反向传播算法

前面我们介绍了三个模型，整个处理的基本流程都是定义模型，读入数据，给出损失函数f，通过梯度下降法更新参数。PyTorch 提供了非常简单的自动求导帮助我们求解导数，对于比较简单的模型，我们也能手动求出参数的梯度，但是对于非常复杂的模型，比如一个 100 层的网络，我们如何能够有效地手动求出这个梯度呢？这里就需要引入反向传播算法，自动求导本质是就是一个反向传播算法。

反向传播算法是一个有效地求解梯度的算法，本质上其实就是一个链式求导法则的应用，然而这个如此简单而且显而易见的方法却是在 Roseblatt 提出感知机算法后将近 30 年才被发明和普及的，对此 Bengio 这样说道："很多看似显而易见的想法只有在事后才变得的显而易见。"

下面我们就来详细将一讲什么是反向传播算法。

链式法则

首先来简单地介绍一下链式法则，考虑一个简单的函数，比如 $f(x, y, z) = (x + y)z$

我们当然可以直接求出这个函数的微分，但是这里我们要使用链式法则，令 $q=x+y$

那么

$f = qz$

对于这两个式子，我们可以分别求出他们的微分

$\frac{\partial f}{\partial q} = z, \frac{\partial f}{\partial z} = q$

同时q是x和y的求和，所以我们能够得到

$\frac{\partial q}{x} = 1, \frac{\partial q}{y} = 1$

我们关心的问题是

$\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}$

链式法则告诉我们如何来计算出他们的值

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial x} \tag{1}$$

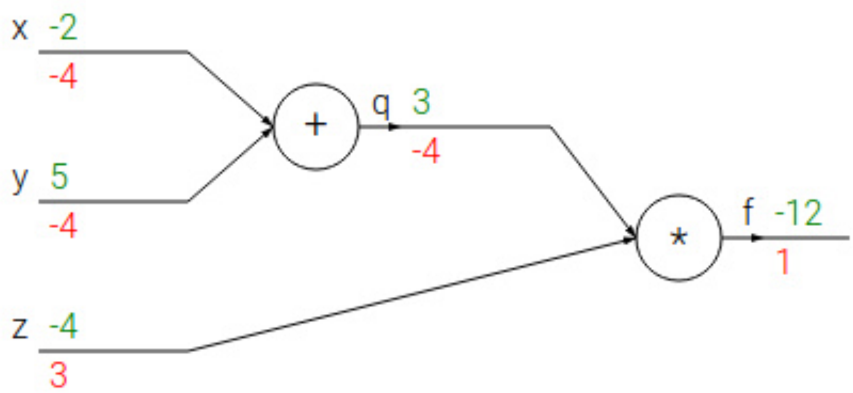
$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial q}\frac{\partial q}{\partial y} \tag{2}$$

$$\frac{\partial f}{\partial z} = q \tag{3}$$

通过链式法则我们知道如果我们需要对其中的元素求导，那么我们可以一层一层求导然后将结果乘起来，这就是链式法则的核心，也是反向传播算法的核心，更多关于链式法则的算法，可以访问这个文档

反向传播算法

了解了链式法则，我们就可以开始介绍反向传播算法了，本质上反向传播算法只是链式法则的一个应用。我们还是使用之前那个相同的例子$q=x+y, f=qz$，通过计算图可以将这个计算过程表达出来



上面绿色的数字表示其数值，下面红色的数字表示求出的梯度，我们可以一步一步看看反向传播算法的实现。首先从最后开始，梯度当然是1，然后计算

$\frac{\partial f}{\partial q} = z = -4, \ \frac{\partial f}{\partial z} = q = 3$

接着我们计算 $\frac{\partial f}{\partial x} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial x} = -4 \times 1 = -4, \ \frac{\partial f}{\partial y} = \frac{\partial f}{\partial q} \frac{\partial q}{\partial y} = -4 \times 1 = -4$

这样一步一步我们就求出了$\nabla f(x, y, z)$。

直观上看反向传播算法是一个优雅的局部过程，每次求导只是对当前的运算求导，求解每层网络的参数都是通过链式法则将前面的结果求出不断迭代到这一层，所以说这是一个传播过程

Sigmoid函数举例

下面我们通过Sigmoid函数来演示反向传播过程在一个复杂的函数上是如何进行的。

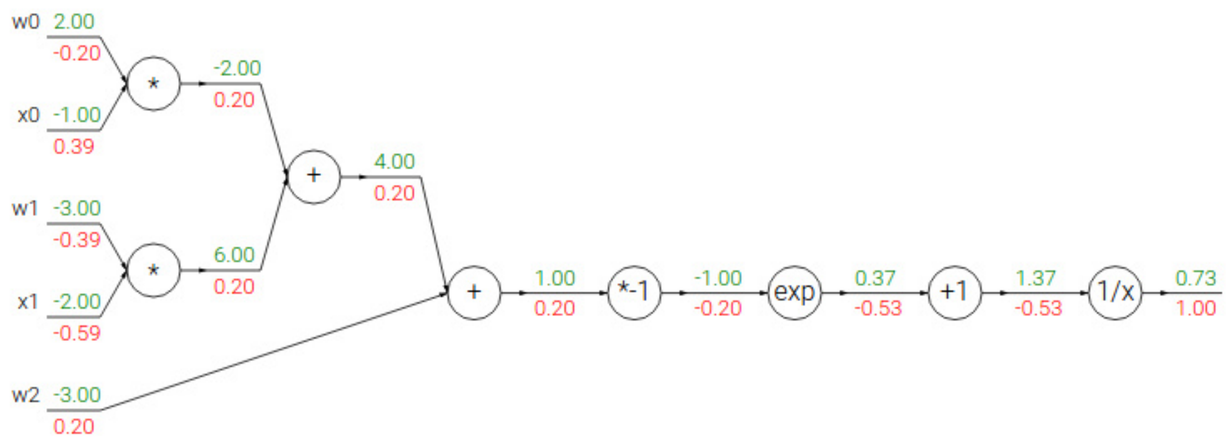
$$f(w, x) = \frac{1}{1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)}} \tag{4}$$

我们需要求解出 $\frac{\partial f}{\partial w_0}, \frac{\partial f}{\partial w_1}, \frac{\partial f}{\partial w_2}$

首先我们将这个函数抽象成一个计算图来表示，即

$$
\begin{aligned}
f(x) &= \frac{1}{x} \\
f_c(x) &= 1 + x \\
f_e(x) &= e^x \\
f_w(x) &= -(w_0 x_0 + w_1 x_1 + w_2)
\end{aligned}
\tag{5}
$$

这样我们就能够画出下面的计算图



同样上面绿色的数子表示数值，下面红色的数字表示梯度，我们从后往前计算一下各个参数的梯度。首先最后面的梯度是1,，然后经过$\frac{1}{x}$这个函数，这个函数的梯度是$-\frac{1}{x^2}$，所以往前传播的梯度是$1 \times -\frac{1}{1.37^2} = -0.53$，然后是$+1$这个操作，梯度不变，接着是$e^x$这个运算，它的梯度就是$-0.53 \times e^{-1} = -0.2$，这样不断往后传播就能够求得每个参数的梯度。