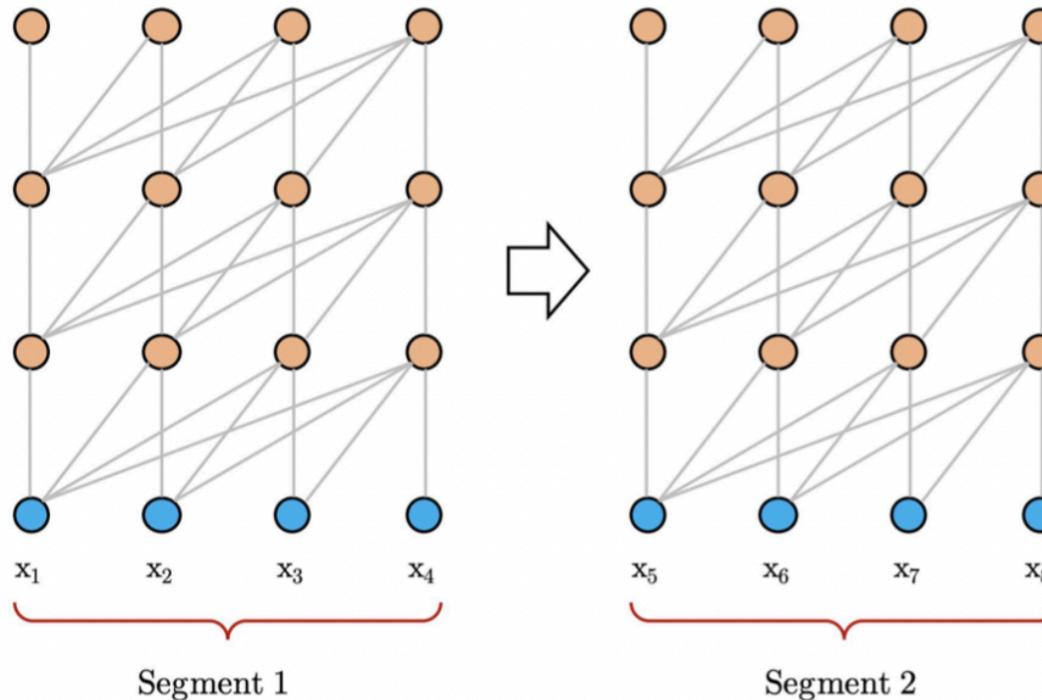


Transformer-XL

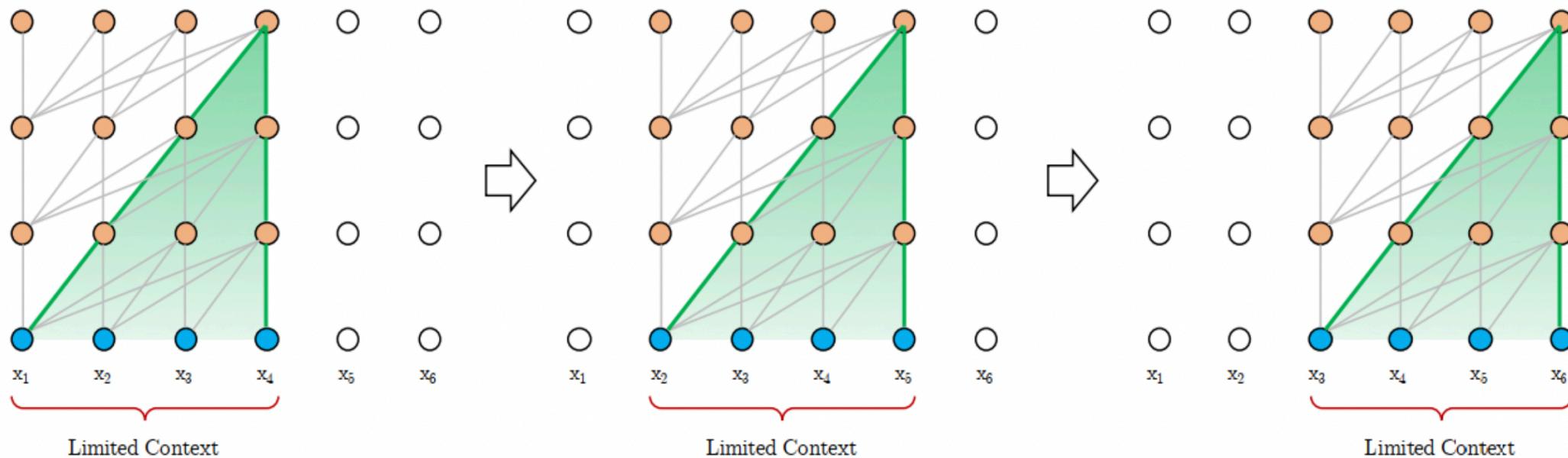
Недостатки vanilla transformer

- Текст должен быть предварительно разделен на несколько сегментов

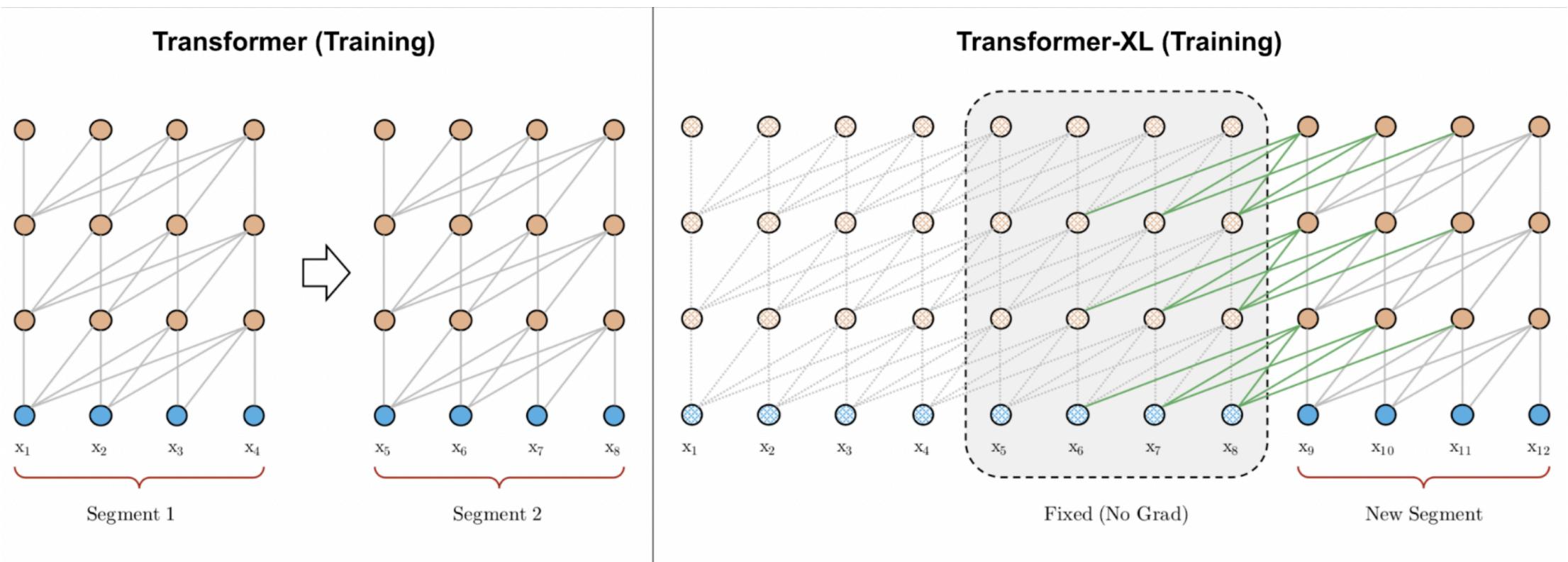


Недостатки vanilla transformer

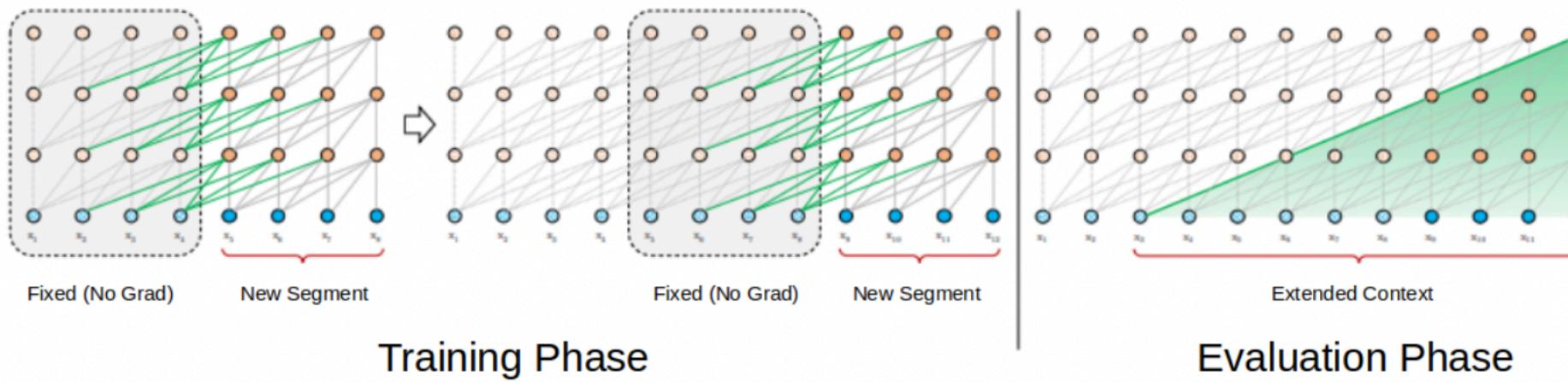
- Дорогая и медленная evaluation



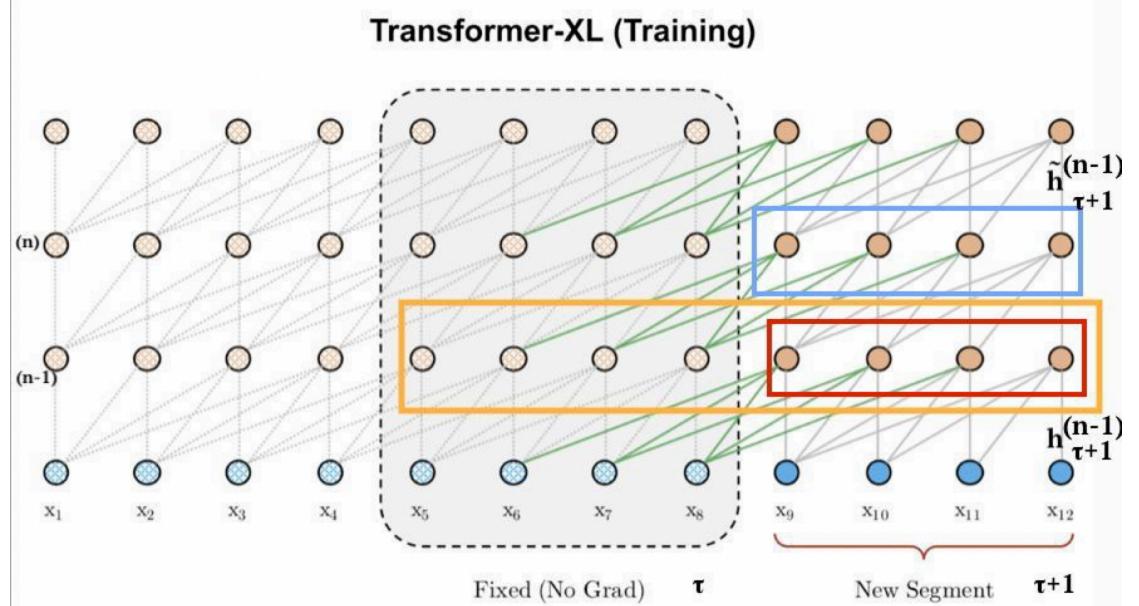
Hidden State Reuse



Hidden State Reuse



$$\begin{aligned}
\tilde{\mathbf{h}}_{\tau+1}^{(n-1)} &= [\text{stop-gradient}(\mathbf{h}_{\tau}^{(n-1)}) \circ \mathbf{h}_{\tau+1}^{(n-1)}] \\
\mathbf{Q}_{\tau+1}^{(n)} &= \mathbf{h}_{\tau+1}^{(n-1)} \mathbf{W}^q \\
\mathbf{K}_{\tau+1}^{(n)} &= \tilde{\mathbf{h}}_{\tau+1}^{(n-1)} \mathbf{W}^k \\
\mathbf{V}_{\tau+1}^{(n)} &= \tilde{\mathbf{h}}_{\tau+1}^{(n-1)} \mathbf{W}^v \\
\mathbf{h}_{\tau+1}^{(n)} &= \text{transformer-layer}(\mathbf{Q}_{\tau+1}^{(n)}, \mathbf{K}_{\tau+1}^{(n)}, \mathbf{V}_{\tau+1}^{(n)})
\end{aligned}$$



Relative Positional Encoding

Было:

$$\begin{aligned} a_{ij} &= \mathbf{q}_i \mathbf{k}_j^\top = (\mathbf{x}_i + \mathbf{p}_i) \mathbf{W}^q ((\mathbf{x}_j + \mathbf{p}_j) \mathbf{W}^k)^\top \\ &= \mathbf{x}_i \mathbf{W}^q \mathbf{W}^k^\top \mathbf{x}_j^\top + \mathbf{x}_i \mathbf{W}^q \mathbf{W}^k^\top \mathbf{p}_j^\top + \mathbf{p}_i \mathbf{W}^q \mathbf{W}^k^\top \mathbf{x}_j^\top + \mathbf{p}_i \mathbf{W}^q \mathbf{W}^k^\top \mathbf{p}_j^\top \end{aligned}$$

Стало:

$$a_{ij}^{\text{rel}} = \underbrace{\mathbf{x}_i \mathbf{W}^q \mathbf{W}_E^{k^\top} \mathbf{x}_j^\top}_{\text{content-based addressing}} + \underbrace{\mathbf{x}_i \mathbf{W}^q \mathbf{W}_R^{k^\top} \mathbf{r}_{i-j}^\top}_{\text{content-dependent positional bias}} + \underbrace{\mathbf{u} \mathbf{W}_E^{k^\top} \mathbf{x}_j^\top}_{\text{global content bias}} + \underbrace{\mathbf{v} \mathbf{W}_R^{k^\top} \mathbf{r}_{i-j}^\top}_{\text{global positional bias}}$$

Relative Positional Encoding

$$\begin{aligned} a_{ij} &= \mathbf{q}_i \mathbf{k}_j^\top = (\mathbf{x}_i + \mathbf{p}_i) \mathbf{W}^q ((\mathbf{x}_j + \mathbf{p}_j) \mathbf{W}^k)^\top \\ &= \mathbf{x}_i \mathbf{W}^q \mathbf{W}^k^\top \mathbf{x}_j^\top + \mathbf{x}_i \mathbf{W}^q \mathbf{W}^k^\top \mathbf{p}_j^\top + \mathbf{p}_i \mathbf{W}^q \mathbf{W}^k^\top \mathbf{x}_j^\top + \mathbf{p}_i \mathbf{W}^q \mathbf{W}^k^\top \mathbf{p}_j^\top \end{aligned}$$



$$a_{ij}^{\text{rel}} = \underbrace{\mathbf{x}_i \mathbf{W}^q \mathbf{W}_E^{k^\top} \mathbf{x}_j^\top}_{\text{content-based addressing}} + \underbrace{\mathbf{x}_i \mathbf{W}^q \mathbf{W}_R^{k^\top} \mathbf{r}_{i-j}^\top}_{\text{content-dependent positional bias}} + \underbrace{\mathbf{u} \mathbf{W}_E^{k^\top} \mathbf{x}_j^\top}_{\text{global content bias}} + \underbrace{\mathbf{v} \mathbf{W}_R^{k^\top} \mathbf{r}_{i-j}^\top}_{\text{global positional bias}}$$

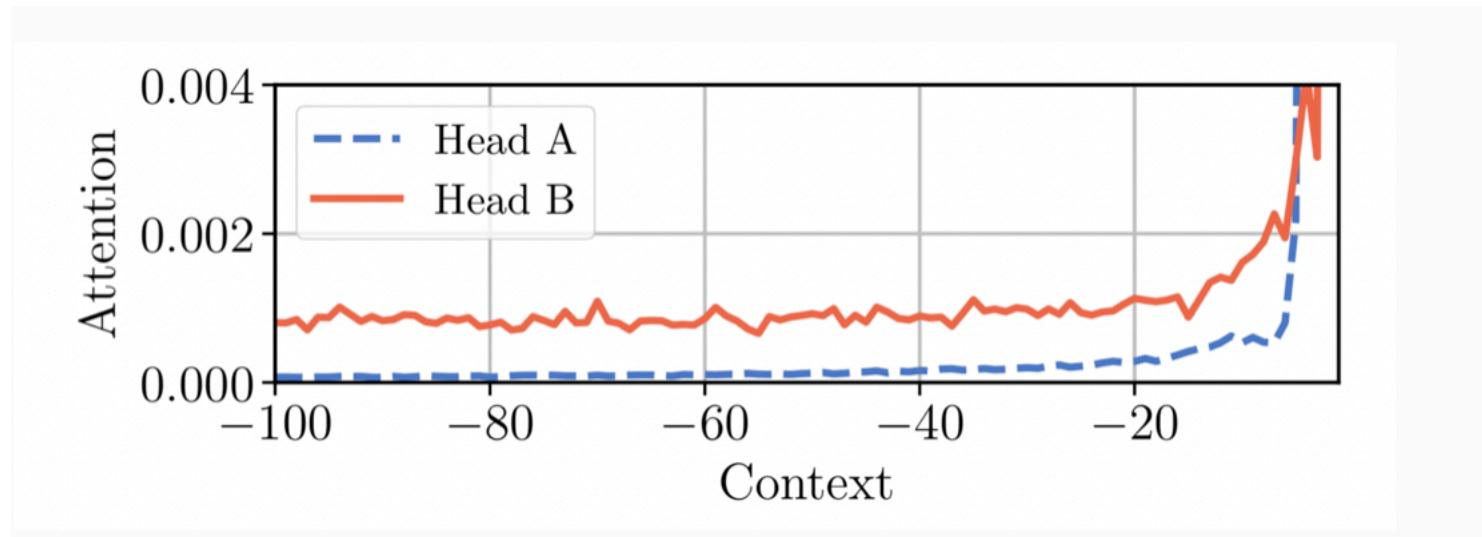
\mathbf{p}_j заменена на relative positional encoding \mathbf{r}_{i-j}

$\mathbf{p}_i \mathbf{W}^q$ заменена на два обучаемых вектора \mathbf{u} (для содержания токена) и \mathbf{v} (позиции токена)

\mathbf{W}^k разделена на \mathbf{W}_E^k (для обучения информации о содержании токена) и \mathbf{W}_R^k (для обучения информации о позиции токена)

Adaptive Attention Span

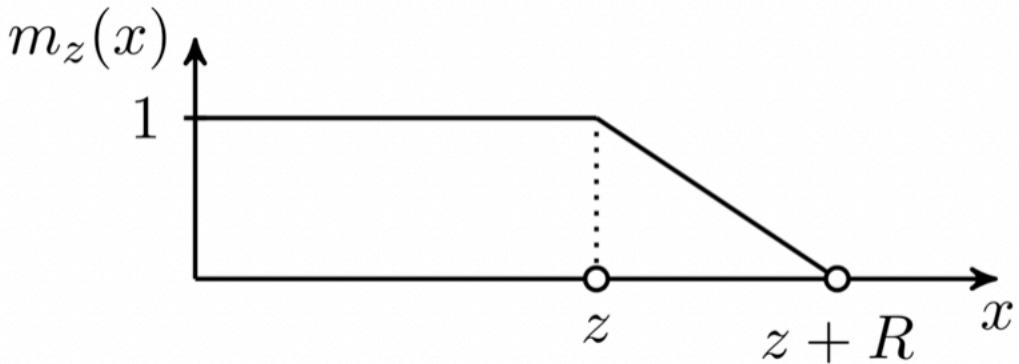
- Разным головам attention может быть нужен attention span разной длины, поэтому можно для каждой головы подбирать длину отдельно.



Adaptive Attention Span

soft mask function:

$$m_z(x) = \text{clamp}\left(\frac{1}{R}(R + z - x), 0, 1\right)$$



$$a_{ij} = \frac{m_z(i - j) \exp(s_{ij})}{\sum_{r=i-s}^{i-1} m_z(i - r) \exp(s_{ir})}$$