Quantum Doubly Stochastic Transformers

Jannis Born Filip Skogh Kahn Rhrissorrakrai Filippo Utro Nico Wagner Aleksandros Sobczyk

presenter: Shen Yuan

中國人民大學 RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院 Gaoling School of Artificial Intelligence

Outline

Outline

Introduction

The dot product self-attention:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{A}\boldsymbol{V} = \text{Softmax}\Big(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\tau}\Big)\boldsymbol{V} \tag{1}$$

where $\boldsymbol{Q} := \boldsymbol{X}\boldsymbol{W}_Q$, $\boldsymbol{K} := \boldsymbol{X}\boldsymbol{W}_K$ and $\boldsymbol{V} := \boldsymbol{X}\boldsymbol{W}_V$. $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{T \times d}$
At the core of the Transformer, the softmax normalizes the attention matrix to be right stochastic. Sinkformer discovered that Transformer attention naturally converge to doubly stochastic matrices (DSMs) over training, i.e. their rows and columns sum to 1.

Introduction

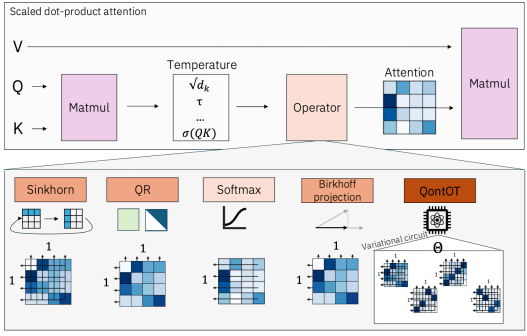


Figure 1. Doubly Stochastic Transformers. Standard scaled dot-product attention applies a Softmax activation on the query-key matrix (*top*). We study different techniques to make attention doubly stochastic attention by replacing the softmax operation (*bottom*). Our proposed Quantum Doubly Stochastic Transformer (QDSFormer) leverages QontOT, a variational quantum circuit with high expressivity.

Outline

Doubly Stochastic Matrices (DSMs)

We denote the n-dimensional vector of ones by $\mathbf{1}_n$ and the $n \times n$ identity matrix as $\boldsymbol{I}_n$. The *Birkhoff polytope* $\Omega_n := \mathcal{N}(\mathbf{1}_n, \mathbf{1}_n)$ defines the convex set of $n \times n$ doubly stochastic matrices (DSMs). A DSM $\boldsymbol{P} \in \Omega_n$ is a nonnegative matrix with row/column sum of 1

$$\boldsymbol{P}\mathbf{1}_n = \mathbf{1}_n, \quad \boldsymbol{P}^\top \mathbf{1}_n = \mathbf{1}_n, \quad \boldsymbol{P}_{i,j} \geq 0. \tag{2}$$

A right stochastic matrix $\boldsymbol{R}$ has row sums of 1

$$\boldsymbol{R}\mathbf{1}_n = \mathbf{1}_n, \quad \boldsymbol{R}_{i,j} \geq 0. \tag{3}$$

and a left stochastic matrix L has column sums of 1

$$\boldsymbol{L}^\top \mathbf{1}_n = \mathbf{1}_n, \quad \boldsymbol{L}_{i,j} \geq 0. \tag{4}$$

Doubly Stochastic Matrices (DSMs)

Birkhoff-von Neumann theorem states that the $n!$ vertices (i.e., extreme points) of the *Birkhoff polytope* Ω_n are permutation matrices, so their entries belong to $\{0, 1\}$.

Notably, every DSM $\boldsymbol{P} \in \Omega_n$ can be decomposed as a convex combination of permutation matrices, that is $\boldsymbol{P} = \sum_{i=1}^{N} \lambda_i \prod_i$ for some probability vector $\boldsymbol{\lambda} \in \Delta_N$, $n \times n$ permutation matrices $\{\prod_i\}$, and the number of extreme points $N \leq n^2$.

Sinkhorn's algorithm

Sinkhorn's algorithm is based on Sinkhorn's theorem, stating that for any square strictly positive matrix $\boldsymbol{R} \in \mathbb{R}_{+}^{T \times T}$, there exist (strictly) positive diagonal matrices $\boldsymbol{P} = \boldsymbol{D}_1, \boldsymbol{D}_2$ s.t., $\boldsymbol{D}_1 \boldsymbol{M} \boldsymbol{D}_2 \in \Omega_T$.

Sinkhorn's algorithmis an approximation procedure that iteratively (K times) normalizes the mass of the rows and the columns of $\boldsymbol{M} := \boldsymbol{Q}\boldsymbol{K}^\top$.

However, it may not converge with few iterations, especially if $\boldsymbol{Q}\boldsymbol{K}^\top$ contains large numeric values. Therefore, the Sinkformer is only an approximately doubly stochastic Transformer.

Projection on the Birkhoff polytope

Alternatively, one can project $\boldsymbol{M}$ directly on Ω_T via $\boldsymbol{P} = \arg\min_{\boldsymbol{X}\in\Omega_T} ||\boldsymbol{X} - \boldsymbol{M}||_F^2$, where the set for $\boldsymbol{X}$ and the objective are convex. We chose to minimize the Frobenius norm here.

The resulting problem is a positive-definite convex quadratic program and can be rewritten as

$$\min_{s.t., \boldsymbol{x}_i \geq 0, \boldsymbol{A}\boldsymbol{x} = \boldsymbol{1}_{2n}} \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{x} - \boldsymbol{q}^\top \boldsymbol{x}, \quad \boldsymbol{A} = \begin{pmatrix} \boldsymbol{1}_n^\top & \boldsymbol{0}^\top & \dots & \boldsymbol{0}^\top \\ \boldsymbol{0}_n^\top & \boldsymbol{1}^\top & \dots & \boldsymbol{0}^\top \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0}^\top & \boldsymbol{0}^\top & \dots & \boldsymbol{1}_n^\top \\ \boldsymbol{I}_n & \boldsymbol{I}_n & \dots & \boldsymbol{I}_n \end{pmatrix} \tag{5}$$

where $\boldsymbol{x} = \text{vec}(\boldsymbol{X}^\top)$, $\boldsymbol{q} = \text{vec}(\boldsymbol{M}^\top)$, $\boldsymbol{A} \in \mathbb{R}^{2n \times n^2}$.

QontOT

QontOT is a parameterized (variational) quantum circuit that was conceived very recently for conditional prediction of optimal transport plans.

For any unitary matrix $\boldsymbol{U} : \boldsymbol{U} \odot \bar{\boldsymbol{U}} \in \Omega_n$. Given the circuit parameters θ (typically in the hundreds) and $p \in \mathbb{R}$, QontOT obtains a DSM via $\boldsymbol{U}(p;\theta) \odot \bar{\boldsymbol{U}}(p;\theta)$.

Furthermore, QontOT requires the DSM dimension n to be a power of 2. While this may be prohibitive within a Transformer (because sequence length T may differ), it can be mitigated by padding. Padding to powers of two is a common technique to maximize hardware efficiency.

QR Decomposition

As highlighted above, any unitary $\boldsymbol{U}$ can provide a DSM by taking $\boldsymbol{U} \odot \bar{\boldsymbol{U}}$. While there are many ways to obtain a basis, we choose a QR decomposition $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{R}$, in which case $\boldsymbol{R}$ is upper triangular.

QR has the advantage that, when implemented with Gram-Schmidt, it is differentiable if $\boldsymbol{M}$ is full-rank. However note that for long-context applications the rank is rarely full because the query and key matrix have $d = \frac{d_{embed}}{n_{heads}}$ rows and typically $\boldsymbol{M} \in \mathbb{R}^{T \times T}$ has rank $\min\{d, T\}$, implying that $\boldsymbol{M}$ only has full rank when $d \geq T$.

Soundness and Completeness

- **Soundness**: Does the operator always produce a DSM? Given that $\boldsymbol{U} \odot \bar{\boldsymbol{U}} \in \Omega_n$, QontOT always yields a DSM. Similarly for the QR decomposition. Instead, Sinkhorn's algorithm (SA) may fail to produce a DSM if the input matrix is not positive.
- **Completeness**: Can the operator produce all possible DSMs?

Empirical analysis

For a $n \times n$ matrix and a discretiztion step $d \in \mathcal{N}_+$, we sample each column from a discretized n-dimensional hybercube with d^n points, yielding d^{n^2} unique matrices. For $n = 4$ and $d = 3$ we obtain $3^{16} \approx 43M$ matrices and computed the DSM for each input, before rounding to third decimal place.

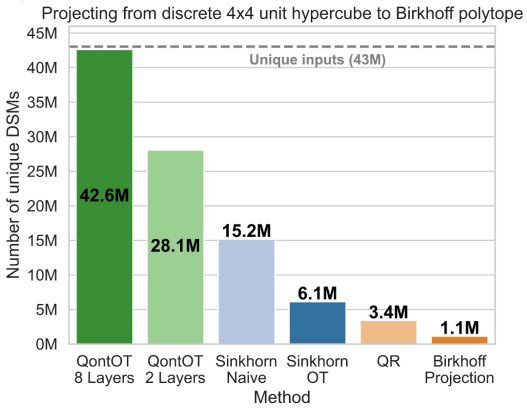


Figure 2. Number of unique DSMs obtained with each operator after exhaustively iterating over a discretized unit hypercube. With only 8 layers, QontOT produces a unique DSM for every possible input, unlike all other methods.

Empirical analysis

A powerful operator needs to possess two further characteristics:

- The information has to be preserved. Obtaining unique DSMs is useless if they destroy information from the input matrix. To assess this, we measured the Frobenius norm of the residuals between input and output matrix.
- The low entropy has to be avoided because it causes vanishing gradients and destabilizes Transformer training.

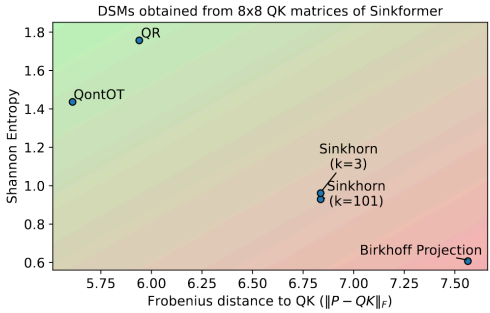


Figure 3. Shannon entropy of different flavors of doubly stochastic attention against a measure of distance preservation – the Frobenius norm of the difference between unnormalized attention $\mathbf{QK}$ and their obtained DSM $\mathbf{P}$. QontOT uses 2 circuit layers.

Outline

Experimental Setup

Running the circuit on quantum hardware requires $\Omega(n^2/\epsilon^2)$ shots to obtain satisfactory sampling error. Assuming a precision of $\epsilon = 0.01$, this is in the order of $640k$ shots per sample. Since quantum hardware operates on kHz frequency, execution and online optimization on hardware is unfortunately not (yet) feasible. Therefore, we perform exact statevector simulation with *Qiskit* and implement three circuit training strategies:

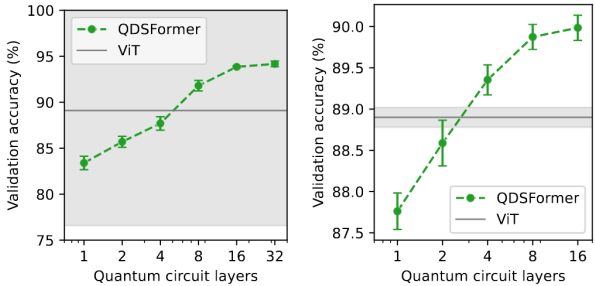
- **Differentiable:** Joint optimization of circuit and Transformer parameters through backpropagation, akin to integrating the circuit as a neural network layer.
- **Mixed:** A mixed strategy where Transformer training is interleaved with 200 steps of gradient-free circuit optimization with *Nevergrad* on a per-epoch basis.
- **Static:** The circuit is used in pure inference mode with parameters obtained from a 24-qubit DSM prediction experiment on quantum hardware.

Datasets

We evaluate the different ViTs on MNIST, Fashion MNIST, seven datasts from the MedMNIST benchmark and a compositional task requiring multistep reasoning.

In that task, a 2×2 grid contains two MNIST digits (upper left and lower right) and two FashionMNIST items (upper right and lower left). If the digits have equal value, the label is the upper right fashion item, otherwise it is the bottom left fashion item. Performance typically ramps up quickly to $\sim 50\%$ because the model learns to attend one (and only one) of the FashionMNIST images. Upon continued training with a long saturation phase, a ViT suddenly grasps the relationship of the MNIST digits to the classification task and then climbs rapidly to a $90 - 95\%$ accuracy. The moment of abrupt improvement is called "Eureka moment".

Experiments



(a) 1-ViT-Layer on MNIST. **(b)** 2-ViT-Layer on FashionMNIST.

Figure 4. Comparison of ViT and QDSFormer while varying the circuit depth. Mean/std from 5 trainings are shown. Within (**a**) and (**b**) all models use the same number of trainable parameters.

Experiments

Table 1. L-layered ViT validation accuracy on FashionMNIST for different attention methods. QontOT uses 16 circuit layers. Mean/std computed from 5 trainings.

L	**Softmax**	**Softmax**$_{\sigma^2}$	**QR**	**QontOT**	**Sinkhorn**
1	$\underline{86.5}_{\pm 0.19}$	$75.3_{\pm 4.61}$	$\mathbf{87.1}_{\pm 0.26}$	$85.6_{\pm 0.10}$	$84.2_{\pm 3.64}$
2	$88.9_{\pm 0.12}$	$84.6_{\pm 2.11}$	$\underline{89.3}_{\pm 0.07}$	$\mathbf{90.0}_{\pm 0.15}$	$89.1_{\pm 0.73}$
3	$\underline{89.4}_{\pm 0.28}$	$86.3_{\pm 2.69}$	$\underline{89.4}_{\pm 0.11}$	$\mathbf{90.3}_{\pm 0.13}$	$\underline{89.4}_{\pm 0.77}$
4	$\underline{89.7}_{\pm 0.29}$	$87.1_{\pm 1.15}$	$89.5_{\pm 0.07}$	$\mathbf{90.3}_{\pm 0.14}$	$89.1_{\pm 1.08}$

Table 2. Identical to Table 1 but for MNIST.

L	**Softmax**	**Softmax**$_{\sigma^2}$	**QR**	**QontOT**	**Sinkhorn**
1	$89.1_{\pm 12.5}$	$66.7_{\pm 22.5}$	$\mathbf{96.6}_{\pm 0.10}$	$93.9_{\pm 0.11}$	$\underline{94.3}_{\pm 1.97}$
2	$98.1_{\pm 0.33}$	$93.0_{\pm 4.57}$	$\underline{98.3}_{\pm 0.13}$	$\mathbf{98.4}_{\pm 0.05}$	$98.2_{\pm 0.27}$
3	$\underline{98.6}_{\pm 0.11}$	$97.7_{\pm 0.65}$	$\underline{98.6}_{\pm 0.13}$	$\mathbf{98.7}_{\pm 0.06}$	$\underline{98.6}_{\pm 0.12}$
4	$\mathbf{98.8}_{\pm 0.10}$	$97.9_{\pm 0.71}$	$\underline{98.7}_{\pm 0.11}$	$\mathbf{98.8}_{\pm 0.07}$	$97.9_{\pm 1.57}$

Experiments

Table 3. Test accuracy for different MedMNIST datasets across five attention types in a 2-layer ViT. QontOT uses 16 circuit layers.

MedMNIST dataset	Softmax	Softmax$_{\sigma^2}$	QR	QontOT	Sinkhorn
OCT	**64.4**$_{\pm 1.6}$	43.6$_{\pm 3.0}$	62.5$_{\pm 0.9}$	61.6$_{\pm 0.6}$	55.1$_{\pm 5.2}$
Pneumonia	84.2$_{\pm 0.8}$	84.7$_{\pm 2.0}$	84.3$_{\pm 0.7}$	**86.1**$_{\pm 1.0}$	83.0$_{\pm 1.5}$
Tissue	60.0$_{\pm 0.2}$	49.4$_{\pm 1.2}$	59.0$_{\pm 0.1}$	**60.6**$_{\pm 0.1}$	56.9$_{\pm 2.0}$
OrganA	78.8$_{\pm 0.5}$	73.6$_{\pm 1.7}$	78.4$_{\pm 0.6}$	**81.2**$_{\pm 0.3}$	77.0$_{\pm 2.5}$
OrganC	79.8$_{\pm 0.5}$	71.7$_{\pm 7.3}$	79.6$_{\pm 0.3}$	**82.7**$_{\pm 0.5}$	79.7$_{\pm 1.0}$
OrganS	64.4$_{\pm 0.6}$	59.3$_{\pm 0.9}$	62.6$_{\pm 0.8}$	**68.1**$_{\pm 0.6}$	63.5$_{\pm 0.9}$
Breast	79.6$_{\pm 2.0}$	78.2$_{\pm 2.2}$	**81.3**$_{\pm 2.9}$	80.0$_{\pm 1.1}$	80.1$_{\pm 0.8}$
Mean	73.0	65.8	72.5	**74.3**	70.8

Experiments

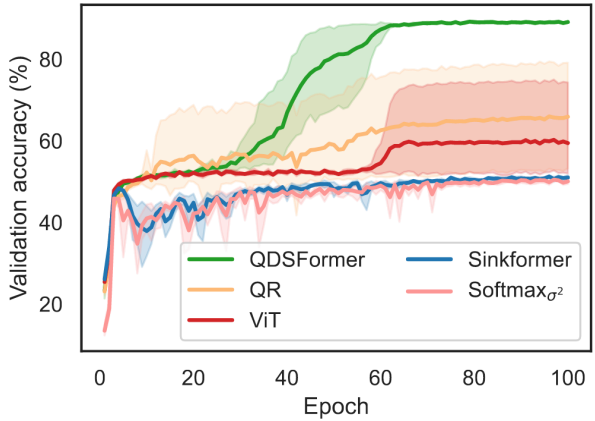


Figure 5. Validation accuracy for each training epoch, highlighting the abrupt learning referred to as Eureka Moment (EM) on the compositional dataset. Confidence bounds from 5 runs.

Outline

Conclusion

- This paper proposed the quantum doubly stochastic Transformer by connecting the centerpiece of a novel variational quantum circuit (devised for optimal transport) with the Transformer, through doubly-stochastic attention improves performance in Transformers.
- However, this method requires a large amount of data for training and is not suitable for larger models.