Spectral Tensor Train Parameterization of Deep Learning Layers

Anton Obukhov[1], Maxim Rakhuba[2], Alexander Liniger[1], Zhiwu Huang[1], Stamatios Georgoulis[1], Dengxin Dai[1], Luc Van Gool[1,3]
[1]ETH Zurich [2]HSE University [3]KU Leuven

STTP is a weight matrix W parameterization based on SVD and Tensor Train decompositions:

1. Low-rank in an unconventional way;
2. Unique and non-redundant;
3. Embedded spectral properties.

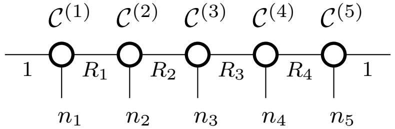
→ **network compression** & **training stability**, shown in image classification and GAN settings.

Recap and notation

Tensor Train (TT) decomposition is a representation of a tensor $W \in \mathbb{R}^{n_1 \times \ldots \times n_D}$ via D TT-cores $\mathcal{C}^{(i)} \in \mathbb{R}^{R_{i-1} \times n_i \times R_i}$ with TT-rank $(R_0, \ldots, R_D)$:

$$W_{i_1,\ldots,i_D} = \sum_{\beta_0,\ldots,\beta_D=1}^{R_0,\ldots,R_D} \mathcal{C}^{(1)}_{\beta_0,i_1,\beta_1} \cdot \mathcal{C}^{(2)}_{\beta_1,i_2,\beta_2} \cdots \mathcal{C}^{(D)}_{\beta_{D-1},i_D,\beta_D}$$

Tensor diagram notation is a technique for visualizing products like the one above (for $D = 5$):



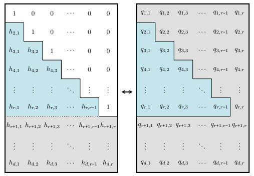
Stiefel manifold contains orthonormal frames of size $d \times r$, denoted as d — r in tensor diagrams:

$$\mathrm{St}(d, r) \equiv \{X \in \mathbb{R}^{d \times r} : X^\top X = I_r\}$$

Compact SVD of a matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ is $W = U\Sigma V^\top$, where $U \in \mathrm{St}(d_{\text{out}}, r)$, $V \in \mathrm{St}(d_{\text{in}}, r)$, and Σ is diagonal denoted as r — r . The rank $1 \le r \le \min(d_{\text{out}}, d_{\text{in}})$ defines the approximation precision.

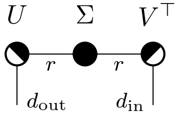
Redundancy in question arises, e.g., in SVD with $\Sigma = I_r$: $UV^\top = (UP)(VP)^\top$ for any $P \in \mathrm{St}(r, r)$. In this context we consider $\mathrm{St}_{\mathsf{U}}(d, r) \subset \mathrm{St}(d, r)$.

Householder parameterization of $Q \in \mathrm{St}(d, r)$ requires $dr - r(r+1)/2$ parameters $h_{i,j}$ organized in a lower-triangular matrix of size $d \times r$. We introduce parameterization of a submanifold $\mathrm{St}_{\mathsf{U}}(d, r)$ with zeros in blue areas, requiring $(d - r)r$ parameters.



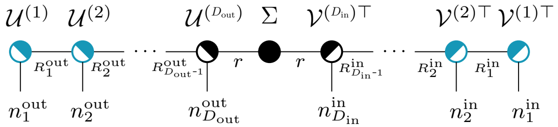
SVDP – a parameterization $W = U\Sigma V^\top$ with Householder parameterization of U, V and r parameters for Σ. Constraining Σ leads to *embedded spectral properties* (e.g., $\|\Sigma\|_\infty \to$ Lipschitz map).

When Σ is constrained to I_r, using St_{U} on one of U, V *removes parameterization redundancy*, which is further used to derive STTP.



STTP promotes *low-rank* in U (similarly in V) by

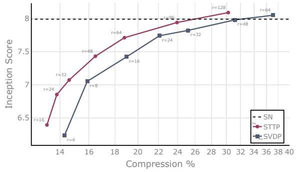
1. Factorizing $d_{\text{out}} = \prod n_i^{\text{out}}$;
2. Setting $\widetilde{U} = \texttt{reshape}(U, [n_1^{\text{out}}, \ldots, n_{D_{\text{out}}}^{\text{out}}, r])$;
3. Representing $\widetilde{U}$ with TT-cores $\mathcal{U}^{(\cdot)}$;
4. Parameterizing $\mathcal{U}^{(\cdot)}$ as elements of St_{U} (St for those adjacent to Σ). Tensor diagram for W:

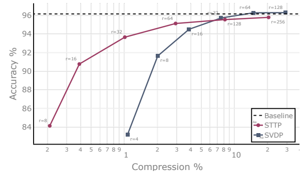


Thus, STTP offsets sparsity into U, V, which permits larger rank r than SVDP given the same budget of parameters → more expressive CNN layers.

CNN compression with STTP

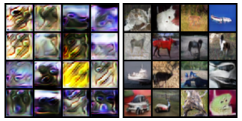
In each setting, all layers share the same rank r. SNGAN generator (left) and Wide ResNet image classifier (right) show higher performance with STTP in a wide range of compression settings.



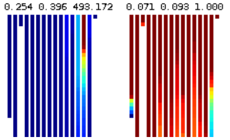


GAN training stability

STTP prevents spectral collapse (left) with fewer parameters and results in samples of a wide variety (right).



Spectral constraints are made easy due to the exposed Σ (image: all layers' Σ w/o and with $\|\Sigma\|_\infty = 1$).



TLDR: stable training of low-rank compressed neural networks with application to GAN and beyond. Updates and code release will be announced on Twitter.

Paper: arXiv 2103.04217
Project: obukhov.ai/sttp
Twitter: AntonObukhov1