Università della Svizzera italiana
Faculty of Informatics
Institute of Computational Science ICS

Pruning Highway Networks For Reducing Memory Usage

Dhananjay Tomar

Faculty of Informatics, Universitá della Svizzera italiana (Lugano)

Introduction

Motivation

While neural networks are getting better with time, they are also getting bigger. Size of ResNet-152 is 240 Mb and that of VGG16 is 552 Mb. This makes it difficult to port them to mobile devices which have less storage capacity, memory and processing power available.

Running a model on cloud means taking care of

- **User Privacy**: Data received from user has privacy issues to take care of.
- **Bandwidth Issues**: High bandwidth is not available everywhere in the world.
- **Budget**: Depending on the number of users, servers tend to be costly.

Highway Networks

Highway nets make it possible to train very deep networks via skip connections and achieve near SOTA results. For an input $\boldsymbol{x}$, a highway layer besides computing a non-linear transform $\boldsymbol{H}$ on $\boldsymbol{x}$ as $\boldsymbol{y} = \boldsymbol{H}(\boldsymbol{x}, \boldsymbol{W_H})$, also computes an additional non-linear transform $\boldsymbol{T}(\boldsymbol{x}, \boldsymbol{W_T})$. Output $\boldsymbol{y}$ of a highway layer is then given by

$$\boldsymbol{y} = \boldsymbol{H}.\boldsymbol{T} + (1 - \boldsymbol{T}).\boldsymbol{x}$$

T is called transform gate and it controls how much information is transformed with $\boldsymbol{H}$ and how much information from the previous layer is carried forward.

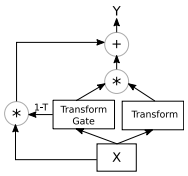

Figure 1: Highway Layer

Architecture Used

The architecture follows the architecture of wide residual networks with width 4. The first hidden layer is a convolutional layer having 16 filters. The rest of the hidden layers are highway blocks divided into 3 stages having 6 blocks each.

- Each block in stage 1 has 64 filters.
- Each block in stage 2 has 128 filters.
- Each block in stage 3 has 256 filters.

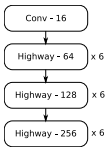

Figure 2: Architecture of the whole network

Each highway block consists of a batch normalization layer, followed by the transform layer and transform gate layer.

The first highway block in each stage uses stride of size 2 in transform layer to down-sample image.

We provide $\boldsymbol{H}(\boldsymbol{x}, \boldsymbol{W_H})$ as input to the transform gate layer instead of $\boldsymbol{x}$ which means that when $\boldsymbol{H}$ is pruned, the corresponding weights in $\boldsymbol{T}$ can also be pruned.

Dataset Used

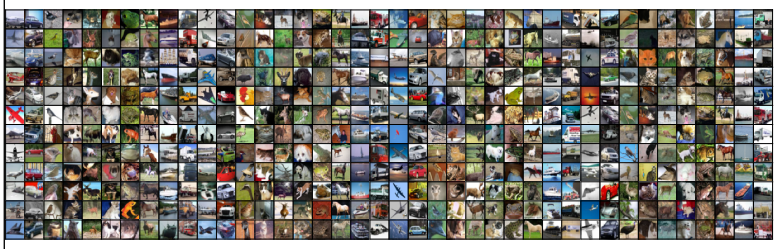

Figure 3: Some random images from CIFAR-10

"The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images."[3]

The 10 classes are - airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck. "The classes are completely mutually exclusive. There is no overlap between automobiles and trucks. 'Automobile' includes sedans, SUVs, things of that sort. 'Truck' includes only big trucks. Neither includes pickup trucks."[3]

Method

- The idea is to incentivize the network to use less transformations.
- Once the network learns to not rely much on certain nodes, they can be pruned. More concretely if the value of $\boldsymbol{T}$ is very low for a node, then the corresponding node in $\boldsymbol{H}$ can be pruned along with $\boldsymbol{T}$ because lower $\boldsymbol{T}$ means output $\boldsymbol{y}$ mostly consists of $\boldsymbol{x}$.
- We select a threshold value for $\boldsymbol{T}$ below which it is pruned.

Results

We train the network on CIFAR-10 dataset and are able to prune the network by about **85%**, reducing it's size from 34 Mb to 5.2 Mb. This massive reduction however comes with a cost of 1.72% loss in accuracy.

The proposed approach is orthogonal to previously proposed approaches for compressing networks and therefore highway network's size can be further reduced by using those approaches on top the proposed approach.

Heat maps for transform gates

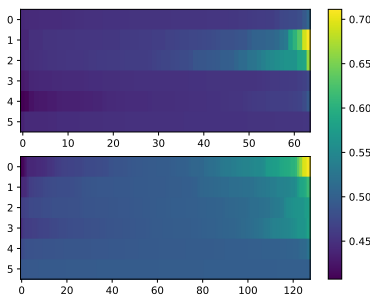

Figure 4: When network **IS NOT** incentivized to use less transforms

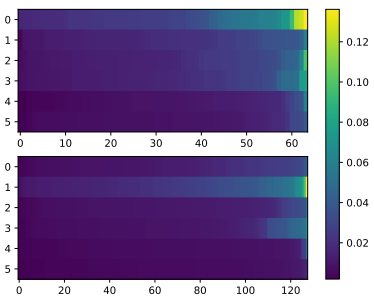

Figure 5: When network **IS** incentivized to use less transforms

Graph for results

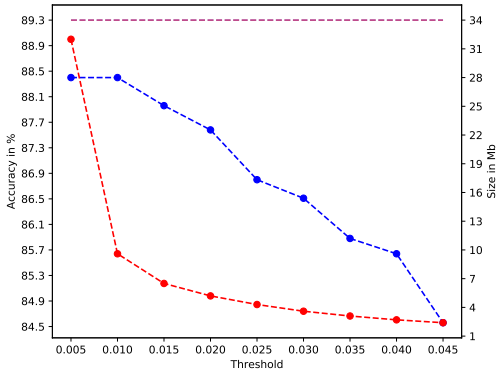

Figure 6: Accuracy (blue) and size (red) for different threshold values

References

[1] Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In NIPS, 2015.
[2] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In BMVC, 2016.
[3] https://www.cs.toronto.edu/ kriz/cifar.html