优化器

概念定义

深度学习中的优化器主要用于调整模型参数，以最小化损失函数。不同的优化器通过不同的方式更新权重，从而影响训练的速度和效果。

常见的优化器

梯度下降 (Gradient Descent, GD)

梯度下降是最基本的优化算法。其核心思想是沿着损失函数梯度的反方向更新参数，以期望找到全局最小值。参数更新公式如下：

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta J(\theta_t)$$

其中，θ 是待优化参数；$J(\theta)$ 是损失函数；η 是学习率(Learning Rate)，控制更新步长。

随机梯度下降 (Stochastic Gradient Descent, SGD)

随机梯度下降(SGD)是梯度下降的一种改进。与传统的GD不同，SGD每次只使用一个样本来估计梯度并更新参数。这使得更新过程更加随机，但也能更快地收敛，并可能跳出局部最优解。其计算公式与GD相同，但由于使用单个样本，因此梯度的估计更不稳定。

小批量梯度下降 (Mini-batch Gradient Descent)

结合了GD和SGD的优点，小批量梯度下降每次使用一个小批量的数据来估计梯度，既保持了稳定性也提高了速度。其计算公式与GD相同，但是基于小批量样本的平均梯度进行更新。

动量法 (Momentum)

动量方法通过加入先前梯度的指数加权平均值来加速SGD在相关方向上的更新，并抑制震荡。其更新公式如下：

$$g_{t+1} = \nabla_\theta J(\theta_t) + \lambda \theta_t$$

$$v_{t+1} = \mu v_t + (1 - \tau) g_{t+1}$$

$$\theta_{t+1} = \theta_t - \eta v_{t+1}$$

其中， λ 是权重衰减系数(L2范数)，有助于提高模型的泛化性能； v_t 是 t 时刻的动量； μ 是动量因子，控制历史梯度的影响程度，取值通常在 0 和 1 之间； τ 是阻尼系数； η 是学习率。

AdaGrad (Adaptive Gradient Algorithm)

AdaGrad是一种自适应学习率的优化方法，它根据参数的梯度大小来调整学习率，适合稀疏数据(如NLP)。其更新公式如下：

$$g_{t+1} = \nabla_\theta J(\theta_t) + \lambda \theta_t$$

$$G_t = G_{t-1} + g_{t+1}^2$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{G_t + \epsilon}} g_{t+1}$$

其中， λ 是权重衰减系数； G_t 是梯度的累积平方和； ϵ 是一个很小的数，防止分母为零。

RMSprop (Root Mean Square Propagation)

RMSprop是AdaGrad的一种改进，它通过对过去梯度的指数加权平均值来调整学习率，从而解决了AdaGrad在非凸优化问题中的问题。其更新公式如下：

$$g_{t+1} = \nabla_\theta J(\theta_t) + \lambda \theta_t$$

$$E[g^2]_t = \beta E[g^2]_{t-1} + (1 - \beta) g_{t+1}^2$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} g_{t+1}$$

其中， λ 是权重衰减系数； $E[g^2]_t$ 是梯度的指数加权移动平均； β 是控制历史梯度影响程度的超参数。

Adam (Adaptive Moment Estimation)

Adam是一种结合了动量法和RMSprop的优化方法，它结合了两者的优点，能够自适应地调整学习率。其更新公式如下：

$$g_{t+1} = \nabla_\theta J(\theta_t) + \lambda \theta_t$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_{t+1}$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_{t+1}^2$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}}\hat{m}_t$$

其中， λ 是权重衰减系数； m_t 和 v_t 分别是梯度的一阶和二阶动量估计； β_1 和 β_2 是超参数，控制动量的影响程度； $\hat{m}_t$ 和 $\hat{v}_t$ 是修正后的一阶和二阶动量估计； ϵ 是一个很小的数，防止分母为零。

AdamW (Adam with Weight Decay)

AdamW是Adam的一种改进，它在Adam的基础上加入了权重衰减(Weight Decay)，用于正则化。其更新公式如下：

$$g_{t+1} = \nabla_\theta J(\theta_t)$$

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_{t+1}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_{t+1}^2$$

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}}\hat{m}_t - \eta\lambda\theta_t$$

其中， λ 是权重衰减系数(从通过作用于 g_{t+1} 间接作用于 θ_{t+1} 变为直接租用于 θ_{t+1})。前面所述的优化器，权重衰减是通过作用于 g_t 实现的，在Adam中，这会导致权重衰减的梯度也会随着 g_t 除以分母(累积平方梯度)。当累积平方梯度(分母)过大时，权重衰减的作用就会被大大地削弱。因此，在AdamW中，权重衰减直接作用于 θ_{t+1} ，更有利于提升模型的泛化性能。