(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 107145940 B
(45)授权公告日 2021.02.12

(21)申请号 201611226122.8

(22)申请日 2016.12.27

(65)同一申请的已公布的文献号
申请公布号 CN 107145940 A

(43)申请公布日 2017.09.08

(30)优先权数据
62/301,734 2016.03.01 US
15/172,457 2016.06.03 US

(73)专利权人 谷歌有限责任公司
地址 美国加利福尼亚州

(72)发明人 塔拉·N·赛纳特
维卡斯·辛德瓦尼

(74)专利代理机构 中原信达知识产权代理有限责任公司 11219
代理人 周亚荣 安翔

(51)Int.Cl.
G06N 3/08(2006.01)
G06N 3/04(2006.01)

(56)对比文件
US 6701236 B2,2004.03.02
US 2004/0199482 A1,2004.10.07
US 2016/035344 A1,2016.02.04
CN 105184369 A,2015.12.23
Vikas Sindhwani etc..“Structured Transforms for Small-Footprint Deep Learning”.《arXiv:1510.01722v1[stat.ML]》.2015,

审查员 李玥

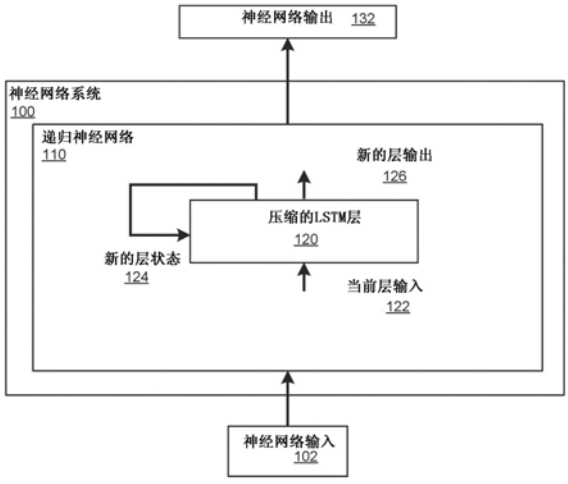
权利要求书3页 说明书8页 附图5页

(54)发明名称

压缩的递归神经网络模型

(57)摘要

本发明公开了压缩的递归神经网络模型。本发明提供用于利用压缩选通函数来实施长短期记忆层的方法、系统、和装置，包括在计算机存储介质上编码的计算机程序。系统之一包括第一LSTM层，该第一LSTM层具有门，门被配置成针对多个时间步长中的每个时间步长，通过使门输入矢量乘以门参数矩阵来生成相应的中间门输出矢量。门中的至少一个门的门参数矩阵是结构化矩阵或者由压缩的参数矩阵和投影矩阵定义。通过将压缩的LSTM层包括在递归神经网络中，递归神经网络被配置成能够更有效地处理数据并且使用更少的数据存储。具有压缩的LSTM层的递归神经网络可以被有效地训练以实现可比得上全尺寸的(例如，未压缩的)递归神经网络的误字率。



CN 107145940 B

1.一种神经网络系统,所述神经网络系统被配置为接收图像数据、互联网资源数据、文件数据、个性化用户推荐数据、文本数据或说出的语句数据作为输入,并且生成对应的图像数据、互联网资源数据、文件数据、个性化用户推荐数据、文本数据或说出的语句数据作为输出,所述神经网络系统包括:

递归神经网络,所述递归神经网络由一个或多个计算机实现,其中,所述递归神经网络被配置成在多个时间步长中的每个时间步长处接收相应的神经网络输入,并且生成在所述多个时间步长中的每个时间步长处的相应的神经网络输出,所述相应的神经网络输入是图像数据、互联网资源数据、文件数据、个性化用户推荐数据、文本数据或说出的语句数据,并且所述相应的神经网络输出是图像属于特定对象类别的可能性分数、互联网资源是关于特定话题的可能性分数、文件是关于特定话题的可能性分数、个性化用户推荐由用户响应的可能性分数、目标语言的文本段是源语言的文本段的适当翻译的可能性分数、或文本段是说出的语句的正确转录的可能性分数,并且其中,所述递归神经网络包括:

第一长短期记忆(LSTM)层,其中,所述第一LSTM层被配置成,针对所述多个时间步长中的每个时间步长,通过将多个门应用于当前层输入、当前层状态和当前层输出来生成新的层状态和新的层输出,所述多个门中的每个门被配置成,针对所述多个时间步长中的每个时间步长,通过使门输入矢量乘以门参数矩阵来生成相应的中间门输出矢量,并且

其中,所述多个门中的至少一个门的所述门参数矩阵是类托普利兹结构化矩阵。

2.根据权利要求1所述的神经网络系统,其中,所述递归神经网络包括第二LSTM层,其中,所述第二LSTM层被配置成,针对所述多个时间步长中的每个时间步长,通过将第二多个门应用于第二当前层输入、第二当前层状态和第二当前层输出来生成第二新的层状态和第二新的层输出,所述第二多个门中的每个门被配置成,针对所述多个时间步长中的每个时间步长,通过使第二门输入矢量乘以第二门参数矩阵来生成相应的第二中间门输出矢量,并且

其中,所述第二多个门中的至少一个门的所述门参数矩阵由压缩的参数矩阵和投影矩阵定义。

3.根据权利要求2所述的神经网络系统,其中,所述第一LSTM层和所述第二LSTM层中的每一个是在层的有序堆叠中的多个LSTM层中的一个。

4.根据权利要求3所述的神经网络系统,其中,所述第一LSTM层在所述堆叠中比所述第二LSTM层低。

5.根据权利要求1-4中任一项所述的神经网络系统,其中,所述多个门中的每个门被配置成,针对所述多个时间步长中的每个时间步长,将相应的选通函数应用于所述相应的中间门输出矢量的每个分量,以生成相应的最终门输出矢量。

6.根据权利要求1-4中任一项所述的神经网络系统,其中,所述神经网络是声学模型。

7.根据权利要求1-4中任一项所述的神经网络系统,其中,所述神经网络是语音识别模型。

8.根据权利要求1-4中任一项所述的神经网络系统,其中,所述神经网络被压缩了所述神经网络的未压缩版本的至少75%。

9.根据权利要求1-4中任一项所述的神经网络系统,其中,所述神经网络的误字率在所述神经网络的未压缩版本的误字率的0.3%内。

10.一种神经网络系统,所述神经网络系统被配置为接收图像数据、互联网资源数据、文件数据、个性化用户推荐数据、文本数据或说出的语句数据作为输入,并且生成对应的图像数据、互联网资源数据、文件数据、个性化用户推荐数据、文本数据或说出的语句数据作为输出,所述神经网络系统包括:

递归神经网络,所述递归神经网络由一个或多个计算机实现,其中,所述递归神经网络被配置成在多个时间步长中的每个时间步长处接收相应的神经网络输入,并且生成在所述多个时间步长中的每个时间步长处的相应的神经网络输出,所述相应的神经网络输入是图像数据、互联网资源数据、文件数据、个性化用户推荐数据、文本数据或说出的语句数据,并且所述相应的神经网络输出是图像属于特定对象类别的可能性分数、互联网资源是关于特定话题的可能性分数、文件是关于特定话题的可能性分数、个性化用户推荐由用户响应的可能性分数、目标语言的文本段是源语言的文本段的适当翻译的可能性分数、或文本段是说出的语句的正确转录的可能性分数,并且其中,所述递归神经网络包括:

第一长短期记忆(LSTM)层,其中,所述第一LSTM层被配置成,针对所述多个时间步长中的每个时间步长,通过将多个门应用于当前层输入、当前层状态和当前层输出来生成新的层状态和新的层输出,所述多个门中的每个门被配置成,针对所述多个时间步长中的每个时间步长,通过使门输入矢量乘以门参数矩阵来生成相应的中间门输出矢量,并且

其中,所述多个门中的至少一个门的所述门参数矩阵由压缩的参数矩阵和投影矩阵定义。

11.根据权利要求10所述的神经网络系统,其中,所述多个门中的每个门被配置成,针对所述多个时间步长中的每个时间步长,将相应的选通函数应用于所述相应的中间门输出矢量的每个分量,以生成相应的最终门输出矢量。

12.根据权利要求10所述的神经网络系统,其中,所述神经网络是声学模型。

13.根据权利要求10所述的神经网络系统,其中,所述神经网络是语音识别模型。

14.根据权利要求10所述的神经网络系统,其中,所述神经网络被压缩了所述神经网络的未压缩版本的至少75%。

15.根据权利要求10所述的神经网络系统,其中,所述神经网络的误字率在所述神经网络的未压缩版本的误字率的0.3%内。

16.编码有计算机程序产品的一个或多个非瞬时计算机存储介质,所述计算机程序产品包括指令,所述指令在由神经网络系统的一个或多个计算机执行时使所述一个或多个计算机执行实现递归神经网络的操作,所述神经网络系统被配置为接收图像数据、互联网资源数据、文件数据、个性化用户推荐数据、文本数据或说出的语句数据作为输入,并且生成对应的图像数据、互联网资源数据、文件数据、个性化用户推荐数据、文本数据或说出的语句数据作为输出,所述递归神经网络被配置成在多个时间步长中的每个时间步长处接收相应的神经网络输入,并且生成在所述多个时间步长中的每个时间步长处的相应的神经网络输出,所述相应的神经网络输入是图像数据、互联网资源数据、文件数据、个性化用户推荐数据、文本数据或说出的语句数据,并且所述相应的神经网络输出是图像属于特定对象类别的可能性分数、互联网资源是关于特定话题的可能性分数、文件是关于特定话题的可能性分数、个性化用户推荐由用户响应的可能性分数、目标语言的文本段是源语言的文本段的适当翻译的可能性分数、或文本段是说出的语句的正确转录的可能性分数,并且其中,所

述递归神经网络包括：

第一长短期记忆(LSTM)层，其中，所述第一LSTM层被配置成，针对所述多个时间步长中的每个时间步长，通过将多个门应用于当前层输入、当前层状态和当前层输出来生成新的层状态和新的层输出，所述多个门中的每个门被配置成，针对所述多个时间步长中的每个时间步长，通过使门输入矢量乘以门参数矩阵来生成相应的中间门输出矢量，并且

其中，所述多个门中的至少一个门的所述门参数矩阵是类托普利兹结构化矩阵。

17.编码有计算机程序产品的一个或多个非瞬时计算机存储介质，所述计算机程序产品包括指令，所述指令在由神经网络系统的一个或多个计算机执行时使所述一个或多个计算机执行实现递归神经网络的操作，所述神经网络系统被配置为接收图像数据、互联网资源数据、文件数据、个性化用户推荐数据、文本数据或说出的语句数据作为输入，并且生成对应的图像数据、互联网资源数据、文件数据、个性化用户推荐数据、文本数据或说出的语句数据作为输出，所述递归神经网络被配置成在多个时间步长中的每个时间步长处接收相应的神经网络输入，并且生成在所述多个时间步长中的每个时间步长处的相应的神经网络输出，所述相应的神经网络输入是图像数据、互联网资源数据、文件数据、个性化用户推荐数据、文本数据或说出的语句数据，并且所述相应的神经网络输出是图像属于特定对象类别的可能性分数、互联网资源是关于特定话题的可能性分数、文件是关于特定话题的可能性分数、个性化用户推荐由用户响应的可能性分数、目标语言的文本段是源语言的文本段的适当翻译的可能性分数、或文本段是说出的语句的正确转录的可能性分数，并且其中，所述递归神经网络包括：

第一长短期记忆(LSTM)层，其中，所述第一LSTM层被配置成，针对所述多个时间步长中的每个时间步长，通过将多个门应用于当前层输入、当前层状态和当前层输出来生成新的层状态和新的层输出，所述多个门中的每个门被配置成，针对所述多个时间步长中的每个时间步长，通过使门输入矢量乘以门参数矩阵来生成相应的中间门输出矢量，并且

其中，所述多个门中的至少一个门的所述门参数矩阵由压缩的参数矩阵和投影矩阵定义。

压缩的递归神经网络模型

技术领域

[0001] 本申请涉及压缩的递归神经网络模型。

背景技术

[0002] 本说明书涉及神经网络架构和压缩的神经网络。

[0003] 神经网络是采用一个或多个层的非线性单元来针对所接收的输入预测输出的机器学习模型。除了输出层之外，一些神经网络还包括一个或多个隐藏层(hidden layer)。每个隐藏层的输出用作对网络中的下一层的输入，即，下一隐藏层或者输出层。网络的每一层根据相应的参数集合的当前值来从所接收的输入生成输出。例如针对时间序列问题或者序列到序列学习而设计的那些神经网络(递归神经网络(RNN))的一些神经网络包含递归环路，该递归环路允许存储器以隐藏状态变量的形式保留在数据输入之间的层内。RNN的变型，长短期记忆(LSTM)神经网络，包括用于控制在数据输入之间的数据持久性的每个层内多个门(gate)。一些神经网络(例如，针对时间序列问题或者序列到序列学习而设计的那些神经网络)包含递归环路，该递归环路允许存储器以隐藏状态变量的形式保留在数据输入之间的层内。

发明内容

[0004] 本说明书描述了涉及递归神经网络架构的技术。一般而言，递归神经网络包括被压缩的至少一个长短期记忆(LSTM)层。LSTM层具有至少一个门，该至少一个门具有压缩的参数矩阵。可以通过用类托普利兹结构化矩阵代替在LSTM层中的门参数矩阵中的一个或多个，或者通过用压缩的参数矩阵和投影矩阵重新定义门参数矩阵，来对LSTM层进行压缩。可选地，可以通过用类托普利兹结构化矩阵代替在LSTM层中的门参数矩阵中的一个来对一个LSTM层进行压缩，并且可以通过用压缩的参数矩阵和投影矩阵，重新定义门参数矩阵，来代替在另一LSTM层中的门参数矩阵中的一个，来对另一LSTM层进行压缩。

[0005] 对于待配置为执行特定操作或者动作的一个或多个计算机的系统，意味着该系统在其上安装有在运行时使该系统执行该操作或者动作的软件、固件、硬件、或者其组合。对于待配置为执行特定操作或者动作的一个或多个计算机程序，意味着该一个或多个程序包括指令，该指令在由数据处理装置执行时使该装置执行该操作或者动作。

[0006] 可以将本说明书中描述的主题实现为具体实施例，从而实现以下优点中的一个或多个。可以通过将压缩的LSTM层包括在递归神经网络中来提高递归神经网络的性能。具体地，通过将压缩的LSTM层包括在递归神经网络中，递归神经网络被配置为能够更有效地处理数据并且使用更少的数据存储。具有压缩的LSTM层的递归神经网络可以被有效地训练为实现可比得上全尺寸的(例如，未压缩的)递归神经网络的误字率。

[0007] 在附图和以下描述中陈述了本说明书中描述的主题的一个或多个实施例的细节。本主题的其它特征、方面和优点通过说明书、附图和权利要求书将变得显而易见。

附图说明

[0008] 图1示出了示例神经网络系统。

[0009] 图2A和图2B示出了示例性结构化矩阵。

[0010] 图3是用于对当前层输入进行处理以生成下一层输出的示例性过程的流程图。

[0011] 图4是用于将门应用于门输入矢量以生成门输出矢量的示例性过程的流程图。

[0012] 图5是用于对包括饱和LSTM层的递归神经网络进行训练的示例性过程的流程图。

[0013] 在各个附图中，相同的附图标记和标志指示相同的元件。

具体实施方式

[0014] 图1示出了示例性神经网络系统100。该神经网络系统100是实现为在一个或多个位置中的一个或多个计算机上的计算机程序的系统的示例，其中实现下面描述的系统、组件和技术。

[0015] 神经网络系统100是一种机器学习系统，该机器学习系统在多个时间步长中的每个时间步长处接收相应的神经网络输入，并且生成在时间步长中的每个时间步长处的相应的神经网络输出。即，在多个时间步长中的每个时间步长处，神经网络系统100接收神经网络输入，并且对神经网络输入进行处理以生成神经网络输出。例如，在给定时间步长处，神经网络系统100可以接收神经网络输入102，并且生成神经网络输出132。

[0016] 神经网络系统100可以将所生成的神经网络输出存储在输出数据存储库中，或者提供神经网络输出以用于一些其它直接目的。

[0017] 神经网络系统100可以被配置为接收任何种类的数字数据输入，并且基于该输入来生成任何种类的分数或者分类输出。

[0018] 例如，如果对神经网络系统100的输入是图像或者已经从图像提取到的特征，则由神经网络系统100针对给定图像生成的输出可以是对象类别集合中的每一个的分数，其中，每个分数表示图像包含属于该类别的对象的图像的估计的可能性。

[0019] 作为另一示例，如果对神经网络系统100的输入是互联网资源(例如，web页面)、文档、或者文档的一部分、或者从互联网资源、文档、或者文档的一部分提取到的特征，则由神经网络系统100针对给定互联网资源、文档、或者文档的一部分所生成的输出可以是话题集合中的每一个的分数，其中，每个分数表示互联网资源、文档、或者文档的一部分与该话题有关的估计的可能性。

[0020] 作为另一示例，如果对神经网络系统100的输入是对用户的个性化推荐的特征，例如，表征推荐的上下文的特征，例如，表征用户所采取的先前动作的特征，则由神经网络系统100生成的输出可以是内容项集合中的每一个的分数，其中，每个分数表示用户将积极响应被推荐内容项的估计的可能性。在这些示例中的一些中，神经网络系统100是将内容推荐提供给用户的增强式学习系统的一部分。

[0021] 作为另一示例，如果对神经网络系统100的输入是一种语言的文本，则由神经网络系统100生成的输出可以是另一语言的文本段集合中的每个文本段的分数，其中，每个分数表示该另一语言的文本段是输入文本成为该另一语言的适当翻译的估计的可能性。

[0022] 作为另一示例，如果对神经网络系统100的输入是说出的语句的特征，则由神经网络系统100生成的输出可以是文本段集合中的每个文本段的分数，每个分数表示该文本段

是对语句的正确转录的估计的可能性。

[0023] 作为另一示例，如果对神经网络系统100的输入是图像，则由神经网络系统100生成的输出可以是文本段集合中的每个文本段的分数，每个分数表示该文本段是存在于输入图像中的文本的估计的可能性。

[0024] 具体地，神经网络系统100包括递归神经网络110，该递归神经网络110进而包括压缩的长短期记忆(LSTM)层120。递归神经网络110被配置成，在时间步长中的每个时间步长处接收神经网络输入，并且对该神经网络输入进行处理以生成在该时间步长处的神经网络输出。

[0025] 除了压缩的LSTM层120之外，递归神经网络110还可以包括一个或多个其它组件，例如，其它压缩的LSTM层、传统LSTM层、其它递归神经网络层、其它非递归神经网络层等。

[0026] 例如，递归神经网络100可以是深度LSTM网络，该深度LSTM网络包括输入层、彼此上下布置在有序堆叠中的包括压缩的LSTM层120的多个LSTM层、以及输出层，该输出层在每个时间步长处接收来自在该堆叠中的最高LSTM层的层输出，并且可选地接收来自在该堆叠中的其它LSTM层的层输出，并且对该层输出进行处理以生成在该时间步长处的神经网络输出132。

[0027] 压缩的LSTM层120被配置成，在时间步长中的每个时间步长处接收当前层输入122，并且对当前层输入122、当前层状态和当前层输出进行处理，以生成新的层输出126并且对当前层状态进行更新以生成新的层状态124。

[0028] 根据递归神经网络110的配置，当前层输入122可以是神经网络输入102或者由递归神经网络110的不同组件生成的输出。

[0029] 另外，针对在第一步长之后的每个时间步长，当前层状态是在前一个时间步长处生成的新的层状态，并且当前层输出是来自前一个时间步长的新的层输出。针对第一时间步长，当前层状态可以是预定初始层状态，并且当前层输出可以是预定初始层输出。

[0030] 根据递归神经网络110的配置，新的层输出126可以被提供作为对在递归神经网络110中的另一LSTM层的输入，作为对不同类型的神经网络组件(例如，对输出层或者不同类型的神经网络层)的输入，或者可以被提供作为递归神经网络110的神经网络输出132。

[0031] 具体地，压缩的LSTM层120将多个门应用于当前层输入122、当前层状态和当前层输出，以生成新的层输出126并且对当前层状态进行更新以生成新的层状态124，其中，门中的至少一个包括压缩的权重矩阵。例如，可以通过用结构化矩阵(“结构化矩阵压缩”)代替在层中的门参数矩阵中的一个或多个，或者通过用压缩的参数矩阵和投影矩阵(“投影压缩”)重新定义门参数矩阵，来对层堆叠中的至少一个层进行压缩。门可以包括但不限于，例如，输入门、遗忘门(forget gate)、元胞状态门(cell state gate)、或者输出门。另外，每个门可以包括层间参数矩阵和递归参数矩阵两者。

[0032] 结构化矩阵是可以用少于mn个参数来描述的m×n矩阵。图2A示出了结构化矩阵的示例。例如，结构化矩阵包括以下一般类别：托普利兹(Toeplitz)矩阵200、范蒙德(Vandermonde)矩阵202和柯西(Cauchy)矩阵204。具体地，托普利兹矩阵200是参数沿对角线关联的矩阵。即，托普利兹矩阵200沿其对角线中的每个对角线具有常数值。当相同的属性适用于反对角线时，矩阵200被称为汉克尔矩阵。范蒙德矩阵202是通过对在矩阵的第二列中的条目的元素求幂来定义第3列到第n列中的矩阵条目的矩阵。类似地，柯西矩阵204是

可以完全由两个矢量(U和V)定义的矩阵。柯西矩阵204的每个元素a_{ij}由$\frac{1}{(u_i - v_j)}$表示。

[0033] 使用这种结构化矩阵来表示在压缩的LSTM层中的门矩阵可以降低对LSTM网络的记忆要求,因为这种结构化矩阵可以用少于mn个参数来完全描述。另外,结构化矩阵可以加快对LSTM网络的训练和处理,因为该结构化矩阵允许更快地执行矩阵矢积和梯度计算。

[0034] 可以将上述一般类别的结构化矩阵修改为用于在压缩的LSTM层120中使用的类结构化矩阵。例如,类托普利兹矩阵是对托普利兹矩阵的推广(generalization),该推广包括托普利兹矩阵的乘积矩阵和逆矩阵、及其线性组合。如图2B所示,可以将类托普利兹矩阵参数化,作为r循环(circulant)矩阵和反循环(skew-circulant)矩阵的乘积之和。

[0035] 再次参照图1,为了方便起见,将使用类托普利兹矩阵为例来讨论在递归神经网络110的压缩的LSTM层120中使用结构化矩阵。可以使用位移秩(即,相加的乘积的数量,如图2B的循环矩阵和反循环矩阵之和所示)来控制类托普利兹矩阵的复杂性。低位移秩与高度结构化矩阵(诸如,循环矩阵和托普利兹矩阵及其逆矩阵)对应。高位移秩可以用于对渐增的非结构化矩阵进行建模。在某些示例中,位移秩可以用于控制压缩方案的计算复杂性、存储要求和建模能力。在某些示例中,可以基于应用要求来调整位移秩。

[0036] 在某些实施方式中,针对在特定压缩的LSTM层120中的所有门,将类托普利兹矩阵结构应用于递归和层间参数矩阵。在某些实施方式中,将类托普利兹矩阵结构应用于在层堆叠中的低阶的层(例如,层1和层2)。

[0037] 在用于递归神经网络层的投影压缩模型中,通过用大小为m×r的压缩的递归参数矩阵和r×n的投影矩阵代替特定层(例如,层l)中的大小为m×n的未压缩的递归参数矩阵,来产生压缩的LSTM层120。另外,用大小为m×r的压缩的层间矩阵和相同的投影矩阵代替下一高阶LSTM层(例如,层l+1)中的大小同样为m×n的对应层间参数矩阵。此外,压缩的递归和层间矩阵以及投影矩阵的相应秩比对应的递归和层间参数矩阵的秩小。低秩投影矩阵在两个对应的层之间共享。可以将投影压缩模型应用于一个或多个不同的门。在某些实施方式中,将投影压缩模型应用于在层堆叠中的高阶层(例如,层2-层N)。

[0038] 在某些实施方式中,可以使用结构化矩阵(或者类结构化矩阵)来对层或者层集合进行压缩,并且可以使用投影矩阵来对层或者层集合进行压缩。例如,可以通过用类托普利兹矩阵代替门参数矩阵来对递归神经网络110的低阶层或者层集合(例如,层1和层2)进行压缩,并且可以通过使用投影矩阵来对高阶层或者层集合(例如,层2-层N)进行压缩。

[0039] 在某些实施方式中,上述压缩技术可以导致用于LSTM神经网络的参数的至少75%的减少。在某些实施方式中,在系统是语音识别模型的情况下,上述压缩技术可以导致在将误字率保持在系统的未压缩版本的误字率(WER)的0.3%内的同时,LSTM神经网络压缩的至少75%的减少。在某些示例中,上述压缩技术可以导致使LSTM神经网络的压缩的范围为75%-83%,对应WER范围为0.3%-2.3%。

[0040] 为了将递归神经网络110配置成生成神经网络输出,神经网络系统100对递归神经网络110进行训练以确定递归神经网络110的参数的训练值,包括确定饱和LSTM层120的参数的训练值。下面参照图5更详细地描述了对递归神经网络进行训练。

[0041] 图3是用于对当前层输入进行处理以生成下一层输出的示例性过程300的流程图。为了方便起见,将过程300描述为由饱和LSTM层执行,该饱和LSTM层由位于一个或多个位置

的一个或多个计算机的系统来实施。例如，根据本说明书被适当地编程的在神经网络系统中的饱和LSTM层(例如，图1的神经网络系统100的压缩的LSTM层120)可以执行过程300。

[0042] LSTM层将遗忘门应用于门输入矢量，以生成遗忘门输出矢量(步骤302)。下面将参照图4更详细地描述将门应用于门输入矢量。

[0043] 在某些实施方式中，LSTM层通过将当前层输入和当前层输出进行连结(concatenate)来生成门输入矢量。在某些其它实施方式中，LSTM层是窥视孔LSTM层，该窥视孔LSTM层通过将当前层输入、当前层输出和当前层状态连结来生成门输入矢量。

[0044] LSTM层将输入门应用于门输入矢量，以生成输入门输出矢量(步骤304)。下面将参照图4更详细地描述将门应用于门输入矢量。在某些实施方式中，输入门包括结构化参数矩阵，例如，类托普利兹结构化参数矩阵。在某些实施方式中，输入门包括压缩的递归或者层间矩阵以及对应的投影矩阵。

[0045] 在某些实施方式中，代替于应用输入门以用于生成输入门输出矢量之外，系统还将遗忘门输出矢量用作输入门输出矢量。即，在某些实施方式中，输入门与遗忘门相同。

[0046] LSTM层将输出门应用于门输入矢量，以生成输出门输出矢量(步骤306)。下面将参照图4更详细地描述将门应用于门输入矢量。在某些实施方式中，输出门包括结构化参数矩阵，例如，类托普利兹结构化参数矩阵。在某些实施方式中，输出门包括压缩的递归或者层间矩阵以及对应的投影矩阵。

[0047] LSTM层从当前层输入和当前层输出生成中间元胞状态更新矢量(步骤308)。具体地，LSTM层通过使用具有作为挤压函数的激活函数的神经网络层来对当前层输入和当前层输出进行处理，以生成中间元胞状态更新矢量。

[0048] 通常，挤压函数是将接收到的输入映射到范围-1至1(不包括-1和1)的函数。例如，挤压函数可以是双曲正切函数。

[0049] LSTM层将中间元胞状态更新矢量和输入门输出矢量进行组合，以生成最终元胞状态更新矢量(步骤310)。具体地，LSTM层计算在中间元胞状态更新矢量与输入门输出矢量之间的逐点乘法，以生成最终元胞状态更新矢量。

[0050] LSTM层将当前元胞状态和遗忘门输出矢量进行组合，以生成中间新的元胞状态(步骤312)。具体地，LSTM层计算在当前元胞状态和遗忘输出矢量之间的逐点乘法，以生成中间新的元胞状态。在某些实施方式中，遗忘门包括结构化参数矩阵，例如，类托普利兹结构化参数矩阵。在某些实施方式中，遗忘门包括压缩的递归或者层间矩阵以及对应的投影矩阵。

[0051] LSTM层将中间新的元胞状态和最终元胞状态更新矢量进行组合(例如，求和)，以生成最终新的元胞状态(步骤314)。

[0052] LSTM层从最终新的元胞状态生成新的层输出(步骤316)。为了生成新的层输出，LSTM层将挤压函数应用于最终新的元胞状态的每个分量，以生成中间新的层输出。

[0053] LSTM层然后将输出门输出矢量和中间新的层输出进行组合，以生成新的层输出。具体地，LSTM层在输出门输出矢量与中间新的层输出之间执行逐点乘法，以生成新的层输出。

[0054] 除了使用最终新的层状态来生成新的层输出之外，LSTM层维持最终新的元胞状态连同新的层输出，在供在后续时间步长使用。

[0055] 图4是用于将门应用于门输入矢量以生成门输出矢量的示例性过程400的流程图。为了方便起见，将过程400描述为由饱和LSTM层执行，该饱和LSTM层由位于一个或多个位置的一个或多个计算机的系统来实施。例如，根据本说明书被适当地编程的在神经网络系统中的压缩的LSTM层(例如，图1的神经网络系统100的压缩的LSTM层120)可以执行过程400。

[0056] LSTM层确定门输入矢量(步骤402)。

[0057] LSTM层根据参数集从门输入矢量生成相应的中间门输出矢量(步骤404)。在某些实施方式中，LSTM层在参数矩阵与门输入矢量之间执行矩阵乘法，并且然后向矩阵乘法的输出添加偏置矢量以生成中间门输出矢量，其中，门中的每个门具有不同的参数矩阵和偏置矢量。即，在LSTM层具有不同的输入门、遗忘门和输出门的实施方式中，这些门中的每个门将具有彼此不同的参数矩阵和偏置矢量。

[0058] LSTM层将选通函数应用于相应的中间门输出矢量的每个分量，以生成最终门输出矢量(步骤406)。

[0059] 通常，选通函数是将接收到的输入映射到范围0至1(不包括0和1)的函数。例如，选通函数可以是S型函数。

[0060] 然而，针对LSTM层的门中的至少一个门，步骤404中所提及的参数矩阵是压缩的参数矩阵。用压缩的参数矩阵代替未压缩的参数矩阵。LSTM层然后在压缩的参数矩阵与门输入矢量之间执行矩阵乘法。例如，压缩的矩阵可以替代输入门、遗忘门、元胞状态、或者输出门中的任何一个中的参数矩阵。在某些实施方式中，将压缩的参数矩阵应用于在LSTM层中的多个门。例如，可以将压缩的参数矩阵应用于输入门和输出门两者。作为另一示例，可以将压缩的参数矩阵应用于输入门、输出门和遗忘门。

[0061] 在某些实施方式中，压缩参数矩阵是类结构化矩阵，例如，类托普利兹结构化矩阵。在某些实施方式中，用压缩的参数矩阵和对应的投影矩阵来重新定义未压缩的门参数。

[0062] 图5是用于对包括压缩的LSTM层的递归神经网络进行训练的示例性过程500的流程图。为了方便起见，将过程500描述为由位于一个或多个位置的一个或多个计算机的系统来执行。例如，根据本说明书被适当地编程的神经网络系统(例如，图1的神经网络系统100)可以执行过程500。

[0063] 系统获取用于对递归神经网络进行训练的训练数据(步骤502)。该训练数据包括多个训练示例对，其中，每个训练示例对包括训练神经网络输入和该训练神经网络输入的目标神经网络输出。

[0064] 系统在训练数据上对递归神经网络进行训练，以通过使目标函数最优化(即，最大化或者最小化)，从参数的初始值确定递归神经网络的参数的训练值(步骤504)。在训练期间，系统对压缩的矩阵的值赋予约束条件，从而使其继续满足对压缩的矩阵的要求。例如，针对类托普利兹结构化矩阵，系统可以赋予约束条件，从而使每个压缩的矩阵的条目总是类托普利兹的，或者，针对投影模型压缩的矩阵，系统可以调节投影矩阵和压缩的参数矩阵的值，而不是直接调节参数矩阵的值。

[0065] 系统通过传统机器学习训练技术(例如，具有随着时间反向传播的随机梯度下降训练技术)使目标函数最优化，来对递归神经网络进行训练。即，系统可以执行训练技术的多次迭代，以通过调节递归神经网络的参数的值来对目标函数进行优化。

[0066] 可以在数字电子电路系统中、在有形地体现的计算机软件或者固件中、在包括本

说明书中公开的结构及其结构等同物的计算机硬件中、或者在它们中的一个或多个的组合中，实施本说明书中描述的主题和功能操作的实施例。可以将本说明书中描述的主题的实施例实施为一个或多个计算机程序，即，编码在有形非瞬时程序载体上的由数据处理设备执行或者控制该数据处理设备的操作的计算机程序指令的一个或多个模块。替选地或附加地，程序指令可以编码在人工生成的为了对用于传输对合适的接收器设备供数据处理设备执行的信息进行编码而生成的传播信号(例如，机器生成的电气、光学或者电磁信号)上。计算机存储介质可以是机器可读存储设备、机器可读存储基板、随机或者串行存取存储器设备、或者它们中的一个或多个的组合。

[0067] 术语“数据处理设备”指数据处理硬件，并且涵盖用于处理数据的各种装置、设备和机器，包括例如可编程处理器、计算机、或者多个处理器或者计算机。所述装置还可以是或者进一步包括专用逻辑电路系统，例如，FPGA(现场可编程门阵列)或者ASIC(专用集成电路)。除了硬件之外，所述装置可以可选地包括为计算机程序创建执行环境的代码，例如，构成处理器固件、协议栈、数据库管理系统、操作系统、或者它们中的一个或多个的组合的代码。

[0068] 可以用任何形式的编程语言(包括编译语言或者解译语言、陈述性语言或者程序语言)来编写计算机程序(也称为程序、软件、软件应用、模块、软件模块、脚本或者代码)，并且可以以任何形式(包括独立程序或者模块、组件、子例程、或者适合用于计算环境的其它单元)来部署该计算机程序。计算机程序可以但并非必须与文件系统中的文件对应。可以将程序存储在保持其它程序或者数据(例如，存储在标记语言文档中的一个或多个脚本)的文件的一部分中，或者存储在专用于所探讨的程序的单个文件中，或者存储在多个协作文件(例如，存储一个或多个模块、子程序、或者部分代码的多个文件)中。可以将计算机程序部署为在一个计算机上执行或者在位于一个站点处或者跨越多个站点分布并且通过通信网络互相连接的多个计算机上执行。

[0069] 本说明书中所描述的过程和逻辑流可以由一个或多个可编程计算机执行，该一个或多个可编程计算机执行一个或多个计算机程序以通过操作输入数据和生成输出来执行功能。也可以由专用逻辑电路系统(例如，FPGA(现场可编程门阵列)或者ASIC(专用集成电路))执行过程和逻辑流程，并且也可以将所述装置实施为FPGA或者ASIC。

[0070] 适合计算机程序的执行的计算机包括例如可以是基于通用微处理器或者专用微处理器或者两者、或者任何其它种类的中央处理单元。一般来说，中央处理单元接收来自只读存储器或者随机存取存储器或者两者的指令和数据。计算机的关键元件是用于实现或者执行指令的中央处理单元和用于存储指令和数据的一个或多个存储器。一般而言，计算机还将包括用于存储数据的一个或多个海量存储设备(例如，磁盘、磁光盘、或者光盘)，或者计算机可以操作地耦合到海量存储设备以接收来自海量存储设备的数据或者将数据传输到该海量存储设备或者两者。然而，计算机无需具有这种设备。此外，计算机可以嵌入在另一设备中，例如，移动电话、个人数字助理(PDA)、移动音频或者视频播放器、游戏机、全球定位系统(GPS)接收器、或者便携式存储设备(例如，通用串行总线(USB)闪存驱动)，仅举几例。

[0071] 适合于存储计算机程序指令和数据的计算机可读介质包括所有形式的非易失性存储器、介质和存储器设备，包括例如半导体存储器设备(例如，EPROM、EEPROM和闪速存储

器设备)、磁盘(例如,内部硬盘或者可移动盘)、磁光盘、CD-ROM盘和DVD-ROM盘。处理器和存储器可以由专用逻辑电路系统补充或者可以并入专用逻辑电路系统中。

[0072] 为了提供与用户的交互,可以在具有显示设备以及键盘和指点设备的计算机上实施本说明书中所描述的主题的实施例,显示设备用于向用户显示信息,例如,CRT(阴极射线管)或者LCD(液晶显示器)监视器;以及键盘和指点设备例如鼠标或者轨迹球,用户可以通过键盘和指点设备来将输入提供给计算机。其它种类的设备也可以用于提供与用户的交互;例如,提供给用户的反馈可以是任何形式的传感反馈,例如,视觉反馈、听觉反馈或者触觉反馈;并且可以用任何形式来接收来自用户的输入(包括声输入、语音输入或者触觉输入)。另外,计算机可以通过将文档发送到由用户所使用的设备并且接收来自该设备的文档(例如,通过响应于从web浏览器接收到的请求来将web页面发送对在用户的客户端上的web浏览器)来与用户交互。

[0073] 可以在计算系统中实施本说明书中所描述的主题的实施例,所述计算系统包括后台组件(例如,数据服务器),或者包括中间件组件(例如,应用服务器),或者包括前端组件(例如,具有关系图形用户界面或者web浏览器的客户端计算机,用户可以通过所述关系图形用户界面或者所述web浏览器来与本发明中所描述的主题的实施方式交互),或者包括这样的后台组件、中间件组件或者前端组件中的一个或多个的任何组合。可以通过任何形式或者介质的数字数据通信(例如,通信网络)来将系统的部件相互连接。通信网络的示例包括局域网(“LAN”)和广域网(“WAN”),例如,互联网。

[0074] 计算系统可以包括客户端和服务器。客户端和服务器通常彼此远离并且典型地通过通信网络交互。客户端和服务器的关系借助在相应的计算机上运行并且彼此具有客户端-服务器关系的计算机程序而产生。

[0075] 虽然本说明书包含了许多具体实施细节,但是不应所述将这些细节视为对任何发明或者可能被要求保护的内容的范围的限制,而是作为针对特定发明的特定实施例的特有特征的描述。在本说明书中在单独实施例的上下文中描述的某些特征也可以组合地实施在单个实施例中。相反,在单个实施例的上下文中描述的各种特征也可以单独地或者按照任何合适的子组合实施在多个实施例中。此外,虽然上面可能将特征描述为以某些组合来起作用并且甚至最初同样地对该特征这样要求,但是在一些情况下可以从组合中删除来自所要求保护的组合的一个或多个特征,并且所要求保护的组合可以涉及子组合或者子组合的变形。

[0076] 同样,虽然在附图中按照特定次序示出了操作,但是不应该将其理解为需要按照所示的特定次序或者按照顺序次序来执行这种操作,或者需要执行所有图示的操作,以实现期望的结果。在某些情况下,多任务处理和并行处理可以是有利的。此外,不应将在上述实施例中的各种系统模块和组件的分离理解为在所有实施例中需要这种分离,并且应理解,所描述的程序部件和系统通常可以一起集成在单个软件产品中或者封装到多个软件产品中。

[0077] 已经描述了本主题的具体实施例。其它实施例在随附权利要求书的范围内。例如,在权利要求书中所记述的动作可以按照不同的次序执行并且仍然可以实现期望结果。作为一个示例,在附图中所示的过程不一定需要所示的特定次序或者顺序次序来完成期望结果。在某些实施方式中,多任务处理和并行处理可能是有利的。

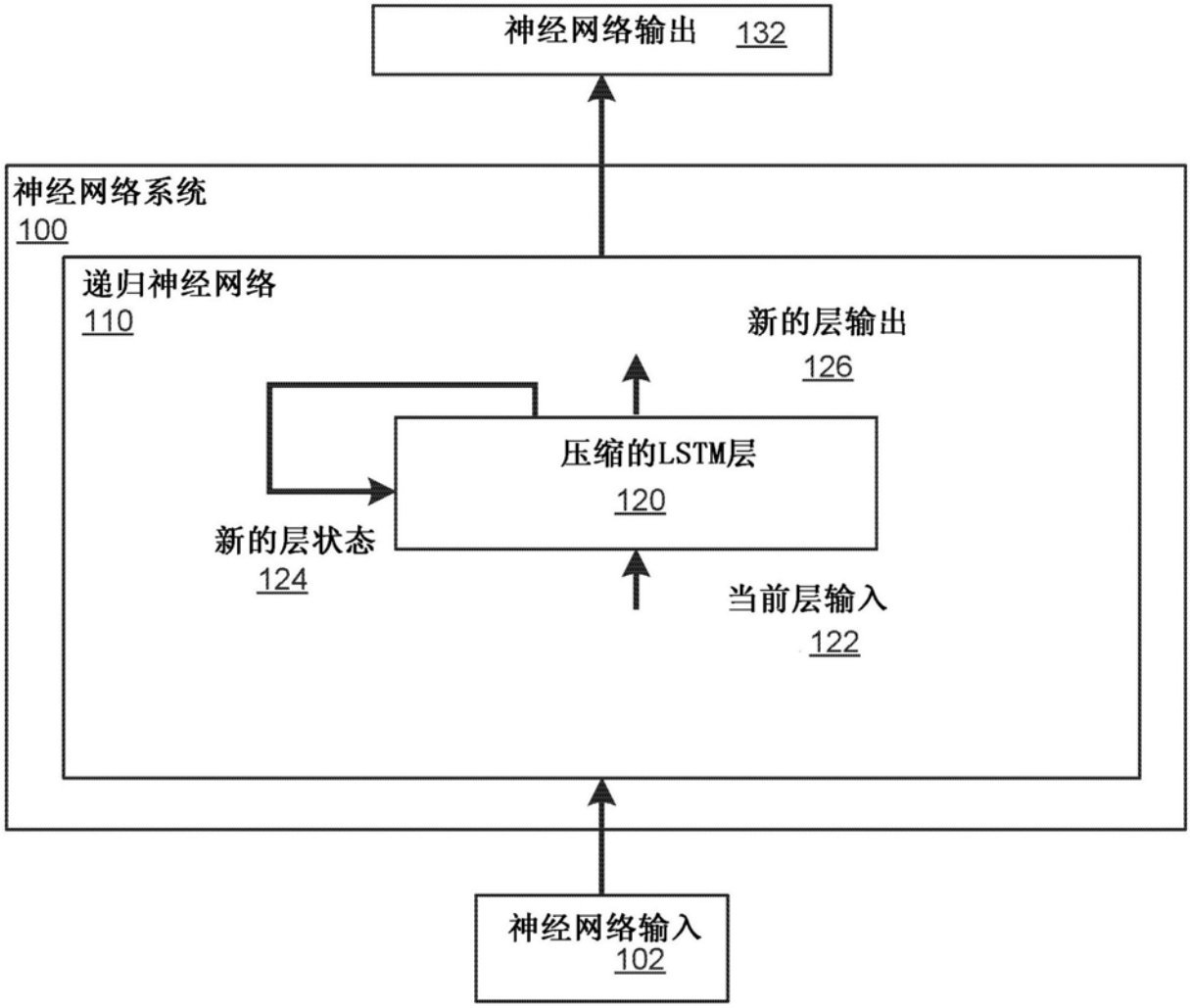

神经网络输出 132
神经网络系统
100
递归神经网络
110
新的层输出
126
压缩的LSTM层
120
新的层状态
124
当前层输入
122
神经网络输入
102

图1

托普利兹：$T(t)_{ij} = t_{i-j}$

200

$$\begin{bmatrix} t_0 & t_{-1} & \cdots & t_{1-n} \\ t_1 & t_0 & \cdots & \vdots \\ \vdots & \vdots & \vdots & t_{-1} \\ t_{n-1} & \vdots & t_1 & t_0 \end{bmatrix}$$

范蒙德：$V(v)_{ij} = v_i^j$

202

$$\begin{bmatrix} 1 & v_0 & \cdots & v_0^{n-1} \\ 1 & v_1 & \cdots & v_1^{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & v_{n-1} & \cdots & v_{n-1}^{n-1} \end{bmatrix}$$

柯西：$C(u,v)_{ij} = (u_i - v_j)^{-1}$

204

$$\begin{bmatrix} \frac{1}{(u_0 - v_0)} & \cdots & \cdots & \frac{1}{(u_0 - v_{n-1})} \\ \frac{1}{(u_1 - v_0)} & \cdots & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \frac{1}{(u_{n-1} - v_0)} & \cdots & \cdots & \frac{1}{(u_{n-1} - v_{n-1})} \end{bmatrix}$$

图2A

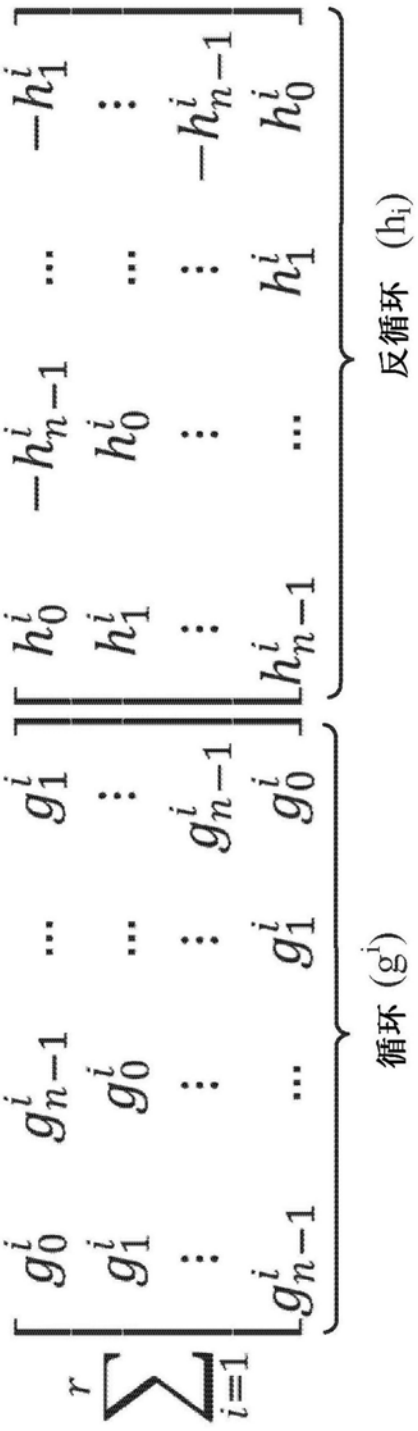


图2B

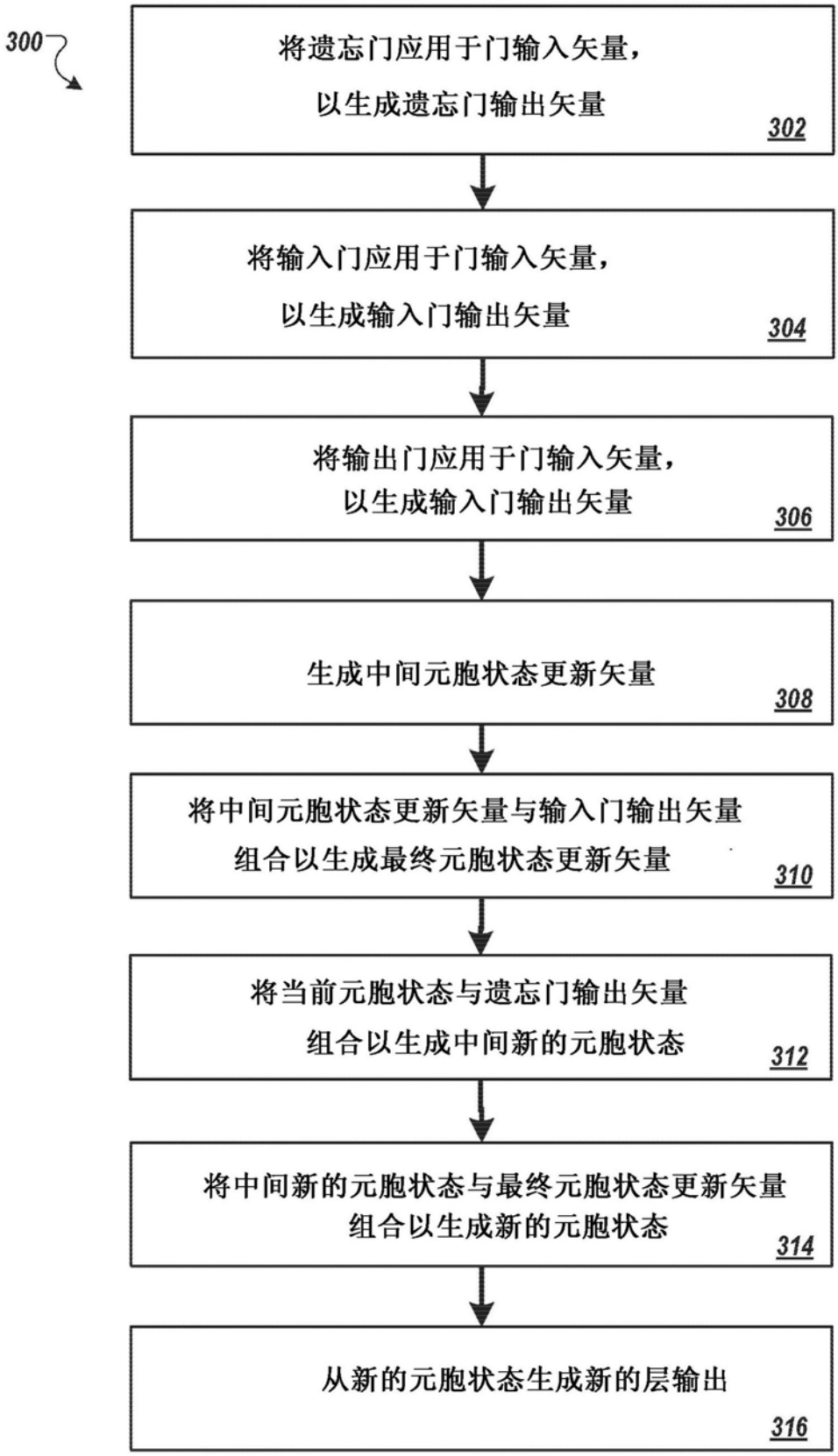

300
将遗忘门应用于门输入矢量，
以生成遗忘门输出矢量
302
将输入门应用于门输入矢量，
以生成输入门输出矢量
304
将输出门应用于门输入矢量，
以生成输入门输出矢量
306
生成中间元胞状态更新矢量
308
将中间元胞状态更新矢量与输入门输出矢量
组合以生成最终元胞状态更新矢量
310
将当前元胞状态与遗忘门输出矢量
组合以生成中间新的元胞状态
312
将中间新的元胞状态与最终元胞状态更新矢量
组合以生成新的元胞状态
314
从新的元胞状态生成新的层输出
316

图3

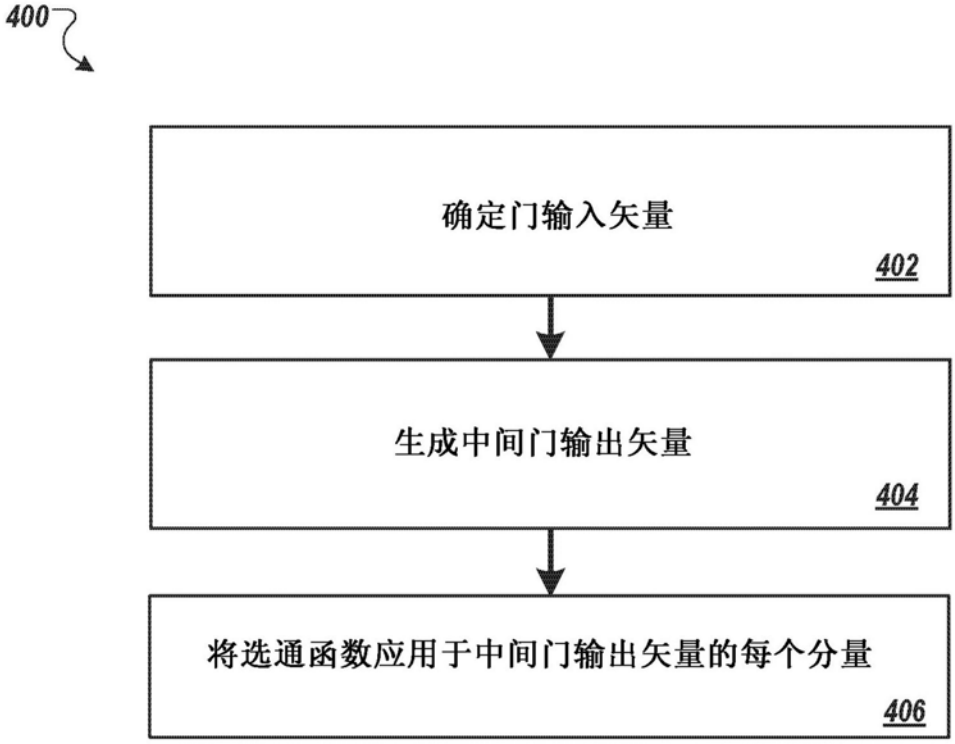

400
确定门输入矢量
402
生成中间门输出矢量
404
将选通函数应用于中间门输出矢量的每个分量
406

图4

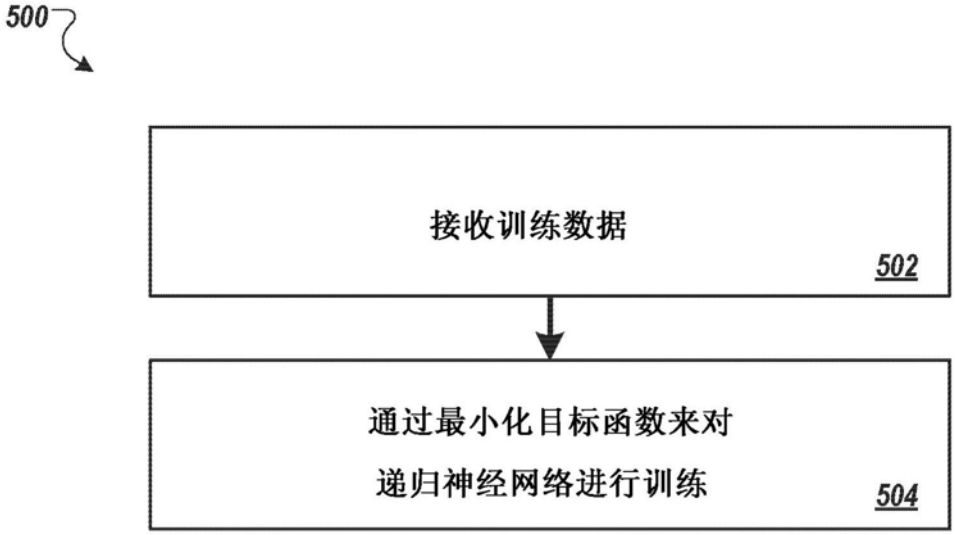

500
接收训练数据
502
通过最小化目标函数来对
递归神经网络进行训练
504

图5