Team#26 project

"Adaptive Mixture of Low-Rank Factorizations for Compact Neural Modeling" (ICLR 2019)

Yuriy Gabuev, Van Khachatryan, Stanislav Tsepa
Denis Zuenko, Aleksandr Rubashevskii

Skoltech
20 December, 2018

Outline

- Key concepts
- Experiment reproduction
- Conclusion
- References

Problem

- Modern NNs have large weight matrices, most are not suitable for mobile deployment
- Low-rank factorization is the popular instrument to reduce matrix size
- Large weight matrix W can be represented as a product of two small rank-d matrices

$$W = UV^\top \qquad U \in \mathbb{R}^{m\times d}, V \in \mathbb{R}^{n\times d}$$

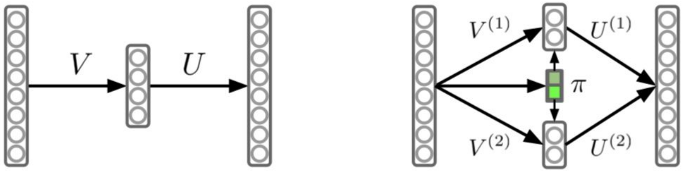
- Large complexity decrease: $O(d(m+n)) \ll O(mn)$
- Problem - loss of information when projecting onto low-dimensional space (Linear bottleneck)

Adaptive mixture of low-rank factorizations

- Make decompositions ~~great again!~~ data-dependent

$$W(h) = \sum_{k=1}^{K} \pi_k(h) U^{(k)} \left(V^{(k)}\right)^\top \qquad \pi(\cdot) : \mathbb{R}^n \to \mathbb{R}^K$$

- Replace large matvec with adaptive mixture of low-rank matvecs



(a) regular low-rank (b) adaptive low-rank

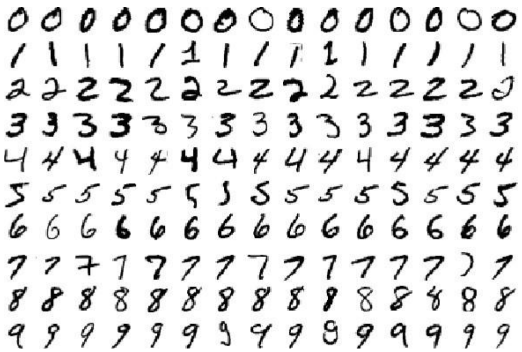
- Strictly speaking, **not a decomposition**, but a new learnable module

- $\pi(\cdot)$ is a small non-linear data-dependent function, e.g. $\pi(h) = \sigma(P(\mathrm{pool}(h)),\ P \in \mathbf{R}^{K \times n_{\mathrm{pooled}}}$

Experiment reproduction

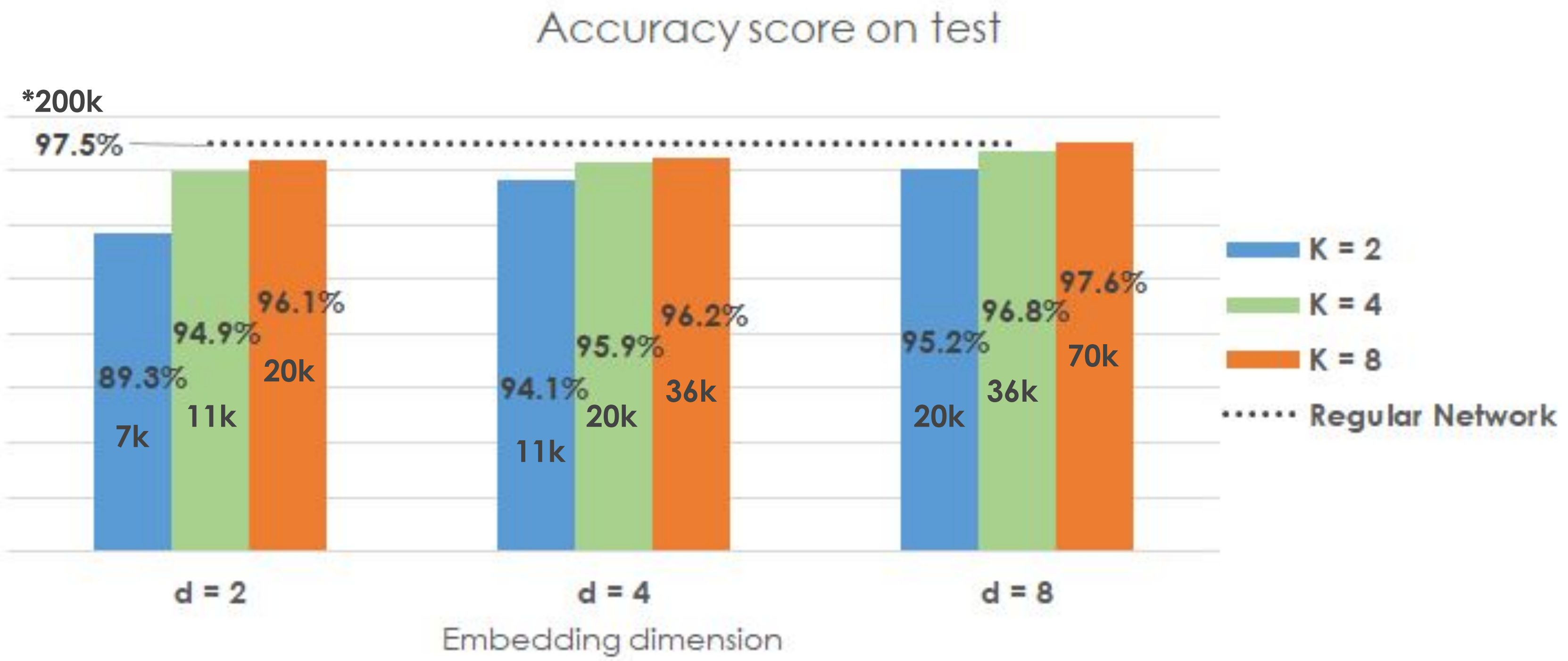
MLP (multi-layer perceptron)

- Digit recognition on MNIST dataset
- Simple one-layer MLP of 300 hidden units, input and output sizes are 784 and 10, respectively
- Rank-d matrices with d = 2, 4, 8
- K = 2, 4, 8
- Computed mixed weights with x (784 x 1) reduced to x (28 x 1)
- Accuracy of adaptive versions of low-rank factorization is 89-97.5%, depending on (K, d)



Experiment reproduction

MLP (multi-layer perceptron): Results

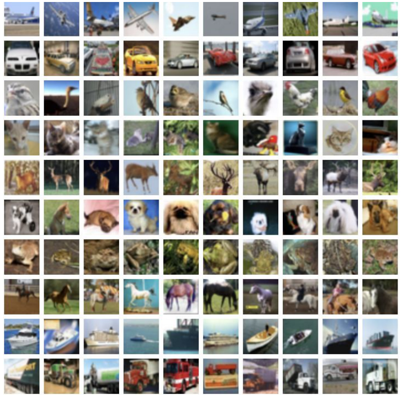


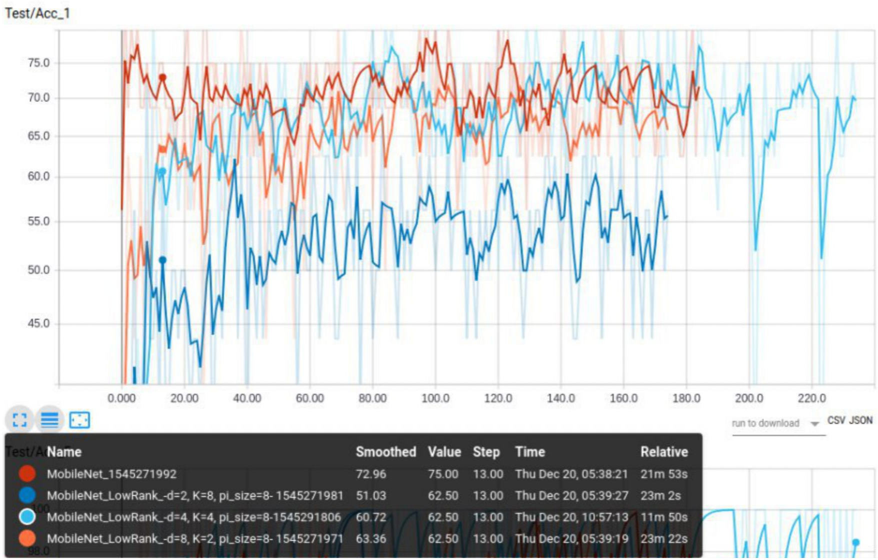
*number of parameters

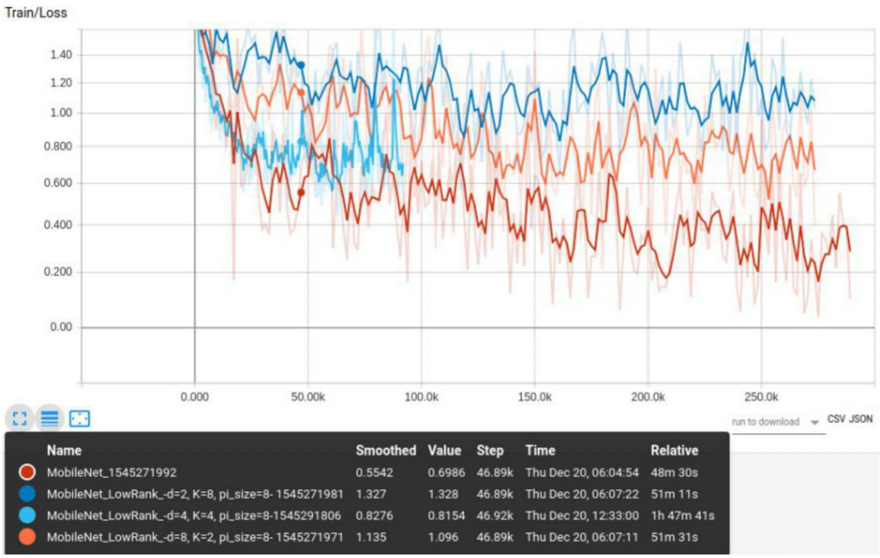
Experiment reproduction

Convolutional Neural Networks

- CIFAR-10
- MobileNet, pointwise convolutions replaced with adaptive low-rank approximation
- Reducing of parametres -> reducing of accuracy
- The best accuracy for MobileNet is 96.875
- The best accuracy for MobileNet-low-rank is 90.625







	best acc Top-1	Params
MobileNet-CIFAR10	96.875	121.61k
MobileNet-CIFAR10 low-rank, (2,8)	81.25	23.1k
MobileNet-CIFAR10 low-rank, (8,2)	90.625	46.52k
MobileNet-CIFAR10 low-rank, (4,4)	81.25	46.7k

Conclusion

- adaptive low-rank factorization is an original method proved to have results much better than regular low-rank decomposition
- it achieves up to 60-80% compression (dependent on the model) without significant decrease of accuracy
- in contrast to regular low-rank methods it learns non-linear low-rank manifolds due to learnable

$\pi(\cdot)$

References

[1] T. Chen, J. Lin, T. Lin, C. Wang, D. Zhou, S. Han, Adaptive Mixture of Low-Rank Factorizations for Compact, 2018
[2] François Chollet. Xception: Deep learning with depth-wise separable convolutions. *arXiv preprint*, 2016.

Code

https://github.com/zuenko/ALRF