

$$x_2 \longrightarrow \cdots \longrightarrow x_{k-1} \longrightarrow x_k \longrightarrow x_{k+1} \longrightarrow \cdots \longrightarrow x_N$$

Decoder

$$\cdots \longrightarrow t_{k-1} \longrightarrow t_k \longrightarrow \cdots$$

Attention
layer

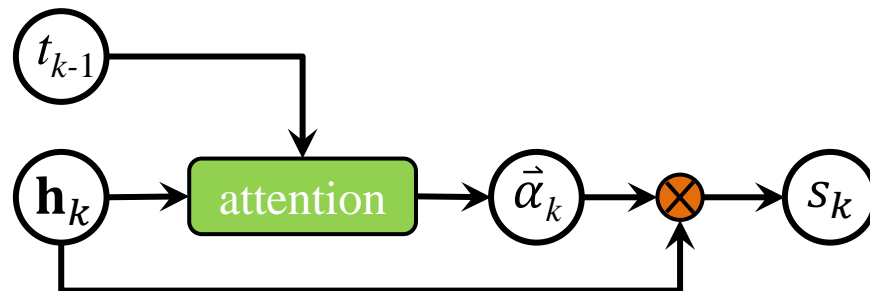
Attention function

Encoder

$$h_1 \longrightarrow \cdots \longrightarrow h_{k-2} \longrightarrow h_{k-1} \longrightarrow h_k \longrightarrow \cdots \longrightarrow h_{N-1}$$

$$x_1 \longrightarrow \cdots \longrightarrow x_{k-2} \longrightarrow x_{k-1} \longrightarrow x_k \longrightarrow \cdots \longrightarrow x_{N-1}$$

Attention layer



Attention layer (with coverage)

