CEViT-DeepFake-Scanner: 基于ConvNext、EfficientNet和ViT的AI生成人脸检测系统

摘要

随着深度学习技术的飞速发展，生成对抗网络（GAN）等人工智能模型在生成高度逼真的人脸图像方面取得了突破性进展。然而，这些技术的滥用也带来了严重的社会和安全隐患，特别是在虚假信息传播和身份盗用等方面。为了应对这一挑战，本研究提出了CEViT-DeepFake-Scanner，一个基于ConvNext、EfficientNet和ViT模型的深度伪造检测系统。该系统通过巧妙结合三种先进的卷积神经网络模型，能够有效地区分真实人脸图像和AI生成的合成图像。本文详细介绍了CEViT-DeepFake-Scanner的系统架构、数学模型、数据集准备、模型训练策略以及全面的评估方法。实验结果表明，该系统在检测AI生成的人脸方面表现出色，在准确率、精确率、召回率和F1分数等多个评估指标上均达到了极高的水平。

1. 引言

近年来，深度学习技术，尤其是生成对抗网络（GAN）的快速发展，使得生成高质量的虚假图像变得越来越容易以假乱真。这些生成图像在视觉上与真实图像几乎无法区分，给社会带来了诸多潜在威胁，如虚假新闻传播、身份盗用、金融欺诈等。因此，开发高效、准确的深度伪造检测系统成为了一个迫切需要解决的重要研究课题。

本研究提出的CEViT-DeepFake-Scanner系统，巧妙地结合了ConvNext、EfficientNet和ViT三种先进的卷积神经网络模型。ConvNext是一种新型的卷积神经网络架构，通过引入多尺度特征融合和注意力机制，具有较强的特征提取能力。EfficientNet则通过高效的模型缩放策略，在保持高精度的同时显著减少了计算资源的消耗。ViT（Vision Transformer）模型通过自注意力机制有效捕捉图像中的全局依赖关系，补充了卷积神经网络在局部特征提取方面的不足。通过融合这三种模型的优势，CEViT-DeepFake-Scanner能够在检测深度伪造图像方面表现出色，同时保持较低的计算复杂度。

2. 系统架构

CEViT-DeepFake-Scanner的整体架构设计包括数据预处理模块、特征提取模块、分类模块和后处理模块。每个模块都经过精心设计，以实现最佳的检测性能。

2.1 数据预处理

数据预处理是深度学习模型训练过程中的关键步骤，对模型的最终性能有着重要影响。在CEViT-DeepFake-Scanner中，我们采用了以下预处理步骤：

1. **图像尺寸调整**：将输入图像统一调整为384x384像素的分辨率。这一尺寸在保证图像细节的同时，能够有效减少计算资源的消耗。

2. **标准化处理**：对图像进行标准化，将像素值缩放到[0,1]范围内，并进行均值和标准差的归一化。具体使用的均值和标准差为：

   - 均值：[0.485, 0.456, 0.406]
   - 标准差：[0.229, 0.224, 0.225]

   标准化的数学表达式如下：

   $$x_{normalized} = \frac{x - \mu}{\sigma}$$

   其中，x 是原始像素值，μ 是均值，σ 是标准差。

3. **数据增强**：在训练过程中，我们采用了一系列数据增强技术，包括随机水平翻转、随机垂直翻转、随机旋转、颜色抖动、随机裁剪和随机灰度转换。这些增强方法有助于增加数据的多样性，有效防止模型过拟合，提高模型的泛化能力。

2.2 特征提取

特征提取模块是CEViT-DeepFake-Scanner的核心部分，主要由ConvNext、EfficientNet和ViT三个模型组成，每个模型负责从输入图像中提取不同层次和类型的特征。

2.2.1 ConvNext

ConvNext是一种基于卷积神经网络的新型架构，旨在结合Transformer的优势与卷积网络的高效性。具体特点包括：

1. **深度可分离卷积**：使用深度可分离卷积替代标准卷积，减少参数量和计算复杂度。深度可分离卷积的数学表达式如下：

   $$Y = (X * K_d) * K_p$$

   其中，X 是输入特征图，K_d 是深度卷积核，K_p 是逐点卷积核。

2. **注意力机制**：引入自注意力机制，增强模型对重要特征的关注。自注意力机制的数学表达式为：

   $$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

   其中，Q、K、V 分别是查询、键和值矩阵，d_k 是键的维度。

2.2.2 EfficientNet

EfficientNet通过复合缩放方法在网络的深度、宽度和分辨率之间进行平衡，从而在保持高精度的同时显著减少计算资源的消耗。其缩放公式如下：

$$
\begin{aligned}
depth: d &= \alpha^{\phi} \\
width: w &= \beta^{\phi} \\
resolution: r &= \gamma^{\phi}
\end{aligned}
$$

其中，ϕ 是缩放系数，α、β、γ 是通过网格搜索确定的常数。我们选用了EfficientNet-B0变体作为特征提取器，因其在参数量和计算复杂度上的优势适合实时检测任务。

2.2.3 ViT

ViT（Vision Transformer）模型通过自注意力机制有效捕捉图像中的全局依赖关系，弥补了卷积神经网络在捕捉长距离依赖关系上的不足。具体特点包括：

1. **自注意力机制**：ViT利用自注意力机制来建模图像中不同区域之间的关系，提高了对复杂特征的识别能力。
2. **高效的特征融合**：通过与ConvNext和EfficientNet的结合，ViT增强了整体模型的特征表达能力，提升了检测准确性。

2.3 特征融合与分类

分类模块通过将ConvNext、EfficientNet和ViT提取的特征进行融合，形成一个高维特征向量，随后通过全连接层和Sigmoid激活函数完成分类。具体实现步骤如下：

1. **特征融合**：将三个模型的输出特征进行拼接，形成一个包含不同特征的高维向量。
2. **特征降维与组合**：通过一个全连接层将高维特征向量映射到一个较低维度的空间，再通过另一个线性层组合特征，输出两个类别的概率。
3. **Sigmoid激活**：使用Sigmoid函数将输出映射到[0,1]区间，表示图像为伪造的概率。Sigmoid函数的数学表达式为：

$$
\sigma(x) = \frac{1}{1 + e^{-x}}
$$

在模型训练过程中，采用了加权的交叉熵损失函数（CrossEntropyLoss），结合WeightedRandomSampler对数据进行采样，以应对数据类别不平衡的问题。

2.4 后处理

后处理模块对分类结果进行进一步处理，以提高结果的可解释性和实用性。主要包括以下步骤：

1. **阈值判断**：对分类概率进行阈值判断，如果某个图像的伪造概率超过预设阈值（例如0.5），则判定其为伪造图像。
2. **结果输出**：将判断结果保存为CSV文件，包含文件名和对应的分类结果（0表示真实，1表示AI生成）。

3. 数据集准备

在本研究中，我们使用了一个包含真实和AI生成人脸图像的数据集。数据集的组织结构如下：

- 数据目录
  - 真实人脸子目录
  - AI生成人脸子目录

我们的FaceDataset类负责加载和预处理这些图像。主要的数据集准备步骤包括：

1. **数据加载**：从指定目录加载真实和AI生成的人脸图像，确保只读取合法的图片文件（如.jpg, .png等）。
2. **标签分配**：为真实人脸分配标签0，为AI生成人脸分配标签1。
3. **数据划分**：将数据集按照8:2的比例随机划分为训练集和验证集，以确保模型在不同数据上的泛化能力。
4. **数据增强**：对训练集应用包括随机水平翻转、随机垂直翻转、随机旋转、颜色抖动、随机裁剪和随机灰度转换等数据增强技术，从而增加数据多样性，防止模型过拟合。

4. 模型训练

在模型训练过程中，我们采用了一系列先进的优化策略和技术，以提高训练效率和模型性能。

4.1 优化器与损失函数选择

我们选择了Adam优化器来优化模型参数，初始学习率设置为0.0001。Adam优化器结合了动量和自适应学习率调整，适合处理稀疏梯度和非平稳目标。具体更新规则如下：

$$
\begin{aligned}
m_t &= \beta_1 m_{t-1} + (1-\beta_1) g_t \\
v_t &= \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \\
\hat{m}_t &= \frac{m_t}{1-\beta_1^t} \\
\hat{v}_t &= \frac{v_t}{1-\beta_2^t} \\
\theta_t &= \theta_{t-1} - \alpha \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}
\end{aligned}
$$

其中，m_t 和 v_t 分别是梯度的一阶矩和二阶矩估计，β_1 和 β_2 是衰减率，α 是学习率，ϵ 是一个小常数。

为了应对训练过程中的类别不平衡问题，我们采用了加权交叉熵损失函数（CrossEntropyLoss）并结合WeightedRandomSampler对数据进行采样。具体来说，CrossEntropyLoss通过为不同类别赋予不同的权重，来缓解模型对多数类别的偏好；而WeightedRandomSampler根据样本权重进行采样，进一步平衡训练过程中的类别分布。

4.2 学习率调度

为了动态调整学习率，避免在训练后期出现梯度消失或过大的问题，我们采用了**CosineAnnealingLR**学习率调度策略。该策略在训练过程中按照余弦函数方式逐渐降低学习率，使模型在训练后期能够更精细地调整参数。具体实现如下：

$$
\text{scheduler} = \text{CosineAnnealingLR}(\text{optimizer}, T_{max} = \text{epochs}, \eta_{min} = 1e-6)
$$

其中，T_{max}是调度周期，η_{min}是学习率的下限。

4.3 混合精度训练

为加速训练过程并减少内存占用，我们采用了混合精度训练（Mixed Precision Training），使用FP16（半精度浮点数）进行部分计算，同时保持FP32（单精度浮点数）的模型权重。这种方法在不显著降低模型精度的情况下，大幅提高了训练速度和计算效率。

4.4 梯度裁剪与早停机制

在训练过程中，为防止梯度爆炸，我们引入了梯度裁剪（Gradient Clipping），具体实现为将梯度的L2范数限制在1.0以内。此外，为了防止模型过拟合，我们实施了早停（Early Stopping）机制。当验证集的ROC-AUC在连续多个epoch（设定为4次）内未能提升时，训练将提前停止，保存当前最佳模型。

4.5 模型训练过程

模型训练持续30个epoch，每个epoch后都会在验证集上评估模型性能，并保存模型检查点。在训练过程中，我们记录了训练损失、验证损失、准确率、精确率、召回率、F1分数和ROC-AUC等多个评估指标，以全面衡量模型的性能。

训练过程的主要步骤如下：

1. **输入数据**：从训练数据加载器中获取批量数据，包括图像和对应的标签。
2. **前向传播**：将输入图像通过特征提取模型（ConvNext、EfficientNet、ViT）提取特征，并进行特征融合和分类。
3. **损失计算**：通过加权交叉熵损失函数计算预测结果与真实标签之间的差距。
4. **反向传播与参数更新**：利用Adam优化器和混合精度训练策略，更新模型参数。
5. **模型评估**：在验证集上评估模型性能，记录各项评估指标，并依据ROC-AUC进行模型保存和早停判断。

5. 实验结果与分析

我们在384x384像素的图像数据集上进行了全面的实验评估。以下是详细的实验结果：

- 损失（Loss）：0.0173
- 准确率（Accuracy）：0.9941
- 精确率（Precision）：0.9985
- 召回率（Recall）：0.9942
- F1分数（F1 Score）：0.9964
- ROC-AUC：0.9992

这些评估指标的计算公式如下：

- **准确率**：$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
- **精确率**：$Precision = \frac{TP}{TP+FP}$
- **召回率**：$Recall = \frac{TP}{TP+FN}$
- **F1分数**：$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$
- **ROC-AUC**：表示分类器区分正负样本的能力，值越接近1表示性能越好。

其中，TP是真正例，TN是真负例，FP是假正例，FN是假负例。

实验结果表明，CEViT-DeepFake-Scanner在检测AI生成的人脸方面表现出色，具有极高的准确性和可靠性。特别是在精确率和ROC-AUC方面的卓越表现，意味着该系统在实际应用中能够最

大限度地减少误报，并对伪造图像进行有效检测，这对于维护社会信任和防止错误指控至关重要。

6. 结论

本文提出了CEViT-DeepFake-Scanner，一个基于ConvNext、EfficientNet和ViT模型的深度伪造检测系统。通过巧妙结合三种先进的卷积神经网络模型，CEViT-DeepFake-Scanner在检测AI生成的人脸方面取得了显著效果。实验结果表明，该系统在准确率、精确率、召回率、F1分数和ROC-AUC等多个评估指标上均达到了极高的水平。

未来的研究可以进一步优化模型架构，引入更多类型的特征提取器，或采用更先进的训练策略，以提升系统在更复杂环境下的检测能力。此外，扩展数据集的多样性和规模，以及在更多实际应用场景中的部署和测试，将有助于验证和提升CEViT-DeepFake-Scanner的实用性和鲁棒性。

致谢

感谢参与本研究的所有团队成员及提供数据支持的机构和个人。