激活函数

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

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

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:

$$GELU\left(x
ight)pprox0.5*x*\left(1+ anh\left(\sqrt{rac{2}{\pi}}*\left(x+0.044715*x^3
ight)
ight)
ight)$$

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:

$$g\left(x
ight)=0.5*x*\left(1+ anh\left(\sqrt{rac{2}{\pi}}*\left(x+0.044715*x^3
ight)
ight)
ight)$$

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.

$$g\left(x
ight) = 0.5*x*\left(1+ anh\left(\sqrt{rac{2}{\pi}}*\left(x+0.044715*x^3
ight)
ight)
ight)$$

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 $g\left(x\right)$ 的梯度 $\frac{\partial L}{\partial a}$:

计算 tanh(x) 的导数

$$egin{aligned} rac{\mathrm{d} anh\left(x
ight)}{\mathrm{d} x} &= rac{\mathrm{d} \left(rac{e^x-e^{-x}}{e^x+e^{-x}}
ight)}{\mathrm{d} x} \ &= rac{\left(e^x+e^{-x}
ight)\left(e^x+e^{-x}
ight)-\left(e^x-e^{-x}
ight)\left(e^x-e^{-x}
ight)}{\left(e^x+e^{-x}
ight)^2} \ &= 1- anh^2\left(x
ight) \end{aligned}$$

设:

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

则:

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

得到:

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

最终得到 $\frac{\partial L}{\partial x}$:

$$egin{align*} rac{\partial L}{\partial x} &= rac{\partial L}{\partial g} rac{\mathrm{d}g}{\mathrm{d}x} \ &= rac{\partial L}{\partial g} * \left[0.5 * \left(1 + anh \left(u \left(x
ight)
ight)
ight) + 0.5 * x * \left(1 - anh^2 \left(u \left(x
ight)
ight)
ight) * rac{\mathrm{d}u \left(x
ight)}{\mathrm{d}x}
ight] \end{aligned}$$