Residual Connections Encourage Iterative Inference

Stanisław Jastrzębski*, Devansh Arpit*, Nicolas Ballas, Vikas Verma, Tong Che, Yoshua Bengio

facebook Artificial Intelligence Research

Introduction

Residual networks have become a prominent architecture in deep learning. A recent view argues that Resnets perform iterative refinement of features. We attempt to further expose properties of this aspect. To this end, we study Resnets both analytically and empirically. We formalize the notion of iterative refinement in Resnets by showing that residual connections naturally encourage features of residual blocks to move along the negative gradient of loss as we go from one block to the next
We focus on Residual Networks:

$$\mathbf{h_L} = \mathbf{h}_{L-1} + F(\mathbf{h}_L) \quad (1)$$

Loss **expands** as:

$$\begin{aligned}\mathcal{L}(\mathbf{h}_L) &= \mathcal{L}(\mathbf{h}_{L-1} + F_{L-1}(\mathbf{h}_{L-1})) \quad (2)\\ &= \mathcal{L}(\mathbf{h}_{L-1}) + F_{L-1}(\mathbf{h}_{L-1}).\frac{\partial \mathcal{L}(\mathbf{h}_{L-1})}{\partial \mathbf{h}_{L-1}}\end{aligned}$$

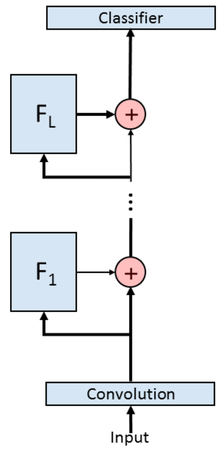

Fig. 1: Resnet Block

Main Result - Iterative Inference

Residual Connections Encourage Iterative Inference

$$\mathcal{L}(\mathbf{h}_L) = \mathcal{L}(\mathbf{h}_i) + \sum_{j=i}^{L-1} \langle F_j(\mathbf{h}_j), \frac{\partial \mathcal{L}(\mathbf{h}_j)}{\partial \mathbf{h}_j} \rangle \quad (3)$$

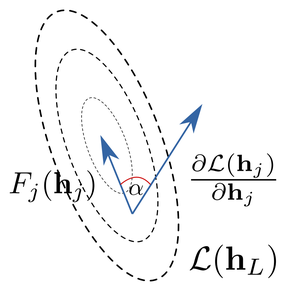

Fig. 2:

Output of the residual block is aligned with the derivative of the loss with respect to the hidden state

Experiments: Cosine Loss

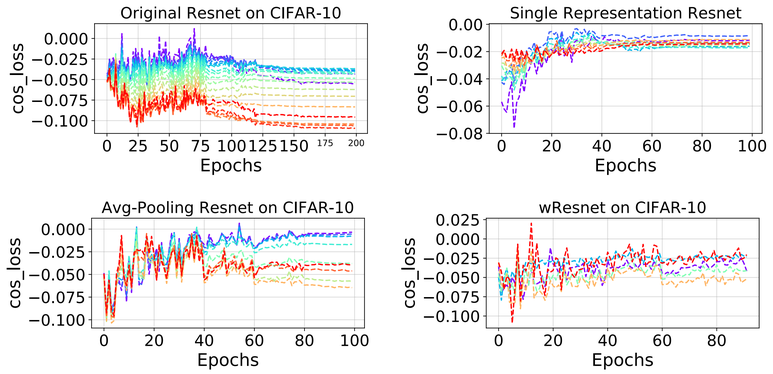

Fig. 3:

Average cos loss between residual block output and loss derivative for original Resnet, single representation Resnet, avg-pooling Resnet, and wideResnet on CIFAR-10.

Resistence to Dropping Blocks

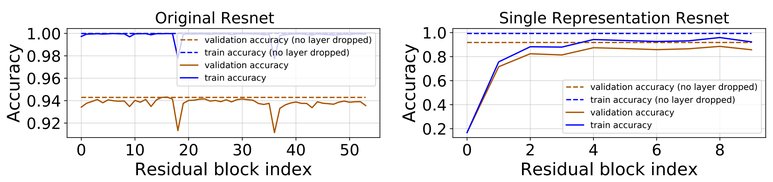

Fig. 4:

Final prediction accuracy after dropping blocks for original Resnet (left), single representation Resnet (right) on CIFAR-10.

L_2 Norm of Input Change

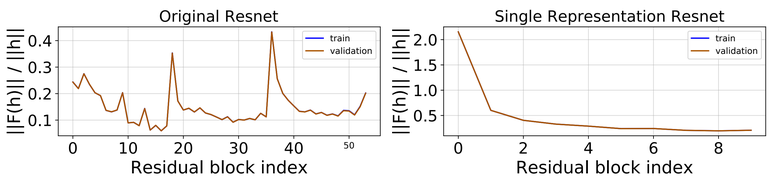

Fig. 5:

Average ratio of ℓ^2 norm of output of residual block to the norm of the input of residual block for original Resnet (left), single representation Resnet (right).

Intermediate Accuracy

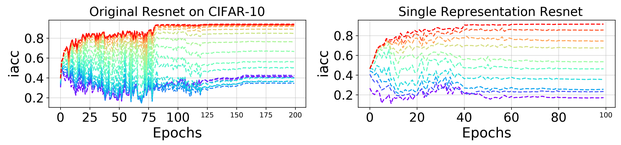

Fig. 6:

Intermediate accuracy measured by applying final classifier at the given hidden state for original Resnet (left), single representation Resnet (right) on CIFAR-10.

Unrolling To More Steps

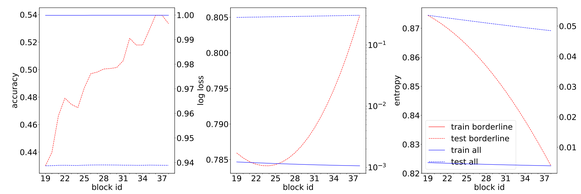

Fig. 7:

Accuracy, loss and entropy for Resnet-110 with last block unrolled for 20 *additional* steps (with an appropriate scaling).

Sharing Residual Blocks

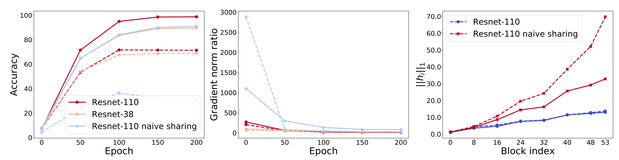

Fig. 8:

Resnet-110 with naively shared top 13 layers of each block compared with unshared Resnet-38. Shared Resnet-110 heavily overfits.

Conclusions

We have shown that residual networks implement iterative inference in a formal way. There remain many open questions, e.g.: Can we find ways to fully share residual blocks? Can we prove similar results for other architectures, e.g. using attention?

Acknowledgments: DA was supported by IVADO. SJ was partially supported by Grant No. DI 2014/016644 from Ministry of Science and Higher Education, Poland.