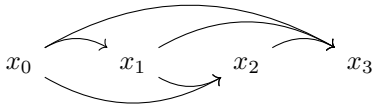
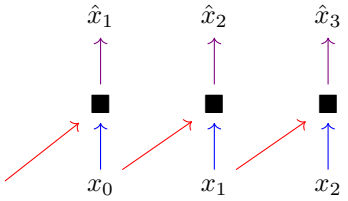
autoregressive network

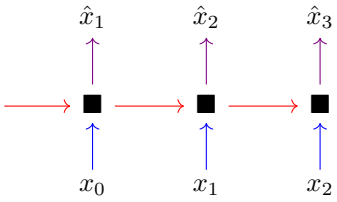
$$x_{0\ldots i} \rightarrow x_{i+1}$$



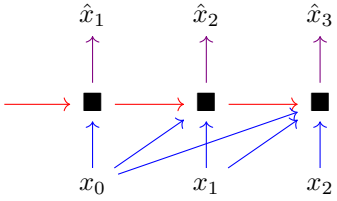
convolution



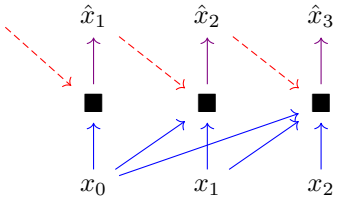
recurrent



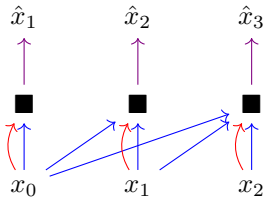
recurrent attention



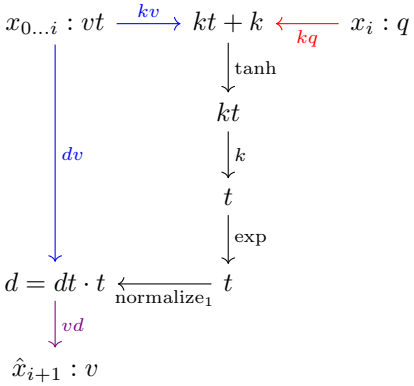
recurrent attention redirected



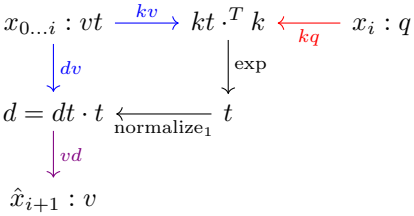
transformer



additive attention



multiplicative attention



dv is unnecessary. since $vd \cdot (dv \cdot vt) \cdot t = (vd \cdot dv) \cdot (vt \cdot t)$, the model could simply learn vd as $vd \cdot dv$.

kv is unnecessary. since the model could simply learn kq as $vk \cdot kq$.

$$
\begin{aligned}
v, k, q &: \mathbb{N}_{>0} \\
V &: \mathbb{R}^v \\
Q &: \mathbb{R}^q \\
K^V &: \mathbb{R}^v \to \mathbb{R}^k \\
K^Q &: \mathbb{R}^q \to \mathbb{R}^k \\
V^K &= \text{transpose}\, K^V \\
Q^K &= \text{transpose}\, K^Q \\
(K^V \cdot V) \cdot (K^Q \cdot Q) &= (Q^K K^V \cdot V) \cdot Q \\
&= V \cdot (V^K K^Q \cdot Q)
\end{aligned}
$$