Problem_1

February 25, 2023

1 Problem 1: Basics of Neural Networks

- Learning Objective: In this problem, you are asked to implement a basic multi-layer fully connected neural network from scratch, including forward and backward passes of certain essential layers, to perform an image classification task on the CIFAR100 dataset. You need to implement essential functions in different indicated python files under directory `lib`.
- Provided Code: We provide the skeletons of classes you need to complete. Forward checking and gradient checkings are provided for verifying your implementation as well.
- TODOs: You are asked to implement the forward passes and backward passes for standard layers and loss functions, various widely-used optimizers, and part of the training procedure. And finally we want you to train a network from scratch on your own. Also, there are inline questions you need to answer. See `README.md` to set up your environment.

[54]:
```
from lib.mlp.fully_conn import *
from lib.mlp.layer_utils import *
from lib.datasets import *
from lib.mlp.train import *
from lib.grad_check import *
from lib.optim import *
import numpy as np
import matplotlib.pyplot as plt

%matplotlib inline
plt.rcParams['figure.figsize'] = (10.0, 8.0) # set default size of plots
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'

# for auto-reloading external modules
# see http://stackoverflow.com/questions/1907993/
 ↪autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

1.1 Loading the data (CIFAR-100 with 20 superclasses)

In this homework, we will be classifying images from the CIFAR-100 dataset into the 20 superclasses. More information about the CIFAR-100 dataset and the 20 superclasses can be found here.

Download the CIFAR-100 data files here, and save the `.mat` files to the `data/cifar100` directory.

Load the dataset.

[2]:
```python
data = CIFAR100_data('data/cifar100/')
for k, v in data.items():
    if type(v) == np.ndarray:
        print ("Name: {} Shape: {}, {}".format(k, v.shape, type(v)))
    else:
        print("{}: {}".format(k, v))
label_names = data['label_names']
mean_image = data['mean_image'][0]
std_image = data['std_image'][0]
```

```
Name: data_train Shape: (40000, 32, 32, 3), <class 'numpy.ndarray'>
Name: labels_train Shape: (40000,), <class 'numpy.ndarray'>
Name: data_val Shape: (10000, 32, 32, 3), <class 'numpy.ndarray'>
Name: labels_val Shape: (10000,), <class 'numpy.ndarray'>
Name: data_test Shape: (10000, 32, 32, 3), <class 'numpy.ndarray'>
Name: labels_test Shape: (10000,), <class 'numpy.ndarray'>
label_names: ['aquatic_mammals', 'fish', 'flowers', 'food_containers',
'fruit_and_vegetables', 'household_electrical_devices', 'household_furniture',
'insects', 'large_carnivores', 'large_man-made_outdoor_things',
'large_natural_outdoor_scenes', 'large_omnivores_and_herbivores',
'medium_mammals', 'non-insect_invertebrates', 'people', 'reptiles',
'small_mammals', 'trees', 'vehicles_1', 'vehicles_2']
Name: mean_image Shape: (1, 1, 1, 3), <class 'numpy.ndarray'>
Name: std_image Shape: (1, 1, 1, 3), <class 'numpy.ndarray'>
```

1.2 Implement Standard Layers

You will now implement all the following standard layers commonly seen in a fully connected neural network (aka multi-layer perceptron, MLP). Please refer to the file `lib/mlp/layer_utils.py`. Take a look at each class skeleton, and we will walk you through the network layer by layer. We provide results of some examples we pre-computed for you for checking the forward pass, and also the gradient checking for the backward pass.

1.3 FC Forward [2pt]

In the class skeleton `flatten` and `fc` in `lib/mlp/layer_utils.py`, please complete the forward pass in function `forward`. The input to the `fc` layer may not be of dimension (batch size, features size), it could be an image or any higher dimensional data. We want to convert the input to have a shape of (batch size, features size). Make sure that you handle this dimensionality issue.

[3]:
```
%reload_ext autoreload

# Test the fc forward function
input_bz = 3 # batch size
input_dim = (7, 6, 4)
output_dim = 4

input_size = input_bz * np.prod(input_dim)
weight_size = output_dim * np.prod(input_dim)

flatten_layer = flatten(name="flatten_test")
single_fc = fc(np.prod(input_dim), output_dim, init_scale=0.02, name="fc_test")

x = np.linspace(-0.1, 0.4, num=input_size).reshape(input_bz, *input_dim)
w = np.linspace(-0.2, 0.2, num=weight_size).reshape(np.prod(input_dim),␣
 ↪output_dim)
b = np.linspace(-0.3, 0.3, num=output_dim)

single_fc.params[single_fc.w_name] = w
single_fc.params[single_fc.b_name] = b

out = single_fc.forward(flatten_layer.forward(x))

correct_out = np.array([[0.63910291, 0.83740057, 1.03569824, 1.23399591],
                        [0.61401587, 0.82903823, 1.04406058, 1.25908294],
                        [0.58892884, 0.82067589, 1.05242293, 1.28416997]])

# Compare your output with the above pre-computed ones.
# The difference should not be larger than 1e-8
print ("Difference: ", rel_error(out, correct_out))
```

```
Difference:  4.026016656214849e-09
```

1.4 FC Backward [2pt]

Please complete the function `backward` as the backward pass of the `flatten` and `fc` layers. Follow the instructions in the comments to store gradients into the predefined dictionaries in the attributes of the class. Parameters of the layer are also stored in the predefined dictionary.

[4]:
```
%reload_ext autoreload

# Test the fc backward function
inp = np.random.randn(15, 2, 2, 3)
w = np.random.randn(12, 15)
b = np.random.randn(15)
dout = np.random.randn(15, 15)

flatten_layer = flatten(name="flatten_test")
```

```python
x = flatten_layer.forward(inp)
single_fc = fc(np.prod(x.shape[1:]), 15, init_scale=5e-2, name="fc_test")
single_fc.params[single_fc.w_name] = w
single_fc.params[single_fc.b_name] = b

dx_num = eval_numerical_gradient_array(lambda x: single_fc.forward(x), x, dout)
dw_num = eval_numerical_gradient_array(lambda w: single_fc.forward(x), w, dout)
db_num = eval_numerical_gradient_array(lambda b: single_fc.forward(x), b, dout)

out = single_fc.forward(x)
dx = single_fc.backward(dout)
dw = single_fc.grads[single_fc.w_name]
db = single_fc.grads[single_fc.b_name]
dinp = flatten_layer.backward(dx)

# The error should be around 1e-9
print("dx Error: ", rel_error(dx_num, dx))
# The errors should be around 1e-10
print("dw Error: ", rel_error(dw_num, dw))
print("db Error: ", rel_error(db_num, db))
# The shapes should be same
print("dinp Shape: ", dinp.shape, inp.shape)
```

```
dx Error:  1.3252312577124126e-09
dw Error:  6.33961240581919e-10
db Error:  5.974585394398448e-11
dinp Shape:  (15, 2, 2, 3) (15, 2, 2, 3)
```

1.5 GeLU Forward [2pt]

In the class skeleton `gelu` in `lib/mlp/layer_utils.py`, please complete the `forward` pass.

GeLU is a smooth version of ReLU and it's used in pre-training LLMs such as GPT-3 and BERT.

$$\mathrm{GeLU}(x) = x\Phi(x) \approx 0.5x(1 + \tanh(\sqrt{2/\pi}(x + 0.044715x^3)))$$

Where $\Phi(x)$ is the CDF for standard Gaussian random variables. You should use the approximate version to compute forward and backward pass.

[5]:
```python
%reload_ext autoreload

# Test the leaky_relu forward function
x = np.linspace(-1.5, 1.5, num=12).reshape(3, 4)
gelu_f = gelu(name="gelu_f")

out = gelu_f.forward(x)
correct_out = np.array([[-0.10042842, -0.13504766, -0.16231757, -0.1689214 ],
                        [-0.13960493, -0.06078651,  0.07557713,  0.26948598],
```

```
                     [ 0.51289678,  0.79222788,  1.09222506,  1.39957158]])

# Compare your output with the above pre-computed ones.
# The difference should not be larger than 1e-7
print ("Difference: ", rel_error(out, correct_out))
```

```
Difference:  1.8037541876132445e-08
```

1.6 GeLU Backward [2pt]

Please complete the `backward` pass of the class `gelu`.

[6]:
```
%reload_ext autoreload

# Test the relu backward function
x = np.random.randn(15, 15)
dout = np.random.randn(*x.shape)
gelu_b = gelu(name="gelu_b")

dx_num = eval_numerical_gradient_array(lambda x: gelu_b.forward(x), x, dout)

out = gelu_b.forward(x)
dx = gelu_b.backward(dout)

# The error should not be larger than 1e-4, since we are using an approximate␣
↪version of GeLU activation.
print ("dx Error: ", rel_error(dx_num, dx))
```

```
dx Error:  1.5818228671769437e-05
```

1.7 Dropout Forward [2pt]

In the class `dropout` in `lib/mlp/layer_utils.py`, please complete the `forward` pass.
Remember that the dropout is **only applied during training phase**, you should pay attention to this while implementing the function. ##### Important Note1: The probability argument input to the function is the "keep probability": probability that each activation is kept. ##### Important Note2: If the keep_prob is set to 1, make it as no dropout.

[7]:
```
%reload_ext autoreload

x = np.random.randn(100, 100) + 5.0

print ("----------------------------------------------------------------")
for p in [0, 0.25, 0.50, 0.75, 1]:
    dropout_f = dropout(keep_prob=p)
    out = dropout_f.forward(x, True)
    out_test = dropout_f.forward(x, False)
```

```python
    # Mean of output should be similar to mean of input
    # Means of output during training time and testing time should be similar
    print ("Dropout Keep Prob = ", p)
    print ("Mean of input: ", x.mean())
    print ("Mean of output during training time: ", out.mean())
    print ("Mean of output during testing time: ", out_test.mean())
    print ("Fraction of output set to zero during training time: ", (out == 0).
↪mean())
    print ("Fraction of output set to zero during testing time: ", (out_test ==␣
↪0).mean())
    print ("----------------------------------------------------------------")
```

```
----------------------------------------------------------------
Dropout Keep Prob =  0
Mean of input:  5.011213817030235
Mean of output during training time:  5.011213817030235
Mean of output during testing time:  5.011213817030235
Fraction of output set to zero during training time:  0.0
Fraction of output set to zero during testing time:  0.0
----------------------------------------------------------------
Dropout Keep Prob =  0.25
Mean of input:  5.011213817030235
Mean of output during training time:  5.069143649755907
Mean of output during testing time:  5.011213817030235
Fraction of output set to zero during training time:  0.7465
Fraction of output set to zero during testing time:  0.0
----------------------------------------------------------------
Dropout Keep Prob =  0.5
Mean of input:  5.011213817030235
Mean of output during training time:  4.9803442697715345
Mean of output during testing time:  5.011213817030235
Fraction of output set to zero during training time:  0.5036
Fraction of output set to zero during testing time:  0.0
----------------------------------------------------------------
Dropout Keep Prob =  0.75
Mean of input:  5.011213817030235
Mean of output during training time:  4.991042544054093
Mean of output during testing time:  5.011213817030235
Fraction of output set to zero during training time:  0.2528
Fraction of output set to zero during testing time:  0.0
----------------------------------------------------------------
Dropout Keep Prob =  1
Mean of input:  5.011213817030235
Mean of output during training time:  5.011213817030235
Mean of output during testing time:  5.011213817030235
Fraction of output set to zero during training time:  0.0
Fraction of output set to zero during testing time:  0.0
```

1.8 Dropout Backward [2pt]

Please complete the `backward` pass. Again remember that the dropout is only applied during training phase, handle this in the backward pass as well.

[8]:
```
%reload_ext autoreload

x = np.random.randn(5, 5) + 5
dout = np.random.randn(*x.shape)

keep_prob = 0.75
dropout_b = dropout(keep_prob, seed=100)
out = dropout_b.forward(x, True, seed=1)
dx = dropout_b.backward(dout)
dx_num = eval_numerical_gradient_array(lambda xx: dropout_b.forward(xx, True,
 ↪seed=1), x, dout)

# The error should not be larger than 1e-10
print ('dx relative error: ', rel_error(dx, dx_num))
```

```
dx relative error:  3.003116148710785e-11
```

1.9 Testing cascaded layers: FC + GeLU [2pt]

Please find the `TestFCGeLU` function in `lib/mlp/fully_conn.py`. You only need to complete a few lines of code in the TODO block. Please design an `Flatten -> FC -> GeLU` network where the parameters of them match the given x, w, and b. Please insert the corresponding names you defined for each layer to param_name_w, and param_name_b respectively. Here you only modify the param_name part, the `_w`, and `_b` are automatically assigned during network setup

[9]:
```
%reload_ext autoreload

x = np.random.randn(3, 5, 3)  # the input features
w = np.random.randn(15, 5)   # the weight of fc layer
b = np.random.randn(5)       # the bias of fc layer
dout = np.random.randn(3, 5) # the gradients to the output, notice the shape

tiny_net = TestFCGeLU()

###################################################
# TODO: param_name should be replaced accordingly #
###################################################
tiny_net.net.assign("fc_w", w)
tiny_net.net.assign("fc_b", b)
###################################################
#                 END OF YOUR CODE                #
```

```
##################################################

out = tiny_net.forward(x)
dx = tiny_net.backward(dout)

##################################################
# TODO: param_name should be replaced accordingly #
##################################################
dw = tiny_net.net.get_grads("fc_w")
db = tiny_net.net.get_grads("fc_b")
##################################################
#                END OF YOUR CODE                #
##################################################

dx_num = eval_numerical_gradient_array(lambda x: tiny_net.forward(x), x, dout)
dw_num = eval_numerical_gradient_array(lambda w: tiny_net.forward(x), w, dout)
db_num = eval_numerical_gradient_array(lambda b: tiny_net.forward(x), b, dout)

# The errors should not be larger than 1e-7
print ("dx error: ", rel_error(dx_num, dx))
print ("dw error: ", rel_error(dw_num, dw))
print ("db error: ", rel_error(db_num, db))
```

```
dx error:  7.543362505633382e-07
dw error:  1.950545363407876e-06
db error:  1.1551142197069544e-06
```

1.10 SoftMax Function and Loss Layer [2pt]

In the `lib/mlp/layer_utils.py`, please first complete the function `softmax`, which will be used in the function `cross_entropy`. Then, implement `corss_entropy` using `softmax`. Please refer to the lecture slides of the mathematical expressions of the cross entropy loss function, and complete its forward pass and backward pass. You should also take care of `size_average` on whether or not to divide by the batch size.

[10]:
```
%reload_ext autoreload

num_classes, num_inputs = 6, 100
x = 0.001 * np.random.randn(num_inputs, num_classes)
y = np.random.randint(num_classes, size=num_inputs)

test_loss = cross_entropy()

dx_num = eval_numerical_gradient(lambda x: test_loss.forward(x, y), x,␣
 ↪verbose=False)

loss = test_loss.forward(x, y)
dx = test_loss.backward()
```

```python
# Test softmax_loss function. Loss should be around 1.792
# and dx error should be at the scale of 1e-8 (or smaller)
print ("Cross Entropy Loss: ", loss)
print ("dx error: ", rel_error(dx_num, dx))
```

```
Cross Entropy Loss:  1.7916466149075227
dx error:  8.601545042846586e-09
```

1.11 Test a Small Fully Connected Network [2pt]

Please find the `SmallFullyConnectedNetwork` function in `lib/mlp/fully_conn.py`. Again you only need to complete few lines of code in the TODO block. Please design an `FC --> GeLU --> FC` network where the shapes of parameters match the given shapes. Please insert the corresponding names you defined for each layer to param_name_w, and param_name_b respectively. Here you only modify the param_name part, the `_w`, and `_b` are automatically assigned during network setup.

[11]:
```python
%reload_ext autoreload

seed = 1234
np.random.seed(seed=seed)

model = SmallFullyConnectedNetwork()
loss_func = cross_entropy()

N, D, = 4, 4  # N: batch size, D: input dimension
H, C  = 30, 7 # H: hidden dimension, C: output dimension
std = 0.02
x = np.random.randn(N, D)
y = np.random.randint(C, size=N)

print ("Testing initialization ... ")

###################################################
# TODO: param_name should be replaced accordingly  #
###################################################
w1_std = abs(model.net.get_params("fc1_w").std() - std)
b1 = model.net.get_params("fc1_b").std()
w2_std = abs(model.net.get_params("fc2_w").std() - std)
b2 = model.net.get_params("fc2_b").std()
###################################################
#                 END OF YOUR CODE                 #
###################################################

assert w1_std < std / 10, "First layer weights do not seem right"
assert np.all(b1 == 0), "First layer biases do not seem right"
assert w2_std < std / 10, "Second layer weights do not seem right"
```

```python
assert np.all(b2 == 0), "Second layer biases do not seem right"
print ("Passed!")

print ("Testing test-time forward pass ... ")
w1 = np.linspace(-0.7, 0.3, num=D*H).reshape(D, H)
w2 = np.linspace(-0.2, 0.2, num=H*C).reshape(H, C)
b1 = np.linspace(-0.6, 0.2, num=H)
b2 = np.linspace(-0.9, 0.1, num=C)

###################################################
# TODO: param_name should be replaced accordingly  #
###################################################
model.net.assign("fc1_w", w1)
model.net.assign("fc1_b", b1)
model.net.assign("fc2_w", w2)
model.net.assign("fc2_b", b2)
###################################################
#                 END OF YOUR CODE                 #
###################################################

feats = np.linspace(-5.5, 4.5, num=N*D).reshape(D, N).T
scores = model.forward(feats)
correct_scores = np.asarray([[-2.33881897, -1.92174121, -1.50466344, -1.
 ↪08758567, -0.6705079, -0.25343013,  0.16364763],
                             [-1.57214916, -1.1857013 , -0.79925345, -0.
 ↪41280559, -0.02635774, 0.36009011,  0.74653797],
                             [-0.80178618, -0.44604469, -0.0903032 ,  0.
 ↪26543829,  0.62117977, 0.97692126,  1.33266275],
                             [-0.00331319,  0.32124836,  0.64580991,  0.
 ↪97037146,  1.29493301, 1.61949456,  1.94405611]])
scores_diff = np.sum(np.abs(scores - correct_scores))
assert scores_diff < 1e-6, "Your implementation might be wrong!"
print ("Passed!")

print ("Testing the loss ...",)
y = np.asarray([0, 5, 1, 4])
loss = loss_func.forward(scores, y)
dLoss = loss_func.backward()
correct_loss = 2.4248995879903195
assert abs(loss - correct_loss) < 1e-10, "Your implementation might be wrong!"
print ("Passed!")

print ("Testing the gradients (error should be no larger than 1e-6) ...")
din = model.backward(dLoss)
for layer in model.net.layers:
    if not layer.params:
        continue
```

```
    for name in sorted(layer.grads):
        f = lambda _: loss_func.forward(model.forward(feats), y)
        grad_num = eval_numerical_gradient(f, layer.params[name], verbose=False)
        print ('%s relative error: %.2e' % (name, rel_error(grad_num, layer.
↪grads[name])))
```

```
Testing initialization …
Passed!
Testing test-time forward pass …
Passed!
Testing the loss …
Passed!
Testing the gradients (error should be no larger than 1e-6) …
fc1_b relative error: 1.31e-08
fc1_w relative error: 2.81e-08
fc2_b relative error: 4.01e-10
fc2_w relative error: 2.50e-08
```

1.12 Test a Fully Connected Network regularized with Dropout [2pt]

Please find the `DropoutNet` function in `fully_conn.py` under `lib/mlp` directory. For this part you don't need to design a new network, just simply run the following test code. If something goes wrong, you might want to double check your dropout implementation.

[12]:
```
%reload_ext autoreload

seed = 1234
np.random.seed(seed=seed)

N, D, C = 3, 15, 10
X = np.random.randn(N, D)
y = np.random.randint(C, size=(N,))

for keep_prob in [0, 0.25, 0.5]:
    np.random.seed(seed=seed)
    print ("Dropout p =", keep_prob)
    model = DropoutNet(keep_prob=keep_prob, seed=seed)
    loss_func = cross_entropy()
    output = model.forward(X, True, seed=seed)
    loss = loss_func.forward(output, y)
    dLoss = loss_func.backward()
    dX = model.backward(dLoss)
    grads = model.net.grads

    print ("Error of gradients should be around or less than 1e-3")
    for name in sorted(grads):
        if name not in model.net.params.keys():
            continue
```

```
        f = lambda _: loss_func.forward(model.forward(X, True, seed=seed), y)
        grad_num = eval_numerical_gradient(f, model.net.params[name],␣
↪verbose=False, h=1e-5)
        print ("{} relative error: {}".format(name, rel_error(grad_num,␣
↪grads[name])))
    print ()
```

```
Dropout p = 0
Error of gradients should be around or less than 1e-3
fc1_b relative error: 2.851654987740154e-07
fc1_w relative error: 3.7626907492775348e-06
fc2_b relative error: 1.3390330536574157e-08
fc2_w relative error: 3.08748753596947e-05
fc3_b relative error: 2.5814305918756386e-10
fc3_w relative error: 2.7022952286094135e-06

Dropout p = 0.25
Error of gradients should be around or less than 1e-3
fc1_b relative error: 3.22303229981011e-07
fc1_w relative error: 2.7844020031010643e-06
fc2_b relative error: 1.490984961643268e-07
fc2_w relative error: 4.5315183533700345e-05
fc3_b relative error: 6.679255248099083e-11
fc3_w relative error: 7.93702122628948e-07

Dropout p = 0.5
Error of gradients should be around or less than 1e-3
fc1_b relative error: 9.415776936845159e-07
fc1_w relative error: 1.0482378119758737e-06
fc2_b relative error: 1.549901840006352e-08
fc2_w relative error: 7.918616789113957e-06
fc3_b relative error: 2.2391181687448885e-10
fc3_w relative error: 1.103440520082865e-05
```

1.13 Training a Network

In this section, we defined a `TinyNet` class for you to fill in the TODO block in `lib/mlp/fully_conn.py`. * Here please design a two layer fully connected network with Leaky ReLU activation (`Flatten --> FC --> GeLU --> FC`). * You can adjust the number of hidden neurons, batch_size, epochs, and learning rate decay parameters. * Please read the `lib/train.py` carefully and complete the TODO blocks in the `train_net` function first. Codes in "Test a Small Fully Connected Network" can be helpful. * Implement SGD in `lib/optim.py`, you will be asked to complete weight decay and Adam in the later sections.

[13]:
```
# Arrange the data
data_dict = {
```

```
    "data_train": (data["data_train"], data["labels_train"]),
    "data_val": (data["data_val"], data["labels_val"]),
    "data_test": (data["data_test"], data["labels_test"])
}
```

[14]:
```
print("Data shape:", data["data_train"].shape)
print("Flattened data input size:", np.prod(data["data_train"].shape[1:]))
print("Number of data classes:", max(data['labels_train']) + 1)
```

```
Data shape: (40000, 32, 32, 3)
Flattened data input size: 3072
Number of data classes: 20
```

1.13.1 Now train the network to achieve at least 30% validation accuracy [5pt]

You may only adjust the hyperparameters inside the TODO block

[15]:
```
%autoreload
```

[62]:
```
%reload_ext autoreload

seed = 123
np.random.seed(seed=seed)

model = TinyNet()
loss_f = cross_entropy()
optimizer = SGD(model.net, 0.1)

results = None
#############################################################################
# TODO: Use the train_net function you completed to train a network         #
#############################################################################

batch_size = 500
epochs = 25
lr_decay = 0.99
lr_decay_every = 10

#############################################################################
#                              END OF YOUR CODE                             #
#############################################################################
results = train_net(data_dict, model, loss_f, optimizer, batch_size, epochs,
                    lr_decay, lr_decay_every, show_every=10000, verbose=True)
opt_params, loss_hist, train_acc_hist, val_acc_hist = results
```

```
  4%|
| 3/80 [00:00<00:07, 10.88it/s]

(Iteration 1 / 2000) Average loss: 3.833604996172152
```

```
100%|
       | 80/80 [00:06<00:00, 13.19it/s]

(Epoch 1 / 25) Training Accuracy: 0.274075, Validation Accuracy: 0.2514

100%|
       | 80/80 [00:05<00:00, 15.90it/s]

(Epoch 2 / 25) Training Accuracy: 0.311875, Validation Accuracy: 0.2807

100%|
       | 80/80 [00:05<00:00, 15.27it/s]

(Epoch 3 / 25) Training Accuracy: 0.328825, Validation Accuracy: 0.2883

100%|
       | 80/80 [00:05<00:00, 13.84it/s]

(Epoch 4 / 25) Training Accuracy: 0.3545, Validation Accuracy: 0.3063

100%|
       | 80/80 [00:05<00:00, 14.83it/s]

(Epoch 5 / 25) Training Accuracy: 0.365575, Validation Accuracy: 0.3093

100%|
       | 80/80 [00:05<00:00, 15.19it/s]

(Epoch 6 / 25) Training Accuracy: 0.3847, Validation Accuracy: 0.3124

100%|
       | 80/80 [00:05<00:00, 14.74it/s]

(Epoch 7 / 25) Training Accuracy: 0.382075, Validation Accuracy: 0.31

100%|
       | 80/80 [00:05<00:00, 14.91it/s]

(Epoch 8 / 25) Training Accuracy: 0.40835, Validation Accuracy: 0.3239

100%|
       | 80/80 [00:05<00:00, 15.05it/s]

(Epoch 9 / 25) Training Accuracy: 0.406325, Validation Accuracy: 0.324

100%|
       | 80/80 [00:05<00:00, 15.41it/s]

(Epoch 10 / 25) Training Accuracy: 0.422475, Validation Accuracy: 0.3251
Decaying learning rate of the optimizer to 0.099

100%|
       | 80/80 [00:05<00:00, 14.27it/s]

(Epoch 11 / 25) Training Accuracy: 0.431225, Validation Accuracy: 0.33

100%|
       | 80/80 [00:09<00:00,  8.78it/s]
```

```
(Epoch 12 / 25) Training Accuracy: 0.436125, Validation Accuracy: 0.3306
100%|
      | 80/80 [00:05<00:00, 15.23it/s]
(Epoch 13 / 25) Training Accuracy: 0.423875, Validation Accuracy: 0.3143
100%|
      | 80/80 [00:04<00:00, 16.27it/s]
(Epoch 14 / 25) Training Accuracy: 0.45195, Validation Accuracy: 0.3244
100%|
      | 80/80 [00:04<00:00, 17.26it/s]
(Epoch 15 / 25) Training Accuracy: 0.4599, Validation Accuracy: 0.3309
100%|
      | 80/80 [00:04<00:00, 17.28it/s]
(Epoch 16 / 25) Training Accuracy: 0.472375, Validation Accuracy: 0.3329
100%|
      | 80/80 [00:04<00:00, 17.30it/s]
(Epoch 17 / 25) Training Accuracy: 0.4714, Validation Accuracy: 0.3243
100%|
      | 80/80 [00:04<00:00, 17.25it/s]
(Epoch 18 / 25) Training Accuracy: 0.468175, Validation Accuracy: 0.317
100%|
      | 80/80 [00:04<00:00, 16.45it/s]
(Epoch 19 / 25) Training Accuracy: 0.498, Validation Accuracy: 0.3354
100%|
      | 80/80 [00:05<00:00, 15.43it/s]
(Epoch 20 / 25) Training Accuracy: 0.505625, Validation Accuracy: 0.3263
Decaying learning rate of the optimizer to 0.09801
100%|
      | 80/80 [00:05<00:00, 14.68it/s]
(Epoch 21 / 25) Training Accuracy: 0.506825, Validation Accuracy: 0.3291
100%|
      | 80/80 [00:05<00:00, 14.12it/s]
(Epoch 22 / 25) Training Accuracy: 0.507525, Validation Accuracy: 0.3282
100%|
      | 80/80 [00:05<00:00, 15.14it/s]
(Epoch 23 / 25) Training Accuracy: 0.520525, Validation Accuracy: 0.3322
```

```
100%|
       | 80/80 [00:05<00:00, 15.61it/s]

(Epoch 24 / 25) Training Accuracy: 0.5386, Validation Accuracy: 0.3347

100%|
       | 80/80 [00:05<00:00, 15.62it/s]

(Epoch 25 / 25) Training Accuracy: 0.535025, Validation Accuracy: 0.3308
```

[63]:

```python
# Take a look at what names of params were stored
print (opt_params.keys())
```

```
dict_keys(['fc1_w', 'fc1_b', 'fc2_w', 'fc2_b'])
```

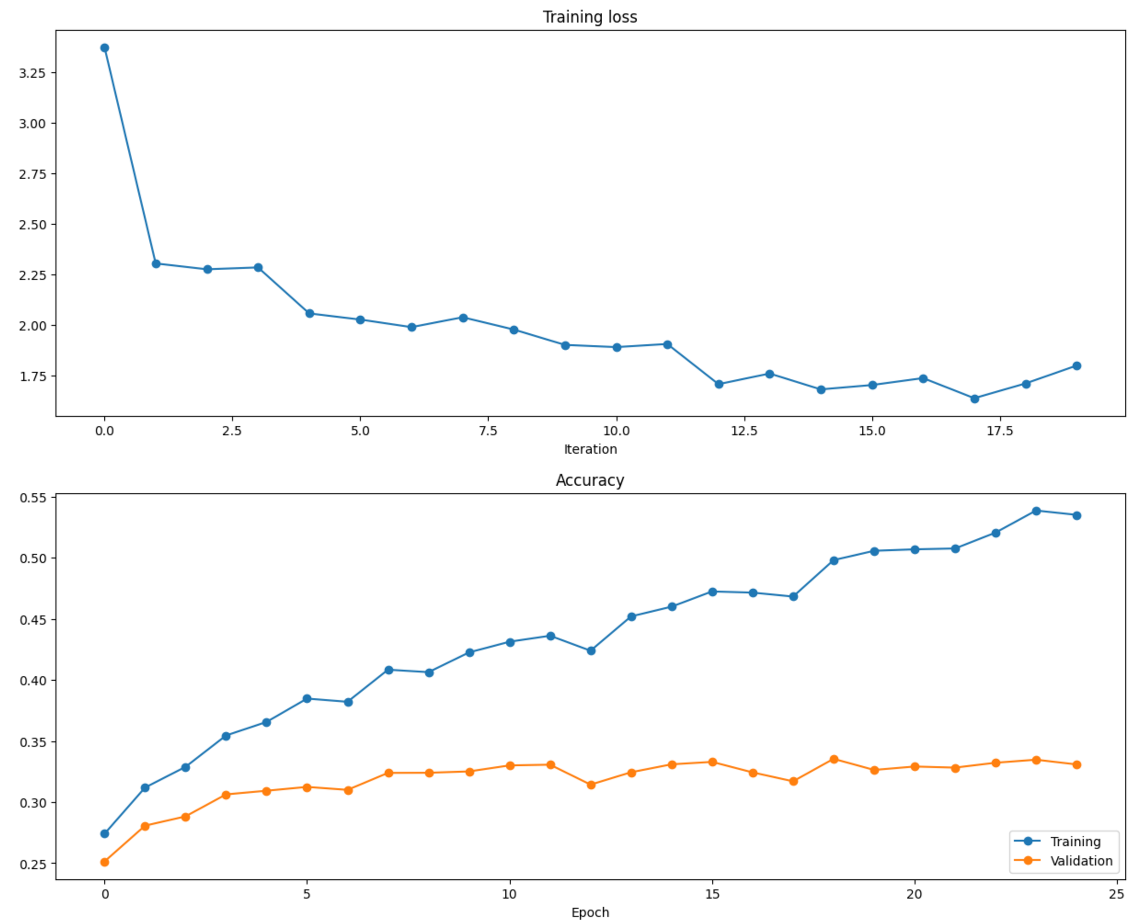
[64]:

```python
# Demo: How to load the parameters to a newly defined network
model = TinyNet()
model.net.load(opt_params)
val_acc = compute_acc(model, data["data_val"], data["labels_val"])
print ("Validation Accuracy: {}%".format(val_acc*100))
test_acc = compute_acc(model, data["data_test"], data["labels_test"])
print ("Testing Accuracy: {}%".format(test_acc*100))
```

```
Loading Params: fc1_w Shape: (3072, 200)
Loading Params: fc1_b Shape: (200,)
Loading Params: fc2_w Shape: (200, 20)
Loading Params: fc2_b Shape: (20,)
Validation Accuracy: 33.08%
Testing Accuracy: 32.67%
```

[65]:

```python
# Plot the learning curves
plt.subplot(2, 1, 1)
plt.title('Training loss')
loss_hist_ = loss_hist[1::100] # sparse the curve a bit
plt.plot(loss_hist_, '-o')
plt.xlabel('Iteration')

plt.subplot(2, 1, 2)
plt.title('Accuracy')
plt.plot(train_acc_hist, '-o', label='Training')
plt.plot(val_acc_hist, '-o', label='Validation')
plt.xlabel('Epoch')
plt.legend(loc='lower right')
plt.gcf().set_size_inches(15, 12)
plt.show()
```



1.14 Different Optimizers and Regularization Techniques

There are several more advanced optimizers than vanilla SGD, and there are many regularization tricks. You'll implement them in this section. Please complete the TODOs in the `lib/optim.py`.

1.15 SGD + Weight Decay [2pt]

The update rule of SGD plus weigh decay is as shown below:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta J(\theta_t) - \lambda \theta_t$$

Update the `SGD()` function in `lib/optim.py`, and also incorporate weight decay options.

[20]:
```
%reload_ext autoreload

# Test the implementation of SGD with Momentum
seed = 1234
np.random.seed(seed=seed)
```

```
N, D = 4, 5
test_sgd = sequential(fc(N, D, name="sgd_fc"))

w = np.linspace(-0.4, 0.6, num=N*D).reshape(N, D)
dw = np.linspace(-0.6, 0.4, num=N*D).reshape(N, D)

test_sgd.layers[0].params = {"sgd_fc_w": w}
test_sgd.layers[0].grads = {"sgd_fc_w": dw}


test_sgd_wd = SGD(test_sgd, 1e-3, 1e-4)
test_sgd_wd.step()

updated_w = test_sgd.layers[0].params["sgd_fc_w"]


expected_updated_w = np.asarray([
       [-0.39936   , -0.34678632, -0.29421263, -0.24163895, -0.18906526],
       [-0.13649158, -0.08391789, -0.03134421,  0.02122947,  0.07380316],
       [ 0.12637684,  0.17895053,  0.23152421,  0.28409789,  0.33667158],
       [ 0.38924526,  0.44181895,  0.49439263,  0.54696632,  0.59954   ]])


print ('The following errors should be around or less than 1e-6')
print ('updated_w error: ', rel_error(updated_w, expected_updated_w))
```

```
The following errors should be around or less than 1e-6
updated_w error:  8.677112905190533e-08
```

1.16 Comparing SGD and SGD with Weight Decay [2pt]

Run the following code block to train a multi-layer fully connected network with both SGD and SGD plus Weight Decay. You are expected to see Weight Decay have better validation accuracy than vinilla SGD.

[21]:
```
seed = 1234

# Arrange a small data
num_train = 20000
small_data_dict = {
    "data_train": (data["data_train"][:num_train], data["labels_train"][:
 ↪num_train]),
    "data_val": (data["data_val"], data["labels_val"]),
    "data_test": (data["data_test"], data["labels_test"])
}

reset_seed(seed=seed)
model_sgd      = FullyConnectedNetwork()
```

```
loss_f_sgd     = cross_entropy()
optimizer_sgd  = SGD(model_sgd.net, 0.01)
print ("Training with Vanilla SGD...")
results_sgd = train_net(small_data_dict, model_sgd, loss_f_sgd, optimizer_sgd,␣
↪batch_size=100,
                        max_epochs=50, show_every=10000, verbose=True)

reset_seed(seed=seed)
model_sgdw     = FullyConnectedNetwork()
loss_f_sgdw    = cross_entropy()
optimizer_sgdw = SGD(model_sgdw.net, 0.01, 1e-4)
print ("\nTraining with SGD plus Weight Decay...")
results_sgdw = train_net(small_data_dict, model_sgdw, loss_f_sgdw,␣
↪optimizer_sgdw, batch_size=100,
                         max_epochs=50, show_every=10000, verbose=True)

opt_params_sgd,  loss_hist_sgd,  train_acc_hist_sgd,  val_acc_hist_sgd  =␣
↪results_sgd
opt_params_sgdw, loss_hist_sgdw, train_acc_hist_sgdw, val_acc_hist_sgdw =␣
↪results_sgdw

plt.subplot(3, 1, 1)
plt.title('Training loss')
plt.xlabel('Iteration')

plt.subplot(3, 1, 2)
plt.title('Training accuracy')
plt.xlabel('Epoch')

plt.subplot(3, 1, 3)
plt.title('Validation accuracy')
plt.xlabel('Epoch')

plt.subplot(3, 1, 1)
plt.plot(loss_hist_sgd, 'o', label="Vanilla SGD")
plt.subplot(3, 1, 2)
plt.plot(train_acc_hist_sgd, '-o', label="Vanilla SGD")
plt.subplot(3, 1, 3)
plt.plot(val_acc_hist_sgd, '-o', label="Vanilla SGD")

plt.subplot(3, 1, 1)
plt.plot(loss_hist_sgdw, 'o', label="SGD with Weight Decay")
plt.subplot(3, 1, 2)
plt.plot(train_acc_hist_sgdw, '-o', label="SGD with Weight Decay")
plt.subplot(3, 1, 3)
plt.plot(val_acc_hist_sgdw, '-o', label="SGD with Weight Decay")
```

```
for i in [1, 2, 3]:
  plt.subplot(3, 1, i)
  plt.legend(loc='upper center', ncol=4)
plt.gcf().set_size_inches(15, 15)
plt.show()
```

```
Training with Vanilla SGD…

  2%|
| 4/200 [00:00<00:05, 33.15it/s]

(Iteration 1 / 10000) Average loss: 3.3332154539088985

100%|
       | 200/200 [00:07<00:00, 26.74it/s]

(Epoch 1 / 50) Training Accuracy: 0.15095, Validation Accuracy: 0.1474

100%|
       | 200/200 [00:07<00:00, 27.58it/s]

(Epoch 2 / 50) Training Accuracy: 0.18815, Validation Accuracy: 0.1805

100%|
       | 200/200 [00:07<00:00, 28.13it/s]

(Epoch 3 / 50) Training Accuracy: 0.2107, Validation Accuracy: 0.2029

100%|
       | 200/200 [00:07<00:00, 27.65it/s]

(Epoch 4 / 50) Training Accuracy: 0.2314, Validation Accuracy: 0.212

100%|
       | 200/200 [00:08<00:00, 24.66it/s]

(Epoch 5 / 50) Training Accuracy: 0.23915, Validation Accuracy: 0.2197

100%|
       | 200/200 [00:07<00:00, 28.15it/s]

(Epoch 6 / 50) Training Accuracy: 0.2552, Validation Accuracy: 0.2298

100%|
       | 200/200 [00:07<00:00, 28.41it/s]

(Epoch 7 / 50) Training Accuracy: 0.26645, Validation Accuracy: 0.2403

100%|
       | 200/200 [00:06<00:00, 30.28it/s]

(Epoch 8 / 50) Training Accuracy: 0.27555, Validation Accuracy: 0.2414

100%|
       | 200/200 [00:07<00:00, 27.93it/s]

(Epoch 9 / 50) Training Accuracy: 0.28185, Validation Accuracy: 0.2413
```

```
100%|
      | 200/200 [00:07<00:00, 26.58it/s]
(Epoch 10 / 50) Training Accuracy: 0.2944, Validation Accuracy: 0.252
100%|
      | 200/200 [00:06<00:00, 30.08it/s]
(Epoch 11 / 50) Training Accuracy: 0.29735, Validation Accuracy: 0.2543
100%|
      | 200/200 [00:05<00:00, 33.66it/s]
(Epoch 12 / 50) Training Accuracy: 0.3021, Validation Accuracy: 0.2587
100%|
      | 200/200 [00:06<00:00, 31.92it/s]
(Epoch 13 / 50) Training Accuracy: 0.31105, Validation Accuracy: 0.2641
100%|
      | 200/200 [00:06<00:00, 31.63it/s]
(Epoch 14 / 50) Training Accuracy: 0.3168, Validation Accuracy: 0.2653
100%|
      | 200/200 [00:06<00:00, 29.36it/s]
(Epoch 15 / 50) Training Accuracy: 0.3217, Validation Accuracy: 0.2681
100%|
      | 200/200 [00:06<00:00, 30.48it/s]
(Epoch 16 / 50) Training Accuracy: 0.3307, Validation Accuracy: 0.2699
100%|
      | 200/200 [00:07<00:00, 28.39it/s]
(Epoch 17 / 50) Training Accuracy: 0.33835, Validation Accuracy: 0.2696
100%|
      | 200/200 [00:10<00:00, 18.98it/s]
(Epoch 18 / 50) Training Accuracy: 0.34565, Validation Accuracy: 0.2737
100%|
      | 200/200 [00:07<00:00, 26.40it/s]
(Epoch 19 / 50) Training Accuracy: 0.3495, Validation Accuracy: 0.2729
100%|
      | 200/200 [00:08<00:00, 22.47it/s]
(Epoch 20 / 50) Training Accuracy: 0.35565, Validation Accuracy: 0.2758
100%|
      | 200/200 [00:07<00:00, 27.00it/s]
(Epoch 21 / 50) Training Accuracy: 0.35825, Validation Accuracy: 0.2729
```

```
100%|
      | 200/200 [00:08<00:00, 24.67it/s]

(Epoch 22 / 50) Training Accuracy: 0.36895, Validation Accuracy: 0.278

100%|
      | 200/200 [00:07<00:00, 26.70it/s]

(Epoch 23 / 50) Training Accuracy: 0.3734, Validation Accuracy: 0.2783

100%|
      | 200/200 [00:07<00:00, 25.65it/s]

(Epoch 24 / 50) Training Accuracy: 0.3756, Validation Accuracy: 0.2768

100%|
      | 200/200 [00:07<00:00, 25.86it/s]

(Epoch 25 / 50) Training Accuracy: 0.38495, Validation Accuracy: 0.278

100%|
      | 200/200 [00:07<00:00, 25.59it/s]

(Epoch 26 / 50) Training Accuracy: 0.38415, Validation Accuracy: 0.2757

100%|
      | 200/200 [00:07<00:00, 27.29it/s]

(Epoch 27 / 50) Training Accuracy: 0.40365, Validation Accuracy: 0.2804

100%|
      | 200/200 [00:06<00:00, 29.53it/s]

(Epoch 28 / 50) Training Accuracy: 0.40105, Validation Accuracy: 0.2812

100%|
      | 200/200 [00:07<00:00, 26.50it/s]

(Epoch 29 / 50) Training Accuracy: 0.40885, Validation Accuracy: 0.2773

100%|
      | 200/200 [00:07<00:00, 25.79it/s]

(Epoch 30 / 50) Training Accuracy: 0.4163, Validation Accuracy: 0.2803

100%|
      | 200/200 [00:07<00:00, 27.61it/s]

(Epoch 31 / 50) Training Accuracy: 0.41745, Validation Accuracy: 0.2838

100%|
      | 200/200 [00:07<00:00, 28.25it/s]

(Epoch 32 / 50) Training Accuracy: 0.42125, Validation Accuracy: 0.2758

100%|
      | 200/200 [00:07<00:00, 27.60it/s]

(Epoch 33 / 50) Training Accuracy: 0.433, Validation Accuracy: 0.2777
```

```
100%|
      | 200/200 [00:07<00:00, 25.36it/s]
(Epoch 34 / 50) Training Accuracy: 0.4322, Validation Accuracy: 0.2782
100%|
      | 200/200 [00:07<00:00, 26.89it/s]
(Epoch 35 / 50) Training Accuracy: 0.44095, Validation Accuracy: 0.2753
100%|
      | 200/200 [00:08<00:00, 24.39it/s]
(Epoch 36 / 50) Training Accuracy: 0.4517, Validation Accuracy: 0.2783
100%|
      | 200/200 [00:06<00:00, 29.99it/s]
(Epoch 37 / 50) Training Accuracy: 0.4583, Validation Accuracy: 0.2759
100%|
      | 200/200 [00:07<00:00, 27.04it/s]
(Epoch 38 / 50) Training Accuracy: 0.4637, Validation Accuracy: 0.2815
100%|
      | 200/200 [00:07<00:00, 26.65it/s]
(Epoch 39 / 50) Training Accuracy: 0.4642, Validation Accuracy: 0.2808
100%|
      | 200/200 [00:07<00:00, 25.33it/s]
(Epoch 40 / 50) Training Accuracy: 0.47055, Validation Accuracy: 0.2784
100%|
      | 200/200 [00:06<00:00, 28.81it/s]
(Epoch 41 / 50) Training Accuracy: 0.4684, Validation Accuracy: 0.2747
100%|
      | 200/200 [00:07<00:00, 27.61it/s]
(Epoch 42 / 50) Training Accuracy: 0.4795, Validation Accuracy: 0.2758
100%|
      | 200/200 [00:08<00:00, 24.08it/s]
(Epoch 43 / 50) Training Accuracy: 0.48745, Validation Accuracy: 0.2793
100%|
      | 200/200 [00:08<00:00, 24.56it/s]
(Epoch 44 / 50) Training Accuracy: 0.49715, Validation Accuracy: 0.2751
100%|
      | 200/200 [00:07<00:00, 26.77it/s]
(Epoch 45 / 50) Training Accuracy: 0.49545, Validation Accuracy: 0.2736
```

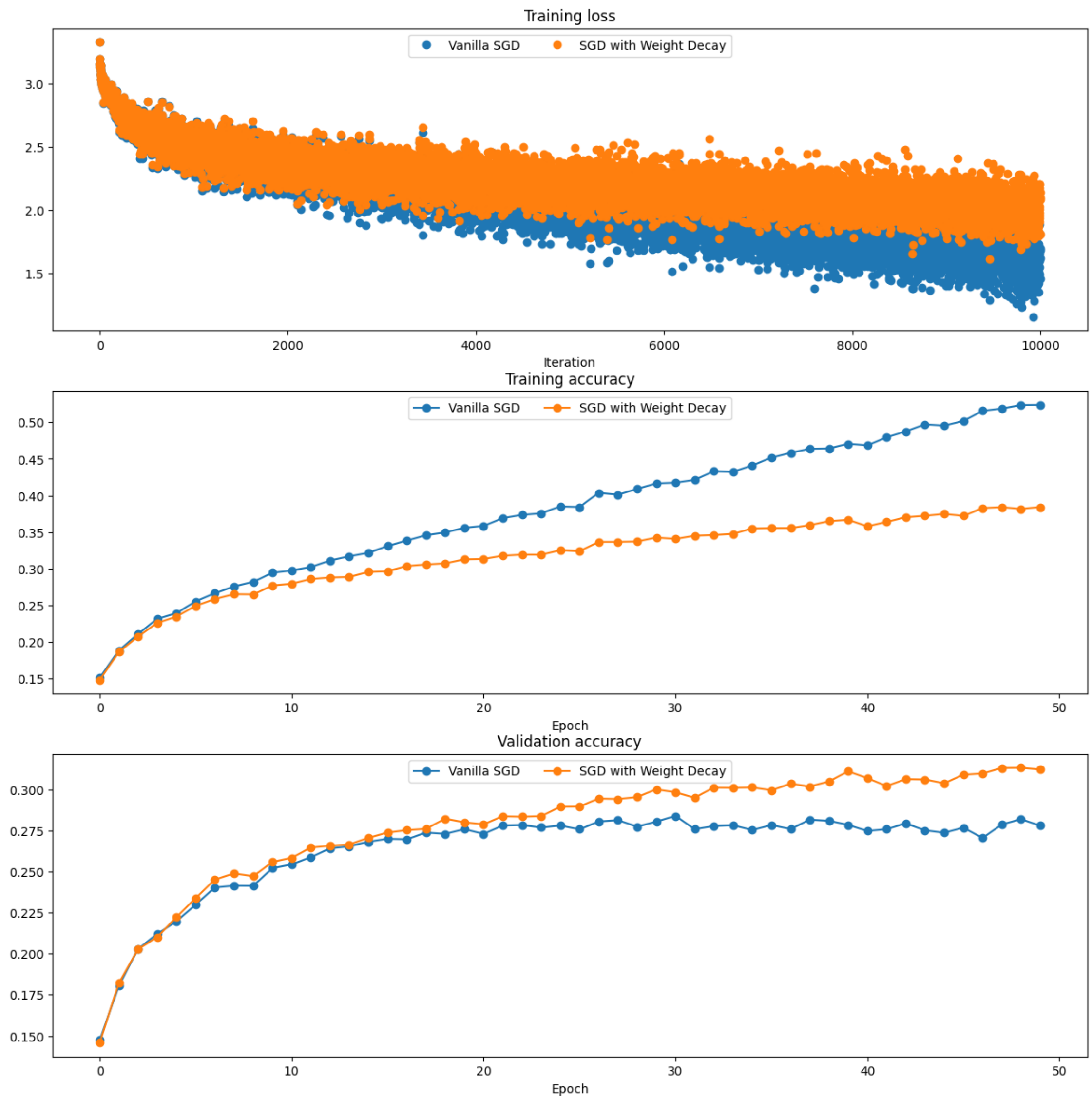
```
100%|
      | 200/200 [00:08<00:00, 24.78it/s]

(Epoch 46 / 50) Training Accuracy: 0.50175, Validation Accuracy: 0.2767

100%|
      | 200/200 [00:07<00:00, 25.71it/s]

(Epoch 47 / 50) Training Accuracy: 0.51565, Validation Accuracy: 0.2704

100%|
      | 200/200 [00:08<00:00, 23.97it/s]

(Epoch 48 / 50) Training Accuracy: 0.51875, Validation Accuracy: 0.2786

100%|
      | 200/200 [00:09<00:00, 21.18it/s]

(Epoch 49 / 50) Training Accuracy: 0.5235, Validation Accuracy: 0.2818

100%|
      | 200/200 [00:08<00:00, 24.90it/s]

(Epoch 50 / 50) Training Accuracy: 0.52375, Validation Accuracy: 0.2779

Training with SGD plus Weight Decay…
  1%|
| 2/200 [00:00<00:13, 14.97it/s]

(Iteration 1 / 10000) Average loss: 3.3332154539088985

100%|
      | 200/200 [00:08<00:00, 23.88it/s]

(Epoch 1 / 50) Training Accuracy: 0.148, Validation Accuracy: 0.1458

100%|
      | 200/200 [00:09<00:00, 20.98it/s]

(Epoch 2 / 50) Training Accuracy: 0.186, Validation Accuracy: 0.1822

100%|
      | 200/200 [00:07<00:00, 27.74it/s]

(Epoch 3 / 50) Training Accuracy: 0.2073, Validation Accuracy: 0.2027

100%|
      | 200/200 [00:08<00:00, 24.97it/s]

(Epoch 4 / 50) Training Accuracy: 0.22575, Validation Accuracy: 0.2101

100%|
      | 200/200 [00:14<00:00, 14.11it/s]

(Epoch 5 / 50) Training Accuracy: 0.2345, Validation Accuracy: 0.2223

100%|
      | 200/200 [00:07<00:00, 25.41it/s]
```

```
(Epoch 6 / 50) Training Accuracy: 0.24915, Validation Accuracy: 0.2338
100%|
      | 200/200 [00:07<00:00, 26.56it/s]
(Epoch 7 / 50) Training Accuracy: 0.2584, Validation Accuracy: 0.2451
100%|
      | 200/200 [00:09<00:00, 21.61it/s]
(Epoch 8 / 50) Training Accuracy: 0.2651, Validation Accuracy: 0.2488
100%|
      | 200/200 [00:07<00:00, 25.98it/s]
(Epoch 9 / 50) Training Accuracy: 0.2648, Validation Accuracy: 0.2471
100%|
      | 200/200 [00:10<00:00, 18.75it/s]
(Epoch 10 / 50) Training Accuracy: 0.27685, Validation Accuracy: 0.2558
100%|
      | 200/200 [00:07<00:00, 27.01it/s]
(Epoch 11 / 50) Training Accuracy: 0.2792, Validation Accuracy: 0.2583
100%|
      | 200/200 [00:07<00:00, 27.05it/s]
(Epoch 12 / 50) Training Accuracy: 0.28575, Validation Accuracy: 0.2646
100%|
      | 200/200 [00:07<00:00, 25.80it/s]
(Epoch 13 / 50) Training Accuracy: 0.2879, Validation Accuracy: 0.2657
100%|
      | 200/200 [00:06<00:00, 31.54it/s]
(Epoch 14 / 50) Training Accuracy: 0.28865, Validation Accuracy: 0.2664
100%|
      | 200/200 [00:05<00:00, 33.80it/s]
(Epoch 15 / 50) Training Accuracy: 0.29545, Validation Accuracy: 0.2705
100%|
      | 200/200 [00:06<00:00, 29.24it/s]
(Epoch 16 / 50) Training Accuracy: 0.2964, Validation Accuracy: 0.2737
100%|
      | 200/200 [00:05<00:00, 34.02it/s]
(Epoch 17 / 50) Training Accuracy: 0.30345, Validation Accuracy: 0.2752
100%|
      | 200/200 [00:05<00:00, 37.26it/s]
```

```
(Epoch 18 / 50) Training Accuracy: 0.30555, Validation Accuracy: 0.276
100%|
      | 200/200 [00:05<00:00, 39.14it/s]
(Epoch 19 / 50) Training Accuracy: 0.30715, Validation Accuracy: 0.2821
100%|
      | 200/200 [00:05<00:00, 38.39it/s]
(Epoch 20 / 50) Training Accuracy: 0.31265, Validation Accuracy: 0.2799
100%|
      | 200/200 [00:05<00:00, 38.64it/s]
(Epoch 21 / 50) Training Accuracy: 0.31315, Validation Accuracy: 0.2787
100%|
      | 200/200 [00:05<00:00, 35.18it/s]
(Epoch 22 / 50) Training Accuracy: 0.31755, Validation Accuracy: 0.2836
100%|
      | 200/200 [00:05<00:00, 35.17it/s]
(Epoch 23 / 50) Training Accuracy: 0.3192, Validation Accuracy: 0.2833
100%|
      | 200/200 [00:05<00:00, 33.84it/s]
(Epoch 24 / 50) Training Accuracy: 0.31905, Validation Accuracy: 0.2837
100%|
      | 200/200 [00:06<00:00, 32.29it/s]
(Epoch 25 / 50) Training Accuracy: 0.32525, Validation Accuracy: 0.2894
100%|
      | 200/200 [00:05<00:00, 33.93it/s]
(Epoch 26 / 50) Training Accuracy: 0.3238, Validation Accuracy: 0.2895
100%|
      | 200/200 [00:05<00:00, 35.65it/s]
(Epoch 27 / 50) Training Accuracy: 0.33645, Validation Accuracy: 0.2944
100%|
      | 200/200 [00:05<00:00, 33.59it/s]
(Epoch 28 / 50) Training Accuracy: 0.33645, Validation Accuracy: 0.2941
100%|
      | 200/200 [00:06<00:00, 31.06it/s]
(Epoch 29 / 50) Training Accuracy: 0.33695, Validation Accuracy: 0.2953
100%|
      | 200/200 [00:05<00:00, 34.08it/s]
```

```
(Epoch 30 / 50) Training Accuracy: 0.3425, Validation Accuracy: 0.3
100%|
      | 200/200 [00:06<00:00, 29.37it/s]
(Epoch 31 / 50) Training Accuracy: 0.3406, Validation Accuracy: 0.2982
100%|
      | 200/200 [00:06<00:00, 32.87it/s]
(Epoch 32 / 50) Training Accuracy: 0.34505, Validation Accuracy: 0.2949
100%|
      | 200/200 [00:05<00:00, 33.62it/s]
(Epoch 33 / 50) Training Accuracy: 0.34595, Validation Accuracy: 0.3011
100%|
      | 200/200 [00:05<00:00, 33.67it/s]
(Epoch 34 / 50) Training Accuracy: 0.34755, Validation Accuracy: 0.301
100%|
      | 200/200 [00:06<00:00, 31.28it/s]
(Epoch 35 / 50) Training Accuracy: 0.3548, Validation Accuracy: 0.3012
100%|
      | 200/200 [00:06<00:00, 31.33it/s]
(Epoch 36 / 50) Training Accuracy: 0.3552, Validation Accuracy: 0.2995
100%|
      | 200/200 [00:05<00:00, 33.93it/s]
(Epoch 37 / 50) Training Accuracy: 0.35525, Validation Accuracy: 0.3034
100%|
      | 200/200 [00:05<00:00, 38.26it/s]
(Epoch 38 / 50) Training Accuracy: 0.3593, Validation Accuracy: 0.3017
100%|
      | 200/200 [00:06<00:00, 32.58it/s]
(Epoch 39 / 50) Training Accuracy: 0.3648, Validation Accuracy: 0.3048
100%|
      | 200/200 [00:05<00:00, 34.98it/s]
(Epoch 40 / 50) Training Accuracy: 0.36665, Validation Accuracy: 0.311
100%|
      | 200/200 [00:05<00:00, 34.01it/s]
(Epoch 41 / 50) Training Accuracy: 0.35765, Validation Accuracy: 0.3068
100%|
      | 200/200 [00:05<00:00, 35.09it/s]
```

```
(Epoch 42 / 50) Training Accuracy: 0.36375, Validation Accuracy: 0.302
100%|
      | 200/200 [00:06<00:00, 31.15it/s]
(Epoch 43 / 50) Training Accuracy: 0.3702, Validation Accuracy: 0.3062
100%|
      | 200/200 [00:06<00:00, 32.96it/s]
(Epoch 44 / 50) Training Accuracy: 0.37215, Validation Accuracy: 0.306
100%|
      | 200/200 [00:06<00:00, 32.69it/s]
(Epoch 45 / 50) Training Accuracy: 0.37475, Validation Accuracy: 0.3037
100%|
      | 200/200 [00:06<00:00, 29.93it/s]
(Epoch 46 / 50) Training Accuracy: 0.37205, Validation Accuracy: 0.3089
100%|
      | 200/200 [00:06<00:00, 32.51it/s]
(Epoch 47 / 50) Training Accuracy: 0.3827, Validation Accuracy: 0.3097
100%|
      | 200/200 [00:06<00:00, 33.24it/s]
(Epoch 48 / 50) Training Accuracy: 0.38395, Validation Accuracy: 0.313
100%|
      | 200/200 [00:05<00:00, 33.36it/s]
(Epoch 49 / 50) Training Accuracy: 0.38155, Validation Accuracy: 0.3131
100%|
      | 200/200 [00:05<00:00, 35.27it/s]
(Epoch 50 / 50) Training Accuracy: 0.38415, Validation Accuracy: 0.3121
```



1.17 SGD with L1 Regularization [2pts]

With L1 Regularization, your regularized loss becomes $\tilde{J}_{_1}(\theta)$ and it's defined as

$$\tilde{J}_{_1}(\theta) = J(\theta) + \lambda\|\theta\|_{\ell_1}$$

where

$$\|\theta\|_{\ell_1} = \sum_{l=1}^{n}\sum_{k=1}^{n_l}|\theta_{l,k}|$$

Please implmemt TODO block of `apply_l1_regularization` in `lib/layer_utils`. Such regularization funcationality is called after gradient gathering in the `backward` process.

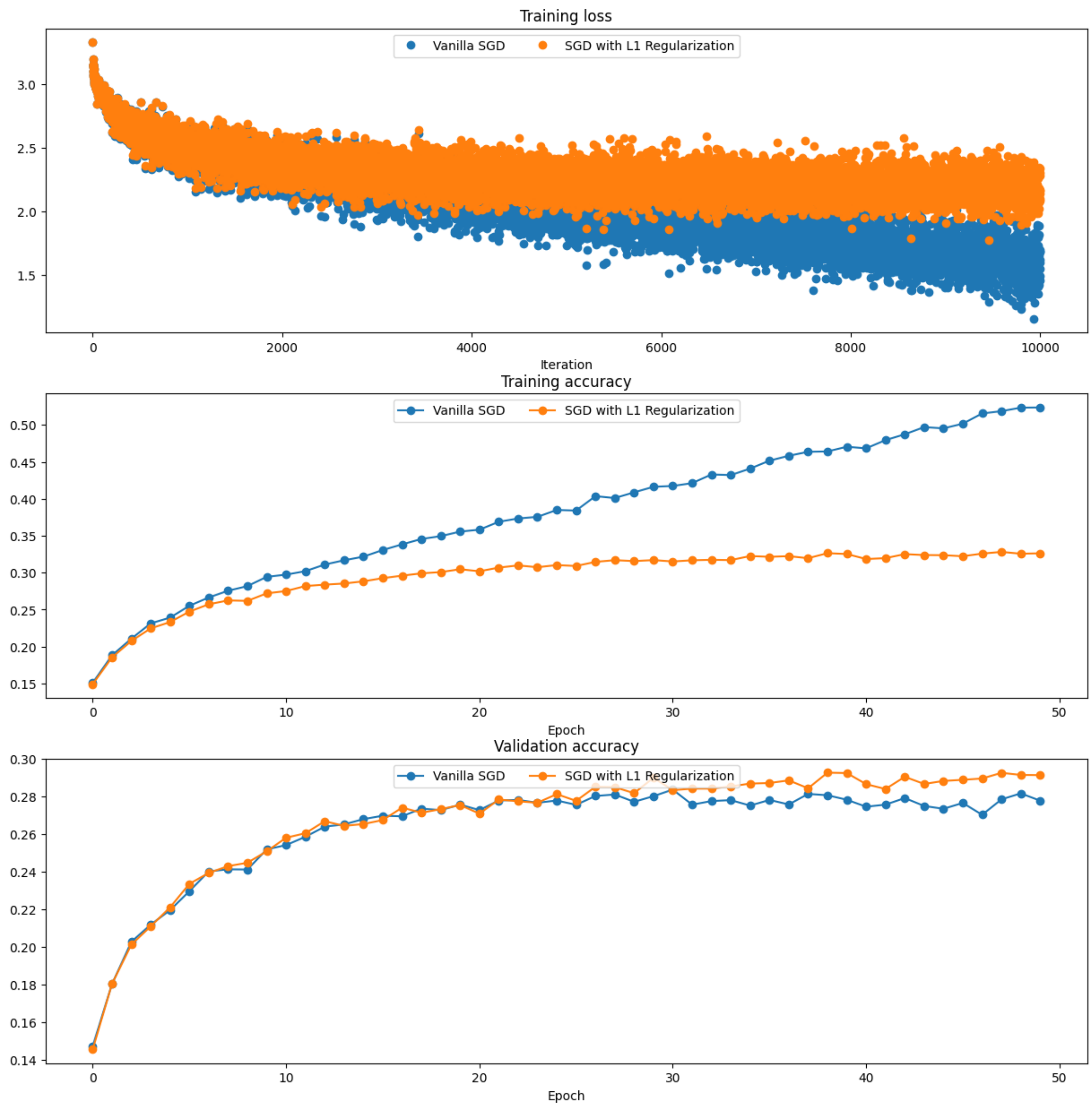
[55]:

```python
reset_seed(seed=seed)
model_sgd_l1     = FullyConnectedNetwork()
loss_f_sgd_l1    = cross_entropy()
optimizer_sgd_l1 = SGD(model_sgd_l1.net, 0.01)

print ("\nTraining with SGD plus L1 Regularization...")
results_sgd_l1 = train_net(small_data_dict, model_sgd_l1, loss_f_sgd_l1,␣
↪optimizer_sgd_l1, batch_size=100,
                         max_epochs=50, show_every=10000, verbose=True,␣
↪regularization="l1", reg_lambda=1e-3)

opt_params_sgd_l1, loss_hist_sgd_l1, train_acc_hist_sgd_l1,␣
↪val_acc_hist_sgd_l1= results_sgd_l1

plt.subplot(3, 1, 1)
plt.title('Training loss')
plt.xlabel('Iteration')

plt.subplot(3, 1, 2)
plt.title('Training accuracy')
plt.xlabel('Epoch')

plt.subplot(3, 1, 3)
plt.title('Validation accuracy')
plt.xlabel('Epoch')

plt.subplot(3, 1, 1)
plt.plot(loss_hist_sgd, 'o', label="Vanilla SGD")
plt.subplot(3, 1, 2)
plt.plot(train_acc_hist_sgd, '-o', label="Vanilla SGD")
plt.subplot(3, 1, 3)
plt.plot(val_acc_hist_sgd, '-o', label="Vanilla SGD")

plt.subplot(3, 1, 1)
plt.plot(loss_hist_sgd_l1, 'o', label="SGD with L1 Regularization")
plt.subplot(3, 1, 2)
plt.plot(train_acc_hist_sgd_l1, '-o', label="SGD with L1 Regularization")
plt.subplot(3, 1, 3)
plt.plot(val_acc_hist_sgd_l1, '-o', label="SGD with L1 Regularization")

for i in [1, 2, 3]:
  plt.subplot(3, 1, i)
  plt.legend(loc='upper center', ncol=4)
plt.gcf().set_size_inches(15, 15)
plt.show()
```

```
Training with SGD plus L1 Regularization…

100%|
        | 200/200 [00:05<00:00, 33.94it/s]
100%|
        | 200/200 [00:06<00:00, 31.82it/s]
100%|
        | 200/200 [00:06<00:00, 33.15it/s]
100%|
        | 200/200 [00:06<00:00, 31.41it/s]
100%|
        | 200/200 [00:05<00:00, 34.08it/s]
100%|
        | 200/200 [00:06<00:00, 33.17it/s]
100%|
        | 200/200 [00:05<00:00, 33.92it/s]
100%|
        | 200/200 [00:05<00:00, 33.97it/s]
100%|
        | 200/200 [00:06<00:00, 32.44it/s]
100%|
        | 200/200 [00:06<00:00, 32.94it/s]
100%|
        | 200/200 [00:05<00:00, 34.48it/s]
100%|
        | 200/200 [00:06<00:00, 31.84it/s]
100%|
        | 200/200 [00:05<00:00, 33.87it/s]
100%|
        | 200/200 [00:05<00:00, 33.75it/s]
100%|
        | 200/200 [00:05<00:00, 34.41it/s]
100%|
        | 200/200 [00:05<00:00, 34.49it/s]
100%|
        | 200/200 [00:05<00:00, 34.50it/s]
100%|
        | 200/200 [00:05<00:00, 34.28it/s]
100%|
        | 200/200 [00:05<00:00, 33.81it/s]
100%|
        | 200/200 [00:05<00:00, 34.48it/s]
100%|
        | 200/200 [00:05<00:00, 34.45it/s]
100%|
        | 200/200 [00:07<00:00, 25.34it/s]
100%|
```

```
      | 200/200 [00:08<00:00, 24.93it/s]
100%|
      | 200/200 [00:08<00:00, 23.37it/s]
100%|
      | 200/200 [00:11<00:00, 17.44it/s]
100%|
      | 200/200 [00:05<00:00, 34.57it/s]
100%|
      | 200/200 [00:05<00:00, 35.17it/s]
100%|
      | 200/200 [00:06<00:00, 28.82it/s]
100%|
      | 200/200 [00:06<00:00, 29.31it/s]
100%|
      | 200/200 [00:06<00:00, 29.17it/s]
100%|
      | 200/200 [00:06<00:00, 30.58it/s]
100%|
      | 200/200 [00:06<00:00, 30.28it/s]
100%|
      | 200/200 [00:06<00:00, 29.33it/s]
100%|
      | 200/200 [00:05<00:00, 33.84it/s]
100%|
      | 200/200 [00:08<00:00, 23.99it/s]
100%|
      | 200/200 [00:09<00:00, 20.73it/s]
100%|
      | 200/200 [00:08<00:00, 24.69it/s]
100%|
      | 200/200 [00:08<00:00, 23.30it/s]
100%|
      | 200/200 [00:07<00:00, 25.90it/s]
100%|
      | 200/200 [00:07<00:00, 26.82it/s]
100%|
      | 200/200 [00:06<00:00, 29.68it/s]
100%|
      | 200/200 [00:07<00:00, 28.44it/s]
100%|
      | 200/200 [00:07<00:00, 27.00it/s]
100%|
      | 200/200 [00:07<00:00, 28.33it/s]
100%|
      | 200/200 [00:07<00:00, 26.85it/s]
100%|
      | 200/200 [00:07<00:00, 27.79it/s]
100%|
```

```
      | 200/200 [00:06<00:00, 29.16it/s]
100%|
      | 200/200 [00:06<00:00, 30.23it/s]
100%|
      | 200/200 [00:07<00:00, 27.99it/s]
100%|
      | 200/200 [00:06<00:00, 30.86it/s]
```



1.18 SGD with L2 Regularization [2pts]

With L2 Regularization, your regularized loss becomes $\tilde{J}_2(\theta)$ and it's defined as

$$\tilde{J}_2(\theta) = J(\theta) + \lambda\|\theta\|_{\ell_2}$$

where

$$\|\theta\|_{\ell_2} = \sum_{l=1}^{n} \sum_{k=1}^{n_l} \theta_{l,k}^2$$

Similarly, implmemt TODO block of `apply_l2_regularization` in `lib/layer_utils`. For SGD, you're also asked to find the λ for L2 Regularization such that it achives the EXACTLY SAME effect as weight decay in the previous cells. As a reminder, learning rate is the same as previously, and the weight decay paramter was 1e-4.

[61]:
```
reset_seed(seed=seed)
model_sgd_l2     = FullyConnectedNetwork()
loss_f_sgd_l2    = cross_entropy()
optimizer_sgd_l2 = SGD(model_sgd_l2.net, 0.01)
#####################################################################
#### Find lambda for L2 regularization so that                  ####
#### it achieves EXACTLY THE SAME learning curve as weight decay ####
l2_lambda = 0.005
#####################################################################

print ("\nTraining with SGD plus L2 Regularization...")
results_sgd_l2 = train_net(small_data_dict, model_sgd_l2, loss_f_sgd_l2,␣
 ↪optimizer_sgd_l2, batch_size=100,
                           max_epochs=50, show_every=10000, verbose=True,␣
 ↪regularization="l2", reg_lambda=l2_lambda)

opt_params_sgd_l2, loss_hist_sgd_l2, train_acc_hist_sgd_l2, val_acc_hist_sgd_l2␣
 ↪= results_sgd_l2

plt.subplot(3, 1, 1)
plt.title('Training loss')
plt.xlabel('Iteration')

plt.subplot(3, 1, 2)
plt.title('Training accuracy')
plt.xlabel('Epoch')

plt.subplot(3, 1, 3)
plt.title('Validation accuracy')
plt.xlabel('Epoch')

plt.subplot(3, 1, 1)
plt.plot(loss_hist_sgdw, 'o', label="SGD with Weight Decay")
plt.subplot(3, 1, 2)
plt.plot(train_acc_hist_sgdw, '-o', label="SGD with Weight Decay")
plt.subplot(3, 1, 3)
plt.plot(val_acc_hist_sgdw, '-o', label="SGD with Weight Decay")

plt.subplot(3, 1, 1)
```

```
plt.plot(loss_hist_sgd_l1, 'o', label="SGD with L1 Regularization")
plt.subplot(3, 1, 2)
plt.plot(train_acc_hist_sgd_l1, '-o', label="SGD with L1 Regularization")
plt.subplot(3, 1, 3)
plt.plot(val_acc_hist_sgd_l1, '-o', label="SGD with L1 Regularization")

plt.subplot(3, 1, 1)
plt.plot(loss_hist_sgd_l2, 'o', label="SGD with L2 Regularization")
plt.subplot(3, 1, 2)
plt.plot(train_acc_hist_sgd_l2, '-o', label="SGD with L2 Regularization")
plt.subplot(3, 1, 3)
plt.plot(val_acc_hist_sgd_l2, '-o', label="SGD with L2 Regularization")

for i in [1, 2, 3]:
  plt.subplot(3, 1, i)
  plt.legend(loc='upper center', ncol=4)
plt.gcf().set_size_inches(15, 15)
plt.show()
```

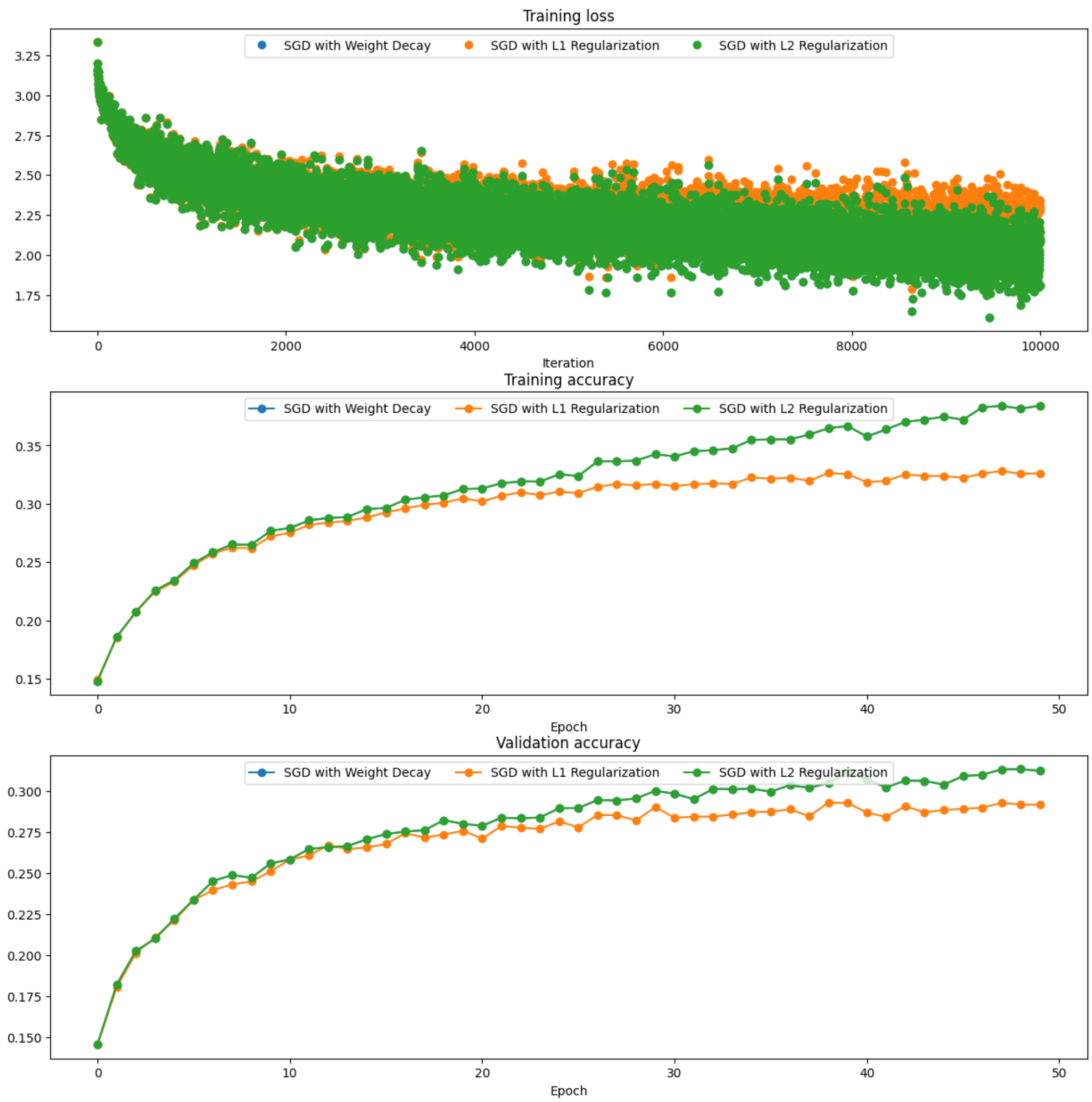
```
Training with SGD plus L2 Regularization…

  2%|
| 4/200 [00:00<00:06, 32.52it/s]

(Iteration 1 / 10000) Average loss: 3.3332154539088985

100%|
      | 200/200 [00:05<00:00, 34.36it/s]

(Epoch 1 / 50) Training Accuracy: 0.148, Validation Accuracy: 0.1458

100%|
      | 200/200 [00:05<00:00, 34.49it/s]

(Epoch 2 / 50) Training Accuracy: 0.186, Validation Accuracy: 0.1822

100%|
      | 200/200 [00:06<00:00, 33.07it/s]

(Epoch 3 / 50) Training Accuracy: 0.2073, Validation Accuracy: 0.2027

100%|
      | 200/200 [00:06<00:00, 31.41it/s]

(Epoch 4 / 50) Training Accuracy: 0.22575, Validation Accuracy: 0.2101

100%|
      | 200/200 [00:06<00:00, 32.70it/s]

(Epoch 5 / 50) Training Accuracy: 0.2345, Validation Accuracy: 0.2223

100%|
      | 200/200 [00:05<00:00, 34.67it/s]
```

```
(Epoch 6 / 50) Training Accuracy: 0.24915, Validation Accuracy: 0.2338
100%|
      | 200/200 [00:05<00:00, 38.62it/s]
(Epoch 7 / 50) Training Accuracy: 0.2584, Validation Accuracy: 0.2451
100%|
      | 200/200 [00:05<00:00, 38.77it/s]
(Epoch 8 / 50) Training Accuracy: 0.2651, Validation Accuracy: 0.2488
100%|
      | 200/200 [00:05<00:00, 37.47it/s]
(Epoch 9 / 50) Training Accuracy: 0.2648, Validation Accuracy: 0.2471
100%|
      | 200/200 [00:06<00:00, 32.82it/s]
(Epoch 10 / 50) Training Accuracy: 0.27685, Validation Accuracy: 0.2558
100%|
      | 200/200 [00:05<00:00, 37.14it/s]
(Epoch 11 / 50) Training Accuracy: 0.2792, Validation Accuracy: 0.2583
100%|
      | 200/200 [00:05<00:00, 36.90it/s]
(Epoch 12 / 50) Training Accuracy: 0.28575, Validation Accuracy: 0.2646
100%|
      | 200/200 [00:05<00:00, 38.13it/s]
(Epoch 13 / 50) Training Accuracy: 0.2879, Validation Accuracy: 0.2657
100%|
      | 200/200 [00:06<00:00, 33.11it/s]
(Epoch 14 / 50) Training Accuracy: 0.28865, Validation Accuracy: 0.2664
100%|
      | 200/200 [00:05<00:00, 34.03it/s]
(Epoch 15 / 50) Training Accuracy: 0.29545, Validation Accuracy: 0.2705
100%|
      | 200/200 [00:06<00:00, 31.77it/s]
(Epoch 16 / 50) Training Accuracy: 0.2964, Validation Accuracy: 0.2737
100%|
      | 200/200 [00:06<00:00, 33.22it/s]
(Epoch 17 / 50) Training Accuracy: 0.30345, Validation Accuracy: 0.2752
100%|
      | 200/200 [00:06<00:00, 32.84it/s]
```

```
(Epoch 18 / 50) Training Accuracy: 0.30555, Validation Accuracy: 0.276
100%|
      | 200/200 [00:06<00:00, 32.42it/s]
(Epoch 19 / 50) Training Accuracy: 0.30715, Validation Accuracy: 0.2821
100%|
      | 200/200 [00:06<00:00, 30.86it/s]
(Epoch 20 / 50) Training Accuracy: 0.31265, Validation Accuracy: 0.2799
100%|
      | 200/200 [00:06<00:00, 32.16it/s]
(Epoch 21 / 50) Training Accuracy: 0.31315, Validation Accuracy: 0.2787
100%|
      | 200/200 [00:06<00:00, 32.12it/s]
(Epoch 22 / 50) Training Accuracy: 0.31755, Validation Accuracy: 0.2836
100%|
      | 200/200 [00:06<00:00, 32.99it/s]
(Epoch 23 / 50) Training Accuracy: 0.3192, Validation Accuracy: 0.2833
100%|
      | 200/200 [00:07<00:00, 25.62it/s]
(Epoch 24 / 50) Training Accuracy: 0.31905, Validation Accuracy: 0.2837
100%|
      | 200/200 [00:06<00:00, 32.24it/s]
(Epoch 25 / 50) Training Accuracy: 0.32525, Validation Accuracy: 0.2894
100%|
      | 200/200 [00:05<00:00, 33.76it/s]
(Epoch 26 / 50) Training Accuracy: 0.3238, Validation Accuracy: 0.2895
100%|
      | 200/200 [00:05<00:00, 33.76it/s]
(Epoch 27 / 50) Training Accuracy: 0.33645, Validation Accuracy: 0.2944
100%|
      | 200/200 [00:05<00:00, 33.84it/s]
(Epoch 28 / 50) Training Accuracy: 0.33645, Validation Accuracy: 0.2941
100%|
      | 200/200 [00:05<00:00, 34.26it/s]
(Epoch 29 / 50) Training Accuracy: 0.33695, Validation Accuracy: 0.2953
100%|
      | 200/200 [00:05<00:00, 34.12it/s]
```

```
(Epoch 30 / 50) Training Accuracy: 0.3425, Validation Accuracy: 0.3
100%|
      | 200/200 [00:06<00:00, 32.85it/s]
(Epoch 31 / 50) Training Accuracy: 0.3406, Validation Accuracy: 0.2982
100%|
      | 200/200 [00:06<00:00, 33.16it/s]
(Epoch 32 / 50) Training Accuracy: 0.34505, Validation Accuracy: 0.2949
100%|
      | 200/200 [00:06<00:00, 33.02it/s]
(Epoch 33 / 50) Training Accuracy: 0.34595, Validation Accuracy: 0.3011
100%|
      | 200/200 [00:05<00:00, 34.06it/s]
(Epoch 34 / 50) Training Accuracy: 0.34755, Validation Accuracy: 0.301
100%|
      | 200/200 [00:06<00:00, 28.88it/s]
(Epoch 35 / 50) Training Accuracy: 0.3548, Validation Accuracy: 0.3012
100%|
      | 200/200 [00:05<00:00, 36.06it/s]
(Epoch 36 / 50) Training Accuracy: 0.3552, Validation Accuracy: 0.2995
100%|
      | 200/200 [00:06<00:00, 33.00it/s]
(Epoch 37 / 50) Training Accuracy: 0.35525, Validation Accuracy: 0.3034
100%|
      | 200/200 [00:06<00:00, 31.42it/s]
(Epoch 38 / 50) Training Accuracy: 0.3593, Validation Accuracy: 0.3017
100%|
      | 200/200 [00:05<00:00, 37.43it/s]
(Epoch 39 / 50) Training Accuracy: 0.3648, Validation Accuracy: 0.3048
100%|
      | 200/200 [00:05<00:00, 37.28it/s]
(Epoch 40 / 50) Training Accuracy: 0.36665, Validation Accuracy: 0.311
100%|
      | 200/200 [00:05<00:00, 34.42it/s]
(Epoch 41 / 50) Training Accuracy: 0.35765, Validation Accuracy: 0.3068
100%|
      | 200/200 [00:05<00:00, 37.55it/s]
```

```
(Epoch 42 / 50) Training Accuracy: 0.36375, Validation Accuracy: 0.302
100%|
      | 200/200 [00:05<00:00, 36.56it/s]
(Epoch 43 / 50) Training Accuracy: 0.3702, Validation Accuracy: 0.3062
100%|
      | 200/200 [00:06<00:00, 31.43it/s]
(Epoch 44 / 50) Training Accuracy: 0.37215, Validation Accuracy: 0.306
100%|
      | 200/200 [00:05<00:00, 36.08it/s]
(Epoch 45 / 50) Training Accuracy: 0.37475, Validation Accuracy: 0.3037
100%|
      | 200/200 [00:05<00:00, 33.59it/s]
(Epoch 46 / 50) Training Accuracy: 0.37205, Validation Accuracy: 0.3089
100%|
      | 200/200 [00:05<00:00, 37.73it/s]
(Epoch 47 / 50) Training Accuracy: 0.3827, Validation Accuracy: 0.3097
100%|
      | 200/200 [00:05<00:00, 39.10it/s]
(Epoch 48 / 50) Training Accuracy: 0.38395, Validation Accuracy: 0.313
100%|
      | 200/200 [00:05<00:00, 36.40it/s]
(Epoch 49 / 50) Training Accuracy: 0.38155, Validation Accuracy: 0.3131
100%|
      | 200/200 [00:05<00:00, 37.06it/s]
(Epoch 50 / 50) Training Accuracy: 0.38415, Validation Accuracy: 0.3121
```


Training loss
SGD with Weight Decay
SGD with L1 Regularization
SGD with L2 Regularization
Iteration
Training accuracy
SGD with Weight Decay
SGD with L1 Regularization
SGD with L2 Regularization
Epoch
Validation accuracy
SGD with Weight Decay
SGD with L1 Regularization
SGD with L2 Regularization
Epoch

1.19 Adam [2pt]

The update rule of Adam is as shown below:

$$
\begin{aligned}
t &= t + 1 \\
g_t &: \text{gradients at update step } t \\
m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t \\
v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \\
\hat{m}_t &= m_t / (1 - \beta_1^t) \\
\hat{v}_t &= v_t / (1 - \beta_2^t) \\
\theta_{t+1} &= \theta_t - \frac{\eta \, \hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}
\end{aligned}
$$

Complete the `Adam()` function in `lib/optim.py` Important Notes: 1) t must be updated before everything else 2) β_1^t is β_1 exponentiated to the t'th power 3) You should also enable weight decay in Adam, similar to what you did in SGD

[57]:
```
%reload_ext autoreload

seed = 1234
np.random.seed(seed=seed)

# Test Adam implementation; you should see errors around 1e-7 or less
N, D = 4, 5
test_adam = sequential(fc(N, D, name="adam_fc"))

w = np.linspace(-0.4, 0.6, num=N*D).reshape(N, D)
dw = np.linspace(-0.6, 0.4, num=N*D).reshape(N, D)
m = np.linspace(0.6, 0.9, num=N*D).reshape(N, D)
v = np.linspace(0.7, 0.5, num=N*D).reshape(N, D)

test_adam.layers[0].params = {"adam_fc_w": w}
test_adam.layers[0].grads = {"adam_fc_w": dw}

opt_adam = Adam(test_adam, 1e-2, 0.9, 0.999, t=5)
opt_adam.mt = {"adam_fc_w": m}
opt_adam.vt = {"adam_fc_w": v}
opt_adam.step()

updated_w = test_adam.layers[0].params["adam_fc_w"]
mt = opt_adam.mt["adam_fc_w"]
vt = opt_adam.vt["adam_fc_w"]

expected_updated_w = np.asarray([
  [-0.40094747, -0.34836187, -0.29577703, -0.24319299, -0.19060977],
```

```
  [-0.1380274,  -0.08544591, -0.03286534,  0.01971428,  0.0722929],
  [ 0.1248705,   0.17744702,  0.23002243,  0.28259667,  0.33516969],
  [ 0.38774145,  0.44031188,  0.49288093,  0.54544852,  0.59801459]])
expected_v = np.asarray([
  [ 0.69966,     0.68908382,  0.67851319,  0.66794809,  0.65738853,],
  [ 0.64683452,  0.63628604,  0.6257431,   0.61520571,  0.60467385,],
  [ 0.59414753,  0.58362676,  0.57311152,  0.56260183,  0.55209767,],
  [ 0.54159906,  0.53110598,  0.52061845,  0.51013645,  0.49966,   ]])
expected_m = np.asarray([
  [ 0.48,        0.49947368,  0.51894737,  0.53842105,  0.55789474],
  [ 0.57736842,  0.59684211,  0.61631579,  0.63578947,  0.65526316],
  [ 0.67473684,  0.69421053,  0.71368421,  0.73315789,  0.75263158],
  [ 0.77210526,  0.79157895,  0.81105263,  0.83052632,  0.85      ]])

print ('The following errors should be around or less than 1e-7')
print ('updated_w error: ', rel_error(expected_updated_w, updated_w))
print ('mt error: ', rel_error(expected_m, mt))
print ('vt error: ', rel_error(expected_v, vt))
```

```
The following errors should be around or less than 1e-7
updated_w error:  1.1395691798535431e-07
mt error:  4.214963193114416e-09
vt error:  4.208314038113071e-09
```

1.20 Comparing the Weight Decay v.s. L2 Regularization in Adam [5pt]

Run the following code block to compare the plotted results between effects of weight decay and L2 regularization on Adam. Are they still the same? (we can make them the same as in SGD, can we also do it in Adam?)

[58]:
```
seed = 1234
reset_seed(seed)
model_adam_wd       = FullyConnectedNetwork()
loss_f_adam_wd      = cross_entropy()
optimizer_adam_wd   = Adam(model_adam_wd.net, lr=1e-4, weight_decay=1e-6)


print ("Training with AdamW...")
results_adam_wd = train_net(small_data_dict, model_adam_wd, loss_f_adam_wd,␣
 ↪optimizer_adam_wd, batch_size=100,
                        max_epochs=50, show_every=10000, verbose=False)

reset_seed(seed)
model_adam_l2       = FullyConnectedNetwork()
loss_f_adam_l2      = cross_entropy()
optimizer_adam_l2 = Adam(model_adam_l2.net, lr=1e-4)
reg_lambda_l2 = 1e-4
print ("\nTraining with Adam + L2...")
```

```
results_adam_l2 = train_net(small_data_dict, model_adam_l2, loss_f_adam_l2,␣
 ↪optimizer_adam_l2, batch_size=100,
                         max_epochs=50, show_every=10000, verbose=False,␣
 ↪regularization='l2', reg_lambda=reg_lambda_l2)

opt_params_adam_wd, loss_hist_adam_wd, train_acc_hist_adam_wd,␣
 ↪val_acc_hist_adam_wd  = results_adam_wd
opt_params_adam_l2, loss_hist_adam_l2, train_acc_hist_adam_l2,␣
 ↪val_acc_hist_adam_l2 = results_adam_l2

plt.subplot(3, 1, 1)
plt.title('Training loss')
plt.xlabel('Iteration')

plt.subplot(3, 1, 2)
plt.title('Training accuracy')
plt.xlabel('Epoch')

plt.subplot(3, 1, 3)
plt.title('Validation accuracy')
plt.xlabel('Epoch')

plt.subplot(3, 1, 1)
plt.plot(loss_hist_sgd, 'o', label="Vanilla SGD")
plt.subplot(3, 1, 2)
plt.plot(train_acc_hist_sgd, '-o', label="Vanilla SGD")
plt.subplot(3, 1, 3)
plt.plot(val_acc_hist_sgd, '-o', label="Vanilla SGD")

plt.subplot(3, 1, 1)
plt.plot(loss_hist_sgdw, 'o', label="SGD with Weight Decay")
plt.subplot(3, 1, 2)
plt.plot(train_acc_hist_sgdw, '-o', label="SGD with Weight Decay")
plt.subplot(3, 1, 3)
plt.plot(val_acc_hist_sgdw, '-o', label="SGD with Weight Decay")

plt.subplot(3, 1, 1)
plt.plot(loss_hist_adam_wd, 'o', label="Adam with Weight Decay")
plt.subplot(3, 1, 2)
plt.plot(train_acc_hist_adam_wd, '-o', label="Adam with Weight Decay")
plt.subplot(3, 1, 3)
plt.plot(val_acc_hist_adam_wd, '-o', label="Adam with Weight Decay")

plt.subplot(3, 1, 1)
plt.plot(loss_hist_adam_l2, 'o', label="Adam with L2")
plt.subplot(3, 1, 2)
plt.plot(train_acc_hist_adam_l2, '-o', label="Adam with L2")
```

```
plt.subplot(3, 1, 3)
plt.plot(val_acc_hist_adam_l2, '-o', label="Adam with L2")

for i in [1, 2, 3]:
  plt.subplot(3, 1, i)
  plt.legend(loc='upper center', ncol=4)
plt.gcf().set_size_inches(15, 15)
plt.show()
```

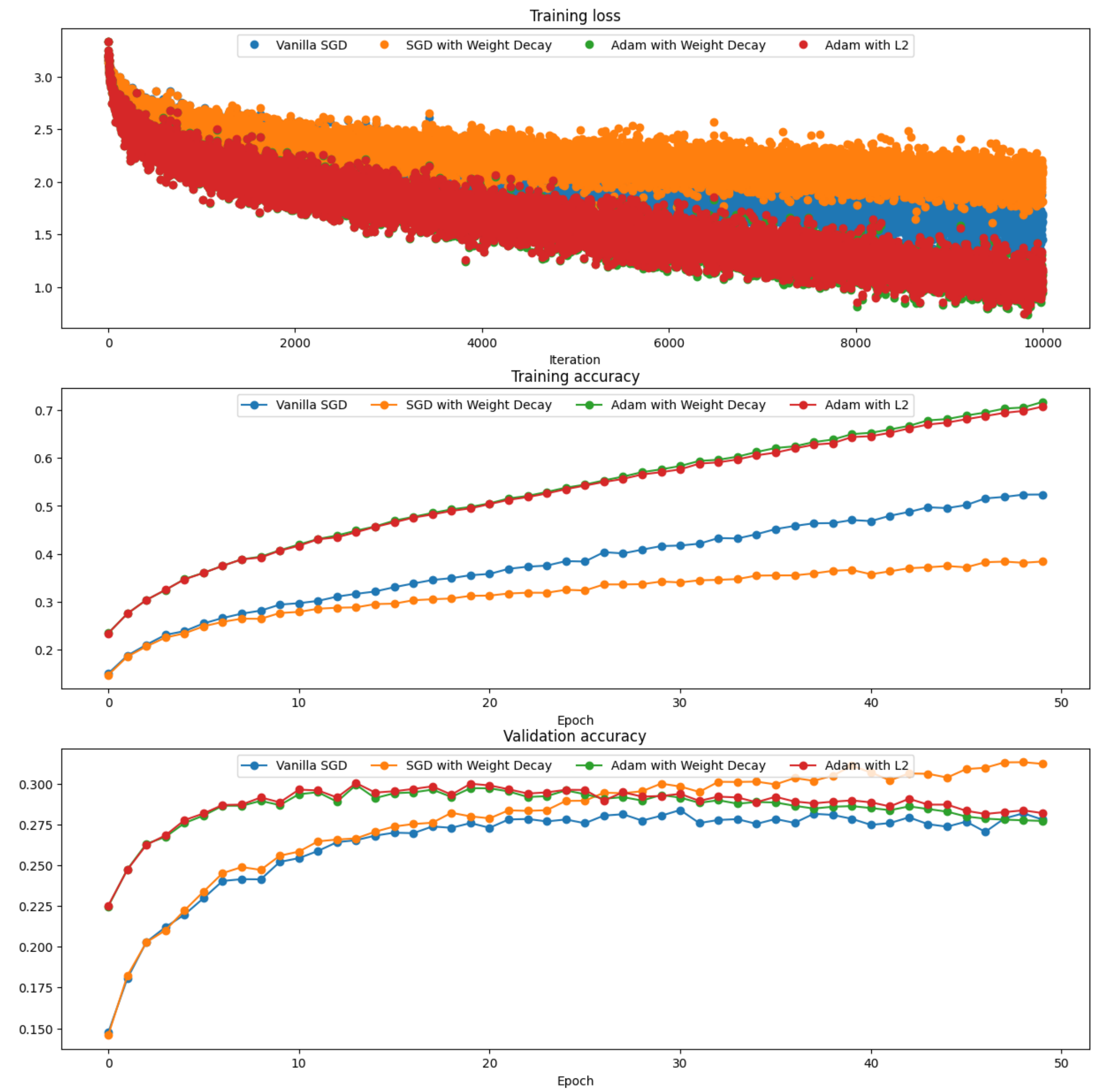
```
Training with AdamW…

100%|
      | 200/200 [00:05<00:00, 33.63it/s]
100%|
      | 200/200 [00:05<00:00, 33.65it/s]
100%|
      | 200/200 [00:06<00:00, 33.20it/s]
100%|
      | 200/200 [00:05<00:00, 34.00it/s]
100%|
      | 200/200 [00:06<00:00, 32.71it/s]
100%|
      | 200/200 [00:05<00:00, 33.68it/s]
100%|
      | 200/200 [00:06<00:00, 32.84it/s]
100%|
      | 200/200 [00:05<00:00, 33.42it/s]
100%|
      | 200/200 [00:05<00:00, 33.53it/s]
100%|
      | 200/200 [00:05<00:00, 33.92it/s]
100%|
      | 200/200 [00:08<00:00, 24.39it/s]
100%|
      | 200/200 [00:06<00:00, 31.62it/s]
100%|
      | 200/200 [00:05<00:00, 33.34it/s]
100%|
      | 200/200 [00:05<00:00, 34.28it/s]
100%|
      | 200/200 [00:05<00:00, 34.97it/s]
100%|
      | 200/200 [00:05<00:00, 34.07it/s]
100%|
      | 200/200 [00:05<00:00, 34.42it/s]
100%|
      | 200/200 [00:05<00:00, 35.92it/s]
100%|
```

```
      | 200/200 [00:05<00:00, 37.09it/s]
100%|
      | 200/200 [00:05<00:00, 33.56it/s]
100%|
      | 200/200 [00:06<00:00, 33.25it/s]
100%|
      | 200/200 [00:06<00:00, 32.75it/s]
100%|
      | 200/200 [00:06<00:00, 31.02it/s]
100%|
      | 200/200 [00:06<00:00, 32.61it/s]
100%|
      | 200/200 [00:06<00:00, 32.30it/s]
100%|
      | 200/200 [00:06<00:00, 31.45it/s]
100%|
      | 200/200 [00:06<00:00, 31.87it/s]
100%|
      | 200/200 [00:08<00:00, 24.17it/s]
100%|
      | 200/200 [00:06<00:00, 32.31it/s]
100%|
      | 200/200 [00:05<00:00, 33.39it/s]
100%|
      | 200/200 [00:06<00:00, 32.74it/s]
100%|
      | 200/200 [00:05<00:00, 35.09it/s]
100%|
      | 200/200 [00:05<00:00, 34.17it/s]
100%|
      | 200/200 [00:05<00:00, 33.34it/s]
100%|
      | 200/200 [00:05<00:00, 36.38it/s]
100%|
      | 200/200 [00:05<00:00, 35.55it/s]
100%|
      | 200/200 [00:05<00:00, 36.03it/s]
100%|
      | 200/200 [00:05<00:00, 33.98it/s]
100%|
      | 200/200 [00:05<00:00, 34.20it/s]
100%|
      | 200/200 [00:06<00:00, 31.95it/s]
100%|
      | 200/200 [00:06<00:00, 31.52it/s]
100%|
      | 200/200 [00:06<00:00, 30.08it/s]
100%|
```

```
      | 200/200 [00:05<00:00, 36.42it/s]
100%|
      | 200/200 [00:05<00:00, 34.76it/s]
100%|
      | 200/200 [00:05<00:00, 35.10it/s]
100%|
      | 200/200 [00:06<00:00, 32.22it/s]
100%|
      | 200/200 [00:06<00:00, 32.98it/s]
100%|
      | 200/200 [00:06<00:00, 32.62it/s]
100%|
      | 200/200 [00:06<00:00, 32.90it/s]
100%|
      | 200/200 [00:06<00:00, 33.26it/s]

Training with Adam + L2…

100%|
      | 200/200 [00:06<00:00, 31.82it/s]
100%|
      | 200/200 [00:06<00:00, 32.11it/s]
100%|
      | 200/200 [00:06<00:00, 32.05it/s]
100%|
      | 200/200 [00:06<00:00, 33.04it/s]
100%|
      | 200/200 [00:05<00:00, 33.34it/s]
100%|
      | 200/200 [00:06<00:00, 32.76it/s]
100%|
      | 200/200 [00:05<00:00, 34.11it/s]
100%|
      | 200/200 [00:05<00:00, 34.06it/s]
100%|
      | 200/200 [00:07<00:00, 27.12it/s]
100%|
      | 200/200 [00:07<00:00, 28.42it/s]
100%|
      | 200/200 [00:06<00:00, 30.46it/s]
100%|
      | 200/200 [00:06<00:00, 31.20it/s]
100%|
      | 200/200 [00:06<00:00, 31.45it/s]
100%|
      | 200/200 [00:06<00:00, 31.82it/s]
100%|
      | 200/200 [00:06<00:00, 30.08it/s]
```

```
100%|
      | 200/200 [00:06<00:00, 32.43it/s]
100%|
      | 200/200 [00:06<00:00, 31.81it/s]
100%|
      | 200/200 [00:06<00:00, 32.17it/s]
100%|
      | 200/200 [00:06<00:00, 31.93it/s]
100%|
      | 200/200 [00:06<00:00, 32.61it/s]
100%|
      | 200/200 [00:06<00:00, 32.36it/s]
100%|
      | 200/200 [00:06<00:00, 31.83it/s]
100%|
      | 200/200 [00:06<00:00, 30.34it/s]
100%|
      | 200/200 [00:06<00:00, 31.83it/s]
100%|
      | 200/200 [00:06<00:00, 32.05it/s]
100%|
      | 200/200 [00:09<00:00, 20.97it/s]
100%|
      | 200/200 [00:12<00:00, 16.52it/s]
100%|
      | 200/200 [00:09<00:00, 20.41it/s]
100%|
      | 200/200 [00:06<00:00, 31.84it/s]
100%|
      | 200/200 [00:06<00:00, 30.98it/s]
100%|
      | 200/200 [00:12<00:00, 16.04it/s]
100%|
      | 200/200 [00:06<00:00, 30.56it/s]
100%|
      | 200/200 [00:06<00:00, 31.89it/s]
100%|
      | 200/200 [00:06<00:00, 32.13it/s]
100%|
      | 200/200 [00:06<00:00, 31.58it/s]
100%|
      | 200/200 [00:08<00:00, 23.50it/s]
100%|
      | 200/200 [00:06<00:00, 30.75it/s]
100%|
      | 200/200 [00:06<00:00, 29.58it/s]
100%|
      | 200/200 [00:06<00:00, 32.65it/s]
```

```
100%|
     | 200/200 [00:06<00:00, 31.45it/s]
100%|
     | 200/200 [00:24<00:00,  8.13it/s]
100%|
     | 200/200 [00:12<00:00, 16.48it/s]
100%|
     | 200/200 [00:06<00:00, 30.93it/s]
100%|
     | 200/200 [00:06<00:00, 30.99it/s]
100%|
     | 200/200 [00:06<00:00, 32.15it/s]
100%|
     | 200/200 [00:06<00:00, 29.93it/s]
100%|
     | 200/200 [00:06<00:00, 32.17it/s]
100%|
     | 200/200 [00:06<00:00, 31.12it/s]
100%|
     | 200/200 [00:06<00:00, 32.09it/s]
100%|
     | 200/200 [00:06<00:00, 31.05it/s]
```



1.20.1 Inline Answer

Weight decay and L2 regularization are not the same in Adam, because Adam calculates the L2 penalty differently (L2 influenced by squareroot) than SGD and it's impossible to convert one to the other by calculating the lambda value vs. learning rate like we did to SGD.

2 Submission

Please prepare a PDF document `problem_1_solution.pdf` in the root directory of this repository with all plots and inline answers of your solution. Concretely, the document should contain the following items in strict order: 1. Training loss / accuracy curves for the simple neural network training with > 30% validation accuracy 2. Plots for comparing vanilla SGD to SGD + Weight Decay, SGD + L1 and SGD + L2 3. "Comparing different Regularizations" plots

Note that you still need to submit the jupyter notebook with all generated solutions. We will randomly pick submissions and check that the plots in the PDF and in the notebook are equivalent.

Problem_2

February 25, 2023

1 Problem 2: Incorporating CNNs

- Learning Objective: In this problem, you will learn how to deeply understand how Convolutional Neural Networks work by implementing one.
- Provided Code: We provide the skeletons of classes you need to complete. Forward checking and gradient checkings are provided for verifying your implementation as well.
- TODOs: you will implement a Convolutional Layer and a MaxPooling Layer to improve on your classification results in part 1.

[57]:
```
from lib.mlp.fully_conn import *
from lib.mlp.layer_utils import *
from lib.mlp.train import *
from lib.cnn.layer_utils import *
from lib.cnn.cnn_models import *
from lib.datasets import *
from lib.grad_check import *
from lib.optim import *
import numpy as np
import matplotlib.pyplot as plt

%matplotlib inline
plt.rcParams['figure.figsize'] = (10.0, 8.0) # set default size of plots
plt.rcParams['image.interpolation'] = 'nearest'
plt.rcParams['image.cmap'] = 'gray'

# for auto-reloading external modules
# see http://stackoverflow.com/questions/1907993/
 ↪autoreload-of-modules-in-ipython
%load_ext autoreload
%autoreload 2
```

```
The autoreload extension is already loaded. To reload it, use:
  %reload_ext autoreload
```

1.1 Loading the data (CIFAR-100 with 20 superclasses)

In this homework, we will be classifying images from the CIFAR-100 dataset into the 20 superclasses. More information about the CIFAR-100 dataset and the 20 superclasses can be found here.

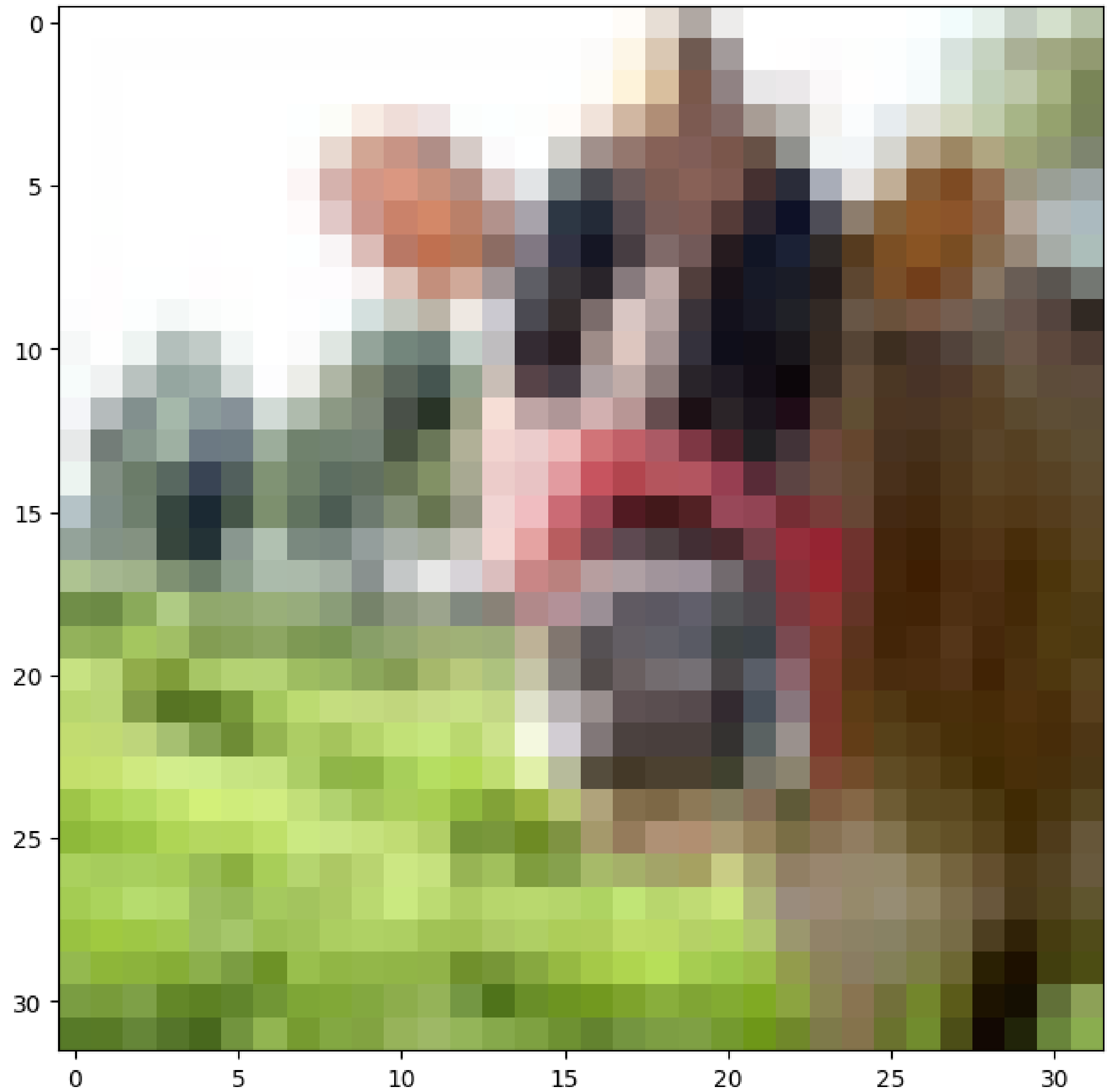
Download the CIFAR-100 data files here, and save the `.mat` files to the `data/cifar100` directory.

[2]:
```
data = CIFAR100_data('data/cifar100/')
for k, v in data.items():
    if type(v) == np.ndarray:
        print ("Name: {} Shape: {}, {}".format(k, v.shape, type(v)))
    else:
        print("{}: {}".format(k, v))
label_names = data['label_names']
mean_image = data['mean_image'][0]
std_image = data['std_image'][0]
```

```
Name: data_train Shape: (40000, 32, 32, 3), <class 'numpy.ndarray'>
Name: labels_train Shape: (40000,), <class 'numpy.ndarray'>
Name: data_val Shape: (10000, 32, 32, 3), <class 'numpy.ndarray'>
Name: labels_val Shape: (10000,), <class 'numpy.ndarray'>
Name: data_test Shape: (10000, 32, 32, 3), <class 'numpy.ndarray'>
Name: labels_test Shape: (10000,), <class 'numpy.ndarray'>
label_names: ['aquatic_mammals', 'fish', 'flowers', 'food_containers',
'fruit_and_vegetables', 'household_electrical_devices', 'household_furniture',
'insects', 'large_carnivores', 'large_man-made_outdoor_things',
'large_natural_outdoor_scenes', 'large_omnivores_and_herbivores',
'medium_mammals', 'non-insect_invertebrates', 'people', 'reptiles',
'small_mammals', 'trees', 'vehicles_1', 'vehicles_2']
Name: mean_image Shape: (1, 1, 1, 3), <class 'numpy.ndarray'>
Name: std_image Shape: (1, 1, 1, 3), <class 'numpy.ndarray'>
```

[4]:
```
idx = 0
image_data = data['data_train'][idx]
image_data = ((image_data*std_image + mean_image) * 255).astype(np.int32)
plt.imshow(image_data)
label = label_names[data['labels_train'][idx]]
print("Label:", label)
```

```
Label: large_omnivores_and_herbivores
```



1.2 Convolutional Neural Networks

We will use convolutional neural networks to try to improve on the results from Problem 1. Convolutional layers make the assumption that local pixels are more important for prediction than far-away pixels. This allows us to form networks that are robust to small changes in positioning in images.

1.2.1 Convolutional Layer Output size calculation [2pts]

As you have learned, two important parameters of a convolutional layer are its stride and padding. To warm up, we will need to calculate the output size of a convolutional layer given its stride and padding. To do this, open the `lib/cnn/layer_utils.py` file and fill out the TODO section in the `get_output_size` function in the ConvLayer2D class.

Implement your function so that it returns the correct size as indicated by the block below.

[25]:
```
%reload_ext autoreload

input_image = np.zeros([32, 28, 28, 3]) # a stack of 32 28 by 28 rgb images

in_channels = input_image.shape[-1] #must agree with the last dimension of the␣
↪input image
k_size = 4
n_filt = 16

conv_layer = ConvLayer2D(in_channels, k_size, n_filt, stride=2, padding=3)
output_size = conv_layer.get_output_size(input_image.shape)

print("Received {} and expected [32, 16, 16, 16]".format(output_size))
```

```
Received [32, 16, 16, 16] and expected [32, 16, 16, 16]
```

1.2.2 Convolutional Layer Forward Pass [5pts]

Now, we will implement the forward pass of a convolutional layer. Fill in the TODO block in the `forward` function of the ConvLayer2D class.

[33]:
```
%reload_ext autoreload

# Test the convolutional forward function
input_image = np.linspace(-0.1, 0.4, num=1*8*8*1).reshape([1, 8, 8, 1]) # a␣
↪single 8 by 8 grayscale image
in_channels, k_size, n_filt = 1, 5, 2

weight_size = k_size*k_size*in_channels*n_filt
bias_size = n_filt



single_conv = ConvLayer2D(in_channels, k_size, n_filt, stride=1, padding=0,␣
↪name="conv_test")

w = np.linspace(-0.2, 0.2, num=weight_size).reshape(k_size, k_size,␣
↪in_channels, n_filt)
b = np.linspace(-0.3, 0.3, num=bias_size)

single_conv.params[single_conv.w_name] = w
single_conv.params[single_conv.b_name] = b

out = single_conv.forward(input_image)
```

```
print("Received output shape: {}, Expected output shape: (1, 4, 4, 2)".
 ↪format(out.shape))

correct_out = np.array([[
   [[-0.03874312, 0.57000324],
   [-0.03955296, 0.57081309],
   [-0.04036281, 0.57162293],
   [-0.04117266, 0.57243278]],

  [[-0.0452219, 0.57648202],
   [-0.04603175, 0.57729187],
   [-0.04684159, 0.57810172],
   [-0.04765144, 0.57891156]],

  [[-0.05170068, 0.5829608 ],
   [-0.05251053, 0.58377065],
   [-0.05332038, 0.5845805 ],
   [-0.05413022, 0.58539035]],

  [[-0.05817946, 0.58943959],
   [-0.05898931, 0.59024943],
   [-0.05979916, 0.59105928],
   [-0.06060901, 0.59186913]]]])

# Compare your output with the above pre-computed ones.
# The difference should not be larger than 1e-7
print ("Difference: ", rel_error(out, correct_out))
```

```
Received output shape: (1, 4, 4, 2), Expected output shape: (1, 4, 4, 2)
Difference:  5.110565335399418e-08
```

1.2.3 Conv Layer Backward [5pts]

Now complete the backward pass of a convolutional layer. Fill in the TODO block in the `backward` function of the ConvLayer2D class. Check you results with this code and expect differences of less than 1e-6.

[34]:
```
%reload_ext autoreload

# Test the conv backward function
img = np.random.randn(15, 8, 8, 3)
w = np.random.randn(4, 4, 3, 12)
b = np.random.randn(12)
dout = np.random.randn(15, 4, 4, 12)

single_conv = ConvLayer2D(input_channels=3, kernel_size=4, number_filters=12,␣
 ↪stride=2, padding=1, name="conv_test")
single_conv.params[single_conv.w_name] = w
```

```
single_conv.params[single_conv.b_name] = b

dimg_num = eval_numerical_gradient_array(lambda x: single_conv.forward(img),
 ↪img, dout)
dw_num = eval_numerical_gradient_array(lambda w: single_conv.forward(img), w,
 ↪dout)
db_num = eval_numerical_gradient_array(lambda b: single_conv.forward(img), b,
 ↪dout)

out = single_conv.forward(img)

dimg = single_conv.backward(dout)
dw = single_conv.grads[single_conv.w_name]
db = single_conv.grads[single_conv.b_name]

# The error should be around 1e-6
print("dimg Error: ", rel_error(dimg_num, dimg))
# The errors should be around 1e-8
print("dw Error: ", rel_error(dw_num, dw))
print("db Error: ", rel_error(db_num, db))
# The shapes should be same
print("dimg Shape: ", dimg.shape, img.shape)
```

```
dimg Error:  1.0635775073105403e-08
dw Error:  2.6069110013156756e-08
db Error:  5.531236483488697e-10
dimg Shape:  (15, 8, 8, 3) (15, 8, 8, 3)
```

1.3 Max pooling Layer

Now we will implement maxpooling layers, which can help to reduce the image size while preserving the overall structure of the image.

1.3.1 Forward Pass max pooling [5pts]

Fill out the TODO block in the `forward` function of the MaxPoolingLayer class.

[38]:
```
# Test the convolutional forward function
input_image = np.linspace(-0.1, 0.4, num=64).reshape([1, 8, 8, 1]) # a single 8
 ↪by 8 grayscale image

maxpool= MaxPoolingLayer(pool_size=4, stride=2, name="maxpool_test")
out = maxpool.forward(input_image)

print("Received output shape: {}, Expected output shape: (1, 3, 3, 1)".
 ↪format(out.shape))

correct_out = np.array([[
```

```
   [[0.11428571],
    [0.13015873],
    [0.14603175]],

   [[0.24126984],
    [0.25714286],
    [0.27301587]],

   [[0.36825397],
    [0.38412698],
    [0.4       ]]]])

# Compare your output with the above pre-computed ones.
# The difference should not be larger than 1e-7
print ("Difference: ", rel_error(out, correct_out))
```

```
Received output shape: (1, 3, 3, 1), Expected output shape: (1, 3, 3, 1)
Difference:  1.8750000280978013e-08
```

1.3.2 Backward Pass Max pooling [5pts]

Fill out the `backward` function in the MaxPoolingLayer class.

[40]:
```
img = np.random.randn(15, 8, 8, 3)

dout = np.random.randn(15, 3, 3, 3)

maxpool= MaxPoolingLayer(pool_size=4, stride=2, name="maxpool_test")

dimg_num = eval_numerical_gradient_array(lambda x: maxpool.forward(img), img,
 ↪dout)

out = maxpool.forward(img)
dimg = maxpool.backward(dout)

# The error should be around 1e-8
print("dimg Error: ", rel_error(dimg_num, dimg))
# The shapes should be same
print("dimg Shape: ", dimg.shape, img.shape)
```

```
dimg Error:  3.2762917712940654e-12
dimg Shape:  (15, 8, 8, 3) (15, 8, 8, 3)
```

1.3.3 Test a Small Convolutional Neural Network [3pts]

Please find the `TestCNN` class in `lib/cnn/cnn_models.py`. Again you only need to complete few lines of code in the TODO block. Please design a Convolutional –> Maxpool –> flatten –> fc network where the shapes of parameters match the given shapes. Please insert the corresponding names you defined for each layer to param_name_w, and param_name_b respectively. Here you only modify the param_name part, the _w, and _b are automatically assigned during network setup.

[45]:
```
%reload_ext autoreload

seed = 1234
np.random.seed(seed=seed)

model = TestCNN()
loss_func = cross_entropy()

B, H, W, iC = 4, 8, 8, 3 #batch, height, width, in_channels
k = 3 #kernel size
oC, Hi, O = 3, 27, 5 # out channels, Hidden Layer input, Output size
std = 0.02
x = np.random.randn(B,H,W,iC)
y = np.random.randint(O, size=B)

print ("Testing initialization ... ")

###################################################
# TODO: param_name should be replaced accordingly  #
###################################################
w1_std = abs(model.net.get_params("conv_w").std() - std)
b1 = model.net.get_params("conv_b").std()
w2_std = abs(model.net.get_params("fc_w").std() - std)
b2 = model.net.get_params("fc_b").std()
###################################################
#                 END OF YOUR CODE                #
###################################################

assert w1_std < std / 10, "First layer weights do not seem right"
assert np.all(b1 == 0), "First layer biases do not seem right"
assert w2_std < std / 10, "Second layer weights do not seem right"
assert np.all(b2 == 0), "Second layer biases do not seem right"
print ("Passed!")

print ("Testing test-time forward pass ... ")
w1 = np.linspace(-0.7, 0.3, num=k*k*iC*oC).reshape(k,k,iC,oC)
w2 = np.linspace(-0.2, 0.2, num=Hi*O).reshape(Hi, O)
b1 = np.linspace(-0.6, 0.2, num=oC)
b2 = np.linspace(-0.9, 0.1, num=O)

###################################################
# TODO: param_name should be replaced accordingly  #
###################################################
```

```python
model.net.assign("conv_w", w1)
model.net.assign("conv_b", b1)
model.net.assign("fc_w", w2)
model.net.assign("fc_b", b2)
###################################################
#                 END OF YOUR CODE                #
###################################################

feats = np.linspace(-5.5, 4.5, num=B*H*W*iC).reshape(B,H,W,iC)
scores = model.forward(feats)
correct_scores = np.asarray([[-13.85107294, -11.52845818,  -9.20584342,  -6.
↪88322866,  -4.5606139 ],
 [-11.44514171, -10.21200524 , -8.97886878 , -7.74573231 , -6.51259584],
 [ -9.03921048,  -8.89555231 , -8.75189413 , -8.60823596,  -8.46457778],
 [ -6.63327925 , -7.57909937 , -8.52491949 , -9.4707396 , -10.41655972]])
scores_diff = np.sum(np.abs(scores - correct_scores))
assert scores_diff < 1e-6, "Your implementation might be wrong!"
print ("Passed!")

print ("Testing the loss ...",)
y = np.asarray([0, 2, 1, 4])
loss = loss_func.forward(scores, y)
dLoss = loss_func.backward()
correct_loss = 4.56046848799693
assert abs(loss - correct_loss) < 1e-10, "Your implementation might be wrong!"
print ("Passed!")

print ("Testing the gradients (error should be no larger than 1e-6) ...")
din = model.backward(dLoss)
for layer in model.net.layers:
    if not layer.params:
        continue
    for name in sorted(layer.grads):
        f = lambda _: loss_func.forward(model.forward(feats), y)
        grad_num = eval_numerical_gradient(f, layer.params[name], verbose=False)
        print ('%s relative error: %.2e' % (name, rel_error(grad_num, layer.
↪grads[name])))
```

```
Testing initialization …
Passed!
Testing test-time forward pass …
Passed!
Testing the loss …
Passed!
Testing the gradients (error should be no larger than 1e-6) …
conv_b relative error: 3.90e-09
conv_w relative error: 9.26e-10
```

```
fc_b relative error: 1.33e-10
fc_w relative error: 3.89e-07
```

1.3.4 Training the Network [25pts]

In this section, we defined a `SmallConvolutionalNetwork` class for you to fill in the TODO block in `lib/cnn/cnn_models.py`.

Here please design a network with at most two convolutions and two maxpooling layers (you may use less). You can adjust the parameters for any layer, and include layers other than those listed above that you have implemented (such as fully-connected layers and non-linearities). You are also free to select any optimizer you have implemented (with any learning rate).

You will train your network on CIFAR-100 20-way superclass classification. Try to find a combination that is able to achieve 40% validation accuracy.

Since the CNN takes significantly longer to train than the fully connected network, it is suggested to start off with fewer filters in your Conv layers and fewer intermediate fully-connected layers so as to get faster initial results.

[46]:
```
# Arrange the data
data_dict = {
    "data_train": (data["data_train"], data["labels_train"]),
    "data_val": (data["data_val"], data["labels_val"]),
    "data_test": (data["data_test"], data["labels_test"])
}
```

[47]:
```
print("Data shape:", data_dict["data_train"][0].shape)
print("Flattened data input size:", np.prod(data["data_train"].shape[1:]))
print("Number of data classes:", max(data['labels_train']) + 1)
```

```
Data shape: (40000, 32, 32, 3)
Flattened data input size: 3072
Number of data classes: 20
```

[62]:
```
%reload_ext autoreload

seed = 123
np.random.seed(seed=seed)

model = SmallConvolutionalNetwork()
loss_f = cross_entropy()


results = None
#############################################################################
# TODO: Use the train_net function you completed to train a network        #
# You may only adjust the hyperparameters within this block                 #
#############################################################################
```

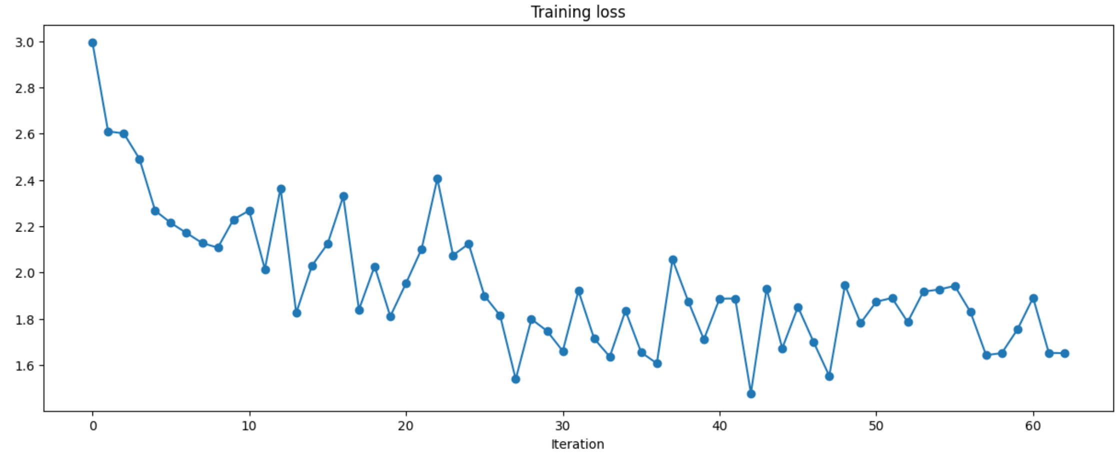
```python
optimizer = Adam(model.net, 1e-3)
batch_size = 64
epochs = 10
lr_decay = .999
lr_decay_every = 10
# regularization = "none"
# reg_lambda = 0.01
######################################################################
#                             END OF YOUR CODE                       #
######################################################################
results = train_net(data_dict, model, loss_f, optimizer, batch_size, epochs,
                    lr_decay, lr_decay_every, show_every=4000, verbose=True,
 ↪regularization=regularization, reg_lambda=reg_lambda)
opt_params, loss_hist, train_acc_hist, val_acc_hist = results
```

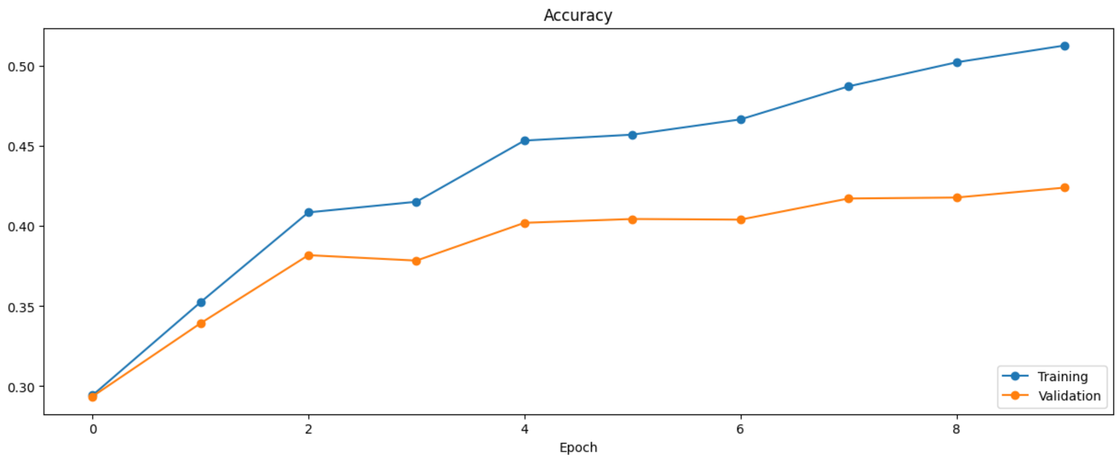
```
  0%|
| 1/625 [00:00<09:21,  1.11it/s]

(Iteration 1 / 6250) Average loss: 2.995702137939523

100%|
      | 625/625 [09:27<00:00,  1.10it/s]

(Epoch 1 / 10) Training Accuracy: 0.294275, Validation Accuracy: 0.2936

100%|
      | 625/625 [09:27<00:00,  1.10it/s]

(Epoch 2 / 10) Training Accuracy: 0.352425, Validation Accuracy: 0.3393

100%|
      | 625/625 [09:32<00:00,  1.09it/s]

(Epoch 3 / 10) Training Accuracy: 0.408625, Validation Accuracy: 0.3819

100%|
      | 625/625 [09:31<00:00,  1.09it/s]

(Epoch 4 / 10) Training Accuracy: 0.415275, Validation Accuracy: 0.3785

100%|
      | 625/625 [09:36<00:00,  1.08it/s]

(Epoch 5 / 10) Training Accuracy: 0.4535, Validation Accuracy: 0.4021

100%|
      | 625/625 [09:31<00:00,  1.09it/s]

(Epoch 6 / 10) Training Accuracy: 0.4572, Validation Accuracy: 0.4045

 40%|
| 251/625 [03:48<05:46,  1.08it/s]

(Iteration 4001 / 6250) Average loss: 2.036589462845751
```

```
100%|
      | 625/625 [09:32<00:00,  1.09it/s]

(Epoch 7 / 10) Training Accuracy: 0.466725, Validation Accuracy: 0.4041

100%|
      | 625/625 [09:28<00:00,  1.10it/s]

(Epoch 8 / 10) Training Accuracy: 0.48735, Validation Accuracy: 0.4173

100%|
      | 625/625 [09:28<00:00,  1.10it/s]

(Epoch 9 / 10) Training Accuracy: 0.5024, Validation Accuracy: 0.4179

100%|
      | 625/625 [09:27<00:00,  1.10it/s]

(Epoch 10 / 10) Training Accuracy: 0.5129, Validation Accuracy: 0.4241
```

Run the code below to generate the training plots.

[63]:
```
%reload_ext autoreload

opt_params, loss_hist, train_acc_hist, val_acc_hist = results

# Plot the learning curves
plt.subplot(2, 1, 1)
plt.title('Training loss')
loss_hist_ = loss_hist[1::100]  # sparse the curve a bit
plt.plot(loss_hist_, '-o')
plt.xlabel('Iteration')

plt.subplot(2, 1, 2)
plt.title('Accuracy')
plt.plot(train_acc_hist, '-o', label='Training')
plt.plot(val_acc_hist, '-o', label='Validation')
plt.xlabel('Epoch')
plt.legend(loc='lower right')
plt.gcf().set_size_inches(15, 12)

plt.show()
```



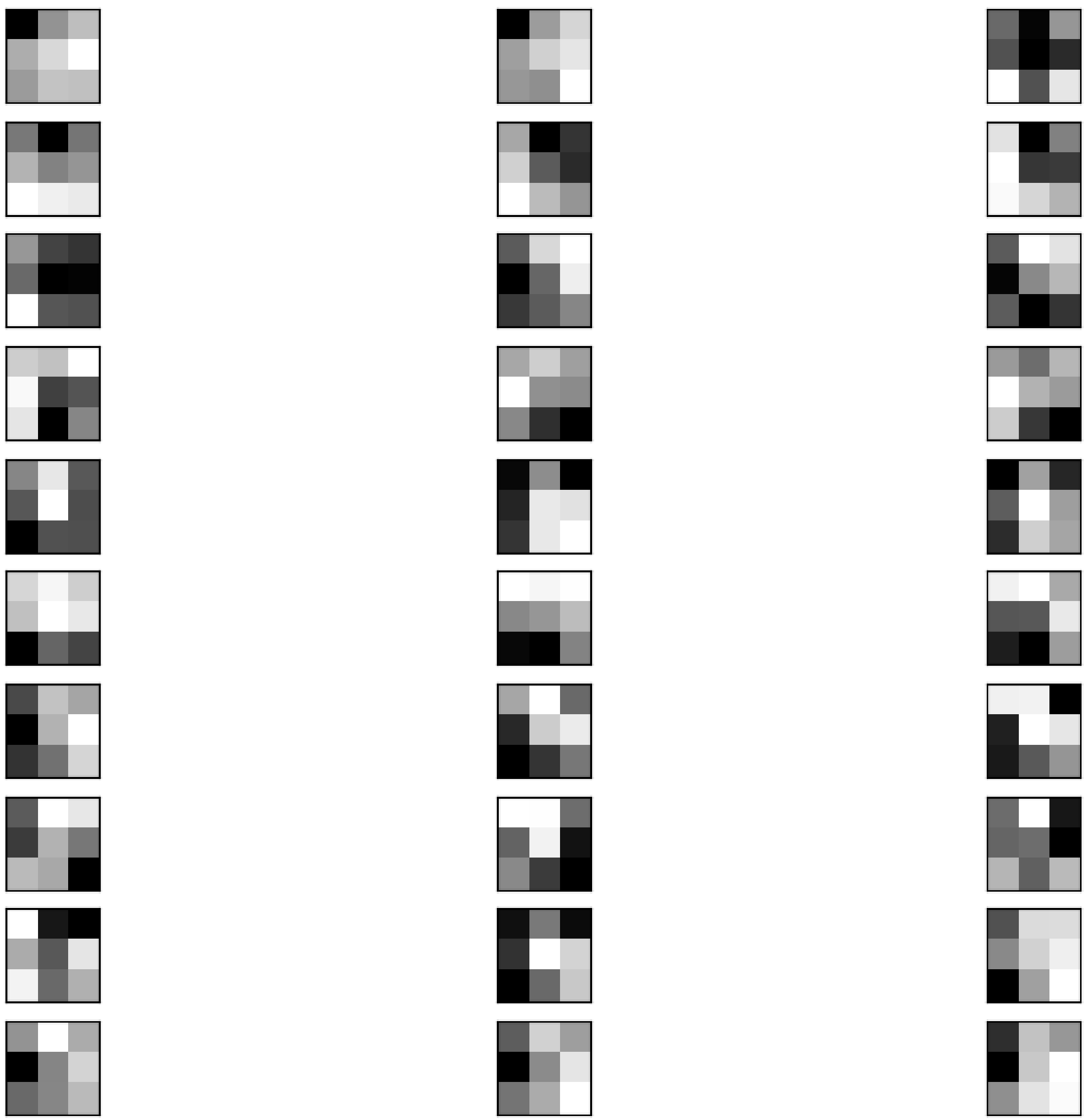


1.3.5 Visualizing Layers [5pts]

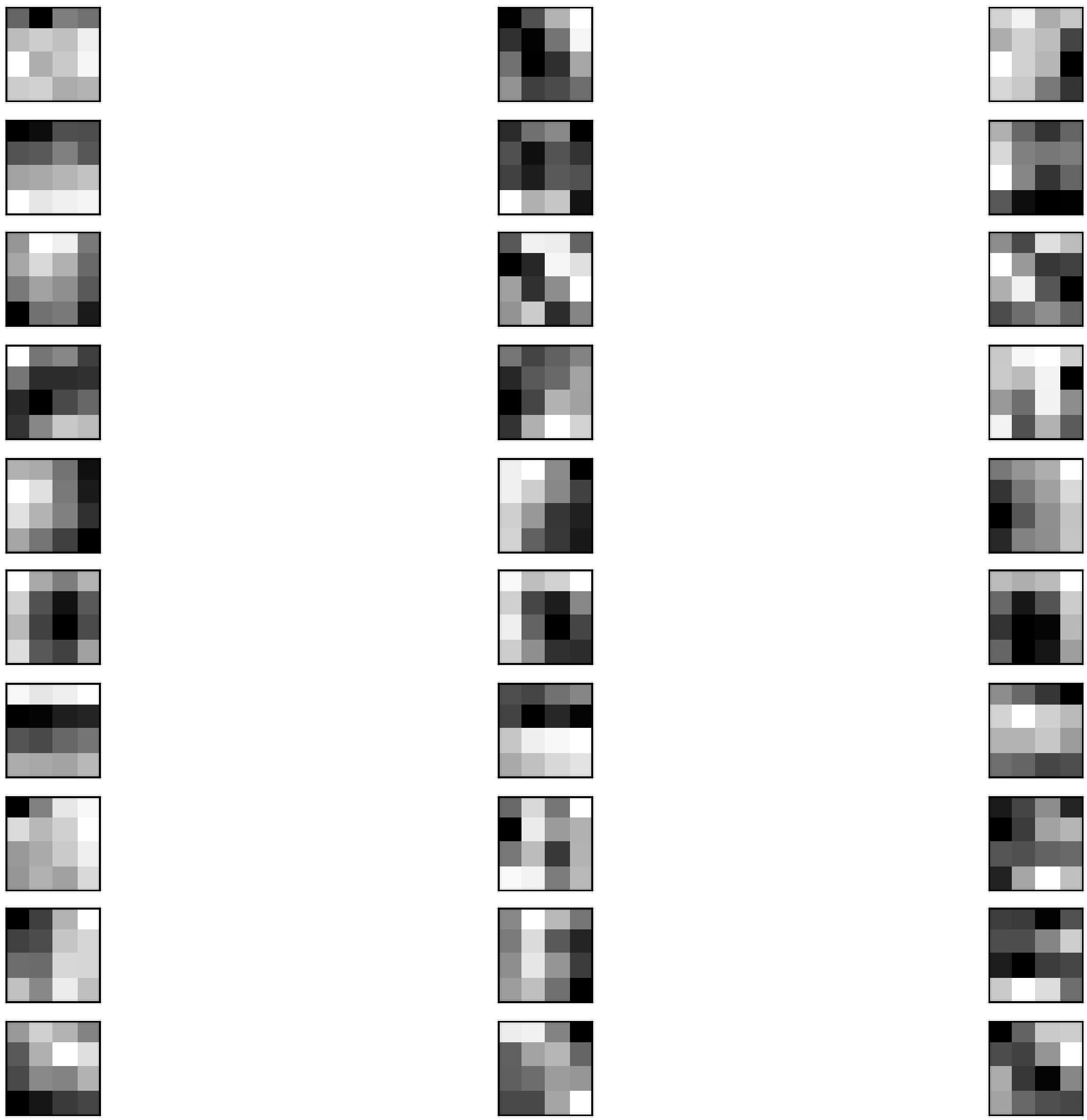
An interesting finding from early research in convolutional networks was that the learned convolutions resembled filters used for things like edge detection. Complete the code below to visualize the filters in the first convolutional layer of your best model.

[80]:
```
im_array = None
nrows, ncols = None, None

###################################################
# TODO: read the weights in the convolutional     #
# layer and reshape them to a grid of images to   #
# view with matplotlib.                           #
###################################################
filters = model.net.get_params("conv1_w")
index = 1
for i in range(10):
    f = filters[:, :, :, i]
```

```
    for j in range(3):
        ax = plt.subplot(10, 3, index)
        ax.set_xticks([])
        ax.set_yticks([])
        plt.imshow(f[:, :, j])
        index += 1
###################################################
#                 END OF YOUR CODE                #
###################################################

# plt.imshow(im_array)
```



[81]:
```
im_array = None
nrows, ncols = None, None

##################################################
# TODO: read the weights in the convolutional     #
# layer and reshape them to a grid of images to   #
# view with matplotlib.                           #
##################################################
filters = model.net.get_params("conv2_w")
index = 1
for i in range(10):
    f = filters[:, :, :, i]
    for j in range(3):
        ax = plt.subplot(10, 3, index)
        ax.set_xticks([])
        ax.set_yticks([])
        plt.imshow(f[:, :, j])
        index += 1
##################################################
#                END OF YOUR CODE                 #
##################################################

# plt.imshow(im_array)
```



Inline Question: Comment below on what kinds of filters you see. Include your response in your submission [5pts] Here we have ten filters (Vertical) for the RGB channels (Horizontal) for the two convolution layers. The lighter pixels indicate a lighter weight, and vice versa.

From both layers of filters, we can see most filters have higher weights around the edges, while a few have filter weights in the center portion. This indicates that the majority of the filters are looking for some sort of edges around the objects during processing for classifications.

Another interesting finding is that, for the same filter, the weights across three different color channels vary filter by filter. Some filters have similar weights for the three different color channels, such as the second 3x3 filter, some have similar weights for two of the channels and a different, or even opposite weight distribution for the third channel, such as the fifth 4x4 filter, and some just have different weight distribution across three channels.

I think these different weight distributions across color channels enable some of the filter to detect the edges of the object against the background, and some other filters to detect other patterns for certain objects.

1.4 Extra-Credit: Analysis on Trained Model [5pts]

For extra credit, you can perform some additional analysis of your trained model. Some suggested analyses are: 1. Plot the confusion matrix of your model's predictions on the test set. Look for trends to see which classes are frequently misclassified as other classes (e.g. are the two vehicle superclasses frequently confused with each other?). 2. Implement BatchNorm and analyze how the models train with and without BatchNorm. 3. Introduce some small noise in the labels, and investigate how that affects training and validation accuracy.

You are free to choose any analysis question of interest to you. We will not be providing any starter code for the extra credit. Include your extra-credit analysis as the final section of your report pdf, titled "Extra Credit".

2 Submission

Please prepare a PDF document `problem_2_solution.pdf` in the root directory of this repository with all plots and inline answers of your solution. Concretely, the document should contain the following items in strict order: 1. Training loss / accuracy curves for CNN training 2. Visualization of convolutional filters 3. Answers to inline questions about convolutional filters

Note that you still need to submit the jupyter notebook with all generated solutions. We will randomly pick submissions and check that the plots in the PDF and in the notebook are equivalent.

[ ]:

```

```