Summary of the paper "Visualizing the Loss Landscape of Neural Nets"

In this paper the researchers have used variety of visualization methods to explore the shape of loss functions and the effect of loss landscape on generalization. They initiated the exploration by making side-by-side comparison between the loss functions using a simple filter normalization method. Then they applied a range of visualizations to present the effect of network architecture on loss landscape and training parameters on the shape of minimizers.

At first, this study reveals faults in a number of visualization methods for loss functions. 1-Dimensional Linear Interpolation is one of the visualization methods to plot loss function. However, this method has few disadvantages. For example, it is difficult to visualize non convexities using this method and it doesn't consider batch normalization too. To prove the issue, they first trained a CIFAR–10 classifier using a 9-layer VGG network and plot 1-D linear interpolation**.** It shows that the result actually fails to visualize the sharpness of minimizers and doesn't show the consistency between sharpness and generalization error. Another approach was considered to plot the loss function with two direction vectors. This approach also failed to capture non-convexity of loss surfaces because of computational burden hence low resolution.

Secondly, they introduced Filter-Wise Normalization as a solution to the low resolution problem mentioned above. Before plotting loss function using Filter-Wise Normalized direction, they removed Scale Invariance. To obtain such directions, they used Gaussian direction vector d that is dimension compatible with θ. Then they normalized each filter in d to have same norm as θ. For training, they repeated the same classifier as 1-D Linear interpolation. This experiment was sufficient enough to make a visual comparison between minimizers as it gave feasibility to make side-by-side comparison and to show that sharpness correlates well with generalization error when the method is applied. These comparisons were more subtle than how it appeared in un-normalized and layer normalized plots.

Then, they used visualization of the loss surface to show how sufficiently deep neural networks transition from nearly convex to being highly chaotic. To conduct this experiment, they trained a range of networks where they considered three classes of neural networks; ResNets , VGG-like and Wide ResNets. From the contour plotting, it was visible that when skip connection is not applied, as the depth of network increases, the loss surface transitions from convex to chaotic. This is problematic as chaotic landscapes apparently lead to worse training and test error. Whereas, more convex landscapes have lower error values. This answers the questions that why for some neural network architectures non convexity is problematic and why for some architectures it is not. Besides this, the visible partitioning between chaotic and convex regions in the plot shows importance of good initialization strategies. They conducted another experiment to make a comparison between Wide-ResNet-56 on CIFAR-10 (with and without skip connection). From this experiment, they came up with three interesting facts. One, increased network width prevents chaotic behavior. Two, skip connections widen minimizers. Three, sharpness correlates extremely well with test errors. Finally, both of the experiments described above was sufficient enough to strengthen the fact again that filter normalization is a natural way to visualize loss function geometry.

After that, they discussed a way of reassuring non convexity of loss function. This is important as they are dealing with dramatic dimension reduction (1-D, 2-D). The way is to compute principal curvatures where non negative curvatures represent truly convex function and negative curvatures represent non-convex function. They further showed that minimum and maximum eigenvalues of Hessian reveals if the function has any hidden non-convexity.

Finally, they explored methods for visualizing the trajectories of different optimizers. Here they pointed out that random directions method of visualizing trajectories fails as the visualization suffers from orthogonality of random directions in high dimension. As a solution to this issue, they introduced an approach based on Principal Component Analysis (PCA) that captures the measurement of variance and shows a path of low dimension.

This study gives insights into the choice that are made in deep learning research. However, many of the choices are made upon theoretical knowledge and complex assumptions. To make further progress in this era, it is important to understand the structure of the neural network architectures. Effective visualization methods can definitely help doing so.