Rectifying Magnitude Neglect in Linear Attention

Supplementary Material

A. Detailed Proofs

Relationship between β/γ and β_{new}/γ_{new}. After the magnitude of $\phi(Q_i)$ is scaled by a factor of $a > 1$, we have:

$$\begin{aligned} \gamma_{new} &= \frac{a\phi(Q_i)\sum_{m=1}^{N}\phi(K_m)^T}{N} = a\gamma, \\ \beta_{new} - 1 &= \frac{1}{a\phi(Q_i)\sum_{m=1}^{N}\phi(K_m)^T} = \frac{1}{a}(\beta - 1); \end{aligned} \tag{1}$$

Based on the above two equations, we derive the following result:

$$\begin{aligned} \gamma_{new} &= a\gamma, \\ \beta_{new} &= \frac{\beta + a - 1}{a}; \end{aligned} \tag{2}$$

Proof of $\frac{a\beta}{a+\beta-1} > 1$. Since $\beta > 1$ and $a > 1$, we have:

$$(a-1)(\beta-1) > 0, \tag{3}$$

Expanding the equation, we obtain:

$$\begin{aligned} a\beta - (a + \beta - 1) &> 0, \\ a\beta &> a + \beta - 1; \end{aligned} \tag{4}$$

Since $a + \beta > 1 + 1 = 2 > 1$, so $a + \beta - 1 > 0$. Based on Eq. 4, we have:

$$\frac{a\beta}{a+\beta-1} > 1 \tag{5}$$

Proof of $p_m > p$. We define:

$$\begin{aligned} A_m &= \beta\phi(Q_i)\phi(K_m)^T, \\ A_n &= \beta\phi(Q_i)\phi(K_n)^T; \end{aligned} \tag{6}$$

Where we assume that Q_i allocate more attention to K_m, thus $A_m > A_n$. Consider a function of x:

$$f(x) = \frac{A_m - \gamma x}{A_n - \gamma x}; \tag{7}$$

The derivative of the function $f(x)$ can be expressed as:

$$f'(x) = \frac{\gamma(A_m - A_n)}{(A_n - \gamma x)^2} > 0; \tag{8}$$

Since we only consider the positive attention scores, we have $\frac{A_n}{\gamma} > \frac{a\beta}{a+\beta-1} > 1$, the function $f(x)$ is monotonically increasing. Thus $p_m = f(\frac{a\beta}{a+\beta-1}) > f(1) = p$.

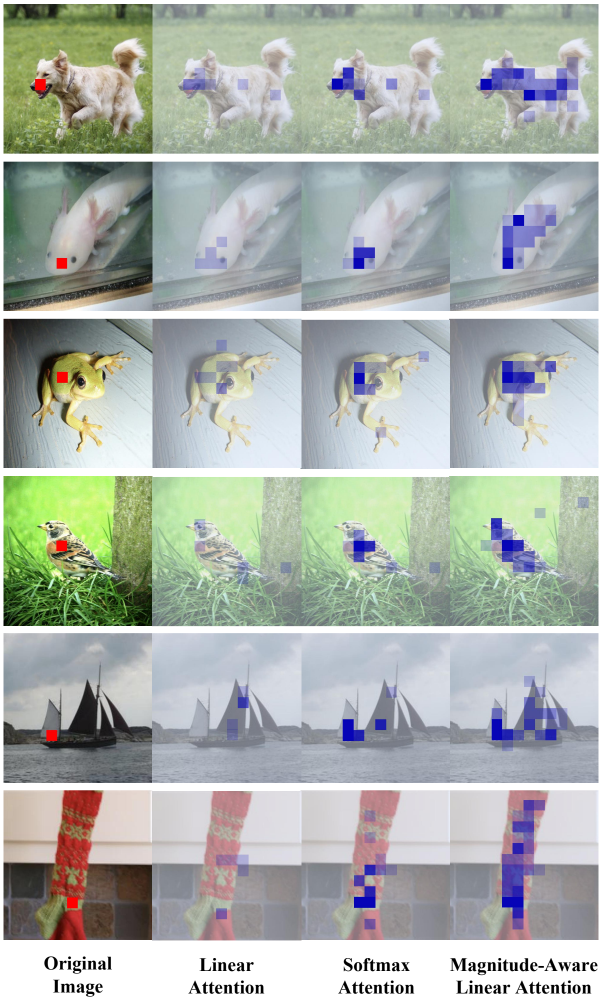

Figure 1. The distribution of attention scores from DeiT-T. Feature corresponding to the red block is used as query.

B. Visualization on Natural Images

As shown in Fig. 1, we present the distribution trends of attention scores on natural images. It can be observed that the attention scores of Linear Attention are overly smooth, whereas Softmax Attention is spiky and excessively focuses on local information. Magnitude-Aware Linear Attention effectively balances the characteristics of both.

References