Qwen3-Next

zero-centered RMSNorm

$$\text{Qwen3-MoE-RMSNorm(x)} = \frac{x}{\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2 + \epsilon}} \cdot w, \quad w := 1$$

$$\text{Qwen3-next-RMSNorm(x)} = \frac{x}{\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2 + \epsilon}} \cdot (1 + w), \quad w := 0$$

Qwen3-next归一化之后相应参数能够很好约束在零中心, 确保训练初期稳定性.

Gated DeltaNet (75%)

1. Linear

$$\begin{gathered} Q, K, V, Z = W_{qkzv} h \\ b, a = W_{ba} h \end{gathered}$$

2. Conv

$$\begin{gathered} X = \text{concat}(Q, K, V) \\ \hat{X} = \text{Conv1D}(X_{t-k+1:t}), k = 4 \\ \hat{X} = \text{SiLU}(\hat{X}) \\ Q, K, V = \text{split}(\hat{X}) \end{gathered}$$

其中,

$$\sigma = \text{SiLU}(x) = \frac{x}{1 + e^{-x}}$$

3. Gate Param

$$\begin{gathered} \beta_t = \sigma(b_t) \in (0, 1) \\ \alpha_t = e^{g_t} \\ g_t = -exp(A_{log}) \cdot \text{softplus}(a_t + \Delta_t) \end{gathered}$$

其中,

$$\text{softplus}(x) = log(1 + e^x)$$

4. Gate Delta Rule

$$h = g \cdot h_{t-1} + f(Q, K, V, \beta)$$

具体地，

$$\begin{aligned} S_t &= \alpha_t S_{t-1}(I - \beta_t k_t k_t^T) + \beta_t v_t k_t^T \\ &= \alpha_t S_{t-1} + \beta_t (v_t - \alpha_t S_{t-1} k_t) k_t^T \end{aligned}$$

5. Output Gate

$$h_t^{out} = w \cdot \frac{h_t^{core}}{\sqrt{\frac{1}{d}\sum_{i=1}^{d}(h_{t,i}^{core})^2 + \epsilon}} \cdot \mathrm{SiLU}(Z_t)$$

6. Linear

$$h_t = W_o \cdot h_t^{core}$$

Linear Attention

$$\begin{aligned} O = softmax(QK^T)V, \\ \rightarrow \quad\quad O = (QK^T)V, \\ \rightarrow \quad\quad O = Q(K^TV) \\ O(n^2) \rightarrow O(n) \end{aligned}$$

具体来说，

$$\begin{aligned} o_t &= \sum_{j=1}^{t} v_j(k_j^T q_t) \\ &= \sum_{j=1}^{t} (v_j k_j^T) q_t \\ &= (\sum_{j=1}^{t} v_j k_j^T) \cdot q_t \end{aligned}$$

$$\text{记} \quad o_t = S_t \cdot q_t, \quad \text{则} \\ S_t = S_{t-1} + v_t k_t^T$$

1. 计算速度线性

2. 只需存储S

考虑到历史信息等权相加的特点，Retentive Network引入遗忘

$$S_t = \lambda S_{t-1} + v_t k_t^T \\ O = (QK^T \odot \Gamma)V \\ \Gamma_{i,j} = \begin{cases} 0 & \text{if } i < j \\ 1 & \text{if } i = j \\ \prod_{\tau=j+1}^{i} \gamma_\tau & \text{if } i > j \end{cases}$$

从测试时训练(Test Time Training)的角度

$$\begin{aligned} o_t = S_t q_t \\ S_t = S_{t-1} + v_t k_t^T \end{aligned}$$

把S_t视作优化目标,

$$\text{令} \quad f(S_{t-1}; k_t) = S_{t-1} k_t$$

根据余弦距离定义损失函数:

$$\mathcal{L}(f(S_{t-1}; k_t), v_t) = -v_t \cdot (S_{t-1} k_t)$$

$$\begin{aligned} S_t &= S_{t-1} - \eta \nabla_{S_{t-1}} \mathcal{L}(f(S_{t-1}; k_t), v_t) \\ &= S_{t-1} + v_t k_t^T \end{aligned}$$

RetNet加入了正则项:

$$
\begin{aligned}
S_t &= \lambda S_{t-1} + v_t k_t^T \\
\mathcal{L} &= -v \cdot (Sk) + \frac{1-\gamma}{2} ||S||_F^2
\end{aligned}
$$

DeltaNet

根据欧氏距离(平方损失)定义损失函数:

$$
\begin{aligned}
\mathcal{L}(f(S_{t-1}; k_t), v_t) &= \nabla_{S_{t-1}} \frac{1}{2} ||S_{t-1}k_t - v_t||^2 \\
&= (S_{t-1}k_t - v_t) \cdot k_t^T
\end{aligned}
$$

$$
S_t = S_{t-1} - \eta_t \cdot (S_{t-1}k_t - v_t) \cdot k_t^T
$$

由

$$
\eta_t \cdot (S_{t-1}k_t - v_t) \cdot k_t^T = (S_{t-1}(\sqrt{\eta_t}k_t) - (\sqrt{\eta_t}v_t))(\sqrt{\eta_t}k_t)^T
$$

仅考虑$\eta_t = 1$,

$$
\begin{aligned}
S_t &= S_{t-1} - (S_{t-1}k_t - v_t) \cdot k_t^T \\
&= S_{t-1} - S_{t-1}k_t k_t^T + v_t k_t^T \\
&= S_{t-1}(I - k_t k_t^T) + v_t k_t^T
\end{aligned}
$$

$$
S_t^{standard\_attn} = S_{t-1} + v_t k_t^T
$$

理解为先移除模型对k_t的旧认知，然后根据(k_t, v_t)补充新认知.

Gated DeltaNet

综上，S_t的两种形式:

$$
\begin{aligned}
S_t &= S_{t-1} + (v_t - S_{t-1}k_t)k_t^T \\
&= S_{t-1}(I - k_t k_t^T) + v_t k_t^T
\end{aligned}
$$

Delta Rule

$$
\begin{aligned}
S_t &= S_{t-1} - \underbrace{(S_{t-1}k_t)}_{v_t^{old}} k_t^T + \underbrace{(\beta_t v_t + (1-\beta_t)S_{t-1}k_t)}_{v_t^{new}} k_t^T \\
&= S_{t-1}(I - \beta_t k_t k_t^T) + \beta_t v_t k_t^T
\end{aligned}
$$

GDN

$$
\begin{aligned}
S_t &= \alpha_t S_{t-1}(I - \beta_t k_t k_t^T) + \beta_t v_t k_t^T \\
&= \alpha_t S_{t-1} + \beta_t (v_t - \alpha_t S_{t-1} k_t) k_t^T
\end{aligned}
$$