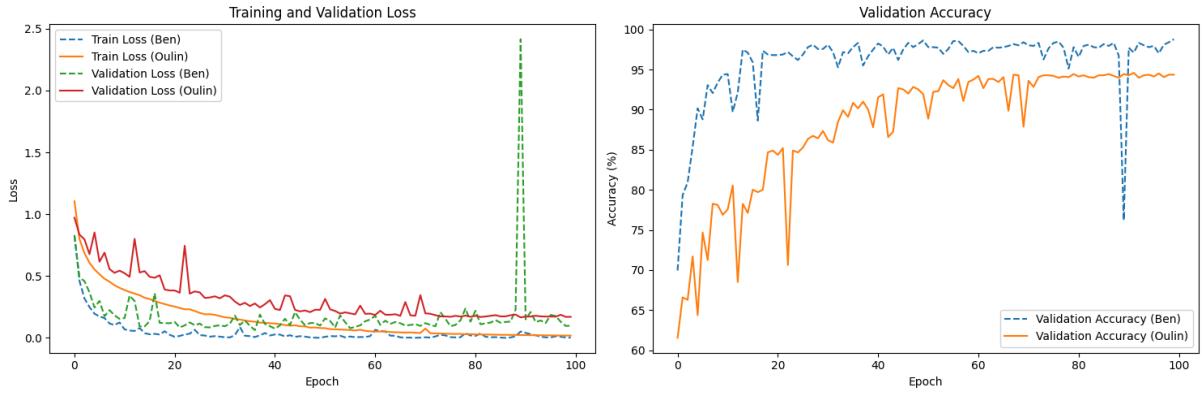
I compared Ben's code with mine, and I summarize the differences in our CNN model below:

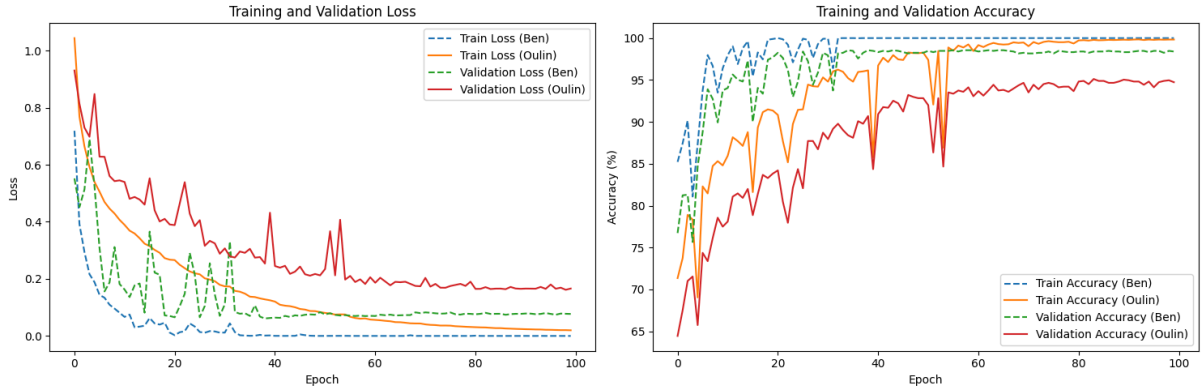
	Ben's CNN	
Total Convolution layers	4	2
Kernel size	5	3
1st layer features	1→16	1→16
2nd layer features	16→32	16→8
3rd layer features	32→64	
4th layer features	64→128	
Activation Function	LeakyReLu	Tanh
MaxPooling (halving)	Yes	Yes
BatchNorm	Yes	No
Dropout	Last linear layer	No
# of Parameters	2400k	260k
Optimizer	Adam + Scheduler*	SGD

*I'm not sure what this does

After running both on Colab, I compare the loss and accuracy below.



I cleaned up and retrained the model again. Now recording traiing set accuracy (on top of validaion accuracy), and produce the plots below.



We should perhaps discuss together in our next meeting. Furthermore, I list some additional strategies to improve the model:

- L2 Regularization
- Dropout in each layer to avoid memorization
- Batch normalization at every step (Ben's model already does this)
- Play with width / depth of the model