(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 116134454 A
(43)申请公布日 2023.05.16

(21)申请号 202180054067.1
(22)申请日 2021.09.09
(30)优先权数据
63/076,335 2020.09.09 US
17/469,573 2021.09.08 US
(85)PCT国际申请进入国家阶段日
2023.03.01
(86)PCT国际申请的申请数据
PCT/CN2021/117532 2021.09.09
(87)PCT国际申请的公布数据
WO2022/052997 EN 2022.03.17
(71)申请人 华为技术有限公司
地址 518129 广东省深圳市龙岗区坂田华为总部办公楼
(72)发明人 佩曼·帕斯班 吴伊萌 梅赫迪·雷扎霍利扎德
(74)专利代理机构 北京同立钧成知识产权代理有限公司 11205
专利代理师 蔡维华 刘芳
(51)Int.Cl.
G06N 3/096(2023.01)
G06N 3/0455(2023.01)
G06N 3/084(2023.01)

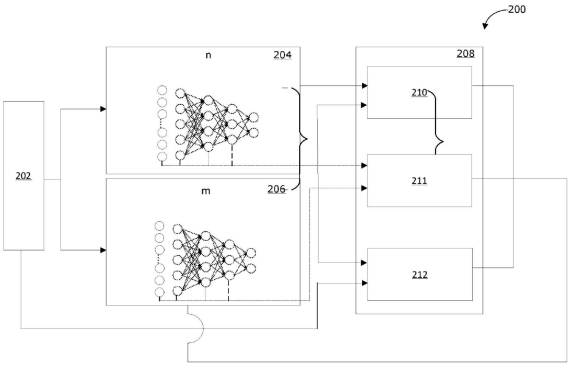
权利要求书4页 说明书17页 附图11页

(54)发明名称

用于使用知识蒸馏训练神经网络模型的方法和系统

(57)摘要

描述了用于将神经模型的训练知识从复杂模型(教师)转移到较不复杂模型(学生)的无关组合知识蒸馏(CKD)方法。除了训练学生以生成近似于教师最终输出和训练输入的真值两者的最终输出外,该方法还通过训练学生的隐藏层来最大化知识转移,以生成输出,该输出近似于映射到针对给定训练输入的学生隐藏层中的每一层的教师隐藏层的子集的表征。



CN 116134454 A

1.一种从具有多个教师隐藏层的教师神经模型向具有多个学生隐藏层的学生神经模型进行知识蒸馏的方法,所述方法包括:

训练所述教师神经模型,其中所述教师神经模型被配置为接收输入并生成教师输出;以及

在多个训练输入上训练所述学生神经模型,其中所述学生神经模型也被配置为接收输入并生成对应的输出,包括:

使用所述教师神经模型处理每个训练输入,以针对所述训练输入生成所述教师输出,所述多个教师隐藏层中的每一层生成教师隐藏层输出;

将所述多个教师隐藏层的子集映射到所述多个学生隐藏层中的每一层;

计算映射到所述多个学生隐藏层中的每一层的所述多个教师隐藏层的子集的所述教师隐藏层输出的表征;以及

训练所述学生以针对所述训练输入中的每一个生成近似于针对所述训练输入的所述教师输出的学生输出,其中,针对所述训练输入中的每一个,所述多个学生隐藏层中的每一层被训练为生成学生隐藏层输出,所述学生隐藏层输出近似于映射到所述多个学生隐藏层中的每一层的所述多个教师隐藏层的子集的所述表征。

2.根据权利要求1所述的方法,还包括训练所述学生,以针对所述训练输入中的每一个,生成近似于所述训练输入的真值的学生输出。

3.根据权利要求2所述的方法,其中训练所述学生以生成近似于针对所述训练输入的所述教师输出的学生输出还包括:

计算所述学生输出和所述教师输出之间的知识蒸馏KD损失;

计算所述学生输出和所述真值之间的标准损失;

计算所述多个学生隐藏层中的每一层和映射到所述多个学生层中的每一层的所述教师隐藏层的子集之间的组合KD CKD损失;

将总损失计算为所述KD损失、所述标准损失和所述CKD损失的加权平均;以及

调整所述学生的参数,以最小化所述总损失。

4.根据权利要求3所述的方法,其中所述CKD损失通过以下方式计算:

$$\mathcal{L}_{CKD} = \sum_{h_i^S \in H^S} MSE(h_i^S, f_i^T)$$

其中

$\mathcal{L}_{CKD}$是所述CKD损失;

MSE()是均方误差函数;

h_i^S是所述学生的第i个隐藏层;

f_i^T是通过$f_i^T = F(H^T(i))$计算的所述教师的第i个隐藏层,其中$H^T(i) = \{h_j^T | j \in M(i)\}$;

F()是由所述教师的第一隐藏层(h_1^T)和第三隐藏层(h_3^T)提供的融合函数,F()的输出被映射到所述学生的第二隐藏层(h_2^S);

H^S和H^T分别是所述学生和所述教师的所有隐藏层的集合;

$H^T(i)$是被选择以待映射到所述学生的第i个隐藏层的所述教师的隐藏层的子集;并且

M()是映射函数,所述映射函数采用引用所述学生的隐藏层的索引,并为所述教师返回

一组索引。

5.根据权利要求4所述的方法，其中所述融合函数F()包括级联操作，后跟线性映射层。

6.根据权利要求5所述的方法，所述融合函数F()由以下定义：

$$F(h_1^T, h_3^T) = mul(W, [h_1^T; h_3^T]) + b$$

其中

";"是级联算子；

mul()是矩阵乘法运算；并且

W和b是可学习参数。

7.根据权利要求4所述的方法，其中所述映射函数M()定义用于组合所述教师的隐藏层的组合策略。

8.根据权利要求7所述的方法，其中所述组合策略为重叠组合、常规组合、跳过组合和交叉组合中的任一种。

9.根据权利要求1至8中任一项所述的方法，其中所述映射还包括，定义用于将所述教师隐藏层映射到所述多个学生隐藏层中的每一层的组合策略。

10.根据权利要求9所述的方法，其中所述组合策略为重叠组合、常规组合、跳过组合和交叉组合中的任一种。

11.根据权利要求1至10中任一项所述的方法，其中所述映射还包括，针对所述学生隐藏层中的每一层，将注意力权重分配给所述多个教师隐藏层的子集中的每一个的所述教师隐藏层输出。

12.根据权利要求3所述的方法，其中所述CKD损失通过以下方式计算：

$$L_{CKD^*}(H^S, H^T) = \sum_{h_i^S \in H^S} MSE\,(h_i^S, f_i^{*T})$$

其中

L_{CKD^*}是所述CKD损失；

MSE()是均方误差函数；

h_i^S是所述学生的第i个隐藏层；并且

f_i^{*T}是针对所述教师的第i个隐藏层的所述教师的隐藏层(H^T)的基于注意力的组合表征。

13.根据权利要求12所述的方法，其中当学生和教师具有相同的维度($|h_i^S| = |h_j^T|$)时，f_i^{*T}通过以下方式计算：

$$f_i^{*T} = \sum_{h_j^T \in H^T} \epsilon_{ij} h_j^T$$

其中

ϵ_{ij}是注意力权重，指示所述教师的所述第j个隐藏层(h_j^T)有多少对所述学生的第i个隐藏层(h_i^S))的知识蒸馏过程有贡献；

h_j^T是所述教师的所述第j个隐藏层；

H^T是所述教师的所有隐藏层的集合。

14.根据权利要求12所述的方法，其中当学生和教师具有不同的维度（$|h_i^S| \neq |h_j^T|$），，f_i^{*T}通过以下方式计算：

$$f_i^{*T} = \sum_{h_j^T \in H^T} \epsilon_{ij}(W_i h_j^T)$$

其中

ϵ_{ij}是注意力权重，指示所述教师的第j个隐藏层（h_j^T）有多少对所述学生的第i个隐藏层（h_i^S））的知识蒸馏过程有贡献；

h_j^T是所述教师的第j个隐藏层；

H^T是所述教师的所有隐藏层的集合；

$W_i \in \mathbb{R}^{|h_i^S| \times |h_j^T|}$是所述教师的第i个隐藏层的权重值。

15.根据权利要求14所述的方法，其中所有注意力权重$W\epsilon_{ij_i}$的总和为1。

16.根据权利要求13所述的方法，其中所述注意力权重（ϵ_{ij}）通过以下方式计算：

$$\epsilon_{ij} = \frac{e^{\varphi_{ij}}}{\sum_{h_k^T \in H^T} e^{\varphi_{ik}}}$$

其中$\varphi_{ij} = \Phi(h_i^S, h_j^T)$，其中$\Phi(h_i^S, h_j^T)$是所述教师的隐藏层和由所述教师的隐藏层生成的输出的能量函数，其中φ_{ij}确定由所述学生的第i个隐藏层（h_i^S）生成的输出和由所述教师的第j个隐藏层（h_j^T））生成的输出之间的能量评分，其中所述能量评分指示两个生成的输出之间的相似性。

17.根据权利要求16所述的方法，其中所述能量函数$\Phi(h_i^S, h_j^T)$计算为所述学生的第i个隐藏层（h_i^S）的输出和由所述教师的第j个隐藏层（h_j^T）生成的输出的加权值的点积：

$\Phi(h_i^S, h_j^T) \equiv < h_i^S, h_j^T >$。

18.根据权利要求14所述的方法，其中所述注意力权重（ϵ_{ij}）通过以下方式计算：

$$\epsilon_{ij} = \frac{e^{\varphi_{ij}}}{\sum_{h_k^T \in H^T} e^{\varphi_{ik}}}$$

其中$\varphi_{ij} = \Phi(h_i^S, h_j^T)$，其中$\Phi(h_i^S, h_j^T)$是所述学生的隐藏层和由所述教师的隐藏层生成的输出的能量函数，其中φ_{ij}确定由所述学生的第i个隐藏层（h_i^S）生成的输出和由所述教师的第j个隐藏层（h_j^T））生成的输出之间的能量评分，其中所述能量评分指示两个生成的输出之间的相似性。

19.根据权利要求17所述的方法，其中所述能量函数$\Phi(h_i^S, h_j^T)$计算为所述学生的第i个隐藏层（h_i^S）的输出和由所述教师的第j个隐藏层（h_j^T）生成的输出的加权值的点积：

$\Phi(h_i^S, h_j^T) \equiv < h_i^S, W_i h_j^T >$。

20.一种从各自具有多个教师隐藏层的多个教师神经模型向具有多个学生隐藏层的学

生神经模型进行知识蒸馏的方法,所述方法包括:

推断所述多个教师神经模型,其中所述多个教师神经模型中的每一个被配置为接收输入并生成教师输出;以及

在多个训练输入上训练所述学生神经模型,其中所述学生神经模型也被配置为接收输入并生成学生输出,包括:

使用所述多个教师神经模型处理每个训练输入,以针对所述训练输入生成多个教师输出,所述多个教师神经模型中的每一个的多个教师隐藏层中的每一层生成教师隐藏层输出;

将所述多个教师神经模型的所述多个教师隐藏层的子集映射到所述多个学生隐藏层中的每一层;

计算所述多个教师隐藏层的子集的教师隐藏层输出的表征,所述多个教师隐藏层的子集映射到所述多个学生隐藏层中的每一层;以及

训练所述学生以针对所述训练输入中的每一个生成近似于针对所述训练输入的所述教师输出的学生输出,其中所述多个学生隐藏层中的每一层针对所述训练输入中的每一个,被训练为生成学生隐藏层输出,所述学生隐藏层输出近似于映射到所述多个学生隐藏层中的每一层的所述多个教师隐藏层的子集的所述表征。

21.一种处理系统,包括:

处理设备;

存储器,所述存储器存储有指令,所述指令当由所述处理设备执行时,使所述处理系统执行根据权利要求1至20中任一项所述的方法。

22.一种计算机可读介质,包括指令,所述指令当由处理系统的处理设备执行时,使所述处理系统执行根据权利要求1至20中任一项所述的方法。

用于使用知识蒸馏训练神经网络模型的方法和系统

[0001] 相关申请的交叉申请

[0002] 本申请要求2020年9月9日提交的名称为“用于使用知识蒸馏训练神经网络模型的方法和系统(METHOD AND SYSTEM FOR TRAINING A NEURAL NETWORK MODEL USING KNOWLEDGE DISTILLATION)”、申请号为63/076,335的美国临时专利申请和2021年9月8日提交的名称为“用于使用知识蒸馏训练神经网络模型的方法和系统(METHOD AND SYSTEM FOR TRAINING A NEURAL NETWORK MODEL USING KNOWLEDGE DISTILLATION)”、申请号为17/469,573的美国专利申请的在先申请优先权和权益,这些申请的内容通过引用并入本文。

技术领域

[0003] 本申请涉及用于训练机器学习模型的方法和系统,尤其涉及用于使用知识蒸馏训练深度神经网络的方法和系统。

背景技术

[0004] 机器学习模型针对每个接收的输入推断(即预测)特定输出。推断的(即预测的)特定输出可以以la可以属于的形式出现。例如,机器学习模型可以基于接收的图像推断(即预测)特定输出,推断的(即预测的)输出包括一组类别中的每个类别的概率分数,其中每个分数表示图像类似于属于该特定类别的对象的概率。

[0005] 机器学习模型是使用学习算法进行学习的,如随机梯度下降。使用此类技术学习的机器学习模型是近似于该输入到输出过程的深度人工神经网络。用于近似机器学习模型的深度人工神经网络包括输入层、一个或多个隐藏层、以及输出层,其中所有隐藏层都具有参数,并且非线性应用于这些参数。用于近似机器学习模型的深度人工神经网络通常被称为神经网络模型。

[0006] 知识蒸馏(Knowledge distillation,KD)是神经网络压缩技术,通过该技术,复杂神经网络模型的学习参数或知识被转移到较不复杂的神经网络模型,该神经网络模型能够以较少的计算资源成本和时间作出与复杂模型相当的推断(即预测)。在此,复杂神经网络模型是指具有相对高数量的计算资源(如GPU/CPU功率和计算机内存空间)的神经网络模型和/或包括相对高数量的隐藏层的那些神经网络模型。为了KD的目的,复杂神经网络模型有时被称为教师神经网络模型(teacher neural network model,T)或简称教师。教师的典型缺点是,其可能需要显著的计算资源,这些计算资源在消费电子设备(如移动通信设备或边缘计算设备)中可能不可用。此外,由于教师神经网络模型本身的复杂度,教师神经网络模型通常需要显著量的时间来推断(即预测)针对输入的特定输出,并且因此教师神经网络模型可能不适合部署到消费计算设备以在其中使用。因此,KD技术应用于提取或蒸馏教师神经网络模型的学习参数或知识,并将此类知识传授给具有更快的推断时间和降低的计算资源和内存空间成本的较不复杂的神经网络模型,这可能会在消费计算设备(如边缘设备)上花费更少的精力。较不复杂的神经网络模型通常被称为学生神经网络模型(student neural network model,S)或简称学生。

[0007] 现有技术的KD技术仅考虑针对接收的输入的特定输出的最终推断(即预测)以计算损失函数,因此现有技术的KD技术不能处理从教师的隐藏层到学生的隐藏层的知识转移。因此,可以提高KD技术的准确性,尤其是对于具有多个深度隐藏层的教师和学生神经网络模型。

[0008] 患者知识蒸馏(patient knowledge distillation,PKD)关注该问题,并引入了层到层的成本函数,还被称为内部层映射。教师神经网络的隐藏层的输出还用于训练学生神经网络模型的一个或多个隐藏层,而不是仅匹配学生和教师神经网络模型的推断的(即预测的)输出。隐藏层可以指神经网络中的内部层。具体地,PKD选择由教师神经网络模型的隐藏层生成的输出的子集,并使用由教师神经网络模型的隐藏层生成的输出来训练学生神经网络的一个或多个隐藏层,如图1所示。具体地,图1示出了示意图,其中具有n=3个内部层的教师神经网络模型100(图1右侧所示的神经网络)与PKD一起用于训练具有m=2个内部层的学生神经网络110。当n>m时,跳过由虚线指示的所示实施例中的教师神经网络神经网络模型的内部层之一,使得教师神经网络模型的剩余内部层中的每个内部层直接用于训练对应的学生神经网络模型110的内部层之一。如图1所示,不仅由学生神经网络模型110和教师神经网络模型100推断(即预测)的最终输出用于计算PKD中KD损失的损失,而且教师和学生神经网络模型的内部层的输出也被匹配,从而学生神经网络模型110可以从教师神经网络模型100内部的信息流中学习。

[0009] 然而,不存在明确的方法来决定跳过教师的哪些隐藏层,以及保留教师的哪些隐藏层以进行蒸馏。因此,当将n层教师神经网络模型提取成m层学生神经网络模型时,跳过的隐藏层可能会存在显著的信息损失。当n>>m时,信息损失变得更明显。

[0010] 因此,期望提供知识蒸馏方法的改进,以最小化教师神经网络模型的跳过的内部(即隐藏)层的信息损失。

发明内容

[0011] 本公开提供用于使用知识蒸馏(KD)训练深度神经网络模型的方法和系统,该方法和系统将教师神经网络模型的跳过的内部层的信息损失最小化。

[0012] 在一些方面,本公开描述了KD方法,其将教师神经网络模型的内部(即隐藏)层映射到学生神经网络模型的对应的内部层(即隐藏层),以最小化信息损失。

[0013] 在一些方面,本公开描述了KD方法,其包括教师模型和学生模型的内部层(即隐藏层)的自动映射和层选择。

[0014] 在一些方面,本公开描述了KD方法,其对层映射采用组合方法,其中n层教师神经网络模型的一个或多个内部(即隐藏)层映射到m层学生神经网络模型的一个或多个层(即隐藏层),其中n>m,从而可以最小化信息损失。

[0015] 在一些方面,本公开描述了KD方法,其采用组合方法将教师神经网络模型的一个或多个隐藏层映射到学生神经网络模型的内部层(即,隐藏层),与教师和学生神经网络模型的架构无关。例如,本公开的KD方法的各个方面使得能够将知识从教师神经网络模型的内部(即隐藏)层(如变压器模型(transformer model))蒸馏到学生神经网络模型的一个或多个层(如递归神经网络模型)。

[0016] 在一些方面,本公开描述了KD方法,其可以在使用通用语言理解评估(general

language understanding evaluation,GLUE)基准评估的训练的变压器的双向编码器表示(bidirectional encoder representations for transformers,BERT)深度学习学生神经网络模型的性能方面进行改进。

[0017] 在一些方面,本公开描述了KD方法,其可以在神经机器翻译模型方面进行改进。

[0018] 在一些方面,本公开描述了知识蒸馏方法,其可能能够将多个教师神经网络模型的内部层(即隐藏层)映射到单个学生神经网络模型。映射内部层(即隐藏层)涉及将教师神经网络模型的一个或多个内部(隐藏)层与单个学生神经网络模型的一个内部(隐藏)层相关联。

[0019] 在一些方面,本文描述的KD方法可以用于训练可以部署到边缘设备的学生神经网络模型。

[0020] 在一些方面,本文描述的方法可以聚合不同的信息源,包括多语言/多域语言理解/处理/翻译模型。

[0021] 在一些方面,本文描述的KD方法可以是与任务无关的,使得其可以适用于训练针对任何特定任务的模型,包括如对象分类等计算机视觉任务。

[0022] 在一些方面,对于服务外部用户的服务器端模型,可以通过本文描述的KD方法组合多个模型,最终训练的教师模型可以被上传到服务器。通过能够从不同的教师神经网络模型执行知识蒸馏,该方法可以免疫任何对抗性攻击。

[0023] 根据本公开的第一方面的第一实施例,提供了从具有多个教师隐藏层的教师机器学习模型向具有多个学生隐藏层的学生机器学习模型进行知识蒸馏的方法。该方法包括训练教师机器学习模型,其中该教师机器学习模型被配置为接收输入并生成教师输出,以及在多个训练输入上训练学生机器学习模型,其中学生机器学习模型也被配置为接收输入并生成对应的输出。训练学生机器学习模型包括使用教师机器学习模型处理每个训练输入,以针对训练输入生成教师输出,其中多个教师隐藏层中的每一层生成教师隐藏层输出。训练学生机器学习模型还包括将多个教师隐藏层的子集映射到多个学生隐藏层中的每一层。训练学生机器学习模型还包括计算多个教师隐藏层的子集的教师隐藏层输出的表征,该多个教师隐藏层的子集映射到多个学生隐藏层中的每一层。训练学生机器学习模型还包括训练学生以针对训练输入中的每一个生成近似于针对训练输入的教师输出的学生输出,其中多个学生隐藏层中的每一层针对训练输入中的每一个,被训练为生成学生隐藏层输出,该学生隐藏层输出近似于映射到多个学生隐藏层中每一个的多个教师隐藏层的子集的表征。

[0024] 在第一方面的第一实施例的一些示例或所有示例中,该方法还包括训练学生,以针对训练输入中的每一个,生成近似于训练输入的真值的学生输出。

[0025] 在第一方面的第一实施例的一些示例或所有示例中,训练学生以生成近似于针对训练输入的教师输出的学生输出还包括:计算学生输出和教师输出之间的知识蒸馏(KD)损失;计算学生输出和上述真值之间的标准损失;计算多个学生隐藏层中的每一层和映射到多个学生层中的每一层的教师隐藏层的子集之间的组合KD(combinatorial KD,CKD)损失;将总损失计算为KD损失、标准损失和CKD损失的加权平均;以及,调整学生的参数,以最小化总损失。

[0026] 在一些示例中,CKD损失通过以下方式计算:

[0027] $$\mathcal{L}_{CKD} = \sum_{h_i^S \in H^S} MSE(h_i^S, f_i^T)$$

[0028] 其中，$\mathcal{L}_{CKD}$是CKD损失，MSE()是均方误差函数，h_i^S是学生的第i个隐藏层，f_i^T是通过$f_i^T = F(H^T(i))$计算的教师的第i个隐藏层，其中$H^T(i) = \{h_j^T | j \in M(i)\}$，F()是教师的第一隐藏层（$h_1^T$）和第三隐藏层（$h_3^T$）提供的融合函数，F()的输出被映射到学生的第二隐藏层（h_2^S），H^S和H^T分别是学生和教师的所有隐藏层的集合，$H^T(i)$是被选择以待映射到学生的第i个隐藏层的教师隐藏层的子集，并且M()是映射函数，其采用引用学生的隐藏层的索引，并为教师返回一组索引。

[0029] 在一些示例或所有示例中，融合函数F()包括级联操作，后跟线性映射层。

[0030] 在一些示例或所有示例中，融合函数F()由以下定义：

[0031] $$F(h_1^T, h_3^T) = mul(W, [h_1^T; h_3^T]) + b$$

[0032] 其中，“;”是级联算子，mul()是矩阵乘法运算，并且W和b是可学习参数。在该示例中，我们仅考虑了两个层，即来自教师侧的层3和1，但这可以扩展到任何数量的层。

[0033] 在一些示例或所有示例中，映射函数M()定义了用于组合教师的隐藏层的组合策略。

[0034] 在一些示例或所有示例中，上述映射还包括，定义用于将该教师隐藏层映射到多个学生隐藏层中的每一层的组合策略。

[0035] 在一些示例或所有示例中，该组合策略为重叠组合、常规组合、跳过组合和交叉组合中的任一种。

[0036] 在一些示例或所有示例中，上述映射还包括，针对学生隐藏层中的每一层，将注意力权重分配给多个教师隐藏层的子集中的每一个的教师隐藏层输出。

[0037] 在一些示例中，CKD可以因注意力或其它形式的组合来增强，在本文中标注为CKD^*。CKD^*损失通过以下方式计算：

[0038] $$L_{CKD^*}(H^S, H^T) = \sum_{h_i^S \in H^S} MSE(h_i^S, f_i^{*T})$$

[0039] 其中L_{CKD^*}是CKD^*损失，MSE()是均方误差函数，h_i^S是学生的第i个隐藏层，并且f_i^{*T}是组合版本，其可以是针对教师的第i个隐藏层的基于注意力的或其它形式的组合的教师的隐藏层（H^T）。

[0040] 在一些示例中，当学生和教师具有相同的维度（$|h_i^S| = |h_j^T|$）时，f_i^{*T}通过以下方式计算：

[0041] $$f_i^{*T} = \sum_{h_j^T \in H^T} \epsilon_{ij} h_j^T$$

[0042] 其中，$\in_{ij}$是注意力权重，指示教师的第j个隐藏层（h_j^T）有多少对学生的第i个隐藏层（h_i^S））的知识蒸馏过程有贡献，h_j^T是教师的第j个隐藏层，并且H^T是教师的所有隐藏层的集合。

[0043] 在一些示例中，当学生和教师具有不同的维度（$|h_i^s| \neq |h_j^T|$），，f_i^{*T}通过以下方式计算：

[0044] $$f_i^{*T} = \sum_{h_j^T \in H^T} \epsilon_{ij}(W_i h_j^T)$$

[0045] 其中，$\in_{ij}$是注意力权重，指示教师的第j个隐藏层（h_j^T）有多少对学生的第i个隐藏层（h_i^s））的知识蒸馏过程有贡献，h_j^T是教师的第j个隐藏层，H^T是教师的所有隐藏层的集合，并且$W_i \in \mathbb{R}^{|h_i^s| \times |h_j^T|}$是注意维度失配的权重值。

[0046] 注意力权重（$\in_{ij}$）的总和为1。

[0047] 在一些示例中，注意力权重（$\in_{ij}$）通过以下方式计算：

[0048] $$\epsilon_{ij} = \frac{e^{\varphi_{ij}}}{\sum_{h_k^T \in H^T} e^{\varphi_{ik}}}$$

[0049] 其中$\varphi_{ij} = \Phi(h_i^s, h_j^T)$，其中$\Phi(h_i^s, h_j^T)$是能量函数，其中$\varphi_{ij}$确定由学生的第i个隐藏层（$h_i^s$）生成的输出和由教师的第j个隐藏层（$h_j^T$））生成的输出之间的能量评分，其中该能量评分指示两个生成的输出之间的相似性。

[0050] 在一些示例中，能量函数$\Phi(h_i^s, h_j^T)$计算为学生的第i个隐藏层（h_i^s）的输出和由教师的第j个隐藏层（h_j^T）生成的输出的加权值的点积：

[0051] $\Phi(h_i^s, h_j^T) \equiv < h_i^s, h_j^T >$。

[0052] 在一些示例中，注意力权重（$\in_{ij}$）通过以下方式计算：

[0053] $$\epsilon_{ij} = \frac{e^{\varphi_{ij}}}{\sum_{h_k^T \in H^T} e^{\varphi_{ik}}}$$

[0054] 其中$\varphi_{ij} = \Phi(h_i^s, h_j^T)$，其中$\Phi(h_i^s, h_j^T)$是能量函数，其中$\varphi_{ij}$确定由学生的第i个隐藏层（$h_i^s$）生成的输出和由教师的第j个隐藏层（$h_j^T$））生成的输出之间的能量评分，其中该能量评分指示两个生成的输出之间的相似性。

[0055] 在一些示例中，能量函数$\Phi(h_i^s, h_j^T)$计算为学生的第i个隐藏层（h_i^s）的输出和由教师的第j个隐藏层（h_j^T）生成的输出的加权值的点积：

[0056] $\Phi(h_i^s, h_j^T) \equiv < h_i^s, W_i h_j^T >$。

[0057] 根据本公开的第一方面的第二实施例，提供了从各自具有多个教师隐藏层的多个教师机器学习模型向具有多个学生隐藏层的学生机器学习模型进行知识蒸馏的方法。该方法包括：训练多个教师机器学习模型，其中该多个教师机器学习模型中的每一个被配置为接收输入并生成教师输出；以及，在多个训练输入上训练学生机器学习模型，其中该学生机器学习模型也被配置为接收输入并生成学生输出。训练学生机器学习模型包括使用多个教师机器学习模型处理每个训练输入，以针对训练输入生成多个教师输出，多个教师机器学习模型中的每一个的多个教师隐藏层中的每一层生成教师隐藏层输出。训练学生机器学习

模型还包括将多个教师机器学习模型的多个教师隐藏层的子集映射到多个学生隐藏层中的每一层。训练学生机器学习模型还包括:计算多个教师隐藏层的子集的教师隐藏层输出的表征,该多个教师隐藏层的子集映射到多个学生隐藏层中的每一层。训练学生机器学习模型还包括:训练学生以针对训练输入中的每一个生成近似于针对训练输入的教师输出的学生输出,其中多个学生隐藏层中的每一层针对训练输入中的每一个,被训练为生成学生隐藏层输出,该学生隐藏输出近似于映射到多个学生隐藏层中的每一层的多个教师隐藏层的子集的表征。

[0058] 根据本公开的另一方面,提供了计算设备,该计算设备包括处理器和存储器,该存储器上有形地存储有用于由处理器执行的可执行指令。响应于处理器的执行,可执行指令使计算设备执行上述和本文描述的方法。

[0059] 根据本公开的另一方面,提供了非瞬时性机器可读存储介质,其上有形地存储有用于由计算设备的处理器执行的可执行指令。响应于处理器的执行,可执行指令使计算设备执行上述和本文描述的方法。

[0060] 对于本领域的普通技术人员来说,在审阅以下特定实施方式的描述后,本公开的其它方面和特征将变得显而易见。

附图说明

[0061] 现在将通过示例的方式参考示出本申请示例实施例的附图。

[0062] 图1示出了现有技术PKD方法的示意图;

[0063] 图2示出了根据本公开的用于使用KD训练神经网络模型的机器学习系统的框图;

[0064] 图3A示出了还被称为交叉组合的第一组合策略;

[0065] 图3B示出了还被称为常规组合的第二组合策略;

[0066] 图3C示出了还被称为跳过组合的第三组合策略;

[0067] 图3D示出了还被称为重叠组合的第四组合策略;

[0068] 图4示出了根据本公开的不具有注意力的示例知识蒸馏方法的流程图;

[0069] 图5示出了在图4的步骤406处确定最终损失值的示例方法的流程图;

[0070] 图6示出了图4中方法的部分的示例伪代码;

[0071] 图7示出了具有注意力的增强CKD的高级示意性架构;

[0072] 图8示出了根据本公开的具有注意力的示例知识蒸馏方法的流程图;

[0073] 图9示出了表格,示出了在执行通用语言理解评估(GLUE)基准时各种KD模型的模拟结果;

[0074] 图10示出了根据本公开的用于从多个教师向一个教师进行知识蒸馏的示例方法的流程图;

[0075] 图11示出了用于计算注意力权重的示例方法的流程图,该示例方法可以在图10所示出的方法的步骤1006处实现;并且

[0076] 图12示出了可以用于实现本文公开的实施例的示例简化处理系统的框图。

具体实施方式

[0077] 本公开结合附图进行,其中示出了技术方案的实施例。然而,可以使用许多不同的

实施例,并且因此不应将描述解释为限于本文中阐述的实施例。相反,提供这些实施例使得本申请将是彻底和完整的。在可能的情况下,在附图和以下描述中使用相同的附图标记来指代相同的元件,并且在替代实施例中使用素数表示法来指示相同的元件、操作或步骤。所示系统和设备的功能元件的单独框或所示分离不一定需要这些功能的物理分离,因为这些元件之间的通信可以在没有任何此类物理分离的情况下通过消息传递、函数调用、共享内存空间等方式发生。因此,尽管本文为了便于解释,它们被示出为分离的,但是这些功能不需要在物理或逻辑上分离的平台中实现。不同的设备可以具有不同的设计,使得尽管一些设备在固定功能硬件中实现一些功能,但其它设备可以在可编程处理器中实现这些功能,该处理器具有从机器可读存储介质获得的代码。最后,以单数提及的元件可以是复数,反之亦然,除非上下文明确或固有地指示。

[0078] 本文中阐述的实施例表示足以实践请求保护的主题的信息,并示出了实践此类主题的方法。根据附图阅读以下描述之后,本领域技术人员会理解请求保护的主题的概念,并会认识到这些概念的应用在本文中并没有特别提及。应当理解,这些概念和应用落入本公开和所附权利要求书的范围之内。

[0079] 此外,应当理解,本文公开的执行指令的任何模块、组件或设备可以包括或以其它方式接入一个或多个非瞬时性计算机/处理器可读存储介质,该介质用于存储信息,如计算机/处理器可读指令、数据结构、程序模块和/或其它数据。非瞬时性计算机/处理器可读存储介质的示例的非穷举式清单包括磁带盒、磁带、磁盘存储或其它磁存储设备,光盘,如光盘只读存储器(compact disc read-only memory,CD-ROM)、数字视频盘或数字多功能盘(即digital versatile disc,DVD)、蓝光盘™或其它光存储器,在任何方法或技术中实现的易失性和非易失性、可移动和不可移动介质,随机存取存储器(random-access memory,RAM),只读存储器(read-only memory,ROM),电可擦除可编程只读存储器(electrically erasable programmable read-only memory,EEPROM),闪存或其它存储技术。任何此类非瞬时性计算机/处理器存储介质可以是设备的一部分,或可访问该设备或可与该设备连接。用于实现本文描述的应用或模块的计算机/处理器可读/可执行指令可以由此类非瞬时性计算机/处理器可读存储介质存储或以其它方式保存。

[0080] 以下是下面描述中可能使用的首字母缩略词和相关定义的部分列表:

[0081] NMT:神经机器翻译(neural machine translation)

[0082] BERT:变压器的双向编码器表示(bidirectional encoder representation from transformers)

[0083] KD:知识蒸馏(knowledge distillation)

[0084] PKD:患者知识蒸馏(patient knowledge distillation)

[0085] S:学生(student)

[0086] T:老师(teacher)

[0087] RKD:常规知识蒸馏(regular knowledge distillation)

[0088] CKD:组合知识蒸馏(combinatorial knowledge distillation)

[0089] RC:常规组合(regular combination)

[0090] OC:重叠组合(overlap combination)

[0091] SC:跳过组合(skip combination)

[0092] CC:交叉组合(cross combination)

[0093] 本文描述了组合知识蒸馏(CKD)方法,用于改进作为在训练学生神经网络模型期间跳过教师神经网络模型的一个或多个隐藏层的结果的信息损失。教师神经网络模型和学生神经网络模型被训练用于特定任务,如对象分类、神经机器翻译等。本描述还描述了知识蒸馏方法的示例实施例,该方法可以使得为了KD的目的,将教师神经网络模型的一个或多个隐藏层映射到单个学生神经网络模型的隐藏层。映射涉及将教师神经网络模型的一个或多个层关联到学生神经网络模型的隐藏层。

[0094] 传统的不经KD训练神经网络模型(即学生神经网络模型)涉及根据等式(1)以负对数似然损失函数的形式最小化交叉熵函数($\mathcal{L}_{nll}$):

[0095] $$\mathcal{L}_{nll}(\theta_S) = -\sum_{v=1}^{|V|} 1(y=v)\log p(y=v|x;\theta_S) \qquad (1)$$

[0096] 1(y=v)是指示函数,如果神经网络模型y针对训练数据样本生成的推断的(即预测的)输出等于真值v,则输出“1”,否则其输出“0”。变量(x,y)是包括在包括几个训练数据样本的训练数据集中的训练数据样本的元组,其中x是输入,并且y是真值输出。参数θ_S和|V|分别是神经网络模型的参数集合和输出数量。该损失可以被称为标准损失。

[0097] 然而,当执行传统的神经网络模型训练时,当指示函数1(y=v)返回零时,神经网络模型不会接收到由神经网络模型推断(即预测)的针对不正确输出的任何反馈。没有负反馈来惩罚由神经网络模型推断(即预测)的不正确输出y,可能需要更长的时间来训练神经网络模型。通过将KD应用于训练神经网络模型的过程可部分解决该问题,其中呈等式(1)形式的损失函数被扩展了附加项,如等式(2)所示出的:

[0098] $$\mathcal{L}_{KD}(\theta_T,\theta_S) = -\sum_{v=1}^{|V|} q(y=v|x;\theta_T)\times\log p(y=v|x;\theta_S) \qquad (2)$$

[0099] 其中,由神经网络模型推断(即预测)的输出由于其自身相对于真值的标准损失而受到惩罚,但也由于由$q(y=v|x;\theta_T)$给出的教师模型的隐藏层生成的输出而受到损失,这可以被称为KD损失。在等式(2)中,损失函数或KD损失的第一分量,即$q((y=v|x;\theta_T))$,当其与由教师模型的softmax函数作出的推断(即预测)(也被称为软标签)进行比较时通常被称为软损失。其余的损失分量,如标准损失,被称为硬损失。因此,在典型KD方法中,根据等式(3),总体损失函数包括至少两个损失项,即标准损失和KD损失:

[0100] $$\mathcal{L} = \alpha\,\mathcal{L}_{nll} + (1-\alpha)\,\mathcal{L}_{KD} \qquad (3)$$

[0101] 其中α是具有值0≤α≤1的超参数。超参数是神经网络模型外部的配置,并且其值无法从数据中估计。

[0102] 图2示出了根据本公开的用于使用KD训练神经网络模型的机器学习系统200的框图,也可以被称为组合知识蒸馏(CKD)。机器学习系统200可以由电子设备(未示出)的一个或多个物理处理单元实现,例如由一个或多个处理单元执行计算机可读指令(其可以存储在机器人电子设备的存储器中)以执行本文描述的方法。

[0103] 在所示的实施例中,包括输入张量(x)和对应的输出张量(y)的元组202可以提供给n层教师神经网络模型204,以下称为教师204,以及m层学生神经网络模型206,以下称为学生206,其中n>m。元组202是训练数据样本,其是包括多个元组202的训练数据集的一部分。通常,教师204和学生206中的每一个都被配置为基于元组202的输入张量(x)和神经网

络的参数集合推断(即预测)输出张量(y)的神经网络。教师204可以是复杂神经网络模型。学生206可以不如教师204复杂(n＞m或具有较少隐藏层和较少模型参数),使得学生206需要比教师204更少的计算资源成本,并且可以在相当短的时间内推断(即预测)针对特定任务的输出。

[0104] 在一些实施例中,教师204可以使用监督或无监督学习算法在包括多个元组202的训练数据集上训练,以学习教师204的参数和学生206的参数。教师204可以被训练以用于分类任务,使得教师204的推断的(即预测的)输出是包括一组类别中每个类别的概率值的张量。例如,如果输入张量(x)包括包含手写数字的图像,则推断的(即预测的)输出(y)可以包括属于类别中每个类别的手写数字的概率评分,如数字“0”至“9”。在一些实施例中,教师204在训练学生206之前被训练。在一些其它实施例中,教师204和学生模型206被同时训练。

[0105] 在一些实施例中,教师204是单个神经网络模型。在一些其它实施例中,教师204是集成神经网络模型,其是已经单独地训练的多个单独神经网络模型的编译,其中单个神经网络模型的推断的(即预测的)输出被组合以生成教师204的输出。在一些其他实施例中,集成神经网络模型中的神经网络模型包括推断(即预测)针对一组类别中的每个类别的输出的分类模型,以及仅针对类别的相应子集生成评分的一个或多个专业模型。

[0106] 在所示的实施例中,推断的(即预测的)输出以及教师204和学生206的隐藏层的输出被提供给损失计算模块208。如图所示,损失计算模块208包括KD损失计算子模块210、CKD损失计算子模块211和标准损失计算子模块212。KD损失计算子模块210比较教师204和学生206的推断的(即预测的)输出,以计算KD损失值,如下文更详细描述的。CKD损失计算子模块211将教师204的隐藏层的子集映射到学生206的隐藏层,并确定由学生206的隐藏层生成的输出和由映射到学生206的隐藏层的教师204的隐藏层的子集生成的输出的表征之间的CKD损失。标准损失计算子模块212将学生206的推断的(即预测的)输出与元组202的真值进行比较,以计算标准损失值,如下文更详细地描述的。损失计算模块208还通过损失函数计算KD损失值、CKD损失值和标准损失值,以生成交叉熵值,该交叉熵值被反向传播到学生206,用于调整学生206的参数(即学生神经网络模型的参数)。在学生206被训练之后,其可以被部署到计算设备(未示出)上,并用于进行预测。在一些实施例中,部署学生206的计算设备是能够以更短的运行时间执行学生206的低容量、低复杂度计算设备。在一些实施例中,部署学生206的计算设备是移动通信设备,如智能手机、智能手表、笔记本电脑或平板电脑。

[0107] 如上所述,在PKD中,找到教师204的可跳过隐藏层是主要挑战之一。由于教师204的隐藏层和学生206的隐藏层之间不存在一一对应关系,因此在蒸馏过程中,如PKD的现有技术跳过了教师204的隐藏层中的一些隐藏层。因此,利用CKD的机器学习系统200可能能够融合或组合教师204的隐藏层,例如通过CKD损失计算子模块211,并受益于存储在教师204的所有隐藏层中的所有或大多数学习参数。

[0108] 在一些实施例中,根据本公开的CKD损失函数可以数学地表述为等式(4A):

[0109] $$\mathcal{L}_{CKD} = \sum_{h_i^S \in H^S} MSE(h_i^S, f_i^T) \qquad (4A)$$

[0110] 其中,f_i^T是通过$f_i^T = F(H^T(i))$计算的教师的第i个隐藏层,其中$H^T(i) = \{h_j^T | j \in M(i)\}$,F()是融合函数,$H^S$和$H^T$分别指示学生206和教师204的所有隐藏层的集合。参数$H^T(i)$是选择待映射到学生206的第i个隐藏层的教师204的隐藏层的子集。函

数MSE()是均方误差函数，并且h_i^S是学生206的第i个隐藏层。MSE只是CKD损失函数的许多可能实现之一，并且可以使用任何其它合适的方法，如有效的矩阵范数。

[0111] 在PKD中，f_i^T是教师204的第i个隐藏层，而相比之下，在根据本公开的CKD的一些实施例中，f_i^T是根据等式(4B)通过融合函数F()应用于教师204的隐藏层的选择子集的组合的结果：

[0112] $$f_i^T = F\left(H^T(i)\right);\ H^T(i) = \left\{h_j^T \middle| j \in M(i)\right\} \qquad (4B)$$

[0113] 在一些实施例中，教师204的所选子集$H^T(i)$经由映射函数M()定义，该映射函数采用引用学生206的隐藏层的索引，并为教师204返回一组索引。基于M()返回的索引，教师204的对应的隐藏层被组合以用于知识蒸馏过程。作为非限制性示例，针对索引为2，函数M(2)可以返回索引{1,3}。因此，融合函数F()由教师204的第一(h_1^T)隐藏层和第三(h_3^T)隐藏层提供，并且F()的输出被映射到学生206的第二(h_2^S)隐藏层。

[0114] 在一些实施例中，融合函数F()包括级联操作，后跟线性映射层。在上述其中索引=2且函数M(2)可以返回索引{1,3}的示例中，融合函数F()可以呈以下形式：

[0115] $$F(h_1^T, h_3^T) = mul(W, [h_1^T; h_3^T]) + b$$

[0116] 其中，“;”是级联算子，mu()是矩阵乘法运算，并且W和b是可学习参数。h_1^T、h_3^T、和h_2^S全部都是d维向量，其中d可以是任何实正整数，因为CKD能够经由mul()函数处理维度失配。

[0117] 映射函数M()定义了用于组合教师204的隐藏层的组合策略。图3A至图3D示出了可以通过映射函数M()实现的一些示例组合策略。特别地，图3A至图3D中的每个图示出了包括5个隐藏层的教师204和包括2个隐藏层的学生206之间的层组合策略。图3A示出了还被称为交叉组合(cross combination,CC)的第一组合策略，其中教师204的每第m隐藏层被组合用于蒸馏到学生206的m层中的对应的一个层。在所示的示例中，针对包括2个隐藏层的学生206，教师204的第一隐藏层(h_1^T)、第三隐藏层(h_3^T)和第五隐藏层(h_5^T)被组合用于蒸馏到学生206的第一隐藏层(h_1^S)，并且教师204的第二隐藏层(h_2^T)和第四隐藏层(h_4^T)被组合用于蒸馏到学生206的第二隐藏层(h_2^S)。图3B示出了还被称为常规组合(regular combination,RC)的第二组合策略，其中教师204的近似相等数量的连续隐藏层被组合用于蒸馏到学生206的对应的隐藏层。在所示的实施例中，教师204的前三个隐藏层(h_1^T, h_2^T, h_3^T))被组合用于蒸馏到学生206的第一隐藏层(h_1^S))，并且教师204的第四隐藏层和第五隐藏层(h_4^T, h_5^T)被组合用于蒸馏到学生206的第二隐藏层(h_2^S))。应当理解，针对n层教师204(即包括n个隐藏层的教师204)和m层学生206(即包括m个隐藏层的学生206)，其中n是m的倍数，教师204的偶数隐藏层可以针对学生206的每个隐藏层组合。可替代地，如果n不是m的精确倍数，则学生206的隐藏层的选择数量可以与教师204的更多组合隐藏层相关联以进行蒸馏。图3C示出了还被称为跳过组合(skip combination,SC)的第三组合策略，其中跳过教师204的一些隐藏层以进行蒸馏。在一些实施例中，可以跳过教师204的每第(m+1)个隐藏层。在一些其它实施例中，跳过教师204的一个或多个隐藏层，使得教师204的相等数量的隐藏层被组合以用于蒸馏到学生206的隐藏层之一。在一些其它实施例中，隐藏层可以以常规或非常

规的间隔跳过，并且然后可以细分和组合剩余的隐藏层以进行蒸馏。可以应用确定跳过间隔的任何其它方法。在所示的示例中，针对具有2个隐藏层的学生206，跳过教师204的第三隐藏层（h_3^T），使得教师204的第一隐藏层（h_1^T）和第二隐藏层（h_2^T）被组合用于蒸馏到学生206的第一隐藏层（h_1^S），并且教师204的第四隐藏层（h_4^T）和第五隐藏层（h_5^T）被组合用于蒸馏到学生206的第二隐藏层（h_2^S）。图3D示出了还被称为重叠组合(overlap combination，OC)的第四组合策略，其中教师204的一个或多个隐藏层被组合成用于蒸馏到学生206的多个隐藏层的多组隐藏层。在所示的实施例中，教师204的第三隐藏层（h_3^T）被组合用于蒸馏学生206的第一隐藏层和第二隐藏层（h_1^S, h_2^S）两者。除了本文描述的四种组合策略之外，还可以应用任何其它合适的组合策略。CKD中的组合策略的类型可以在从教师204的隐藏层的不同配置中蒸馏方面提供灵活性。组合策略可以手动确定(即，不具有注意力)，或由机器学习系统200自动确定(即，具有注意力)。

[0118] 图4示出了根据本公开的不具有注意力的示例知识蒸馏方法400的流程图。

[0119] 在步骤402处，在初始化损失值之后，如通过将它们设置为0(例如$\mathcal{L}_{CKD} \leftarrow 0$)，预定义的层映射函数M()用于确定教师204的哪些隐藏层待被组合并与学生206的哪些隐藏层相关联。在此，组合层意味着组合由这些隐藏层生成的输出。

[0120] 在步骤404处，针对学生206的每一层h_i^S，通过级联由待组合的教师204的隐藏层生成的输出并通过mul(W,Concat(H^T))+b应用线性映射来计算教师204的组合隐藏层的表征f_1^T。

[0121] 在步骤406处，学生204和教师206均被提供有相同的输入元组202，并且部分地基于基于元组202的输入和教师204的隐藏层的输出的教师204的推断的(即预测的)输出来计算最终损失值。

[0122] 图5示出了在步骤406处确定最终损失值的示例方法500的流程图。

[0123] 在步骤502处，例如根据等式(4)，针对教师204的隐藏层中的每一层计算均方误差(MSE)（h_i^S, f_i^T）损失(也被称为$\mathcal{L}_{CKD}$)。应当理解，基于MSE损失的L_{CKD}是该损失函数的许多可能实现之一，并且可以使用任何有效的矩阵范数来代替MSE。

[0124] 在步骤504处，针对学生206和教师204两者，例如根据等式(2)计算KD损失（$\mathcal{L}_{KD}$）。

[0125] 在步骤506处，针对学生，例如根据等式(1)计算负对数似然损失（$\mathcal{L}_{nll}$）。

[0126] 在步骤508处，根据等式(5)，将最终损失计算为在步骤502处、504处和506处确定的损失的加权值：

[0127] $$\mathcal{L} = \alpha \mathcal{L}_{nll} + \beta \mathcal{L}_{KD} + \eta \mathcal{L}_{CKD} \qquad (5)$$

[0128] 其中α,β,和η是示出每个损失贡献的系数。

[0129] 返回参考图4，在步骤408处，相对于损失值的梯度被计算并被用于更新学生206的对推断的(即预测的)输出y有贡献的所有参数。

[0130] 在步骤410处，所有损失值再次被初始化回其相应的初始值，如零，并重复步骤404至410以进行进一步迭代，直到满足完成条件，如损失值降到可接受的阈值以下。图6示出了方法400的部分的示例伪代码。

[0131] 在一些情况下，手动定义的组合策略可能不呈现针对教师204的不同隐藏层定义

功能M(i)的最优组合方法。因此，本公开的一些其他实施例提供了具有注意力的增强CKD方法，其可以自动定义用于组合教师204的隐藏层的最优策略。在至少一个方面，基于注意力的KD解决了搜索待组合的教师204的隐藏层的代价高昂的问题。具体地，PKD和其它类似的方法可能需要训练几个学生206以搜索在训练期间应跳过的教师204的隐藏层，并且在教师204是深度神经网络模型的情况下，(如果有的话)寻找最佳/最优解决方案可能是耗时的。

[0132] 在一些实施例中，可以在学生206的相应的隐藏层中的每个隐藏层(H^S)和教师204的隐藏层(H^T)之间学习注意力权重。每个注意力权重可以是有多少教师204的隐藏层对学生206的给定隐藏层的知识蒸馏过程有贡献的指示。然后，机器学习系统200可以优化注意力权重，以试图实现教师204的隐藏层和学生206的隐藏层之间的最优知识蒸馏。

[0133] 图7示出了具有注意力的增强CKD的高级示意性架构，或本文可以被标注为CKD*。如图7所示出的，组合策略可能不需要经由M()手动定义用于CKD*中的知识蒸馏，而是CKD*将教师204的所有隐藏层(h_1^T至h_5^T)考虑在内，并将注意力权重(ϵ_{11}、ϵ_{12}、ϵ_{13}、ϵ_{14}、和ϵ_{15})分配到教师204的隐藏层(h_1^T至h_5^T))中的每一层。注意力权重指示由教师204的隐藏层生成的特定输出的贡献量，该输出在蒸馏期间待用于学生206的给定隐藏层。教师204的每个隐藏层的输出张量可以应用其对应的注意力权重，以计算教师204的所有隐藏层的输出的加权平均。加权平均是可以连接到学生206的隐藏层以用于在其之间进行知识蒸馏的张量。总损失根据等式(6)表示：

[0134] $$\mathcal{L}_{CKD^*}(H^S, H^T) = \sum_{h_i^S \in H^S} MSE\,(h_i^S, f_i^{*T}) \qquad (6)$$

[0135] 其中MSE损失是损失函数$\mathcal{L}_{CKD^*}()$的许多可能实现之一，并且也可以使用任何其它合适的损失函数，如KL散度。f_i^{*T}是针对教师204的第i个隐藏层的教师204的隐藏层(H^T)的基于注意力的组合表征，并且，针对学生206和教师204具有相同维度($|h_i^S| = |h_j^T|$)的实施例，f_i^{*T}可以根据等式7A确定：

[0136] $$f_i^{*T} = \sum_{h_j^T \in H^T} \epsilon_{ij} h_j^T \qquad (7A)$$

[0137] 可替代地，针对学生206和教师204具有不同维度($|h_i^S| \neq |h_j^T|$)的实施例，f_i^{*T}可以根据等式7B确定：

[0138] $$f_i^{*T} = \sum_{h_j^T \in H^T} \epsilon_{ij} (W_i h_j^T) \qquad (7B)$$

[0139] 其中，$W_i \in \mathbb{R}^{|h_i^S| \times |h_j^T|}$是教师204的第i个隐藏层的权重值，指示教师204的不同隐藏层(h_j^T)在知识传输过程中对学生206的每个特定隐藏层(h_i^S)的贡献量，以及因此相对重要性。注意力权重($\in_{ij}$)总和应为1，并可以根据等式(8A)计算：

[0140] $$\epsilon_{ij} = \frac{e^{\varphi_{ij}}}{\sum_{h_k^T \in H^T} e^{\varphi_{ik}}} \qquad (8A)$$

[0141] 能量函数$\Phi(h_i^S, h_j^T)$是根据等式(8B)的函数：

[0142] $\varphi_{ij} = \Phi\left(h_i^s, h_j^T\right)$ (8B)

[0143] 其中，φ_{ij}确定由学生206的第i个隐藏层h_i^s生成的输出和由教师204的第j个隐藏层h_j^T生成的输出之间的能量评分，其中能量评分指示两个生成的输出之间的相似性。

[0144] 在学生206和教师204具有相同维度($|h_i^s| = |h_j^T|$)的一些实施例中，能量函数$\Phi(h_i^s, h_j^T)$是学生206的第i个隐藏层h_i^s的输出和由教师204的第j个隐藏层h_j^T生成的输出的点积，根据等式(9A)：

[0145] $\Phi\left(h_i^s, h_j^T\right) \equiv < h_i^s, h_j^T >$ (9A)

[0146] 可替代地，在其中学生和教师之间存在维度失配($|h_i^s| \neq |h_j^T|$)的实施例中，能量函数Φ(,)可以根据等式(9B)计算为学生206的第i个隐藏层h_i^s的输出和由教师204的第j个隐藏层h_j^T生成的输出的加权值的点积；

[0147] $\Phi\left(h_i^s, h_j^T\right) \equiv < h_i^s, W_i h_j^T >$ (9B)

[0148] 基于点积的能量函数可以允许待通过附加映射层处理的两个隐藏层的输出之间的任何潜在维度失配，如下文更详细地描述的。

[0149] 图8示出了根据本公开的具有注意力的示例知识蒸馏方法800的流程图。

[0150] 在步骤802处，初始化损失值，如通过将其设置为0(例如$\mathcal{L}_{CKD} \leftarrow 0$)。与方法400不同，层映射函数M()不需要被明确地定义。相反，针对学生206的每个隐藏层，权重值被分配给教师204的隐藏层的子集。在一些实施例中，教师204的隐藏层的子集可以包括教师204的所有隐藏层。从概念上讲，权重值可以用作教师204的隐藏层到学生206的隐藏层中的每一层的隐式映射。

[0151] 在步骤804处，针对每个学生层h_l^s，通过根据等式(6)将每个教师隐藏层输出乘以其对应的权重值，计算教师隐藏层f_1^T子集的加权平均作为教师隐藏层的表征。

[0152] 在步骤806处，学生206和教师204均被提供有相同元组202的输入向量(x)，并且例如根据方法500，部分地基于教师和学生206的基于元组202的输入向量(x)的推断的(即预测的)输出计算最终CKD^*损失值。

[0153] 在步骤808处，计算相对于损失值的梯度，并且计算的梯度用于更新学生206的所有参数。

[0154] 在步骤810处，所有损失值再次初始化回其相应的初始值，如零，并重复步骤804至810进行进一步迭代，直到满足完成条件，如损失值降到可接受的阈值以下。

[0155] 应当理解，尽管本文描述了利用教师204的所有隐藏层的加权平均将知识转移到学生206的每个隐藏层的CKD^*的实施例，但CKD^*的其它实施例可以利用教师204的部分隐藏层的加权平均。另外，可以存在重叠的教师的隐藏层的子集，这些隐藏层被组合用于学生206的不同隐藏层。例如，教师204的隐藏层中的一半可以用于在没有重叠的情况下组合学生206的隐藏层中的一半。可替代地，教师204的隐藏层中的三分之二(即教师204的隐藏层中的第一1/3隐藏层和第二1/3隐藏层)可以用于组合学生的隐藏层中的一半，并且教师204的隐藏层中的另外三分之二(第二1/3隐藏层和第三1/3隐藏层)可以用于在部分重叠的情况下组合学生206的隐藏层中的另一半。

[0156] 有利地，上述CKD*方法可以实现自动教师隐藏层映射选择，其可以实现最优知识转移。在一些实施例中，CKD*方法可以在BERT和神经机器翻译模型的性能方面进行改进。图9示出了表格，示出在执行本领域已知的通用语言理解评估(GLUE)基准时各种KD模型的模拟结果，包括具有105,000个数据点的问题自然语言推断(question natural language inference，QNLI)、具有3,700个数据点的微软研究释义语料库(microsoft research paraphrase corpus，MRPC)和具有2,500个数据点的识别文本蕴涵(recognizing textual entailment，RTE)。在图9所示出的表格中，“T”表示教师，其是基于BERT的模型，具有12个隐藏层，12个注意力头，并且隐藏尺寸为768。所有学生206都是BERT_4模型，具有4个隐藏层，12个注意力头，并且隐藏尺寸为768。在学生206中，“NKD”表示无KD，“KD”表示常规的现有技术KD。表格的最后三列分别是应用了具有无注意力重叠(即T[1,2,3,4]->S1、T[5,6,7,8]->S2、T[9,10,11,12]->S3)、部分注意力重叠(即T[1,2,3,4,5]->S1、T[5,6,7,8,9]->S2、T[9,10,11,12]->S3)，和完全注意力重叠(即，教师204的所有12个隐藏层用于梳理学生206的每个隐藏层)的CKD*的学生。如图所示，在QNLI基准中，具有无注意力重叠的CKD*实现了评分为87.11，优于其余学生，在RTE基准中，具有完全注意力重叠的CKD*实现了评分为67.15，优于所有其他学生206。在MRPC基准中，具有完全注意力重叠的CKD*评分为80.72，优于所有其他学生。

[0157] 在一些其他实施例中，本文描述的基于注意力的CKD*方法也可以应用于多教师/多任务KD场景。具体地，CKD*方法可以被应用来组合来自不同教师204的不同隐藏层，以将知识蒸馏到一个学生206中，而不是通过注意力来组合单个教师204的不同隐藏层。至少一些现有技术KD方法在不同的训练迭代中迭代多个教师204，并独立地考虑它们，如Clark，K.、Luong，M.T.、Khandelwal，U.、Manning，C.D.和Le，Q.V.的“瓶颈注意力模块!用于自然语言理解的再生多任务网络(Bam!born-again multi-task networks for natural language understanding)(2019)”中公开的。例如，用K个不同的训练数据集 $\mathcal{D}_q = \{(x_i^q, y_i^q)\}_{i=1}^{N_q}, 1 \le q \le K$ 训练K个不同的教师204，其中N_q指定第q教师204的训练数据集中的训练数据样本的数量。上述用于训练具有多个教师204的学生206的现有技术方案，并且特别是第q教师T_q和学生S之间的KD损失($\mathcal{L}_{KD}^q$)可以在数学上表征如下：

[0158] $$\mathcal{L}_{KD} = \sum_{q=1}^{K} \mathcal{L}_{KD}^q = \sum_{q=1}^{K} \sum_{x_i^q \in \mathcal{D}_q} L_{KD}\left(T_q(x_i^q), S(x_i^q)\right)$$

[0159] 其中，项$T_q(x_i^q)$和$S(x_i^q)$分别给出第q教师T_q和学生S的推断的(即预测的)输出。每个教师的KD损失值T_q($\mathcal{L}_{KD}^q$)可以确定为：

[0160] $$\mathcal{L}_{KD}^q = \sum_{x_i^q \in \mathcal{D}_q} L_{KD}(T_q(x_i^q), S(x_i^q))$$

[0161] 其中L_{KD}可以是任何损失函数，如库尔贝克·莱布勒(Kullback-Leibler，KL)散度或均方误差。为了计算$L_{KD}(T_q(x_i^q), S(x_i^q))$，训练数据集的数据样本$\{(x_i^q)\}_{i=1}^{N_q}$被发送到第q教师$T_q$和学生S，以获得其相应的推断的(即预测的)输出。

[0162] 对于根据本公开的实施例，上述多教师根据等式(10)通过CKD/CKD*方法扩展：

[0163] $$\mathcal{L}_{KD}^{*q}=\sum_{p=1}^{p=K}\sum_{x_i^q\in\mathcal{D}_q}\epsilon_{pq}L_{KD}(T_p(x_i^q),S(x_i^q)) \qquad (10)$$

[0164] 其中，每个教师T_p的隐藏层的权重值$\in_{pq}$可以根据等式(11A)确定：

[0165] $$\epsilon_{pq}=\frac{e^{\varphi_{pq}}}{\sum_{i=1}^{K}e^{\varphi_{iq}}} \qquad (11A)$$

[0166] 在一些实施例中，Φ(.)函数是两个输入向量(x)用于测量两个输入向量(x)之间的能量或相似性的点积，如可以根据等式11(B)计算的：

[0167] $$\varphi_{pq}=\Phi\left(T_p(x_i^q),S(x_i^q)\right);\ x_i^q\in\mathcal{D}_q \qquad (11B)$$

[0168] 在一些其他实施例中，该Φ(.)函数是神经网络或用于测量两个输入向量(x)的能量的任何其它合适的函数。

[0169] 图10示出了根据本公开的用于从多个教师204到一个学生206的知识蒸馏的示例方法1000的流程图。

[0170] 在步骤1002处，初始化每个教师204的参数，包括每个教师204的隐藏层的注意力权重。

[0171] 在步骤1004处，将训练数据集提供给每个教师204，如$\mathcal{D}_q=\{(x_i^q,y_i^q)\}_{i=1}^{N_q},1\le q\le K$，根据等式(12)计算基于针对每个教师204计算的损失的加权KD损失：

[0172] $$\mathcal{L}_{KD}^{*q}=\sum_{p=1}^{p=K}\sum_{x_i^q\in\mathcal{D}_q}\epsilon_{pq}\,L_{KD}(T_p(x_i^q),S(x_i^q)) \qquad (12)$$

[0173] 其中L_{KD}可以是任何合适的损失函数，包括KL散度或均方误差(mean squared error，MSE)。为了计算$L_{KD}(T_q(x_i^q),S(x_i^q))$，将训练数据集$\{(x_i^q)\}_{i=1}^{N_q}$的数据样本作为输入提供给第q教师204和学生206，以获得第q教师204和学生206的推断的(即预测的)输出。

[0174] 在步骤1006处，针对每个训练数据集q，计算所有K名教师204的注意力权重(即，$\{\in_{1q},\in_{2q},\cdots,\in_{Kq}\}$的集合，其中$\sum_{i=1}^{K}\epsilon_{1q}=1$)。图11示出了可以在步骤1006处实现的用于计算注意力权重的示例方法1100的流程图。

[0175] 在步骤1102处，将训练数据集的x_i^q作为输入提供给所有教师204和学生206，以获得其相应的推断的(即，预测的)输出。

[0176] 在步骤1104处，使用函数$\varphi_{pq}=\Phi\left(T_p(x_i^q),S(x_i^q)\right)$计算学生206的输出向量$S(x_i^q)$和每个教师204的输出向量$T_p(x_i^q)$之间的能量/相似性。在一些实施例中，函数Φ可以是其两个向量输入之间的简单点积函数。

[0177] 在步骤1106处，基于与教师204的推断的(即，预测的)输出的相似性，例如使用$\epsilon_{pq}=\frac{e^{\varphi_{pq}}}{\sum_{i=1}^{K}e^{\varphi_{iq}}}$，来计算应分配给每个教师204的权重$\in_{pq}$。

[0178] 返回参考图10，在步骤1008处，针对所有K名教师204的权重KD损失$\mathcal{L}_{KD}^{q}$被总计为教

师204在每个时间步骤处的总KD损失。应当理解，尽管本文描述的实施例包括$\mathcal{L}_{KD}$损失，但在一些其它实施例中，取决于每个问题的设计，训练除$\mathcal{L}_{KD}$的损失函数也是适用的。

[0179] 在步骤1010处，计算相对于教师204的参数的总KD损失梯度，并更新学生206的参数。

[0180] 在一些实施例中，方法1000将如本文描述的CKD*扩展到多个教师知识蒸馏场景。这可以允许多任务蒸馏、多语言蒸馏、多检查点蒸馏和任何其它期望从多个教师蒸馏到单个学生的应用。

[0181] 参考图12，可以用于实现本文公开的实施例的示例简化处理系统1200的框图，并提供了更高级别的实现示例。教师204和学生206以及包括在机器学习系统200中的其它功能可以使用示例处理系统1200或处理系统1200的变体来实现。处理系统1200可以是终端，例如桌面终端、平板电脑、笔记本电脑、AR/VR或车载终端，或可以是服务器、云端或任何合适的处理系统。可以使用适合于实现本公开中描述的实施例的其它处理系统，这些系统可以包括与下面讨论的那些组件不同的组件。虽然图12示出了每个组件的单个实例，但是在处理系统1200中可能存在每个组件的多个实例。

[0182] 处理系统1200可以包括一个或多个处理设备1202，如图形处理单元、处理器、微处理器、专用集成电路(application-specific integrated circuit，ASIC)、现场可编程门阵列(field-programmable gate array，FPGA)、专用逻辑电路、加速器、张量处理单元(tensor processing unit，TPU)、神经处理单元(neural processing unit，NPU)，或它们的组合。处理系统1200还可以包括一个或多个输入/输出(input/output，I/O)接口1204，其可以实现与一个或多个适当的输入设备1214和/或输出设备1216介接。处理系统1200可以包括一个或多个网络接口1206用于与网络进行有线通信或无线通信。

[0183] 处理系统1200还可以包括一个或多个存储单元1208，该一个或多个存储单元可以包括如固态驱动器、硬盘驱动器、磁盘驱动器和/或光盘驱动器等大容量存储单元。处理系统1200可以包括一个或多个存储器1210，该一个或多个存储器可以包括易失性或非易失性存储器(例如，闪存、随机存取存储器(random access memory，RAM)和/或只读存储器(read-only memory，ROM))。存储器1210的非瞬时性存储器可以存储用于由处理设备1202执行的指令，如用于执行本公开中描述的示例，例如用于机器学习系统200的CKD/CKD*指令和数据1212。存储器1210可以包括其它软件指令，如用于实现处理系统1200和其它应用/功能的操作系统。在一些示例中，一个或多个数据集和/或模块可以由外部存储器(例如，与处理系统1200进行有线通信或无线通信的外部驱动器)提供，或可以由瞬时性或非瞬时性计算机可读介质提供。非瞬时性计算机可读介质的示例包括RAM、ROM、可擦除可编程ROM(EPROM)、电可擦除可编程ROM(electrically erasable programmable ROM，EEPROM)、闪存、CD-ROM或其它便携式存储装置。

[0184] 处理系统1200还可以包括总线1218，该总线提供处理系统1200的组件之间的通信，包括处理设备1202、I/O接口1204、网络接口1206、存储单元1208和/或存储器1210。总线1218可以是任何合适的总线架构，例如包括存储器总线、外围总线或视频总线。

[0185] 教师204和学生206的计算可以由处理系统1200的任何合适的处理设备1202或其变体执行。此外，教师204和学生206可以是任何合适的神经网络模型，包括如递归神经网络模型、长短期记忆(long short-term memory，LSTM)神经网络模型等变体。

[0186] 综述

[0187] 尽管本公开可以描述具有一定顺序的步骤的方法和过程,但是可以适当地省略或改变方法和过程中的一个或多个步骤。在适当情况下,一个或多个步骤可以按所描述的顺序以外的顺序执行。

[0188] 尽管本公开可以在方法方面至少部分地进行描述,但本领域普通技术人员将理解,本公开还针对用于执行所描述方法的至少一些方面和特征的各种组件,无论是通过硬件组件、软件或这两者的任何组合。相应地,本公开的技术方案可以以软件产品的形式体现。合适的软件产品可以存储在预先记录的存储设备或其它类似的非易失性或非瞬时性计算机可读介质中,例如包括DVD、CD-ROM、USB闪存盘、可移动硬盘或其它存储介质。软件产品包括有形地存储在其上的指令,这些指令使得处理设备(例如,个人计算机、服务器或网络设备)能够执行本文中公开的方法的示例。

[0189] 在不脱离权利要求书的主题的前提下,本公开可以通过其它特定形式来体现。所描述的示例实施例在所有方面都应被视为仅是示意性的,而非限制性的。可以组合从一个或多个上述实施例中选择的特征,以创建未明确描述的可选实施例,在本公开的范围内可以理解适合于此类组合的特征。

[0190] 还公开了所公开范围内的所有值和子范围。此外,尽管本文所公开和示出的系统、设备和过程可以包括特定数量的元件/组件,但可以修改这些系统、设备和组件,以包括更多或更少此类元件/组件。例如,尽管所公开的任何元件/组件可以引用为单数,但是可以修改本文所公开的实施例以包括多个此类元件/组件。本文描述的主题旨在覆盖和涵盖所有适当的技术变更。

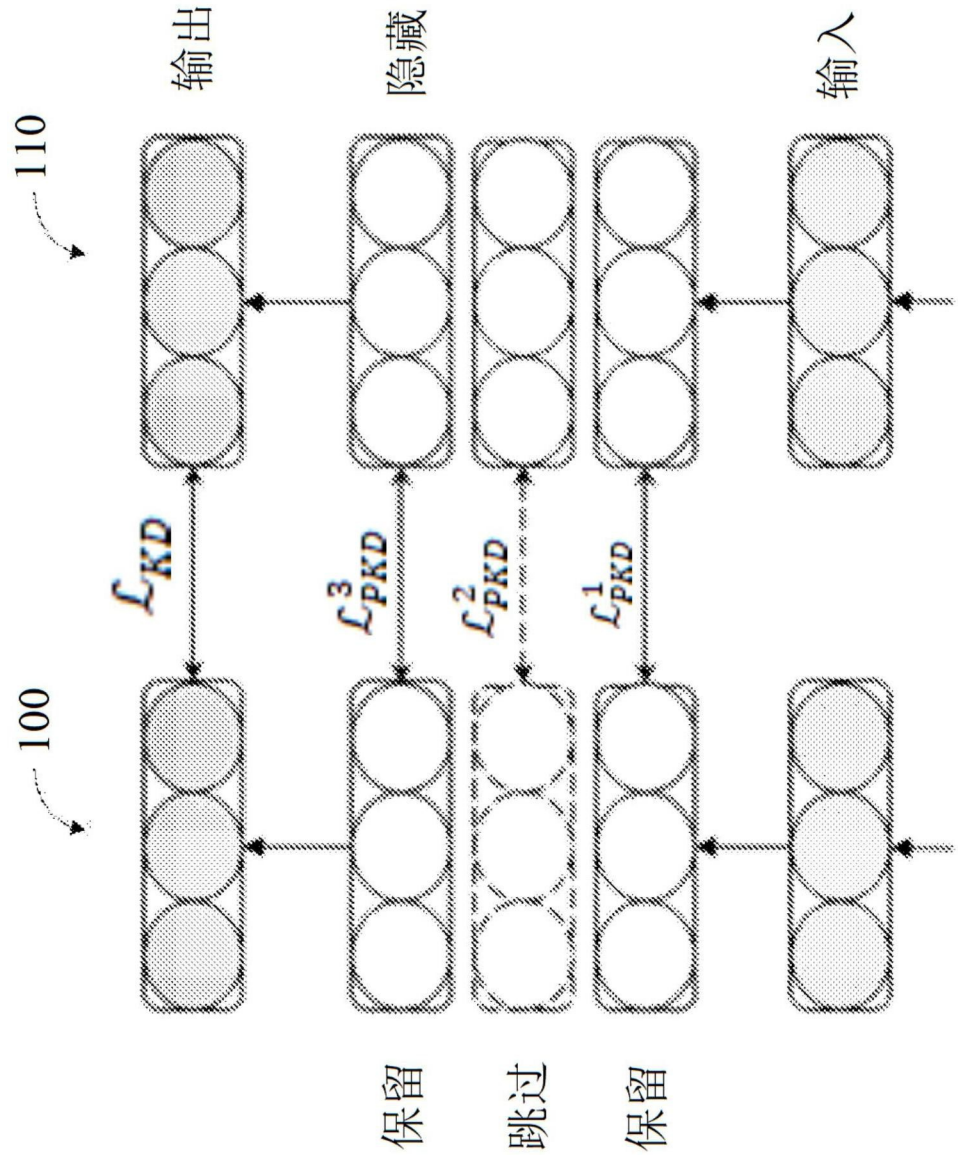

输出
隐藏
输入
110
100
$\mathcal{L}_{KD}$
$\mathcal{L}_{PKD}^{3}$
$\mathcal{L}_{PKD}^{2}$
$\mathcal{L}_{PKD}^{1}$
保留
跳过
保留

图1现有技术

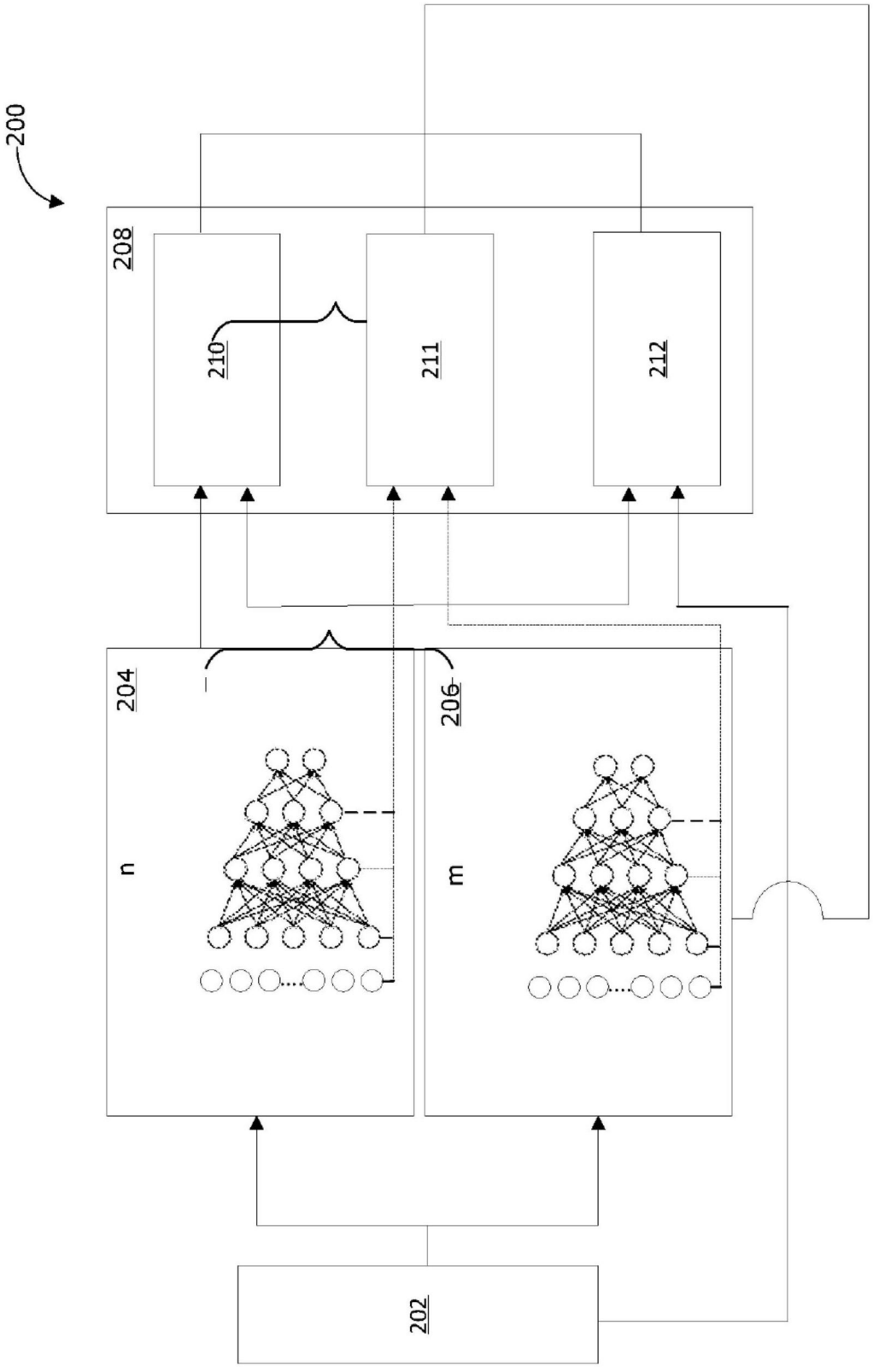


图2

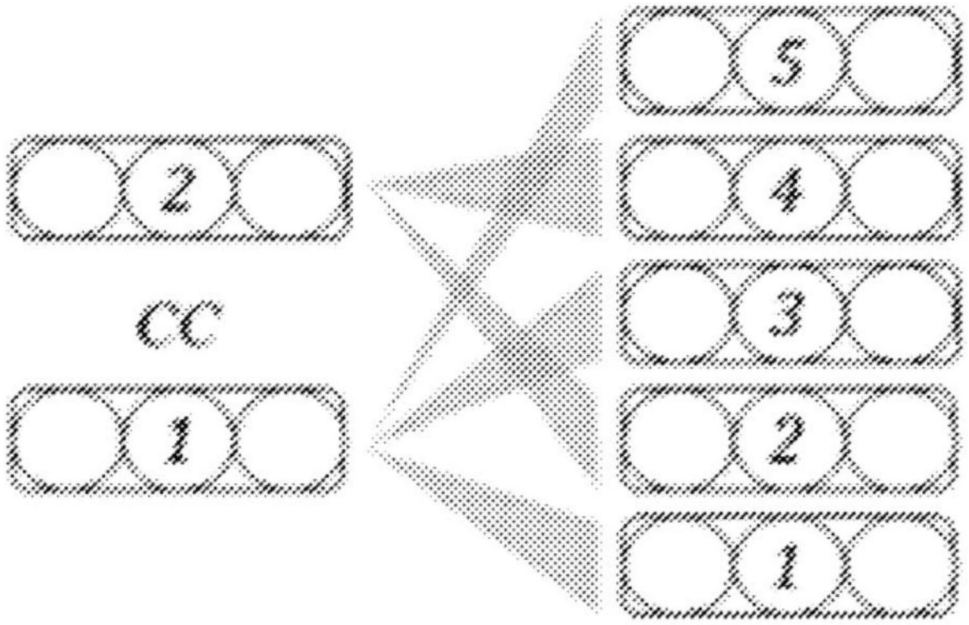


图3A

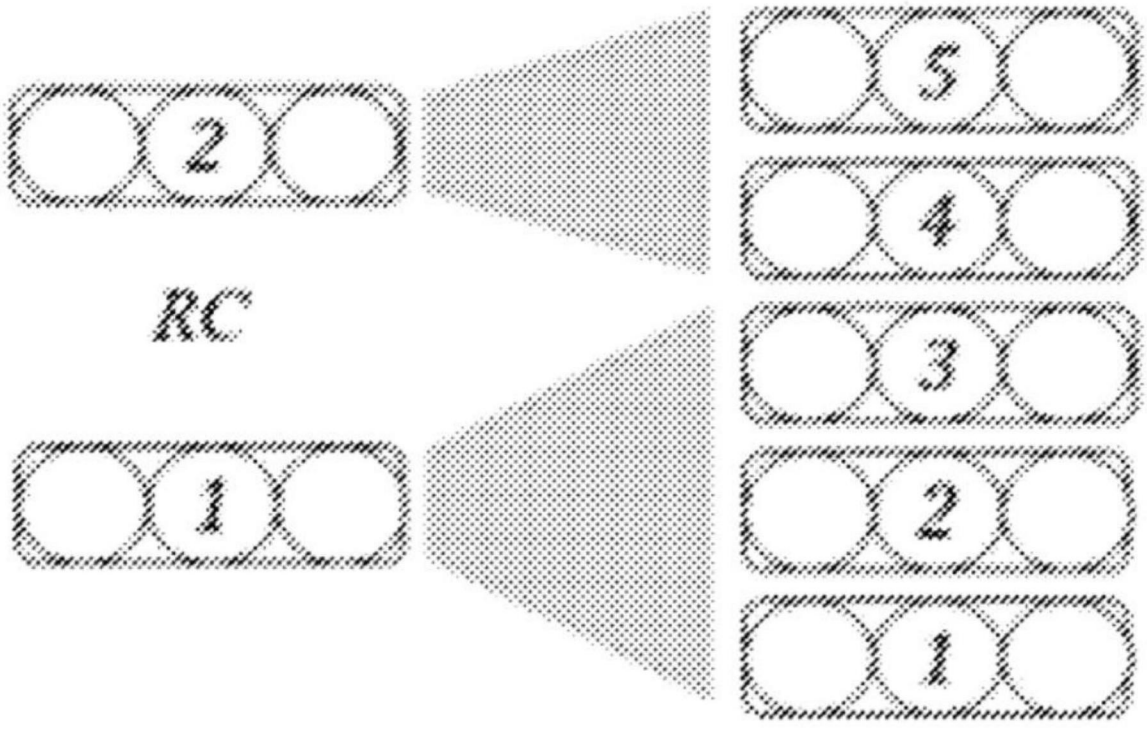


图3B

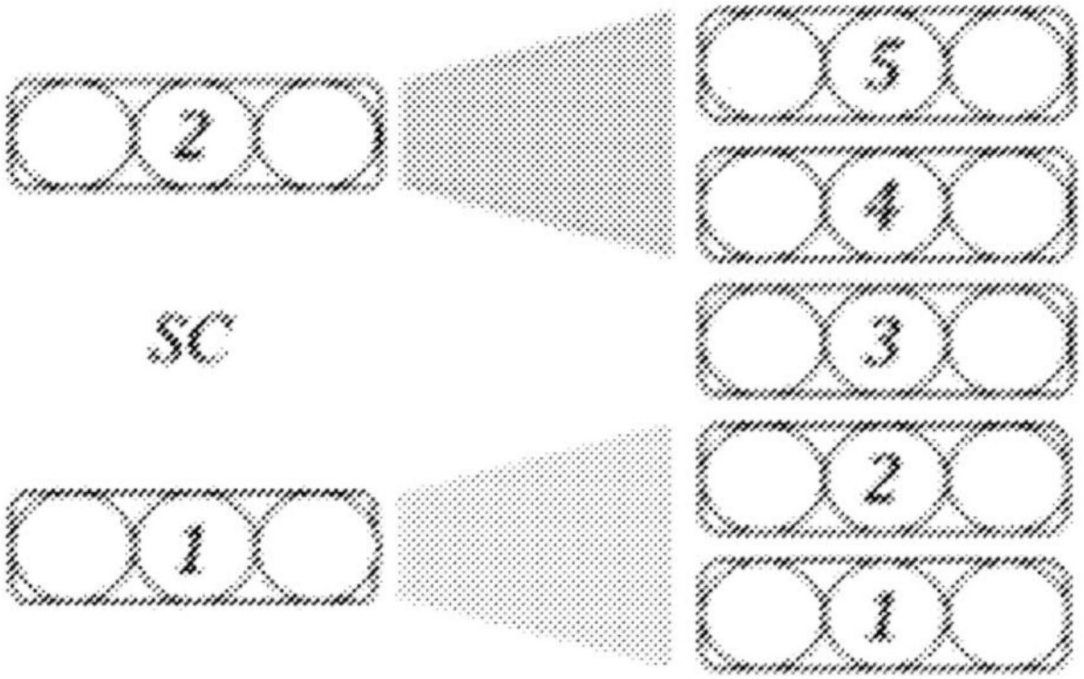


图3C

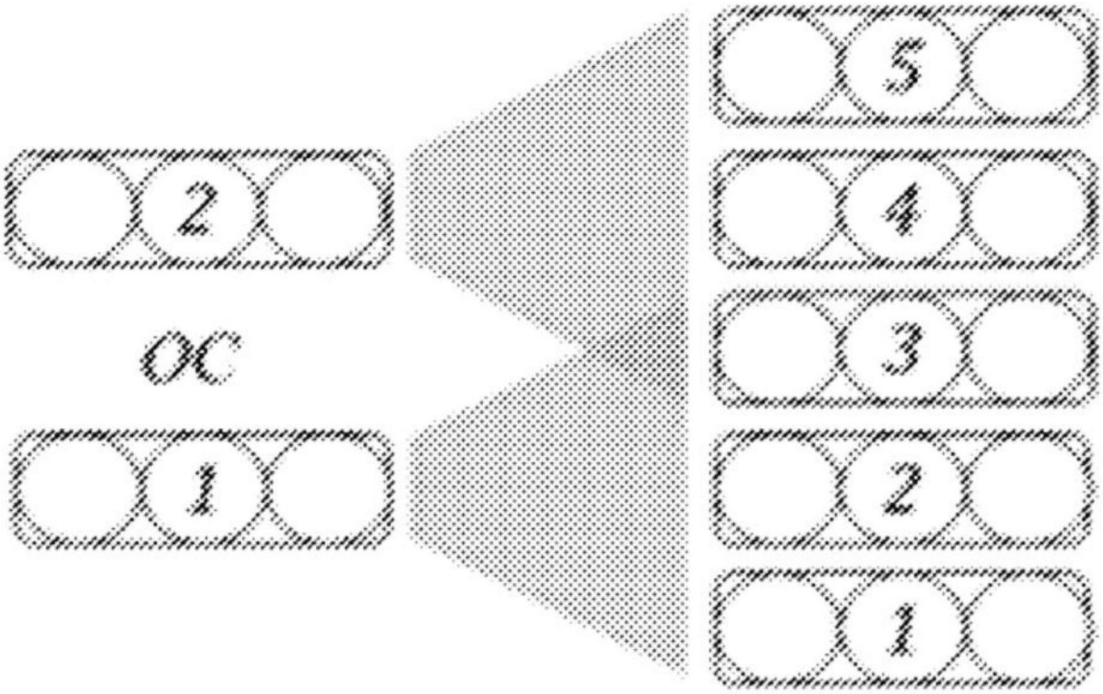


图3D

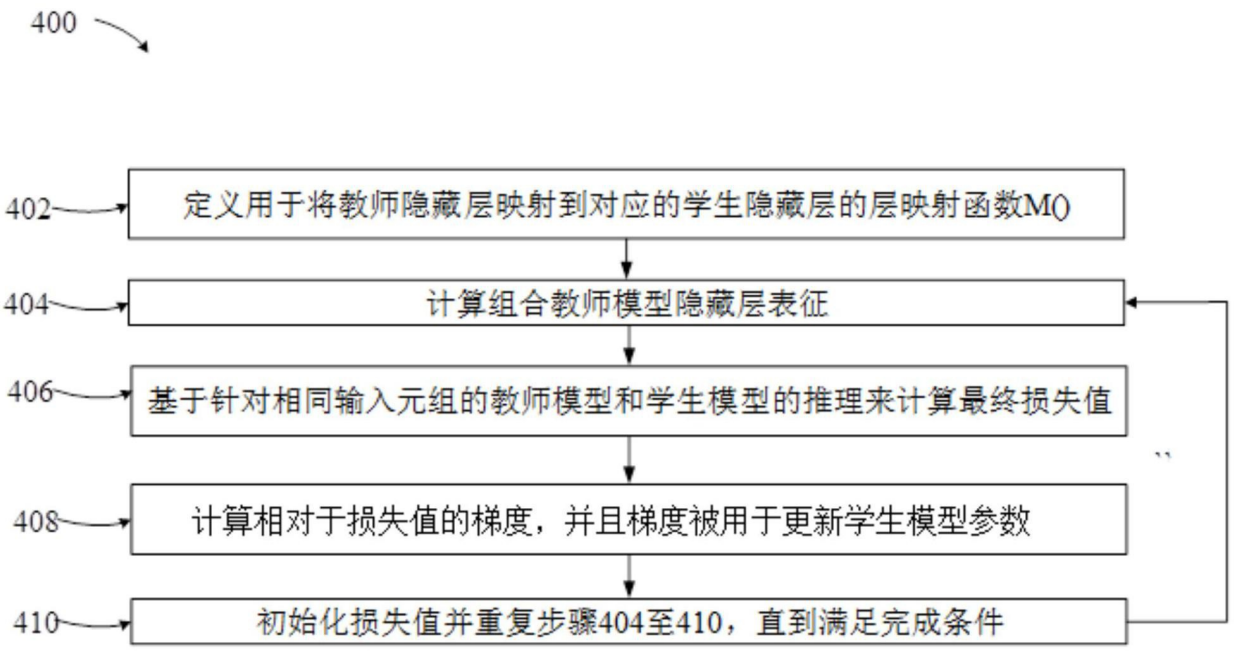

400
402
定义用于将教师隐藏层映射到对应的学生隐藏层的层映射函数M()
404
计算组合教师模型隐藏层表征
406
基于针对相同输入元组的教师模型和学生模型的推理来计算最终损失值
408
计算相对于损失值的梯度，并且梯度被用于更新学生模型参数
410
初始化损失值并重复步骤404至410，直到满足完成条件

图4

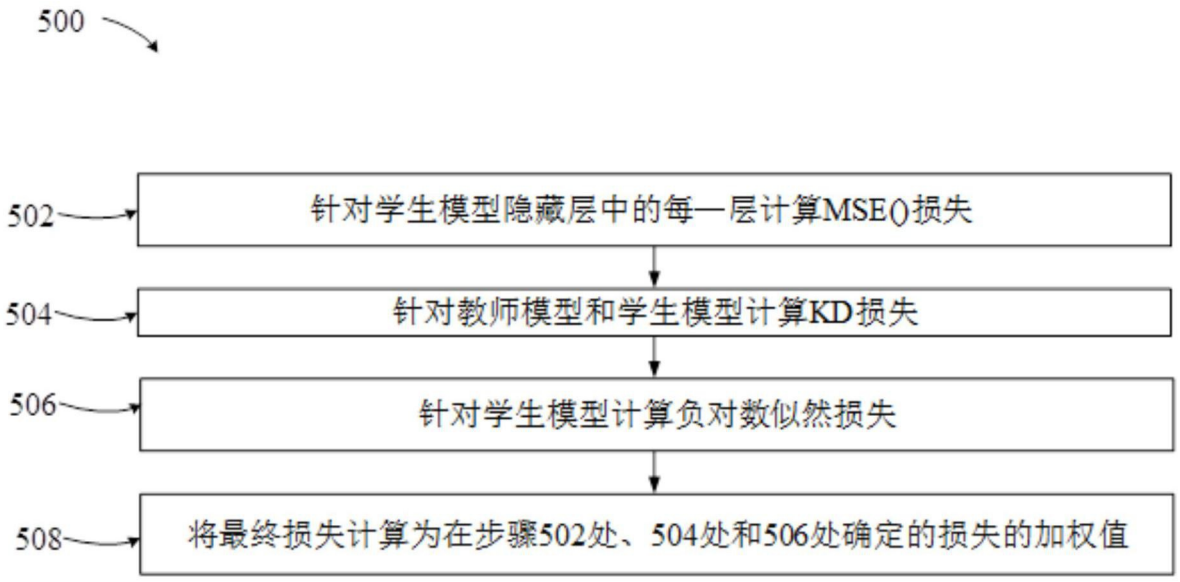

500
502
针对学生模型隐藏层中的每一层计算MSE()损失
504
针对教师模型和学生模型计算KD损失
506
针对学生模型计算负对数似然损失
508
将最终损失计算为在步骤502处、504处和506处确定的损失的加权值

图5

算法1：组合知识蒸馏（CKD/CKD*）

Function $Fusion(h_i^S, H^T, kd)$:
- ▷ h_i^S: i-th layer of S,
- ▷ H^T: a subset of T's layers which are selected to map to h_i^S
- ▷ kd: Kd type (CKD or CKD*)

if $kd == CKD$ **then**
- $f_i^T \leftarrow mul(W, Concat(H^T)) + b_i$

else
- $\varphi_{ij} = \Phi(h_i^S, h_j^T)$;
- $\alpha_j \leftarrow \frac{e^{\varphi_{ij}}}{\sum_{h_j^T \in H^T} e^{\varphi_{ij}}}$;
- $f_i^T \leftarrow \sum_{h_j^T \in H^T} \alpha_j h_j^T$;

end

return f_i^T;

End Function

Function $ProposedKD(S, T, X, M, e, kd)$:
- ▷ S: student network
- ▷ T: teacher network
- ▷ (X,Y): set of input-output training pairs
- ▷ M: mapping function
- ▷ e: training epochs
- ▷ kd: kd type (CKD or CKD*)

for $i = 1$ *to* e **do**
- $\mathcal{L} \leftarrow 0, \mathcal{L}_{ce} \leftarrow 0, \mathcal{L}_{KD} \leftarrow 0, \mathcal{L}_{CKD} \leftarrow 0$;
- **for** (x, y) *in* (X, Y) **do**
  - $\ell_{CKD} \leftarrow 0$;
  - **for** l *in* H^S **do**
    - $f_l^T \leftarrow Fusion(h_l^S, M(H^T), kd)$;
    - $\ell_{CKD} \leftarrow \ell_{CKD} + MSE(h_l^S, f_l^T)$;
  - **end**
  - $\mathcal{L}_{CKD} \leftarrow \mathcal{L}_{CKD} + \ell_{CKD}$;
  - $\mathcal{L}_{KD} \leftarrow \mathcal{L}_{KD} + \mathcal{L}_{KD}(\theta_S, \theta_T)$;
  - $\mathcal{L}_{ce} \leftarrow \mathcal{L}_{ce} + \mathcal{L}_{ce}(\theta_S)$;
  - $\mathcal{L} \leftarrow \alpha\mathcal{L}_{ce} + \beta\mathcal{L}_{KD} + \eta\mathcal{L}_{CKD}$;
- **end**

end

return $\mathcal{L}$;

End Function

图6

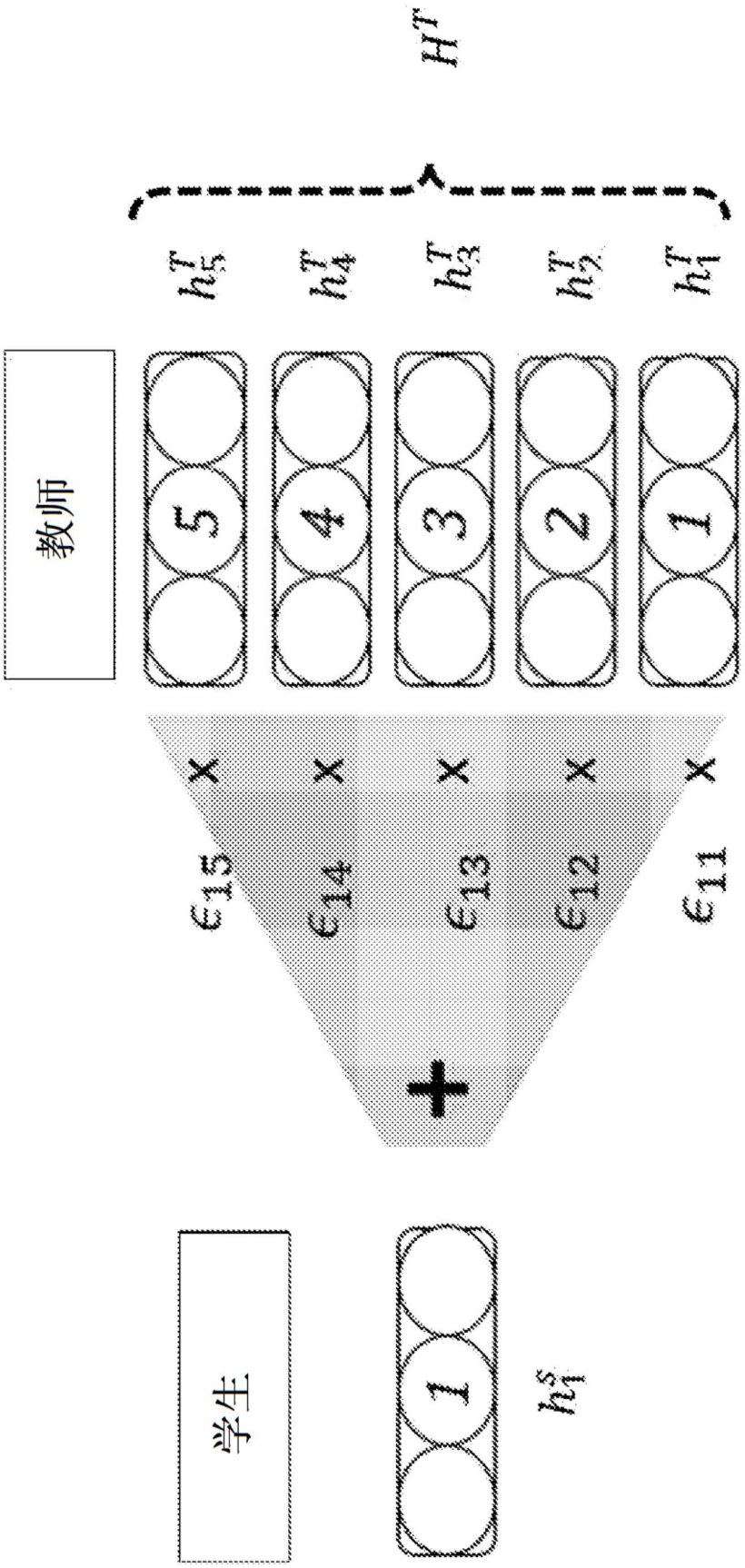

H^T
h_5^T
h_4^T
h_3^T
h_2^T
h_1^T
教师
5
4
3
2
1
×
×
×
×
×
ϵ_{15}
ϵ_{14}
ϵ_{13}
ϵ_{12}
ϵ_{11}
+
学生
1
h_1^S

图7

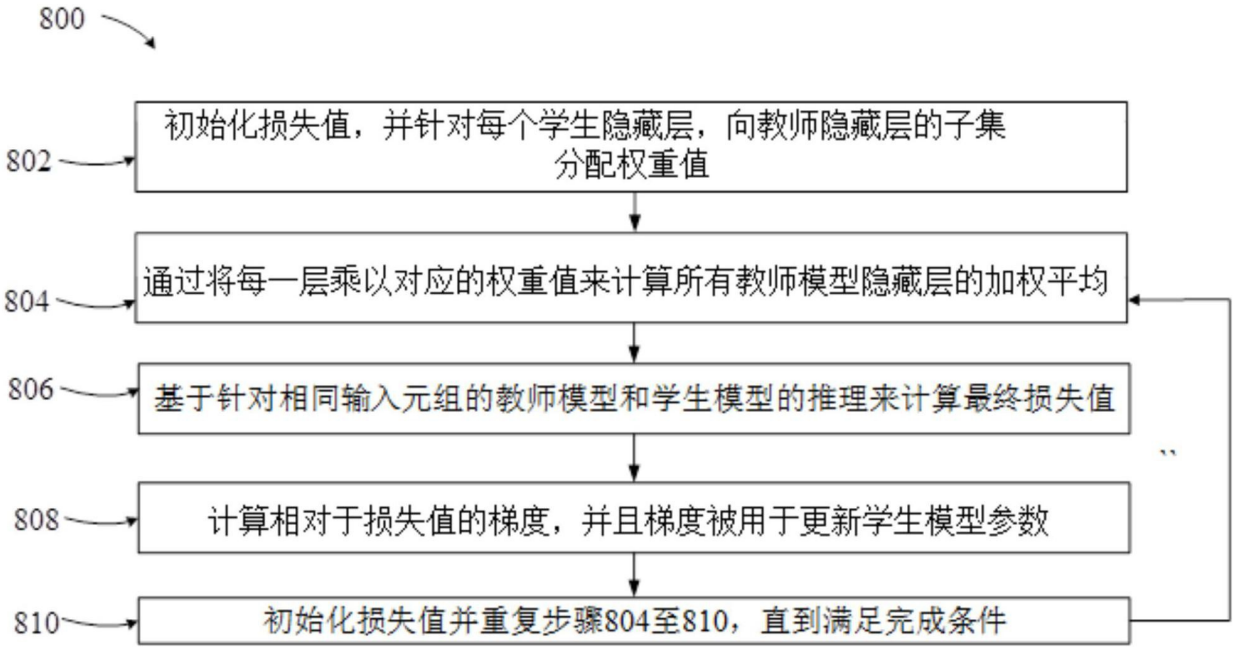

800
802
初始化损失值，并针对每个学生隐藏层，向教师隐藏层的子集分配权重值
804
通过将每一层乘以对应的权重值来计算所有教师模型隐藏层的加权平均
806
基于针对相同输入元组的教师模型和学生模型的推理来计算最终损失值
808
计算相对于损失值的梯度，并且梯度被用于更新学生模型参数
810
初始化损失值并重复步骤804至810，直到满足完成条件

图8

数据集	T	NKD	KD	PKD	注意力（没有重叠）	注意力（部分重叠）	注意力（完全重叠）
QNLI（105K）	91.25	85.13	86.77	86.64	87.11	86.95	87.02
MRPC（3.7K）	86.76	77.70	79.41	80.15	79.66	79.90	80.72
RTE（2.5K）	68.23	61.73	65.34	65.70	65.70	66.43	67.15

图9

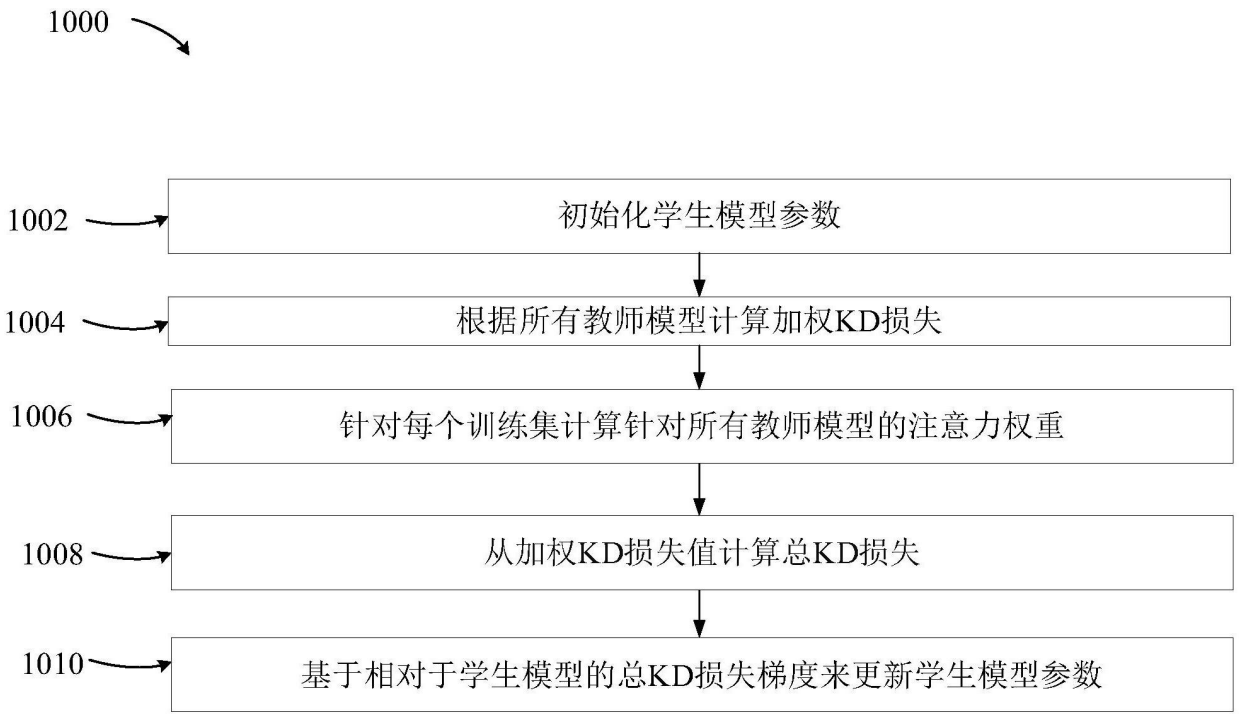

1000
1002
初始化学生模型参数
1004
根据所有教师模型计算加权KD损失
1006
针对每个训练集计算针对所有教师模型的注意力权重
1008
从加权KD损失值计算总KD损失
1010
基于相对于学生模型的总KD损失梯度来更新学生模型参数

图10

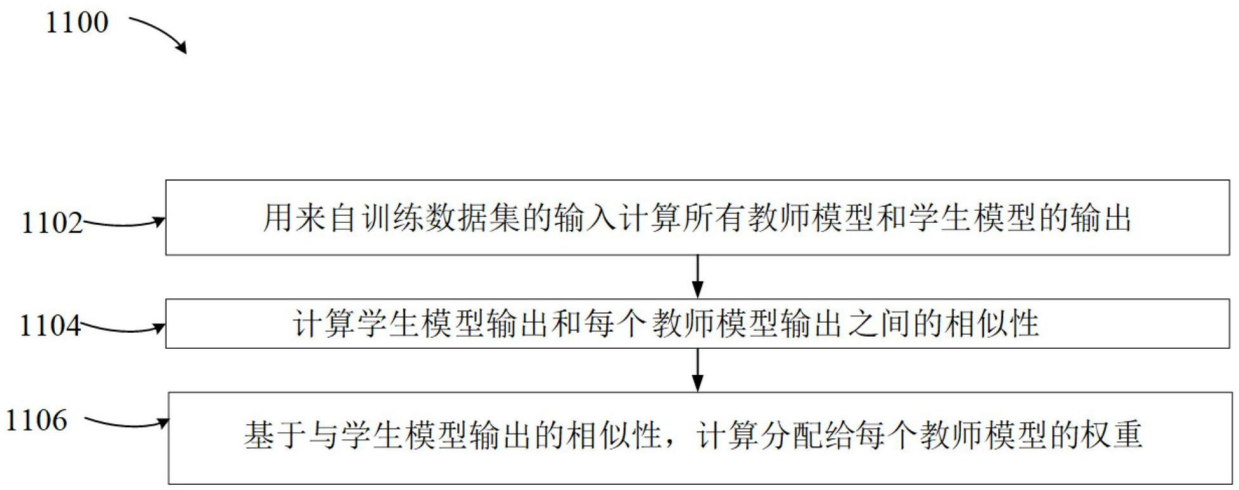

1100
1102
用来自训练数据集的输入计算所有教师模型和学生模型的输出
1104
计算学生模型输出和每个教师模型输出之间的相似性
1106
基于与学生模型输出的相似性，计算分配给每个教师模型的权重

图11

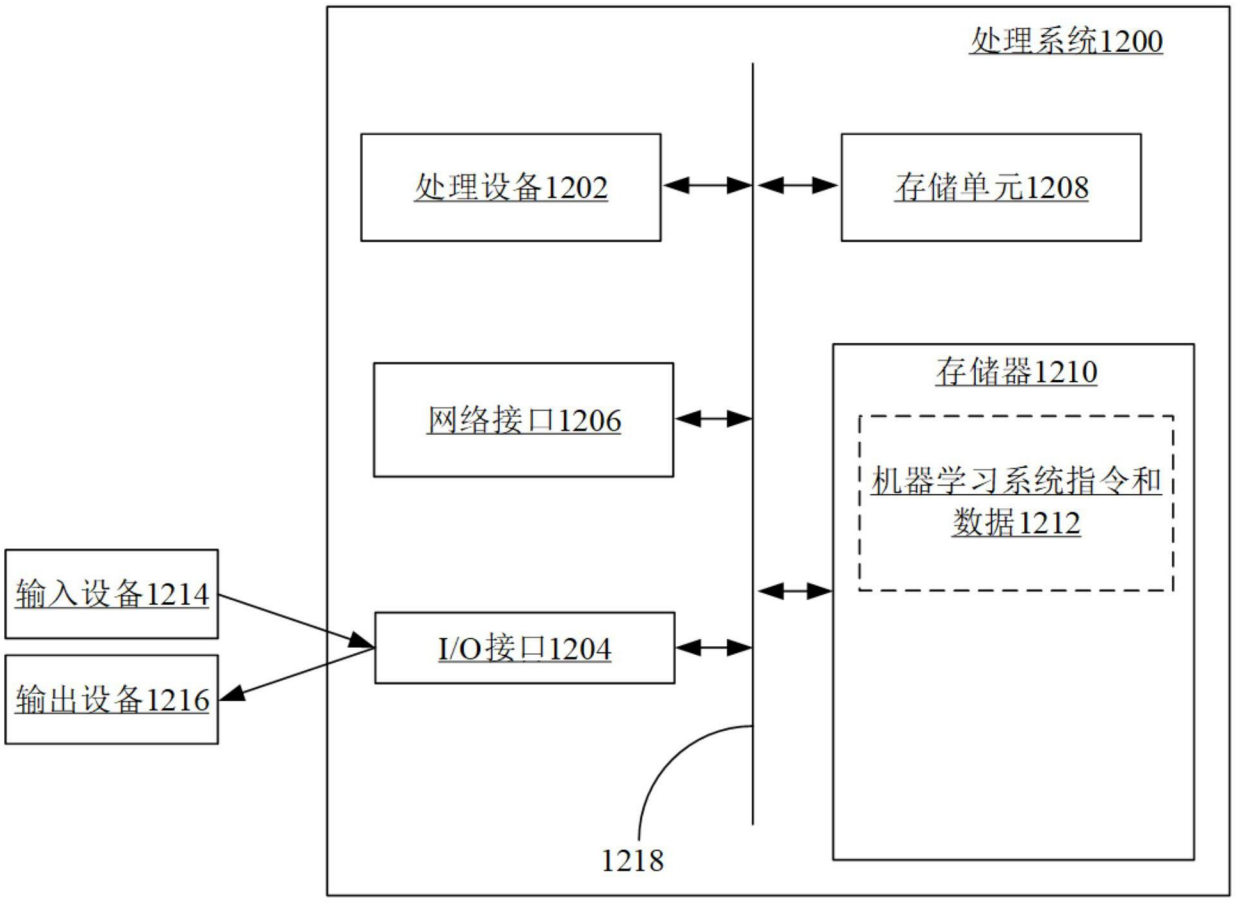

处理系统1200
处理设备1202
存储单元1208
网络接口1206
存储器1210
机器学习系统指令和数据1212
输入设备1214
输出设备1216
I/O接口1204
1218

图12