Rethinking the Usage of Batch Normalization and Dropout in the Training of Deep Neural Networks

Aliaksei Rak

Generating independent components

Theorem 1:

Let g_i denote a family of independent random variables generated from Bernoulli distribution with its mean being p. Let $\overline{x_i} = g_i x_i$.

Then we have:

- $I(\hat{x}_i; \hat{x}_j) = p^2 I(x_i; x_j)$
- $H(\hat{x}_i) = pH(x_i) + \epsilon_p$

Correlation coefficient:

- $\hat{c}_{ij} = pc_{ij}$

Residual connections

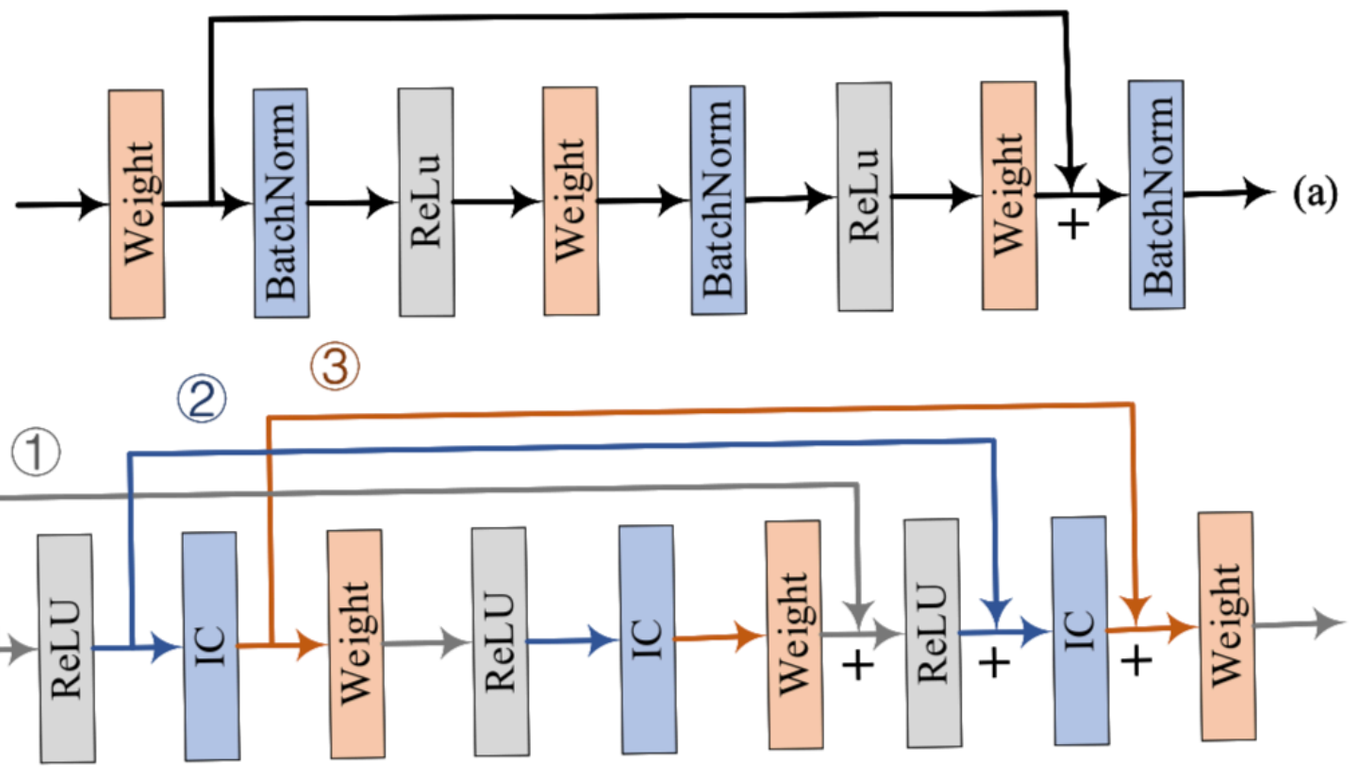


Figure 3. (a) The classical ResNet architecture, where $'+'$ denotes summation. (b) Three proposed ResNet architectures reformulated with the IC layer.

Results

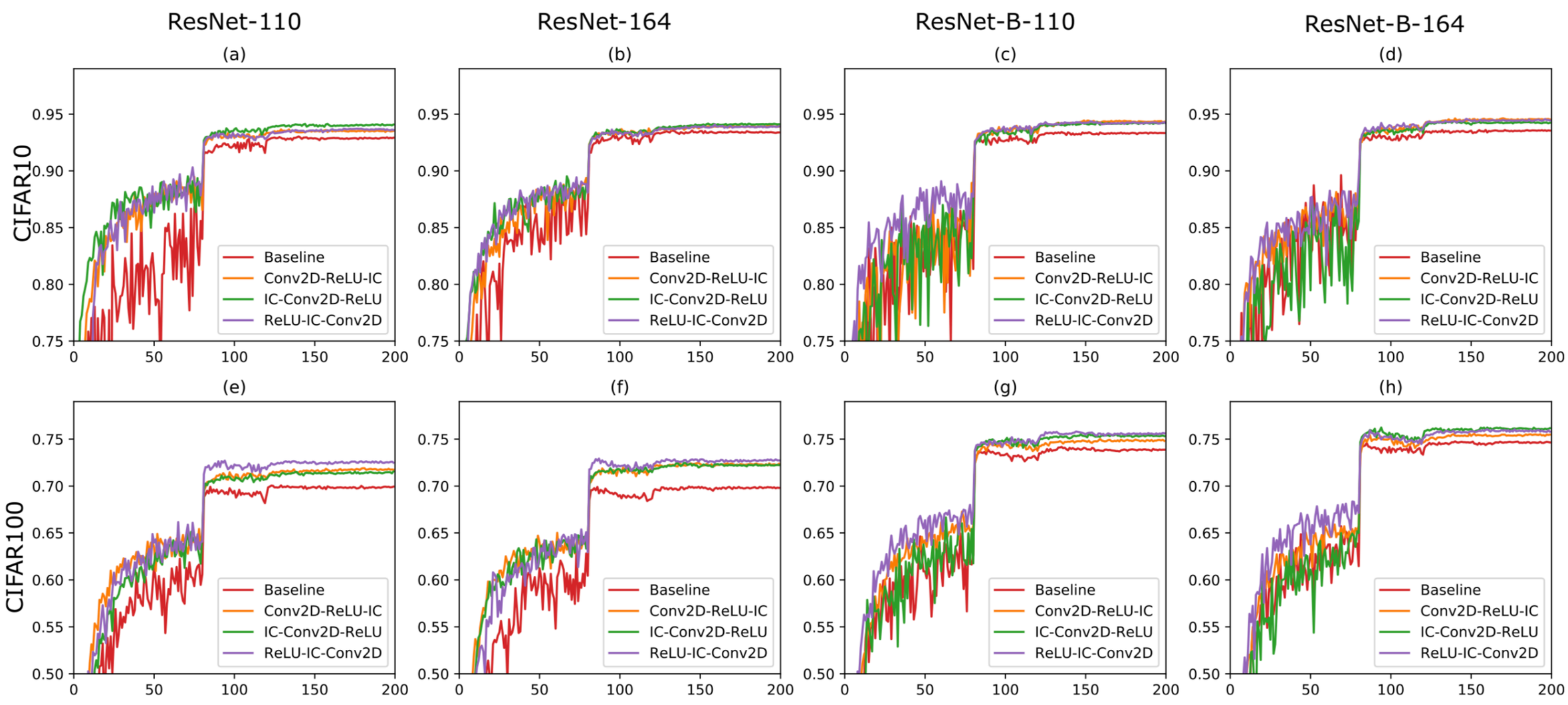
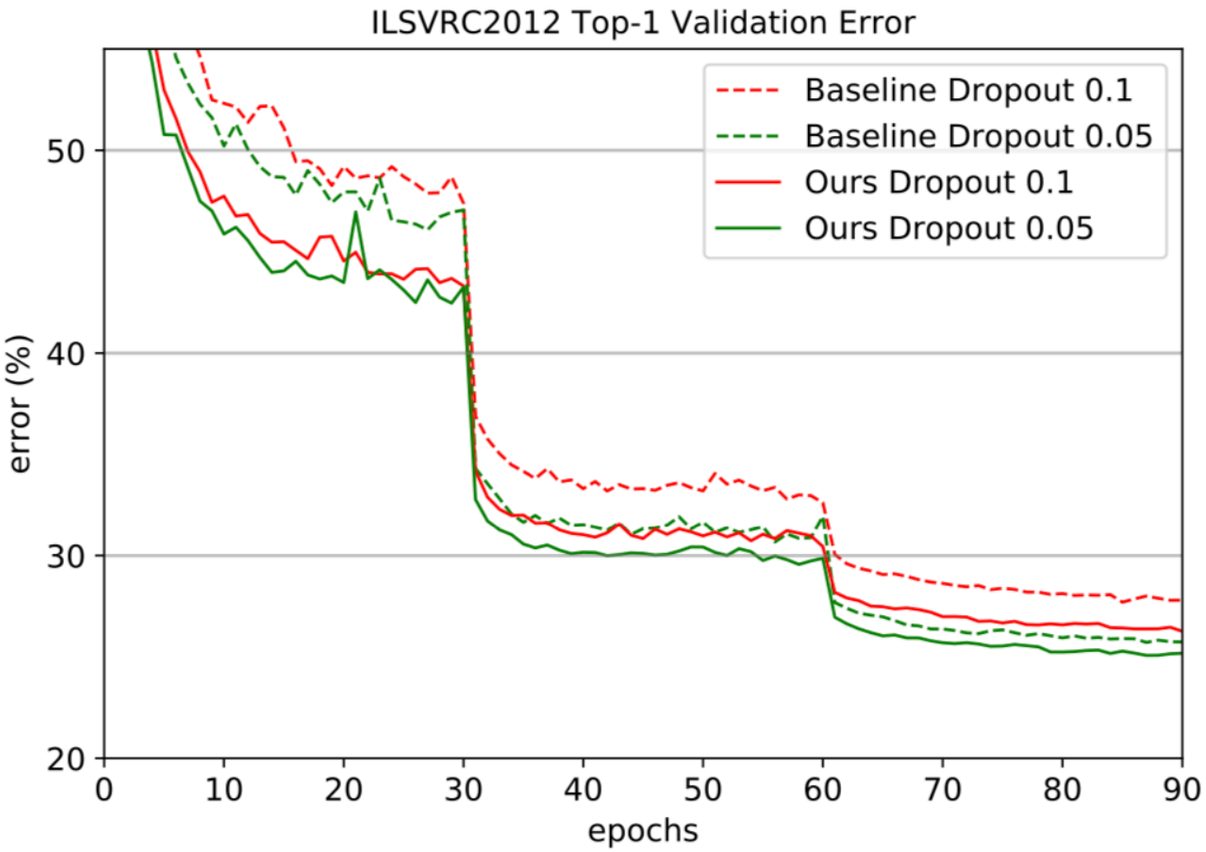


Figure 4. The testing accuracy of implementing ResNet and ResNet-B with the IC layer on the CIFAR10/100 datasets with respect to the training epochs. (a) ResNet110 on CIFAR 10. (b) ResNet164 on CIFAR 10. (c) ResNet-B 110 on CIFAR 10. (d) ResNet-B 164 on CIFAR 10. (e) ResNet110 on CIFAR 100. (f) ResNet164 on CIFAR 100. (g) ResNet-B 110 on CIFAR 100. (h) ResNet-B 164 on CIFAR 100.

Results



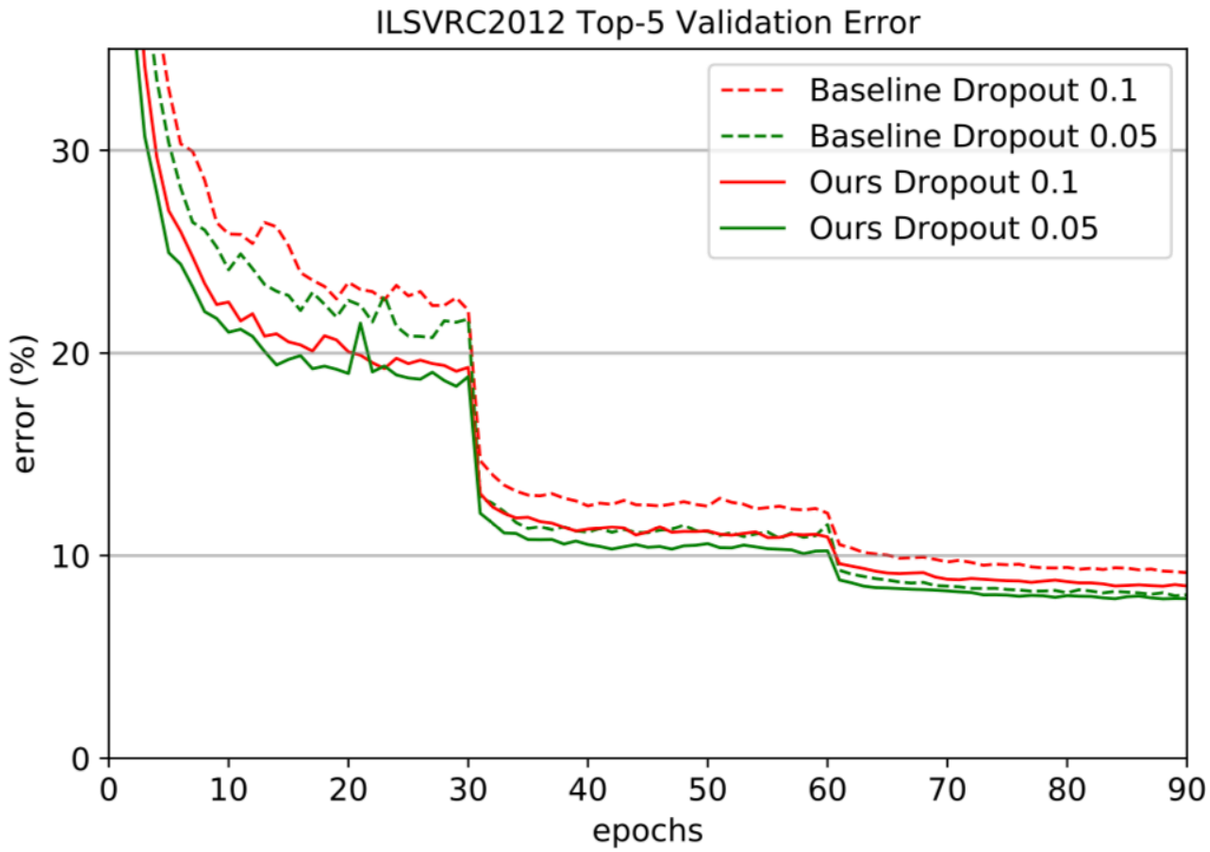


Figure 5. Top-1 and Top-5 (1-crop testing) error on ImageNet validation.