题目（中英文题目一致）字体为 2 号黑体 (全文除特别声明外，外文统一用 Times New Roman)

作者名 [1)]　作者名 [2),3)]　作者名 [3)](* 字体为 3 号仿宋 * 作者)

[1)](单位全名部门 (系) 全名, 市 (或直辖市) 国家名邮政编码) * 字体为 6 号宋体 * 单位
[2)](单位全名部门 (系) 全名, 市 (或直辖市) 国家名邮政编码)* 中英文单位名称、作者姓名须一致 *
[3)](单位全名部门 (系) 全名, 市 (或直辖市) 国家名邮政编码)

摘　要　* 中文摘要内容置于此处 (英文摘要中要有这些内容)，字体为小 5 号宋体。摘要贡献部分，要有数据支持，不要出现 "... 大大提高"、"... 显著改善" 等描述，正确的描述是 "比…提高 X%"、"在…上改善 X%"。* 摘要

关键词　* 关键词（中文关键字与英文关键字对应且一致，应有 5-7 个关键词)；关键词；关键词；关键词 *

中图法分类号 TP　　DOI **号**: * 投稿时不提供 DOI 号

Title * (**中英文题目一致**) **字体为** 4 **号** Times New Roman, **加粗** * Title

NAME Name-Name[1)] NAME Name[2)] NAME Name-Name[3)] * **字体为** 5 **号** Times new Roman*Name

[1)](Department of ****, University, City ZipCode, China) * **字体为** 6 **号** Times new Roman* Depart.Correspond
[2)](Department of ****, University, City ZipCode)* **中国不写国家名** *
[3)](Department of ****, University, City ZipCode, country)* **外国写国家名** *

Abstract　(500 **英文单词，内容包含中文摘要的内容**). **字体为** Times new Roman, **字号** 5 **号** * Abstract

Do not modify the amount of space before and after the artworks. One- or two-column format artworks are preferred. and Tables, create a new break line and paste the resized artworks where desired. Do not modify the amount of space before and after the artworks. One- or two-column format artworks are preferred. All Schemes, Equations, Figures, and Tables should be mentioned in the text consecutively and numbered with Arabic numerals, and appear below where they are mentioned for the first time in the main text. To insert Schemes, Equations, Figures, and Tables, create a new break line and paste the resized artworks where desired. Do not modify the amount of space before and after the artworks. One- or two-column format artworks are preferred.Do not modify the amount of space before and after the artworks. One- or two-column format artworks are preferred. and Tables, create a new break line and paste the resized artworks where desired. Do not modify the amount of space before and after the artworks. One- or two-column format artworks are preferred. All Schemes, Equations, Figures, and Tables should be mentioned in the text consecutively and numbered with Arabic numerals, and appear below where they are mentioned for the first time in the main text.

Keywords　**中文关键字与英文关键字对应且一致，不要用英文缩写**); key word; key word; key word* * **字体为** 5 **号** Times new Roman * Key words

1 长短时记忆网络原理

1.1 长短时记忆网络的前向计算

前面描述的开关是怎样在算法中实现的呢？这就用到了门（gate）的概念。门实际上就是一层全连接层，它的输入是一个向量，输出是一个 0 到 1 之间的实数向量。假设 W 是门的权重向量，是偏置项，那么门可以表示为：

$$g(\mathrm{x}) = \sigma(W\mathrm{x} + \mathrm{b})$$

门的使用，就是用门的输出向量按元素乘以我们需要控制的那个向量。因为门的输出是 0 到 1 之间的实数向量，那么，当门输出为 0 时，任何向量与之相乘都会得到 0 向量，这就相当于🚪都不能通过；输出为 1 时，任何向量与之相乘都不会有任何改变，这就相当于啥都可以通过。因为 σ (也就是 sigmoid 函数) 的值域是 $(0,1)$，所以门的状态都是半开半闭的。

LSTM 用两个门来控制单元状态 c 的内容，一个是遗忘门（forget gate），它决定了上一时刻的单元状态 c_{t-1} 有多少保留到当前时刻 c_t；另个个是输入门（input gate），它决定了当前时刻网络的输入 x_t 有多少保存到单元状态 c_t。LSTM 用输出门 (output gate) 来控制单元状态 c_t 有多少输出到 LSTM 的当前输出值 h_t。

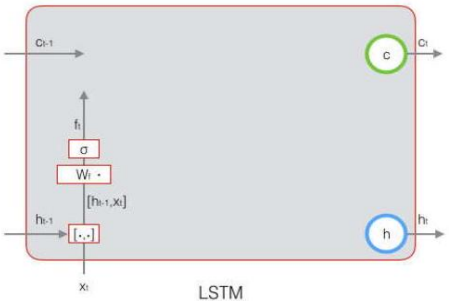
我们先来看一下遗忘门：

$$\mathrm{f}_t = \sigma\left(W_f \cdot [\mathrm{h}_{t-1}, \mathrm{x}_t] + \mathrm{b}_f\right)$$

上式中，W_f 是遗忘门的权重矩阵，$[\mathrm{h}_{t-1}, \mathrm{x}_t]$ 表示把两个向量连接成一个更长的向量，b_f 是遗忘门的偏置项，σ 是 sigmoid 函数。如果输入的维度是 d_x，隐藏层的维度是 d_h，单元状态的维度是 d_c（通常 $d_c = d_h$），则遗忘门的权重矩阵 W_f 维度是 $d_c \times (d_h + d_x)$。事实上，权重矩阵 W_f 都是两个矩阵拼接而成的：个个是 W_{fh}，它对应着输入项 h_{t-1}，其维度为 $d_c \times d_h$；个是 W_{fx}，它对应着输入项 x_t，其维度为 $d_c \times d_x$。W_f 可以写为：

$$[W_f]\begin{bmatrix}\mathrm{h}_{t-1}\\ \mathrm{x}_t\end{bmatrix} = \begin{bmatrix}W_{fh} & W_{fx}\end{bmatrix}\begin{bmatrix}\mathrm{h}_{t-1}\\ \mathrm{x}_t\end{bmatrix} = W_{fh}\mathrm{h}_{t-1} + W_{fx}\mathrm{x}_t$$

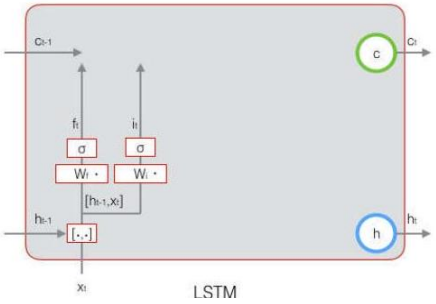
下图显示了遗忘门的计算：



接下来看看输入门：

$$\mathrm{i}_t = \sigma\left(W_i \cdot [\mathrm{h}_{t-1}, \mathrm{x}_t] + \mathrm{b}_i\right)$$

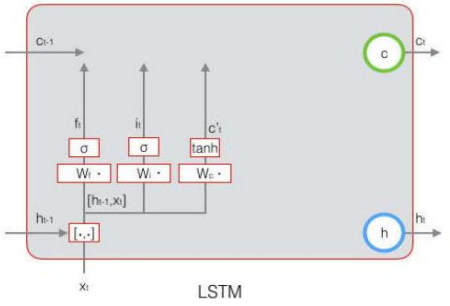
上式中，W_i 是输入门的权重矩阵，b_i 是输入门的偏置项。下图表示了输入门的计算：



接下来，我们计算用于描述当前输入的单元状态 $\boldsymbol{c}_t$，它是根据上一次的输出和本次输入来计算的：

$$\tilde{\mathrm{c}}_t = \tanh\left(W_c \cdot [\mathrm{h}_{t-1}, \mathrm{x}_t] + \mathrm{b}_c\right)$$

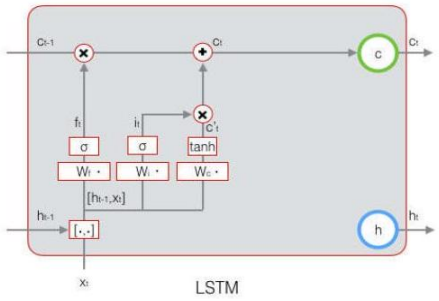
下图是 $\tilde{c}_t$ 的计算：



现在，我们计算当前时刻的单元状态 c_t。它是由上一次的单元状态 c_{t-1} 按元素乘以遗忘门 f_t，再用当前输入的单元状态 $\tilde{c}_t$ 按元素乘以输入 i_t，再将两个积加和产生的：

$$\mathrm{c}_t = f_t \circ \mathrm{c}_{t-1} + i_t \circ \tilde{\mathrm{c}}_t$$

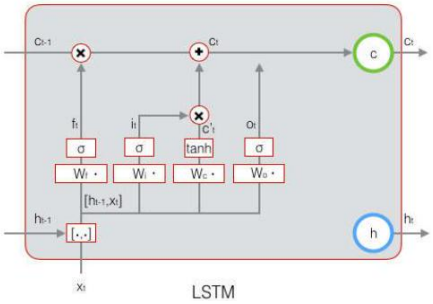
符号 $\circ$ 表示按元素乘。下图是 c_t 的计算：



这样，我们就把 LSTM 关于当前的记忆 $\tilde{c}_t$ 和长期的记忆 c_{t-1} 组合在一起，形成了新的单元状态 c_t 。由于遗忘门的控制，它可以保存很久很久之前的信息，由于输入门的控制，它又可以避免当前无关紧要的内容进入记忆。下面，我们要看看输出门，它控制了长期记忆对当前输出的影响:

$$\mathrm{o}_t = \sigma\left(W_o \cdot [\mathrm{h}_{t-1}, \mathrm{x}_t] + \mathrm{b}_o\right)$$

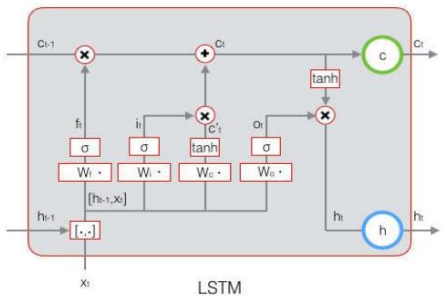
下图表示输出门的计算:



LSTM 最终的输出，是由输出门和单元状态共同确定的:

$$\mathrm{h}_t = \mathrm{o}_t \circ \tanh\left(\mathrm{c}_t\right)$$

下图表示 LSTM 最终输出的计算:



式 1 到式 6 就是 LSTM 前向计算的全部公式。

1.2 长短时记忆网络的训练

熟悉我们这个系列文章的同学都清楚，训练部分往往比前向计算部分复杂多了。LSTM 的前向计算都这么复杂，那么，可想而知，它的训练算法一定是非常非常复杂的。现在只有做几次深呼吸，再一头扎进公式海洋吧。

1.3 LSTM 训练算法框架

LSTM 的训练算法仍然是反向传播算法，对于这个算法，我们已经非常熟悉了。主要有下面三个步骤:

(1) 前向计算每个神经元的输出值，对于 LSTM 来说，即 f_t i_t c_t o_t h_t 五个向量的值。计算方法已经在上一节中描述过了。

(2) 反向计算每个神经元的误差项 δ 值。与循环神经网络一样，LSTM 误差项的反向传播也是包括两个方向: 一个是沿时间的反向传播，即从当前时刻开始，计算每个时刻的误差项；一个是将误差项向上一层传播。

1.4 用于公式和符号的说明

首先，我们对推导中用到的一些公式、符号做一下必要的说明。

接下来的推导中，我们设定 gate 的激活函数为 sigmoid 函数，输出的激活函数为 tanh 函数。他们的导数分别为:

$$\begin{aligned}
\sigma(z) &= y = \frac{1}{1+e^{-z}} \\
\sigma'(z) &= y(1-y) \\
\tanh(z) &= y = \frac{e^z - e^{-z}}{e^z + e^{-z}} \\
\tanh'(z) &= 1 - y^2
\end{aligned}$$

从上面可以看出，sigmoid 和 tanh 函数的导数都是原函数的函数。这样，我们一旦计算原函数的值，就可以用它来计算出导数的值。

LSTM 需要学习的参数共有 8 组，分别是: 遗忘门的权重矩阵 W_f 和偏置项 b_f 、输入门的权重矩阵 W_i 和偏置项 b_i 、输出门的权重矩阵 W_o 和我们解释一下按元素乘 @ 符号。当 O 作用于两个向量时，运算如下:

$$\mathrm{a} \circ \mathrm{b} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \cdots \\ a_n \end{bmatrix} \circ \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \cdots \\ b_n \end{bmatrix} = \begin{bmatrix} a_1 b_1 \\ a_2 b_2 \\ a_3 b_3 \\ \cdots \\ a_n b_n \end{bmatrix}$$

当 作用于一个向量和一个矩阵时，运算如下:

$$\mathrm{a}\circ X=\begin{bmatrix}a_1\\a_2\\a_3\\\cdots\\a_n\end{bmatrix}\circ\begin{bmatrix}x_{11}&x_{12}&x_{13}&\cdots&x_{1n}\\x_{21}&x_{22}&x_{23}&\cdots&x_{2n}\\x_{31}&x_{32}&x_{33}&\cdots&x_{3n}\\&&\cdots&&\\x_{n1}&x_{n2}&x_{n3}&\cdots&x_{nn}\end{bmatrix}$$
$$=\begin{bmatrix}a_1x_{11}&a_1x_{12}&a_1x_{13}&\cdots&a_1x_{1n}\\a_2x_{21}&a_2x_{22}&a_2x_{23}&\cdots&a_2x_{2n}\\a_3x_{31}&a_3x_{32}&a_3x_{33}&\cdots&a_3x_{3n}\\a_nx_{n1}&a_nx_{n2}&a_nx_{n3}&\cdots&a_nx_{nn}\end{bmatrix}$$

当 作用于两个矩阵时，两个矩阵对应位置的元素相乘。按元素乘可以在某些情况下简化矩阵和向量运算。例如，当一个对角矩阵右乘一个矩阵时，相当于用对角矩阵的对角线组成的向量按元素乘那个矩阵:

$$\mathrm{diag[a]}X=\mathrm{a}\circ X$$

当一个行向量右乘一个对角矩阵时，相当于这个行向量按元素乘那个矩阵对角线组成的向量:

$$\mathrm{a}^T\,\mathrm{diag[b]}=\mathrm{a}\circ\mathrm{b}$$

上面这两点，在我们后续推导中会多次用到。

在 t 时刻，LSTM 的输出值为 h_t 。我们定义时刻的误差项 δ_t 为:

$$\delta_t\overset{\text{def}}{=}\frac{\partial E}{\partial \mathrm{h}_t}$$

注意，和前面几篇文章不同，我们这里假设误差项是损失函数对输出值的导数，而不是对加权输入 net t_t^l 的导数。因为 LSTM 有四个加权输入，分别对应 f_t i_t c_t o_t ，我们希望往上一层传递一个误差项而不是四个。但我们仍然需要定义出这四个加权输入，以及他们对应的误差项。

$$\begin{aligned}\mathrm{net}_{f,t}&=W_f\left[\mathrm{h}_{t-1},\mathrm{x}_t\right]+\mathrm{b}_f\\&=W_{fh}\mathrm{h}_{t-1}+W_{fx}\mathrm{x}_t+\mathrm{b}_f\\\mathrm{net}_{i,t}&=W_i\left[\mathrm{h}_{t-1},\mathrm{x}_t\right]+\mathrm{b}_i\\&=W_{ih}\mathrm{h}_{t-1}+W_{ix}\mathrm{x}_t+\mathrm{b}_i\\\mathrm{net}_{\tilde{c},t}&=W_c\left[\mathrm{h}_{t-1},\mathrm{x}_t\right]+\mathrm{b}_c\\&=W_{ch}\mathrm{h}_{t-1}+W_{cx}\mathrm{x}_t+\mathrm{b}_c\\\mathrm{net}_{o,t}&=W_o\left[\mathrm{h}_{t-1},\mathrm{x}_t\right]+\mathrm{b}_o\\&=W_{oh}\mathrm{h}_{t-1}+W_{ox}\mathrm{x}_t+\mathrm{b}_o\end{aligned}$$

$$\delta_{f,t}\overset{\text{def}}{=}\frac{\partial E}{\partial \mathrm{net}_{f,t}}$$
$$\delta_{i,t}\overset{\text{def}}{=}\frac{\partial E}{\partial \mathrm{net}^{i,t}}$$
$$\delta_{\tilde{c},t}\overset{\text{def}}{=}\frac{\partial E}{\partial \mathrm{net}_{\tilde{c},t}}$$
$$\delta_{o,t}\overset{\text{def}}{=}\frac{\partial E}{\partial \mathrm{net}_{o,t}}$$

1.5 误差项沿时间的反向传递

沿时间反向传递误差项，就是要计算出 t-1 时刻的误差项 δ_{t-1} 。

$$\begin{aligned}\delta_{t-1}^T&=\frac{\partial E}{\partial \mathrm{h}_{t-1}}\\&=\frac{\partial E}{\partial \mathrm{h}_t}\frac{\partial \mathrm{h}_t}{\partial \mathrm{h}_{t-1}}\\&=\delta_t^T\frac{\partial \mathrm{h}_t}{\partial \mathrm{h}_{t-1}}\end{aligned}$$

我们知道，$\frac{\partial \mathrm{h}_t}{\partial \mathrm{h}_{t-1}}$ 是一个 Jacobian 矩阵。如果隐藏层 h 的维度是 N 的话，那么它就是一个 $N\times N$ 矩阵。为了求出它，我们列出 h_t 的计算公式，即前面的式 6 和式 4 :

$$\mathrm{h}_t=\mathrm{o}_t\circ\tanh\left(\mathrm{c}_t\right)$$
$$\mathrm{c}_t=\mathrm{f}_t\circ\mathrm{c}_{t-1}+\mathrm{i}_t\circ\tilde{\mathrm{c}}_t$$

显然，o_t f_t i_t $\tilde{\mathrm{c}}_t$ 都是 h_{t-1} 的函数，那么，利用全导数公式可得:

$$\begin{aligned}\delta_t^T\frac{\partial \mathrm{h}_t}{\partial \mathrm{h}_{t-1}}&=\delta_t^T\frac{\partial \mathrm{h}_t}{\partial \mathrm{o}_t}\frac{\partial \mathrm{o}_t}{\partial \mathrm{net}_{o,t}}\frac{\partial \mathrm{net}_{o,t}}{\partial \mathrm{h}_{t-1}}\\&+\delta_t^T\frac{\partial \mathrm{h}_t}{\partial \mathrm{c}_t}\frac{\partial \mathrm{c}_t}{\partial \mathrm{f}_t}\frac{\partial \mathrm{f}_t}{\partial \mathrm{net}_{f,t}}\frac{\partial \mathrm{net}_{f,t}}{\partial \mathrm{h}_{t-1}}\\&+\delta_t^T\frac{\partial \mathrm{h}_t}{\partial \mathrm{c}_t}\frac{\partial \mathrm{c}_t}{\partial \mathrm{i}_t}\frac{\partial \mathrm{i}_t}{\partial \mathrm{net}_{i,t}}\frac{\partial \mathrm{net}_{i,t}}{\partial \mathrm{h}_{t-1}}\\&+\delta_t^T\frac{\partial \mathrm{h}_t}{\partial \mathrm{c}_t}\frac{\partial \mathrm{c}_t}{\partial \tilde{\mathrm{c}}_t}\frac{\partial \tilde{\mathrm{c}}_t}{\partial \mathrm{net}_{\tilde{c},t}}\\&=\delta_{o,t}^T\frac{\partial \mathrm{net}_{o,t}}{\partial \mathrm{h}_{t-1}}+\delta_{f,t}^T\frac{\partial \mathrm{net}_{f,t}}{\partial \mathrm{h}_{t-1}}\\&+\delta_{i,t}^T\frac{\partial \mathrm{net}_{i,t}}{\partial \mathrm{h}_{t-1}}+\delta_{\tilde{c},t}^T\frac{\partial \mathrm{net}_{\tilde{c},t}}{\partial \mathrm{h}_{t-1}}\end{aligned}$$

下面，我们要把式 7 中的每个偏导数都求出来。根据式 6，我们可以求出:

$$\frac{\partial \mathrm{h_t}}{\partial \mathrm{o}_t} = \mathrm{diag}\left[\tanh\left(\mathrm{c}_t\right)\right]$$
$$\frac{\partial \mathrm{h_t}}{\partial \mathrm{c}_t} = \mathrm{diag}\left[\mathrm{o}_t \circ \left(1 - \tanh\left(\mathrm{c}_t\right)^2\right)\right]$$

根据式 4，我们可以求出:

$$\frac{\partial \mathrm{c}_t}{\partial \mathrm{f_t}} = \mathrm{diag}\left[\mathrm{c}_{t-1}\right]$$
$$\frac{\partial \mathrm{c}_t}{\partial \mathrm{i_t}} = \mathrm{diag}\left[\tilde{\mathrm{c}}_t\right]$$
$$\frac{\partial \mathrm{c}_t}{\partial \tilde{\mathrm{c}}_\mathrm{t}} = \mathrm{diag}\left[\mathrm{i}_t\right]$$

因为:

$$\begin{aligned}
\mathrm{o}_t &= \sigma\left(\mathrm{net}_{o,t}\right) \\
\mathrm{net}_{o,t} &= W_{oh}\mathrm{h}_{t-1} + W_{ox}\mathrm{x}_t + \mathrm{b}_o \\
\mathrm{f}_t &= \sigma\left(\mathrm{net}_{f,t}\right) \\
\mathrm{net}_{f,t} &= W_{fh}\mathrm{h}_{t-1} + W_{fx}\mathrm{x}_t + \mathrm{b}_f \\
\mathrm{i}_t &= \sigma\left(\mathrm{net}_{i,t}\right) \\
\mathrm{net}_{i,t} &= W_{ih}\mathrm{h}_{t-1} + W_{ix}\mathrm{x}_t + \mathrm{b}_i \\
\tilde{\mathrm{c}}_t &= \tanh\left(\mathrm{net}_{\tilde{c},t}\right) \\
\mathrm{net}_{\tilde{c},t} &= W_{ch}\mathrm{h}_{t-1} + W_{cx}\mathrm{x}_t + \mathrm{b}_c
\end{aligned}$$

我们很容易得出:

$$\frac{\partial \mathrm{o}_t}{\partial \mathrm{net}_{o,t}} = \mathrm{diag}\left[\mathrm{f}_t \circ (1 - \mathrm{o}_t)\right]$$
$$\frac{\partial \mathrm{net}_{o,t}}{\partial \mathrm{h_{t-1}}} = W_{oh}$$
$$\frac{\partial \mathrm{f_t}}{\partial \mathrm{net}_{f,t}} = \mathrm{diag}\left[\mathrm{f}_t \circ (1 - \mathrm{f}_t)\right]$$
$$\frac{\partial \mathrm{net}_{f,t}}{\partial \mathrm{h}_{t-1}} = W_{fh}$$
$$\frac{\partial \mathrm{i}_t}{\partial \mathrm{net}_{i,t}} = \mathrm{diag}\left[\mathrm{i}_t \circ (1 - \mathrm{i}_t)\right]$$
$$\frac{\partial \mathrm{net}_{i,t}}{\partial \mathrm{h}_{t-1}} = W_{ih}$$
$$\frac{\partial \tilde{\mathrm{c}}_t}{\partial \mathrm{net}_{\tilde{,},t}} = \mathrm{diag}\left[1 - \tilde{\mathrm{c}}_t^2\right]$$
$$\frac{\partial \mathrm{net}_{\tilde{,},t}}{\partial \mathrm{h}_{t-1}} = W_{ch}$$

将上述偏导数带入到式 7，我们得到:

$$\begin{aligned}
\delta_{t-1} &= \delta_{o,t}^T \frac{\partial \mathrm{net}_{o,t}}{\partial \mathrm{h_{t-1}}} + \delta_{f,t}^T \frac{\partial \mathrm{net}_{f,t}}{\partial \mathrm{h_{t-1}}} + \delta_{i,t}^T \frac{\partial \mathrm{net}_{i,t}}{\partial \mathrm{h_{t-1}}} + \delta_{\tilde{c},t}^T \frac{\partial \mathrm{net}_{\tilde{c},t}}{\partial \mathrm{h_{t-1}}} \\
&= \delta_{o,t}^T W_{oh} + \delta_{f,t}^T W_{fh} + \delta_{i,t}^T W_{ih} + \delta_{\tilde{c},t}^T W_{ch} \quad (\text{式8})
\end{aligned}$$

根据 $\delta_{o,t}$ $\delta_{f,t}$ $\delta_{i,t}$ $\delta_{\tilde{c},t}$ 的定义，可知:

$$\delta_{o,t}^T = \delta_t^T \circ \tanh\left(\mathrm{c}_t\right) \circ \mathrm{o}_t \circ (1 - \mathrm{o}_t) \quad (\text{式9})$$
$$\delta_{f,t}^T = \delta_t^T \circ \mathrm{o}_t \circ \left(1 - \tanh\left(\mathrm{c}_t\right)^2\right) \circ \mathrm{c}_{t-1} \circ \mathrm{f}_t \circ (1 - \mathrm{f}_t)$$
$$\delta_{i,t}^T = \delta_t^T \circ \mathrm{o}_t \circ \left(1 - \tanh\left(\mathrm{c}_t\right)^2\right) \circ \tilde{\mathrm{c}}_t \circ \mathrm{i}_t \circ (1 - \mathrm{i}_t)$$
$$\delta_{\tilde{c},t}^T = \delta_t^T \circ \mathrm{o}_t \circ \left(1 - \tanh\left(\mathrm{c}_t\right)^2\right) \circ \mathrm{i}_t \circ \left(1 - \tilde{\mathrm{c}}^2\right)$$

式 8 到式 12 就是将误差沿时间反向传播一个时刻的公式。有了它，我们可以写出将误差项向前传递到任意 k 时刻的公式:

$$\delta_k^T = \prod_{j=k}^{t-1} \delta_{o,j}^T W_{oh} + \delta_{f,j}^T W_{fh} + \delta_{i,j}^T W_{ih} + \delta_{\tilde{c},j}^T W_{ch}$$

1.6 将误差项传递到上一层

我们假设当前为第 I 层，定义 I-1 层的误差项是误差函数对 |-1 层加权输入的导数，即:

$$\delta_t^{l-1} \overset{\text{def}}{=} \frac{\partial E}{\mathrm{net}_t^{l-1}}$$

本次 LSTM 的输入 x_t 由下面的公式计算:

$$\mathrm{x}_t^l = f^{l-1}\left(\mathrm{net}_t^{l-1}\right)$$

上式中，f^{l-1} 表示第 $\mathrm{I}-1$ 层的激活函数。

因为 net $^l_{f,t}$ $\mathrm{net}_{i,t}^l$ $\mathrm{net}_{\tilde{c},t}^l$ $\mathrm{net}_{o,t}^l$ 都是 x_t 的函数，x_t 又是 net_t^{l-1} 的函数，因此，要求出 E 对 net t_t^{l-1} 的导数，就需要使用全导数公式:

$$\begin{aligned}
\frac{\partial E}{\partial \mathrm{net}_t^{l-1}} &= \frac{\partial E}{\partial \mathrm{net}_{\mathrm{f,t}}^1} \frac{\partial \mathrm{net}_{\mathrm{f,t}}^1}{\partial \mathrm{x}_t^l} \frac{\partial \mathrm{x}_t^l}{\partial \mathrm{net}_{\mathrm{t}}^{1-1}} + \frac{\partial E}{\partial \mathrm{net}_{\mathrm{i,t}}^1} \frac{\partial \mathrm{net}_{\mathrm{i,t}}^1}{\partial \mathrm{x}_t^l} \frac{\partial \mathrm{x}_t^l}{\partial \mathrm{net}_{\mathrm{t}}^{1-1}} + \frac{\partial}{\partial \mathrm{n}} \\
&= \delta_{f,t}^T W_{fx} \circ f'\left(\mathrm{net}_t^{l-1}\right) + \delta_{i,t}^T W_{ix} \circ f'\left(\mathrm{net}_t^{l-1}\right) \\
&+ \delta_{\tilde{c},t}^T W_{cx} \circ f'\left(\mathrm{net}_t^{l-1}\right) + \delta_{o,t}^T W_{ox} \circ f'\left(\mathrm{net}_t^{l-1}\right) \\
&= \left(\delta_{f,t}^T W_{fx} + \delta_{i,t}^T W_{ix} + \delta_{\tilde{c},t}^T W_{cx} + \delta_{o,t}^T W_{ox}\right) \circ f'\left(\mathrm{net}_t^{l-1}\right)
\end{aligned}$$

式 14 就是将误差传递到上一层的公式。

1.7 权重梯度的计算

对于 W_{fh} W_{ih} W_{ch} W_{oh} 的权重梯度，我们知道它的梯度是各个时刻梯度之和（证明过程请参考文章零基础入门深度学习 (5) - 循环神经网络），我们首先求出它们在 t 时刻的梯度，然后再求出他们最终的梯度。

我们已经求得了误差项 $\delta_{o,t}$ $\delta_{f,t}$ $\delta_{i,t}$ $\delta_{\tilde{c},t}$ ，很容

易求出 t 时刻的 W_{oh} 的 W_{ih} 的 W_{fh} 的 W_{ch}：

$$\begin{aligned}\frac{\partial E}{\partial W_{oh,t}} &= \frac{\partial E}{\partial \mathrm{net}_{o,t}}\frac{\partial \mathrm{net}_{o,t}}{\partial W_{oh,t}} \\ &= \delta_{o,t}\mathrm{h}_{t-1}^T \\ \frac{\partial E}{\partial W_{fh,t}} &= \frac{\partial E}{\partial \mathrm{net}_{f,t}}\frac{\partial \mathrm{net}_{f,t}}{\partial W_{fh,t}} \\ &= \delta_{f,t}\mathrm{h}_{t-1}^T \\ \frac{\partial E}{\partial W_{ih,t}} &= \frac{\partial E}{\partial \mathrm{net}_{i,t}}\frac{\partial \mathrm{net}_{i,t}}{\partial W_{ih,t}} \\ &= \delta_{i,t}\mathrm{h}_{t-1}^T \\ \frac{\partial E}{\partial W_{ch,t}} &= \frac{\partial E}{\partial \mathrm{net}_{\tilde{c},t}}\frac{\partial \mathrm{net}_{\tilde{c},t}}{\partial W_{ch,t}} \\ &= \delta_{\tilde{c},t}\mathrm{h}_{t-1}^T\end{aligned}$$

将各个时刻的梯度加在一起，就能得到最终的梯度:

$$\begin{aligned}\frac{\partial E}{\partial W_{oh}} &= \sum_{j=1}^{t}\delta_{o,j}\mathrm{h}_{j-1}^T \\ \frac{\partial E}{\partial W_{fh}} &= \sum_{j=1}^{t}\delta_{f,j}\mathrm{h}_{j-1}^T \\ \frac{\partial E}{\partial W_{ih}} &= \sum_{j=1}^{t}\delta_{i,j}\mathrm{h}_{j-1}^T \\ \frac{\partial E}{\partial W_{ch}} &= \sum_{j=1}^{t}\delta_{\tilde{c},j}\mathrm{h}_{j-1}^T\end{aligned}$$

对于偏置项 b_f b_i b_c b_o 的梯度，也是将各个时刻的梯度加在一起。下面是各个时刻的偏置项梯度:

$$\begin{aligned}\frac{\partial E}{\partial \mathrm{b}_{o,t}} &= \frac{\partial E}{\partial \mathrm{net}_{o,t}}\frac{\partial \mathrm{net}_{o,t}}{\partial \mathrm{b}_{o,t}} \\ &= \delta_{o,t} \\ \frac{\partial E}{\partial \mathrm{b}_{f,t}} &= \frac{\partial E}{\partial \mathrm{net}_{f,t}}\frac{\partial \mathrm{net}_{f,t}}{\partial \mathrm{b}_{f,t}} \\ &= \delta_{f,t} \\ \frac{\partial E}{\partial \mathrm{b}_{i,t}} &= \frac{\partial E}{\partial \mathrm{net}_{i,t}}\frac{\partial \mathrm{net}_{i,t}}{\partial \mathrm{b}_{i,t}} \\ &= \delta_{i,t} \\ \frac{\partial E}{\partial \mathrm{b}_{c,t}} &= \frac{\partial E}{\partial \mathrm{net}_{\tilde{c},t}}\frac{\partial \mathrm{net}_{\tilde{c},t}}{\partial \mathrm{b}_{c,t}} \\ &= \delta_{\tilde{c},t}\end{aligned}$$

下面是最终的偏置项梯度，即将各个时刻的偏置项梯度加在一起:

$$\begin{aligned}\frac{\partial E}{\partial \mathrm{b}_o} &= \sum_{j=1}^{t}\delta_{o,j} \\ \frac{\partial E}{\partial \mathrm{b}_i} &= \sum_{j=1}^{t}\delta_{i,j} \\ \frac{\partial E}{\partial \mathrm{b}_f} &= \sum_{j=1}^{t}\delta_{f,j} \\ \frac{\partial E}{\partial \mathrm{b}_c} &= \sum_{j=1}^{t}\delta_{\tilde{c},j}\end{aligned}$$

对于 W_{fx} W_{ix} W_{cx} W_{ox} 的权重梯度，只需要根据相应的误差项直接计算即可:

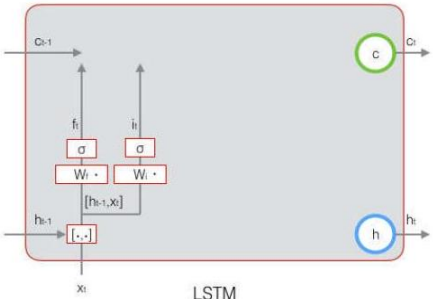
$$\begin{aligned}\frac{\partial E}{\partial W_{ox}} &= \frac{\partial E}{\partial \mathrm{net}_{o,t}}\frac{\partial \mathrm{net}_{o,t}}{\partial W_{ox}} \\ &= \delta_{o,t}\mathrm{x}_t^T \\ \frac{\partial E}{\partial W_{fx}} &= \frac{\partial E}{\partial \mathrm{net}_{f,t}}\frac{\partial \mathrm{net}_{f,t}}{\partial W_{fx}} \\ &= \delta_{f,t}\mathrm{x}_t^T \\ \frac{\partial E}{\partial W_{ix}} &= \frac{\partial E}{\partial \mathrm{net}_{i,t}}\frac{\partial \mathrm{net}_{i,t}}{\partial W_{ix}} \\ &= \delta_{i,t}\mathrm{x}_t^T \\ \frac{\partial E}{\partial W_{cx}} &= \frac{\partial E}{\partial \mathrm{net}_{\tilde{c},t}}\frac{\partial \mathrm{net}_{\tilde{c},t}}{\partial W_{cx}} \\ &= \delta_{\tilde{c},t}\mathrm{x}_t^T\end{aligned}$$

当然，LSTM 存在着相当多的变体，读者可以在互联网上找到很多资料。因为大家已经熟悉了基本 LSTM 的算法，因此理解这些变体比较容易，因此本文就不再赘述了。

2 正弦波预测

为了演示 LSTM 神经网络在预测时间序列中的用途，让我们从我们能想到的最基本的东西开始，那就是时间序列：可靠的正弦波。让我们创建我们需要的数据来模拟这个函数的许多振荡，以便 LSTM 网络进行训练。

代码的数据文件夹中提供的数据包含我们创建的 sinewave.csv 文件，其中包含 5001 个正弦波时间段，幅度和频率为 1（角频率为 6.28）和时间增量为 0.01。绘制时的结果如下所示：



现在我们有了数据，我们实际上想要实现什么？好吧，只是我们希望 LSTM 从我们将提供给它的一组数据窗口大小的数据中学习正弦波，并希望我们可以让 LSTM 预测该系列中接下来的 N 步，它会继续输出正弦波。

我们将首先将数据从 CSV 文件转换并加载到 pandas 数据帧，然后将其用于输出将馈送 LSTM 的 numpy 数组。Keras LSTM 层的工作方式是采用 3 维（N、W、F）的 numpy 数组，其中 N 是训练序列的数量，W 是序列长度，F 是每个序列的特征数。我们选择使用 50 的序列长度（读取窗口大小），这样网络就可以瞥见每个序列的正弦波形状，因此希望能够自学根据之前收到的窗口。

为了加载这些数据，我们在代码中创建了一个 DataLoader 类来为数据加载层提供抽象。您会注意到，在初始化 DataLoader 对象时，会传入文件名，以及一个拆分变量，该变量确定用于训练与测试的数据百分比，以及一个允许选择一列或多列数据的列变量用于单维或多维分析。

附录 X.

* **附录内容置于此处，字体为小 5 号宋体。附录内容包括：详细的定理证明、公式推导、原始数据等** *

表 X　表说明 * 表说明采用黑体 *

* 示例表格 *	* 第 1 行为表头，表头要有内容 *

First A. Author * **计算机学报第 1 作者提供照片电子图片，尺寸为 1 寸。英文作者介绍内容包括：出生年，学位（或目前学历），职称，主要研究领域（与中文作者介绍中的研究方向一致）。* * 字体为小 5 号 Times New Roman***

2.1　二级标题 * 字体为 5 号黑体 * 标题 2

2.1.1　三级标题 * 字体为 5 号宋体 * 标题 3 *

示例图片

(请插入当时做图时的矢量版，如有当时的文件，例如 Visio,origin,matlab, smartdraw,Execl,powerpoint 等各种软件作的图，图字用6号宋体，外文Times new roman，图中文字尽量用翻译成中文)

如插入图为截图，必将原作图文件(如*.vsd,*.opj, *.fig,*.sdr,*.eps,*.emf, *.wmf,*.ps 等后缀名)随修改稿压缩后传过来排版。

图 X　图片说明 * 字体为小 5 号，图片应为黑白图，图中的子图要有子图说明 *

过程 X.　过程名称

* **《计算机学报》的方法过程描述字体为小 5 号宋体，IF、THEN 等伪代码关键词全部用大写字母，变量和函数名称用斜体** *

算法 Y．　算法名称.

输入：… …

输出：… …

* **《计算机学报》的算法描述字体为小 5 号宋体，IF、THEN 等伪代码关键词全部用大写字母，变量和函数名称用斜体** *

致　谢　* 致谢内容.* 致谢

参 考 文 献

Second B. Author * **英文作者介绍内容包括：出生年，学位（或目前学历），职称，主要研究领域（与中文作者介绍中的研究方向一致）。* * 字体为小 5 号** Times New Roman*

Background

* 论文背景介绍为**英文**，字体为小 5 号 Times New Roman 体 *

论文后面为 400 单词左右的英文背景介绍。介绍的内容包括:

本文研究的问题属于哪一个领域的什么问题。该类问题目前国际上解决到什么程度。

本文将问题解决到什么程度。

课题所属的项目。

项目的意义。

本研究群体以往在这个方向上的研究成果。

本文的成果是解决大课题中的哪一部分，如果涉及 863\973 以及其项目、基金、研究计划，注意这些项目的英文名称应书写正确。