Neural network loss landscape

Maksim Zabelin

Visualizing the Loss Landscape of Neural Nets

Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, Tom Goldstein

Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, Andrew Gordon Wilson

Loss function

$$L(W) = \frac{1}{N}\sum_{i=1}^{N} L_i(f(x_i, W), y_i)$$

- why are we able to minimize highly non-convex neural loss functions?
- why do the resulting minima generalize?
- how loss function geometry affects generalization in neural nets?

The Basics of Loss Function Visualization

1-Dimensional Linear Interpolation

choose two parameter vectors and plot the values of the loss function along the line connecting these two points.

$$\theta(\alpha) = (1-\alpha)\theta + \alpha\theta'$$

Finally, we plot the function

$$f(\alpha) = L(\theta(\alpha))$$

Weaknesses:

- is difficult to visualize non-convexities
- this method does not consider batch normalization or invariance symmetries in the network

The Basics of Loss Function Visualization

Contour Plots & Random Directions

chooses a center point and chooses two direction vectors, then plots a function

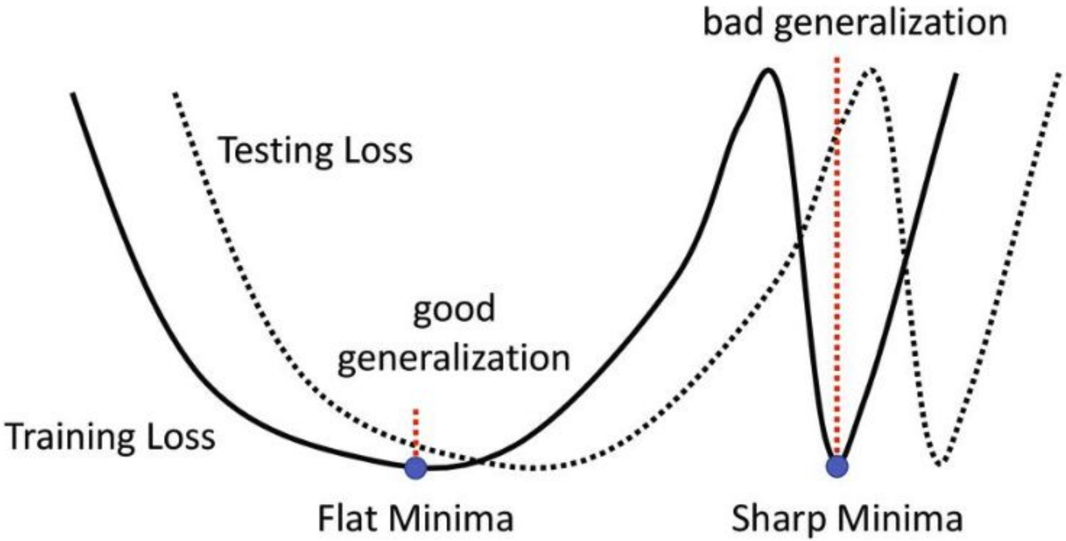
$$f(\alpha) = L(\theta^* + \alpha\delta) \text{ in the 1D (line) case,}$$

$$f(\alpha, \beta) = L(\theta^* + \alpha\delta + \beta\eta) \text{ in the 2D (surface) case}$$

Weaknesses:

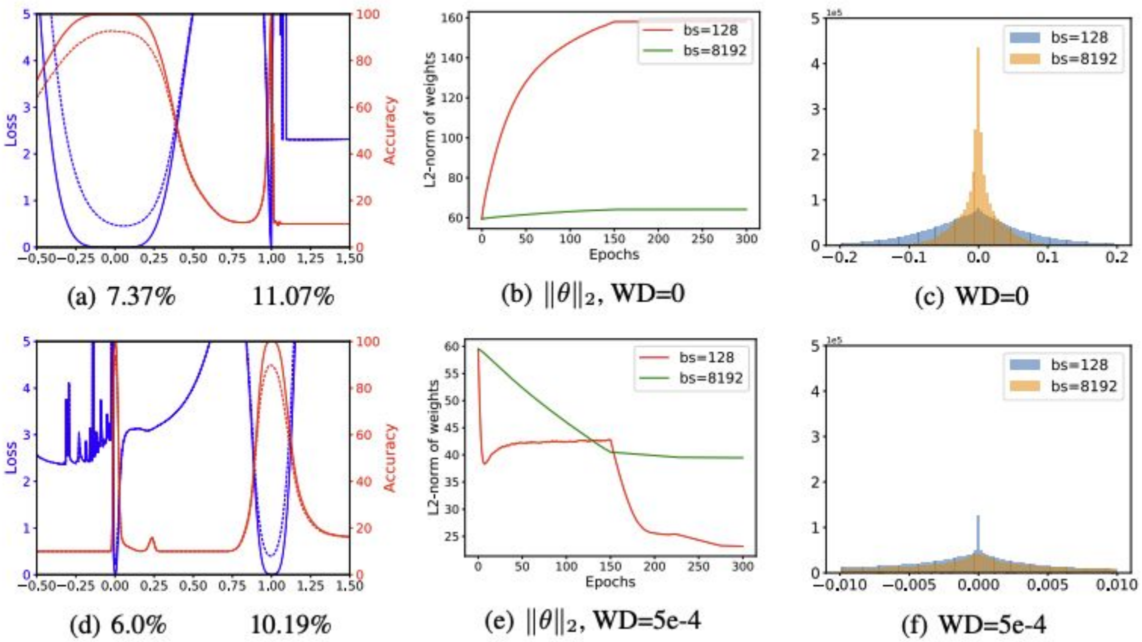
- it fails to capture the intrinsic geometry of loss surfaces
- cannot be used to compare the geometry of two different minimizers or two different networks.

The Sharp vs Flat Dilemma



Large batch – Sharp Minima

The Sharp vs Flat Dilemma Experiment



(a) 7.37% 11.07%

(b) $\|\theta\|_2$, WD=0

(c) WD=0

(d) 6.0% 10.19%

(e) $\|\theta\|_2$, WD=5e-4

(f) WD=5e-4

Scale invariance

$$X \rightarrow C * W1(X) \rightarrow 1/C * W2(C * W1(X)) = W2(W1(X))$$

Batch normalization

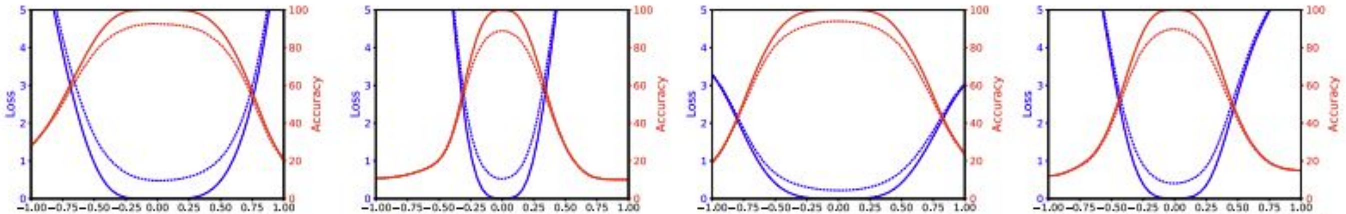
X → C * W1(X) → BN(C * W1(X)) → BN(W1(X))

Filter-Wise Normalization

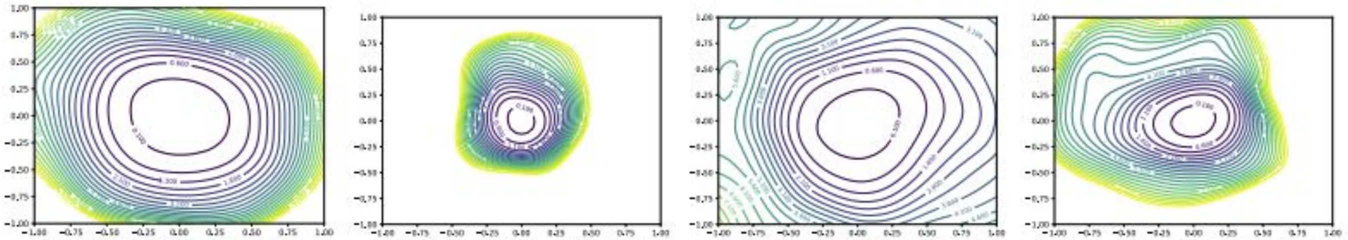
$$d_{i,j} \leftarrow \frac{d_{i,j}}{\|d_{i,j}\|} \|\theta_{i,j}\|$$

where $d_{i,j}$ represents the j-th filter of the i-th layer of d, and $\|\cdot\|$ denotes the Frobenius norm

The Sharp vs Flat Dilemma Experiment

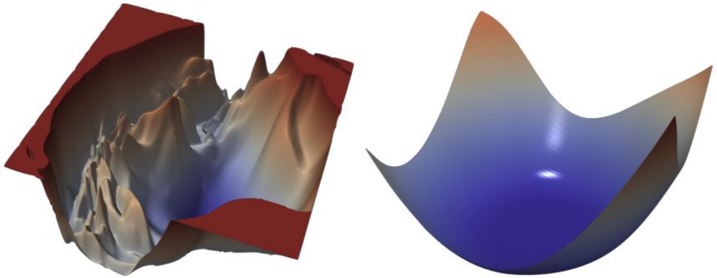


(a) 0.0, 128, 7.37% (b) 0.0, 8192, 11.07% (c) 5e-4, 128, 6.00% (d) 5e-4, 8192, 10.19%

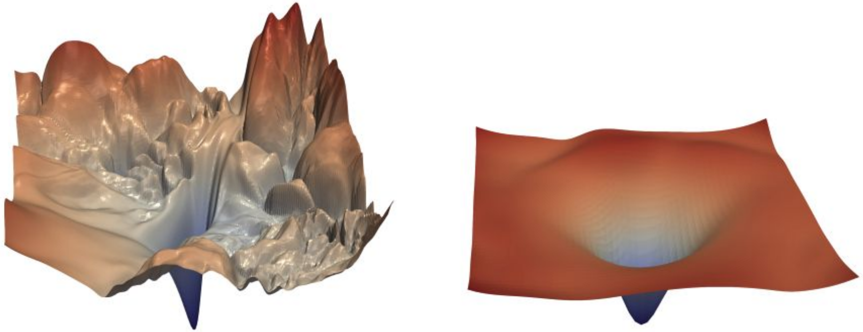


(e) 0.0, 128, 7.37% (f) 0.0, 8192, 11.07% (g) 5e-4, 128, 6.00% (h) 5e-4, 8192, 10.19%

What Makes Neural Networks Trainable?



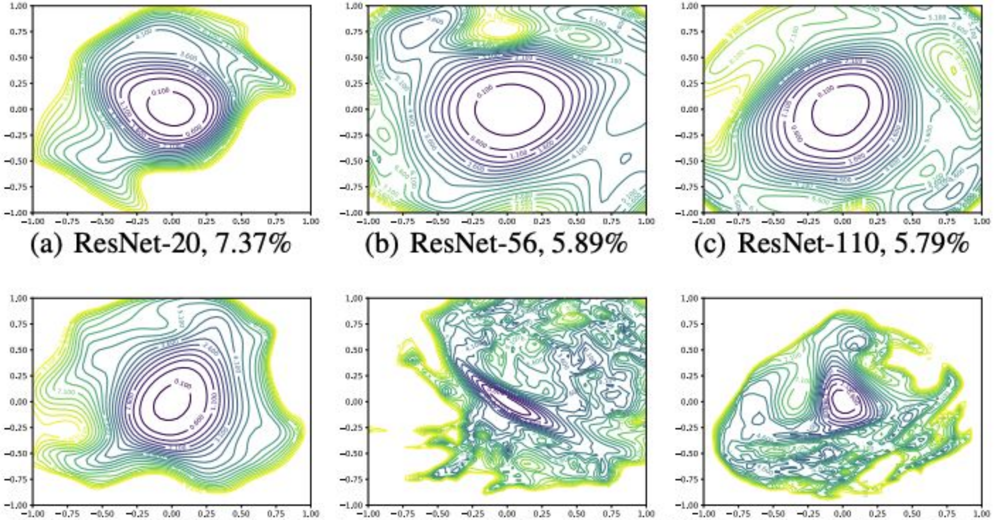
(a) ResNet-110, no skip connections (b) DenseNet, 121 layers



(a) without skip connections (b) with skip connections

- Do loss functions have significant non-convexity at all?
- If prominent non-convexities exist, why are they not problematic in all situations?
- Why are some architectures easy to train, and why are results so sensitive to the initialization?

The Effect of Network Depth

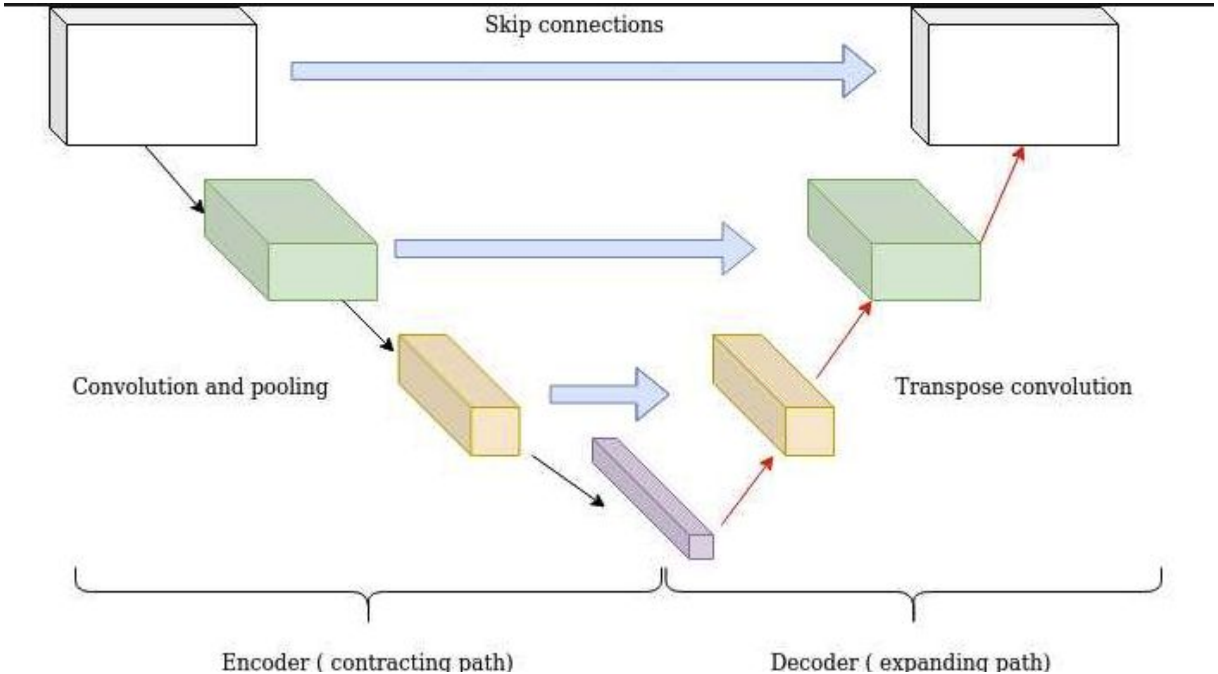


(a) ResNet-20, 7.37% (b) ResNet-56, 5.89% (c) ResNet-110, 5.79%

(d) ResNet-20-NS, 8.18% (e) ResNet-56-NS, 13.31% (f) ResNet-110-NS, 16.44%

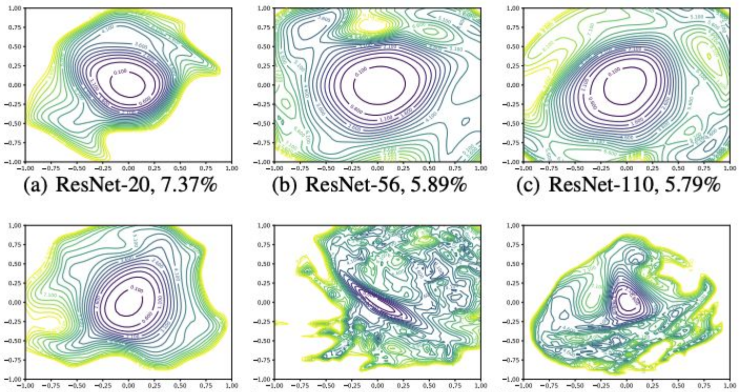
Figure 5: 2D visualization of the loss surface of ResNet and ResNet-noshort with different depth.

Shortcut and skip connections



Shortcut and skip connections

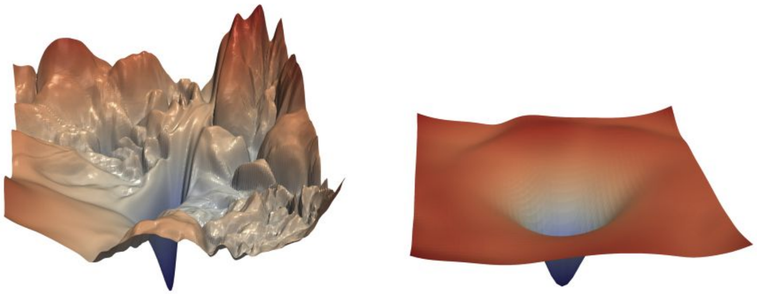
The Effect of Network Depth



(a) ResNet-20, 7.37% (b) ResNet-56, 5.89% (c) ResNet-110, 5.79%

(d) ResNet-20-NS, 8.18% (e) ResNet-56-NS, 13.31% (f) ResNet-110-NS, 16.44%

Figure 5: 2D visualization of the loss surface of ResNet and ResNet-noshort with different depth.



(a) without skip connections (b) with skip connections

Wide Models vs Thin Models

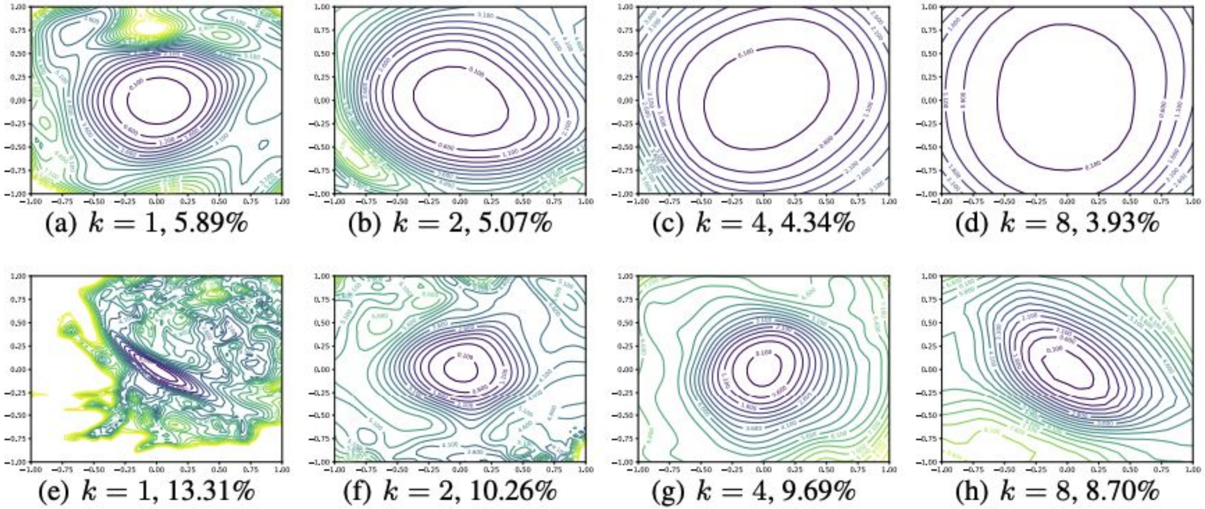


Figure 6: Wide-ResNet-56 on CIFAR-10 both with shortcut connections (top) and without (bottom). The label $k = 2$ means twice as many filters per layer. Test error is reported below each figure.

Are we really seeing convexity?

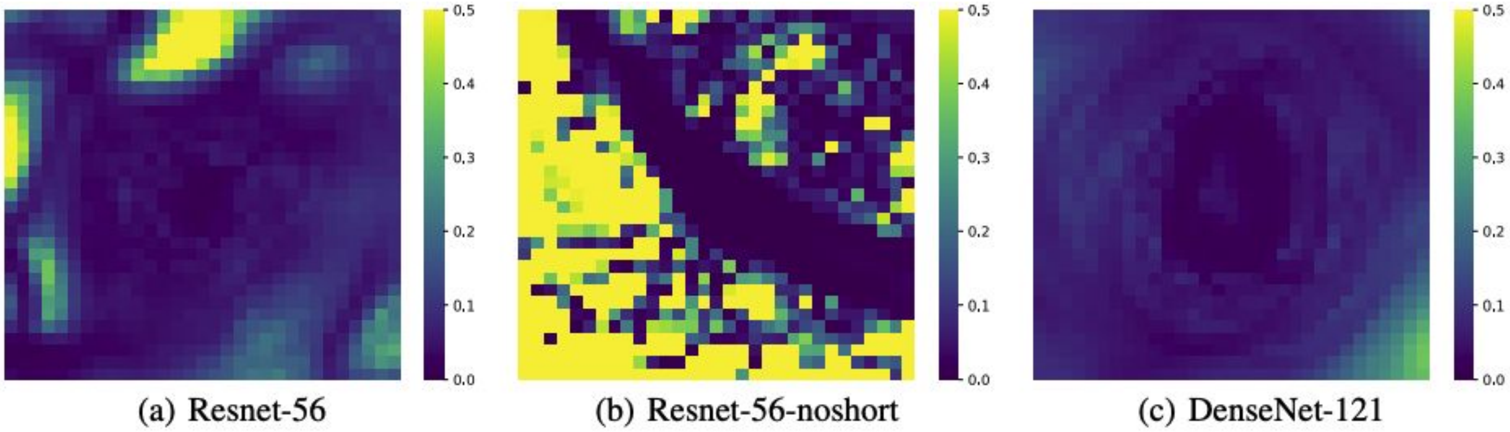


Figure 7: For each point in the filter-normalized surface plots, we calculate the maximum and minimum eigenvalue of the Hessian, and map the ratio of these two.

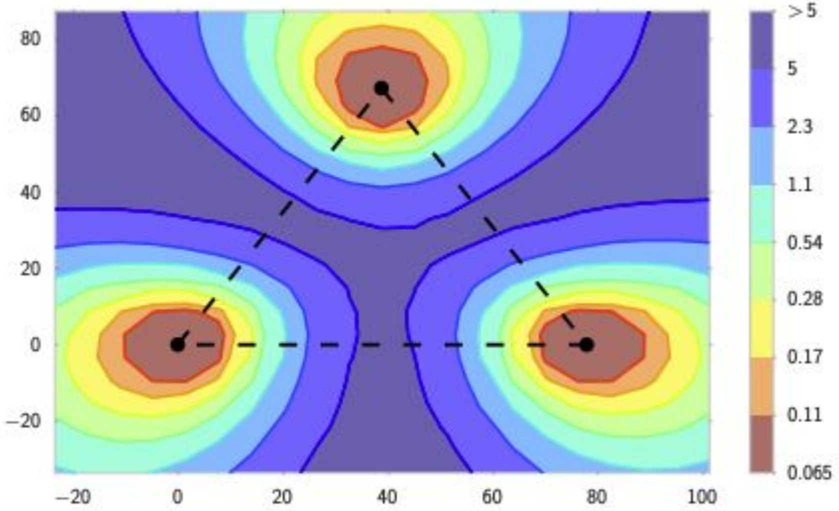
Conclusion

- Flat loss landscape = better generalization
- Shortcut and skip connections affect on trainability
- Deep network = landscape is chaotic
- Wide network = landscape is flatter

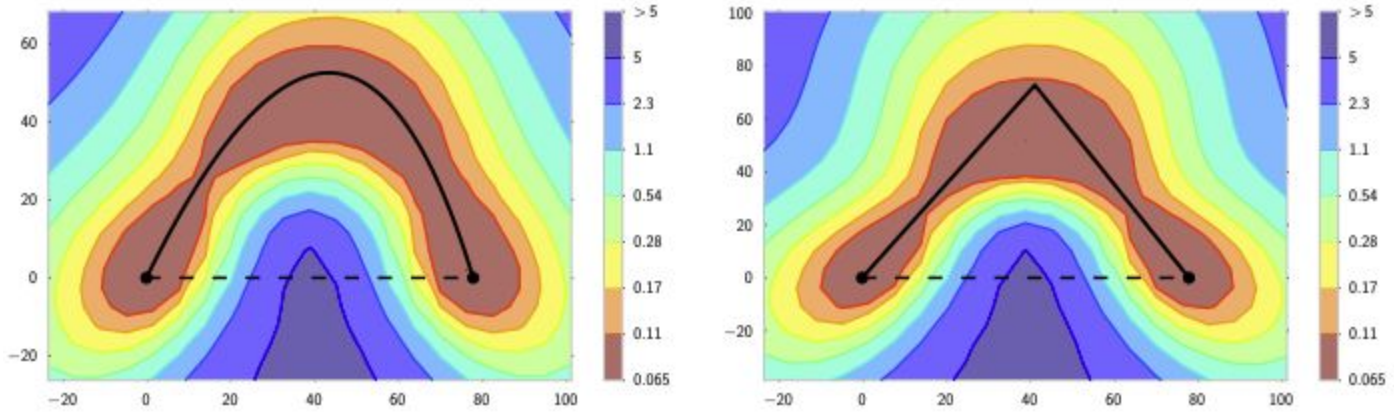
Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, Andrew Gordon Wilson

Finding Paths between Modes



Finding Paths between Modes



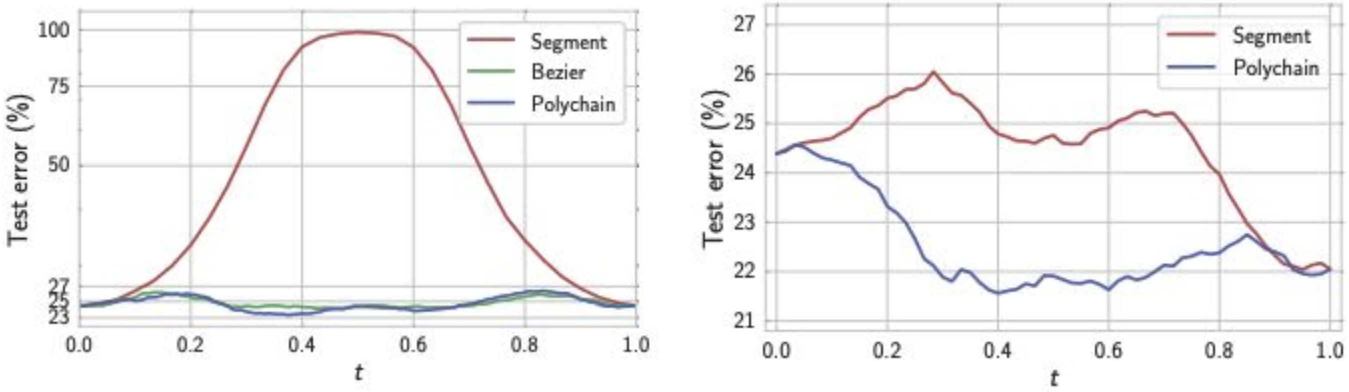
Finding Paths between Modes

two sets of weights $\hat{w}_1$ and $\hat{w}_2$ in $\mathbb{R}^{|net|}$

let $\phi_\theta : [0, 1] \to \mathbb{R}^{|net|}$ be a continuous piecewise smooth such that $\phi_\theta(0) = \hat{w}_1, \;\; \phi_\theta(1) = \hat{w}_2.$

find the parameters θ that minimize $\ell(\theta) = \int_0^1 \mathcal{L}(\phi_\theta(t))dt = \mathbb{E}_{t\sim U(0,1)}\mathcal{L}(\phi_\theta(t))$

Finding Paths between Modes



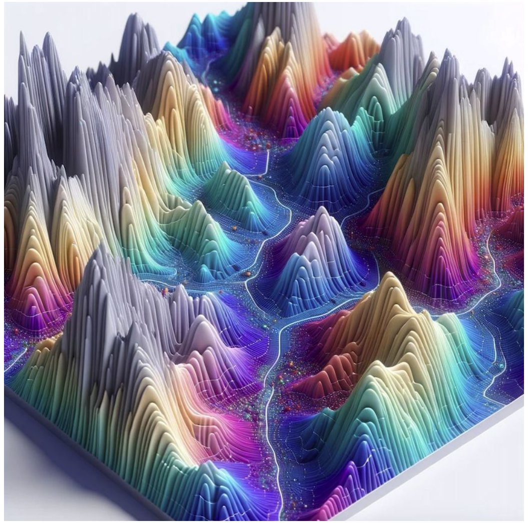
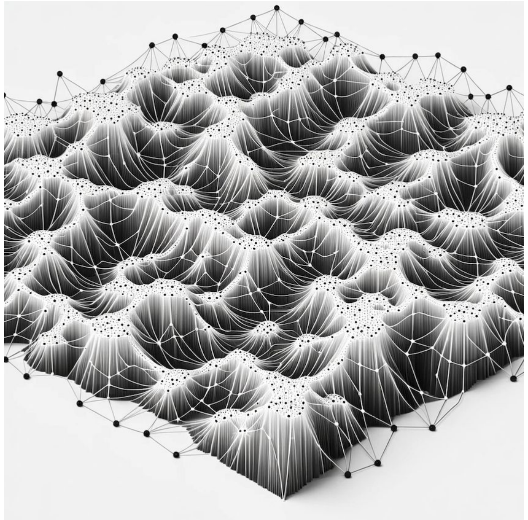
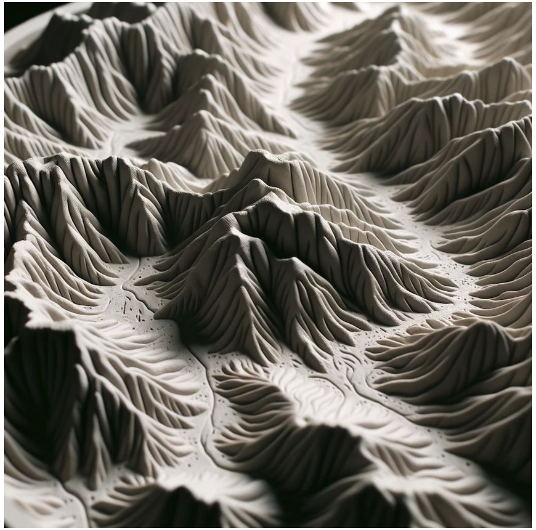
Conclusion

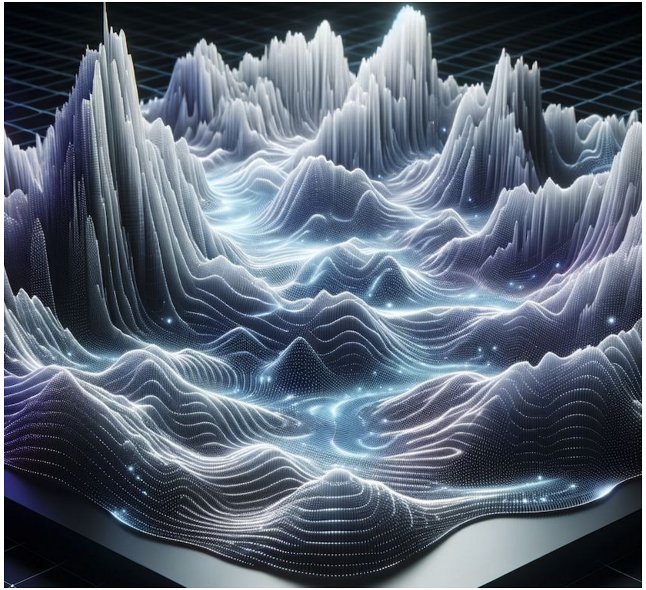
- Local minimas are connected
- This lines not unique
- Can use in the Fast Geometric Ensembling

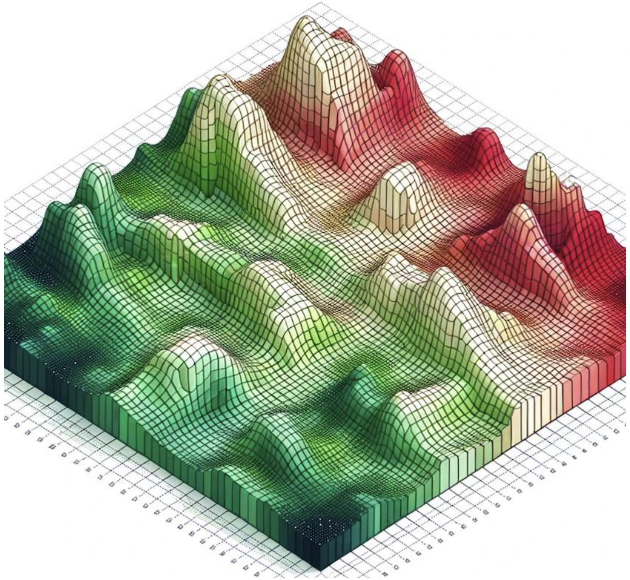
https://proceedings.neurips.cc/paper/2018/hash/a41b3bb3e6b050b6c9067c67f663b915-Abstract.html

https://asset-pdf.scinapse.io/prod/2963384892/2963384892.pdf

Красивые картинки





Здесь могла быть ваша реклама