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

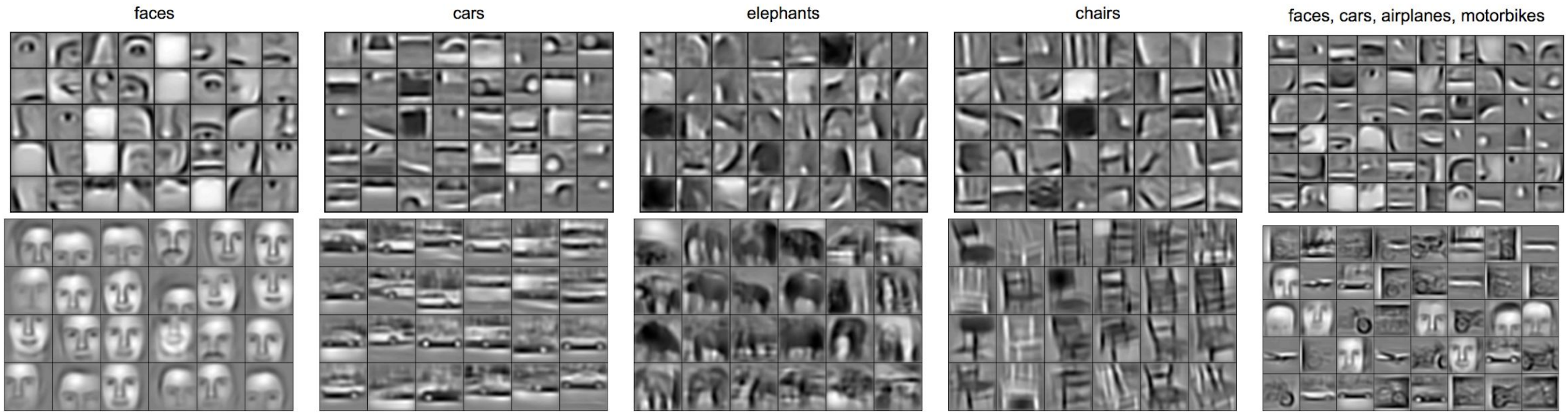
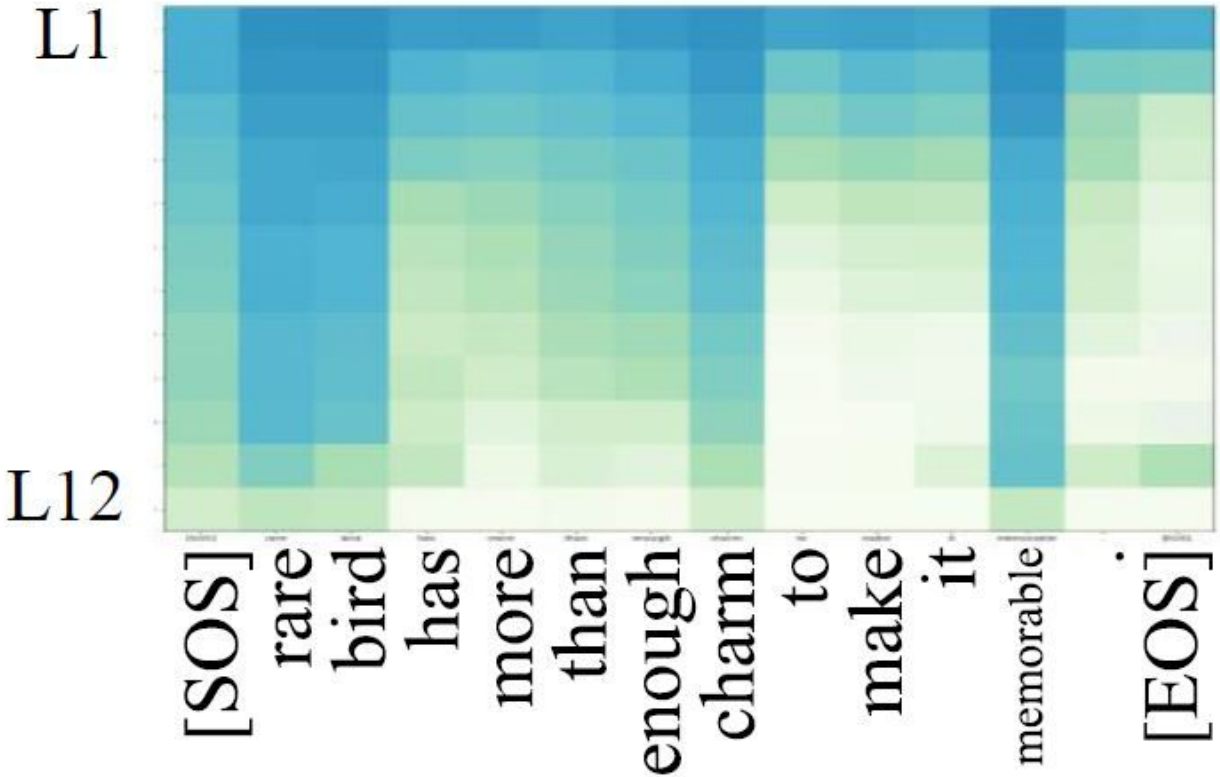


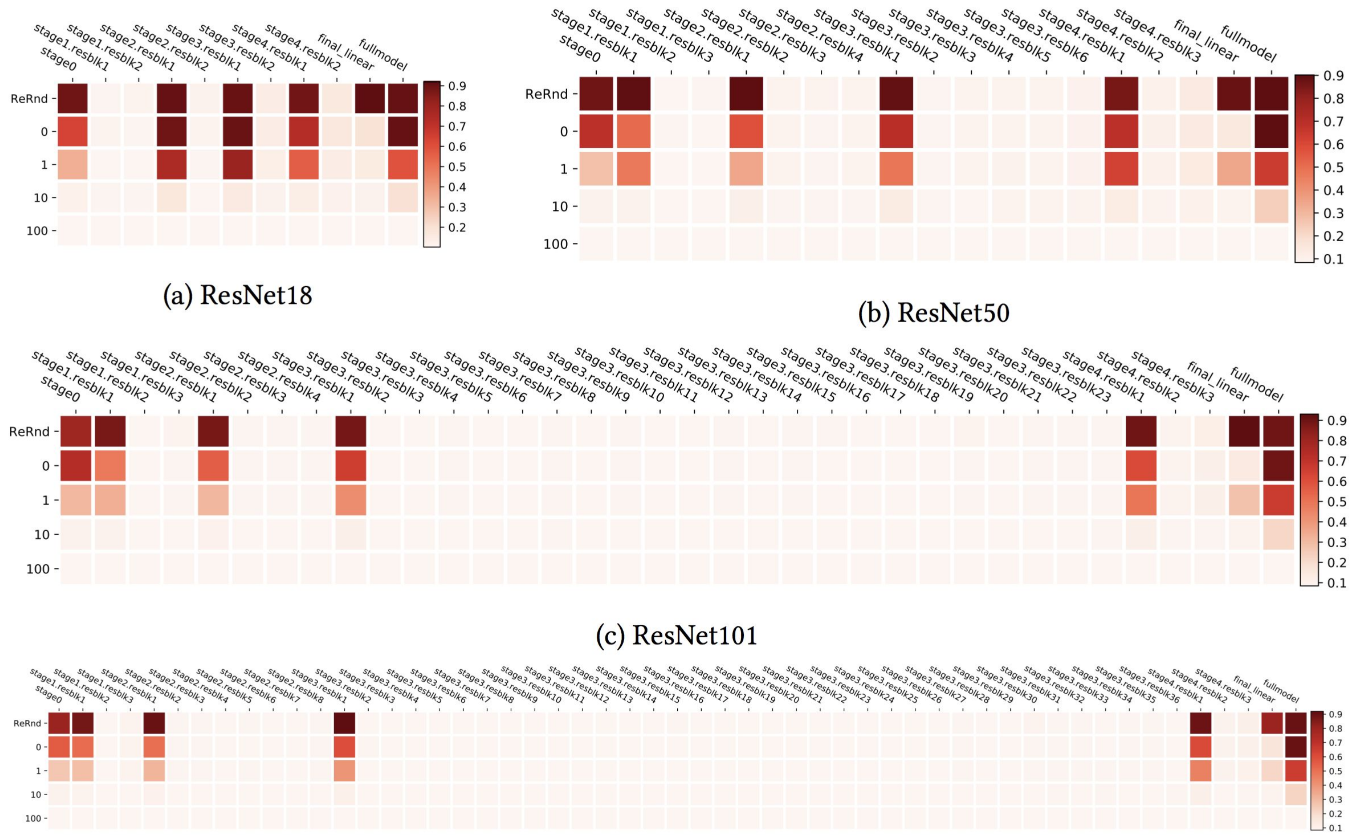
Figure 3. Columns 1-4: the second layer bases (top) and the third layer bases (bottom) learned from specific object categories. Column 5: the second layer bases (top) and the third layer bases (bottom) learned from a mixture of four object categories (faces, cars, airplanes, motorbikes).

Honglak Lee et al. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. *ICML*, 2009.

BERT



Chaoyu Guan et al. Towards a deep and unified understanding of deep neural models in NLP. *ICML*, 2019.



(a) ResNet18

(b) ResNet50

(c) ResNet101

(d) ResNet152

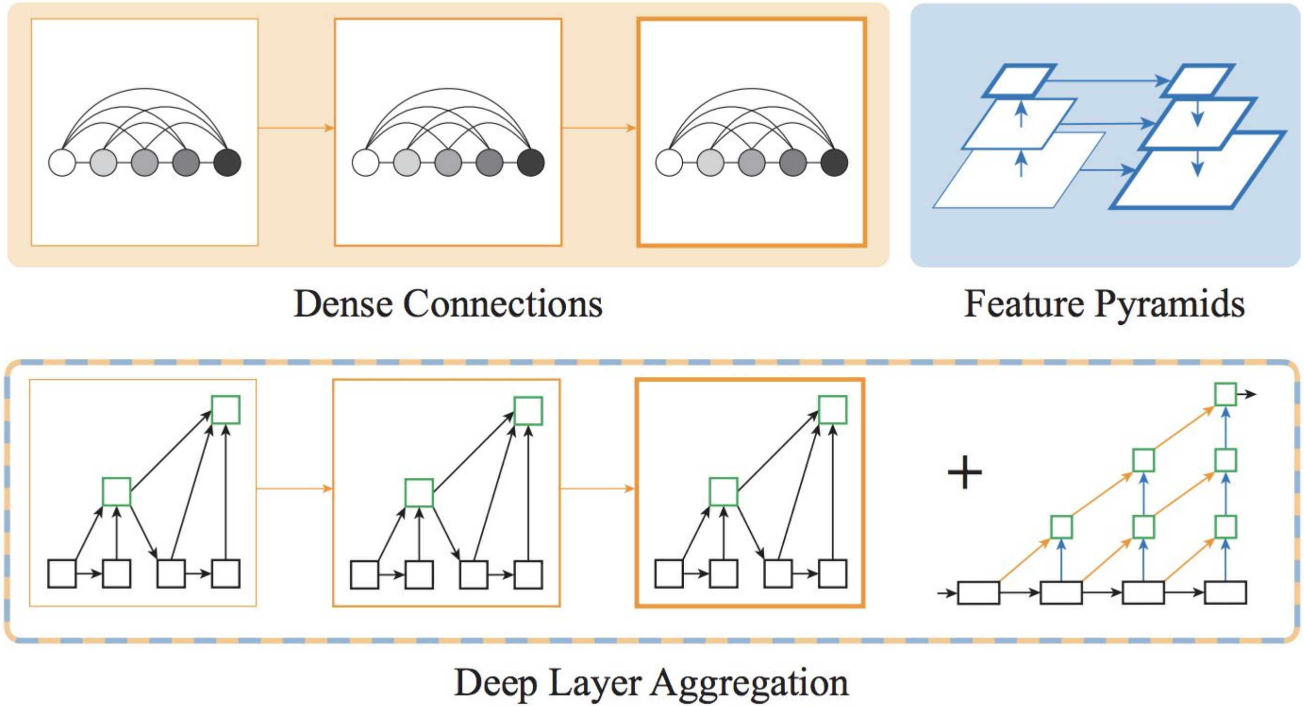
Figure 4: **Whole-layer robustness for residual blocks of** ResNet**s trained on CIFAR10.**

Chiyuan Zhang et al. Are all layers created equal? *arXiv/1902.01996*, 2019.

Motivation

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

Layer Aggregation



Fisher Yu et al. Deep layer aggregation. *CVPR*, 2018.

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.

Matthew E. Peters et al. Deep contextualized word representations. *NAACL*, 2018.

Approaches

- We obtain the output of all L layer:

$$X_1, X_2, \ldots, X_L$$

- Assume the dimentionality 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:

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

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

Vaswani et al. Attention is all you need. *NeurIPS*, 2017.

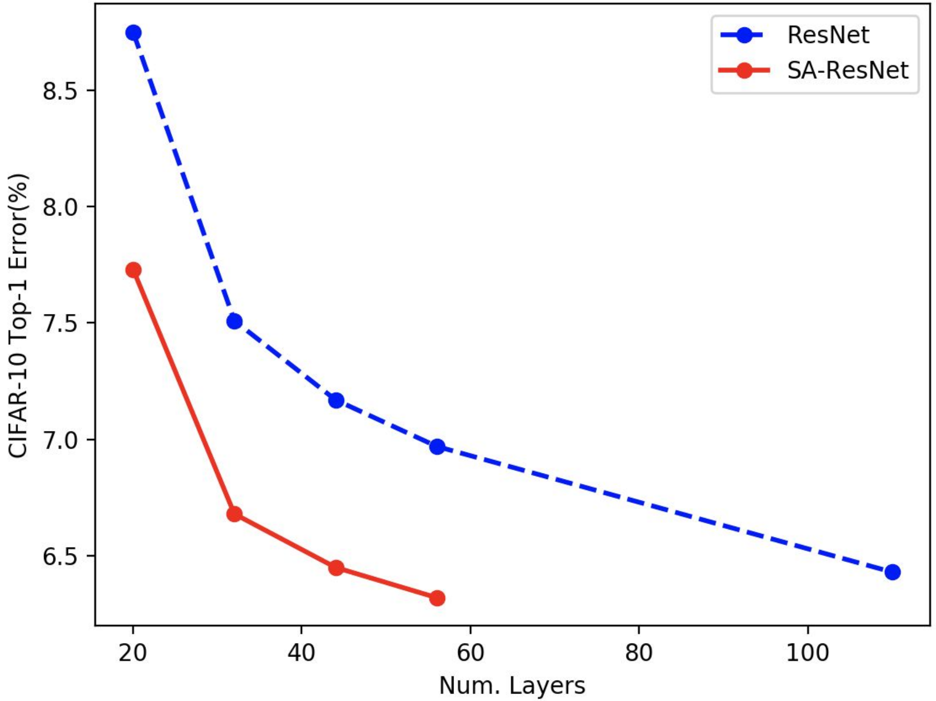
Experiments

- 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.

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?