Diverse optima searching

Project proposal

Ensembling is a popular machine learning technique that helps to improve prediction quality and decrease uncertainty. However, in order to achieve best results with this technique we need to ensemble diverse networks.

The conventional methods of ensembling rely on training multiple models each with a different initialization. Garipov et al. in https://arxiv.org/abs/1802.10026 propose a method called Fast Geometric Ensembling. It allows for creating ensembles using only one trained network with a modest computational overhead. The ensembles produced by this method seem to achieve better quality for a given training budget than a single network or an ensemble of independently-trained networks. Different method for achieving robust model is to average not models predictions but weights themselves (which is equivalent to the conventional ensembling in case of linear regression) described in https://arxiv.org/pdf/1803.05407 by the same cohort of authors.

First method relies on the assumption that local minima of the loss function for complex networks are connected via curves of relatively low loss values – a hypothesis that was shown to hold true for a variety of architectures and loss functions but so far lacking a rigorous proof.

Second method constructs an ensemble by averaging weights that occur while training for multiple epochs without loss improvement, but traversing through above-mentioned low-loss valleys whose centers tend to have the most flat surface. It is shown in experiments that models with parameters at low curvature regions tend to generalize better.

Our hypothesis is that there exist local optima that aren’t connected by such curves. Moreover, we believe that creating ensembles of networks corresponding to these ‘disjoint’ optima will lead to better diversity thus improving accuracy and reducing uncertainty.

For this project we want to:

1. Implement the curve-fitting algorithm described in the paper.

2. Find the local optima that cannot be connected by low-loss curves.

3. Perform an all-round comparison between these disjoint ensembles, conventional ensembles and ensembles produced by FGE and SWA.

All of the experiments will be conducted using CIFAR-10 dataset and a ResNet or VGG architecture.