hw5

February 18, 2025

1 Homework 5

We have learned about the basics of using machine learning and deep learning for many computer vision problems, including object classification, semantic segmentation, object detection, etc. In this assignment, we will be building a framework for object classification using PyTorch.

Topics you will be learning in this assignment: * Defining datasets in PyTorch; * Defining models in PyTorch; * Specifying the training procedure; * Training and evaluating a model; * Tuning hyper-parameters.

2 0. Setup

[ ]:
```
# import os

# if not os.path.exists("CS131_release"):
#     # Clone the repository if it doesn't already exist
#     !git clone https://github.com/StanfordVL/CS131_release.git

# %cd CS131_release/winter_2025/hw5_release/
```

[ ]:
```
# # Install the necessary dependencies
# # (restart your runtime session if prompted to, and then re-run this cell)
# !pip install -r requirements.txt
```

2.1 1. Object classification on CIFAR10 with a simple ConvNet.

As one of the most famous datasets in computer vision, CIFAR10 is an object classfication dataset that consists of 60000 color RGB images in 10 classes, with 6000 images per class. The images are all at a resolution of 32x32. In this section, we will be working out a framework that is able to tell us what object there is in a given image, using a simple Convolution Neural Network.

2.1.1 1.1 Data preparation.

The most important ingredient in a deep learning recipe is arguably data - what we feed into the model largely determines what we get out of it. In this part, let's prepare our data in a format that will be best useable in the rest of the framework.

For vision, there's a useful package called `torchvision` that defines data loaders for common datasets as well as various image transformation operations. Let's first load and normalize the training and testing dataset using `torchvision`.

As a quick refresher question. Why do we want to split our data into training and testing sets?

You answer here: We split our data into training and testing sets so that we can train the model on one portion of the data and then evaluate its performance on a completely separate set of data. This approach helps us determine whether the model is generalizing well to new, unseen data or simply memorizing the training data (i.e., overfitting).

[1]:

```python
import torch
import torchvision
import torchvision.transforms as transforms
import numpy as np
import random

# set random seeds
torch.manual_seed(131)
np.random.seed(131)
random.seed(131)
```

From this step, you want to create two dataloaders `trainloader` and `testloader` from which we will query our data. You might want to familiarize yourself with PyTorch data structures for this. Specifically, `torch.utils.data.Dataset` and `torch.utils.data.DataLoader` might be helpful here. `torchvision` also provides convenient interfaces for some popular datasets including CIFAR10, so you may find `torchvision.datasets` helpful too.

When dealing with image data, oftentimes we need to do some preprocessing to convert the data to the format we need. In this problem, the main preprocessing we need to do is normalization. Specifically, let's normalize the image to have 0.5 mean and 0.5 standard deviation for each of the 3 channels. Feel free to add in other transformations you may find necessary. `torchvision.transforms` is a good point to reference.

Why do we want to normalize the images beforehand?

HINT: consider the fact that the network we developed will be deployed to a large number of images.

You answer here: Normalization standardizes the scale of pixel values, which makes the training process more stable and efficient. This ensures that when the network is deployed on large datasets, it processes inputs consistent with the training data distribution, thereby improving performance and reliability.

[2]:

```python
trainloader = None
testloader = None
batch_size = 4

### YOUR CODE HERE
# Define a transformation to convert images to tensors and normalize them
transform = transforms.Compose([
    transforms.ToTensor(),
    transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))
])

trainset = torchvision.datasets.CIFAR10(root='./data', train=True,
                                        download=True, transform=transform)
trainloader = torch.utils.data.DataLoader(trainset, batch_size=batch_size,
                                          shuffle=True, num_workers=2)

testset = torchvision.datasets.CIFAR10(root='./data', train=False,
                                       download=True, transform=transform)
testloader = torch.utils.data.DataLoader(testset, batch_size=batch_size,
                                         shuffle=False, num_workers=2)
### END YOUR CODE

# these are the 10 classes we have in CIFAR10
classes = ('plane', 'car', 'bird', 'cat',
           'deer', 'dog', 'frog', 'horse', 'ship', 'truck')

# running this block will take a few minutes to download the dataset if you␣
 ↪haven't done so
```

Let's plot out some training images to see what we are dealing with:

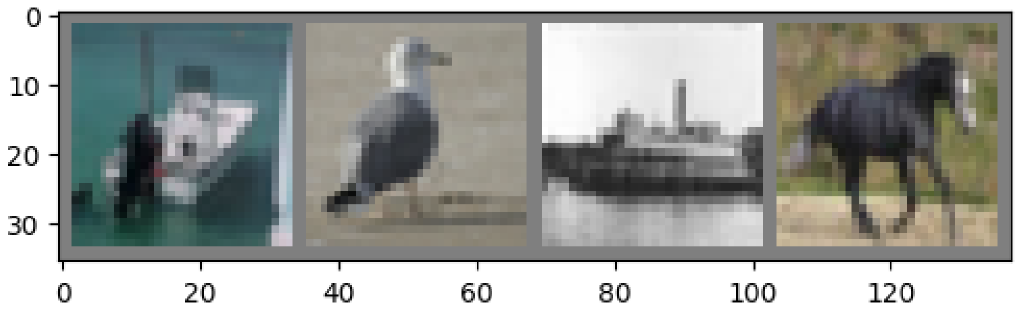
[3]:

```
import matplotlib.pyplot as plt

def imshow(img):
    img = img / 2 + 0.5     # unnormalize
    npimg = img.numpy()
    plt.imshow(np.transpose(npimg, (1, 2, 0)))
    plt.show()

# get some random training images
dataiter = iter(trainloader)
images, labels = next(dataiter)

# show images
imshow(torchvision.utils.make_grid(images))
# print labels
print(' '.join('%5s' % classes[labels[j]] for j in range(batch_size)))
```



```
ship  bird  ship horse
```

2.1.2 1.2 Model definition.

Now we have the data ready, the next step is to define the model that we want to train on these data. Since CIFAR10 is a small dataset, we'll just build a very simple Convolutional Neural Network for our problem. The architecture of it will be (in order):

- 2D convolution: output feature channel number = 6, kernel size = 5x5, stride = 1, no padding;
- 2D max pooling: kernel size = 2x2, stride = 2;
- 2D convolution: output feature channel number = 16, kernel size = 5x5, stride = 1, no padding;
- 2D max pooling: kernel size = 2x2, stride = 2;
- Fully-connected layer: output feature channel number = 120;
- Fully-connected layer: output feature channel number = 84;
- Fully-connected layer: output feature channel number = 10 (number of classes).

Implement the `__init__()` and `forward()` functions in `Net`. As a good practice, `__init__()` generally defines the network architecture and `forward()` takes the runtime input `x` and passes through the network defined in `__init__()`, and returns the output.

[4]:

```python
import torch
import torch.nn as nn
import torch.nn.functional as F

class Net(nn.Module):
    def __init__(self):
        super().__init__()
        ### YOUR CODE HERE
        self.conv1 = nn.Conv2d(3, 6, kernel_size=5)
        self.conv2 = nn.Conv2d(6, 16, kernel_size=5)
        self.pool = nn.MaxPool2d(kernel_size=2, stride=2)
        self.fc1 = nn.Linear(16 * 5 * 5, 120)
        self.fc2 = nn.Linear(120, 84)
        self.fc3 = nn.Linear(84, 10)
        ### END YOUR CODE
    def forward(self, x):
        ### YOUR CODE HERE
        x = self.pool(F.relu(self.conv1(x)))
        x = self.pool(F.relu(self.conv2(x)))
        # Flatten the output tensor
        x = x.view(-1, 16 * 5 * 5)
        x = F.relu(self.fc1(x))
        x = F.relu(self.fc2(x))
        x = self.fc3(x)
        ### END YOUR CODE
        return x

net = Net()

# Set device to GPU if available, else CPU
device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")
net.to(device)
```

```
[4]: Net(
       (conv1): Conv2d(3, 6, kernel_size=(5, 5), stride=(1, 1))
       (conv2): Conv2d(6, 16, kernel_size=(5, 5), stride=(1, 1))
       (pool): MaxPool2d(kernel_size=2, stride=2, padding=0, dilation=1,
     ceil_mode=False)
       (fc1): Linear(in_features=400, out_features=120, bias=True)
       (fc2): Linear(in_features=120, out_features=84, bias=True)
       (fc3): Linear(in_features=84, out_features=10, bias=True)
     )
```

2.1.3 1.3 Loss and optimizer definition.

Okay we now have the model too! The next step is to train the model on the data we have prepared. But before that , we first need to define a loss function and an optimization procedure, which specifies how well our model does and how the training process is carried out, respectively. We'll be using Cross Entropy loss as our loss function and Stochastic Gradient Descent as our optimization algorithm. We will not cover them in detail here but you are welcome to read more on it. (this article and this article from CS231n would be a great point to start).

PyTorch implements very convenient interfaces for loss functions and optimizers, which we have put for you below.

```
[5]: import torch.optim as optim
     import torch.nn as nn

     criterion = nn.CrossEntropyLoss()
     optimizer = optim.SGD(net.parameters(), lr=0.001, momentum=0.9)
```

2.1.4 1.4 Kick start training.

What we have done so far prepares all the necessary pieces for actual training, and now let's kick start the training process! Running this training block should take just several minutes on your CPU.

[6]:

```
import torch

epoch_num = 2
for epoch in range(epoch_num):  # loop over the dataset multiple times
    running_loss = 0.0
    for i, data in enumerate(trainloader, 0):
        # get the inputs; data is a list of [inputs, labels]
        inputs, labels = data[0].to(device), data[1].to(device)

        # zero the parameter gradients
        optimizer.zero_grad()

        # forward + backward + optimize
        outputs = net(inputs)
        loss = criterion(outputs, labels)
        loss.backward()
        optimizer.step()

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

print('Finished Training')
```

```
[1,  2000] loss: 2.272
[1,  4000] loss: 1.904
[1,  6000] loss: 1.674
[1,  8000] loss: 1.587
[1, 10000] loss: 1.522
[1, 12000] loss: 1.463
[2,  2000] loss: 1.384
[2,  4000] loss: 1.355
[2,  6000] loss: 1.357
[2,  8000] loss: 1.332
[2, 10000] loss: 1.302
[2, 12000] loss: 1.265
Finished Training
```

The last step of training is to save the trained model locally to a checkpoint:

[7]:

```
PATH = './cifar_net.pth'
torch.save(net.state_dict(), PATH)
```

2.1.5 1.5 Test the trained model on the test data.

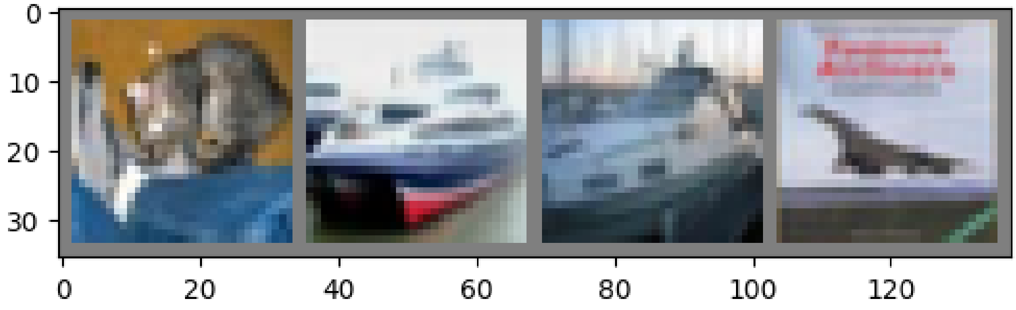
Remember earlier we split the data into training and testing set? Now we'll be using the testing split to see how our model performs on unseen data. We'll check this by predicting the class label that the neural network outputs, and comparing it against the ground-truth.

Let's first examine some data from the testing set:

[8]:

```
dataiter = iter(testloader)
images, labels = next(dataiter)

# print images
imshow(torchvision.utils.make_grid(images))
print('GroundTruth: ', ' '.join('%5s' % classes[labels[j]] for j in range(4)))
```



```
GroundTruth:    cat  ship  ship plane
```

Now, let's load in our saved model checkpoint and get its output:

[9]:

```
# load in model checkpoint
net = Net()
net.to(device)
net.load_state_dict(torch.load(PATH))
```

[9]:

```
<All keys matched successfully>
```

[12]:

```
# First, get the output from the model by passing in `images`;
# Next, think about what the model outputs mean / represent, and convert it to␣
 ↪the predicted class index (`predicted`);
# Finally, output the predicted class label (already done for you).

predicted = []
### YOUR CODE HERE
```

```
images = images.to(device)
outputs = net(images)
_, predicted = torch.max(outputs, 1)
### END YOUR CODE
print('Predicted: ', ' '.join('%5s' % classes[predicted[j]]
                              for j in range(4)))
```

```
Predicted:    cat   car  ship plane
```

How does your prediction look like? Does that match your expectation? Write a few sentences to describe what you got and provide some analysis if you have any.

You answer here: It looks like the predictions are reasonably well: three out of four are correct, which indicates that the model is making reasonable guesses even if not perfect. For example, the model predicted that the boat to a car suggests that while the network is capturing some key features, there might be ambiguity in the learned representations or challenges in distinguishing similar-looking objects.

Besides inspecting these several examples, let's also look at how the network performs on the entire testing set by calculating the percentage of correctly classified examples.

```
[14]: correct = 0
total = 0
# since we're not training, we don't need to calculate the gradients for our␣
 ↪outputs
with torch.no_grad():
    for data in testloader:
        images, labels = data
        # Similar to the previous question, calculate model's output and the␣
 ↪percentage as correct / total
        ### YOUR CODE HERE
        images = images.to(device)
        labels = labels.to(device)
        outputs = net(images)

        # The class with the highest score is our prediction
        _, predicted_labels = torch.max(outputs, 1)

        # Update total count and correct predictions count
        total += labels.size(0)
        correct += (predicted_labels == labels).sum().item()
        ### END YOUR CODE

print('Accuracy of the network on the 10000 test images: %d %%' % (
    100 * correct / total))
```

```
Accuracy of the network on the 10000 test images: 54 %
```

What accuracy did you get? Compared to random guessing, does your model perform significantly better?

You answer here: The accuracy I obtained is around 54%. Compared to random guessing—which on a 10-class problem would give an expected accuracy of about 10%—this result indicates that the model is performing significantly better than chance.

Let's do some analysis to gain more insights of the results. One analysis we can carry out is the accuracy for each class, which can tell us what classes our model did well, and what classes our model did poorly.

[15]:

```
# prepare to count predictions for each class
correct_pred = {classname: 0 for classname in classes}
total_pred = {classname: 0 for classname in classes}

with torch.no_grad():
    for data in testloader:
        images, labels = data
        # repeat what you did previously, but now for each class
        ### YOUR CODE HERE
        images, labels = images.to(device), labels.to(device)

        # Get outputs from the model
        outputs = net(images)

        # The class with the highest score is our prediction
        _, predicted_labels = torch.max(outputs, 1)

        # Update total count and correct predictions count
        for i in range(len(labels)):
            label = labels[i].item()  # ground truth class index
            pred = predicted_labels[i].item()  # predicted class index
            # Increment total predictions for the actual class
            total_pred[classes[label]] += 1
            # Increment correct predictions if prediction matches the label
            if pred == label:
                correct_pred[classes[label]] += 1
        ### END YOUR CODE

# print accuracy for each class
for classname, correct_count in correct_pred.items():
    accuracy = 100 * float(correct_count) / total_pred[classname]
    print("Accuracy for class {:5s} is: {:.1f} %".format(classname,
                                                       accuracy))
```

```
Accuracy for class plane is: 65.8 %
Accuracy for class car   is: 79.5 %
Accuracy for class bird  is: 22.1 %
Accuracy for class cat   is: 41.4 %
Accuracy for class deer  is: 47.5 %
Accuracy for class dog   is: 50.7 %
Accuracy for class frog  is: 67.5 %
Accuracy for class horse is: 54.0 %
Accuracy for class ship  is: 72.8 %
Accuracy for class truck is: 48.1 %
```

2.1.6 1.6 Hyper-parameter tuning.

An important phase in deep learning framework is hyper-parameter search. Hyper-parameters generally refer to those parameters that are **not** automatically optimized during the learning process, e.g., model architecture, optimizer, learning rate, batch size, training length, etc. Tuning these hyper-parameters could often lead to significant improvement of your model performance.

Your job in this section is to identify the hyper-parameters and tune them to improve the model performance as much as possible. You might want to refer to PyTorch documentation or other online resources to gain an understanding of what these hyper-parameters mean. Some of the options you might want to look into are: * Model architecture (number of layers, layer size, feature number, etc.); * Loss and optimizer (including loss function, regularization, learning rate, learning rate decay, etc.); * Training configuration (batch size, epoch number, etc.). These are by no means a complete list, but is supposed to give you an idea of the hyper-parameters. You are encouraged to identify and tune more.

Report in detail what you did in this section. Which of them improved model performance, and which did not?

You answer here: We tuned key hyperparameters such as the learning rate, momentum, batch size, and number of training epochs using grid search. Generally, training for longer epoch increase the performance and too large learning rate reduce the performance.

[17]:
```
from itertools import product

# Define the grid of hyper-parameters
learning_rates = [0.001, 0.01, 0.1]
batch_sizes = [4, 8, 16]
epoch_nums = [2, 5, 10]

device = torch.device("cuda:0" if torch.cuda.is_available() else "cpu")
criterion = nn.CrossEntropyLoss()

results = {}

# Grid search over all hyper-parameter combinations
for lr, bs, epoch_num in product(learning_rates, batch_sizes, epoch_nums):
    # Setup dataloaders with current batch size
    trainloader = torch.utils.data.DataLoader(trainset, batch_size=bs,␣
 ↪shuffle=True, num_workers=10)
    testloader = torch.utils.data.DataLoader(testset, batch_size=bs,␣
 ↪shuffle=False, num_workers=10)

    # Initialize the model and move it to the appropriate device
    net = Net().to(device)
```

```python
    optimizer = optim.SGD(net.parameters(), lr=lr, momentum=0.9)

    # Train the model for a given number of epochs
    net.train()
    for epoch in range(epoch_num):
        running_loss = 0.0
        for i, data in enumerate(trainloader, 0):
            inputs, labels = data[0].to(device), data[1].to(device)
            optimizer.zero_grad()
            outputs = net(inputs)
            loss = criterion(outputs, labels)
            loss.backward()
            optimizer.step()
            running_loss += loss.item()

    # Evaluate the model on the test set
    net.eval()
    correct = 0
    total = 0
    with torch.no_grad():
        for data in testloader:
            images, labels = data[0].to(device), data[1].to(device)
            outputs = net(images)
            _, predicted = torch.max(outputs, 1)
            total += labels.size(0)
            correct += (predicted == labels).sum().item()

    accuracy = 100 * correct / total
    results[(lr, bs, epoch_num)] = accuracy
    print(f"lr: {lr}, batch_size: {bs}, epochs: {epoch_num}, accuracy: {accuracy:.2f}%")

# Find the best hyper-parameter configuration
best_params = max(results, key=results.get)
print("\nBest hyper-parameters:")
print(f"Learning Rate: {best_params[0]}, Momentum: {best_params[1]}, Batch Size: {best_params[2]}, Epochs: {best_params[3]}")
print(f"Best Accuracy: {results[best_params]:.2f}%")
```

```
lr: 0.001, batch_size: 4, epochs: 2, accuracy: 56.31%
lr: 0.001, batch_size: 4, epochs: 5, accuracy: 60.70%
lr: 0.001, batch_size: 4, epochs: 10, accuracy: 61.81%
lr: 0.001, batch_size: 8, epochs: 2, accuracy: 54.31%
lr: 0.001, batch_size: 8, epochs: 5, accuracy: 60.84%
lr: 0.001, batch_size: 8, epochs: 10, accuracy: 64.73%
lr: 0.001, batch_size: 16, epochs: 2, accuracy: 46.75%
lr: 0.001, batch_size: 16, epochs: 5, accuracy: 58.44%
```

```
lr: 0.001, batch_size: 16, epochs: 10, accuracy: 63.46%
lr: 0.01, batch_size: 4, epochs: 2, accuracy: 26.95%
lr: 0.01, batch_size: 4, epochs: 5, accuracy: 26.61%
lr: 0.01, batch_size: 4, epochs: 10, accuracy: 22.48%
lr: 0.01, batch_size: 8, epochs: 2, accuracy: 43.50%
lr: 0.01, batch_size: 8, epochs: 5, accuracy: 45.66%
lr: 0.01, batch_size: 8, epochs: 10, accuracy: 38.97%
lr: 0.01, batch_size: 16, epochs: 2, accuracy: 53.55%
lr: 0.01, batch_size: 16, epochs: 5, accuracy: 55.75%
lr: 0.01, batch_size: 16, epochs: 10, accuracy: 56.14%
lr: 0.1, batch_size: 4, epochs: 2, accuracy: 10.00%
lr: 0.1, batch_size: 4, epochs: 5, accuracy: 10.00%
lr: 0.1, batch_size: 4, epochs: 10, accuracy: 10.00%
lr: 0.1, batch_size: 8, epochs: 2, accuracy: 10.00%
lr: 0.1, batch_size: 8, epochs: 5, accuracy: 10.00%
lr: 0.1, batch_size: 8, epochs: 10, accuracy: 10.00%
lr: 0.1, batch_size: 16, epochs: 2, accuracy: 10.00%
lr: 0.1, batch_size: 16, epochs: 5, accuracy: 10.00%
lr: 0.1, batch_size: 16, epochs: 10, accuracy: 10.00%

Best hyper-parameters:
Learning Rate: 0.001, Batch Size: 8, Epochs: 10
Best Accuracy: 64.73%
```

2.2 2. Extra Credit: further improve your model performance

You have just tried tuning the hyper-parameters to improve your model performance. It's a very important part but not all! In this section, you are encouraged to read online to explore other options to further enhance your model. You may or may not need additional compute resources depending on what you do. But if you do need GPUs, Google Colab could be a great point to start.

Since this a free-form section, you should report here in detail what you have done, and feel free to submit any additional files if needed (e.g., additional code files). We'll be grading based on the effort you spend and the performance you achieved.

[ ]:

```python
from torchvision.datasets import CIFAR10
from torch.utils.data import DataLoader
from tqdm import tqdm

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

transform = transforms.Compose([
    transforms.Resize(224),
    transforms.ToTensor(),
    transforms.Normalize((0.5, 0.5, 0.5), (0.5, 0.5, 0.5))
])
```

```
trainset = CIFAR10(root='./data', train=True, download=True,␣
 ↪transform=transform)
testset = CIFAR10(root='./data', train=False, download=True,␣
 ↪transform=transform)

trainloader = DataLoader(trainset, batch_size=512, shuffle=True,␣
 ↪num_workers=10, pin_memory=True)
testloader = DataLoader(testset, batch_size=512, shuffle=False, num_workers=10,␣
 ↪pin_memory=True)

dinov2 = torch.hub.load("facebookresearch/dinov2", "dinov2_vitb14",␣
 ↪pretrained=True)
dinov2.to(device)
dinov2.eval()

# Freeze DINOv2 parameters
for param in dinov2.parameters():
    param.requires_grad = False

class DINOv2Classifier(nn.Module):
    def __init__(self, feature_dim=768, num_classes=10):
        super().__init__()
        self.fc = nn.Linear(feature_dim, num_classes)

    def forward(self, x):
        features = dinov2(x)  # shape: (batch_size, feature_dim)
        logits = self.fc(features)
        return logits

model = DINOv2Classifier(feature_dim=768, num_classes=10).to(device)
criterion = nn.CrossEntropyLoss()
# optimizer = optim.SGD(model.parameters(), lr=0.01, momentum=0.9)
optimizer = torch.optim.AdamW(model.parameters(), lr=1e-3, weight_decay=1e-2,)

epoch_num = 5
for epoch in range(epoch_num):
    model.train()
    running_loss = 0.0
    for images, labels in tqdm(trainloader, desc=f"Epoch {epoch+1}/{epoch_num}␣
 ↪Training", leave=False):
        images, labels = images.to(device), labels.to(device)
        optimizer.zero_grad()
        outputs = model(images)
        loss = criterion(outputs, labels)
        loss.backward()
        optimizer.step()
        running_loss += loss.item()
    avg_loss = running_loss / len(trainloader)
    print(f"Epoch [{epoch+1}/{epoch_num}], Loss: {avg_loss:.4f}")

model.eval()
correct = 0
total = 0
with torch.no_grad():
    for images, labels in tqdm(testloader, desc="Evaluating", leave=False):
        images, labels = images.to(device), labels.to(device)
        outputs = model(images)
        _, predicted = torch.max(outputs, 1)
        total += labels.size(0)
        correct += (predicted == labels).sum().item()

accuracy = 100 * correct / total
print(f'Accuracy of the DINOv2-based classifier on the test images: {accuracy:.
 ↪2f} %')
```

```
Using cache found in /orion/u/yixinli/hub/facebookresearch_dinov2_main
/orion/u/yixinli/hub/facebookresearch_dinov2_main/dinov2/layers/swiglu_ffn.py:51
: UserWarning: xFormers is not available (SwiGLU)
  warnings.warn("xFormers is not available (SwiGLU)")
/orion/u/yixinli/hub/facebookresearch_dinov2_main/dinov2/layers/attention.py:33:
UserWarning: xFormers is not available (Attention)
  warnings.warn("xFormers is not available (Attention)")
/orion/u/yixinli/hub/facebookresearch_dinov2_main/dinov2/layers/block.py:40:
UserWarning: xFormers is not available (Block)
  warnings.warn("xFormers is not available (Block)")
```

```
Epoch [1/5], Loss: 0.1802
```

```
Epoch [2/5], Loss: 0.0676
```

```
Epoch [3/5], Loss: 0.0587
```

```
Epoch [4/5], Loss: 0.0529
```

```
Epoch [5/5], Loss: 0.0489
```

```
Accuracy of the DINOv2-based classifier on the test images: 97.52 %
```