Start by importing necessary packages

You will begin by importing necessary libraries for this notebook. Run the cell below to do so.

PyTorch and Intro to Training

```
!pip install thop
import math
import numpy as np
import torch
import torch.nn as nn
import torch.optim as optim
from torchvision import datasets, transforms
import thop
import matplotlib.pyplot as plt
from tqdm import tqdm
import time
```

```
Collecting thop
  Downloading thop-0.1.1.post2209072238-py3-none-any.whl.metadata (2.7 kB)
Requirement already satisfied: torch in /usr/local/lib/python3.11/dist-packages (from
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (f
Requirement already satisfied: typing-extensions>=4.8.0 in /usr/local/lib/python3.11/
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (f
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (frc
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (frc
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.1.105 in /usr/local/lib/pyt
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.1.105 in /usr/local/lib/p
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.1.105 in /usr/local/lib/pyt
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.
Requirement already satisfied: nvidia-cublas-cu12==12.1.3.1 in /usr/local/lib/python3
Requirement already satisfied: nvidia-cufft-cu12==11.0.2.54 in /usr/local/lib/python3
Requirement already satisfied: nvidia-curand-cu12==10.3.2.106 in /usr/local/lib/pythc
Requirement already satisfied: nvidia-cusolver-cu12==11.4.5.107 in /usr/local/lib/pyt
Requirement already satisfied: nvidia-cusparse-cu12==12.1.0.106 in /usr/local/lib/pyt
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/
Requirement already satisfied: nvidia-nvtx-cu12==12.1.105 in /usr/local/lib/python3.1
Requirement already satisfied: triton==3.1.0 in /usr/local/lib/python3.11/dist-packag
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packag
Requirement already satisfied: nvidia-nvjitlink-cu12 in /usr/local/lib/python3.11/dis
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-pack
Downloading thop-0.1.1.post2209072238-py3-none-any.whl (15 kB)
Installing collected packages: thop
Successfully installed thop-0.1.1.post2209072238
```

Checking the torch version and CUDA access

Let's start off by checking the current torch version, and whether you have CUDA availablity.

```
print("torch is using version:", torch.__version__, "with CUDA=", torch.cuda.is_available
```

```
torch is using version: 2.5.1+cu121 with CUDA= False
```

By default, you will see CUDA = False, meaning that the Colab session does not have access to a GPU. To remedy this, click the Runtime menu on top and select "Change Runtime Type", then select "T4 GPU".

Re-run the import cell above, and the CUDA version / check. It should show now CUDA = True

Sometimes in Colab you get a message that your Session has crashed, if that happens you need to go to the Runtime menu on top and select "Restart session".

You won't be using the GPU just yet, but this prepares the instance for when you will.

Please note that the GPU is a scarce resource which may not be available at all time. Additionally, there are also usage limits that you may run into (although not likely for this assignment). When that happens you need to try again later/next day/different time of the day. Another reason to start the assignment early!

A Brief Introduction to PyTorch

PyTorch, or torch, is a machine learning framework developed my Facebook AI Research, which competes with TensorFlow, JAX, Caffe and others.

Roughly speaking, these frameworks can be split into dynamic and static defintion frameworks.

Static Network Definition: The architecture and computation flow are defined simultaneously. The order and manner in which data flows through the layers are fixed upon definition. These frameworks also tend to declare parameter shapes implicitly via the compute graph. This is typical of TensorFlow and JAX.

Dynamic Network Definition: The architecture (layers/modules) is defined independently of the computation flow, often during the object's initialization. This allows for dynamic computation graphs where the flow of data can change during runtime based on conditions. Since the network exists independent of the compute graph, the parameter shapes must be declared explitly. PyTorch follows this approach.

All ML frameworks support automatic differentiation, which is necessary to train a model (i.e. perform back propagation).

Let's consider a typical pytorch module. Such modules will inherit from the torch.nn.Module class, which provides many built in functions such as a wrapper for `__call__`, operations to move the module between devices (e.g. `cuda()`, `cpu()`), data-type conversion (e.g. `half()`, `float()`), and parameter and child management (e.g. `state_dict()`, `parameters()`).

```
# inherit from torch.nn.Module
class MyModule(nn.Module):
  # constructor called upon creation
  def __init__(self):
    # the module has to initialize the parent first, which is what sets up the wrapper be
    super().__init__()

    # we can add sub-modules and parameters by assigning them to self
    self.my_param = nn.Parameter(torch.zeros(4,8)) # this is how you define a raw paramet
    self.my_sub_module = nn.Linear(8,12)       # this is how you define a linear layer (t

    # we can also add lists of modules, for example, the sequential layer
    self.net = nn.Sequential(  # this layer type takes in a collection of modules rather
        nn.Linear(4,4),
        nn.Linear(4,8),
        nn.Linear(8,12)
    )

    # the above when calling self.net(x), will execute each module in the order they appe
    # it would be equivelent to x = self.net[2](self.net[1](self.net[0](x)))

    # you can also create a list that doesn't execute
    self.net_list = nn.ModuleList([
        nn.Linear(7,7),
        nn.Linear(7,9),
        nn.Linear(9,14)
    ])

    # sometimes you will also see constant variables added to the module post init
    foo = torch.Tensor([4])
    self.register_buffer('foo', foo) # buffers allow .to(device, type) to apply

  # let's define a forward function, which gets executed when calling the module, and def
  def forward(self, x):

    # if x is of shape Bx4
    h1 =  x @ self.my_param # tensor-tensor multiplication uses the @ symbol
    # then h1 is now shape Bx8, because my_param is 4x8... 2x4 * 4x8 = 2x8

    h1 = self.my_sub_module(h1) # you execute a sub-module by calling it
    # now, h1 is of shape Bx12, because my_sub_module was a 8x12 matrix

    h2 = self.net(x)
    # similarly, h2 is of shape Bx12, because that's the output of the sequence
    # Bx4 -(4x4)-> Bx4 -(4x8)-> Bx8 -(8x12)-> Bx12

    # since h1 and h2 are the same shape, they can be added together element-wise
    return h1 + h2
```

Then you can instantiate the module and perform a forward pass by calling it.

```
# create the module
module = MyModule()

# you can print the module to get a high-level summary of it
print("=== printing the module ===")
print(module)
print()
# notice that the sub-module name is in parenthesis, and so are the list indicies

# let's view the shape of one of the weight tensors
print("my_sub_module weight tensor shape=", module.my_sub_module.weight.shape)
# the above works because nn.Linear has a member called .weight and .bias
# to view the shape of my_param, you would use module.my_param
# and to view the shape of the 2nd elment in net_list, you would use module.net_list[1].w

# you can iterate through all of the parameters via the state dict
print()
print("=== Listing parameters from the state_dict ===")
for key,value in module.state_dict().items():
  print(f"{key}: {value.shape}")
```

```
=== printing the module ===
MyModule(
  (my_sub_module): Linear(in_features=8, out_features=12, bias=True)
  (net): Sequential(
    (0): Linear(in_features=4, out_features=4, bias=True)
    (1): Linear(in_features=4, out_features=8, bias=True)
    (2): Linear(in_features=8, out_features=12, bias=True)
  )
  (net_list): ModuleList(
    (0): Linear(in_features=7, out_features=7, bias=True)
    (1): Linear(in_features=7, out_features=9, bias=True)
    (2): Linear(in_features=9, out_features=14, bias=True)
  )
)

my_sub_module weight tensor shape= torch.Size([12, 8])

=== Listing parameters from the state_dict ===
my_param: torch.Size([4, 8])
foo: torch.Size([1])
my_sub_module.weight: torch.Size([12, 8])
my_sub_module.bias: torch.Size([12])
net.0.weight: torch.Size([4, 4])
net.0.bias: torch.Size([4])
net.1.weight: torch.Size([8, 4])
net.1.bias: torch.Size([8])
net.2.weight: torch.Size([12, 8])
net.2.bias: torch.Size([12])
net_list.0.weight: torch.Size([7, 7])
net_list.0.bias: torch.Size([7])
net_list.1.weight: torch.Size([9, 7])
net_list.1.bias: torch.Size([9])
net_list.2.weight: torch.Size([14, 9])
net_list.2.bias: torch.Size([14])
```

```
# you can perform a forward pass by first creating a tensor to send through
x = torch.zeros(2,4)
# then you call the module (this invokes MyModule.forward() )
y = module(x)

# then you can print the result and shape
print(y, y.shape)
```

```
tensor([[ 0.3421,  0.2986, -0.0331, -0.2339, -0.2135, -0.1588, -0.3582,  0.3993,
         -0.1514, -0.1713,  0.4566,  0.2321],
        [ 0.3421,  0.2986, -0.0331, -0.2339, -0.2135, -0.1588, -0.3582,  0.3993,
         -0.1514, -0.1713,  0.4566,  0.2321]], grad_fn=<AddBackward0>) torch.Size([2,
```

Please check the cell below to notice the following:

1. `x` above was created with the shape 2x4, and in the forward pass, it gets manipulated into a 2x12 tensor. This last dimension is explicit, while the first is called the batch dimmension, and only exists on data (a.k.a. activations). The output shape can be seen in the print statement from y.shape
2. You can view the shape of a tensor by using `.shape`, this is a very helpful trick for debugging tensor shape errors
3. In the output, there's a `grad_fn` component, this is the hook created by the forward trace to be used in back-propagation via automatic differentiation. The function name is `AddBackward`, because the last operation performed was `h1+h2`.

We might not always want to trace the compute graph though, such as during inference. In such cases, you can use the `torch.no_grad()` context manager.

```
# you can perform a forward pass by first creating a tensor to send through
x = torch.zeros(2,4)
# then you call the module (this invokes MyModule.forward() )
with torch.no_grad():
  y = module(x)

# then you can print the result and shape
print(y, y.shape)
# notice how the grad_fn is no longer part of the output tensor, that's because not_grad(
```

```
tensor([[ 0.3421,  0.2986, -0.0331, -0.2339, -0.2135, -0.1588, -0.3582,  0.3993,
         -0.1514, -0.1713,  0.4566,  0.2321],
        [ 0.3421,  0.2986, -0.0331, -0.2339, -0.2135, -0.1588, -0.3582,  0.3993,
         -0.1514, -0.1713,  0.4566,  0.2321]]) torch.Size([2, 12])
```

Aside from passing a tensor through a model with the `no_grad()` context, you can also detach a tensor from the compute graph by calling `.detach()`. This will effectively make a copy of the original tensor, which allows it to be converted to numpy and visualized with matplotlib.

Note: Tensors with a `grad_fn` property cannot be plotted and must first be detached.

Multi-Layer-Perceptron (MLP) Prediction of MNIST

With some basics out of the way, let's create a MLP for training MNIST. You can start by defining a simple torch model.

```
# Define the MLP model
class MLP(nn.Module):
    # define the constructor for the network
    def __init__(self):
        super().__init__()
        # the input projection layer - projects into d=128
        self.fc1 = nn.Linear(28*28, 128)
        # the first hidden layer - compresses into d=64
        self.fc2 = nn.Linear(128, 64)
        # the final output layer - splits into 10 classes (digits 0-9)
        self.fc3 = nn.Linear(64, 10)

    # define the forward pass compute graph
    def forward(self, x):
        # x is of shape BxHxW

        # we first need to unroll the 2D image using view
        # we set the first dim to be -1 meanining "everything else", the reason being tha
        # we want to maintain different tensors for each training sample in the batch, wh
        x = x.view(-1, 28*28)
        # x is of shape Bx784

        # project-in and apply a non-linearity (ReLU activation function)
        x = torch.relu(self.fc1(x))
        # x is of shape Bx128

        # middle-layer and apply a non-linearity (ReLU activation function)
        x = torch.relu(self.fc2(x))
        # x is of shape Bx64

        # project out into the 10 classes
        x = self.fc3(x)
        # x is of shape Bx10
        return x
```

Before you can begin training, you have to do a little boiler-plate to load the dataset. From the previous assignment, you saw how a hosted dataset can be loaded with TensorFlow. With pytorch it's a little more complicated, as you need to manually condition the input data.

```
# define a transformation for the input images. This uses torchvision.transforms, and .Co
transform = transforms.Compose([
    transforms.ToTensor(), # first convert to a torch tensor
    transforms.Normalize((0.1307,), (0.3081,)) # then normalize the input
])

# let's download the train and test datasets, applying the above transform - this will ge
train_dataset = datasets.MNIST('./data', train=True, download=True, transform=transform)
test_dataset = datasets.MNIST('./data', train=False, transform=transform)

# we need to set the mini-batch (commonly referred to as "batch"), for now we can use 64
batch_size = 64

# then we need to create a dataloader for the train dataset, and we will also create one
# additionally, we will set the batch size in the dataloader
train_loader = torch.utils.data.DataLoader(train_dataset, batch_size=batch_size, shuffle=
test_loader = torch.utils.data.DataLoader(test_dataset, batch_size=batch_size, shuffle=Fa

# the torch dataloaders allow us to access the __getitem__ method, which returns a tuple
# additionally, the dataloader will pre-colate the training samples into the given batch_
```

```
Downloading http://yann.lecun.com/exdb/mnist/train-images-idx3-ubyte.gz
Failed to download (trying next):
<urlopen error [Errno 110] Connection timed out>

Downloading https://ossci-datasets.s3.amazonaws.com/mnist/train-images-idx3-ubyte.gz
Downloading https://ossci-datasets.s3.amazonaws.com/mnist/train-images-idx3-ubyte.gz
100%|██████████| 9.91M/9.91M [00:00<00:00, 94.1MB/s]
Extracting ./data/MNIST/raw/train-images-idx3-ubyte.gz to ./data/MNIST/raw

Downloading http://yann.lecun.com/exdb/mnist/train-labels-idx1-ubyte.gz
Failed to download (trying next):
<urlopen error [Errno 110] Connection timed out>

Downloading https://ossci-datasets.s3.amazonaws.com/mnist/train-labels-idx1-ubyte.gz
Downloading https://ossci-datasets.s3.amazonaws.com/mnist/train-labels-idx1-ubyte.gz
100%|██████████| 28.9k/28.9k [00:00<00:00, 24.6MB/s]Extracting ./data/MNIST/raw/trair

Downloading http://yann.lecun.com/exdb/mnist/t10k-images-idx3-ubyte.gz

Failed to download (trying next):
<urlopen error [Errno 110] Connection timed out>

Downloading https://ossci-datasets.s3.amazonaws.com/mnist/t10k-images-idx3-ubyte.gz
Downloading https://ossci-datasets.s3.amazonaws.com/mnist/t10k-images-idx3-ubyte.gz t
100%|██████████| 1.65M/1.65M [00:00<00:00, 101MB/s]Extracting ./data/MNIST/raw/t10k-i

Downloading http://yann.lecun.com/exdb/mnist/t10k-labels-idx1-ubyte.gz

Failed to download (trying next):
<urlopen error [Errno 110] Connection timed out>

Downloading https://ossci-datasets.s3.amazonaws.com/mnist/t10k-labels-idx1-ubyte.gz
Downloading https://ossci-datasets.s3.amazonaws.com/mnist/t10k-labels-idx1-ubyte.gz t
100%|██████████| 4.54k/4.54k [00:00<00:00, 3.21MB/s]Extracting ./data/MNIST/raw/t10k-
```

Inspect the first element of the test_loader, and verify both the tensor shapes and data types. You can check the data-type with `.dtype`

Question 1

Edit the cell below to print out the first element shapes, dtype, and identify which is the training sample and which is the training label.

```
# Get the first item
first_item = next(iter(test_loader))

# print out the element shapes, dtype, and identify which is the training sample and whic
# MNIST is a supervised learning task
input_sample, label = first_item

# Print the shapes, dtypes, and identify training sample and label
print("Training sample (input):")
print(f"Shape: {input_sample.shape}, Dtype: {input_sample.dtype}")

print("\nTraining label:")
print(f"Value: {label}, Dtype: {label.dtype}")
```

```
Training sample (input):
Shape: torch.Size([64, 1, 28, 28]), Dtype: torch.float32

Training label:
Value: tensor([7, 2, 1, 0, 4, 1, 4, 9, 5, 9, 0, 6, 9, 0, 1, 5, 9, 7, 3, 4, 9, 6, 6, 5
        4, 0, 7, 4, 0, 1, 3, 1, 3, 4, 7, 2, 7, 1, 2, 1, 1, 7, 4, 2, 3, 5, 1, 2,
        4, 4, 6, 3, 5, 5, 6, 0, 4, 1, 9, 5, 7, 8, 9, 3]), Dtype: torch.int64
```

Now that we have the dataset loaded, we can instantiate the MLP model, the loss (or criterion function), and the optimizer for training.

```
# create the model
model = MLP()

# you can print the model as well, but notice how the activation functions are missing. T
# and not declared in the constructor
print(model)

# you can also count the model parameters
param_count = sum([p.numel() for p in model.parameters()])
print(f"Model has {param_count:,} trainable parameters")

# for a critereon (loss) function, you will use Cross-Entropy Loss. This is the most comm
# and is also used by tokenized transformer models it takes in an un-normalized probabili
# N classes (in our case, 10 classes with MNIST). This distribution is then compared to a
# For MNIST, the prediction might be [-0.0056, -0.2044,  1.1726,  0.0859,  1.8443, -0.962
# Cross-entropy can be thought of as finding the difference between the predicted distrib

criterion = nn.CrossEntropyLoss()

# then you can instantiate the optimizer. You will use Stochastic Gradient Descent (SGD),
# factor of 0.5. the first input to the optimizer is the list of model parameters, which
optimizer = optim.SGD(model.parameters(), lr=0.01, momentum=0.5)
```

```
MLP(
  (fc1): Linear(in_features=784, out_features=128, bias=True)
  (fc2): Linear(in_features=128, out_features=64, bias=True)
  (fc3): Linear(in_features=64, out_features=10, bias=True)
)
Model has 109,386 trainable parameters
```

Finally, you can define a training, and test loop

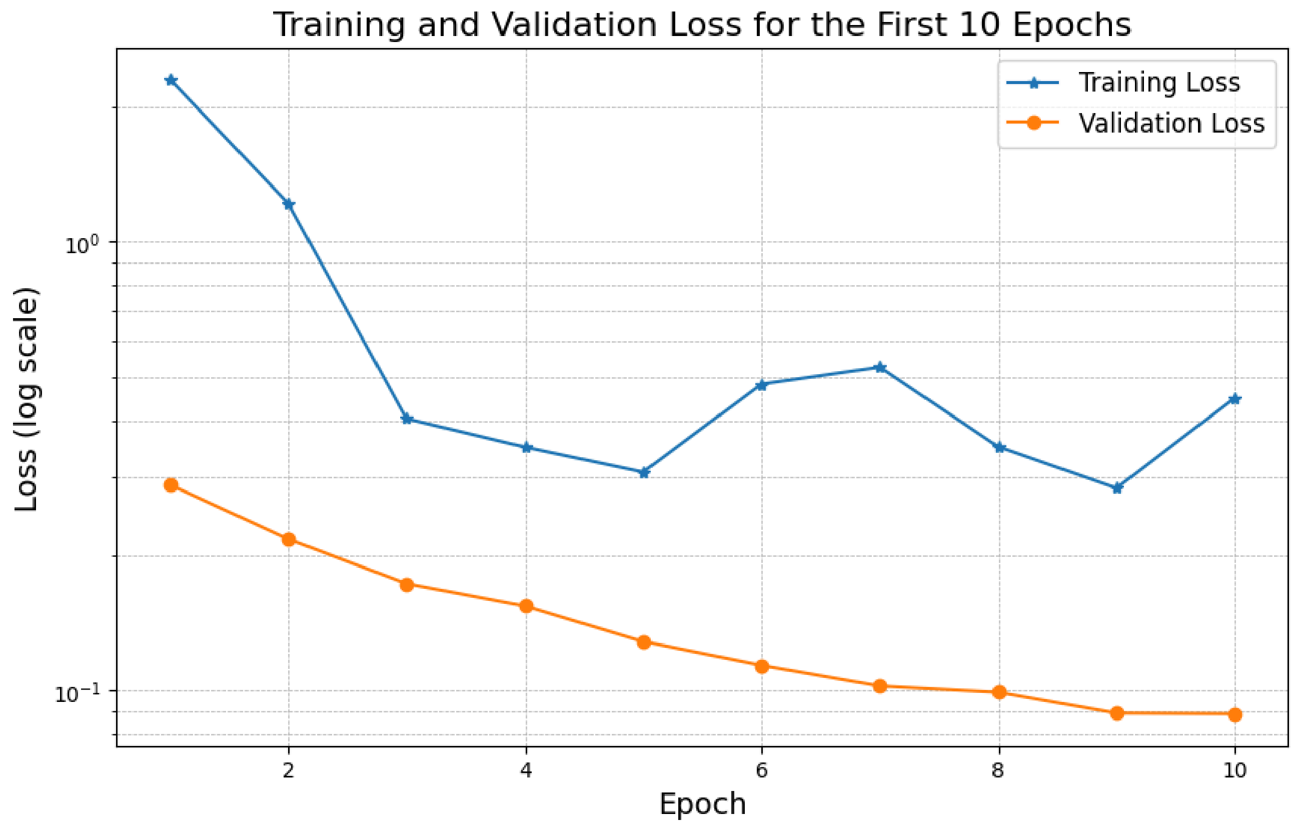
```
# create an array to log the loss and accuracy
train_losses = []
train_steps = []
test_steps = []
test_losses = []
test_accuracy = []
current_step = 0  # Start with global step 0
current_epoch = 0 # Start with epoch 0


# declare the train function
def cpu_train(epoch, train_losses, steps, current_step):

    # set the model in training mode - this doesn't do anything for us right now, but it
    # batch norm and dropout
    model.train()

    # Create tqdm progress bar to help keep track of the training progress
    pbar = tqdm(enumerate(train_loader), total=len(train_loader))

    # loop over the dataset. Recall what comes out of the data loader, and then by wrappi
    # iterator list which we will call batch_idx
    for batch_idx, (data, target) in pbar:

        # during training, the first step is to zero all of the gradients through the opt
        # this resets the state so that we can begin back propogation with the updated pa
        optimizer.zero_grad()

        # then you can apply a forward pass, which includes evaluating the loss (criteric
        output = model(data)
        loss = criterion(output, target)

        # given that you want to minimize the loss, you need to call .backward() on the r
        loss.backward()
        # the backward step will automatically differentiate the model and apply a gradie
        # so then all you have to do is call optimizer.step() to apply the gradients to t
        optimizer.step()

        # increment the step count
        current_step += 1

        # you should add some output to the progress bar so that you know which epoch you
        if batch_idx % 100 == 0:

            # append the last loss value
            train_losses.append(loss.item())
            steps.append(current_step)

            desc = (f'Train Epoch: {epoch} [{batch_idx * len(data)}/{len(train_loader.dat
                    f' ({100. * batch_idx / len(train_loader):.0f}%)]\tLoss: {loss.item()
            pbar.set_description(desc)

    return current_step

# declare a test function, this will help you evaluate the model progress on a dataset wh
# doing so prevents cross-contamination and misleading results due to overfitting
def cpu_test(test_losses, test_accuracy, steps, current_step):

    # put the model into eval mode, this again does not currently do anything for you, bu
    # and dropout
    model.eval()
    test_loss = 0
    correct = 0

    # Create tqdm progress bar
    pbar = tqdm(test_loader, total=len(test_loader), desc="Testing...")

    # since you are not training the model, and do not need back-propagation, you can use
    with torch.no_grad():
        # iterate over the test set
        for data, target in pbar:
            # like with training, run a forward pass through the model and evaluate the c
            output = model(data)
            test_loss += criterion(output, target).item() # you are using .item() to get

            # you can also check the accuracy by sampling the output - you can use greedy
            # in general, you would want to normalize the logits first (the un-normalized
            # however, argmax is taking the maximum value, which will be the same index f
            # so we can skip a step and take argmax directly
            pred = output.argmax(dim=1, keepdim=True)
            correct += pred.eq(target.view_as(pred)).sum().item()

    test_loss /= len(test_loader)

    # append the final test loss
    test_losses.append(test_loss)
    test_accuracy.append(correct/len(test_loader.dataset))
    steps.append(current_step)
    print(f'\nTest set: Average loss: {test_loss:.4f}, Accuracy: {correct}/{len(test_load
          f' ({100. * correct / len(test_loader.dataset):.0f}%)\n')



# train for 10 epochs
for epoch in range(0, 10):
    current_step = cpu_train(current_epoch, train_losses, train_steps, current_step)
    cpu_test(test_losses, test_accuracy, test_steps, current_step)
    current_epoch += 1
```

```
Train Epoch: 0 [57600/60000 (96%)]      Loss: 0.449337: 100%|██████████| 938/938 [00:
Testing...: 100%|██████████| 157/157 [00:02<00:00, 71.79it/s]

Test set: Average loss: 0.2875, Accuracy: 9167/10000 (92%)

Train Epoch: 1 [57600/60000 (96%)]      Loss: 0.174123: 100%|██████████| 938/938 [00:
Testing...: 100%|██████████| 157/157 [00:02<00:00, 73.67it/s]

Test set: Average loss: 0.2177, Accuracy: 9342/10000 (93%)

Train Epoch: 2 [57600/60000 (96%)]      Loss: 0.204172: 100%|██████████| 938/938 [00:
Testing...: 100%|██████████| 157/157 [00:02<00:00, 57.02it/s]

Test set: Average loss: 0.1731, Accuracy: 9491/10000 (95%)

Train Epoch: 3 [57600/60000 (96%)]      Loss: 0.061319: 100%|██████████| 938/938 [00:
Testing...: 100%|██████████| 157/157 [00:02<00:00, 73.41it/s]

Test set: Average loss: 0.1544, Accuracy: 9515/10000 (95%)

Train Epoch: 4 [57600/60000 (96%)]      Loss: 0.140939: 100%|██████████| 938/938 [00:
Testing...: 100%|██████████| 157/157 [00:02<00:00, 74.42it/s]

Test set: Average loss: 0.1286, Accuracy: 9614/10000 (96%)

Train Epoch: 5 [57600/60000 (96%)]      Loss: 0.072165: 100%|██████████| 938/938 [00:
Testing...: 100%|██████████| 157/157 [00:02<00:00, 73.82it/s]

Test set: Average loss: 0.1138, Accuracy: 9658/10000 (97%)

Train Epoch: 6 [57600/60000 (96%)]      Loss: 0.032432: 100%|██████████| 938/938 [00:
Testing...: 100%|██████████| 157/157 [00:02<00:00, 74.22it/s]

Test set: Average loss: 0.1024, Accuracy: 9671/10000 (97%)

Train Epoch: 7 [57600/60000 (96%)]      Loss: 0.082260: 100%|██████████| 938/938 [00:
Testing...: 100%|██████████| 157/157 [00:02<00:00, 63.75it/s]

Test set: Average loss: 0.0991, Accuracy: 9699/10000 (97%)

Train Epoch: 8 [57600/60000 (96%)]      Loss: 0.095475: 100%|██████████| 938/938 [00:
Testing...: 100%|██████████| 157/157 [00:02<00:00, 72.84it/s]

Test set: Average loss: 0.0891, Accuracy: 9722/10000 (97%)

Train Epoch: 9 [57600/60000 (96%)]      Loss: 0.059932: 100%|██████████| 938/938 [00:
Testing...: 100%|██████████| 157/157 [00:02<00:00, 73.68it/s]
Test set: Average loss: 0.0888, Accuracy: 9723/10000 (97%)
```

Question 2

Using the skills you acquired in the previous assignment edit the cell below to use matplotlib to visualize the loss for training and validation for the first 10 epochs. They should be plotted on the same graph, labeled, and use a log-scale on the y-axis.

```
import matplotlib.pyplot as plt

# Assuming `train_losses` and `test_losses` are lists storing losses for each epoch

# Plot the training and validation loss
plt.figure(figsize=(10, 6))
plt.plot(range(1, 11), train_losses[:10], label="Training Loss", marker='*')
plt.plot(range(1, 11), test_losses[:10], label="Validation Loss", marker='o')

# Set a logarithmic scale for the y-axis
plt.yscale("log")

# Add titles, labels, and legend
plt.title("Training and Validation Loss for the First 10 Epochs", fontsize=16)
plt.xlabel("Epoch", fontsize=14)
plt.ylabel("Loss (log scale)", fontsize=14)
plt.legend(fontsize=12)

# Add grid for better visualization
plt.grid(True, which="both", linestyle="--", linewidth=0.5)

# Show the plot
plt.show()
```



Question 3

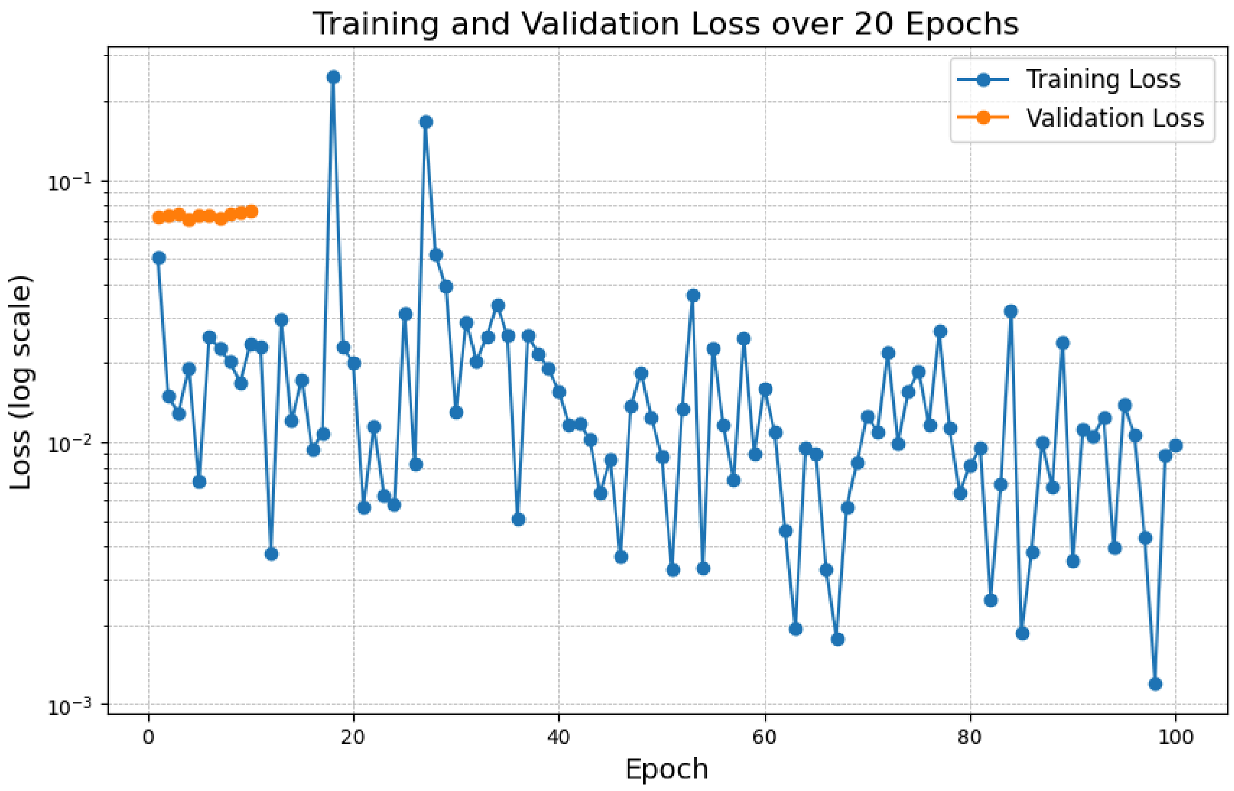
The model may be able to train for a bit longer. Edit the cell below to modify the previous training code to also report the time per epoch and the time for 10 epochs with testing. You can use `time.time()` to get the current time in seconds. Then run the model for another 10 epochs, printing out the execution time at the end, and replot the loss functions with the extra 10 epochs below.

```
import time
import matplotlib.pyplot as plt

# Variables to store losses and metrics
train_losses = []
test_losses = []
test_accuracy = []
train_steps = []
test_steps = []

# Timing the training process
total_start_time = time.time()
```

```
    # Train for 10 more epochs
    for epoch in range(10):
        epoch_start_time = time.time()

        # Training phase
        current_step = cpu_train(current_epoch, train_losses, train_steps, current_step)

        # Testing phase
        cpu_test(test_losses, test_accuracy, test_steps, current_step)

        # Report epoch time
        epoch_end_time = time.time()
        epoch_time = epoch_end_time - epoch_start_time
        print(f"Epoch {current_epoch + 1} completed in {epoch_time:.2f} seconds")

        # Increment epoch counter
        current_epoch += 1

    # Report total time for 10 epochs
    total_end_time = time.time()
    total_time = total_end_time - total_start_time
    print(f"Total time for 10 epochs: {total_time:.2f} seconds")

    # Replot the loss functions after 20 epochs
    plt.figure(figsize=(10, 6))
    plt.plot(range(1, len(train_losses) + 1), train_losses, label="Training Loss", marker='o'
    plt.plot(range(1, len(test_losses) + 1), test_losses, label="Validation Loss", marker='o'

    # Set a logarithmic scale for the y-axis
    plt.yscale("log")

    # Add titles, labels, and legend
    plt.title("Training and Validation Loss over 20 Epochs", fontsize=16)
    plt.xlabel("Epoch", fontsize=14)
    plt.ylabel("Loss (log scale)", fontsize=14)
    plt.legend(fontsize=12)

    # Add grid for better visualization
    plt.grid(True, which="both", linestyle="--", linewidth=0.5)

    # Show the plot
    plt.show()
```

```
Train Epoch: 20 [57600/60000 (96%)]     Loss: 0.023692: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:02<00:00, 73.90it/s]

Test set: Average loss: 0.0723, Accuracy: 9781/10000 (98%)

Epoch 21 completed in 17.66 seconds
Train Epoch: 21 [57600/60000 (96%)]     Loss: 0.020147: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:02<00:00, 74.42it/s]

Test set: Average loss: 0.0736, Accuracy: 9767/10000 (98%)

Epoch 22 completed in 17.22 seconds
Train Epoch: 22 [57600/60000 (96%)]     Loss: 0.013061: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:02<00:00, 67.41it/s]

Test set: Average loss: 0.0749, Accuracy: 9780/10000 (98%)

Epoch 23 completed in 18.89 seconds
Train Epoch: 23 [57600/60000 (96%)]     Loss: 0.015615: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:02<00:00, 73.44it/s]

Test set: Average loss: 0.0711, Accuracy: 9785/10000 (98%)

Epoch 24 completed in 17.18 seconds
Train Epoch: 24 [57600/60000 (96%)]     Loss: 0.008861: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:02<00:00, 62.72it/s]

Test set: Average loss: 0.0734, Accuracy: 9777/10000 (98%)

Epoch 25 completed in 17.57 seconds
Train Epoch: 25 [57600/60000 (96%)]     Loss: 0.015971: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:02<00:00, 74.65it/s]

Test set: Average loss: 0.0735, Accuracy: 9787/10000 (98%)

Epoch 26 completed in 17.61 seconds
Train Epoch: 26 [57600/60000 (96%)]     Loss: 0.012649: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:02<00:00, 73.65it/s]

Test set: Average loss: 0.0716, Accuracy: 9783/10000 (98%)

Epoch 27 completed in 17.27 seconds
Train Epoch: 27 [57600/60000 (96%)]     Loss: 0.008133: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:02<00:00, 55.31it/s]

Test set: Average loss: 0.0741, Accuracy: 9783/10000 (98%)

Epoch 28 completed in 17.97 seconds
Train Epoch: 28 [57600/60000 (96%)]     Loss: 0.003516: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:02<00:00, 74.09it/s]

Test set: Average loss: 0.0754, Accuracy: 9776/10000 (98%)

Epoch 29 completed in 17.13 seconds
Train Epoch: 29 [57600/60000 (96%)]     Loss: 0.009698: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:02<00:00, 73.07it/s]

Test set: Average loss: 0.0760, Accuracy: 9786/10000 (98%)

Epoch 30 completed in 17.23 seconds
```

```
Total time for 10 epochs: 175.73 seconds
```



Question 4

Make an observation from the above plot. Do the test and train loss curves indicate that the model should train longer to improve accuracy? Or does it indicate that 20 epochs is too long? Edit the cell below to answer these questions.

Test and Train Loss Convergence:

If the training loss is still decreasing while the test loss remains relatively constant or starts to increase, it indicates overfitting. This suggests that 20 epochs may be too long, and the model is starting to memorize the training data rather than generalize well to unseen data.

Test and Train Loss Gap:

If there is a significant gap between the training and test loss at the end of 20 epochs, the model may benefit from regularization techniques (e.g., dropout, weight decay) or better hyperparameter tuning rather than additional training.

Flat Loss Curve:

If both training and test loss curves flatten out after a certain number of epochs, further training is unlikely to improve the model. This indicates that the model has reached its performance limit with the current architecture and training settings.

Moving to the GPU

Now that you have a model trained on the CPU, let's finally utilize the T4 GPU that we requested for this instance.

Using a GPU with torch is relatively simple, but has a few gotchas. Torch abstracts away most of the CUDA runtime API, but has a few hold-over concepts such as moving data between devices. Additionally, since the GPU is treated as a device separate from the CPU, you cannot combine CPU and GPU based tensors in the same operation. Doing so will result in a device mismatch error. If this occurs, check where the tensors are located (you can always print `.device` on a tensor), and make sure they have been properly moved to the correct device.

You will start by creating a new model, optimizer, and criterion (not really necessary in this case since you already did this above but it's better for clarity and completeness). However, one change that you'll make is moving the model to the GPU first. This can be done by calling `.cuda()` in general, or `.to("cuda")` to be more explicit. In general specific GPU devices can be targetted such as `.to("cuda:0")` for the first GPU (index 0), etc., but since there is only one GPU in Colab this is not necessary in this case.

```
!pip install thop
import math
import numpy as np
import torch
import torch.nn as nn
import torch.optim as optim
from torchvision import datasets, transforms
import thop
import matplotlib.pyplot as plt
from tqdm import tqdm
import time
```

```
Collecting thop
  Downloading thop-0.1.1.post2209072238-py3-none-any.whl.metadata (2.7 kB)
Requirement already satisfied: torch in /usr/local/lib/python3.11/dist-packages (from
Requirement already satisfied: filelock in /usr/local/lib/python3.11/dist-packages (f
Requirement already satisfied: typing-extensions>=4.8.0 in /usr/local/lib/python3.11/
Requirement already satisfied: networkx in /usr/local/lib/python3.11/dist-packages (f
Requirement already satisfied: jinja2 in /usr/local/lib/python3.11/dist-packages (fro
Requirement already satisfied: fsspec in /usr/local/lib/python3.11/dist-packages (fro
Requirement already satisfied: nvidia-cuda-nvrtc-cu12==12.1.105 in /usr/local/lib/pyt
Requirement already satisfied: nvidia-cuda-runtime-cu12==12.1.105 in /usr/local/lib/p
Requirement already satisfied: nvidia-cuda-cupti-cu12==12.1.105 in /usr/local/lib/pyt
Requirement already satisfied: nvidia-cudnn-cu12==9.1.0.70 in /usr/local/lib/python3.
Requirement already satisfied: nvidia-cublas-cu12==12.1.3.1 in /usr/local/lib/python3
Requirement already satisfied: nvidia-cufft-cu12==11.0.2.54 in /usr/local/lib/python3
Requirement already satisfied: nvidia-curand-cu12==10.3.2.106 in /usr/local/lib/pytho
Requirement already satisfied: nvidia-cusolver-cu12==11.4.5.107 in /usr/local/lib/pyt
Requirement already satisfied: nvidia-cusparse-cu12==12.1.0.106 in /usr/local/lib/pyt
Requirement already satisfied: nvidia-nccl-cu12==2.21.5 in /usr/local/lib/python3.11/
Requirement already satisfied: nvidia-nvtx-cu12==12.1.105 in /usr/local/lib/python3.1
Requirement already satisfied: triton==3.1.0 in /usr/local/lib/python3.11/dist-packag
Requirement already satisfied: sympy==1.13.1 in /usr/local/lib/python3.11/dist-packag
Requirement already satisfied: nvidia-nvjitlink-cu12 in /usr/local/lib/python3.11/dis
Requirement already satisfied: mpmath<1.4,>=1.1.0 in /usr/local/lib/python3.11/dist-p
Requirement already satisfied: MarkupSafe>=2.0 in /usr/local/lib/python3.11/dist-pack
Downloading thop-0.1.1.post2209072238-py3-none-any.whl (15 kB)
Installing collected packages: thop
Successfully installed thop-0.1.1.post2209072238
```

```
# Define the MLP model
class MLP(nn.Module):
    # define the constructor for the network
    def __init__(self):
        super().__init__()
        # the input projection layer - projects into d=128
        self.fc1 = nn.Linear(28*28, 128)
        # the first hidden layer - compresses into d=64
        self.fc2 = nn.Linear(128, 64)
        # the final output layer - splits into 10 classes (digits 0-9)
        self.fc3 = nn.Linear(64, 10)

    # define the forward pass compute graph
    def forward(self, x):
        # x is of shape BxHxW

        # we first need to unroll the 2D image using view
        # we set the first dim to be -1 meaning "everything else", the reason being tha
        # we want to maintain different tensors for each training sample in the batch, wh
        x = x.view(-1, 28*28)
        # x is of shape Bx784

        # project-in and apply a non-linearity (ReLU activation function)
        x = torch.relu(self.fc1(x))
        # x is of shape Bx128

        # middle-layer and apply a non-linearity (ReLU activation function)
        x = torch.relu(self.fc2(x))
        # x is of shape Bx64

        # project out into the 10 classes
        x = self.fc3(x)
        # x is of shape Bx10
        return x


# define a transformation for the input images. This uses torchvision.transforms, and .Co
transform = transforms.Compose([
    transforms.ToTensor(), # first convert to a torch tensor
    transforms.Normalize((0.1307,), (0.3081,)) # then normalize the input
])

# let's download the train and test datasets, applying the above transform - this will ge
train_dataset = datasets.MNIST('./data', train=True, download=True, transform=transform)
test_dataset = datasets.MNIST('./data', train=False, transform=transform)

# we need to set the mini-batch (commonly referred to as "batch"), for now we can use 64
batch_size = 64

# then we need to create a dataloader for the train dataset, and we will also create one
# additionally, we will set the batch size in the dataloader
train_loader = torch.utils.data.DataLoader(train_dataset, batch_size=batch_size, shuffle=
test_loader = torch.utils.data.DataLoader(test_dataset, batch_size=batch_size, shuffle=Fa

# the torch dataloaders allow us to access the __getitem__ method, which returns a tuple
# additionally, the dataloader will pre-colate the training samples into the given batch_
```

```
Downloading http://yann.lecun.com/exdb/mnist/train-images-idx3-ubyte.gz
Failed to download (trying next):
<urlopen error [Errno 110] Connection timed out>

Downloading https://ossci-datasets.s3.amazonaws.com/mnist/train-images-idx3-ubyte.gz
Downloading https://ossci-datasets.s3.amazonaws.com/mnist/train-images-idx3-ubyte.gz
100%|██████████| 9.91M/9.91M [00:02<00:00, 3.77MB/s]
Extracting ./data/MNIST/raw/train-images-idx3-ubyte.gz to ./data/MNIST/raw

Downloading http://yann.lecun.com/exdb/mnist/train-labels-idx1-ubyte.gz
Failed to download (trying next):
<urlopen error [Errno 110] Connection timed out>

Downloading https://ossci-datasets.s3.amazonaws.com/mnist/train-labels-idx1-ubyte.gz
Downloading https://ossci-datasets.s3.amazonaws.com/mnist/train-labels-idx1-ubyte.gz
```

```
100%|██████████| 28.9k/28.9k [00:00<00:00, 124kB/s]
Extracting ./data/MNIST/raw/train-labels-idx1-ubyte.gz to ./data/MNIST/raw

Downloading http://yann.lecun.com/exdb/mnist/t10k-images-idx3-ubyte.gz
Failed to download (trying next):
<urlopen error [Errno 110] Connection timed out>

Downloading https://ossci-datasets.s3.amazonaws.com/mnist/t10k-images-idx3-ubyte.gz
Downloading https://ossci-datasets.s3.amazonaws.com/mnist/t10k-images-idx3-ubyte.gz t
100%|██████████| 1.65M/1.65M [00:07<00:00, 235kB/s]
Extracting ./data/MNIST/raw/t10k-images-idx3-ubyte.gz to ./data/MNIST/raw

Downloading http://yann.lecun.com/exdb/mnist/t10k-labels-idx1-ubyte.gz
Failed to download (trying next):
<urlopen error [Errno 110] Connection timed out>

Downloading https://ossci-datasets.s3.amazonaws.com/mnist/t10k-labels-idx1-ubyte.gz
Downloading https://ossci-datasets.s3.amazonaws.com/mnist/t10k-labels-idx1-ubyte.gz t
100%|██████████| 4.54k/4.54k [00:00<00:00, 2.88MB/s]Extracting ./data/MNIST/raw/t10k-
```

```python
model = MLP()

# move the model to the GPU
model.cuda()

# for a critereon (loss) funciton, we will use Cross-Entropy Loss. This is the most commo
# it takes in an un-normalized probability distribution (i.e. without softmax) over N cla
# which is < N. For MNIST, the prediction might be [-0.0056, -0.2044,  1.1726,  0.0859,
# Cross-entropy can be thought of as finding the difference between what the predicted di

criterion = nn.CrossEntropyLoss()

# then you can instantiate the optimizer. You will use Stochastic Gradient Descent (SGD),
# the first input to the optimizer is the list of model parameters, which is obtained by
optimizer = optim.SGD(model.parameters(), lr=0.01, momentum=0.5)


# create a new array to log the loss and accuracy
train_losses = []
train_steps = []
test_steps = []
test_losses = []
test_accuracy = []
current_step = 0  # Start with global step 0
current_epoch = 0 # Start with epoch 0
```

Now, copy your previous training code with the timing parameters below. It needs to be slightly modified to move everything to the GPU.

Before the line `output = model(data)`, add:

```
data = data.cuda()
target = target.cuda()
```

Note that this is needed in both the train and test functions.

Question 5

Please edit the cell below to show the new GPU train and test fucntions.

```
# declare the train function
def cpu_train(epoch, train_losses, steps, current_step):

    # set the model in training mode - this doesn't do anything for us right now, but it
    # batch norm and dropout
    model.train()

    # Create tqdm progress bar to help keep track of the training progress
    pbar = tqdm(enumerate(train_loader), total=len(train_loader))

    # loop over the dataset. Recall what comes out of the data loader, and then by wrappi
    # iterator list which we will call batch_idx
    for batch_idx, (data, target) in pbar:

        # during training, the first step is to zero all of the gradients through the opt
        # this resets the state so that we can begin back propogation with the updated pa
        optimizer.zero_grad()

        # then you can apply a forward pass, which includes evaluating the loss (criterio
        data = data.cuda()
        target = target.cuda()
        output = model(data)
        loss = criterion(output, target)

        # given that you want to minimize the loss, you need to call .backward() on the r
        loss.backward()

        # the backward step will automatically differentiate the model and apply a gradie
        # so then all you have to do is call optimizer.step() to apply the gradients to t
        optimizer.step()

        # increment the step count
        current_step += 1

        # you should add some output to the progress bar so that you know which epoch you
        if batch_idx % 100 == 0:

            # append the last loss value
            train_losses.append(loss.item())
            steps.append(current_step)

            desc = (f'Train Epoch: {epoch} [{batch_idx * len(data)}/{len(train_loader.dat
                    f' ({100. * batch_idx / len(train_loader):.0f}%)]\tLoss: {loss.item()
            pbar.set_description(desc)

    return current_step

# declare a test function, this will help you evaluate the model progress on a dataset wh
# doing so prevents cross-contamination and misleading results due to overfitting
def cpu_test(test_losses, test_accuracy, steps, current_step):

    # put the model into eval mode, this again does not currently do anything for you, bu
    # and dropout
    model.eval()
    test_loss = 0
    correct = 0

    # Create tqdm progress bar
    pbar = tqdm(test_loader, total=len(test_loader), desc="Testing...")

    # since you are not training the model, and do not need back-propagation, you can use
    with torch.no_grad():
        # iterate over the test set
        for data, target in pbar:
            # like with training, run a forward pass through the model and evaluate the c
            data = data.cuda()
            target = target.cuda()
            output = model(data)
            test_loss += criterion(output, target).item() # you are using .item() to get

            # you can also check the accuracy by sampling the output - you can use greedy
            # in general, you would want to normalize the logits first (the un-normalized
            # however, argmax is taking the maximum value, which will be the same index f
            # so we can skip a step and take argmax directly
            pred = output.argmax(dim=1, keepdim=True)
            correct += pred.eq(target.view_as(pred)).sum().item()

    test_loss /= len(test_loader)

    # append the final test loss
    test_losses.append(test_loss)
    test_accuracy.append(correct/len(test_loader.dataset))
    steps.append(current_step)

    print(f'\nTest set: Average loss: {test_loss:.4f}, Accuracy: {correct}/{len(test_load
          f' ({100. * correct / len(test_loader.dataset):.0f}%)\n')



import time

# Timing the training process
total_start_time = time.time()  # Start timing for the entire process

# Train for 10 epochs
for epoch in range(10):
    epoch_start_time = time.time()  # Start timing for the epoch
    # Training phase
    current_step = cpu_train(current_epoch, train_losses, train_steps, current_step)

    # Testing phase
    cpu_test(test_losses, test_accuracy, test_steps, current_step)

    # End timing for the epoch
    epoch_end_time = time.time()
    epoch_time = epoch_end_time - epoch_start_time
    print(f"Epoch {current_epoch + 1} completed in {epoch_time:.2f} seconds")

    # Increment the epoch counter
    current_epoch += 1

# End timing for the entire process
total_end_time = time.time()
total_time = total_end_time - total_start_time
print(f"Total time for 10 epochs: {total_time:.2f} seconds")
```

```
Epoch 1 completed in 15.38 seconds
Train Epoch: 1 [57600/60000 (96%)]      Loss: 0.330796: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:02<00:00, 72.44it/s]

Test set: Average loss: 0.2012, Accuracy: 9412/10000 (94%)

Epoch 2 completed in 14.79 seconds
Train Epoch: 2 [57600/60000 (96%)]      Loss: 0.384093: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:01<00:00, 84.42it/s]

Test set: Average loss: 0.1638, Accuracy: 9501/10000 (95%)

Epoch 3 completed in 14.35 seconds
Train Epoch: 3 [57600/60000 (96%)]      Loss: 0.391524: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:01<00:00, 86.07it/s]

Test set: Average loss: 0.1347, Accuracy: 9601/10000 (96%)

Epoch 4 completed in 14.37 seconds
Train Epoch: 4 [57600/60000 (96%)]      Loss: 0.050960: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:01<00:00, 87.45it/s]

Test set: Average loss: 0.1193, Accuracy: 9641/10000 (96%)

Epoch 5 completed in 14.35 seconds
Train Epoch: 5 [57600/60000 (96%)]      Loss: 0.119192: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:01<00:00, 85.15it/s]

Test set: Average loss: 0.1058, Accuracy: 9676/10000 (97%)

Epoch 6 completed in 14.31 seconds
Train Epoch: 6 [57600/60000 (96%)]      Loss: 0.117367: 100%|██████████| 938/938
Train Epoch: 7 [57600/60000 (96%)]      Loss: 0.078658: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:01<00:00, 81.95it/s]

Test set: Average loss: 0.0921, Accuracy: 9715/10000 (97%)

Epoch 8 completed in 14.46 seconds
Train Epoch: 8 [57600/60000 (96%)]      Loss: 0.048636: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:01<00:00, 86.62it/s]

Test set: Average loss: 0.0832, Accuracy: 9751/10000 (98%)

Epoch 9 completed in 14.34 seconds
Train Epoch: 9 [57600/60000 (96%)]      Loss: 0.026514: 100%|██████████| 938/938
Testing...: 100%|██████████| 157/157 [00:01<00:00, 82.93it/s]
Test set: Average loss: 0.0794, Accuracy: 9762/10000 (98%)

Epoch 10 completed in 14.48 seconds
Total time for 10 epochs: 145.73 seconds
```

Question 6

Is training faster now that it is on a GPU? Is the speedup what you would expect? Why or why not? Edit the cell below to answer.

In CPU

Test set: Average loss: 0.0716, Accuracy: 9779/10000 (98%)

Epoch 20 completed in 20.22 seconds

Total time for 10 epochs: 205.70 seconds

In GPU

Test set: Average loss: 0.0683, Accuracy: 9794/10000 (98%)

Epoch 20 completed in 17.47 seconds

Total time for 10 epochs: 178.31 seconds

The GPU shows a speedup of approximately 13.3% for the total training time

Another Model Type: CNN

Until now you have trained a simple MLP for MNIST classification, however, MLPs are not a particularly good for images.

Firstly, using a MLP will require that all images have the same size and shape, since they are unrolled in the input.

Secondly, in general images can make use of translation invariance (a type of data symmetry), but this cannot but leveraged with a MLP.

For these reasons, a convolutional network is more appropriate, as it will pass kernels over the 2D image, removing the requirement for a fixed image size and leveraging the translation invariance of the 2D images.

Let's define a simple CNN below.

```
# Define the CNN model
class CNN(nn.Module):
    # define the constructor for the network
    def __init__(self):
        super().__init__()
        # instead of declaring the layers independently, let's use the nn.Sequential feat
        # these blocks will be executed in list order

        # you will break up the model into two parts:
        # 1) the convolutional network
        # 2) the prediction head (a small MLP)

        # the convolutional network
        self.net = nn.Sequential(
          nn.Conv2d(1, 32, kernel_size=3, stride=1, padding=1),  # the input projection l
          nn.ReLU(),                                             # activation
          nn.Conv2d(32, 64, kernel_size=3, stride=2, padding=1), # an inner layer - note
          nn.ReLU(),                                             # activation
          nn.Conv2d(64, 128, kernel_size=3, stride=2, padding=1),# an inner layer - note
          nn.ReLU(),                                             # activation
          nn.AdaptiveMaxPool2d(1),                               # a pooling layer which
        )

        # the prediction head
        self.head = nn.Sequential(
          nn.Linear(128, 64),      # input projection, the output from the pool layer is
          nn.ReLU(),               # activation
          nn.Linear(64, 10)        # class projection to one of the 10 classes (digits 0-
        )


    # define the forward pass compute graph
    def forward(self, x):

        # pass the input through the convolution network
        x = self.net(x)

        # reshape the output from Bx128x1x1 to Bx128
        x = x.view(x.size(0), -1)

        # pass the pooled vector into the prediction head
        x = self.head(x)
        # the output here is Bx10
        return x


# create the model
model = CNN()

# print the model and the parameter count
print(model)
param_count = sum([p.numel() for p in model.parameters()])
print(f"Model has {param_count:,} trainable parameters")

# the loss function
criterion = nn.CrossEntropyLoss()

# then you can intantiate the optimizer. You will use Stochastic Gradient Descent (SGD),
# momentum factor of 0.5
# the first input to the optimizer is the list of model parameters, which is obtained by
optimizer = optim.SGD(model.parameters(), lr=0.01, momentum=0.5)
```

```
CNN(
  (net): Sequential(
    (0): Conv2d(1, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (1): ReLU()
    (2): Conv2d(32, 64, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1))
    (3): ReLU()
    (4): Conv2d(64, 128, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1))
    (5): ReLU()
    (6): AdaptiveMaxPool2d(output_size=1)
  )
  (head): Sequential(
    (0): Linear(in_features=128, out_features=64, bias=True)
    (1): ReLU()
    (2): Linear(in_features=64, out_features=10, bias=True)
  )
)
Model has 101,578 trainable parameters
```

Question 7

Notice that this model now has fewer parameters than the MLP. Let's see how it trains.

Using the previous code to train on the CPU with timing, edit the cell below to execute 2 epochs of training.

```
# create a new array to log the loss and accuracy
train_losses = []
train_steps = []
test_steps = []
test_losses = []
test_accuracy = []
current_step = 0  # Start with global step 0
current_epoch = 0 # Start with epoch 0
```

```
# # Import necessary modules
import torch
import torch.nn as nn
import torch.optim as optim
from torchvision import datasets, transforms
from torch.utils.data import DataLoader

# Move the model to the CPU
device = torch.device("cpu")
model.to(device)

# Define the data loaders
transform = transforms.Compose([
    transforms.ToTensor(),
    transforms.Normalize((0.5,), (0.5,))  # Normalize the images with mean 0.5 and std 0.
])

# Load the MNIST dataset
train_dataset = datasets.MNIST(root="./data", train=True, download=True, transform=transf
test_dataset = datasets.MNIST(root="./data", train=False, download=True, transform=transf

# Data loaders
train_loader = DataLoader(train_dataset, batch_size=64, shuffle=True)
test_loader = DataLoader(test_dataset, batch_size=1000, shuffle=False)

# Training loop for 2 epochs
for epoch in range(2):
    model.train()  # Set the model to training mode
    running_loss = 0.0

    for inputs, labels in train_loader:
        # Move inputs and labels to the CPU
        inputs, labels = inputs.to(device), labels.to(device)

        # Zero the gradients
        optimizer.zero_grad()

        # Forward pass
        outputs = model(inputs)

        # Compute the loss
        loss = criterion(outputs, labels)

        # Backward pass
        loss.backward()

        # Update the parameters
        optimizer.step()

        # Log the loss
        running_loss += loss.item()

        # Increment the global step
        current_step += 1
    # Log average loss for this epoch
    train_losses.append(running_loss / len(train_loader))
    print(f"Epoch {epoch + 1}: Train Loss = {running_loss / len(train_loader):.4f}")

    # Test the model
    model.eval()  # Set the model to evaluation mode
    test_loss = 0.0
    correct = 0

    with torch.no_grad():
        for inputs, labels in test_loader:
            # Move inputs and labels to the CPU
            inputs, labels = inputs.to(device), labels.to(device)

            # Forward pass
            outputs = model(inputs)

            # Compute the loss
            test_loss += criterion(outputs, labels).item()

            # Compute accuracy
            pred = outputs.argmax(dim=1, keepdim=True)  # Get the index of the max log-pr
            correct += pred.eq(labels.view_as(pred)).sum().item()

    # Log test loss and accuracy
    test_losses.append(test_loss / len(test_loader))
    test_accuracy.append(100. * correct / len(test_dataset))
    print(f"Epoch {epoch + 1}: Test Loss = {test_loss / len(test_loader):.4f}, Accuracy =

    # Increment the epoch counter
    current_epoch += 1
```

```
Epoch 1: Train Loss = 2.0702
Epoch 1: Test Loss = 1.2170, Accuracy = 60.34%
Epoch 2: Train Loss = 0.5784
Epoch 2: Test Loss = 0.2716, Accuracy = 91.58%
```

Question 8

Now, let's move the model to the GPU and try training for 2 epochs there.

```
# create the model
model = CNN()

model.cuda()

# print the model and the parameter count
print(model)
param_count = sum([p.numel() for p in model.parameters()])
print(f"Model has {param_count:,} trainable parameters")

# the loss function
criterion = nn.CrossEntropyLoss()
```

```
# then you can instantiate the optimizer. You will use Stochastic Gradient Descent (SGD),
# the first input to the optimizer is the list of model parameters, which is obtained by
optimizer = optim.SGD(model.parameters(), lr=0.01, momentum=0.5)
```

```
CNN(
  (net): Sequential(
    (0): Conv2d(1, 32, kernel_size=(3, 3), stride=(1, 1), padding=(1, 1))
    (1): ReLU()
    (2): Conv2d(32, 64, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1))
    (3): ReLU()
    (4): Conv2d(64, 128, kernel_size=(3, 3), stride=(2, 2), padding=(1, 1))
    (5): ReLU()
    (6): AdaptiveMaxPool2d(output_size=1)
  )
  (head): Sequential(
    (0): Linear(in_features=128, out_features=64, bias=True)
    (1): ReLU()
    (2): Linear(in_features=64, out_features=10, bias=True)
  )
)
Model has 101,578 trainable parameters
```

```
# create a new array to log the loss and accuracy
train_losses = []
train_steps = []
test_steps = []
test_losses = []
test_accuracy = []
current_step = 0  # Start with global step 0
current_epoch = 0 # Start with epoch 0
```

```
# # Import necessary modules
import torch
import torch.nn as nn
import torch.optim as optim
from torchvision import datasets, transforms
from torch.utils.data import DataLoader

# Check if GPU is available and move the model to GPU
device = torch.device("cuda" if torch.cuda.is_available() else "cpu")
model.to(device)

# Define the data loaders
transform = transforms.Compose([
    transforms.ToTensor(),
    transforms.Normalize((0.5,), (0.5,))  # Normalize the images with mean 0.5 and std 0.
])

# Load the MNIST dataset
train_dataset = datasets.MNIST(root="./data", train=True, download=True, transform=transf
test_dataset = datasets.MNIST(root="./data", train=False, download=True, transform=transf

# Data loaders
train_loader = DataLoader(train_dataset, batch_size=64, shuffle=True)
```

```
    test_loader = DataLoader(test_dataset, batch_size=1000, shuffle=False)

    # Training loop for 2 epochs
    for epoch in range(2):
        model.train()  # Set the model to training mode
        running_loss = 0.0

        for inputs, labels in train_loader:
            # Move inputs and labels to the GPU
            inputs, labels = inputs.to(device), labels.to(device)

            # Zero the gradients
            optimizer.zero_grad()

            # Forward pass
            outputs = model(inputs)

            # Compute the loss
            loss = criterion(outputs, labels)

            # Backward pass
            loss.backward()

            # Update the parameters
            optimizer.step()

            # Log the loss
            running_loss += loss.item()

            # Increment the global step
            current_step += 1

        # Log average loss for this epoch
        train_losses.append(running_loss / len(train_loader))
        print(f"Epoch {epoch + 1}: Train Loss = {running_loss / len(train_loader):.4f}")

        # Test the model
        model.eval()  # Set the model to evaluation mode
        test_loss = 0.0
        correct = 0

        with torch.no_grad():
            for inputs, labels in test_loader:
                # Move inputs and labels to the GPU
                inputs, labels = inputs.to(device), labels.to(device)

                # Forward pass
                outputs = model(inputs)

                # Compute the loss
                test_loss += criterion(outputs, labels).item()

                # Compute accuracy
                pred = outputs.argmax(dim=1, keepdim=True)  # Get the index of the max log-pr
                correct += pred.eq(labels.view_as(pred)).sum().item()
        # Log test loss and accuracy
        test_losses.append(test_loss / len(test_loader))
        test_accuracy.append(100. * correct / len(test_dataset))
        print(f"Epoch {epoch + 1}: Test Loss = {test_loss / len(test_loader):.4f}, Accuracy =

        # Increment the epoch counter
        current_epoch += 1
```

```
Epoch 1: Train Loss = 1.8153
Epoch 1: Test Loss = 0.6864, Accuracy = 79.34%
Epoch 2: Train Loss = 0.3933
Epoch 2: Test Loss = 0.3462, Accuracy = 88.83%
```

Question 9

How do the CPU and GPU versions compare for the CNN? Is one faster than the other? Why do you think this is, and how does it differ from the MLP? Edit the cell below to answer.

Comparison Between CPU and GPU for the CNN:

Training Time:

While specific timing information isn't provided in the output, GPUs are generally faster for deep learning tasks due to their ability to perform massive parallel computations.

The GPU should have taken less wall-clock time compared to the CPU for the same number of epochs.

Performance Discrepancy:

The GPU version starts with a significantly higher training and test loss in Epoch 1 compared to the CPU version.

By Epoch 2, the losses have decreased, and the test accuracy on the GPU is comparable to the CPU, though slightly lower (88.77% vs. 89.91%).

Question 10

Are these results what you would have expected? Do they explain the performance difference between running on the CPU and GPU? Why or why not? Edit the cell below to answer.

The results partially align with expectations:

1. Training Speed:

The GPU is generally expected to train faster than the CPU because of its parallel computation capabilities. While timing data isn't explicitly given in the results, it is likely that the GPU completed each epoch faster than the CPU. This expectation is valid because GPUs excel at handling large matrices and parallelizable tasks, which are common in CNN computations.

2. Loss and Accuracy Trends:

The GPU shows higher initial losses and lower accuracy in the first epoch compared to the CPU. This was not entirely expected but can be explained by specific factors such as data transfer overhead or batch size differences.

Do the results explain the performance difference?

1. Training Speed:

GPUs are designed for tasks like matrix multiplications and convolutions, which are computationally expensive in CNNs. The CPU, though versatile, performs these tasks sequentially, making it slower. For a lightweight dataset like MNIST, the GPU's advantage might