

Previous Works

- Performer (Choromanski et al.,) uses a finite-dimensional map φ to approximate exponential
 - Vectors with larger norms require φ with larger dimension

- Other works consider arbitrary φ instead of first defining $\text{sim}(\cdot, \cdot)$
 - $\varphi(x) \doteq \text{elu}(x) + 1, \varphi(x) \doteq \text{relu}(x)$
 - Model quality is worse compared to softmax

• Is softmax necessary? Do any other functions with similar properties work?

- Consider $\text{sim}(q, k) = \langle q, k \rangle^p$ where $p \geq 2$ is an even integer
 - Always ≥ 0
 - Increases as $\langle q, k \rangle$ goes up



Previous Works

- Performer (Choromanski et al.,) uses a finite-dimensional map φ to approximate exponential
 - Vectors with larger norms require φ with larger dimension
- Other works consider arbitrary φ instead of first defining $\text{sim}(\cdot, \cdot)$
 - $\varphi(x) \doteq \text{elu}(x) + 1, \varphi(x) \doteq \text{relu}(x)$
 - Model quality is worse compared to softmax
- Is softmax necessary? Do any other functions with similar properties work?
- Consider $\text{sim}(q, k) = \langle q, k \rangle^p$ where $p \geq 2$ is an even integer
 - Always ≥ 0
 - Increases as $\langle q, k \rangle$ goes up

Comparisons

Perplexities on Wiki-40B

