
Self-Attentive Layer Aggregation

Liangchen Luo

Peking University, Beijing
luolc.witty@gmail.com

June, 2019

Outline

- Background and motivation
- Approaches
- Experiments
- Future work

Background - Layers are not created equal

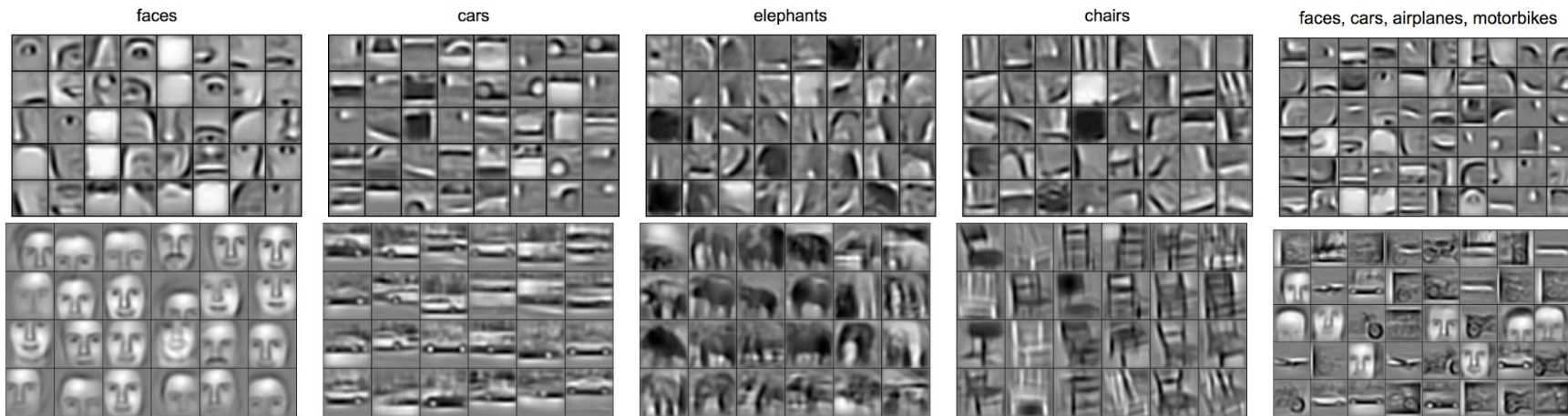
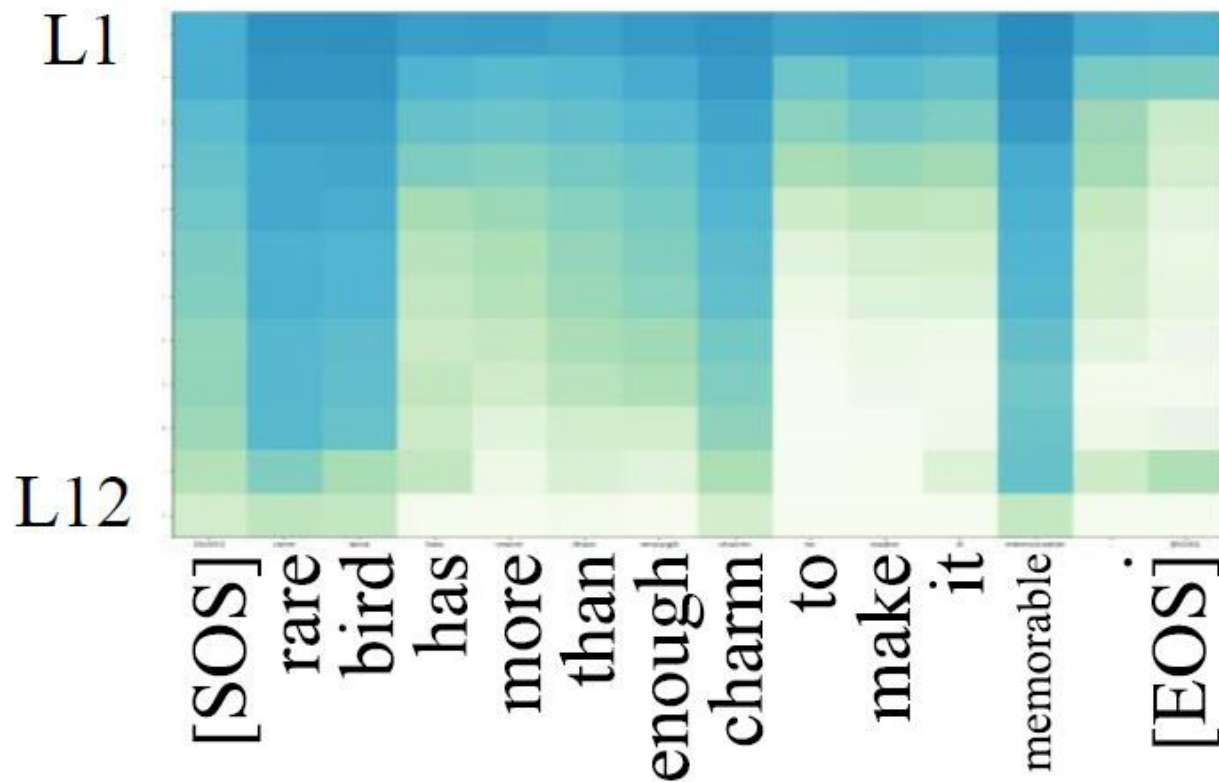


Figure 3. Columns 1-4: the second layer bases (top) and the third layer bases (bottom) learned from specific object categories. Column 5: the second layer bases (top) and the third layer bases (bottom) learned from a mixture of four object categories (faces, cars, airplanes, motorbikes).

BERT



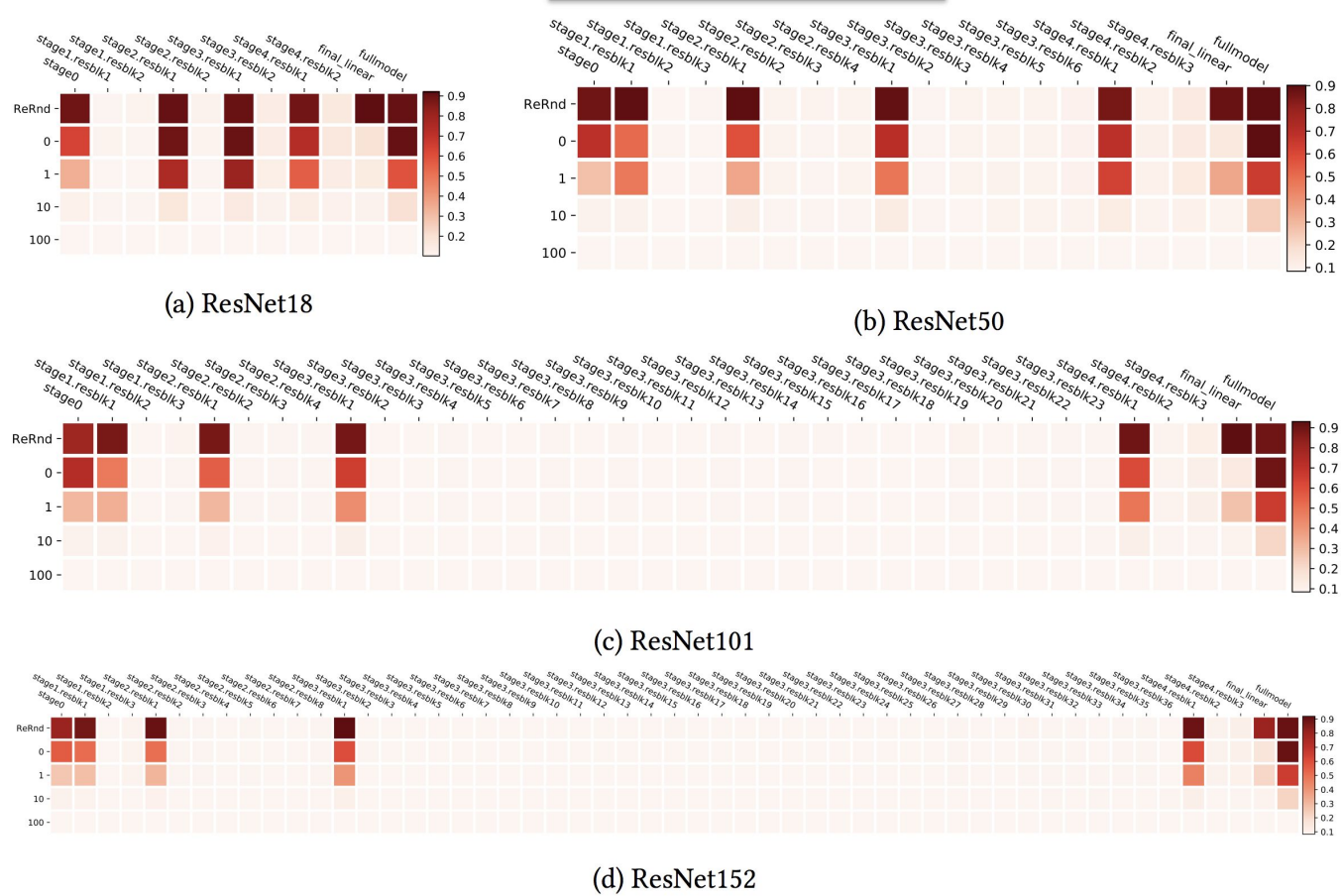
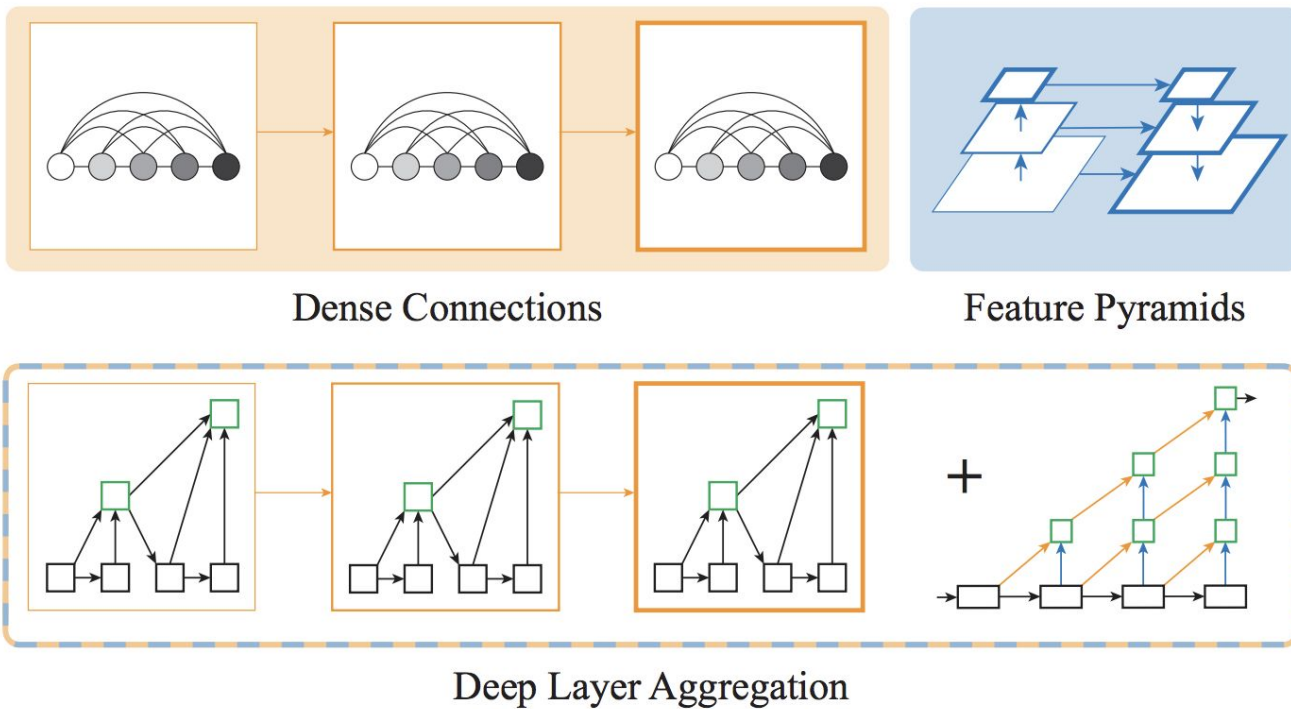


Figure 4: Whole-layer robustness for residual blocks of ResNets trained on CIFAR10.

Motivation

- Using the output of the last layer → Inefficient
- How to fuse the information from all layers? → Layer aggregation

Layer Aggregation



Layer Aggregation

- ELMo (Peters et al., 2018): As a result, the biLM provides **three layers of representations for each input token**, including those outside the training set due to the purely character input. In contrast, traditional word embedding methods only provide one layer of representation for tokens in a fixed vocabulary.

Approaches

- We obtain the output of all L layer:

$$X_1, X_2, \dots, X_L$$

- Assume the dimensionality is the same (otherwise use linear transformation)
- Feed the layer outputs into a self-attention (Vaswani et al., 2017) module, where Q , K , V are computed with three linear transformations, respectively:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- Use the output of self-attention module at index- L for downstream applications.

Experiments

- We test on CIFAR-10 (Krizhevsky and Hinton, 2009);
- with ResNets (Kaiming et al., 2016);
- trained with AdaBound optimizer (Luo et al., 2019).

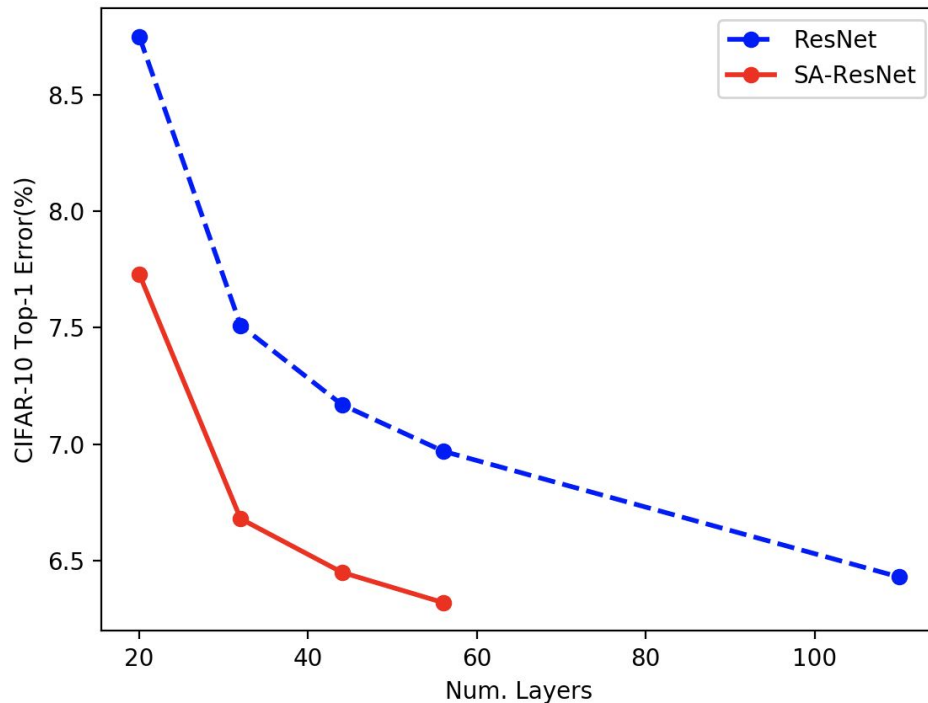
Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. *Tech. report*, 2009.

Kaiming He et al. Deep residual learning for image recognition. *CVPR*, 2016.

Liangchen Luo et al. Adaptive gradient methods with dynamic bound of learning rate. *ICLR*, 2019.

Results

#layers	ResNet	SA-ResNet
20	8.75	7.73
32	7.51	6.68
44	7.17	6.45
56	6.97	6.32
110	6.43	-



It is worth noticing that the testing error of the 56-layer SA-ResNet is comparable to that of the 110-layer ResNet on CIFAR10. These results indicate that the Self-Attentive architecture can greatly compress ResNet without losing the performance.

Future work (ongoing work)

- A neural ODE view
- Using recurrent units

Feel free to contact me if you are interested in further details.

Any questions?