环**

9: 识别 $\mathcal{Z} \leftarrow \bigcup_{t \in \mathcal{T}} [z_t, z_{t+1}]$

10: $p_{\text{selective}} \leftarrow$ 公式 (14)

输出： $p_{\text{selective}}$

算法算法1展示了我们如何识别$\mathcal{Z} = \{z \in \mathbb{R} \mid \mathcal{M}_{\boldsymbol{X}(z),\boldsymbol{X}^{\text{ref}}(z)} = \mathcal{M}_{\text{obs}}\}$。我们只需按照 z的顺序检查 $[z_0, z_1]$、$[z_1, z_2]$、……、$[z_{T-1}, z_T]$，并利用式（16）判断$\mathcal{M}_{\boldsymbol{X}(z)} = \mathcal{M}_{\boldsymbol{X}(z_{\text{obs}})}$是否位于区间内。然后，式（13）中的截断区域 $\mathcal{Z}$ 为 $\mathcal{Z} = \bigcup_{t \in [T] \mid \mathcal{E}_{\boldsymbol{X}(z),\ \boldsymbol{X}^{\text{ref}}(z)} = \mathcal{E}_{\text{obs}} \text{ 对于 } z \in [z_t, z_{t+1}]} [z_t, z_{t+1}]$.

5 实现：自动条件化

我们算法的瓶颈在于算法1的第3行，其中， z_{t+1} 必须通过考虑复杂的训练好的卷积神经网络中的所有相关分段线性组件来求得。困难不仅在于计算成本，还在于实现成本。若要在深度神经网络中朴素地实现条件选择性推断，就必须将网络每一层的所有操作刻画为选择事件，并逐一实现每个特定操作[Duy et al.，2022] 。为克服这一困难，我们引入了一种称为自动条件化的模块化实现方案，其概念与自动微分 [Baydin et al.，2018]类似。借助该方案，我们无需额外的实现成本，即可对一大类卷积神经网络进行条件选择性推断。

自动条件化的基本思想是添加一种机制，为网络中的每个分段线性组件 f（例如 Keras框架中的层 API）计算并维护区间$z \in [L_k^f, U_k^f]$。这样，当分段线性函数 f 由多个分段线性组件拼接和/或复合而成时，我们便能自动计算其区间 $[L_k^f, U_k^f]$。如果 f 由两个分段线性函数 f_1 和 f_2拼接而成，我们可以轻松得到 $[L_k^f, U_k^f] = [L_{k_1}^{f_1}, U_{k_1}^{f_1}] \cap [L_{k_2}^{f_2}, U_{k_2}^{f_2}]$。然而，如果 f是由两个分段线性函数 f_1 和 f_2复合而成，则区间的计算如下引理所述。

引理2。　　考虑两个分段线性函数的复合，即 $f(\boldsymbol{X}(z)) = (f_2 \circ$

$f_1)(\boldsymbol{X}(z))$。给定 z 的一个实数值，可以计算出 f_2 输入域中的区间 $[L_k^{f_2}, U_k^{f_2}]$，其计算方式为

$$L_{k_2}^{f_2} = \max_{j:(\Delta_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1})_j<0} \frac{(\boldsymbol{\delta}_{k_2}^{f_2})_j - (\Delta_{k_2}^{f_2}\boldsymbol{\beta}^{f_1})_j}{(\Delta_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1})_j}, \qquad U_{k_2}^{f_2} = \min_{j:(\Delta_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1})_j>0} \frac{(\boldsymbol{\delta}_{k_2}^{f_2})_j - (\Delta_{k_2}^{f_2}\boldsymbol{\beta}^{f_1})_j}{(\Delta_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1})_j},$$

其中，$\boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z$ 是 f_1 的输出（即 f_2的输入）。此外，$\Delta_{k_2}^{f_2}$ 和 $\boldsymbol{\delta}_{k_2}^{f_2}$ 通过验证 $\boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z$的取值获得。然后，复合函数的区间可按如下方式得到：$[L_k^f, U_k^f] = [L_{k_1}^{f_1}, U_{k_1}^{f_1}] \cap [L_{k_2}^{f_2}, U_{k_2}^{f_2}]$

证明见附录 A.3。这里，变量 $\boldsymbol{\beta}^{f_k}$ 和 $\boldsymbol{\gamma}^{f_k}$ 可以通过各层递归计算如下：

$$\boldsymbol{\beta}^{f_{k+1}} = \Psi_k^{f_k}\boldsymbol{\beta}^{f_k} + \boldsymbol{\psi}_k^{f_k} \quad \text{and} \quad \boldsymbol{\gamma}^{f_{k+1}} = \Psi_k^{f_k}\boldsymbol{\gamma}^{f_k}.$$

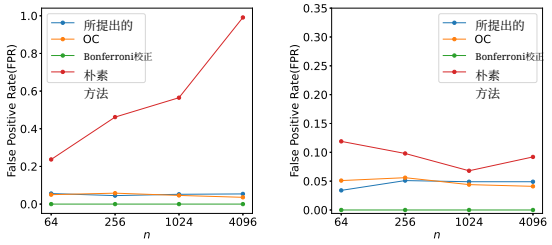
引理2表明，$\boldsymbol{X}$（z）递减的区间可以通过这些层向前传播。这意味着，式（15）中分段线性函数当前线性段的下界$L_k^{\mathcal{A}_i}$和上界 $U_k^{\mathcal{A}_i}$，可以通过对相关分段线性组件的区间进行前向传播自动计算得到。

6 实验

我们仅重点介绍主要结果。更多详情（比较方法、网络结构等）见附录 A.4。

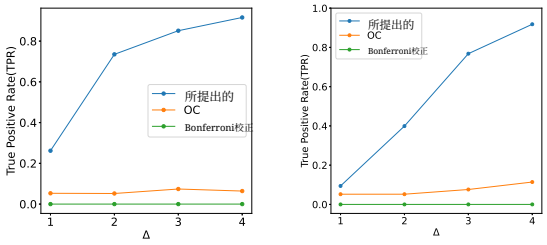
实验设置。 我们将所提出的方法与朴素方法、过度条件化（OC）方法以及 Bonferroni 校正进行了比较。为考察假阳性率（FPR），我们考虑了 1000 张零假设图像 $\boldsymbol{X} = (X_1, ..., X_n)$ 和 1000 张参考图像 $\boldsymbol{X}^{\text{ref}} = (X_1^{\text{ref}}, ..., x_n^{\text{ref}})$，其中$\boldsymbol{s} = \boldsymbol{s}^{\text{ref}} = \boldsymbol{0}$和 $\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^{\text{ref}} \sim \mathbb{N}(\boldsymbol{0}, I_n)$，对每个 $n \in \{64, 256, 1024, 4096\}$ 均如此。为考察真正率（ TPR ），我们设定 $n = 256$，并生成了 1,000 张图像，其中 $s_i =$信号适用于任意 $i \in \mathcal{S}$，其中$\mathcal{S}$ 是“真实”显著区域，其位置随机确定。 $s_i = 0$ 适用于任意 $i \notin \mathcal{S}$，且 $\boldsymbol{\varepsilon} \sim \mathbb{N}(\boldsymbol{0}, I_n)$。我们设定 $\Delta \in \{1, 2, 3, 4\}$。参考图像以与 FPR 情况相同的方式生成。在所有实验中，我们在均值零假设检验中将选择显著区域的阈值设为 $\tau = 0$，在全局零假设检验中设为 $\tau = 5$。我们将显著性水平设为 $\alpha = 0.05$。所有实验均使用 CAM 作为显著性方法。

数值结果。 假阳性率控制的结果见图 2。所提出的方法、OC方法和Bonferroni校正，在均值检验和全局零假设检验两种情况下均成功控制了假阳性率，而



（a）均值零假设检验　　（b）全局零假设检验

图 2：假阳性率（FPR）比较。



（a）均值零假设检验　　（b）全局零假设检验

图3：真正率（假阳性率）对比。

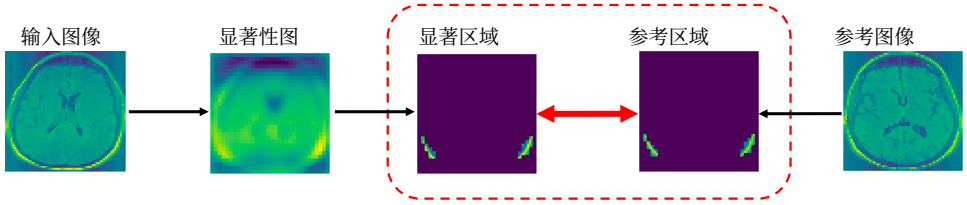


图4：无肿瘤图像的均值零假设检验（$p_{\text{naive}} = \mathbf{0.00}$，　$p_{\text{selective}} = \mathbf{0.78}$）。

其他方法则未能做到。由于朴素方法未能控制假阳性率，我们不再考虑其真正率。真正率比较结果见图 3。所提出的方法在所有情况下的真正率均最高。Bonferroni方法的真正率最低，因为它考虑了所有可能假设的数量，因此较为保守。OC方法的真正率也低于所提出的方法，因为它考虑了若干额外条件，从而导致真正率损失。

真实数据实验。我们考察了从 Buda等人 使用的数据集中提取的脑部图像数据集 [2019], ，其中分别包含 939 张有肿瘤图像和 941 张无肿瘤图像。均值零假设检验的结果见图4和图5。全局零假设检验的结果见图6和图7。即使图像中没有肿瘤区域，朴素方法的 p-值仍然很小，这表明朴素方法的 p-值无法用于量化基于DNN的显著区域的可靠性。所提方法成功识别出了假阳性和真阳性检测。

7 结论

在本研究中，我们基于条件选择性推断的概念，提出了一种用于对深度神经网络驱动的显著区域的重要性进行统计推断的新方法。我们还提出了一种新颖的算法，能够高效、灵活地对显著区域进行条件选择性推断。我们在

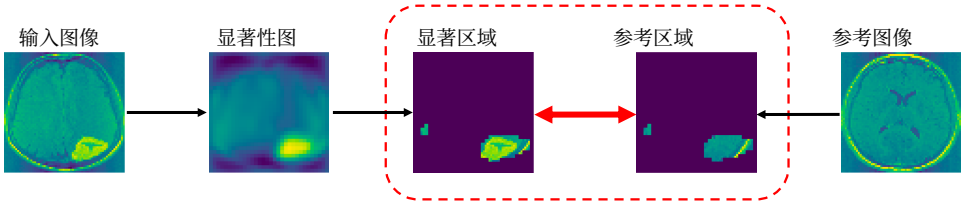


图 5：含肿瘤图像的均值零假设检验（$p_{\text{naive}} = \mathbf{0.00}$, $p_{\text{selective}} = \mathbf{1.92 \times 10^{-4}}$）。

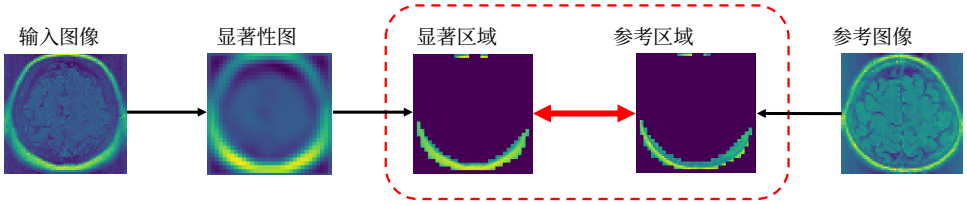


图 6：不含肿瘤图像的全局零假设检验（$p_{\text{naive}} = \mathbf{0.03}$, $p_{\text{selective}} = \mathbf{0.46}$）

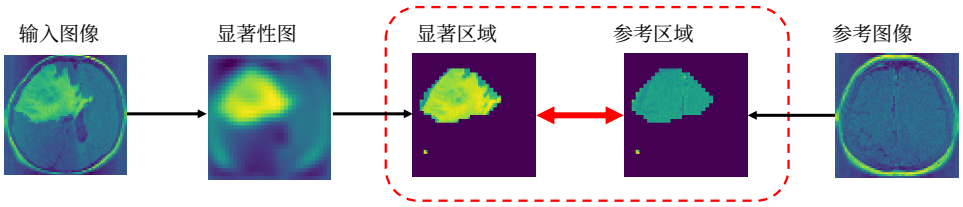


Figure 7：含肿瘤图像的全局零假设检验（$p_{\text{naive}} = \mathbf{0.00}$, $p_{\text{selective}} = \mathbf{1.51 \times 10^{-3}}$）.

合成数据集和真实世界数据集上进行了实验，以展示所提方法的性能。

致谢

本研究部分获得文部科学省科学研究费助成事业（20H00601）、日本科学技术振兴机构CREST（JPMJCR21D3）、日本科学技术振兴机构Moonshot研发计划（JPMJMS2033-05）、日本科学技术振兴机构AIP加速研究（JPMJCR21U2）、新能源与产业技术综合开发机构（JPNP18002、JPNP20006）以及理化学研究所先进智能项目的资助。

参考文献

S. Bach、A. Binder、G. Montavon、F. Klauschen、K.-R. M¨uller 和 W. Samek。通过逐层相关性传播解释非线性分类器决策的逐像素解释。*PloS one*，10(7)：e0130140，2015年。

A. G. Baydin、B. A. Pearlmutter、A. A. Radul 和 J. M. Siskind。机器学习中的自动微分：综述。《机器学习研究杂志》，18：1–43，2018。

M. Buda、A. Saha 和 M. A. Mazurowski。深度学习算法自动提取的形状特征与低级别胶质瘤基因组亚型的关联。《生物与医学计算机》，109：218–225，2019。C. Burns、J. Thomason 和 W. Tansey。通过假设检验解释黑箱模型。载于*2020*年 *ACM-IMS* 数据科学基础会议论文集，第47–57页，2020年。S. Chen 和 J. Bien。针对异常值剔除进行校正的有效推断。《计算与图形统计学杂志》，第1–12页，2019年。D. Das、V. N. L. Duy、H. Hanada、K. Tsuda 和 I. Takeuchi。针对稀疏高阶交互模型的快速且更强大的选择性推断。*arXiv* 预印本 *arXiv:2106.04929*，2021年。A.-K. Dombrowski、M. Alber、C. Anders、M. Ackermann、K.-R. M¨uller 和 P. Kessel。解释可能受到操纵，问题根源在于几何。载于神经信息处理系统进展，第13589–13600页，2019年。F. Doshi-Velez 和 B. Kim。迈向可解释机器学习的严谨科学。*arXiv* 预印本 *arXiv:1702.08608*，2017年。A. Dosovitskiy 和 T. Brox。利用卷积网络反演视觉表征。载于*IEEE*计算机视觉与模式识别会议论文集，第4829–4837页，2016年。V. N. L. Duy 和 I. Takeuchi。通过参数规划，为广义套索实现更强大的条件选择性推断。*arXiv* 预印本 *arXiv:2105.04920*，2021年。V. N. L. Duy 和 I. Takeuchi。通过选择性推断，利用动态时间规整对时间序列相似性进行精确统计推断。*arXiv* 预印本 *arXiv:2202.06593*，2022年。V. N. L. Duy、H. Toda、R. Sugiyama 和 I. Takeuchi。通过动态规划进行选择性推断，计算最优变点的有效p值。载于神经信息处理系统进展，第11356–11367页，2020年。V. N. L. Duy、S. Iwazaki 和 I. Takeuchi。通过选择性推断量化基于神经网络的图像分割的统计显著性。神经信息处理系统进展，2022年。W. Fithian、J. Taylor、R. Tibshirani 和 R. Tibshirani。选择性序贯模型选择。*arXiv* 预印本 *arXiv:1512.02565*，2015年。R. C. Fong 和 A. Vedaldi。通过有意义的扰动解释黑箱模型。载于*IEEE*国际计算机视觉会议论文集，第3429–3437页，2017年。

A. Ghorbani、A. Abid 和 J. Zou。神经网络的诠释很脆弱。载于*AAAI*人工智能会议论文集，第 33 卷，第 3681–3688 页，2019年。J. Heo、S. Joo 和 T. Moon。通过对抗性模型操纵愚弄神经网络诠释。载于 神经信息处理系统进展，第 2925–2936 页，2019年。S. Hyun、M. G'sell 和 R. J. Tibshirani。广义套索路径的精确选择后推断。电子统计学杂志，12(1):1053–1097，2018年。S. Hyun、K. Z. Lin、M. G'Sell 和 R. J. Tibshirani。用于变点检测算法的选择后推断及其在拷贝数变异数据中的应用。生物统计学，77(3):1037–1049，2021年。P.-J. Kindermans、S. Hooker、J. Adebayo、M. Alber、K. T. Sch¨utt、S. D¨ahne、D. Erhan 和 B. Kim。显著性方法的（不）可靠性。*arXiv* 预印本 *arXiv:1711.00867*，2017年。V. N. Le Duy 和 I. Takeuchi。采用参数规划方法实现更强大且更通用的套索选择性推断。载于人工智能与统计国际会议，第 901–909 页。PMLR，2021年。J. D. Lee、D. L. Sun、Y. Sun 和 J. E. Taylor。精确的选择后推断及其在套索中的应用。《统计年刊》，44(3):907–927，2016。K. Liu、J. Markovic 和 R. Tibshirani。更强大的选择后推断及其在套索中的应用。*arXiv* 预印本 *arXiv:1801.09037*，2018年。S. M. Lundberg 和 S.-I. Lee。诠释模型预测的统一方法。载于 神经信息处理系统进展，第 4765–4774 页，2017年。A. Mahendran 和 A. Vedaldi。通过反演深层图像表征来理解它们。载于*IEEE*计算机视觉与模式识别会议论文集，第 5188–5196 页，2015年。D. A. Melis 和 T. Jaakkola。借助自解释神经网络实现稳健的可解释性。载于神经信息处理系统进展，第 7775–7784 页，2018年。S. Panigrahi、J. Taylor 和 A. Weinstein。线性模型中的贝叶斯选择后推断。*arXiv* 预印本 *arXiv:1605.08824*，28，2016年。M. T. Ribeiro、S. Singh 和 C. Guestrin。"我为什么应该信任你？"解释任意分类器的预测。载于 第 *22* 届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集，第 1135–1144 页，2016年。

R. R. Selvaraju、M. Cogswell、A. Das、R. Vedantam、D. Parikh 和 D. Batra。Grad-CAM：通过基于梯度的定位实现深度网络的视觉解释。载于*IEEE*计算机视觉国际会议论文集，第 618–626 页，2017年。K. Simonyan、A. Vedaldi 和 A. Zisserman。深入卷积网络内部：图像分类模型和显著性图的可视化。*arXiv* 预印本 *arXiv:1312.6034*，2013年。K. Sugiyama、V. N. L. Duy 和 I. Takeuchi。采用同伦延拓方法，为逐步特征选择实现更强大且更通用的选择性推断。载于第 *38* 届国际机器学习会议论文集，2021a。

R. Sugiyama、H. Toda、V. N. L. Duy、Y. Inatsu 和 I. Takeuchi。通过选择性推断对多维多个变点进行有效且精确的统计推断。*arXiv* 预印本 *arXiv:2110.08989*，2021b。

S. Suzumura、K. Nakagawa、Y. Umezu、K. Tsuda 和 I. Takeuchi。稀疏高阶交互模型的选择性推断。载于第*34*届国际机器学习会议论文集，第*70*卷，第3338–3347页。JMLR.org，2017年。

K. Tanizaki、N. Hashimoto、Y. Inatsu、H. Hontani 和 I. Takeuchi。通过选择性推断计算图像分割的有效p值。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第9553–9562页，2020年。

X. Tian 和 J. Taylor。带随机化响应的选择性推断。《统计年刊》，46 (2):679–710，2018年。R. J. Tibshirani、J. Taylor、R. Lockhart 和 R. Tibshirani。序贯回归程序的精确选择后推断。《美国统计学会杂志》，111(514):600–620，2016年。T. Tsukurimichi、Y. Inatsu、V. N. L. Duy 和 I. Takeuchi。使用分段线性同伦延拓进行稳健回归和异常值检测的条件选择性推断。*arXiv* 预印本 *arXiv:2104.10840*，2021年。F. Yang、R. F. Barber、P. Jain 和 J. Lafferty。组稀疏线性模型的选择性推断。载于神经信息处理系统进展，第 2469–2477 页，2016年。M. D. Zeiler 和 R. Fergus。卷积网络的可视化与理解。载于欧洲计算机视觉会议，第 818–833 页。施普林格，2014年。

X. Zhang、N. Wang、H. Shen、S. Ji、X. Luo 和 T. Wang。火线下的可解释深度学习。载于第 *29* 届 {*USENIX*} 安全研讨会（{*USENIX*} 安全 *20*），2020年。

B. Zhou、A. Khosla、A. Lapedriza、A. Oliva 和A. Torralba。学习用于判别性定位的深度特征。载于*IEEE*计算机视觉与模式识别会议论文集，第 2921–2929 页，2016 年。

A 附录

A.1 引理 1 的证明

在均值零假设检验中，根据式 (10) 的第二个条件，我们有

$$\begin{aligned}
\Omega_{\boldsymbol{X},\boldsymbol{X}^{\mathrm{ref}}} &= \Omega_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}} \\
\Leftrightarrow \quad \left(I_{2n} - \frac{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top}}{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top}\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}}\right)\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{pmatrix} &= \Omega_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}} \\
\Leftrightarrow \quad \begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{pmatrix} &= \Omega_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}} + \frac{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}}{\|\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}\|^2}\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{pmatrix}.
\end{aligned}$$

通过定义 $\boldsymbol{a} = \boldsymbol{q}_{\boldsymbol{X}_{\mathrm{obs}},\ \boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}}$、$\boldsymbol{b} = \frac{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}}{\|\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}\|^2}$、$z = \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top}\left(\begin{smallmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{smallmatrix}\right)$，我们得到引理 1 中的结果。

在全局零假设检验中，根据式 (10) 中的第二个条件，

$$\begin{aligned}
\mathcal{U}_{\boldsymbol{X},\boldsymbol{X}^{\mathrm{ref}}} &= \mathcal{U}_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}} \\
\Leftrightarrow P^{\perp}_{\mathcal{M}_{\boldsymbol{X}}}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{pmatrix} &= \mathcal{U}_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}} \\
\Leftrightarrow (I_{2_n} - P_{\mathcal{M}_{\boldsymbol{X}}})\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{pmatrix} &= \mathcal{U}_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}} \\
\Leftrightarrow \begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{pmatrix} &= \mathcal{U}_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}} + \mathcal{V}_{\boldsymbol{X}_{\mathrm{obs}},\boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}}\sigma^{-1}\left\|P_{\mathcal{M}_{\boldsymbol{X}}}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{pmatrix}\right\|.
\end{aligned}$$

通过定义 $\boldsymbol{a} = \mathcal{U}_{\boldsymbol{X}_{\mathrm{obs}},\ \boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}}$, $\boldsymbol{b} = \mathcal{V}_{\boldsymbol{X}_{\mathrm{obs}},\ \boldsymbol{X}^{\mathrm{ref}}_{\mathrm{obs}}}$, $z = \sigma^{-1}\left\|P_{\mathcal{M}_{\boldsymbol{X}}}\left(\begin{smallmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\mathrm{ref}}\end{smallmatrix}\right)\right\|$，我们得到引理 1 的结论。

A.2 分段线性函数的示例

示例 o　　训练好的卷积神经网络中包含 $\boldsymbol{X} \in \mathbb{R}^2$ 个分段线性组件，具体如下：

$ReLU$: 设 f 为 ReLU 函数。那么，对于任意 $k \in [4]$, $K(f) = 4$, $\boldsymbol{\psi}_k = (0\ 0)^\top$,

$$\Psi_1^f = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix},\ \mathcal{P}_1^f = \left\{ \boldsymbol{X} : \begin{array}{l} X_1 < 0, \\ X_2 < 0 \end{array} \right\}, \quad \Psi_2^f = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},\ \mathcal{P}_2^f = \left\{ \boldsymbol{X} : \begin{array}{l} X_1 < 0, \\ X_2 \geq 0 \end{array} \right\},$$
$$\Psi_3^f = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},\ \mathcal{P}_3^f = \left\{ \boldsymbol{X} : \begin{array}{l} X_1 \geq 0, \\ X_2 < 0 \end{array} \right\}, \quad \Psi_4^f = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ \mathcal{P}_4^f = \left\{ \boldsymbol{X} : \begin{array}{l} X_1 \geq 0, \\ X_2 \geq 0 \end{array} \right\}.$$

这也可以类似地扩展到 Leaky ReLU 的情况。

最大池化：考虑 $f(\boldsymbol{X}) = \max\{X_1, X_2\}$。然后，对于任意 $k \in [2]$, 它可以表示为具有 $K(f) = 2$, $\boldsymbol{\psi}_k = (0)$ 的分段线性函数,

$$\Psi_1^f = \begin{pmatrix} 1 & 0 \end{pmatrix},\ \mathcal{P}_1^f = \{\boldsymbol{X} : X_1 \geq X_2\}, \quad \Psi_2^f = \begin{pmatrix} 0 & 1 \end{pmatrix},\ \mathcal{P}_2^f = \{\boldsymbol{X} : X_1 < X_2\}.$$

卷积与矩阵向量乘法：在神经网络中，权重矩阵与前一层输出相乘的结果，再加上偏置向量，是一个线性函数。在卷积神经网络中，卷积操作显然是一个线性函数。

上采样：考虑 f 是对 $\boldsymbol{X} \in \mathbb{R}^2$ 进行上采样操作，则它可以表示为具有 $K(f) = 1, \boldsymbol{\psi}_1 = (0\ 0\ 0\ 0)^\top$ 的分段线性函数,

$$\Psi_1^f = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}^\top, \quad \mathcal{P}_1^f = \mathbb{R}^2.$$

Sigmoid 和双曲正切：如果有任何必须使用非分段线性激活函数的特定需求，我们可以对这些函数采用分段线性近似方法。

A.3 引理2的证明

在 f_1 处，给定一个实数值 z，输入为 $\boldsymbol{\beta}^{f_0} + \boldsymbol{\gamma}^{f_0} z = \boldsymbol{a}_{1:n} + \boldsymbol{b}_{1:n} z$。通过检查该输入的值，我们可以轻松得到多胞体

$$\{\Delta_{k_1}^{f_1}(\boldsymbol{\beta}^{f_0} + \boldsymbol{\gamma}^{f_0} z) \leq \boldsymbol{\delta}_{k_1}^{f_1}\}, \quad k_1 \in [K(f_1)],$$

即 $\boldsymbol{\beta}^{f_0} + \boldsymbol{\gamma}^{f_0}\ z$所属的多胞体。基于得到的多胞体，我们可以计算区间 $[L_{k_1}^{f_1}, U_{k_1}^{f_1}]$,

$$L_{k_1}^{f_1} = \max_{j:(\Delta_{k_1}^{f_1}\boldsymbol{\gamma}^{f_0})_j < 0} \frac{(\boldsymbol{\delta}_{k_1}^{f_1})_j - (\Delta_{k_1}^{f_1}\boldsymbol{\beta}^{f_0})_j}{(\Delta_{k_1}^{f_1}\boldsymbol{\gamma}^{f_0})_j} \quad \text{and} \quad U_{k_1}^{f_1} = \min_{j:(\Delta_{k_1}^{f_1}\boldsymbol{\gamma}^{f_0})_j > 0} \frac{(\boldsymbol{\delta}_{k_1}^{f_1})_j - (\Delta_{k_1}^{f_1}\boldsymbol{\beta}^{f_0})_j}{(\Delta_{k_1}^{f_1}\boldsymbol{\gamma}^{f_0})_j}.$$

此外，基于得到的多胞体，我们可以轻松得到 $\Psi_{k_1}^{f_1}$ 和 $\boldsymbol{\psi}_{k_1}^{f_1}, k_1 \in [K(f_1)]$。因此，第一层在 z 处的输出可定义为

$$\begin{aligned} f_1(z) &= \Psi_{k_1}^{f_1}(\boldsymbol{\beta}^{f_0} + \boldsymbol{\gamma}^{f_0} z) + \boldsymbol{\psi}_{k_1}^{f_1} \\ &= \boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z, \end{aligned}$$

其中 $\boldsymbol{\beta}^{f_1} = \Psi_{k_1}^{f_1}\boldsymbol{\beta}^{f_0} + \boldsymbol{\psi}_{k_1}^{f_1}$ 和 $\boldsymbol{\gamma}^{f_1} = \Psi_{k_1}^{f_1}\boldsymbol{\gamma}^{f_0}$。接下来，我们将 $\boldsymbol{\beta}^{f_1}, \boldsymbol{\gamma}^{f_1}$ 输入到 f_2中。

在 2nd 层，输入同样为 $\boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z$。通过检查该输入的值，我们可以轻松获得多胞体

$$\{\Delta_{k_2}^{f_2}(\boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z) \leq \boldsymbol{\delta}_{k_2}^{f_2}\}, \quad k_2 \in [K(f_2)],$$

所属的 $\boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z$。基于获得的多胞体，我们可以计算区间 $[L_{k_2}^{f_2}, U_{k_2}^{f_2}]$,

$$L_{k_2}^{f_2} = \max_{j:(\Delta_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1})_j<0} \frac{(\boldsymbol{\delta}_{k_2}^{f_2})_j - (\Delta_{k_2}^{f_2}\boldsymbol{\beta}^{f_1})_j}{(\Delta_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1})_j} \quad \text{and} \quad U_{k_2}^{f_2} = \min_{j:(\Delta_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1})_j>0} \frac{(\boldsymbol{\delta}_{k_2}^{f_2})_j - (\Delta_{k_2}^{f_2}\boldsymbol{\beta}^{f_1})_j}{(\Delta_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1})_j}.$$

此外，基于获得的多胞体，我们可以轻松得到 $\Psi_{k_2}^{f_2}$ 和 $\boldsymbol{\psi}_{k_2}^{f_2}$, $k_2 \in [K(f_2)]$。因此，第一层在 z 处的输出可以定义为

$$\begin{aligned} f_2(z) &= \Psi_{k_2}^{f_2}(\boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z) + \boldsymbol{\psi}_{k_2}^{f_2} \\ &= \boldsymbol{\beta}^{f_2} + \boldsymbol{\gamma}^{f_2} z, \end{aligned}$$

其中 $\boldsymbol{\beta}^{f_2} = \Psi_{k_2}^{f_2}\boldsymbol{\beta}^{f_1} + \boldsymbol{\psi}_{k_2}^{f_2}$ 和 $\boldsymbol{\gamma}^{f_2} = \Psi_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1}$。

A.4 实验细节。

比较方法。 我们将所提方法与以下方法进行了比较：

- 朴素方法：使用经典的 z 检验计算朴素的 p 值。
- Bonferroni校正：考虑所有可能假设的数量，以校正选择偏差。 p 值通过 $p_{\text{Bonferroni校正}} = \min(1, p_{\text{朴素方法}} * 2^n)$ 计算。
- 过度条件化（OC）：进一步根据所有节点观测到的激活状态和非激活状态进行条件化。该方法的局限在于，过度条件化会导致其统计功效较低。

网络结构。 在所有实验中，我们都使用图 8 所示的网络结构。

脑部图像数据集上的实验设置。 我们研究了从 Buda等人所用数据集中提取的脑部图像数据集 [2019], , 其中包含 941 张无肿瘤图像（C1）和 939 张有肿瘤图像（C2）。我们从 C1 中选取了 50 张图像作为参考图像。我们使用 C1 中的 841 张图像和 C2 中的 889 张图像进行深度神经网络训练。C1 和 C2 中剩余的图像用于展示所提出的选择性 p 值的优势。

更多结果　脑部图像数据集上的结果。更多结果见图 11、12、9 和 10

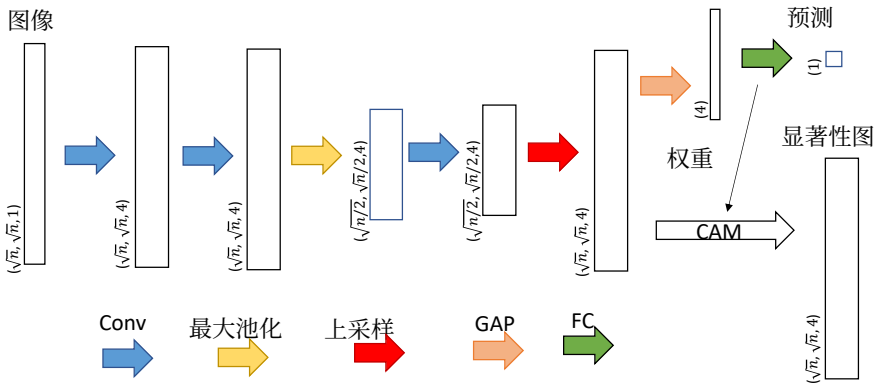
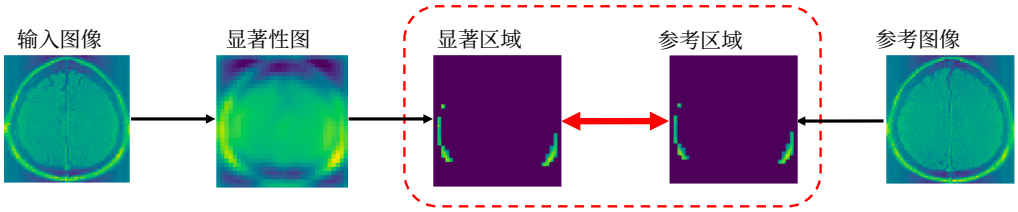
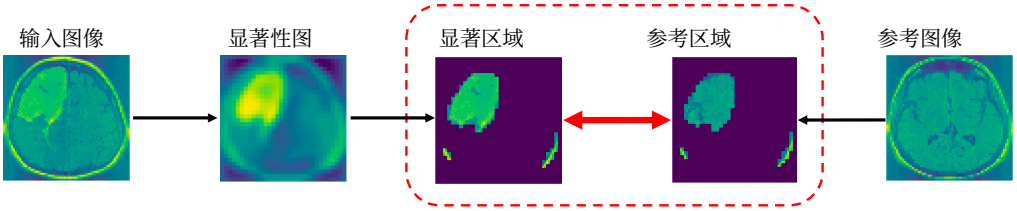


图8：网络结构。



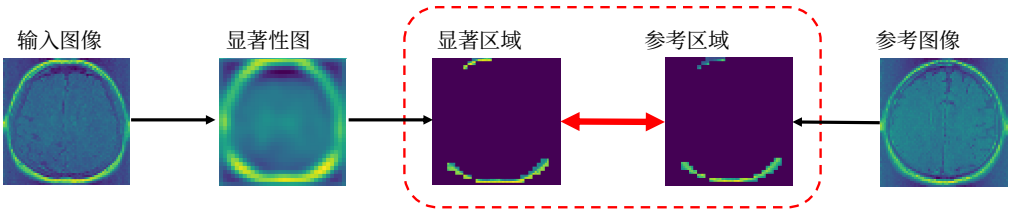
(a) $p_{\text{naive}} = \mathbf{0.01}$, $p_{\text{selective}} = \mathbf{0.47}$

图9：对无肿瘤图像中的显著区域进行推断（均值零假设检验）。



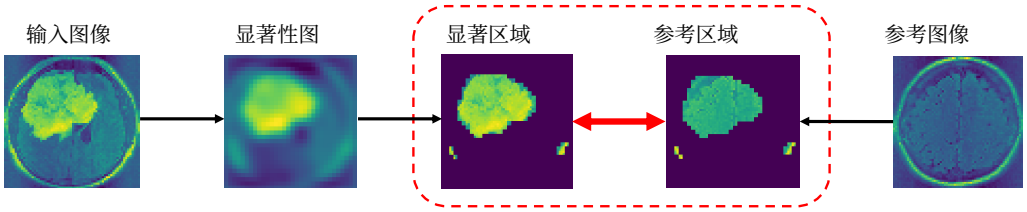
(a) $p_{\text{naive}} = \mathbf{0.00}$, $p_{\text{selective}} = \mathbf{2.82 \times 10^{-4}}$

图10：对存在肿瘤的图像中的显著区域进行推断（均值零假设检验）。



(a) $p_{\text{naive}} = \mathbf{3.00 \times 10^{-4}}$, $p_{\text{selective}} = \mathbf{0.29}$

图11：无肿瘤图像中显著区域的推断（全局零假设检验）。



(a) $p_{\text{naive}} = \mathbf{0.00}$, $p_{\text{selective}} = \mathbf{2.66 \times 10^{-20}}$

图12：存在肿瘤的图像中显著区域的推断（全局零假设检验）。