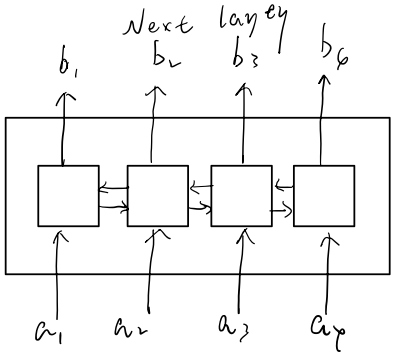
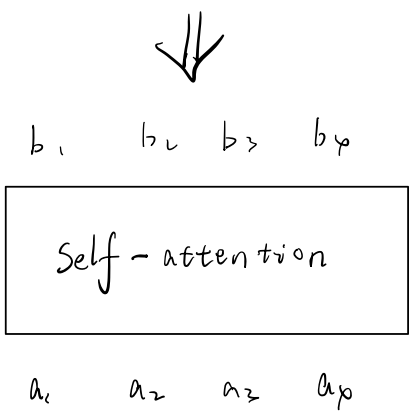
Transformer —unsupervised→ BERT

Sequence to Sequence

RNN

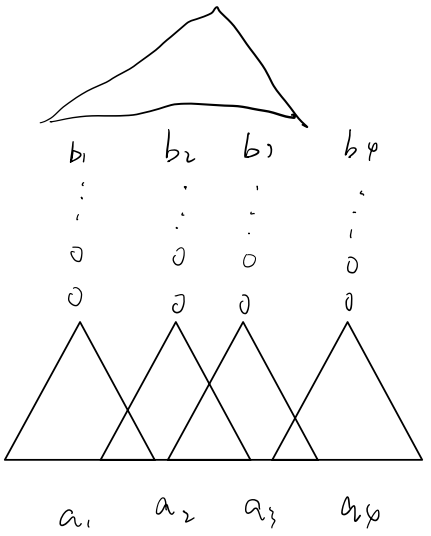


Hard to parallel



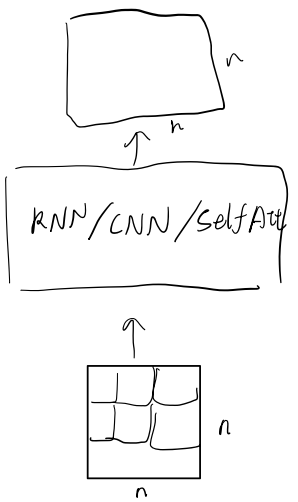
bi is obtained based on the whole input sequence

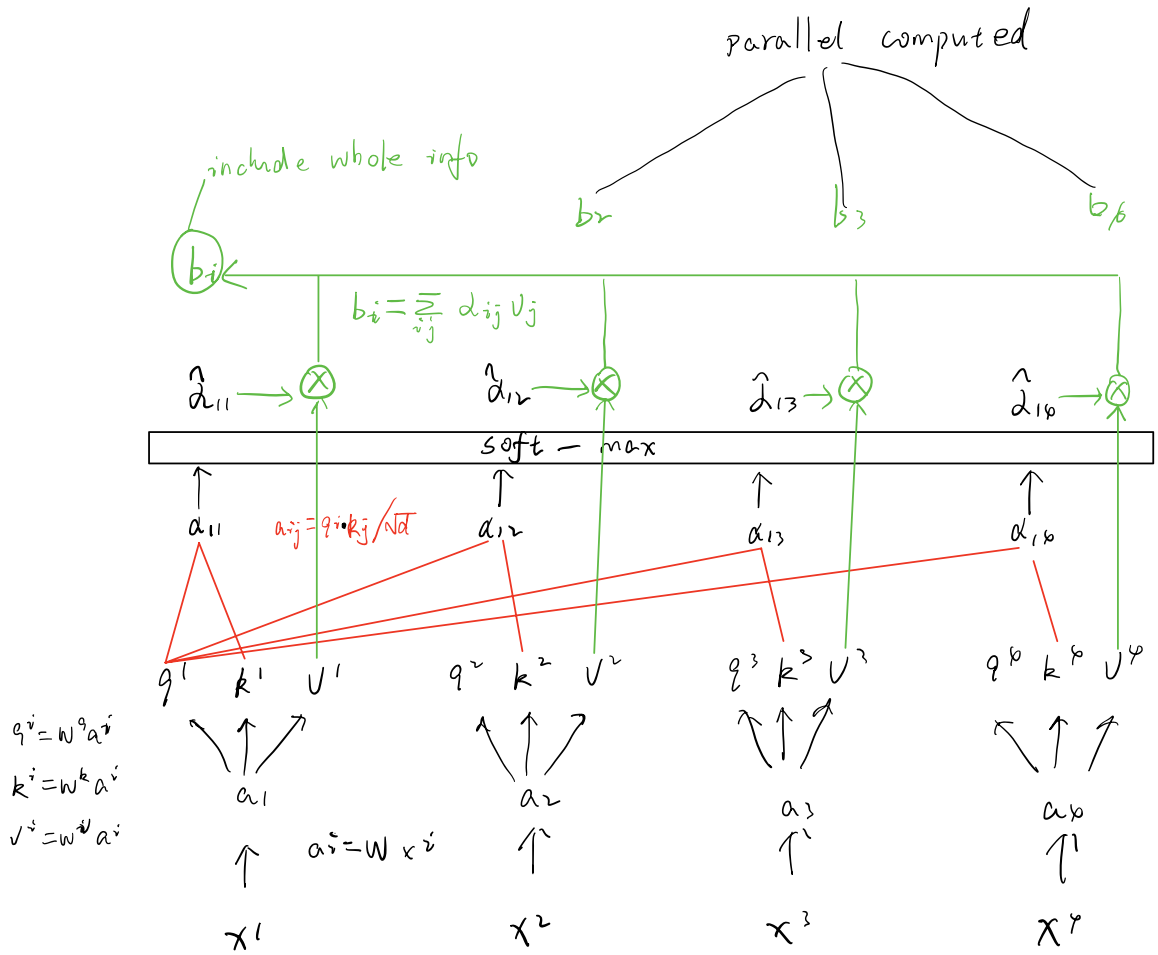
can be parallelly computed



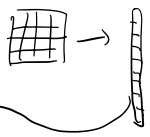
GNN relave RNN

视野受限，多层才能看全




parallel computed
include whole info
b_1
b_2
b_3
b_4
b_i
$b_i = \sum_j \alpha_{ij} v_j$
$\hat{\alpha}_{11}$
$\hat{\alpha}_{12}$
$\hat{\alpha}_{13}$
$\hat{\alpha}_{14}$
soft - max
α_{11}
$\alpha_{ij} = q_i \cdot k_j / \sqrt{d}$
α_{12}
α_{13}
α_{14}
q^1 k^1 v^1
q^2 k^2 v^2
q^3 k^3 v^3
q^4 k^4 v^4
$q^i = W^q a^i$
$k^i = W^k a^i$
$v^i = W^v a^i$
a_1
a_2
a_3
a_4
$a^i = W x^i$
x^1
x^2
x^3
x^4

How to parallelly computed ?



$$\underset{Q}{\begin{bmatrix} q^1 & q^2 & q^3 & q^4 \end{bmatrix}} = W^q \underset{I}{\begin{bmatrix} a^1 & a^2 & a^3 & a^4 \end{bmatrix}}$$

1×256 1×64 1×64 1×64 64×64 1×64 1×64 1×64 1×256

$$Q = W^q I$$
$$K = W^k I$$
$$V = W^v I$$

$$\alpha_{11} = k^1 q^1$$
$$\alpha_{12} = k^2 q^1$$
$$\alpha_{13} = k^3 q^1$$
$$\alpha_{14} = k^4 q^1$$

$$\begin{pmatrix} \alpha_{i1} \\ \alpha_{i2} \\ \alpha_{i3} \\ \alpha_{i4} \end{pmatrix} = \begin{pmatrix} k^1 \\ k^2 \\ k^3 \\ k^4 \end{pmatrix} q^i$$

$$\alpha^{ij} \in R^{1\times 1}$$
$$k^i \in R^{16\times 1}$$
$$q^j \in R^{16\times 1}$$

$$\begin{matrix} \hat{\alpha}_{11} & \hat{\alpha}_{12} & \hat{\alpha}_{13} & \hat{\alpha}_{14} \\ \hat{\alpha}_{21} & \hat{\alpha}_{22} & \hat{\alpha}_{23} & \hat{\alpha}_{24} \\ \hat{\alpha}_{31} & \hat{\alpha}_{32} & \hat{\alpha}_{33} & \hat{\alpha}_{34} \\ \hat{\alpha}_{41} & \hat{\alpha}_{42} & \hat{\alpha}_{43} & \hat{\alpha}_{44} \end{matrix} \Leftarrow \text{soft max} \quad \begin{matrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} \\ \alpha_{31} & \alpha_{32} & \alpha_{33} & \alpha_{34} \\ \alpha_{41} & \alpha_{42} & \alpha_{43} & \alpha_{44} \end{matrix} = \begin{matrix} k_1 \\ k_2 \\ k_3 \\ k_4 \end{matrix} \cdot \begin{matrix} q^1 & q^2 & q^3 & q^4 \end{matrix}$$

$\hat{A}$ A K^T Q

$$\begin{matrix} b_1 & b_2 & b_3 & b_4 \end{matrix} = \begin{matrix} v_1 & v_2 & v_3 & v^4 \end{matrix} \begin{matrix} \hat{\alpha}_{11} & \hat{\alpha}_{12} & \hat{\alpha}_{13} & \hat{\alpha}_{14} \\ \hat{\alpha}_{21} & \hat{\alpha}_{22} & \hat{\alpha}_{23} & \hat{\alpha}_{24} \\ \hat{\alpha}_{31} & \hat{\alpha}_{32} & \hat{\alpha}_{33} & \hat{\alpha}_{34} \\ \hat{\alpha}_{41} & \hat{\alpha}_{42} & \hat{\alpha}_{43} & \hat{\alpha}_{44} \end{matrix}$$

Q V $\hat{A}$

b_1 b_2 b_3 b_4 $\Rightarrow$ O

Self-attention

a_1 a_2 a_3 a_4 $\Rightarrow$ I

q	$=$	W^q	I
k	$=$	W^k	I
v	$=$	W^v	I

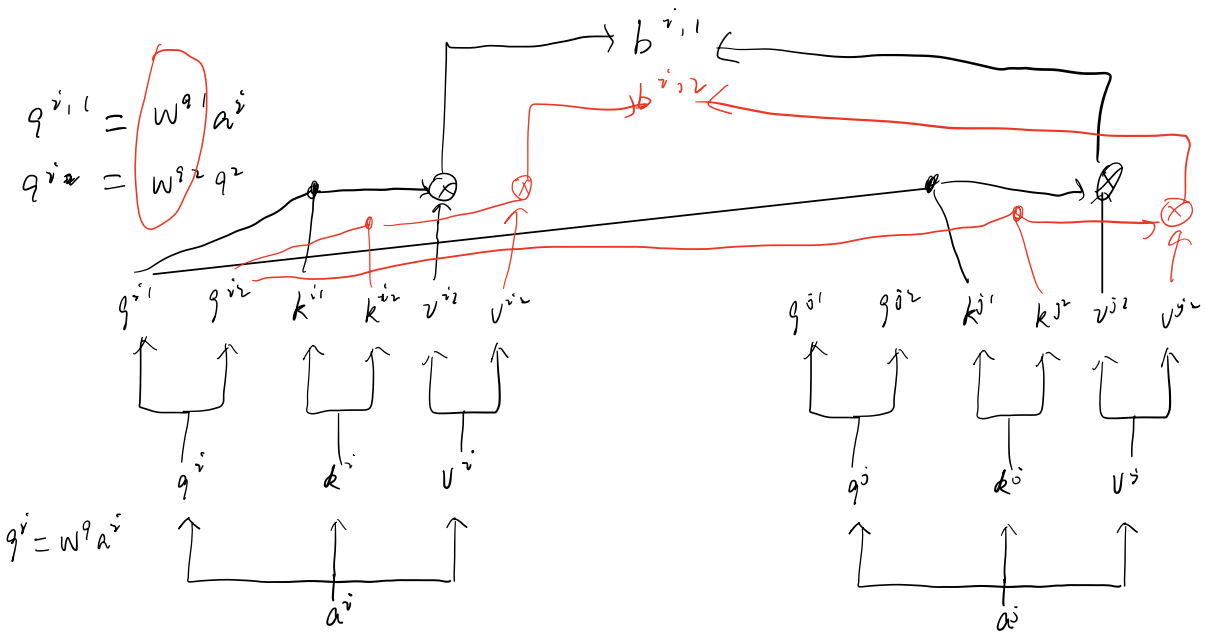
$\hat{A} \Leftarrow A = K^T \; Q$

$O = V \; \hat{A}$

Multi-head Self-attention

不同 head

不同关注点 (local vs long time)



$b^i = W^O \begin{bmatrix} b^{i,1} \\ b^{i,2} \end{bmatrix}$ 降维

Position Encoding

1. no position info in self-attention

2. each position has a unique position vector e^i not learned from data