激活函数

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

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

其中, $\Phi\left(x\right)$ 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:

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

其中 $\tanh\left(x\right)$ 表达式:

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

我们定义 $g\left(x\right)$ 为 GELU的近似公式:

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

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.

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

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

计算 $\tanh\left(x\right)$ 的导数

$$\frac{\mathrm{d}\tanh\left(x\right)}{\mathrm{d}x}=\frac{\mathrm{d}\left(\frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}\right)}{\mathrm{d}x}$$

$$=\frac{\left(e^{x}+e^{-x}\right)\left(e^{x}+e^{-x}\right)-\left(e^{x}-e^{-x}\right)\left(e^{x}-e^{-x}\right)}{\left(e^{x}+e^{-x}\right)^{2}}$$

$$=1-\tanh^{2}\left(x\right)$$

设:

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

则:

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

得到:

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

最终得到 $\frac{\partial L}{\partial x}$:

$$\frac{\partial L}{\partial x}=\frac{\partial L}{\partial g}\frac{\mathrm{d}g}{\mathrm{d}x}$$

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