(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 111133453 A
(43)申请公布日 2020.05.08

(21)申请号 201780094924.4
(22)申请日 2017.08.04
(85)PCT国际申请进入国家阶段日
2020.03.16
(86)PCT国际申请的申请数据
PCT/CN2017/096004 2017.08.04
(87)PCT国际申请的公布数据
WO2019/024083 EN 2019.02.07
(71)申请人 诺基亚技术有限公司
地址 芬兰埃斯波
(72)发明人 汪萌
(74)专利代理机构 北京市金杜律师事务所 11256
代理人 酆迅
(51)Int.Cl.
G06N 3/08(2006.01)

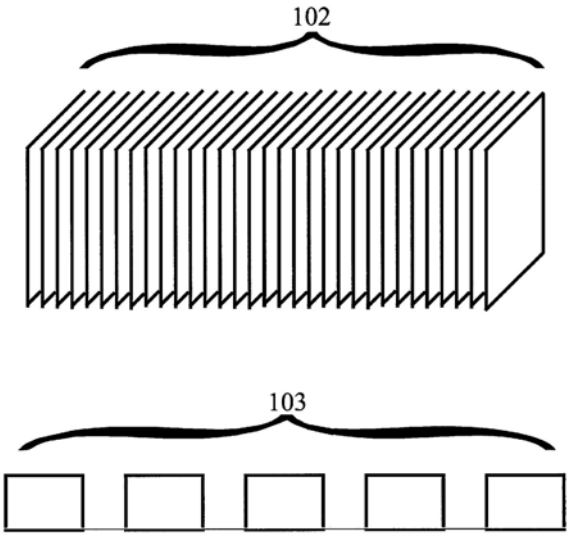
权利要求书2页 说明书13页 附图4页

(54)发明名称

人工神经网络

(57)摘要

提供了一种装置，该装置包括至少一个处理核心、至少一个存储器，该至少一个存储器包括计算机程序代码，该至少一个存储器和计算机程序代码被配置为与至少一个处理核一起使装置至少：从第一顺序输入获得来自第一循环神经网络的第一输出，该第一顺序输入具有第一模态；从第二顺序输入获得来自第二循环神经网络的第二输出，该第二顺序输入具有第二模态；以及处理第一输出和第二输出以获得第一顺序输入和第二顺序输入的相关性。



CN 111133453 A

1.一种装置,包括至少一个处理核心、至少一个存储器,所述至少一个存储器包括计算机程序代码,所述至少一个存储器和所述计算机程序代码被配置为与所述至少一个处理核心一起使所述装置至少:

-从第一顺序输入获得来自第一循环神经网络的第一输出,所述第一顺序输入具有第一模态;

-从第二顺序输入获得来自第二循环神经网络的第二输出,所述第二顺序输入具有第二模态;以及

-处理所述第一输出和所述第二输出以获得所述第一顺序输入和所述第二顺序输入的相关性。

2.根据权利要求1所述的装置,其中所述第一循环神经网络和所述第二循环神经网络均包括多个长短期记忆LSTM块。

3.根据权利要求1或2所述的装置,其中所述第一循环神经网络包括第一数目的长短期记忆LSTM块,并且所述第二循环神经网络包括第二数目的LSTM块,所述第一数目和所述第二数目不相等。

4.根据权利要求1至3中任一项所述的装置,其中所述至少一个存储器和所述计算机程序代码被配置为与所述至少一个处理核心一起使所述装置:通过在获得内积以获得所述相关性之前将所述第一输出和所述第二输出绘制为具有相等的维度,来处理所述第一输出和所述第二输出。

5.根据权利要求4所述的装置,其中所述至少一个存储器和所述计算机程序代码还被配置为与所述至少一个处理核心一起使所述装置通过采用线性变换来将所述输出绘制为具有相等的维度。

6.根据权利要求1至5中任一项所述的装置,其中所述第一模态和所述第二模态彼此不同。

7.根据权利要求1至6中任一项所述的装置,其中所述第一模态从以下列表中被选择:视频、文本语句、语音、手写文本、立体视频、图像序列、飞行器控制状态和飞行器系统输出。

8.根据权利要求1至7中任一项所述的装置,其中所述第二模态从以下列表中被选择:视频、文本语句、语音、手写文本、立体视频、图像序列、飞行器控制状态和飞行器系统输出。

9.根据权利要求1至8中任一项所述的装置,其中顺序输入包括以下输入,所述输入包括输入元素序列。

10.一种方法,包括:

-从第一顺序输入获得来自第一循环神经网络的第一输出,所述第一顺序输入具有第一模态;

-从第二顺序输入获得来自第二循环神经网络的第二输出,所述第二顺序输入具有第二模态;以及

-处理所述第一输出和所述第二输出以获得所述第一顺序输入和所述第二顺序输入的相关性。

11.根据权利要求10所述的方法,其中所述第一循环神经网络和所述第二循环神经网络均包括多个长短期记忆LSTM块。

12.根据权利要求10或11所述的方法,其中所述第一循环神经网络包括第一数目的长

短期记忆LSTM块，并且所述第二循环神经网络包括第二数目的LSTM块，所述第一数目和所述第二数目不相等。

13.根据权利要求10至12中任一项所述的方法，其中所述第一输出和所述第二输出通过以下被处理：在获得内积以获得所述相关性之前，将所述第一输出和所述第二输出绘制为具有相等的维度。

14.根据权利要求13所述的方法，其中所述输出通过采用线性变换被绘制为具有相等的维度。

15.根据权利要求10至14中任一项所述的方法，其中所述第一模态和所述第二模态彼此不同。

16.根据权利要求10至15中任一项所述的方法，其中所述第一模态从以下列表中被选择：视频、文本语句、语音、手写文本、立体视频、图像序列和业务模式。

17.根据权利要求10至16中任一项所述的方法，其中所述第二模态从以下列表中被选择：视频、文本语句、语音、手写文本、立体视频、图像序列和业务模式。

18.根据权利要求10至17中任一项所述的方法，其中顺序输入包括以下输入，所述输入包括输入元素序列。

19.一种装置，包括：

-用于从第一顺序输入获得来自第一循环神经网络的第一输出的部件，所述第一顺序输入具有第一模态；

-用于从第二顺序输入获得来自第二循环神经网络的第二输出的部件，所述第二顺序输入具有第二模态；以及

-用于处理所述第一输出和所述第二输出以获得所述第一顺序输入和所述第二顺序输入的相关性的部件。

20.一种非瞬态计算机可读介质，其上存储有计算机可读指令集，所述计算机可读指令集在由至少一个处理器执行时使装置至少：

-从第一顺序输入获得来自第一循环神经网络的第一输出，所述第一顺序输入具有第一模态；

-从第二顺序输入获得来自第二循环神经网络的第二输出，所述第二顺序输入具有第二模态；以及

-处理所述第一输出和所述第二输出以获得所述第一顺序输入和所述第二顺序输入的相关性。

21.一种计算机程序，被配置为使根据权利要求1至9中至少一项的方法被执行。

人工神经网络

技术领域

[0001] 本发明涉及人工神经网络,并且得出输入序列之间的相关性

背景技术

[0002] 人工神经网络可以以不同的方式被配置。通常,在本文中简称为神经网络的人工神经网络包括人工神经元,该人工神经元在本文中将同样简称为神经元。例如,神经元可以被配置为接收输入、施加权重以及基于输入和权重提供输出。神经元可以以软件来实现,或者附加地或备选地,神经网络可以包括基于硬件的人工神经元。

[0003] 神经网络可以基于其属性和配置而被分类为不同的类型。前馈神经网络是其中神经元之间的连接不形成循环的网络,即数据单向流动。前馈型神经网络的示例是卷积神经网络CNN,其中滤波器被用于对先前层中的数据应用卷积运算,以获得用于后续层的数据。CNN可以包括全连接的层,其中每个神经元与先前层中的每个神经元相连接。例如,CNN已经在图像分类中被采用。CNN在有限的设定中已经达到了与人类相当的分类性能水平。

[0004] 另一方面,循环神经网络(recurrent neural network)RNN是其中神经元之间的连接形成有向循环的网络。这些网络由于其内部存储器而进行对序列数据的处理,并且RNN可以对时间的线性进程进行操作,将先前的时间步长和隐藏的表示组合为当前时间步长的表示。例如,RNN已经被用于语音识别和连接的未分段的笔迹识别中。

[0005] 递归神经网络(recursive neural network)在结构上递归地应用权重的集合,以针对可能具有变化大小的输入结构产生结构化的预测。循环神经网络是递归神经网络的一种形式。

发明内容

[0006] 本发明由独立权利要求的特征限定。一些具体实施例在从属权利要求中被定义。

[0007] 根据本发明的第一方面,提供了一种装置,该装置包括至少一个处理核心、至少一个存储器,该至少一个存储器包括计算机程序代码,该至少一个存储器和计算机程序代码被配置为,与至少一个处理核心一起使装置至少:从第一顺序输入获得来自第一循环神经网络的第一输出,该第一顺序输入具有第一模态;从第二顺序输入获得来自第二循环神经网络的第二输出,该第二顺序输入具有第二模态;以及处理第一输出和第二输出以获得第一顺序输入和第二顺序输入的相关性。

[0008] 第一方面的各种实施例可以包括来自以下项目符号列表的至少一个特征:

[0009] • 第一循环神经网络和第二循环神经网络均包括多个长短期记忆LSTM块

[0010] • 第一循环神经网络包括第一数目的长短期记忆LSTM块,并且第二循环神经网络包括第二数目的LSTM块,第一数目和第二数目不相等

[0011] • 至少一个存储器和计算机程序代码被配置为,与至少一个处理核心一起使装置:通过在获得内积以获得相关性之前,对第一输出和第二输出进行绘制具有相等的维度,来处理第一输出和第二输出

[0012] • 至少一个存储器和计算机程序代码被配置为与至少一个处理核一起，使装置通过采用线性变换来对输出进行绘制以具有相等的维度

[0013] • 第一模态和第二模态彼此相同

[0014] • 第一模态从以下列表中被选择：视频、文本语句、语音、手写文本、立体声视频、图像序列、飞行器控制状态和飞行器系统输出

[0015] • 第二模态从以下列表中被选择：视频、文本语句、语音、手写文本、立体声视频、图像序列、飞行器控制状态和飞行器系统输出

[0016] • 顺序输入包括如下输入，该输入包括输入元素序列。

[0017] 根据本发明的第二方面，提供了一种方法，包括：从第一顺序输入获得来自第一循环神经网络的第一输出，该第一顺序输入具有第一模态；从第二顺序输入获得来自第二循环神经网络的第二输出，该第二顺序输入具有第二模态；以及处理第一输出和第二输出以获得第一顺序输入和第二顺序输入的相关性。

[0018] 第二方面的各种实施例可以包括来自以下项目符号列表的至少一个特征：

[0019] • 第一循环神经网络和第二循环神经网络均包括多个长短期记忆LSTM块

[0020] • 第一循环神经网络包括第一数目的长短期记忆LSTM块，并且第二循环神经网络包括第二数目的LSTM块，第一数目和第二数目不相等

[0021] • 在获得内积以获得相关性之前，第一输出和第二输出通过调第一输出和第二输出进行绘制以具有相等的维度而被处理

[0022] • 输出通过采用线性变换被绘制为具有相等的维度

[0023] • 第一模态和第二模态彼此相同

[0024] • 第一模态从以下列表中选择：视频、文本语句、语音、手写文本、立体声视频、图像序列和业务模式

[0025] • 从以下列表中选择第二模态：视频、文字语句、语音、手写文本、立体声视频、图像序列和业务模式

[0026] • 顺序输入包括一个包括一系列输入元素的输入。

[0027] 根据本发明的第三方面，提供一种装置，该装置包括用于从第一顺序输入获得来自第一循环神经网络的第一输出的部件，该第一顺序输入具有第一模态；用于从第二顺序输入获得来自第二循环神经网络的第二输出的部件，该第二顺序输入具有第二模态；以及用于处理第一输出和第二输出以获得第一顺序输入和第二顺序输入的相关性的部件。

[0028] 根据本发明的第四方面，提供了一种非瞬态计算机可读介质，其上存储有计算机可读指令的集合，该计算机可读指令的集合在由至少一个处理器执行时，使装置至少：从第一顺序输入获得来自第一循环神经网络的第一输出，该第一顺序输入具有第一模态；从第二顺序输入获得来自第二循环神经网络的第二输出，该第二顺序输入具有第二模态；以及处理第一输出和第二输出以获得第一顺序输入和第二顺序输入的相关性。

[0029] 根据本发明的第五方面，提供一种计算机程序，该计算机程序被配置为使得根据第二方面的方法被执行。

附图说明

[0030] 图1图示了与本发明的至少一些实施例一起可用的顺序输入；

[0031] 图2A图示了LSTM块；

[0032] 图2B图示了根据本发明的至少一些实施例的示例系统；

[0033] 图3图示了能够支持本发明的至少一些实施例的示例装置；

[0034] 图4A图示了编码器-解码器架构；

[0035] 图4B图示了LSTM块中的数据流，以及

[0036] 图5是根据本发明的至少一些实施例的方法的流程图。

具体实施方式

[0037] 本文描述了人工神经网络架构，其可用于获得描述输入数据的不同序列之间的相关水平的相关系数，该输入数据不需要具有相同的模态。例如，输入序列可以包括不同媒体格式或媒体类型的数据，诸如视频序列和文本语句。在得出相关系数之前，循环神经网络被用于表征每个输入序列。

[0038] 图1图示了与本发明的至少一些实施例一起可用的顺序输入。第一输入序列102被示于图的较上部，并且第二输入序列103被示于图的较下部。例如，第一输入序列102可以包括视频剪辑，该视频剪辑包括多个静止视频帧。第二输入序列103可以包括例如语句，该语句包括自然语言的单词序列，诸如中文或法语。

[0039] 视频剪辑可以通过将每个静止视频帧存储在存储介质上而被存储。由于分开存储每个帧可能需要大量的存储容量，因此视频剪辑通常以经编码的格式被存储，其中静止帧不是每个分别被存储，而是参照彼此被存储。这可以大大节省存储容量，因为根据视频的内容，顺序的帧通常共享其内容的一部分，因此不必要重复存储这种共享的内容。当视频剪辑以经编码的格式被存储时，可以通过执行解码操作来从经编码的剪辑生成剪辑的个体静止视频帧。

[0040] 神经网络可以被用于对静止图像进行分类。卷积前馈网络可以被用于确定，例如，特定的静止图像是否描绘了特定类别的对象，例如，从行车记录仪生成的静止图像中是否存在行人。另一方面，循环神经网络RNN可以被用于处理数据序列，诸如自然语言中的语句。例如，从一种自然语言到另一自然语言的机器翻译可以依赖于RNN。近年来，随着基于RNN的解决方案已经被采用，机器翻译的质量已经显著改进。

[0041] 在将视频剪辑的事件转换为自然语言格式时，长短期记忆LSTM、RNN已经被采用。LSTM是RNN块，其可以包括存储单元和三个门。LSTMS可能在LSTM单元内部的循环组件内缺少激活功能。例如，使用基于LSTM的RNN，视频剪辑中两个行星的碰撞可以被绘制为自然语言中的“两个球体碰撞”。然而，迄今为止，这样的系统仅获得了很低的成功率，这使得它不切实际。本发明的至少一些实施例涉及在顺序输入(例如视频剪辑和语句)之间获得相关性的相关但不同的任务。该语句可以以自然语言形式。发明方已经认识到，与将一个顺序输入转换成另一顺序输入，例如转换成另一模态(诸如视频到语句)相比，获得两个顺序输入之间的相关性更容易实现。

[0042] 模态在本文中是指一种数据类型。模态的示例包括视频、文本语句、语音、手写文本、立体声视频、图像序列和业务模式。在一些实施例中，模态可以包括不同的视频编码方法，诸如mpeg4、HEVC和VP9。模式可以包括彩色视频和黑白视频。模式可以包括不同的媒体类型，诸如视频、文本、音频和没有音频的运动图像。

[0043] 现在，给出标准相关分析CCA的简要介绍。给定数据$\mathbf{X} \in \mathbb{R}^{d_1 \times M}$和$\mathbf{Y} \in \mathbb{R}^{d_2 \times M}$的两个视图，其中M是样本的数目，并且$d_1$和$d_2$和是两个视图中的特征维度，CCA旨在将它们投影到统一的特征空间中，在该特征空间中它们最大相关。将两个投影矩阵分别表示为W_X和W_Y，然后CCA的目标可以被写为：

[0044] $$\rho = \max_{\mathbf{W_X},\mathbf{W_Y}} corr(\mathbf{W_X^T X}, \mathbf{W_Y^T Y}) = \max_{\mathbf{W_X},\mathbf{W_Y}} \frac{\mathbf{W_X^T \Sigma_{XY} W_Y}}{\sqrt{\mathbf{W_X^T \Sigma_{XX} W_X}}\sqrt{\mathbf{W_Y^T \Sigma_{YY} W_Y}}}, \quad (1)$$

[0045] 其中Σ_{XX}和Σ_{YY}是集合内协方差矩阵，以及$\mathbf{\Sigma_{XY}}$ $(\mathbf{\Sigma_{XY}} = \mathbf{\Sigma_{XY}^T})$是集合间协方差矩阵。等式(1)等于

[0046] $$\rho = \mathbf{W_X^T \Sigma_{XY} W_Y}, \ s.t. \mathbf{W_X^T \Sigma_{XX} W_X} = \mathbf{W_Y^T \Sigma_{YY} W_Y} = 1. \quad (2)$$

[0047] 通过引入拉格朗日参数，等式(2)可以进一步被写为：

[0048] $$f(\mathbf{W_X}, \mathbf{W_Y}, \boldsymbol{\lambda_X}, \boldsymbol{\lambda_Y}) = \mathbf{W_X^T \Sigma_{XY} W_Y} - \frac{\lambda_X}{2}(\mathbf{W_X^T \Sigma_{XX} W_X} - 1) - \frac{\lambda_Y}{2}(\mathbf{W_Y^T \Sigma_{YY} W_Y} - 1), \quad (3)$$

[0049] 然后关于W_X和W_Y求导，并且使其为零，我们得到

[0050] $\Sigma_{XY}W_Y = \lambda_X \Sigma_{XX} W_X$， (4)

[0051] $\Sigma_{YX}W_Y = \lambda_Y \Sigma_{YY} W_Y$. (5)

[0052] 从等式(4)和等式(5)，以及等式(2)的约束，很容易获得

[0053] $$\lambda_X = \lambda_Y = \mathbf{W_X^T \Sigma_{XY} W_Y}. \quad (6)$$

[0054] 记$\lambda = \lambda_X = \lambda_Y$，并且假定$\Sigma_{XX}$是可逆的，通过考虑等式(5)和等式(6)，我们有：

[0055] $$\mathbf{\Sigma_{YX} \Sigma_{XY}^{-1} \Sigma_{XY} W_Y} = \lambda^2 \mathbf{\Sigma_{YY} W_Y}. \quad (7)$$

[0056] 这是广义特征向量问题。通过解决等式(7)，可以得到W_Y，并且在等式(5)中通过代替W_Y，对应的W_X也可以被获得。利用W_X和W_Y，可以基于它们的经投影的表示容易地计算出不同视图中的两个样本之间的相关性。

[0057] CCA已经被应用于多种任务，诸如多媒体注释和搜索[38]。由于CCA仅执行线性投影，因此在变量之间的关系为非线性的某些复杂情况下，其无法提取有用的特征。常见的方法是使用核方法[2]-[4]、[19]、[20]，以首先将数据投影到较高维度，就像支持向量机SVM中常使用的那些方法一样，并且这种类型的方法是前述的KCCA。由于KCCA的设计遵循CCA的思想，除了核方法外，其还可以容易地被转换为类似的特征向量问题。

[0058] 尽管KCCA实现了非线性，但是其依赖于固定的核，因此KCCA可以学习的表示固有地被限制。另一方面，利用新传入的训练样本对KCCA进行再训练将导致严重的计算成本。最近提出的DCCA[5]旨在解决该问题。DCCA无需将相关性计算转换为特征向量问题，而是利用特定的深层网络对数据的每个模态进行建模，并且投影可以通过梯度下降法被学习。杨等人在[40]提出了学习时间数据表示的方法，以用于多模型融合。然而，当与本发明的实施例相比时，他们的工作具有不同的目标。他们的目标是融合两种模态以完成分类任务。相反，本方法的某些实施例侧重于跨模态匹配。列出的参考文献能够学习联合表示，并且可以被用于执行跨模态匹配，但是它是次优的，因为它不能直接优化匹配。另外，它要求两种模态

中的序列具有相同的长度。这是一个严格的限制，这是因为其目标只是将两个模态组合在一起。有关CCA及其扩展技术的更多详细信息可以参阅[2]中的调研。

[0059] 图2A图示了LSTM块。循环神经网络RNN已经被研究了数十年[24]-[26]，并且与已知的卷积神经网络CNN[27]-[30]相比，它们能够存储过去状态的表示。这种机制对于涉及顺序数据的任务是重要的。然而，当对长期动态进行建模时，由于随时间流回的指数变化的误差信号，利用循环策略的网络很容易遇到梯度消失或梯度爆炸的问题。长短期记忆LSTM[18]架构通过包含存储单元和三个门来解决此问题。这些实现学习何时忘记过去的状态以及何时在给定新的输入信息的情况下更新当前状态。LSTM的基本结构如图2A中所示，对应的公式如下：

[0060] $i_t=\sigma(W^{(i)}H)$， (8)

[0061] $o_t=\sigma(W^{(o)}H)$， (9)

[0062] $f_t=\sigma(W^{(f)}H)$， (10)

[0063] $m_t=\phi(W^{(m)}H)$， (11)

[0064] $c_t=f_t\odot c_{t-1}+i_t\odot m_t$， (12)

[0065] $h_t=o_t\odot\phi(c_t)$， (13)

[0066] 其中i_t、o_t、f_t分别是输入门、输出门、遗忘门，并且它们被设计用于控制来自不同连接的误差流。标记m_t也可以被看作是特殊门，其被用于调制输入。H是当前单元的输入和先前隐藏单元的输出的串接。四个门的权重分别被表示为$W^{(i)}$、$W^{(o)}$、$W^{(f)}$和$W^{(m)}$。标记$\odot$代表逐元素乘积，并且存储单元是两部分的总和，先前记忆由f_t调制，并且当前输入由i_t调制。标记h_t是当前隐藏单元的输出。LSTM中常用的两个激活函数是逻辑sigmoid函数和双曲正切函数，其分别被表示为σ和ϕ。

[0067] 由于在序列到序列学习中的强大功能，LSTM已经广泛地被用于例如语音识别[31]、机器翻译[32]、图像字幕[33]和动作识别[34]。

[0068] 在过去的几年中，LSTM已经广泛地被应用于序列到序列学习问题。通常，如[33]中所述，此类任务可以被分为三组，即顺序输入与固定输出、固定输入与顺序输出、顺序输入与顺序输出。本发明的实施例针对第三个任务，其已经被证明是最具挑战性的任务。由于输入时间步长通常与输出时间步长不同，因此学习每个时间步长的从输入到输出的直接映射是有挑战性的。为了解决这个问题，最近的工作[35]、[36]，[36]利用编码器-解码器策略，参见图4A。编码器被用于生成序列的固定长度表示，其以后可以由解码器展开为任何长度的其他序列。

[0069] 编码器-解码器类型的解决方案可以被应用于诸如语音识别、机器翻译等任务。

[0070] 图2B示出了根据本发明的至少一些实施例的示例系统。上方部分被用于对一个序列x进行建模，并且左下部分被用于对另一序列y进行建模。两个序列的输出通过相关函数210进行交互。相关函数210可以被配置为生成相关系数作为输出，其在图2B中被示出为向下的箭头。在此，神经网络的输入是两个序列$\{x_1,x_2,\ldots,x_m\}$和$\{y_1,y_2,\ldots,y_n\}$，输出是它们的相关系数。输入序列可以具有不同的模态。尽管在图2B中示出了LSTM块，但是其他类型的RNN块在本发明的不同实施例中可以是可用的。本发明的原理在本文中被称为深度循环相关神经网络DRCNN。

[0071] 通常，DRCNN学习针对每个输入序列的非线性投影，同时保留其内的内部依存性。

令{$x_1,x_2,\ldots,x_m$}为第一个视图中的输入序列，$\mathbf{x}_i \in \mathbb{R}^{d_1}$是第i个时间步长的特征向量，令{$y_1,y_2,\ldots,y_n$}为第二个视图中的对应输入序列，其中$\mathbf{y}_j \in \mathbb{R}^{d_2}$是第j个时间步长的特征向量。换言之，输入序列不需要具有相同的长度。由于LSTM单元($\theta_1,\theta'_1,\theta_2,\theta'_2$)的权重通过时间共享，因此第一视图(m)中的时间步进数目不必等于第二视图(n)中的时间步数。假定f_1和f_2是将每个序列的输入映射到输出的函数，然后两个序列z_1和z_2的输出可以被表示为：

[0072] $z_1=f_1(x_1,x_2,\ldots,x_m,\theta_1,\theta'_1)$， (14)

[0073] $z_2=f_2(y_1,y_2,\ldots,y_n,\theta_2,\theta'_2)$. (15)

[0074] 由于我们需要计算数据的两个视图之间的相关性，因此它们的维度必须被绘制为相同的。这样，线性变换可以被用于每个LSTM序列的顶部，如下：

[0075] $$\mathbf{p}_1=\boldsymbol{\alpha}_1^T\mathbf{z}_1, \quad (16)$$

[0076] $$\mathbf{p}_2=\boldsymbol{\alpha}_2^T\mathbf{z}_2. \quad (17)$$

[0077] 在此p_1和p_2可以被视为两个序列的表示。它们的维度由参数α_1和α_2控制。假定$\mathbf{p}_1 \in \mathbb{R}^d$， $\mathbf{p}_2 \in \mathbb{R}^d$，并且样本总数为M。然后，用于相关性学习的数据的两个视图可以分别被表示为$\mathbf{P}_1^* = \mathbb{R}^{d\times M}$和$\mathbf{P}_2^* = \mathbb{R}^{d\times M}$。这样，训练DRCNN的目标可以被写成：

[0078] $$\rho = \max_{\boldsymbol{\theta}_1,\boldsymbol{\theta}'_1,\boldsymbol{\theta}_2,\boldsymbol{\theta}'_2} corr(\mathbf{P}_1^*,\mathbf{P}_2^*), \quad (18)$$

[0079] 也即，使两个视图的表示的相关性最大化。为了训练DRCNN，可以遵循网络训练中的一般实践，并且可以采用小批量随机梯度下降法。DRCNN框架可以包括两个关键组件，即序列学习组件和相关性学习组件，其可以分别被优化。

[0080] 关于首先优化相关性学习组件，假定训练批量大小为N，并且$\mathbf{P}_1 = (\mathbf{p}_1^{(1)},\mathbf{p}_1^{(2)},\ldots,\mathbf{p}_1^{(N)}) \in \mathbb{R}^{d\times N}$和$\mathbf{P}_2 = (\mathbf{p}_2^{(1)},\mathbf{p}_2^{(2)},\ldots,\mathbf{p}_2^{(N)}) \in \mathbb{R}^{d\times N}$是小批量中来自两个视图的数据。$\mathbf{1} \in \mathbb{R}^{N\times N}$是全1矩阵。然后，$p_1$和$p_2$的中心数据矩阵可以分别被表示为$\bar{\mathbf{P}}_1 = \mathbf{P}_1 - \frac{1}{N}\mathbf{P}_1\mathbf{1}$和$\bar{\mathbf{P}}_2 = \mathbf{P}_2 - \frac{1}{N}\mathbf{P}_2\mathbf{1}$。令$\boldsymbol{\Sigma}_{11} = \frac{1}{N-1}\bar{\mathbf{P}}_1\bar{\mathbf{P}}'_1$和$\boldsymbol{\Sigma}_{22} = \frac{1}{N-1}\bar{\mathbf{P}}_2\bar{\mathbf{P}}'_2$为两个视图的集合内协方差矩阵，并且令$\boldsymbol{\Sigma}_{12} = \frac{1}{N-1}\bar{\mathbf{P}}_1\bar{\mathbf{P}}'_2$为对应的集合间协方差矩阵。定义$\mathbf{S} = \boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$，并且根据[5]，等式(16)等同于：

[0081] $\rho=||S||_{tr}=tr(S'S)^{1/2}$. (19)

[0082] 假定S的特征值分解为S＝UDV′，然后类似于深度CCA的优化(参见[5]的附录)，可以关于p_1计算corr(P_1,P_2)的梯度：

[0083] $$\frac{\partial corr(\mathbf{P}_1,\mathbf{P}_2)}{\partial \mathbf{P}_1} = \frac{1}{N-1}(2\boldsymbol{\nabla}_{11}\bar{\mathbf{P}}_1 + \boldsymbol{\nabla}_{12}\bar{\mathbf{P}}_2), \quad (20)$$

[0084] 其中

[0085] $$\boldsymbol{\nabla}_{11} = -\frac{1}{2}\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}\mathbf{U}\mathbf{D}\mathbf{U}'\boldsymbol{\Sigma}_{11}^{-\frac{1}{2}}, \quad (21)$$

[0086] 以及

[0087] $$\nabla_{12} = -\frac{1}{2}\Sigma_{11}^{-\frac{1}{2}}UV'\Sigma_{22}^{-\frac{1}{2}}. \quad (22)$$

[0088] 由于采用小批量随机梯度下降方法来训练网络，因此训练误差应当向后传播到每个样本，并且corr(P1,P2)关于$\mathbf{p}_1^{(i)}$的梯度可以被计算为：

[0089] $$\frac{\partial corr(\mathbf{P}_1,\mathbf{P}_2)}{\partial \mathbf{p}_1^{(i)}} = \frac{1}{N-1}\left(2\nabla_{11}(\bar{\mathbf{P}}_1)_{(:,i)} + \nabla_{12}(\bar{\mathbf{P}}_2)_{(:,i)}\right) \quad (23)$$

[0090] 其中$(\bar{\mathbf{P}}_1)_{(:,i)}$和$(\bar{\mathbf{P}}_2)_{(:,i)}$分别是$\bar{\mathbf{P}}_1$和$\bar{\mathbf{P}}_2$的第i列。

[0091] 一旦$\frac{\partial corr(\mathbf{P}_1,\mathbf{P}_2)}{\partial \mathbf{p}_1^{(i)}}$被计算，我们就可以相对于每个LSTM单元的输入来计算目标函数的梯度。注意到，在两个视图中优化目标的方法可能几乎相同。因此，为简洁起见，我们仅在第一个视图中示出优化过程。

[0092] 然后考虑优化序列学习组件，从等式(8)至(14)，我们可以看到序列学习中用于学习所需的参数恰好是四个门的权重。假定第一个视图中最后的时间步长的第i个样本的输出为h_t(即，$\mathbf{h}_t = \mathbf{z}_1^{(i)}$)，而最后的时间步长的对应输入为H，然后我们可以可视化数据流，如图4B所示。因此，可以使用链式规则如下计算h_t关于H的梯度(参见图4B)：

[0093] $$\frac{\partial \mathbf{h}_t}{\partial \mathbf{H}} = \frac{\partial h_t}{\partial o_t}\frac{\partial o_t}{\partial \mathbf{H}} + \frac{\partial h_t}{\partial c_t}\frac{\partial c_t}{\partial \mathbf{H}} = \frac{\partial h_t}{\partial o_t}\frac{\partial o_t}{\partial \mathbf{H}} + \frac{\partial h_t}{\partial c_t}\left(\frac{\partial c_t}{\partial m_t}\frac{\partial m_t}{\partial \mathbf{H}} + \frac{\partial c_t}{\partial i_t}\frac{\partial i_t}{\partial \mathbf{H}} + \frac{\partial c_t}{\partial f_t}\frac{\partial f_t}{\partial \mathbf{H}}\right) \quad (24)$$

[0094] 从LSTM的公式(等式(8)至(13))，可以计算等式(21)中每个部分的梯度，如下(有关梯度得出的参考也可以在[39]中找到)：

[0095] $$\frac{\partial h_t}{\partial o_t} = \emptyset(c_t),\ \frac{\partial h_t}{\partial c_t} = o_t \odot \emptyset'(c_t), \quad (25)$$

[0096] $$\frac{\partial c_t}{\partial m_t} = i_t,\ \frac{\partial c_t}{\partial i_t} = m_t,\ \frac{\partial c_t}{\partial f} = c_{t-1}, \quad (26)$$

[0097] $$\frac{\partial o_t}{\partial H} = \sigma'(\mathbf{W}^{(o)}\mathbf{H})\mathbf{W}^{(o)},\ \frac{\partial \mathbf{m}_t}{\partial \mathbf{H}} = \sigma'(\mathbf{W}^{(m)}\mathbf{H})\mathbf{W}^{(m)}, \quad (27)$$

[0098] $$\frac{\partial i_t}{\partial \mathbf{H}} = \sigma'(\mathbf{W}^{(i)}\mathbf{H})\mathbf{W}^{(i)},\ \frac{\partial f_t}{\partial \mathbf{H}} = \sigma'(\mathbf{W}^{(f)}\mathbf{H})\mathbf{W}^{(f)}. \quad (28)$$

[0099] 从图2B可以看出，最后时间步长的序列输出正好是相关性学习组件的输入。因此，可以容易地计算最终目标函数关于H的梯度，如下：

[0100] $$\frac{\partial corr(\mathbf{P}_1,\mathbf{P}_2)}{\partial \mathbf{H}} = \frac{\partial corr(\mathbf{P}_1,\mathbf{P}_2)}{\partial \mathbf{p}_1^{(i)}}\frac{\partial \mathbf{p}_1^{(i)}}{\partial \mathbf{z}_1^{(i)}}\frac{\partial \mathbf{z}_1^{(i)}}{\partial \mathbf{H}}, \quad (29)$$

[0101] 其中$\frac{\partial corr(\mathbf{P}_1,\mathbf{P}_2)}{\partial \mathbf{p}_1^{(i)}}$可以如等式(23)中被计算，$\frac{\partial \mathbf{p}_1^{(i)}}{\partial \mathbf{z}_1^{(i)}}$可以很容易地被计算，因为$\mathbf{p}_1^{(i)}$是$\mathbf{z}_1^{(i)}$的线性变换，如等式(16)中所示，并且$\frac{\partial \mathbf{z}_1^{(i)}}{\partial \mathbf{H}}$可以如等式(24)中被计算(为$\mathbf{h}_t = \mathbf{z}_1^{(i)}$)。

由于四个门的参数通过时间共享，因此可以以完全相同的方式使用链规则来容易地计算先前时间步长的梯度。

[0102] 到目前为止，我们已经介绍了DRCNN如何被训练，即所涉及的参数如何被优化。一旦训练完成，就可以通过计算p_1和p_2的内积来计算两个序列的相关系数(参见等式(16)和等式(17))。

[0103] 可以在测试数据中采用DRCNN来确定其功能如何。利用训练好的DRCNN，不同的应用程序场景变得可用，诸如：

[0104] 文本到视频的匹配，其可以促进复杂查询视频搜索。那意味着查询可以是描述性语句，而不是一个或两个简单的单词。视频可以通过计算匹配分数而被排名。

[0105] 视频字幕，其引起了学术界和工业界的兴趣。那意味着，我们为给定的视频生成描述性语句。我们可以通过计算其与视频的匹配分数来对若干候选语句进行排名。

[0106] 视频到视频的匹配，其可以被用于视频推荐和分类。如何计算两个视频的距离也是问题。由于DRCNN可以生成用于视频的固定维度表示，因此我们可以根据该表示来估计视频之间的距离，并且该距离很好地探索了视频的语义和时间信息。

[0107] 在此我们仅讨论视频，但是该方案也可以被扩展到其他媒体文档。实际上，在各种应用中，输入序列可以包括飞行器系统输出，该飞行器系统输出可以与飞行器控制状态相关，以例如警告飞行员飞行器已经进入或正在进入失速或可控制性降低的状态。

[0108] 现在呈现第一应用场景的初步结果，即针对复杂查询视频搜索的初步结果。发明方使用了MSR-VTT数据集[37]，其包含200k个视频语句对。8k个对被随机选择用于实验。4k个对被用于训练DRCNN，并且其余的被用于视频搜索实验。在视频搜索实验中，100个语句被选择并且被用作查询以执行搜索，以及然后检查平均精度AP性能。观察到MAP测量值为0.35。为了比较，本方法与CCA和DCCA方法被比较。在这两种方法中，视频和文本表示通过平均视频帧和单词的特征而被获得。显然，这两种方法不能很好地探索序列的时间信息。结果是，通过CCA和DCCA方法获得的MAP测量值分别为0.19和0.23。因此，DRCNN方法能够明显提高性能。

[0109] 图3示出了能够支持本发明的至少一些实施例的示例设备。示出了设备300，其可以包括例如服务器、台式计算机、膝上型计算机或平板计算机。设备300中包括处理器310，该处理器310可以包括例如单核或多核处理器，其中单核处理器包括一个处理核，并且而多核处理器包括多于一个处理核心。处理器310可以包括多于一个处理器。处理核心可以包括例如由ARM控股制造的Cortex-A8处理核或由超微半导体公司生产的Steamroller处理核心。处理器310可以包括高通骁龙、英特尔凌动和/或英特尔至强处理器中的至少一项。处理器310可以包括至少一个专用集成电路ASIC。处理器310可以包括至少一个现场可编程门阵列FPGA。处理器310可以是用于执行设备300中的方法步骤的部件。处理器310可以至少部分地由计算机指令配置为执行动作。

[0110] 设备300可以包括存储器320。存储器320可以包括随机存取存储器和/或永久性存储器。存储器320可包括至少一个RAM芯片。存储器320可以包括例如固态、磁性、光学和/或全息存储器。存储器320可以是处理器310至少部分可访问的。存储器320可以至少部分被包括在处理器310中。存储器320可以是用于存储信息的部件。存储器320可以包括计算机指令，处理器310被配置为执行该计算机指令。当被配置为使处理器310执行某些动作的计算

机指令被存储在存储器320中，并且设备300整体被配置为使用来自存储器320的计算机指令来在处理器310的指导下运行时，处理器310和/或其至少一个处理核可用被认为被配置为执行所述某些动作。存储器320可以至少部分地被包括在处理器310中。存储器320可以至少部分地在设备300外部，但是对于设备300是可访问的。

[0111] 设备300可以包括发射器330。设备300可以包括接收器340。发射器330和接收器340可以被配置为分别根据至少一种蜂窝或非蜂窝标准来发送和接收信息。发射器330可以包括多于一个发射器。接收器340可以包括多于一个接收器。发射器330和/或接收器340可以被配置成根据以下项来操作：例如，全球移动通信系统GSM、宽带码分多址WCDMA、长期演进、LTE、IS-95、无线局域网WLAN、以太网和/或全球微波接入互操作性WiMAX标准。

[0112] 设备300可以包括近场通信NFC收发器350。NFC收发器350可以支持至少一种NFC技术，诸如NFC、蓝牙、低功耗蓝牙或类似技术。

[0113] 设备300可以包括用户接口UI 360。UI 360可以包括以下中的至少一项：显示器、键盘、触摸屏、被布置为通过使设备300振动来向用户发信号的振动器、扬声器以及麦克风。用户可用能够经由UI 360来操作设备300，例如以配置机器学习参数。

[0114] 设备300可以包括或被布置为接受用户身份模块370。用户身份模块370可以包括例如订户身份模块SIM、可安装在设备300中的卡。用户身份模块370可以包括标识设备300的用户的订阅的信息。用户身份模块370可以包括密码信息，该密码信息可用于验证设备300的用户的身份和/或有助于对所通信的信息的加密以及对设备300的用户经由设备300所产生的通信的计费。

[0115] 处理器310可以被配备有发射器，该发射器被布置为经由设备300内部的电引线从处理器310向设备300中所包括的其他设备输出信息。这样的发射器可以包括串行总线发射器，该串行总线发射器例如被布置为经由至少一根电引线向存储器320输出信息以用于其中的存储装置。作为串行总线的替代，发射器可以包括并行总线发射器。同样地，处理器310可以包括接收器，该接收器被布置为经由设备300内部的电引线从设备300中包括的其他设备接收处理器310中的信息。这样的接收器可以包括串行总线接收器，该串行总线接收器被布置为例如经由至少一根电引线从接收器340接收信息以用于处理器310中的处理。作为对串行总线的备选，该接收器可以包括并行总线接收器。

[0116] 设备300可以包括图3中未示出的另外的设备。例如，在设备300包括智能电话的情况下，其可以包括至少一个数字照相机。一些设备300可以包括背面照相机和正面照相机，其中背面照相机可以意在用于数字摄影，并且正面照相机用于视频电话。设备300可以包括指纹传感器，该指纹传感器被布置为至少部分地认证设备300的用户。在一些实施例中，设备300缺少上述至少一个设备。例如，一些设备300可能缺少NFC收发器350和/或用户身份模块370。

[0117] 处理器310、存储器320、发射器330、接收器340、NFC收发器350、UI 360和/或用户身份模块370可以通过设备300内部的电引线以多种不同方式被互连。例如，前述设备中的每个设备可以分别被连接到设备300内部的主总线，以允许设备交换信息。然而，如本领域技术人员将理解的，这仅是一个示例，并且根据实施例，在不脱离本发明的范围的情况下，可以选择互连前述设备中的至少两个设备的各种方式。

[0118] 图4A示出了以上讨论的编码器-解码器架构。在该术语中，上面的网络对应于“编

码器”部分，下面的网络对应于“解码器”部分。

[0119] 图4B示出了一个LSTM块中的数据流。输出向量h_t受输出门向量o_t和单元状态向量c_t所影响。单元状态向量c_t又受以下中的一项所影响：输入门向量i_t、遗忘门向量f_t以及特殊门m_t。

[0120] 图5是根据本发明的至少一些实施例的方法的流程图。所示出的方法的阶段可以在被配置为根据本发明来执行关联的设备(诸如计算机或服务器)中被执行。

[0121] 阶段510包括从第一顺序输入获得来自第一循环神经网络的第一输出，该第一顺序输入具有第一模态。阶段520包括从第二顺序输入获得来自第二循环神经网络的第二输出，该第二顺序输入具有第二模态。最后，阶段530包括处理第一输出和第二输出以获得第一顺序输入和第二顺序输入的相关性。该处理可以包括获得内积。

[0122] 将理解，所公开的本发明的实施例不限于本文所公开的特定结构、工序或材料，而是扩展至其等同物，如相关领域的普通技术人员将认识到的。还应当理解，本文采用的术语仅被用于描述特定实施例的目的，而非旨在限制。

[0123] 贯穿本说明书中对一个实施例或实施例的引用意味着结合该实施例所述的特定特征、结构或特性被包括在本发明的至少一个实施例中。因此，贯穿该说明书中各处出现的短语“在一个实施例中”或“在实施例中”并不一定全部是指同一实施例。在使用诸如例如大约或基本上的术语来参考数值的情况下，也公开了确切的数值。

[0124] 如在本文中所使用的，出于方便起见，可以在公共列表中呈现多个项、结构要素、组成要素和/或材料。然而，这些列表应当被解释为好像列表的每个成员都被个体地标识为单独且唯一的成员。因此，仅基于其在公共组中的呈现而不具有相反的指示，此类列表的任何个体成员都不应当被解释为相同列表的任何其他成员的实际上的等同物。另外，在本文中可以参考本发明的各种实施例和示例，以及用于其各种组件的备选方案。应当理解，这样的实施例、示例和备选方案不被理解为彼此的实际上等同，而是被认为是本发明的独立和自主的表示。

[0125] 此外，在一个或多个实施例中，所描述的特征、结构或特性可以以任何合适的方式被组合。在先前的描述中，提供了许多具体细节，诸如长度、宽度、形状等的示例，以提供对本发明实施例的透彻理解。然而，相关领域的技术人员将认识到，本发明可以在不具有具体细节中的一项或多项或利用其它方法、组件、材料的情况下被实践。在其他情况下，公知的结构、材料或操作未被详细示出或描述，以避免混淆本发明的方面。

[0126] 尽管前述示例在一个或多个特定应用中说明了本发明的原理，但是对于本领域的普通技术人员而言明显的是，在不进行创造性能力劳动，并且不脱离本发明的原理和概念的情况下，在形式、使用和实现细节方面进行许多修改。相应地，除了由下面提出的权利要求书之外，不旨在限制本发明。

[0127] 动词“包括(to comprise)”和“包括(to include)”在本文档中用作开放的限制，既不排除也不要求未叙述的特征的存在。除非另外明确说明，否则从属权利要求中阐述的特征可以相互自由组合。此外，应当理解，在贯穿本文档中使用“一个(a)”或“一个(an)”，即单数形式，并不排除多个。

[0128] 工业适用性

[0129] 本发明的至少一些实施例在使用人工神经网络来处理顺序输入数据方面找到了

工业应用。

[0130] 缩写表

[0131] CCA 典型相关分析

[0132] CNN 卷积神经网络

[0133] DRCNN 深度循环相关神经网络

[0134] KCCA 核CCA

[0135] RNN 循环神经网络

[0136] 附图标记表

[0137]

102	第一输入序列
103	第二输入序列
210	相关函数
300至370	图3的装置的结构
510至530	图5的方法的阶段

[0138] 引文列表

[0139] [1]H.Hotelling,“Relations between two sets of variates,”Biometrika,vol.28,no.3/4,pp.321-377,1936.

[0140] [2]D.R.Hardoon,S.Szedmak,and J.Shawe-Taylor,“Canonical correlation analysis:An overview with application to learning methods,”Neural computation,vol.16,no.12,pp.2639-2664,2004.

[0141] [3]S.Akaho,“A kernel method for canonical correlation analysis,”arXiv preprint cs/0609071,2006.

[0142] [4]F.R.Bach and M.I.Jordan,“Kernel independent component analysis,”Journal of Machine Learning Research,vol.3,no.Jul,pp.1-48,2002.

[0143] [5]G.Andrew,R.Arora,J.A.Bilmes,and K.Livescu,“Deep canonical correlation analysis.”in Proc.ICML,2013,pp.1247-1255.

[0144] [6]F.Yan and K.Mikolajczyk,“Deep correlation for matching images and text,”in Proc.CVPR,2015,pp.3441-3450.

[0145] [7]Y.Verma and C.Jawahar,“Im2text and text2im:Associating images and texts for cross-modal retrieval.”in Proc.BMVC,vol.1,2014,p.2.

[0146] [8]J.C.Pereira,E.Coviello,G.Doyle,N.Rasiwasia,G.R.Lanckriet,R.Levy,and N.Vasconcelos,“On the role of correlation and abstraction in cross-modal multimedia retrieval,”IEEE Transactions on Pattern Analysis and Machine Intelligence,vol.36,no.3,pp.521-535,2014.

[0147] [9]Y.Gong,Q.Ke,M.Isard,and S.Lazebnik,“A multi-view embedding space for modeling internet images,tags,and their semantics,”International Journal of Computer Vision,vol.106,no.2,pp.210-233,2014.

[0148] [10]K.Choukri and G.Chollet,“Adaptation of automatic speech recognizers to new speakers using canonical correlation analysis techniques,”Computer Speech&Language,vol.1,no.2,pp.95-107,1986.

[0149] [11]R.Arora and K.Livescu,“Multi-view cca-based acoustic features for phonetic recognition across speakers and domains,”in Proc.ICASSP,2013,pp.7135-7139.

[0150] [12]C.Chapdelaine,V.Gouaillier,M.Beaulieu,and L.Gagnon,“Improving video captioning for deaf and hearing-impaired people based on eye movement and attention overload,”in Electronic Imaging 2007,2007,pp.64 921K-64 921K.

[0151] [13]M.Kahn,“Consumer video captioning system,”Mar.5 2002,US Patent App.10/091,098.

[0152] [14]M.E.Sargin,Y.Yemez,E.Erzin,and A.M.Tekalp,“Audiovisual synchronization and fusion using canonical correlation analysis,”IEEE Transactions on Multimedia,vol.9,no.7,pp.1396-1403,2007.

[0153] [15]H.Bredin and G.Chollet,“Audio-visual speech synchrony measure for talking-face identity verification,”in Proc.ICASSP,vol.2,2007,pp.II-233.

[0154] [16]J.Yuan,Z.-J.Zha,Y.-T.Zheng,M.Wang,X.Zhou,and T.-S.Chua,“Learning concept bundles for video search with complex queries,”in Proc.ACM MM,2011,pp.453-462.

[0155] [17]D.Lin,S.Fidler,C.Kong,and R.Urtasun,“Visual semantic search:Retrieving videos via complex textual queries,”in Proc.CVPR,2014,pp.2657-2664.

[0156] [18]S.Hochreiter and J.Schmidhuber,“Long short-term memory,”Neural computation,vol.9,no.8,pp.1735-1780,1997.

[0157] [19]P.L.Lai and C.Fyfe,“Kernel and nonlinear canonical correlation analysis,”International Journal of Neural Systems,vol.10,no.05,pp.365-377,2000.

[0158] [20]K.Fukumizu,F.R.Bach,and A.Gretton,“Statistical consistency of kernel canonical correlation analysis,”Journal of Machine Learning Research,vol.8,no.Feb,pp.361-383,2007.

[0159] [21]J. V´ıa, I. Santamar´ıa, and J.P′erez,“A learning algorithm for adaptive canonical correlation analysis of several data sets,”Neural Networks,vol.20,no.1,pp.139-152,2007.

[0160] [22]A.Sharma,A.Kumar,H.Daume,and D.W.Jacobs,“Generalized multiview analysis:A discriminative latent space,”in Proc.CVPR,2012,pp.2160-2167.

[0161] [23]Y.Yamanishi,J.-P.Vert,A.Nakaya,and M.Kanehisa,“Extraction of correlated gene clusters from multiple genomic data by generalized kernel canonical correlation analysis,”Bioinformatics,vol.19,no.suppl 1,pp.i323-i330,2003.

[0162] [24]F.J.Pineda,“Generalization of back-propagation to recurrent neural networks,”Physical review letters,vol.59,no.19,p.2229,1987.

[0163] [25]P.J.Werbos,“Generalization of backpropagation with application to

a recurrent gas market model,"Neural Networks,vol.1,no.4,pp.339-356,1988.

[0164] [26]R.J.Williams and D.Zipser,"Gradient-based learning algorithms for recurrent networks and their computational complexity,"Backpropagation: Theory,architectures and applications,pp.433-486,1995.

[0165] [27]A.Krizhevsky,I.Sutskever,and G.E.Hinton,"Imagenet classification with deep convolutional neural networks,"in Proc.NIPS,2012.

[0166] [28]C.Szegedy,W.Liu,Y.Jia,P.Sermanet,S.Reed,D.Anguelov,D.Erhan, V.Vanhoucke,and A.Rabinovich,"Going deeper with convolutions,"in Proc.CVPR, 2015.

[0167] [29]K.Simonyan and A.Zisserman,"Very deep convolutional networks for large-scale image recognition,"arXiv preprint arXiv:1409.1556,2014.

[0168] [30]K.He,X.Zhang,S.Ren,and J.Sun,"Deep residual learning for image recognition,"arXiv preprint arXiv:1512.03385,2015.

[0169] [31]A.Graves,A.-r.Mohamed,and G.Hinton,"Speech recognition with deep recurrent neural networks,"in Proc.ICASSP,2013,pp.6645-6649.

[0170] [32]I.Sutskever,O.Vinyals,and Q.V.Le,"Sequence to sequence learning with neural networks,"in Advances in neural information processing systems, 2014,pp.3104-3112.

[0171] [33]J.Donahue,L.Anne Hendricks,S.Guadarrama,M.Rohrbach,S.Venugopalan, K.Saenko,and T.Darrell,"Long-term recurrent convolutional networks for visual recognition and description,"in Proc.CVPR,2015,pp.2625-2634.

[0172] [34]J.Yue-Hei Ng,M.Hausknecht,S.Vijayanarasimhan,O.Vinyals,R.Monga, and G.Toderici,"Beyond short snippets:Deep networks for video classification,"in Proc.CVPR,2015,pp.4694-4702.

[0173] [35]W.Zaremba,I.Sutskever,and O.Vinyals,"Recurrent neural network regularization,"arXiv preprint arXiv:1409.2329,2014.

[0174] [36]N.Srivastava,E.Mansimov,and R.Salakhutdinov,"Unsupervised learning of video representations using lstms,"CoRR,abs/1502.04681,vol.2,2015

[0175] [37]J.Xu,T.Mei,T.Yao,Y.Rui."MSR-VTT:A Large Video Description Dataset for Bridging Video and Language,"in Proc.CVPR 2016.

[0176] [38]Nikhil Rasiwasi,Jose Costa Pereira,Emanuele Coviello,Gabriel Doyle,Gert R.G.Lanckriet,Roger Levy,Nuno Vasconcelos,"A New Approach to Cross-Modal Multimedia Retrieval,"in Proc.ACM MM,2010.

[0177] [39]Sepp Hochreiter and Jurgen Schmidhuber,"long short-term memory," Neural Computations,vol.9,no.8,pp.1735-1780,1997.

[0178] [40]Yang et al.,"Deep Multimodal Representation Learning from Temporal Data,"ArXiv,https://arxiv.org/pdf/1704.03152.pdf

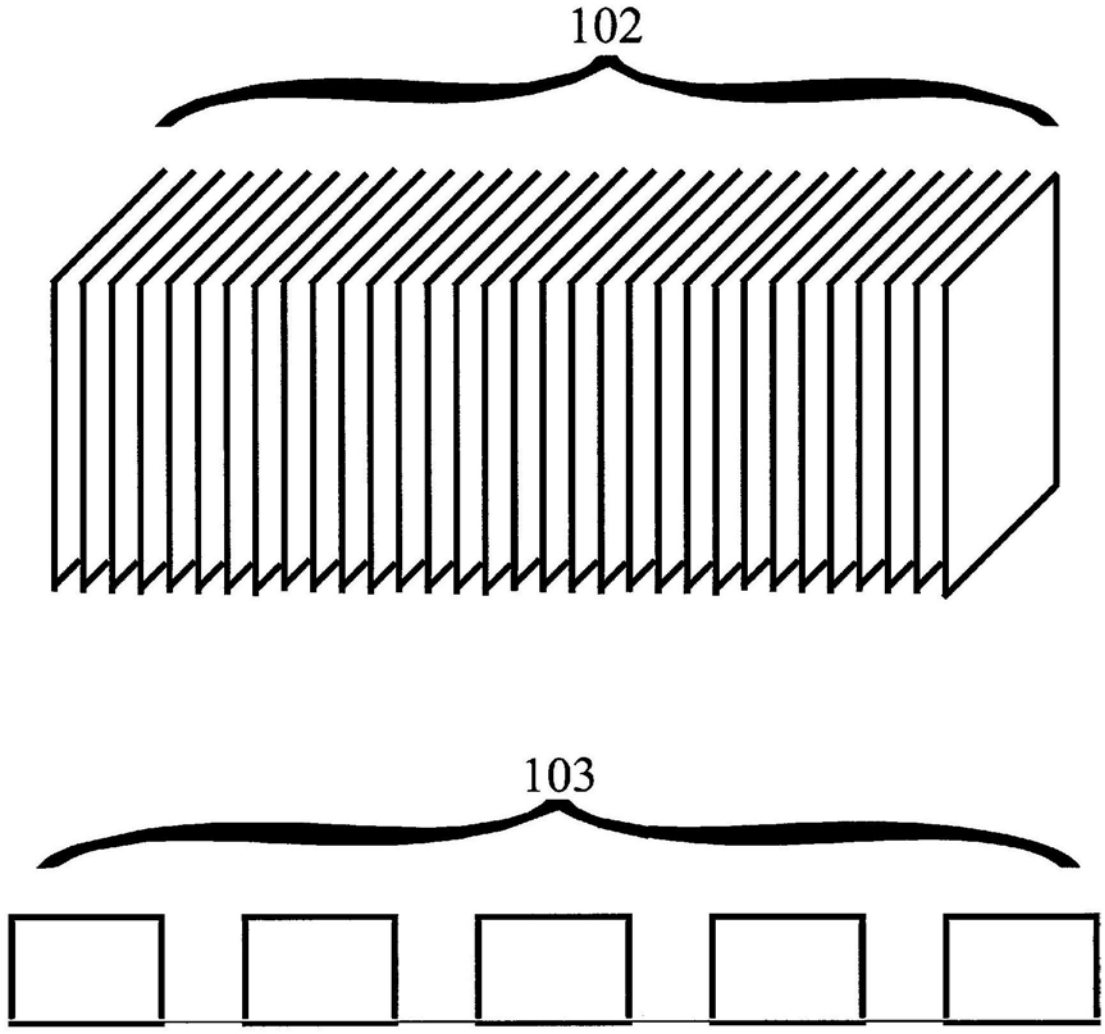


图1

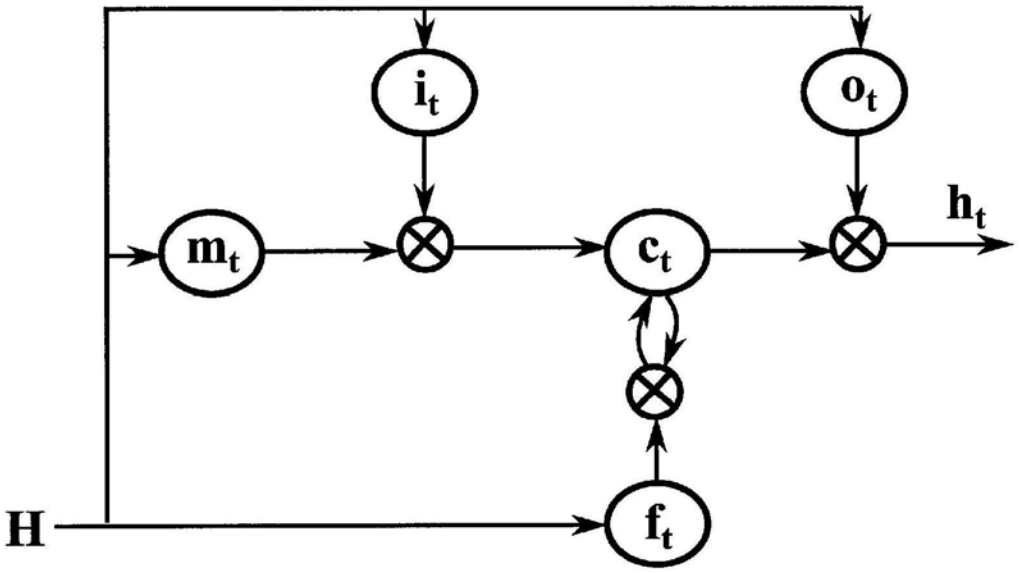

i_t
o_t
h_t
m_t
c_t
H
f_t

图2A

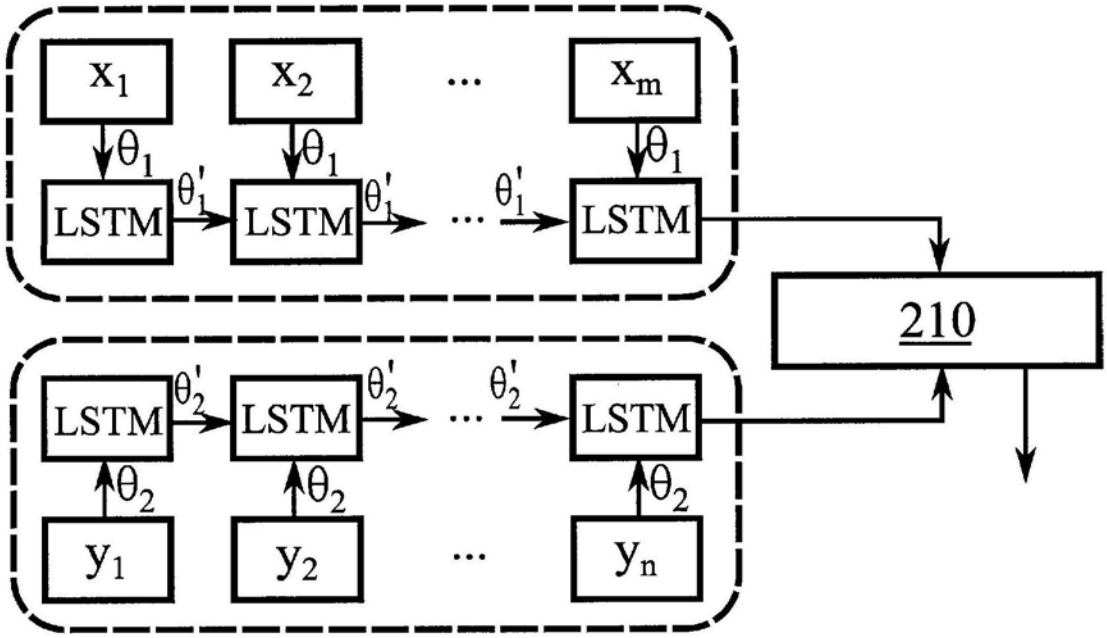

x_1
x_2
x_m
θ_1
θ_1'
LSTM
210
θ_2'
θ_2
y_1
y_2
y_n

图2B

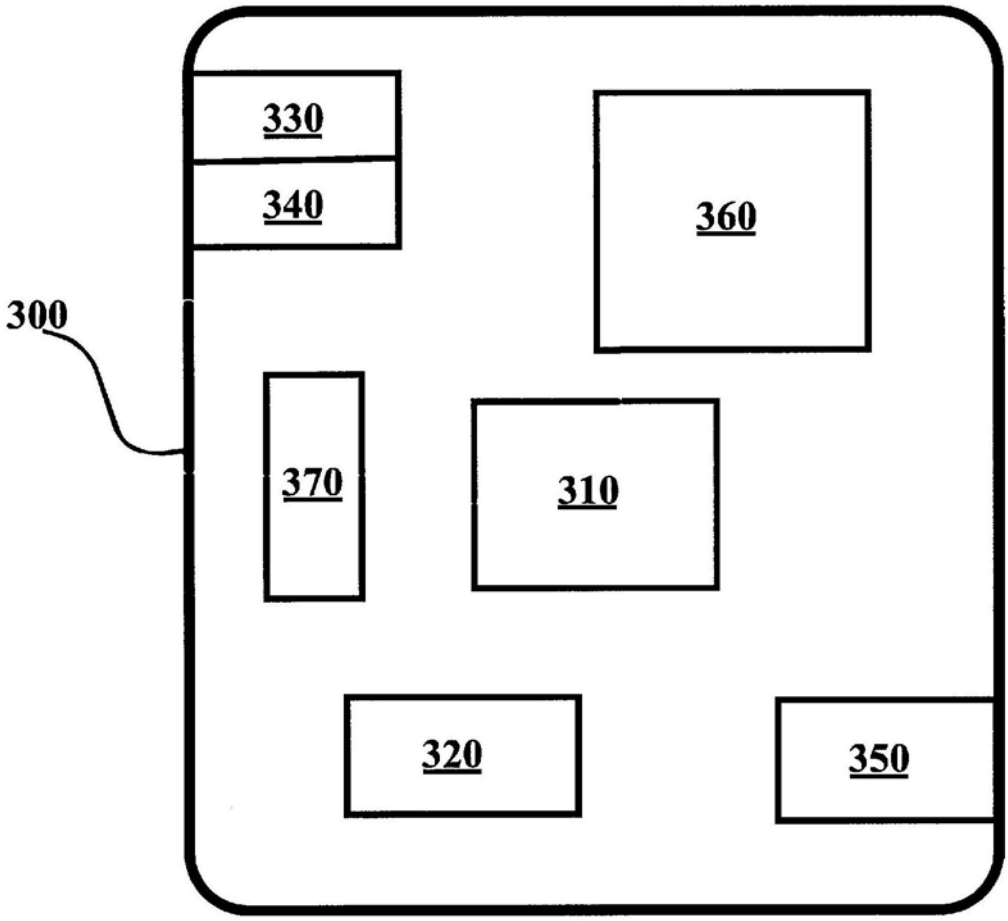

330
340
360
300
370
310
320
350

图3

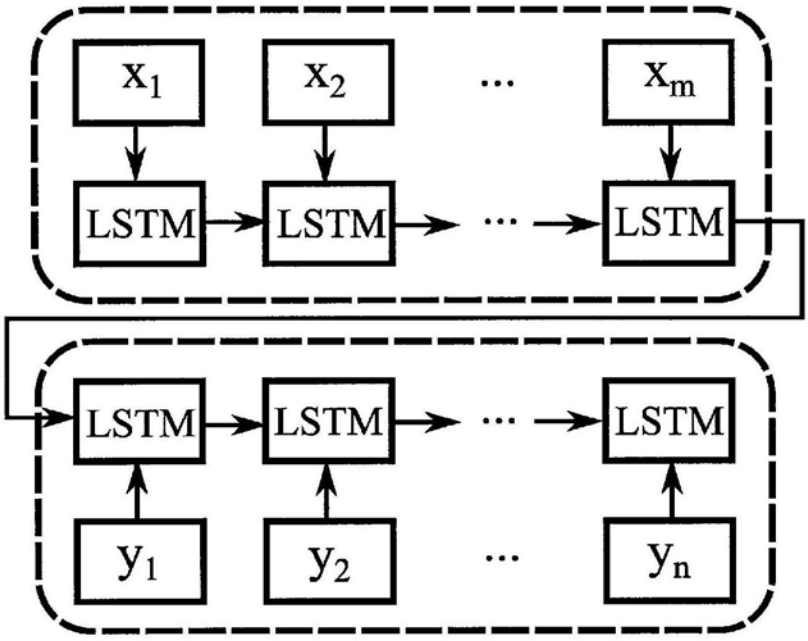

x1
x2
...
xm
LSTM
LSTM
...
LSTM
LSTM
LSTM
...
LSTM
y1
y2
...
yn

图4A

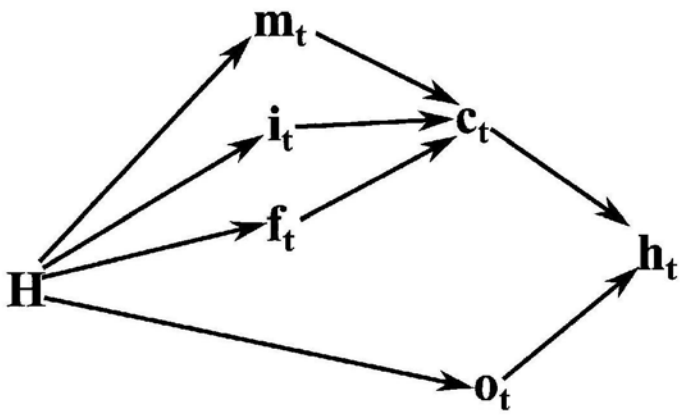

m_t
i_t
c_t
f_t
h_t
H
o_t

图4B

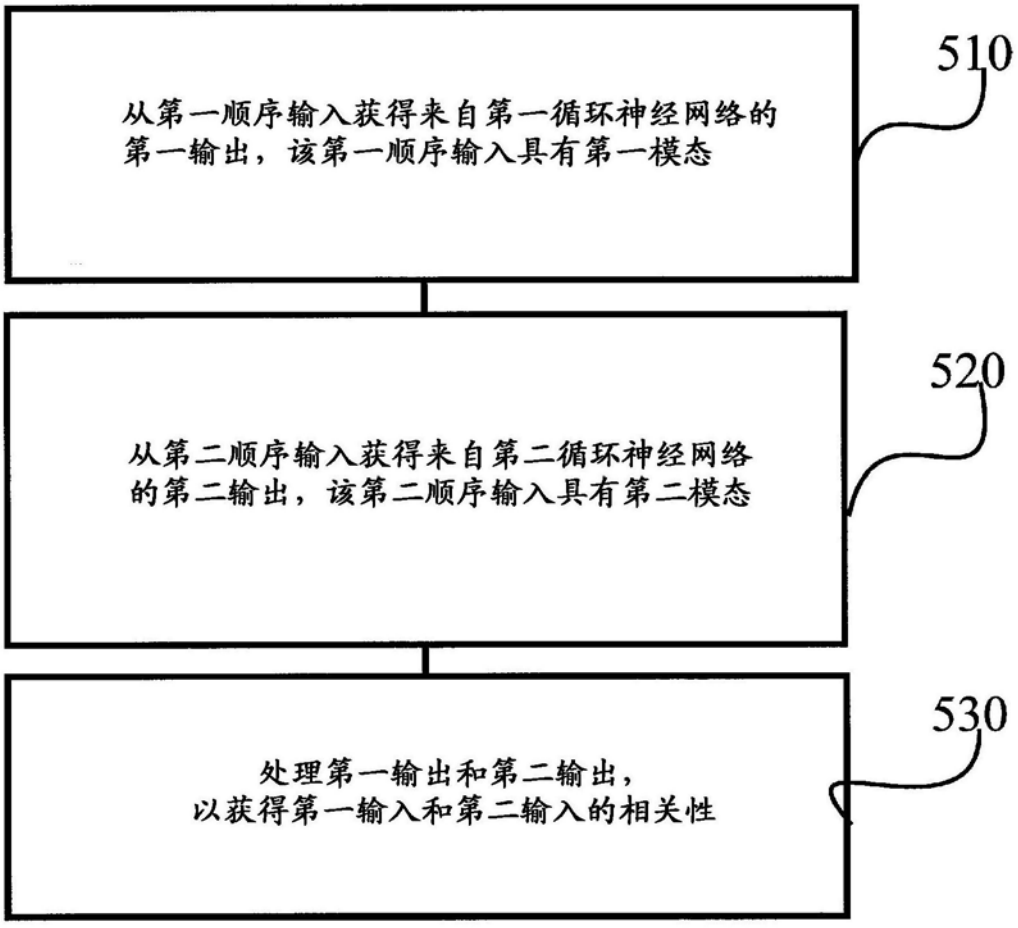

510
从第一顺序输入获得来自第一循环神经网络的第一输出，该第一顺序输入具有第一模态
520
从第二顺序输入获得来自第二循环神经网络的第二输出，该第二顺序输入具有第二模态
530
处理第一输出和第二输出，以获得第一输入和第二输入的相关性

图5