How can we think about the high-dimensional parameter spaces of neural networks? One hypothesis is that the good solutions lie within a hyper-annulus ("Goldilocks Zone"). This project verifies the existence of the Goldilocks Zone when fitting to the CIFAR-10 dataset. It gives simple geometric arguments that explain some of the observed behaviors.

In modern neural networks, the number of parameters $D \sim 10^{13}$ is extremely large. The parameters live in a $D$-dimensional vector space. When a network is trained, it searches for solutions by tracing out a trajectory $\hat{r}(t)$ in parameter-space. A good solution has low loss $\langle F \rangle$ and high generalization accuracy. Some regions of parameter-space may contain more good solutions than others.

Neural networks are heavily overparameterized. It is possible to find good solutions when the training trajectory is restricted to random $d$-dimensional hyperplanes—even if $d < D$.

Further work has demonstrated that the accuracy achieved on a hyperplane depends on the radial distance $r = \| \hat{r} \|$.

For fully connected networks trained on MNIST, good solutions appear to be common in the hyper-annulus $r_1 = W[r_i]$, $|r|^2 = \sum_{a,b}(W_{ab})^2$, where $W(r_i)$ are the weight matrices.

Random $d$-dimensional hyperplane $\hat{r} = \bar{a} + P\theta$, $\theta \in \mathbb{R}^d$.

Random $d$-dimensional hypersphere $\hat{r} = P\theta$, $\|\hat{r}\|_2 = r_0$, $\theta \in \mathbb{R}^d$.

Fix a random $\bar{a} \in \mathbb{R}^d$ and $P \in \mathbb{R}^{d \times d}$ with orthonormal columns, and only train the $d$ parameters $\theta_i$.

The Goldilocks Zone and Geometric Features of High-Dimensional Parameter Spaces

Jeffrey Chang, Department of Physics, Stanford University
jefjiang@stanford.edu

Parameter Space

$D = 656,810$ dimensions!

Accuracy on random $d$-dim 1 hyperplanes

Accuracy on random $d$-dimensional hyperspheres

Radial trajectory of different initializations ($d = D$)

Findings

- Fully-connected neural networks trained on CIFAR-10 exhibit a Goldilocks Zone.
- When initialized at $r_0 < r_1$, the radius grows as $r \propto \sqrt{t}$, and a good solution is found near $r = r_1$ after 15 epochs of training.
- When initialized at $r_0 > r_1$, the radius does not change appreciably over training, and no good solution is found.
- If the radius is constrained to a fixed $r_0$, a good solution can only be found if $r = r_1$.

Some of these observations can be explained by the peculiar properties of high-dimensional spaces.

1. There is much more volume where $r$ is large.

$$\int_0^r d^d r = \int_0^r r^{d-1} dr$$

So all else being equal, regions of large $r$ are more likely to contain solutions.

2. The radius increases along the vast majority of directions in high-dimensional space.

- Consider a random walk $\hat{r}(t) = \hat{r}_0 + \hat{s}$, where $\hat{s}_i \sim \mathcal{N}(0, \sigma^2 t)$.
- Then $\langle \| \hat{r}(t) \|^2 \rangle = r_0^2 + D\sigma^2 t$.

In this light, it is unsurprising that the radius of the training trajectory grows as $r \propto \sqrt{t}$ when $r_0 < r_1$.

Further Questions

- Why are solutions hard to find when $r_0 > r_1$?
- How similar is a training trajectory to a random walk? What about for $d < D$?

Acknowledgements

The author would like to thank Stanislav Fort for inspiring this project and for providing stimulating discussions.

Stanford University