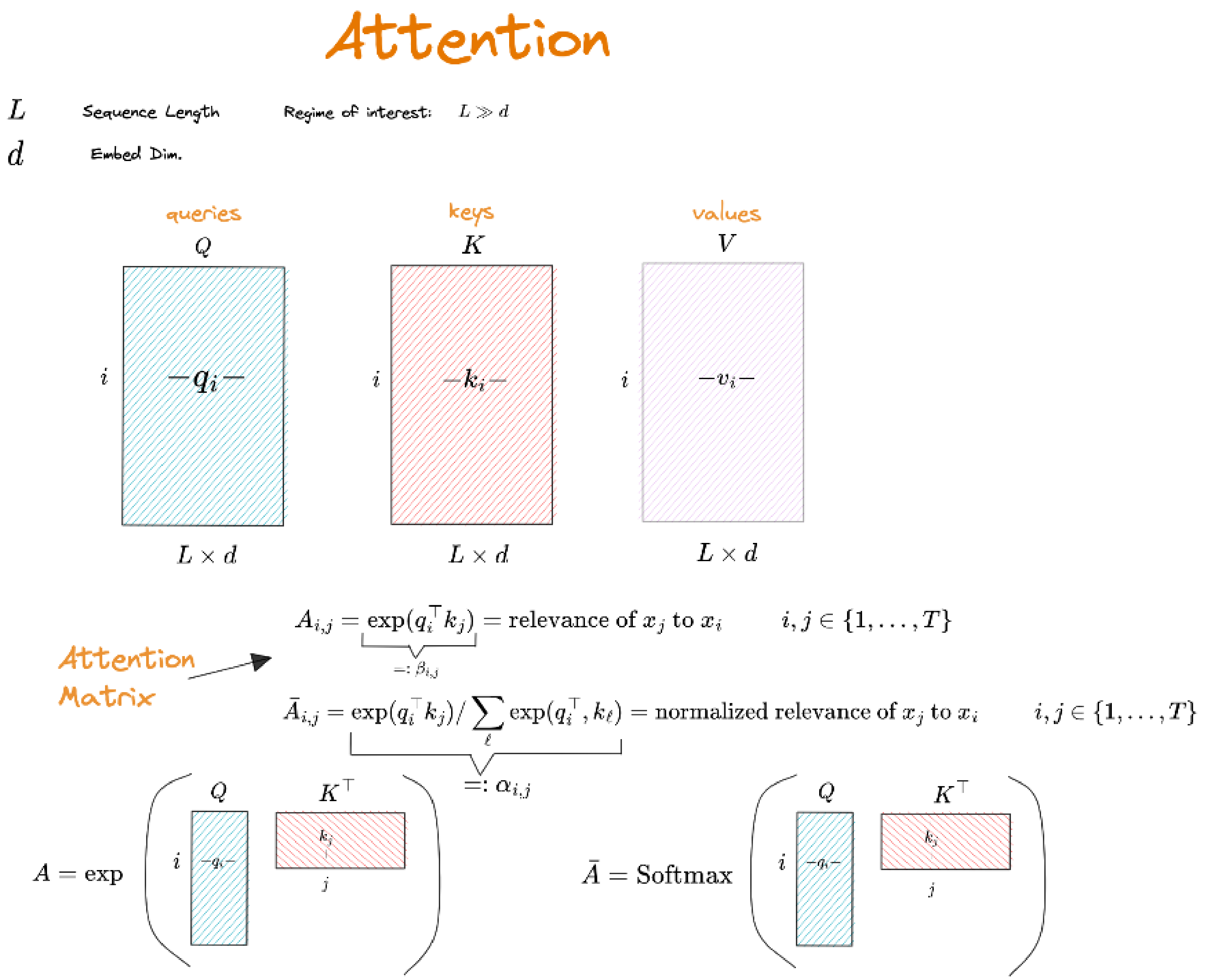
Performer

Attention: Recap



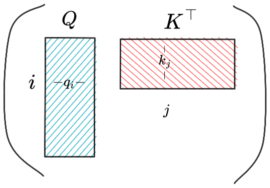
$$A_{i,j} = \underbrace{\exp(q_i^\top k_j)}_{=:\,\beta_{i,j}} = \text{relevance of } x_j \text{ to } x_i \qquad i,j \in \{1,\dots,T\}$$

$$\bar{A}_{i,j} = \underbrace{\exp(q_i^\top k_j)/\sum_\ell \exp(q_i^\top, k_\ell)}_{=:\,\alpha_{i,j}} = \text{normalized relevance of } x_j \text{ to } x_i \qquad i,j \in \{1,\dots,T\}$$

$$A = \exp\left(Q K^\top \right) \qquad \bar{A} = \text{Softmax}\left(Q K^\top \right)$$

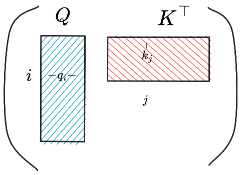
Attention (cont.)

$A =$



$$\begin{pmatrix} \beta_{1,1} & \cdots & \beta_{1,L} \\ & \vdots & \\ \beta_{L,L} & \cdots & \beta_{L,1} \end{pmatrix}$$

$\bar{A} = \text{Softmax}$



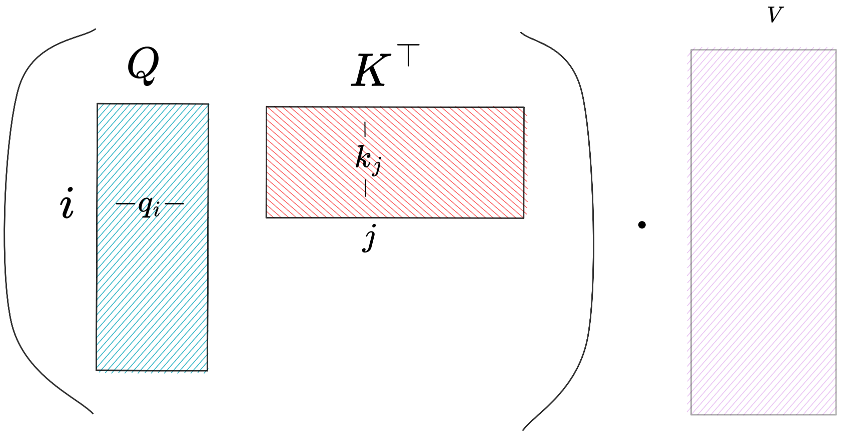
$$= \begin{pmatrix} \alpha_{1,L} & \cdots & \alpha_{1,L} \\ & \vdots & \\ \alpha_{L,1} & \cdots & \alpha_{L,L} \end{pmatrix}$$

$$\beta_{i,j} = \frac{\alpha_{i,j}}{\sum_\ell \alpha_{i,\ell}} \longrightarrow \bar{A} = D^{-1}A \qquad D = \begin{pmatrix} \sum_\ell \beta_{1,\ell} & & \\ & \ddots & \\ & & \sum_\ell \beta_{L,\ell} \end{pmatrix}$$

$$D = A 1_L = Q K^\top 1_L$$

Attention (cont.)

$Z = \text{Softmax}$



$$= \bar{A}V = D^{-1}AV$$

Computation complexity: $O(L^2 d)$

(prohibitive for long sequences)

space complexity: $O(L^2)$

Efficient transformers

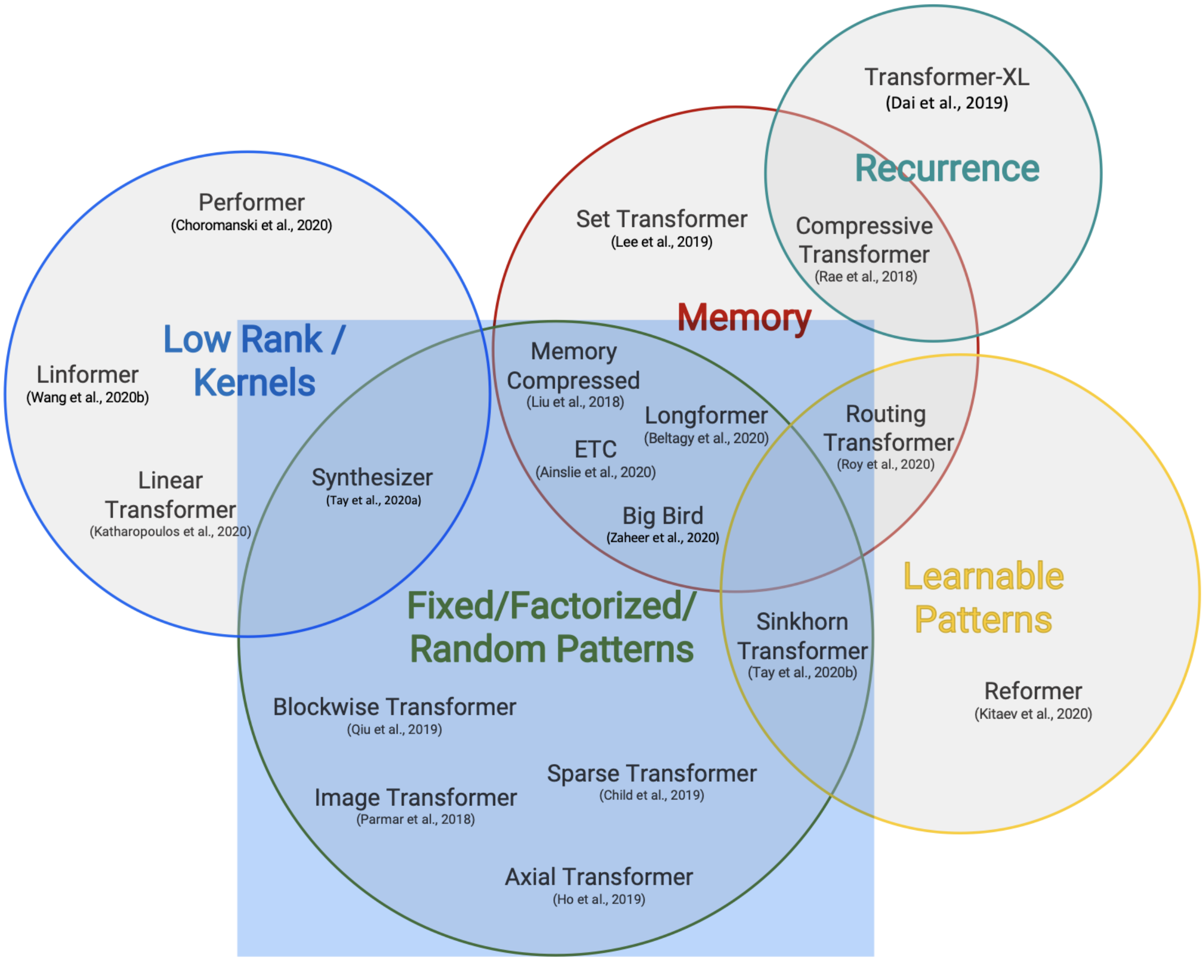
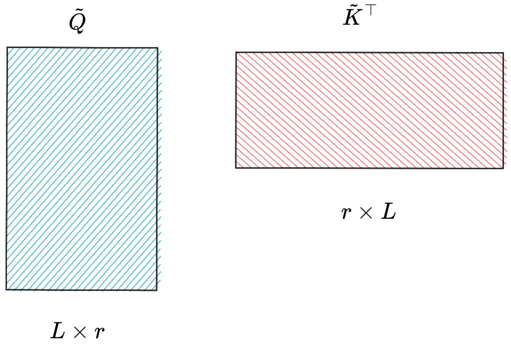


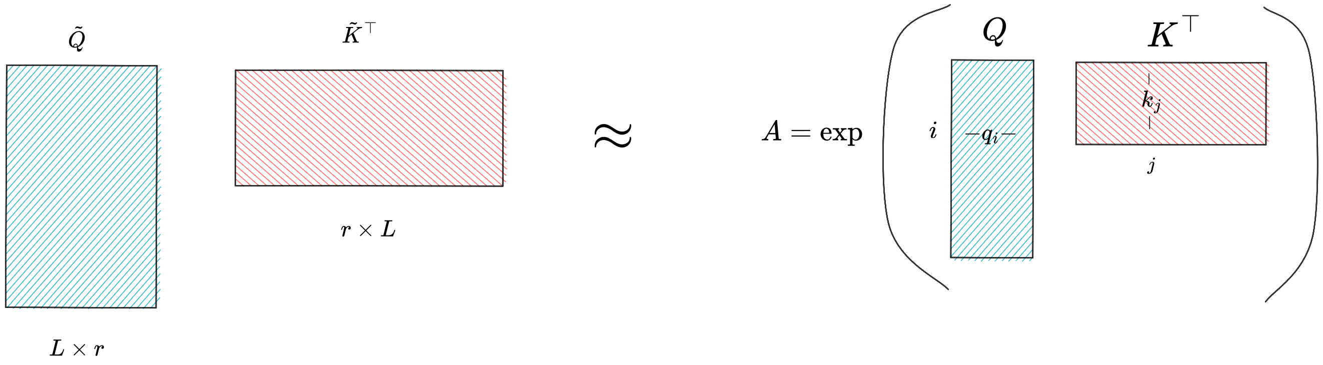
Figure 2: Taxonomy of Efficient Transformer Architectures.

Performer

main idea: use random features to construct

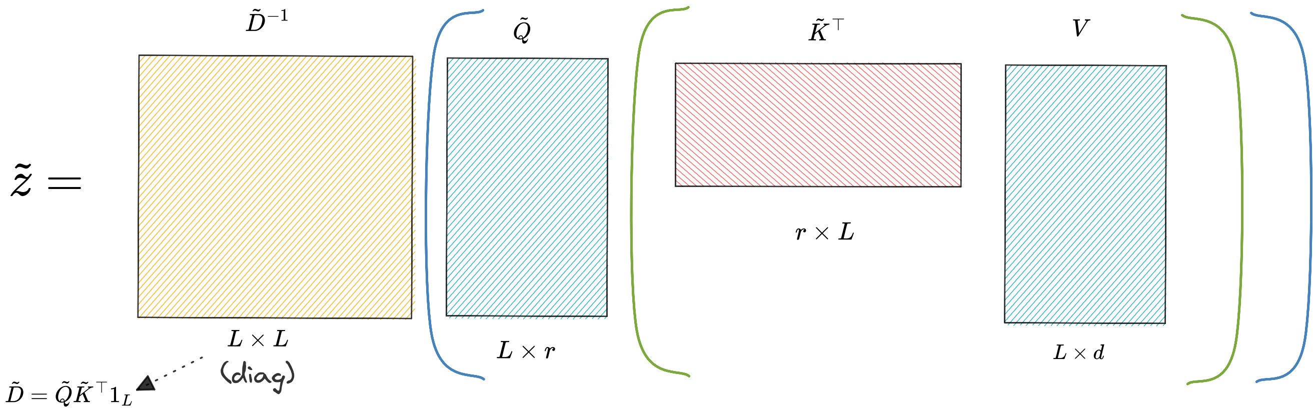


with $r \ll L$ such that



Performer (cont.)

computational and space savings



Computation complexity: $O(Lr^2)$

space complexity: $O(Lr)$

Random Features

$k(x, y) = \exp(x^\top y)$ Kernel/similarity function

$A_{i,j} = \exp(q_i^\top k_j) = k(q_i, k_j)$

Construction for r=1

draw $w_1, \ldots, w_r \sim \mathcal{N}(0, I)$

$$x \xmapsto{\phi_w} \exp\left(w^\top x - \|x\|^2/2\right)$$

Lemma

$$\mathbb{E}[\phi_w(x)^\top \phi_w(y)] = \exp(x^\top y)$$

Amplification (r>1)

draw $w_1, \ldots, w_r \sim \mathcal{N}(0, I)$ $r = \tilde{\Theta}(d/\epsilon^2)$

$$x \xmapsto{\phi} \frac{1}{\sqrt{r}}(\phi_{w_1}(x), \ldots, \phi_{w_r}(x))$$

Theorem

W.h.p. for all x,y (simulataneously)

$$|\phi(x)^\top \phi(y) - k(x, y)| \leq \epsilon$$

$$\tilde{Q} = \begin{pmatrix} — & \phi(q_1) & — \\ & \circ & \\ & \circ & \\ & \circ & \\ — & \phi(q_L) & — \end{pmatrix} \qquad \tilde{K} = \begin{pmatrix} — & \phi(k_1) & — \\ & \circ & \\ & \circ & \\ & \circ & \\ — & \phi(k_L) & — \end{pmatrix} \longrightarrow \|\tilde{Q}\tilde{K}^\top - A\|_\infty \leq \epsilon$$

Random Features: Proof

$k(x, y) = \exp(x^\top y)$ Kernel/similarity function

draw $w_1, \dots, w_r \sim \mathcal{N}(0, I)$

$$x \xmapsto{\phi_w} \exp\left(w^\top x - \|x\|^2/2\right)$$

Lemma

$$\mathbb{E}[\phi_w(x)^\top \phi_w(y)] = \exp(x^\top y)$$

Proof sketch

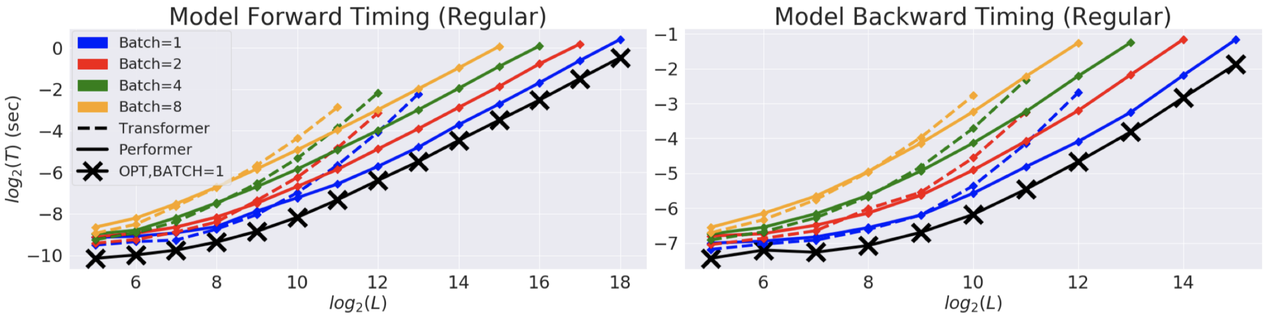
$$\mathbb{E}_w[\phi_w(x)^\top \phi_w(y)] \approx \int_w \phi_w(x)\phi_w(y)\exp(-\|w\|^2/2)\,dw$$

$$= \int_w \exp(-\|x\|^2/2)\ \exp(-\|y\|^2/2)\ \exp(w^\top(x+y))\exp(-\|w\|^2/2)\,dw$$

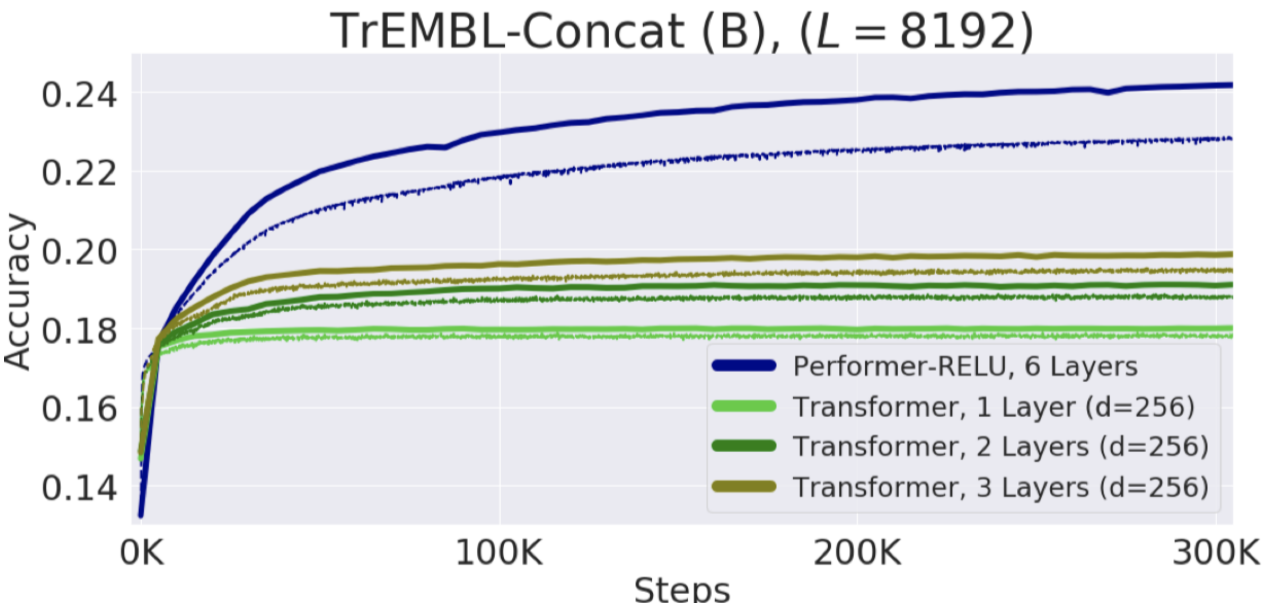
$$= \int_w \exp(\|w - (x+y)\|^2)\ \exp(x^\top y)\,dw$$

$$= \exp(x^\top y)\underbrace{\int_w \exp(\|w - (x+y)\|^2)\,dw}_{\approx 1}$$

Experiments



TrEMBL: predicting interactions among groups of proteins by concatenating protein sequences


TrEMBL-Concat (B), ($L = 8192$)
Accuracy
Steps
0K
100K
200K
300K
Performer-RELU, 6 Layers
Transformer, 1 Layer (d=256)
Transformer, 2 Layers (d=256)
Transformer, 3 Layers (d=256)