论文阅读报告:
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

刘昭阳 25215133
中山大学数学学院 (珠海)

2025 年 10 月 15 日

摘要

摘要：深度卷积神经网络（CNN）本质上是一个将高维输入流形映射到低维类别概率单纯形的复杂非线性算子。尽管其在计算机视觉任务中表现出卓越的泛化能力，但其决策曲面的局部几何性质往往难以解析，导致了所谓的“黑盒”问题。本文基于 Ramprasad Selvaraju 等人的论文 *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*，从微积分和线性代数的视角深入剖析了 Grad-CAM 算法。本文首先通过泰勒级数展开论证了梯度作为特征重要性度量的理论基础；其次，推导了 Grad-CAM 作为 CAM 一般化形式的数学证明；进一步，引入哈达玛积（构建了高分辨率的 Guided Grad-CAM；最后，探讨了该方法在图像描述和视觉问答等广义映射任务中的泛化应用及忠实度评估。

关键词：可解释性；梯度分析；泰勒展开；哈达玛积；流形学习

目录

1 引言：非线性映射的可视化挑战 4

1.1 问题的数学表述 4

1.2 现有方法的局限性 4

2 Grad-CAM 的数学原理 4

2.1 梯度的全局平均与特征权重 4

2.2 Grad-CAM 是 CAM 的推广 5

2.3 热力图生成与 ReLU 约束 5

3 高分辨率可视化：Guided Grad-CAM 5

3.1 融合策略 6

4 泛化应用：超越分类任务 6

4.1 图像描述 7

4.2 视觉问答 7

5 实验评估与性质分析 7

5.1 弱监督定位 7

5.2 忠实度验证 7

5.3 反事实解释 8

6 个人认为启发与衍生想法 8

6.1 梯度视角对深度学习的启示 8

6.2 可解释性与泛化性的平衡 8

6.3 渐进式精细化的设计哲学 8

6.4 泛化边界的拓展 9

6.5 解释的代价与局限 9

6.6 未来研究方向的思考 9

6.7 个人感悟 9

7 结论 **10**

1 引言：非线性映射的可视化挑战

1.1 问题的数学表述

设深度神经网络为一个复合函数 $F : \Omega \to \mathbb{R}^K$，其中 $\Omega \subset \mathbb{R}^{H\times W\times 3}$ 为输入图像空间，K 为类别数。对于输入张量 $\mathbf{x} \in \Omega$，网络输出得分类别向量 $\mathbf{y} = F(\mathbf{x})$。我们的目标是寻找一个解释算子 E，对于特定的目标类别 c，生成一个显著性映射（Saliency Map）$L^c \in \mathbb{R}^{H\times W}$，使得 $L^c(i,j)$ 能够反映输入空间位置 (i,j) 对输出标量 y^c 的一阶敏感度或贡献度。

1.2 现有方法的局限性

早期的反向传播方法（如 Saliency Maps）计算 $\nabla_{\mathbf{x}} y^c$，即输出关于输入的梯度。虽然这在数学上精确描述了局部敏感度，但由于 ReLU 激活函数的非凸性和高频梯度的存在，导致生成的热力图在视觉上充满噪声。另一方面，CAM (Class Activation Mapping) 方法假设网络末端存在全局平均池化（GAP）层，即 $y^c = \sum_k w_k^c \left(\frac{1}{Z}\sum_{i,j} A_{ij}^k\right)$。这种线性约束限制了其在 VGG、ResNet 等一般化网络架构中的应用。

2 Grad-CAM 的数学原理

Grad-CAM 的核心思想是利用反向传播计算的梯度信息来近似特征通道的重要性权重，从而解除对特定网络结构的依赖。

2.1 梯度的全局平均与特征权重

考虑网络中的最后一个卷积层，其输出特征图为 $A \in \mathbb{R}^{u\times v\times K'}$，其中 A^k 表示第 k 个通道的特征矩阵。根据多元微积分的链式法则，我们计算目标分数 y^c 相对于特征图 A^k 的梯度 $\frac{\partial y^c}{\partial A^k}$。为了获得该通道的全局重要性标量 α_k^c，我们对梯度矩阵进行全局平均池化：

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

其中 $Z = u \times v$。**数学直觉：**这一步利用了一阶泰勒展开的系数。如果我们将 y^c 视为 A^k 的函数，那么 α_k^c 近似了 y^c 在 A^k 方向上的平均变化率。

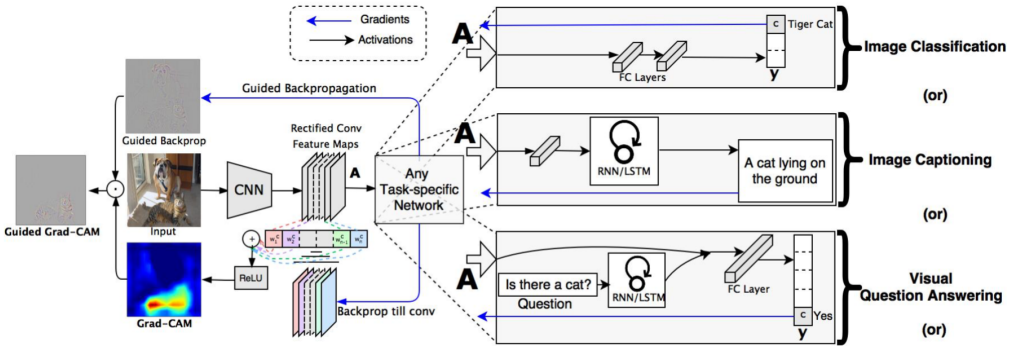

图 1: Grad-CAM 算法架构图。通过反向传播计算梯度，经 GAP 得到权重，再与特征图加权组合，最终经 ReLU 生成热力图。

2.2 Grad-CAM 是 CAM 的推广

我们可以证明，当网络结构满足 CAM 的假设（GAP + FC）时，Grad-CAM 计算出的权重 α_k^c 与 CAM 的训练权重 w_k^c 是等价的（相差一个归一化常数）。设 $F_k = \frac{1}{Z}\sum_{i,j} A_{ij}^k$，且 $y^c = \sum_k w_k^c F_k$。则：

$$\frac{\partial y^c}{\partial A_{ij}^k} = \frac{\partial y^c}{\partial F_k}\frac{\partial F_k}{\partial A_{ij}^k} = w_k^c \cdot \frac{1}{Z} \tag{2}$$

代入 α_k^c 的定义：

$$\alpha_k^c = \sum_{i,j} \frac{1}{Z}\left(\frac{w_k^c}{Z}\right) = \frac{w_k^c}{Z} \tag{3}$$

这证明了 Grad-CAM 是 CAM 的严格数学推广，适用于任意可微的 CNN 架构。

2.3 热力图生成与 ReLU 约束

利用计算出的权重 α_k^c，我们对特征图进行加权线性组合，并应用 ReLU 激活函数：

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{4}$$

ReLU 的必要性：我们只关注对类别 c 有**正向贡献**的特征。若 $\sum \alpha_k^c A_{ij}^k < 0$，说明该区域的激活会降低目标类别的置信度，这在定位任务中属于干扰项，应当被滤除。

3 高分辨率可视化：Guided Grad-CAM

Grad-CAM 生成的热力图 $L_{\text{Grad-CAM}}^c$ 分辨率较低（通常为 14×14 或 7×7），为了获得像素级的精细可视化，论文提出了 Guided Grad-CAM。

3.1 融合策略

设 Guided Backpropagation 生成的梯度图为 $M_{\text{Guided}} \in \mathbb{R}^{H \times W}$。该方法通过修改反向传播规则（仅传递正梯度和正激活值），能够捕捉高频细节，但缺乏类别判别性。我们将 Grad-CAM 热力图通过双线性插值（Bilinear Interpolation）上采样至 $H \times W$，记为 L^c_{up}。Guided Grad-CAM 定义为二者的哈达玛积（Hadamard Product）：

$$L^c_{\text{Guided Grad-CAM}} = M_{\text{Guided}} \odot L^c_{\text{up}} \tag{5}$$

这种逐元素乘法操作在数学上相当于施加了一个空间掩码（Spatial Mask），利用 Grad-CAM 的类别定位能力过滤掉了 Guided Backprop 中的背景噪声，保留了与目标类别相关的纹理细节。

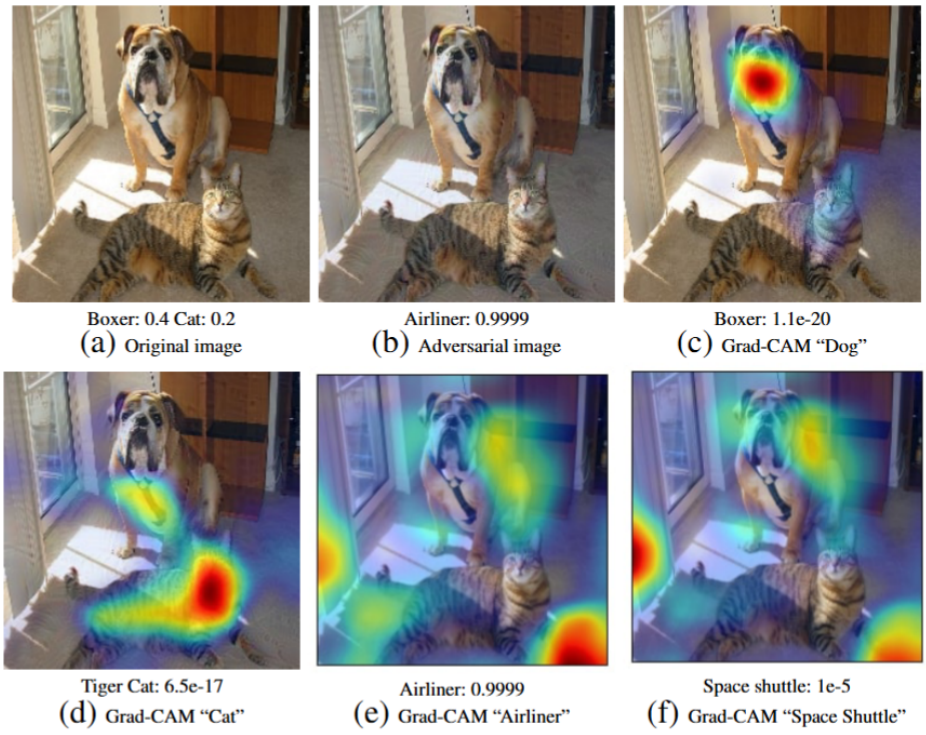

图 2: Grad-CAM 可视化示例。展示了原图、对抗样本以及针对不同类别（如狗、猫、飞机等）生成的 Grad-CAM 热力图，显示了模型对不同对象的定位能力。

4 泛化应用：超越分类任务

Grad-CAM 的数学推导仅依赖于标量 y^c 的存在，而不依赖于 y^c 的具体物理意义。因此，它可以无缝扩展到其他视觉任务。

4.1 图像描述

在图像描述任务中，模型输出是一个词序列。对于生成的第 t 个词 w_t，我们可以定义目标标量为该词的对数概率：$y^{w_t} = \log P(w_t|\mathbf{x}, w_0, \ldots, w_{t-1})$。计算 $\nabla_A y^{w_t}$ 并应用 Grad-CAM 算法，即可得到模型在生成该词时关注的图像区域。这验证了模型在序列生成过程中的空间注意力机制。

4.2 视觉问答

对于 VQA 任务，输入为图像 $\mathbf{x}$ 和问题 $\mathbf{q}$，输出为答案 a 的分数 y^a。同样地，通过计算 $\nabla_A y^a$，Grad-CAM 能够揭示模型是如何根据不同的问题 $\mathbf{q}$ 调整其在图像 $\mathbf{x}$ 上的关注区域。实验表明，即使对于同一张图像，不同的问题会导致完全不同的梯度分布，从而生成截然不同的热力图。

5 实验评估与性质分析

5.1 弱监督定位

为了量化 Grad-CAM 的定位能力，论文在 ILSVRC-15 数据集上进行了实验。通过对热力图 L^c 设定阈值并提取最大连通域的边界框，Grad-CAM 在 VGG-16 上实现了 56.51% 的 Top-1 定位误差，优于传统的反向传播方法。这表明梯度加权的特征图能够有效地近似物体的空间分布。

5.2 忠实度验证

如何证明 Grad-CAM 解释的正确性？论文设计了遮挡实验。定义遮挡函数 $\Phi(\mathbf{x}, L^c)$，将图像中对应 L^c 高响应的区域像素置零。计算分数下降率：

$$\Delta y = y^c(\mathbf{x}) - y^c(\Phi(\mathbf{x}, L^c)) \tag{6}$$

实验结果显示，遮挡 Grad-CAM 识别的重要区域会导致模型分数的显著下降，这在统计上证明了 Grad-CAM 识别的区域确实是模型决策的关键支撑集（Support Set）。

5.3 反事实解释

通过计算负梯度的加权和：

$$L^c_{\text{Counterfactual}} = \text{ReLU}\left(\sum_k (-\alpha_k^c) A^k\right) \tag{7}$$

我们可以可视化那些“阻止”模型预测为类别 c 的区域。这在数学上对应于寻找梯度下降最快的方向，即改变哪些区域可以最大程度地提升类别 c 的概率（如果当前概率较低）。

6 个人认为启发与衍生想法

6.1 梯度视角对深度学习的启示

Grad-CAM 的核心贡献在于将**梯度**这一微积分工具从优化领域重新定义为**解释工具**。这给了我一个重要的启发：不仅要使用梯度来指导参数更新，更要用梯度来理解模型的内部决策逻辑。在深度学习中，梯度不再仅仅是优化的导向，而是一扇观察神经网络” 思维过程” 的窗口。

6.2 可解释性与泛化性的平衡

传统的 CAM 方法因其依赖于特定的网络结构（全局平均池化），面临着泛化性不足的问题。Grad-CAM 通过充分利用反向传播的通用性，突破了这一限制。这启发我，**数学的约束有时会化为设计的束缚**——放宽某些假设（如不再假设 GAP 层的存在），反而能得到更具普适性的算法。这在其他领域也有启示：有时候，放弃某些” 便利的假设”，用更基础、更通用的原理重新推导，往往能得到更强大的方法。

6.3 渐进式精细化的设计哲学

Grad-CAM → Guided Grad-CAM 的进化过程体现了一种**逐层精细化**的设计思路。从低分辨率但具有语义判别性的热力图，到通过哈达玛积融合高分辨率的梯度细节，这种分层的组合策略值得借鉴。在其他计算机视觉和机器学习问题中，我们也许可以类似地将” 粗粒度的判别性特征” 与” 细粒度的纹理信息” 进行融合。

6.4 泛化边界的拓展

Grad-CAM 从图像分类扩展到图像描述、视觉问答的能力，本质上源于其**对于" 目标标量"** 的最小化依赖。这提示我们，很多在特定任务上设计的方法，其实可以通过识别其数学结构中的" 本质" 部分，推广到看似完全不同的应用场景。例如，对于任何输出是连续标量的神经网络任务，Grad-CAM 似乎都可能适用。

6.5 解释的代价与局限

虽然 Grad-CAM 提供了直观的可视化，但梯度基的可解释性方法本身有**一阶近似**的局限——它只捕捉了目标函数在特定点的局部敏感度。对于非凸、多模态的损失曲面，梯度可能被局部的陡峭梯度所主导，而忽略全局的结构信息。换句话说，Grad-CAM 可以回答" 对哪些像素敏感"，但难以回答" 为什么选择这个决策而不是另一个决策"，这是根本上的信息不完整性。

6.6 未来研究方向的思考

(1) **高阶导数的利用**：二阶或更高阶导数（如海森矩阵）是否能捕捉更多的决策逻辑？

(2) **对抗鲁棒性**：对抗样本会导致 Grad-CAM 生成的热力图发生剧烈变化。这是否反映了模型对高频纹理的过度依赖，还是表明梯度本身在对抗环境下的不稳定性？

(3) **全局 vs 局部**：Grad-CAM 强调全局平均池化来得到权重，但像素级的梯度-特征图交互是否存在更复杂的空间相关性结构，用简单的加权求和无法充分表达？

(4) **因果解释**：梯度表示关联性（Correlation），如何过渡到因果性（Causality）的解释？例如，通过因果干预（Causal Intervention）来验证 Grad-CAM 识别的区域是否真正" 因果地" 影响了决策？

(5) **多模态模型的扩展**：对于 Vision-Language 模型（如 CLIP），Grad-CAM 如何在跨模态的特征空间中进行解释？

6.7 个人感悟

通过深入学习 Grad-CAM，我意识到**可解释性本质上是人与机器之间的一场" 沟通"**。好的解释不仅要数学上严谨，更要在直观性和完整性之间找到平衡。Grad-CAM 的成功之处在于它用最小化的数学工具（梯度 + 加权求和），通过充分利用反向传播的已有基础设施，生成了视觉上容易理解的结果。这提醒我，在追求复杂的理论时，有时候最优雅的解决方案往往来自于对已有工具的深刻理解和巧妙重组。

7 结论

本文从数学分析的角度重构了 Grad-CAM 算法。作为一种基于梯度的后处理方法，Grad-CAM 巧妙地利用了深度网络的微分性质，将高层的语义信息（梯度）与底层的空间信息（特征图）相结合。它不仅提供了一种通用的可视化工具，更为理解深度神经网络这一复杂非线性系统的内部运作机制提供了坚实的数学窗口。