



(12) 发明专利申请

(10) 申请公布号 CN 119096250 A

(43) 申请公布日 2024. 12. 06

(21) 申请号 202380025622.7

(22) 申请日 2023.03.07

(30) 优先权数据

63/317,538 2022.03.07 US

(85) PCT国际申请进入国家阶段日

2024.09.05

(86) PCT国际申请的申请数据

PCT/EP2023/055753 2023.03.07

(87) PCT国际申请的公布数据

W02023/170067 EN 2023.09.14

(71) 申请人 渊慧科技有限公司

地址 英国

(72) 发明人 阿德里亚·雷卡森斯康丁恩特

詹森·加陈·林 王路宇

让-巴普蒂斯特·阿莱拉克

安德鲁·库尔特·耶格莱

若昂·卡雷拉 波利娜·吕克

安托万·米奇

卢卡斯·德弗雷塔斯斯马伊拉

罗斯·赫姆斯莱 安德鲁·西塞曼

(74) 专利代理机构 中原信达知识产权代理有限
责任公司 11219

专利代理师 邓聪惠 周亚荣

(51) Int.Cl.

G06N 3/0455 (2006.01)

G06N 3/096 (2006.01)

G06N 3/0895 (2006.01)

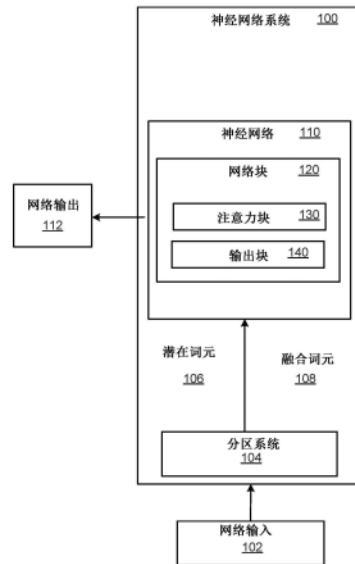
权利要求书3页 说明书15页 附图7页

(54) 发明名称

使用分区注意力处理网络输入

(57) 摘要

用于使用实现分区注意力的神经网络来处理网络输入的方法、系统和装置,包括编码在计算机存储介质上的计算机程序。



1. 一种由一个或多个计算机执行的方法,所述方法包括:

获取第一网络输入;

确定所述第一网络输入的多个不相交的分区;

针对每个分区,从该分区生成该分区的潜在词元的相应集合;

生成融合潜在词元的集合;

使用神经网络处理每个分区的所述潜在词元的相应集合和所述融合潜在词元的集合以生成表征所述第一网络输入的网络输出,

其中,所述神经网络包括神经网络块序列,所述神经网络块序列包括:(i)一个或多个注意力块和(ii)输出块,

其中,每个注意力块执行操作,所述操作包括:

针对每个分区,通过针对该分区应用仅注意该分区的所述潜在词元的相应集合的对应注意力机制来更新该分区的所述潜在词元的相应集合;以及

通过应用对应融合注意力机制来更新所述融合潜在词元的集合,所述对应融合注意力机制注意所述分区中的每个分区的所述潜在词元的相应集合和所述融合潜在词元的集合;以及

其中,所述输出块执行操作,所述操作包括:

在所述潜在词元的相应集合和所述融合潜在词元使用所述一个或多个注意力块被更新之后,处理所述融合潜在词元中的至少一个或多个融合潜在词元以生成表征所述第一网络输入的所述网络输出。

2. 根据权利要求1所述的方法,其中,每个不相交的分区对应于来自多个模态的不同模态。

3. 根据权利要求2所述的方法,其中,所述模态包括图像模态、视频模态或两者。

4. 根据权利要求2或权利要求3所述的方法,其中,所述模态包括音频模态。

5. 根据权利要求2至4中任一项所述的方法,其中,所述模态包括文本模态。

6. 根据权利要求2至5中任一项所述的方法,进一步包括:

接收仅包括所述多个模态中的第一模态的第二输入;

生成与所述第一模态相对应的分区的潜在词元的集合;

使用所述神经网络处理与所述第一模态相对应的所述分区的所述潜在词元的集合,包括:

针对每个注意力块,通过针对所述分区应用仅注意所述分区的所述潜在词元的相应集合的对应注意力机制来更新与所述模态相对应的所述分区的所述潜在词元的集合;以及

在所述对应模态的所述潜在词元的集合使用所述一个或多个注意力块被更新之后,处理来自所述对应模态的所述潜在词元的集合的一个或多个潜在词元以生成表征所述第二网络输入的网络输出。

7. 根据任一前述权利要求所述的方法,其中,所述融合潜在词元在所述神经网络的训练期间被学习。

8. 根据任一前述权利要求所述的方法,其中,针对每个分区,从该分区生成该分区的潜在词元的相应集合包括:

使用该分区的编码器神经网络来处理所述第一网络输入的该分区,以生成该分区的所

述潜在词元的相应集合。

9. 根据权利要求8所述的方法, 其中, 针对每个分区, 从该分区生成该分区的潜在词元的相应集合进一步包括:

使用位置编码来修改通过使用所述编码器神经网络处理所述第一网络输入的该分区而生成的每个潜在词元。

10. 根据权利要求8或权利要求9所述的方法, 其中, 针对每个分区, 从该分区生成该分区的潜在词元的相应集合进一步包括:

利用一个或多个经学习的词元来增强通过使用所述编码器神经网络处理所述第一网络输入的该分区而生成的所述潜在词元。

11. 根据任一前述权利要求所述的方法, 其中:

针对每个分区, 该分区的所述对应注意力机制是具有分区特定掩蔽的自注意力机制, 所述分区特定掩蔽将所述注意力机制限制为仅注意该分区的所述潜在词元的相应集合; 以及

针对所述融合潜在词元, 所述对应注意力机制是不被掩蔽的自注意力机制, 以允许所述注意力机制注意所述分区中的每个分区的所述潜在词元的相应集合和所述融合潜在词元的相应集合。

12. 根据权利要求11所述的方法, 其中, 用于所述分区和所述融合潜在词元的所述对应注意力机制共享参数。

13. 根据任一前述权利要求所述的方法, 其中, 处理所述融合潜在词元中的至少一个或多个融合潜在词元以生成表征所述第一网络输入的所述网络输出包括:

根据仅从所述融合潜在词元选择一个或多个潜在词元生成融合网络输出。

14. 根据任一前述权利要求所述的方法, 其中, 处理所述融合潜在词元中的至少一个或多个融合潜在词元以生成表征所述第一网络输入的所述网络输出包括:

根据从所述融合潜在词元和所述分区中的每个分区的所述潜在词元的集合中选择一个或多个潜在词元生成总体网络输出。

15. 根据任一前述权利要求所述的方法, 其中, 处理所述融合潜在词元中的至少一个或多个融合潜在词元以生成表征所述第一网络输入的所述网络输出包括:

生成两个或更多个候选网络输出, 所述两个或更多个候选网络输出包括使用所述融合潜在词元中的一个或多个融合潜在词元生成的至少一个候选网络输出; 以及

组合所述两个或更多个候选网络输出以生成最终网络输出。

16. 一种由一个或多个计算机执行的训练神经网络的方法, 所述方法包括:

获取第一网络输入;

确定所述第一网络输入的多个不相交的分区;

针对每个分区, 从该分区生成该分区的潜在词元的相应集合;

生成融合潜在词元的集合;

使用神经网络处理每个分区的所述潜在词元的相应集合和所述融合潜在词元的集合以生成表征所述第一网络输入的网络输出,

其中, 所述神经网络包括神经网络块序列, 所述神经网络块序列包括: (i) 一个或多个注意力块和(ii) 输出块,

其中,每个注意力块执行操作,所述操作包括:

针对每个分区,通过针对该分区应用仅注意该分区的所述潜在词元的相应集合的对应注意力机制来更新该分区的所述潜在词元的相应集合;以及

通过应用对应融合注意力机制来更新所述融合潜在词元的集合,所述对应融合注意力机制注意所述分区中的每个分区的所述潜在词元的相应集合和所述融合潜在词元的集合;以及

其中,所述输出块执行操作,所述操作包括:

在所述潜在词元的相应集合和所述融合潜在词元使用所述一个或多个注意力块被更新之后,处理所述融合潜在词元中的至少一个或多个融合潜在词元以生成表征所述第一网络输入的所述网络输出;以及

将网络输出与所述第一网络输入的预期网络输出进行比较以确定损失并基于所述损失来训练神经网络。

17.根据权利要求16所述的方法,其中,生成所述分区的所述潜在词元的相应集合包括:利用编码器神经网络来处理所述第一网络输入,所述编码器神经网络被训练以生成每个分区的所述潜在词元的相应集合。

18.根据权利要求17所述的方法,其中,

其中,基于所述损失来训练所述神经网络包括训练所述编码器神经网络。

19.根据权利要求16至18中任一项所述的方法,其中,所述融合潜在词元的集合在所述训练期间被学习,并且其中,基于所述损失来训练所述神经网络包括更新所述融合潜在词元的集合。

20.根据权利要求16至19中任一项所述的方法,进一步包括使用无监督数据集使用自监督学习目标或无监督学习目标对所述神经网络进行预训练。

21.根据权利要求16至20中任一项的方法,其中,确定损失包括使用无监督损失函数或自监督损失函数作为辅助损失函数。

22.一种由一个或多个计算机执行的方法,所述方法包括:

获取第一单模态网络输入;

使用根据权利要求16至21中任一项训练的神经网络来处理所述单模态输入以生成表征所述第一单模态网络输入的网络输出。

23.一种系统,包括一个或多个计算机和一个或多个存储设备,所述一个或多个存储设备存储指令,所述指令在由所述一个或多个计算机执行时,使所述一个或多个计算机执行根据权利要求1至22中任一项所述的相应方法的所述操作。

24.一个或多个计算机存储介质,所述一个或多个计算机存储介质存储指令,所述指令在由一个或多个计算机执行时,使所述一个或多个计算机执行根据权利要求1至22中任一项所述的相应方法的所述操作。

使用分区注意力处理网络输入

[0001] 相关申请的交叉引用

[0002] 本申请要求于2022年3月7日提交的美国临时申请No. 63/317,538的优先权。现有申请的公开内容被认为是本申请的公开内容的一部分并且通过引用并入本申请的公开内容中。

背景技术

[0003] 本说明书涉及使用神经网络处理输入。

[0004] 神经网络是采用一层或多层非线性单元来预测接收到的输入的输出的机器学习模型。一些神经网络除了输出层之外还包括一个或多个隐藏层。每个隐藏层的输出用作网络中的下一层——即,下一个隐藏层或输出层——的输入。网络的每个层根据参数的相应集合的当前值输入从接收到的输入生成输出。

发明内容

[0005] 本说明书描述了一种被实现为一个或多个位所中的一个或多个计算机上的计算机程序的系统,该系统处理被分区为多个不相交分区的网络输入以生成用于机器学习任务的网络输出。

[0006] 例如,网络输入可以是单个单模态张量,并且每个不相交分区可以是张量的不同的非重叠区域。

[0007] 作为另一示例,网络输入可以是包括多个不同模态的多模态输入,并且每个分区可以是多个模态中的一个不同模态。

[0008] 本说明书中描述的主题可以在特定实施例中实现,以便实现以下优点中的一个或多个优点。

[0009] 基于注意力的模型对于多模态处理是有吸引力的,因为来自多个模态的输入可以被连结并馈送到单个主干网络——因此需要非常少的融合工程。然而,所产生的表示贯穿网络是完全纠缠的,这在若干场景中可能是有问题的。

[0010] 例如,对比学习已经被证明是用于利用未标记的数据来改进各种任务上的下游性能的有效技术。在训练期间,多模态对比自监督学习需要每个模态的独立特征来操作,否则学习坍塌(collapse)。然而,因为表示是纠缠的,所以不存在适合用作到对比的输入的独立特征。

[0011] 此外,在推断时,当一个模态缺失时,这些模型不能有效地执行单模态任务或处理输入。

[0012] 本说明书描述了用于控制如何在基于注意力的神经网络内路由来自每个模态的输入以便保持模型模态特定——即仅基于来自单个模态的数据——的内部表示的部分的技术。具体地,本说明书描述了针对每个模态仅在该模态(而不是其他模态)的潜在向量上使用注意力来更新该模态的潜在向量集合而使用来自所有模态的信息来更新融合潜在向量的集合的技术。

[0013] 这允许系统有效地并入对比预训练,以便改进各种下游多模态任务的性能。此外,这允许系统有效地执行单模态推断,或者更一般地,即使来自一个模态的数据丢失,仍然为任务生成准确的输出。

[0014] 在下面的附图和描述中阐述了本说明书的主题的一个或多个实施例的细节。根据说明书、附图和权利要求,主题的其他特征、方面和优点将变得显而易见。

附图说明

[0015] 图1是示例神经网络系统的图。

[0016] 图2是用于使用神经网络处理网络输入的示例过程的流程图。

[0017] 图3示出了注意力块的示例架构。

[0018] 图4示出了注意力块的另一示例架构。

[0019] 图5示出了注意力块的又一示例架构。

[0020] 图6示出了输出块的操作。

[0021] 图7示出了神经网络在训练期间和推断期间两者的各种使用。

[0022] 各个附图中相同的附图标记和名称指示相同的元件。

具体实施方式

[0023] 图1是示例神经网络系统100的图。神经网络系统100是被实现为一个或多个位所中的一个或多个计算机上的计算机程序的系统的示例,其中可以实现下面描述的系统、组件和技术。

[0024] 神经网络系统100处理被分区为多个不相交分区104的网络输入102,以生成用于机器学习任务的网络输出112。

[0025] 例如,网络输入102可以是图像,并且不相交的分区可以是图像的不同的非重叠区域。在该示例中,网络输出112可以是标识图像中描绘对象的区域的对象检测输出、向图像中的每个像素指派相应类别标签的图像分割输出、或将图像分类为来自对象类的集合的一个或多个对象类的图像分类输出。

[0026] 作为另一示例,网络输入102可以是视频,并且不相交分区可以是来自视频的视频帧的不同集合或来自视频的不同空间-时间区域。在该示例中,网络输出112可以是任何适当的视频理解输出,例如,识别视频中正在执行的一个或多个动作的动作辨识输出、识别视频涉及的一个或多个主题的主题分类输出、识别视频帧中描绘对象的区域的对象检测输出、将相应的类别标签指派给视频的视频帧中的一个或多个视频帧中的每个像素的视频分割输出等。

[0027] 作为另一示例,网络输入102可以是来自测量物理环境的特性的传感器的其他输入。不相交分区可以是例如时间上相异的测量或空间中不同区域的测量的不同集合。传感器可以是任何适当的传感器,并且仅作为示例,网络输入102可以包括光、温度、雷达、声纳、LIDAR(激光雷达)、触觉反馈、电阻、电压、电流、加速度、化学构成等的测量。在该示例中,网络输出112可以是表征物理环境的任何输出,例如,识别环境的一个或多个特性的辨识输出、将相应类别标签指派给一系列环境中的每个环境的类别化输出等。任务可以包括例如响应于特性的识别或类别的指派而选择要由机械智能体执行的动作,或者可以用于检测或

预测故障状况或其他状况。

[0028] 作为另一示例,网络输入102可以是测量机械智能体的位置或状态——即本体感受或姿态——的传感器输入的。不相交分区可以是例如机械智能体的不同部分(例如,致动器、肢体)的不同测量或时间上相异的测量的不同集合。在该示例中,网络输出112可以是表征机械智能体的状态的任何输出、将相应类别标签指派给机械智能体的每个部分或机械智能体的一系列姿态的类别化输出。

[0029] 作为另一示例,网络输入102可以是包括多个不同模态的多模态输入,并且每个分区可以是多个模态中的一个不同模态。通常,多模态处理任务可以与先前针对构成多模态组合的任何数据类型描述的任何任务相对应。例如,当将任务应用于组合先前已经描述了任务的数据和另一类型的数据的多模态数据时,可以提高先前描述的任务的准确性。例如,当处理多种不同类型(模态)的数据时,可以改进对象或事件的检测或分类。

[0030] 例如,多模态输入可以包括上述任何输入中的两个或更多个输入,诸如图像、视频、文本和音频。

[0031] 在这些情况下,机器学习任务可以是任何适当的多模态任务,并且网络输出112可以包括任何适当的多模态任务输出。

[0032] 例如,当输入102包括视频或图像和文本时,网络输出112可以是回答由文本提出的关于视频或图像的问题的文本序列,或者是如文本中所描述的已经编辑的经编辑的视频或图像。

[0033] 作为另一示例,当输入102包括视频和音频时,音频可以是用于视频的音频音轨,并且网络输出112可以是用于分类任务的输出,分类任务例如,对视频的主题进行分类、对音频音轨中的音频事件类进行分类、对音频中发出噪声的一种或多种类型的对象进行分类、对视频中正在执行的一个或多个动作进行分类等。

[0034] 作为另一示例,当输入102包括视频和音频时,音频可以是用于视频的音频音轨,并且网络输出112可以是用于语音辨识任务的输出,例如,可以转录存在于音频音轨中的语音,用于语音隔离任务,例如,可以输出用于音频音轨中的一个或多个扬声器的隔离语音信号,等。

[0035] 特别地,系统100获取网络输入102并且使用分区引擎104确定网络输入的多个不相交分区。

[0036] 如果网络输入102是多模态输入,则多个不相交分区可以是输入的不同模态。

[0037] 如果网络输入102是单个模态的张量,则多个不相交分区104可以是张量的不同部分。

[0038] 针对每个分区104,系统100从该分区生成分区的潜在词元106的相应集合。例如,系统100可以使用对应的编码器神经网络来处理分区以生成潜在词元。

[0039] 如本说明书中所使用的,“词元”是具有固定维度的数值的向量或其他有序合集,即,有序集合中的值的数量跨不同词元是恒定的。

[0040] 系统100还生成融合潜在词元108的集合。

[0041] 例如,融合潜在词元108可以各自由预定值组成或可以在训练期间学习。

[0042] 系统100使用神经网络110处理每个分区的潜在词元106的相应集合和融合潜在词元108的集合,以生成表征网络输入102的网络输出112。

[0043] 神经网络110包括神经网络块120的序列,该序列包括(i)一个或多个注意力块130和(ii)输出块140。

[0044] 每个注意力块130使用分区注意力来更新潜在词元,即,针对每个分区和融合词元使用相应的注意力机制。

[0045] 下面将参考图2和3更详细地描述使用分区注意力更新潜在词元。

[0046] 当更新潜在词元时,每个注意力块130还可以执行附加操作,例如归一化操作、跳过连接操作、前馈层操作等中的一个或多个操作。

[0047] 在使用一个或多个注意力块130更新潜在词元106的相应集合和融合潜在词元108之后,输出块140处理潜在词元106、108中的至少一个或多个潜在词元,以生成表征网络输入102的网络输出112。

[0048] 在一些实施方式中,输出块140(从潜在词元106、108中的至少一个或多个潜在词元)生成单个网络输出,然后使用单个网络输出作为网络输出112。例如,输出块140可以生成仅从一个或多个融合词元(而不是任何分区的任何词元)生成的融合输出或从所有潜在词元生成的组合(“全局”)输出作为单个输出。作为另一示例,当需要分区特定输出时,例如,当来自其他分区的数据丢失时,输出块140可以生成针对指定分区的分区输出作为单个输出,该分区输出仅从分区的词元中的一个或多个词元(而不是融合词元或任何其他分区的任何词元)生成。

[0049] 在一些其他实施方式中,输出块140生成多个候选网络输出,然后组合候选网络输出,例如,对网络输出求平均,以生成最终网络输出112。

[0050] 例如,输出块140可以生成以下中的两个或更多个:仅从融合词元中的一个或多个融合词元(而不是任何分区的任何词元)生成的融合输出、仅从分区的词元中的一个或多个词元(而不是融合词元或任何其他分区的任何词元)生成的每个分区的相应分区输出、或从所有潜在词元生成的组合(“全局”)输出。

[0051] 下面参考图6描述由输出块140执行以生成输出的操作。

[0052] 图2是用于处理网络输入以生成网络输出的示例过程200的流程图。为了方便起见,过程200将被描述为由位于一个或多个位所的一个或多个计算机的系统执行。例如,根据本说明书适当编程的神经网络系统,例如图1中描绘的神经网络系统100,可以执行过程200。

[0053] 系统获取网络输入(步骤202)。

[0054] 系统确定网络输入的多个不相交分区(步骤204)。

[0055] 针对每个分区,系统从分区生成分区的潜在词元的相应集合(步骤206)。

[0056] 例如,针对每个分区,系统可以使用分区的编码器神经网络来处理网络输入的分区,以生成分区的潜在词元的相应集合。

[0057] 作为特定示例,系统可以将分区划分为补片,然后使用分区的编码器神经网络来处理每个补片以生成潜在词元。

[0058] 当每个分区来自相同模态时,系统可以针对每个模态使用相同的编码器神经网络。当分区来自不同模态时,系统可以针对不同分区使用不同的编码器神经网络。例如,编码器神经网络可以包括仅包括例如单个线性层或单个卷积层的单个神经网络层的卷积神经网络、多层感知器(MLP)或浅神经网络中的任何一个。

[0059] 可选地,系统可以使用位置编码来修改通过使用编码器神经网络处理网络输入的分区而生成的每个潜在词元,例如,通过将每个潜在词元与对应的位置编码求和或平均。位置编码通常为每个分区的每个补片标识该补片在该分区内的位置以及该补片属于哪个分区。例如,系统可以将所有模态的潜在词元(以及可选地将融合潜在词元)布置为序列,并且向每个潜在词元指派与序列中的潜在的位置相对应的位置编码。

[0060] 位置编码可以与神经网络的训练联合学习,或者可以是预定的,例如,可以是正弦或傅里叶位置编码。可替代地,位置编码可以是学习位置编码和预定位置编码的组合,例如总和。

[0061] 系统生成融合潜在词元的集合(步骤208)。例如,可以在神经网络的训练期间联合学习融合潜在词元的集合。如上所述,系统还可以可选地利用位置编码来修改融合潜在词元。

[0062] 系统使用神经网络处理每个分区的潜在词元的相应集合和融合潜在词元的集合以生成表征网络输入的网络输出(步骤210)。

[0063] 如上所述,神经网络通常包括神经网络块序列,该神经网络块序列包括(i)一个或多个注意力块和(ii)输出块。

[0064] 每个注意力块更新潜在词元,即,每个分区的潜在词元的集合和融合潜在词元。

[0065] 作为更新潜在词元的一部分,针对每个分区,块通过针对该分区应用仅注意该分区的潜在词元的相应集合的对应注意力机制来更新该分区的潜在词元的相应集合。

[0066] 注意力块还通过应用对应的融合注意力机制来更新融合潜在词元的集合,该对应的融合注意力机制注意分区中的每个分区的潜在词元的相应集合和融合潜在词元的集合。

[0067] 注意力块的对应注意力机制的一个示例在下面参考图3描述。

[0068] 用于注意力块的对应注意力机制的另一示例在下面参考图4描述。

[0069] 用于注意力块的对应注意力机制的又一示例在下面参考图5描述。

[0070] 在使用一个或多个注意力块更新潜在词元的相应集合和融合潜在词元之后,输出块处理融合潜在词元中的至少一个或多个融合潜在词元以生成表征网络输入的网络输出。

[0071] 在一些实施方式中,输出块通过应用交叉注意力机制来生成网络输出。

[0072] 通过交叉注意力生成网络输出的示例在下面参考图6描述。

[0073] 图3示出了注意力块130的示例架构300。

[0074] 在图3的示例中,输入是包括两种模态的多模态输入:音频和视频。因此,到注意力块130的输入包括用于音频模态的潜在词元的集合302、融合潜在词元304和用于视频模态的潜在词元306的集合。

[0075] 在图3的示例中,每个分区和融合潜在词元的对应注意力机制是自注意力机制,即,分区和融合潜在词元的对应注意力机制共享参数并且可以被并行地计算。然而,注意力块并入掩蔽以调节分区之间的信息流。具体地,注意力块并入掩蔽以确保针对每个分区,仅使用来自该分区的词元(而不是任何其他分区的词元或融合潜在词元)来更新分区词元,而使用来自所有分区的所有词元和潜在词元来更新潜在融合词元。

[0076] 具体地,在图3的示例中,针对每个分区,该分区的对应注意力机制是具有分区特定掩蔽的自注意力机制,该分区特定掩蔽限制注意力机制仅注意分区的潜在词元的相应集合,并且对于融合潜在词元,对应注意力机制是不被掩蔽以允许注意力机制注意分区中的

每个分区的潜在词元的相应集合和融合潜在词元的相应集合的自注意力机制。

[0077] 具体地,注意力块可以通过将掩蔽二进制张量 m 引入到注意力权重的每个计算中,例如,在注意力块130的每个注意力头内,来实现上述注意力机制。该掩蔽张量被应用于标准注意力操作 $o_i = \sum_j a_{ij} v_i$, 其变为 $o_i = \sum_j \hat{a}_{ij} v_i$, 其中 v_i 是词元 i 的值向量, a_{ij} 是词元 i 的查询 q 与注意力被应用的 j 个位置的键 k 之间的注意力权重, 并且 $\hat{a}_{ij}$ 是:

$$\hat{a}_{ij} = \frac{m_{ij} \exp(\frac{q_i^T k_j}{\sqrt{D}})}{\sum_{\{j', m'_{ij}=1\}} \exp(\frac{q_i^T k_{j'}}{\sqrt{D}})}$$

[0078] 其中 D 是常数值,例如查询的维度、键向量和值向量,并且其中如果 j 是融合词元之一,则 m_{ij} 等于1,否则仅在 i 和 j 来自同一分区时, m_{ij} 等于1,否则等于0。

[0079] 因此,每个分区的每个潜在词元仅使用该分区的潜在词元来更新,而融合潜在词元使用融合潜在词元和所有分区的潜在词元来更新。

[0080] 图4示出了注意力块130的另一示例架构400。

[0081] 在图4的示例中,输入是包括两个模态的多模态输入。因此,到注意力块130的输入包括用于第一模态的潜在词元的集合、融合潜在词元和用于第二模态的潜在词元的集合。

[0082] 在图4的示例中,每个分区和融合潜在词元的对应注意力机制是不同的自注意力机制。

[0083] 具体地,注意力块使用仅注意每个分区中的潜在词元的潜在词元的对应自注意力机制来更新该分区的潜在词元。

[0084] 对于融合潜在词元,自注意力机制包括两个单独的机制:(i) 自注意力机制,其通过在每个融合潜在词元上的自注意力来更新该融合潜在词元;以及(ii) 交叉注意力机制,其使用(如通过(i)更新的)融合潜在词元以在那些词元已经被其对应的自注意力机制更新之后交叉注意到分区的潜在词元中。

[0085] 因此,每个分区的每个潜在词元仅使用该分区的潜在词元来更新,而融合潜在词元使用融合潜在词元和所有分区的潜在词元来更新。

[0086] 构造如图4所示的注意力块允许不同类型的自注意力用于分区中的每个分区,例如,以利用底层数据中的不同结构。例如,如图4所示,一个分区可以对应于音频模态,而另一个分区可以对应于视频模态。系统可以将一个注意力机制用于音频模态,而将不同的注意力机制用于视频模态。

[0087] 在图4的示例中,由于音频频谱图的输入空间的2D结构,系统使用2D Swin Transformer块作为音频模态的注意力机制。对于视频模态,系统使用3D Swin Transformer块作为注意力机制来利用视频帧的输入空间的3D结构。

[0088] 通常,Swin Transformer块不同于标准自注意力块,因为针对每个词元,Swin Transformer块仅将自注意力操作应用于附近词元而不是分区中的所有词元。2D Swin Transformer块将自注意力操作应用于附近词元的(a, b)补片,即,2维(2D)补片。另一方面,3D Swin Transformer块将自注意力操作应用于附近词元的(a, b, c)补片,即,3维(3D)补片。在Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling

up capacity and resolution(Swin transformer v2:扩大容量和分辨率). arXiv preprint arXiv:2111.09883, 2021中更详细地描述了Swin Transformer块。图5示出了注意力块130的又一示例架构500。

[0089] 类似于图3和图4的示例,输入是包括两种模态的多模态输入。因此,对注意力块的输入包括用于第一模态的潜在词元集合、融合潜在词元和用于第二模态的潜在词元集合。

[0090] 在图5的示例中,每个分区和融合潜在词元的对应注意力机制基于分层感知器(HiP)注意力块。

[0091] HiP注意力块通过将到块的输入分成多个组来操作,并且针对每个组,将交叉注意力应用于该组的学习查询向量的集合以生成潜在向量的集合,然后仅在该组的潜在向量内操作自注意力。然后合并潜在向量以生成例如下一HiP注意力块的下一网络组件的输入。通过在分层架构中布置这样的块,分层感知器神经网络可以将组融合在一起,以便聚合关于输入的信息和全局原因。在Joao Carreira, Skanda Koppula, Daniel Zoran, Adria Recasens, Catalin Ionescu, Olivier Henaff, Evan Shelhamer, Relja Arandjelovic, Matt Botvinick, Oriol Vinyals, et al. Hierarchical perceiver(分层感知器). arXiv preprint arXiv:2202.10890, 2022中更详细地描述了HiP块。示例架构500包括用于每个分区的分开的HiP注意力块。也就是说,用于每个分区的对应注意力机制被实现为用于每个分区的分开的HiP块。

[0092] 因此,注意力块130使用分区的HiP注意力块来处理每个分区的潜在词元,以更新分区的潜在词元。

[0093] 对于融合词元,注意力块130包括两个分开的HiP块。块130仅使用第一HiP块处理融合潜在词元以更新融合潜在词元。在第一HiP块之后,注意力块130将分区的潜在词元与融合潜在词元连结以生成组合序列,并且使用HiP块处理组合序列以更新融合潜在词元,而不是分区的潜在词元。

[0094] 因此,融合潜在词元的对应注意力机制被实现为对融合潜在词元进行操作的第一HiP块和对组合序列进行操作的第二HiP块。

[0095] 因此,每个分区的每个潜在词元仅使用该分区的潜在词元来更新,即,通过由分区的HiP块处理,而融合潜在词元使用融合潜在词元和所有分区的潜在词元来更新,即,借助由融合潜在词元的第二HiP块处理的组合序列。

[0096] 系统可以通过修改由不同注意力块130内的HiP块使用的块的数量来实现跨整个输入的分层融合。例如,当神经网络110包括五个注意力块130的序列时,组的数量可以如下:(32;4;1;1;1)。因此,在该示例中,最后三个注意力块130内的HiP块对由HiP块接收作为输入的潜在词元的整个集合进行操作。

[0097] 图6示出了由输出块140执行的操作的示例600。

[0098] 在图6的示例中,输出块140通过利用交叉注意力来生成网络输出。

[0099] 具体地,在图6的示例中,对于任何给定输入,输出块140可以生成三种不同类型的输出630中的一个或多个:(1)仅使用单个分区的潜在词元生成的分区特定输出,(2)仅使用融合潜在词元生成的融合输出,或(3)使用所有潜在词元——即融合潜在词元和所有分区的潜在词元——生成的全局输出。

[0100] 在一些情况下,输出块140仅被配置为生成融合输出或全局输出,而不是两者。

[0101] 类似地,在一些情况下,在训练之后,输出块140仅被配置为生成融合输出或全局输出中的一个或多个,而不是任何分区特定的输出。

[0102] 因此,输出块140维护每个分区的、融合潜在词元的以及全局输出的学习查询向量610的相应集合和针对每个输出630的相应输出头620。学习查询向量610与神经网络110的训练联合学习。

[0103] 为了生成给定分区的输出,输出块140使用分区的学习查询向量将交叉注意力应用到分区的潜在标记中以更新学习查询向量。然后,输出块140可以使用将一个或多个适当的学习变换应用于一个或多个更新的学习查询向量的分区特定头部,以将学习查询向量映射到机器学习任务的输出。例如,输出块140可以使用一个或多个线性层、激活函数层、softmax层或上述的某种组合来处理每个学习查询向量。

[0104] 为了生成融合输出,输出块140使用融合潜在词元的学习查询向量将交叉注意力应用到融合潜在词元中以更新学习查询向量。然后,输出块140可以使用融合头,该融合头将一个或多个适当的学习变换应用于一个或多个更新的学习查询向量,以将学习查询向量映射到机器学习任务的输出。例如,输出块140可以使用一个或多个线性层、激活函数层、softmax层或上述的某种组合来处理每个学习查询向量。

[0105] 为了生成全局输出,输出块140使用针对全局输出的学习查询向量将交叉注意力应用到所有潜在标记中以更新学习查询向量。然后,输出块140可以使用全局头,该全局头将一个或多个适当的学习变换应用于一个或多个更新的学习查询向量,以将学习查询向量映射到机器学习任务的输出。例如,输出块140可以使用一个或多个线性层、激活函数层、softmax层或上述的某种组合来处理每个学习查询向量。

[0106] 在图6的示例中,任务是需要将输入分类为527个类别中的一个或多个类别的分类任务。因此,每个输出是分数——例如概率——的 1×527 向量。类似地,学习查询向量的每个集合包括单个查询向量,并且每个输出头将对应的单个更新的学习查询向量映射到分数的 1×527 向量。

[0107] 因此,分区特定输出仅取决于单个给定分区,而融合输出和全局输出并入来自整个输入的信息。

[0108] 在使用神经网络110执行机器学习任务之前,训练系统训练神经网络110以执行任务,即,确定神经网络的参数的训练值,即,序列中的注意力块和输出块的训练值,以及可选地,用于生成潜在词元的编码器神经网络。例如,训练系统可以使用常规机器学习技术在任务的训练数据上从头开始训练神经网络以最小化任务的损失函数,例如交叉熵损失、负对数似然损失等。作为另一示例,训练系统可以首先在无监督目标上预训练神经网络,然后在任务的训练数据上微调神经网络。作为又一示例,训练系统可以通过半监督学习在任务的未标记数据和训练数据两者上训练神经网络。

[0109] 在训练期间,训练系统可以并入任何数量的技术来改进训练过程的速度、有效性或两者。例如,系统可以使用丢弃、标签平滑或两者来减少过度拟合。作为另一示例,系统可以使用并行训练神经网络的多个实例的分布式架构来执行训练。

[0110] 此外,如上所述,系统可以首先通过自监督或无监督学习在大的无监督数据集上预训练神经网络,然后在任务特定训练数据上微调神经网络以优化任务的损失函数。

[0111] 特别地,由于神经网络的架构,神经网络特别适于使用无监督目标或自监督目标

来训练,作为预训练步骤或作为监督训练的辅助损失,即使当到神经网络的输入是多模态的。

[0112] 另外,在训练之后,神经网络110可以用于处理仅包括单个分区的两个输入,例如,单模态输入和多模态输入。

[0113] 图7示出了神经网络110的各种使用的示例。

[0114] 如图7所示,在图中称为“Zorro”的神经网络可以用于训练之后的多模态推断740和单模态推断730两者,即使神经网络仅被训练为执行多模态任务或单模态任务(而不是两者)。

[0115] 为了将神经网络用于多模态推断740,即,其中输入包括两个或更多个模态,每个模态被指派到不同的分区中,系统使用输出块来生成融合输出、全局输出或两者。进一步可选地,系统还可以为每个分区生成相应的分区特定输出。然后,系统可以使用融合输出或全局输出作为网络输出,或者当生成多个输出时,如上所述组合输出以生成网络输出。

[0116] 为了将神经网络用于单模态推断730,即,其中输入仅包括一个模态,并且该模态被指派给单个分区,系统可以如下进行。

[0117] 具体地,系统可以接收仅包括多个模态中的第一模态的网络输入。如上所述,系统可以为与第一模态相对应的分区生成潜在词元的集合。然后,系统可以使用神经网络通过以下方式来处理与第一模态相对应的分区的潜在词元的集合:针对每个注意力块,通过针对该分区应用仅注意该分区的潜在词元的集合的对应注意力机制来更新与模态相对应的分区的潜在词元的相应集合。在使用一个或多个注意力块更新对应模态的潜在词元的集合之后,系统可以使用输出块处理来自对应模态的潜在词元的集合的一个或多个潜在词元,以生成表征网络输入的网络输出。因此,神经网络不需要针对其他模态或针对融合词元执行注意力块的操作。

[0118] 在训练神经网络期间,系统可以使用神经网络来执行对比预训练(无监督任务)710和监督训练720两者。

[0119] 为了对包括训练输入和标签的给定训练示例执行监督训练,系统可以使用神经网络来生成训练输入的融合输出和模态特定输出(以及可选地,全局输出)中的每一个。然后,系统可以计算标签与所有生成的输出之间的相应损失,并使用所有相应损失的梯度来更新神经网络的参数。这可以给出比仅使用单个多模态输出更丰富的训练信号,并且还可以允许神经网络在训练之后有效地执行单模态推断。

[0120] 多模态对比方法通过将多个模态对准到公共嵌入空间中来学习表示。与单模态方法相反,多模态对比方法使用不同的模态作为视图,而不是产生数据的多个视图。一个重要要求是生成嵌入的神经网络的两个组件不共享信息。如果跨模态共享信息,则自监督训练可以容易地折叠或收敛到平凡解决方案。

[0121] 用于多模态感知的常规模型通常为多个输入产生单个输出。这对于受监督的应用是足够的,但是阻止使用这些对比技术。然而,神经网络生成单模态和多模态输出两者的能力使得能够有效地使用自监督对比损失。

[0122] 对于具有例如噪声对比估计损失的对比损失的训练,系统针对每个模态将最终线性投影(每个模态不同)应用于由输出块针对该模态执行的交叉注意力的输出,以针对每个模态产生最终嵌入向量。然后,系统可以使用这些最终嵌入向量来计算对比损失,例如,噪

声对比估计损失或测量嵌入之间的相似性的其他对比学习损失。为了训练特定于融合表示或输出的任何参数(例如,融合交叉注意力或融合权重,如果模型具有每个模态的单独的权重),系统可以为每个模态添加相应的融合对比损失。给定模态的融合损失是将从如上所述的融合潜在向量计算的最终融合嵌入和给定模态的最终嵌入作为输入的对比损失。例如,当存在两个输入模态时,总体对比损失可以是两个输入模态之间的对比损失与两个输入模态的融合对比损失之间的总和或加权总和。

[0123] 当在监督、非监督或自监督训练期间学习融合潜在向量时,系统还可以例如通过将损失函数的梯度反向传播通过神经网络到融合潜在向量中来更新融合潜在向量。

[0124] 当系统还与神经网络的训练联合训练一个或多个编码器神经网络时,系统还可以通过将损失函数的梯度反向传播通过神经网络到编码器神经网络中来更新编码器神经网络。

模型	预训练	监督预训练	训练集合	模态	mAP
AST [22]	IN-21k	✓	AudioSet	A	45.9
GBlend [45]	IG-65	✓	AudioSet	A+V	37.8
Attn Audio-Visual [16]	ImageNet	✓	AudioSet	A+V	46.2
MBT [30]	IN-21k	✓	AudioSet-500k	A+V	52.1
Zorro	YT8M	✓	AudioSet	A+V	50.6
Zorro	YT8M	✓	AudioSet-500k	A+V	51.6
ERANN [44]	✗	✗	AudioSet	A	45.0
Perceiver [24]	✗	✗	AudioSet	A+V	44.2
Zorro	✗	✗	AudioSet	A+V	44.8
Zorro	AudioSet	✗	AudioSet-500k	A+V	47.5
Zorro	YT8M	✗	AudioSet	A+V	48.8
Zorro	YT8M	✗	AudioSet-500k	A+V	48.8

[0126] 表1

模型	预训练	监督预训练	模态	前1	前5
ResNet-50 [14]	✗	✗	A	51.0	76.4
AudioSlowFast [25]	✗	✗	A	52.5	78.1
MBT [30]	ImageNet-21k	✓	A+V	64.1	85.6
Zorro	YT8M	✗	A+V	64.2	85.5
Zorro	YT8M	✓	A+V	66.2	86.3

[0127] 表2

[0128] 表1和表2示出了所描述的神经网络(“Zorro”)在需要处理音频(A)、视频(V)或需要处理A和V的多模态输入的各种任务上以及在各种自监督预训练(“pre-training”)和监督预训练(“Sup. Pre-training”)下的性能。在表1和表2中,当处理音频和视频两者时,任务是音频-视频分类,并且当仅处理音频时,任务是音频分类。从表1和表2可以看出,Zorro实现了在前1精度(“top-1”)和前5精度(“top-5”)两方面在各种机制下在各种任务上匹配或超过各种基线(在表中所有模型未标记为“Zorro”)的结果。也就是说,所描述的架构足够灵活以在各种训练机制中的任何一种中生成高质量结果,而不需要对架构进行修改。此外,在训练之后,即使训练数据都是多模态的,所描述的架构也可以有效地用于执行单模态处理。

[0129] 本说明书结合系统和计算机程序组件使用术语“配置”。对于要被配置为执行特定操作或动作的一个或多个计算机的系统,意味着系统已经在其上安装了软件、固件、硬件或它们的组合,其在操作中使系统执行操作或动作。对于要被配置为执行特定操作或动作的

一个或多个计算机程序,意味着一个或多个程序包括指令,所述指令在由数据处理装置执行时使得装置执行操作或动作。

[0130] 在本说明书中描述的主题和功能操作的实施例可以在数字电子电路中、在有形地体现的计算机软件或固件中、在计算机硬件中——包括在本说明书中公开的结构及其结构等同物——或在它们中的一个或多个的组合中实现。本说明书中描述的主题的实施例可以被实现为一个或多个计算机程序,例如,编码在有形非暂时性存储介质上的计算机程序指令的一个或多个模块,用于由数据处理装置执行或控制数据处理装置的操作。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、随机或串行访问存储器设备或它们中的一个或多个的组合。可替代地或附加地,程序指令可以被编码在人工生成的传播信号——例如机器生成的电信号、光信号或电磁信号——上,该传播信号被生成以编码用于传输到合适的接收器装置以供数据处理装置执行的信息。

[0131] 术语“数据处理装置”是指数据处理硬件,并且涵盖用于处理数据的所有种类的装置、设备和机器,包括例如可编程处理器、计算机或多个处理器或计算机。该装置还可以是或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ASIC(专用集成电路)。除了硬件之外,该装置还可以可选地包括创建用于计算机程序的执行环境的代码,例如,构成处理器固件、协议栈、数据库管理系统、操作系统或它们中的一个或多个的组的代码。

[0132] 也可以被称为或描述为程序、软件、软件应用、app、模块、软件模块、脚本或代码的计算机程序可以以任何形式的编程语言编写,包括编译或解释语言、或声明或过程语言;并且它可以以任何形式部署,包括作为独立程序或作为适合于在计算环境中使用的模块、组件、子例程或其他单元。程序可以但不需要对应于文件系统中的文件。程序可以存储在保存其他程序或数据的文件的一部分中,例如存储在标记语言文档中的一个或多个脚本,存储在专用于所讨论的程序的单个文件中,或者存储在多个协调文件中,例如存储一个或多个模块、子程序或代码部分的文件。可以部署计算机程序以在位于一个站点的一个计算机或跨多个站点分布并通过数据通信网络互连的多个计算机上执行。

[0133] 在本说明书中,术语“数据库”广泛地用于指代任何数据合集:数据不需要以任何特定方式结构化或根本不需要结构化,并且其可以存储在一个或多个位所中的存储设备上。因此,例如,索引数据库可以包括多个数据合集,每个数据合集可以被不同地组织和访问。

[0134] 类似地,在本说明书中,术语“引擎”广泛地用于指代被编程为执行一个或多个特定功能的基于软件的系统、子系统或过程。通常,引擎将被实现为安装在一个或多个位所中的一个或多个计算机上的一个或多个软件模块或组件。在一些情况下,一个或多个计算机将专用于特定引擎;在其他情况下,可以在同一计算机或多个计算机上安装和运行多个引擎。

[0135] 本说明书中描述的过程和逻辑流程可以由执行一个或多个计算机程序的一个或多个可编程计算机执行,以通过对输入数据进行操作并生成输出来执行功能。过程和逻辑流程还可以由例如FPGA或ASIC的专用逻辑电路或由专用逻辑电路和一个或多个编程计算机的组合来执行。

[0136] 适于执行计算机程序的计算机可以基于通用或专用微处理器或两者,或任何其他类型的中央处理单元。通常,中央处理单元将从只读存储器或随机存取存储器或两者接收

指令和数据。计算机的基本元件是用于实施或执行指令的中央处理单元以及用于存储指令和数据的一个或多个存储器设备。中央处理单元和存储器可以由专用逻辑电路补充或并入专用逻辑电路中。通常,计算机还将包括一个或多个大容量存储设备,或可操作地耦合以从用于存储数据的一个或多个大容量存储设备——例如磁盘、磁光盘或光盘——接收数据或向其传送数据或两者都有。然而,计算机不需要具有这样的设备。此外,计算机可嵌入于另一设备中,例如移动电话、个人数字助理(PDA)、移动音频或视频播放器、游戏控制台、全球定位系统(GPS)接收器或便携式存储设备,例如通用串行总线(USB)快闪驱动器,仅举几例。

[0137] 适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和闪存设备;磁盘,例如内部硬盘或可移动盘;磁光盘;以及CD ROM和DVD-ROM盘。

[0138] 为了提供与用户的交互,本说明书中描述的主题的实施例可以在计算机上实现,该计算机具有用于向用户显示信息的显示设备,例如CRT(阴极射线管)或LCD(液晶显示器)监视器,以及用户可以通过其向计算机提供输入的键盘和指示设备,例如鼠标或轨迹球。其他类型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感觉反馈,例如视觉反馈、听觉反馈或触觉反馈;并且可以以任何形式接收来自用户的输入,包括声学、语音或触觉输入。此外,计算机可以通过向用户使用的设备发送文档和从用户使用的设备接收文档来与用户交互;例如,通过响应于从web浏览器接收的请求向用户设备上的web浏览器发送网页。此外,计算机可以通过向个人设备——例如,正在运行消息传送应用的智能电话——发送文本消息或其他形式的消息并且从用户接收响应消息作为回报来与用户交互。

[0139] 用于实现机器学习模型的数据处理装置还可以包括例如专用硬件加速器单元,用于处理机器学习训练或生产的公共和计算密集型部分,例如推断、工作负载。

[0140] 可以使用例如TensorFlow框架或Jax框架的机器学习框架来实现和部署机器学习模型。

[0141] 本说明书中描述的主题的实施例可以在计算系统中实现,该计算系统包括后端组件,例如作为数据服务器,或者包括中间件组件,,例如应用服务器,或者包括前端组件,例如具有用户可以通过其与本说明书中描述的主题的实施方式进行交互的图形用户界面、web浏览器或应用的客户端计算机,或者一个或多个这样的后端、中间件或前端组件的任何组合。系统的组件可以通过例如通信网络的任何形式或介质的数字数据通信互连。通信网络的示例包括局域网(LAN)和广域网(WAN),例如因特网。

[0142] 计算系统可以包括客户端和服务端。客户端和服务端通常彼此远离,并且典型地通过通信网络进行交互。客户端和服务端的关系借助在相应计算机上运行并且彼此具有客户端-服务端关系的计算机程序而产生。在一些实施例中,服务端将例如HTML页面的数据传输到用户设备,例如,用于向充当客户端的设备交互的用户显示数据和从其接收用户输入。可以在服务端处从设备接收在用户设备处生成的数据,例如用户交互的结果。

[0143] 虽然本说明书包含许多具体实施方式细节,但是这些不应当被解释为对任何发明的范围或对可以要求保护的范围的限制,而是作为可以特定于特定发明的特定实施例的特征的描述。在本说明书中在分开的实施例的上下文中描述的某些特征也可以在单个实施例中组合实现。相反,在单个实施例的上下文中描述的各种特征也可以分开地或以任何合适

的子组合在多个实施例中实现。此外,尽管上面可以将特征描述为以某些组合起作用并且甚至最初如此要求保护,但是在一些情况下,来自要求保护的组合的一个或多个特征可以从组合中去除,并且要求保护的组合可以针对子组合或子组合的变型。

[0144] 类似地,虽然在附图中以特定次序描绘并且在权利要求中叙述了操作,但是这不应被理解为要求以所示的特定次序或顺序次序执行这样的操作,或者执行所有所示的操作以实现期望的结果。在某些情况下,多任务和并行处理可能是有利的。此外,上述实施例中的各种系统模块和组件的分离不应被理解为在所有实施例中都需要这样的分离,并且应当理解,所描述的程序组件和系统通常可以一起集成在单个软件产品中或被封装到多个软件产品中。

[0145] 已经描述了主题的特定实施例。其他实施例在所附权利要求书的范围内。例如,权利要求中记载的动作可以以不同的次序执行并且仍然实现期望的结果。作为一个示例,附图中描绘的过程不一定需要所示的特定次序或顺序来实现期望的结果。在一些情况下,多任务和并行处理可能是有利的。

[0146] 现在阐述描述本文描述的主题的示例的多个语句。

[0147] 1. 一种由一个或多个计算机执行的方法,所述方法包括:

获取第一网络输入;

确定所述第一网络输入的多个不相交分区;

针对每个分区,从该分区生成该分区的潜在词元的相应集合;

生成融合潜在词元的集合;

使用神经网络处理每个分区的潜在词元的相应集合和融合潜在词元集合以生成表征第一网络输入的网络输出,

其中,神经网络包括神经网络块序列,该神经网络块序列包括:(i) 一个或多个注意力块和(ii) 输出块,其中每个注意力块执行操作,所述操作包括:

针对每个分区,通过针对该分区应用仅注意分区的潜在词元的相应集合的对应注意力机制来更新分区的潜在词元的相应集合;以及

通过应用对应的融合注意力机制来更新融合潜在词元集合,该对应的融合注意力机制注意针对分区中的每个分区的潜在词元的相应集合和融合潜在词元的集合;以及

其中,输出块执行操作,所述操作包括:

在使用一个或多个注意力块更新潜在词元的相应集合和融合潜在词元之后,处理潜在词元中的至少一个或多个潜在词元以生成表征第一网络输入的网络输出。

[0148] 2. 根据示例1所述的方法,其中,每个不相交分区对应于来自多个模态的不同模态。

[0149] 3. 根据示例2所述的方法,其中,模态包括图像模态、视频模态或两者。

[0150] 4. 根据示例2或示例3所述的方法,其中,模态包括音频模态。

[0151] 5. 根据示例2至4中任一项所述的方法,其中,模态包括文本模态。

[0152] 6. 根据示例2至5中任一项所述的方法,进一步包括:

接收仅包括多个模态中的第一模态的第二输入;

生成与第一模态相对应的分区的潜在词元的集合;

使用神经网络处理与第一模态相对应的分区的潜在词元集合,包括:

针对每个注意力块,通过针对分区应用仅注意该分区的潜在词元的相应集合的对应注意力机制来更新与模态相对应的分区的潜在词元的集合;以及

在使用一个或多个注意力块更新对应模态的潜在词元集合之后,处理来自对应模态的潜在词元集合的一个或多个潜在词元以生成表征第二网络输入的网络输出。

[0153] 7. 根据任一前述示例所述的方法,其中,在神经网络的训练期间学习融合潜在词元。

[0154] 8. 根据任一前述示例所述的方法,其中,针对每个分区,从分区生成分区的潜在词元的相应集合包括:

使用分区的编码器神经网络来处理第一网络输入的分区,以生成分区的潜在词元的相应集合。

[0155] 9. 根据示例8所述的方法,其中,针对每个分区,从该分区生成该分区的潜在词元的相应集合进一步包括:

使用位置编码来修改通过使用编码器神经网络处理第一网络输入的分区而生成的每个潜在词元。

[0156] 10. 根据示例8或示例9所述的方法,其中,针对每个分区,从该分区生成该分区的潜在词元的相应集合进一步包括:

利用一个或多个学习词元来增强通过使用编码器神经网络处理第一网络输入的分区而生成的潜在词元。

[0157] 11. 根据任一前述示例所述的方法,其中:

针对每个分区,分区的对应注意力机制是具有分区特定掩蔽的自注意力机制,该分区特定掩蔽将注意力机制限制为仅注意分区的潜在词元的相应集合;以及

对于融合潜在词元,对应注意力机制是自注意力机制,其不被掩蔽以允许注意力机制注意分区中的每个分区的潜在词元的相应集合和融合潜在词元的相应集合。

[0158] 12. 根据示例11所述的方法,其中,用于分区和融合潜在词元的对应注意力机制共享参数。

[0159] 13. 根据任一前述示例所述的方法,其中,处理融合潜在词元中的至少一个或多个融合潜在词元以生成表征第一网络输入的网络输出包括:

根据仅从融合潜在词元选择的一个或多个潜在词元生成融合网络输出。

[0160] 14. 根据任一前述示例所述的方法,其中,处理融合潜在词元中的至少一个或多个融合潜在词元以生成表征第一网络输入的网络输出包括:

根据从分区中的每个分区的潜在词元的集合和融合潜在词元中选择一个或多个潜在词元生成总体网络输出。

[0161] 15. 根据任一前述示例所述的方法,其中,处理融合潜在词元中的至少一个或多个融合潜在词元以生成表征第一网络输入的网络输出包括:

生成两个或更多个候选网络输出,所述两个或更多个候选网络输出包括使用融合潜在词元中的一个或多个融合潜在词元生成的至少一个候选网络输出;以及

组合所述两个或更多个候选网络输出以生成最终网络输出。

[0162] 16. 一种由一个或多个计算机执行的训练神经网络的方法,所述方法包括:

获取第一网络输入;

确定所述第一网络输入的多个不相交分区；

针对每个分区,从分区生成分区的潜在词元的相应集合；

生成融合潜在词元的集合；

使用神经网络处理每个分区的潜在词元的相应集合和融合潜在词元的集合以生成表征第一网络输入的网络输出，

其中,神经网络包括神经网络块序列,该神经网络块序列包括:(i)一个或多个注意力块和(ii)输出块，

其中,每个注意力块执行操作,该操作包括：

针对每个分区,通过针对该分区应用仅注意该分区的潜在词元的相应集合的对应注意力机制来更新分区的潜在词元的相应集合；以及

通过应用对应的融合注意力机制来更新融合潜在词元集合,该对应的融合注意力机制注意分区中的每个分区的潜在词元的相应集合和融合潜在词元的集合；以及

其中,输出块执行操作,所述操作包括：

在使用一个或多个注意力块更新潜在词元的相应集合和融合潜在词元之后,处理融合潜在词元中的至少一个或多个融合潜在词元以生成表征第一网络输入的网络输出。

[0163] 将网络输出与第一网络输入的预期网络输出进行比较以确定损失并基于损失来训练神经网络。

[0164] 17. 根据示例16所述的方法,其中,生成分区的潜在词元的相应集合和/或生成融合潜在词元集合包括:利用编码器神经网络来处理第一网络输入,所述编码器神经网络被训练以生成每个分区的潜在词元的相应集合和/或生成融合潜在词元集合；以及

其中,基于损失训练神经网络包括训练编码器神经网络。

[0165] 18. 根据示例16或17所述的方法,进一步包括使用自监督或无监督学习目标使用无监督数据集预训练神经网络。

[0166] 19. 根据示例16、示例17或18的方法,其中,确定损失包括使用无监督损失函数或自监督损失函数作为辅助损失函数。

[0167] 20. 一种由一个或多个计算机执行的方法,所述方法包括：

获取第一单模态网络输入；

使用根据示例16至19中任一项训练的神经网络来处理所述单模态输入以生成表征所述第一单模态网络输入的网络输出。

[0168] 21. 一种系统,所述系统包括一个或多个计算机和存储指令的一个或多个存储设备,所述指令在由所述一个或多个计算机执行时使得所述一个或多个计算机执行根据示例1至20中任一项的相应方法的操作。

[0169] 22. 一个或多个计算机存储介质,所述一个或多个计算机存储介质存储指令,所述指令在由一个或多个计算机执行时使所述一个或多个计算机执行根据示例1至20中任一项所述的相应方法的操作。

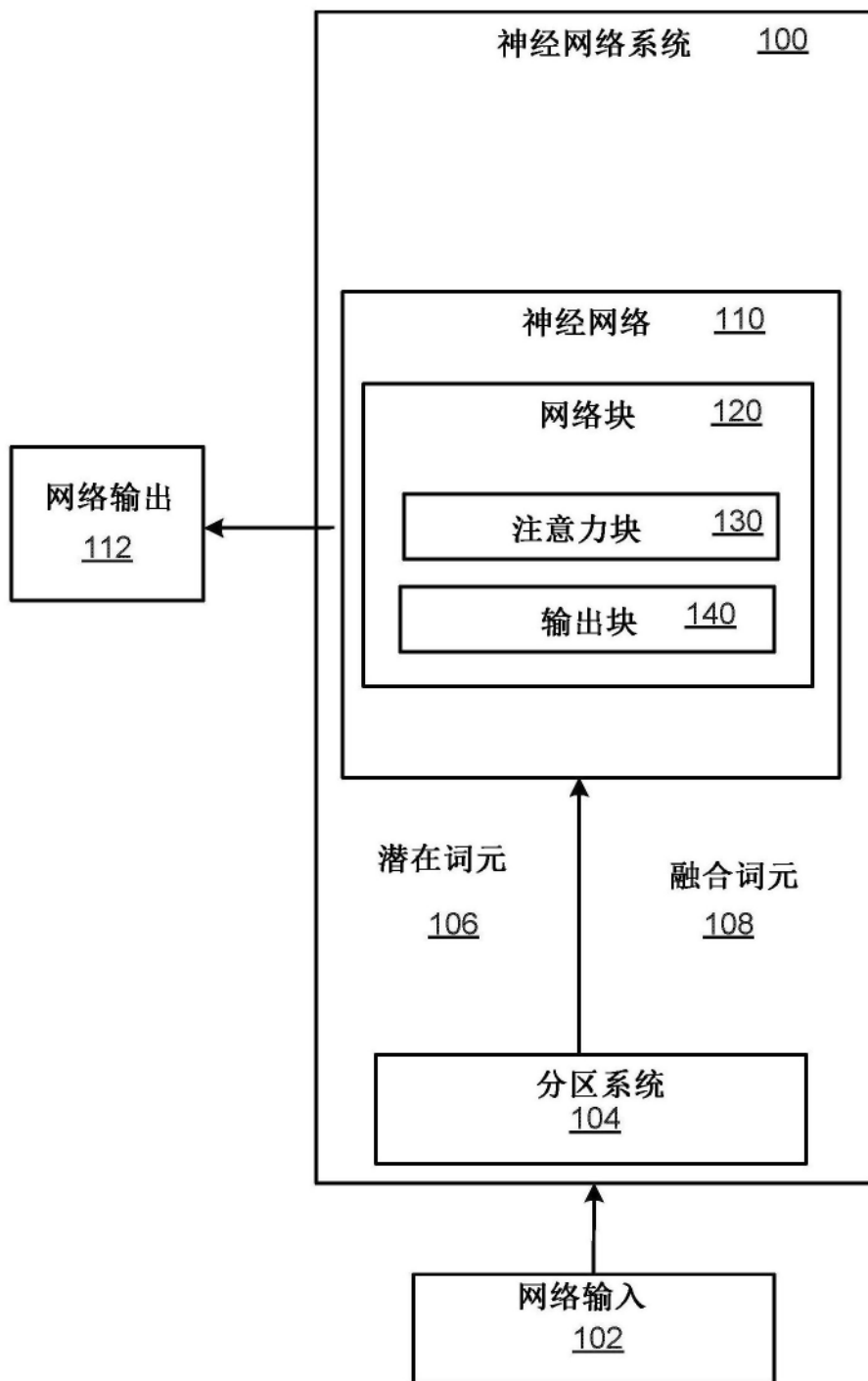


图1

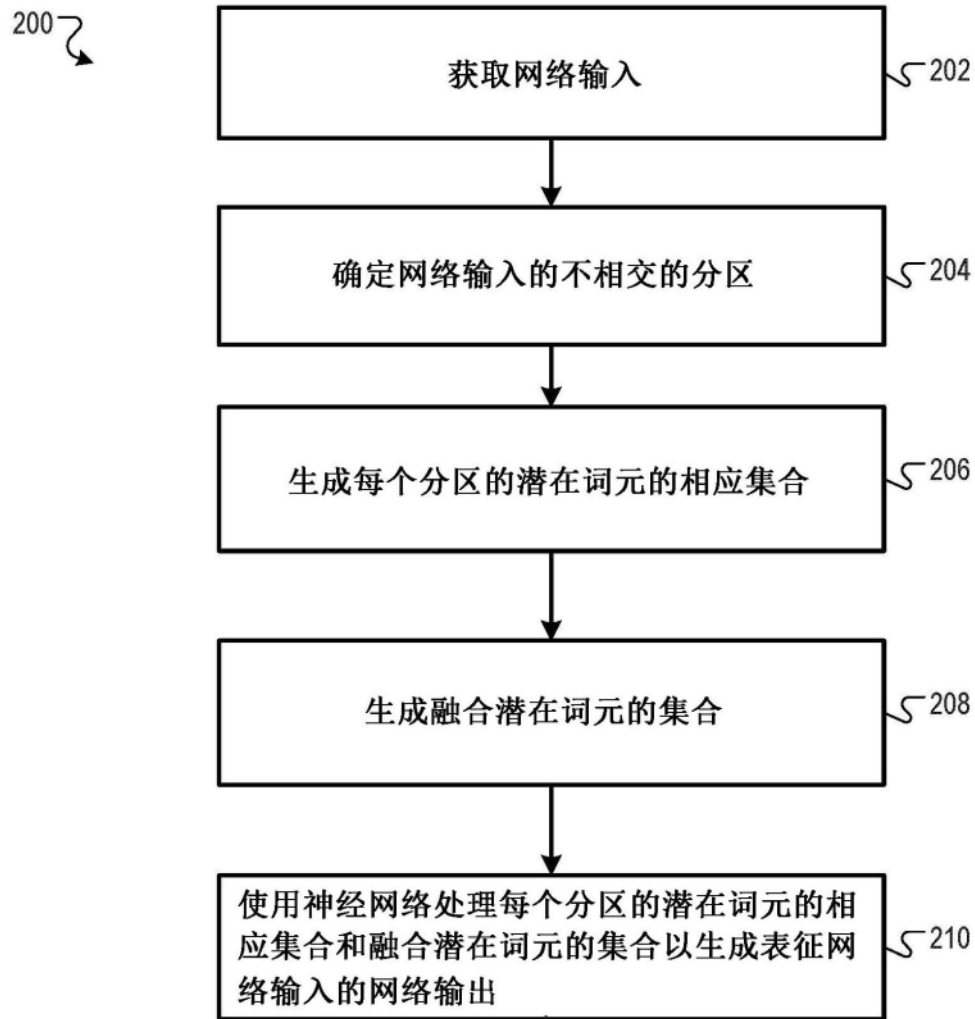


图2

300 ↗

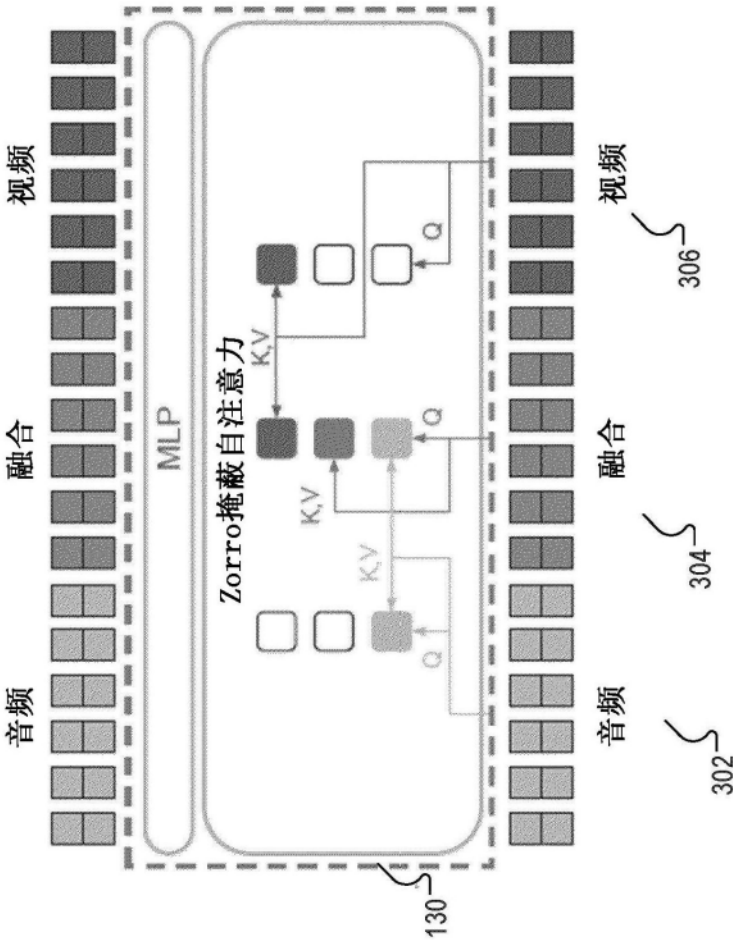


图3

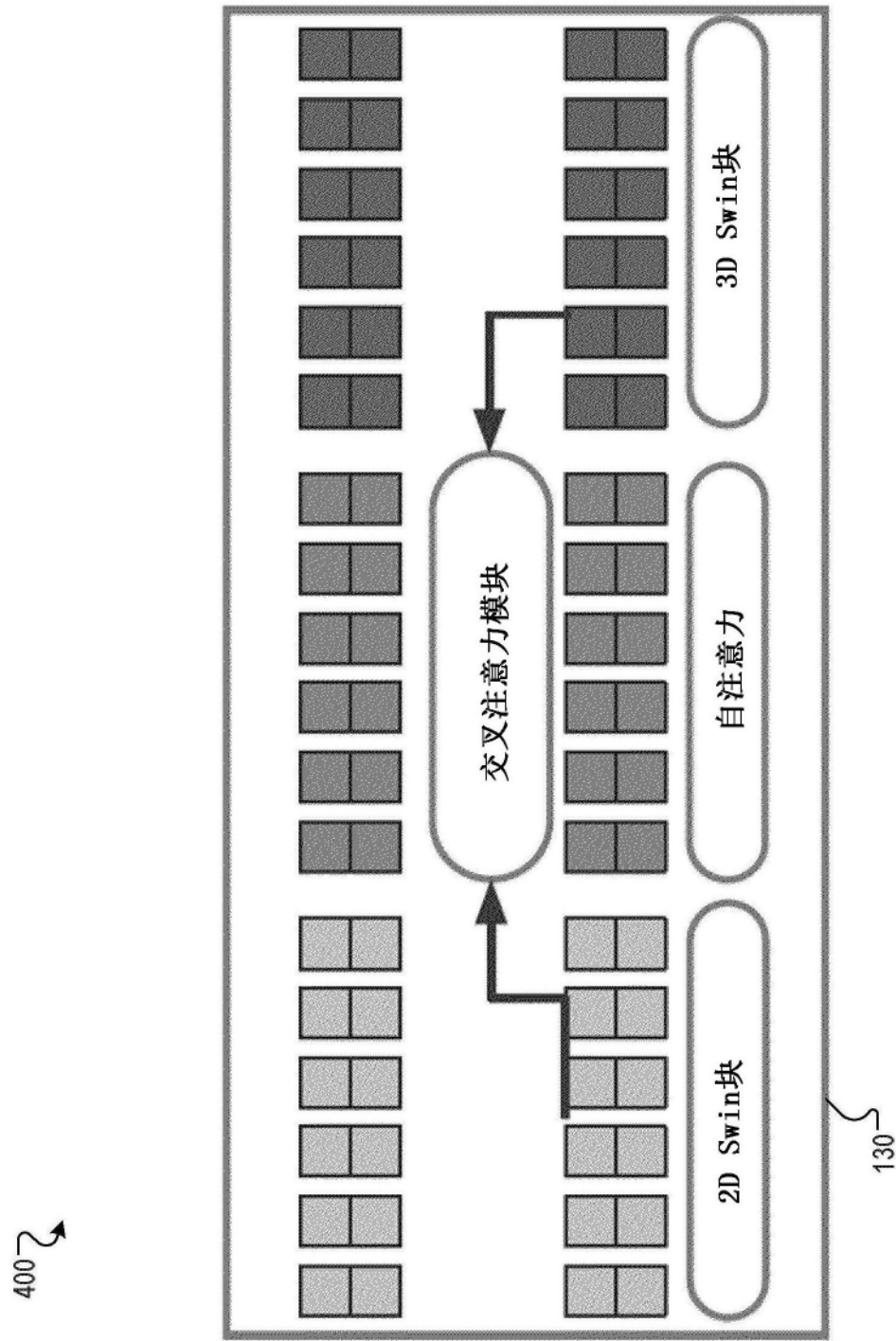


图4

500 ↺

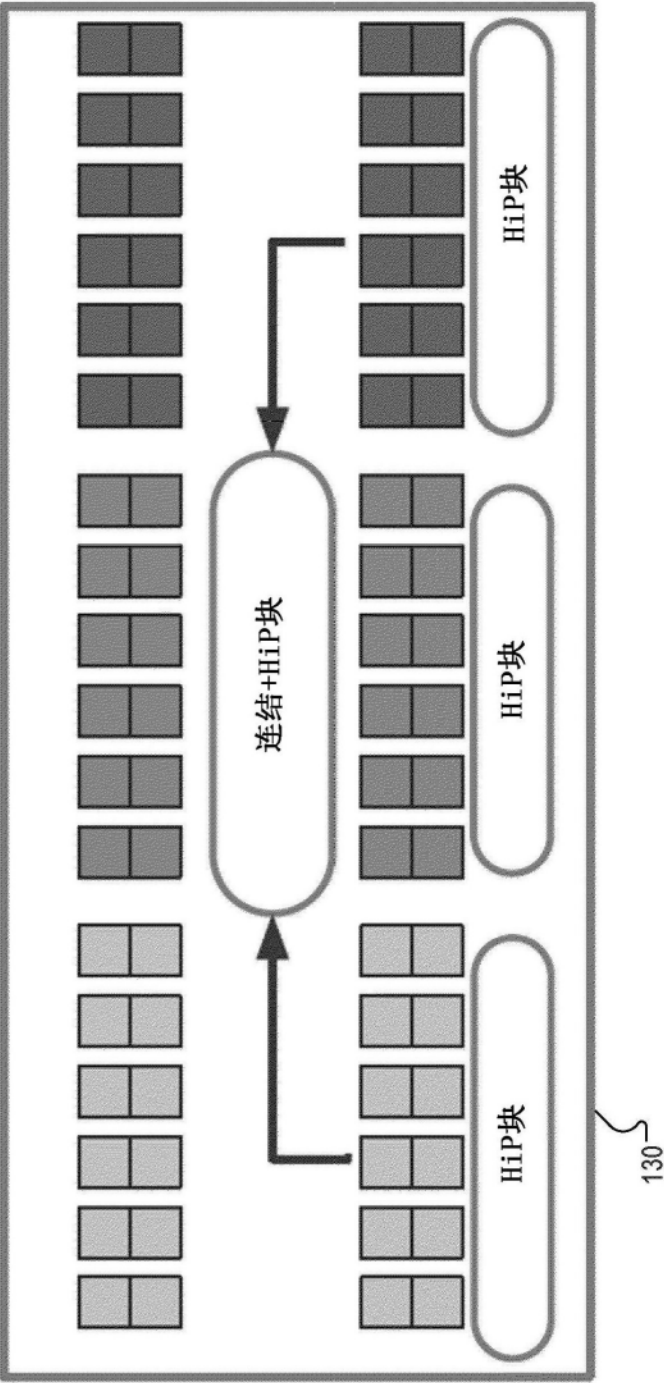


图5

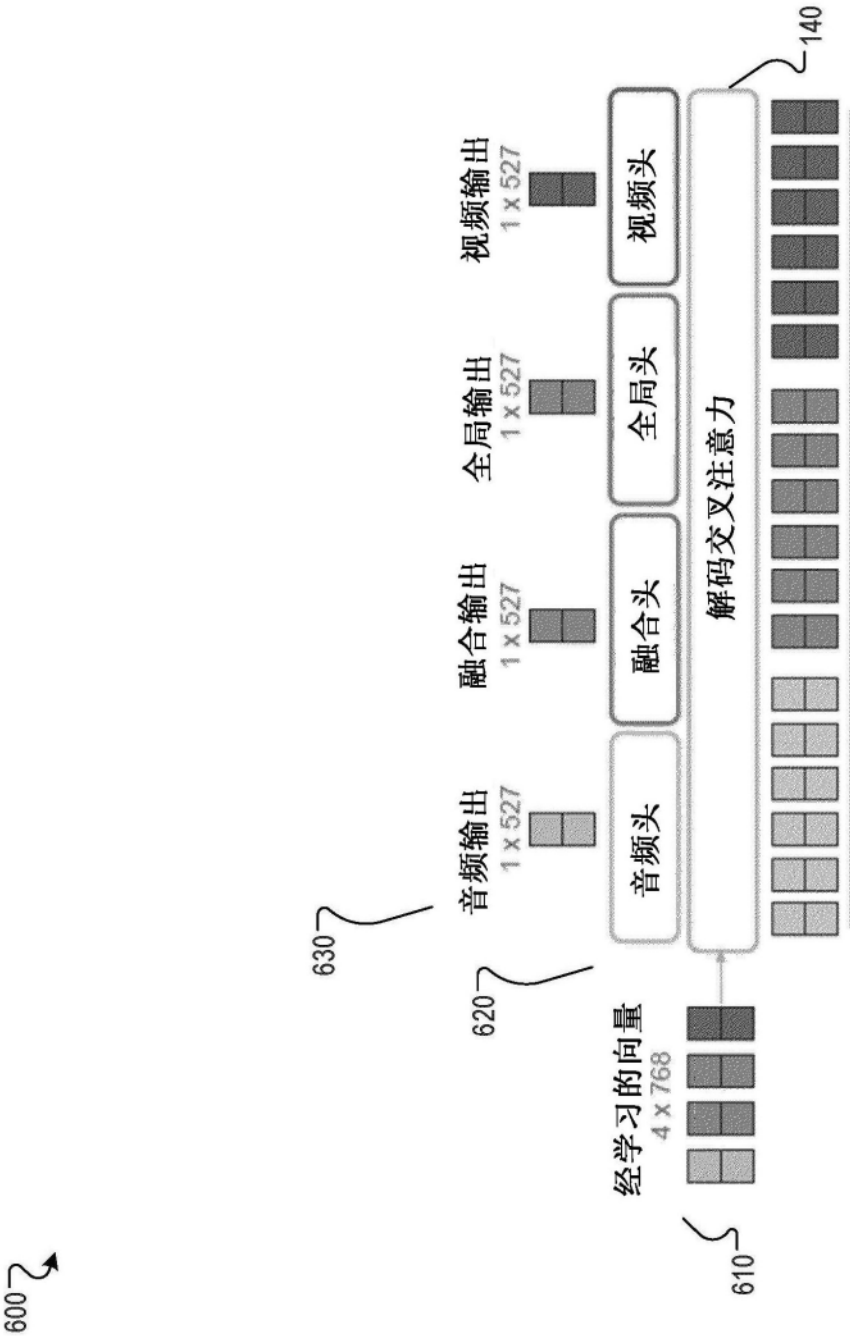


图6

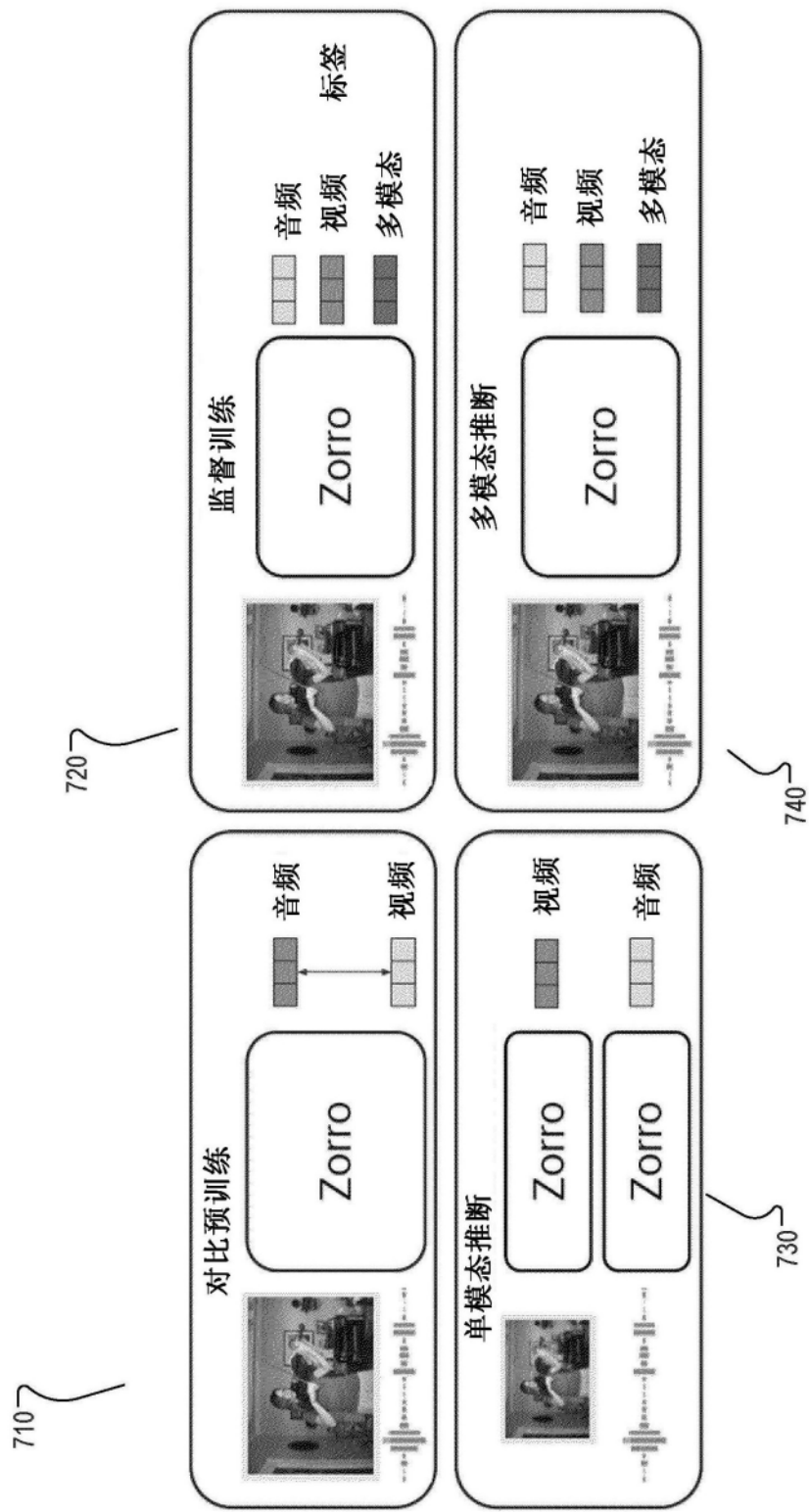


图7