Part A

[512, 256, 128], relu, use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8905 - 0.8930, diverge > 7

[256, 128], relu, use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8918 – 0.8923, diverge > 7

[1024,512,256], relu, use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8885 – 0.8938, diverge > 6 (We will keep with this setting)

[784,392,196], relu, use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8875, diverge > 5, huge diverge >7

[1024,512,256,128], relu, use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8828, diverge > 8

[512, 256, 128], LeakyReLU(0.1), use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8897, diverge > 8

[256, 128], LeakyReLU(0.1), use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8875 – 0.8898, diverge > 8

[1024,512,256], LeakyReLU(0.1), use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8855 – 0.8888, diverge > 6

[784,392,196], LeakyReLU(0.1), use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8850, diverge > 7

[1024,512,256,128], LeakyReLU(0.1), use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8805, diverge > 8

[512, 256, 128], GELU, use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8920, diverge > 5

[256, 128], GELU, use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8920, diverge > 4

[1024,512,256], GELU, use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8897, diverge > 7

[784,392,196], GELU, use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8877, diverge > 6

[1024,512,256,128], GELU, use_batchnorm=true, dropout=0.25, normalize=global

Val_acc = 0.8812 – 0.8887, diverge > 5

[512, 256, 128], relu, use_batchnorm=false, dropout=0.25, normalize=global

Val_acc = 0.8860 – 0.8912, diverge > 13

[256, 128], relu, use_batchnorm=false, dropout=0.25, normalize=global

Val_acc = 0.8868 – 0.8880, diverge > 9

[1024,512,256], relu, use_batchnorm=false, dropout=0.25, normalize=global

Val_acc = 0.8818 – 0.8860, diverge > 6

[784,392,196], relu, use_batchnorm=false, dropout=0.25, normalize=global

Val_acc = 0.8840, diverge > 8, huge diverge > 15

[1024,512,256,128], relu, use_batchnorm=false, dropout=0.25, normalize=global

Val_acc = 0.8852 – 0.8885, diverge > 12

[512, 256, 128], LeakyReLU(0.1), use_batchnorm=false, dropout=0.25, normalize=global

Val_acc = 0.8803 – 0.8893, diverge > 13

[256, 128], LeakyReLU(0.1), use_batchnorm=false, dropout=0.25, normalize=global

Val_acc = 0.8853, diverge > 10

[1024,512,256], LeakyReLU(0.1), use_batchnorm=false, dropout=0.25, normalize=global

Val_acc = 0.8865, diverge > 9 (basically matched not that diverge)

[784,392,196], LeakyReLU(0.1), use_batchnorm=false, dropout=0.25, normalize=global

Val_acc = 0.8802 – 0.8823, diverge > 8

[1024,512,256,128], LeakyReLU(0.1), use_batchnorm=false, dropout=0.25, normalize=global

Val_acc = 0.8798, diverge > 10, huge diverge > 13

[512, 256, 128], GELU, use_batchnorm= false, dropout=0.25, normalize=global

Val_acc = 0.8838 – 0.88.75, diverge > 7, huge diverge > 13

[256, 128], GELU, use_batchnorm= false, dropout=0.25, normalize=global

Val_acc = 0.8898, diverge > 9

[1024,512,256], GELU, use_batchnorm= false, dropout=0.25, normalize=global

Val_acc = 0.8872 – 0.8890, diverge > 10, huge diverge > 12

[784,392,196], GELU, use_batchnorm= false, dropout=0.25, normalize=global

Val_acc = 0.8870, diverge > 9

[1024,512,256,128], GELU, use_batchnorm= false, dropout=0.25, normalize=global

Val_acc = 0.8835, diverge > 8

Part B

[512,256,128], GELU, use_batchnorm=true, dropout=0, normalize=global, weight decay=5e-4

Val_acc = 0.8843, Diverge > 2

[512,256,128], GELU, use_batchnorm=true, dropout=0.25, normalize=global, weight decay=5e-4

Val_acc = 0.8858, diverge > 13

[512,256,128], GELU, use_batchnorm=true, dropout=0.5, normalize=global, weight decay=5e-4

Val_acc = 0.8787, never converge???? Always parallel

[512,256,128], GELU, use_batchnorm=true, dropout=0, normalize=global, weight decay=1e-3

Val_acc = 0.8850, diverge > 2/6

[512,256,128], GELU, use_batchnorm=true, dropout=0.25, normalize=global, weight decay=1e-3

Val_acc = 0.8782, no diverge (this prevents the overfitting, but hurt the val_acc)

[512,256,128], GELU, use_batchnorm=true, dropout=0.5, normalize=global, weight decay=1e-3

Val_acc = 0.8700, never converge???? Always parallel

[512,256,128], GELU, use_batchnorm=true, dropout=0, normalize=global, weight decay=1e-4

Val_acc = 0.8887, diverge > 2

[512,256,128], GELU, use_batchnorm=true, dropout=0.25, normalize=global, weight decay=1e-4

Val_acc = 0.8962, diverge > 5, (although this has overfit later, but it does not hurt val_acc)

[512,256,128], GELU, use_batchnorm=true, dropout=0.5, normalize=global, weight decay=1e-4

Val_acc = 0.8907, loss parallel, but acc same at epoch=12 - 17, 18 – diverge, 19 – close

[512,256,128], GELU, use_batchnorm=true, dropout=0, normalize=global, weight decay=1e-5

Val_acc = 0.8945, diverge > 2

[512,256,128], GELU, use_batchnorm=true, dropout=0.25, normalize=global, weight decay=1e-5

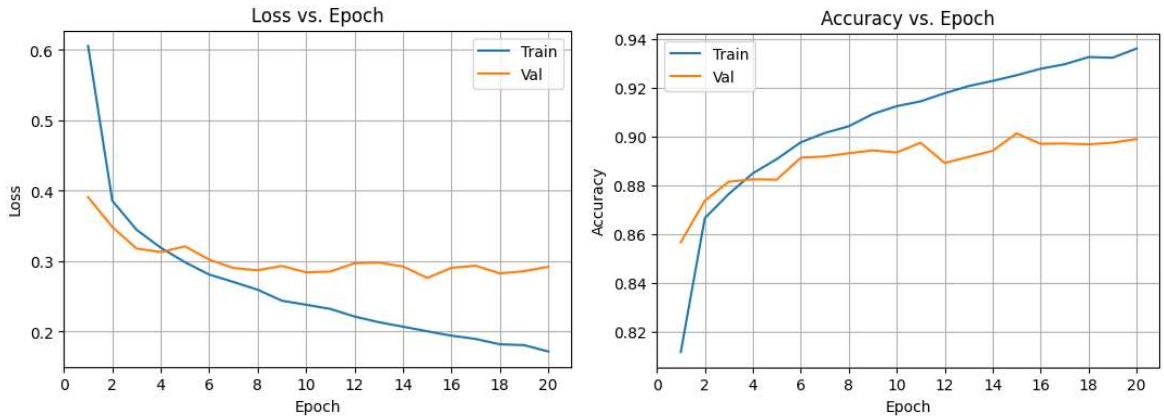
Val_acc = 0.9013, diverge > 4

[512,256,128], GELU, use_batchnorm=true, dropout=0.5, normalize=global, weight decay=1e-5

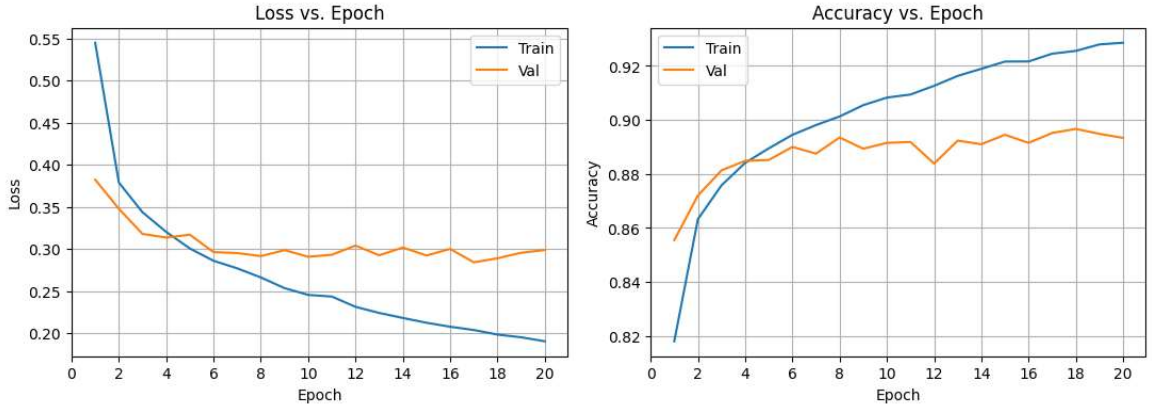
Val_acc = 0.8962, converge at epoch = 12 then diverge again

Part C

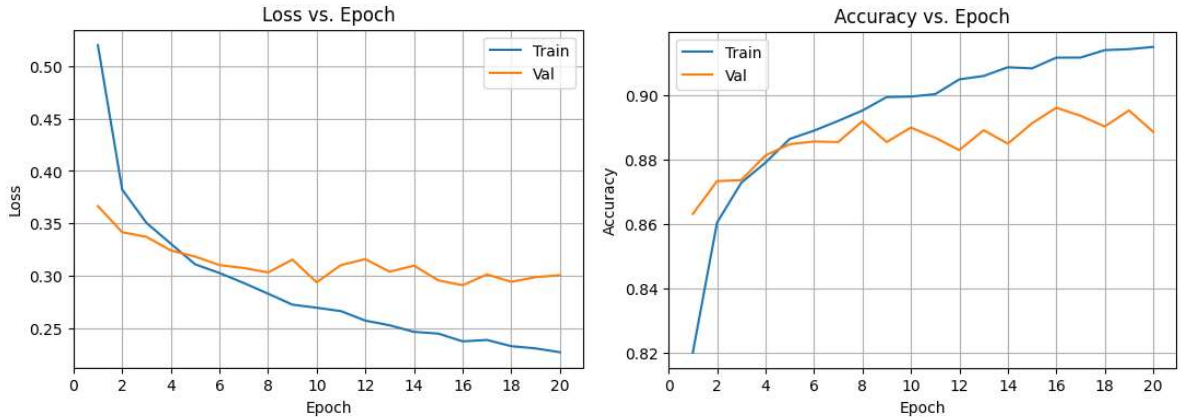
[512,256,128], GELU, use_batchnorm=true, dropout=0.25, normalize=global, weight decay=1e-4, lr=5e-4, optimizer=Adam (slow learning)



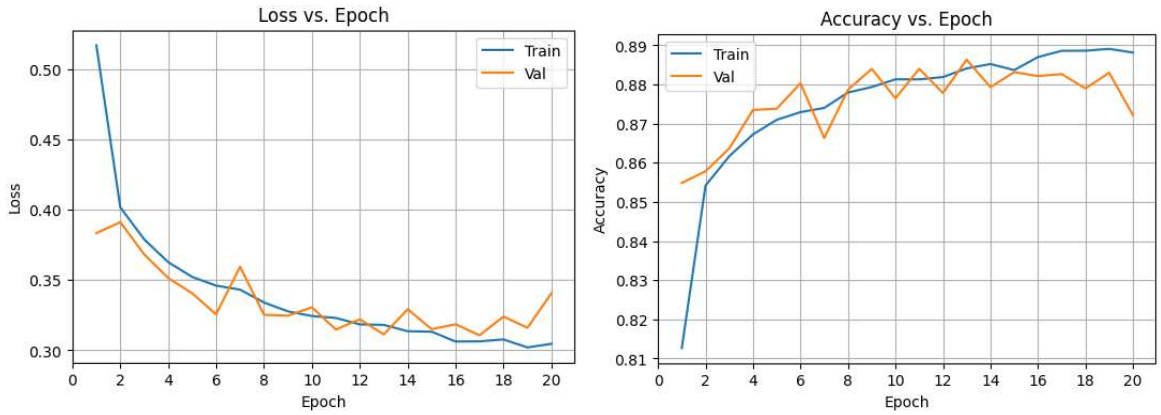
[512,256,128], GELU, use_batchnorm=true, dropout=0.25, normalize=global, weight decay=1e-4, lr=1e-3, optimizer=Adam (I'll keep with this one for part D)



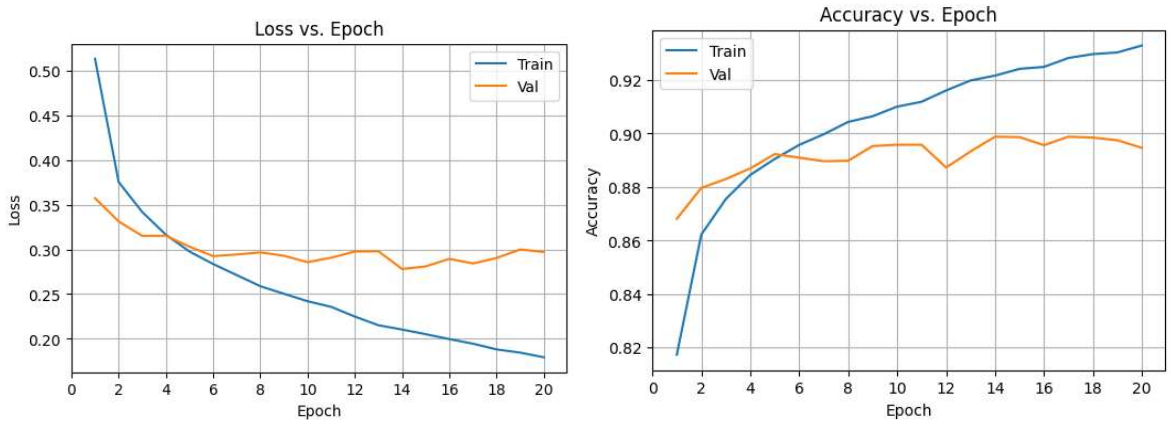
[512,256,128], GELU, use_batchnorm=true, dropout=0.25, normalize=global, weight decay=1e-4, lr=2e-3, optimizer=Adam (a little large, oscillating)



[512,256,128], GELU, use_batchnorm=true, dropout=0.25, normalize=global, weight decay=1e-4, lr=5e-3, optimizer=Adam (too large, keep oscillating)



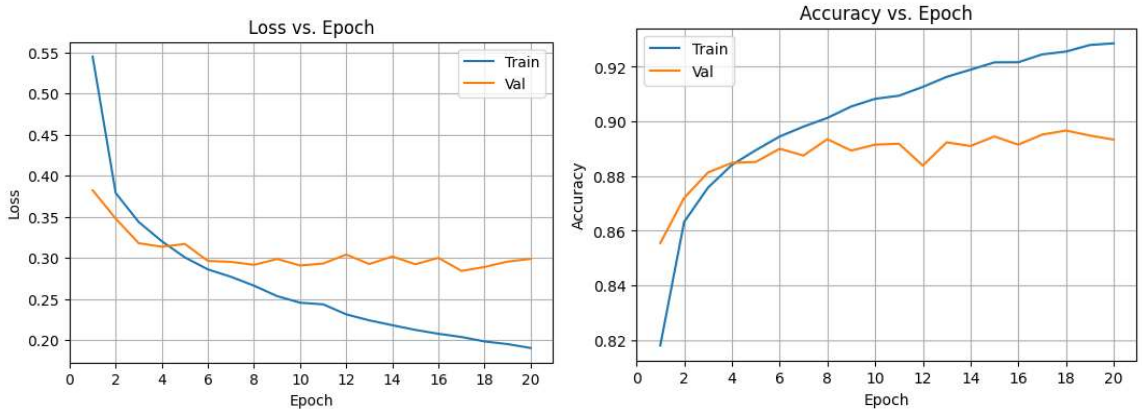
[512,256,128], GELU, use_batchnorm=true, dropout=0.25, normalize=global, weight decay=1e-4, lr=0.05, optimizer=SGD



Part D

[512,256,128], GELU, use_batchnorm=true, dropout=0.25, normalize=global, weight decay=1e-4, lr=1e-3, optimizer=Adam

Val_acc = 0.8967, test_acc = 0.8971



[512,256,128], GELU, use_batchnorm=true, dropout=0.25, normalize=per-image, weight decay=1e-4,

lr=1e-3, optimizer=Adam

Val_acc = 0.8967, test_acc = 0.8999