Outline

- Local optima for modern deep neural networks are connected by very simple curves of near-constant loss and accuracy.
- We propose a simple method to find such curves.
- Mode connectivity holds for a wide range of architectures and hyper-parameter settings, such as batch size, optimizer, and learning rate schedule.
- Inspired by these observations, we propose Fast Geometric Ensembling (FGE). FGE explores the region of low loss and ensembles multiple models from this region.

Loss Surfaces

2D slices of loss surfaces, CIFAR-100, ResNet-164

- Middle and Right: A quadratic Bezier curve, and a polygonal chain with one bend, connecting the lower two optima on the left panel along a path of near-constant loss.

Finding Paths between Modes

- Endpoints: \( \hat{w}_1, \hat{w}_2 \in \mathbb{R}^{[w/c]} \), sets of weights of DNNs
- Loss function: \( \mathcal{L}(w) \)
- Curve parametrization: \( \phi_\theta : [0, 1] \to \mathbb{R}^{[w/c]} \) with parameters \( \theta \)
  \[
  \phi(0) = \hat{w}_1, \quad \phi(1) = \hat{w}_2
  \]
  minimize \( \ell(\theta) = \int_0^1 \mathcal{L}(\phi(t)) dt = \mathbb{E}_{\theta \sim U[0,1]} \mathcal{L}(\phi(t)) \).

Fast Geometric Ensembling

- We run FGE starting from a pretrained model.
- Form ensemble of 4 models in only 5 epochs.
- Achieve 0.36% improvement of top-1 error-rate.

Discussion

- New posterior approximation families for Bayesian deep learning.
- Geometric insights in this paper could be used to accelerate the convergence, stability and accuracy of optimization procedures.

Code

github.com/timgaripov/dnn-mode-connectivity