Summary

We could observe that the base learning rate matters the most for SGD + momentum. Based on the experiments carried out, SGD + momentum takes us much closer to an optimal local minima, as a result, it tends to improve the generalization performance of our model. The behavior of SGD + momentum could be observed on the train vs epochs graph from the real-time semantic segmentation results. The oscillations combined with a good base learning rate can take us much closer to an optimal local minima improving the generalization performance of our model.

In theory, SGD + momentum trained with a good base learning rate can traverse through the rugged landscape to locate a flatter minima of the loss function. On the other hand, Adam tends to locate a sharper minima i.e. the surrounding parameters near the optimal parameters have varying losses making loss sensitive to the nudge in parameters. Flatter minimas don’t have parameters with varying losses. Moreover, learning rate schedulers in the form of polynomial decays helps SGD + momentum from overshooting or slower convergence. SGD + momentum remains popular amongst CNN researchers and is the optimizer of choice for computer vision tasks [63].

Training models is easier than ever, with the help of higher-level deep learning frameworks and Google Colab we can retrain, fine-tune and evaluate the semantic segmentation models. The modular functionality of deep learning architectures gives us even more incentive to perform complex operations. To conclude, we experimented with state-of-the-art semantic segmentation architectures from the last couple of years with varying optimizers and different learning rates under the constraints of a free VM instance of Google Colab. The thesis contains a detailed observation of the experimental results, the behaviors of the optimizers and the performance of the retrained models on comparison to the baseline.