

激活函数

GELU (Gaussian Error Linear units) 激活函数的精确定义:

$$GELU(x) = x \cdot P(X \leq x) = x \cdot \Phi(x)$$

其中, $\Phi(x)$ 是标准正态分布的累积分布函数 (CDF)

真正使用的是其近似公式:

$$GELU(x) \approx 0.5 * x * \left(1 + \tanh \left(\sqrt{\frac{2}{\pi}} * (x + 0.044715 * x^3) \right) \right)$$

其中 $\tanh(x)$ 表达式:

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

我们定义 $g(x)$ 为 GELU 的近似公式:

$$g(x) = 0.5 * x * \left(1 + \tanh \left(\sqrt{\frac{2}{\pi}} * (x + 0.044715 * x^3) \right) \right)$$

前向传播

这个激活函数应用于输入的逐元素计算, 只需要将输入张量展平, 对其中的每个元素使用下面的近似公式计算即可. 输出形状与输入形状完全相同.

$$g(x) = 0.5 * x * \left(1 + \tanh \left(\sqrt{\frac{2}{\pi}} * (x + 0.044715 * x^3) \right) \right)$$

反向传播公式

反向传播的目标是计算损失函数 L 对输入 x 的梯度 $\frac{\partial L}{\partial x}$. 我们已经得到损失函数 L 对 GELU 的输出 $g(x)$ 的梯度 $\frac{\partial L}{\partial g}$:

计算 $\tanh(x)$ 的导数

$$\begin{aligned}
\frac{\mathrm{d} \tanh (x)}{\mathrm{d} x} &= \frac{\mathrm{d}\left(\frac{e^x-e^{-x}}{e^x+e^{-x}}\right)}{\mathrm{d} x} \\
&= \frac{\left(e^x+e^{-x}\right)\left(e^x+e^{-x}\right)-\left(e^x-e^{-x}\right)\left(e^x-e^{-x}\right)}{\left(e^x+e^{-x}\right)^2} \\
&= 1-\tanh ^2(x)
\end{aligned}$$

设:

$$u(x)=\sqrt{\frac{2}{\pi}}*(x+0.044715*x^3)$$

则:

$$\frac{\mathrm{d} u(x)}{\mathrm{d} x}=\sqrt{\frac{2}{\pi}}\left(1+0.044715*3*x^2\right)$$

得到:

$$\frac{\mathrm{d} g}{\mathrm{d} x}=0.5*(1+\tanh (u(x)))+0.5*x*(1-\tanh ^2(u(x)))*\frac{\mathrm{d} u(x)}{\mathrm{d} x}$$

最终得到 $\frac{\partial L}{\partial x}$:

$$\begin{aligned}
&\frac{\partial L}{\partial x}=\frac{\partial L}{\partial g} \frac{\mathrm{d} g}{\mathrm{d} x} \\
&=\frac{\partial L}{\partial g} *\left[0.5*(1+\tanh (u(x)))+0.5*x*(1-\tanh ^2(u(x)))*\frac{\mathrm{d} u(x)}{\mathrm{d} x}\right]
\end{aligned}$$