ESPFormer: Doubly-Stochastic Attention with Expected Sliced Transport Plans

VANDERBILT UNIVERSITY

Duke UNIVERSITY

Check out our project page!

Ashkan Shahbazi, Elaheh Akbari, Darian Salehi, Xinran Liu, Navid NaderiAlizadeh, Soheil Kolouri

ICML International Conference On Machine Learning

Motivation

- Self-attention often collapses onto a few tokens, throttling information flow. Making the attention matrix doubly-stochastic restores balance, but existing Sinkhorn-based solutions are slow and memory-intensive. We need a cheaper way to enforce this structure.

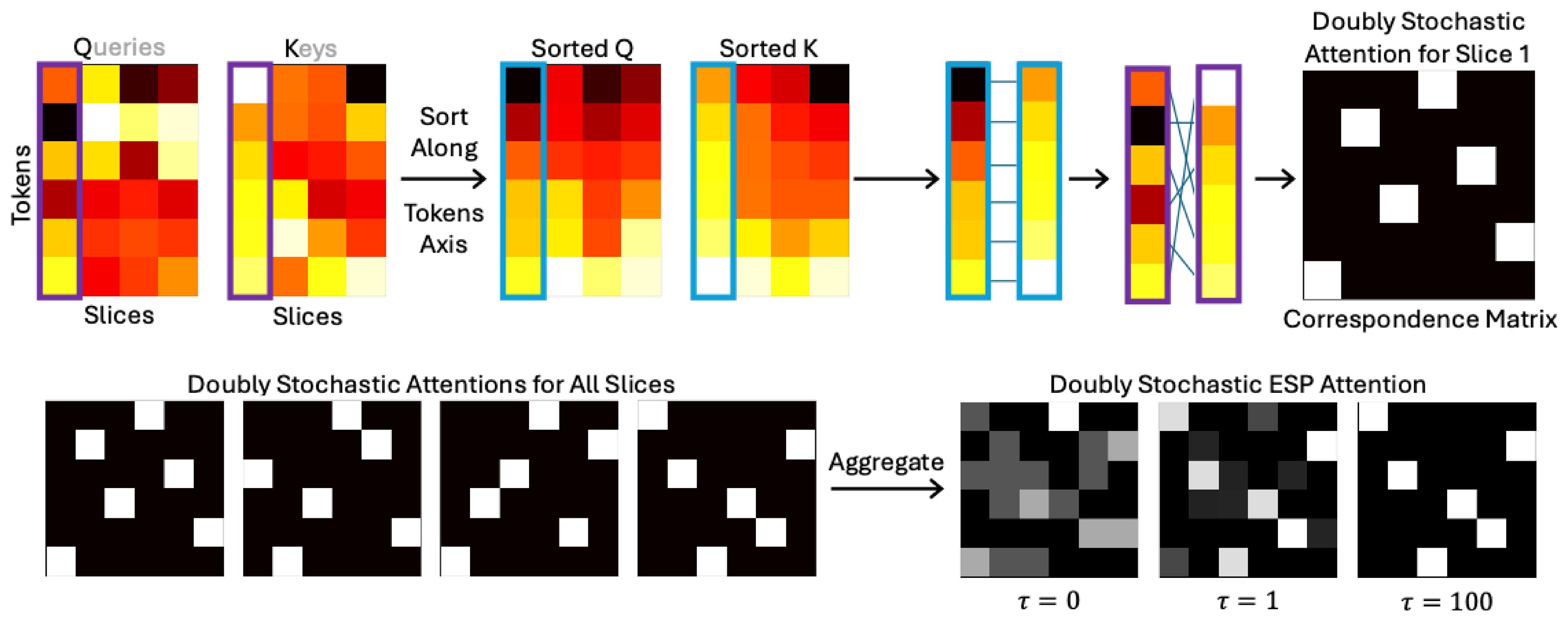
Contributions

- **ESPFormer**: Expected Sliced Transport–based, doubly-stochastic attention with tunable sparsity; annealing → hard sorting yields exact matrices in $\mathcal{O}(mNlogN)$
- Outperforms Vanilla Transformer and Sinkformer in both accuracy and compute; drops straight into pre-trained or differential-attention models with minimal fine-tuning.

Background

- In the space of uniform discrete probability measures supported on N particles in $\mathbb{R}^d$, that is $\mathcal{P}_{(N)}(\mathbb{R}^d) = \left\{\frac{1}{N}\sum_{i=1}^{N}\delta_{x_i} \mid x_i \in \mathbb{R}^d, \forall i \in \{1,\dots,N\}\right\}$, for $\mu^1 = \frac{1}{N}\sum_{i=1}^{N}\delta_{x_i}, \mu^2 = \frac{1}{N}\sum_{j=1}^{N}\delta_{y_j} \in \mathcal{P}_{(N)}(\mathbb{R}^d)$, let $\xi_\theta, \tau_\theta \in S_N$ be the sorted indices such that
$$\theta \cdot x_{\xi_\theta^{-1}(1)} \le \theta \cdot x_{\xi_\theta^{-1}(2)} \le \cdots \le \theta \cdot x_{\xi_\theta^{-1}(N)};$$
$$\theta \cdot y_{\tau_\theta^{-1}(1)} \le \theta \cdot y_{\tau_\theta^{-1}(2)} \le \cdots \le \theta \cdot y_{\tau_\theta^{-1}(N)},$$
the optimal matching from $\theta_\# \mu^1$ to $\theta_\# \mu^2$ is given by
$$\theta \cdot x_{\xi_\theta^{-1}(i)} \longmapsto \theta \cdot y_{\tau_\theta^{-1}(1)}, \forall i \in \{1,\dots,N\},$$ with a unique OT plan $\Lambda_\theta^{\mu^1,\mu^2}$
- Lifting Transport Plans $\Lambda_\theta^{\mu^1,\mu^2}$ lifted to $\gamma_\theta^{\mu^1,\mu^2}$
$$u_\theta^{\mu^1,\mu^2}(x,y) = \frac{p(x)q(y)}{P(\overline{x^\theta})Q(\overline{y^\theta})}\Lambda_\theta^{\mu^1,\mu^2}(\{(\overline{x^\theta},\overline{y^\theta})\})$$
with $\theta_\#\mu^1 = \sum_{\overline{x^\theta}\in R/\sim_\theta} P(\overline{x^\theta})\delta_{\overline{x^\theta}}$ and $\theta_\#\mu^2 = \sum_{\overline{y^\theta}\in R/\sim_\theta} Q(\overline{y^\theta})\delta_{\overline{y^\theta}}$
- **E**xpected **S**liced Transport **P**lan (given $\sigma \in \mathcal{P}(\mathbb{S}^{d-1})$)
$$\bar{\gamma}^{\mu^1,\mu^2} := \mathbb{E}_{\theta\sim\sigma}\left[\gamma_\theta^{\mu^1,\mu^2}\right]$$

ESP Doubly-Stochastic Attention



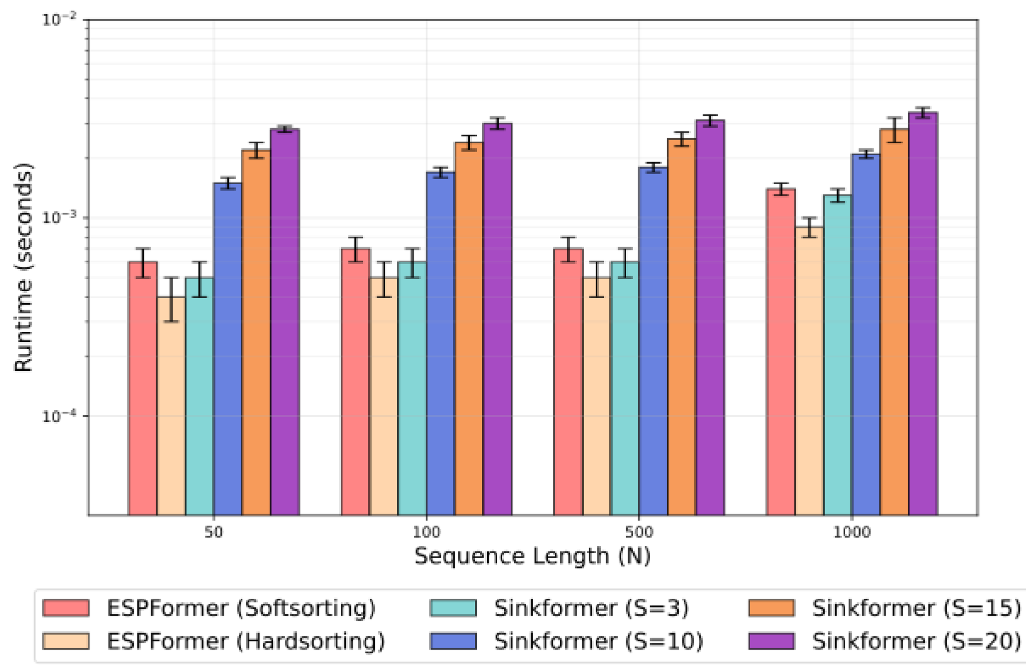
ESP integrates slicing into keys/queries, treating each dimension as a learnable slice. Tokens are (soft) sorted per slice, generating dimension-wise doubly-stochastic correspondence matrices. Aggregating these matrices yields the final attention matrix.

$$\mu^Q = \frac{1}{N}\sum_{i=1}^{N}\delta_{q_i}, \ \mu^K = \frac{1}{N}\sum_{j=1}^{N}\delta_{k_j} \ \rightarrow \text{ESP Attention}(Q,K,V) = V * \bar{\gamma}^{\mu^Q,\mu^K}$$

- $\text{SoftSort}_t^d(v) = \text{softmax}\left(\frac{-d(\text{sort}(v)\,\mathbf{1}^T, \mathbf{1}v^T)}{t}\right)$ is used for differentiability of the Transport plans.
- Keys and Queries are themselves learned, optimizing Θ is unnecessary. We propose using axis-aligned slices by setting $\Theta = I_{m\times m}$.
- A temperature annealing schedule enables the transition from soft to hard sorting during fine-tuning

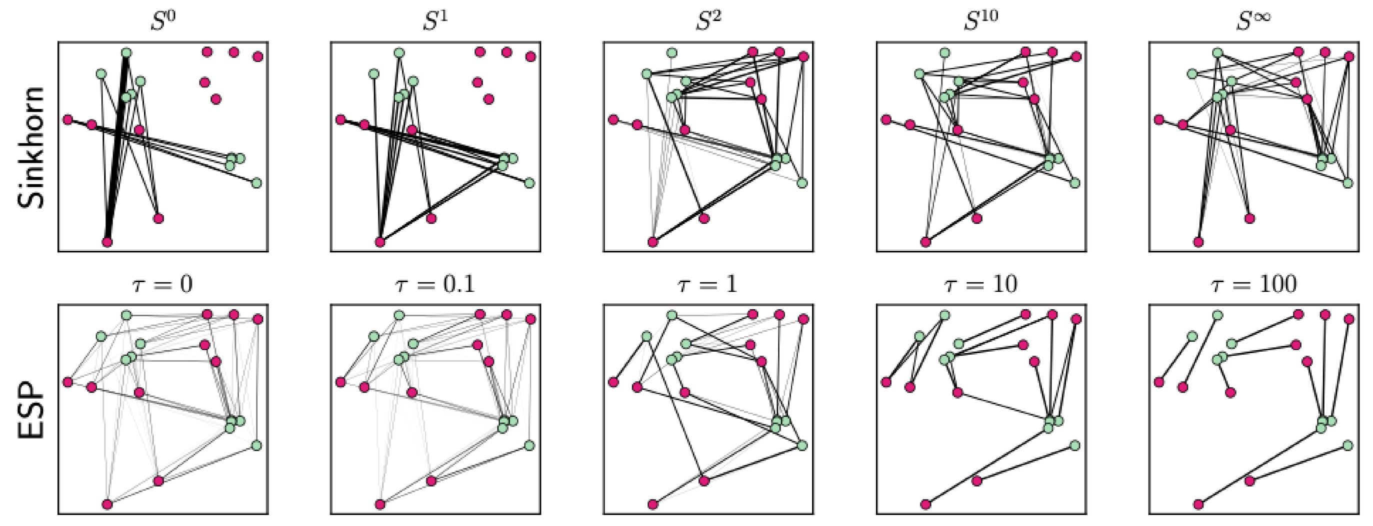
Computational Efficiency

- ESPFormer runtime complexity:
  - Soft Sorting: $\mathcal{O}(mN(N+d))$
  - Hard Sorting: $\mathcal{O}(mNlogN)$
- Sinkformer runtime for S steps: $\mathcal{O}((S+m)N^2)$



Runtime comparison of ESPFormer and Sinkformer (iterations S) for sequence lengths N∈{50,100,500,1000}, averaged over 10 runs.

Numerical Experiments



Attention weights between keys (red) and queries (green) computed by Sinkhorn's algorithm (top) and Expected Sliced Transport Plans (bottom). Sinkhorn at iteration S reduces to classic self-attention. Line width indicates attention weight magnitude.

Data Fraction	Baselines			ESPFormer		
	Sinkformer	DiffTransformer	Transformer	Initial Soft Sort	Sharp Soft Sort	Hard Sort
1%	55.07 ± 3.34	53.78 ± 0.28	49.71 ± 0.31	55.66 ± 3.95	57.86 ± 3.77	$\mathbf{58.52 \pm 3.73}$
10%	69.56 ± 0.32	67.34 ± 0.11	57.25 ± 0.22	71.49 ± 0.43	72.22 ± 0.37	$\mathbf{72.71 \pm 0.36}$
25%	74.56 ± 0.58	74.86 ± 0.17	72.25 ± 0.16	75.40 ± 0.38	75.92 ± 0.31	$\mathbf{75.92 \pm 0.28}$
100%	79.12 ± 0.17	78.85 ± 0.11	78.49 ± 0.09	79.47 ± 0.12	80.61 ± 0.11	$\mathbf{81.23 \pm 0.11}$

Average and standard deviation (over 3 runs) of ESPFormer's classification accuracy (%) vs. baselines on the Cats and Dogs dataset under varying data availability. ESPFormer's performance is reported in three modes: initial soft sort, sharp soft sort, and hard sort.

Model	Best	Median	Mean	Worst
Set Transformer*	87.8	86.3	85.8	84.7
Set DiffTransformer	89.0	88.7	88.7	88.6
Set Sinkformer*	89.1	88.4	88.3	88.1
Set ESPFormer	**89.6**	**89.5**	**89.4**	**89.1**
Point Cloud Transformer*	**93.2**	92.5	92.5	92.3
Point Cloud DiffTransformer	93.1	92.8	**92.7**	**92.6**
Point Cloud Sinkformer*	93.1	92.8	**92.7**	92.5
Point Cloud ESPFormer	**93.2**	**92.9**	**92.7**	**92.6**

Test accuracy (%) on the ModelNet40 dataset over 4 runs.

	Model	Plug-and-Play	Fine-Tune Boost
Transformer	Transformer	33.40	34.61
	Sinkformer	33.36*	34.61
	ESPFormer	33.38*	**34.64**
DiffTransformer	DiffTransformer	33.85*	34.78
	Sinkformer	33.67*	34.81
	ESPFormer	33.72*	**34.83**

Median BLEU scores over 4 runs on IWSLT14 German-to-English for Transformer/DiffTransformer baselines. Results marked * indicate use of an alternate attention module.

Model	Best	Median	Mean	Worst
Transformer	71.50	71.35	71.31	71.10
DiffTransformer	**72.60**	72.35	72.31	72.00
Sinkformer	72.40	72.30	72.23	71.90
ESPFormer	**72.60**	**72.40**	**72.36**	**72.10**

Test accuracy (%) for Sentiment Analysis on TweetEval.

Model	Best	Median	Mean	Worst
Transformer	85.30	85.25	85.25	85.20
DiffTransformer	**85.50**	85.45	85.45	**85.40**
Sinkformer	85.40	85.39	85.37	85.30
ESPFormer	**85.50**	**85.50**	**85.47**	**85.40**

Test accuracy (%) for Sentiment Analysis on IMDb.

	$L = 1$	$L = 8$	$L = 32$	$L = 64$	$L = 128$
			$\tau = 0.1$		
Learnable	74.30 ± 0.48	78.70 ± 0.32	79.10 ± 0.22	78.40 ± 0.26	76.20 ± 0.42
Frozen	66.50 ± 0.52	72.80 ± 0.38	78.30 ± 0.18	79.20 ± 0.30	79.60 ± 0.28
Axis-Aligned	–	–	–	79.47 ± 0.12	–
			$\tau = 1.0$		
Learnable	74.30 ± 0.48	79.07 ± 0.30	78.20 ± 0.35	77.80 ± 0.21	74.10 ± 0.46
Frozen	66.50 ± 0.52	73.11 ± 0.43	77.95 ± 0.25	78.80 ± 0.29	78.40 ± 0.27
Axis-Aligned	–	–	–	78.85 ± 0.31	–
			$\tau = 10$		
Learnable	74.30 ± 0.48	79.15 ± 0.29	78.06 ± 0.27	77.10 ± 0.23	74.15 ± 0.44
Frozen	66.50 ± 0.52	73.45 ± 0.41	76.85 ± 0.24	77.85 ± 0.30	78.10 ± 0.26
Axis-Aligned	–	–	–	77.75 ± 0.32	–

Accuracy (%) over three runs across slicer types, slice counts (L), and inverse temperature (τ).