Time to Perform

Owen Howell[1]

[1]*Department of Electrical Engineering, Northeastern University, Huntington Ave., Boston, MA 02215, USA*

(Dated: November 20, 2022)

We reproduce the results in the original performers paper and talk about some possible extensions.

I. BIDIRECTIONAL ATTENTION

Transformers have found use in a wide variety of machine learning tasks [3]. One of the key ideas in the transformer is the attention mechanism. Let L be the size of an input sequence of tokens. Let d be the latent transformer dimension. Let $Q, K, V \in \mathbb{R}^{L \times d}$. Then, the bidirectional dot-product attention is given by the following,

$$\begin{aligned}
&\text{Att}(Q, K, V) = D^{-1}AV \\
&A = \exp(QK^T) \\
&D = \text{diag}(A\bar{1}) = \text{diag}(\sum_{j=1}^{d} A_{ij})
\end{aligned}$$

for large L, it requires both time consuming and memory expensive to compute Att. The paper [1] proposes using a stochastic kernel trick to approximately compute the attention function. This idea, which utilizes the Johnson-Linderhaus method, has a rich history [2].

II. KERNEL TRICK

Let $\phi : \mathbb{R}^d \to \mathbb{R}^r$ be a random map. We define the kernel as

$$K(x, y) = \mathbb{E}[\phi(x)^T \phi(y)]$$

the validity of this approach hinges on the quantity $\phi(x)^T\phi(y)$ being concentrated around its expectation. In matrix form, this involves approximating the matrix $A \in \mathbb{R}^{d \times d}$ as a stochastic low rank decomposition $A = \mathbb{E}[\phi^T \phi]$ where $\phi \in \mathbb{R}^{d \times r}$ with $r << d$.

1. *Soft-max Kernel*

Consider the soft-max kernel defined as

$$SM(x, y) = \exp(x^T y)$$

We will specifically be interested in approximating the softmax-kernel function. Specifically, we have that

$$SM(x, y) = \mathbb{E}_{\omega \sim N(0, \mathbb{1}_d)}[\exp(\omega^T x - \frac{1}{2}||x||^2) \exp(\omega^T y - \frac{1}{2}||y||^2)]$$

The idea behind the paper [1] is to replace the soft-max kernel with a sample average of the quantity,

$$\mathbb{E}_{\omega \sim N(0, \mathbb{1}_d)}[\exp(\omega^T x - \frac{1}{2}||x||^2) \exp(\omega^T y - \frac{1}{2}||y||^2)]$$

The validity of this approach depends on how concentrated $\exp(\omega^T x - \frac{1}{2}||x||^2) \exp(\omega^T y - \frac{1}{2}||y||^2)$ is around the expectation $\mathbb{E}_{\omega \sim N(0, \mathbb{1}_d)}[\exp(\omega^T x - \frac{1}{2}||x||^2) \exp(\omega^T y - \frac{1}{2}||y||^2)]$. Furthermore, the number of samples guaranteed to require convergence should be small enough that the stochastic method is faster then just computing the original attention matrix.

A. Datasets

We will test the performer on a few standard datasets.

---

[1] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers, 2020.

[2] Krzysztof Choromanski, Mark Rowland, and Adrian Weller. The unreasonable effectiveness of structured random orthogonal embeddings, 2017.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.