Continual Learning Project

Term 1 Update

January 8, 2020

1 Latest Experiments

From our most recent discussions, the experiments I ran focused on looking at the way in which the generalisation error and rate of forgetting between the two teachers varied with the amount of overlap between different layers.

There is a lot to unpack so I highlight one observation in the plots below. I have added the raw tensorboard logs to the google drive here:

https://drive.google.com/drive/folders/1SDVxsGVrHZesUncODQXMhKiUPnOCzW3p

if you want to look in more detail.

1.1 Periodic - ReLU

In this setup the teacher is changed with a fixed period of 5000 training steps. I ran the experiment with five different random seeds. This section shows results with the ReLU non-linearity.

1.1.1 Independent Teachers

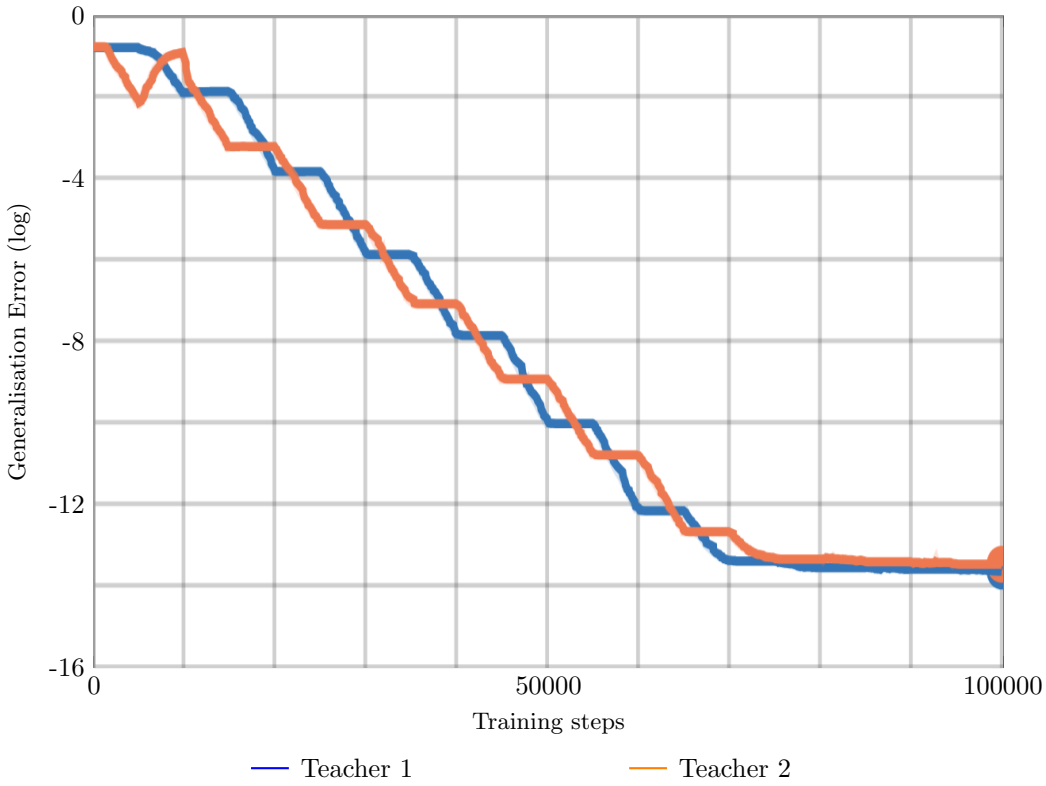


Figure 1: At the start of training, the rate of forgetting is very large. Very quickly however this levels off and performance on the network not currently being used to train the student stays level.

1.1.2 Overlap 25-50%

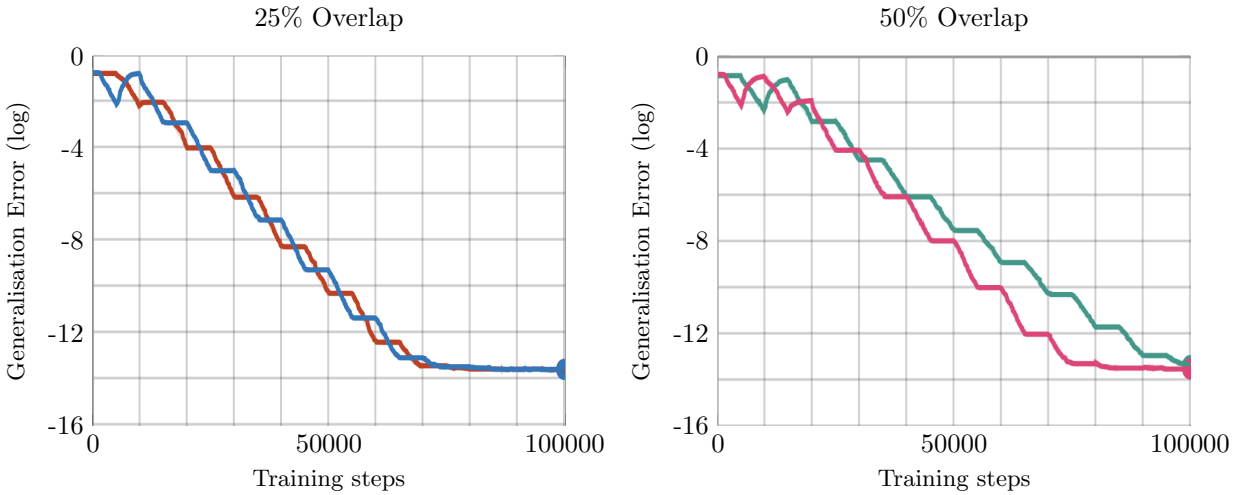


Figure 2: With 25 and 50% overlaps they are not clearly distinguishable from the independent case although the 50% case seems to separate itself a little already. Oddly it takes longer to stop forgetting than in the independent case.

1.1.3 Overlap 75%

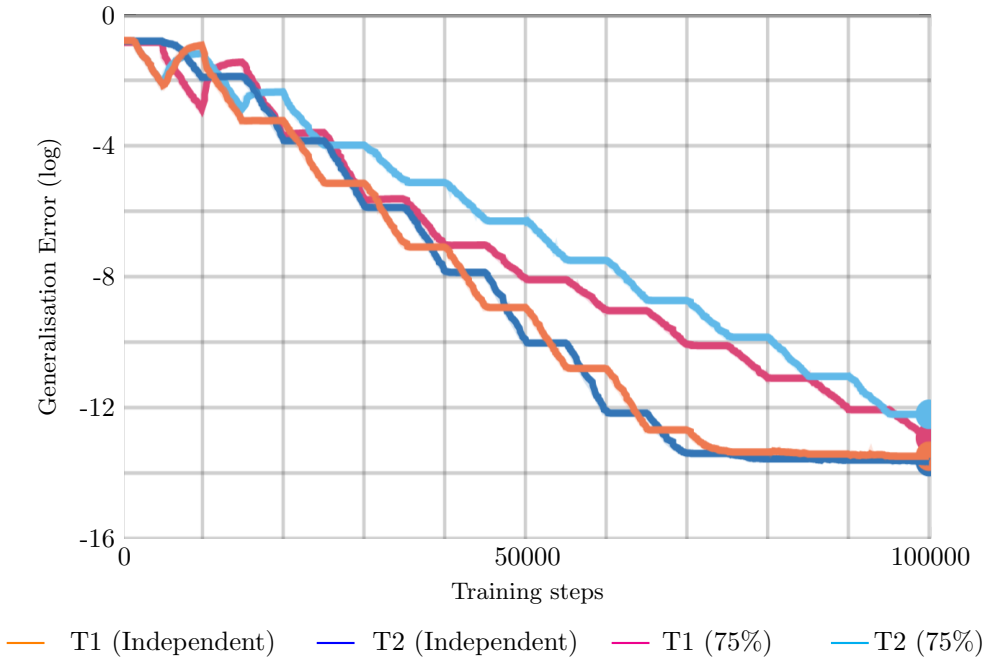


Figure 3: Once we reach 75% there is a clear split from the independent case. There is some co-learning happening but again it starts to take longer to level out when teachers are swapped.

1.1.4 Overlap 100%

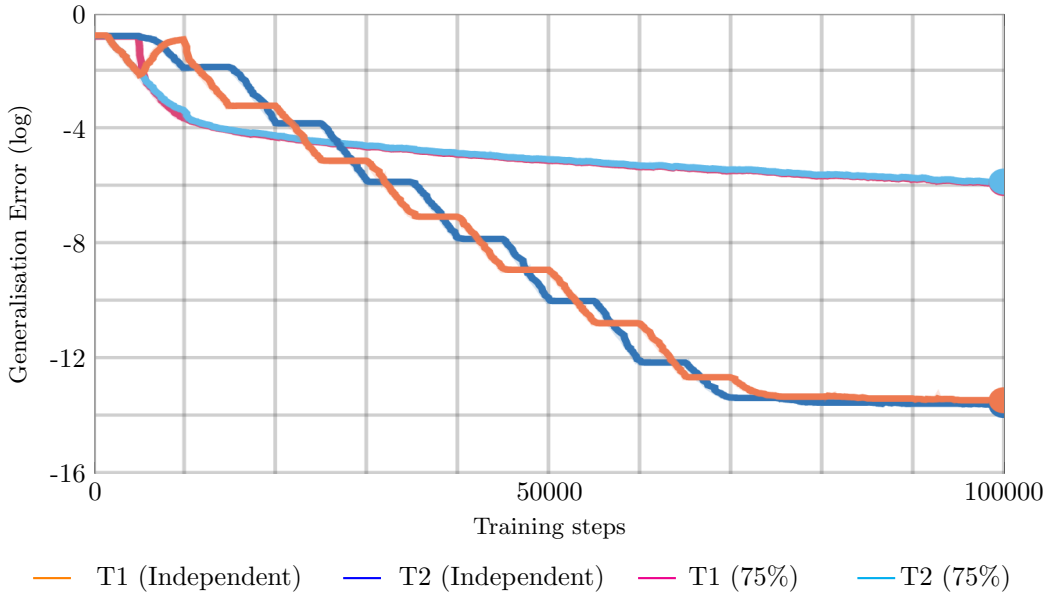
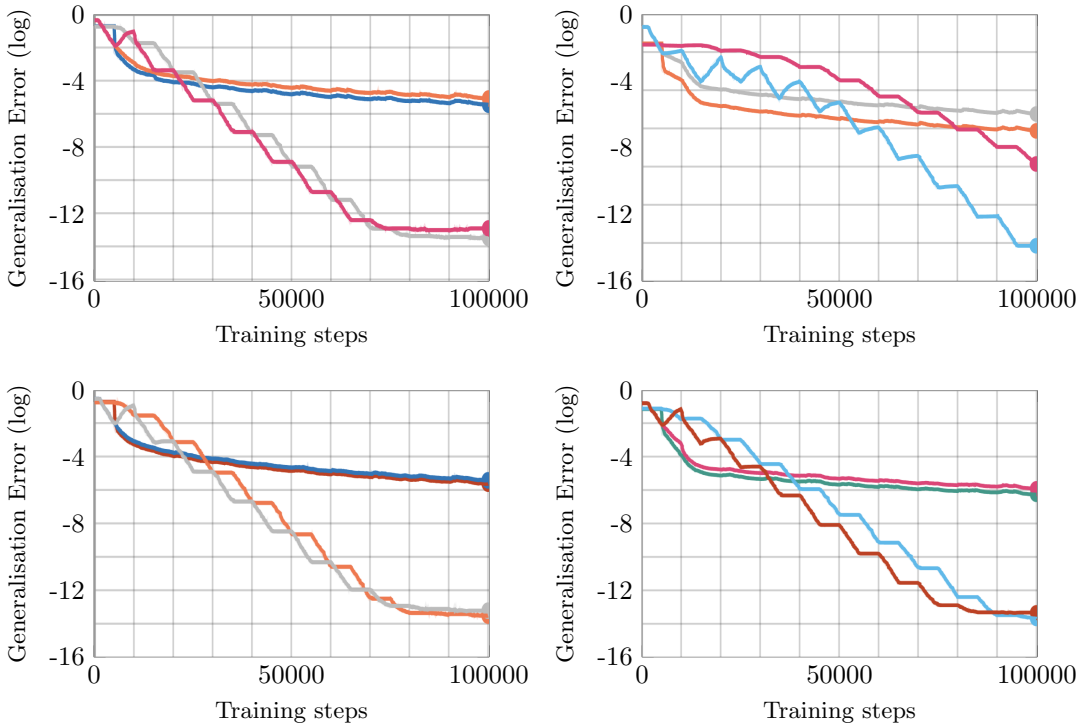


Figure 4: There is a strange phenomenon when the overlap reaches the limit of 100%, which is already hinted at above. While there is no discernible forgetting the overall learning is much slower and saturates.

This phenomenon of saturation with 100% overlap of hidden weights holds for all the random seeds I tested:



1.2 Periodic - Sigmoid

In this setup the teacher is changed with a fixed period of 5000 training steps. I ran the experiment with five different random seeds. This section shows results with the Sigmoid non-linearity. I've only shown the plots for independent and full overlap settings, i.e. the two extremes. Other plots like above are on the google drive. With the sigmoid non-linearity, there is no proper convergence.

1.2.1 Independent Teachers

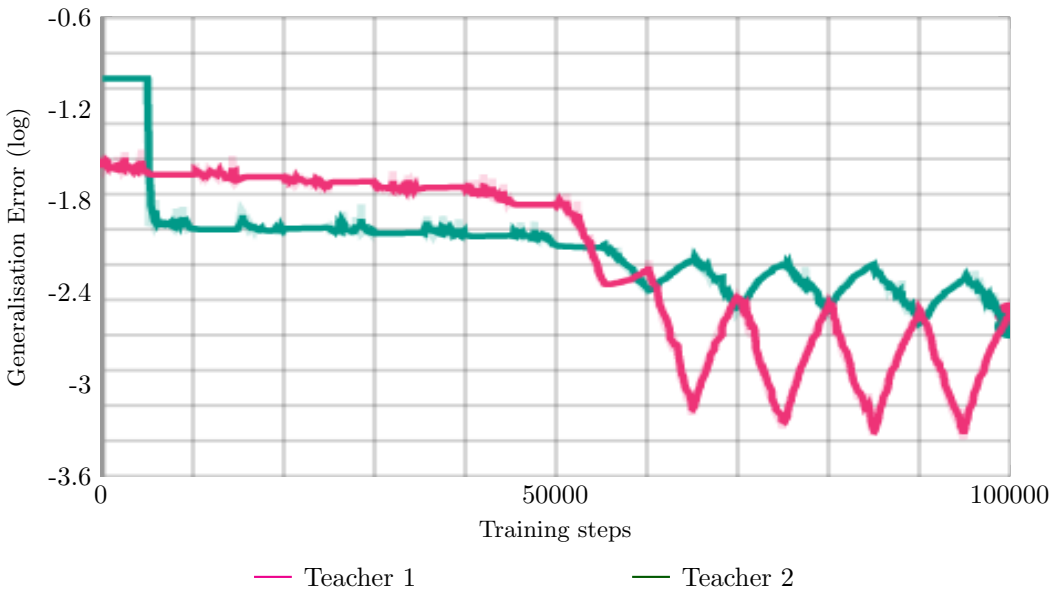


Figure 5

1.2.2 Overlap 100%

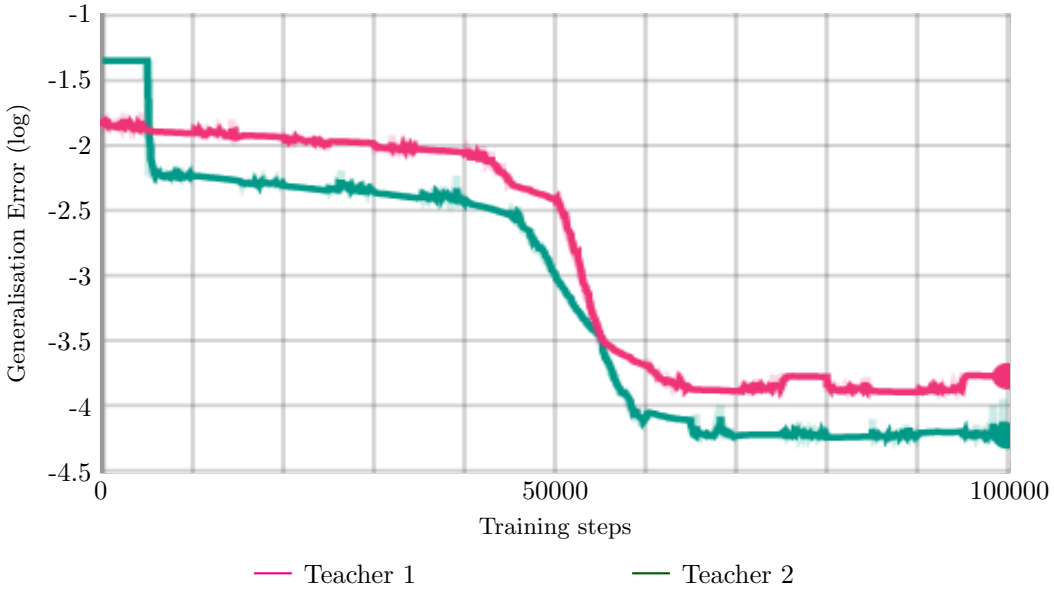


Figure 6

1.3 Periodic - Linear

In this setup the teacher is changed with a fixed period of 5000 training steps. I ran the experiment with five different random seeds. This section shows results without a non-linearity. Again only independent and full overlap shown here.

1.3.1 Independent Teachers

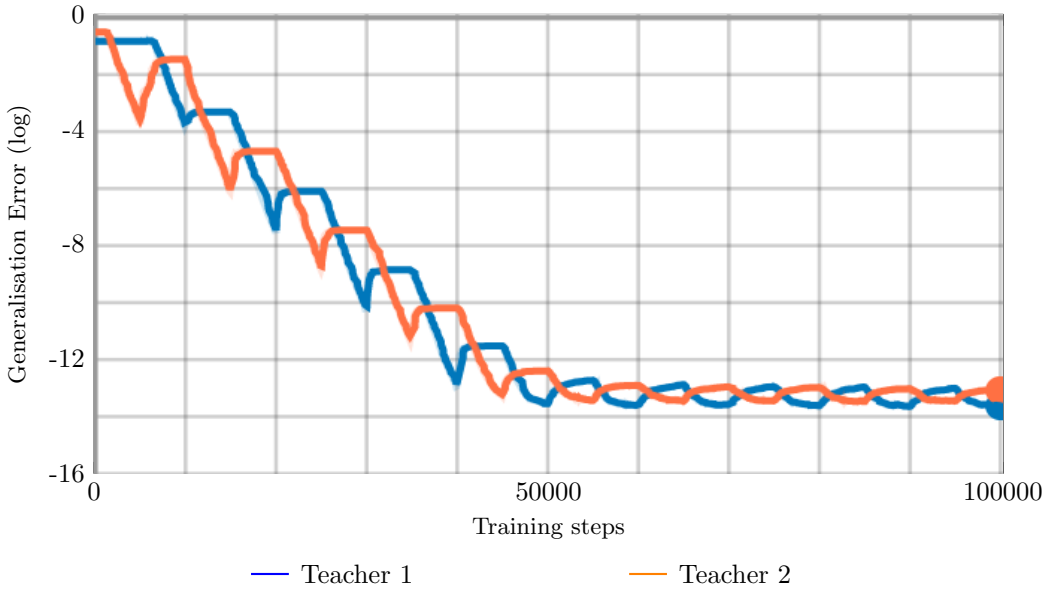


Figure 7

1.3.2 Overlap 100%

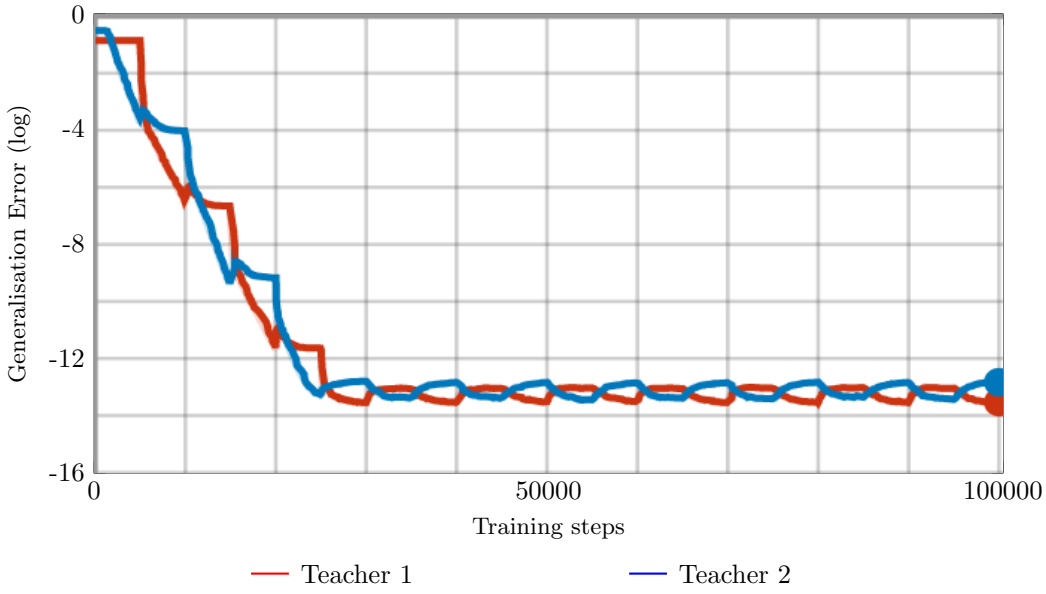


Figure 8

2 Experimental Details

Table 1: Hyperparameters

Hyperparameter	Value
Optimiser	SGD
Loss	MSE
Non-linearity	ReLU
Teacher Architecture	Input: $500 \to 1 \to 1$
Student Architecture	Input: $500 \to 2 \to 1$
Learning Rate	0.2
Test Batch Size	50000
Train Batch Size	1
Teacher Initialisation STD	1
Student Initialisation STD	0.001
Noise	None

Currently an overlap of X% is defined in the following way: Teacher 1 is randomly initialised. Teacher 2 is initialised such that X% of it's hidden weights are exact copies of the corresponding weights in teacher 1, and the rest are also randomly initialised.