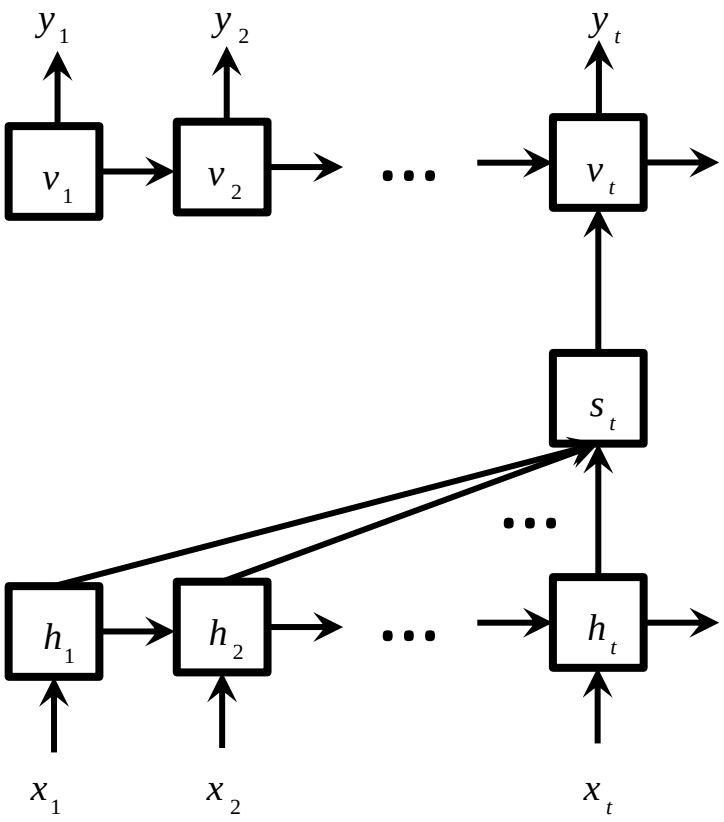
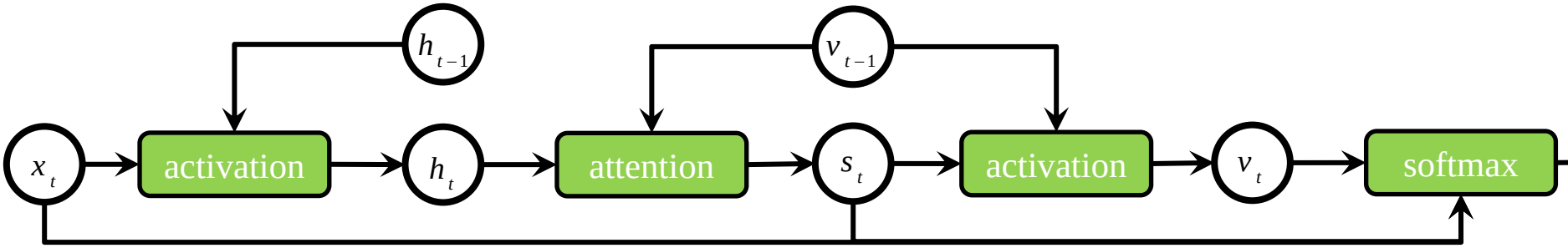
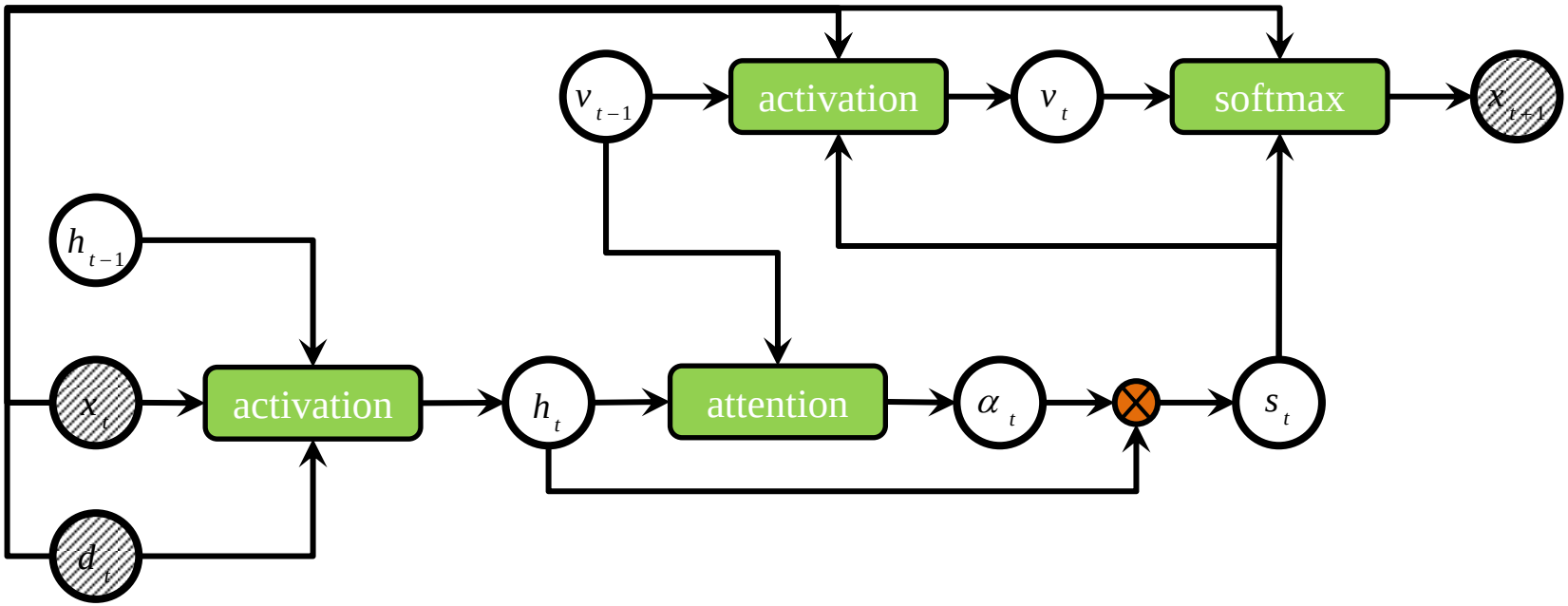
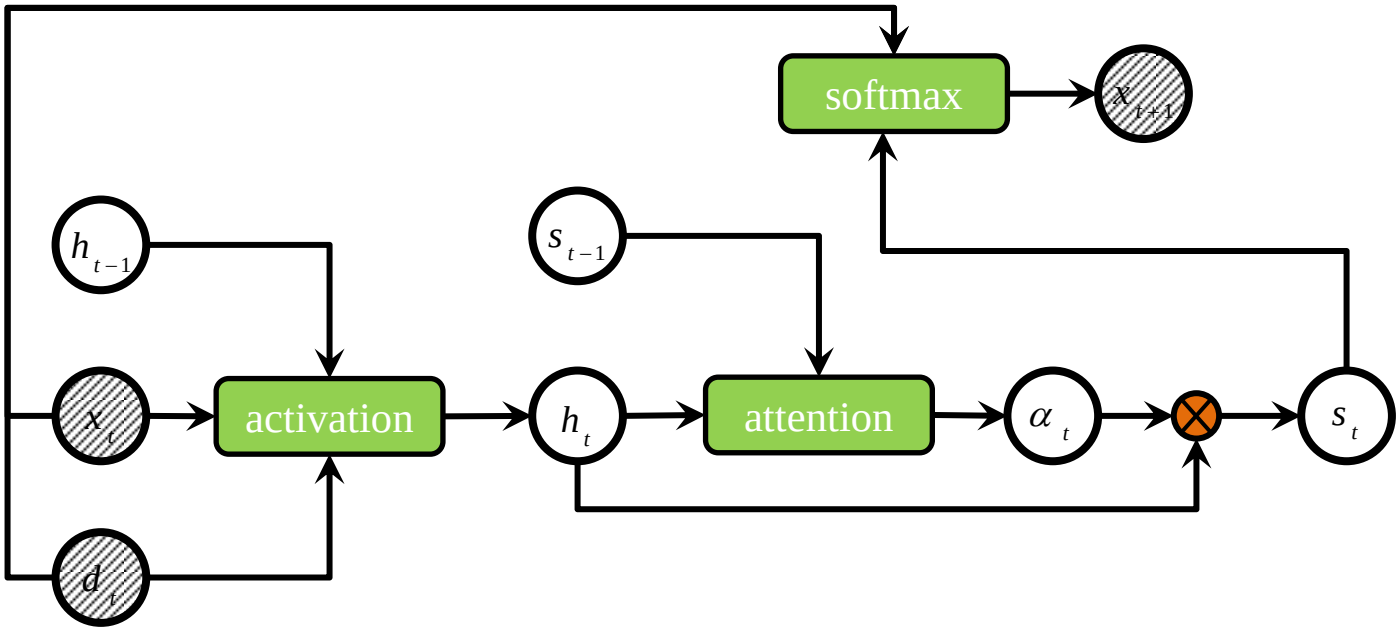
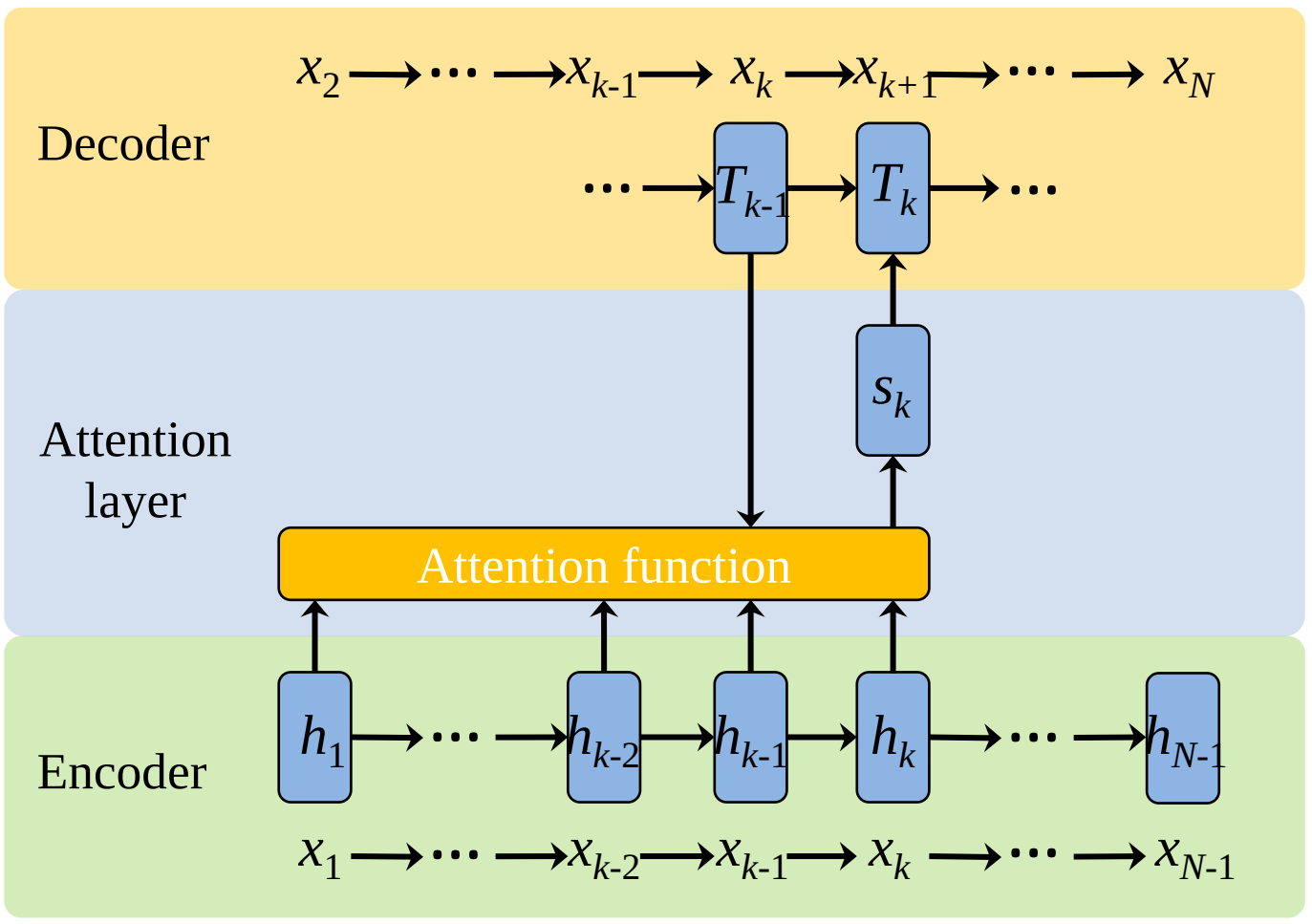
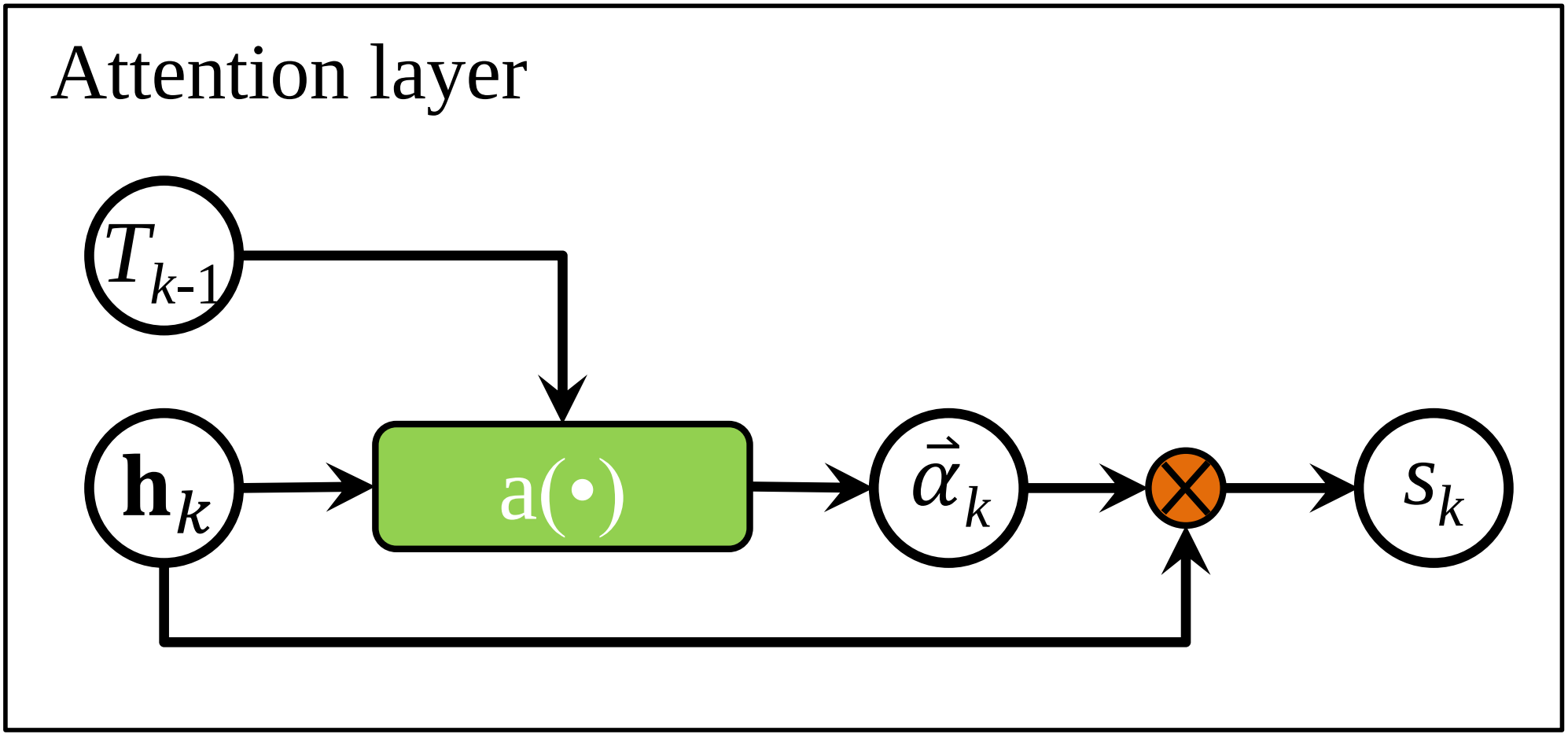

y_1
y_2
y_t
v_1
v_2
v_t
s_t
h_1
h_2
h_t
x_1
x_2
x_t


h_{t-1}
v_{t-1}
x_t
activation
h_t
attention
s_t
activation
v_t
softmax

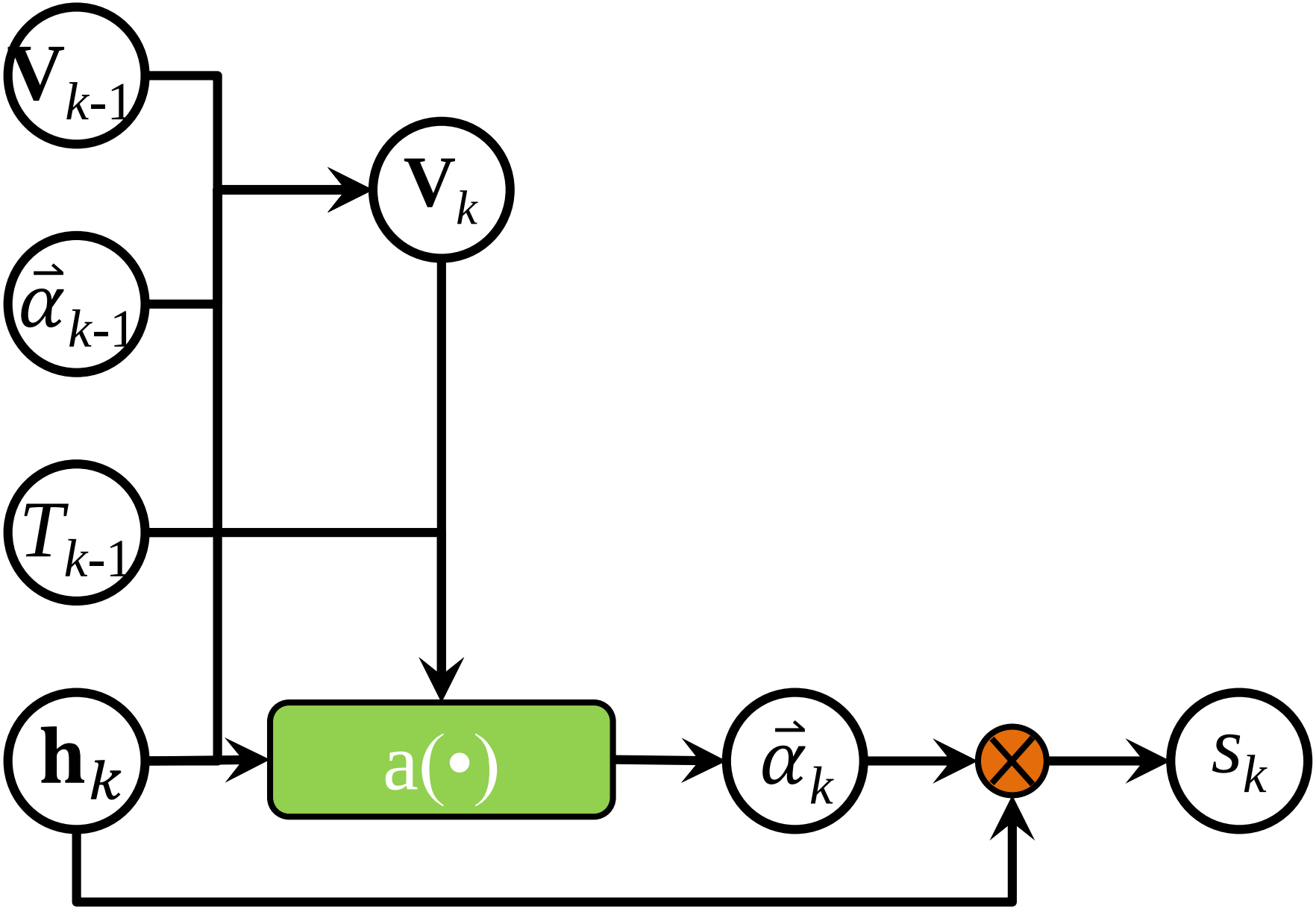

v_{t-1}
activation
v_t
softmax
x_{t+1}
h_{t-1}
x_t
activation
h_t
attention
α_t
s_t
d_t

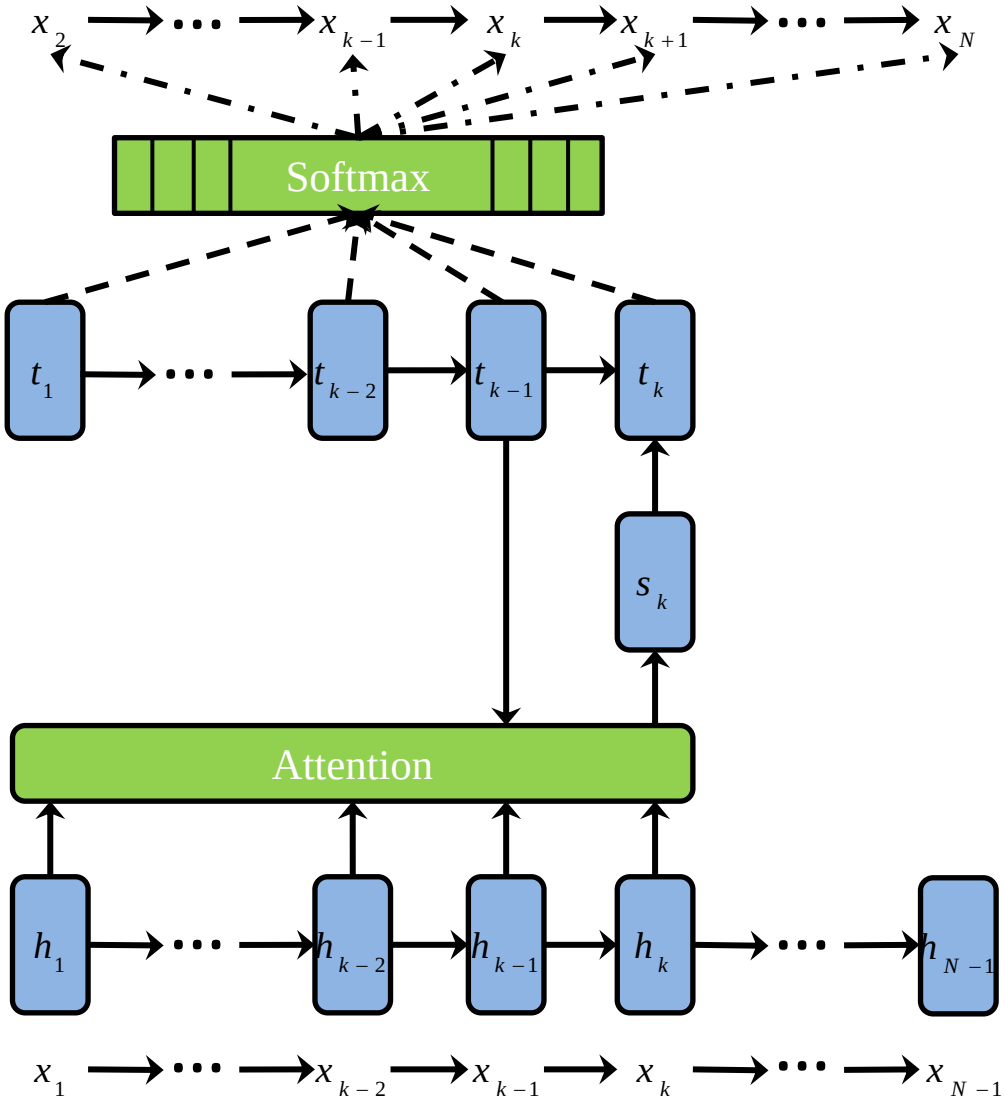

softmax
x_{t+1}
h_{t-1}
s_{t-1}
x_t
activation
h_t
attention
α_t
s_t
d_t


Decoder
x_2 ⋯ x_{k-1} x_k x_{k+1} ⋯ x_N
⋯ T_{k-1} T_k ⋯
Attention layer
s_k
Attention function
Encoder
h_1 ⋯ h_{k-2} h_{k-1} h_k ⋯ h_{N-1}
x_1 ⋯ x_{k-2} x_{k-1} x_k ⋯ x_{N-1}


Attention layer
T_{k-1}
$\mathbf{h}_k$
a(•)
$\vec{\alpha}_k$
s_k

Attention layer (with coverage)




x_2 ... x_{k-1} x_k x_{k+1} ... x_N
Softmax
t_1 ... t_{k-2} t_{k-1} t_k
s_k
Attention
h_1 ... h_{k-2} h_{k-1} h_k ... h_{N-1}
x_1 ... x_{k-2} x_{k-1} x_k ... x_{N-1}