脑肿瘤分类中的深度学习架构对比分析：ResNet 与 Vision Transformer

周炜杰

20231610

GitHub 仓库地址：https://github.com/Double-Z-wj/Machine_Learning_Task
数据集地址： https://www.kaggle.com/datasets/masoudnickparvar/brain-tumor-mri-dataset/data

1 摘要

脑部肿瘤的精准与快速诊断对于制定治疗方案至关重要。随着深度学习技术的发展，卷积神经网络（CNN）已成为医学影像分析的主流方法，而新兴的视觉 Transformer（ViT）架构凭借全局注意力机制发起了挑战。为了明确两种架构在小样本医疗影像场景下的适用性，本项目在脑部肿瘤 MRI 数据集（Brain Tumor MRI Dataset）上，对比评估了 ResNet-18 与 ViT-Tiny 的性能表现。

实验结果表明，尽管 ViT-Tiny 的参数量仅为 ResNet-18 的一半，但在准确率、推理速度与安全性上，CNN 架构均占据显著优势。ResNet-18 取得了 99.69% 的最高准确率，优于 ViT-Tiny 的 99.01%。在推理效率方面，得益于卷积操作的硬件优化，ResNet 的单样本推理时间仅为 1.95ms，比 ViT 的 4.50ms 快了约 2.3 倍。

更重要的，错误模式分析发现 ViT 的误判样本构成了 ResNet 误判样本的超集，且 ViT 出现了更多高风险的漏诊。本研究证实：在缺乏海量预训练数据的医疗诊断任务中，具备归纳偏置的 CNN 依然是比 Transformer 更稳健、更快速且更安全的临床选择。

2 任务与方法

2.1 任务重述

本实验聚焦脑肿瘤 MRI 图像多分类任务，核心目标是基于脑部 MRI 影像分类系统，并探讨不同深度学习架构在医学影像任务上的效率与行为差异。任务本质为全监督的图像分类任务。实验的目的是探讨在少样本，低算力的情况下，轻量的 ResNet 与 ViT 的性能异同。

2.2 使用的数据集与数据处理

实验采用 Kaggle 公开数据集 Brain Tumor MRI Dataset（Masoud Nickparvar, 2021），该数据集为临床真实脑部 MRI 影像，无人工合成样本。该数据集包含胶质瘤 (Glioma)、脑膜瘤 (Meningioma)、垂体瘤 (Pituitary) 以及无肿瘤 (No Tumor) 四种典型的脑部生理状态类别，涵盖了多种切面角度（如轴状面、冠状面），具有较高的临床代表性。

本项目建立了一套标准化的数据预处理流水线。首先，将所有不同分辨率的原始 MRI 图像统一调整为 224*224 像素，以适配标准模型的输入维度。随后，

对图像进行张量化转换，并在 RGB 通道上执行归一化处理。为了增强模型的泛化能力并抑制过拟合现象，在训练阶段引入了数据增强策略，包括随机水平翻转和随机旋转，模拟真实的影像拍摄差异，强迫模型学习具有旋转不变性的病理特征。

2.3 模型架构设计

选取了 ResNet 和 ViT 模型进行对比。其中，ResNet 使用 ResNet-18，加载 ImageNet 预训练权重，替换原 fc 层（1000 类）为 4 分类线性层。ResNet 为传统的卷积神经网络，具备局部性和平移不变性的归纳偏置。ViT 使用 ViT-Tiny。ViT-Tiny 是轻量化 Vision Transformer，使用自注意力机制建模全局关系。替换原 head 层（1000 类）为 4 分类线性层。但 ViT 缺乏 CNN 的空间归纳偏置，在小规模数据集上不一定更加有效。

2.4 训练超参数设置

两类模型使用完全一致的训练配置，消除超参数差异带来的影响。两个模型均加载了在 ImageNet 数据集上预训练的权重，仅对全连接分类层进行重置和微调。训练过程中，采用交叉熵损失函数作为目标函数，优化器选用 Adam 算法，初始学习率设定为1×10^{-4}，以确保在预训练权重的基础上进行平稳的参数更新。模型训练共进行 10 个 Epoch，并在每个 Epoch 结束后在独立的测试集上进行验证，保存验证准确率最高的模型权重用于最终评估。

3 实验结果分析

为了全面评估模型的综合性能，本项目摒弃了单一的准确率指标，构建了涵盖性能、效率与可解释性的三维评价体系：

分类性能指标：除了总体准确率 (Accuracy) 外，通过计算混淆矩阵 (Confusion Matrix) 以及各类别的精确率 (Precision)、召回率 (Recall) 和 F1-Score，详细评估模型在特定肿瘤亚型上的诊断能力，重点考察是否存在严重的误诊（即无肿瘤类别的假阳性）、漏诊（有肿瘤类别的假阴性）情况。

计算效率指标：记录模型的参数量以评估存储需求，并评估单样本在 GPU 上的推理时间，以分析模型在临床实时诊断场景下的响应速度。

可解释性分析：引入 Grad-CAM 技术，通过回传梯度生成类激活热力图，可视化模型关注的图像区域，用于验证模型是真正依据病灶区域进行判断，还是错误地依赖了背景噪声。

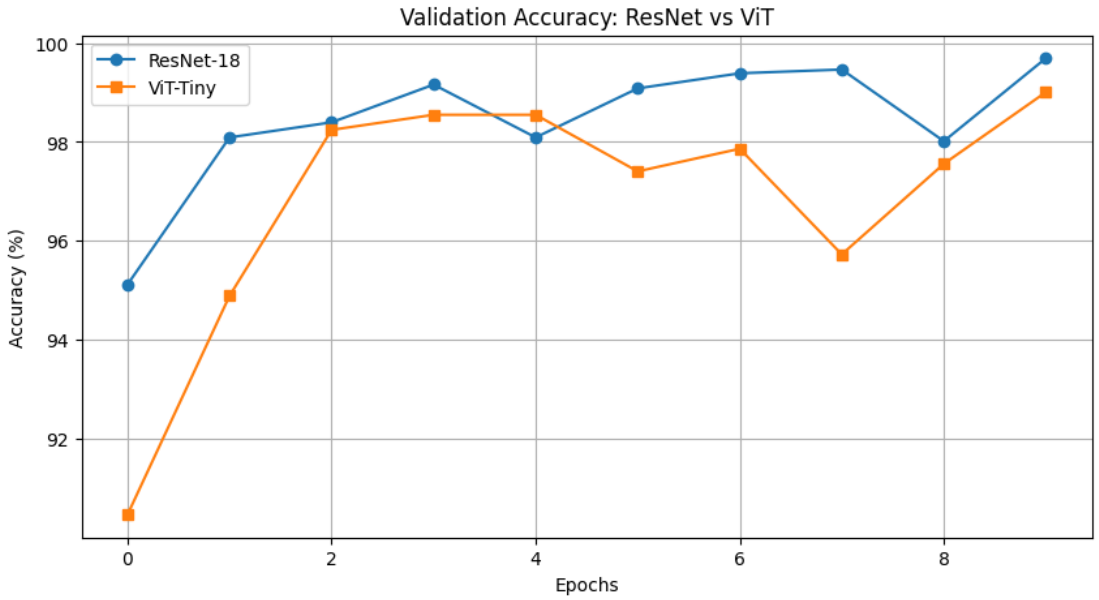
3.1 训练动态对比

验证集准确率如下图所示。ResNet-18 展现了极高的训练稳定性，其验证准确率曲线平滑上升，在早期的 Epoch 即达到 98% 以上的高位，并最终稳定收敛于 99.69% 的峰值。这种的收敛特性主要得益于卷积神经网络固有的归纳偏置，即平移不变性和局部相关性，这使得模型能够利用预训练权重快速适应小样本的医疗影像数据。

相比之下，ViT-Tiny 的训练曲线表现出明显的震荡特性。这印证了 ViT 对数据规模和训练超参数的敏感性。由于缺乏卷积神经网络的先验知识，ViT 需要更多的迭代来通过自注意力机制建立像素间的全局依赖关系，导致其在小样本数据集上的训练稳定性不如 CNN。尽管 ViT 最终也能达到 99.01% 的高准确率，

但从收敛效率和鲁棒性角度来看，ResNet 在本任务中占据明显优势。

总体而言，ResNet-18 以 99.69% 的整体准确率优于 ViT-Tiny 的 99.01%。



3.2 分类性能深度评估

针对“无肿瘤”类别，两者的召回率均达到了 100%，这意味着在测试集的所有健康样本中，没有一例被误诊为肿瘤患者，实现了零假阳性。然而，在肿瘤亚型的区分能力上，ResNet 展现了更强的判别力。具体而言，对于 Glioma 类别，ResNet 实现了 100% 的精确率，而 ViT 的精确率仅为 98%。混淆矩阵显示，ViT 将部分胶质瘤样本误判为脑膜瘤或无肿瘤，这表明在处理纹理特征相似的病灶时，CNN 提取的深层纹理特征比 ViT 的全局特征更具区分度，能够更准确地界定肿瘤的边缘与类别。

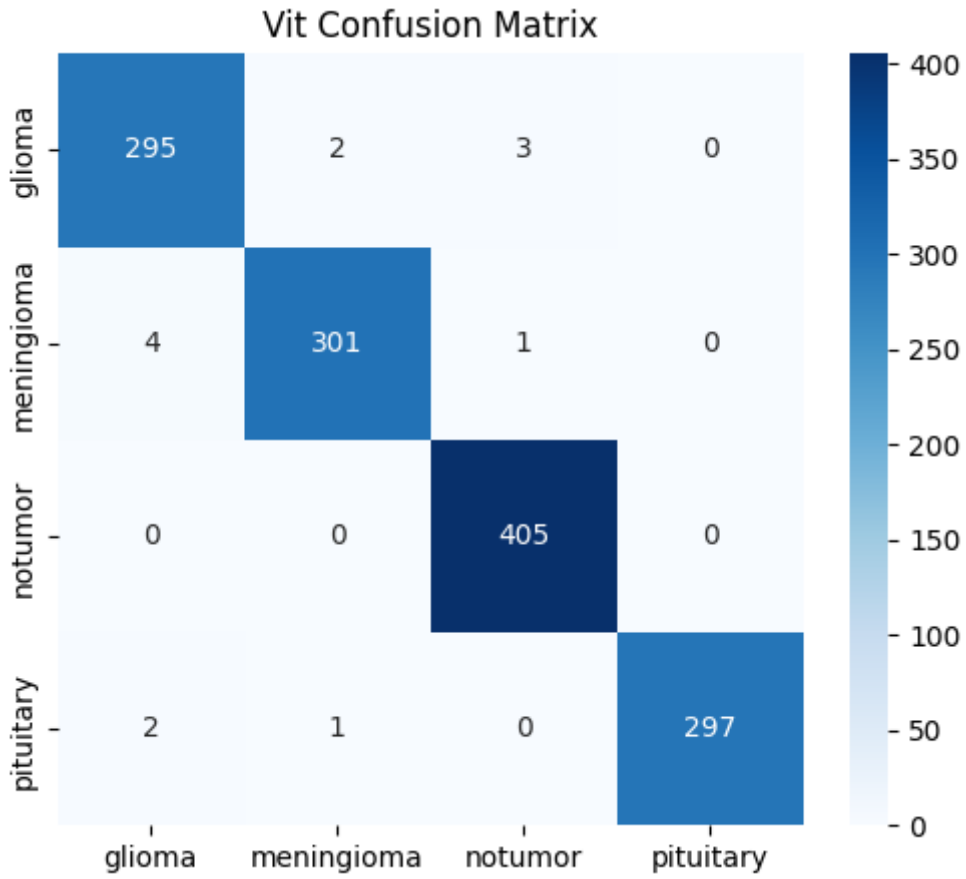
以下是 ResNet-18 分类性能表。

	precision	recall	f1-score	support
glioma	1.000000	0.993333	0.996656	300.000000
meningioma	0.990260	0.996732	0.993485	306.000000
notumor	0.997537	1.000000	0.998767	405.000000
pituitary	1.000000	0.996667	0.998331	300.000000
accuracy	0.996949	0.996949	0.996949	0.996949
macro avg	0.996949	0.996683	0.996810	1311.000000
weighted avg	0.996966	0.996949	0.996951	1311.000000

以下是 ViT-Tiny 分类性能表

	precision	recall	f1-score	support
glioma	0.980066	0.983333	0.981697	300.000000
meningioma	0.990132	0.983660	0.986885	306.000000
notumor	0.990220	1.000000	0.995086	405.000000
pituitary	1.000000	0.990000	0.994975	300.000000
accuracy	0.990084	0.990084	0.990084	0.990084
macro avg	0.990105	0.989248	0.989661	1311.000000
weighted avg	0.990114	0.990084	0.990083	1311.000000

以下是混淆矩阵。

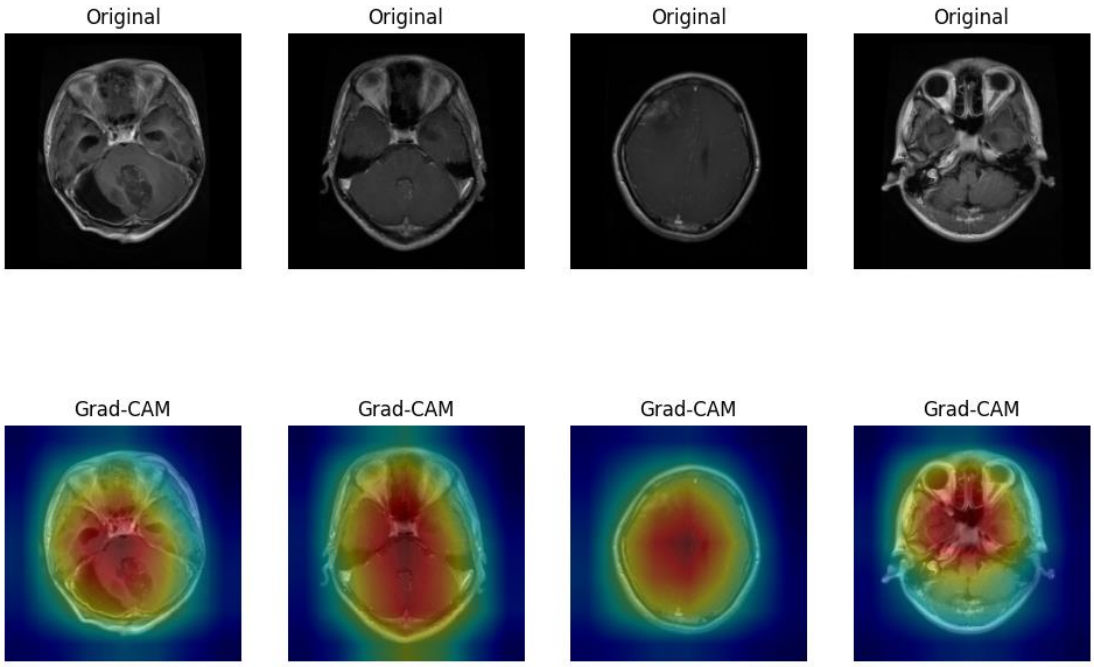


3.3 推理速度分析

从存储空间来看，轻量化的 ViT-Tiny 具有显著优势，其参数量仅为 5.53M，不到 ResNet-18 (11.18M) 的一半。然而，在推理速度层面，ViT-Tiny 却比 ResNet-18 慢。在推理速度的测试实验中，ResNet-18 的单样本推理时间仅为 1.95ms，而参数量更小的 ViT-Tiny 却需要 4.50ms。我们尝试给出一种解释：CNN 的卷积操作在 GPU 上拥有高度成熟的并行计算库（如 cuDNN）支持，且内存访问模式规整。而 Transformer 的自注意力机制涉及复杂的矩阵运算与内存交互，在处理特征图时计算开销较大。因此，ResNet-18 提供了更优的速度。

3.4 可解释性验证

为了验证模型高准确率的来源，本研究利用 Grad-CAM 技术对 ResNet-18 的决策过程进行了可视化。生成的类激活热力图清晰地显示，模型的高响应区域覆盖了脑部的肿瘤实质区域，且能够适应不同位置、不同大小的病灶。热力图并未在颅骨边缘、眼球或图像背景噪声处产生高激活。这种可视化结果不仅在技术层面验证了模型确实学习到了具有病理意义的形态学特征，更为该系统进入临床辅助诊断流程提供了必要的可解释性信任基础，使得医生可以直观地理解 AI 的判读依据。



4 讨论

4.1 为什么 ResNet 在医疗小数据集上表现得更好

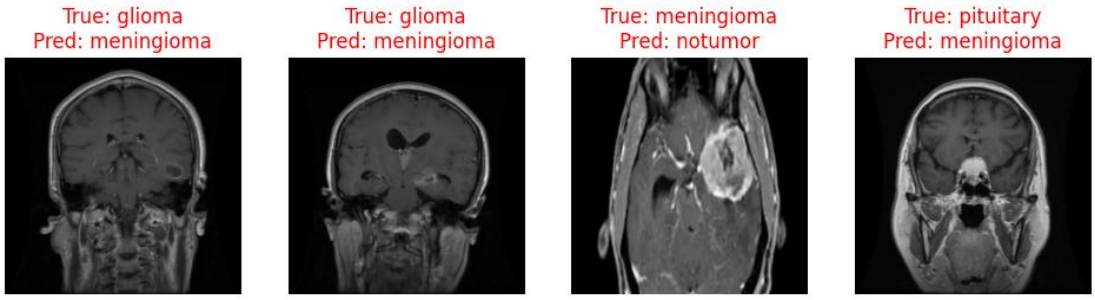
本研究的定量实验结果表明，在小样本医疗影像数据集上，基于卷积神经网络的 ResNet-18 架构在准确率和推理速度上均优于基于 Transformer 的 ViT-Tiny 架构。这一结果有力的证明了在小数据集上归纳偏置的作用。

CNN 固有的平移不变性和局部相关性假设，通过识别局部的异常纹理和边缘来进行诊断。这种特性使得 ResNet 能够在数据量有限的情况下快速收敛并提取鲁棒特征。相比之下，ViT 缺乏这种先验假设，它依赖于通过海量数据训练来学习。在本实验约 7000 张的训练规模下，ViT 难以构建出足以超越 CNN 的分类器。此外，ViT 自注意力机制的计算在缺乏硬件优化的前提下，导致了推理延迟的增加，进一步削弱了其竞争力。

4.2 错误分析

我们提取了 ResNet-18（共 4 例错误）和 ViT-Tiny（其中 5 例错误）的误判样本进行逐一视检与分析。

针对 ResNet-18 的 4 张错误样本分析发现，其主要误判集中在冠状面图像上。如图所示，前两张样本均为冠状面扫描的胶质瘤（Glioma），却被误判为脑膜瘤(Meningioma)。此外,ResNet 出现了一例将脑膜瘤误判为无肿瘤(Notumor)的“假阴性”案例，该图像对比度极低且伪影严重，导致卷积核未能有效提取出病灶特征，这提示 CNN 对图像质量的鲁棒性仍有待提升。



在分析 ViT-Tiny 的错误样本时，我们观察到：ResNet-18 犯过的典型错误，ViT-Tiny 全部重犯了一遍。这表明，这些样本属于数据集中的固有“困难样本”，即无论采用局部卷积还是全局注意力架构，其视觉特征本身就存在高度的模棱两可性。这意味着，在本任务中，引入 Transformer 架构并未能通过其全局视野解决 CNN 的固有短板，二者在困难样本上未能形成优势互补。ViT-Tiny 不仅未能修正 ResNet 的错误，反而引入了更多的错误和误判。

综上所述，ResNet-18 在本任务中表现出了对 ViT-Tiny 不仅在整体准确率上更高，而且在困难样本的鲁棒性上严格优于 ViT。ViT 的注意力机制非但未能提供额外的纠错能力，反而增加了漏诊风险。

5 不足与展望

尽管本实验在定量指标上得出 ResNet 优于 ViT 的结论，但必须承认，该结论是在特定的实验约束下得出的，存在以下局限性：

数据集过于简单：本实验使用的 Brain Tumor MRI 数据集过于简单。分类任务难度较低，ResNet-18 和 ViT-Tiny 均轻松达到了 99% 以上的准确率。在这种简单数据集上，深度学习模型的性能差异容易被掩盖。若应用在真实的临床场景——使用严重类别不平衡或背景复杂的原始数据时，缺乏归纳偏置的 ViT 可能会因为其更强的全局建模能力而表现出不同的鲁棒性，或者 ResNet 的稳定性优势可能会进一步扩大。

仅对比轻量模型：受限于算力和任务简单，本研究仅对比了轻量级模型（ResNet-18 vs ViT-Tiny）。ViT 的优势通常在参数量巨大时才能通过 Scaling Law 体现出来。ViT-Tiny 可能因容量不足无法充分拟合特征。

针对上述不足，未来的研究可以从以下维度深入：

使用更复杂的数据集：在更具挑战性的数据集，更有挑战性的任务上进行验证。在处理复杂数据和任务时，Transformer 的全局注意力机制可能展现出比 CNN 更强的优势。

扩展模型规模：对比 ResNet-50/101 与 ViT-Base/Large 在长周期训练下的表现。

引入新的模型对比： Mamba 架构兼具 CNN 的线性计算复杂度和 Transformer

的全局上下文建模能力。未来的工作应引入 Mamba 与 ResNet、ViT 进行横向对比。

6 结论

本项目通过在脑肿瘤 MRI 图像上的实验，系统的探讨了卷积神经网络与视觉 Transformer 的性能差异。通过定量指标与定性分析，得出以下结论：

在准确率指标上，ResNet-18 达到了 99.69%，不仅在数值上超越了 ViT-Tiny，且训练收敛曲线更为平滑稳定。这证明了 CNN 固有的平移不变性和局部感知特性（归纳偏置）天然地契合医学影像中纹理特征显著、样本规模有限的特点。相比之下，ViT 缺乏归纳偏置，在小数据上难以构建出超越 CNN 的特征提取能力。

尽管 ViT-Tiny 在模型体积上具有轻量化优势，但在实际 GPU 推理速度上，ResNet-18 反更快。底层硬件优化成熟的 CNN 架构，在低算力情境下是更好的选择。

通过分析错误判例发现，ViT 的错误样本不仅完全包含了 ResNet 的错误样本，还额外产生了性质更为严重的漏诊错误。从临床安全性的角度评估，ResNet-18 更加可靠。

通过 Grad-CAM 可视化分析，进一步佐证了 ResNet 的优越性。模型的高响应区域覆盖了病灶区域，证明了 ResNet 的高准确率源于对病理特征的真实捕捉。

综上所述，尽管 Vision Transformer 代表了深度学习的更前沿趋势，但在本项目的脑肿瘤诊断任务中，经典的 ResNet-18 架构在精度、速度与安全性三个维度上实现了对 ViT-Tiny 的全面超越。在很多医学影像处理的情境下，卷积神经网络仍有不可替代的地位。对于此类特定领域的应用，卷积神经网络依然是更好的的工程解决方案。