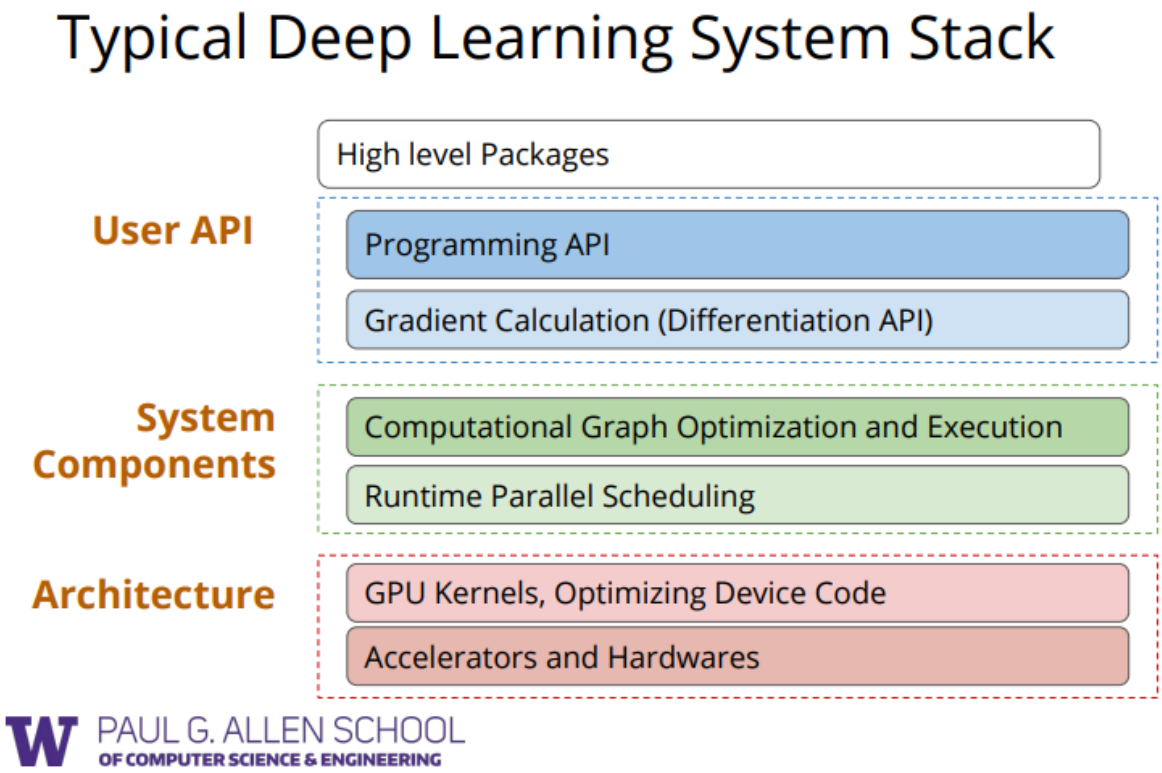
思路

内容

参考资料：

- CSE 599W： Systems for ML博客
- CSE 599W： Systems for ML



- 调用API
  - python
  - c++/cuda
  - 其他硬件加速方法
- 自动求导方法（`autodiff`）
- 内存共享优化
- 网络结构表示方法
- 计算图执行和优化

程序求导的四种方法

参考资料：

- 自动微分(Automatic Differentiation)简介
- Automatic Differentiation in Machine Learning: a Survey
- Dual Numbers & Automatic Differentiation

手动求导 Manual Derivatives

这种求导方法在传统计算机视觉模型中比较常用，也就是模型方法会定义一个能量函数之类的量。需要优化的变量则通过对能量函数进行理论求导之后再在代码中实现。很明显，这种方法几乎没有什么可拓展性。

数值微分 Numerical Differentiation

主要利用导数的定义：

$$\frac{\partial f(x)}{\partial x_i} \approx \frac{f(x + he_i) - f(x)}{h} \tag{1}$$

这样输出量$f(x)$的梯度$\nabla f = \left(\frac{\partial f}{\partial x_i}, \cdots, \frac{\partial f}{\partial x_n}\right)$，其中$e_i$是第$i$个元素为1其他为0的单位向量，$h$是一个很小的步长。这种方法比较容易实现，但是存在比较多的问题。

第一，这种方法只能近似。误差来源主要有两个，第一个是截断误差（truncate error），这是式(???)造成的，主要是由于$h \neq 0$引起的；另一个误差来源是舍入误差（round-off error），主要是由于计算机本身表示上无法完全与理论相等，$f(x + he_i)$与$f(x)$在表示时存在误差。当$h \to 0$时，截断误差趋于0，但是舍入误差占主导；而随着h增大，截断误差则占据主导。

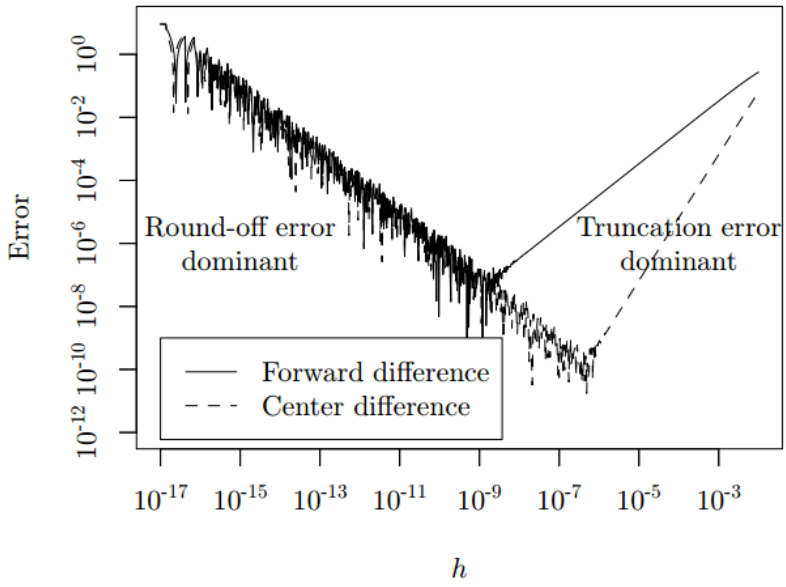


Figure 3: Error in the forward (Eq. 1) and center difference (Eq. 2) approximations as a function of step size h, for the derivative of the truncated logistic map $f(x) = 64x(1-x)(1-2x)^2(1-8x+8x^2)^2$. Plotted errors are computed using $E_{\text{forward}}(h, x_0) = \left|\frac{f(x_0+h)-f(x_0)}{h} - \frac{d}{dx}f(x)\big|_{x_0}\right|$ and $E_{\text{center}}(h, x_0) = \left|\frac{f(x_0+h)-f(x_0-h)}{2h} - \frac{d}{dx}f(x)\big|_{x_0}\right|$ at $x_0 = 0.2$.

一种改进方法是不用式(???)的前向方式，改为中心式的：

$$\frac{\partial f(x)}{\partial x_i} \approx \frac{f(x + he_i) - f(x - he_i)}{2h} + O(h^2) \tag{2}$$

这能去掉一阶截断误差（当然更高阶的截断误差仍然存在）。由式(???)，每次计算一个方向的梯度就要执行两次函数f。对于一个n维的输入变量和一个$f : \mathbb{R}^n \to \mathbb{R}^m$，计算一个雅可比矩阵需要执行$2mn$次$f$函数。

第二个问题是，各个维度的敏感度不同，步长h不能很好的确定。如果x本身的量级与h差不多，这种方法就会造成问题。

第三个问题，也是这种求导方法最主要的问题就是计算的复杂度。当n增大到成千上万时，计算这一梯度就成了主要的问题。相比于第一个误差问题，在深度学习的语境下，这种误差的容忍度较高。

符号微分 Symbolic Differentiation

符号求导在现在的一些数学软件如`Mathematica/Maple`中已经应用了，比如针对复合函数：

$$\frac{d}{dx}(f(x)+g(x)) = \frac{d}{dx}f(x) + \frac{d}{dx}g(x) \tag{3}$$

$$\frac{d}{dx}f(x)g(x) = \left(\frac{d}{dx}f(x)\right)g(x) + f(x)\left(\frac{d}{dx}g(x)\right) \tag{4}$$

$$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2} \tag{5}$$

符号微分旨在为人提供一种直观的闭式解的自动微分。因此如果能将问题转化为一个纯数学符号问题，那么也就能用这类符号微分方法进行自动求解了。

符号微分自然也存在问题。其一是带求解问题必须能转化为一个数学符号表达式；其二，更重要的问题是，随着复合函数嵌套层数的增加，符号微分会遇到所谓的“表达式膨胀”（expression swell）问题：

Table 1: Iterations of the logistic map $l_{n+1} = 4l_n(1-l_n)$, $l_1 = x$ and the corresponding derivatives of l_n with respect to x, illustrating expression swell.

n	l_n	$\frac{d}{dx}l_n$	$\frac{d}{dx}l_n$ (Simplified form)
1	x	1	1
2	$4x(1-x)$	$4(1-x)-4x$	$4-8x$
3	$16x(1-x)(1-2x)^2$	$16(1-x)(1-2x)^2-16x(1-2x)^2-64x(1-x)(1-2x)$	$16(1-10x+24x^2-16x^3)$
4	$64x(1-x)(1-2x)^2(1-8x+8x^2)^2$	$128x(1-x)(-8+16x)(1-2x)^2(1-8x+8x^2)+64(1-x)(1-2x)^2(1-8x+8x^2)^2-64x(1-2x)^2(1-8x+8x^2)^2-256x(1-x)(1-2x)(1-8x+8x^2)^2$	$64(1-42x+504x^2-2640x^3+7040x^4-9984x^5+7168x^6-2048x^7)$

如果不加处理，为了计算嵌套函数的梯度，可能需要多次执行同一个表达式，这就造成实际所需的符号表达式将呈指数级增长，比如中间一列。事实上，我们可以看到$n=4$时的导数中有很多基本表达式在之前也出现过，我们可以保留一些中间结果避免再次计算。

自动微分 Auto Differentiation

自动微分技术可以看成是在执行一个计算机程序，只不过其中一步可能是对某些公式进行求导。由于所有数学计算最终都可以被分解为有限个基本操作，并且这些基本运算的梯度是已知的，通过链式法则对这些导数进行运算和组合就能计算出完整的结果。这些基本算子包括：二值逻辑运算，单元符号转换运算，超越函数（比如指数），对数函数和三角函数等。现在的深度学习框架都是使用AD方法实现自动求导的。

自动微分技术包括两种模式：前向模式（forward mode / tangent linear mode）和反向模式（reverse mode / cotangent linear mode）。假定一个函数$f(x_1,x_2)=\ln(x_1)+x_1x_2-\sin(x_2)$，并定义：

- 变量$v_{i-n}=x_i, i=1,\cdots,n$为输入变量；
- 变量$v_i\ i=1,\cdots,l$是中间变量；
- 变量$y_{m-i}=v_{l-i},\ i=m-1,\cdots,0$为输出变量。

现在通过对这一函数的求导过程来解释AD的前向和反向模式。

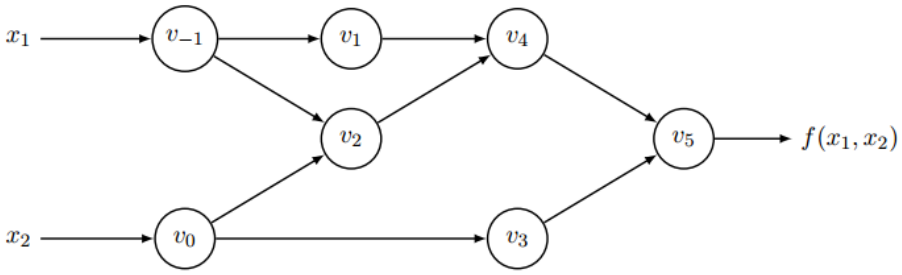


Figure 4: Computational graph of the example $f(x_1, x_2) = \ln(x_1) + x_1 x_2 - \sin(x_2)$. See the primal trace in Tables 2 or 3 for the definitions of the intermediate variables $v_{-1} \dots v_5$.

前向模式

前向模式的思路比较简单直接：根据计算图，我们利用链式法则自前向后逐个计算各中间变量相对于输入变量的导数：

$$\dot{v}_i = \frac{\partial v_i}{\partial x_1} \tag{6}$$

Table 2: Forward mode AD example, with $y = f(x_1, x_2) = \ln(x_1) + x_1 x_2 - \sin(x_2)$ evaluated at $(x_1, x_2) = (2, 5)$ and setting $\dot{x}_1 = 1$ to compute $\frac{\partial y}{\partial x_1}$. The original forward evaluation of the primals on the left is augmented by the tangent operations on the right, where each line complements the original directly to its left.

Forward Primal Trace			Forward Tangent (Derivative) Trace		
v_{-1}	$= x_1$	$= 2$	$\dot{v}_{-1}$	$= \dot{x}_1$	$= 1$
v_0	$= x_2$	$= 5$	$\dot{v}_0$	$= \dot{x}_2$	$= 0$
v_1	$= \ln v_{-1}$	$= \ln 2$	$\dot{v}_1$	$= \dot{v}_{-1}/v_{-1}$	$= 1/2$
v_2	$= v_{-1} \times v_0$	$= 2 \times 5$	$\dot{v}_2$	$= \dot{v}_{-1} \times v_0 + \dot{v}_0 \times v_{-1}$	$= 1 \times 5 + 0 \times 2$
v_3	$= \sin v_0$	$= \sin 5$	$\dot{v}_3$	$= \dot{v}_0 \times \cos v_0$	$= 0 \times \cos 5$
v_4	$= v_1 + v_2$	$= 0.693 + 10$	$\dot{v}_4$	$= \dot{v}_1 + \dot{v}_2$	$= 0.5 + 5$
v_5	$= v_4 - v_3$	$= 10.693 + 0.959$	$\dot{v}_5$	$= \dot{v}_4 - \dot{v}_3$	$= 5.5 - 0$
y	$= v_5$	$= 11.652$	$\dot{y}$	$= \dot{v}_5$	$= \mathbf{5.5}$

给定一个数学表达式$f(x)$，它可以用一系列算子（加减乘除三角函数指数对数等）的组合表示。前向计算中每一步都对应一步函数计算和一步导数计算（即执行f和计算梯度同时进行），导数计算的依据则来自于函数。通过合适的数据表示方法，我们只需编写这些基础算子的前向计算和求导过程即可。

这一思路推广到多维数据和多维函数$f : \mathbb{R}^n \to \mathbb{R}^m$，其中$n$是输入变量的维度，$m$是输出变量的维度。求解其雅可比矩阵的每个元素时，可以在每个前向AD中设置为$\dot{\mathbf{x}} = \mathbf{e}_i$（即只有第$i$个元素为1，其他为0的单位向量）作为输入进行计算：

$$\dot{y}_j = \left. \frac{\partial y_j}{\partial x_i} \right|_{\mathbf{x}=\mathbf{a}}, \quad j = 1, \cdots, m \tag{7}$$

那么整个雅可比矩阵为：

$$\mathbf{J}_f = \left. \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \right|_{\mathbf{x}=\mathbf{a}} \tag{8}$$

或者可以初始化$\dot{\mathbf{x}} = \mathbf{r}$，用矩阵形式来计算：

$$\mathbf{J}_f\mathbf{r} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} \tag{9}$$

这种前向表示对$f:\mathbb{R}\to\mathbb{R}^m$类型的函数比较高效和直接，只需要执行$f$一次即可；但对于另一种极端形式$f:\mathbb{R}^n\to\mathbb{R}$，则需要执行$n$次$f$函数的流程。对于一个$f:\mathbb{R}\to\mathbb{R}^m$的映射，求解其导数需要$n\ c\ \mathrm{ops}(f)$的运算时间（其中$c<6$，一般取$c\sim[2,3]$）。我们知道实际使用中，输入的维度$n$往往远大于输出的维度$m$（即$n\gg m$），所以这使得AD的前向模式并不那么好用；而AD反向模式则能使运算时间降为$m\ c\ \mathrm{ops}(f)$。

前向模式的实现：二元数求导法 Dual Number

AD的前向模式可以使用二元数求导法来方便的实现。

二元数是实数的一种推广。二元数引入了一个“二元数单位”ε，满足$\varepsilon\neq 0$且$\varepsilon^2=0$。每个二元数都具有$z=a+b\varepsilon$的形式（其中a和b是实数）。这种表达形式可以看成是对一般实数的一阶展开（$\varepsilon\neq 0$），更高阶的数据则被消除掉了（$\varepsilon^2=0$）。根据泰勒展开，函数$f(x)$可表达为：

$$f(x)=f(x_0)+f'(x_0)(x-x_0)+\cdots+\frac{f^{(n)}(x_0)}{n!}(x-x_0)^n+R_n(x) \tag{10}$$

所以如果忽略二阶项及更高阶项（$n\geq 2$），$f(x)$在$x=x_0+\varepsilon$处满足：

$$f(x)=f(x_0)+f'(x_0)\varepsilon \tag{11}$$

二元数系数b即可看成是某函数$f(x)$在$x=a$处的导数。我们要做的就是为每个实数绑定一个二元数系数，并根据常用函数的求导法则更新该系数，就能获得任意复合函数在$x=a$处的导数了。

假定两个二元数分别为$x=a+b\varepsilon$和$y=c+d\varepsilon$，二元数的运算法则如下：

- 加法：

$$\begin{aligned} x+y &= (a+b\varepsilon)+(c+d\varepsilon) \\ &= (a+c)+(b+d)\varepsilon \end{aligned} \tag{12}$$

- 减法：

$$\begin{aligned} x-y &= (a+b\varepsilon)-(c+d\varepsilon) \\ &= (a-c)+(b-d)\varepsilon \end{aligned} \tag{13}$$

- 乘法：

$$\begin{aligned} x\times y &= (a+b\varepsilon)\times(c+d\varepsilon) \\ &= (ac+cd)+(bc+ad)\varepsilon+bd\varepsilon^2 \\ &= (ac+cd)+(bc+ad)\varepsilon \end{aligned} \tag{14}$$

- 除法：

$$\begin{aligned} \frac{x}{y} &= \frac{a+b\varepsilon}{c+d\varepsilon} \\ &= \frac{(a+b\varepsilon)(c-d\varepsilon)}{(c+d\varepsilon)(c-d\varepsilon)} \\ &= \frac{ac+(bc-ad)\varepsilon}{c^2} \\ &= \frac{a}{c}+\frac{bc-ad}{c^2}\varepsilon \end{aligned} \tag{15}$$

- 幂：

$$\begin{aligned} x^y &= (a+b\varepsilon)^{c+d\varepsilon} \\ &= a^c+\varepsilon\left(b(ca^{c-1})+d(a^c\ln a)\right) \end{aligned} \tag{16}$$

特别的，当指数为实数时：

$$\begin{aligned} x^y &= (a+b\varepsilon)^c \\ &= a^c+(ca^{c-1})b\varepsilon \end{aligned} \tag{17}$$

当底数为实数时：

$$
\begin{aligned}
x^y &= a^{c+d\varepsilon} \\
&= a^c + d(a^c \ln a)\varepsilon
\end{aligned} \tag{18}
$$

- 三角函数：

$$\sin(a+b\varepsilon) = \sin(a) + \cos(a)b\varepsilon \tag{19}$$
$$\cos(a+b\varepsilon) = \cos(a) - \sin(a)b\varepsilon \tag{20}$$
$$\tan(a+b\varepsilon) = \tan(a) + \frac{1}{\cos(a)^2}b\varepsilon \tag{21}$$
$$\arctan(a+b\varepsilon) = \arctan(a) + \frac{1}{1+a^2}b\varepsilon \tag{22}$$

- 对数函数：

$$\log_s(a+b\varepsilon) = log_s(a) + \frac{1}{\ln(s)a}b\varepsilon \tag{23}$$

一般的，令一个实数a对应的一个二元数为$a+\varepsilon$，则复合函数$F = f_1(f_2(f_3(\dots f_n(x)\dots)))$在$x=a$处的导数为：

$$F'|_{x=a} = \text{Dual}\left(F\left(a+\varepsilon\right)\right) \tag{24}$$

因此，我们只需要编写一些针对二元数的基础运算法则和函数即可。需要注意的是，我们并不需要实际给ε进行赋值，只要记住它与虚数单位类似，是一个独立的单位即可。这里用python给个简单的实现：

```python
import numpy as np
import math


class DualNumber:
    def __init__(self, x, y):
        self.real = x
        self.dual = y

    def __str__(self):
        rpr = '{}+{}e'.format(self.real, self.dual)
        return rpr

    def __repr__(self):
        return self.__str__()

    def __add__(self, other):
        if isinstance(other, DualNumber):
            real = self.real + other.real
            dual = self.dual + other.dual
        elif np.isscalar(other):
            real = self.real + other
            dual = self.dual
        else:
            raise TypeError('The other operator should be a scalar or a {}'.format(self.__class__.__name__))
        return DualNumber(real, dual)

    def __radd__(self, other):
        return self.__add__(other)

    def __sub__(self, other):
        if isinstance(other, DualNumber):
            real = self.real - other.real
            dual = self.dual - other.dual
        elif np.isscalar(other):
            real = self.real - other
            dual = self.dual
        else:
```

```
            raise TypeError('The other operator should be a scalar or a {}'.format(self.__class__.__name__))
        return DualNumber(real, dual)

    def __rsub__(self, other):
        if isinstance(other, DualNumber):
            real = other.real - self.real
            dual = other.dual - self.dual
        elif np.isscalar(other):
            real = other.real - self.real
            dual = - self.dual
        else:
            raise TypeError('The other operator should be a scalar or a {}'.format(self.__class__.__name__))
        return DualNumber(real, dual)

    def __mul__(self, other):
        if isinstance(other, DualNumber):
            real = self.real * other.real
            dual = self.dual * other.real + self.real * other.dual
        elif np.isscalar(other):
            real = self.real * other
            dual = self.dual * other
        else:
            raise TypeError('The other operator should be a scalar or a {}'.format(self.__class__.__name__))
        return DualNumber(real, dual)

    def __rmul__(self, other):
        return self.__mul__(other)

    def __truediv__(self, other):
        if isinstance(other, DualNumber):
            if other.real == 0:
                raise ValueError
            real = self.real / other.real
            dual = (self.dual - self.real / other.real * other.dual) / other.real
        elif np.isscalar(other):
            if other == 0:
                raise ValueError
            real = self.real / other
            dual = self.dual / other
        else:
            raise TypeError('The other operator should be a scalar or a {}'.format(self.__class__.__name__))
        return DualNumber(real, dual)

    def __pow__(self, power, modulo=None):
        real = math.pow(self.real, power)
        dual = self.dual * power * math.pow(self.real, power-1)
        return DualNumber(real, dual)

    def __abs__(self):
        real = abs(self.real)
        dual = np.sign(self.real)
        return DualNumber(real, dual)

    @staticmethod
    def sin(a):
        real = math.sin(a.real)
```

```
        dual = a.dual * math.cos(a.real)
        return DualNumber(real, dual)

    @staticmethod
    def cos(a):
        real = math.cos(a.real)
        dual = - a.dual * math.sin(a.real)
        return DualNumber(real, dual)

    @staticmethod
    def tan(a):
        real = math.tan(a.real)
        x = math.cos(a.real)
        dual = a.dual / (x * x)
        return DualNumber(real, dual)

    @staticmethod
    def atan(a):
        real = math.atan(a.real)
        x = a.real
        dual = a.dual / (1. + x*x)
        return DualNumber(real, dual)

    @staticmethod
    def sqrt(a):
        real = math.sqrt(a.real)
        dual = .5 * a.dual / real
        return DualNumber(real, dual)

    @staticmethod
    def exp(a):
        real = math.exp(a.real)
        dual = a.dual * math.exp(a.real)
        return DualNumber(real, dual)

    @staticmethod
    def log(a, base=math.e):
        real = math.log(a.real, base)
        dual = 1. / a.real / math.log(base) * a.dual
        return DualNumber(real, dual)
```

反向模式

反向传播BP可以看成AD反向模式的一种特例。不同于前向模式，反向模式需要计算输出对于每个中间变量v_i的梯度伴随量：

$$\bar{v}_i = \frac{\partial y_j}{\partial v_i} \tag{25}$$

这一导数表征着输出变量y_j对于中间变量v_i的敏感程度。在BP算法中，y就是最后的损失函数值了。

在反向模式中，导数是通过一个两阶段的过程计算出来的。在第一个阶段中，我们执行函数f的计算，获得所有的中间变量v_i，并且在计算图中记录变量之间的依赖性和相关性；在第二阶段中，输出对输入的导数是通过反方向从输出到输入传播梯度伴随量得到的：

Table 3: Reverse mode AD example, with $y = f(x_1, x_2) = \ln(x_1) + x_1 x_2 - \sin(x_2)$ evaluated at $(x_1, x_2) = (2, 5)$. After the forward evaluation of the primals on the left, the adjoint operations on the right are evaluated in reverse (cf. Figure 1). Note that both $\frac{\partial y}{\partial x_1}$ and $\frac{\partial y}{\partial x_2}$ are computed in the same reverse pass, starting from the adjoint $\bar{v}_5 = \bar{y} = \frac{\partial y}{\partial y} = 1$.

Forward Primal Trace			Reverse Adjoint (Derivative) Trace		
v_{-1}	$= x_1$	$= 2$	$\bar{x}_1 = \bar{v}_{-1}$		$= \mathbf{5.5}$
v_0	$= x_2$	$= 5$	$\bar{x}_2 = \bar{v}_0$		$= \mathbf{1.716}$
v_1	$= \ln v_{-1}$	$= \ln 2$	$\bar{v}_{-1} = \bar{v}_{-1} + \bar{v}_1 \frac{\partial v_1}{\partial v_{-1}}$	$= \bar{v}_{-1} + \bar{v}_1 / v_{-1}$	$= 5.5$
v_2	$= v_{-1} \times v_0$	$= 2 \times 5$	$\bar{v}_0 = \bar{v}_0 + \bar{v}_2 \frac{\partial v_2}{\partial v_0}$	$= \bar{v}_0 + \bar{v}_2 \times v_{-1}$	$= 1.716$
			$\bar{v}_{-1} = \bar{v}_2 \frac{\partial v_2}{\partial v_{-1}}$	$= \bar{v}_2 \times v_0$	$= 5$
v_3	$= \sin v_0$	$= \sin 5$	$\bar{v}_0 = \bar{v}_3 \frac{\partial v_3}{\partial v_0}$	$= \bar{v}_3 \times \cos v_0$	$= -0.284$
v_4	$= v_1 + v_2$	$= 0.693 + 10$	$\bar{v}_2 = \bar{v}_4 \frac{\partial v_4}{\partial v_2}$	$= \bar{v}_4 \times 1$	$= 1$
			$\bar{v}_1 = \bar{v}_4 \frac{\partial v_4}{\partial v_1}$	$= \bar{v}_4 \times 1$	$= 1$
v_5	$= v_4 - v_3$	$= 10.693 + 0.959$	$\bar{v}_3 = \bar{v}_5 \frac{\partial v_5}{\partial v_3}$	$= \bar{v}_5 \times (-1)$	$= -1$
			$\bar{v}_4 = \bar{v}_5 \frac{\partial v_5}{\partial v_4}$	$= \bar{v}_5 \times 1$	$= 1$
y	$= v_5$	$= 11.652$	$\bar{v}_5 = \bar{y}$	$= 1$	

同样以函数$f(x_1, x_2) = \ln(x_1) + x_1 x_2 - \sin(x_2)$为例。前馈过程与AD前向模式中的情况一样（左列），但是求导则与之前的顺序相反，是从输出变量开始的。由于定义了$y = v_5$，所以$\bar{v}_5 = \frac{\partial y}{\partial v_5} = 1$；而$v_5$是由$v_3$和$v_4$两个变量计算得到的，并且：

$$\frac{\partial y}{\partial v_3} = \frac{\partial y}{\partial v_5}\frac{\partial v_5}{\partial v_3} = \bar{v}_5 \frac{\partial v_5}{\partial v_3} \tag{26}$$

$$\frac{\partial y}{\partial v_4} = \frac{\partial y}{\partial v_5}\frac{\partial v_5}{\partial v_4} = \bar{v}_5 \frac{\partial v_5}{\partial v_4} \tag{27}$$

所以通过$\bar{v}_5$和$\frac{\partial v_5}{\partial v_3}$可以计算得到伴随量$\bar{v}_3$。$\bar{v}_4$类似。不难看出，这一过程就是本质上就是机器学习中的反向传播方法。只是此处输出y是变量（标量或矢量、矩阵），而不仅仅可以是机器学习中的损失函数值（标量）。另外值得一提的是，计算完$\bar{v}_3$和$\bar{v}_4$后，$\bar{v}_5$也就完成了任务，离开了其作用域（红线之间的几行为对应变量的作用域），可以在内存中释放掉。这可能也是PyTorch的`loss.backward()`实现中，一个结点完成反传后计算图被释放掉的原因。

对于输出到多个结点的中间变量，如v_0与v_2/v_3都相关，其梯度为：

$$\begin{aligned}\frac{\partial y}{\partial v_0} &= \frac{\partial y}{\partial v_2}\frac{\partial v_2}{\partial v_0} + \frac{\partial y}{\partial v_1}\frac{\partial v_1}{\partial v_0} \\ &= \bar{v}_2 \frac{\partial v_2}{\partial v_0} + \bar{v}_1 \frac{\partial v_1}{\partial v_0}\end{aligned} \tag{28}$$

具体实现时，一般使用多步增量模式：

$$\bar{v}_0 = 0 \tag{29}$$

$$\bar{v}_0 = \bar{v}_0 + \bar{v}_2 \frac{\partial v_2}{\partial v_0} \tag{30}$$

$$\bar{v}_0 = \bar{v}_0 + \bar{v}_1 \frac{\partial v_1}{\partial v_0} \tag{31}$$

上文中我们提到前向模式中，如果$f : \mathbb{R}^n \to \mathbb{R}$，那么计算所有针对输入变量的导数需要执行$n$次$f$函数流程；而在反向模式中，$f$函数流程执行次数则变为了$m$，即输出变量的维度。一次流程即可算出某个输出变量针对所有输入变量的导数：

$$\nabla y_i = (\frac{\partial y_i}{\partial x_1}, \cdots, \frac{\partial y_i}{\partial x_n}) \tag{32}$$

在$n \gg m$的情况下， AD的反向模式能够有效降低执行计算量。反向模式也可以用矩阵向量化表达为：

$$\mathbf{J}_f^T \mathbf{r} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{y_m}{x_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{y_m}{x_n} \end{bmatrix} \begin{bmatrix} r_1 \\ \vdots \\ r_m \end{bmatrix} \tag{33}$$

其中初始化$\bar{\mathbf{y}} = \mathbf{r}$。

AD反向模式也有自身的缺陷，就是在最坏情况下导致计算所需内存空间增加（与反馈过程中的操作数量成比例）。如何优化和高效利用内存是一个比较热门的研究方向。

Table 4: Evaluation times of the Helmholtz free energy function and its gradient (Figure 5). Times are given relative to that of the original function with both (1) $n = 1$ and (2) n corresponding to each column. (For instance, reverse mode AD with $n = 43$ takes approximately twice the time to evaluate relative to the original function with $n = 43$.) Times are measured by averaging a thousand runs on a machine with Intel Core i7-4785T 2.20 GHz CPU and 16 GB RAM, using DiffSharp 0.5.7. The evaluation time for the original function with $n = 1$ is 0.0023 ms.

	n, number of variables							
	1	8	15	22	29	36	43	50
f, original								
Relative $n = 1$	1	5.12	14.51	29.11	52.58	84.00	127.33	174.44
∇f, numerical diff.								
Relative $n = 1$	1.08	35.55	176.79	499.43	1045.29	1986.70	3269.36	4995.96
Relative n in column	1.08	6.93	12.17	17.15	19.87	23.64	25.67	28.63
∇f, forward AD								
Relative $n = 1$	1.34	13.69	51.54	132.33	251.32	469.84	815.55	1342.07
Relative n in column	1.34	2.66	3.55	4.54	4.77	5.59	6.40	7.69
∇f, reverse AD								
Relative $n = 1$	1.52	11.12	31.37	67.27	113.99	174.62	254.15	342.33
Relative n in column	1.52	2.16	2.16	2.31	2.16	2.07	1.99	1.96

自动微分实现类型

Table 5: Survey of AD implementations. Tools developed primarily for machine learning are highlighted in bold.

Language	Tool	Type	Mode	Institution / Project	Reference	URL
AMPL	AMPL	INT	F, R	Bell Laboratories	Fourer et al. (2002)	http://www.ampl.com/
C, C++	ADIC	ST	F, R	Argonne National Laboratory	Bischof et al. (1997)	http://www.mcs.anl.gov/research/projects/adic/
	ADOL-C	OO	F, R	Computational Infrastructure for Operations Research	Walther and Griewank (2012)	https://projects.coin-or.org/ADOL-C
C++	Ceres Solver	LIB	F	Google		http://ceres-solver.org/
	CppAD	OO	F, R	Computational Infrastructure for Operations Research	Bell and Burke (2008)	http://www.coin-or.org/CppAD/
	FADBAD++	OO	F, R	Technical University of Denmark	Bendtsen and Stauning (1996)	http://www.fadbad.com/fadbad.html
	Mxyzptlk	OO	F	Fermi National Accelerator Laboratory	Ostiguy and Michelotti (2007)	
C#	AutoDiff	LIB	R	George Mason Univ., Dept. of Computer Science	Shtof et al. (2013)	http://autodiff.codeplex.com/
F#, C#	**DiffSharp**	OO	F, R	Maynooth University, Microsoft Research Cambridge	Baydin et al. (2016a)	http://diffsharp.github.io
Fortran	ADIFOR	ST	F, R	Argonne National Laboratory	Bischof et al. (1996)	http://www.mcs.anl.gov/research/projects/adifor/
	NAGWare	COM	F, R	Numerical Algorithms Group	Naumann and Riehme (2005)	http://www.nag.co.uk/nagware/Research/ad_overview.asp
	TAMC	ST	R	Max Planck Institute for Meteorology	Giering and Kaminski (1998)	http://autodiff.com/tamc/
Fortran, C	COSY	INT	F	Michigan State Univ., Biomedical and Physical Sci.	Berz et al. (1996)	http://www.bt.pa.msu.edu/index_cosy.htm
	Tapenade	ST	F, R	INRIA Sophia-Antipolis	Hascoët and Pascual (2013)	http://www-sop.inria.fr/tropics/tapenade.html
Haskell	ad	OO	F, R	Haskell package		http://hackage.haskell.org/package/ad
Java	ADiJaC	ST	F, R	University Politehnica of Bucharest	Slusanschi and Dumitrel (2016)	http://adijac.cs.pub.ro
	Deriva	LIB	R	Java & Clojure library		https://github.com/lambder/Deriva
Julia	JuliaDiff	OO	F, R	Julia packages	Revels et al. (2016a)	http://www.juliadiff.org/
Lua	**torch-autograd**	OO	R	Twitter Cortex		https://github.com/twitter/torch-autograd
MATLAB	ADiMat	ST	F, R	Technical University of Darmstadt, Scientific Comp.	Willkomm and Vehreschild (2013)	http://adimat.sc.informatik.tu-darmstadt.de/
	INTLab	OO	F	Hamburg Univ. of Technology, Inst. for Reliable Comp.	Rump (1999)	http://www.ti3.tu-harburg.de/rump/intlab/
	TOMLAB/MAD	OO	F	Cranfield University & Tomlab Optimization Inc.	Forth (2006)	http://tomlab.biz/products/mad
Python	ad	OO	R	Python package		https://pypi.python.org/pypi/ad
	autograd	OO	F, R	Harvard Intelligent Probabilistic Systems Group	Maclaurin (2016)	https://github.com/HIPS/autograd
	Chainer	OO	R	Preferred Networks	Tokui et al. (2015)	https://chainer.org/
	PyTorch	OO	R	PyTorch core team	Paszke et al. (2017)	http://pytorch.org/
	Tangent	ST	F, R	Google Brain	van Merriënboer et al. (2017)	https://github.com/google/tangent
Scheme	R6RS-AD	OO	F, R	Purdue Univ., School of Electrical and Computer Eng.		https://github.com/qobi/R6RS-AD
	Scmutils	OO	F	MIT Computer Science and Artificial Intelligence Lab.	Sussman and Wisdom (2001)	http://groups.csail.mit.edu/mac/users/gjs/6946/refman.txt
	Stalingrad	COM	F, R	Purdue Univ., School of Electrical and Computer Eng.	Pearlmutter and Siskind (2008)	http://www.bcl.hamilton.ie/~qobi/stalingrad/

目前大部分AD方法大致可分为以下几种：

- 基础算子型（elemental）：这类实现方法主要通过将任意函数分解为有限个基础AD算子，并用基础AD算子替代基础数学算子来实现自动微分。在没有运算重载符的语言环境中，这种方法比较适用；

- 编译和源码转换型（compilers and source code transformation）：用另一种语法或扩展语言编写运算，然后再转换到原始编程语言上，比如用数学标记来表达目标函数和约束，再用解释器或编译器分析为编程语言；
- 运算符重载型（operator overloading）：现代编程语言支持运算符重载，这使得基础算子型的AD方法更加容易实现。

名称	编程语言	实现方法	支持模式	地址
torch-autograd	Lua	运算符重载	反向	https://github.com/twitter/torch-autograd
autograd	Python	运算符重载	前向/反向	https://github.com/HIPS/autograd
Chainer	Python	运算符重载	反向	https://chainer.org/
PyTorch	Python	运算符重载	反向	https://pytorch.org/
Tangent	Python	源码转换	前向/反向	https://github.com/google/tangent

源码转换型AD方法原理

参考资料：

- [python] ast模块
- ast --- 抽象语法树

Tangent库通过对Python抽象语法树的修改，为部分系统数学运算以及numpy部分基础运算添加了自定义的求导函数并自动生成代码。具体代码尚未完全理解，这里自己简单记录下原理，并附上一些简单的代码辅助说明。

我们先得理解Python代码的执行过程：

> 语法分析 $\Longrightarrow$ 具体语法树 $\Longrightarrow$ 抽象语法树 $\Longrightarrow$ 控制流图 $\Longrightarrow$ 字节码 $\Longrightarrow$ 执行

Tangent库就用到了gast库（以ast库作为基础）对抽象语法树进行读取和补充。所以其中的关键就是如何利用ast和抽象语法树。

看一个简单的例子。先在代码中嵌入expr这一段Python代码，其中包括一个add函数，用来计算两个输入的和，然后执行并print：

```
>>> import ast
>>> expr = """
... def add(x,y):
...     return x + y
... print(add(3,4))
... """
>>> expr_ast = ast.parse(expr)
>>> expr_ast
<_ast.Module object at 0x7f61f2a58ac8>
```

expr经过ast模块解析后，得到的抽象语法树如下：

```
>>> ast.dump(expr_ast)
Module(
    body=[
        FunctionDef(
            name='add',
            args=arguments(
                args=[
                    arg(arg='x', annotation=None),
```

```
                    arg(arg='y', annotation=None)
                ],
                vararg=None,
                kwonlyargs=[],
                kw_defaults=[],
                kwarg=None,
                defaults=[]
            ),
            body=[
                Return(
                    value=BinOp(
                        left=Name(id='x', ctx=Load()),
                        op=Add(),
                        right=Name(id='y', ctx=Load()))
                )
            ],
            decorator_list=[],
            returns=None
        ),
        Expr(
            value=Call(
                func=Name(id='print', ctx=Load()),
                args=[
                    Call(
                        func=Name(id='add', ctx=Load()),
                        args=[Num(n=3), Num(n=4)],
                        keywords=[]
                    )
                ],
                keywords=[]
            )
        )
    ]
)
```

可以看到，expr中自定义的函数在抽象语法树中位于FunctionDef这个field中，而其中具体算子（+）位于FunctionDef.body.Return.value中，名为BinOp，具体操作为Add()。接下来我们通过ast的转换模块，将BinOp这个域中的Add()函数修改为乘法（ast.Mult()）。

定义一个转换类，将ast中的结点进行修改。由于目标结点是BinOp，所以在其中定义一个visit_BinOp函数，并将其中op域替换为ast.Mult()：

```
>>> class Transformer(ast.NodeTransformer):
...     def visit_BinOp(self, node):
...         node.op = ast.Mult()
...         return node
...
>>> trans = Transformer()
```

执行一下原始expr中的代码，$3+4=7$，+号执行的是加法：

```
>>> exec(compile(expr_ast, '<string>', 'exec'))
7
```

接下来我们替换掉其中的加法：

```
>>> modified = trans.visit(expr_ast)  # visit会调用所有visit_<classname>的方法
>>> ast.dump(modified)
```

```
Module(
    body=[
        FunctionDef(
            name='add',
            args=arguments(
                args=[
                    arg(arg='x', annotation=None),
                    arg(arg='y', annotation=None)
                ],
                vararg=None,
                kwonlyargs=[],
                kw_defaults=[],
                kwarg=None,
                defaults=[]
            ),
            body=[
                Return(
                    value=BinOp(
                        left=Name(id='x', ctx=Load()),
                        op=Mult(),
                        right=Name(id='y', ctx=Load())
                    )
                )
            ],
            decorator_list=[],
            returns=None
        ),
        Expr(
            value=Call(
                func=Name(id='print', ctx=Load()),
                args=[
                    Call(
                        func=Name(id='add', ctx=Load()),
                        args=[Num(n=3), Num(n=4)],
                        keywords=[])],
                keywords=[]
            )
        )
    ]
)
```

可以看到，在第22行，原来BinOp域里的op已经被替换为了x乘法。执行一下新的抽象语法树：

```
>>> exec(compile(modified, '<string>', 'exec'))
12
```

结果变成了$3 \times 4 = 12$。

这个例子说明，我们能够通过ast模块注入并修改源代码。此处再给出一个例子，调用numpy.add（也就是numpy.ndarray的加法）然后通过ast注入修改为了减法。由于numpy.add并非系统函数，所以抽象语法树有些不同：

```
import ast
expr = """
import numpy as np
def add(x, y):
    out = np.add(x, y)
    return out
a = np.zeros((1,3))
b = np.ones((1,3))
print(add(a, b))
"""

expr_ast = ast.parse(expr)
print(ast.dump(expr_ast))
```

获得上述代码的抽象语法树为：

```
Module(
    body=[
        Import(
            names=[alias(name='numpy', asname='np')]
        ),
        FunctionDef(
            name='add',
            args=arguments(
                args=[arg(arg='x', annotation=None),
                      arg(arg='y', annotation=None)],
                vararg=None,
                kwonlyargs=[],
                kw_defaults=[],
                kwarg=None,
                defaults=[]
            ),
            body=[
                Assign(
                    targets=[Name(id='out', ctx=Store())],
                    value=Call(
                        func=Attribute(
                            value=Name(id='np', ctx=Load()),
                            attr='add',
                            ctx=Load()
                        ),
                        args=[Name(id='x', ctx=Load()),
                              Name(id='y', ctx=Load())],
                        keywords=[]
                    )
                ),
                Return(
                    value=Name(id='out', ctx=Load())
                )
            ],
            decorator_list=[],
            returns=None
        ),
        Assign(
            targets=[Name(id='a', ctx=Store())],
            value=Call(
                func=Attribute(
                    value=Name(id='np', ctx=Load()),
                    attr='ones',
                    ctx=Load()
```

```
                ),
                args=[Tuple(elts=[Num(n=1), Num(n=3)], ctx=Load())],
                keywords=[]
            )
        ),
        Assign(
            targets=[Name(id='b', ctx=Store())],
            value=Call(
                func=Attribute(
                    value=Name(id='np', ctx=Load()),
                    attr='ones',
                    ctx=Load()
                ),
                args=[Tuple(elts=[Num(n=1), Num(n=3)], ctx=Load())],
                keywords=[])
        ),
        Expr(
            value=Call(
                func=Name(id='print', ctx=Load()),
                args=[
                    Call(
                        func=Name(id='add', ctx=Load()),
                        args=[Name(id='a', ctx=Load()), Name(id='b', ctx=Load())],
                        keywords=[]
                    )
                ],
                keywords=[]
            )
        )
    ]
)
```

注入目标numpy.add位于第23行Attribute.attr内，所以修改的点就在这里：

```
class EvilTransformer(ast.NodeTransformer):
    def visit_Attribute(self, node):
        if node.attr == 'add':
            node.attr = 'subtract'  # numpy中减法为numpy.subtract
        return node

trans = EvilTransformer()
new_ast = trans.visit(expr_ast)
exec(compile(new_ast, '<string>', 'exec'))
```

理论上，注入前，expr执行的结果应该是[[1.,1.,1.]]；注入后，输出结果就会变成[[-1.,-1.,-1.]]，成功将numpy.add改成了numpy.subtract。

注意，虽然trans.visit返回了“新”的抽象语法树变量new_ast，实际上该函数是对expr_ast直接进行了修改。所以new_ast和expr_ast是同一个变量的两个别名。

ast的作用在于，假如我们用某些算子编写了一个网络（函数），我们都够借助ast模块获得这一网络结构的抽象语法树，其中每个独立的结点都是一个运算语句。参考语句所使用的算子形式，编写对应的自动求导规则也就成了可能。

Tangent自动微分库

最后，简单分析下Tangent库是如何借助ast库来完成源码转换型的自动微分的。Tangent库中一些关键的函数有：

- tangent.grads.create_register

```
def create_register(dict_):
  def register(key):
    def _(f):
      dict_[key] = f
      return f
    return _
  return register
```

这一函数用于生成一个作为装饰器的注册机。注册信息保存在dict_这一变量中。通过这类注册机，可以自定义某特定函数的微分函数。需要注意的是，**不需要该微分函数能运行，它只是作为模板用于自动生成真正微分函数的代码。**

- tangent.grad
  - 用于对某个函数进行求微分；
  - 适用于$f:\mathbb{R}^n \to \mathbb{R}$类型的函数；
  - 会检查输入是否为标量。
- tangent.autodiff
  - `autodiff(f, mode='forward')`：AD前向模式，调用的关键函数：ForwardAD
  - `autodiff(f, mode='reverse')`：AD反向模式，调用的关键函数：ReverseAD

对于一个自定义的数学函数，Tangent库将会分析函数代码，并根据注册机内的微分函数模板，进行解析并生成一个对应的微分函数的ast抽象语法树，然后通过astor包将该抽象语法树转化回python源码。

autograd自动微分库

autograd自动微分库基于基础运算的重载，主要重载的是numpy包和scipy包。

自动求导实现思路

我们可以基于基础算子方法和运算符重载方法，在Python下实现一个自动微分的库。首先我们需要自定义一种数据类（例如Tensorflow中的Variable和PyTorch中的tensor数据类），库内所有基础算子将对其进行运算，并支持反向传播过程。在此基础上，我们只需要实现对该数据类的基础算子定义和运算符重载即可。

基础数据类型

为了简单起见，我们先利用numpy的ndarray作为基础，将其封装一层即作为自定义数据类，并将其命名为Zhangliang（“张量”拼音）：

```
class Zhangliang(BaseZhangliang):
    def __init__(self, data, dtype=np.float64, requires_grad=False):
        if isinstance(data, Zhangliang):
            data = data.values
        elif np.isscalar(data):
            data = [data]

        self._zhi = np.array(data, dtype=dtype)
        self.requires_grad = requires_grad
        self._tidu = np.zeros_like(self._zhi)

```

```
    def assign_value(self, new_value):
        self._zhi = new_value

    def update_grad(self, grad_value):
        self._tidu += grad_value

    @property
    def grad(self):
        if not self.requires_grad:
            raise AttributeError('Tensor requires no gradient.')
        else:
            return self._tidu

    @property
    def values(self):
        return self._zhi

    @property
    def shape(self):
        return self._zhi.shape

    @property
    def ndim(self):
        return self._zhi.ndim

    @property
    def dtype(self):
        return self._zhi.dtype

    @property
    def size(self):
        return self._zhi.size

    def __iter__(self):
        return self._zhi.__iter__()

    def __len__(self):
        return len(self._zhi)

    def __getitem__(self, item):
        return self._zhi[item]

    def __repr__(self):
        return self._zhi.__repr__()

    def __str__(self):
        return self._zhi.__str__()

    @classmethod
    def zeros(cls, shape, dtype=np.float64, requires_grad=False):
        zeros_ = np.zeros(shape, dtype=dtype)
        return cls(zeros_, requires_grad=requires_grad)

    @classmethod
    def ones(cls, shape, dtype=np.float64, requires_grad=False):
        ones_ = np.ones(shape, dtype=dtype)
        return cls(ones_, requires_grad=requires_grad)

    @classmethod
    def zeros_like(cls, data, dtype=np.float64, requires_grad=False):
        shape = data.shape
```

```
        zeros_ = np.zeros(shape, dtype=dtype)
        return cls(zeros_, requires_grad=requires_grad)

    @classmethod
    def ones_like(cls, data, dtype=np.float64, requires_grad=False):
        shape = data.shapes
        ones_ = np.ones(shape, dtype=dtype)
        return cls(ones_, requires_grad=requires_grad)

    @classmethod
    def array(cls, data, requires_grad=False):
        if isinstance(data, Zhangliang):
            return cls(data.values, dtype=data.dtype,
requires_grad=requires_grad)
        elif np.isscalar(data):
            return cls([data], dtype=np.int32, requires_grad=requires_grad)
        elif isinstance(data, (list, tuple)):
            return cls(data, dtype=np.float64, requires_grad=requires_grad)
        elif isinstance(data, collections.Iterable):
            data = np.array(data)
            return cls(data, dtype=np.float64, requires_grad=requires_grad)
        else:
            raise TypeError

    @classmethod
    def linspace(cls, start, stop, num):
        data = np.linspace(start, stop, num)
        return cls(data, dtype=data.dtype, requires_grad=False)

    @classmethod
    def arange(cls, start, stop=None, step=1):
        if stop is None:
            stop = start
            start = 0
        data = np.arange(start, stop, step)
        return cls(data, dtype=data.dtype, requires_grad=False)
```

目前定义的Zhangliang类仅包括其值（Zhangliang.zhi，本质就是numpy.ndarray），以及一些基础函数，尚未包括运算符重载。回顾下PyTorch和Tensorflow中的数据类，它们都支持a+b形式的调用。这在Python中是遵循了协议接口，调用其数据类的__add__方法；另一方面，在PyTorch和Tensorflow我们还能看到tf.add(a,b)和torch.add(a,b)形式的调用，这说明两个框架也存在着独立的基础算子。综合这两点，实际上我们只需要实现独立的基础算子，然后在Zhangliang.__add__方法中调用add算子即可：

```
# 定义基础算子，并重载运算符
# 这里只展示基础四则运算

class Zhangliang(object):
     # 省略上述已有内容

     def __add__(self, other):
        return zl_add(self, other)

     def __radd__(self, other):
        return zl_add(other, self)

     def __sub__(self, other):
        return zl_sub(self, other)

     def __rsub__(self, other):
        return zl_sub(other, self)
```

```
18
19      def __truediv__(self, other):
20          return zl_truediv(self, other)
21
22
23  def zl_add(a, b):
24      if isinstance(a, numbers.Real):
25          value = a + b.zhi
26      elif isinstance(b, numbers.Real):
27          value = a.zhi + b
28      else:
29          value = a.zhi + b.zhi
30      return Zhangliang(value)
31
32
33  def zl_sub(a, b):
34      if isinstance(a, numbers.Real):
35          value = a - b.zhi
36      elif isinstance(b, numbers.Real):
37          value = a.zhi - b
38      else:
39          value = a.zhi - b.zhi
40      return Zhangliang(value)
41
42
43  def zl_mul(a, b):
44      if isinstance(a, numbers.Real):
45          value = a * b.zhi
46      elif isinstance(b, numbers.Real):
47          value = a.zhi * b
48      else:
49          value = a.zhi * b.zhi
50      return Zhangliang(value)
51
52
53  def zl_truediv(a, b):
54      if isinstance(a, numbers.Real):
55          value = a / b.zhi
56      elif isinstance(b, numbers.Real):
57          if b == 0:
58              raise ValueError('0 cannot be divisor.')
59          value = a.zhi / b
60      else:
61          value = a.zhi / b.zhi
62      return Zhangliang(value)
63
64  # 省略以下
```

定义这些运算比较简单，甚至求对应的微分过程也比较容易实现。但问题是：

- **如何注册前向和反向运算?**
- **我们在编写前馈运算时，怎么跟踪和记录运算过程?**
- **为了确定反传的执行顺序，怎么通过上述跟踪记录来获得计算图?**
- **如何进行反传?**

第一个问题：如何注册前向和反向函数?

第一个问题相对而言比较简单：我们定义两个字典：

```
forward_func = dict()
backward_func = dict()
```

然后前馈函数和反馈函数各自以{算子名：前馈/反馈函数}对的方式进行注册。我们通过Python装饰器来进行注册，为此，先定义两个注册机：

```
def create_register(dict_):
    def register(key):
        def _(fn):
            dict_[key] = fn
            return fn
        return _
    return register

forward_func = {}
forward_register = create_register(forward_func)
backward_func = {}
backward_register = create_register(backward_func)
```

那么只需要在基础算子定义前加上装饰器即可完成注册，比如以下代码可以将zl_add函数注册到前馈函数库中并注册为{'add': zl_add}：

```
@forward_register(key='add')
def zl_add(a, b):
    pass
```

第二个问题：如何跟踪和记录运算过程?

即使注册了函数，我们仍需要跟踪和记录前馈过程的运算过程。假设有一个跟踪机tracer，在每次调用基础算子时，都需要记录算子的Zhangliang输入和输出，以及算子自身的类型。这一需求同样可以通过装饰器来实现：

```
# create_tracer将定义一个计算图`graph_`，具体类型稍后介绍
# trave_with_name则是真正的装饰器函数，将使用`op_name`来追踪被装饰函数
# warp是对原函数的封装，传入为原函数引用
# eval_fn是真正执行原函数，并且在执行后将输入输出记录到计算图中

def create_tracer(graph_):
    def trace_with_name(op_name):
        def warp(fn):
            def eval_fn(*args, **kwargs):
                output = fn(*args, **kwargs)

                # 将输入args和输出output记录到计算图graph_中
                # 也要记录算子的配置kwargs

                return output
            return eval_fn
        return warp
    return trace_with_name

graph = {}
```

```
trace = create_tracer(graph)
```

将trace装饰器放在前馈函数前，就能在每次调用该函数时将输入和输出记录到计算图graph_中。

不过此时又出现了一个问题：前馈函数注册机forward_func和跟踪器trace两个装饰器有点不同，forward_func是在前馈函数定义时调用一次（只需要一次即可），而trace则是每次调用前馈函数时都要调用。连着使用两个装饰器会导致每次调用前馈函数时都注册一次；另外，两个有点麻烦，不如一个装饰器简便。基于上述理由，我们将trace修改为：

```
1  def create_tracer(graph_: Graph):
2      def trace_with_name(op_name):
3          @forward_func(op_name=op_name)
4          def warp(fn):
5              def eval_fn(*args, **kwargs):
6                  output = fn(*args, **kwargs)
7
8                  # 将输入args和输出output记录到计算图graph_中
9                  # 也要记录算子的配置kwargs
10
11                 return output
12             return eval_fn
13         return warp
14     return trace_with_name
15
16
17 graph = Graph()
18 trace = create_tracer(graph)
```

注意到第3行，forward_func装饰器现在装饰了原函数的封装函数warp。这样在添加trace装饰器的时候，就会调用一次注册机，原函数即在装饰trace时完成了注册。

因此，剩下的目标就是如何定义计算图，以便在调用前馈函数时将各张量和算子记录下来。

第三个问题：怎么通过上述跟踪记录来获得计算图?

计算图中需要包括三类元素：

- 运算过程中的常量；
- 运算过程中的张量、中间变量；
- 运算过程中的算子。

一个基本的观察是：每个基础算子可能有若干个输入，但是只会有一个输出。所以每个算子都会绑定一个输出的中间变量/张量。我们定义一个节点类：

```
1  # 节点类，每个节点对应一个算子，以及一个输出的中间变量
2  # input_list:        输入的`Zhangliang`
3  # input_list_id:     每个输入`Zhangliang`的id
4  # output:            输出的`Zhangliang`
5  # op_type:           节点算子类型，对应于注册机内的关键字
6  # input_kwargs:      算子配置
7  # op_id:             节点编号
8  class Node(object):
9      def __init__(self, input_args, input_kwargs, output, op_type):
10         self.input_list = tuple(input_args)
11         self.input_list_id = tuple([id(an_input) for an_input in self.input_list])
12         self.output = output
```

```
        self.op_type = op_type
        self.input_kwargs = input_kwargs
        self.op_id = -1

    def set_id(self, id_value):
        self.op_id = id_value

    @property
    def name(self):
        if self.op_id < 0:
            raise ValueError('Node not added to graph.')
        node_name = '{}_{}'.format(self.op_type, self.op_id)
        return node_name
```

节点类将记录每个被追踪的算子的输入张量，输入参数，算子类型以及输出张量。通过张量的先后关系，我们再定义一个计算图类用于记录每个节点，并根据输入输出关系计算反传时的拓扑顺序。计算图类定义如下：

```
"""
计算图类
成员变量：
  _op_count：每种不同类型的算子的计数
  _nodes_by_name：按照添加顺序排列的节点字典；会给每个节点一个名字；通过节点名进行索引
  _nodes_by_id：按照输出Zhangliang的id对节点进行索引；在反传时用于找出对应的节点和算子
  _topo：计算图的反向拓扑，每个节点映射到其父节点，通过节点名进行索引
成员函数：
  is_initialized：反传时用于判断是否已经进行拓扑排序
  is_leaf：判断某个Zhangliang所在节点是否为叶节点（即最终的输出）
  get_node_by_output_tensor：对外接口，获取Zhangliang对应的节点
  get_parents：对外接口，获取Zhangliang对应节点的父节点
  append_node：添加节点，会为每个节点一个名字
  toposort：按照拓扑顺序对节点进行排列
  clear_graph：清除计算图
"""

class Graph:
    def __init__(self):
        self._op_count = dict()
        self._nodes_by_name = OrderedDict()
        self._nodes_by_id = OrderedDict()
        self._topo = OrderedDict()

    def is_initialized(self):
        return len(self._topo) != 0

    def is_leaf(self, tensor):
        node = self.get_node_by_output_tensor(tensor)
        return list(self._topo.items())[0][0] == node.name

    def get_node_by_output_tensor(self, tensor):
        query_id = id(tensor)
        node = self._nodes_by_id[query_id]
        return node

    def get_parents(self, node):
        if not self.is_initialized():
            self.toposort()
        parent_name = self._topo[node.name]
        parent_nodes = [self._nodes_by_name[p] for p in parent_name]
        return parent_nodes
```

```
43
44      def append_node(self, node: Node):
45          node_type = node.op_type
46          count = self._op_count.setdefault(node_type, 0)
47          node.set_id(count)
48          self._op_count[node_type] += 1
49
50          # Index node by the op name
51          self._nodes_by_name[node.name] = node
52
53          # Index node by the output id
54          self._nodes_by_id[id(node.output)] = node
55
56      def toposort(self):
57          for k, node_ in reversed(self._nodes_by_name.items()):
58              parents = []
59              for j, node_b in reversed(self._nodes_by_name.items()):
60                  output = node_b.output
61                  if id(output) in node_.input_list_id:
62                      parents.append(j)
63                  if len(parents) == len(node_.input_list):
64                      break
65              self._topo[k] = parents
66
67      def clear_graph(self):
68          self._op_count.clear()
69          self._nodes_by_name.clear()
70          self._nodes_by_id.clear()
71          self._topo.clear()
```

然后在trace中调用计算图：

```
1  def create_tracer(graph_: Graph):
2      def trace_with_name(op_name):
3          @func_register(op_name=op_name)
4          def warp(fn):
5              def eval_fn(*args, **kwargs):
6                  output = fn(*args, **kwargs)
7                  new_node = Node(input_args=args, input_kwargs=kwargs,
   output=output, op_type=op_name)
8                  graph_.append_node(new_node)
9                  return output
10             return eval_fn
11         return warp
12     return trace_with_name
```

至此，我们完成了自动微分的部分必需内容。

到此为止，在调用注册的函数执行张量运算时，计算图会将所有算子都记录下来，有时候比较不方便，比如类似tensorflow和pytorch都有只执行前馈不加入计算图的功能函数（tensorflow的stop_gradient函数和pytorch的no_grad上下文函数）；另外，如果我们编写一些算子函数可能也会需要这一功能（比如编写卷积层时可能会用到张量基本运算，但是将卷积这一过程都分解记录为张量的基础算子不是非常合适）。所以我们需要额外的一个功能，就是编写类似于no_grad的上下文函数。首先，为Graph计算图类添加一个**当前是否记入节点**的标志：

```python
class Graph:
    def __init__(self):
        # 省略了这部分内容
        self._ctx_requires_grad = True

    def is_grad_enabled(self):
        return self._ctx_requires_grad

    def set_grad_enable(self, enabled=True):
        self._ctx_requires_grad = enabled

    # 省略了其他函数
```

然后编写上下文函数（参考pytorch的no_grad函数）：

```python
class no_grad(object):
    def __init__(self):
        self.prev_state = graph.is_grad_enabled()

    def __enter__(self):
        self.prev_state = graph.is_grad_enabled()
        graph.set_grad_enable(False)

    def __exit__(self, exc_type, exc_val, exc_tb):
        graph.set_grad_enable(self.prev_state)
        return False


class has_grad(object):
    def __init__(self):
        self.prev_state = graph.is_grad_enabled()

    def __enter__(self):
        self.prev_state = graph.is_grad_enabled()
        graph.set_grad_enable(True)

    def __exit__(self, exc_type, exc_val, exc_tb):
        graph.set_grad_enable(self.prev_state)
        return False
```

第四个问题：如何进行反传?

假定我们已经获得了所有算子节点的拓扑顺序，也编写了每个算子反传函数，那么如何完成计算图的反传？实际上，在深度学习语境下，所有计算图最终只有一个输出节点，即loss。这是唯一的一个叶节点，从这个叶节点开始，根据拓扑顺序，我们可以依次使用反向模式进行传播。所以这也是计算图Graph类会有一个is_leaf函数的原因，叶节点总是位于反向拓扑的第一个位置。

调用形式上，pytorch使用了tensor.backward()的形式，即只需调用最终输出节点的反传函数，便可对整个计算图中的节点进行反传。我们模仿这一调用形式，并在Zhangliang类中进行实现：

```python
class Zhangliang(BaseZhangliang):
    # 省略其他

    def release(self):
        self._tidu = None
```

```
    def backward(self, retain_graph=False):
        # 检查计算图是否已经完成拓扑排序
        if not graph.is_initialized():
            graph.toposort()

        # 检查当前节点是否为叶节点。
        # 如果是叶节点且支持梯度，该Zhangliang是不会有梯度输入的，所以更新其梯度值为1；
        # 如果是叶节点但不支持梯度，那么这个函数不应该被调用，报错；
        # 如果不是叶节点且不支持梯度，说明到了某个输出点或分离点，直接返回不报错
        if graph.is_leaf(self) and self.requires_grad:
            self.update_grad(1.)
        elif graph.is_leaf(self) and (not self.requires_grad):
            raise AttributeError('Zhangliang does not requires grad.')
        elif (not graph.is_leaf(self)) and (not self.requires_grad):
            return

        # 通过Zhangliang的id获得对应的节点
        node = graph.get_node_by_output_tensor(self)
        # 调用节点对应的反传函数，将本输出Zhangliang的梯度反传到输入Zhangliang中
        node.backprop()

        # 默认不保持计算图，完成反传后释放本Zhangliang梯度数据所占内存
        if not retain_graph:
            self.release()

        # 获得节点的父节点
        parents = graph.get_parents(node)
        # 尾递归调用输入Zhangliang的backward方法继续进行反传
        for node_in in parents:
            o = node_in.output
            o.backward(retain_graph)

        # 再次判断是否为叶节点。是的话，清除当前计算图，准备下次前馈
        if graph.is_leaf(self):
            graph.clear_graph()
```

反传函数可选传入参数retain_graph=False，同样模仿pytorch中tensor.backward接口，默认为False，即在某节点完成反传后将其梯度数据所占内存释放。注释解释了backward中的每一步的用途。这里用到了Node类的反传，其实就是根据算子类别自动调用对应的反传函数：

```
class Node(object):
    # 省略其他

    def backprop(self):
        grad_fn = grad_lib[self.op_type]
        grad_fn(self.output, *self.input_list, **self.input_kwargs)
```

至此我们完成了梯度反传的过程。为了进行测试，我们定义函数为上文的$f(x_1,x_2)=\log(x_1)+x_1x_2-\sin(x_2)$，输入值为$(x_1,x_2)=(2,5)$，然后调用输出Zhangliang的backward函数，完成后查看x_1和x_2的梯度值：

```
"""
测试用例1：x1和x2均为Zhangliang
"""
x1 = Zhangliang(2, requires_grad=True)
x2 = Zhangliang(5, requires_grad=True)

f = log(x1) + x1*x2 - sin(x2)
```

```
f.backward()
print("Test function f=log(x1)+x1*x2-sin(x2), with initial values x1=2, x2=5.\n"
      "\tOracle grad: g_x1 = {:.5f}, g_x2 = {:.5f}\n"
      "\tResult grad: g_x1 = {:.5f}, g_x2 = {:.5f}".
      format(5.5, 1.716, x1.grad[0], x2.grad[0]))

"""
测试用例2：x1为张量，x2为常数
"""
x1 = Zhangliang(2, requires_grad=True)
x2 = 5

f = log(x1) + x1 * x2 - sin(x2)
f.backward()
print("Test function f=log(x1)+x1*x2-sin(x2), with initial values x1=2, x2=5.\n"
      "\tOracle grad: g_x1 = {:.5f}\n"
      "\tResult grad: g_x1 = {:.5f}".
      format(5.5, x1.grad[0]))

"""
测试用例3：x1为常数，x2为张量
"""
x1 = 2
x2 = Zhangliang(5, requires_grad=True)
f = log(x1) + x1 * x2 - sin(x2)
f.backward()
print("Test function f=log(x1)+x1*x2-sin(x2), with initial values x1=2, x2=5.\n"
      "\tOracle grad: g_x2 = {:.5f}\n"
      "\tResult grad: g_x2 = {:.5f}".
      format(1.716, x2.grad[0]))

"""
测试用例4：no_grad环境
"""
x1 = Zhangliang(2, requires_grad=True)
x2 = Zhangliang(5, requires_grad=True)

with no_grad():
    f = log(x1) + x1 * x2 - sin(x2)

try:
    f.backward()
    print('This line should not be print.')
except:
    print('Backprop is disabled in `no_grad` situation.')

"""
测试用例5：has_grad环境
"""
x1 = Zhangliang(2, requires_grad=True)
x2 = Zhangliang(5, requires_grad=True)

with no_grad():
    with has_grad():
        f = log(x1) + x1 * x2 - sin(x2)

try:
    f.backward()
    print("Test function f=log(x1)+x1*x2-sin(x2), with initial values x1=2,
x2=5.\n"
          "\tOracle grad: g_x1 = {:.5f}, g_x2 = {:.5f}\n"
```

```
            "\tResult grad: g_x1 = {:.5f}, g_x2 = {:.5f}".
            format(5.5, 1.716, x1.grad[0], x2.grad[0]))
except:
    print('This line should not be print.')
```

可以看到输出结果：

```
test_tensor_ops.py .Test function f=log(x1)+x1*x2-sin(x2), with initial values x1=2, x2=5.
    Oracle grad: g_x1 = 5.50000, g_x2 = 1.71600
    Result grad: g_x1 = 5.50000, g_x2 = 1.71634
Test function f=log(x1)+x1*x2-sin(x2), with initial values x1=2, x2=5.
    Oracle grad: g_x1 = 5.50000
    Result grad: g_x1 = 5.50000
Test function f=log(x1)+x1*x2-sin(x2), with initial values x1=2, x2=5.
    Oracle grad: g_x2 = 1.71600
    Result grad: g_x2 = 1.71634
Backprop is disabled in `no_grad` situation.
Test function f=log(x1)+x1*x2-sin(x2), with initial values x1=2, x2=5.
    Oracle grad: g_x1 = 5.50000, g_x2 = 1.71600
    Result grad: g_x1 = 5.50000, g_x2 = 1.71634
```

计算库实现探究

张量广播规律

张量计算过程中会有各种广播问题。由于自定义数据结构（Zhangliang）实际上是对numpy.ndarray的又一次封装，前馈过程的广播（broadcast）可由numpy内置运算确定，张量各维度需满足一定的关系才能完成广播。目前观察到张量在各个库（numpy，tensorflow和pytorch）内存在两类广播形式，我们这里分别称为"元素型"和"矩阵型"（下表中的n下标从1开始）：

	a 大小	b 大小	输出大小	反传时 b 应该缩减的维度（下标从 1 开始）
元素型 a+b a*b a-b a/b	$(n_1,n_2,\cdots,n_K)$	$(1,)$	$(n_1,n_2,\cdots,n_K)$	$(1,2,\cdots,K-1,K)$
		$\left(\underbrace{1,1,\cdots,1}_{K-1},n_K\right)$	$(n_1,n_2,\cdots,n_K)$	$(1,2,\cdots,K-2,K-1)$
		$\left(n_1,\underbrace{1,\cdots,1}_{j-2},n_j,\underbrace{1,\cdots,1}_{K-j-1},n_K\right)$	$(n_1,n_2,\cdots,n_K)$	$(2,3,\cdots,j-1,j+1,\cdots,K-1)$
		$(n_1,n_2,\cdots,n_K)$	$(n_1,n_2,\cdots,n_K)$	空
		$(n_1,)$	无法广播	
		(n_{K-1},n_K)	$(n_1,n_2,\cdots,n_K)$	$(1,2,\cdots,K-2)$
矩阵型 matmul(a,b)	$(n_1,n_2,\cdots,n_K)$	$(1,)$	无法广播	
		$(n_K,)$	$(n_1,n_2,\cdots,n_{K-1})$	$(1,2,\cdots,K-2)$
		(n_K,m)	$(n_1,n_2,\cdots,n_{K-1},m)$	$(1,2,\cdots,K-2)$
		$\left(\underbrace{1,\cdots,1}_{<K-2},n_K,m\right)$	$(n_1,n_2,\cdots,n_{K-1},m)$	$(1,2,\cdots,K-2)$
		(n_{K-2},n_K,m)	$(n_1,n_2,\cdots,n_{K-1},m)$	$(1,2,\cdots,K-3)$
		$\left(n_1,\underbrace{1,\cdots,1}_{j-2},n_j,\underbrace{1,\cdots,1}_{K-j-1},n_K,m\right)$	$(n_1,n_2,\cdots,n_{K-1},m)$	$(2,3,\cdots,j-1,j+1,\cdots,K-2)$
		$\left(\underbrace{1,\cdots,1}_{<K-3},n_{K-2},n_K,m\right)$	$(n_1,n_2,\cdots,n_{K-1},m)$	$(1,2,\cdots,K-3)$

可以看到不同的广播类型在反传时需要有不同的维度缩减策略。但事实上我们可以大致总结出不同广播类型需要满足的条件。

假定两个变量的大小分别为：$(a_1,a_2,\ldots,a_m)$以及$(b_1,b_2,\ldots,b_n)$。不失对称性，假定$n<m$。那么元素级运算的广播需满足：

$$(a_1,a_2,\ldots,a_m)\sim(\underbrace{1,1,\ldots,1}_{n-m},b_1,b_2,\ldots,b_n) \tag{34}$$

也即，在广播时，numpy会尝试将维度较小的那个变量进行维度扩增（填充1），扩增至m。如果扩增后的两个维度互容，那么则允许广播。

类似的，矩阵型的广播条件为：

$$(a_1, a_2, \dots, a_m) \sim (\underbrace{1, 1, \dots, 1}_{n-m}, b_1, b_2, \dots b_{n-1}, b_n) \tag{35}$$

$$\text{s.t.} \quad a_m = b_{n-1} \tag{36}$$

确定了广播条件，我们就能在反传时确定输出梯度如何反馈到输入张量。我们举个例子说明反传时的维度缩减情况。

元素级运算

假定两个矩阵a大小为2×1，b大小为2×2：

$$a = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad b = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} \tag{37}$$

两者的四则运算为（以加法为例）：

$$c = a \oplus b = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \oplus \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} x_1 + y_{11} & x_1 + y_{12} \\ x_2 + y_{21} & x_2 + y_{22} \end{bmatrix} = \begin{bmatrix} \tilde{y}_{11} & \tilde{y}_{12} \\ \tilde{y}_{21} & \tilde{y}_{22} \end{bmatrix} \tag{38}$$

输出大小为2×2。在反向传播时，损失值传到变量c应该同样为2×2大小。那么传回变量a和b时：

$$\frac{\partial l}{\partial x_1} = \frac{\partial l}{\partial \tilde{y}_{11}} \frac{\partial \tilde{y}_{11}}{\partial x_1} + \frac{\partial l}{\partial \tilde{y}_{12}} \frac{\partial \tilde{y}_{12}}{\partial x_1} = \frac{\partial l}{\partial \tilde{y}_{11}} + \frac{\partial l}{\partial \tilde{y}_{12}} \tag{39}$$

$$\frac{\partial l}{\partial y_{11}} = \frac{\partial l}{\partial \tilde{y}_{11}} \frac{\partial \tilde{y}_{11}}{\partial y_{11}} = \frac{\partial l}{\partial \tilde{y}_{11}} \tag{40}$$

记上层传播到张量c的Jacobian为：

$$\nabla_c l = \begin{bmatrix} \dfrac{\partial l}{\partial \tilde{y}_{11}} & \dfrac{\partial l}{\partial \tilde{y}_{12}} \\ \dfrac{\partial l}{\partial \tilde{y}_{21}} & \dfrac{\partial l}{\partial \tilde{y}_{22}} \end{bmatrix} \tag{41}$$

那么：

$$\nabla_a l = \text{ReducedSum}(\nabla_c l, dim = 1) \tag{42}$$

$$\nabla_b l = \nabla_c l \tag{43}$$

元素乘法类似：

$$c = a + b = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \otimes \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} x_1 y_{11} & x_1 y_{12} \\ x_2 y_{21} & x_2 y_{22} \end{bmatrix} = \begin{bmatrix} \tilde{y}_{11} & \tilde{y}_{12} \\ \tilde{y}_{21} & \tilde{y}_{22} \end{bmatrix} \tag{44}$$

$$\frac{\partial l}{\partial x_1} = \frac{\partial l}{\partial \tilde{y}_{11}} \frac{\partial \tilde{y}_{11}}{\partial x_1} + \frac{\partial l}{\partial \tilde{y}_{12}} \frac{\partial \tilde{y}_{12}}{\partial x_1} = \frac{\partial l}{\partial \tilde{y}_{11}} y_{11} + \frac{\partial l}{\partial \tilde{y}_{12}} y_{12} \tag{45}$$

$$\frac{\partial l}{\partial y_{11}} = \frac{\partial l}{\partial \tilde{y}_{11}} \frac{\partial \tilde{y}_{11}}{\partial y_{11}} = \frac{\partial l}{\partial \tilde{y}_{11}} x_1 \tag{46}$$

所以：

$$\nabla_a l = \text{ReducedSum}(\nabla_c l \otimes b, dim = 1) \tag{47}$$

$$\nabla_b l = \nabla_c l \otimes a \tag{48}$$

这里的`dim`参数序号从0开始。

矩阵级运算

假定两个矩阵，a大小为2×1，b大小为$4 \times 1 \times 3$：

$$a = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad b = \begin{bmatrix} [\, y_{11} & y_{12} & y_{13} \,] \\ [\, y_{21} & y_{22} & y_{23} \,] \\ [\, y_{31} & y_{32} & y_{33} \,] \\ [\, y_{41} & y_{42} & y_{43} \,] \end{bmatrix} \tag{49}$$

由于矩阵乘法实际参与运算的维度是最后两维，所以两者的矩阵乘结果为$4 \times 2 \times 3$大小：

$$
\begin{aligned}
c = a \times b &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \times \begin{bmatrix} \begin{bmatrix} y_{11} & y_{12} & y_{13} \end{bmatrix} \\ \begin{bmatrix} y_{21} & y_{22} & y_{23} \end{bmatrix} \\ \begin{bmatrix} y_{31} & y_{32} & y_{33} \end{bmatrix} \\ \begin{bmatrix} y_{41} & y_{42} & y_{43} \end{bmatrix} \end{bmatrix} \\
&= \begin{bmatrix} \begin{bmatrix} x_1 y_{11} & x_1 y_{12} & x_1 y_{13} \\ x_2 y_{11} & x_2 y_{12} & x_2 y_{13} \end{bmatrix} \\ \begin{bmatrix} x_1 y_{21} & x_1 y_{22} & x_1 y_{23} \\ x_2 y_{21} & x_2 y_{22} & x_2 y_{23} \end{bmatrix} \\ \begin{bmatrix} x_1 y_{31} & x_1 y_{32} & x_1 y_{33} \\ x_2 y_{31} & x_2 y_{32} & x_2 y_{33} \end{bmatrix} \\ \begin{bmatrix} x_1 y_{41} & x_1 y_{42} & x_1 y_{43} \\ x_2 y_{41} & x_2 y_{42} & x_2 y_{43} \end{bmatrix} \end{bmatrix} = \begin{bmatrix} \begin{bmatrix} z_{111} & z_{112} & z_{113} \\ z_{211} & z_{212} & z_{213} \end{bmatrix} \\ \begin{bmatrix} z_{121} & z_{122} & z_{123} \\ z_{221} & z_{222} & z_{223} \end{bmatrix} \\ \begin{bmatrix} z_{131} & z_{132} & z_{133} \\ z_{231} & z_{232} & z_{233} \end{bmatrix} \\ \begin{bmatrix} z_{141} & z_{142} & z_{143} \\ z_{241} & z_{242} & z_{243} \end{bmatrix} \end{bmatrix}
\end{aligned} \tag{50}
$$

于是不难得到：

$$
\frac{\partial l}{\partial x_1} = \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{\partial l}{\partial z_{1ij}} \frac{\partial z_{1ij}}{\partial x_1} = \sum_{i=1}^{4} \sum_{j=1}^{3} \frac{\partial l}{\partial z_{1ij}} y_{ij} \tag{51}
$$

$$
\frac{\partial l}{\partial y_{11}} = \sum_{i=1}^{2} \frac{\partial l}{\partial z_{i11}} \frac{\partial z_{i11}}{\partial y_{11}} = \sum_{i=1}^{2} \frac{\partial l}{\partial z_{i11}} x_i \tag{52}
$$

因此，类似的，如果记：

$$
\nabla_c l = \left[\frac{\partial l}{\partial z_{ijk}} \right]_{4 \times 2 \times 3} \tag{53}
$$

那么反传时的梯度为：

$$
\begin{aligned}
&\nabla_a l = \mathrm{ReducedSum}(\nabla_c l \times b^T, dim = 0) \\
&\nabla_b l = a^T \times \nabla_c l
\end{aligned} \tag{54}
$$

这里的转置是对张量最后两个维度进行转换。

从上述的例子中可以发现，我们需要根据输入张量和输出张量之间的维度差别，从而找出对每个输入应该如何变换其数据形状才能计算出其对应的梯度。

部分不可导函数的微分近似

函数	近似微分
abs	
max/min	
maximum/minimum	

浮点数精度问题

在编写测试脚本时发现一个问题：`np.float32`精度比较低，经常导致测试通不过。改成`np.float64`稍微好一点。但是实际使用时，神经网络其实可以允许一定的截断误差，而且这些截断误差可能有利于网络的训练（比如跳出局部最优点）。因此，测试时使用`np.float64`，而实际使用时`np.float32`即可。

算子实现

这一节探讨下各种算子的实现问题，也会简单分析下各种算子的梯度计算问题。梯度计算这一过程，事实上大部分都能在网上找到计算公式，但是自己实现过程中还是需要推导一下，这一过程中也能更好地理解其工程考虑。

Sigmoid算子

假定输入为x（可以是标量或者张量），那么输出为：

$$z_i = \frac{1}{1+e^{-x_i}} \tag{55}$$

记$y_i = e^{x_i}$，那么：

$$z_i = \frac{y_i}{1+y_i} \tag{56}$$

于是有：

$$\begin{aligned}
\frac{\partial z_i}{\partial x_i} &= \frac{\partial z_i}{\partial y_i}\frac{\partial y_i}{\partial x_i} && (57)\\
&= \frac{1+y_i-y_i}{(1+y_i)^2}e^{x_i} && (58)\\
&= \frac{1}{(1+y_i)^2}y_i && (59)\\
&= \frac{1}{1+y_i}\frac{y_i}{1+y_i} && (60)\\
&= (1-z_i)z_i && (61)
\end{aligned}$$

具体实现还有一点需要考虑：式(???)适用于x为正的情况，当x为负且绝对数值较大时，e^{-x}会造成数值溢出，此时较好的实现应该是：

$$z = \frac{e^x}{1+e^x} \tag{62}$$

因此，需要根据x内每个点的数值情况分别使用不同的计算方法，以保证不会溢出。

Softmax算子

同样假定输入为x，那么输出为：

$$z_i = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{63}$$

记$y_i = e^{x_i}$，那么：

$$z_i = \frac{y_i}{\sum_j y_j} \tag{64}$$

于是$j \neq i$时，$\frac{\partial y_j}{\partial x_i} = 0$，有：

$$\begin{aligned}
\frac{\partial z_i}{\partial x_i} &= \sum_j \frac{\partial z_i}{\partial y_j}\frac{\partial y_j}{\partial x_i}\\
&= \frac{\partial z_i}{\partial y_i}\frac{\partial y_i}{\partial x_i}\\
&= \frac{\sum_j y_j - y_i}{(\sum_j y_j)^2}e^{x_i}\\
&= \left(\frac{1}{\sum_j y_j} - \frac{y_i}{(\sum_j y_j)^2}\right)y_i\\
&= \frac{y_i}{\sum_j y_j} - \left(\frac{y_i}{\sum_j y_j}\right)^2\\
&= (1-z_i)z_i
\end{aligned} \tag{65}$$

softmax算子同样有溢出的问题，但是比较好处理。一个通常的方法是找出对应维度的最大值，然后减去这个最大值，保证e^x中的x小于0：

$$\tilde{x} = x - \max(x, dim) \tag{66}$$

$$z = \frac{e^{\tilde{x}}}{\text{ReduceSum}(e^{\tilde{x}}, dim)} \tag{67}$$

卷积算子

参考资料：

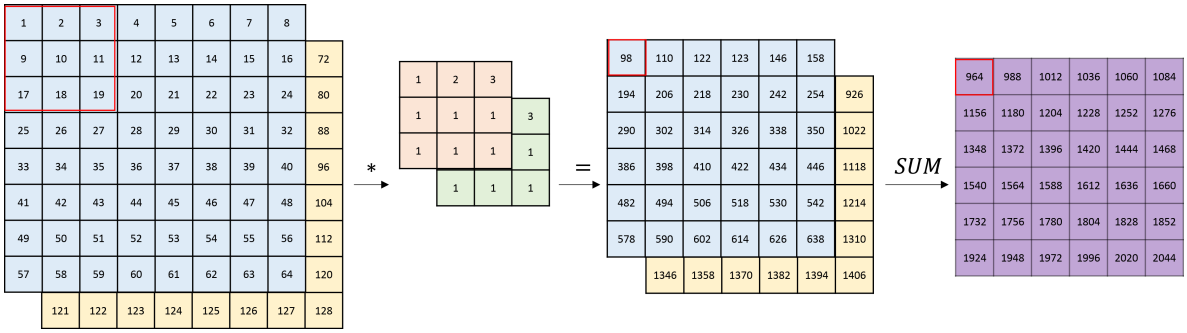
- FAST CONVOLUTIONAL NETS WITH fbfft : A GPU PERFORMANCE EVALUATION
- Fast Algorithms for Convolutional Neural Networks
- Leonardo的博客
- Sahnimanas的博客
- Jonathan Ragan-Kelley的博士论文
- 贾扬清的备忘录

卷积作为现代神经网络中最重要的算子，其实现需要特定的优化。卷积过程本质上仍可看成是对图像每个局部区域内的信息进行矩阵乘法，因此实现卷积算子时力求将这一个过程进行加速和优化。这其中又涉及到算法具体实现时的各种问题，计算调度、缓存调度、并行计算等（事实上其他算子也有这些问题，只是在卷积算子中这些问题更加突出）。由于笔者能力有限，暂时未完成这一部分。

目前各大框架对卷积的实现有四种方法：

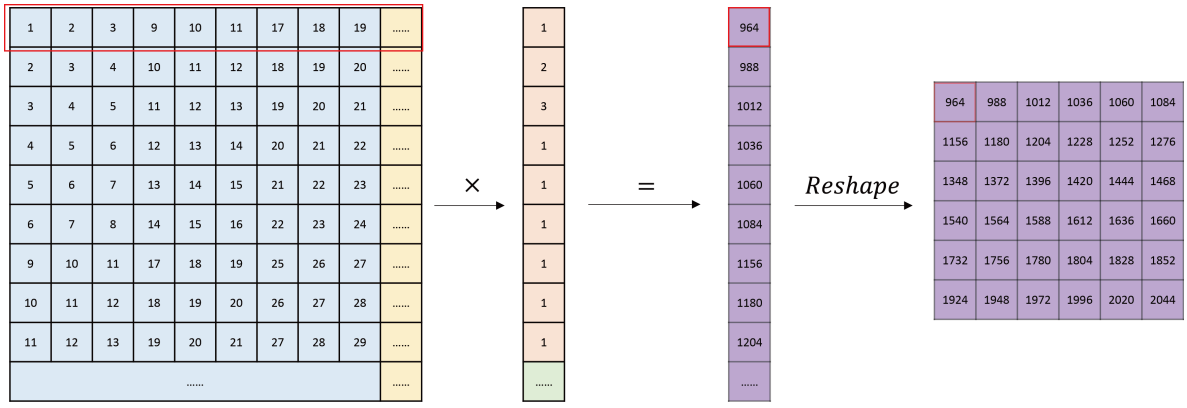
- 直接计算
- 通过im2col方法，将卷积完全转化为两个矩阵的乘法
- 通过FFT进行计算
- 通过winograd方法

直接计算



朴素计算方法比较直观，给定一个cin通道数的特征图/图像，一个卷积核对每个通道内的数据进行滑动窗口内元素乘法加和运算，如此生成一个cin通道的中间特征图，再将各通道数据进行求和获得最终的输出。这种运算可以用于$cout$个不同的卷积核同时进行运算，最终形成$cout$通道的输出特征图。 很明显，朴素计算方法没有任何的优化，假定输入特征图大小为$S \times n \times n \times f$，卷积核大小$f' \times f \times k \times k$，步长1，没有$padding = 1$，输出特征图大小为$S \times n \times n \times f'$，实际上乘法执行次数为$Sff'n^2k^2$，加法执行次数为$Sff'n^2k^2$。我们知道计算机里面乘法的运算时间比加法大很多。将“一次乘法+一次加法”称为一个“FLOP”，那么朴素计算所需的复杂度可表示为$O(Sff'n^2k^2)$ FLOPs。

通过im2col方法



展开的特征图　　展开的卷积核

观察卷积运算过程，其实就是每个位置上的特征数据与对应位置的核参数进行相乘后，再对局部区域进行求和的操作，再对各通道求和，本质就可看成是矩阵/向量之间的乘法。所以`im2col`的思路就是将特征图局部和卷积核进行展开，处理成矩阵或向量来完成运算。这样展开成大矩阵，通过矩阵乘法对卷积运算进行加速。我们知道矩阵乘法已经有了加速的方法，因此卷积的效率可以进一步提高。上图中展示了一个具体过程。

通过FFT进行计算

卷积过程顾名思义，来自于信号处理邻域。那么在信号处理中常用的傅里叶变换自然也可用于与实现卷积的计算，而且比朴素方法更加高效：

$$y_{(s,j)} = \sum_{i \in f} x_{(s,i)} \star w_{(j,i)} = \sum_{i \in f} \mathcal{F}^{-1}\left(\mathcal{F}(x_{(s,i)}) \circ \mathcal{F}(w_{(j,i)})^*\right)$$

其中$\circ$是点对点乘积，$*$是共轭。相比于朴素算法的$O(Sff'n^2k^2)$复杂度，通过FFT计算卷积的复杂度只需要$O(Sff'n^2 + (Sf + ff' + Sf')n^2 \log n)$。基于FFT的卷积方法通常将空间维度$n$分解为若干个基础卷积算子的组合，比如$2 \times 2$，$3 \times 3$，$5 \times 5$，$7 \times 7$，计算这些组合的结果后再计算总的输出结果。如果$n$不能分解为这些质数卷积核的组合，那么就比$n$大且能分解为上述质数组合的最快的点数$n'$的卷积，然后再取对应的值。由于FFT的性质，在卷积核较小问题规模也较小时，基于FFT的卷积性能反而可能不如朴素方法；但随着问题规模的增大，或者卷积核较大时，基于FFT的卷积性能要远远好于朴素卷积方法，以及`im2col`方法。

通过winograd方法

`Winograd`则是另一种优化思路，通过优化计算的流程来加快卷积过程。通常对于较小的卷积核，用winograd方法计算会比较快。

优化方法

参考资料：

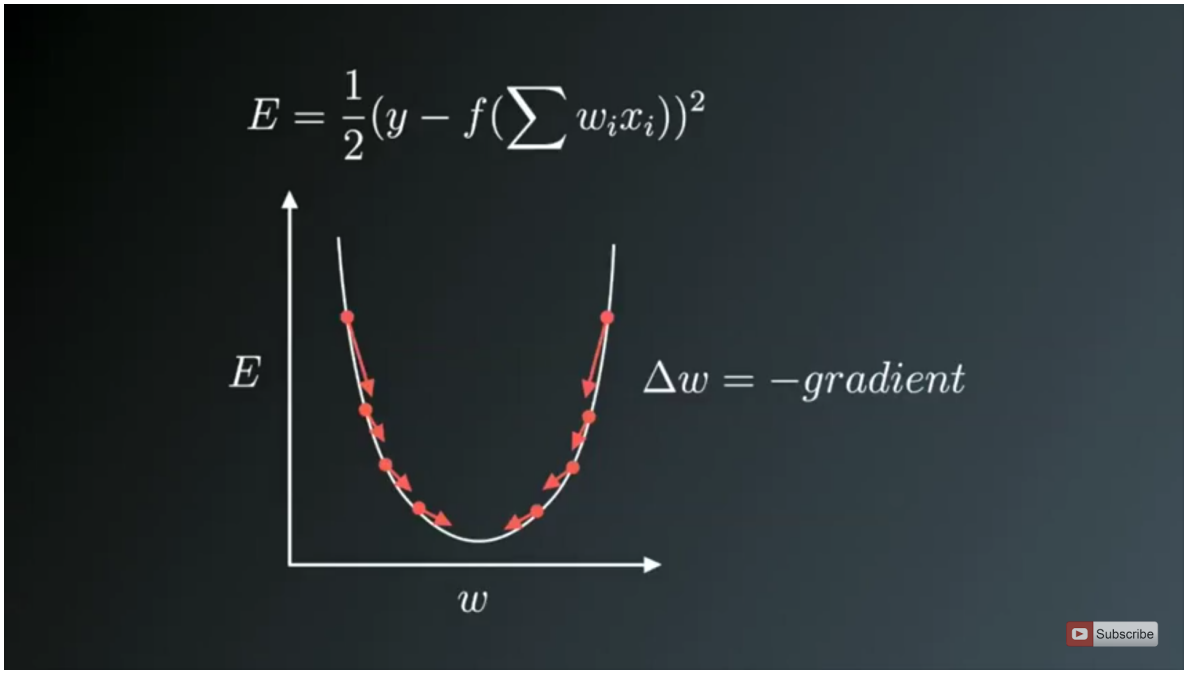
- Rprop
- Caffe Solver
- 深度学习最全优化方法总结比较（SGD，Adagrad，Adadelta，Adam，Adamax，Nadam）
- Types of Optimization Algorithms used in Neural Networks and Ways to Optimize Gradient Descent
- PyTorch与caffe中SGD算法实现的一点小区别

常见的网络优化方法包括以下若干种：

方法 方法	优化规则 优化规则	备注 备注
Momentum SGD	$v_{t+1} = \mu v_t - \alpha \nabla L_w$ $w_{t+1} = w_t + v_{t+1}$	μ是动量，α是学习率
AdaGrad	$w_{t+1} = w_t - \frac{\alpha}{\sqrt{\sum_{\tau=1}^{t} (\nabla L_w)_\tau^2}} \nabla L_w$	
AdaDelta	$v_{t+1} = -\frac{RMS[v]_t}{RMS[g]_{t+1}} g_{t+1}$ $w_{t+1} = w_t + v_{t+1}$	RMS是root of mean squared
Adam	$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla L_w$ $v_t = \beta_2 v_{t-1} + (1-\beta_2)(\nabla L_w)^2$ $w_{t+1} = w_t - \alpha \frac{\sqrt{1-\beta_2^t}}{1-\beta_1^t} \frac{m_t}{\sqrt{v_t}+\epsilon}$	自动矩估计
Nesterov	$v_{t+1} = \mu v_t - \alpha \nabla L_{w_t + \mu v_t}$ $w_{t+1} = w_t + v_{t+1}$	
RMSprop	$MS(w_t) = \delta MS(w_{t-1}) + (1-\delta)\nabla L_w^2$ $w_{t+1} = w_t - \alpha \frac{\nabla L_w}{\sqrt{MS(w_t)}}$	
Rprop	$v_{t+1} = -\alpha_{t+1} * sgn(\nabla L_w)$ $w_{t+1} = w_t + v_{t+1}$ $\alpha_{t+1} = \begin{cases} \min(\alpha_t * a, \alpha_{max}),\ if\ \nabla_{w_{t+1}} L * \nabla_{w_t} L > 0 \\ \max(\alpha_t * b, \alpha_{min}),\ if\ \nabla_{w_{t+1}} L * \nabla_{w_t} L < 0 \\ \alpha_t,\ otherwise \end{cases}$	$a > 1 > b$，典型值$a = 1.2,\ b = .5$

随机梯度下降SGD

简单回顾下随机梯度下降方法。在没有解析解的情况下，我们知道优化一个函数往往是通过一些其他方法，其中梯度下降方法就是最重要的一种。其目的是从某个点开始，沿着当前点的负梯度方向逐步找到函数的极小值点。



对于机器学习任务，可能有大量数据用于学习$\{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$。学习时，一种方法就是每次都使用所有数据对模型参数进行调整：

$$\Delta w = -\sum_{i=1}^{N} \nabla_w L(x_i, y_i) \tag{68}$$
$$w_{t+1} = w_t + \alpha \Delta w \tag{69}$$

若干次迭代优化达到收敛即可。当N较小时，上式方法仍比较高效；但随着N的逐渐增大，同时计算所有样本对这一过程成为了训练的瓶颈。无论是内存、显存或者算力，都无法支持如此大量数据同时进行训练和迭代；而逐批输入数据再汇总则具有巨大的时间开销，也是不可取的。因此随机梯度下降方法也就应运而生。它是对上式的一种极端简化：每次迭代只取一个样本计算梯度。

$$\Delta w = -\nabla_w L(x_i, y_i) \tag{70}$$
$$w_{t+1} = w_t + \alpha \Delta w \tag{71}$$

而这个样本也是每次迭代随机选取的。由于SGD不知道每次使用的是什么数据，或者之前是否已使用过，这种随机性以及一个独立样本内包含的噪声干扰能够一定程度上提升模型的泛化能力。SGD的收敛性有Robbins-Siegmund定理保证，只需要满足$\sum_t \alpha_t^2 < \infty$以及$\sum_t \alpha_t = \infty$。

但上述极端简化方法仍存在一个问题：样本噪声在提升鲁棒性和泛化性的同时，也使得SGD收敛得非常慢。所以在实现上，现代的SGD了吸收了两种极端情况的各自优点，将单例更新修改为了批数据更新：

$$\Delta w = -\sum_{i=1}^{B} \nabla_w L(x_i, y_i) \tag{72}$$
$$w_{t+1} = w_t + \alpha \Delta w \tag{73}$$

其中B是批大小。基于批的SGD既提高了单例SGD的收敛速度，也保留了样本的随机性使得模型更具泛化性能。当然SGD本身又有可改进之处。

二阶优化方法

在介绍SGD的各种改进方法之前，有必要介绍下二阶的优化方法。我们称普通的梯度下降方法是一阶的，因为它们只用到了损失函数$\mathcal{L}$对参数w的一阶导数。事实上，优化方法可以使用更高阶的信息，比如二阶优化方法（Newton法、拟Newton法等）用到了Hessian矩阵来调整演化的方向：

$$\Delta w = -\frac{1}{|\text{diag}(H_t)| + \mu} \nabla_w L \tag{74}$$

通常，高阶优化方法比一阶方法具有更快的收敛速度，当然也是有代价的：需要计算$\mathcal{L}$对参数w的二阶导数（Hessian矩阵），这是一个时间复杂度和空间复杂度都很高的过程。

Momentum SGD （带动量的SGD）

SGD存在一个问题是会产生振荡，因此收敛更难更慢。动量方法是SGD最常用最有效的一种改进，其思想是增强梯度持续指向方向的演化速度，减缓梯度符号变化之处的演化进程。做到这两点只需要一个小小的改进：

$$\Delta w_{t+1} = \rho \Delta w_t - \alpha \sum_{i=1}^{B} \nabla_w L(x_i, y_i) \tag{75}$$
$$w_{t+1} = w_t + \Delta w_{t+1} \tag{76}$$

其中$\rho \in [0, 1]$是动量参数。通过引入动量ρ，梯度演化方法将受到历史演化方向的影响。在梯度主方向，演化速度会累计提升加速；而在异常点，新的梯度方向影响较小，也就避免了参数在某个点附近振荡，加速了优化过程。

这里需要提到一点，事实上根据FAIR的论文，朴素的动量SGD的学习率不是立即生效的，因为学习率α只是被应用在了新的梯度上；当α发生变化时，动量方向仍旧保持为前一时刻的方向，并且需要很长时间才能调整到新的方向。所以PyTorch在实现优化器时，使用的不是(???)和(???)，而是调整为：

$$\Delta w_{t+1} = \rho \Delta w_t - \sum_{i=1}^{B} \nabla_w L(x_i, y_i) \tag{77}$$
$$w_{t+1} = w_t + \alpha \Delta w_{t+1} \tag{78}$$

这样在α改变时演化立即就能生效。这还带来一个额外的好处，就是在实施**权重衰减**时也能避免衰减项主导演化速度和方向。考察权重衰减项$\frac{\lambda}{2}||w||^2$，通常这一项是单独加入更新中的。按照式(???)，加入衰减后的权重更新变为：

$$\Delta w_{t+1} = \rho \Delta w_t - \alpha \sum_{i=1}^{B} \nabla_w L(x_i, y_i) - \lambda w_t \tag{79}$$

$$w_{t+1} = w_t + \Delta w_{t+1} \tag{80}$$

那么在更新权重时，如果学习率很小，衰减项将占据主导地位从而影响收敛和网络性能。如果按照PyTorch的实现方法，加入衰减的权重更新则变为了：

$$\Delta w_{t+1} = \rho \Delta w_t - \sum_{i=1}^{B} \nabla_w L(x_i, y_i) - \lambda w_t \tag{81}$$

$$w_{t+1} = w_t + \alpha \Delta w_{t+1} \tag{82}$$

可以看到，在PyTorch版本的SGD中，衰减项对演化速度和方向的影响要小于原实现。

Nesterov Accelerated Gradient

Yurii Nesterov观察到动量方法的一个问题：动量SGD无法预见到极值点应该减小演化速度。当优化达到某个极值点时，优化速度应该减小并趋于0；而由于动量的存在，动量SGD需要很长时间才能减小演化速度，甚至错过极值点。Nesterov的解决方法简单直接：相比于计算当前参数的梯度，我们可以前瞻一步查看下一步参数的梯度方向。如果这个未来的梯度方向与当前方向是反向的，那么应该减小当前的演化速度：

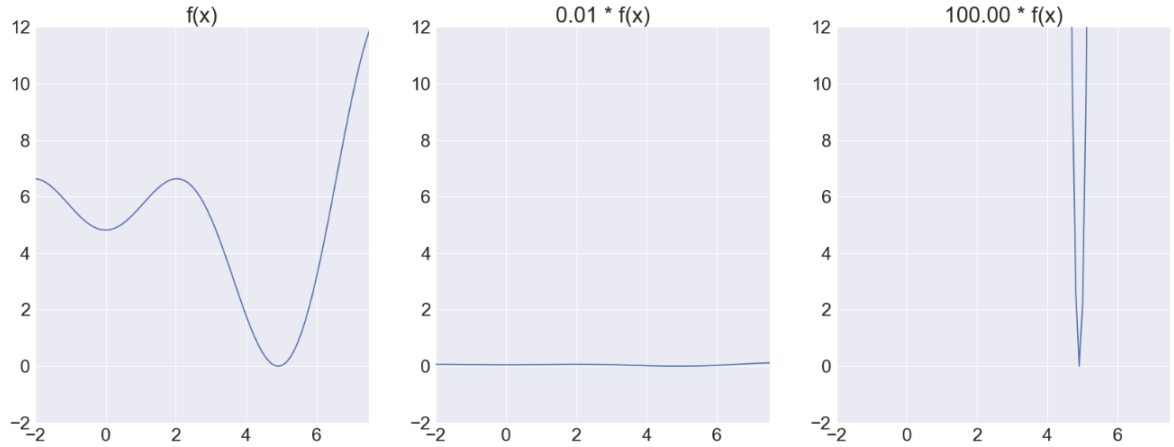
$$\Delta w_{t+1} = \rho \Delta w_t - \alpha \sum_{i=1}^{B} \nabla_{w+\rho \Delta w} L(x_i, y_i) \tag{83}$$

$$w_{t+1} = w_t + \Delta w_{t+1} \tag{84}$$

这一方法能够根据梯度大小和符号自适应调整演化速度。

Resilient Propagation (Rprop)

弹性反向传播基于一个有趣的发现：梯度下降优化过程很可能与梯度绝对数值大小无关，而与梯度的符号有关。为了说明这一点，定义一个函数f，然后相应地定义另外两个函数，都是对f赋予一个scale量：



Three functions with the same optima but vastly different gradients

第二、第三个图与原函数f相差一个尺度量。显然三者有一个相同的最优点，但是三个函数的各点的梯度绝对数值却也差了一个scale。于是，在不知道理论函数与实际函数的尺度情况下，确定演化步长α是个比较困难的事情。基于梯度的绝对数值来确定学习率就可能导致优化过程无法收敛到极值点。Rprop方法提出使用梯度符号来进行优化，并且提出一种自适应学习率的方法来调整演化步长：

$$v_{t+1} = -\alpha_{t+1} * sgn\left(\nabla L_w\right) \tag{85}$$
$$w_{t+1} = w_t + v_{t+1} \tag{86}$$
$$\alpha_{t+1} = \begin{cases} \min(\alpha_t * a, \alpha_{max}),\ if\ \nabla_{w_{t+1}} L * \nabla_{w_t} L > 0 \\ \max(\alpha_t * b, \alpha_{min}),\ if\ \nabla_{w_{t+1}} L * \nabla_{w_t} L < 0 \\ \alpha_t,\ otherwise \end{cases} \tag{87}$$

Rprop的另一个优点是，能够根据每个参数梯度的数值大小自适应调整参数的学习率。换言之，由于量级不同，我们可能很难找到一个全局的学习率适合于所有参数，但是Rprop使得每个不同的参数都具有自己的学习率，这就能够克服梯度量级不同的问题。当然Rprop也有自己的缺陷，比如Rprop只适用于大批次数据，而对小批次数据不太适合。这是因为不同批次数据的梯度符号可能会不同，小批次数据的梯度符号变动较频繁，导致Rprop学习无法收敛。

RMSprop

正因为Rprop无法用于小批次数据，Tieleman提出了RMSprop将Rprop的思想应用于小批次数据。其想法也很简单：Rprop对每个batch都会除以一个不同的数，那么为什么不对相邻mini-batch除以不同的scale呢（并且每批次的scale只是略有不同）？ 实现这个思想的方法也很简单，只需要维护一个浮动的统计数据：

$$MS(w_t) = \delta MS(w_{t-1}) + (1-\delta)\nabla L_w^2 \tag{88}$$
$$w_{t+1} = w_t - \alpha \frac{\nabla L_w}{\sqrt{MS(w_t)}} \tag{89}$$

其中δ默认取0.9（Tieleman论文）或者0.99（Caffe默认）

AdaGrad

AdaGrad是另一种自适应学习率的方法。类似与RMSprop，其梯度被除以了一个尺度量：

$$MS(w_t) = MS(w_{t-1}) + (\nabla L_w)^2 \tag{90}$$
$$w_{t+1} = w_t - \alpha \frac{\nabla L_w}{\sqrt{MS(w_t)}} \tag{91}$$

注意到第一式与RMSprop中略有不同。在RMSprop中，浮动量$MS(w_t)$是由手动设定的数值δ控制更新的，为固定值；而AdaGrad中是按照迭代次数t进行累计。这么做也有一个额外的优点，即学习率自主地随着时间减小，类似于退火技术。但是这也带来几个问题：

- AdaGrad方法对于初始值比较敏感。比如，初始值具有较大的梯度值，那么后续所有的学习率实际上都被缩小了，造成学习缓慢；如果选择手动增大初始学习率，这反而又造成AdaGrad方法对初始学习率敏感了。
- 学习率将随着训练过程的进行一直减小，直到无法再学习。

AdaDelta

AdaDelta着重针对AdaGrad的上述两个缺点，做了两点改进：

- 将梯度平方累计过程改为在一个窗口的时间；
- 一阶方法忽略了数值单位，所以使用近似二阶数据来弥补

事实上第一点与RMSprop的表达式是一致的，即：

$$E[g^2]_t = \rho E[g^2]_{t-1} + (1-\rho) g_t^2 \tag{92}$$
$$RMS[g]_t = \sqrt{E[g^2]_t + \varepsilon} \tag{93}$$
$$v_t = -\frac{\alpha}{RMS[g]_t} g_t \tag{94}$$

其中$g_t = \nabla_{w_t} L$为t时刻的梯度。第二点，在一阶优化方法中，梯度和参数的度量单位都是被忽略的，都是有问题的：

$$\text{units of } \Delta w \propto \text{units of } \nabla_w L \propto \frac{\partial \mathcal{L}}{\partial w} \propto \frac{1}{\text{units of } x} \tag{95}$$

相比之下二阶方法则保证了度量单位的统一：

$$\text{units of } \Delta w \propto H^{-1}\nabla_w L \propto \frac{\frac{\partial \mathcal{L}}{\partial w}}{\frac{\partial^2 \mathcal{L}}{\partial w^2}} \propto \text{units of } x \tag{96}$$

可以看到，一阶方法和二阶方法在度量单位上相差了两级。而由于：

$$\Delta w = \frac{\frac{\partial \mathcal{L}}{\partial w}}{\frac{\partial^2 \mathcal{L}}{\partial w^2}} \Rightarrow \frac{1}{\frac{\partial^2 \mathcal{L}}{\partial w^2}} = \frac{\Delta w}{\frac{\partial \mathcal{L}}{\partial w}} \tag{97}$$

我们只需要在现有梯度更新的过程中加入类似于二阶导数倒数的scale即可：

$$v_{t+1} = -\frac{RMS[v]_t}{RMS[g]_{t+1}} g_{t+1} \tag{98}$$

$$w_{t+1} = w_t + v_{t+1} \tag{99}$$

Adaptive Moments Esimation (Adam)

Adam方法集合了AdaGrad的优点（较适用于稀疏梯度）与RMSprop的优点，具有诸多优点：

- 对于梯度的赋值不敏感
- 演化步长由超参数控制
- 不需要稳态优化目标
- 能够适用于稀疏梯度
- 自动步长退火

Adam方法也很简单：直接用浮动数据（running average）来拟合和估计一阶矩和二阶矩，然后计算梯度方向：

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{100}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \tag{101}$$

$$w_{t+1} = w_t - \alpha \frac{\sqrt{1-\beta_2^t}}{1-\beta_1^t} \frac{m_t}{\sqrt{v_t}+\epsilon} \tag{102}$$

这一形式与RMSprop的动量版本非常相似。不同之处在于，RMSprop的动量版本是依靠梯度的幅值来调整动量，而Adam则是直接维护和估计了梯度的二阶矩；另外RMSprop缺少有偏修正量（bias-correction term），在稀疏梯度的应用中会造成较大步长而使得优化过程发散。

AdaMax

AdaMax是Adam的一个变体。Adam实际上在梯度更新时使用了一个L2范数进行scale，这可以推广到Lp范数。而当$p \to \infty$时就得到了：

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \tag{103}$$

$$u_t = \lim_{p\to\infty} (v_t)^{1/p} = \max(\beta_2 \cdot u_{t-1}, |g_t|) \tag{104}$$

$$w_{t+1} = w_t - \alpha \frac{1}{1-\beta_1^t} \frac{m_t}{u_t+\epsilon} \tag{105}$$

其他知识

hook技术

参考资料：

- PyTorch 学习笔记（六）：PyTorch hook 和关于 PyTorch backward 过程的理解
- pytorch中的钩子（Hook）有何作用?

PyTorch使用了一种hook方法来捕捉模型在前馈和反馈时的中间数据。由于AD的设计，调用损失的backward方法后各结点的梯度逐个计算完成并释放计算图，所以无法通过模型来获得中间结果的一些数据，所以使用了**钩子**（hook）技术来抓取这些数据保存到一个新的变量中。

PyTorch中有四种钩子：

- torch.tensor.register_hook(self, hook)
  - 用于tensor；
  - 在每次计算完tensor的梯度时都会调用其中的钩子；
  - 不能修改数据，只能获得一个新的变量用于保存新的梯度（通过tensor.grad获取）；
  - 钩子签名必须为：hook(grad) -> Tensor or None。
- torch.nn.Module.register_forward_pre_hook(self, hook)
  - 用于Module；
  - 在每次调用forward函数**前**都会调用其中的钩子，主要用于各类Norm模块/层；
  - 可以修改输入数据；
  - 钩子签名必须为：hook(module, input) -> None or modified input
- torch.nn.Module.register_forward_hook(self, hook)
  - 用于Module；
  - 在每次调用forward函数后，backward之前都会调用其中的钩子；
  - 可以修改输出数据，也可以原址修改输入数据，但是不会影响前馈结果（因为执行在forward之后）；
  - 钩子签名必须为：hook(module, input, output) -> None or modified output
- torch.nn.Module.register_backward_hook(self, hook)
  - 用于Module；
  - 在每次计算完输入数据的梯度后都会调用其中的钩子；
  - 不能修改数据，但是可以返回一个新变量包含其中的梯度数据（通过Module.grad_input获取）；
  - 钩子签名必须为：hook(module, grad_input, grad_output) -> Tensor or None；其中grad_input和grad_output可以是tuple。

我们举个例子来说明各种钩子的作用。首先，定义一个简单的模块，其中包含一个大小为3x1的参数：

In [1]:
```
import torch
import torch.nn as nn
```

In [2]:
```
class TestModule(nn.Module):
    def __init__(self):
        super(TestModule, self).__init__()
        self.w = torch.rand((3,1), requires_grad=True)

    def forward(self, x):
        return torch.matmul(x, self.w)
```

随后我们定义三个钩子函数，分别用于tensor.register_hook，Module.register_forward_hook和Module.register_backward_hook，并且读取相应的数据：

In [3]:
```
grad_list = []
def tensor_hook(grad):
    grad_list.append(grad)

forward_in_list = []
forward_out_list = []
def forward_hook(module, input, output):
    forward_in_list.extend(list(input))
    forward_out_list.extend(list(output))

backward_grad_in_list = []
backward_grad_out_list = []
def backward_hook(module, grad_input, grad_output):
    backward_grad_in_list.extend(list(grad_input))
    backward_grad_out_list.extend(list(grad_output))
```

我们定义运算引入中间变量y：

In [4]:
```
x = torch.rand((1,3), requires_grad=True)
y = x + 2
model = TestModule()
```

先来观察下各个数据：

```
In [5]: x
Out[5]: tensor([[0.1545, 0.0325, 0.8274]], requires_grad=True)

In [6]: y
Out[6]: tensor([[2.1545, 2.0325, 2.8274]], grad_fn=<AddBackward0>)

In [7]: model.w
Out[7]: tensor([[0.7566],
                [0.4152],
                [0.7712]], requires_grad=True)
```

注册钩子：

```
In [8]: y.register_hook(tensor_hook)
        model.register_forward_hook(forward_hook)
        model.register_backward_hook(backward_hook)
Out[8]: <torch.utils.hooks.RemovableHandle at 0x7f945eb3def0>
```

然后我们调用模块，并反传：

```
In [9]: z = model(y)
        z.backward()
```

来看下`y.grad`，发现是`NoneType`：

```
In [11]: type(y.grad)
Out[11]: NoneType
```

但是y的钩子函数捕捉到了数据，放在了`grad_list`这个列表中。各钩子捕捉到的数据：

```
In [12]: grad_list
Out[12]: [tensor([[0.7566, 0.4152, 0.7712]])]

In [13]: forward_in_list
Out[13]: [tensor([[2.1545, 2.0325, 2.8274]], grad_fn=<AddBackward0>)]

In [14]: forward_out_list
Out[14]: [tensor([4.6544], grad_fn=<SelectBackward>)]

In [15]: backward_grad_in_list
Out[15]: [tensor([[0.7566, 0.4152, 0.7712]]), tensor([[2.1545],
                  [2.0325],
                  [2.8274]])]

In [16]: backward_grad_out_list
Out[16]: [tensor([[1.]])]
```

由此可以看到，通过对不同阶段的数据使用钩子，我们可以容易得获得中间变量/模块的数值/梯度等数据，并在其他任务中进行分析和处理。

个人认为钩子函数主要适用于对中间变量、特征图的数值和梯度的提取，这在对抗样本、迁移学习等邻域可能较为常用。而对于`torch.tensor`定义的变量（比如上例中的`x`）和模块参数（上例中的`model.w`），无需使用钩子技术。

资源

内容	网址	备注
自动微分社区	http://www.autodiff.org/	有关自动微分的内容