JDS

Principal Component Networks: Utilizing Low-Rank Activation Structure to Reduce Parameters Early in Training

ROGER WALEFFE, University of Wisconsin-Madison, USA
THEODOROS REKATSINAS*, ETH Zurich, Switzerland

PROBLEM STATEMENT

Many recent results show that large neural networks can lead to improved generalization. Yet, training these large models comes with increased computational costs. In an effort to address this issue, several works have show that large networks contain subnetworks that can be trained in isolation to nearly the same accuracy as the full model. Existing approaches for subnetwork identification, however, face two main challenges: they are either 1) computationally expensive and lead to sparse architectures that can not be executed efficiently on modern hardware or 2) produce dense architectures that are hardware friendly but lead to lower accuracy compared to the full network. These challenges limit the practical potential for end-to-end training cost reductions. Motivated by this, we focus on finding an efficient procedure for discovering small, dense subnetworks contained in large neural networks that can train to the same accuracy as the full model.

METHODS

We study network activations, rather than network weights, and observe that hidden layer activations in large networks exist primarily in low-dimensional subspaces an order of magnitude smaller than the actual model width. We also find that these subspaces can be identified early in training. Based on these observations, we show that after only a few training epochs, large networks can be efficiently transformed into small, dense networks—which we term Principal Component Networks (PCNs)—that exhibit comparable generalization performance. PCNs compress individual layers by first using Principal Component Analysis to find the high-variance directions that describe the layer's input and output activations; then only the linear combinations of weights defined by the found high-variance subspaces are used to construct the transformed PCN layers.

RESULTS

We empirically validate training PCNs on CIFAR-10 and ImageNet for VGG and ResNet style architectures. PCNs consistently reduce parameter counts of large models with little accuracy loss. For example, on ImageNet, PCNs can reduce the parameters of a wide ResNet-50 from 98M to 62M *without any accuracy loss*. Our results also show that 1) PCNs achieve higher end-model accuracy than existing structured pruning methods and 2) partially training and then compressing wide models may lead to a more resource-efficient method for obtaining high-accuracy versus standard training of deep networks.

SIGNIFICANCE

The observations and methods described in our work offer the potential to reduce the computational costs of high-accuracy deep neural network training. Moreover, we also believe that our observation regarding hidden layer activations may have implications for improved understanding of the feature representations learned by neural networks.

Keywords: model pruning, compression, activations, sparsity, resource-efficient training

Authors' addresses: Roger Waleffe, waleffe@wisc.edu, University of Wisconsin-Madison, Madison, Wisconsin, USA; Theodoros Rekatsinas, theo.rekatsinas@inf.ethz.ch, ETH Zurich, Zurich, Switzerland.

Research Article

*Currently at Apple