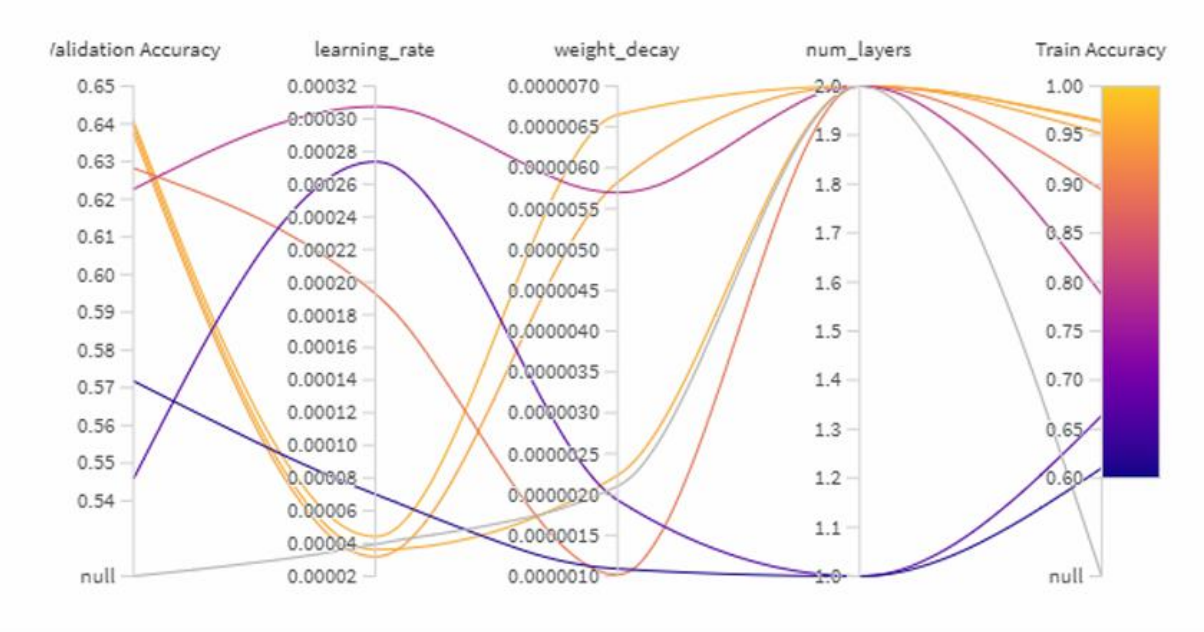
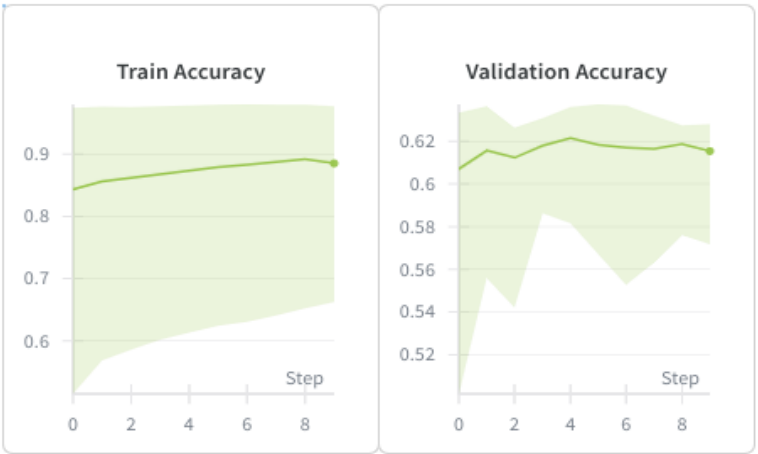


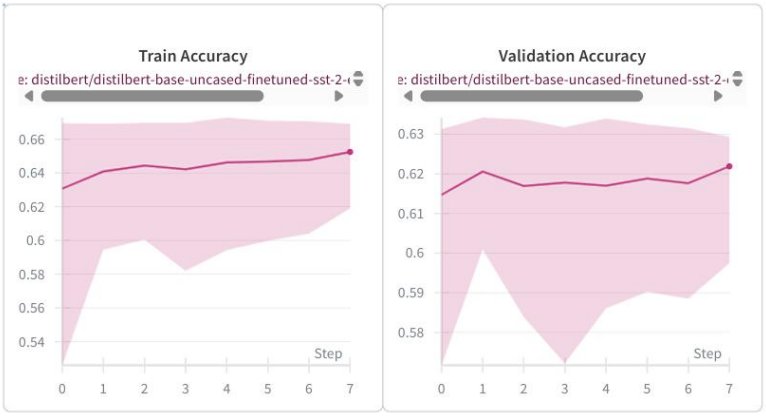
achieved with a smaller **learning rate**, an increased **weight decay**, and a larger number of **layers**. This makes intuitive sense, as training more layers with a smaller learning step allows for a more granular and refined learning process.

However, a trade-off is observed with this configuration: it leads to **overfitting**. The model performs exceptionally well on the training data but fails to generalize to new, unseen data.

Interestingly, the purple line in the graph presents an exception to this trend. It shows good results despite an increased learning rate, which suggests the model is effectively escaping a **local maximum** in the loss function, allowing it to find a better, more generalized solution.



An example of training that led to overfitting over time.



An example of training that did not result in overfitting over time.