大作业 赖勋豪 2301213297

1. 模型介绍

本次大作业，使用了两种网络结构，卷积神经网络和视觉 Transformer。其中，卷积神经网络（Convolutional Neural Networks，CNN）主要特点为使用卷积核进行卷积操作以捕获原始数据的特征，视觉Transformer（Vision Transformer，ViT）是借鉴了自然语言处理中的 Transformer 架构，将输入图像切分为一系列的小图像块，通过自注意力操作来捕获信息。

为了训练的高效，选择了 NormFree-Net 作为训练的卷积网络；对于 Transformer，则选择了 patch size 为 8 的原始 Vision Transformer 结构。其中，NFnet 是一种没有 Normalization 层的模型，ViT 使用了论文中的原始版本。

2. 实验设置

实验在 CIFAR100 数据集上进行，利用 timm 库中的模型来初始化网络。使用的网络如下：

卷积神经网络：

```
model = timm.create_model(
            "dm_nfnet_f0.dm_in1k",
            pretrained=True,
            num_classes=100
        )
```

Vision Transformer：

```
model = timm.create_model(
            "vit_base_patch8_224.augreg_in21k_ft_in1k",
            pretrained=True,
            num_classes=100,
            img_size=(32, 32),
        )
```

训练默认使用 AdamW 优化器和 cosine 学习率衰减和线性学习率 warmup，使用 Batch size 为 256，学习率为 5e-4，权重衰减为 1e-2，训练 25 个 epochs。具体参数设置如下：

```
@dataclass
class TrainingArguments:
    seed: int = 3407
    device: str = "cuda"
    image_size: Tuple[int] = (32, 32)
    epochs: int = 25
    warmup_epochs: int = 5
    learning_rate: float = 5e-4
```

```
    weight_decay: float = 1e-2
    batch_size: int = 256
    checkpoint: str = None
    save_epochs: int = -1
    save_dir: str = "checkpoint"
    optimizer: str = "adamw"
    scheduler: str = "cosine"
```

在上述默认参数的基础上，实验分别对卷积网络和 ViT 进行了训练。此外，还探究了几种不同因素对模型训练的影响。

（1）将优化器更改为 SGDM

（2）将学习率调度更改为线性衰减

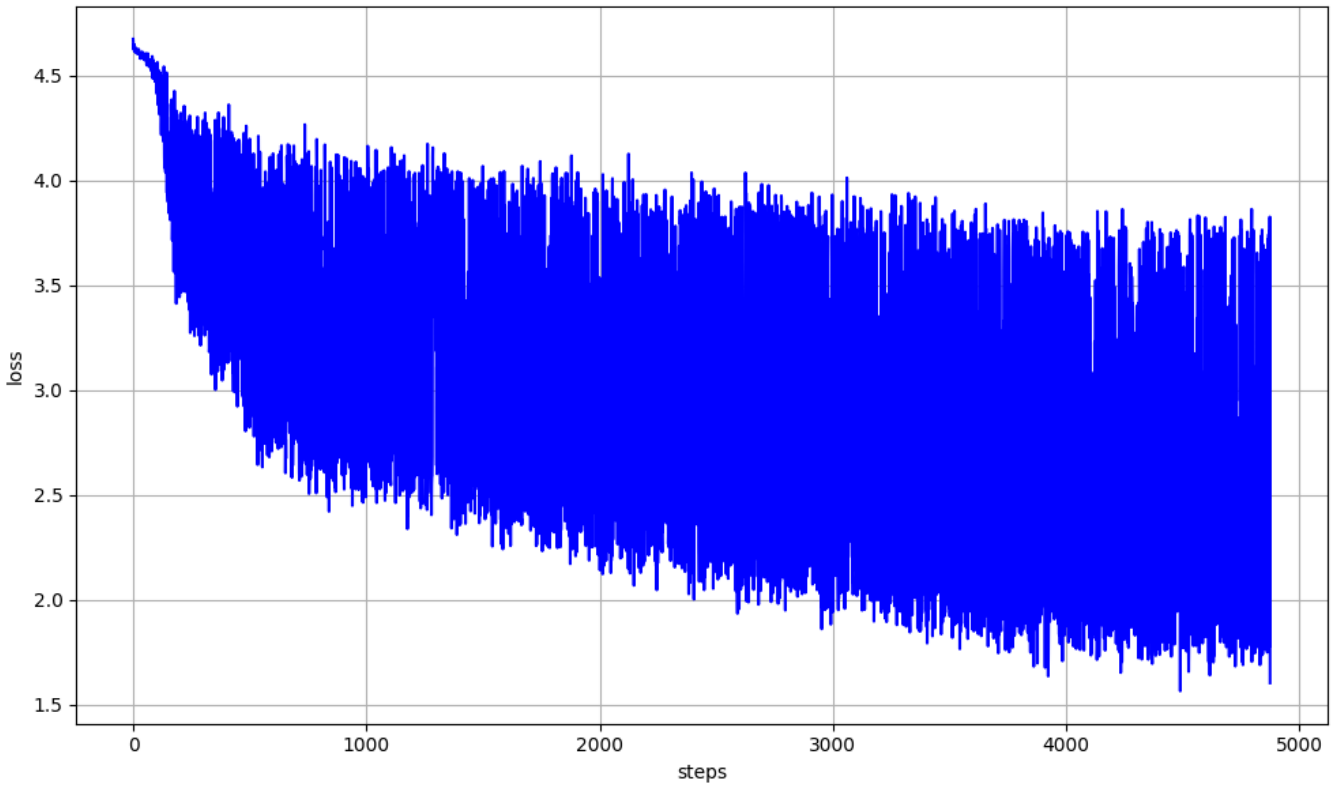
3. 实验结果

3.1 默认训练设置

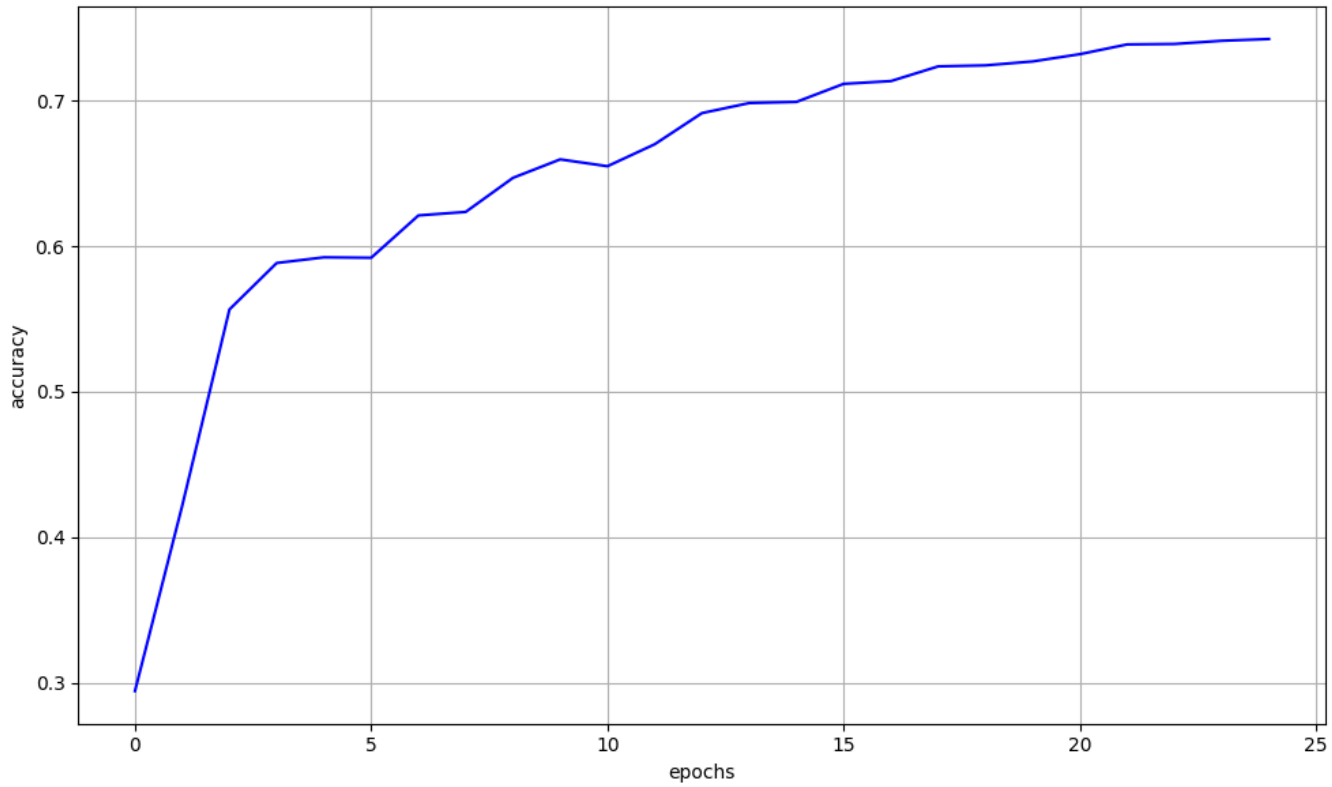
训练的最终损失和准确率如下表：

模型	训练损失	准确率(%)
NfNet	1.6041	74.23
ViT	1.7612	73.17

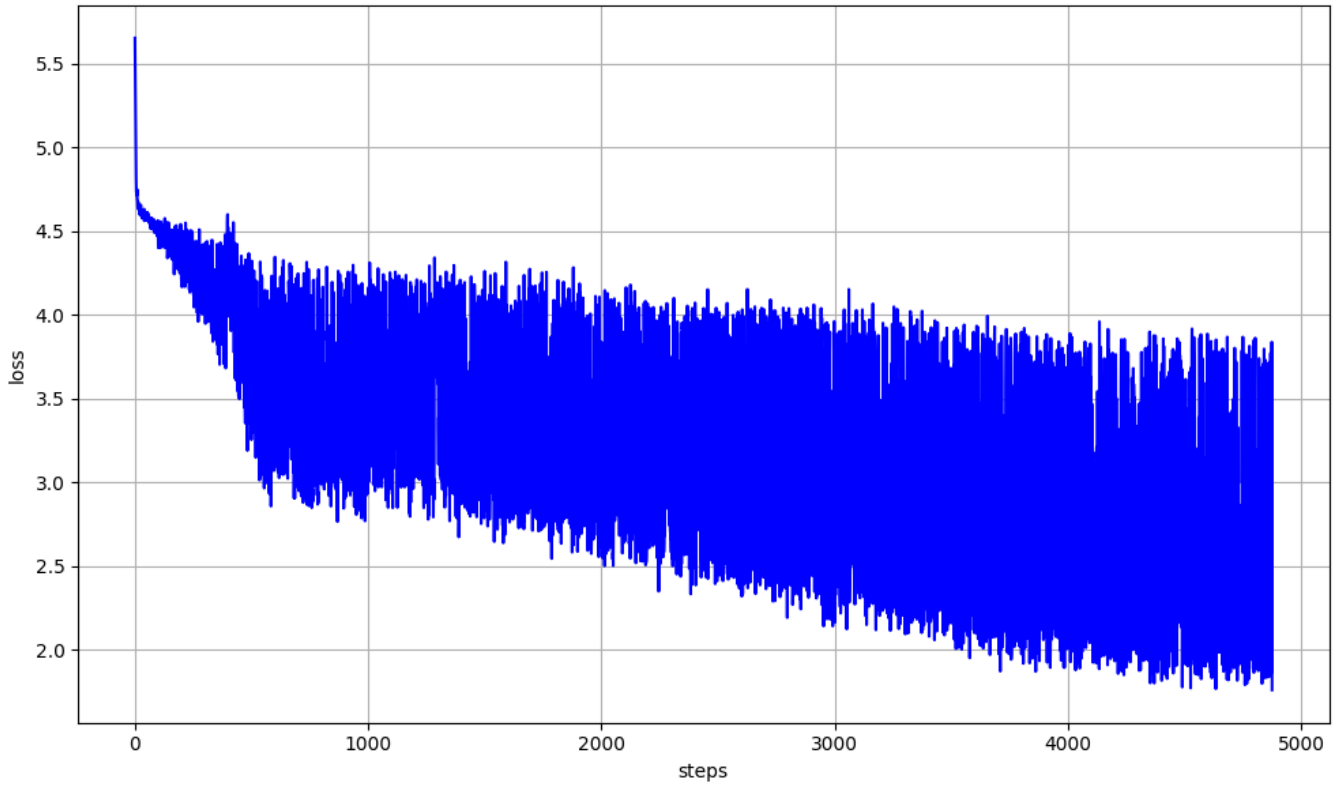
使用 NFnet 进行训练的训练损失变化如下：



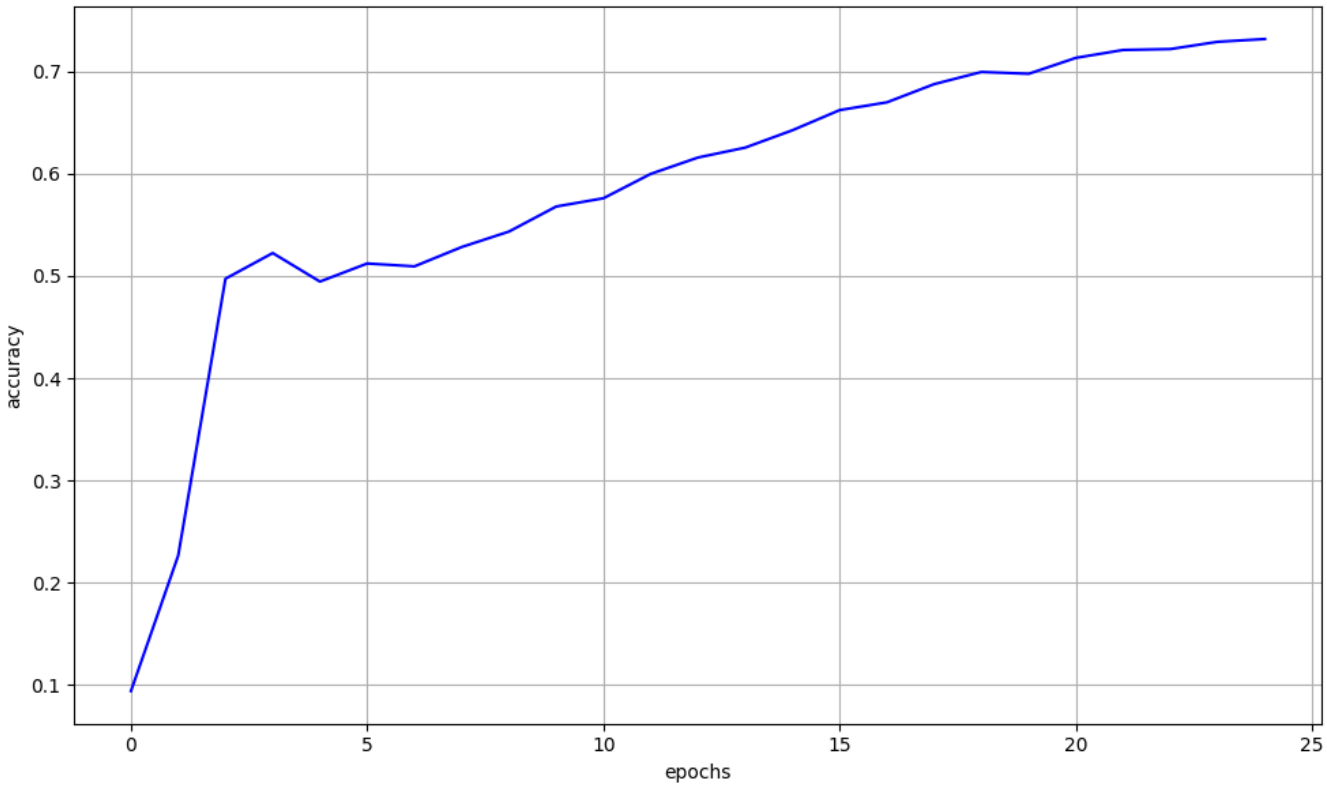
在验证集上的准确率随着训练的变化如下：



使用 ViT 进行训练的训练损失变化如下：



在验证集上的准确率随着训练的变化如下：

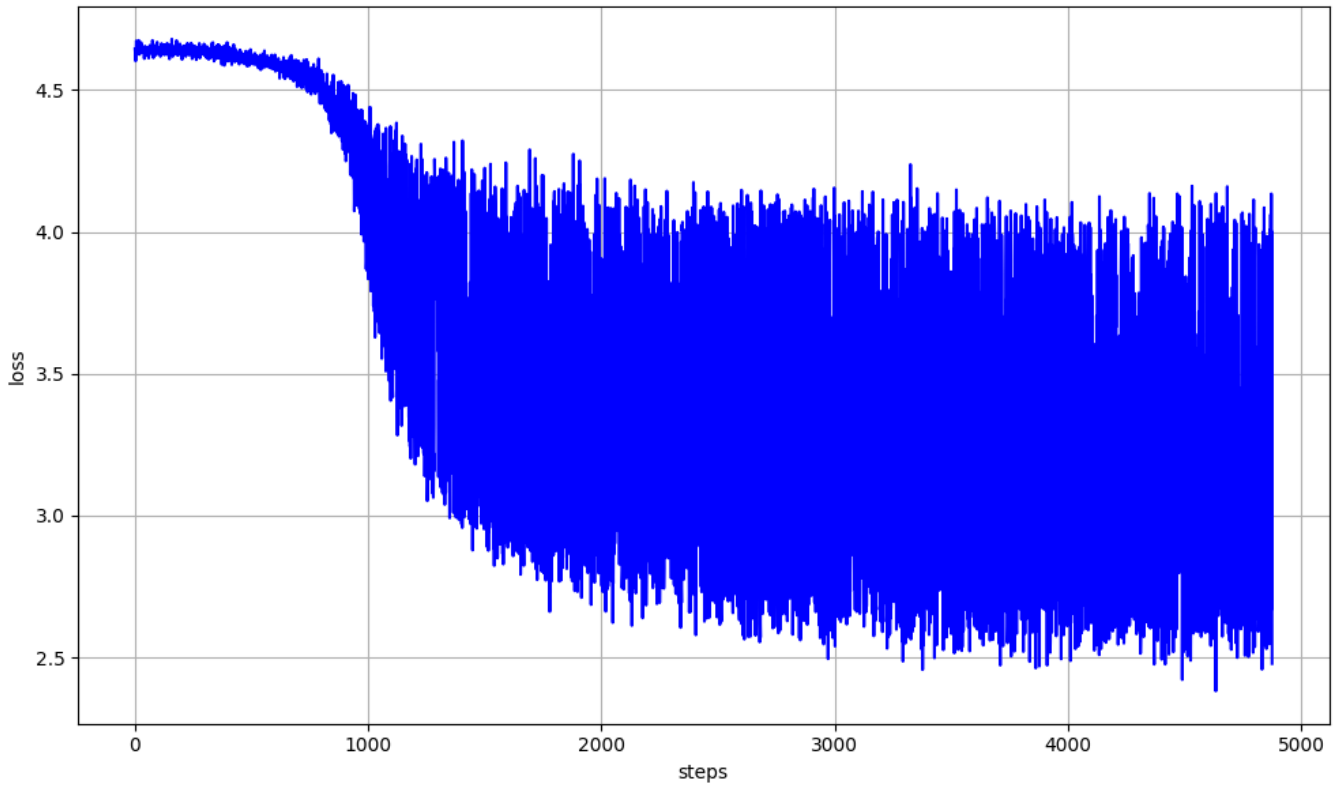


3.2 不同训练超参

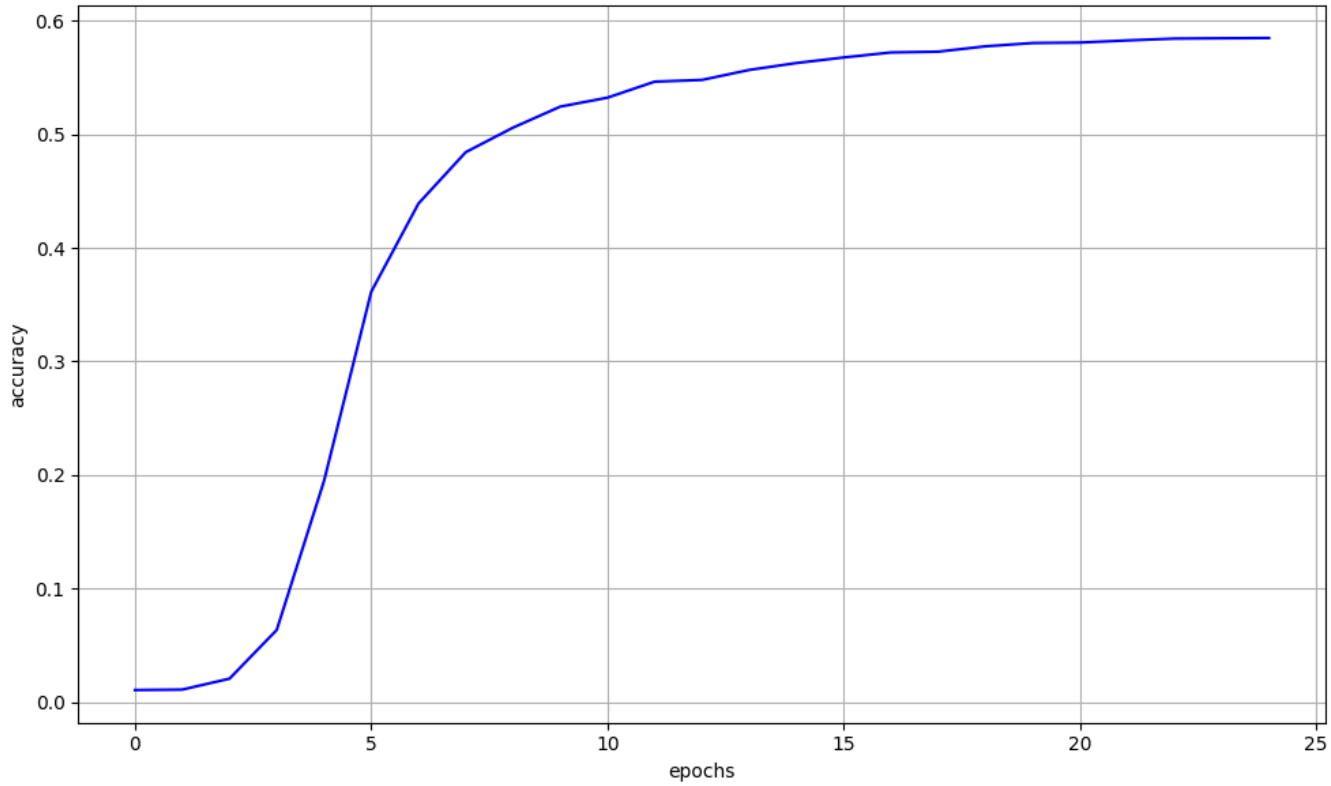
使用不同超参数训练的最终损失和准确率如下表：

模型	优化器	学习率调度	训练损失	准确率(%)
NfNet	AdamW	warmup+cosine	1.6041	74.23
NfNet	SGDM	warmup+cosine	2.4778	58.47
NfNet	AdamW	linear_decay	1.6641	73.36

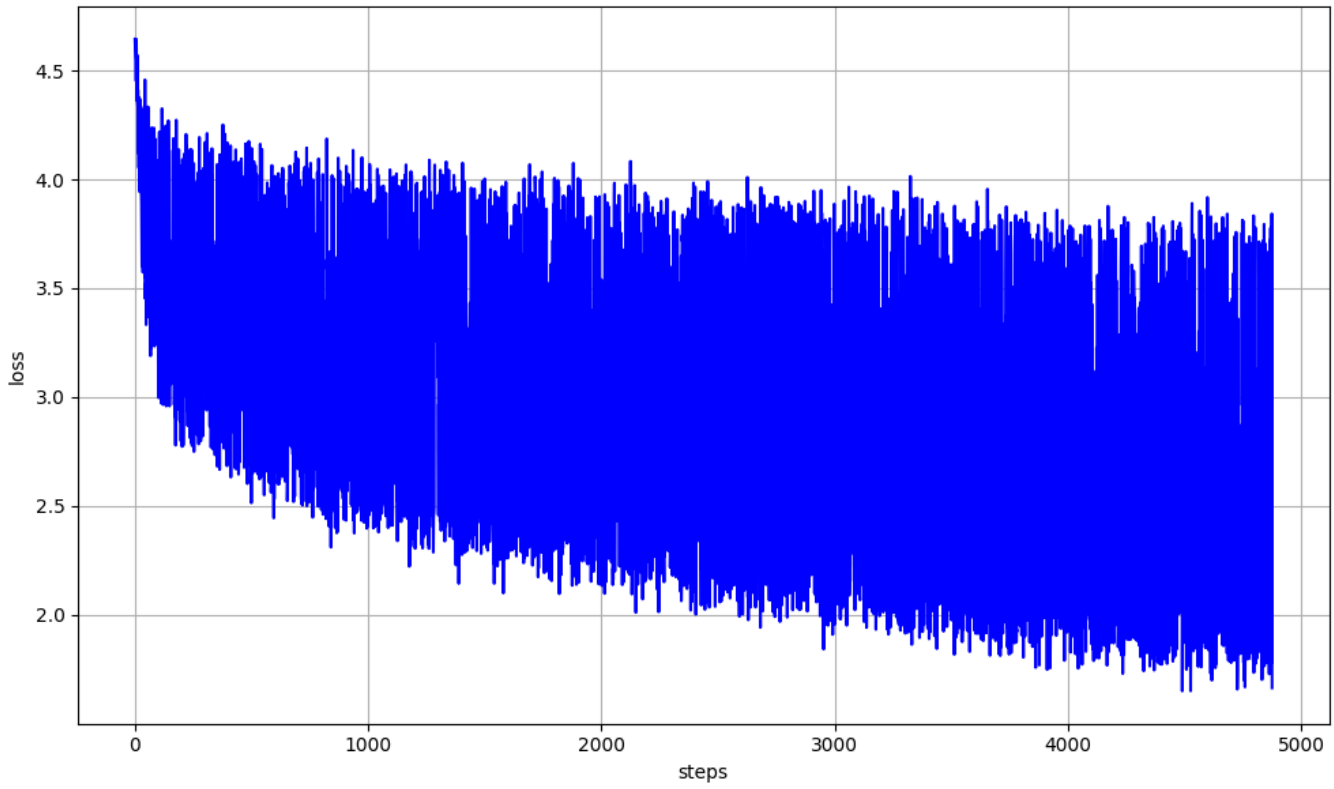
使用 NFnet + SGDM 进行训练的训练损失变化如下：



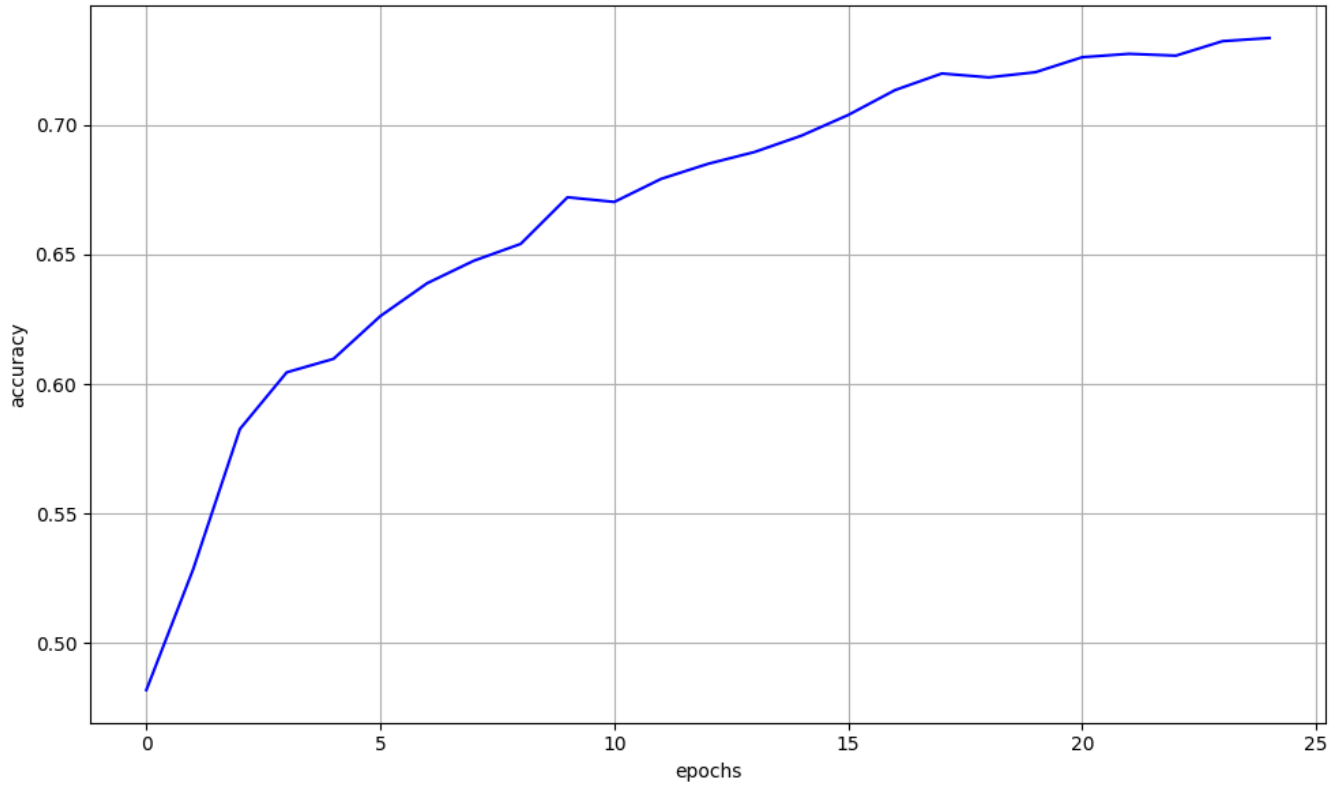
在验证集上的准确率随着训练的变化如下：



使用 NFnet + linear_decay 进行训练的训练损失变化如下：



在验证集上的准确率随着训练的变化如下：



3.3 总结

由于训练资源限制，只能进行少量 epochs 的训练，但在使用 ImageNet 上训练过的模型权重作为初始化时，依然能得到不错的结果。最终最高准确率达到了 74.23%。然而，因为 ImageNet 上训练的模型是以 224x224 或更大图像大小设计的，在直接迁移到 32x32 大小的 CIFAR100 数据集上，无法发挥出最好的效果。

对比 ViT 和 CNN，由于 CNN 存在天生的归纳偏置，更加容易收敛，最终的学习效果也略好与 ViT。但是在更大的数据集和更长的训练下，ViT 依然有着更好的可扩展性。

此外，对优化器和学习率调度进行的实验发现，SGDM 优化器相比 AdamW 优化器收敛更慢，在同样的 epochs 下难以得到更好的效果。学习率的 linear_decay 相比 warmup+cosine_decay，效果略差，但区别不大。

4. 附加评分点

（1）使用 timm 的模型
（2）使用 timm 的数据增强
（3）使用 timm 的 Loss
（4）探索 Optimizer 对训练的影响
（5）探索 Scheduler 对训练的影响
（6）训练了 CNN 和 ViT 并进行比较