**说明文档**

**前言**

**文档共分为三部分：第一部分讲解代码运行逻辑；第二部分介绍模型原理和创新点。第三部分为训练参数和策略。**

第一部分详细解释了代码的运行，输出CSV文件的位置，以及需要导入的库。

第二部分详细解释了算法模型的原理，和我们的想法以及创新点。

第三部分介绍了模型训练的参数和数据增强的策略。

**目 录**

一、代码运行讲解 1

二、模型原理和创新点 3

1、AI生成图像特点 3

2、对AI生成图像鉴别技术的研究 4

3、算法设计思路 5

4、创新点 8

三、训练参数和策略 8

**一、代码运行讲解**

**代码运行环境为：cuda版本12.4，pytorch版本2.4.1+cu124。**

首先必须要导入必要的库，所需库的版本等信息会在requirements.txt文件中进行详细介绍。

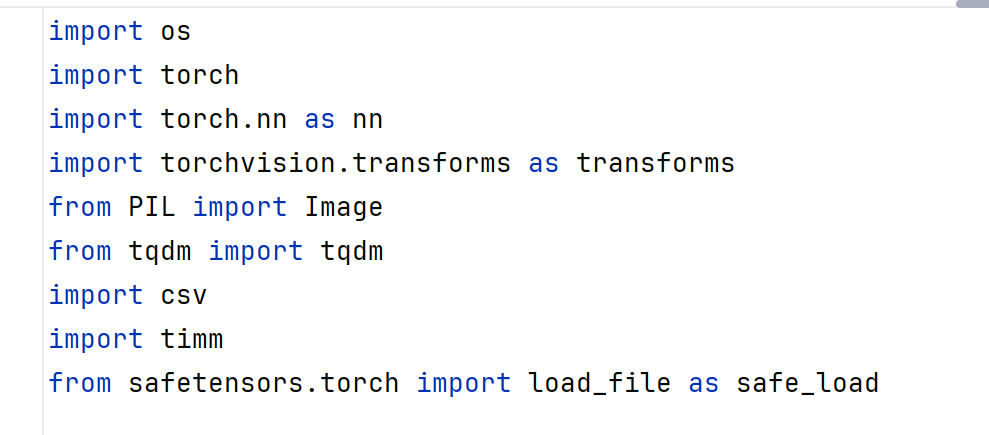
torch==2.4.1+cu124

torchvision==0.19.1+cu124

PIL==10.2.0

timm==1.0.9

safetensors==safetensors



图表 1 需要导入的库

必须导入预训练模型convnext\_large,我们将模型文件存放在了src文件夹中，可以自动导入。



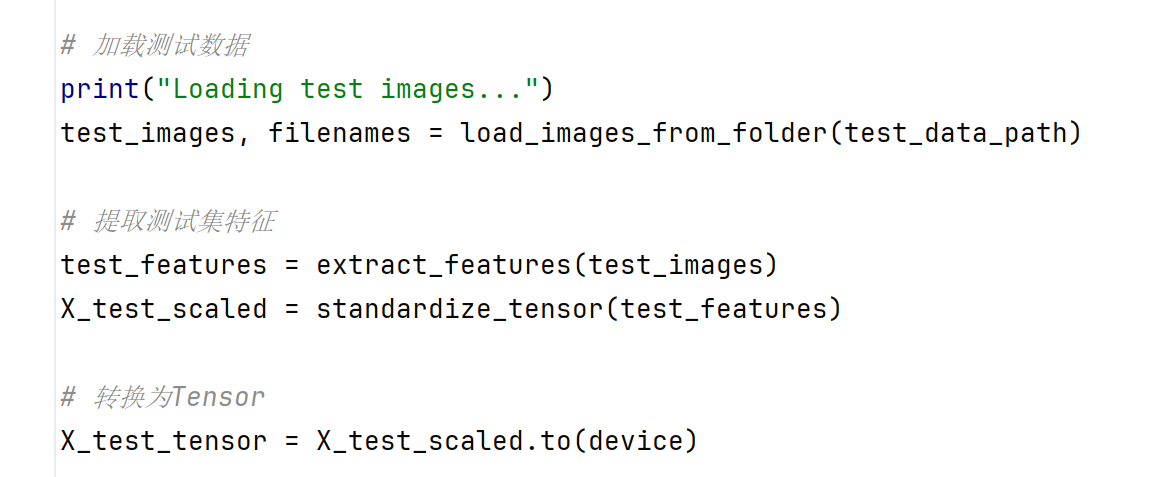
图表 2 导入预训练模型

****

图表 3 存放在src中的预训练模型权重

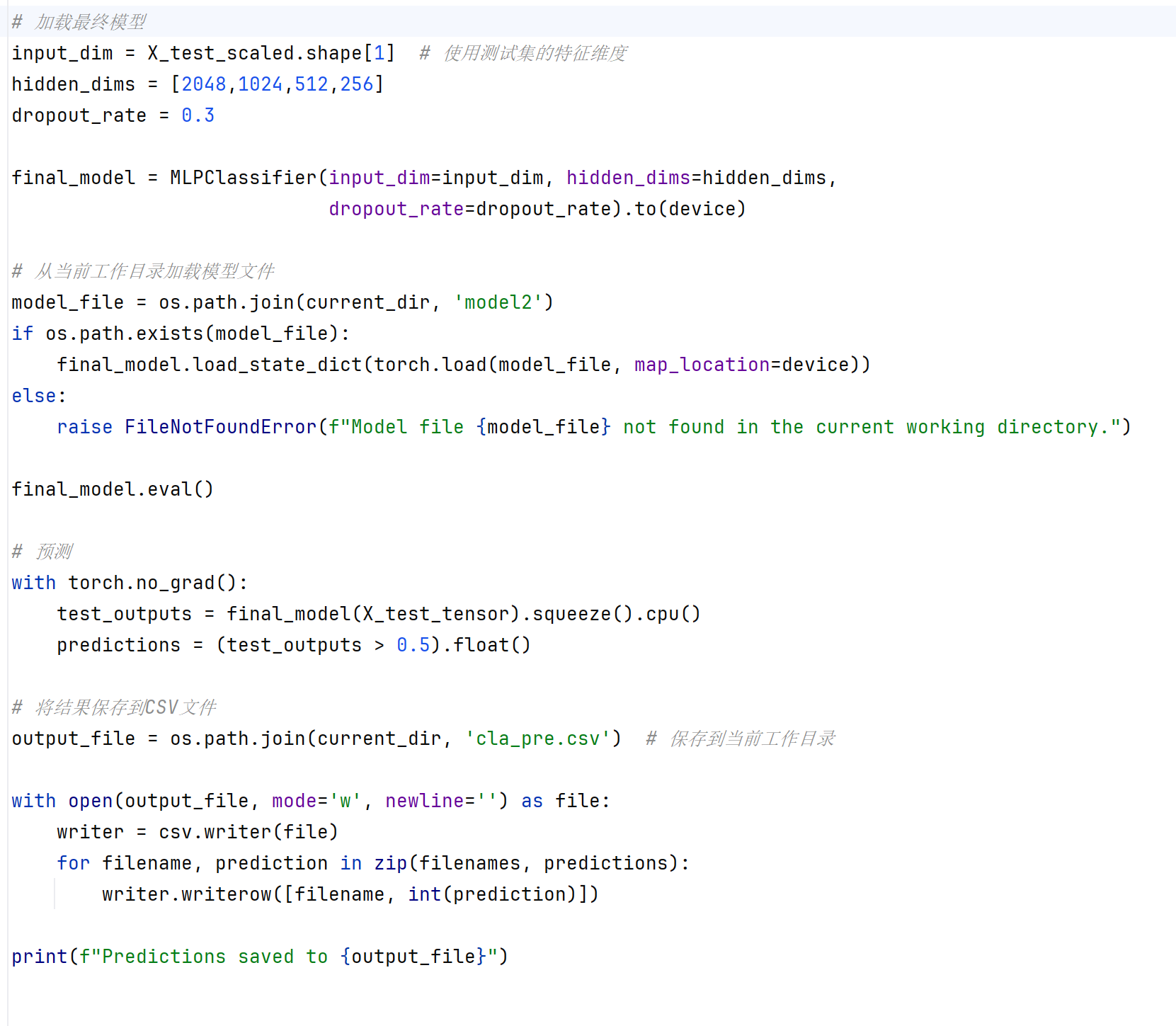
接着代码会**读取系统根目录下的testdata文件**，提取特征。





图表 4 读取测试集并提取特征

最后会从**src文件夹中加载**我们训练的分类器模型，然后进行分类，并**输出CSV文件到src文件夹**中。（根据我们的理解，./cla\_pre.csv的意思，其中./的意思就是当前运行脚本的目录下，就是main.py文件的同级目录下，也就是src目录下，所以我们将csv文件输出到src文件夹下，我们的理解是这样的。）



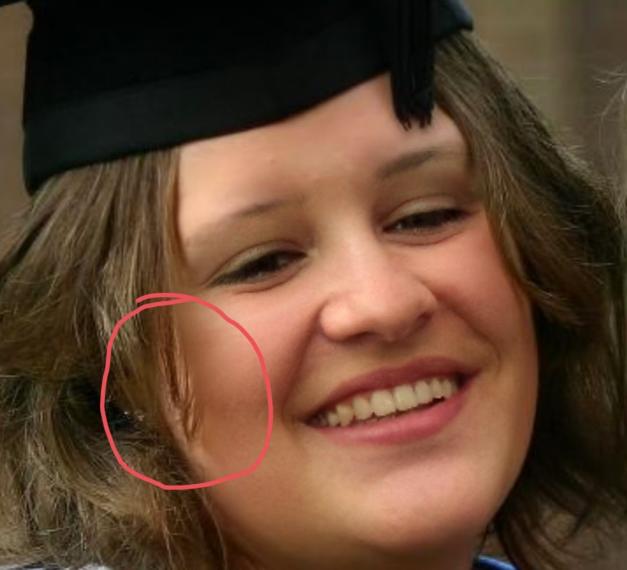
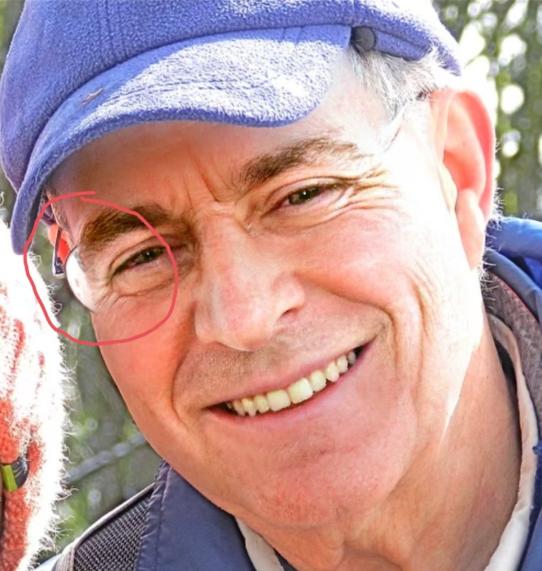
图表 5 加载分类器MLP分类并输出CSV文件

**二、模型原理和创新点**

1、AI生成图像特点

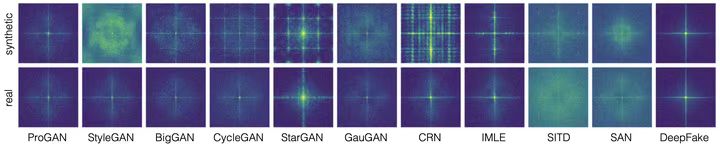
知己知彼才能百战百胜，要想准确鉴别AI生成图像，首先要了解AI生成图像的特点。

通过对AI生成人脸技术原理的深入研究发现，无论是何种生成模型，总是会在生成的图像上留下算法痕迹；在空域上的主要表现为AI生成算法在生成或者换脸的过程中会在图像上留下伪影，这些伪影主要包括面部扭曲、不同分辨率、边界色差等特征，在鼻子、嘴巴、脸的轮廓边缘上表现最为明显。



图表 6 空域上的伪影

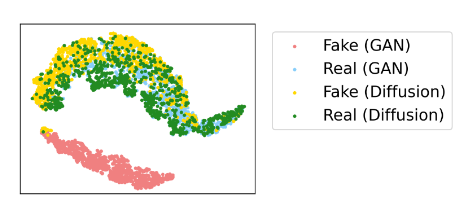
在频域上的主要表现为真实图像与GAN的伪造图像在高频成分上有着显著差异，GAN的伪造图像往往会展现出更多的高频信息，这种差异普遍认为是卷积核的上采样操作放大了高频信息，而其他模型生成的图像在各个频段中与真实图像也有一定的差异。



图表 7 频域上的差异

2、对AI生成图像鉴别技术的研究

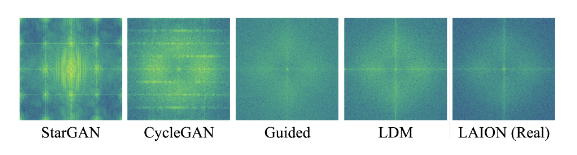
在上述的AI生成图像特征中可以看到，在空域和频域上确实和真实图像存在着一些差异，所以在AI生成图像鉴别的技术中，一种最为常见的方法是将假图像检测视为一个基于学习的问题，将一组真实的和伪造的图像作为训练集。然后训练一个深度网络来执行真实与伪造的二分类任务。这种模型迄今为止确实在检测同一生成模型家族中不同算法生成的假图像中表现出了很好的泛化能力；但是根据Utkarsh Ojha等的研究表明，这种泛化能力仅仅局限于在同一生成模型家族中，如果使用GAN模型训练的鉴别模型去鉴别扩散模型的生成图像则会导致效果非常不好。根据Utkarsh Ojha的研究表明，分类器很容易抓住区分伪造图像和真实图像的低级图像伪影，会使分类器的决策边界不均匀地绑定在伪造图像类别上，所以导致了这种跨生成模型的泛化能力较差的问题。



图表 8 特征空间分布

以一种常见分类器为例，(i) FGAN，由ProGAN生成的伪造图像组成；(ii) RGAN，用于训练ProGAN的真实图像；(iii) FDiffusion，由潜在扩散模型[46]生成的伪造图像；(iv) RDiffusion，用于训练潜在扩散模型的真实图像；真实与伪造分类器是在(i)和(ii)上训练的。在图8中可以看到，这种分类器在通过使用GAN图像训练后，在特征空间中这四种图像分布形成了两个明显的聚类，这种分类器可以很容易地区分FGAN和其他三种，但学到的真实类却没有自己的特征，而是被分类器用来形成一个汇类，容纳所有不是FGAN的内容。

根据现有研究表明，GAN模型生成的图像在频域上有着相似的独特模式，能够凭借这个特征实现对真实与GAN图像分类，但是如图9所示，在在扩散模型（Guided [23] 和 LDM [46]）生成的伪造图像中，这种模式却消失了，反而与真实图像的模式非常相似，所以导致分类器抓住了这种模式特征，反而忽略了真实图像的特征，导致在不存在这种模式的扩散模型上的泛化能力非常差。



图表 9 频谱特征

3、算法设计思路

为了解决传统分类器倾向于捕捉单一生成模型伪造图像的特有指纹特征，而忽视真实图像本质特性的问题，本研究对Utkarsh Ojha的工作进行了深入探讨，并通过复现多种模型进行了详尽的实验分析。通过对比不同模型的实验数据，我们旨在寻找一种更为有效的解决方案，以克服现有模型在面对未知生成源图像时的局限性。

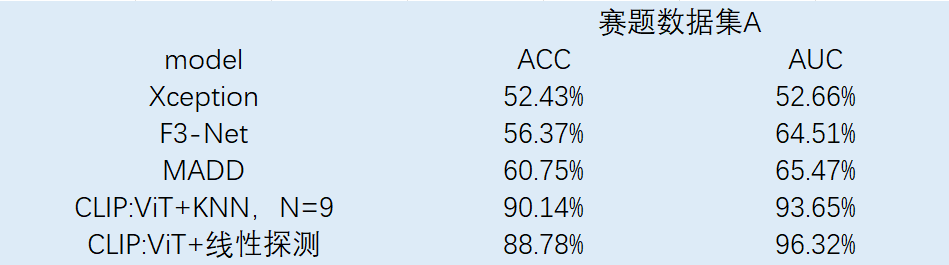
Utkarsh Ojha的研究主要采用的鉴别模型是CLIP:ViT+KNN或一个简单的线性探测：

**最近邻分类：**给定预训练的CLIP:ViT视觉编码器，使用其最后一层ϕ将整个训练数据映射到它们的特征表示（768维）。由此产生的特征库是ϕbank = {ϕR ∪ ϕF}，其中ϕR = {ϕr1, ϕr2, ..., ϕrN} 和 ϕF = {ϕf1, ϕf2, ..., ϕfN}。在测试时，一张图像x首先被映射到其特征表示ϕx。使用余弦距离作为度量d，分别找到它在真实(ϕR)和伪造(ϕF)特征库中的最近邻居。基于最近邻居的距离，给出预测——真实:0, 伪造:1。

**线性分类：**取预训练的CLIP:ViT编码器，在其顶部添加一个带有Sigmoid激活函数的单一线性层，并且只训练这个新的分类层ψ，用于二元的真实与伪造分类，使用二元交叉熵损失。

同时我们也复现了三种传统的AI生成图像鉴别模型。

实验中统一采用LFW作为真实数据集、一种StyleGAN和一种LDM扩散模型作为伪造数据集进行训练。



图表 10 实验对比

通过实验观察，传统的鉴别模型在训练过程中同时优化特征提取与分类任务，这导致其在处理来自未知生成源的图像时表现不佳。具体表现为，此类模型倾向于捕捉特定生成算法的“指纹”特征，而忽视了真实图像的本质特性，从而错误地将由其他生成模型产生的图像识别为真实图像。

针对这一问题，Utkarsh Ojha在其论文中提出了一种的方法，即分类任务应在未受过专门训练以区分两类图像的特征空间中执行。此策略旨在防止模型过度依赖于某一类图像的特征标识，进而提高模型的泛化能力。实验结果表明，该方法有效缓解了上述挑战。

因此，我们提出了一种全新的鉴别模式，该模式将鉴别模型划分为两个独立的过程：特征提取过程和分类过程。在特征提取阶段，我们采用预训练模型，且不对该模型进行任何与AI生成图像相关的额外训练，而是直接利用其原始状态下的特征提取能力。而在分类阶段，则使用包含AI生成图像的训练集对该阶段的模型进行专门训练，以优化其在鉴别任务上的表现。

简而言之，这一模式可以概括为：“预训练特征提取模型+分类器”模式。通过将特征提取与分类任务解耦，我们旨在构建一个既能有效捕捉图像间细微差异，又具有良好泛化能力的鉴别系统。

在此框架下，预训练特征提取模型与分类器的选择尤为关键。前者需具备从AI生成图像与真实图像中高效提取差异特征的能力，后者则应通过充分训练，达到高精度鉴别AI生成图像的目标。

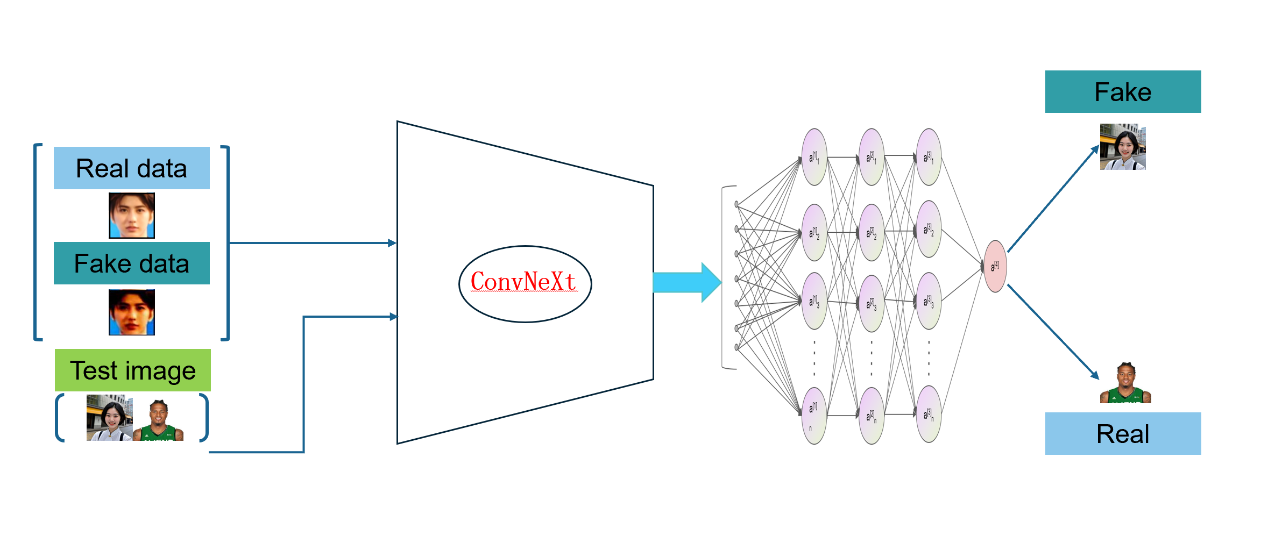
在Utkarsh Ojha论文中，特征提取器选定为ViT-L/14，分类器选择为KNN和一个简单的线性探测。但是通过上文的图10可以看出，虽然这种模型相比于传统的鉴别模型泛化能力得到了很大的提升，但是在面对赛题任务的时候，仍然无法达到很高的鉴别精度。

虽然我们的省赛模型（CLIP:ViT+MLP）在省赛上取得了较好的成绩，但是在面对国赛验证集的时候，却无法到达理想的效果，鉴于此，我们团队对省赛中表现优异的模型（CLIP:ViT+MLP）进行了深入分析，发现分类器性能已接近极限，因此我们开始探索更有效的特征提取解决方案。

通过对比测试多种预训练模型，我们最终确定ConvNeXt为此次竞赛最理想的特征提取器选择。ConvNeXt创造性地融合了传统卷积神经网络（CNN）与现代Transformer架构的优点，借助深度可分离卷积、多尺度特征提取机制以及LayerNorm和MLP层的设计，不仅实现了高效的特征表示，还在多个视觉基准任务上取得了卓越的成绩。其快速收敛、稳定训练及高效推理等特性，使得ConvNeXt成为了本次任务中超越ViT-L/14等其他模型的最佳选择。

为了进一步提升模型的准确率（ACC），我们在原有省赛模型基础上，构建了一个新的“预训练特征提取器+分类器”的体系结构，并针对性地优化了特征提取器与分类器的选择。最终我们的选择是**ConvNeXt+MLP**。

在**特征提取器**方面：我们选择了ConvNeXt作为特征提取模块，因为它在融合传统CNN与现代Transformer架构方面的独特优势，使其在特征提取能力上超越了单一的CNN或ViT模型。

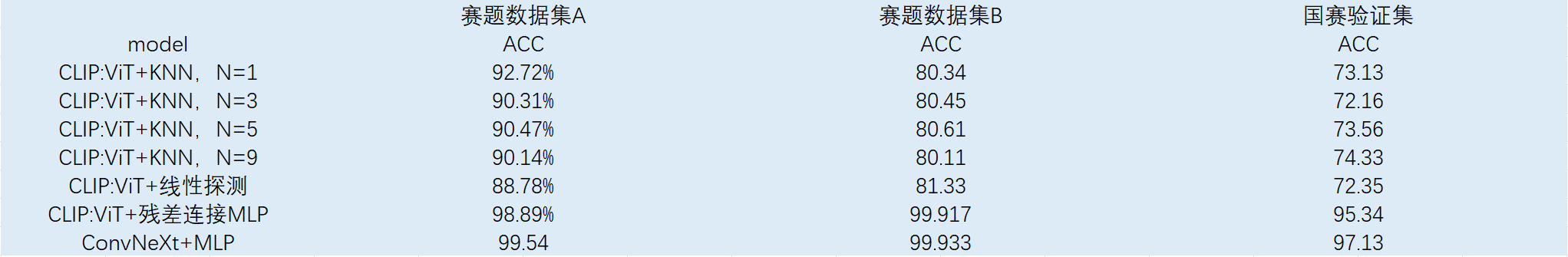
在**分类器**方面：为了改善分类器的表现，我们放弃了较为简单的KNN和线性探测方法，转而采用多层感知机（MLP）作为核心分类器，通过引入Dropout和L2正则化技术防止过拟合现象的发生。经过广泛的超参数调优，我们确定了MLP的最佳配置包括四层隐藏层，每层维度分别为2048、1024、512和256，Dropout率为0.3，学习率为0.001，L2正则化系数为0.001，批处理大小设为128，训练周期为20轮。最终，模型通过一个Sigmoid激活函数输出预测概率，并依据设定的阈值（0.5）作出最终判断，即概率值高于阈值视为AI生成图像，反之则认定为真实图像。该模型的整体架构如图13所示。

图表 11 整体结构图

4、创新点

**（1）结构上的创新**：基于对Utkarsh Ojha等人发表的《Towards Universal Fake Image Detectors that Generalize Across Generative Models》论文的深入总结与分析，我们进一步提出并验证了一种“预训练特征提取器+分类器”模式的有效性。通过广泛的实验对比，我们证明了该模式在处理未知数据源的AI生成图像鉴别任务中展现出显著优于传统鉴别模型的性能，特别是在准确率方面取得了显著提升。

**（2）模型上的创新：**尽管“预训练特征提取器+分类器”模式在总体上提供了比传统鉴别模型更好的性能，但对于特征提取器和分类器的具体选择仍需慎重考虑。针对本次竞赛的具体需求，Ojha等人论文中推荐的模型组合已不足以满足高精度鉴别的要求。为此，我们在特征提取器和分类器的选择与设计上进行了创新。具体而言，我们通过一系列实验验证了ConvNeXt作为特征提取器的优越性和创新性，同时自主研发了一种多层感知机（MLP）作为分类器，其在实验中显著超越了Ojha论文中提到的分类器性能。因此，我们提出了全新的“**ConvNeXt+MLP**”模型，这不仅体现了我们在模型架构上的创新，同时也展示了该模型在实际应用中的优越性能。

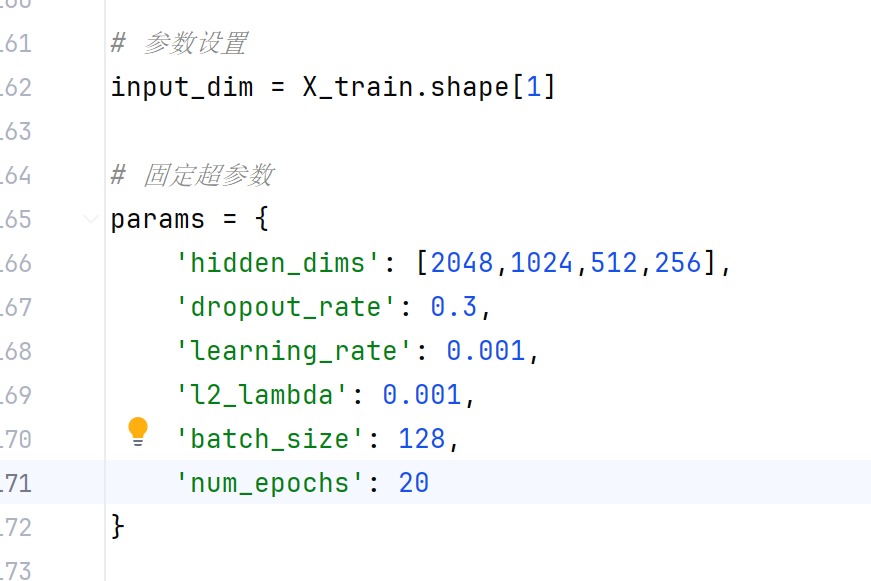


图表 12 模型性能对比

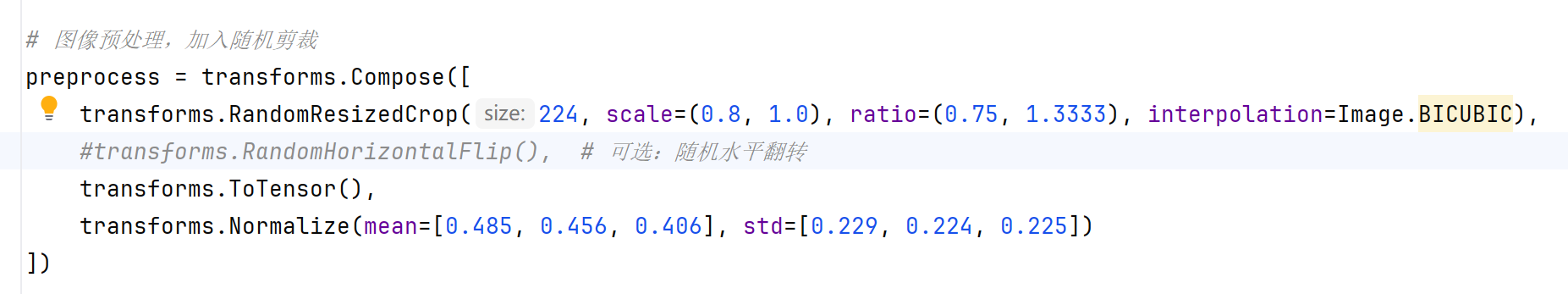
通过多次实验对比，我们发现“**ConvNeXt+MLP**”模型在各类赛题数据集上的鉴别性能显著优于论文中提及的两种分类方法（CLIP:VIT+knn和线性探测）以及我们在省级比赛中所使用的模型。如图14所示，“**ConvNeXt+MLP**”模型在所有测试数据集上的准确率（ACC）均显著高于上述方法，这不仅证明了我们提出的模型具有更强的泛化能力和更高的性能，同时也表明CLIP:VIT并非最优的特征提取器选择。我们的研究表明，ConvNeXt作为特征提取器时，能够提供更加丰富和有效的特征表示，这对于提高AI生成图像的鉴别准确性至关重要。综上所述，“**ConvNeXt+MLP**”模型不仅在性能上超越了现有方法，而且在创新性方面也具有明显优势。

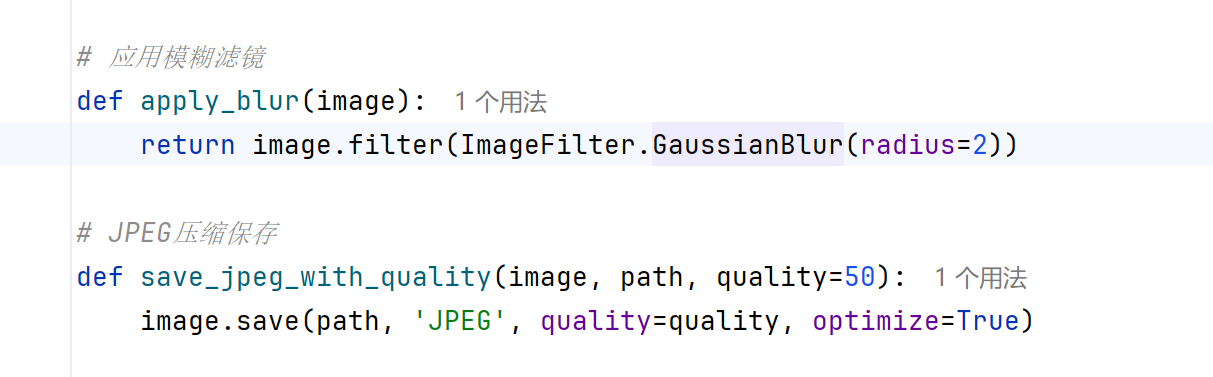
**三、训练参数和策略**

根据多次训练测试，选择的超参数为4层隐藏层，'hidden\_dims': [2048, 1024, 512, 256],'dropout\_rate': 0.3,'learning\_rate': 0.001,'l2\_lambda': 0.001,'batch\_size': 128,'num\_epochs': 20。通过dropout和L2正则化来防止过拟合，使用二元交叉熵函数作为损失函数。并通过使用随机剪裁、高斯模糊+jpeg压缩等数据增强技术处理训练数据。



图表 13 超参数设置





图表 14 数据增强