Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs

Timur Garipov[1,2] Pavel Izmailov[3] Dmitrii Podoprikhin[4]
Dmitry Vetrov[5] Andrew Gordon Wilson[3]

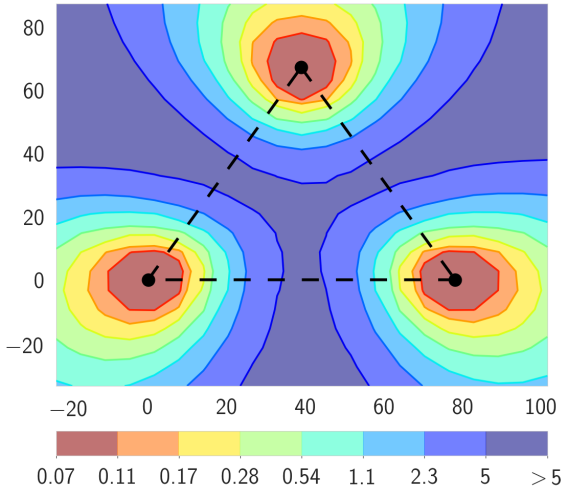
[1]Samsung AI Center in Moscow, [2]Skolkovo Institute of Science and Technology, [3]Cornell University,
[4]Samsung-HSE Laboratory, [5]National Research University Higher School of Economics

Neural Information Processing Systems
Montreal, Canada

December 4, 2018

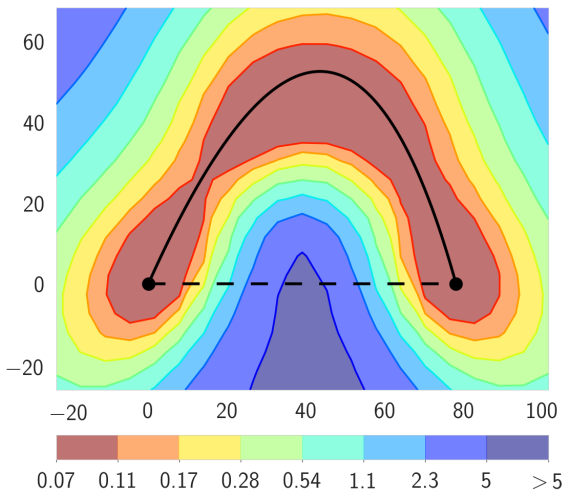
Loss Surfaces

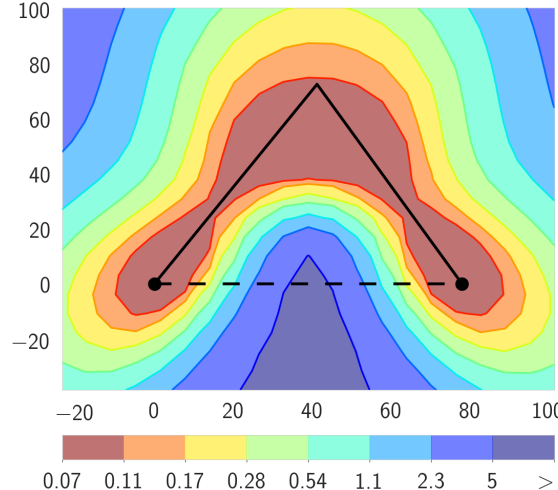
ResNet-164, CIFAR-100



Loss Surfaces

ResNet-164, CIFAR-100





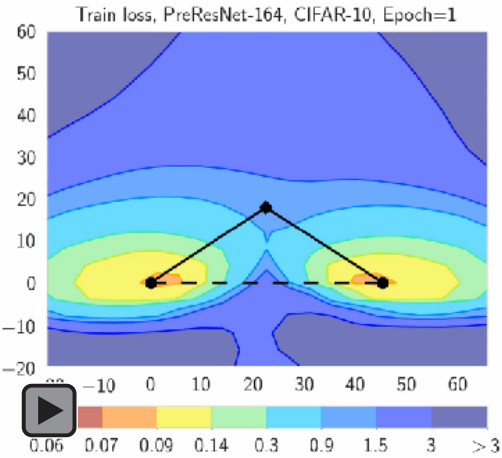
Finding Paths between Modes

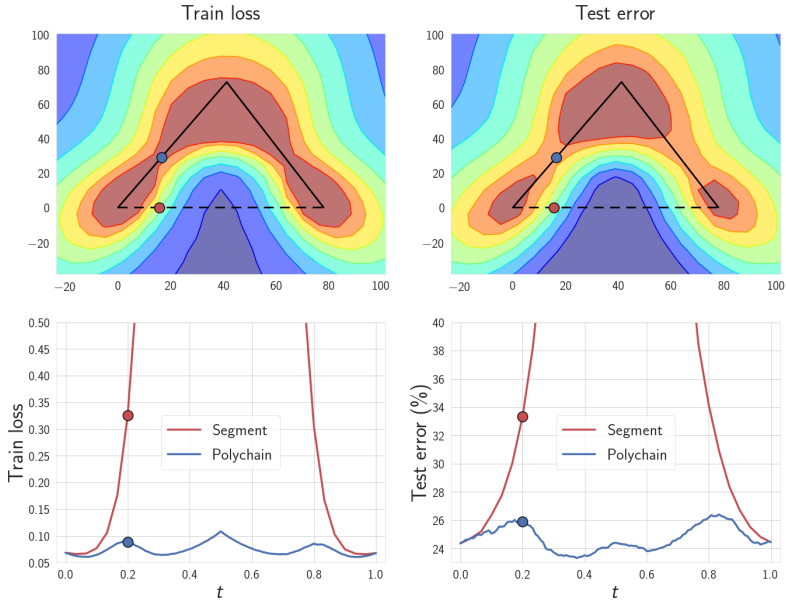
- Weights of pretrained networks: $\widehat{w}_1, \widehat{w}_2 \in \mathbb{R}^{|net|}$
- Define parametric curve: $\phi_\theta(\cdot)\ [0,1] \to \mathbb{R}^{|net|}$

$$\phi_\theta(0) = \widehat{w}_1, \quad \phi_\theta(1) = \widehat{w}_2$$

- DNN loss function: $\mathcal{L}(w)$
- Minimize averaged loss w.r.t. θ

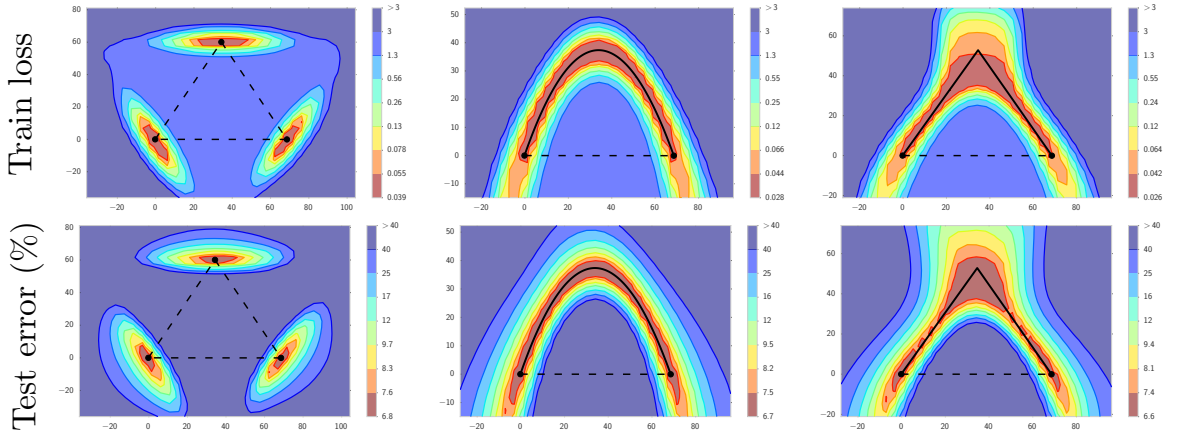
$$\underset{\theta}{\text{minimize}}\ \ell(\theta) = \int_0^1 \mathcal{L}(\phi_\theta(t))dt = \mathbb{E}_{t\sim U(0,1)}\mathcal{L}(\phi_\theta(t))$$

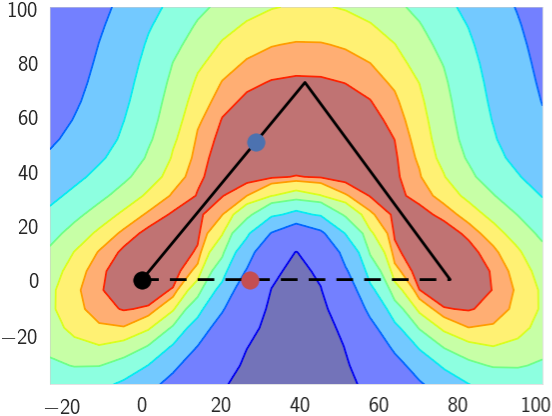


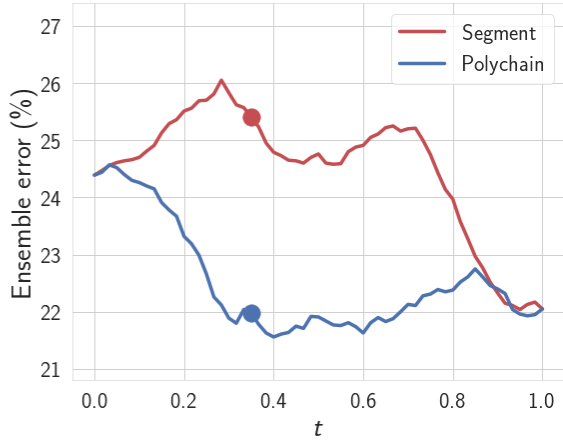

Train loss
Test error
100
80
60
40
20
0
−20
−20
0
20
40
60
80
100
Train loss
Test error (%)
Segment
Polychain
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
40
38
36
34
32
30
28
26
24
0.0
0.2
0.4
0.6
0.8
1.0
t

Loss Surfaces

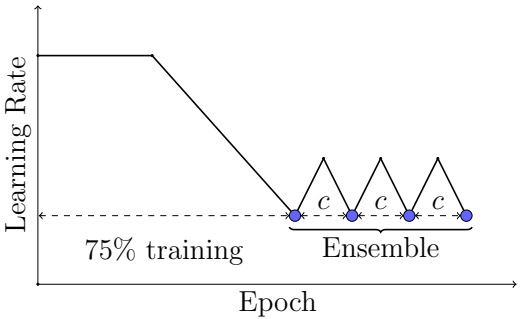
VGG-16, CIFAR-10

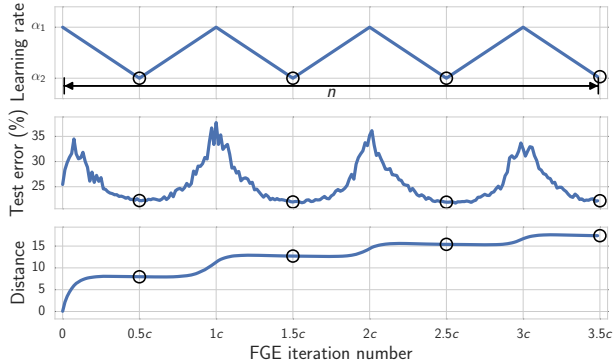



100
80
60
40
20
0
−20
−20
0
20
40
60
80
100

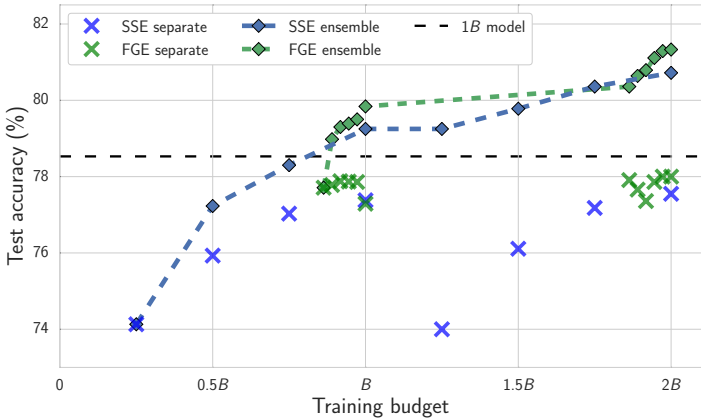

Segment
Polychain
Ensemble error (%)
27
26
25
24
23
22
21
0.0
0.2
0.4
0.6
0.8
1.0
t

Fast Geometric Ensembles (FGE)





Ensembling Results



SSE = Huang et al., ("Snapshot ensembles: Train 1, get m for free"), ICLR 2017

Summary

- Local optima are connected by simple curves.
- To find these curves we minimize loss uniformly in expectation over a path from one mode to another.
- We are inspired by these insights to propose a fast ensembling algorithm.

PyTorch code released for both mode connectivity and FGE

Come to our poster #162!