Problem 1: Basics of Neural Networks

Problem 1: Basics of Neural Networks
- Training loss / accuracy curves for the simple neural network training with > 30% validation accuracy
- Plots for comparing vanilla SGD to SGD + Weight Decay, SGD + L1 and SGD + L2
- "Comparing different Regularizations with Adam" plots

Training loss / accuracy curves for the simple neural network training with > 30% validation accuracy

Set hyperparameter:

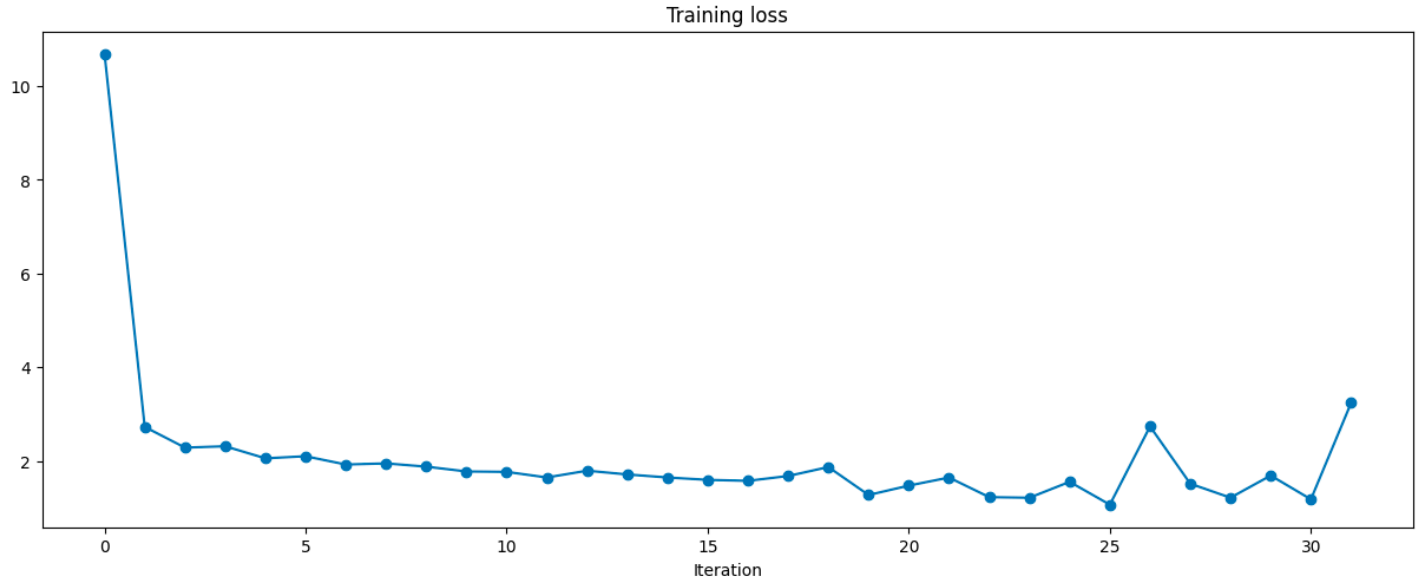
```
batch_size = 256
epochs = 20
lr_decay = 0.95
lr_decay_every = 50
```

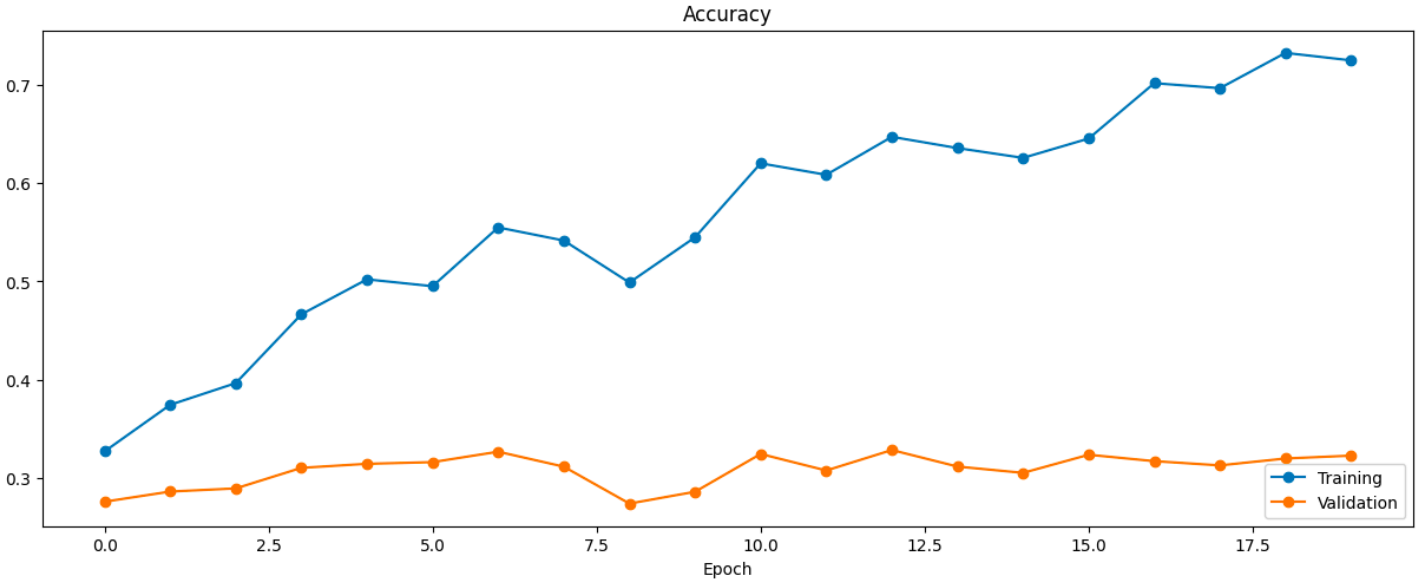
Change Structure:

```
class TinyNet(Module):
    def __init__(self, keep_prob=0, dtype=np.float32, seed=None):
        """ Some comments """
        self.net = sequential(
            ########## TODO: ##########
            flatten(name='flat'),
            fc(input_dim=3072, output_dim=1024, init_scale=5e-2, name='fc1'),
            gelu(name="gelu1"),
            fc(1024, 20, init_scale=5e-2, name='fc2')
            ########### END ###########
        )
```

The accuracy meets the requirements:

```
Validation Accuracy: 32.300000000000004%
Testing Accuracy: 33.03%
```

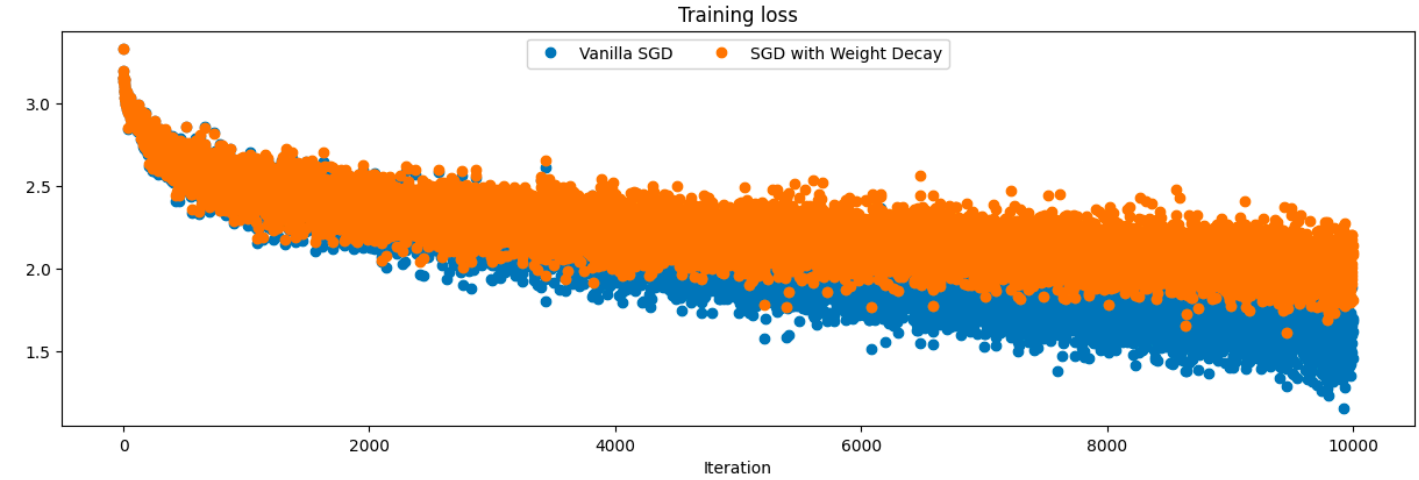


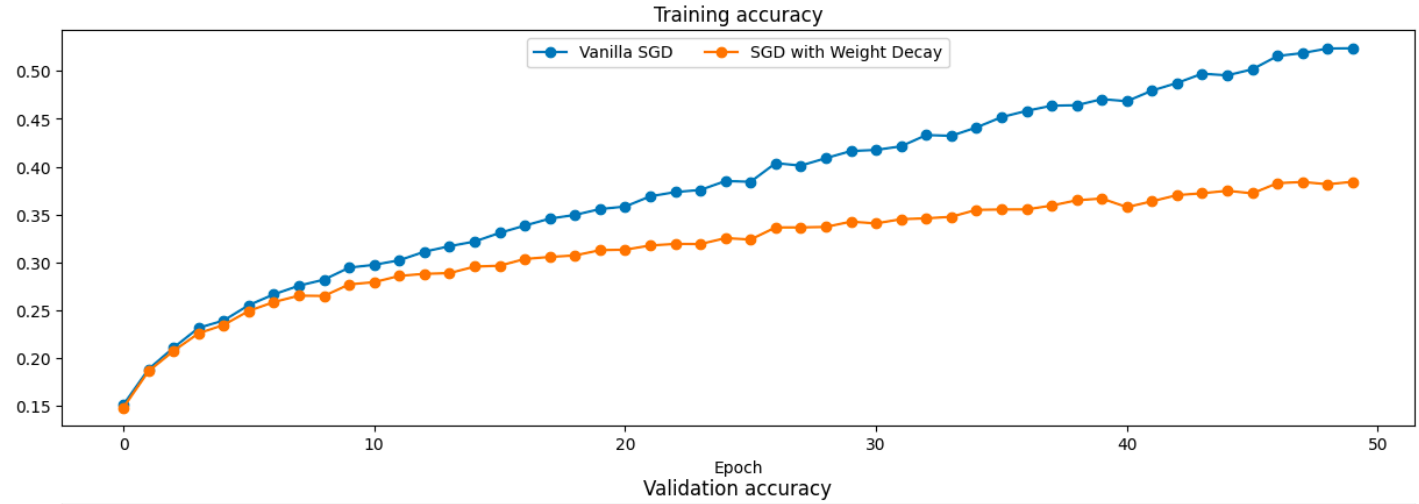


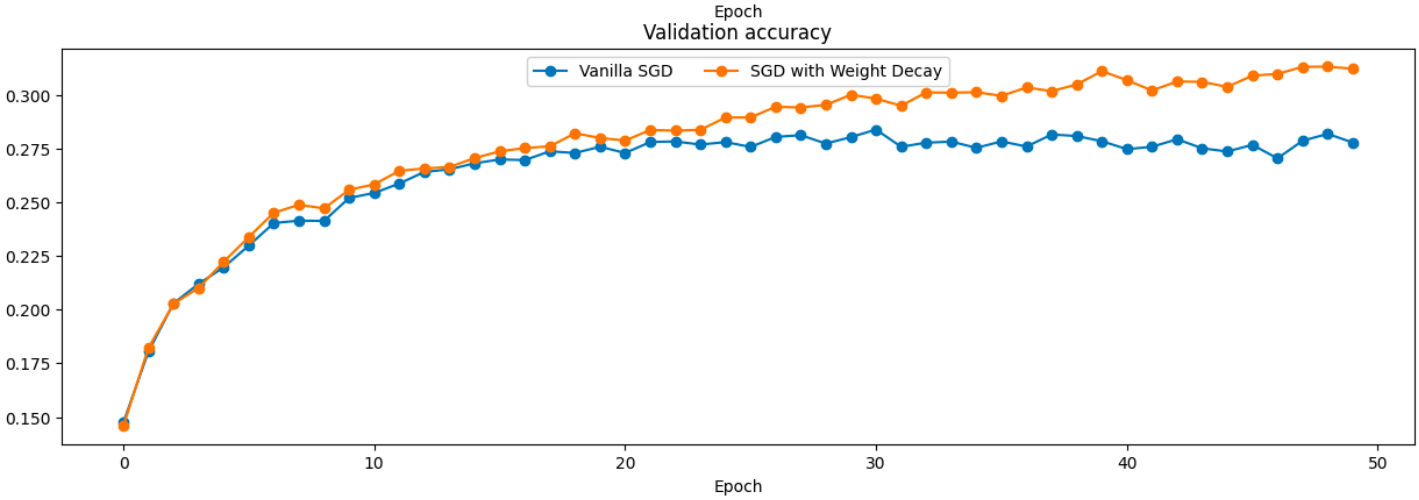
Plots for comparing vanilla SGD to SGD + Weight Decay, SGD + L1 and SGD + L2

vanilla SGD and SGD + Weight Decay:

As shown in figure, SGD with weight decay has relatively higher loss, lower training accuracy, but has higher validation accuracy than vanilla SGD.

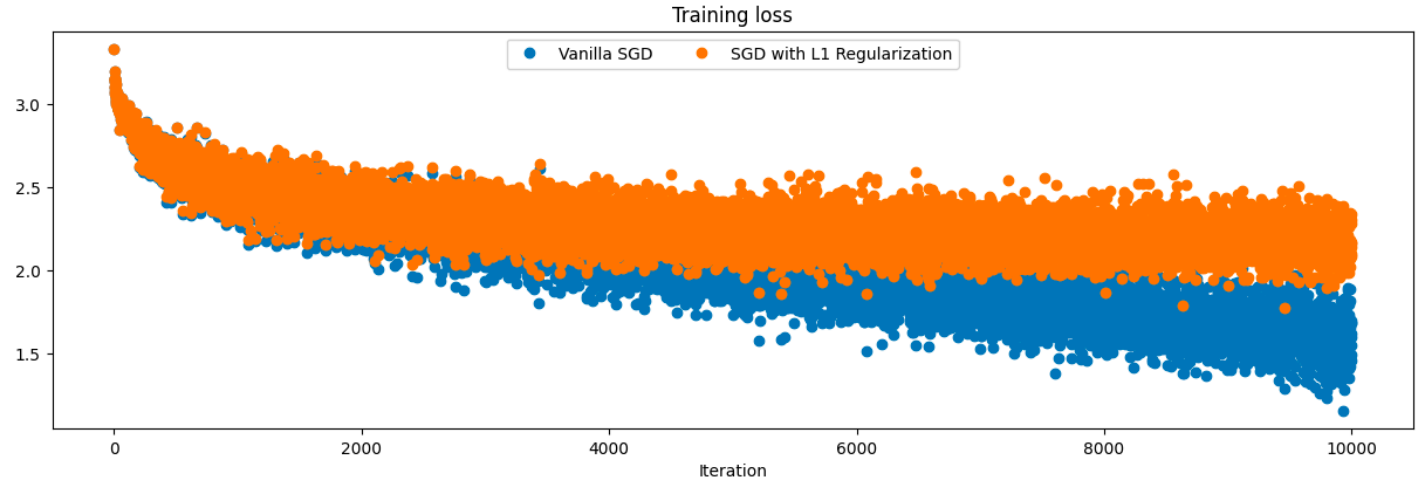


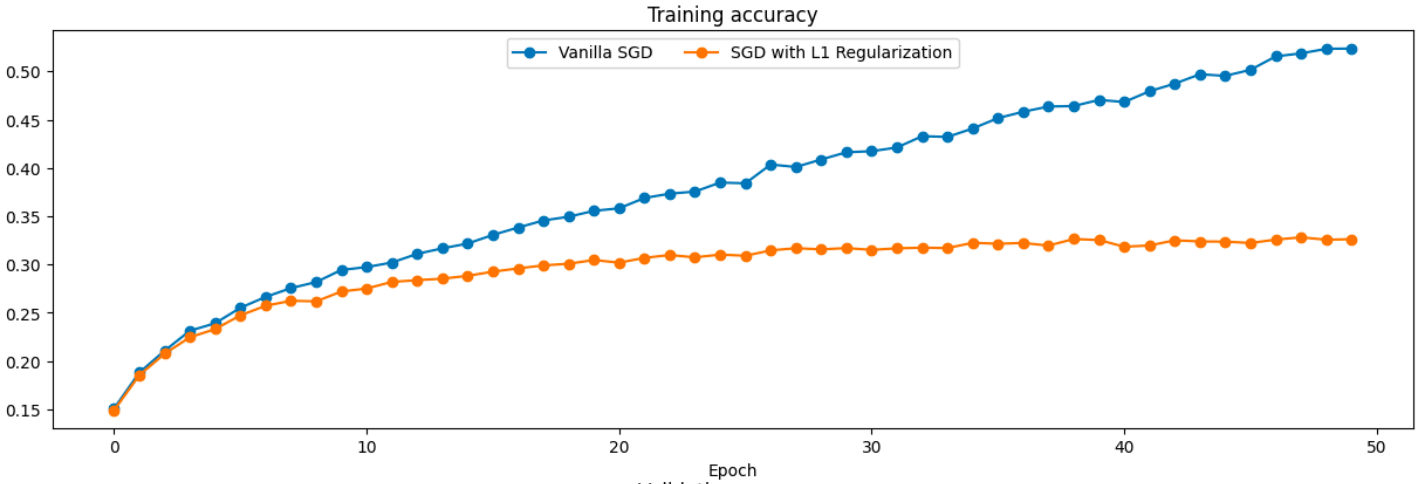


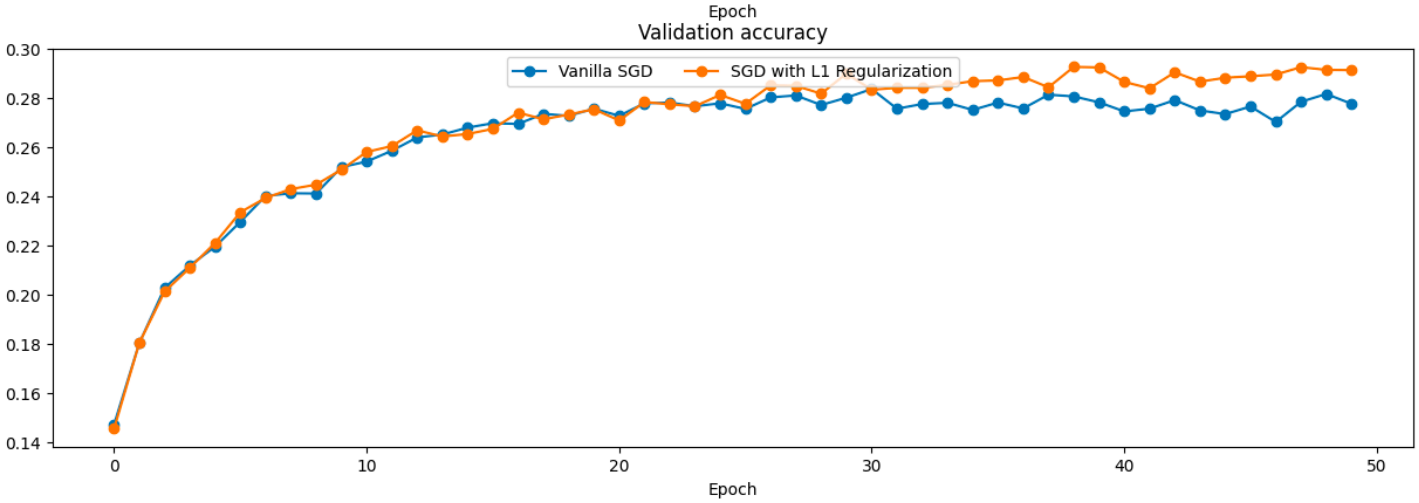


vanilla SGD and SGD + L1:

As shown in figure, SGD with L1 regularization has relatively higher loss, lower training accuracy, but higher validation accuracy than vanilla SGD.



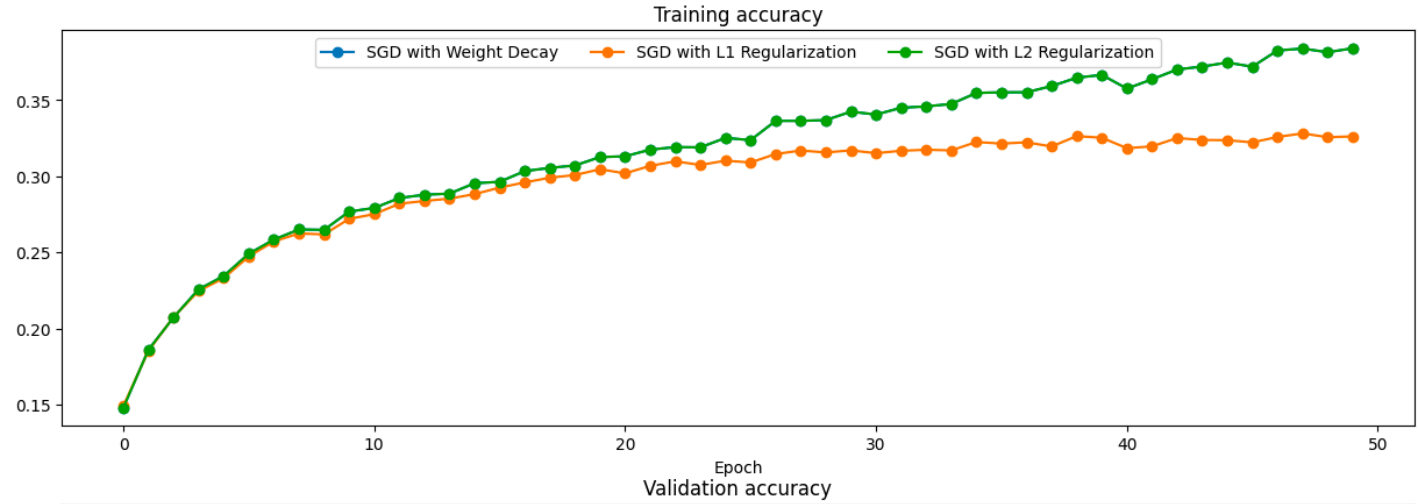


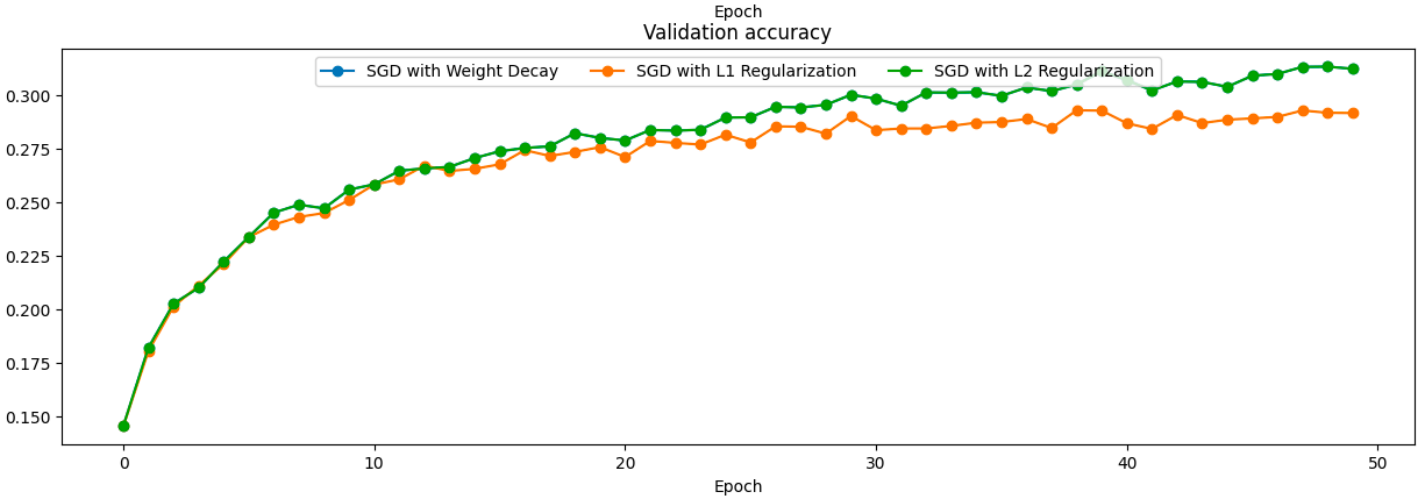


SGD + Weight Decay, SGD + L1 and SGD + L2:

```
l2_lambda = 0.5e-2 # I use 2 * lambda * theta as l2 regularization
```

As shown in figure, SGD + L2 achives the EXACTLY SAME effect as weight decay. They both have better training and validation accuracy than SGD + L1.

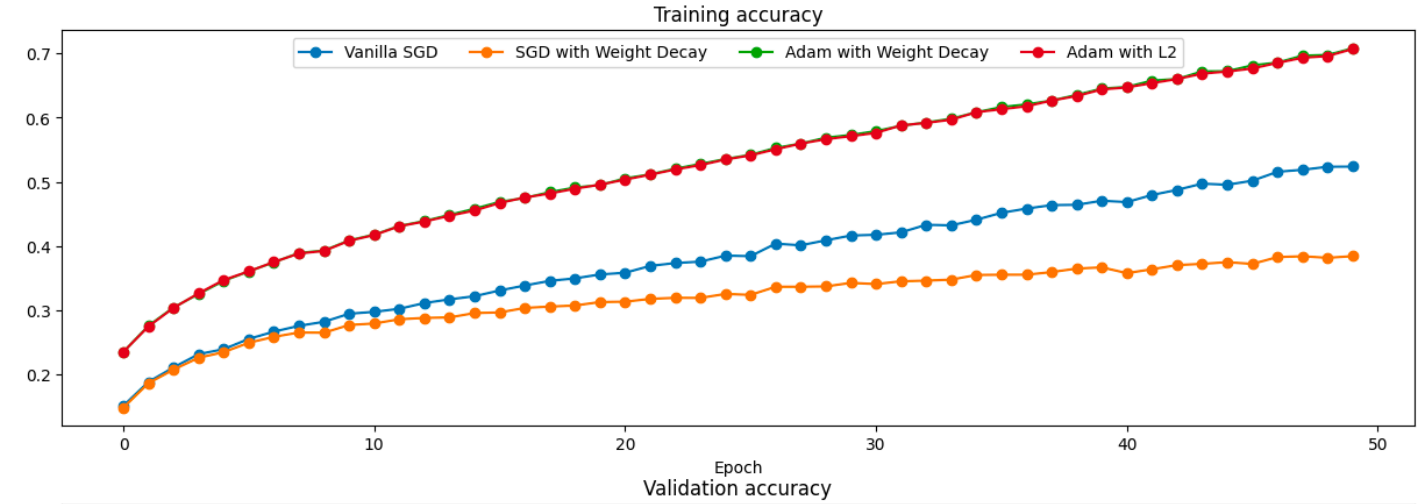


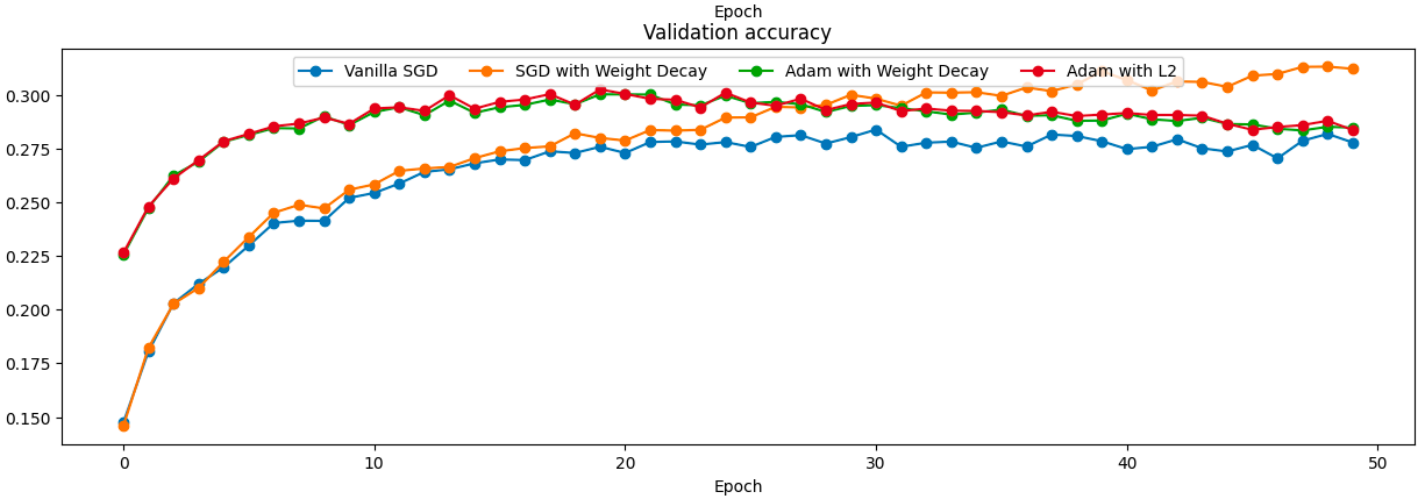


"Comparing different Regularizations with Adam" plots

AdamW use $weight_decay = 1e-6$, Adam with L2 use $\lambda_{L_2} = 1e-4$ strength. There is not much difference between their effect. Both Adam have lower loss, and higher training accuracy than two kinds of SGD. But with epoch increasing, SGD with weight decay achieve highest validation loss.

Training loss
Vanilla SGD
SGD with Weight Decay
Adam with Weight Decay
Adam with L2
3.0
2.5
2.0
1.5
1.0
0
2000
4000
6000
8000
10000
Iteration


Training accuracy
Vanilla SGD
SGD with Weight Decay
Adam with Weight Decay
Adam with L2
0.7
0.6
0.5
0.4
0.3
0.2
0
10
20
30
40
50
Epoch


Validation accuracy
Vanilla SGD
SGD with Weight Decay
Adam with Weight Decay
Adam with L2
0.300
0.275
0.250
0.225
0.200
0.175
0.150
0
10
20
30
40
50
Epoch