Assignment 2: Convolutional Neural Networks with Pytorch

For this assignment, we're going to use one of most popular deep learning frameworks: PyTorch. And build our way through Convolutional Neural Networks.

What is PyTorch?

PyTorch is a system for executing dynamic computational graphs over Tensor objects that behave similarly as numpy ndarray. It comes with a powerful automatic differentiation engine that removes the need for manual back-propagation.

Why?

- Our code will now run on GPUs! Much faster training. When using a framework like PyTorch or TensorFlow you can harness the power of the GPU for your own custom neural network architectures without having to write CUDA code directly (which is beyond the scope of this class).
- We want you to be ready to use one of these frameworks for your project so you can experiment more efficiently than if you were writing every feature you want to use by hand.
- We want you to stand on the shoulders of giants! TensorFlow and PyTorch are both excellent frameworks that will make your lives a lot easier, and now that you understand their guts, you are free to use them :)
- We want you to be exposed to the sort of deep learning code you might run into in academia or industry.

PyTorch versions

This notebook assumes that you are using **PyTorch version >=1.0**. In some of the previous versions (e.g. before 0.4), Tensors had to be wrapped in Variable objects to be used in autograd; however Variables have now been deprecated. In addition 1.0 also separates a Tensor's datatype from its device, and uses numpy-style factories for constructing Tensors rather than directly invoking Tensor constructors.

If you are running on datahub, you shouldn't face any problem.

You can also find the detailed PyTorch API doc here. If you have other questions that are not addressed by the API docs, the PyTorch forum is a much better place to ask than StackOverflow.

Table of Contents

This assignment has 5 parts. You will learn PyTorch on **three different levels of abstraction**, which will help you understand it better and prepare you for the final project.

1. Part I, Preparation: we will use CIFAR-100 dataset.
2. Part II, Barebones PyTorch: **Abstraction level 1**, we will work directly with the lowest-level PyTorch Tensors.
3. Part III, PyTorch Module API: **Abstraction level 2**, we will use `nn.Module` to define arbitrary neural network architecture.
4. Part IV, PyTorch Sequential API: **Abstraction level 3**, we will use `nn.Sequential` to define a linear feed-forward network very conveniently.
5. Part V. ResNet10 Implementation: we will implement ResNet10 from scratch given the architecture details
6. Part VI, CIFAR-100 open-ended challenge: please implement your own network to get as high accuracy as possible on CIFAR-100. You can experiment with any layer, optimizer, hyperparameters or other advanced features.

Here is a table of comparison:

API	Flexibility	Convenience
Barebone	High	Low
`nn.Module`	High	Medium

API	Flexibility	Convenience
`nn.Sequential`	Low	High

Part I. Preparation

First, we load the CIFAR-100 dataset. This might take a couple minutes the first time you do it, but the files should stay cached after that.

In [ ]:
```
# Add official website of pytorch

import torch
import torch.nn as nn
import torch.optim as optim
from torch.utils.data import DataLoader
from torch.utils.data import sampler

import torchvision.datasets as dset
import torchvision.transforms as T

import numpy as np
```

In [ ]:
```
NUM_TRAIN = 49000
batch_size= 64

# The torchvision.transforms package provides tools for preprocessing data
# and for performing data augmentation; here we set up a transform to
# preprocess the data by subtracting the mean RGB value and dividing by the
# standard deviation of each RGB value; we've hardcoded the mean and std.

#================================================================================#
# You should try changing the transform for the training data to include    #
# data augmentation such as RandomCrop and HorizontalFlip                    #
# when running the final part of the notebook where you have to achieve     #
# as high accuracy as possible on CIFAR-100.                                #
# Of course you will have to re-run this block for the effect to take place #
#================================================================================#
train_transform = transform = T.Compose([
                T.ToTensor(),
                T.Normalize((0.5071, 0.4867, 0.4408), (0.2675, 0.2565, 0.2761))
            ])

# We set up a Dataset object for each split (train / val / test); Datasets load
# training examples one at a time, so we wrap each Dataset in a DataLoader which
# iterates through the Dataset and forms minibatches. We divide the CIFAR-100
# training set into train and val sets by passing a Sampler object to the
# DataLoader telling how it should sample from the underlying Dataset.
cifar100_train = dset.CIFAR100('./datasets/cifar100', train=True, download=True,
                             transform=train_transform)
loader_train = DataLoader(cifar100_train, batch_size=batch_size, num_workers=2,
                          sampler=sampler.SubsetRandomSampler(range(NUM_TRAIN))) # 49000 train data

cifar100_val = dset.CIFAR100('./datasets/cifar100', train=True, download=True,
                            transform=transform)
loader_val = DataLoader(cifar100_val, batch_size=batch_size, num_workers=2,
                        sampler=sampler.SubsetRandomSampler(range(NUM_TRAIN, 50000))) # 1000 val data

cifar100_test = dset.CIFAR100('./datasets/cifar100', train=False, download=True,
                             transform=transform)
loader_test = DataLoader(cifar100_test, batch_size=batch_size, num_workers=2)
```

```
Files already downloaded and verified
Files already downloaded and verified
Files already downloaded and verified
```

You have an option to **use GPU by setting the flag to True below** (recommended). It is not necessary to use GPU for this assignment. Note that if your computer does not have CUDA enabled, `torch.cuda.is_available()` will return False and this notebook will fallback to CPU mode. **You can run on GPU on datahub.**

The global variables `dtype` and `device` will control the data types throughout this assignment.

In [ ]:
```
USE_GPU = True
num_class = 100
dtype = torch.float32 # we will be using float throughout this tutorial

if USE_GPU and torch.cuda.is_available():
    device = torch.device('cuda')
else:
    device = torch.device('cpu')

# Constant to control how frequently we print train loss
print_every = 100

print('using device:', device)
```

```
using device: cuda
```

Part II. Barebones PyTorch (10% of Grade)

PyTorch ships with high-level APIs to help us define model architectures conveniently, which we will cover in Part II of this tutorial. In this section, we will start with the barebone PyTorch elements to understand the autograd engine better. After this exercise, you will come to appreciate the high-level model API more.

We will start with a simple fully-connected ReLU network with two hidden layers and no biases for CIFAR-100 classification. This implementation computes the forward pass using operations on PyTorch Tensors, and uses PyTorch autograd to compute gradients. It is important that you understand every line, because you will write a harder version after the example.

When we create a PyTorch Tensor with `requires_grad=True`, then operations involving that Tensor will not just compute values; they will also build up a computational graph in the background, allowing us to easily backpropagate through the graph to compute gradients of some Tensors with respect to a downstream loss. Concretely if x is a Tensor with `x.requires_grad == True` then after backpropagation `x.grad` will be another Tensor holding the gradient of x with respect to the scalar loss at the end.

PyTorch Tensors: Flatten Function

A PyTorch Tensor is conceptionally similar to a numpy array: it is an n-dimensional grid of numbers, and like numpy PyTorch provides many functions to efficiently operate on Tensors. As a simple example, we provide a `flatten` function below which reshapes image data for use in a fully-connected neural network.

Recall that image data is typically stored in a Tensor of shape N x C x H x W, where:

- N is the number of datapoints
- C is the number of channels
- H is the height of the intermediate feature map in pixels
- W is the height of the intermediate feature map in pixels

This is the right way to represent the data when we are doing something like a 2D convolution, that needs spatial understanding of where the intermediate features are relative to each other. When we use fully connected affine layers to process the image, however, we want each datapoint to be represented by a single vector -- it's no longer useful to segregate the different channels, rows, and columns of the data. So, we use a "flatten" operation to collapse the `C x H x W` values per representation into a single long vector. The flatten function below first reads in the N, C, H, and W values from a given batch of data, and then returns a "view" of that data. "View" is analogous to numpy's "reshape" method: it reshapes x's dimensions to be N x ??, where ?? is allowed to be anything (in this case, it will be C x H x W, but we don't need to specify that explicitly).

In [ ]:
```
def flatten(x):
    N = x.shape[0] # read in N, C, H, W
    return x.view(N, -1)  # "flatten" the C * H * W values into a single vector per image

def test_flatten():
    x = torch.arange(12).view(2, 1, 3, 2)
    print('Before flattening: ', x)
    print('After flattening: ', flatten(x))
```

```
test_flatten()
```

```
Before flattening:  tensor([[[[ 0,  1],
          [ 2,  3],
          [ 4,  5]]],


        [[[ 6,  7],
          [ 8,  9],
          [10, 11]]]])
After flattening:  tensor([[ 0,  1,  2,  3,  4,  5],
        [ 6,  7,  8,  9, 10, 11]])
```

Barebones PyTorch: Two-Layer Network

Here we define a function `two_layer_fc` which performs the forward pass of a two-layer fully-connected ReLU network on a batch of image data. After defining the forward pass we check that it doesn't crash and that it produces outputs of the right shape by running zeros through the network.

You don't have to write any code here, but it's important that you read and understand the implementation.

In [ ]:
```
import torch.nn.functional as F  # useful stateless functions

def two_layer_fc(x, params):
    """
    A fully-connected neural networks; the architecture is:
    NN is fully connected -> ReLU -> fully connected layer.
    Note that this function only defines the forward pass;
    PyTorch will take care of the backward pass for us.

    The input to the network will be a minibatch of data, of shape
    (N, d1, ..., dM) where d1 * ... * dM = D. The hidden layer will have H units, # minibatch -> the batch used for
    and the output layer will produce scores for C classes.

    Inputs:
    - x: A PyTorch Tensor of shape (N, d1, ..., dM) giving a minibatch of
      input data.
    - params: A list [w1, w2] of PyTorch Tensors giving weights for the network;
      w1 has shape (D, H) and w2 has shape (H, C).

    Returns:
    - scores: A PyTorch Tensor of shape (N, C) giving classification scores for
      the input data x.
    """
    # first we flatten the image
    x = flatten(x)  # shape: [batch_size, C x H x W]

    w1, w2 = params

    # Forward pass: compute predicted y using operations on Tensors. Since w1 and
    # w2 have requires_grad=True, operations involving these Tensors will cause
    # PyTorch to build a computational graph, allowing automatic computation of
    # gradients. Since we are no longer implementing the backward pass by hand we
    # don't need to keep references to intermediate values.
    # you can also use `.clamp(min=0)`, equivalent to F.relu()
    x = F.relu(x.mm(w1))
    x = x.mm(w2)
    return x


def two_layer_fc_test():
    hidden_layer_size = 42
    x = torch.zeros((64, 50), dtype=dtype)  # minibatch size 64, feature dimension 50
    w1 = torch.zeros((50, hidden_layer_size), dtype=dtype) # need require_grad_() to cast into tensor type with gra
    w2 = torch.zeros((hidden_layer_size, num_class), dtype=dtype)
    scores = two_layer_fc(x, [w1, w2])
    print(scores.size())  # you should see [64, 100]

two_layer_fc_test()
```

```
torch.Size([64, 100])
```

Barebones PyTorch: Three-Layer ConvNet

Here you will complete the implementation of the function `three_layer_convnet`, which will perform the forward pass of a three-layer convolutional network. Like above, we can immediately test our implementation by passing zeros through the network. The network should have the following architecture:

1. A convolutional layer (with bias) with `channel_1` filters, each with shape `KW1 x KH1`, and zero-padding of two
2. ReLU nonlinearity
3. A convolutional layer (with bias) with `channel_2` filters, each with shape `KW2 x KH2`, and zero-padding of one
4. ReLU nonlinearity
5. Fully-connected layer with bias, producing scores for C classes.

Note that we have **no softmax activation** here after our fully-connected layer: this is because PyTorch's cross entropy loss performs a softmax activation for you, and by bundling that step in makes computation more efficient.

HINT: For convolutions: https://pytorch.org/docs/stable/nn.functional.html#torch.nn.functional.conv2d; pay attention to the shapes of convolutional filters!

In [ ]:
```
filters = torch.ones(8, 4, 3, 3)
inputs = torch.ones(1, 4, 5, 5)
b=F.conv2d(inputs, filters, padding=1)
# b
```

In [ ]:
```
def three_layer_convnet(x, params):
    """
    Performs the forward pass of a three-layer convolutional network with the
    architecture defined above.

    Inputs:
    - x: A PyTorch Tensor of shape (N, 3, H, W) giving a minibatch of images
    - params: A list of PyTorch Tensors giving the weights and biases for the
      network; should contain the following:
      - conv_w1: PyTorch Tensor of shape (channel_1, 3, KH1, KW1) giving weights
        for the first convolutional layer
      - conv_b1: PyTorch Tensor of shape (channel_1,) giving biases for the first
        convolutional layer (outputting with shape (N, channel_1, H, W))
      - conv_w2: PyTorch Tensor of shape (channel_2, channel_1, KH2, KW2) giving
        weights for the second convolutional layer
      - conv_b2: PyTorch Tensor of shape (channel_2,) giving biases for the second
        convolutional layer. Then we get (N, channel_2, H, W)
      - fc_w: PyTorch Tensor giving weights for the fully-connected layer. Can you
        figure out what the shape should be? (channel_2* H* W, C)
      - fc_b: PyTorch Tensor giving biases for the fully-connected layer. Can you
        figure out what the shape should be? (C,)

    Returns:
    - scores: PyTorch Tensor of shape (N, C) giving classification scores for x
    """
    conv_w1, conv_b1, conv_w2, conv_b2, fc_w, fc_b = params
    scores = None
    ################################################################################
    # TODO: Implement the forward pass for the three-layer ConvNet.                #
    ################################################################################
    # *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
    x = F.conv2d(x, weight=conv_w1, bias=conv_b1, padding=2)
    x = F.relu(x)
    x = F.conv2d(x, weight=conv_w2, bias=conv_b2, padding=1)
    x = F.relu(x)
    x = flatten(x)
    # print(x.shape, fc_w.shape, fc_b.shape)
    scores = torch.matmul(x, fc_w) + fc_b
    pass

    # *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
    ################################################################################
    #                                 END OF YOUR CODE                             #
    ################################################################################
    return scores
```

After defining the forward pass of the ConvNet above, run the following cell to test your implementation.

When you run this function, scores should have shape (64, 100).

In [ ]:
```
def three_layer_convnet_test():
    x = torch.zeros((64, 3, 32, 32), dtype=dtype)  # minibatch size 64, image size [3, 32, 32]

    conv_w1 = torch.zeros((6, 3, 5, 5), dtype=dtype)  # [out_channel, in_channel, kernel_H, kernel_W]
    conv_b1 = torch.zeros((6,))  # out_channel
    conv_w2 = torch.zeros((9, 6, 3, 3), dtype=dtype)  # [out_channel, in_channel, kernel_H, kernel_W]
    conv_b2 = torch.zeros((9,))  # out_channel

    # you must calculate the shape of the tensor after two conv layers, before the fully-connected layer
    fc_w = torch.zeros((9 * 32 * 32, num_class))
    fc_b = torch.zeros(num_class)

    scores = three_layer_convnet(x, [conv_w1, conv_b1, conv_w2, conv_b2, fc_w, fc_b])
    print(scores.size())  # you should see [64, 100]
three_layer_convnet_test()
```

```
torch.Size([64, 100])
```

Barebones PyTorch: Initialization

Let's write a couple utility methods to initialize the weight matrices for our models.

- `random_weight(shape)` initializes a weight tensor with the Kaiming normalization method.
- `zero_weight(shape)` initializes a weight tensor with all zeros. Useful for instantiating bias parameters.

The `random_weight` function uses the Kaiming normal initialization method, described in:

He et al, *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*, ICCV 2015, https://arxiv.org/abs/1502.01852

In [ ]:
```
def random_weight(shape):
    """
    Create random Tensors for weights; setting requires_grad=True means that we
    want to compute gradients for these Tensors during the backward pass.
    We use Kaiming normalization: sqrt(2 / fan_in)
    """
    if len(shape) == 2:  # FC weight
        fan_in = shape[0]
    else:
        fan_in = np.prod(shape[1:]) # conv weight [out_channel, in_channel, kH, kW]
    # randn is standard normal distribution generator.
    w = torch.randn(shape, device=device, dtype=dtype) * np.sqrt(2. / fan_in)
    w.requires_grad = True
    return w

def zero_weight(shape):
    return torch.zeros(shape, device=device, dtype=dtype, requires_grad=True)

# create a weight of shape [3 x 5]
# you should see the type `torch.cuda.FloatTensor` if you use GPU.
# Otherwise it should be `torch.FloatTensor`
random_weight((3, 5))
```

Out[ ]:
```
tensor([[-0.7134,  0.4247,  0.1595,  0.8472,  1.2556],
        [ 0.1591,  0.9924, -0.8986,  1.8867,  1.1273],
        [ 0.5633,  0.5287,  0.8450, -1.7805, -0.8970]], device='cuda:0',
       requires_grad=True)
```

Barebones PyTorch: Check Accuracy

When training the model we will use the following function to check the accuracy of our model on the training or validation sets.

When checking accuracy we don't need to compute any gradients; as a result we don't need PyTorch to build a computational graph for us when we compute scores. To prevent a graph from being built we scope our computation under a `torch.no_grad()` context manager.

In [ ]:
```
def check_accuracy_part2(loader, model_fn, params):
    """
    Check the accuracy of a classification model.
```

```
    Inputs:
    - loader: A DataLoader for the data split we want to check
    - model_fn: A function that performs the forward pass of the model,
      with the signature scores = model_fn(x, params)
    - params: List of PyTorch Tensors giving parameters of the model

    Returns: The accuracy of the model
    """
    split = 'val' if loader.dataset.train else 'test'
    print('Checking accuracy on the %s set' % split)
    num_correct, num_samples = 0, 0
    with torch.no_grad():
        for x, y in loader:
            x = x.to(device=device, dtype=dtype)  # move to device, e.g. GPU
            y = y.to(device=device, dtype=torch.int64)
            scores = model_fn(x, params)
            _, preds = scores.max(1)
            num_correct += (preds == y).sum()
            num_samples += preds.size(0)
        acc = float(num_correct) / num_samples
        print('Got %d / %d correct (%.2f%%)' % (num_correct, num_samples, 100 * acc))
    return acc
```

BareBones PyTorch: Training Loop

We can now set up a basic training loop to train our network. We will train the model using stochastic gradient descent without momentum. We will use `torch.functional.cross_entropy` to compute the loss; you can read about it here.

The training loop takes as input the neural network function, a list of initialized parameters (`[w1, w2]` in our example), and learning rate.

In [ ]:
```
def train_part2(model_fn, params, learning_rate):
    """
    Train a model on CIFAR-10.

    Inputs:
    - model_fn: A Python function that performs the forward pass of the model.
      It should have the signature scores = model_fn(x, params) where x is a
      PyTorch Tensor of image data, params is a list of PyTorch Tensors giving
      model weights, and scores is a PyTorch Tensor of shape (N, C) giving
      scores for the elements in x.
    - params: List of PyTorch Tensors giving weights for the model
    - learning_rate: Python scalar giving the learning rate to use for SGD

    Returns: The accuracy of the model
    """
    for t, (x, y) in enumerate(loader_train):
        # Move the data to the proper device (GPU or CPU)
        x = x.to(device=device, dtype=dtype)
        y = y.to(device=device, dtype=torch.long)

        # Forward pass: compute scores and loss
        scores = model_fn(x, params)
        loss = F.cross_entropy(scores, y)

        # Backward pass: PyTorch figures out which Tensors in the computational
        # graph has requires_grad=True and uses backpropagation to compute the
        # gradient of the loss with respect to these Tensors, and stores the
        # gradients in the .grad attribute of each Tensor.
        loss.backward()

        # Update parameters. We don't want to backpropagate through the
        # parameter updates, so we scope the updates under a torch.no_grad()
        # context manager to prevent a computational graph from being built.
        with torch.no_grad():
            for w in params:
                w -= learning_rate * w.grad

                # Manually zero the gradients after running the backward pass
                w.grad.zero_()

        if (t + 1) % print_every == 0:
```

```
            print('Iteration %d, loss = %.4f' % (t + 1, loss.item()))
            check_accuracy_part2(loader_val, model_fn, params)
            print()
    return check_accuracy_part2(loader_val, model_fn, params)
```

BareBones PyTorch: Train a Two-Layer Network

Now we are ready to run the training loop. We need to explicitly allocate tensors for the fully connected weights, `w1` and `w2`.

Each minibatch of CIFAR has 64 examples, so the tensor shape is `[64, 3, 32, 32]`.

After flattening, `x` shape should be `[64, 3 * 32 * 32]`. This will be the size of the first dimension of `w1`. The second dimension of `w1` is the hidden layer size, which will also be the first dimension of `w2`.

Finally, the output of the network is a 100-dimensional vector that represents the probability distribution over 100 classes.

You don't need to tune any hyperparameters but you should see accuracies above 15% after training for one epoch.

In [ ]:
```
hidden_layer_size = 4000
learning_rate = 1e-2

w1 = random_weight((3 * 32 * 32, hidden_layer_size)) # fc layer-1 weight (D (C*H*W), H)
w2 = random_weight((hidden_layer_size, num_class)) # fc layer-2 weight (H, C)

train_part2(two_layer_fc, [w1, w2], learning_rate)
```

```
Iteration 100, loss = 4.1599
Checking accuracy on the val set
Got 73 / 1000 correct (7.30%)

Iteration 200, loss = 3.8234
Checking accuracy on the val set
Got 102 / 1000 correct (10.20%)

Iteration 300, loss = 3.9627
Checking accuracy on the val set
Got 113 / 1000 correct (11.30%)

Iteration 400, loss = 4.0484
Checking accuracy on the val set
Got 129 / 1000 correct (12.90%)

Iteration 500, loss = 3.2232
Checking accuracy on the val set
Got 146 / 1000 correct (14.60%)

Iteration 600, loss = 3.7111
Checking accuracy on the val set
Got 151 / 1000 correct (15.10%)

Iteration 700, loss = 3.5815
Checking accuracy on the val set
Got 152 / 1000 correct (15.20%)

Checking accuracy on the val set
Got 169 / 1000 correct (16.90%)
```

Out[ ]:
```
0.169
```

BareBones PyTorch: Training a ConvNet

In the below cell you should use the functions defined above to train a three-layer convolutional network on CIFAR. The network should have the following architecture:

1. Convolutional layer (with bias) with 32 5x5 filters, with zero-padding of 2
2. ReLU
3. Convolutional layer (with bias) with 16 3x3 filters, with zero-padding of 1
4. ReLU
5. Fully-connected layer (with bias) to compute scores for 100 classes

You should initialize your weight matrices using the `random_weight` function defined above, and you should initialize your bias vectors using the `zero_weight` function above.

You don't need to tune any hyperparameters, but if everything works correctly you should achieve an accuracy above **12% after one epoch**.

In [ ]:
```
learning_rate = 3e-3

channel_1 = 32
channel_2 = 16

conv_w1 = None
conv_b1 = None
conv_w2 = None
conv_b2 = None
fc_w = None
fc_b = None

################################################################################
# TODO: Initialize the parameters of a three-layer ConvNet.                    #
################################################################################
# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
H, W, C = 32, 32, 100
conv_w1 = random_weight((channel_1, 3, 5, 5))
conv_b1 = zero_weight((channel_1))
conv_w2 = random_weight((channel_2, channel_1, 3, 3))
conv_b2 = zero_weight((channel_2))
fc_w = random_weight((channel_2*H*W, C))
fc_b = zero_weight((C))

# print(conv_w1, conv_b1)
# print(conv_w2, conv_b2)
# *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
################################################################################
#                                END OF YOUR CODE                              #
################################################################################

params = [conv_w1, conv_b1, conv_w2, conv_b2, fc_w, fc_b]
train_part2(three_layer_convnet, params, learning_rate)
```

```
Iteration 100, loss = 4.2168
Checking accuracy on the val set
Got 63 / 1000 correct (6.30%)

Iteration 200, loss = 4.0626
Checking accuracy on the val set
Got 88 / 1000 correct (8.80%)

Iteration 300, loss = 4.2101
Checking accuracy on the val set
Got 113 / 1000 correct (11.30%)

Iteration 400, loss = 3.8881
Checking accuracy on the val set
Got 115 / 1000 correct (11.50%)

Iteration 500, loss = 3.9755
Checking accuracy on the val set
Got 126 / 1000 correct (12.60%)

Iteration 600, loss = 3.7519
Checking accuracy on the val set
Got 135 / 1000 correct (13.50%)

Iteration 700, loss = 3.9760
Checking accuracy on the val set
Got 145 / 1000 correct (14.50%)

Checking accuracy on the val set
Got 135 / 1000 correct (13.50%)
```

Out[ ]:
```
0.135
```

Part III. PyTorch Module API (10% of Grade)

Barebone PyTorch requires that we track all the parameter tensors by hand. This is fine for small networks with a few tensors, but it would be extremely inconvenient and error-prone to track tens or hundreds of tensors in larger networks.

PyTorch provides the `nn.Module` API for you to define arbitrary network architectures, while tracking every learnable parameters for you. In Part II, we implemented SGD ourselves. PyTorch also provides the `torch.optim` package that implements all the common optimizers, such as RMSProp, Adagrad, and Adam. It even supports approximate second-order methods like L-BFGS! You can refer to the doc for the exact specifications of each optimizer.

To use the Module API, follow the steps below:

1. Subclass `nn.Module`. Give your network class an intuitive name like `TwoLayerFC`.
2. In the constructor `__init__()`, define all the layers you need as class attributes. Layer objects like `nn.Linear` and `nn.Conv2d` are themselves `nn.Module` subclasses and contain learnable parameters, so that you don't have to instantiate the raw tensors yourself. `nn.Module` will track these internal parameters for you. Refer to the doc to learn more about the dozens of builtin layers. **Warning**: don't forget to call the `super().__init__()` first!
3. In the `forward()` method, define the *connectivity* of your network. You should use the attributes defined in `__init__` as function calls that take tensor as input and output the "transformed" tensor. Do *not* create any new layers with learnable parameters in `forward()`! All of them must be declared upfront in `__init__`.

After you define your Module subclass, you can instantiate it as an object and call it just like the NN forward function in part II.

Module API: Two-Layer Network

Here is a concrete example of a 2-layer fully connected network:

In [ ]:
```
class TwoLayerFC(nn.Module):
    def __init__(self, input_size, hidden_size, num_classes):
        super().__init__()
        # assign layer objects to class attributes
        self.fc1 = nn.Linear(input_size, hidden_size)
        # nn.init package contains convenient initialization methods
        # http://pytorch.org/docs/master/nn.html#torch-nn-init
        nn.init.kaiming_normal_(self.fc1.weight)
        self.fc2 = nn.Linear(hidden_size, num_classes)
        nn.init.kaiming_normal_(self.fc2.weight)

    def forward(self, x):
        # forward always defines connectivity
        x = flatten(x)
        scores = self.fc2(F.relu(self.fc1(x)))
        return scores

def test_TwoLayerFC():
    input_size = 50
    x = torch.zeros((64, input_size), dtype=dtype)  # minibatch size 64, feature dimension 50
    model = TwoLayerFC(input_size, 42, num_class)
    scores = model(x)
    print(scores.size())  # you should see [64, 100]
test_TwoLayerFC()
```

```
torch.Size([64, 100])
```

Module API: Three-Layer ConvNet

It's your turn to implement a 3-layer ConvNet followed by a fully connected layer. The network architecture should be the same as in Part II:

1. Convolutional layer with `channel_1` 5x5 filters with zero-padding of 2
2. ReLU
3. Convolutional layer with `channel_2` 3x3 filters with zero-padding of 1

4. ReLU
5. Fully-connected layer to `num_classes` classes

You should initialize the weight matrices of the model using the Kaiming normal initialization method.

HINT: http://pytorch.org/docs/stable/nn.html#conv2d

After you implement the three-layer ConvNet, the `test_ThreeLayerConvNet` function will run your implementation; it should print `(64, 10)` for the shape of the output scores.

In [ ]:
```
class ThreeLayerConvNet(nn.Module):
    def __init__(self, in_channel, channel_1, channel_2, num_classes):
        super().__init__()

        self.conv_1 = nn.Conv2d(in_channel, channel_1, (5,5), padding=2)
        nn.init.kaiming_normal_(self.conv_1.weight)
        self.conv_2 = nn.Conv2d(channel_1, channel_2, (3,3), padding=1)
        nn.init.kaiming_normal_(self.conv_2.weight)

        self.fc1 = nn.Linear(65536, num_classes)
        self.relu = nn.ReLU(inplace=True)



    def forward(self, x):
        scores = None
        ########################################################################
        # TODO: Implement the forward function for a 3-layer ConvNet. you      #
        # should use the layers you defined in __init__ and specify the        #
        # connectivity of those layers in forward()                            #
        ########################################################################
        # *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
        x = self.relu(self.conv_1(x))
        x = self.relu(self.conv_2(x))
        x = flatten(x)
        scores = self.fc1(x)
        # *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
        ########################################################################
        #                               END OF YOUR CODE                       #
        ########################################################################
        return scores


def test_ThreeLayerConvNet():
    x = torch.zeros((64, 3, 32, 32), dtype=dtype)  # minibatch size 64, image size [3, 32, 32]
    model = ThreeLayerConvNet(in_channel=3, channel_1=32, channel_2=64, num_classes=num_class)
    scores = model(x)
    print(scores.size())  # you should see [64, 100]
test_ThreeLayerConvNet()
```

```
torch.Size([64, 100])
```

Module API: Check Accuracy

Given the validation or test set, we can check the classification accuracy of a neural network.

This version is slightly different from the one in part II. You don't manually pass in the parameters anymore.

In [ ]:
```
def check_accuracy_part34(loader, model):
    if loader.dataset.train:
        print('Checking accuracy on validation set')
    else:
        print('Checking accuracy on test set')
    num_correct = 0
    num_samples = 0
    model.eval()  # set model to evaluation mode
    with torch.no_grad():
        for x, y in loader:
            x = x.to(device=device, dtype=dtype)  # move to device, e.g. GPU
            y = y.to(device=device, dtype=torch.long)
            scores = model(x)
            _, preds = scores.max(1)
            num_correct += (preds == y).sum()
```

```
            num_samples += preds.size(0)
        acc = float(num_correct) / num_samples
        print('Got %d / %d correct (%.2f)' % (num_correct, num_samples, 100 * acc))
    return acc
```

Module API: Training Loop

We also use a slightly different training loop. Rather than updating the values of the weights ourselves, we use an Optimizer object from the `torch.optim` package, which abstract the notion of an optimization algorithm and provides implementations of most of the algorithms commonly used to optimize neural networks.

In [ ]:
```
def train_part34(model, optimizer, epochs=1):
    """
    Train a model on CIFAR-10 using the PyTorch Module API.

    Inputs:
    - model: A PyTorch Module giving the model to train.
    - optimizer: An Optimizer object we will use to train the model
    - epochs: (Optional) A Python integer giving the number of epochs to train for

    Returns: The accuracy of the model
    """
    model = model.to(device=device)  # move the model parameters to CPU/GPU
    for e in range(epochs):
        for t, (x, y) in enumerate(loader_train):
            model.train()  # put model to training mode
            x = x.to(device=device, dtype=dtype)  # move to device, e.g. GPU
            y = y.to(device=device, dtype=torch.long)

            scores = model(x)
            loss = F.cross_entropy(scores, y)

            # Zero out all of the gradients for the variables which the optimizer
            # will update.
            optimizer.zero_grad()

            # This is the backwards pass: compute the gradient of the loss with
            # respect to each  parameter of the model.
            loss.backward()

            # Actually update the parameters of the model using the gradients
            # computed by the backwards pass.
            optimizer.step()

            if (t + 1) % print_every == 0:
                print('Epoch %d, Iteration %d, loss = %.4f' % (e, t + 1, loss.item()))
                check_accuracy_part34(loader_val, model)
                print()
    return check_accuracy_part34(loader_val, model)
```

Module API: Train a Two-Layer Network

Now we are ready to run the training loop. In contrast to part II, we don't explicitly allocate parameter tensors anymore.

Simply pass the input size, hidden layer size, and number of classes (i.e. output size) to the constructor of `TwoLayerFC`.

You also need to define an optimizer that tracks all the learnable parameters inside `TwoLayerFC`.

You don't need to tune any hyperparameters, but you should see model accuracies above 8% after training for one epoch.

In [ ]:
```
hidden_layer_size = 4000
learning_rate = 1e-3
model = TwoLayerFC(3 * 32 * 32, hidden_layer_size, num_class)
optimizer = optim.SGD(model.parameters(), lr=learning_rate)

train_part34(model, optimizer)
```

```
Epoch 0, Iteration 100, loss = 4.6942
Checking accuracy on validation set
Got 32 / 1000 correct (3.20)

Epoch 0, Iteration 200, loss = 4.6001
Checking accuracy on validation set
Got 61 / 1000 correct (6.10)

Epoch 0, Iteration 300, loss = 4.3677
Checking accuracy on validation set
Got 69 / 1000 correct (6.90)

Epoch 0, Iteration 400, loss = 4.2413
Checking accuracy on validation set
Got 84 / 1000 correct (8.40)

Epoch 0, Iteration 500, loss = 4.3595
Checking accuracy on validation set
Got 82 / 1000 correct (8.20)

Epoch 0, Iteration 600, loss = 4.3201
Checking accuracy on validation set
Got 91 / 1000 correct (9.10)

Epoch 0, Iteration 700, loss = 4.4071
Checking accuracy on validation set
Got 97 / 1000 correct (9.70)

Checking accuracy on validation set
Got 95 / 1000 correct (9.50)
```

Out[ ]: 0.095

Module API: Train a Three-Layer ConvNet

You should now use the Module API to train a three-layer ConvNet on CIFAR. This should look very similar to training the two-layer network! You don't need to tune any hyperparameters, but you should achieve **accuracy above 14% after training for one epoch**.

You should train the model using stochastic gradient descent without momentum.

In [ ]:
```python
learning_rate = 1e-3
channel_1 = 32
channel_2 = 64

model = None
optimizer = None
################################################################################
# TODO: Instantiate your ThreeLayerConvNet model and a corresponding optimizer #
################################################################################
# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
model = ThreeLayerConvNet(3, channel_1, channel_2, num_class)
optimizer = optim.SGD(params=model.parameters(), lr=learning_rate)
# *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
################################################################################
#                                END OF YOUR CODE
################################################################################

train_part34(model, optimizer, epochs=1)
```

```
Epoch 0, Iteration 100, loss = 4.0954
Checking accuracy on validation set
Got 78 / 1000 correct (7.80)

Epoch 0, Iteration 200, loss = 3.8737
Checking accuracy on validation set
Got 122 / 1000 correct (12.20)

Epoch 0, Iteration 300, loss = 3.9155
Checking accuracy on validation set
Got 135 / 1000 correct (13.50)

Epoch 0, Iteration 400, loss = 3.6420
Checking accuracy on validation set
Got 151 / 1000 correct (15.10)

Epoch 0, Iteration 500, loss = 3.5827
Checking accuracy on validation set
Got 151 / 1000 correct (15.10)

Epoch 0, Iteration 600, loss = 3.8418
Checking accuracy on validation set
Got 162 / 1000 correct (16.20)

Epoch 0, Iteration 700, loss = 3.9383
Checking accuracy on validation set
Got 159 / 1000 correct (15.90)

Checking accuracy on validation set
Got 160 / 1000 correct (16.00)
```

Out[ ]: 0.16

Part IV. PyTorch Sequential API (10% of Grade)

Part III introduced the PyTorch Module API, which allows you to define arbitrary learnable layers and their connectivity.

For simple models like a stack of feed forward layers, you still need to go through 3 steps: subclass `nn.Module`, assign layers to class attributes in `__init__`, and call each layer one by one in `forward()`. Is there a more convenient way?

Fortunately, PyTorch provides a container Module called `nn.Sequential`, which merges the above steps into one. It is not as flexible as `nn.Module`, because you cannot specify more complex topology than a feed-forward stack, but it's good enough for many use cases.

Sequential API: Two-Layer Network

Let's see how to rewrite our two-layer fully connected network example with `nn.Sequential`, and train it using the training loop defined above.

Again, you don't need to tune any hyperparameters here, but you shoud achieve above 17% accuracy after one epoch of training.

In [ ]:
```python
# We need to wrap `flatten` function in a module in order to stack it
# in nn.Sequential
class Flatten(nn.Module):
    def forward(self, x):
        return flatten(x)

hidden_layer_size = 4000
learning_rate = 1e-2

model = nn.Sequential(
    Flatten(),
    nn.Linear(3 * 32 * 32, hidden_layer_size),
    nn.ReLU(),
    nn.Linear(hidden_layer_size, num_class),
)

# you can use Nesterov momentum in optim.SGD
optimizer = optim.SGD(model.parameters(), lr=learning_rate,
                      momentum=0.9, nesterov=True)
```

```
train_part34(model, optimizer)
```

```
Epoch 0, Iteration 100, loss = 4.1383
Checking accuracy on validation set
Got 118 / 1000 correct (11.80)

Epoch 0, Iteration 200, loss = 3.9598
Checking accuracy on validation set
Got 142 / 1000 correct (14.20)

Epoch 0, Iteration 300, loss = 3.5631
Checking accuracy on validation set
Got 140 / 1000 correct (14.00)

Epoch 0, Iteration 400, loss = 3.8697
Checking accuracy on validation set
Got 151 / 1000 correct (15.10)

Epoch 0, Iteration 500, loss = 3.6436
Checking accuracy on validation set
Got 155 / 1000 correct (15.50)

Epoch 0, Iteration 600, loss = 3.6651
Checking accuracy on validation set
Got 153 / 1000 correct (15.30)

Epoch 0, Iteration 700, loss = 3.7589
Checking accuracy on validation set
Got 170 / 1000 correct (17.00)

Checking accuracy on validation set
Got 202 / 1000 correct (20.20)
```

Out[ ]: 0.202

Sequential API: Three-Layer ConvNet

Here you should use `nn.Sequential` to define and train a three-layer ConvNet with the same architecture we used in Part III:

1. Convolutional layer (with bias) with 32 5x5 filters, with zero-padding of 2
2. ReLU
3. Convolutional layer (with bias) with 16 3x3 filters, with zero-padding of 1
4. ReLU
5. Fully-connected layer (with bias) to compute scores for 100 classes

You should initialize your weight matrices using the `random_weight` function defined above, and you should initialize your bias vectors using the `zero_weight` function above.

You should optimize your model using stochastic gradient descent with Nesterov momentum 0.9.

Again, you don't need to tune any hyperparameters but you should see **accuracy above 14% after one epoch** of training.

In [ ]:
```
channel_1 = 32
channel_2 = 16
learning_rate = 1e-3

model = None
optimizer = None

################################################################################
# TODO: Rewrite the 2-layer ConvNet with bias from Part III with the           #
# Sequential API.                                                              #
################################################################################
# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
model = nn.Sequential(
    nn.Conv2d(3, 32, (5,5), padding=2),
    nn.ReLU(),
    nn.Conv2d(32, 16, (3,3), padding=1),
    nn.ReLU(),
```

```
    Flatten(),
    nn.Linear(16*32*32, num_class)
)

def init_weights(m):
    # print(type(m))
    if (type(m) in [nn.Conv2d]):
        zr_b = zero_weight(m.bias.shape)
        km_w = random_weight(m.weight.shape)
        m.bias.data = zr_b
        m.weight.data = km_w
    elif (type(m) in [nn.Linear]):
        zr_b = zero_weight(m.bias.shape)
        num_out, num_in = m.weight.shape
        linear_w_shape = (num_in, num_out) # reshape to match fan-in-dimension-first needs
        km_w = random_weight(linear_w_shape)
        km_w = km_w.view((num_out, num_in))
        m.bias.data = zr_b
        m.weight.data = km_w
    return 0
model.apply(init_weights)
optimizer = optim.SGD(params=model.parameters(), lr=learning_rate, momentum=0.9, nesterov=True)

# *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
################################################################################
#                                END OF YOUR CODE
################################################################################

train_part34(model, optimizer, epochs=1)
```

```
Epoch 0, Iteration 100, loss = 4.0649
Checking accuracy on validation set
Got 89 / 1000 correct (8.90)

Epoch 0, Iteration 200, loss = 4.2042
Checking accuracy on validation set
Got 121 / 1000 correct (12.10)

Epoch 0, Iteration 300, loss = 3.9313
Checking accuracy on validation set
Got 126 / 1000 correct (12.60)

Epoch 0, Iteration 400, loss = 3.5301
Checking accuracy on validation set
Got 141 / 1000 correct (14.10)

Epoch 0, Iteration 500, loss = 3.4397
Checking accuracy on validation set
Got 169 / 1000 correct (16.90)

Epoch 0, Iteration 600, loss = 3.3897
Checking accuracy on validation set
Got 187 / 1000 correct (18.70)

Epoch 0, Iteration 700, loss = 3.1412
Checking accuracy on validation set
Got 197 / 1000 correct (19.70)

Checking accuracy on validation set
Got 227 / 1000 correct (22.70)
```

Out[ ]: 0.227

Part V. Resnet10 Implementation (35% of Grade)

In this section, you will use the tools introduced above to implement the Resnet architecture. The Resnet architecture was introduced in: https://arxiv.org/pdf/1512.03385.pdf and it has become one of the most popular architectures used for computer vision. The key feature of the resnet architecture is the presence of skip connections which allow for better gradient flow even for very deep networks. Therefore, unlike vanilla CNNs introduced above, we can effectively build Resnets models having more than 100 layers. However, for the purposes of this exercise we will be using a smaller Resnet-10 architecture shown in the diagram below.

In the architecture above, the downsampling is performed in conv5_1. We recommend using the adam optimzer for training Resnet. You should see about *45%* accuracy in 10 epochs. The template below is based on the Module API but you are allowed to use other Pytorch APIs if you prefer.

In [ ]:
```
################################################################################
# TODO: Implement the forward function for the Resnet specified        #
# above. HINT: You might need to create a helper class to              #
# define a Resnet block and then use that block here to create         #
# the resnet layers i.e. conv2_x, conv3_x, conv4_x and conv5_x         #
################################################################################

# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
import torch.nn.functional as F  # useful stateless functions
def flatten(x):
    N = x.shape[0] # read in N, C, H, W
    return x.view(N, -1)  # "flatten" the C * H * W values into a single vector per image

class ResNetBlock(nn.Module):
    def __init__(self, kernel_size=3, in_channel=64, out_channel=64, stride1=1, stride2=1, padding1=1, padding2=1,
        super(ResNetBlock, self).__init__()
        self.conv1 = nn.Conv2d(in_channels=in_channel, out_channels=out_channel, kernel_size=kernel_size, stride=st
        self.conv2 = nn.Conv2d(in_channels=out_channel, out_channels=out_channel, kernel_size=kernel_size, stride=s
        self.batchNorm = batchNorm
        if self.batchNorm:
          self.bn1 = nn.BatchNorm2d(out_channel)
          self.bn2 = nn.BatchNorm2d(out_channel)
        self.do_projection = do_projection
        if self.do_projection:
            self.shortcut = nn.Sequential(nn.Conv2d(in_channels=in_channel, out_channels=out_channel, kernel_size=s
                                          nn.BatchNorm2d(out_channel))
    def forward(self, x):
        identity = x.clone()
        if self.do_projection:
            identity = self.shortcut(identity)
        x = self.conv1(x)
        if self.batchNorm:
          x = self.bn1(x)
        x = F.relu(x)
        x = self.conv2(x)
        if self.batchNorm:
          x = self.bn2(x)
        x = x + identity
        x = F.relu(x)
        return x

class ResNet(nn.Module):
    def __init__(self, num_classes=100, batchNorm=True):
        super(ResNet, self).__init__()
        self.conv1 = nn.Conv2d(3, 64, (7,7), stride=2, padding=3) # output size 16*16
        self.batchNorm = batchNorm
        if self.batchNorm:
          self.bn1 = nn.BatchNorm2d(64)
        self.maxpool1 = nn.MaxPool2d(kernel_size=3, stride=2, padding=1) # output size 8*8, 64 channels
        self.conv2 = ResNetBlock(kernel_size=3, in_channel=64, out_channel=64, do_projection=False, batchNorm=batcl
        self.conv3 = ResNetBlock(kernel_size=3, in_channel=64, out_channel=128, do_projection=True, batchNorm=batcl
        self.conv4 = ResNetBlock(kernel_size=3, in_channel=128, out_channel=256, do_projection=True, batchNorm=bat
        self.conv5 = ResNetBlock(kernel_size=3, in_channel=256, stride1=2, padding1=1, stride2=1, padding2=1, out_
                                 do_projection=True, sc_kernel_size=1, sc_stride=2, sc_padding=0, batchNorm=batchN
        self.avgpool1 = nn.AdaptiveAvgPool2d((1, 1))
        self.fc1 = nn.Linear(1*1*512, num_classes)
    def forward(self, x): # input size N*3*32*32
        x = self.conv1(x)
        if self.batchNorm:
          x = self.bn1(x)
        x = F.relu(x)
        x = self.maxpool1(x)
        x = self.conv2(x)
        x = self.conv3(x)
        x = self.conv4(x)
        x = self.conv5(x)
        x = self.avgpool1(x)
        x = flatten(x)
        x = self.fc1(x) # returning raw linear outputs for 100 classes
        return x
```

```
################################################################################
#                               END OF YOUR CODE                               #
################################################################################
```

In [ ]:
```
learning_rate = 1e-3

model = None
optimizer = None

################################################################################
# TODO: Instantiate and train Resnet-10.                                       #
################################################################################
# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
model = ResNet(batchNorm=False)
optimizer = optim.Adam(params=model.parameters(), lr=learning_rate)

# *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
################################################################################
#                                  END OF YOUR CODE
################################################################################

print_every = 700
train_part34(model, optimizer, epochs=10)
print_every = 100
```

```
Epoch 0, Iteration 700, loss = 2.8744
Checking accuracy on validation set
Got 222 / 1000 correct (22.20)

Epoch 1, Iteration 700, loss = 2.5381
Checking accuracy on validation set
Got 356 / 1000 correct (35.60)

Epoch 2, Iteration 700, loss = 2.0900
Checking accuracy on validation set
Got 366 / 1000 correct (36.60)

Epoch 3, Iteration 700, loss = 1.8775
Checking accuracy on validation set
Got 413 / 1000 correct (41.30)

Epoch 4, Iteration 700, loss = 1.6261
Checking accuracy on validation set
Got 450 / 1000 correct (45.00)

Epoch 5, Iteration 700, loss = 1.3894
Checking accuracy on validation set
Got 470 / 1000 correct (47.00)

Epoch 6, Iteration 700, loss = 1.5636
Checking accuracy on validation set
Got 455 / 1000 correct (45.50)

Epoch 7, Iteration 700, loss = 1.1394
Checking accuracy on validation set
Got 481 / 1000 correct (48.10)

Epoch 8, Iteration 700, loss = 1.1861
Checking accuracy on validation set
Got 495 / 1000 correct (49.50)

Epoch 9, Iteration 700, loss = 0.8012
Checking accuracy on validation set
Got 463 / 1000 correct (46.30)

Checking accuracy on validation set
Got 465 / 1000 correct (46.50)
```

BatchNorm

Now you will also introduce the Batch-Normalization layer within the Resnet architecture implemented above. Please add a batch normalization layer after each conv in your network before applying the activation function (i.e. the order should be conv->BatchNorm->Relu). Please read the section 3.4 from the Resnet paper (https://arxiv.org/pdf/1512.03385.pdf.

Feel free to re-use the Resnet class that you have implemented above by introducing a boolean flag for batch normalization.

After trying out batch-norm, please discuss the performance comparison between Resnet with BatchNorm and without BatchNorm and possible reasons for why one performs better than the other.

In [ ]:
```
learning_rate = 1e-3

model = None
optimizer = None

################################################################################
# TODO: InstantiateResnet with BatchNorm                                       #
################################################################################
# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
model = ResNet(batchNorm=True)
optimizer = optim.Adam(params=model.parameters(), lr=learning_rate)


# *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
################################################################################
#                                 END OF YOUR CODE
################################################################################

print_every = 700
train_part34(model, optimizer, epochs=10)
print_every = 100
```

```
Epoch 0, Iteration 700, loss = 3.2622
Checking accuracy on validation set
Got 228 / 1000 correct (22.80)

Epoch 1, Iteration 700, loss = 2.4321
Checking accuracy on validation set
Got 304 / 1000 correct (30.40)

Epoch 2, Iteration 700, loss = 2.6425
Checking accuracy on validation set
Got 377 / 1000 correct (37.70)

Epoch 3, Iteration 700, loss = 1.9153
Checking accuracy on validation set
Got 410 / 1000 correct (41.00)

Epoch 4, Iteration 700, loss = 1.7098
Checking accuracy on validation set
Got 465 / 1000 correct (46.50)

Epoch 5, Iteration 700, loss = 2.2182
Checking accuracy on validation set
Got 485 / 1000 correct (48.50)

Epoch 6, Iteration 700, loss = 1.1542
Checking accuracy on validation set
Got 511 / 1000 correct (51.10)

Epoch 7, Iteration 700, loss = 1.0563
Checking accuracy on validation set
Got 519 / 1000 correct (51.90)

Epoch 8, Iteration 700, loss = 1.3591
Checking accuracy on validation set
Got 523 / 1000 correct (52.30)

Epoch 9, Iteration 700, loss = 1.2013
Checking accuracy on validation set
Got 523 / 1000 correct (52.30)

Checking accuracy on validation set
Got 507 / 1000 correct (50.70)
```

Discussion on BatchNorm

The validation accuracy after BatchNorm (50.7%) turns out to be higher than that without BatchNorm (46.5%). This is probably because BatchNorm can act as a regularizer adding noise to each mini-batch of images by normalizing, and thus reducing overfitting and improving the model's generalizability to unseen data. It may also because BatchNorm stabilizes training by mitigating the problems of vanishing or exploding gradients in the network, facilitating faster and better convergence to the local optimum.

Batch Size

In this exercise, we will study the effect of batch size on performance of ResNet (with BatchNorm).

Specifically, you should try batch sizes of 32, 64 and 128 and describe the effect of varying batch size. You should print the validation accuracy of using each batch size in different rows.

After trying out different batch size, please discuss the effect of different batch sizes and possible reasons for that (either they are showing some trend or not).

In [ ]:
```
print_every = 9999
batch_sizes = [32, 64, 128]
learning_rate = 1e-3
model = None
optimizer = None

################################################################################
# TODO: Try Resnet with different batch sizes. Hint: You will need to          #
#  create a new dataloader with appropriate batch size for each experiment.    #
```

```
# You will also need to store the final accuracy for each experiment           #
################################################################################
# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
final_acc = []

# training with different batch sizes
for batch_size in batch_sizes:
  loader_train = DataLoader(cifar100_train, batch_size=batch_size, num_workers=2,
                            sampler=sampler.SubsetRandomSampler(range(NUM_TRAIN))) # 49000 train data

  loader_val = DataLoader(cifar100_val, batch_size=batch_size, num_workers=2,
                          sampler=sampler.SubsetRandomSampler(range(NUM_TRAIN, 50000))) # 1000 val data

  loader_test = DataLoader(cifar100_test, batch_size=batch_size, num_workers=2)
  model = ResNet(batchNorm=True)
  optimizer = optim.Adam(params=model.parameters(), lr=learning_rate)
  acc = train_part34(model, optimizer, epochs=10)
  final_acc.append(acc)

# print out accuracies for different batch sizes
print()
for j in range(len(batch_sizes)):
  batch_size, acc = batch_sizes[j], final_acc[j]
  print(f"Validation accuracy for batch_size={batch_size}: {acc}")

# *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
################################################################################
#                                END OF YOUR CODE
################################################################################
print_every = 100
```

```
Checking accuracy on validation set
Got 502 / 1000 correct (50.20)
Checking accuracy on validation set
Got 514 / 1000 correct (51.40)
Checking accuracy on validation set
Got 511 / 1000 correct (51.10)

Validation accuracy for batch_size=32: 0.502
Validation accuracy for batch_size=64: 0.514
Validation accuracy for batch_size=128: 0.511
```

Discuss effect of Batch Size

The general but not too obvious trend is that as the batch size increases, the overall validation accuracy increases:

We see from the results that the network with batch_size=128/64 gives a slightly higher validation accuracy. This is probably because the large batch size allows a more stable convergence to the local optimum by computing the gradient using more data within each batch.

The results with batch_size=32 is slightly poorer, which is probably because of the noisy gradients caused by the batches which contain too few samples, which leads to worse generalization.

The results with batch_size=32/64/128 actually don't differ much. This may be due to the usage of Adam, which uses adaptive learning rate techniques to make gradients move more stably and to mitigate the effects of batch size variations. It may also be due to the batch normalizations since it can avoid too aggressive gradients caused by the variations on batch size and hence stabilize training.

Part VI. CIFAR-100 open-ended challenge (25% of Grade)

In this section, you can experiment with whatever ConvNet architecture you'd like on CIFAR-100 **except Resnet** because we already tried it.

Now it's your job to experiment with architectures, hyperparameters, loss functions, and optimizers to train a model that achieves **at least 48%** accuracy on the CIFAR-100 **validation** set within 10 epochs. You can use the check_accuracy and train functions from above. You can use either `nn.Module` or `nn.Sequential` API.

Describe what you did at the end of this notebook.

Here are the official API documentation for each component.

- Layers in torch.nn package: http://pytorch.org/docs/stable/nn.html
- Activations: http://pytorch.org/docs/stable/nn.html#non-linear-activations
- Loss functions: http://pytorch.org/docs/stable/nn.html#loss-functions
- Optimizers: http://pytorch.org/docs/stable/optim.html

Things you might try:

- **Filter size**: Above we used 5x5; would smaller filters be more efficient?
- **Adam Optimizer**: Above we used SGD optimizer, would an Adam optimizer do better?
- **Number of filters**: Above we used 32 filters. Do more or fewer do better?
- **Pooling vs Strided Convolution**: Do you use max pooling or just stride convolutions?
- **Batch normalization**: Try adding spatial batch normalization after convolution layers and vanilla batch normalization after affine layers. Do your networks train faster? You can also try out LayerNorm and GroupNorm.
- **Network architecture**: Can you do better with a deep network? Good architectures to try include:
  - [conv-relu-pool]xN -> [affine]xM -> [softmax or SVM]
  - [conv-relu-conv-relu-pool]xN -> [affine]xM -> [softmax or SVM]
  - [batchnorm-relu-conv]xN -> [affine]xM -> [softmax or SVM]
- **Global Average Pooling**: Instead of flattening and then having multiple affine layers, perform convolutions until your image gets small (7x7 or so) and then perform an average pooling operation to get to a 1x1 image picture (1, 1 , Filter#), which is then reshaped into a (Filter#) vector. This is used in Google's Inception Network (See Table 1 for their architecture).
- **Regularization**: Add l2 weight regularization, or perhaps use Dropout.

Tips for training

For each network architecture that you try, you should tune the learning rate and other hyperparameters. When doing this there are a couple important things to keep in mind:

- If the parameters are working well, you should see improvement within a few hundred iterations
- Remember the coarse-to-fine approach for hyperparameter tuning: start by testing a large range of hyperparameters for just a few training iterations to find the combinations of parameters that are working at all.
- Once you have found some sets of parameters that seem to work, search more finely around these parameters. You may need to train for more epochs.
- You should use the validation set for hyperparameter search, and save your test set for evaluating your architecture on the best parameters as selected by the validation set.

Want more improvements?

There are many other features you can implement to try and improve your performance.

- Alternative optimizers: you can try Adam, Adagrad, RMSprop, etc.
- Alternative activation functions such as leaky ReLU, parametric ReLU, ELU, or MaxOut.
- Model ensembles
- Data augmentation
- New Architectures
  - DenseNets where inputs into previous layers are concatenated together.

Have fun and may the gradients be with you!

In [ ]:
```
################################################################################
# TODO:                                                                        #
# Experiment with any architectures, optimizers, and hyperparameters.          #
# Achieve AT LEAST 48% accuracy on the *validation set* within 10 epochs.      #
#                                                                              #
# Note that you can use the check_accuracy function to evaluate on either      #
# the test set or the validation set, by passing either loader_test or         #
# loader_val as the second argument to check_accuracy. You should not touch    #
# the test set until you have finished your architecture and  hyperparameter   #
# tuning, and only run the test set once at the end to report a final value.   #
################################################################################
```

```
model = None
optimizer = None

# *****START OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
import torch.nn.functional as F  # useful stateless functions
learning_rate = 3e-3
def flatten(x):
    N = x.shape[0] # read in N, C, H, W
    return x.view(N, -1)  # "flatten" the C * H * W values into a single vector per image

class BottleneckBlock(nn.Module):
    def __init__(self, in_channel=64, mid_channels=64, out_channel=64, stride1=1, stride2=1, padding1=0, padding2=1
        super(BottleneckBlock, self).__init__()
        self.conv1 = nn.Conv2d(in_channels=in_channel, out_channels=mid_channels, kernel_size=1, stride=stride1, pa
        self.conv2 = nn.Conv2d(in_channels=mid_channels, out_channels=mid_channels, kernel_size=3, stride=stride2,
        self.conv3 = nn.Conv2d(in_channels=mid_channels, out_channels=out_channel, kernel_size=1, stride=stride1, p
        self.bn1 = nn.BatchNorm2d(mid_channels)
        self.bn2 = nn.BatchNorm2d(mid_channels)
        self.bn3 = nn.BatchNorm2d(out_channel)
        self.do_projection = ((in_channel != out_channel) or dn_sample)
        if self.do_projection:
            self.shortcut = nn.Sequential(nn.Conv2d(in_channels=in_channel, out_channels=out_channel, kernel_size=1
                                          nn.BatchNorm2d(out_channel))
    def forward(self, x):
        identity = x.clone()
        if self.do_projection:
            identity = self.shortcut(identity)
        x = self.bn1(self.conv1(x))
        x = F.relu(x)
        x = self.bn2(self.conv2(x))
        x = F.relu(x)
        x = self.bn3(self.conv3(x))
#         print(x.shape, identity.shape)
        x = x + identity
        x = F.relu(x)
        return x

class ResNetBottleneck(nn.Module):
    def __init__(self, num_classes=100):
        super(ResNetBottleneck, self).__init__()
        self.conv1 = nn.Conv2d(3, 64, (7,7), stride=2, padding=3) # output size 16*16
        self.bn1 = nn.BatchNorm2d(64)
        self.maxpool1 = nn.MaxPool2d(kernel_size=3, stride=2, padding=1) # output size 8*8, 64 channels
        self.conv2_1 = BottleneckBlock(in_channel=64, mid_channels=64, out_channel=256) # output size 8*8, 64 chann
        self.conv2_2 = BottleneckBlock(in_channel=256, mid_channels=64, out_channel=256) # output size 8*8, 64 chan
        self.dropout1 = nn.Dropout(0.2)
        self.conv3_1 = BottleneckBlock(in_channel=256, mid_channels=128, out_channel=512) # output size 8*8, 128 ch
        self.conv3_2 = BottleneckBlock(in_channel=512, mid_channels=128, out_channel=512) # output size 8*8, 128 ch
        self.dropout2 = nn.Dropout(0.2)
#         self.maxpool2 = nn.MaxPool2d(2, stride=2)
        self.conv4_1 = BottleneckBlock(in_channel=512, mid_channels=256, out_channel=1024) # output size 8*8, 256 c
        self.conv4_2 = BottleneckBlock(in_channel=1024, mid_channels=256, out_channel=1024) # output size 8*8, 256
        self.dropout3 = nn.Dropout(0.2)
        self.conv5_1 = BottleneckBlock(in_channel=1024, mid_channels=512, out_channel=2048) # output size 4*4, 512
#         self.conv5_2 = BottleneckBlock(in_channel=2048, mid_channels=512, out_channel=2048, stride1=2, padding1=1
#                                        dn_sample=True, sc_stride=2) # output size 4*4, 512 channels, down-sampled
        self.conv5_2 = BottleneckBlock(in_channel=2048, mid_channels=512, out_channel=2048) # output size 4*4, 512
        self.dropout4 = nn.Dropout(0.2)
        self.maxpool2 = nn.MaxPool2d(2, stride=2)
        self.avgpool1 = nn.AdaptiveAvgPool2d((1, 1))
        self.fc1 = nn.Linear(1*1*2048, num_classes)
    def forward(self, x): # input size N*3*32*32
        x = self.bn1(self.conv1(x))
        x = F.relu(x)
        x = self.maxpool1(x)
        x = self.conv2_1(x)
        x = self.conv2_2(x)
#         x = self.dropout1(x)
        x = self.conv3_1(x)
        x = self.conv3_2(x)
#         x = self.dropout2(x)
        x = self.conv4_1(x)
        x = self.conv4_2(x)
        x = self.dropout3(x)
        x = self.conv5_1(x)
        x = self.conv5_2(x)
```

```
        x = self.dropout4(x)
        x = self.avgpool1(self.maxpool2(x))
        x = flatten(x)
        x = self.fc1(x) # returning raw linear outputs for 100 classes
#         x = self.fc2(x)
        return x

batch_size = 64
loader_train = DataLoader(cifar100_train, batch_size=batch_size, num_workers=2,
                          sampler=sampler.SubsetRandomSampler(range(NUM_TRAIN))) # 49000 train data

loader_val = DataLoader(cifar100_val, batch_size=batch_size, num_workers=2,
                        sampler=sampler.SubsetRandomSampler(range(NUM_TRAIN, 50000))) # 1000 val data

loader_test = DataLoader(cifar100_test, batch_size=batch_size, num_workers=2)

model = ResNetBottleneck()
optimizer = optim.Adam(params=model.parameters(), lr=learning_rate)




# *****END OF YOUR CODE (DO NOT DELETE/MODIFY THIS LINE)*****
################################################################################
#                                 END OF YOUR CODE
################################################################################

# You should get at least 48% accuracy.
print_every = 700
train_part34(model, optimizer, epochs=10)
print_every = 100
```

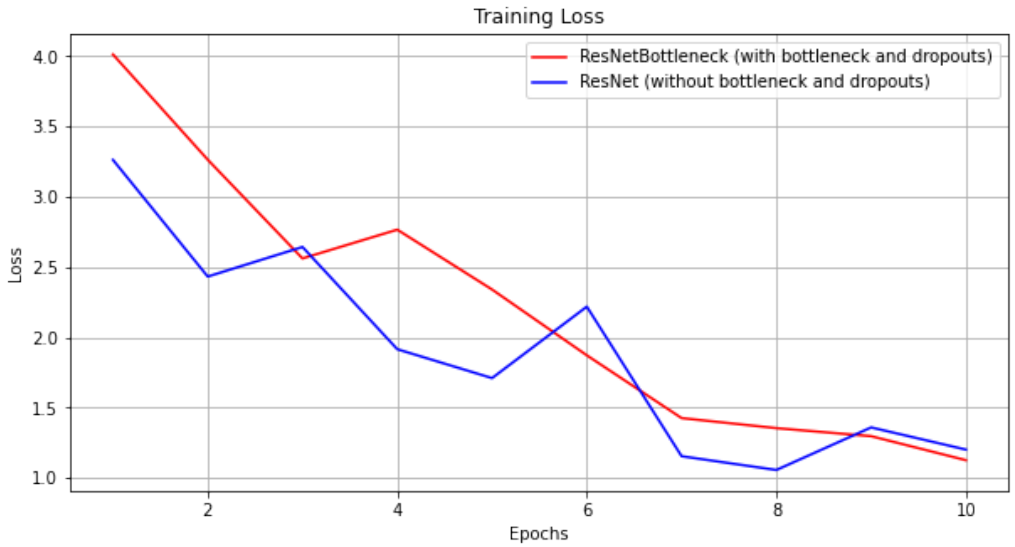
```
Epoch 0, Iteration 700, loss = 4.0120
Checking accuracy on validation set
Got 104 / 1000 correct (10.40)

Epoch 1, Iteration 700, loss = 3.2635
Checking accuracy on validation set
Got 182 / 1000 correct (18.20)

Epoch 2, Iteration 700, loss = 2.5610
Checking accuracy on validation set
Got 290 / 1000 correct (29.00)

Epoch 3, Iteration 700, loss = 2.7657
Checking accuracy on validation set
Got 337 / 1000 correct (33.70)

Epoch 4, Iteration 700, loss = 2.3400
Checking accuracy on validation set
Got 420 / 1000 correct (42.00)

Epoch 5, Iteration 700, loss = 1.8719
Checking accuracy on validation set
Got 433 / 1000 correct (43.30)

Epoch 6, Iteration 700, loss = 1.4254
Checking accuracy on validation set
Got 473 / 1000 correct (47.30)

Epoch 7, Iteration 700, loss = 1.3539
Checking accuracy on validation set
Got 522 / 1000 correct (52.20)

Epoch 8, Iteration 700, loss = 1.2968
Checking accuracy on validation set
Got 508 / 1000 correct (50.80)

Epoch 9, Iteration 700, loss = 1.1259
Checking accuracy on validation set
Got 545 / 1000 correct (54.50)

Checking accuracy on validation set
Got 552 / 1000 correct (55.20)
```

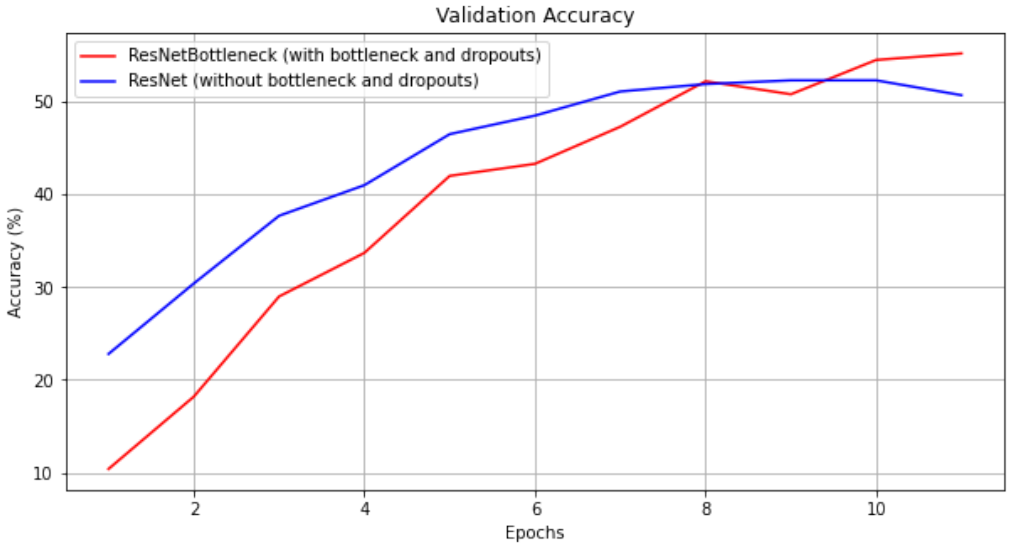
Describe what you did (10% of Grade)

In the cell below you should write an explanation of what you did, any additional features that you implemented, and/or any graphs that you made in the process of training and evaluating your network.

I used the bottleneck architecture for the block design (for deeper networks) to replace the original block design (for shallow networks) in ResNet. I also made the network deeper to improve its capability of feature extractions. In this sense, I built an approximate version of 50-layer ResNet with the bottleneck architecture. In addition, I used 2 dropout layers with rate 0.2 in the middle of the feature extraction network to regularize training. I also added a max pooling layer to down-sample the features from 8*8 to 4*4 instead of using the original convolution layer. I also used a larger learning rate 3e-3 to make convergence faster. The figure below shows that the validation accuracy increases and the loss decreases more stably as the bottleneck architecture with dropouts is adopted.

In [ ]:
```
import matplotlib.pyplot as plt

# Training logs for Model 1 (ResNet with bottleneck and dropouts)
model1_loss = [4.0120, 3.2635, 2.5610, 2.7657, 2.3400, 1.8719, 1.4254, 1.3539, 1.2968, 1.1259]
model1_accuracy = [10.40, 18.20, 29.00, 33.70, 42.00, 43.30, 47.30, 52.20, 50.80, 54.50, 55.20]

# Training logs for Model 2 (ResNet without bottleneck and dropouts)
model2_loss = [3.2622, 2.4321, 2.6425, 1.9153, 1.7098, 2.2182, 1.1542, 1.0563, 1.3591, 1.2013]
model2_accuracy = [22.80, 30.40, 37.70, 41.00, 46.50, 48.50, 51.10, 51.90, 52.30, 52.30, 50.70]

epochs = range(1, len(model1_loss) + 1)

# Plotting Loss
plt.figure(figsize=(10, 5))
plt.plot(epochs, model1_loss, 'r', label='ResNetBottleneck (with bottleneck and dropouts)')
plt.plot(epochs, model2_loss, 'b', label='ResNet (without bottleneck and dropouts)')
plt.title('Training Loss')
plt.xlabel('Epochs')
plt.ylabel('Loss')
plt.legend()
plt.grid(True)
plt.show()

epochs = range(1, len(model1_loss) + 2)

# Plotting Validation Accuracy
plt.figure(figsize=(10, 5))
plt.plot(epochs, model1_accuracy, 'r', label='ResNetBottleneck (with bottleneck and dropouts)')
plt.plot(epochs, model2_accuracy, 'b', label='ResNet (without bottleneck and dropouts)')
plt.title('Validation Accuracy')
plt.xlabel('Epochs')
plt.ylabel('Accuracy (%)')
plt.legend()
plt.grid(True)
plt.show()
```





Test set -- run this only once

Now that we've gotten a result we're happy with, we test our final model on the test set (which you should store in best_model). Think about how this compares to your validation set accuracy.

In [ ]:
```
best_model = model
check_accuracy_part34(loader_test, best_model)
```

```
Checking accuracy on test set
Got 5401 / 10000 correct (54.01)
```

Out[ ]:
```
0.5401
```

The accuracy on the test set is similar to that on the validation set even though their sizes are different. This is because both sets aim at testing the performance of the model on a more general dataset. Both results show that the generalization capability of the model is satisfactory.