Visualizing 3D Loss Landscapes for High-Dimensional Machine Learning Models

Shouvik Mani, Hai Pham, Yang Yang

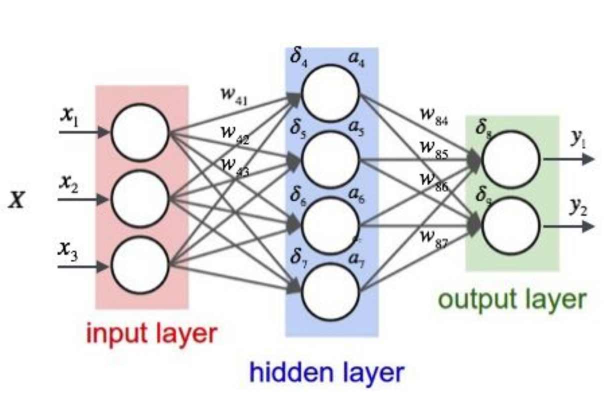
Concept

- The goal of this project is to compute and visualize 3D loss landscapes for high-dimensional models. We propose two methods to reduce the dimensionality of the parameters in a model.
- In both methods, we reduce a model to just two parameters and visualize the loss against those parameters over iterations of gradient descent.
- The resulting 3D loss landscape can be used to diagnose models (e.g. to identify local optima) and also serves as an artistic representation of a complex ML model.

Motivation

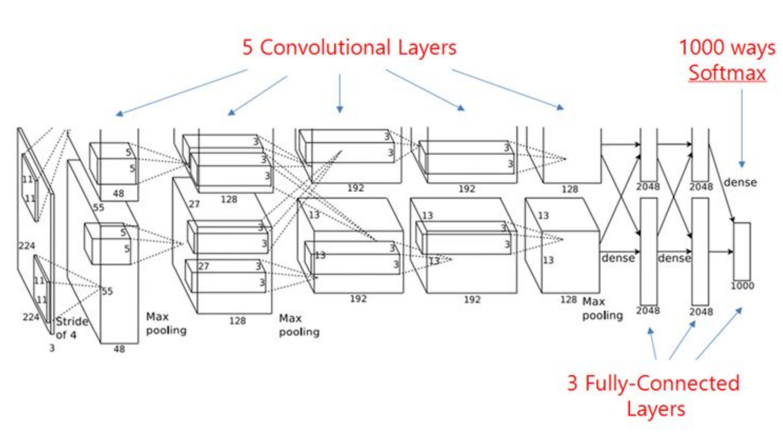
How high-dimensional models like deep neural networks work is still not well understood.

Normal MLP



20 params

AlexNet



62,367,776 params

Literature Review

- In “Visualizing the Loss Landscape of Neural Nets”, Xu et al. introduce a "filter normalization" method to help visualize loss function curvature[1].
- Goodfellow et al. present simple visualization techniques to diagnose whether a neural network is facing local optima.
- Lorch uses PCA to reduce the dimensionality of neural network parameters, but does not investigate visualizing loss landscapes.

References

[1] Li, H., Xu, Z., Taylor, G. and Goldstein, T., 2017. Visualizing the Loss Landscape of Neural Nets. arXiv preprint arXiv:1712.09913.

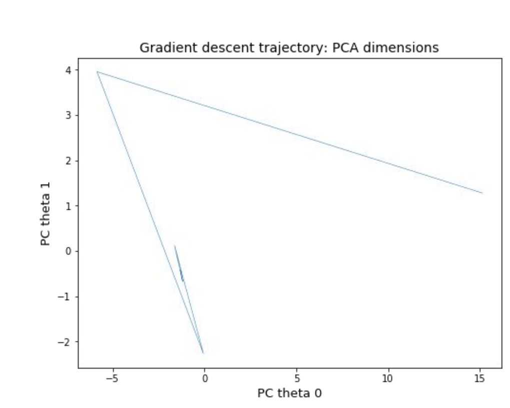
[2] Ian J Goodfellow, Oriol Vinyals, and Andrew M Saxe. Qualitatively characterizing neural network optimization problems. ICLR, 2015.

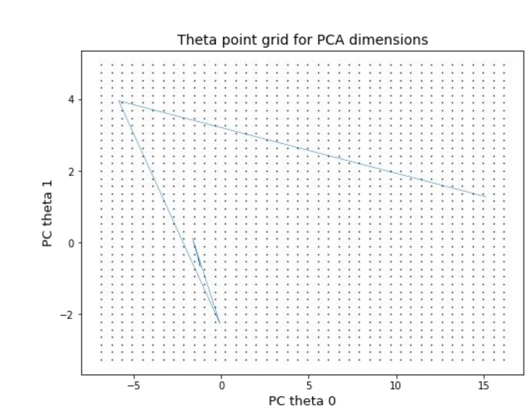
[3] Lorch, E. Visualizing Deep Network Training Trajectories with PCA. IMCL 2016

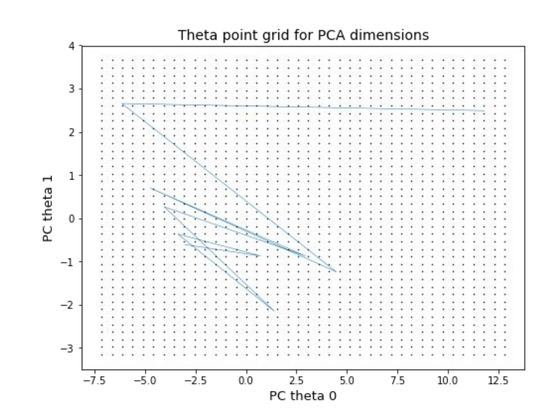
Methodology

- **Approach 1**: Use PCA to reduce the dimensionality of a dataset to two features.
  - Obtain optimal parameters **θ*** through gradient descent
  - Sample a grid of points around **θ*** and plot loss for each point

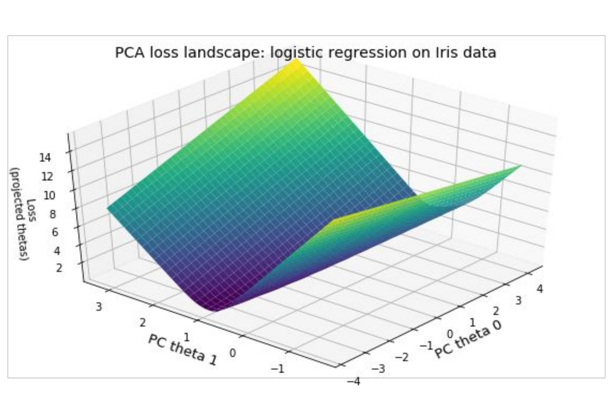
- **Approach 2**: Use PCA to reduce the dimensionality of the parameters of a model during gradient descent.
  - Compute parameters $\theta^{(1)}...\theta^{(t)}$ for t epochs of gradient descent
  - Use PCA to reduce $\theta^{(1)}...\theta^{(t)}$ parameters to two dimensions of highest variance during gradient descent.
  - Sample a grid of points around each parameter $\theta^{(1)}...\theta^{(t)}$. Project sampled points back to the original parameter space through an inverse PCA.

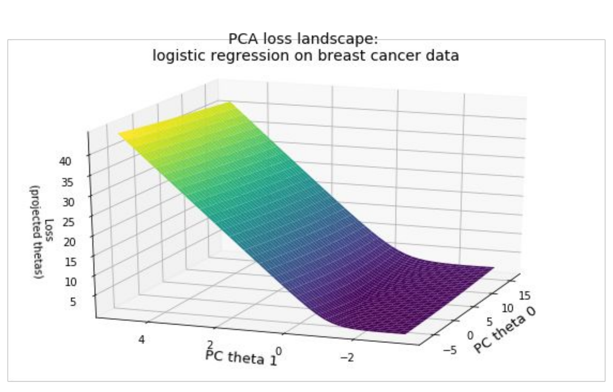


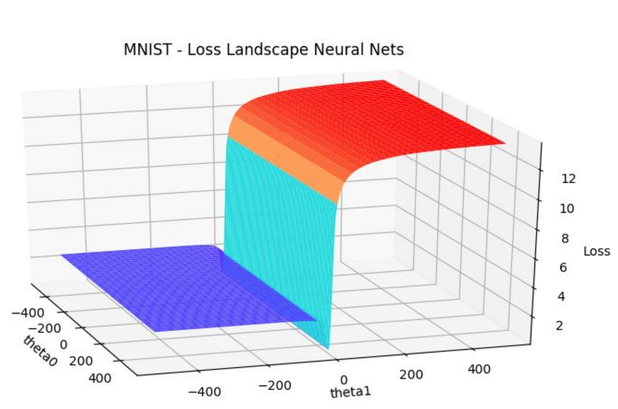




  - Evaluate loss for (projected) sampled points. Plot loss for sampled points against the reduced 2D parameters.







  - Generate Loss Landscapes
    Datasets: IRIS (4D), Breast Cancer (30D), MNIST (784D)
    Models: Logistic Regression, SVM, Neural Network

Result

We deconstructed the high dimensional spaces to a series of metaphorical three-dimensional spaces where a little ball that keeps looking for the lowest point in the loss landscapes with gravity. When the little ball finds the lowest point in one loss landscape, it switches to another loss landscape which constructed by another three dimensions in the high dimensional landscape and keeps dropping to the lowest point. It gives us, human beings, an intuitive idea of how Loss Functions works in the high dimensional spaces of Machine Learning algorithms.