Masterarbeit

Jessica Bühler

6. Oktober 2019

Übersicht

1 Kernel Rescaling

2 Parameter Pruning and Sharing

3 Low Rank Factoriazation

4 Transferred/compact convolutional Filters

5 Knowledge distillation

Fast Training Of Convolutional Neural Networks Via Kernel Rescaling (20% less training time)

Training deep Convolutional Neural Networks (CNN) is a time consuming task that may take weeks to complete. In this article we propose a novel, theoretically founded method for reducing CNN training time without incurring any loss in accuracy. The basic idea is to begin training with a pre-train network using lower-resolution kernels and input images, and then refine the results at the full resolution by exploiting the spatial scaling property of convolutions. We apply our method to the ImageNet winner OverFeat and to the more recent ResNet architec- ture and show a reduction in training time of nearly 20% while test set accuracy is preserved in both cases.

Übersicht

1 Kernel Rescaling

2 Parameter Pruning and Sharing

3 Low Rank Factoriazation

4 Transferred/compact convolutional Filters

5 Knowledge distillation

Übersicht

2 Parameter Pruning and Sharing

- Quantization and Binarization
- Pruning and Sharing
- Structural Matrix

Drawbacks

the accuracy of the binary nets is significantly lowered when dealing with large CNNs such as GoogleNet. Another drawback of such binary nets is that existing bina- rization schemes are based on simple matrix approximations and ignore the effect of binarization on the accuracy loss.

To address this issue, the work in [16] proposed a proximal Newton algorithm with diagonal Hessian approximation that directly minimizes the loss with respect to the binary weights. The work in [17] reduced the time on float point multiplication in the training stage by stochastically binarizing weights and converting multiplications in the hidden state computation to significant

Übersicht

changes.

2 Parameter Pruning and Sharing

- Quantization and Binarization
- Pruning and Sharing
- Structural Matrix

Übersicht

2 Parameter Pruning and Sharing

- Quantization and Binarization
- Pruning and Sharing
- Structural Matrix

Übersicht

1 Kernel Rescaling

2 Parameter Pruning and Sharing

3 Low Rank Factoriazation

4 Transferred/compact convolutional Filters

5 Knowledge distillation

Übersicht

1 Kernel Rescaling

2 Parameter Pruning and Sharing

3 Low Rank Factoriazation

4 Transferred/compact convolutional Filters

5 Knowledge distillation

Übersicht

1 Kernel Rescaling

2 Parameter Pruning and Sharing

3 Low Rank Factoriazation

4 Transferred/compact convolutional Filters

5 Knowledge distillation