MaxUp: A Simple Way to Improve Generalization of Neural Network Training

Seungmin Lee (profile2697@gmail.com; 2020-20866),

Dept. of Electrical and Computer Engineering, Seoul National University

1. Introduction

This paper proposes a simple training strategy called MaxUp. The authors claim that MaxUp is embarrassingly simple and easy to apply while it still enhances performance. The central intuition behind MaxUp is similar to that of adversarial training, which is regularizing networks by inducing them to lie on a smooth loss landscape. MaxUp implements the smoothness regularization by generating a set of randomly augmented samples to calculate the worst-case loss for a given image and letting networks minimize the worst-case loss.

2. Method

Empirical Risk Minimization (ERM) is the standard scheme for training a regular neural network:

$$\min_{\theta} \mathbb{E}_{x \sim \mathcal{D}_n} \left[L(x; \theta) \right], \tag{1}$$

where θ and L denote the parameters of the network and the loss function, respectively, and $\mathcal{D}_n$ indicates the dataset while x is a sampled data from $\mathcal{D}_n$. While ERM is simple, the authors argue that ERM often incurs networks' overfitting.

To ameliorate overfitting, the authors propose MaxUp, which induces the loss landscape around a sample to be smoother. More specifically, MaxUp randomly augments a given image to produce a set of transformed images. Then, MaxUp selects the worst-case loss among the losses calculated with the transformed images. Finally, MaxUp minimizes the worst-case loss as follows:

$$\min_{\theta} \mathbb{E}_{x \sim \mathcal{D}_n} \left[\max_{i \in [m]} L(x'_i; \theta) \right]. \tag{2}$$

where x'_i denotes the i-th augmented sample. Eq. 2 is similar to the adversarial training in that both schemes first maximize loss using an augmentation and minimizes the maximized loss. According to the paper, both training methods share similar smoothing effect.

3. Experiments

The paper conducts experiments on various datasets such as ImageNet, CIFAR10, and CIFAR100. Used network architectures include ResNet and EfficientNet. Across all the settings, MaxUp + CutMix shows slightly improved performance compared to Dropout, MixUp, and CutMix.

4. Personal Opinion

The proposed method requires multiple forward passes, which limits the usefulness of MaxUp. Moreover, MaxUp seems neither novel nor effective.