model_multipleVideos-class_wise_Implementation-Vision_Transformer-large

April 6, 2022

[1]:
```
import torch
import torchvision
import torchvision.transforms as transforms
import torchvision.models as models
import torch.nn as nn
import torch.nn.functional as F
import torch.optim as optim
import time
from itertools import count
import natsort
import datetime
import numpy as np
import os
import math
```

[2]:
```
from torch.utils.data import Dataset, DataLoader, WeightedRandomSampler
import albumentations as A
from albumentations.pytorch import ToTensorV2
import cv2
import glob
import numpy
import random
import pandas as pd
import tqdm
torch.manual_seed(10)
```

[2]:
```
<torch._C.Generator at 0x25fd105f130>
```

[3]:
```
print(f"Is CUDA supported by this system? {torch.cuda.is_available()}")
print(f"CUDA version: {torch.version.cuda}")
# Storing ID of current CUDA device
cuda_id = torch.cuda.current_device()
print(f"ID of current CUDA device: {torch.cuda.current_device()}")
print(f"Name of current CUDA device: {torch.cuda.get_device_name(cuda_id)}")

device = torch.device('cuda:0' if torch.cuda.is_available() else 'cpu')
```

```
print(device)
```

```
Is CUDA supported by this system? True
CUDA version: 11.3
ID of current CUDA device: 0
Name of current CUDA device: NVIDIA GeForce RTX 2070 Super
cuda:0
```

1 Building the dataset

[4]:
```
class SurgicalDataset(Dataset):
    def __init__(self, image_paths, labels, transform=False):
        super(SurgicalDataset, self).__init__()
        self.image_paths = image_paths
        self.labels = labels    #.astype(dtype='int')
        self.transform = transform

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

    def __getitem__(self, idx):
        image_filepath = self.image_paths[idx]
        image = cv2.imread(image_filepath)
        label = self.labels[idx]
        if self.transform is not None:
            image = self.transform(image=image)["image"]

        return image, label
```

[5]:
```
def get_transform(model_name):

    if model_name == 'alexnet':
        transform = A.Compose([
            A.Resize(227, 227),
            A.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5)),
            ToTensorV2(),
        ])

    elif model_name == 'effinet':
        transform = A.Compose([
            A.Resize(224, 224),
            A.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5)),
            ToTensorV2(),
        ])

    elif model_name == 'TransferViT':
```

```
        transform = A.Compose([
            A.Resize(224, 224),
            A.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5)),
            ToTensorV2(),
        ])

    return transform
```

[6]:
```
# Preparing the datasets
# Get images
train_image_paths = []
train_data_path = r"C:
 ↪\Users\panji\EECS6691_Advanced_DL\Assignment2\training_data_images"
train_image_paths.append(glob.glob(train_data_path + '/*'))
# unpack the listed list
train_image_paths1 = [item for sublist in train_image_paths for item in sublist]
train_image_paths1 = natsort.natsorted(train_image_paths1)
print('len(train_image_paths1)', len(train_image_paths1))

# Get labels
df = pd.read_csv("Processed_data.csv")
df1 = df.loc[:,"Phases"].to_numpy()
df2 = df1.tolist()
print('len(df2)', len(df2))

# Preparing the datasets (images and labels)
dataset_train = pd.DataFrame(
    {'Link': train_image_paths1,
     'Label': df2,
    })
dataset_train1 = dataset_train.sample(frac=1, random_state=1)
train_image_paths = dataset_train1.loc[:,"Link"].to_numpy().tolist()
labels = dataset_train1.loc[:,"Label"].to_numpy().tolist()

# manually split the dataset
train_image_paths, valid_image_paths = train_image_paths[:int(0.
 ↪8*len(train_image_paths))], train_image_paths[int(0.
 ↪8*len(train_image_paths)):]
train_labels, valid_labels = labels[:int(0.8*len(labels))], labels[int(0.
 ↪8*len(labels)):]
print('train_labels', len(train_labels))
print('train_image_paths', len(train_image_paths))
print('label distribution in the training data', np.bincount(train_labels))
```

```
len(train_image_paths1) 215057
len(df2) 215057
train_labels 172045
```

```
train_image_paths 172045
label distribution in the training data [  243  8681 22901 41140   952 22305
666 10930   896  2308 44928 12987
  1789  1246    73]
```

2 Weighted Data Sampler

[7]:
```python
# from torch.utils.data import WeightedRandomSampler

# Get labels
df = pd.read_csv("Processed_data.csv")
df1 = df.loc[:,"Phases"].to_numpy()
df2 = df1.tolist()
print('len(df2)', len(df2))

# Preparing the datasets (images and labels)
dataset_train = pd.DataFrame(
    {'Link': train_image_paths1,
     'Label': df2,
    })
dataset_train1 = dataset_train.sample(frac=1, random_state=1)
train_image_paths = dataset_train1.loc[:,"Link"].to_numpy().tolist()
labels = dataset_train1.loc[:,"Label"].to_numpy().tolist()

summary = {i:0 for i in range(15)}
num_classes = 15
total_samples = 0
for i in train_labels:
    total_samples += 1
    summary[i] += 1
print(summary)
print(total_samples)

class_weights = [total_samples/summary[i] for i in range(num_classes)]
weights = [class_weights[train_labels[i]] for i in range(total_samples)]
sampler = WeightedRandomSampler(torch.DoubleTensor(weights), len(weights))
print(len(class_weights))
print(len(weights))
print(len(list(sampler)))
```

```
len(df2) 215057
{0: 243, 1: 8681, 2: 22901, 3: 41140, 4: 952, 5: 22305, 6: 666, 7: 10930, 8:
896, 9: 2308, 10: 44928, 11: 12987, 12: 1789, 13: 1246, 14: 73}
172045
15
172045
172045
```

3 Building the classifier class

[8]:
```
class Classifier():

    def __init__(self, name, model, dataloaders, parameter, use_cuda=False):

        '''
        @name: Experiment name. Will define stored results etc.
        @model: Any models
        @dataloaders: Dictionary with keys train, val and test and␣
↪corresponding dataloaders
        @class_names: list of classes, where the idx of class name corresponds␣
↪to the label used for it in the data
        @use_cuda: whether or not to use cuda
        '''

        self.name = name
        if use_cuda and not torch.cuda.is_available():
            raise Exception("Asked for CUDA but GPU not found")

        self.use_cuda = use_cuda
        self.epoch = parameter['epochs']
        self.lr = parameter['lr']
        self.batch_size = parameter['batch_size']

        self.model = model.to('cuda' if use_cuda else 'cpu') # model.to('cpu')
        self.criterion = nn.CrossEntropyLoss()
        self.optimizer = optim.Adam(self.model.parameters(), lr=self.lr)
        self.train_loader, self.valid_loader = self.
↪get_dataloaders(dataloaders['train_image_paths'],
                                                                    ␣
↪dataloaders['train_labels'],
                                                                    ␣
↪dataloaders['valid_image_paths'],
                                                                    ␣
↪dataloaders['valid_labels'],
                                                                    ␣
↪train_transforms=dataloaders['transforms'],
                                                                     batch_size␣
↪= self.batch_size,
                                                                    ␣
↪shuffle=parameter['shuffle'],
                                                                    sampler =␣
↪dataloaders['sampler'])
        self.class_names = parameter['class_names']
```

```
        self.activations_path = os.path.join('activations', self.name)
        self.kernel_path = os.path.join('kernel_viz', self.name)
        save_path = os.path.join(os.getcwd(), 'models', self.name)
        if not os.path.exists(save_path):
            os.makedirs(save_path)

        if not os.path.exists(self.activations_path):
            os.makedirs(self.activations_path)

        if not os.path.exists(self.kernel_path):
            os.makedirs(self.kernel_path)

        self.save_path = save_path

    def train(self, save=True):
        '''
        @epochs: number of epochs to train
        @save: whether or not to save the checkpoints
        '''
        best_val_accuracy = - math.inf

        for epoch in range(self.epoch):  # loop over the dataset multiple times
            self.model.train()
            t = time.time()
            running_loss = 0.0
            train_acc = 0
            val_accuracy = 0
            correct = 0
            total = 0
            count = 0
            loop = tqdm.tqdm(self.train_loader, total = len(self.train_loader),␣
↪leave = True)

            for img, label in loop:
                # get the inputs; data is a list of [inputs, labels]
                inputs, labels = img.to(device), label.to(device) #img.
↪to(device), label.to(device)

                # zero the parameter gradients
                self.optimizer.zero_grad()

                # forward + backward + optimize
                outputs = self.model(inputs)
                _, predictions = torch.max(outputs, 1)
                loss = self.criterion(outputs, labels)
                loss.backward()
                self.optimizer.step()
```

```
                # print statistics
                running_loss += loss.item()
                total += labels.shape[0]
                correct += (predictions == labels).sum().item()

                count += 1
                if count % 2000 == 1999:    # print every 2000 mini-batches
                    print(f'[{epoch + 1}, {count + 1:5d}] loss: {running_loss /␣
↪2000:.3f}')
                    running_loss = 0.0

            train_acc = 100 * correct / total
            print(f'Epoch:', epoch + 1, f'Training Epoch Accuracy:{train_acc}')

            # evaluate the validation dataset
            self.model.eval()
            correct_pred = {classname: 0 for classname in self.class_names}
            total_pred = {classname: 0 for classname in self.class_names}

            # again no gradients needed
            correct = 0
            total = 0
            with torch.no_grad():
                for data in self.valid_loader:
                    images, labels = data[0].to(device), data[1].to(device)␣
↪#data[0], data[1]
                    outputs = self.model(images)
                    _, predictions = torch.max(outputs, 1)
                    # collect the correct predictions for each class
                    total += labels.shape[0]
                    correct += (predictions == labels).sum().item()

                    for label, prediction in zip(labels, predictions):
                        if label == prediction:
                            correct_pred[classes[label]] += 1
                        total_pred[classes[label]] += 1

            val_accuracy = 100 * correct / total
            print(f'Epoch:', epoch + 1, f'Validation Epoch Accuracy:
↪{val_accuracy}')

            # print the summary for each class
            print('Epoch:', epoch + 1, 'Correct predictions', correct_pred)
            print('Epoch:', epoch + 1, 'Total predictions', total_pred)
            print('Epoch:', epoch + 1, 'Correct predictions', correct_pred)
            print('Epoch:', epoch + 1, 'Total predictions', total_pred)
```

```
            # inspect the time taken to train one epoch
            d = time.time()-t
            print('Fininsh Trainig Epoch', epoch, '!', 'Time used:', d)

            if save:
                torch.save(self.model.state_dict(), os.path.join(self.
↪save_path, f'epoch_{epoch}.pt'))
                if val_accuracy > best_val_accuracy:
                    torch.save(self.model.state_dict(), os.path.join(self.
↪save_path, 'best.pt'))
                    best_val_accuracy = val_accuracy

        print('Done training!')


    def evaluate(self):
        # for evaluating the test dataset if there were any.
        try:
            assert os.path.exists(os.path.join(self.save_path, 'best.pt'))

        except:
            print('Please train first')
            return

        self.model.load_state_dict(torch.load(os.path.join(self.save_path,␣
↪'best.pt')))
        self.model.eval()


    def get_dataloaders(self, train_image_paths, train_labels,␣
↪valid_image_paths, valid_labels, train_transforms=False, batch_size=32,␣
↪shuffle=True, sampler = None):
        train_dataset = SurgicalDataset(train_image_paths,train_labels,␣
↪train_transforms)
        val_dataset = SurgicalDataset(valid_image_paths,valid_labels,␣
↪train_transforms)
        train_loader = DataLoader(train_dataset, batch_size, shuffle, sampler)
        valid_loader = DataLoader(val_dataset, batch_size, shuffle = True)

        return train_loader, valid_loader


    def grad_cam_on_input(self, img):

        try:
```

```
            assert os.path.exists(os.path.join(self.save_path, 'best.pt'))

        except:
            print('It appears you are testing the model without training. Please train first')
            return

        self.model.load_state_dict(torch.load(os.path.join(self.save_path, 'best.pt')))


        self.model.eval()
        img = img.to('cuda' if self.use_cuda else 'cpu')


        out = self.model(img)

        _, pred = torch.max(out, 1)

        predicted_class = self.class_names[int(pred)]
        print(f'Predicted class was {predicted_class}')

        out[:, pred].backward()
        gradients = self.model.get_gradient_activations()

        print('Gradients shape: ', f'{gradients.shape}')

        mean_gradients = torch.mean(gradients, [0, 2, 3]).cpu()
        activations = self.model.get_final_conv_layer(img).detach().cpu()

        print('Activations shape: ', f'{activations.shape}')

        for idx in range(activations.shape[1]):
            activations[:, idx, :, :] *= mean_gradients[idx]

        final_heatmap = np.maximum(torch.mean(activations, dim=1).squeeze(), 0)

        final_heatmap /= torch.max(final_heatmap)

        return final_heatmap

    def trained_kernel_viz(self):

        all_layers = [0, 3]
        all_filters = []
        for layer in all_layers:
```

```
            filters = self.model.conv_model[layer].weight
            all_filters.append(filters.detach().cpu().clone()[:8, :8, :, :])

        for filter_idx in range(len(all_filters)):

            filter = all_filters[filter_idx]
            print(filter.shape)
            filter = filter.contiguous().view(-1, 1, filter.shape[2], filter.
↪shape[3])
            image = show_img(make_grid(filter))
            image = 255 * image
            cv2.imwrite(os.path.join(self.kernel_path,␣
↪f'filter_layer{all_layers[filter_idx]}.jpg'), image)


    def activations_on_input(self, img):

        img = img.to('cuda' if self.use_cuda else 'cpu')

        all_layers = [0,3,6,8,10]
        all_viz = []

        # looking at the outputs of the relu
        for each in all_layers:

            current_model = self.model.conv_model[:each+1]
            current_out = current_model(img)
            all_viz.append(current_out.detach().cpu().clone()[:, :64, :, :])

        for viz_idx in range(len(all_viz)):

            viz = all_viz[viz_idx]
            viz = viz.view(-1, 1, viz.shape[2], viz.shape[3])
            image = show_img(make_grid(viz))
            image = 255 * image
            cv2.imwrite(os.path.join(self.activations_path,␣
↪f'sample_layer{all_layers[viz_idx]}.jpg'), image)
```

4 Build and train models

[9]:
```
from prettytable import PrettyTable

def count_parameters(model):
    table = PrettyTable(["Modules", "Parameters"])
    total_params = 0
```

```
    for name, parameter in model.named_parameters():
        if not parameter.requires_grad: continue
        params = parameter.numel()
        table.add_row([name, params])
        total_params+=params
    print(table)
    print(f"Total Trainable Params: {total_params}")
    return total_params
example_model =  models.vit_l_32(pretrained=True) #vit_b_32 = models.
 ↪vit_b_32(pretrained=True)
count_parameters(example_model)
```

[10]:
```
# example_model = models.vit_l_32(pretrained=True)
# count_parameters(example_model)
```

[11]:
```
# vit_l_16 = models.vit_l_16(pretrained=True)
# count_parameters(vit_l_16)
```

[12]:
```
example_model =  models.vit_l_32(pretrained=True) #vit_b_32 = models.
 ↪vit_b_32(pretrained=True)
count_parameters(example_model)
```

```
+-----------------------------------------------------------------+------------+
|                             Modules                             | Parameters |
+-----------------------------------------------------------------+------------+
|                           class_token                           |    1024    |
|                         conv_proj.weight                        |  3145728   |
|                          conv_proj.bias                         |    1024    |
|                      encoder.pos_embedding                      |   51200    |
|           encoder.layers.encoder_layer_0.ln_1.weight            |    1024    |
|            encoder.layers.encoder_layer_0.ln_1.bias             |    1024    |
|  encoder.layers.encoder_layer_0.self_attention.in_proj_weight   |  3145728   |
|   encoder.layers.encoder_layer_0.self_attention.in_proj_bias    |    3072    |
| encoder.layers.encoder_layer_0.self_attention.out_proj.weight   |  1048576   |
|  encoder.layers.encoder_layer_0.self_attention.out_proj.bias    |    1024    |
|           encoder.layers.encoder_layer_0.ln_2.weight            |    1024    |
|            encoder.layers.encoder_layer_0.ln_2.bias             |    1024    |
|       encoder.layers.encoder_layer_0.mlp.linear_1.weight        |  4194304   |
|        encoder.layers.encoder_layer_0.mlp.linear_1.bias         |    4096    |
|       encoder.layers.encoder_layer_0.mlp.linear_2.weight        |  4194304   |
|        encoder.layers.encoder_layer_0.mlp.linear_2.bias         |    1024    |
|           encoder.layers.encoder_layer_1.ln_1.weight            |    1024    |
|            encoder.layers.encoder_layer_1.ln_1.bias             |    1024    |
|  encoder.layers.encoder_layer_1.self_attention.in_proj_weight   |  3145728   |
|   encoder.layers.encoder_layer_1.self_attention.in_proj_bias    |    3072    |
| encoder.layers.encoder_layer_1.self_attention.out_proj.weight   |  1048576   |
|  encoder.layers.encoder_layer_1.self_attention.out_proj.bias    |    1024    |
|           encoder.layers.encoder_layer_1.ln_2.weight            |    1024    |
```

encoder.layers.encoder_layer_1.ln_2.bias	1024
encoder.layers.encoder_layer_1.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_1.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_1.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_1.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_2.ln_1.weight	1024
encoder.layers.encoder_layer_2.ln_1.bias	1024
encoder.layers.encoder_layer_2.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_2.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_2.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_2.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_2.ln_2.weight	1024
encoder.layers.encoder_layer_2.ln_2.bias	1024
encoder.layers.encoder_layer_2.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_2.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_2.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_2.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_3.ln_1.weight	1024
encoder.layers.encoder_layer_3.ln_1.bias	1024
encoder.layers.encoder_layer_3.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_3.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_3.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_3.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_3.ln_2.weight	1024
encoder.layers.encoder_layer_3.ln_2.bias	1024
encoder.layers.encoder_layer_3.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_3.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_3.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_3.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_4.ln_1.weight	1024
encoder.layers.encoder_layer_4.ln_1.bias	1024
encoder.layers.encoder_layer_4.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_4.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_4.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_4.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_4.ln_2.weight	1024
encoder.layers.encoder_layer_4.ln_2.bias	1024
encoder.layers.encoder_layer_4.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_4.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_4.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_4.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_5.ln_1.weight	1024
encoder.layers.encoder_layer_5.ln_1.bias	1024
encoder.layers.encoder_layer_5.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_5.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_5.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_5.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_5.ln_2.weight	1024

encoder.layers.encoder_layer_5.ln_2.bias	1024
encoder.layers.encoder_layer_5.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_5.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_5.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_5.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_6.ln_1.weight	1024
encoder.layers.encoder_layer_6.ln_1.bias	1024
encoder.layers.encoder_layer_6.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_6.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_6.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_6.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_6.ln_2.weight	1024
encoder.layers.encoder_layer_6.ln_2.bias	1024
encoder.layers.encoder_layer_6.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_6.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_6.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_6.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_7.ln_1.weight	1024
encoder.layers.encoder_layer_7.ln_1.bias	1024
encoder.layers.encoder_layer_7.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_7.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_7.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_7.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_7.ln_2.weight	1024
encoder.layers.encoder_layer_7.ln_2.bias	1024
encoder.layers.encoder_layer_7.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_7.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_7.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_7.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_8.ln_1.weight	1024
encoder.layers.encoder_layer_8.ln_1.bias	1024
encoder.layers.encoder_layer_8.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_8.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_8.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_8.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_8.ln_2.weight	1024
encoder.layers.encoder_layer_8.ln_2.bias	1024
encoder.layers.encoder_layer_8.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_8.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_8.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_8.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_9.ln_1.weight	1024
encoder.layers.encoder_layer_9.ln_1.bias	1024
encoder.layers.encoder_layer_9.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_9.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_9.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_9.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_9.ln_2.weight	1024

encoder.layers.encoder_layer_9.ln_2.bias	1024
encoder.layers.encoder_layer_9.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_9.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_9.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_9.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_10.ln_1.weight	1024
encoder.layers.encoder_layer_10.ln_1.bias	1024
encoder.layers.encoder_layer_10.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_10.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_10.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_10.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_10.ln_2.weight	1024
encoder.layers.encoder_layer_10.ln_2.bias	1024
encoder.layers.encoder_layer_10.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_10.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_10.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_10.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_11.ln_1.weight	1024
encoder.layers.encoder_layer_11.ln_1.bias	1024
encoder.layers.encoder_layer_11.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_11.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_11.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_11.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_11.ln_2.weight	1024
encoder.layers.encoder_layer_11.ln_2.bias	1024
encoder.layers.encoder_layer_11.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_11.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_11.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_11.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_12.ln_1.weight	1024
encoder.layers.encoder_layer_12.ln_1.bias	1024
encoder.layers.encoder_layer_12.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_12.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_12.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_12.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_12.ln_2.weight	1024
encoder.layers.encoder_layer_12.ln_2.bias	1024
encoder.layers.encoder_layer_12.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_12.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_12.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_12.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_13.ln_1.weight	1024
encoder.layers.encoder_layer_13.ln_1.bias	1024
encoder.layers.encoder_layer_13.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_13.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_13.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_13.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_13.ln_2.weight	1024

encoder.layers.encoder_layer_13.ln_2.bias	1024
encoder.layers.encoder_layer_13.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_13.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_13.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_13.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_14.ln_1.weight	1024
encoder.layers.encoder_layer_14.ln_1.bias	1024
encoder.layers.encoder_layer_14.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_14.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_14.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_14.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_14.ln_2.weight	1024
encoder.layers.encoder_layer_14.ln_2.bias	1024
encoder.layers.encoder_layer_14.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_14.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_14.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_14.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_15.ln_1.weight	1024
encoder.layers.encoder_layer_15.ln_1.bias	1024
encoder.layers.encoder_layer_15.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_15.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_15.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_15.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_15.ln_2.weight	1024
encoder.layers.encoder_layer_15.ln_2.bias	1024
encoder.layers.encoder_layer_15.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_15.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_15.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_15.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_16.ln_1.weight	1024
encoder.layers.encoder_layer_16.ln_1.bias	1024
encoder.layers.encoder_layer_16.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_16.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_16.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_16.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_16.ln_2.weight	1024
encoder.layers.encoder_layer_16.ln_2.bias	1024
encoder.layers.encoder_layer_16.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_16.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_16.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_16.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_17.ln_1.weight	1024
encoder.layers.encoder_layer_17.ln_1.bias	1024
encoder.layers.encoder_layer_17.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_17.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_17.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_17.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_17.ln_2.weight	1024

encoder.layers.encoder_layer_17.ln_2.bias	1024
encoder.layers.encoder_layer_17.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_17.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_17.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_17.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_18.ln_1.weight	1024
encoder.layers.encoder_layer_18.ln_1.bias	1024
encoder.layers.encoder_layer_18.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_18.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_18.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_18.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_18.ln_2.weight	1024
encoder.layers.encoder_layer_18.ln_2.bias	1024
encoder.layers.encoder_layer_18.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_18.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_18.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_18.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_19.ln_1.weight	1024
encoder.layers.encoder_layer_19.ln_1.bias	1024
encoder.layers.encoder_layer_19.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_19.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_19.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_19.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_19.ln_2.weight	1024
encoder.layers.encoder_layer_19.ln_2.bias	1024
encoder.layers.encoder_layer_19.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_19.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_19.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_19.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_20.ln_1.weight	1024
encoder.layers.encoder_layer_20.ln_1.bias	1024
encoder.layers.encoder_layer_20.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_20.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_20.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_20.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_20.ln_2.weight	1024
encoder.layers.encoder_layer_20.ln_2.bias	1024
encoder.layers.encoder_layer_20.mlp.linear_1.weight	4194304
encoder.layers.encoder_layer_20.mlp.linear_1.bias	4096
encoder.layers.encoder_layer_20.mlp.linear_2.weight	4194304
encoder.layers.encoder_layer_20.mlp.linear_2.bias	1024
encoder.layers.encoder_layer_21.ln_1.weight	1024
encoder.layers.encoder_layer_21.ln_1.bias	1024
encoder.layers.encoder_layer_21.self_attention.in_proj_weight	3145728
encoder.layers.encoder_layer_21.self_attention.in_proj_bias	3072
encoder.layers.encoder_layer_21.self_attention.out_proj.weight	1048576
encoder.layers.encoder_layer_21.self_attention.out_proj.bias	1024
encoder.layers.encoder_layer_21.ln_2.weight	1024

```
|           encoder.layers.encoder_layer_21.ln_2.bias            |    1024    |
|      encoder.layers.encoder_layer_21.mlp.linear_1.weight       |  4194304   |
|       encoder.layers.encoder_layer_21.mlp.linear_1.bias        |    4096    |
|      encoder.layers.encoder_layer_21.mlp.linear_2.weight       |  4194304   |
|       encoder.layers.encoder_layer_21.mlp.linear_2.bias        |    1024    |
|          encoder.layers.encoder_layer_22.ln_1.weight           |    1024    |
|           encoder.layers.encoder_layer_22.ln_1.bias            |    1024    |
| encoder.layers.encoder_layer_22.self_attention.in_proj_weight  |  3145728   |
|  encoder.layers.encoder_layer_22.self_attention.in_proj_bias   |    3072    |
| encoder.layers.encoder_layer_22.self_attention.out_proj.weight |  1048576   |
|  encoder.layers.encoder_layer_22.self_attention.out_proj.bias  |    1024    |
|          encoder.layers.encoder_layer_22.ln_2.weight           |    1024    |
|           encoder.layers.encoder_layer_22.ln_2.bias            |    1024    |
|      encoder.layers.encoder_layer_22.mlp.linear_1.weight       |  4194304   |
|       encoder.layers.encoder_layer_22.mlp.linear_1.bias        |    4096    |
|      encoder.layers.encoder_layer_22.mlp.linear_2.weight       |  4194304   |
|       encoder.layers.encoder_layer_22.mlp.linear_2.bias        |    1024    |
|          encoder.layers.encoder_layer_23.ln_1.weight           |    1024    |
|           encoder.layers.encoder_layer_23.ln_1.bias            |    1024    |
| encoder.layers.encoder_layer_23.self_attention.in_proj_weight  |  3145728   |
|  encoder.layers.encoder_layer_23.self_attention.in_proj_bias   |    3072    |
| encoder.layers.encoder_layer_23.self_attention.out_proj.weight |  1048576   |
|  encoder.layers.encoder_layer_23.self_attention.out_proj.bias  |    1024    |
|          encoder.layers.encoder_layer_23.ln_2.weight           |    1024    |
|           encoder.layers.encoder_layer_23.ln_2.bias            |    1024    |
|      encoder.layers.encoder_layer_23.mlp.linear_1.weight       |  4194304   |
|       encoder.layers.encoder_layer_23.mlp.linear_1.bias        |    4096    |
|      encoder.layers.encoder_layer_23.mlp.linear_2.weight       |  4194304   |
|       encoder.layers.encoder_layer_23.mlp.linear_2.bias        |    1024    |
|                      encoder.ln.weight                         |    1024    |
|                       encoder.ln.bias                          |    1024    |
|                      heads.head.weight                         |  1024000   |
|                       heads.head.bias                          |    1000    |
+---------------------------------------------------------------+------------+
Total Trainable Params: 306535400
```

[12]:
```
306535400
```

[13]:
```
print(example_model)
```

```
VisionTransformer(
  (conv_proj): Conv2d(3, 1024, kernel_size=(32, 32), stride=(32, 32))
  (encoder): Encoder(
    (dropout): Dropout(p=0.0, inplace=False)
    (layers): Sequential(
      (encoder_layer_0): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
```

```
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_1): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_2): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_3): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
```

```
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_4): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_5): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_6): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
```

```
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_7): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_8): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_9): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
```

```
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_10): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_11): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_12): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
```

```
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_13): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_14): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_15): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
```

```
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_16): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_17): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_18): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
```

```
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_19): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_20): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_21): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
```

```
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_22): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
      (encoder_layer_23): EncoderBlock(
        (ln_1): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (self_attention): MultiheadAttention(
          (out_proj): NonDynamicallyQuantizableLinear(in_features=1024,
out_features=1024, bias=True)
        )
        (dropout): Dropout(p=0.0, inplace=False)
        (ln_2): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
        (mlp): MLPBlock(
          (linear_1): Linear(in_features=1024, out_features=4096, bias=True)
          (act): GELU()
          (dropout_1): Dropout(p=0.0, inplace=False)
          (linear_2): Linear(in_features=4096, out_features=1024, bias=True)
          (dropout_2): Dropout(p=0.0, inplace=False)
        )
      )
    )
    (ln): LayerNorm((1024,), eps=1e-06, elementwise_affine=True)
  )
```

```
  (heads): Sequential(
    (head): Linear(in_features=1024, out_features=1000, bias=True)
  )
)
```

[14]:
```
print(example_model.heads)
```

```
Sequential(
  (head): Linear(in_features=1024, out_features=1000, bias=True)
)
```

[15]:
```
for name, param in example_model.named_parameters():
    number = name.split('.')
    print(number)
    #if number[0] == 'layers':
        #print(number[1].split('_')[2])
        #print(number[2])
```

```
['class_token']
['conv_proj', 'weight']
['conv_proj', 'bias']
['encoder', 'pos_embedding']
['encoder', 'layers', 'encoder_layer_0', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_0', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_0', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_0', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_0', 'self_attention', 'out_proj', 'weight']
['encoder', 'layers', 'encoder_layer_0', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_0', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_0', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_0', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_0', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_0', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_0', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_1', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_1', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_1', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_1', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_1', 'self_attention', 'out_proj', 'weight']
['encoder', 'layers', 'encoder_layer_1', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_1', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_1', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_1', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_1', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_1', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_1', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_2', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_2', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_2', 'self_attention', 'in_proj_weight']
```

```
['encoder', 'layers', 'encoder_layer_2', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_2', 'self_attention', 'out_proj', 'weight']
['encoder', 'layers', 'encoder_layer_2', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_2', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_2', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_2', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_2', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_2', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_2', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_3', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_3', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_3', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_3', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_3', 'self_attention', 'out_proj', 'weight']
['encoder', 'layers', 'encoder_layer_3', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_3', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_3', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_3', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_3', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_3', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_3', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_4', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_4', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_4', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_4', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_4', 'self_attention', 'out_proj', 'weight']
['encoder', 'layers', 'encoder_layer_4', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_4', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_4', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_4', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_4', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_4', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_4', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_5', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_5', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_5', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_5', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_5', 'self_attention', 'out_proj', 'weight']
['encoder', 'layers', 'encoder_layer_5', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_5', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_5', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_5', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_5', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_5', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_5', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_6', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_6', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_6', 'self_attention', 'in_proj_weight']
```

```
['encoder', 'layers', 'encoder_layer_6', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_6', 'self_attention', 'out_proj', 'weight']
['encoder', 'layers', 'encoder_layer_6', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_6', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_6', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_6', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_6', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_6', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_6', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_7', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_7', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_7', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_7', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_7', 'self_attention', 'out_proj', 'weight']
['encoder', 'layers', 'encoder_layer_7', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_7', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_7', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_7', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_7', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_7', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_7', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_8', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_8', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_8', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_8', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_8', 'self_attention', 'out_proj', 'weight']
['encoder', 'layers', 'encoder_layer_8', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_8', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_8', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_8', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_8', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_8', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_8', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_9', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_9', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_9', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_9', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_9', 'self_attention', 'out_proj', 'weight']
['encoder', 'layers', 'encoder_layer_9', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_9', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_9', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_9', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_9', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_9', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_9', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_10', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_10', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_10', 'self_attention', 'in_proj_weight']
```

```
['encoder', 'layers', 'encoder_layer_10', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_10', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_10', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_10', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_10', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_10', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_10', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_10', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_10', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_11', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_11', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_11', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_11', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_11', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_11', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_11', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_11', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_11', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_11', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_11', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_11', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_12', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_12', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_12', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_12', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_12', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_12', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_12', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_12', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_12', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_12', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_12', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_12', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_13', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_13', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_13', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_13', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_13', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_13', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_13', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_13', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_13', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_13', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_13', 'mlp', 'linear_2', 'weight']
```

```
['encoder', 'layers', 'encoder_layer_13', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_14', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_14', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_14', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_14', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_14', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_14', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_14', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_14', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_14', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_14', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_14', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_14', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_15', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_15', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_15', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_15', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_15', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_15', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_15', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_15', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_15', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_15', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_15', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_15', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_16', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_16', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_16', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_16', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_16', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_16', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_16', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_16', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_16', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_16', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_16', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_16', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_17', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_17', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_17', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_17', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_17', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_17', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_17', 'ln_2', 'weight']
```

```
['encoder', 'layers', 'encoder_layer_17', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_17', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_17', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_17', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_17', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_18', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_18', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_18', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_18', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_18', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_18', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_18', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_18', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_18', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_18', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_18', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_18', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_19', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_19', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_19', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_19', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_19', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_19', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_19', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_19', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_19', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_19', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_19', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_19', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_20', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_20', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_20', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_20', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_20', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_20', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_20', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_20', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_20', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_20', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_20', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_20', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_21', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_21', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_21', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_21', 'self_attention', 'in_proj_bias']
```

```
['encoder', 'layers', 'encoder_layer_21', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_21', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_21', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_21', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_21', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_21', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_21', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_21', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_22', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_22', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_22', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_22', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_22', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_22', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_22', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_22', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_22', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_22', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_22', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_22', 'mlp', 'linear_2', 'bias']
['encoder', 'layers', 'encoder_layer_23', 'ln_1', 'weight']
['encoder', 'layers', 'encoder_layer_23', 'ln_1', 'bias']
['encoder', 'layers', 'encoder_layer_23', 'self_attention', 'in_proj_weight']
['encoder', 'layers', 'encoder_layer_23', 'self_attention', 'in_proj_bias']
['encoder', 'layers', 'encoder_layer_23', 'self_attention', 'out_proj',
'weight']
['encoder', 'layers', 'encoder_layer_23', 'self_attention', 'out_proj', 'bias']
['encoder', 'layers', 'encoder_layer_23', 'ln_2', 'weight']
['encoder', 'layers', 'encoder_layer_23', 'ln_2', 'bias']
['encoder', 'layers', 'encoder_layer_23', 'mlp', 'linear_1', 'weight']
['encoder', 'layers', 'encoder_layer_23', 'mlp', 'linear_1', 'bias']
['encoder', 'layers', 'encoder_layer_23', 'mlp', 'linear_2', 'weight']
['encoder', 'layers', 'encoder_layer_23', 'mlp', 'linear_2', 'bias']
['encoder', 'ln', 'weight']
['encoder', 'ln', 'bias']
['heads', 'head', 'weight']
['heads', 'head', 'bias']
```

[18]:
```python
class TransferViT_l_32(nn.Module):
    def __init__(self):
        super().__init__()
        self.vit = models.vit_l_32(pretrained=True)
        #self.conv_layer = self.get_conv_proj()
        self.vit.heads = self.get_fc_layers()
        #self.vit = self.get_ViT_encoder()
```

```
        #self.fc_model = self.get_fc_layers()
        self.activate_training_layers()

    def activate_training_layers(self):
#         for name, param in self.conv_layer.named_parameters():
#             # for all of these layers set param.requires_grad as True
#             param.requires_grad = False

        for name, param in self.vit.named_parameters():
            number = name.split('.')
            # for all layers except the last conv layer, set param.
 ↪requires_grad = False
            if number[0] == 'heads':
#                 if number[1].split('_')[2] == 11 and number[2] == 'mlp':
#                     param.requires_grad = True
#                 else:
                param.requires_grad = True
                print('required_grad = True', number)
            else:
                param.requires_grad = False
                print('required_grad = False', number)

        #for name, param in self.vit.heads.named_parameters():
            # for all of these layers set param.requires_grad as True


#     def get_conv_proj(self):
#         return self.base_ViT.conv_proj

#     def get_ViT_encoder(self):
#         return self.base_ViT.encoder

    def get_fc_layers(self):
        return nn.Sequential(
            nn.Dropout(p=0.5, inplace=False),
            nn.Linear(in_features=1024, out_features=512, bias=True),
            nn.ReLU(inplace=True),
            nn.Dropout(p=0.5, inplace=False),
            nn.Linear(in_features=512, out_features=128, bias=True),
            nn.ReLU(inplace=True),
            nn.Linear(in_features=128, out_features=15, bias=True),
        )

    def forward(self, x):
        #x = self.conv_layer(x)
        x = self.vit(x)
        #x = torch.flatten(x, 1)
```

```
        #x = self.fc_model(x)  #call fully connected layers

        return x
```

[19]:
```
# Train a transfer learning model with Alexnet
name = 'TransferViT_l_32'
classes = [i for i in range(15)]
transforms = get_transform('TransferViT')
dataloaders = {'train_image_paths': train_image_paths, 'train_labels' :␣
↪train_labels, 'valid_image_paths': valid_image_paths, 'valid_labels':
↪valid_labels, 'transforms':transforms, 'sampler':sampler}
parameters = {'lr': 0.001, 'epochs' : 5, 'batch_size':32, 'shuffle':False,␣
↪'class_names':classes}

model = TransferViT_l_32()
classifier = Classifier(name, model, dataloaders, parameters, use_cuda=True)
classifier.train()
```

```
required_grad = False ['class_token']
required_grad = False ['conv_proj', 'weight']
required_grad = False ['conv_proj', 'bias']
required_grad = False ['encoder', 'pos_embedding']
required_grad = False ['encoder', 'layers', 'encoder_layer_0', 'ln_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_0', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_0', 'self_attention',
'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_0', 'self_attention',
'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_0', 'self_attention',
'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_0', 'self_attention',
'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_0', 'ln_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_0', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_0', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_0', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_0', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_0', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_1', 'ln_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_1', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_1', 'self_attention',
'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_1', 'self_attention',
'in_proj_bias']
```

```
required_grad = False ['encoder', 'layers', 'encoder_layer_1', 'self_attention',
'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_1', 'self_attention',
'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_1', 'ln_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_1', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_1', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_1', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_1', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_1', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_2', 'ln_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_2', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_2', 'self_attention',
'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_2', 'self_attention',
'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_2', 'self_attention',
'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_2', 'self_attention',
'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_2', 'ln_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_2', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_2', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_2', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_2', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_2', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_3', 'ln_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_3', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_3', 'self_attention',
'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_3', 'self_attention',
'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_3', 'self_attention',
'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_3', 'self_attention',
'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_3', 'ln_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_3', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_3', 'mlp',
'linear_1', 'weight']
```

```
required_grad = False ['encoder', 'layers', 'encoder_layer_3', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_3', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_3', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_4', 'ln_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_4', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_4', 'self_attention',
'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_4', 'self_attention',
'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_4', 'self_attention',
'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_4', 'self_attention',
'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_4', 'ln_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_4', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_4', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_4', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_4', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_4', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_5', 'ln_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_5', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_5', 'self_attention',
'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_5', 'self_attention',
'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_5', 'self_attention',
'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_5', 'self_attention',
'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_5', 'ln_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_5', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_5', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_5', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_5', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_5', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_6', 'ln_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_6', 'ln_1', 'bias']
```

```
required_grad = False ['encoder', 'layers', 'encoder_layer_6', 'self_attention',
'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_6', 'self_attention',
'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_6', 'self_attention',
'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_6', 'self_attention',
'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_6', 'ln_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_6', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_6', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_6', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_6', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_6', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_7', 'ln_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_7', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_7', 'self_attention',
'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_7', 'self_attention',
'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_7', 'self_attention',
'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_7', 'self_attention',
'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_7', 'ln_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_7', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_7', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_7', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_7', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_7', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_8', 'ln_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_8', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_8', 'self_attention',
'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_8', 'self_attention',
'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_8', 'self_attention',
'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_8', 'self_attention',
'out_proj', 'bias']
```

```
required_grad = False ['encoder', 'layers', 'encoder_layer_8', 'ln_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_8', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_8', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_8', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_8', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_8', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_9', 'ln_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_9', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_9', 'self_attention',
'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_9', 'self_attention',
'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_9', 'self_attention',
'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_9', 'self_attention',
'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_9', 'ln_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_9', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_9', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_9', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_9', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_9', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_10', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_10', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_10',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_10',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_10',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_10',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_10', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_10', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_10', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_10', 'mlp',
'linear_1', 'bias']
```

```
required_grad = False ['encoder', 'layers', 'encoder_layer_10', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_10', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_11', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_11', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_11',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_11',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_11',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_11',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_11', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_11', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_11', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_11', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_11', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_11', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_12', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_12', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_12',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_12',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_12',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_12',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_12', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_12', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_12', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_12', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_12', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_12', 'mlp',
'linear_2', 'bias']
```

```
required_grad = False ['encoder', 'layers', 'encoder_layer_13', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_13', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_13',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_13',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_13',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_13',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_13', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_13', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_13', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_13', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_13', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_13', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_14', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_14', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_14',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_14',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_14',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_14',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_14', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_14', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_14', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_14', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_14', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_14', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_15', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_15', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_15',
```

```
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_15',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_15',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_15',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_15', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_15', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_15', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_15', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_15', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_15', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_16', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_16', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_16',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_16',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_16',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_16',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_16', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_16', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_16', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_16', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_16', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_16', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_17', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_17', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_17',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_17',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_17',
```

```
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_17',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_17', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_17', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_17', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_17', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_17', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_17', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_18', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_18', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_18',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_18',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_18',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_18',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_18', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_18', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_18', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_18', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_18', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_18', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_19', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_19', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_19',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_19',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_19',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_19',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_19', 'ln_2',
```

```
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_19', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_19', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_19', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_19', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_19', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_20', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_20', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_20',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_20',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_20',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_20',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_20', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_20', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_20', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_20', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_20', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_20', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_21', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_21', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_21',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_21',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_21',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_21',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_21', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_21', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_21', 'mlp',
'linear_1', 'weight']
```

```
required_grad = False ['encoder', 'layers', 'encoder_layer_21', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_21', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_21', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_22', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_22', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_22',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_22',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_22',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_22',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_22', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_22', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_22', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_22', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_22', 'mlp',
'linear_2', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_22', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_23', 'ln_1',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_23', 'ln_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_23',
'self_attention', 'in_proj_weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_23',
'self_attention', 'in_proj_bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_23',
'self_attention', 'out_proj', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_23',
'self_attention', 'out_proj', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_23', 'ln_2',
'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_23', 'ln_2', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_23', 'mlp',
'linear_1', 'weight']
required_grad = False ['encoder', 'layers', 'encoder_layer_23', 'mlp',
'linear_1', 'bias']
required_grad = False ['encoder', 'layers', 'encoder_layer_23', 'mlp',
'linear_2', 'weight']
```

```
required_grad = False ['encoder', 'layers', 'encoder_layer_23', 'mlp',
'linear_2', 'bias']
required_grad = False ['encoder', 'ln', 'weight']
required_grad = False ['encoder', 'ln', 'bias']
required_grad = True ['heads', '1', 'weight']
required_grad = True ['heads', '1', 'bias']
required_grad = True ['heads', '4', 'weight']
required_grad = True ['heads', '4', 'bias']
required_grad = True ['heads', '6', 'weight']
required_grad = True ['heads', '6', 'bias']

 37%|
| 1999/5377 [12:53<21:38,  2.60it/s]

[1,  2000] loss: 0.875

 74%|
| 3999/5377 [25:46<08:48,  2.61it/s]

[1,  4000] loss: 0.658

100%|
  | 5377/5377 [34:38<00:00,  2.59it/s]

Epoch: 1 Training Epoch Accuracy:75.0652445581098
Epoch: 1 Validation Epoch Accuracy:76.74369943271645
Epoch: 1 Correct predictions {0: 50, 1: 1983, 2: 4002, 3: 8854, 4: 186, 5: 4675,
6: 133, 7: 1917, 8: 245, 9: 521, 10: 8826, 11: 921, 12: 395, 13: 284, 14: 17}
Epoch: 1 Total predictions {0: 52, 1: 2177, 2: 5799, 3: 10405, 4: 213, 5: 5470,
6: 158, 7: 2843, 8: 247, 9: 587, 10: 11135, 11: 3216, 12: 404, 13: 288, 14: 18}
Epoch: 1 Correct predictions {0: 50, 1: 1983, 2: 4002, 3: 8854, 4: 186, 5: 4675,
6: 133, 7: 1917, 8: 245, 9: 521, 10: 8826, 11: 921, 12: 395, 13: 284, 14: 17}
Epoch: 1 Total predictions {0: 52, 1: 2177, 2: 5799, 3: 10405, 4: 213, 5: 5470,
6: 158, 7: 2843, 8: 247, 9: 587, 10: 11135, 11: 3216, 12: 404, 13: 288, 14: 18}
Fininsh Trainig Epoch 0 ! Time used: 2615.5640823841095

 37%|
| 1999/5377 [12:53<21:46,  2.59it/s]

[2,  2000] loss: 0.588

 74%|
| 3999/5377 [25:42<08:50,  2.60it/s]

[2,  4000] loss: 0.563

100%|
  | 5377/5377 [34:32<00:00,  2.59it/s]

Epoch: 2 Training Epoch Accuracy:80.09358016797931
Epoch: 2 Validation Epoch Accuracy:78.55249697758765
Epoch: 2 Correct predictions {0: 49, 1: 1998, 2: 4455, 3: 8534, 4: 190, 5: 4808,
6: 136, 7: 1769, 8: 246, 9: 538, 10: 9467, 11: 902, 12: 395, 13: 283, 14: 17}
Epoch: 2 Total predictions {0: 52, 1: 2177, 2: 5799, 3: 10405, 4: 213, 5: 5470,
```

```
6: 158, 7: 2843, 8: 247, 9: 587, 10: 11135, 11: 3216, 12: 404, 13: 288, 14: 18}
Epoch: 2 Correct predictions {0: 49, 1: 1998, 2: 4455, 3: 8534, 4: 190, 5: 4808,
6: 136, 7: 1769, 8: 246, 9: 538, 10: 9467, 11: 902, 12: 395, 13: 283, 14: 17}
Epoch: 2 Total predictions {0: 52, 1: 2177, 2: 5799, 3: 10405, 4: 213, 5: 5470,
6: 158, 7: 2843, 8: 247, 9: 587, 10: 11135, 11: 3216, 12: 404, 13: 288, 14: 18}
Fininsh Trainig Epoch 1 ! Time used: 2607.7863669395447

 37%|
| 1999/5377 [12:51<21:50,  2.58it/s]

[3,  2000] loss: 0.538

 74%|
| 3999/5377 [25:38<08:43,  2.63it/s]

[3,  4000] loss: 0.530

100%|
  | 5377/5377 [34:28<00:00,  2.60it/s]

Epoch: 3 Training Epoch Accuracy:81.43683338661397
Epoch: 3 Validation Epoch Accuracy:76.75997396075513
Epoch: 3 Correct predictions {0: 50, 1: 2045, 2: 4279, 3: 8769, 4: 185, 5: 4805,
6: 135, 7: 2202, 8: 244, 9: 553, 10: 7960, 11: 1092, 12: 396, 13: 284, 14: 17}
Epoch: 3 Total predictions {0: 52, 1: 2177, 2: 5799, 3: 10405, 4: 213, 5: 5470,
6: 158, 7: 2843, 8: 247, 9: 587, 10: 11135, 11: 3216, 12: 404, 13: 288, 14: 18}
Epoch: 3 Correct predictions {0: 50, 1: 2045, 2: 4279, 3: 8769, 4: 185, 5: 4805,
6: 135, 7: 2202, 8: 244, 9: 553, 10: 7960, 11: 1092, 12: 396, 13: 284, 14: 17}
Epoch: 3 Total predictions {0: 52, 1: 2177, 2: 5799, 3: 10405, 4: 213, 5: 5470,
6: 158, 7: 2843, 8: 247, 9: 587, 10: 11135, 11: 3216, 12: 404, 13: 288, 14: 18}
Fininsh Trainig Epoch 2 ! Time used: 2605.05811047554

 37%|
| 1999/5377 [12:52<21:27,  2.62it/s]

[4,  2000] loss: 0.510

 74%|
| 3999/5377 [25:39<08:50,  2.60it/s]

[4,  4000] loss: 0.509

100%|
  | 5377/5377 [34:27<00:00,  2.60it/s]

Epoch: 4 Training Epoch Accuracy:82.0535325060304
Epoch: 4 Validation Epoch Accuracy:78.18283269785177
Epoch: 4 Correct predictions {0: 50, 1: 2014, 2: 4271, 3: 8793, 4: 191, 5: 4846,
6: 138, 7: 2232, 8: 244, 9: 535, 10: 8629, 11: 985, 12: 400, 13: 283, 14: 17}
Epoch: 4 Total predictions {0: 52, 1: 2177, 2: 5799, 3: 10405, 4: 213, 5: 5470,
6: 158, 7: 2843, 8: 247, 9: 587, 10: 11135, 11: 3216, 12: 404, 13: 288, 14: 18}
Epoch: 4 Correct predictions {0: 50, 1: 2014, 2: 4271, 3: 8793, 4: 191, 5: 4846,
6: 138, 7: 2232, 8: 244, 9: 535, 10: 8629, 11: 985, 12: 400, 13: 283, 14: 17}
Epoch: 4 Total predictions {0: 52, 1: 2177, 2: 5799, 3: 10405, 4: 213, 5: 5470,
```

```
6: 158, 7: 2843, 8: 247, 9: 587, 10: 11135, 11: 3216, 12: 404, 13: 288, 14: 18}
Fininsh Trainig Epoch 3 ! Time used: 2602.013230085373

 37%|
| 1999/5377 [12:54<21:27,  2.62it/s]

[5,  2000] loss: 0.504

 74%|
| 3999/5377 [25:39<08:46,  2.62it/s]

[5,  4000] loss: 0.495

100%|
  | 5377/5377 [34:41<00:00,  2.58it/s]

Epoch: 5 Training Epoch Accuracy:82.50050858786945
Epoch: 5 Validation Epoch Accuracy:81.49818655258997
Epoch: 5 Correct predictions {0: 51, 1: 2032, 2: 3500, 3: 9769, 4: 186, 5: 4936,
6: 133, 7: 1903, 8: 245, 9: 527, 10: 9819, 11: 1256, 12: 398, 13: 282, 14: 17}
Epoch: 5 Total predictions {0: 52, 1: 2177, 2: 5799, 3: 10405, 4: 213, 5: 5470,
6: 158, 7: 2843, 8: 247, 9: 587, 10: 11135, 11: 3216, 12: 404, 13: 288, 14: 18}
Epoch: 5 Correct predictions {0: 51, 1: 2032, 2: 3500, 3: 9769, 4: 186, 5: 4936,
6: 133, 7: 1903, 8: 245, 9: 527, 10: 9819, 11: 1256, 12: 398, 13: 282, 14: 17}
Epoch: 5 Total predictions {0: 52, 1: 2177, 2: 5799, 3: 10405, 4: 213, 5: 5470,
6: 158, 7: 2843, 8: 247, 9: 587, 10: 11135, 11: 3216, 12: 404, 13: 288, 14: 18}
Fininsh Trainig Epoch 4 ! Time used: 2621.1499495506287
Done training!
```

5 Data Augmentations for scarse data

[ ]:
```
# find out the ratio of different labels/data

# we decide to augment the scarser data in the following scheme: label 0, +␣
↪400, label 14 + 500, label 6 + 200
```