

CapsFormer: 融合 CapsNet 与 Transformer 的医学影像几何感知分类模型

郭万玲¹ 刘海川¹

GUO Wanling LIU Haichuan

摘要

医学影像分类在临床诊断中发挥着重要作用,可以显著提升疾病的诊断效率与准确性。然而,传统方法和现有深度学习技术在处理医学影像的复杂性和几何特征时仍存在局限。基于此,文章提出了一种 CapsNet 几何感知能力与 Transformer 全局上下文建模能力的 CapsFormer 架构,以提高医学影像分类的准确性。实验结果表明, CapsFormer 在分类准确率、 F_1 分数和 AUC-ROC 等关键指标上表现较好。消融实验进一步表明, CapsNet 模块精准建模几何特征, Transformer 模块有效整合全局信息,二者协同作用提升了模型的鲁棒性和分类性能。CapsFormer 不仅为医学影像分类提供有效解决方案,还展现出在复杂临床场景中的广阔应用前景,推动智能医疗诊断的发展。

关键词

CapsNet; Transformer; 医学影像分析; 深度学习; 图像识别

doi: 10.3969/j.issn.1672-9528.2025.12.011

0 引言

医学影像分类是医学图像处理领域的核心任务之一,自动分析医学影像能够辅助医生快速而准确地诊断疾病,提升临床决策的质量和效率。近年来,随着医学影像数据量的快速增长以及疾病诊断复杂性的增加,自动化医学影像分类技术在临床诊断中显得十分重要。然而,医学影像分类面临诸多挑战。这些挑战包括影像中病变区域的复杂性、病灶分布的多样性、图像噪声、模态差异以及标注数据不足等问题。这些因素对分类模型的特征提取能力和泛化能力提出了更高的要求。

本文提出的 CapsFormer 模型融合了 CapsNet 的局部几何感知能力和 Transformer 的全局上下文建模能力,以提高医学影像分类的准确性并增强对图像几何特征的感知能力。CapsFormer 通过 CapsNet 模块提取医学影像的局部几何特征, Transformer 模块捕捉这些特征之间的全局依赖关系,全连接层来进行分类预测。

本文的主要贡献如下:

(1) 设计并实现了 CapsFormer 架构,有效结合 CapsNet 的几何感知能力和 Transformer 的全局建模能力,为医学影像分类提供了新的解决方案。

(2) 探究不同学习率和批次对模型性能的影响,为医学影像分类任务提供超参数优化的参考。

(3) 在医学影像数据集中评估模型性能,证明了 CapsFormer 的优越性能。

(4) 通过消融实验来分析不同组件对分类性能的贡献,证明融合架构的优势。

1 相关工作

1.1 医学影像分类

医学影像分类主要是基于卷积神经网络(CNN)的传统方法,如 Krizhevsky 等^[1]提出的 AlexNet 通过多层卷积结构提升了图像分类的性能,He 等^[2]提出的 ResNet 通过残差连接解决了深层网络的训练问题,使其在肺结节检测、肺炎分类等任务中表现出色。然而,这些方法在处理类别不平衡和复杂几何特征时仍存在不足。

1.2 CapsNet

CapsNet 由 Sabour 等^[3]于 2017 年提出,通过动态路由机制来替代传统神经网络的池化操作,能够捕捉图像的空间层次关系与几何特征。CapsNet 在医学影像分类中已有应用,例如 Afshar 等^[4]利用 CapsNet 实现脑肿瘤类型的分类,成功地捕捉了肿瘤的形状和边界特征。在胸部 X 光片分类中, CapsNet 能够识别病灶的几何特性,如结节的大小和分布。CapsNet 的计算复杂度较高,且对全局特征的建模能力有限,限制了其在复杂任务中的表现。

1.3 Transformer 架构

Transformer 由 Vaswani 等^[5]提出,最初是应用于自然语

1. 重庆移通学院公共大数据安全技术重庆市重点实验室
重庆 401520

言处理领域，其核心是基于自注意力机制的全局建模能力。Dosovitskiy 等^[6]将 Transformer 引入到计算机视觉，提出了 Vision Transformer，通过将图像分块并建模全局依赖关系，提升了分类性能。Transformer 在医学影像分类中的应用仍处于探索阶段，且其对局部几何特征的感知能力较弱。

1.4 混合架构的探索

近年来，研究人员开始探索混合架构，结合不同类型的神经网络以发挥各自的优势。例如，CNN-Transformer 混合架构将 CNN 的局部特征提取能力与 Transformer 的全局建模能力结合，在计算机视觉任务中表现优异，但在医学影像领域的应用有限。

CapsNet 和 Transformer 的结合是一个有前景但尚未被充分探索的方向。CapsNet 的几何感知能力和 Transformer 的全局建模能力是互补的，结合后可以创建更加优异的医学影像分类系统。本文提出的 CapsFormer 正是基于这一观察，融合这两种架构的优势，为医学影像分类提供新的解决方案。

2 方法

2.1 CapsFormer 架构概述

CapsFormer 模型融合了 CapsNet 和 Transformer 架构，旨在同时捕捉医学影像的局部几何特征和全局上下文信息。主要包括 CapsNet 模块、Transformer 模块和全连接分类层。胶囊网络通过动态路由提取局部几何特征，而 Transformer 通过自注意力机制来捕捉全局依赖关系。这种结合使模型能够同时感知疾病特征的形状、位置关系和上下文信息，为医学影像分类发挥了重要的作用。

2.2 CapsNet 模块

CapsNet 模块主要包含 3 个组件：初始卷积层、Primary-Caps 层和 DigitCaps 层。初始卷积层使用 9×9 卷积核和步长 1，将 3 通道输入转换为 128 通道特征图。PrimaryCaps 层使用 9×9 卷积核和步长 2，生成 32 个 4 维胶囊。DigitCaps 层使用动态路由算法将 PrimaryCaps 的输出映射到 2 个 16 维类别胶囊。动态路由过程通过迭代来更新耦合系数。

2.3 Transformer 模块

Transformer 模块接收 CapsNet 的输出，通过线性投影层将 16 维胶囊向量映射到 128 维特征空间。再通过四层 Transformer 编码器处理这些特征，每层包含多头自注意力机制和前馈网络。实验设置了 8 个注意力头，维度为 128，以平衡计算效率和模型表达能力。每个 Transformer 层还包含层归一化和残差连接，以稳定训练过程。

2.4 分类层和损失函数

Transformer 的输出经过展平后传入全连接层，映射到二

分类的 logits。考虑到医学影像数据集中常见的类别不平衡问题，实验采用 Focal Loss 作为损失函数。通过降低高置信度样本的权重，使模型更加关注难以分类的样本，缓解类别不平衡问题。

3 实验

3.1 数据集

实验使用 3 个公开医学影像数据集按照 8:2 的比例随机划分为训练集和测试集来评估 CapsFormer 的性能，3 个数据集具有不同的类别分布和疾病特征，能够有效地评估 CapsFormer 在不同医学影像分类任务中的性能。

(1) TB-CXR (Tuberculosis-TB-Chest-Xray): 正类样本 700，正常样本 3 500，比例 5:1。

(2) CXP (Chest-Xray-Pneumonia): 正类样本 3 875，正常样本 1 341，比例 1:2.9。

(3) PedCXR (Pediatric Chest X-ray Pneumonia): 正类样本 4 281，正常样本 1 575，比例 1:2.7。

3.2 数据预处理

为增强模型的泛化能力，对数据集进行了以下的预处理：

(1) 调整大小至 $128 \text{ px} \times 128 \text{ px}$ ，降低计算复杂度，同时保留关键病灶特征。

(2) 随机水平翻转（概率 0.5），模拟患者姿势变化，增加数据的多样性。

(3) 随机旋转 ($\pm 10^\circ$)，模拟成像设备角度差异。

(4) 标准化（均值 [0.485, 0.456, 0.406]，标准差 [0.229, 0.224, 0.225]），加速模型收敛。

3.3 评估指标

准确率 (Accuracy) 衡量正确分类的样本比例，反映总体性能；精度 (Precision) 计算真阳性占正类预测的比例，评估正类预测的准确性；召回率 (Recall) 表示真阳性占实际正类的比例，衡量正类覆盖率； F_1 分数为精度与召回率的调和平均值，综合评估类别不平衡场景下的表现；AUC-ROC 为接收者操作特征曲线下面积，反映模型区分正负类的能力。

3.4 实验设置

本实验是在 Kaggle 平台上进行，以确保结果的可比性和可重复性。硬件配置包括 NVIDIA Tesla P100 GPU、Intel Xeon 处理器和 16 GB DDR4 RAM。软件环境基于 Ubuntu 18.04 LTS 操作系统，使用 Python 3.8 作为编程语言，结合 PyTorch 1.9.0 深度学习框架，以充分利用 GPU 加速能力。还依赖 NumPy、Pandas、Scikit-learn 等库。实验设置如下：

(1) 主实验：评估完整的 CapsFormer 在 3 个数据集上的性能，来记录训练过程。

(2) 超参数调优实验：通过网格搜索测试学习率（0.001、0.000 1、0.000 01）和批次大小（16、32、64），记录不同参数组合下的验证集表现，以确定最优的配置。

(3) 消融实验：分别测试单独使用 CapsNet、Transformer、CapsFormer + Focal Loss、CapsNet + Focal Loss、CapsFormer + Cross Entropy、CapsNet + Cross Entropy 的变体，以分析各组件的贡献。

4 实验结果与分析

4.1 主实验结果与分析

表 1 总结了模型在这些数据集上的关键性能指标，包括准确率、精度、召回率、 F_1 分数以及 AUC-ROC 值。

表 1 CapsFormer 模型在医学影像分类数据集上的性能指标

数据集	准确率 /%	精度	召回率	F_1 分数	AUC-ROC
TB-CXR	99.17	0.995 5	0.956 5	0.974 5	0.982 2
CXP	97.70	0.984 3	0.984 3	0.984 3	0.995 4
PedCXR	93.69	0.988 2	0.617 6	0.790 2	0.978 3

CapsFormer 在 TB-CXR 数据集上取得了较好的结果，反映了模型在肺结核分类任务中的卓越性能。在 CXP 数据集上，模型各项指标均保持在较高水平，显示出对成人肺炎的分类能力。对于 PedCXR 数据集，虽然准确率略低，但模型仍取得了良好的结果，表明其在儿童肺炎任务中具有较好的泛化能力。这些结果超越了许多现有的方法，如在类似数据集上，Wang 等^[7] 报告 94.5% 准确率和 Rajaraman 等^[8] 提出 96.2% 准确率。

CapsFormer 的出色表现主要是因为：CapsNet 的几何感知能力与 Transformer 的全局建模能力相互补充，使模型能够同时理解局部病理特征和它们的空间关系；CapsNet 的动态路由算法使模型能够学习病变特征的层次结构，这对于医学影像分析十分重要；Transformer 的自注意力机制能够有效捕捉远程依赖关系，帮助模型理解病变特征之间的上下文关联；Focal Loss 的应用使模型能够有效地处理数据集中的类别不平衡问题，能够平衡学习少数类和多数类的特征，提高整体分类性能。

4.2 超参数调优实验结果与分析

如图 1~2 所示，本文在 3 个公开数据集上进行了超参数调优实验。通过分阶段网格搜索策略，重点分析了学习率和批次大小对模型性能的影响。由实验可知，CapsFormer 模型在医学影像分类任务中的最佳配置为学习率 = 0.000 1 和批次大小 = 32。模型在 3 个公开数据集上均取得了优异性能，验证了 CapsFormer 融合 CapsNet 与 Transformer 的设计理念在医学影像分类中的有效性。CapsFormer 在保持高分类准确率

的同时，展现出较好的泛化能力和模型稳定性，为医学影像分析提供了新的技术路径。

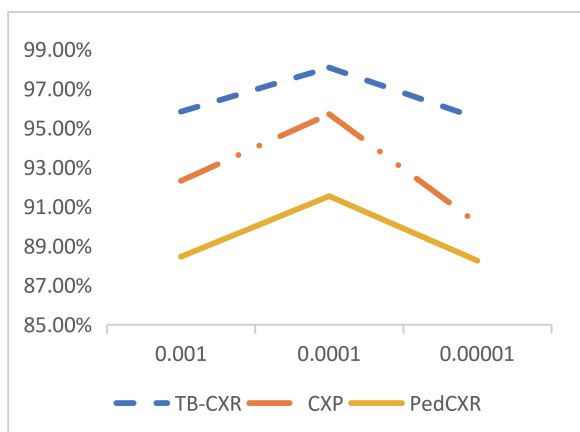


图 1 CapsFormer 模型在医学影像数据集上的学习率调优

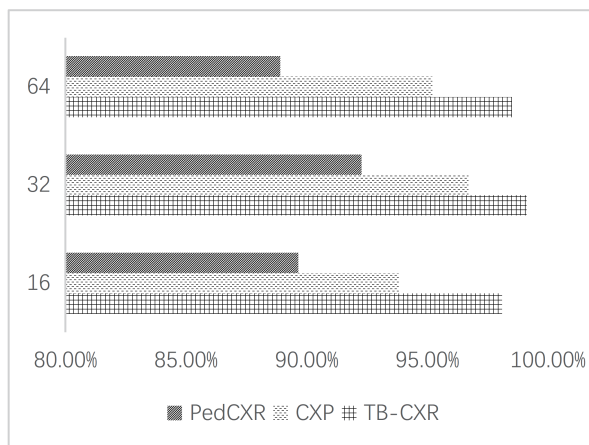


图 2 CapsFormer 模型在学习率 0.000 1 下的批次大小调优准确率

4.3 消融实验结果与分析

如图 3 所示，本文在 3 个公开数据集进行了一系列消融实验，来考察 Focal Loss 和 Cross Entropy 两种损失函数对模型性能的影响。

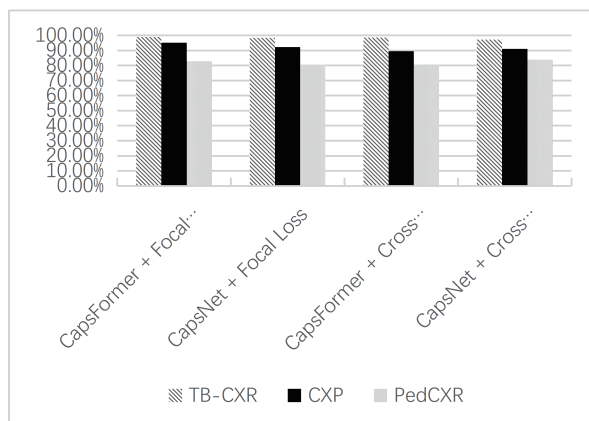


图 3 不同模型配置在 3 个医学影像数据集上的分类准确率

Focal Loss 在多数情况下优于 Cross Entropy，尤其是与

CapsFormer 结合时,能有效应对类别不平衡问题。然而,损失函数的最优选择依赖于数据集特性,如儿科肺炎数据更适配 Cross Entropy 与 CapsNet 的组合。CapsFormer 通过融合 Transformer 与 CapsNet 的优势,在医学影像分类中展现潜力,未来可进一步优化以应对更复杂任务。

5 结论与未来工作

5.1 结论

本文提出了一种融合 CapsNet 的几何感知能力与 Transformer 的全局上下文建模能力的新型医学影像分类架构 CapsFormer,为医学影像分类任务提供了一种有效的解决方案。CapsFormer 充分利用 CapsNet 对空间层次关系的敏感性和 Transformer 对长距离依赖关系的建模优势,在公开医学影像数据集上取得了优异的分类性能。不仅解决了传统 CNN 在空间建模方面的局限性,也弥补了单独使用 Transformer 时对局部几何细节感知不足的问题,为医学影像分类提供了一种有效的解决方案,其优秀的分类性能和对几何特征的感知能力使其具有广阔的临床应用前景,为计算机辅助诊断系统的发展提供了新的技术路径。

5.2 未来工作

虽然 CapsFormer 在医学影像分类任务中表现出色,但仍存在进一步的改进和拓展的空间。未来工作可以探索 CapsFormer 与其他模态医学数据的融合策略,构建更加全面的诊断模型;当前的 CapsFormer 架构计算复杂度较高,限制了其在资源受限环境中的应用。未来可以借鉴 MobileViT 等^[9]工作,探索胶囊网络与轻量级 Transformer 的结合,降低模型参数量和计算成本,使其更适用于临床实时应用;医学影像面临严重的域偏移问题,如不同设备、不同医院间的图像差异。未来可以结合域自适应技术^[10],增强 CapsFormer 在跨域场景中的鲁棒性,减少模型在新数据源上的性能下降。这些未来的研究方向将进一步发挥 CapsFormer 在医学影像分析中的潜力,推动智能医疗技术向更加精准、高效和可信的方向发展。

参考文献:

- [1] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017,60(6):84-90.
- [2] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 770-778.

- [3] SABOUR S, FROSST N, HINTON G E. Dynamic routing between capsules[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 3859-3869.
- [4] AFSHAR P, MOHAMMADI A, PLATANIOTIS K N. Brain tumor type classification via capsule networks[C]//2018 25th IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2018: 3129-3133.
- [5] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc, 2017: 5998-6008.
- [6] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[C/OL]//International Conference on Learning Representations (ICLR). 2021[2025-08-19]. <https://openreview.net/pdf?id=YicbFdNTTy>.
- [7] WANG X S, PENG Y F, LU L, et al. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases[C]//Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics. Berlin: Springer, 2017: 2097-2106.
- [8] RAJARAMAN S, CANDEMIR S, KIM I, et al. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs[J]. Applied sciences, 2018, 8(10): 1715.
- [9] MEHTA S, RASTEGARI M. MobileViT: light-weight, general-purpose, and mobile-friendly vision transformer[EB/OL].(2022-03-04)[2025-11-03].<https://doi.org/10.48550/arXiv.2110.02178>.
- [10] TALEB A, LOETZSCH W, DANZ N, et al. 3D self-supervised methods for medical imaging[EB/OL].(2020-11-02)[2025-11-06].<https://doi.org/10.48550/arXiv.2006.03829>.

【作者简介】

郭万玲(2004—),女,陕西商洛人,本科,研究方向:人工智能。

刘海川(1992—),男,重庆忠县人,博士,研究方向:多模态、人工智能。

(收稿日期:2025-06-12 修回日期:2025-12-03)