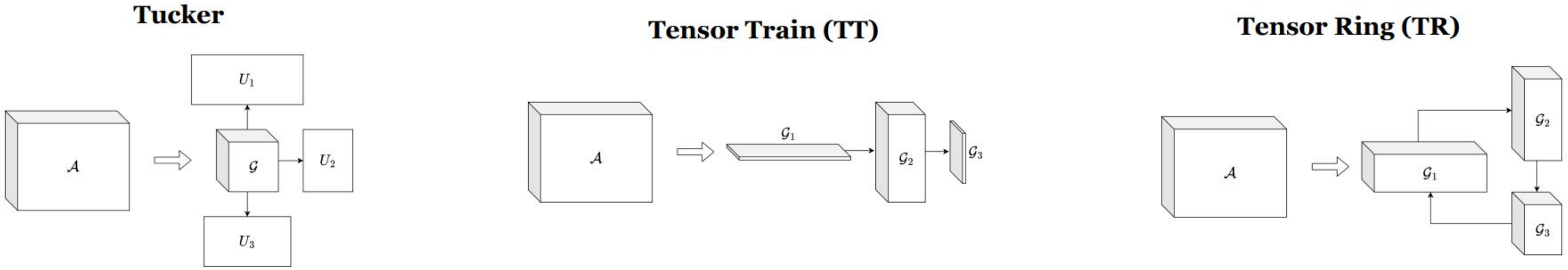
Skoltech

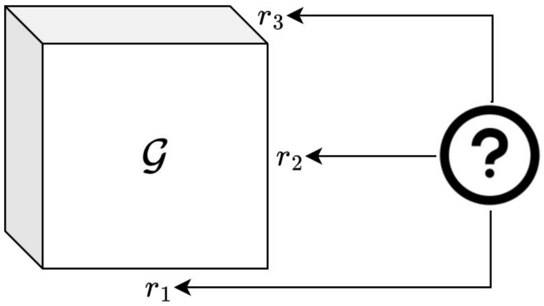
MARS: Masked Automatic Ranks Selection in Tensor Decomposition

Paper Review

UFO Team

Motivation&Problem: Tensor decomposition methods are effective in compression and acceleration of neural networks. However, to achieve balance between compression and performance we need to carefully select tensor ranks.



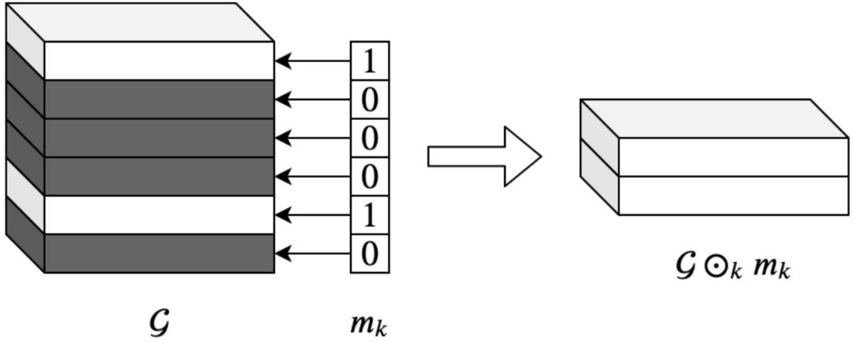


- Typical hyperparameter selection techniques, like cross-validation, are poorly suited for the choice of multiple tensor ranks.
- Existing approaches for core ranks selection are not general. They also could require significant computational overhead.

Skoltech

Proposed method

- Ranks can be represented by binary masks over tensor dimensions



- Bayesian model for these binary masks is proposed, MAP estimate is considered

$$p\left(Y, \boldsymbol{m}, \boldsymbol{G} \mid X\right)=\prod_{i=1}^{N} p\left(y_i \mid x_i, \boldsymbol{G} \odot \boldsymbol{m}\right) p(\boldsymbol{m}) p(\boldsymbol{G})$$

- Learned masks are binarized and applied to core tensors, resulting in compressed model

$$\mathcal{G}_k \odot \boldsymbol{m} := \mathcal{G}_k \odot_{k_1} m_{k_1} \cdots \odot_{k_p} m_{k_p}$$

Model

$$p\left(Y, \boldsymbol{m}, \boldsymbol{G} \mid X\right) = \prod_{i=1}^{N} p\left(y_i \mid x_i, \boldsymbol{G} \odot \boldsymbol{m}\right) p\left(\boldsymbol{m}\right) p\left(\boldsymbol{G}\right)$$

Prior over masks is assumed to be Bernoulli with the success parameter π

$$p\left(\boldsymbol{m}\right) = p\left(\boldsymbol{m} \mid \pi\right) = \prod_{k} \prod_{s=1}^{r_k} \pi^{m_k(s)} (1-\pi)^{1-m_k(s)}$$

MAP estimate is considered. This discrete optimization problem can be reduced to continuous optimization

$$\sum_{i=1}^{N} \log p\left(y_i \mid x_i, \boldsymbol{G} \odot \boldsymbol{m}\right) + \log p\left(\boldsymbol{m}\right) + \\ + \log p\left(\boldsymbol{G}\right) \longrightarrow \max_{\boldsymbol{m}, \boldsymbol{G}}$$

Skoltech

Model

Under the assumption that q(m) is a factorized Bernoulli distribution, problem from previous slide is equivalent to

$$\mathbb{E}_{\boldsymbol{m}\sim q_{\boldsymbol{\phi}}(\boldsymbol{m})}\left[\sum_{i=1}^{N}\log p\left(y_i \mid x_i, \boldsymbol{G}\odot\boldsymbol{m}\right)\right]+$$
$$+\sum_{k}\sum_{s=1}^{r_k}\left[\phi_k(s)\log\pi+(1-\phi_k(s))\log(1-\pi)\right]+$$
$$+\log p\left(\boldsymbol{G}\right)\longrightarrow\max_{\boldsymbol{\phi},\boldsymbol{G}}.$$

After applying reparameterization trick, we can use stochastic gradient descent to get MAP estimate of G (cores) and ϕ (probability parameter from Bernoulli distribution of masks).

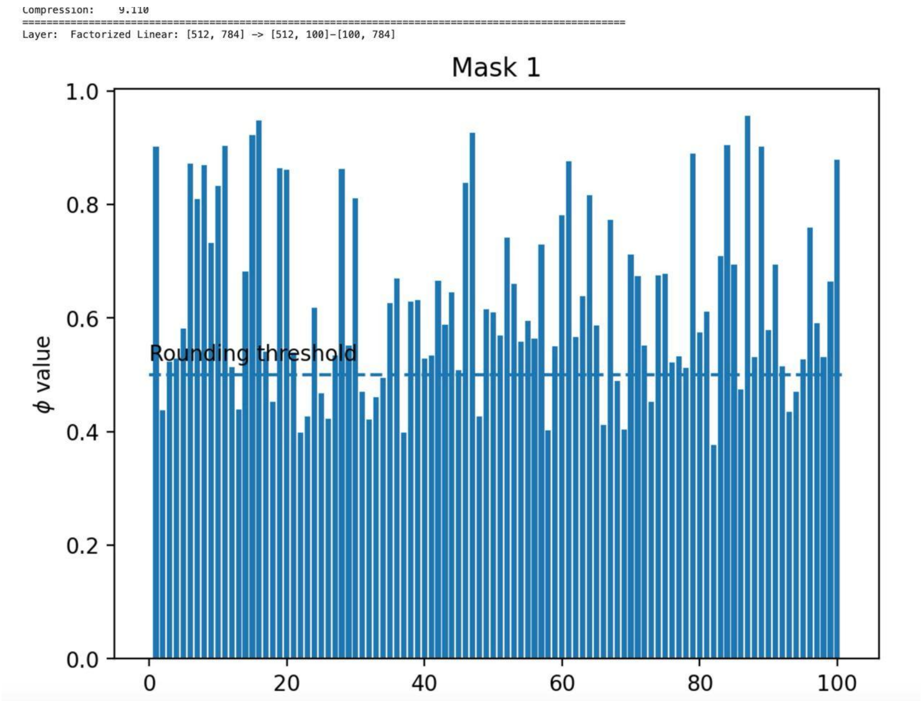
Finally, binary masks are obtained by rounding ϕ

Mode-k Hadamard product between cores and corresponding masks gives compressed tensors.

Reproducing paper results

https://github.com/xiyori/mars-reproducibility

- The code provided by the authors only reproduces 1 experiment with a simple fully-connected model
- We implement 2 more experiments: LeNet-5 on MNIST and ResNet-110 on CIFAR10
- Our results align with the paper sufficiently well



Reproducing paper results: LeNet-5

- Tucker decomposition for convolution
- Low-rank factorization for linear (Skeleton decomposition!)

$$A = UV^T,\ U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r}$$

Model	Compression	Accuracy	Speed-up (inference)	Slow-down (training)
		Original results		
Baseline	$1\times$	99.2%	$1\times$	-
MARS + Tucker	$10 \pm 0.8\times$	$99.0 \pm 0.07\%$	$1.19 \pm 0.01\times$	-
5-ensemble	$2\times$	99.5%	faster in parallel	-
		Our reproduction		
Baseline	$1\times$	99.3%	$1\times$	$1\times$
MARS + Tucker	$7.1\times$	$99.0 \pm 0.08\%$	$1.32 \pm 0.04\times$	$1.04\times$
5-ensemble	$1.4\times$	99.1%	faster in parallel	$5.2\times$

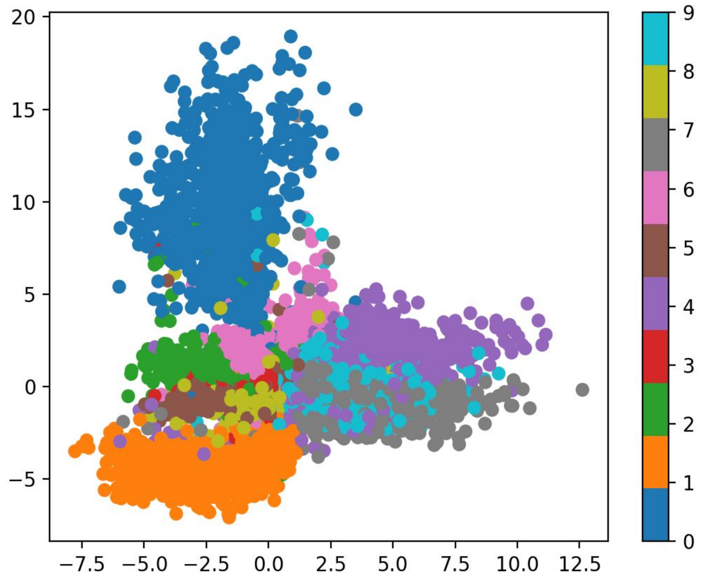
Reproducing paper results: ResNet-110

- Tensor Train for convolutions in 2nd and 3rd blocks
- Naive: 2d + 1 tensors r_{k-1} x n_k x r_k
- Proper: similar number of tensors r_{k-1} x n_k x m_k x r_k, inspired by Garipov et al. (2016)

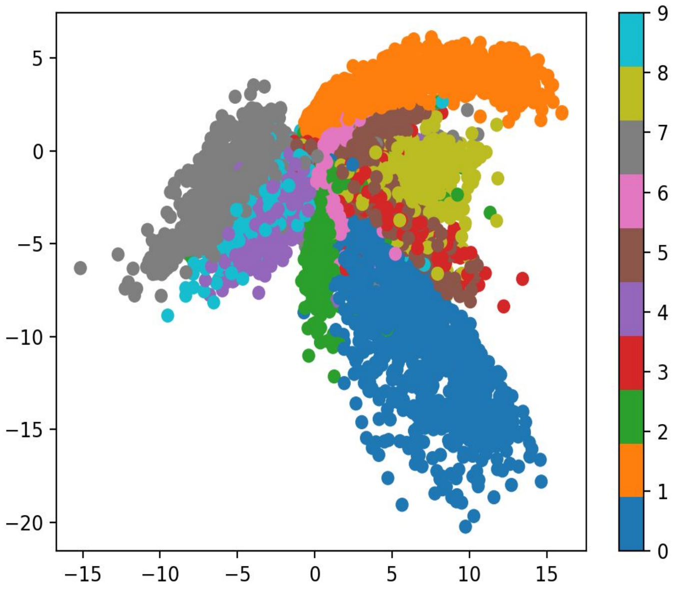
Model	Compression	Accuracy	Slow-down (inference)	Slow-down (training)
Original results				
Baseline	1×	92.6%	-	-
MARS (naive)	7.0×	90.7%	-	-
MARS (proper)	5.5×	91.1%	-	-
Our reproduction				
Baseline	1×	92.3%	1×	1×
MARS (naive)	at least 2.7×	at least 89%	1.60×	2.9×
MARS (proper)	at least 2.3×	?%	1.54×	2.8×

Skoltech

MARS for AutoEncoders



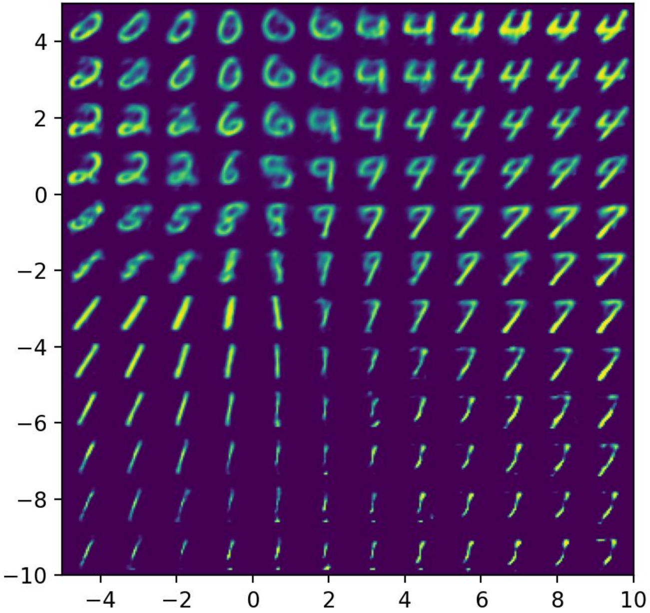
Factorized autoencoders latent space
Compression: 3.819



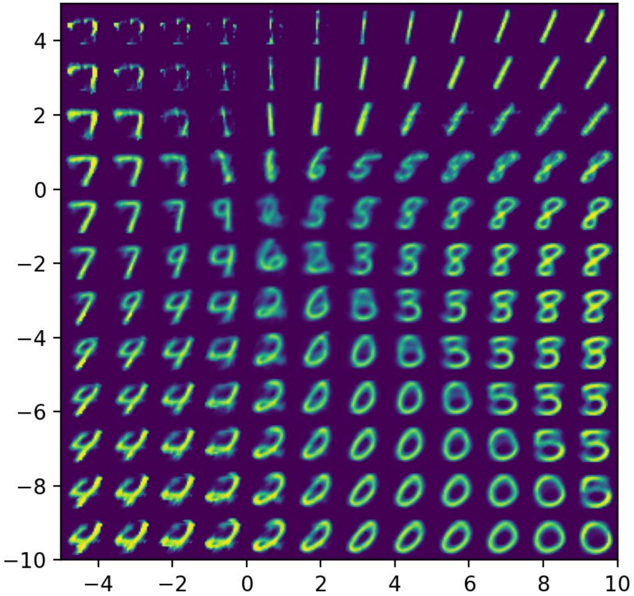
Base autoencoders latent space's latent space

Skoltech

MARS for AutoEncoders

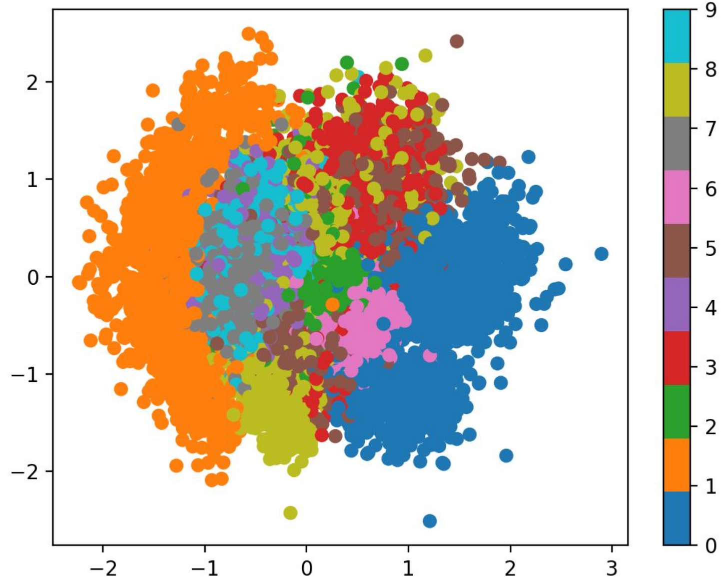


Factorized autoencoder reconstruction
Inference speedup: 1.11

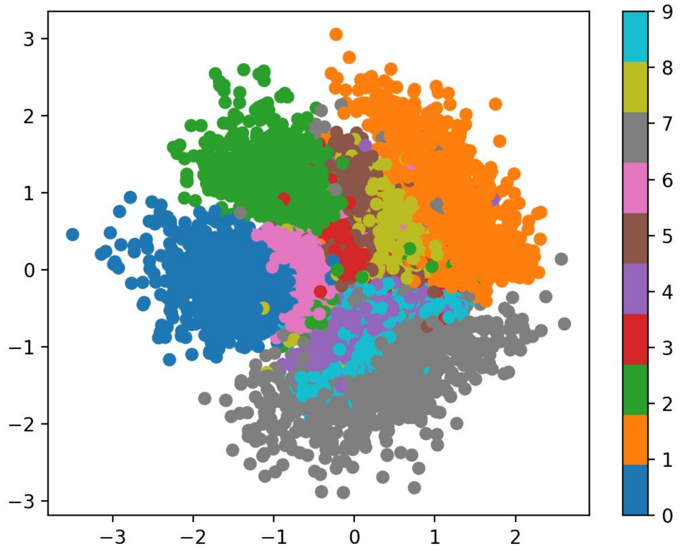


Base autoencoder reconstruction

MARS for VAE



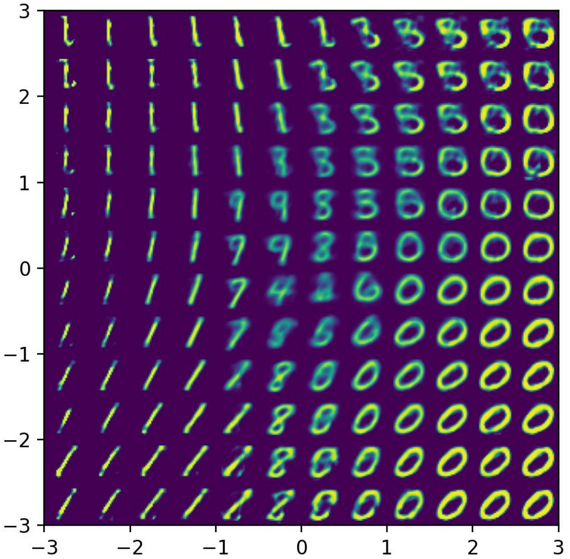
Factorized VAE latent space
Compression: 5.383



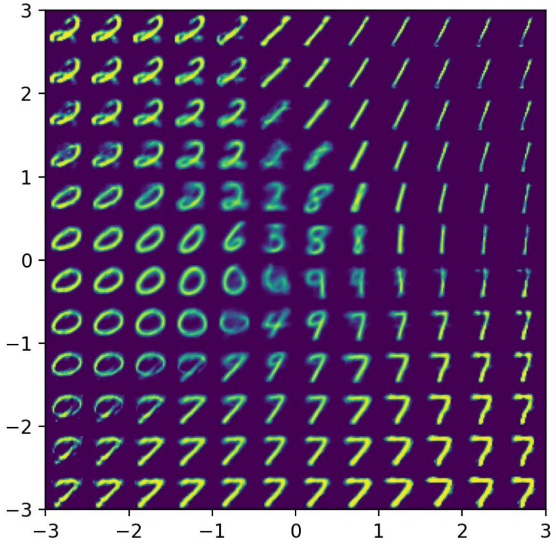
Base VAE latent space

Skoltech

MARS for VAE



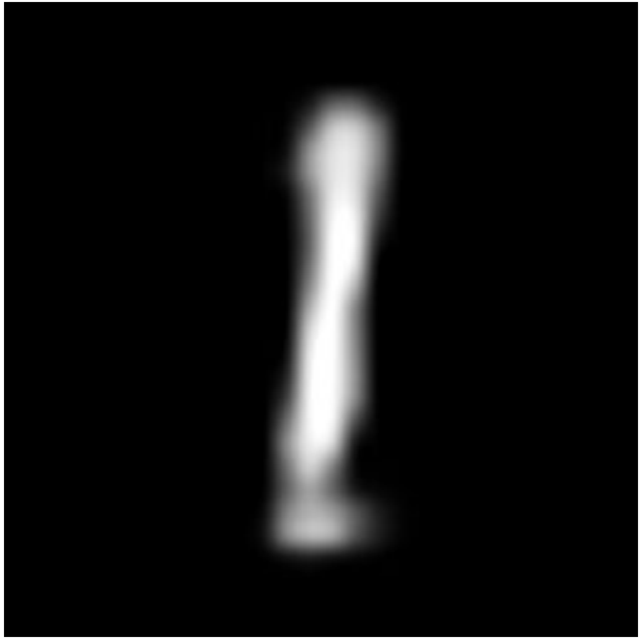
Factorized VAE reconstruction
Inference speedup: 1.12



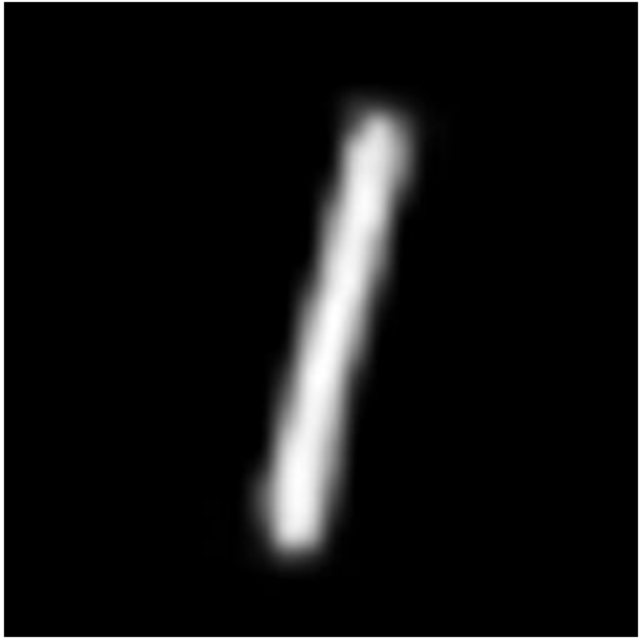
Base VAE reconstruction

Skoltech

MARS for VAE



Factorized VAE transition



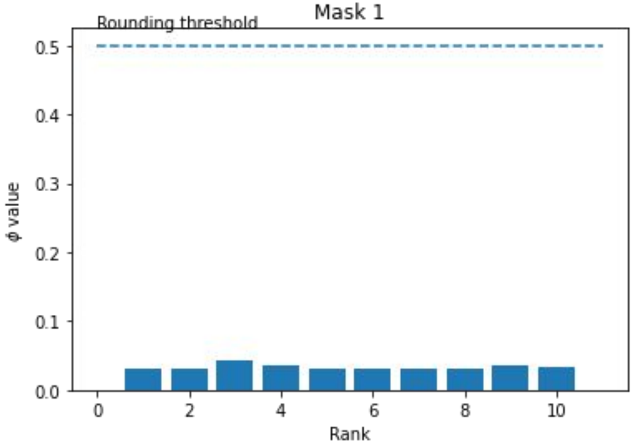
Base VAE transition

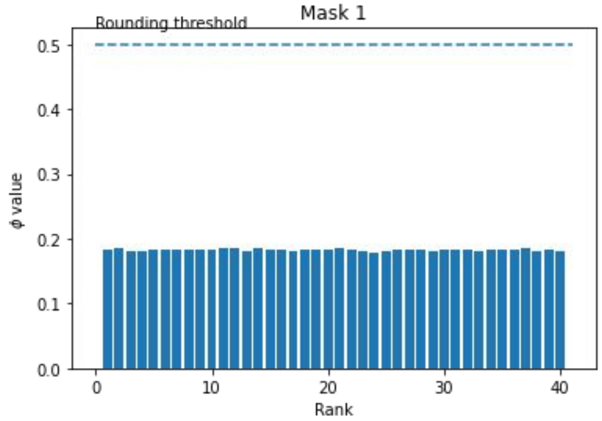
MARS for U-Net

Approach for U-Net:
- Replace ordinary convolution blocks with either ones based on TT or Tucker decomposition (similar to famous approach with MobileNet convolutions)

As a result, size of model decreases from the start

Challenges:
- Hard to make large model to “take off”
- During training MARS more likely to not find any structure inside of models at all and give a zero mask
- Training process take 2x-3x more time, than for ordinary U-Net, because requires some time after convergence to find a mask





Conclusion

Model Advantages

- **Universal** - applicable to various tensorized models
- **Sensible** - closely rank approximation
- **Effective** - much better than manual selection of ranks and no worse than specialized rank selection schemes
- **Efficient** - no extra computational cost
- **Scalable** - easy tensorization of ResNet-110
- **Consistent** - learned masks probabilities are close to hard values {0, 1}

Model Disadvantages

- In practice it is difficult to choose hyperparameters
- Ensemble learning approach is quite long
- Model is working long for TensorTrain rank selection

Our Team

Elfat Sabitov
Experimental applications

Petr Sychev
Theory understanding, preso

Foma Shipilov
Experiments reproduction

Petr Kushnir
Theory understanding, preso

Sergey Kushneryuk
Experiments with CNN