

Deep Learning Practical Tutorial - Github Project Reproduction and Application

School of Optics and Photonics, BIT, FuYu

目录

1	预备知识	3
1.1	参考教程	3
1.2	主要过程	3
1.2.1	数据	4
1.2.2	模型	6
1.2.3	训练	9
1.2.4	可视化	10
2	图像分类	11
3	视频动作识别	11
4	实战目标	11
5	总结	12
A	过程记录	13
A.1	数据处理	13
A.2	确认模型	14
A.3	训练设置	16
A.4	训练	16

摘要

该文档总结了数个已经复现的深度学习项目，包括代码实现、数据集、训练细节等内容，旨在为大家提供一个全面的参考资源，帮助其更好地理解和应用深度学习技术。

1 预备知识

1.1 参考教程

此处我们只讲解代码实现层面上的基本知识，更多理论知识请参考相关教材和论文。在这里我们给出几个可供参考的学习资料：

- [Dive into Deep Learning中文版](#)
- [CS231n课程笔记斯坦福大学计算机视觉课程](#)
- [小土堆教程：PyTorch深度学习快速入门教程](#)

如果是初学者，建议先从基本知识入手，理解神经网络的基本概念、常用的网络结构（如卷积神经网络、循环神经网络等）、损失函数和优化算法等内容。掌握这些基础知识后，再逐步深入到具体的项目实现中。参考Dive into Deep Learning和CS231n（CS231n的课后作业可不做，可以参考本文后续的实际操作）。当然不想看电子书的话，可以搜索在B站李宏毅、吴恩达等老师的教程。最后，小土堆教程给出了更加细致的代码实现细节，讲解了在实际写代码时遇到的问题和解决方法，例如张量大小的计算、如何修改模块等。我建议小土堆的教程可以多研究一下，特别是**Dataset**、**Dataloader**、**现有模型的使用和修改**等章节，值得参考。同时，提供之前完成的各种简单例程和代码片段，方便大家理解和实践。

在实际的实现过程中，我们多多少少会遇到一些奇怪的问题和无从下手的情况，下面我将通过该文档展示2025年复现的项目，逐步总结一些常见的问题和解决思路，供大家参考。

1.2 主要过程

在初步掌握上述知识后，我们可以选择一个感兴趣的项目进行复现。建议从简单的项目开始，例如图像分类或文本分类等任务。通过阅读相关


```
train_dataset = MyDataSet(images_path=train_images_path,
                           images_class=train_images_label,
                           transform=data_transform["train"])
```

这里我们定义了一个数据增强的transform，通过一系列操作将图像进行预处理，再将其传入自定义的MyDataSet类中，完成数据集的构建。无论是图像分类、目标检测还是语义分割等任务，数据集的构建过程都是类似的，只是具体的数据增强方法和数据集格式可能有所不同。有的代码是直接train.py文件中直接定义并且使用DataLoader加载数据集的，也有代码是将其归为一个单独的dataset.py文件中进行定义，在我们下面调用Dataset方法时候直接使用，这就是不同项目的代码风格不同而已，逻辑是一样的。

注意，我们自定义的MyDataSet类需要继承torch.utils.data.Dataset，并实现__len__和__getitem__方法，以能被DataLoader使用。这里getitem输出数据集中每一个数据的具体内容和它的下标，len输出数据集的长度（这里不理解的话可以参考前面给出的小土堆中的相关讲解）。这个类是随着每个项目的不同而不同的，我们需要根据具体的数据集来编写相应的Dataset类。我们这里给出一个简单的例子：

```
class MyDataSet(Dataset):
    def __init__(self, images_path: list, images_class: list, transform
                 =None):
        self.images_path = images_path
        self.images_class = images_class
        self.transform = transform

    def __len__(self):
        return len(self.images_path)

    def __getitem__(self, item):
        img = Image.open(self.images_path[item])
        if img.mode != 'RGB':
            raise ValueError("image: {} isn't RGB mode.".format(self.
                                                                    images_path[item]))
        label = self.images_class[item]

        if self.transform is not None:
            img = self.transform(img)

        return img, label
```

这里定义了MyDataSet类，用于加载图像数据集。该类接受图像路径列表

和对应的标签列表，并在`getitem`方法中读取图像文件，应用预处理变换，并返回图像和标签。有的数据增强或预处理是写入到这个方法中的，具体可以参考下面讲述的视频动作识别部分的实现，他的`VideoDataset`类中就包含了`crop`方法来进行数据增强。接下来我们只需要使用`DataLoader`来加载数据集即可：

```
train_loader = torch.utils.data.DataLoader(train_dataset,
                                           batch_size=batch_size,
                                           shuffle=True,
                                           pin_memory=True,
                                           num_workers=nw,
                                           collate_fn=train_dataset.collate_fn)
```

这里我们使用`DataLoader`来加载训练数据集，设置了批量大小、是否打乱数据、使用的线程数等参数。`collate_fn`参数用于指定如何将多个样本合并成一个批次，这个函数可以根据具体需求进行自定义。此时，我们就已经完成了数据集的构建和加载，在后续的过程中我们就可以使用创建好的训练集、验证集、测试集对整个模型进行训练。

1.2.2 模型

构建网络模型是深度学习项目中的核心部分。从代码层面上讲，一般的项目会有`network`、`model`、`models`等文件或文件夹来存放模型相关的代码。我们需要根据具体的任务选择合适的模型架构，并在代码中实现该模型。那么我们给出一个最为简单的代码进行说明：

```
import torch
import torch.nn as nn

class SimpleNet(nn.Module):
    def __init__(self):
        super(SimpleNet, self).__init__()
        self.fc1 = nn.Linear(784, 128)
        self.relu = nn.ReLU()
        self.fc2 = nn.Linear(128, 10)

    def forward(self, x):
        x = x.view(-1, 784)
        x = self.fc1(x)
        x = self.relu(x)
        x = self.fc2(x)
        return x
```

```
model = SimpleNet()
```

这里我们只需要说两个地方：

1、继承nn.Module类：在我们定义神经网络模型时，我们需要创建一个类并继承nn.Module类，以便能够使用PyTorch提供的各种功能和方法。

2、定义前向传播方法：在模型类中，我们需要实现forward方法，定义数据在网络中的前向传播过程。这是模型的核心部分，决定了输入数据如何经过各个层次进行计算并输出结果。这个一定是需要根据项目自身根据自己的输出要求来定义的。

在我们实际的代码实现中，模型的定义可能会更加复杂，包含多个层次、不同类型的层（如卷积层、池化层、全连接层等）以及各种激活函数和正则化方法，有的项目代码甚至会给出多层嵌套，多个文件来回调用，虽然减小了代码维护的难度，但是给我们带来了很大理解难度。但是我们最终关心的往往是网络的中间层的形状和最终输出，我们可以通过自定义测试代码来确定：

```
if __name__ == "__main__":
    inputs = torch.rand(1, 3, 16, 112, 112)
    net = C3D(num_classes=101, pretrained=True)

    outputs = net.forward(inputs)
    print(outputs.size())
```

这里我们以后续说明的视频动作识别中的C3D模型为例，创建了一个随机输入张量，并通过模型的forward方法进行前向传播，最后打印输出的形状。通过这种方式，我们可以验证模型的定义是否正确，并了解各层之间的数据流动情况。

如果我们正处在模型的更改阶段，我们可以通过打印每一层的输出形状来帮助我们理解模型的结构和数据流动情况（当然最直接的方法是对代码进行调试，具体的方法后续有机会补充）。例如：

```
def forward(self, x):
    print("Input shape:", x.shape)
    x = self.conv1(x)
    print("After conv1:", x.shape)
    x = self.pool1(x)
    print("After pool1:", x.shape)
    return x
```

这里我们在forward方法中添加了打印语句，输出每一层操作后的张量形状。通过这种方式，我们可以清晰地看到数据在网络中的变化，帮助我们理解模型的结构和调试潜在的问题。当然，我们在运行代码的时候，如果我们因为加入了某些模块导致模型无法正常运行，会发生某两个向量形状不匹配的情况（有时候报错信息也会直接告诉我们两个向量形状是多少，方便我们定位），我们也可以通过这种打印形状的方式来定位问题所在。到此如果对向量的形状计算不理解的话，可以参考前面给出的小土堆中的相关讲解，也可以参考[这个视频](#)，了解具体特征图张量的计算方法。那么在我们进行模型修改的时候，我们就可以通过这种方式来帮助我们理解模型的结构和数据流动情况，从而更好地进行模型的修改和优化。

在我们定义好模型后，我们在实例化模型时，还需要考虑模型的初始化、加载预训练权重等问题，这些内容在不同的项目中可能会有所不同，我们需要根据具体的项目要求来进行相应的操作。这里我们给出一个示例：

```
def create_model(num_classes: int = 21843, has_logits: bool = True)
    :
    model = VisionTransformer(img_size=224,
                              patch_size=16,
                              embed_dim=768,
                              depth=12,
                              num_heads=12,
                              representation_size=768 if has_logits
                                                                else
                                                                None,
                              num_classes=num_classes)

    return model

model = create_model(num_classes=args.num_classes, has_logits=False)
                    ).to(device)
```

这里我们定义了一个create_model函数，用于创建Vision Transformer模型，并根据传入的参数设置模型的类别数和是否包含logits层。然后我们实例化模型并将其移动到指定的设备上（如GPU）。在实际的项目中，我们可能还需要加载预训练权重，这可以通过torch.load()函数来实现，具体的加载方式会根据预训练权重的格式和存储位置有所不同。这个实例就是稍微复杂的案例，将模型的创建封装在一个函数中，方便我们根据不同的需求来创建模型实例。我们需要注意的点就是：加载哪个定义好的模型、是否需要加载预训练权重、模型的输入输出设定等，一般这个会在readme中有说明，如果没有相关说明，或者GitHub项目给出有给出预训练权重下载地址，我们

就需要果断放弃该项目，去找其他的项目。最后还需要注意的点是否所有参数都需要训练、是否需要冻结某些层，这些内容会根据具体的项目要求有所不同，我们需要根据实际情况来进行相应的操作。

1.2.3 训练

在我们完成数据集的构建和模型的定义后，接下来就是训练模型的过程了。首先，我们需要一些必要的定义，这些内容在不同的项目中可能会有所不同。这里我们给出一个简单的示例：

```
criterion = nn.CrossEntropyLoss()
optimizer = optim.SGD(train_params, lr=lr, momentum=0.9,
                      weight_decay=5e-4)
scheduler = optim.lr_scheduler.StepLR(optimizer, step_size=10,
                                       gamma=0.1)
```

这里我们定义了损失函数、优化器和学习率调度器。损失函数用于衡量模型预测结果与真实标签之间的差距，优化器用于更新模型参数以最小化损失函数，学习率调度器用于调整学习率以提高训练效果。在这里我们需要注意的点是，optimizer可能有多个不同的优化器可供选择，例如Adam、RMSprop等，如果我们自行搭建了网络模型或者加入了自己的模块，大多数情况可以不改变原有代码的优化器，但是小部分情况我们需要根据具体的任务和模型选择合适的优化器，可以自行查阅不同的优化器之间的差别。另外，optimizer的参数同样需要注意，我们在这里给他传递我们需要训练的模型的参数，学习率等，一般会在训练模型的方法或者调用文件的传参对这些参数进行设置，可以参考项目中已给出的文件说明（readme等文件）。同样地，损失函数也有很多种类，例如均方误差、交叉熵等，我们需要根据具体的任务选择合适的损失函数。最后，学习率调度器也是一个重要的环节，可以帮助我们更好地控制学习率的变化，提高训练效果。不同的学习率调度器有不同的调整策略，我们需要根据具体的任务和模型选择合适的学习率调度器。这里可以自行查阅相关资料了解不同的学习率调度器之间的差别和使用方法。

完成了这些步骤，我们就要开始写循环进行网络的训练，下面我们给出最主要的训练代码：

```
scheduler.step()
model.train()
optimizer.zero_grad()
```

```
outputs = model(inputs)
probs = nn.Softmax(dim=1)(outputs)
preds = torch.max(probs, 1)[1]
loss = criterion(outputs, labels)
loss.backward()
optimizer.step()
```

这里我们首先调用`scheduler.step()`来更新学习率，然后将模型设置为训练模式。接着，我们清零优化器的梯度，进行前向传播计算模型输出，并通过损失函数计算损失值。然后，我们进行反向传播计算梯度，最后调用优化器的`step()`方法来更新模型参数。在实际的训练过程中，我们通常会将上述代码放在一个循环中，遍历整个训练数据集多次，以不断优化模型参数，提高模型性能。在实际使用过程中，无论是有的项目将训练过程写成类似`train_one_epoch()`的方法，还是直接写在`train.py`文件中进行训练，逻辑是一样的，我们都需要经历上述的过程。在具体的实现过程中，可能有一些细节需要注意，例如批量大小、迭代次数等，这些内容会根据具体的项目要求有所不同，我们需要根据实际情况来进行相应的操作。

1.2.4 可视化

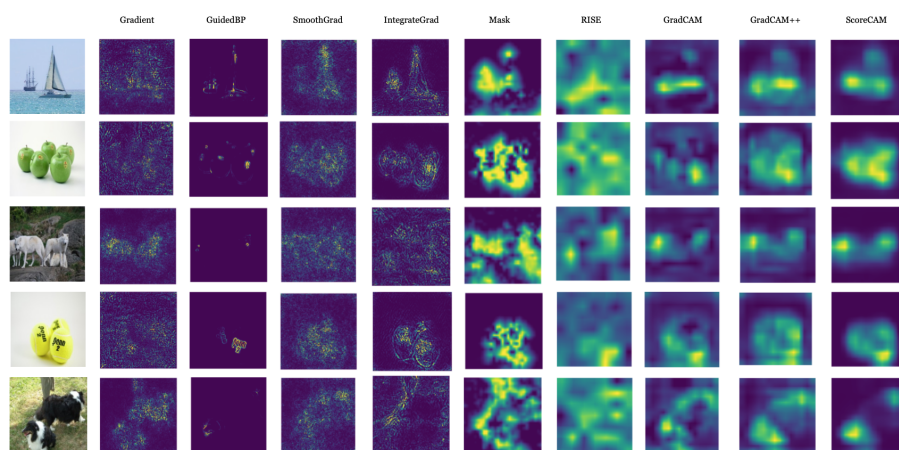


图 1: 不同可视化方法效果示例

可视化的部分在不同的项目中可能会有所不同，我们需要根据具体的项目要求来进行相应的操作。这里，我给出一个已经复现成功的[可视化示例](#)，访问这个链接下载GitHub项目，然后按照readme中的说明进行操作即

可。里面已经包含了几个较为简单的网络模型的可视化实现细节，大家可以参考学习。

我们这里的可视化一般用于对网络模型的注意力区域进行分析，帮助我们理解模型的决策过程。通过可视化，我们可以看到模型在处理输入数据时关注的区域，从而更好地理解模型的行为和性能。在实际的项目中，我们可能需要对模型的中间层的输出进行可视化，帮助我们理解模型的决策过程，如何实现可以参考每个可视化工具的具体实现和使用方法。有时我们需要根据具体的项目要求来选择合适的可视化方法和工具，并进行相关的操作。

2 图像分类

在编辑该章节时，我们已经成功复现[github项目](#)中有关ViT的部分。

这里我们就不再过多讲述上面已经提到的各种实现细节，我们在这里讨论在实现过程中遇到的一些问题和解决思路。

3 视频动作识别

在编辑该章节时，我们已经成功复现[github项目](#),成功实现了包括并不限于C3D模型在UCF-101和HMDB-51数据集上的动作识别任务。

这里我们就不再过多讲述上面已经提到的各种实现细节，我们在这里讨论在实现过程中遇到的一些问题和解决思路。

4 实战目标

这里给大家提供一个可行的实战目标，供大家参考和实践。

首先，复现上面提到的图像分类的项目。为了确保可复现性，大家访问上面给出的[github链接](#)并且找到pytorch_classification_vision_transformer,这里给出了ViT的代码（其他的网络模型大家可以在完成ViT模型的复现后根据自己的目标或者兴趣自行进行复现，由于我只复现了ViT模型，所以真敢保证这个模型的可行性，大家在判断其他模型是否可行时最简单的方法就是看流程是否完整，包括数据集、训练超参数等，另外还需观察是否给出模型预训练链接，都有的话是很有保障复现的）。

成功复现后再尝试对模型进行修改，可以使用上面2.1提到的模块，也可以自行添加其他的模块或网络层。观察添加前后的模型性能变化，提升或下降都无所谓，修改完模型能收敛就是成功。最后，可以尝试将修改后的模型应用到其他图像分类数据集上，例如CIFAR-10、CIFAR-100等，观察模型在不同数据集上的表现。

到此，我们就完成了一个完整的实战项目，从复现代码到修改模型再到应用新数据集，全面提升了对深度学习项目的理解和实践能力。大家可以根据自己的课题和自己的兴趣，自行选择其他的项目进行复现和实践，过程大同小异，关键是要理解每一步的细节和背后的原理。

如果在看了这一部分仍有困难，请参考附录中的“过程记录”进行实际操作学习。

5 总结

如果大家能成功复现代码，距离成功已经完成了65%的工作，那么剩下的35%才是：理解代码、分析结果、改进模型。在后续过程中，我会在网站上持续更新该文档，希望大家能有所收获。祝科研顺利！

A 过程记录

这里我们带来一个简单的过程记录，供大家参考，具体的训练代码我上传到了[funcCodeExample](#)文件夹中，可以边查看trainvimucf边学习，并且和原来的train.py文件进行比较，可以更好的理解代码的运行流程。

我们这里实现的是将ViedoMamba应用在这里的[视频动作识别项目](#)中，因为我们已经完成了对该项目的复现，所以我们仅需要修改模型部分的代码即可。请大家确保学习该部分时已经可以复现上述的视频动作识别项目。

A.1 数据处理

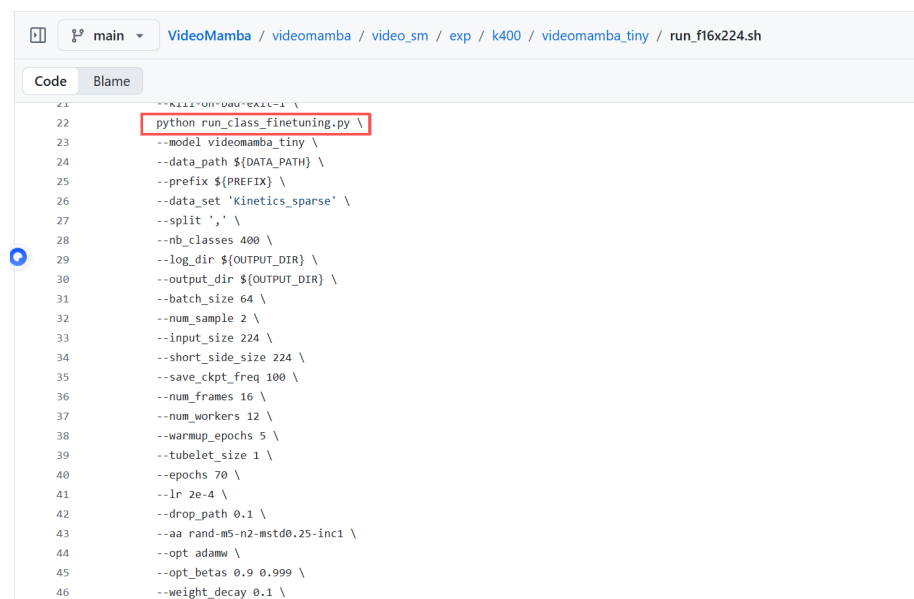
由于我们已经能跑通了这个[视频动作识别项目](#)，所以数据处理部分我们是已经实现了的。由于我们是调用VideoMamba的模型架构代码，所以我们需要确保数据处理部分和VideoMamba的要求是一致的。我们可以找到图2的说明，我们点开script指向了一个微调的启动命令，是一个sh命令，我们可以找到这个文件中对数据处理的超参数input_size,这里就是用的224，当时回到我们的视频动作识别项目中，找到数据处理部分，我们发现它的数据大小并不是这个值，并且他还做了crop用于数据增强，那么我们现在就用两种处理思路：1、直接将数据处理部分改为VideoMamba的要求，这样就能保证数据处理的一致性，2、重新生成视频帧，这样就需要重新跑数据处理代码生成新的数据集。这里我选择了第一种方法，直接将数据处理部分改为VideoMamba的要求，这样省得再跑一遍数据生成（当然，这里是考虑到视频数据集比较大，如果新生成数据集比较简单，我建议重新生成数据集，这样可以减少对代码的改动）。通过阅读代码，我们将resize_height resize_width crop_size改成256、344、224（这样改的原理是由于我们最终需要进行图像增强，所以crop之后的图像大小一定是224×224，我们将原图像使用插值放大到256×344尺寸后增加数据集的鲁棒性），之后，我们将DataLoader的传参改成sh文件中的参数，这样就能保证数据处理的一贯性，到此，我们就完成了新数据集的构建（可以参考[funcCodeExample](#)文件夹中的datasetvimucf文件参考学习）。

这里的需要抓住的核心就是数据集的传输和模型想要的输入一致，找到相关代码确认即可。

A.2 确认模型

- 🔥 2024/03/12: All the code and models are released.
 - [image_sm](#): Single-modality Image Tasks
 - Image Classification: [script](#) & [model](#)
 - [video_sm](#): Single-modality Video Tasks
 - Short-term Video Understanding: [script](#) & [model](#)
 - Long-term Video Understanding: [script](#) & [model](#)
 - Masked Modeling: [script](#), [model](#)
 - [video_mm](#): Multi-modality Video Tasks
 - Video-Text Retrieval: [script](#) & [model](#)

图 2: readme截图



```
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46
"""ALL-ON-ONE-CALL"""
python run_class_finetuning.py \
--model videomamba_tiny \
--data_path ${DATA_PATH} \
--prefix ${PREFIX} \
--data_set 'Kinetics_sparse' \
--split ',' \
--nb_classes 400 \
--log_dir ${OUTPUT_DIR} \
--output_dir ${OUTPUT_DIR} \
--batch_size 64 \
--num_sample 2 \
--input_size 224 \
--short_side_size 224 \
--save_ckpt_freq 100 \
--num_frames 16 \
--num_workers 12 \
--warmup_epochs 5 \
--tubelet_size 1 \
--epochs 70 \
--lr 2e-4 \
--drop_path 0.1 \
--aa rand-m5-n2-mstd0.25-inc1 \
--opt adamw \
--opt_betas 0.9 0.999 \
--weight_decay 0.1 \
```

图 3: 调用文件截图

第一步我们要做的就是确认哪些是VideoMamba的模型架构代码，并且将涉及到的所有文件都下载下来。首先我们需要看的就是readme文件中是否有明显的指向运行模型文件代码的说明：同样的，我们可以找到图2的说明，我们点开script指向了一个微调的启动命令，里面的文件一定会包含一个调用的python文件，如图3所示，还有其他必要的超参数，包括视频帧数、数据集类别数、输入数据尺寸、学习率等，我们后续还需要使用这些参数

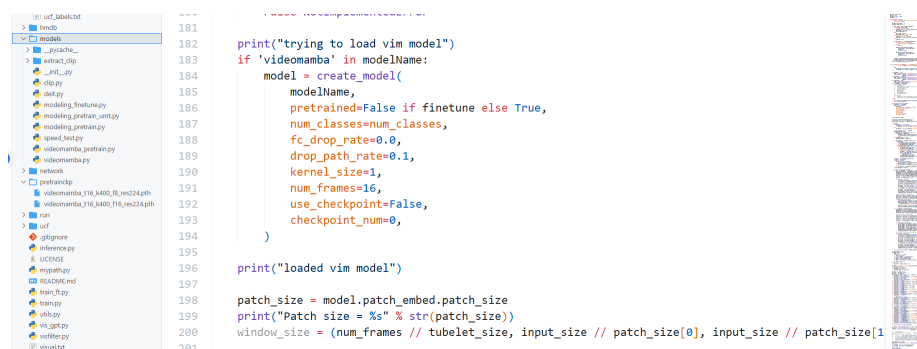


图 4: 复制后截图

进行模型搭建。为了方便理解，我们尽量少的复制文件，将一些必要的文件和参数进行整合即可。我们进入这个run_class_finetuning.py文件，找到构建模型的部分（参考上面1.2的讲解），我们发现它使用了create_model()方法来创建模型，这里使用的是timm库中的方法，这里我们直接将models文件夹进行复制到我们需要加入的项目代码之中，如图4所示，然后同样使用create_model()方法即可。复制过去之后，我们需要考虑的事情就是超参数的设置，我们需要根据上面的sh文件中的参数进行设置，包括类别数、输入尺寸等，这里我使用了一个非常偷懒的方法，不建议学习，但是为了展示过程我还是给出，如图5所示：我们直接对照sh文件将里面的参数直接复制到我们需要调用的方法中即可，这样就避免了传参的问题。当然，这种偷懒的方法是极其不利于维护和更新的，强烈建议自行书写argparse进行传参，有的集成好的函数就在主程序中调用函数时进行方法传参即可。这里可以参考主页中关于这部分传参方法的使用。完成了这些步骤，我们下一步需要完成的是模型的预训练模型的加载。同样的，这里我们需要去看sh文件和run_class_finetuning.py文件，从这两个文件找到线索。我们需要确定的是预训练模型的下载地址和加载方法。我们可以在readme找到预训练模型的下载地址，并且在run_class_finetuning.py文件中的337行开始，到503行，是加载预训练模型的代码，因为第503行出现了我们在1.2中提到的加载模型的代码load_state_dict（如果对这些命令不敏感的话，可以去funcCodeExample文件夹找到里面的trainvimucf代码对这些相关的方法函数进行学习），别看中间100多行代码，如果让我去讲解的话也道不清其中的过程，但我们只需要抓住重点实现代码，把中间这些一块复制过去就好。到此，我们就完成了模型的构建。



图 5: 复制后截图

A.3 训练设置

走到这一步，我们就要去查看原来VideoMamba中对优化器等的是怎么做的了。通过前面的学习，我们知道我们需要确定模型哪些参数是参加训练的。通过查阅代码，我们可以找到一个关键的参数`n_parameters`、`optimizer`，以此为线索我们发现它使用了`create_optimizer`建立的`optimizer`，同样的我们复制这个方法，并且进入这个方法，为了直观我们直接复制方法内容到我们的代码中（建议复制方法并且引用方法，有较高的代码阅读性和可维护性），根据sh文件中的超参数设置优化器，并将原文`optimizer = optim.SGD`备注即可。

A.4 训练

开始训练。