Tensorized State Spaces for Sequential Tensor Networks

Jacob Miller[1,2,3], Guillaume Rabusseau[1,3], and John Terilla[2,4]

[1]*Mila*, [2]*Tunnel Technologies*, [3]*Université de Montréal*, [4]*CUNY Queens College*

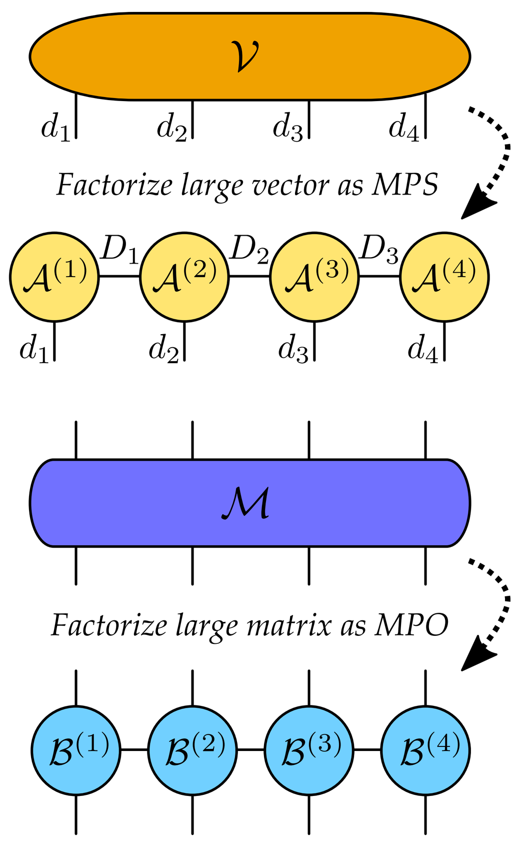
Abstract

Tensor networks have been used as theoretical frameworks for understanding deep learning architectures [2, 3], as well as practical tools for compressing large neural networks via "tensorization" [13]. But tensor networks on their own can serve as expressive models for many machine learning tasks [14, 19, 20], including unsupervised generative modeling [10]. Although general tensor networks impose a large computational overhead, matrix product states (MPS) represent efficient tensor network models which are well-adapted for sequential data, with close ties to weighted finite automata [4].

Although MPS generative models have interesting capabilities not achievable in neural network models [6], their expressive power is limited by area laws upper-bounding the mutual information between disjoint regions of their output sequences [5]. Here we propose a novel MPS architecture which circumvents this limitation through the use of a tensorized hidden state space. Our model can be seen as a second-order recurrent neural network [17], but one possessing a number of hidden units that grows exponentially in the allocated computational resources. Although still in its early development, we view this architecture as a promising merger of theoretical simplicity with significant expressive power, whose application to language modeling is a subject of active investigation.

Matrix Product States/Operators

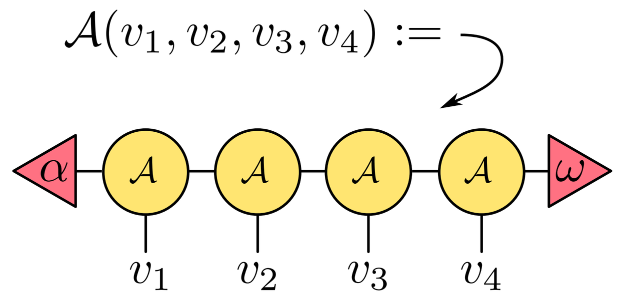
- Given vector $\mathcal{V} \in \mathbb{R}^{d_1 \times \dots \times d_4}$, a **matrix product state** (MPS) approximates $\mathcal{V}$ as contraction of smaller core tensors $\mathcal{A}^{(i)}$, where (e.g.) $\mathcal{A}^{(2)} \in \mathbb{R}^{D_1 \times d_2 \times D_2}$
- The D_i are **bond dimensions** of the MPS, hyperparameters which can be seen as a tensor generalization of matrix rank
- Applying the same procedure to a multi-mode matrix yields a **matrix product operator** (MPO)



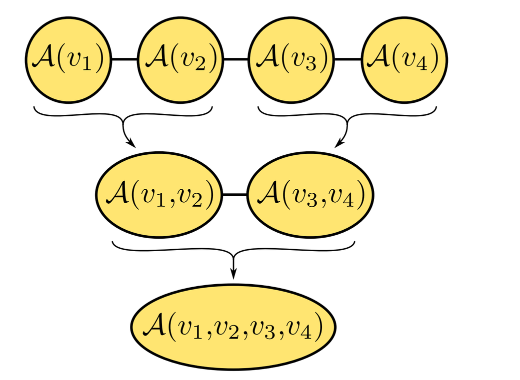
MPS and Recurrent Neural Networks

Translation-invariant MPS, satisfying $\mathcal{A}^{(i)} = \mathcal{A} \in \mathbb{R}^{D \times d \times D}$ can be seen as recurrent models equivalent to linear second-order recurrent neural networks (RNN) [17]

In contrast to standard recurrent models, TI-MPS parallelize well, with inputs of length n requiring only $\mathcal{O}(\log n)$ depth to evaluate (associativity of mat. mult.)

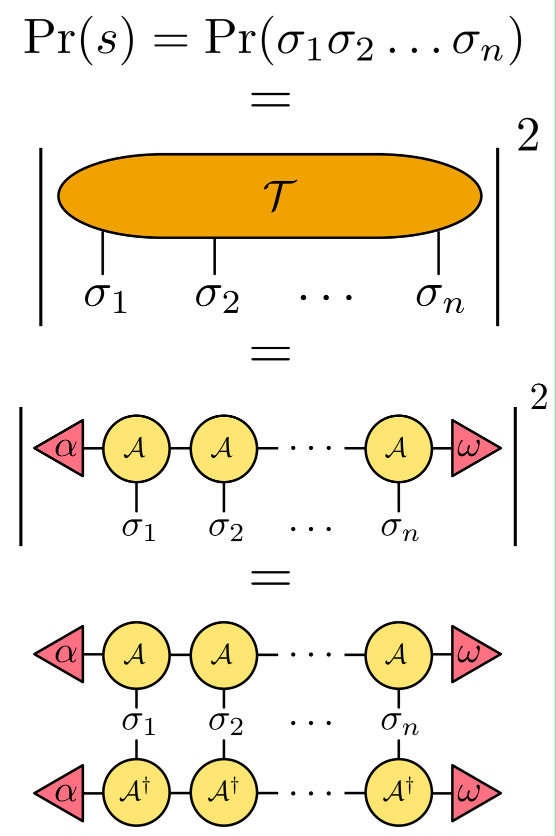


*α and ω are **boundary conditions** used to initialize and terminate a hidden state*



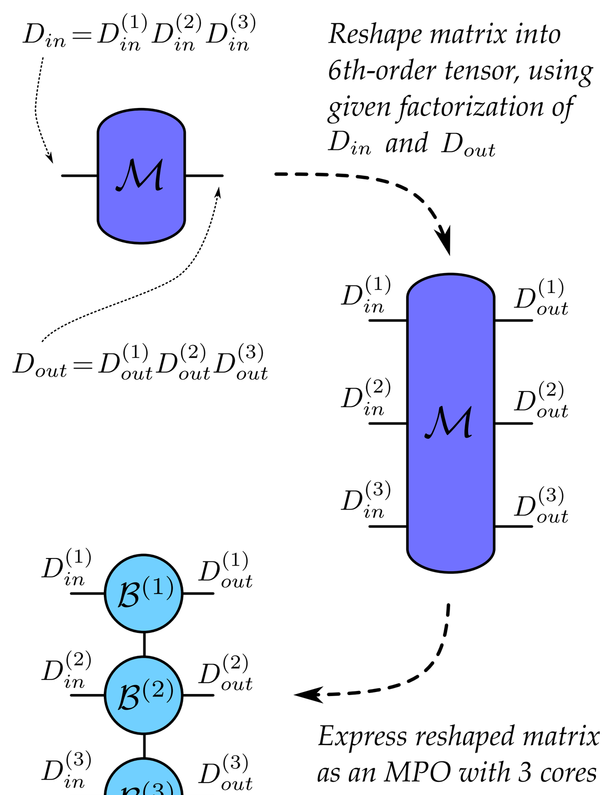
Modeling Sequences with Born Machines

- Distribution of sequences over an alphabet Σ can be modeled using a **Born machine** [10], in terms of a high-order tensor $\mathcal{T}$
- Associate symbols in Σ with orthogonal basis, then define $\Pr(\sigma_1 \dots \sigma_n) = |\mathcal{T}_{\sigma_1 \dots \sigma_n}|^2$
- Representing $\mathcal{T}$ using a TI-MPS is a natural choice for modeling natural language [15, 18], and also permits efficient evaluation and sampling from distribution



Tensorizing Neural Networks

- Neural network layers can be **tensorized** by expressing large weight matrices as MPO's [13]
- Permits significant reduction in number of parameters, with negligible loss in performance
- Neural net structure requires intermediate state vectors to be represented in dense format for application of elementwise nonlinearities (e.g. ReLU)



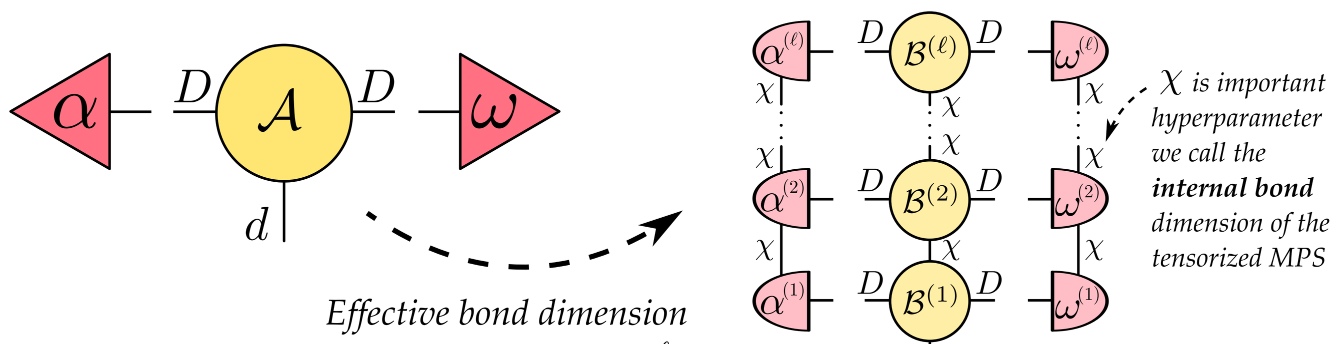
Our Tensorized MPS Architecture

Expressivity of a TI-MPS model is constrained by its bond dimension D, which is in turn limited by $\mathcal{O}(D^2)$ cost of evaluation

Our solution: Replace single core $\mathcal{A}$ by stack of connected cores, and intermediate states by vectors in MPS format

Dense MPS defined by:

Tensorized MPS defined by:



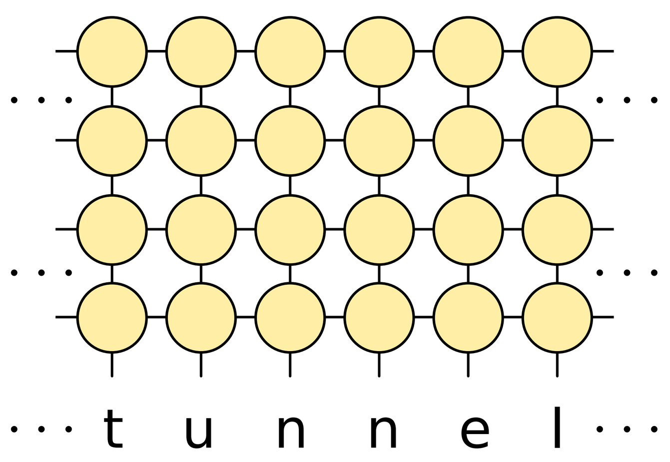
Effective bond dimension increases to $D_{\text{eff}} = D^\ell$

Design Tradeoffs in Architecture

- Exact evaluation of tensorized model requires exponential resources, so approximate contraction schemes are used instead
- Approximate evaluation of model with ℓ cores and internal bond dimension χ requires $\mathcal{O}(\chi^6 \ell)$ operations, constraining χ to take small values (~2-10)
- Model can be trained with stochastic gradient descent, but density matrix renormalization group (**DMRG**) and **spectral learning** algorithms are likely better choices, having strong theoretical guarantees and solid performance in modeling synthetic and real-world data [1, 17, 18, 19]

Connections with Other Models

- Formally, our model can be seen as a second-order RNN whose internal state space has an exponential number of hidden units
- Many standard neural network components aren't allowed in this architecture (e.g. nonlinear functions), but second-order structure circumvents many issues of standard linear RNN's [22]
- Tensorized MPS model is also an example of projected entangled pair state (PEPS), but with quasi-1D shape



Bibliography

[1] TD Bradley, EM Stoudenmire, and J Terilla, arXiv:1910.07425 (2019)
[2] N Cohen, O Sharir, and A Shashua, COLT (2016)
[3] N Cohen and A Shashua, ICLR (2017)
[4] M Droste, WK Manfred, and V Heiko (editors), Handbook of Weighted Automata, Springer (2009)
[5] J Eisert, M Cramer, and MB Plenio, Rev. Mod. Phys. 82, (2010)
[6] AJ Ferris and G Vidal, Phys. Rev. B 85, 165146 (2012)
[7] AJ Gallego and R Orus, arXiv:1708.01525 (2017)
[8] I Glasser, N Pancotti, and JI Cirac, arXiv:1806.05964 (2018)
[9] I Glasser, R Sweke, N Pancotti, J Eisert, and JI Cirac, arXiv:1907.03741 (2019)
[10] ZY Han, J Wang, H Fan, L Wang, and P Zhang, Phys. Rev. X 8, 031012 (2018)
[11] Y Levine, D Yakira, N Cohen, and A Shashua, ICLR (2018)
[12] HJ Liao, JG Liu, L Wang, and T Xiang, arXiv:1903.09650 (2019)
[13] A Novikov, D Podoprikhin, A Osokin, and DP Vetrov, NIPS (2015)
[14] A Novikov, M Trofimov, and I Oseledets, ICLR (2016)
[15] V Pestun, J Terilla, and Y Vlassopoulos, arXiv:1711.01416 (2017)
[16] V Pestun and Y Vlassopoulos, arXiv:1710.10248 (2017)
[17] G Rabusseau, TY Li and D Precup, AISTATS (2019)
[18] J Stokes and J Terilla, arXiv:1902.06888 (2019)
[19] EM Stoudenmire and DJ Schwab, NIPS (2016)
[20] EM Stoudenmire, Quant. Sci. and Tech. 3, 3 (2018)
[21] A Tjandra, S Sakti, and S Nakamura, IJCNN (2017)
[22] YH Wu, S Zhang, Y Zhang, Y Bengio, and RR Salakhutdinov, NIPS (2016)