(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 107145483 A
(43)申请公布日 2017.09.08

(21)申请号 201710269840.1

(22)申请日 2017.04.24

(71)申请人 北京邮电大学
地址 100876 北京市海淀区西土城路10号

(72)发明人 李思 包祖贻 徐蔚然 高升

(51)Int.Cl.
G06F 17/27(2006.01)
G06N 3/04(2006.01)

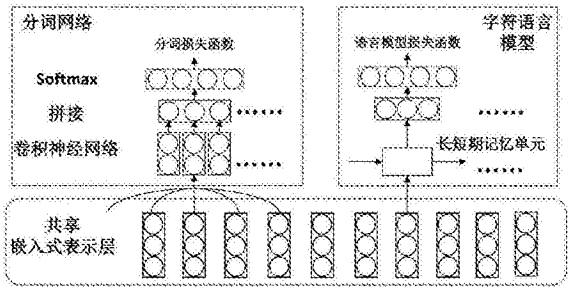
权利要求书2页 说明书5页 附图1页

(54)发明名称

一种基于嵌入式表示的自适应中文分词方法

(57)摘要

本发明实施例公开了一种基于嵌入式表示的自适应中文分词方法。属于信息处理领域。该方法的特征包括：分词网络和字符语言模型共享一个字符的嵌入式表示层。字符的嵌入式表示，一方面通过基于卷积神经网络的分词网络，得到待分词文本的隐多粒度局部特征；再经过一个前向网络层，得到字符的标签概率；最后应用标签推断，得到句子级别上的最优分词结果。另一方面，我们随机抽取未标注的文本，通过一个基于长短期记忆单元(LSTM)循环神经网络的字符语言模型，预测该字符下一个位置的字符，对分词网络进行约束；本发明通过字符语言模型建模中文不同领域文本中的字符共现关系，并通过嵌入式表示将信息传递给分词网络，使得分词的领域迁移能力得到提升，具有很大的实用价值。



CN 107145483 A

1.一种基于嵌入式表示的自适应中文分词方法，其特征在于，所述神经网络包含以下结构和步骤：

训练时：

(1)分词网络和字符语言模型网络共享字符的嵌入式表示层。对输入已标注句子和未标注句子的字符向量参数化：对输入字符进行映射，将离散的字符转化为数值向量，输入的待分词文本即可数值化为各个字符的数值向量连接而成的矩阵；

(2)卷积神经网络提取隐多粒度局部信息：对步骤(1)得到的已标注文本矩阵进行卷积操作，得到文本中各个字符周围的隐多粒度局部特征；

(3)前向神经网络计算各个字符的标签得分：对步骤(2)中得到的隐多粒度局部特征经过一个前向网络得到各个字符的各个标签的概率；

(4)使用标签推断方法得到最优标签序列：对步骤(3)中得到的各个字符的各个标签的概率进行处理，在整个句子层面对各个字符的标签进行推断，得到整个句子上最优的损失函数值和标签序列，即整个句子上最优的分词结果；

(5)长短期记忆单元(LSTM)循环神经网络得到未标注句子各个位置的隐层表示：对步骤(1)中得到的未标注句子的参数表示进行处理，得到句子各个位置的隐层表示；

(6)前向神经网络预测句子下一个字符的概率分布：将步骤(5)中得到的隐层表示送入一个前向神经网络，得到下一个位置字符的概率分布和损失函数值。

(7)组合分词网络损失函数和字符语言模型损失函数，更新网络权值：对步骤(4)中得到的分词网络的损失函数值和步骤(6)字符语言模型的损失函数值进行组合，得到整体的损失函数值，利用误差反向传播算法，更新网络权值。

分词时，仅激活分词网络一侧，执行步骤(1)至步骤(4)即可得到分词结果。

2.如权利要求1所述的方法，其特征在于，所述步骤(1)具体包括：

(1.1)初始化字典向量矩阵以及字符到向量编号的映射索引；

(1.2)对输入文本进行字符切分，通过映射索引将字符映射为向量编号；

(1.3)通过各个字符的向量编号取得字典向量矩阵中各个字符的向量表示；

(1.4)将各个字符向量连接起来，得到输入文本的数值化矩阵。

3.如权利要求1所述方法，其特征在于，所述步骤(2)具体包括：

(2.1)初始化各个卷积核的参数矩阵；

(2.2)按照卷积核的窗口大小，对输入矩阵进行补齐；

(2.3)对补齐后的矩阵，用卷积核进行卷积操作，得到卷积结果；

(2.4)对不同窗口大小的卷积核重复步骤(2.2)和步骤(2.3)，得到各个窗口大小卷积核的卷积结果，即隐多粒度局部特征。

4.如权利要求1所述方法，其特征在于，所述步骤(3)具体包括：

(3.1)初始化前向网络参数；

(3.2)将步骤(2)中得到的输出矩阵中每一个字符对应的信息输入前向神经，得到每一个字符对应各个标签的得分；

(3.3)对每一个字符对应的各个标签的得分输入softmax函数，得到每一个字符各个标签的概率。

5.如权利要求1所述方法，其特征在于，所述步骤(4)具体包括：

(4.1)初始化标签转移矩阵；

(4.2)对步骤(3)中得到的各字符标签概率矩阵补齐开始位置和结束位置；

(4.3)对补齐的标签概率矩阵，根据标签转移矩阵进行维特比译码，得到最优的标签序列。

6.如权利要求1所述方法，其特征在于，所述步骤(5)具体包括：

(5.1)初始化循环神经网络参数；

(5.2)一个前向的循环神经网络单元按照文本正向顺序对步骤(1)的输出矩阵进行处理，得到正向输出矩阵，每个位置的句子隐层表示，即各个字符的上文信息。

7.如权利要求1所述方法，其特征在于，所述步骤(6)具体包括：

(6.1)初始化前向网络参数；

(6.2)将步骤(5)中得到的输出矩阵中每一个字符位置对应的信息输入前向神经，得到每一个字符对应各个标签的得分；

(6.3)对每一个字符对应的各个标签的得分输入softmax函数，得到每一个字符各个标签的概率。

8.如权利要求1所述方法，其特征在于，所述步骤(7)具体包括：

(7.1)将步骤(3)和步骤(6)中得到的分词网络和语言模型网络的损失函数加权相加得到整体网络的损失函数。

(7.2)利用误差反向传播算法，根据整体损失函数更新网络权值。

一种基于嵌入式表示的自适应中文分词方法

技术领域

[0001] 本发明涉及信息处理领域，特别涉及一种基于神经网络中文分词的领域迁移方法。

背景技术

[0002] 中文分词是中文自然语言处理中的基础任务，它的目标是将以中文汉字为组成的序列转换为以中文词语组成的序列。因为中文词语是中文语义表达的基本单元，中文分词是非常重要的基础任务，而且分词系统的性能会直接影响到中文自然语言处理的上层任务，例如，信息检索和机器翻译。

[0003] 在过去的十几年里，中文分词方面有许多研究工作，也取得了很多瞩目的成果。一方面，许多中文分词的标准数据集被建立了起来；另一方面，许多统计学习的分类器被应用到中文分词任务中，目前最普遍的分词方法是把分词任务作为一个有监督的序列标注任务来完成。比较常见的传统分词模型有结构化感知器、条件随机场(CRFs)等。但是这些传统模型都十分依赖人工设计的特征，需要复杂的特征才能取得较好的分词效果。最近，由于神经网络可以自己学习特征以代替复杂的人工设计特征，大大减轻特征工程的负担，许多工作尝试将神经网络应用于中文分词任务。正是由于这些大量的标注数据和不断改进的统计学习模型，中文分词在标准数据集上取得了很好的效果，有些模型在标准数据集上的准确率甚至超过了98％。

[0004] 然而中文分词并不能说是一个已经解决了的任务，由于大量标注的数据主要是新闻语料，这使得在这些数据上训练得到的分词器在例如专利、文学、金融等领域的文本上性能大大下降。这一问题就是著名的领域迁移问题。领域迁移问题，由可得到的资源可进一步细分为两个大类，一个是全监督领域迁移，一个是半监督领域迁移。这两个类别的主要区别在于迁移的目标领域是否有标注数据。全监督领域迁移中，我们有大量的源领域标注数据和少量目标领域标注数据。在半监督领域迁移中，我们有大量源领域标注数据，但是在目标领域我们只能得到无标注的数据。

[0005] 而本发明主要为了解决上述的半监督领域迁移问题，采用了一种基于嵌入式表示的领域迁移方法，利用语言模型建模中文文本字符之间的共现关系，将这一跨领域信息通过嵌入式表示传递给神经网络分词器，得到了较好的领域迁移分词效果。

发明内容

[0006] 为了解决现有的技术问题，本发明提供了一种基于神经网络分词的领域迁移方法。方案如下：

[0007] 训练时，分词网络和语言模型网络同时工作：

[0008] 步骤一，我们将输入的已标注句子和随机抽取的未标注句子的每个字符都映射为字符向量，通过这一步将句子参数化，句子各映射为一个数值矩阵。

[0009] 步骤二，我们使用一个多卷积核的卷积神经网络对参数化的已标注句子进行卷积

操作，不同窗口大小的卷积核从句子中提取到隐多粒度的局部特征。

[0010] 步骤三，将隐多粒度局部特征送入一个前向网络中，得到各个字符序列标注的标签概率。

[0011] 步骤四，在句子层面上，对整个句子中各个字符的标签概率进行维特比解码，得到句子层面的最优分词结果和分词网络的损失函数值。

[0012] 步骤五，未标注的句子送入一个基于长短期记忆单元(LSTM)循环神经网络的字符语言模型。得到各个字符位置的隐层表示。

[0013] 步骤六，将隐层表示送入一个前向网络中，得到各个字符位置的下一个字符的概率分布。得到语言模型网络的损失函数值。

[0014] 步骤七，将分词网络的损失函数值与语言模型网络的损失函数值进行组合，利用反向传播算法回传损失，更新两个网络的参数值。

[0015] 分词时，仅只用分词网络一侧：

[0016] 步骤一，我们将输入句子的每个字符都映射为字符向量，通过这一步将句子参数化，句子映射为一个数值矩阵。

[0017] 步骤二，我们使用一个多卷积核的卷积神经网络对参数化的句子进行卷积操作，不同窗口大小的卷积核从句子中提取到隐多粒度的局部特征。

[0018] 步骤三，将隐多粒度局部特征送入一个前向网络中，得到各个字符序列标注的标签概率。

[0019] 步骤四，在句子层面上，对整个句子中各个字符的标签概率进行维特比解码，得到句子层面的最优分词结果。

附图说明

[0020] 图1是本发明提供的分词领域迁移方法的网络结构图

[0021] 图2为LSTM循环神经网络单元的内部结构图

具体实施方式

[0022] 接下来将对本发明的实施方法作更详细的描述。

[0023] 图1是本发明提供的分词方法的网络结构图，其中包括：

[0024] 训练部分：

[0025] 步骤S1：共享的字符嵌入式表示层，将输入的已标注句子和随机抽取的未标注句子字符向量参数化；

[0026] 步骤S2：卷积神经网络对已标注句子提取隐多粒度局部信息；

[0027] 步骤S3：前向神经网络计算各个字符的标签得分；

[0028] 步骤S4：使用标签推断方法得到最优标签序列和损失函数值；

[0029] 步骤S5：未标注的句子送入一个基于长短期记忆单元(LSTM)循环神经网络的字符语言模型，得到各个字符位置的隐层表示；

[0030] 步骤S6：将隐层表示送入一个前向网络中，得到各个字符位置的下一个字符的概率分布，得到语言模型网络的损失函数值；

[0031] 步骤S7：将分词网络的损失函数值与语言模型网络的损失函数值进行组合，利用

反向传播算法回传损失，更新两个网络的参数值；

[0032] 分词部分，仅使用分词网络部分：

[0033] 步骤S1：共享的字符嵌入式表示层，将输入句子字符向量参数化；

[0034] 步骤S2：卷积神经网络对已标注句子提取隐多粒度局部信息；

[0035] 步骤S3：前向神经网络计算各个字符的标签得分；

[0036] 步骤S4：使用标签推断方法得到最优标签序列；

[0037] 下面将对每个步骤进行具体的说明：

[0038] 步骤S1：向量参数化，为了克服传统one-hot表示法所带来的稀疏性和无关性的问题，本发明首先将句子中的各个字符参数化，通过一个映射字典，将字符映射为不稀疏的向量表示。假设中文汉字一共有C个字符，那么整个映射字典可以表示为一个C*d维的数值矩阵，其中每一个行是一个字符的数值表示，一个d维的数值向量。那么一个句子，就可以表示为句子中每一个字符都映射为向量后组成的数值矩阵。

[0039] $$\mathbf{x}=x_1 \oplus x_2 \oplus \ldots \oplus x_n$$

[0040] 其中x为句子的矩阵表示，x_i为句子中第i个字符映射后的向量，⊕表示向量的连接。

[0041] 在这一步骤中，借鉴去噪自动编码器的思想，本发明引入了dropout的机制，在训练网络时，随机将一部分参数置零，使得参数训练更具有鲁棒性，训练过程更为平滑。

[0042] 步骤S2：使用卷积神经网络层提取隐多粒度局部信息。卷积神经网络擅长于局部特征的提取，并已经被广泛用于中文自然语言处理任务中，如：情感分类、文档分类。不同的卷积核卷积句子，得到不同的局部特征。卷积神经网络提取到的局部特征比传统使用的uni-gram、bi-gram有更好的表现。所以本发明中将多卷积核的卷积神经网络引入中文分词中，用于提取更好的局部特征。

[0043] 对于文本处理中的卷积神经网络而言，一个窗口为w的卷积核可以表示为一个w*d维的矩阵，其中d是文本参数化后的向量维度。则卷积核对窗口内的w个向量的卷积操作，可以表示为：

[0044] $$c_i=f(\mathbf{m}\otimes\mathbf{x}_i+b)$$

[0045] 其中c为提取到的局部特征，⊗表示卷积操作，b是一个偏置项，f是一个非线性函数，例如sigmoid函数、ReLu函数。由于ReLu函数更适合用于深度神经网络中，所以本发明中选择使用的是ReLu函数。

[0046] 而且中文词语的成词规律有很多种，仅仅用一个特征是不能表示的，所以我们对不同的窗口都引入了多个卷积核。假设我们对窗口w引入n个卷积核，则在句子中一个字符周围窗口为w个字符提取到的局部特征就可以表示为各个卷积核卷积提取到的特征的组合。

[0047] $$\mathbf{c}=c_1 \otimes c_2 \otimes \ldots \otimes c_n$$

[0048] 其中c为句子中一个字符周围提取到的特征向量，c_i表示一个卷积核提取到的局部特征。

[0049] 步骤S3：使用前向神经网络计算各个字符的标签得分。在步骤S2中，卷积神经网络得到了隐多粒度局部特征，这一步中的前向网络就是利用之前提取到的局部信息对序列进行标注生成标签概率。以BIES四标签体系为例，则输出标签共有4个，分别表示字符是一个

词语的开头、中间、结尾和当前字是一个单字词语。这个前向网络是一个输入为卷积神经网络输出维度，输出维度是4的全连接网络。前向网络的输入是步骤S2中得到的输出向量，输入是BIES四标签的得分，最后使用softmax函数对输出的标签得分进行概率化，得到四个标签对应的概率。在这一层中，本发明还使用了dropout策略，提升网络的整体性能，防止过拟合。

[0050] 步骤S4：使用标签推断方法得到最优标签序列。本发明将中文分词作为一个序列标注的问题，其中的标注标签并不是相互无关的，以BIES四标签体系为例，B表示字符是一个词语的开始，I表示字符在一个词语的内部，E表示字符是一个词语的结尾，S表示字符是一个单字词语。存在明确的约束关系，标签B之后符合约束的只能是标签I或者E，标签E后面符合约束的只能是标签B或者S，标签I之后符合约束的只能是标签E，标签S之后符合约束的只能是标签B或者S。这些约束关系表明标注标签之间有很强的依赖关系。为了建模这种关系，本发明中加入了标签的跳转得分。同时为了从各个字符的标签概率分布得到句子的最优标签序列。本发明使用标签推断来计算得到整个句子层面上的最优标签路径。路径的得分有两部分组成，一个是标签的跳转得分，一个是标签本身的概率得分。假设标签转移矩阵是A，其中的第i行第j列的元素表示从标签i跳转到标签j的得分。则一个句子上某一个标签路径的得分为：

[0051] $$Score(\mathbf{y}) = \sum_{t=1}^{n} (A_{y_{t-1}y_t} + s(y_t))$$

[0052] 其中$s(y_t)$为该标签本身的概率得分，n为句子长度。本发明使用维特比算法计算得到最优标签路径。网络用最大化对数似然函数进行训练：

[0053] $$\log p(\mathbf{y}|\theta) = score(\mathbf{y}) - \log \sum_{\forall \mathbf{y}' \in \mathbf{Y}(s)} \exp(score(\mathbf{y}'))$$

[0054] 其中Y(s)表示输入句子s的所有可能标签序列。

[0055] 步骤S5：将步骤S1中参数化的未标注句子送入一个基于长短期记忆单元(LSTM)循环神经网络的字符语言模型，得到各个字符位置的隐层表示。循环神经网络擅长于抽取长距离的依赖关系，也被广泛用于自然语言处理各个任务。但是传统循环神经网络由于结构比较简单，很容易出现梯度爆炸和梯度弥散的问题。梯度弥散会使得网络训练变得非常缓慢，梯度爆炸会使得训练变得困难，甚至导致网络发散。而LSTM(长短期记忆)单元通过使用类似门电路的方式控制记忆单元的遗忘和更新，使得循环神经网络能够更有效地学习到长距离的依赖关系。

[0056] 图2给出了一种LSTM单元的单元结构，一个LSTM单元在坐标点t可以描述为：

[0057] $i_t = \sigma(W_i \cdot x_t + U_i \cdot h_{t-1} + b_i)$

[0058] $f_t = \sigma(W_f \cdot x_t + U_f \cdot h_{t-1} + b_f)$

[0059] $$\tilde{C}_t = \tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c)$$

[0060] $$C_t = i_t \odot \tilde{C}_t + f_t \odot C_{t-1}$$

[0061] $o_t = \sigma(W_o \cdot x_t + U_o \cdot h_{t-1} + b_o)$

[0062] $h_t = o_t \odot \tanh(C_t)$

[0063] 其中x是输入，C是记忆单元状态，i、f、o分别是输入门，遗忘门和输出门，σ和tanh

是逻辑斯蒂克函数和双曲正切函数。⊙是数值对位相乘。W、U和b是权重矩阵和偏置项。$\tilde{C}$是计算出来的候选记忆单元状态。记忆单元状态C在输入门、遗忘门的控制下，从候选记忆单元状态和前一时刻的记忆单元状态更新得到。而输出门则控制记忆单元状态的输出。循环神经网络将步骤S1输出的未标注句子的参数表示为输入。字符语言模型神经网络只有一个正向的网络单元，输出各个字符位置的隐层表示。

[0064] 步骤S6：语言模型预测下一个字符。将步骤S5中得到的隐层表示送入一个前向网络中，使用softmax函数得到各个字符位置的下一个字符的概率分布，语言模型使用交叉熵的损失函数，计算得到语言模型网络的损失函数值。

[0065] 步骤S7：损失函数组合，更新参数。将步骤S4中得到的分词网络损失函数值与步骤S6中得到的语言模型网络的损失函数值进行组合：

[0066] $Loss = Loss_{seg} + a \cdot Loss_{lm}$

[0067] 其中$Loss_{seg}$为分词网络损失函数，$Loss_{lm}$为语言模型网络损失函数，α为一个平衡两个网络损失的超参数。最后利用反向传播算法回传损失，更新两个网络的参数值。

[0068] 分词时，仅激活分词网络一侧，执行步骤S1、S2、S3、S4即可得到分词结果。

[0069] 以上结合附图对所提出的一种基于隐多粒度局部特征的中文分词方法及各模块的具体实施方式进行了阐述。通过以上实施方式的描述，所属领域的一般技术人员可以清楚的了解到本发明可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件实现，但前者是更佳的实施方式。基于这样的理解，本发明的技术方案本质上或者说对现有技术做出贡献的部分可以以计算机软件产品的形式体现，该软件产品存储在一个存储介质中，包括若干指令用以使得一台或多台计算机设备执行本发明各个实施例所述的方法。

[0070] 依据本发明的思想，在具体实施方式及应用范围上均会有改变之处。综上所述，本说明书内容不应理解为对本发明的限制。

[0071] 以上所述的本发明实施方式，并不构成对发明保护范围的限定。任何在本发明的精神和原则之内所作的修改、等同替换和改进等，均应包含在本发明的保护范围之内。

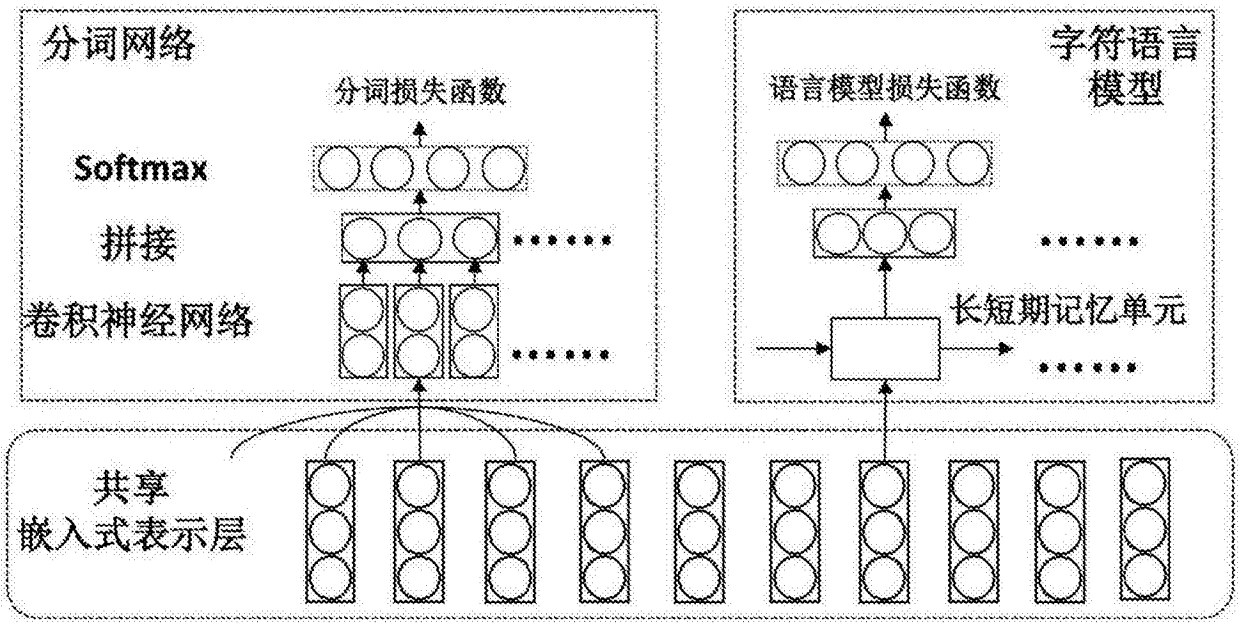

分词网络
分词损失函数
Softmax
拼接
卷积神经网络
字符语言模型
语言模型损失函数
长短期记忆单元
共享
嵌入式表示层

图1

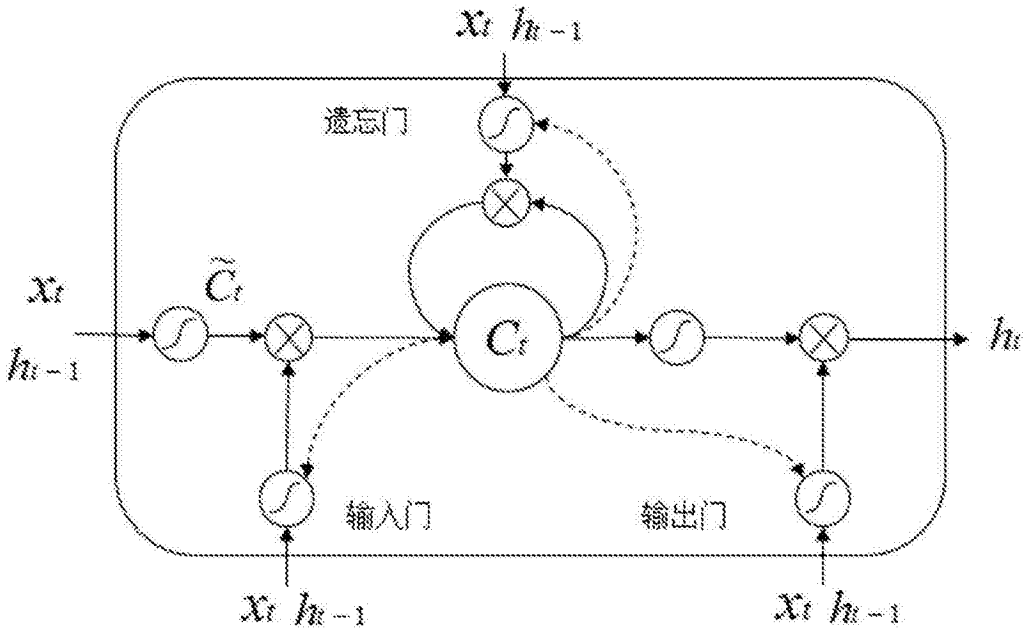

遗忘门
输入门
输出门

图2