Catastrophic Fisher Explosion: Early Phase Fisher Matrix Impacts Generalization

Stanislaw Jastrzebski[1], Devansh Arpit[2], Oliver Astrand[1], Giancarlo Kerg[3], Huan Wang[2], Caiming Xiong[2], Richard Socher[2], Kyunghyun Cho[*,1], Krzysztof Geras[*,1]

The implicit regularization effects of using a large learning rate in SGD can be introduced explicitly by regularizing the trace of the Fisher Information Matrix.

Introduction

Implicit regularization in gradient-based training of deep neural networks (DNNs) remains relatively poorly understood despite being considered a critical component in their empirical success

Our main contribution is to show that the implicit regularization effect of using a large learning rate or a small batch size can be modeled as an implicit penalization of the trace of the FIM.

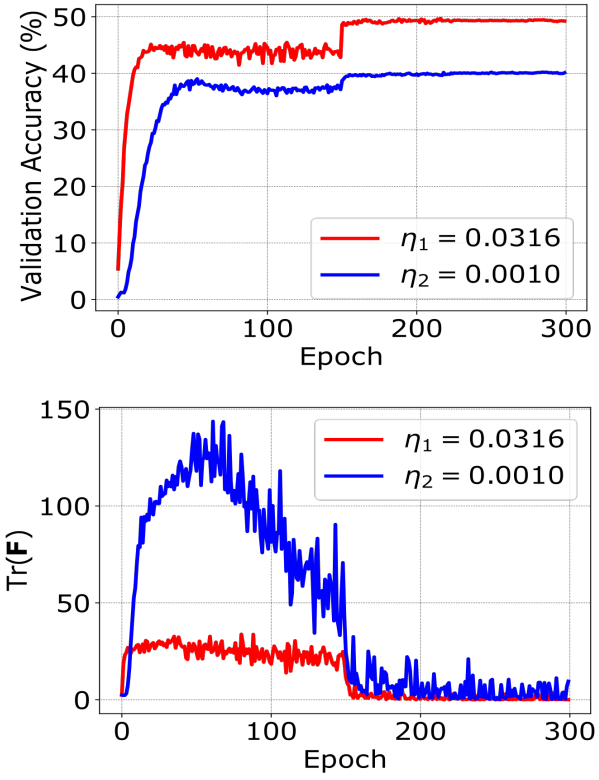

Figure 1. The catastrophic Fisher explosion phenomenon demonstrated for Wide ResNet trained using SGD on the TinyImageNet dataset. Training is done with either a learning rate optimized using grid search, or a small learning rate. Training with the latter leads to large overfitting (top) and a sharp increase in Tr(F) (bottom).

Fisher Penalty

To regularize Tr(F), we add the following term to the loss function:

$$\ell'(\mathbf{x}_{1:B}, y_{1:B}; \theta) = \frac{1}{B}\sum_{i=1}^{B} \ell(\mathbf{x}_i, y_i; \theta) + \alpha \left\| \frac{1}{B}\sum_{i=1}^{B} g(\mathbf{x}_i, \hat{y}_i) \right\|^2$$

Fisher Penalty resembles implicit regularization in SGD

First, we use a learning rate 10-30x smaller than the optimal one, which incurs up to 9% degradation in test accuracy and results in large value of Tr(F).

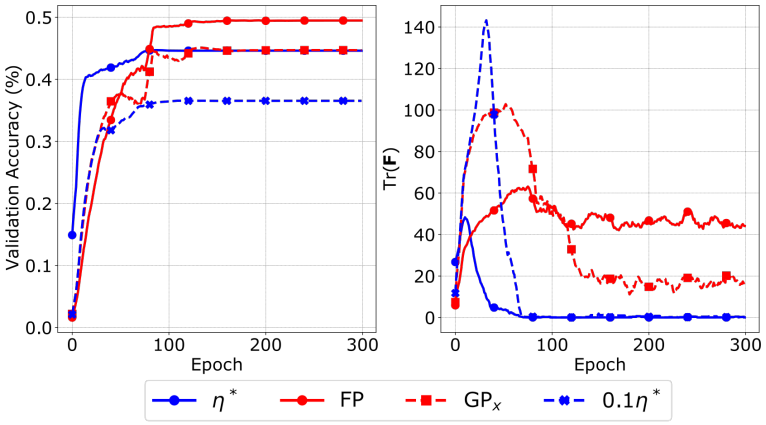

Figure 2. Fisher Penalty applied to VGG11 trained on CIFAR-100. **Left**: Training using small learning rate and Fisher Penalty can result in better generalization than training with large learning rate. **Right**: Using Fisher Penalty results in achieving smaller Tr(F).

Fisher Penalty reduces memorization

We also investigated the impact of penalizing Tr(F) on learning on data with noisy labels.

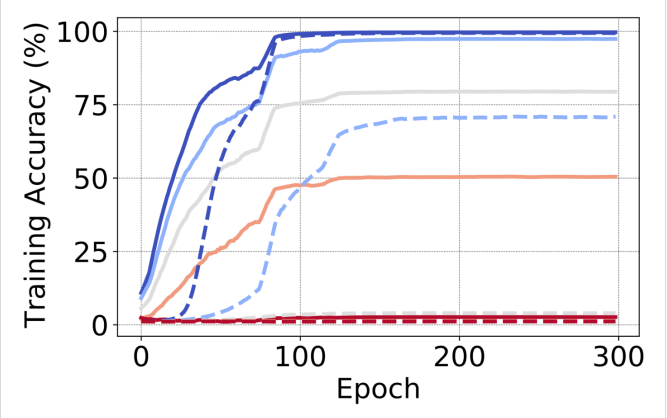

Figure 3. Fisher Penalty applied to VGG11 trained on CIFAR-100 with 20% random labels. Training speed on noisy examples (dotted line) is disproportionally more affected by Fisher Penalty than training speed on clean examples (solid line).