Rethinking Attention with Performers

Attention with linear time

허세민

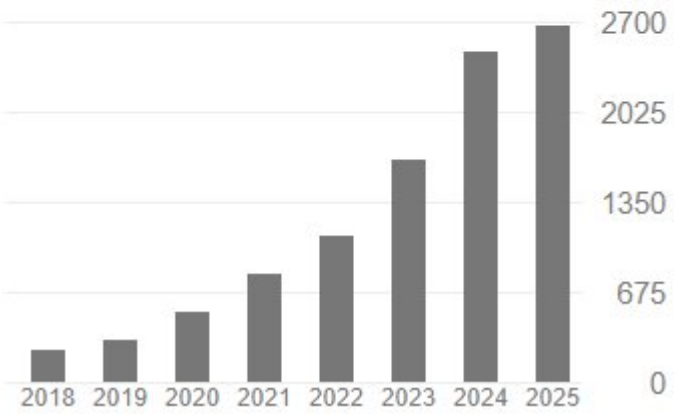
Author

Krzysztof Choromanski

Google DeepMind Robotics & Columbia University
columbia.edu의 이메일 확인됨 - 홈페이지

robotics reinforcement learning efficient Transformers quasi Monte Carlo methods

인용		모두 보기
	전체	2020년 이후
서지정보	10501	9606
h-index	40	37
i10-index	82	76



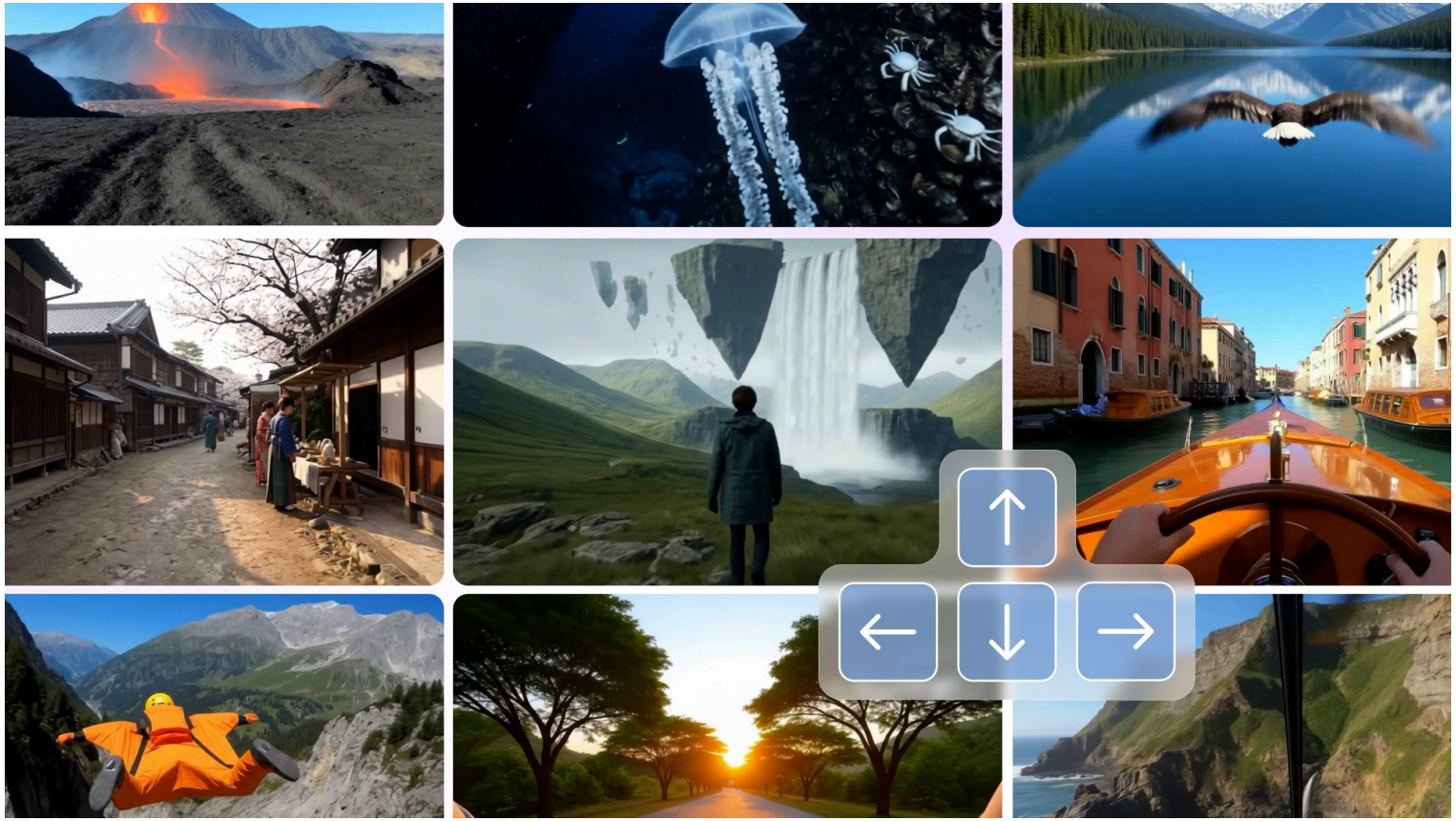
Rethinking attention with performers K Choromanski, V Likhosherstov, D Dohan, X Song, A Gane, T Sarlos, ... arXiv preprint arXiv:2009.14794	2615	2020
Rt-2: Vision-language-action models transfer web knowledge to robotic control B Zitkovich, T Yu, S Xu, P Xu, T Xiao, F Xia, J Wu, P Wohlhart, S Welker, ... Conference on Robot Learning, 2165-2183	1860	2023
Socratic models: Composing zero-shot multimodal reasoning with language A Zeng, M Attarian, B Ichter, K Choromanski, A Wong, S Welker, ... arXiv preprint arXiv:2204.00598	654	2022
Explaining how a deep neural network trained with end-to-end learning steers a car M Bojarski, P Yeres, A Choromanska, K Choromanski, B Firner, L Jackel, ... arXiv preprint arXiv:1704.07911	614	2017
Orthogonal random features FXX Yu, AT Suresh, KM Choromanski, DN Holtmann-Rice, S Kumar Advances in neural information processing systems 29	281	2016
End to end learning for self-driving cars. arXiv 2016 M Bojarski, D Del Testa, D Dworakowski, B Firner, B Flepp, P Goyal, ... arXiv preprint arXiv:1604.07316 103	268	2016
A theoretical and empirical comparison of gradient approximations in derivative-free optimization AS Berahas, L Cao, K Choromanski, K Scheinberg Foundations of Computational Mathematics 22 (2), 507-560	245	2022
Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities G Comanici, E Bieber, M Schaekermann, I Pasupat, N Sachdeva, I Dhillon, ... arXiv preprint arXiv:2507.06261	200	2025
Effective diversity in population based reinforcement learning J Parker-Holder, A Pacchiano, KM Choromanski, SJ Roberts Advances in Neural Information Processing Systems 33, 18050-18062	198	2020
Structured evolution with compact architectures for scalable policy optimization K Choromanski, M Rowland, V Sindhwani, R Turner, A Weller International Conference on Machine Learning, 970-978	170	2018
Es-maml: Simple hessian-free meta learning X Song, W Gao, Y Yang, K Choromanski, A Pacchiano, Y Tang arXiv preprint arXiv:1910.01215	151	2019

Contents

Intro

Towards world Simulation

"DeepMind has ambitious plans to make massive generative models that simulate the world," Brooks wrote Monday morning. "I'm hiring for a new team with this mission."



Titans

Titans: Learning to Memorize at Test Time

Ali Behrouz[†], Peilin Zhong[†], and Vahab Mirrokni[†]

[†]Google Research

{alibehrouz, peilinz, mirrokni}@google.com

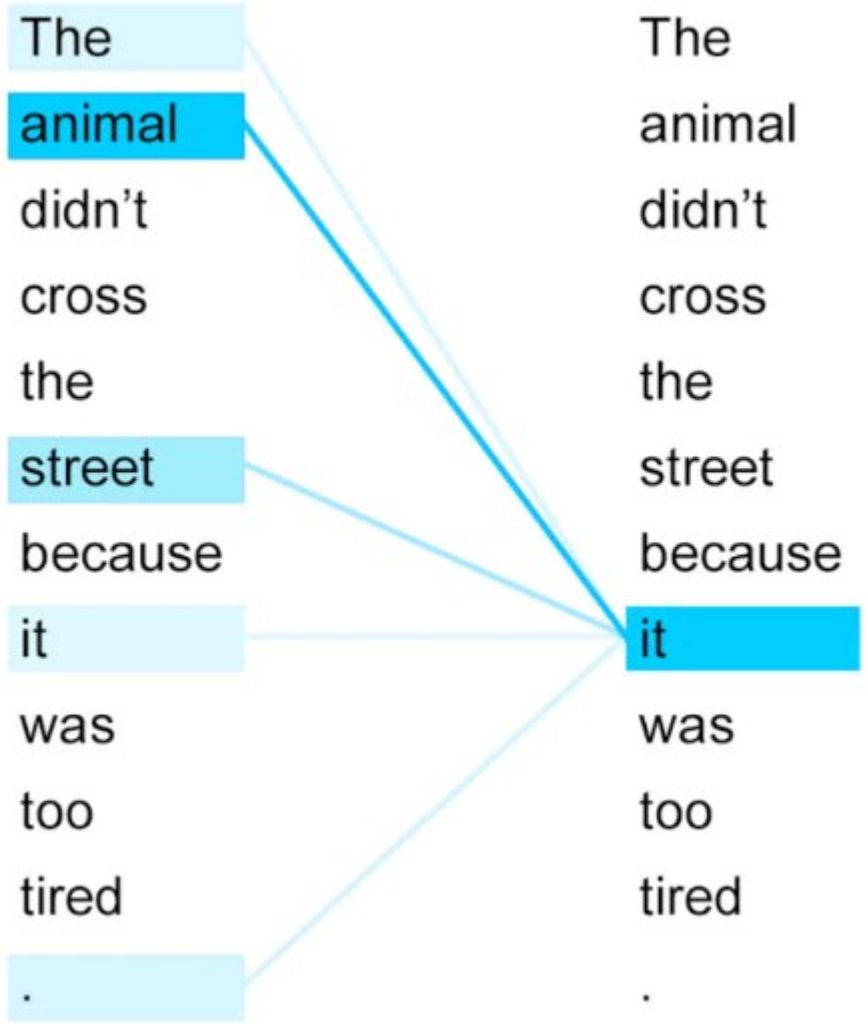
Prior Knowledge

- ✓ Test-Time Training (TTT)
  - 추론 시점(test time)에서 모델이 스스로 적응 할 수 있도록 하는 방법
  - 새로운 환경이나 분포 변화에 맞게 온라인 학습처럼 동작
  - Titan도 추론 시점에서 모델을 적응 시키는 전략을 활용함

- ✓ Linear Attention
  - 긴 시퀀스를 다룰 때, 기존 Attention의 $O(n^2)$복잡도를 줄이려는 접근
  - 시퀀스 길이가 길어져도 효율적으로 처리 가능
  - Titan은 효율적인 attention 구조를 필요로 하며, Linear Attention 아이디어가 그 기반이 됨

Attention

Attention



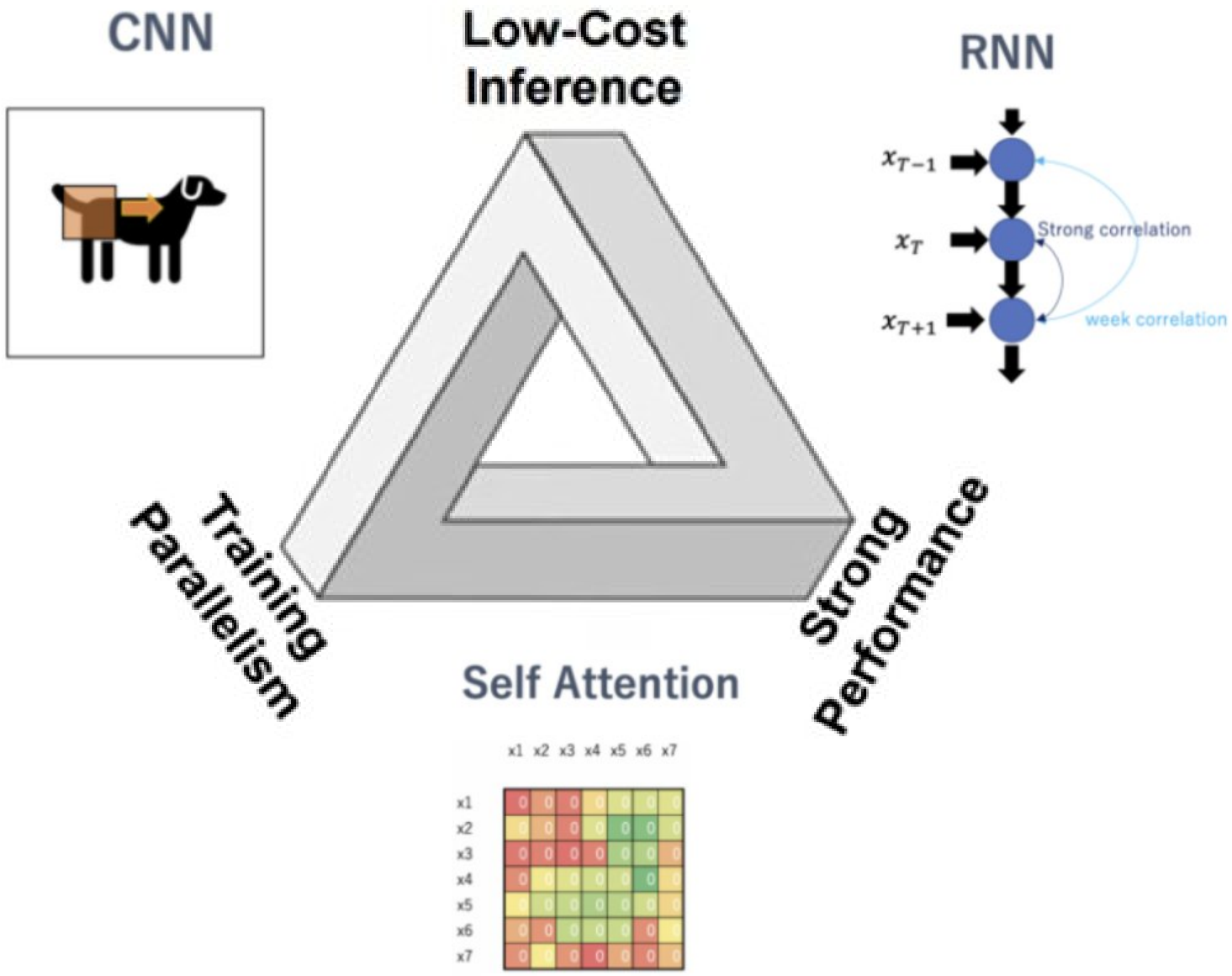
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

Advantages
- 장거리 의존성을 효과적으로 포착할 수 있음
- 학습 과정에서 완전 병렬화가 가능함

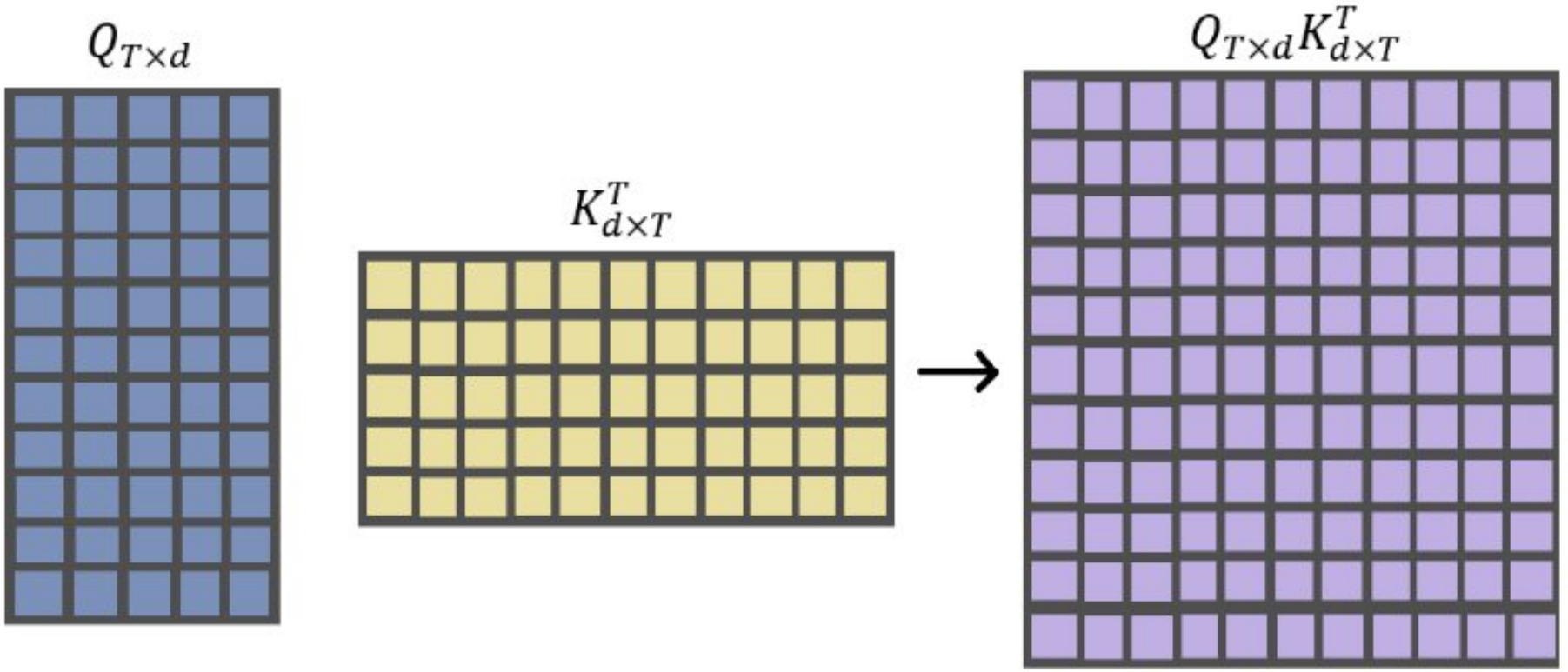
Limitations
- 시간과 메모리 복잡도가 **제곱(Quadratic, $O(n^2)$)**
- **매우 긴 시퀀스**를 다루기 어려움

Alternatives
- FlashAttention
- **Performer**

Penrose Triangle



Complexity



Transformers are RNNs

Linear Attention

$$V_i' = \frac{\sum_{j=1}^{N} \mathrm{sim}\left(Q_i, K_j\right) V_j}{\sum_{j=1}^{N} \mathrm{sim}\left(Q_i, K_j\right)}.$$

Attention

$$V' = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{D}}\right) V.$$

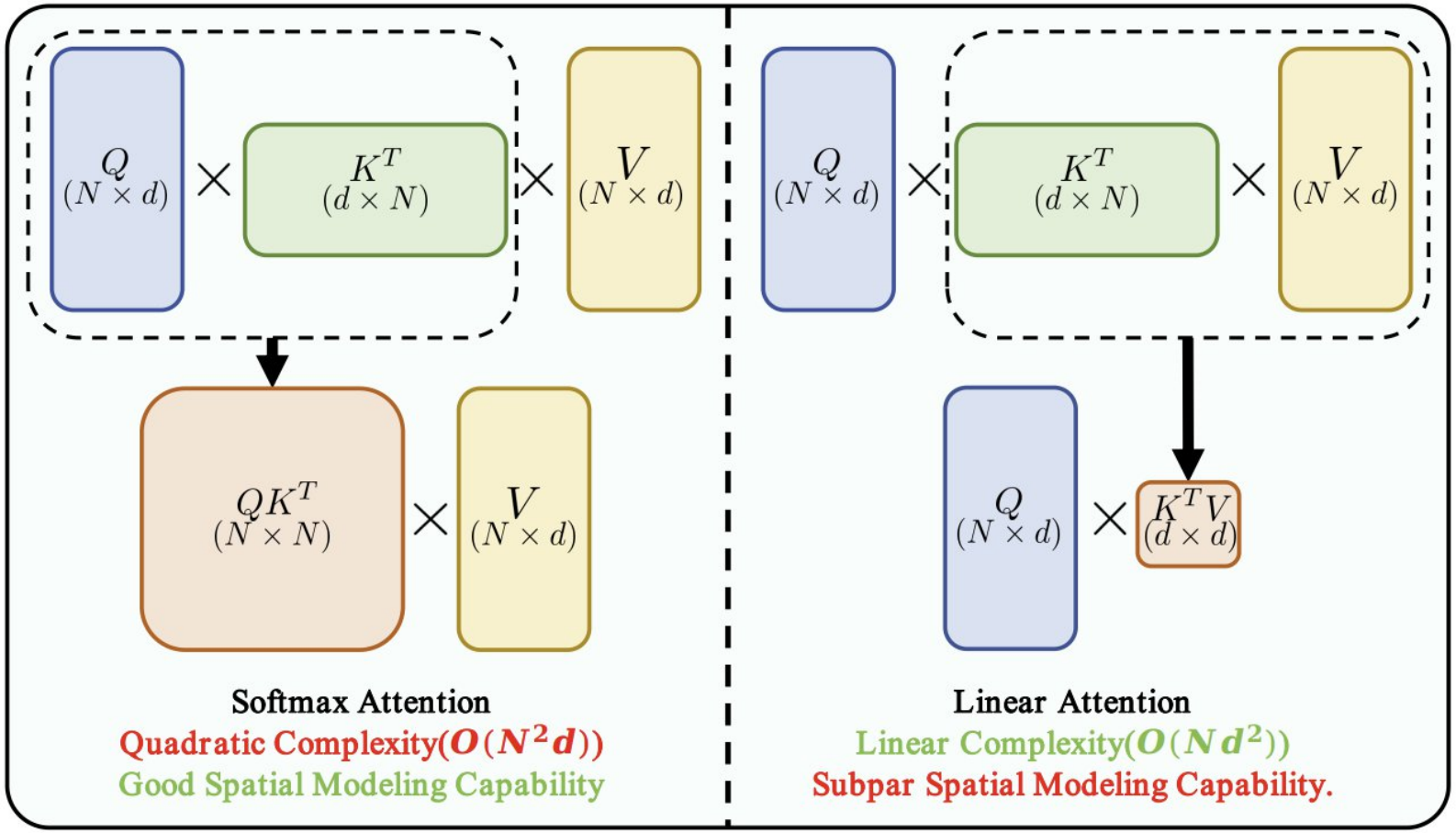
Linear Attention

$$V_i' = \frac{\sum_{j=1}^{N} \phi\left(Q_i\right)^T \phi\left(K_j\right) V_j}{\sum_{j=1}^{N} \phi\left(Q_i\right)^T \phi\left(K_j\right)}$$

$$\downarrow$$

$$V_i' = \frac{\phi\left(Q_i\right)^T \sum_{j=1}^{N} \phi\left(K_j\right) V_j^T}{\phi\left(Q_i\right)^T \sum_{j=1}^{N} \phi\left(K_j\right)}.$$

Linear Attention



$$\text{softmax}\left(\frac{QK^T}{\sqrt{D}}\right)V. \qquad \phi(Q)\left(\phi(K)^T V\right).$$

Transformers are RNNs

$$S_i = \sum_{j=1}^{i} \phi\left(K_j\right) V_j^T,$$

$$Z_i = \sum_{j=1}^{i} \phi\left(K_j\right),$$

$$s_0 = 0,$$

$$z_0 = 0,$$

$$s_i = s_{i-1} + \phi\left(x_i W_K\right)\left(x_i W_V\right)^T,$$

$$z_i = z_{i-1} + \phi\left(x_i W_K\right),$$

$$y_i = f_l\left(\frac{\phi\left(x_i W_Q\right)^T s_i}{\phi\left(x_i W_Q\right)^T z_i} + x_i\right).$$

Transformers are RNNs

Advantages

- 선형 시간/메모리 복잡도: $O\left(n^2\right) \rightarrow O\left(n\right)$

Recurrent formulation 가능 → 긴 시퀀스에서도 순차적으로 업데이트

기존 *Transformer* 대비 긴 문맥 처리 효율 ↑

병렬화 구조를 유지하면서도 *RNN*적 특성 획득

Limitations

*Softmax Attention*의 정확한 표현 불가 → 근사 오차 발생

커널 함수 선택에 따라 성능이 크게 변동 (논문에서는 $\phi\left(x\right) = \mathrm{elu}(x) + 1$일 수 있음)

수치적 안정성 문제: 긴 시퀀스에서 누적

실제 응용에서는 성능 저하 가능성

Performer

Motivation

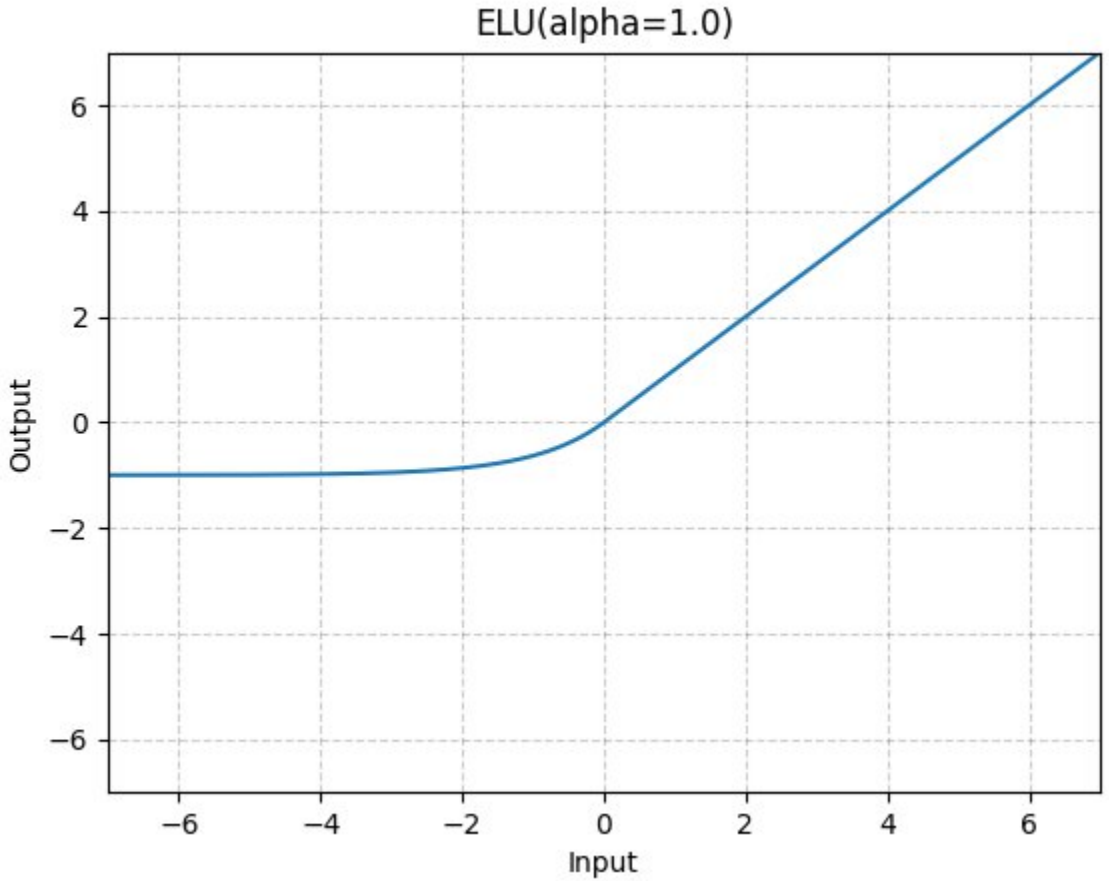
✓ 수치적 불안정성
- 근사된 Attention 값이 발산하거나 collapse 가능

✓ 근사 품질 저하
- Softmax kernel을 정확히 근사하지 못함

✓ 분산 문제
- 근사 분산이 매우 커져서 불안정

$$\phi(x) = \mathrm{elu}(x) + 1$$

$$V_i' = \frac{\phi(Q_i)^T \sum_{j=1}^{N} \phi(K_j) V_j^T}{\phi(Q_i)^T \sum_{j=1}^{N} \phi(K_j)}.$$



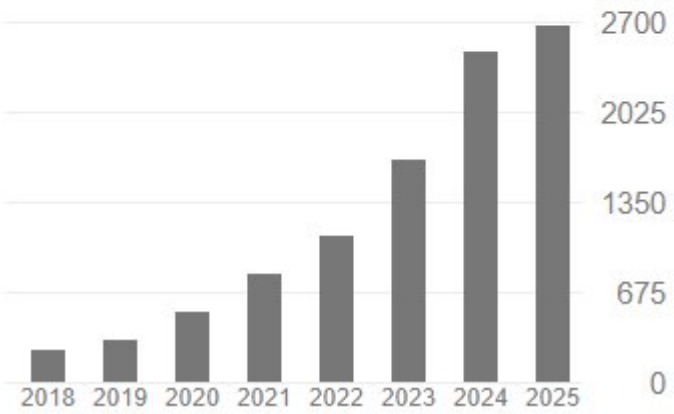
Author

Krzysztof Choromanski

Google DeepMind Robotics & Columbia University
columbia.edu의 이메일 확인됨 - 홈페이지

robotics reinforcement learning efficient Transformers quasi Monte Carlo methods

인용		모두 보기
	전체	2020년 이후
서지정보	10501	9606
h-index	40	37
i10-index	82	76



Rethinking attention with performers K Choromanski, V Likhosherstov, D Dohan, X Song, A Gane, T Sarlos, ... arXiv preprint arXiv:2009.14794	2615	2020
Rt-2: Vision-language-action models transfer web knowledge to robotic control B Zitkovich, T Yu, S Xu, P Xu, T Xiao, F Xia, J Wu, P Wohlhart, S Welker, ... Conference on Robot Learning, 2165-2183	1860	2023
Socratic models: Composing zero-shot multimodal reasoning with language A Zeng, M Attarian, B Ichter, K Choromanski, A Wong, S Welker, ... arXiv preprint arXiv:2204.00598	654	2022
Explaining how a deep neural network trained with end-to-end learning steers a car M Bojarski, P Yeres, A Choromanska, K Choromanski, B Firner, L Jackel, ... arXiv preprint arXiv:1704.07911	614	2017
Orthogonal random features FXX Yu, AT Suresh, KM Choromanski, DN Holtmann-Rice, S Kumar Advances in neural information processing systems 29	281	2016
End to end learning for self-driving cars. arXiv 2016 M Bojarski, D Del Testa, D Dworakowski, B Firner, B Flepp, P Goyal, ... arXiv preprint arXiv:1604.07316 103	268	2016
A theoretical and empirical comparison of gradient approximations in derivative-free optimization AS Berahas, L Cao, K Choromanski, K Scheinberg Foundations of Computational Mathematics 22 (2), 507-560	245	2022
Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities G Comanici, E Bieber, M Schaekermann, I Pasupat, N Sachdeva, I Dhillon, ... arXiv preprint arXiv:2507.06261	200	2025
Effective diversity in population based reinforcement learning J Parker-Holder, A Pacchiano, KM Choromanski, SJ Roberts Advances in Neural Information Processing Systems 33, 18050-18062	198	2020
Structured evolution with compact architectures for scalable policy optimization K Choromanski, M Rowland, V Sindhwani, R Turner, A Weller International Conference on Machine Learning, 970-978	170	2018
Es-maml: Simple hessian-free meta learning X Song, W Gao, Y Yang, K Choromanski, A Pacchiano, Y Tang arXiv preprint arXiv:1910.01215	151	2019

Motivation

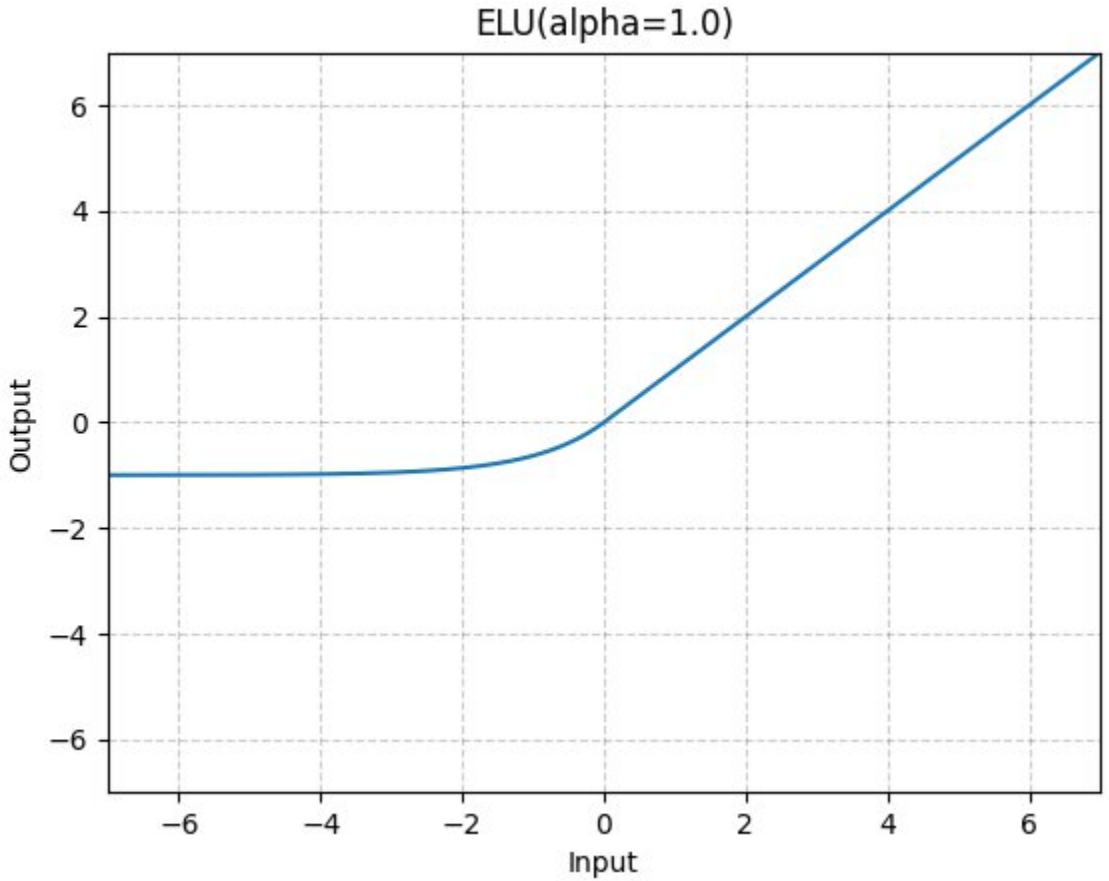
✓ 수치적 불안정성
- 근사된 Attention 값이 발산하거나 collapse 가능

✓ 근사 품질 저하
- Softmax kernel을 정확히 근사하지 못함

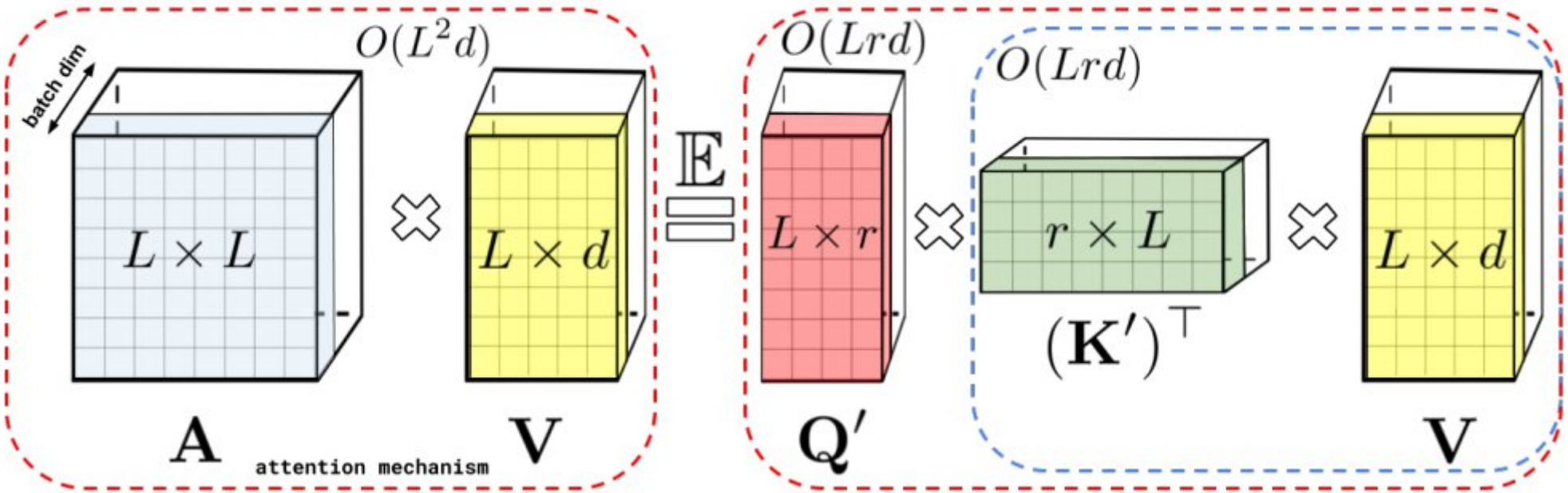
✓ 분산 문제
- 근사 분산이 매우 커져서 불안정

$$\phi(x) = \mathrm{elu}(x) + 1$$

$$V_i' = \frac{\phi(Q_i)^T \sum_{j=1}^{N} \phi(K_j) V_j^T}{\phi(Q_i)^T \sum_{j=1}^{N} \phi(K_j)}.$$



FAVOR+ (FA)


batch dim
$O(L^2 d)$
$L \times L$
$L \times d$
$\mathbf{A}$
attention mechanism
$\mathbf{V}$
$\mathbb{E}$
$O(Lrd)$
$L \times r$
$\mathbf{Q}'$
$O(Lrd)$
$r \times L$
$(\mathbf{K}')^\top$
$L \times d$
$\mathbf{V}$

FAVOR+ (R+)

Lemma 1 (Positive Random Features (PRFs) for Softmax). *For* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, $\mathbf{z} = \mathbf{x} + \mathbf{y}$ *we have:*

$$\mathrm{SM}(\mathbf{x}, \mathbf{y}) = \mathbb{E}_{\omega \sim \mathcal{N}(0, \mathbf{I}_d)} \left[\exp\left(\omega^\top \mathbf{x} - \frac{\|\mathbf{x}\|^2}{2}\right) \exp\left(\omega^\top \mathbf{y} - \frac{\|\mathbf{y}\|^2}{2}\right) \right] = \Lambda \mathbb{E}_{\omega \sim \mathcal{N}(0, \mathbf{I}_d)} \cosh(\omega^\top \mathbf{z}), \quad (7)$$

where $\Lambda = \exp(-\frac{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2}{2})$ *and* cosh *is hyperbolic cosine. Consequently, softmax-kernel admits a positive random feature map unbiased approximation with* $h(\mathbf{x}) = \exp(-\frac{\|\mathbf{x}\|^2}{2})$, $l = 1$, $f_1 = \exp$ *and* $\mathcal{D} = \mathcal{N}(0, \mathbf{I}_d)$ *or:* $h(\mathbf{x}) = \frac{1}{\sqrt{2}} \exp(-\frac{\|\mathbf{x}\|^2}{2})$, $l = 2$, $f_1(u) = \exp(u)$, $f_2(u) = \exp(-u)$ *and the same* $\mathcal{D}$ *(the latter for further variance reduction). We call related estimators:* $\widehat{\mathrm{SM}}^{+}_{m}$ *and* $\widehat{\mathrm{SM}}^{\mathrm{hyp}+}_{m}$.

FAVOR+ (R+)

F.1 Proof of Lemma 1

Proof. We first deduce that for any $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$

$$\mathrm{SM}(\mathbf{x}, \mathbf{y}) = \exp(\boldsymbol{x}^\top \boldsymbol{y}) = \exp(-\|\boldsymbol{x}\|^2/2) \cdot \exp(\|\boldsymbol{x}+\boldsymbol{y}\|^2/2) \cdot \exp(-\|\boldsymbol{y}\|^2/2).$$

Next, let $\boldsymbol{w} \in \mathbb{R}^d$. We use the fact that

$$(2\pi)^{-d/2} \int \exp(-\|\boldsymbol{w}-\boldsymbol{c}\|_2^2/2) d\boldsymbol{w} = 1$$

for any $\boldsymbol{c} \in \mathbb{R}^d$ and derive:

$$\begin{aligned}
\exp(\|\boldsymbol{x}+\boldsymbol{y}\|^2/2) &= (2\pi)^{-d/2} \exp(\|\boldsymbol{x}+\boldsymbol{y}\|^2/2) \int \exp(-\|\boldsymbol{w}-(\boldsymbol{x}+\boldsymbol{y})\|^2/2) d\boldsymbol{w} \\
&= (2\pi)^{-d/2} \int \exp(-\|\boldsymbol{w}\|^2/2 + \boldsymbol{w}^\top(\boldsymbol{x}+\boldsymbol{y}) - \|\boldsymbol{x}+\boldsymbol{y}\|^2/2 + \|\boldsymbol{x}+\boldsymbol{y}\|^2/2) d\boldsymbol{w} \\
&= (2\pi)^{-d/2} \int \exp(-\|\boldsymbol{w}\|^2/2 + \boldsymbol{w}^\top(\boldsymbol{x}+\boldsymbol{y})) d\boldsymbol{w} \\
&= (2\pi)^{-d/2} \int \exp(-\|\boldsymbol{w}\|^2/2) \cdot \exp(\boldsymbol{w}^\top \boldsymbol{x}) \cdot \exp(\boldsymbol{w}^\top \boldsymbol{y}) d\boldsymbol{w} \\
&= \mathbb{E}_{\omega \sim \mathcal{N}(\mathbf{0}_d, \mathbf{I}_d)} [\exp(\omega^\top \boldsymbol{x}) \cdot \exp(\omega^\top \boldsymbol{y})].
\end{aligned}$$

That completes the proof of the first part of the lemma. An identity involving hyperbolic cosine function is implied by the fact that for every $\mathbf{u} \in \mathbb{R}^d$ and $\omega \sim \mathcal{N}(0, \mathbf{I}_d)$ the following is true:

$$\mathbb{E}[\exp(\omega^\top \mathbf{u})] = \sum_{i=0}^{\infty} \frac{\mathbb{E}[(\omega^\top \mathbf{u})^{2i}]}{(2i)!} = \frac{1}{2} \sum_{i=0}^{\infty} \frac{\mathbb{E}[(\omega^\top \mathbf{u})^{2i}] + \mathbb{E}[(-\omega^\top \mathbf{u})^{2i}]}{(2i)!}. \tag{12}$$

The cancellation of the odd moments $\mathbb{E}[(\omega^\top \mathbf{u})^{2i+1}]$ follows directly from the fact that ω is taken from the isotropic distribution (i.e. distribution with pdf function constant on each sphere). That completes the proof. □

FAVOR+ (O)

Theorem 1 (regularized versus softmax-kernel). *Assume that the L_∞-norm of the attention matrix for the softmax-kernel satisfies: $\|\mathbf{A}\|_\infty \leq C$ for some constant $C \geq 1$. Denote by $\mathbf{A}^{\mathrm{reg}}$ the corresponding attention matrix for the regularized softmax-kernel. The following holds:*

$$\inf_{i,j} \frac{\mathbf{A}^{\mathrm{reg}}(i,j)}{\mathbf{A}(i,j)} \geq 1 - \frac{2}{d^{\frac{1}{3}}} + o\left(\frac{1}{d^{\frac{1}{3}}}\right), \text{ and } \sup_{i,j} \frac{\mathbf{A}^{\mathrm{reg}}(i,j)}{\mathbf{A}(i,j)} \leq 1. \tag{9}$$

Furthermore, the latter holds for $d \geq 2$ even if the L_∞-norm condition is not satisfied, i.e. the regularized softmax-kernel is a universal lower bound for the softmax-kernel.

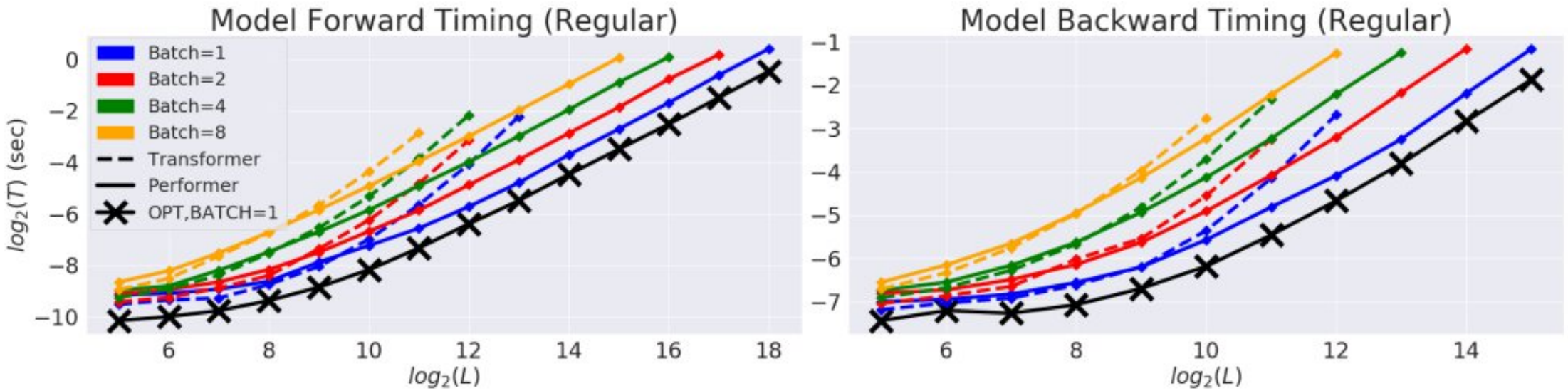
Theorem 2. *If $\widehat{\mathrm{SM}}_m^{\mathrm{ort}+}(\mathbf{x},\mathbf{y})$ stands for the modification of $\widehat{\mathrm{SM}}_m^{+}(\mathbf{x},\mathbf{y})$ with orthogonal random features (and thus for $m \leq d$), then the following holds for any $d > 0$:*

$$\mathrm{MSE}(\widehat{\mathrm{SM}}_m^{\mathrm{ort}+}(\mathbf{x},\mathbf{y})) \leq \mathrm{MSE}(\widehat{\mathrm{SM}}_m^{+}(\mathbf{x},\mathbf{y})) - \frac{2(m-1)}{m(d+2)}\left(\mathrm{SM}(\mathbf{x},\mathbf{y}) - \exp\left(-\frac{\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2}{2}\right)\right)^2. \tag{10}$$

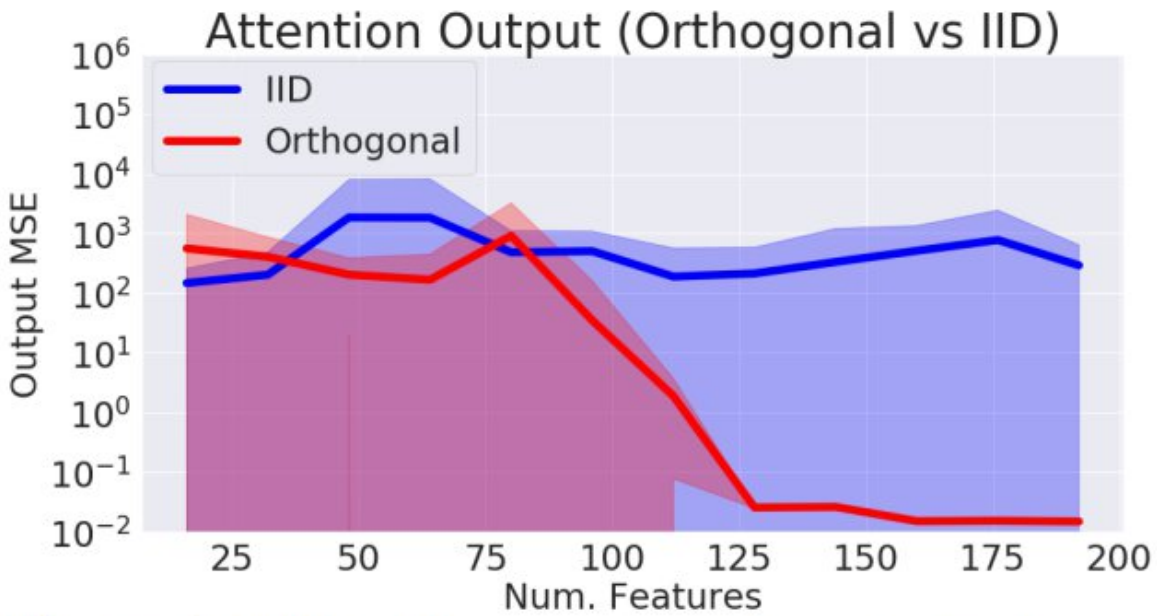
Furthermore, completely analogous result holds for the regularized softmax-kernel SMREG.

Experiments

FAVOR+ (FA)


Model Forward Timing (Regular)
Batch=1
Batch=2
Batch=4
Batch=8
Transformer
Performer
OPT,BATCH=1
$log_2(T)$ (sec)
$log_2(L)$
Model Backward Timing (Regular)
$log_2(L)$

FAVOR+ (OR+)



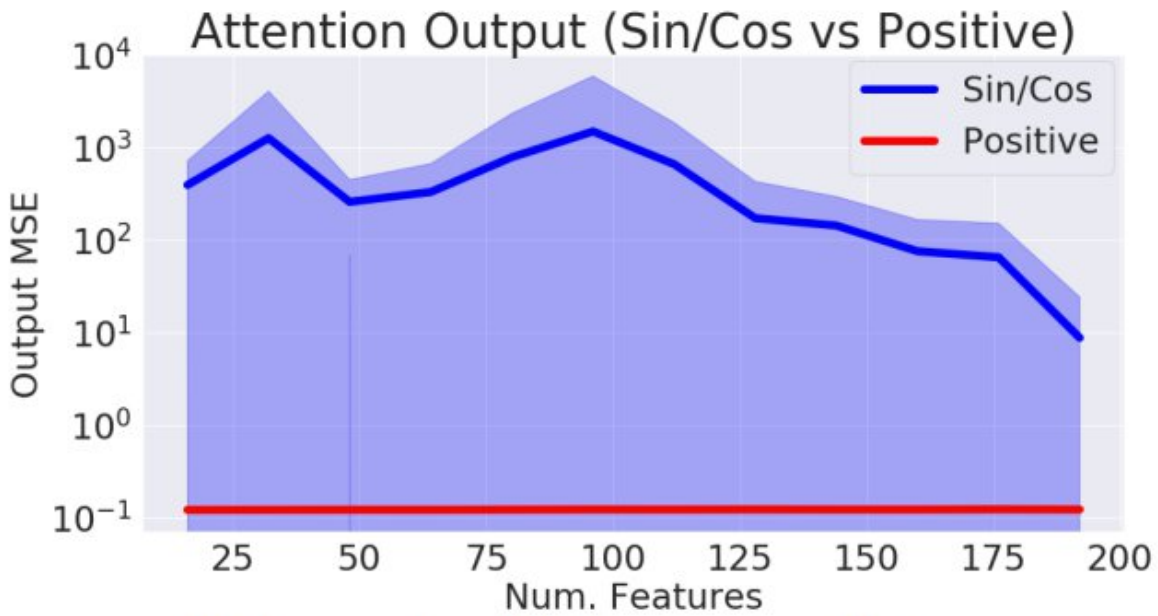


Figure 4: MSE of the approximation output when comparing Orthogonal vs IID features and trigonometric sin/cos vs positive features. We took $L = 4096, d = 16$, and varied the number of random samples m. Standard deviations shown across 15 samples of appropriately normalized random matrix input data.

Conclusion & Broader Impact

Conclusion

✓ 핵심 아이디어
- FAVOR+

✓ 핵심 기여
- Attention을 $O(n^2) \rightarrow O(n)$으로 근사
- 무편향, 안정성, 수렴성을 수학적으로 보장
- 다양한 실험에서 Transformer와 동등하거나 더 나은 성능

✓ 의의와 영향
- **대규모 시퀀스 처리 가능 →** 긴 문장, 유전체 데이터, 영상 등 확장성
- **자원 효율적 →** 메모리/시간 절약, 친환경적 AI 연구 기여 가능
- **연구적 영향 →** 이후 Linear Attention 계열 연구(Linformer, Nyströmformer, FlashAttention 등)의 기초를 마련
- **한계 인식:** approximation error 존재, 특정 task에서는 성능 저하 가능

감사합니다