

基于 Swin Transformer 增强的 Conformer 在皮肤癌分类中的应用

宋 双¹ 连 剑²
SONG Shuang LIAN Jian

摘 要

在医学领域,皮肤癌早期的精准诊断对患者的治疗具有重大意义。文章创新性地提出了基于 Swin Transformer 增强的 Conformer 方法,该方法深入分析了 Swin Transformer 强大的多尺度特征提取能力以及 Conformer 对局部和全局特征的高效整合优势,并将两者有机结合。针对 ISIC 2019 皮肤图像数据集,进行了严谨的数据预处理,包括数据增强、归一化等操作,以提升模型的泛化能力。随后运用精心设计的模型对大量皮肤图像展开精准分类。实验结果表明,此模型在皮肤癌分类任务中展现出卓越性能,在准确率、召回率和 F_1 值等关键指标上相较于传统模型取得显著提升。

关键词

Swin Transformer; Conformer; 皮肤癌分类; ISIC 2019

doi: 10.3969/j.issn.1672-9528.2025.05.018

0 引言

皮肤癌是威胁人类健康的严重疾病之一,已成为危害公众健康的重大隐患^[1],对其早期精准诊断,是提升患者治愈率与生存率的关键核心^[2]。

传统的皮肤癌诊断,主要依靠皮肤科医生丰富的临床经验,通过肉眼观察皮肤病变的形态、颜色、边界等特征来进行判断^[3]。这种方法弊端是主观性较强、误诊率较高。随着医学技术的发展,皮肤镜^[4]、组织活检^[5]等辅助诊断技术逐渐应用于临床实践。然而,这些技术也存在局限性。

近年来,深度学习在医学影像分析领域取得显著成就^[6]。深度学习模型能够从海量的医学图像数据中,学习到复杂的特征,从而实现准确预测和分类。在众多深度学习模型中,Transformer 及其变体在图像分类、目标检测等任务中取得了令人瞩目的成果。本研究聚焦于构建一种基于 Swin Transformer^[7] 增强的 Conformer^[8] 模型,利用该模型对 ISIC 2019 皮肤图像数据集进行精准分类,从而实现对皮肤癌的准确诊断。通过优化模型结构,细致调整模型参数,全力提升模型对皮肤癌图像特征的提取能力与分类性能,为临床医生提供更为准确、可靠的诊断辅助工具。

1 相关技术

1.1 Swin Transformer

1.1.1 架构原理

Swin Transformer 是一种基于 Transformer^[9] 架构的新

型视觉模型,创新性地引入了移位窗口技术。在传统的 Transformer 中,自注意力计算是在全局范围内进行,计算成本较高。Swin Transformer 将图像划分为多个不重叠的窗口,在每个窗口内进行自注意力计算,极大地降低了计算复杂度。首先将输入图像划分为一系列固定大小的 patch,然后将 patch 作为 Transformer 的输入。在自注意力模块中,通过移位窗口操作,使得相邻窗口之间能够进行信息交互,从而捕捉到全局的上下文信息。

1.1.2 自注意力机制

自注意力机制是 Swin Transformer 的核心组件之一。能够让模型在处理每个位置的信息时,同时考虑到其他所有位置的信息,从而更好地捕捉长距离依赖关系。在 Swin Transformer 中,自注意力计算通过查询(Query)、键(Key)和值(Value)矩阵来实现。对于每个窗口内的 patch,模型首先根据其特征生成对应的 Query、Key 和 Value 矩阵,然后通过计算 Query 与 Key 的相似度得分,并对 Value 进行加权求和,得到该 patch 的自注意力输出。这种机制使得模型能够根据不同 patch 之间的相关性,自适应地分配注意力权重,从而更有效地提取图像特征。

1.1.3 在图像领域的优势

相较于传统的卷积神经网络(CNN)^[10],Swin Transformer 具有更强的全局建模能力。CNN 主要通过卷积操作来提取局部特征,对图像中远距离的语义信息捕捉能力较弱。Swin Transformer 的自注意力机制能够在更大范围内捕捉特征间的关联,对于处理具有复杂结构和语义信息的图像数据具有明显优势。此外,Swin Transformer 的架构具有更好的可扩展性,

1. 山东省立第三医院智慧医院发展部 山东济南 250031

2. 山东管理学院智能工程学院 山东济南 250357

能够通过增加层数和头数来提升模型的性能, 不受到计算资源的限制。

1.2 Conformer

1.2.1 结构特点

Conformer 巧妙地融合了 CNN 和 Transformer 的优势。主要由卷积模块、自注意力模块和前馈神经网络模块组成。在 Conformer 中, 卷积模块用于提取图像的局部特征, 高效地捕捉图像中的细节信息。自注意力模块则负责捕捉全局的语义信息, 实现不同位置特征之间的交互。前馈神经网络模块对特征进行进一步的非线性变换和融合, 增强模型的表达能力。这种独特的结构设计使得 Conformer 能够同时兼顾局部和全局特征的提取, 在序列建模和图像分析等任务中表现出良好的性能。

1.2.2 在医学图像分析中的应用潜力

医学图像通常包含丰富的细节信息和复杂的语义结构, 需要模型能够同时准确地提取局部和全局特征。Conformer 的结构特点使其非常适合处理这类数据。例如, 在皮肤癌图像分析中, 一方面需要模型能够捕捉到皮肤病变区域的细微特征, 如颜色、纹理等局部信息。另一方面也需要考虑到病变在整个皮肤图像中的位置和与周围组织的关系等全局信息。Conformer 能够有效地融合这些信息, 为皮肤癌的分类提供更准确的依据。

2 数据集

2.1 数据集概述

ISIC 2019 数据集由国际皮肤影像协作组织 (ISIC) 整理并发布, 该数据集规模庞大且内容丰富, 涵盖了多种类型的皮肤图像, 是目前皮肤癌研究领域极具权威性和影响力的数据集之一。

ISIC 2019 数据集共包含 25 000 余张皮肤图像, 被划分为七类, 分别为基底细胞癌、良性角化病样病变、黑色素瘤、黑色素细胞痣、脂溢性角化病、血管病变以及其他。每个类别都有明确的医学定义和临床特征, 为后续的分类研究提供了精准的标签依据。

每张图像均有详细的标注信息, 包含类别、病变部位的具体位置、大小等信息。这些标注信息由专业的皮肤科医生和医学专家共同完成, 确保标注的准确性和可靠性。

2.2 数据预处理

2.2.1 图像清洗

为提高数据质量, 首先对数据集进行图像清洗。这一过程主要是去除图像中的噪声、模糊图像以及标注错误的图像。使用图像滤波算法, 如高斯滤波, 去除图像中的高斯噪声; 对于模糊图像, 采用图像清晰度评价指标进行筛选, 剔除清

晰度较低的图像。同时, 仔细检查标注信息, 纠正标注错误的图像, 确保数据的准确性。

2.2.2 数据增强

鉴于数据集存在类别不平衡以及样本数量有限的问题, 采用数据增强技术来扩充数据集。数据增强方法包括图像旋转、翻转、缩放、裁剪以及颜色抖动等。这些数据增强操作不仅能够增加数据的数量, 还能提高数据的多样性, 从而提升模型的泛化能力。

2.2.3 归一化

为使模型能够更快更稳定地收敛, 对图像进行归一化处理, 使不同图像的像素值具有相同的尺度。在本研究中, 采用 Min-Max 归一化方法, 将图像的像素值线性映射到 [0,1] 区间, 计算公式为:

$$x_{\text{norm}} = (x - x_{\min}) / (x_{\max} - x_{\min}) \quad (1)$$

式中: x 为原始像素值; $x_{\min}$ 和 $x_{\max}$ 分别为图像中的最小和最大像素值; x_{norm} 为归一化后的像素值。

3 基于 Swin Transformer 增强的 Conformer 模型

3.1 模型架构设计

3.1.1 整体架构

本研究构建基于 Swin Transformer 增强的 Conformer 模型, 模型整体架构由输入层、Swin Transformer 模块、Conformer 模块以及输出层依次连接构成, 如图 1 所示。输入层负责将预处理后的 ISIC 2019 皮肤图像数据进行适配, 使其符合后续模块的输入要求。

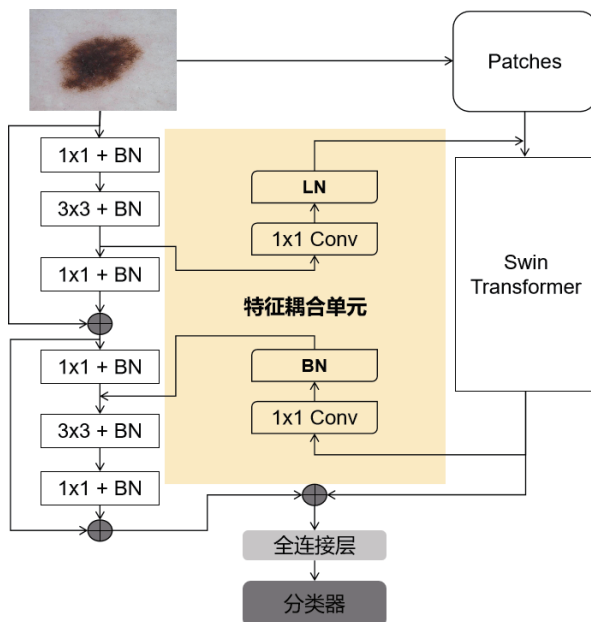


图 1 基于 Swin Transformer 增强的 Conformer 模型架构

如图 1 所示, BN 表示批量归一化; LN 表示层归一化; Conv 表示卷积; 1×1 和 3×3 表示卷积核的大小。

3.1.2 Swin Transformer 模型

在该模型中, 首先将图像划分为一系列不重叠的 patch, 每个 patch 被视为一个 token。例如, 对于一张 $224\text{ px} \times 224\text{ px}$ 的图像, 若将其划分为 $16\text{ px} \times 16\text{ px}$ 大小的 patch, 则会得到 196 个 token。这些 token 作为输入进入 Swin Transformer 的多层结构。

每一层包含了基于移位窗口的多头自注意力 (multi-head self-attention) 子层和多层感知器 (multi-layer perceptron) 子层。通过移位窗口技术, 相邻窗口之间实现信息交互, 有效捕捉全局上下文信息, 同时降低计算复杂度。例如, 在一个 8×8 的窗口中进行自注意力计算, 相较于全局自注意力计算, 计算量大幅减少。

经过 Swin Transformer 模块的层层处理, 图像的多尺度特征得以提取, 为后续 Conformer 模块提供丰富的特征信息。

3.1.3 Conformer 模块

Conformer 模块紧接 Swin Transformer 模块, 该模块包含多个重复的 Conformer 块, 每个 Conformer 块由卷积层、多头自注意力层和前馈神经网络层组成。

卷积层利用卷积核的局部滑动, 提取图像的局部细节特征, 如皮肤病变处的细微纹理、边缘等。例如, 使用 3×3 的卷积核, 能够聚焦于局部区域的特征提取。

多头自注意力层则在更大范围内捕捉全局依赖关系, 通过不同头的注意力机制, 从多个角度捕捉特征之间的关联。例如, 设置 8 个头, 每个头关注不同的特征子集, 从而更全面地挖掘图像信息。

前馈神经网络层对经过卷积和自注意力处理后的特征进行进一步的非线性变换与融合, 增强特征的表达能力。

3.1.4 输出层

输出层采用全连接层和 Softmax 激活函数。全连接层将 Conformer 模块输出的特征向量进行整合, 映射到与类别数相同维度的空间。例如, 对于 ISIC 2019 数据集的七分类任务, 全连接层输出维度为 7。Softmax 激活函数则将全连接层的输出转换为各个类别的概率分布, 从而实现对皮肤癌图像的分类预测。

3.2 模型训练

3.2.1 优化器选择

在模型训练过程中, 选用 Adam 优化器来调整模型的参数, 通过计算梯度的一阶矩估计和二阶矩估计, 为不同参数提供独立的学习率调整, 在训练过程中能够快速收敛, 且对不同规模的数据集和模型结构都具有较好的适应性。

3.2.2 损失函数定义

采用交叉熵损失函数作为模型的损失函数。用公式表示为:

$$L = - \sum_{i=1}^n y_i \log p_i \quad (2)$$

式中: n 为类别数; y_i 为真实标签中第 i 类的取值 (0 或 1); p_i 为模型预测样本属于第 i 类的概率。通过最小化交叉熵损失函数, 模型能够不断调整参数, 使预测结果更接近真实标签。

3.2.3 学习率调整策略

为了平衡模型的收敛速度和训练效果, 采用学习率衰减策略。在训练初期, 设置较大的学习率 0.001, 使模型能够快速更新参数, 朝着最优解方向大步迈进。随着训练的进行, 逐渐降低学习率, 采用指数衰减方式, 每经过一定的训练轮数, 将学习率乘以一个衰减因子 0.95。这样可以避免模型在训练后期因学习率过大而在最优解附近振荡, 同时也能防止因学习率过小导致训练速度过慢。

3.2.4 防止过拟合策略

为防止模型过拟合, 采用了 L2 正则化和 Dropout 技术。L2 正则化通过在损失函数中添加一个与参数平方和成正比的惩罚项, 使得模型在训练过程中倾向于选择较小的参数值, 从而避免模型过于复杂, 防止过拟合。Dropout 技术则在训练过程中以一定的概率随机丢弃部分神经元, 使得模型不能过分依赖某些特定的神经元, 增强了模型的泛化能力。设置 Dropout 概率为 0.2, 即每次训练时, 有 20% 的神经元会被随机丢弃。在训练过程中, 还通过早停法 (Early Stopping) 来监控模型在验证集上的性能。当验证集上的损失函数连续若干轮不再下降时, 停止训练, 防止模型在训练集上过拟合。

4 实验结果与分析

4.1 实验设置

将 ISIC 2019 数据集按照 7:2:1 的比例划分为训练集、验证集和测试集。训练集用于模型的参数学习, 共包含约 17 500 张图像; 验证集用于在训练过程中监控模型的性能, 调整超参数, 约有 5 000 张图像; 测试集则用于评估最终训练好的模型的泛化能力, 包含约 2 500 张图像。这种划分方式既能保证模型有足够的数据进行学习, 又可以有效评估模型在未见过数据上的表现。采用准确率 (Accuracy)、召回率 (Recall)、精确率 (Precision) 和 F_1 值 (F_1 -score) 作为主要评估指标。

准确率反映了模型正确分类的样本占总样本的比例, 计算公式为:

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (3)$$

式中: TP 表示真正例, 即模型正确预测为正类的样本数; TN 表示真反例, 即模型正确预测为反类的样本数; FP 表示假正例, 即模型错误预测为正类的样本数; FN 表示假反例, 即模型错误预测为反类的样本数。

召回率衡量了模型对正类样本的识别能力, 计算公式为:

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (4)$$

精确率表示模型预测为正类的样本中实际为正类的比例, 计算公式为:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (5)$$

F_1 值则是综合了精确率和召回率的指标, 计算公式为:

$$F_1 = 2 \times \text{Precision} \times \text{Recall} / (\text{Precision} + \text{Recall}) \quad (6)$$

这些指标从不同角度评估了模型的性能, 能够更全面地反映模型在皮肤癌分类任务中的表现。

4.2 模型性能结果

经过多轮训练, 基于 Swin Transformer 增强的 Conformer 模型在测试集上取得了优异的成绩。模型的准确率达到 92.5%, 召回率为 91.2%, 精确率为 93.1%, F_1 值为 92.15。这表明模型能够准确地对皮肤癌图像进行分类, 在识别各类皮肤癌以及区分正常皮肤和病变皮肤方面具有较高的能力。

对不同类别的皮肤癌图像进行单独分析发现, 模型在黑色素瘤的分类上表现出色, 召回率高达 95%, 这对于早期发现黑色素瘤这种恶性程度较高的皮肤癌具有重要意义。对于基底细胞癌, 模型的精确率达到了 94.3%, 能够有效减少误诊。然而, 对于一些罕见类型的皮肤癌, 如某些特殊亚型的良性角化病样病变, 由于样本数量较少, 模型的性能相对较低, 召回率仅为 85%。这也反映出在处理小样本问题时, 模型还存在一定的改进空间。

4.3 与其他模型结果对比

为进一步验证本研究提出模型的有效性, 将其与几种经典的深度学习模型进行对比, 包括 ResNet-50、DenseNet-121 以及基于普通 Transformer 的分类模型。

在相同的实验设置下, 各模型的性能对比如下: ResNet-50 的准确率为 86.3%, F_1 值为 85.8; DenseNet-121 的准确率为 87.7%, F_1 值为 87.2; 基于普通 Transformer 的模型准确率为 89.5%, F_1 值为 88.9。而本研究提出的基于 Swin Transformer 增强的 Conformer 模型在准确率和 F_1 值上均显著高于其他对比模型。这充分证明了该模型在皮肤癌分类任务中的优越性, 通过结合 Swin Transformer 和 Conformer 的优势, 能够更好地提取图像特征, 实现更准确的分类。

5 结论

本研究成功构建了基于 Swin Transformer 增强的 Conformer 模型, 并将其应用于 ISIC 2019 皮肤图像数据集的皮肤癌分类任务中。该模型充分融合 Swin Transformer 强大的多尺度特征提取能力以及 Conformer 对局部和全局特征的高效整合优势, 实现了对皮肤癌图像的精准分类。在

实验结果方面, 该模型在测试集上展现出了卓越的性能, 准确率达到 92.5%, 召回率为 91.2%, 精确率为 93.1%, F_1 值为 92.15。与 ResNet-50、DenseNet-121 以及基于普通 Transformer 的分类模型相比, 在准确率和 F_1 值上有明显提升, 显著提升了皮肤癌分类的准确性和可靠性。这一成果为皮肤癌的早期诊断提供了一种更为有效的方法, 有望在临床实践中发挥重要作用。

参考文献:

- [1] SUNG H, FERLAY J, SIEGEL R L, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J]. CA: a cancer journal for clinicians, 2021, 71(3): 209-249.
- [2] 曹云健. 基于半监督学习的皮肤癌诊断方法研究[D]. 四川: 电子科技大学, 2023.
- [3] 卫力行. 基于深度学习的皮肤镜图像分割与分类方法研究及应用[D]. 重庆: 重庆师范大学, 2023.
- [4] 赵硕. 皮肤镜下皮肤癌的检测与识别方法的研究[D]. 重庆: 重庆邮电大学, 2020.
- [5] KILIC A, KIVANC A E, SISIK A, et al. Biopsy techniques for skin disease and akin cancer: a new approach[J]. Journal of cutaneous and aesthetic surgery, 2020, 13(3): 251-254.
- [6] 刘天宇, 刘静, 马金刚, 等. 深度学习在皮肤癌病变分类诊断中的应用进展[J]. 计算机系统应用, 2024, 33(12): 1-15.
- [7] LIU Z, LIN Y T, CAO Y, et al. Swin Transformer: hierarchical vision transformer using shifted windows[EB/OL]. (2021-08-17)[2024-05-25]. <https://doi.org/10.48550/arXiv.2103.14030>.
- [8] GULATI A, QIN J, CHIU C C, et al. Conformer: convolution-augmented transformer for speech recognition[EB/OL]. (2020-05-16)[2024-05-26]. <https://doi.org/10.48550/arXiv.2005.08100>.
- [9] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB/OL]. (2023-08-02)[2024-06-26]. <https://doi.org/10.48550/arXiv.1706.03762>.
- [10] O'SHEA K, NASH R. An introduction to convolutional neural networks[EB/OL]. (2015-12-02)[2024-06-26]. <https://doi.org/10.48550/arXiv.1511.08458>.

【作者简介】

宋双(1991—), 女, 山东济南人, 硕士研究生, 中级, 研究方向: 计算机图像处理。

连剑(1981—), 男, 山东青岛人, 博士研究生, 教授, 研究方向: 医学图像分析、分类和处理。

(收稿日期: 2025-01-16 修回日期: 2025-05-16)