Layer Folding: Neural Network Depth Reduction using Activation Linearization

MOTIVATION

- As deep neural networks become more prevalent, their applicability to resource-constrained devices is limited.

- While modern devices exhibit a high level of parallelism, real-time latency is still highly dependent on networks’ depth.

- Recent works show the width of shallower networks must grow exponentially below a certain depth. However, we presume that neural networks usually exceed this minimum depth to accelerate convergence and incrementally increase accuracy.

- This motivates us to transform pre-trained deep networks that already exploit such advantages into shallower forms.

OBJECTIVES

- Reduce the depth of a pre-trained network with minimal impact on accuracy.

- Provide more efficient alternatives to MobileNet and EfficientNet architectures on the classification task.

- Explore the accuracy-depth and accuracy-latency trade-offs.

METHOD

Removing activations (non-linearities) allows us to merge consecutive linear layers into a single layer. Thus, we focus on removing activations as a method to reduce depth.

We replace each activation $\sigma$ with its learnable parametric counterpart:

$$\sigma_a(x) = ax + (1 - a)\sigma(x)$$

When $a = 0$ we get the original activation, when $a = 1$ we get the identity (essentially removing the activation).

We use an auxiliary loss to encourage each $a$ to become 1:

$$L_c = \sum_{i \in L} (1 - a^2_i)$$

ACCURACY ~ DEPTH

Layer Folding applied on ResNet, VGG, and MobileNetV2 (MNV2) architectures on CIFAR-10 (left) and CIFAR-100 (right). For each network, we gradually remove nonlinear layers.

ACCURACY ~ LATENCY

Latency and FLOPs reduction obtained by applying Layer Folding on MobileNetV2 (MNV2) and EfficientNet (EffNet) on ImageNet.