(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 117216715 A
(43)申请公布日 2023.12.12

(21)申请号 202311167770.0
(22)申请日 2023.09.11
(71)申请人 京东城市(北京)数字科技有限公司
地址 100086 北京市海淀区知春路76号(写字楼)1号楼9层1-7-5号
申请人 京东科技信息技术有限公司
(72)发明人 张振羽 陈蒙 宋阳
(74)专利代理机构 中科专利商标代理有限责任公司 11021
专利代理师 郭梦雅
(51)Int.Cl.
*G06F 18/25*(2023.01)
*G06F 18/2415*(2023.01)
*G06N 3/0464*(2023.01)
*G06N 3/045*(2023.01)
*G06N 3/048*(2023.01)
*G06N 3/08*(2023.01)

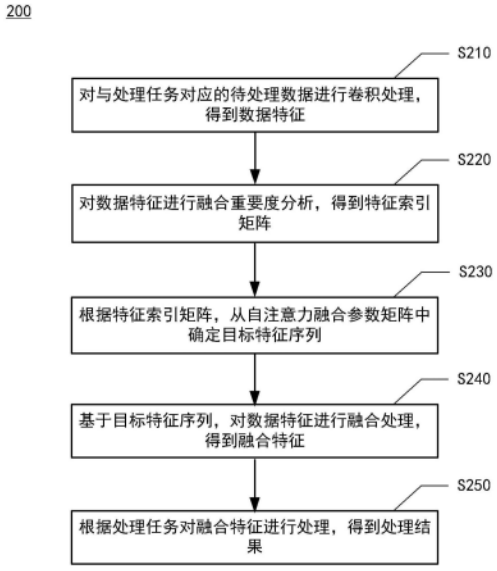
权利要求书4页 说明书18页 附图9页

(54)发明名称

# 数据处理方法、深度学习模型的训练方法及装置

(57)摘要

本公开提供了一种数据处理方法、深度学习模型的训练方法及装置,可以用于人工智能技术领域。该数据处理方法包括:对与处理任务对应的待处理数据进行卷积处理,得到数据特征;对数据特征进行融合重要度分析,得到特征索引矩阵,其中,特征索引矩阵中的每一个元素表征每一个数据特征与全局其他数据特征的融合重要度;根据特征索引矩阵,从自注意力融合参数矩阵中确定目标特征序列,其中,自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵,目标特征序列包括由目标查询特征参数和目标键特征参数组成的多个特征参数对;基于目标特征序列,对数据特征进行融合处理,得到融合特征;根据处理任务对融合特征进行处理,得到处理结果。




CN 117216715 A

1.一种数据处理方法,包括:

对与处理任务对应的待处理数据进行卷积处理,得到数据特征;

对所述数据特征进行融合重要度分析,得到特征索引矩阵,其中,所述特征索引矩阵中的每一个元素表征每一个数据特征与全局其他数据特征的融合重要度;

根据所述特征索引矩阵,从自注意力融合参数矩阵中确定目标特征序列,其中,所述自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵,所述目标特征序列包括由目标查询特征参数和目标键特征参数组成的多个特征参数对;

基于所述目标特征序列,对所述数据特征进行融合处理,得到融合特征;以及

根据所述处理任务对所述融合特征进行处理,得到处理结果。

2.根据权利要求1所述的方法,其中,所述数据特征为N乘以d维的特征矩阵,N、d均为大于1的整数;所述对所述数据特征进行融合重要度分析,得到特征索引矩阵,包括:

对第i个位置的数据特征进行融合重要度分析,得到所述第i个位置的数据特征与全局其他位置的数据特征的注意力掩码概率,其中,i为大于等于1小于等于N乘以d的整数;

在确定所述注意力掩码概率大于等于预定融合概率阈值的情况下,确定所述第i个位置的数据特征的融合重要度为第一预定值;

在确定所述注意力掩码概率大于等于预定融合概率阈值的情况下,确定所述第i个位置的数据特征的融合重要度为第二预定值;以及

根据所述特征矩阵的每一个位置的数据特征的融合重要度对应的预定值,生成所述特征索引矩阵。

3.根据权利要求2所述的方法,其中,所述对第i个位置的数据特征进行融合重要度分析,得到所述第i个位置的数据特征与全局其他位置的数据特征的注意力掩码概率,包括:

基于预定激活函数,对所述第i个位置的数据特征进行处理,得到所述第i个位置的数据特征与全局其他位置的数据特征的注意力掩码概率。

4.根据权利要求1所述的方法,其中,所述根据所述特征索引矩阵,从自注意力融合参数矩阵中确定目标特征序列,包括:

根据所述特征索引矩阵中每一个元素的元素值,分别从查询特征参数矩阵和键特征参数矩阵确定与每一个元素对应的目标查询特征参数和目标键特征参数;

将与每一个元素对应的目标查询特征参数和目标键特征参数,确定为与所述每一个元素对应的目标特征参数对;以及

根据多个所述目标特征参数对,得到所述目标特征序列。

5.根据权利要求1所述的方法,其中,所述基于所述目标特征序列,对所述数据特征进行融合处理,得到融合特征,包括:

根据所述目标特征序列,得到自注意力权重矩阵;以及

基于自注意力机制,根据所述自注意力权重矩阵对所述数据特征进行融合处理,得到所述融合特征。

6.一种深度学习模型的训练方法,包括:

利用预训练模型执行如下操作,其中,所述预训练模型是基于对初始模型进行自注意力剪枝的预训练得到的:

对与样本任务对应的第一样本数据进行卷积处理,得到第一样本数据特征;

对所述第一样本数据特征进行融合重要度分析,得到样本特征索引矩阵,其中,所述样本特征索引矩阵中的每一个元素表征每一个样本数据特征与全局其他样本数据特征的第一样本融合重要度;

基于所述样本特征索引矩阵,从自注意力融合参数矩阵中得到第一目标样本特征序列,其中,所述自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵;所述第一目标样本特征序列包括由第一目标样本查询特征参数和第一目标样本键特征参数组成的多个样本特征参数对;

基于所述第一目标样本特征序列,对所述样本数据特征进行融合处理,得到第一样本融合特征;

根据所述样本任务对所述第一样本融合特征进行处理,得到第一样本处理结果;

基于第一目标损失函数,根据所述第一样本处理结果、所述样本特征索引矩阵和样本标签得到第一损失值,其中,所述样本标签包括与所述样本处理任务对应的第一样本分类标签和目标三元组数据集;其中,所述目标三元组数据集包括注意力掩码矩阵与第一样本数据、第一样本分类标签的映射关系;所述目标三元组数据集是基于对初始模型进行自注意力剪枝的预训练得到的。

基于所述第一损失值,调整所述预训练模型的模型参数,得到经训练的深度学习模型。

7.根据权利要求1所述的方法,其中,所述对所述初始模型进行自注意力剪枝的预训练,包括:

对所述初始模型执行如下操作:

对与所述样本任务对应的第二样本数据进行卷积处理,得到第二样本数据特征;

对与第二样本数据特征对应的注意力特征参数矩阵,得到注意力掩码特征矩阵,其中,所述注意力掩码特征矩阵中每一个元素表征每一个样本数据特征与全局其他样本数据特征的第二样本融合重要度;

根据所述注意力掩码特征矩阵,得到样本特征稀疏度;

基于所述注意力掩码特征矩阵,从自注意力融合参数矩阵中得到第二目标样本特征序列,其中,所述自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵所述第二目标样本特征序列包括由第二目标样本查询特征参数和第二目标样本键特征参数组成的多个样本特征参数对;

根据所述第二目标样本特征序列,得到数据增强分布函数的缩放因子特征值;

基于所述第二目标样本特征序列,对所述第二样本数据特征进行融合处理,得到第二样本融合特征;

根据所述样本任务对所述第二样本融合特征进行处理,得到所述第二样本数据的样本处理结果;

基于第二目标损失函数,根据所述第二样本处理结果、所述注意力掩码特征矩阵、所述数据增强分布函数的缩放因子特征值、预定稀疏度和第二样本分类标签得到第二损失值;以及

基于所述第二损失值,调整所述初始模型的模型参数,得到所述预训练模型和所述目标三元组数据集。

8.根据权利要求7所述的方法,其中,所述对所述第二样本数据特征进行数据分布增强

处理,得到注意力掩码特征矩阵,包括:

基于预定激活函数,对所述第二样本数据特征进行处理,得到注意力分数矩阵;以及

基于预定数据分布增强函数,对所述注意力分数矩阵进行处理,得到所述注意力掩码特征矩阵,其中,所述注意力掩码特征矩阵中的每一个元素表征注意力分数的可信度。

9.根据权利要求7所述的方法,其中,所述根据所述注意力掩码特征矩阵,得到样本特征稀疏度,包括:

根据所述注意力掩码特征矩阵中的每一个元素的注意力分数,确定掩码特征元素在所述注意力掩码特征矩阵中的比例;以及

根据所述比例,确定所述样本特征稀疏度。

10.根据权利要求7所述的方法,其中,所述根据所述第二目标样本特征序列,得到数据增强分布函数的缩放因子特征值,包括:

根据所述第二目标样本特征序列和第一缩放因子参数矩阵,得到第一缩放因子特征值;

根据所述第二目标样本特征序列和第二缩放因子参数矩阵,得到第二缩放因子特征值;以及

根据所述第一缩放因子特征值和所述第二缩放因子特征,得到所述数据增强分布函数的缩放因子特征值。

11.一种数据处理装置,包括:

第一卷积模块,用于对与处理任务对应的待处理数据进行卷积处理,得到数据特征;

第一分析模块,用于对所述数据特征进行融合重要度分析,得到特征索引矩阵,其中,所述特征索引矩阵中的每一个元素表征每一个数据特征与全局其他数据特征的融合重要度;

第一参数确定模块,用于根据所述特征索引矩阵,从自注意力融合参数矩阵中确定目标特征序列,其中,所述自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵,所述目标特征序列包括由目标查询特征参数和目标键特征参数组成的多个特征参数对;

第一特征融合模块,用于基于所述目标特征序列,对所述数据特征进行融合处理,得到融合特征;以及

第一处理模块,用于根据所述处理任务对所述融合特征进行处理,得到处理结果。

12.一种深度学习模型的训练装置,包括:

利用预训练模型执行如下操作,其中,所述预训练模型是基于对初始模型进行自注意力剪枝的预训练得到的:

第二卷积模块,用于对与样本任务对应的第一样本数据进行卷积处理,得到第一样本数据特征;

第二分析模块,用于对所述第一样本数据特征进行融合重要度分析,得到样本特征索引矩阵,其中,所述样本特征索引矩阵中的每一个元素表征每一个样本数据特征与全局其他样本数据特征的第一样本融合重要度;

第二参数确定模块,用于基于所述样本特征索引矩阵,从自注意力融合参数矩阵中确定第一目标样本特征序列,其中,所述自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵;所述第一目标样本特征序列包括由第一目标样本查询特征参数和第一目标

样本键特征参数组成的多个样本特征参数对；

第二特征融合模块，用于基于所述第一目标样本特征序列，对所述样本数据特征进行融合处理，得到第一样本融合特征；

第二处理模块，用于根据所述样本任务对所述第一样本融合特征进行处理，得到第一样本处理结果；

第一损失计算模块，用于基于第一目标损失函数，根据所述第一样本处理结果、所述样本特征索引矩阵和样本标签得到第一损失值，其中，所述样本标签包括与所述样本处理任务对应的第一样本分类标签和目标三元组数据集；其中，所述目标三元组数据集包括注意力掩码矩阵与第一样本数据、第一样本分类标签的映射关系；所述目标三元组数据集是基于对初始模型进行自注意力剪枝的预训练得到的；

第一调整模块，用于基于所述第一损失值，调整所述预训练模型的模型参数，得到经训练的深度学习模型。

13.一种电子设备，包括：

一个或多个处理器；

存储器，用于存储一个或多个程序，

其中，当所述一个或多个程序被所述一个或多个处理器执行时，使得所述一个或多个处理器实现权利要求1至10中任一项所述的方法。

14.一种计算机可读存储介质，其上存储有可执行指令，该指令被处理器执行时使处理器实现权利要求1至10中任一项所述的方法。

15.一种计算机程序产品，包括计算机程序，所述计算机程序在被处理器执行时实现根据权利要求1-10中任一项所述的方法。

# 数据处理方法、深度学习模型的训练方法及装置

## 技术领域

[0001] 本公开涉及人工智能技术领域,具体涉及深度学习、计算机视觉、文本处理等技术领域,更具体地,涉及一种数据处理方法、深度学习模型的训练方法及装置、电子设备和程序产品。

## 背景技术

[0002] 随着人工智能技术的发展,基于自注意力机制进行特征融合的transformer网络被应用到多个领域,例如:计算机视觉、目标检测与跟踪、数据变化趋势预测、语言识别、文本翻译、自然语言理解等等。

[0003] 在实现本公开构思的过程中,发明人发现相关技术中至少存在如下问题:相关技术中的自注意力机制的计算复杂度较高,通常是输入的数据特征长度的平方倍,导致数据的存储量较大,硬件处理速度较低。

## 发明内容

[0004] 有鉴于此,本公开提供了一种数据处理方法、深度学习模型的训练方法及装置、电子设备和程序产品。

[0005] 本公开的一个方面提供了一种数据处理方法,包括:对与处理任务对应的待处理数据进行卷积处理,得到数据特征;对数据特征进行融合重要度分析,得到特征索引矩阵,其中,特征索引矩阵中的每一个元素表征每一个数据特征与全局其他数据特征的融合重要度;根据特征索引矩阵,从自注意力融合参数矩阵中确定目标特征序列,其中,自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵,目标特征序列包括由目标查询特征参数和目标键特征参数组成的多个特征参数对;基于目标特征序列,对数据特征进行融合处理,得到融合特征;以及根据处理任务对融合特征进行处理,得到处理结果。

[0006] 根据本公开的实施例,数据特征为N乘以d维的特征矩阵,N、d均为大于1的整数,对数据特征进行融合重要度分析,得到特征索引矩阵,包括:对第i个位置的数据特征进行融合重要度分析,得到第i个位置的数据特征与全局其他位置的数据特征的注意力掩码概率,其中,i为大于等于1小于等于N乘以d的整数;在确定注意力掩码概率大于等于预定融合概率阈值的情况下,确定第i个位置的数据特征的融合重要度为第一预定值;在确定注意力掩码概率大于等于预定融合概率阈值的情况下,确定第i个位置的数据特征的融合重要度为第二预定值;以及根据特征矩阵的每一个位置的数据特征的融合重要度对应的预定值,生成特征索引矩阵。

[0007] 根据本公开的实施例,对第i个位置的数据特征进行融合重要度分析,得到第i个位置的数据特征与全局其他位置的数据特征的注意力掩码概率,包括:基于预定激活函数,对第i个位置的数据特征进行处理,得到第i个位置的数据特征与全局其他位置的数据特征的注意力掩码概率。

[0008] 根据本公开的实施例,根据特征索引矩阵,从自注意力融合参数矩阵中确定目标

特征序列,包括:根据特征索引矩阵中每一个元素的元素值,分别从查询特征参数矩阵和键特征参数矩阵确定与每一个元素对应的目标查询特征参数和目标键特征参数;将与每一个元素对应的目标查询特征参数和目标键特征参数,确定为与每一个元素对应的目标特征参数对;以及根据多个目标特征参数对,得到目标特征序列。

[0009] 根据本公开的实施例,基于目标特征序列,对数据特征进行融合处理,得到融合特征,包括:根据目标特征序列,得到自注意力权重矩阵;以及基于自注意力机制,根据自注意力权重矩阵对数据特征进行融合处理,得到融合特征。

[0010] 本公开的另一个方面提供了一种深度学习模型的训练方法,包括:利用预训练模型执行如下操作,其中,预训练模型是基于对初始模型进行自注意力剪枝的预训练得到的:对与样本任务对应的第一样本数据进行卷积处理,得到第一样本数据特征;对第一样本数据特征进行融合重要度分析,得到样本特征索引矩阵,其中,样本特征索引矩阵中的每一个元素表征每一个样本数据特征与全局其他样本数据特征的第一样本融合重要度;基于第一样本特征索引矩阵,从自注意力融合参数矩阵中得到第一目标样本特征序列,其中,自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵;第一目标样本特征序列包括由第一目标样本查询特征参数和第一目标样本键特征参数组成的多个样本特征参数对;基于第一目标样本特征序列,对样本数据特征进行融合处理,得到第一样本融合特征;根据样本任务对第一样本融合特征进行处理,得到第一样本处理结果;基于第一目标损失函数,根据第一样本处理结果、样本特征索引矩阵和样本标签得到第一损失值,其中,样本标签包括与样本处理任务对应的第一样本分类标签和目标三元组数据集;其中,目标三元组数据集包括注意力掩码矩阵与第一样本数据、第一样本分类标签的映射关系,其中,目标三元组数据集是基于对初始模型进行自注意力剪枝的预训练得到的;基于第一损失值,调整预训练模型的模型参数,得到经训练的深度学习模型。

[0011] 根据本公开的实施例,对初始模型进行自注意力剪枝的预训练,包括:对初始模型执行如下操作:对与样本任务对应的第二样本数据进行卷积处理,得到第二样本数据特征;对第二样本数据特征进行数据分布增强处理,得到注意力掩码特征矩阵,其中,注意力掩码特征矩阵中每一个元素表征每一个样本数据特征与全局其他样本数据特征的第二样本融合重要度;根据注意力掩码特征矩阵,得到样本特征稀疏度;基于注意力掩码特征矩阵,从自注意力融合参数矩阵中得到第二目标样本特征序列,其中,自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵第二目标样本特征序列包括由第二目标样本查询特征参数和第二目标样本键特征参数组成的多个样本特征参数对;根据第二目标样本特征序列,得到数据增强分布函数的缩放因子特征值;基于第二目标样本特征序列,对第二样本数据特征进行融合处理,得到第二样本融合特征;根据样本任务对第二样本融合特征进行处理,得到第二样本数据的样本处理结果;基于第二目标损失函数,根据第二样本处理结果、注意力掩码特征矩阵、数据增强分布函数的缩放因子特征值、预定稀疏度和第二样本分类标签得到第二损失值;以及基于第二损失值,调整初始模型的模型参数,得到预训练模型和目标三元组数据集,其中,目标三元组数据集中表征第二样本数据特征、第二样本分类标签与注意力掩码特征矩阵的映射关系。

[0012] 根据本公开的实施例,对第二样本数据特征进行数据分布增强处理,得到注意力掩码特征矩阵,包括:基于预定激活函数,对第二样本数据特征进行处理,得到注意力分数

矩阵;以及基于预定数据分布增强函数,对注意力分数矩阵进行处理,得到注意力掩码特征矩阵,其中,注意力掩码特征矩阵中的每一个元素表征注意力分数的可信度。

[0013] 根据本公开的实施例,根据注意力掩码特征矩阵,得到样本特征稀疏度,包括:根据注意力掩码特征矩阵中的每一个元素的注意力分数,确定掩码特征元素在注意力掩码特征矩阵中的比例;以及根据比例,确定样本特征稀疏度。

[0014] 根据本公开的实施例,根据第二目标样本特征序列,得到数据增强分布函数的缩放因子特征值,包括:根据第二目标样本特征序列和第一缩放因子参数矩阵,得到第一缩放因子特征值;根据第二目标样本特征序列和第二缩放因子参数矩阵,得到第二缩放因子特征值;以及根据第一缩放因子特征值和第二缩放因子特征,得到数据增强分布函数的缩放因子特征值。

[0015] 本公开的另一方面提供了一种数据处理装置,包括:第一卷积模块、第一分析模块、第一参数确定模块、第一特征融合模块和第一处理模块。第一卷积模块,用于用于对与处理任务对应的待处理数据进行卷积处理,得到数据特征。第一分析模块,用于对数据特征进行融合重要度分析,得到特征索引矩阵,其中,特征索引矩阵中的每一个元素表征每一个数据特征与全局其他数据特征的融合重要度。第一分析模块,用于用于根据特征索引矩阵,从自注意力融合参数矩阵中确定目标特征序列,其中,自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵,目标特征序列包括由目标查询特征参数和目标键特征参数组成的多个特征参数对。第一特征融合模块,用于基于目标特征序列,对数据特征进行融合处理,得到融合特征。第一处理模块,用于根据处理任务对融合特征进行处理,得到处理结果。

[0016] 根据本公开的另一个方面,提供了一种深度学习模型的训练装置包括:第二卷积模块、第二分析模块、第二参数确定模块、第二特征融合模块、第二处理模块、第一损失计算模块和第一调整模块。

[0017] 根据本公开的实施例,第二卷积模块,用于对与样本任务对应的第一样本数据进行卷积处理,得到样本数据特征。第二分析模块,用于对样本数据特征进行融合重要度分析,得到样本特征索引矩阵,其中,样本特征索引矩阵中的每一个元素表征每一个样本数据特征与全局其他样本数据特征的第一样本融合重要度。第二参数确定模块,用于基于第一样本特征索引矩阵,从自注意力融合参数矩阵中确定第一目标样本特征序列,其中,自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵;第一目标样本特征序列包括由第一目标样本查询特征参数和第一目标样本键特征参数组成的多个样本特征参数对。第二特征融合模块,用于基于第一目标样本特征序列,对样本数据特征进行融合处理,得到第一样本融合特征。第二处理模块,用于根据样本任务对第一样本融合特征进行处理,得到第一样本处理结果。第一损失计算模块,用于基于第一目标损失函数,根据第一样本处理结果、样本特征索引矩阵和样本标签得到第一损失值,其中,样本标签包括与样本处理任务对应的第一样本分类标签和目标三元组数据集;其中,目标三元组数据集包括注意力掩码矩阵与第一样本数据、第一样本分类标签的映射关系;目标三元组数据集是基于对初始模型进行自注意力剪枝的预训练得到的。第一调整模块,用于基于第一损失值,调整预训练模型的模型参数,得到经训练的深度学习模型。

[0018] 本公开的另一方面提供了一种计算机可读存储介质,存储有计算机可执行指令,

上述指令在被执行时用于实现如上描述的方法。

[0019] 本公开的另一方面提供了一种计算机程序产品,上述计算机程序产品包括计算机可执行指令,上述指令在被执行时用于实现如上描述的方法。

[0020] 根据本公开的实施例,因为采用了对数据特征进行融合重要度分析,根据特征索引矩阵中每一个数据特征与全局其他数据特征的融合重要度,从自注意力融合参数矩阵中确定用于数据特征融合的特征参数对的技术手段,所以至少部分地克服了相关技术中基于自注意力机制的参数计算复杂度较高,数据存储量过多,硬件数据处理速度较低的技术问题,从自注意力融合参数矩阵中选择与数据特征强相关的特征参数对,达到了在确保数据处理结果准确度的同时,降低数据存储量,提高计算机硬件的数据处理速度的技术效果。

## 附图说明

[0021] 通过以下参照附图对本公开实施例的描述,本公开的上述以及其他目的、特征和优点将更为清楚,在附图中:

[0022] 图1示意性示出了可以应用本公开的用于数据处理方法、深度学习模型的训练方法及装置的示例性架构图;

[0023] 图2示意性示出了根据本公开实施例的数据处理方法的流程图;

[0024] 图3示意性示出了根据本公开实施例的对数据特征进行融合重要度分析,得到特征索引矩阵的示意图;

[0025] 图4示意性示出了根据本公开实施例的根据特征索引矩阵,从自注意力融合参数矩阵中确定目标特征序列的示意图;

[0026] 图5示意性示出了根据本公开实施例的深度学习模型的训练方法流程图;

[0027] 图6示意性示出了根据本公开实施例的深度学习模型的训练原理示意图;

[0028] 图7示意性示出了根据本公开实施例的数据处理装置的框图;

[0029] 图8示意性示出了根据本公开实施例的深度学习模型的训练装置框图;以及

[0030] 图9示意性示出了根据本公开实施例的适于实现数据处理方法、深度学习模型的训练方法的电子设备900的框图。

## 具体实施方式

[0031] 以下,将参照附图来描述本公开的实施例。但是应该理解,这些描述只是示例性的,而并非要限制本公开的范围。在下面的详细描述中,为便于解释,阐述了许多具体的细节以提供对本公开实施例的全面理解。然而,明显地,一个或多个实施例在没有这些具体细节的情况下也可以被实施。此外,在以下说明中,省略了对公知结构和技术的描述,以避免不必要地混淆本公开的概念。

[0032] 在此使用的术语仅仅是为了描述具体实施例,而并非意在限制本公开。在此使用的术语“包括”、“包含”等表明了所述特征、步骤、操作和/或部件的存在,但是并不排除存在或添加一个或多个其他特征、步骤、操作或部件。

[0033] 在此使用的所有术语(包括技术和科学术语)具有本领域技术人员通常所理解的含义,除非另外定义。应注意,这里使用的术语应解释为具有与本说明书的上下文相一致的含义,而不应以理想化或过于刻板的方式来解释。

[0034] 在使用类似于“A、B和C等中至少一个”这样的表述的情况下,一般来说应该按照本领域技术人员通常理解该表述的含义来予以解释(例如,“具有A、B和C中至少一个的系统”应包括但不限于单独具有A、单独具有B、单独具有C、具有A和B、具有A和C、具有B和C、和/或具有A、B、C的系统等)。

[0035] 基于transformer架构的模型现在已经是NLP(Neuro-Linguistic Programming,自然语言处理)、CV(ComputerVision,计算机视觉)等诸多领域的生成式模型,其核心是一种基于自注意力机制的全局信息融合方法:对于长度为N输入序列表征,首先获取对应的请求($Q\in R^{N\times d}$)和键($K\in R^{N\times d}$)的表征,然后计算Q-K注意力从而利用全局信息更新表征H。在具体应用中,transformer模型的自注意力机制的计算复杂度是序列长度(N)的平方倍,导致对于较长的输入序列计算复杂度迅速提高。在真实工业场景中,数据序列一般较长,自注意力机制的这个特性使得计算实时性较差,阻碍了transformer模型的在应用场景中的高效部署。

[0036] 为了降低自注意力的计算复杂度,目前的方法主要通过稀疏化设计的自注意力机制和层级模型的形式来实现。

[0037] 对于稀疏化设计的自注意力机制主要包括以下几种方式:longformer,performer,sinkhore attention,卷积自注意力混合模型等。Longformer方法中人工设计组合了滑动窗口注意力、局部窗口空洞注意力和个别位置的全局注意力,从而降低了自注意力机的计算复杂度。Performer方法则利用核函数将注意力计算分解到低纬空间进行运算,设计出了一种基于正交随机特征的快速注意力。Sinkhorn attention方法则把全局注意力划分为多个子窗口块内部的注意力计算,同时通过排序网络为每个子窗口块找到可用的需要计算注意力的键,从而使得全局注意力降级为局部注意力。ConvBert方法中将多头自注意力的一半注意力头通过span卷积网络(类似局部注意力)来实现,通过这种混合方案使得计算复杂度降低且内存消耗变少。

[0038] 但是这类方法普遍需要人工提前设计相关的稀疏策略,自注意力计算方式都要提前定义完善,不能根据输入的数据动态调整;此外,这些方法设计的局部注意力都是固定窗口大小。但在真实的应用场景中,这种假设并不适用于真实的应用场景中的待处理图像、文本或语音等数据的数据特征。导致目前的稀疏注意力方法虽然一定程度降低了计算复杂度,但是却难以保证较高的准确率。

[0039] 层级模型主要出现在CV领域,这类方法在模型架构上使用多阶段的金字塔形式的结构,逐步降低输入的尺寸大小,从而使得注意力感受野变小,达到降低计算复杂度的目的。但是这类网络架构多用于序列分类任务,不适合NLP领域中很多很任务,例如序列标注,阅读理解等,导致其应用范围受限制。

[0040] 有鉴于此,本公开实施例提供了一种数据处理方法,包括:对与处理任务对应的待处理数据进行卷积处理,得到数据特征;对数据特征进行融合重要度分析,得到特征索引矩阵,其中,特征索引矩阵中的每一个元素表征每一个数据特征与全局其他数据特征的融合重要度;根据特征索引矩阵,从自注意力融合参数矩阵中确定目标特征序列,其中,自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵,目标特征序列包括由目标查询特征参数和目标键特征参数组成的多个特征参数对;基于目标特征序列,对数据特征进行融合处理,得到融合特征;以及根据处理任务对融合特征进行处理,得到处理结果。以达到

在确保数据处理结果准确度的同时,降低数据存储量,提高计算机硬件的数据处理速度的技术效果。

[0041] 在本公开的实施例中,所涉及的数据(例如,包括但不限于用户个人信息)的收集、更新、分析、处理、使用、传输、提供、公开、存储等方面,均符合相关法律法规的规定,被用于合法的用途,且不违背公序良俗。特别地,对用户个人信息采取了必要措施,防止对用户个人信息数据的非法访问,维护用户个人信息安全、网络安全和国家安全。

[0042] 在本公开的实施例中,在获取或采集用户个人信息之前,均获取了用户的授权或同意。

[0043] 图1示意性示出了可以应用本公开的用于数据处理方法、深度学习模型的训练方法及装置的示例性架构图。需要注意的是,图1所示仅为可以应用本公开实施例的系统架构的示例,以帮助本领域技术人员理解本公开的技术内容,但并不意味着本公开实施例不可以用于其他设备、系统、环境或场景。

[0044] 如图1所示,根据该实施例的系统架构100可以包括第一终端设备101、第二终端设备102、第三终端设备103,网络104和服务器105。网络104用以在第一终端设备101、第二终端设备102、第三终端设备103之间提供通信链路的介质。网络104可以包括各种连接类型,例如有线和/或无线通信链路等等。

[0045] 用户可以使用第一终端设备101、第二终端设备102、第三终端设备103通过网络104与服务器105交互,以接收或发送消息等。第一终端设备101、第二终端设备102、第三终端设备103上可以安装有各种通讯客户端应用,例如购物类应用、网页浏览器应用、搜索类应用、即时通信工具、邮箱客户端和/或社交平台软件等(仅为示例)。

[0046] 第一终端设备101、第二终端设备102、第三终端设备103可以是具有显示屏并且支持网页浏览的各种电子设备,包括但不限于智能手机、平板电脑、膝上型便携计算机和台式计算机等等。

[0047] 服务器105可以是提供各种服务的服务器,例如对用户利用第一终端设备101、第二终端设备102、第三终端设备103所浏览的网站提供支持的后台管理服务器(仅为示例)。后台管理服务器可以对接收到的用户请求等数据进行分析等处理,并将处理结果(例如根据用户请求获取或生成的网页、信息、或数据等)反馈给终端设备。

[0048] 需要说明的是,本公开实施例所提供的数据处理方法或深度学习模型的训练方法也可以由第一终端设备101、第二终端设备102、或第三终端设备103执行,或者也可以由不同于第一终端设备101、第二终端设备102、或第三终端设备103的其他终端设备执行。相应地,本公开实施例所提供的数据处理装置或深度学习模型的训练装置也可以设置于第一终端设备101、第二终端设备102、或第三终端设备103中,或设置于不同于第一终端设备101、第二终端设备102、或第三终端设备103的其他终端设备中。

[0049] 或者,本公开实施例所提供的数据处理方法或深度学习模型的训练方法一般可以由服务器105执行。相应地,本公开实施例所提供的数据处理装置或深度学习模型的训练装置一般可以设置于服务器105中。本公开实施例所提供的数据处理方法或深度学习模型的训练方法也可以由不同于服务器105且能够与第一终端设备101、第二终端设备102、第三终端设备103和/或服务器105通信的服务器或服务器集群执行。相应地,本公开实施例所提供的数据处理装置或深度学习模型的训练装置也可以设置于不同于服务器105且能够与第一

终端设备101、第二终端设备102、第三终端设备103和/或服务器105通信的服务器或服务器集群中。

[0050] 例如,待处理数据可以由第一终端设备101、第二终端设备102或第三终端设备103获取,并由第一终端设备101、第二终端设备102、第三终端设备103向服务器105发送待处理数据。由服务器105执行本公开实施例提供的数据处理方法,得到处理结果,并将处理结果返回第一终端设备101、第二终端设备102或第三终端设备103。

[0051] 应该理解,图1中的终端设备、网络和服务器的数目仅仅是示意性的。根据实现需要,可以具有任意数目的终端设备、网络和服务器。

[0052] 图2示意性示出了根据本公开实施例的数据处理方法的流程图。

[0053] 如图2所示,该方法200包括操作S210～S250。

[0054] 在操作S210,对与处理任务对应的待处理数据进行卷积处理,得到数据特征。

[0055] 在操作S220,对数据特征进行融合重要度分析,得到特征索引矩阵。

[0056] 在操作S230,根据特征索引矩阵,从自注意力融合参数矩阵中确定目标特征序列。

[0057] 在操作S240,基于目标特征序列,对数据特征进行融合处理,得到融合特征。

[0058] 在操作S250,根据处理任务对融合特征进行处理,得到处理结果。

[0059] 根据本公开的实施例,处理任务可以包括与应用场景相对应的处理任务,处理任务与待处理数据的类型是相对应的。

[0060] 例如:在计算机视觉应用场景中,处理任务可以是目标检测任务,待处理的数据的类型可以是图像,对图像进行卷积处理,得到的可以是图像特征。

[0061] 例如:在语音识别的应用场景中,处理任务可以是语音识别任务,待处理的数据的类型可以是音频。对音频进行卷积处理,得到的可以是音频特征。

[0062] 例如:在语义理解的应用场景中,处理任务可以是文本语义识别,待处理的数据的类型可以是文本。对文本进行卷积处理,得到的可以是文本特征。

[0063] 根据本公开的实施例,融合重要度可以表征在数据特征矩阵中某个位置的数据特征是否需要与全局的其他位置的数据特征进行注意力的计算。在特征索引矩阵中每一个元素表征每一个数据特征与全局其他数据特征的融合重要度。

[0064] 根据本公开的实施例,在数据特征矩阵中第i个位置的数据特征与其他位置的数据特征需要进行注意力计算,表征该位置的数据特征与其他位置的数据特征相关性较强,可以通过注意力计算,将相关性较强的数据特征进行融合,以得到融合特征。

[0065] 根据本公开的实施例,自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵。

[0066] 根据本公开的实施例,基于自注意力机制,每一个数据特征对应一个查询特征参数矩阵Q和一个键特征参数矩阵K。相关技术中,通常是将查询特征参数矩阵Q中每一个元素与键特征参数矩阵K元素一次进行一次计算,最终得到自注意力权重。因此,对于一个长度为N的数据特征序列,自注意力权重的计算复杂度是数据特征序列长度的平方倍,计算复杂度非常高。

[0067] 根据本公开的实施例,由于特征索引矩阵中每一个元素可以表征该位置的数据特征是否需要与其他位置的数据特征进行注意力计算,因此,可以定向地从自注意力融合参数矩阵中选择出需要进行注意力计算的数据特征对应的特征参数对,该特征参数对中可以

包括目标查询特征参数和目标键特征参数。

[0068] 例如:对于数据特征矩阵第i个位置的数据特征,基于与该位置对应的特征索引矩阵中第i个位置的索引值,从查询特征参数矩阵和键特征参数矩阵中确定与第i个位置的数据特征对应的特征参数对“$Q_i/K_i$”。

[0069] 根据本公开的实施例,基于目标特征序列,再对数据特征进行自注意力计算,由于只需基于目标特征序列中的特征参数对确定自注意力权重,从而降低了自注意力计算的复杂度。

[0070] 根据本公开的实施例,根据处理任务对融合特征进行处理,可以是将融合特征输入全连接层,全连接层的网络架构可以根据具体的处理任务而定,在此不作具体限定,输出处理结果。例如:处理任务可以是目标检测任务,处理结果可以是待处理的图像的目标对象的类别与预定对象类别的匹配概率以及目标对象所在位置(即:目标检测框)。

[0071] 需要说明的是,在本公开的实施例中,基于特征索引矩阵对特征参数对进行筛选,实现对自注意力参数的剪枝,但是并没有改变transformer网络架构的输入数据和输出数据的类型。

[0072] 根据本公开的实施例,因为采用了对数据特征进行融合重要度分析,根据特征索引矩阵中每一个数据特征与全局其他数据特征的融合重要度,从自注意力融合参数矩阵中确定用于数据特征融合的特征参数对的技术手段,所以至少部分地克服了相关技术中基于自注意力机制的参数计算复杂度较高,数据存储量过多,硬件数据处理速度较低的技术问题,从自注意力融合参数矩阵中选择与数据特征强相关的特征参数对,达到了在确保数据处理结果准确度的同时,降低数据存储量,提高计算机硬件的数据处理速度的技术效果。

[0073] 根据本公开的实施例,数据特征为N乘以d维的特征矩阵,N、d均为大于1的整数。对数据特征进行融合重要度分析,得到特征索引矩阵,可以包括如下操作:对第i个位置的数据特征进行融合重要度分析,得到第i个位置的数据特征与全局其他位置的数据特征的注意力掩码概率,其中,i为大于等于1小于等于N乘以d的整数;在确定注意力掩码概率大于等于预定注意力掩码概率阈值的情况下,确定第i个位置的数据特征的融合重要度为第一预定值;在确定注意力掩码概率大于等于预定注意力掩码概率阈值的情况下,确定第i个位置的数据特征的融合重要度为第二预定值;根据特征矩阵的每一个位置的数据特征的融合重要度对应的预定值,生成特征索引矩阵。

[0074] 根据本公开的实施例,对第i个位置的数据特征进行融合重要度分析,得到第i个位置的数据特征与全局其他位置的数据特征的注意力掩码概率。

[0075] 例如:可以按照式(1)计算注意力掩码概率:

[0076] $$Ind_i = \mathrm{sigmoid}(W_{index} H_{c,i}^T) \quad (1)$$

[0077] 其中,$Ind_i$表示注意力掩码概率,$W_{index}$表示可学习的网络参数,$H_{c,i} \in R^d$,$Ind_i \in R^N$,$H^T_{c,i}$表示第i个位置的数据特征。

[0078] 图3示意性示出了根据本公开实施例的对数据特征进行融合重要度分析,得到特征索引矩阵的示意图。

[0079] 如图3所示,在实施例300中,对数据特征矩阵321中每一个位置的数据特征与其他位置的数据特征进行融合重要度分析,注意力掩码概率矩阵322。在注意力掩码概率矩阵322中,每一个元素表征该位置的数据特征与其他位置的数据特征的注意力掩码概率(即需

要计算注意力的概率)。例如:与第i个位置的数据特征Fi对应的注意力掩码概率为pi。

[0080] 在操作S320中,确定pi是否大于预定注意力掩码概率阈值。若是,则在特征索引矩阵A324中对应位置的融合重要度值为1。若不是,则在特征索引矩阵A323中对应位置的融合重要度值为0。

[0081] 例如:第i个位置的数据特征与其他位置的数据特征需要计算注意力的概率为0.8,注意力掩码概率阈值可以是0.7,则第i个位置的融合重要度为1,表示在进行自注意力计算时,需要从查询特征参数矩阵和键特征参数矩阵中确定与该位置对应的查询特征参数Qi和键特征参数Ki。

[0082] 根据本公开的实施例,由于自注意力机制中,每一个查询特征参数在全局的注意力分布是不同的,基于融合重要度可以筛选出与数据特征强相关的特征参数对,从而有效降低自注意力计算的复杂度,提高特征融合过程的数据处理速度。

[0083] 根据本公开的实施例,根据特征索引矩阵,从自注意力融合参数矩阵中确定目标特征序列,可以包括如下操作:根据特征索引矩阵中每一个元素的元素值,分别从查询特征参数矩阵和键特征参数矩阵确定与每一个元素对应的目标查询特征参数和目标键特征参数;将与每一个元素对应的目标查询特征参数和目标键特征参数,确定为与每一个元素对应的目标特征参数对。根据多个目标特征参数对,得到目标特征序列。

[0084] 例如:可以将特征索引矩阵中的每一个元素的元素值与查询特征参数矩阵和键特征参数矩阵进行点乘,以得到目标特征序列。

[0085] 图4示意性示出了根据本公开实施例的根据特征索引矩阵,从自注意力融合参数矩阵中确定目标特征序列的示意图。

[0086] 如图4所示,在实施例400中,将特征索引矩阵431与查询特征参数矩阵432进行元素点乘,得到查询特征参数序列434。将特征索引矩阵431与键特征参数矩阵433进行元素点乘,得到键特征参数序列435。按照位置对应关系,将查询特征参数序列434与键特征参数序列435进行组合,得到目标特征序列436。

[0087] 根据本公开的实施例,根据特征索引矩阵中每一个元素的元素值,分别从查询特征参数矩阵和键特征参数矩阵确定与每一个元素对应的目标查询特征参数和目标键特征参数,可以仅基于目标查询特征参数和目标键特征参数进行自注意力计算,从而降低了自注意力计算的复杂度,提高数据处理速度。

[0088] 根据本公开的实施例,基于目标特征序列,对数据特征进行融合处理,得到融合特征,可以包括如下操作:根据目标特征序列,得到自注意力权重矩阵;以及基于自注意力机制,根据自注意力权重矩阵对数据特征进行融合处理,得到融合特征。

[0089] 例如:可以基于目标特征序列,按照式(3)进行自注意力计算,得到融合特征:

[0090]
$$\begin{aligned} &\mathrm{Q,K,V=W}_Q H, W_K H, W_V H, \quad W_Q, W_K, W_V \in R^{d\times d} \\ &\beta_{ij}=\frac{Q_i K_j}{\sqrt{d}}, Q_i, K_i, V_i \in R^d \\ &\alpha_i = soft\max(\{\beta_{i1}, \beta_{i2}, ..., \beta_{iN}\}) \in R^N \\ &H' = W_O H + \alpha V, \quad \alpha V = \{\alpha_1 V, \alpha_2 V, ..., \alpha_N V\}, \quad \mathrm{W_O} \in R^{d\times d} \end{aligned} \tag{3}$$

[0091] 其中,$\alpha_i$表示自注意力权重,$W_Q$,$W_K$,$W_V$,$W_O$都表示可学习的网络参数,H′表示融合特

征，$Q_iK_j$表示目标特征序列。

[0092] 根据本公开的实施例，由于仅需要基于目标特征序列中特征参数对进行的注意力加权计算，因此，有效降低了自注意力计算的复杂度，提高了数据处理的速度。

[0093] 图5示意性示出了根据本公开实施例的深度学习模型的训练方法流程图。

[0094] 如图5所示，在实施例500中，利用预训练模型执行如下操作：对与样本任务对应的第一样本数据501进行卷积处理，得到第一样本数据特征502。对第一样本数据特征502进行融合重要度分析，得到样本特征索引矩阵503。基于样本特征索引矩阵503，从自注意力融合参数矩阵中得到第一目标样本特征序列504。基于第一目标样本特征序列504，对第一样本数据特征502进行融合处理，得到第一样本融合特征505。按照样本任务对第一样本融合特征505进行处理，得到第一样本处理结果506。

[0095] 根据本公开的实施例，在实施例500中，从第一样本数据501到第一样本处理结果506的数据处理过程与前文描述的数据处理方法中的数据处理过程相同，样本处理任务、样本数据、样本数据特征等也与前文描述的数据处理方法中的处理任务、待处理数据、数据特征的定义范围相同，在此不作赘述。

[0096] 根据本公开的实施例，第一样本分类标签可以是与样本任务对应的，例如：样本任务可以是目标检测任务，样本图像中可以包括树、人、房子，则对应的样本分类标签可以分别为树、人、房子。

[0097] 根据本公开的实施例，可以基于交叉熵损失函数根据第一样本处理结果506和第一样本分类标签507得到分类损失值508。

[0098] 根据本公开的实施例，样本特征索引矩阵503中每一个元素表示每一个元素表征每一个样本数据特征与全局其他样本数据特征的第一样本融合重要度。

[0099] 在本公开实施例的模型训练过程中可以包括两个阶段，第一个阶段是对自注意力剪枝学习进行预训练的过程，预训练结束，可以得到目标三元组数据集509。目标三元组数据集中包括样本分类标签、样本数据与注意力掩码特征矩阵之间的映射关系。因此，在第二个阶段中，目标三元组数据集可以作为样本特征索引矩阵的标签，用于计算掩码损失值。

[0100] 根据本公开的实施例，可以基于公式(4)计算掩码损失值：

[0101] $$L_{ind}=\sum_{j=1}^{j=N} z_{ij}\log(Ind_{i,j})+(1-z_{ij})\log(1-Ind_{i,j}) \quad (4)$$

[0102] 其中，$z_{ij}$表示在注意力掩码特征矩阵中第i行第j个位置的值；$Ind_{i,j}$表示在样本特征索引矩阵中第i行第j列的值。

[0103] 根据本公开的实施例，在对注意力的特征参数进行筛选的同时，需要样本处理结果的准确度，因此，第一损失值511可以是分类损失值508与掩码损失值510之和，也可以是分类损失值508与掩码损失值510的加权求和的结果。

[0104] 根据本公开的实施例，基于第一损失值511调整模型参数，得到经训练的深度学习模型512。

[0105] 根据本公开的实施例，基于掩码损失和与样本任务对应的样本分类损失共同训练模型，可以根据样本数据的特征，准确预测特征索引矩阵，以便从自注意力参数矩阵中准确确定目标特征参数对，并且可以在降低模型数据处理量的同时，确保模型预测的准确度。

[0106] 根据本公开的实施例，对初始模型进行自注意力剪枝的预训练，包括如下操作：对

与样本任务对应的第二样本数据进行卷积处理,得到第二样本数据特征;对第二样本数据特征进行数据分布增强处理,得到注意力掩码特征矩阵,其中,注意力掩码特征矩阵中每一个元素表征每一个样本数据特征与全局其他样本数据特征的第二样本融合重要度;根据注意力掩码特征矩阵,得到样本特征稀疏度;基于注意力掩码特征矩阵,从自注意力融合参数矩阵中得到第二目标样本特征序列,其中,自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵第二目标样本特征序列包括由第二目标样本查询特征参数和第二目标样本键特征参数组成的多个样本特征参数对;根据第二目标样本特征序列,得到数据增强分布函数的缩放因子特征值;基于第二目标样本特征序列,对第二样本数据特征进行融合处理,得到第二样本融合特征;根据样本任务对第二样本融合特征进行处理,得到第二样本数据的样本处理结果;基于第二目标损失函数,根据第二样本处理结果、注意力掩码特征矩阵、数据增强分布函数的缩放因子特征值、预定稀疏度和第二样本分类标签得到第二损失值;基于第二损失值,调整初始模型的模型参数,得到预训练模型和目标三元组数据集。

[0107] 图6示意性示出了根据本公开实施例的深度学习模型的训练原理示意图。

[0108] 如图6所示,在实施例600中,每个查询特征q都会对应有一个全局的注意力分布,而且这种注意力分布在不同的位置分布也是不同的。因此,可以基于L0正则算法,为每个位置的自注意力权重引入正则优化参数,即融合重要度。

[0109] 根据本公开的实施例,对第二样本数据特征进行数据分布增强处理,得到注意力掩码特征矩阵,可以包括如下操作:基于预定激活函数,对第二样本数据特征进行处理,得到注意力分数矩阵。基于预定数据分布增强函数,对注意力分数矩阵进行处理,得到注意力掩码特征矩阵,其中,注意力掩码特征矩阵中的每一个元素表征注意力分数的可信度。

[0110] 例如:可以按照式(5)计算每一个位置的注意力分数的可信度z:

[0111]
$$u \sim U(0,1)$$
$$s = sigmoid(\frac{1}{b}\log(\frac{u}{1-u}) + \log a)$$
$$\bar{s} = s \times (r - l) + l, \quad r > 0, l < 0$$
$$z = \min(1, \max(0, \bar{s})) \tag{5}$$

[0112] 其中,u表示0~1之间的随机数,b表示预定的超参数,log a表示可学习参数,l、r表示数据增强分布函数(hard-concrete)的上下边界缩放因子值;z表示注意力分数的可信度。

[0113] 根据本公开的实施例,根据注意力掩码特征矩阵,得到样本特征稀疏度,可以包括如下操作:根据注意力掩码特征矩阵中的每一个元素的注意力分数,确定掩码特征元素在注意力掩码特征矩阵中的比例;以及根据比例,确定样本特征稀疏度。

[0114] 例如:注意力掩码特征矩阵中的元素共10x4个,其中,元素值为0的元素为15个,则样本特征稀疏度可以为15/40=37.5%。

[0115] 根据本公开的实施例,在训练阶段1S610中,可以预定每一次训练得到的注意力掩码矩阵中融合度为0的元素占全部元素的比例,即稀疏度,并可以通过设定目标稀疏度限定模型训练的最大训练次数。

[0116] 例如:可以通过涉及一个稀疏度的调度程序,设定在训练总次数的前20%的次数

中,注意力掩码特征矩阵的稀疏度由0线性增至目标稀疏度$S_0$,如式(6)所示:

[0117] $$S_t = \frac{t}{T_{warmup}} S_0 \quad (6)$$

[0118] 其中,T标识最大训练次数,$T_{warmup}$表示稀疏度线性增加的最大步数;$S_0$表示目标稀疏度;$S_t$表示第t个迭代训练次数中得到的注意力掩码特征矩阵的稀疏度。

[0119] 根据本公开的实施例中,在第t次迭代训练过程中,可以根据注意力分数的可信度$z_{ij}$按照式(7)更新自注意力权重$\alpha_{ij}$,并基于更新后的自注意力权重,代入公式(3)中进行注意力计算。

[0120] $$\tilde{\alpha}_{ij} = z_{ij} \times \alpha_{ij} \quad (7)$$

[0121] 根据本公开的实施例,在第t次迭代训练过程中,更新后的自注意力权重对应的$Q_iK_j$可以作为第二目标样本特征序列,得到该轮训练的数据增强分布函数的缩放因子特征值。

[0122] 例如:根据第二目标样本特征序列和第一缩放因子参数矩阵,得到第一缩放因子特征值;根据第二目标样本特征序列和第二缩放因子参数矩阵,得到第二缩放因子特征值;以及根据第一缩放因子特征值和第二缩放因子特征,得到数据增强分布函数的缩放因子特征值。

[0123] 根据本公开的实施例,第一缩放因子特征值可以表征数据增强分布函数的上边界缩放因子l,第二缩放因子特征值可以表征数据增强分布函数的下边界缩放因子r。

[0124] 例如:可以按照式(8)计算数据增强分布函数的缩放因子特征值(l,r):

[0125] $l = W_l(Q_iK_j)$, $r = W_r(Q_iK_j)$ (8)

[0126] 其中,$W_l$、$W_r$为可学习的网络参数,$Q_iK_j$为第二样本目标样本特征序列中的特征参数对。

[0127] 根据本公开的实施例,在预训练的过程中,第二损失值可以包括样本分类损失值、数据增强分布损失值和稀疏度压缩损失值,其中,样本分类损失值与前文描述的相同,在此不作赘述。

[0128] 根据本公开的实施例,由于在数据增强分布函数中上边界的缩放因子应该小于下边界的缩放因子,因此,可以按照式(9)计算数据增强分布损失值:

[0129] $L_h = \max(l,0) - \min(r,0)$ (9)

[0130] 其中,l表示第一缩放因子特征值,r表示第二缩放因子特征值。

[0131] 根据本公开的实施例,可以按照式(10)计算稀疏度压缩损失值:

[0132] $L_s = \lambda_1(E(z) - S_t) + \lambda_2(E(z) - S_t)^2$ (10)

[0133] 其中,$\lambda_1$、$\lambda_2$表示可学习的网络参数,$S_t$表示第t次迭代训练的注意力掩码特征矩阵的稀疏度;E(z)表示公式(5)中计算得到的注意力分数的期望。

[0134] 如图6所示,经过训练阶段-1S610实现了对与序列表征601对应的自注意力参数的有效剪枝,得到目标三元组数据集602。在目标三元组数据集602中可以指明数据特征、样本分类标签与特征参数对的映射关系。

[0135] 在训练阶段-2S602中,将序列表征601进行动态卷积602,得到数据特征,再将数据特征输入特征索引矩阵预测器(N-index预测器)603得到特征索引矩阵。基于预测得到的特

征索引矩阵索引选择对应的特征参数对604进行注意力计算。并利用训练阶段-1中的注意力掩码矩阵对训练阶段-2中特征索引矩阵进行损失计算,以提高特征索引预测的准确度。

[0136] 根据本公开的实施例,通过引入L0正则算法和数据增强分布函数,基于样本分类损失值、数据增强分布损失值和稀疏度压缩损失值调整模型参数,得到样本数据特征、样本分类标签与注意力掩码特征矩阵的映射关系,从而实现了对自注意力参数的有效剪枝。

[0137] 图7示意性示出了根据本公开实施例的数据处理装置的框图。

[0138] 如图7所示,在实施例700中可以包括第一卷积模块710、第一分析模块720、第一参数确定模块730、第一特征融合模块740和第一处理模块750。

[0139] 第一卷积模块710,用于对与处理任务对应的待处理数据进行卷积处理,得到数据特征。

[0140] 第一分析模块720,用于对数据特征进行融合重要度分析,得到特征索引矩阵,其中,特征索引矩阵中的每一个元素表征每一个数据特征与全局其他数据特征的融合重要度。

[0141] 第一参数确定模块730,用于根据特征索引矩阵,从自注意力融合参数矩阵中确定目标特征序列,其中,自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵,目标特征序列包括由目标查询特征参数和目标键特征参数组成的多个特征参数对。

[0142] 第一特征融合模块740,用于基于目标特征序列,对数据特征进行融合处理,得到融合特征。

[0143] 第一处理模块750,用于根据处理任务对融合特征进行处理,得到处理结果。

[0144] 根据本公开的实施例,第一分析模块包括:第一分析子模块、第一确定子模块、第二确定子模块和第一生成子模块。第一分析子模块,用于对第i个位置的数据特征进行融合重要度分析,得到第i个位置的数据特征与全局其他位置的数据特征的注意力掩码概率,其中,i为大于等于1小于等于N乘以d的整数。第一确定子模块,用于在确定注意力掩码概率大于等于预定注意力掩码概率阈值的情况下,确定第i个位置的数据特征的融合重要度为第一预定值。第二确定子模块,用于在确定注意力掩码概率大于等于预定注意力掩码概率阈值的情况下,确定第i个位置的数据特征的融合重要度为第二预定值。第一生成子模块,用于根据特征矩阵的每一个位置的数据特征的融合重要度对应的预定值,生成特征索引矩阵。

[0145] 根据本公开的实施例,第一分析子模块包括第一分析单元,用于基于预定激活函数,对第i个位置的数据特征进行处理,得到第i个位置的数据特征与全局其他位置的数据特征的注意力掩码概率。

[0146] 根据本公开的实施例,第一参数确定模块包括第三确定子模块、第四确定子模块和第一获得子模块。第三确定子模块,用于根据特征索引矩阵中每一个元素的元素值,分别从查询特征参数矩阵和键特征参数矩阵确定与每一个元素对应的目标查询特征参数和目标键特征参数。第四确定子模块,用于将与每一个元素对应的目标查询特征参数和目标键特征参数,确定为与每一个元素对应的目标特征参数对。第一获得子模块,用于根据多个目标特征参数对,得到目标特征序列。

[0147] 根据本公开的实施例,第一特征融合模块包括第二获得子模块、第一融合子模块。第二获得子模块,用于根据目标特征序列,得到自注意力权重矩阵。第一融合子模块,用于

基于自注意力机制,根据自注意力权重矩阵对数据特征进行融合处理,得到融合特征。

[0148] 图8示意性示出了根据本公开实施例的深度学习模型的训练装置的框图。

[0149] 如图8所示,在实施例800中可以包括第二卷积模块810、第二分析模块820、第二参数确定模块830、第二特征融合模块840、第二处理模块890、第一损失计算模块860和第一调整模块870。

[0150] 第二卷积模块810,用于对与样本任务对应的第一样本数据进行卷积处理,得到样本数据特征。

[0151] 第二分析模块820,用于对样本数据特征进行融合重要度分析,得到样本特征索引矩阵,其中,样本特征索引矩阵中的每一个元素表征每一个样本数据特征与全局其他样本数据特征的第一样本融合重要度。

[0152] 第二参数确定模块,用于基于第一样本特征索引矩阵,从自注意力融合参数矩阵中确定第一目标样本特征序列,其中,自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵;第一目标样本特征序列包括由第一目标样本查询特征参数和第一目标样本键特征参数组成的多个样本特征参数对。

[0153] 第二特征融合模块,用于基于第一目标样本特征序列,对样本数据特征进行融合处理,得到第一样本融合特征。

[0154] 第二处理模块,用于根据样本任务对第一样本融合特征进行处理,得到第一样本处理结果。

[0155] 第一损失计算模块,用于基于第一目标损失函数,根据第一样本处理结果、样本特征索引矩阵和样本标签得到第一损失值,其中,样本标签包括与样本处理任务对应的第一样本分类标签和目标三元组数据集;其中,目标三元组数据集包括注意力掩码矩阵与第一样本数据、第一样本分类标签的映射关系;目标三元组数据集是基于对初始模型进行自注意力剪枝的预训练得到的;

[0156] 第一调整模块,用于基于第一损失值,调整预训练模型的模型参数,得到经训练的深度学习模型。

[0157] 根据本公开的实施例,上述训练装置800还包括预训练模块,其中,所述预训练模块包括卷积子模块、数据分布增强子模块、稀疏度计算子模块、参数确定子模块、缩放因子特征获得子模块、特征融合子模块、处理子模块、损失计算子模块和调整子模块。

[0158] 根据本公开的实施例,卷积子模块,用于对与样本任务对应的第二样本数据进行卷积处理,得到第二样本数据特征。数据分布增强子模块,用于对第二样本数据特征进行数据分布增强处理,得到注意力掩码特征矩阵,其中,注意力掩码特征矩阵中每一个元素表征每一个样本数据特征与全局其他样本数据特征的第二样本融合重要度。稀疏度计算子模块,用于根据注意力掩码特征矩阵,得到样本特征稀疏度。参数确定子模块,用于基于注意力掩码特征矩阵,从自注意力融合参数矩阵中得到第二目标样本特征序列,其中,自注意力融合参数矩阵包括查询特征参数矩阵和键特征参数矩阵第二目标样本特征序列包括由第二目标样本查询特征参数和第二目标样本键特征参数组成的多个样本特征参数对。缩放因子特征获得子模块,用于根据第二目标样本特征序列,得到数据增强分布函数的缩放因子特征值。特征融合子模块,用于基于第二目标样本特征序列,对第二样本数据特征进行融合处理,得到第二样本融合特征。特征融合子模块,用于根据样本任务对第二样本融合特征进

行处理,得到第二样本数据的样本处理结果。损失计算子模块,用于基于第二目标损失函数,根据第二样本处理结果、注意力掩码特征矩阵、数据增强分布函数的缩放因子特征值、预定稀疏度和第二样本分类标签得到第二损失值。调整子模块,用于基于第二损失值,调整初始模型的模型参数,得到预训练模型和目标三元组数据集。

[0159] 根据本公开的实施例,数据分布增强子模块可以包括注意力分数计算单元和注意力掩码特征计算单元。其中,注意力分数计算单元,用于基于预定激活函数,对第二样本数据特征进行处理,得到注意力分数矩阵。注意力掩码特征计算单元,用于基于预定数据分布增强函数,对注意力分数矩阵进行处理,得到注意力掩码特征矩阵,其中,注意力掩码特征矩阵中的每一个元素表征注意力分数的可信度。

[0160] 根据本公开的实施例,稀疏度计算子模块包括:掩码特征比例计算单元和稀疏度计算单元。其中,掩码特征比例计算单元,用于根据注意力掩码特征矩阵中的每一个元素的注意力分数,确定掩码特征元素在注意力掩码特征矩阵中的比例。稀疏度计算单元,用于根据比例,确定样本特征稀疏度。

[0161] 根据本公开的实施例,缩放因子特征获得子模块包括第一缩放因子特征获得单元、第二缩放因子特征获得单元和数据增强分布函数的缩放因子特征获得单元。其中,第一缩放因子特征获得单元,用于根据第二目标样本特征序列和第一缩放因子参数矩阵,得到第一缩放因子特征值。第二缩放因子特征获得单元,用于根据第二目标样本特征序列和第二缩放因子参数矩阵,得到第二缩放因子特征值。数据增强分布函数的缩放因子特征获得单元,用于根据第一缩放因子特征值和第二缩放因子特征,得到数据增强分布函数的缩放因子特征值。

[0162] 根据本公开的实施例的模块、子模块、单元、子单元中的任意多个、或其中任意多个的至少部分功能可以在一个模块中实现。根据本公开实施例的模块、子模块、单元、子单元中的任意一个或多个可以被拆分成多个模块来实现。根据本公开实施例的模块、子模块、单元、子单元中的任意一个或多个可以至少被部分地实现为硬件电路,例如现场可编程门阵列(FPGA)、可编程逻辑阵列(PLA)、片上系统、基板上的系统、封装上的系统、专用集成电路(ASIC),或可以通过对电路进行集成或封装的任何其他的合理方式的硬件或固件来实现,或以软件、硬件以及固件三种实现方式中任意一种或以其中任意几种的适当组合来实现。或者,根据本公开实施例的模块、子模块、单元、子单元中的一个或多个可以至少被部分地实现为计算机程序模块,当该计算机程序模块被运行时,可以执行相应的功能。

[0163] 例如,第二卷积模块810、第二分析模块820、第二参数确定模块830、第二特征融合模块840、第二处理模块890、第一损失计算模块860和第一调整模块870中的任意多个可以合并在一个模块/单元/子单元中实现,或者其中的任意一个模块/单元/子单元可以被拆分成多个模块/单元/子单元。或者,这些模块/单元/子单元中的一个或多个模块/单元/子单元的至少部分功能可以与其他模块/单元/子单元的至少部分功能相结合,并在一个模块/单元/子单元中实现。根据本公开的实施例,第二卷积模块810、第二分析模块820、第二参数确定模块830、第二特征融合模块840、第二处理模块890、第一损失计算模块860和第一调整模块870中的至少一个可以至少被部分地实现为硬件电路,例如现场可编程门阵列(FPGA)、可编程逻辑阵列(PLA)、片上系统、基板上的系统、封装上的系统、专用集成电路(ASIC),或可以通过对电路进行集成或封装的任何其他的合理方式等硬件或固件来实现,或以软件、

硬件以及固件三种实现方式中任意一种或以其中任意几种的适当组合来实现。或者,第二卷积模块810、第二分析模块820、第二参数确定模块830、第二特征融合模块840、第二处理模块890、第一损失计算模块860和第一调整模块870中的至少一个可以至少被部分地实现为计算机程序模块,当该计算机程序模块被运行时,可以执行相应的功能。

[0164] 需要说明的是,本公开的实施例中数据处理装置部分与本公开的实施例中数据处理方法部分是相对应的,数据处理装置部分的描述具体参考数据处理方法部分,在此不再赘述。

[0165] 图9示意性示出了根据本公开实施例的适于实现上文描述的方法的电子设备的框图。图9示出的电子设备仅仅是一个示例,不应对本公开实施例的功能和使用范围带来任何限制。

[0166] 如图9所示,根据本公开实施例的电子设备900包括处理器901,其可以根据存储在只读存储器(ROM)902中的程序或者从存储部分908加载到随机访问存储器(RAM)903中的程序而执行各种适当的动作和处理。处理器901例如可以包括通用微处理器(例如CPU)、指令集处理器和/或相关芯片组和/或专用微处理器(例如,专用集成电路(ASIC)),等等。处理器901还可以包括用于缓存用途的板载存储器。处理器901可以包括用于执行根据本公开实施例的方法流程的不同动作的单一处理单元或者是多个处理单元。

[0167] 在RAM 903中,存储有电子设备900操作所需的各种程序和数据。处理器901、ROM 902以及RAM 903通过总线904彼此相连。处理器901通过执行ROM 902和/或RAM 903中的程序来执行根据本公开实施例的方法流程的各种操作。需要注意,所述程序也可以存储在除ROM 902和RAM 903以外的一个或多个存储器中。处理器901也可以通过执行存储在所述一个或多个存储器中的程序来执行根据本公开实施例的方法流程的各种操作。

[0168] 根据本公开的实施例,电子设备900还可以包括输入/输出(I/O)接口905,输入/输出(I/O)接口905也连接至总线904。系统900还可以包括连接至I/O接口905的以下部件中的一项或多项:包括键盘、鼠标等的输入部分906;包括诸如阴极射线管(CRT)、液晶显示器(LCD)等以及扬声器等的输出部分907;包括硬盘等的存储部分908;以及包括诸如LAN卡、调制解调器等的网络接口卡的通信部分909。通信部分909经由诸如因特网的网络执行通信处理。驱动器910也根据需要连接至I/O接口905。可拆卸介质911,诸如磁盘、光盘、磁光盘、半导体存储器等等,根据需要安装在驱动器910上,以便于从其上读出的计算机程序根据需要被安装入存储部分908。

[0169] 根据本公开的实施例,根据本公开实施例的方法流程可以被实现为计算机软件程序。例如,本公开的实施例包括一种计算机程序产品,其包括承载在计算机可读存储介质上的计算机程序,该计算机程序包含用于执行流程图所示的方法的程序代码。在这样的实施例中,该计算机程序可以通过通信部分909从网络上被下载和安装,和/或从可拆卸介质911被安装。在该计算机程序被处理器901执行时,执行本公开实施例的系统中限定的上述功能。根据本公开的实施例,上文描述的系统、设备、装置、模块、单元等可以通过计算机程序模块来实现。

[0170] 本公开还提供了一种计算机可读存储介质,该计算机可读存储介质可以是上述实施例中描述的设备/装置/系统中所包含的;也可以是单独存在,而未装配入该设备/装置/系统中。上述计算机可读存储介质承载有一个或者多个程序,当上述一个或者多个程序被

执行时,实现根据本公开实施例的方法。

[0171] 根据本公开的实施例,计算机可读存储介质可以是非易失性的计算机可读存储介质。例如可以包括但不限于:便携式计算机磁盘、硬盘、随机访问存储器(RAM)、只读存储器(ROM)、可擦式可编程只读存储器(EPROM或闪存)、便携式紧凑磁盘只读存储器(CD-ROM)、光存储器件、磁存储器件、或者上述的任意合适的组合。在本公开中,计算机可读存储介质可以是任何包含或存储程序的有形介质,该程序可以被指令执行系统、装置或者器件使用或者与其结合使用。

[0172] 例如,根据本公开的实施例,计算机可读存储介质可以包括上文描述的ROM 902和/或RAM 903和/或ROM 902和RAM 903以外的一个或多个存储器。

[0173] 本公开的实施例还包括一种计算机程序产品,其包括计算机程序,该计算机程序包含用于执行本公开实施例所提供的方法的程序代码,当计算机程序产品在电子设备上运行时,该程序代码用于使电子设备实现本公开实施例所提供的数据处理方法或深度学习模型的训练方法。

[0174] 在该计算机程序被处理器901执行时,执行本公开实施例的系统/装置中限定的上述功能。根据本公开的实施例,上文描述的系统、装置、模块、单元等可以通过计算机程序模块来实现。

[0175] 在一种实施例中,该计算机程序可以依托于光存储器件、磁存储器件等有形存储介质。在另一种实施例中,该计算机程序也可以在网络介质上以信号的形式进行传输、分发,并通过通信部分909被下载和安装,和/或从可拆卸介质911被安装。该计算机程序包含的程序代码可以用任何适当的网络介质传输,包括但不限于:无线、有线等等,或者上述的任意合适的组合。

[0176] 根据本公开的实施例,可以以一种或多种程序设计语言的任意组合来编写用于执行本公开实施例提供的计算机程序的程序代码,具体地,可以利用高级过程和/或面向对象的编程语言、和/或汇编/机器语言来实施这些计算程序。程序设计语言包括但不限于诸如Java,C++,python,“C”语言或类似的程序设计语言。程序代码可以完全地在用户计算设备上执行、部分地在用户设备上执行、部分在远程计算设备上执行、或者完全在远程计算设备或服务器上执行。在涉及远程计算设备的情形中,远程计算设备可以通过任意种类的网络,包括局域网(LAN)或广域网(WAN),连接到用户计算设备,或者,可以连接到外部计算设备(例如利用因特网服务提供商来通过因特网连接)。

[0177] 附图中的流程图和框图,图示了按照本公开各种实施例的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上,流程图或框图中的每个方框可以代表一个模块、程序段、或代码的一部分,上述模块、程序段、或代码的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。也应当注意,在有些作为替换的实现中,方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如,两个接连地表示的方框实际上可以基本并行地执行,它们有时也可以按相反的顺序执行,这依所涉及的功能而定。也要注意的是,框图或流程图中的每个方框、以及框图或流程图中的方框的组合,可以用执行规定的功能或操作的专用的基于硬件的系统来实现,或者可以用专用硬件与计算机指令的组合来实现。本领域技术人员可以理解,本公开的各个实施例和/或权利要求中记载的特征可以进行多种组合和/或结合,即使这样的组合或结合没有明确记载于本公开中。特别地,在

不脱离本公开精神和教导的情况下,本公开的各个实施例和/或权利要求中记载的特征可以进行多种组合和/或结合。所有这些组合和/或结合均落入本公开的范围。

[0178] 以上对本公开的实施例进行了描述。但是,这些实施例仅仅是为了说明的目的,而并非为了限制本公开的范围。尽管在以上分别描述了各实施例,但是这并不意味着各个实施例中的措施不能有利地结合使用。本公开的范围由所附权利要求及其等同物限定。不脱离本公开的范围,本领域技术人员可以做出多种替代和修改,这些替代和修改都应落在本公开的范围之内。

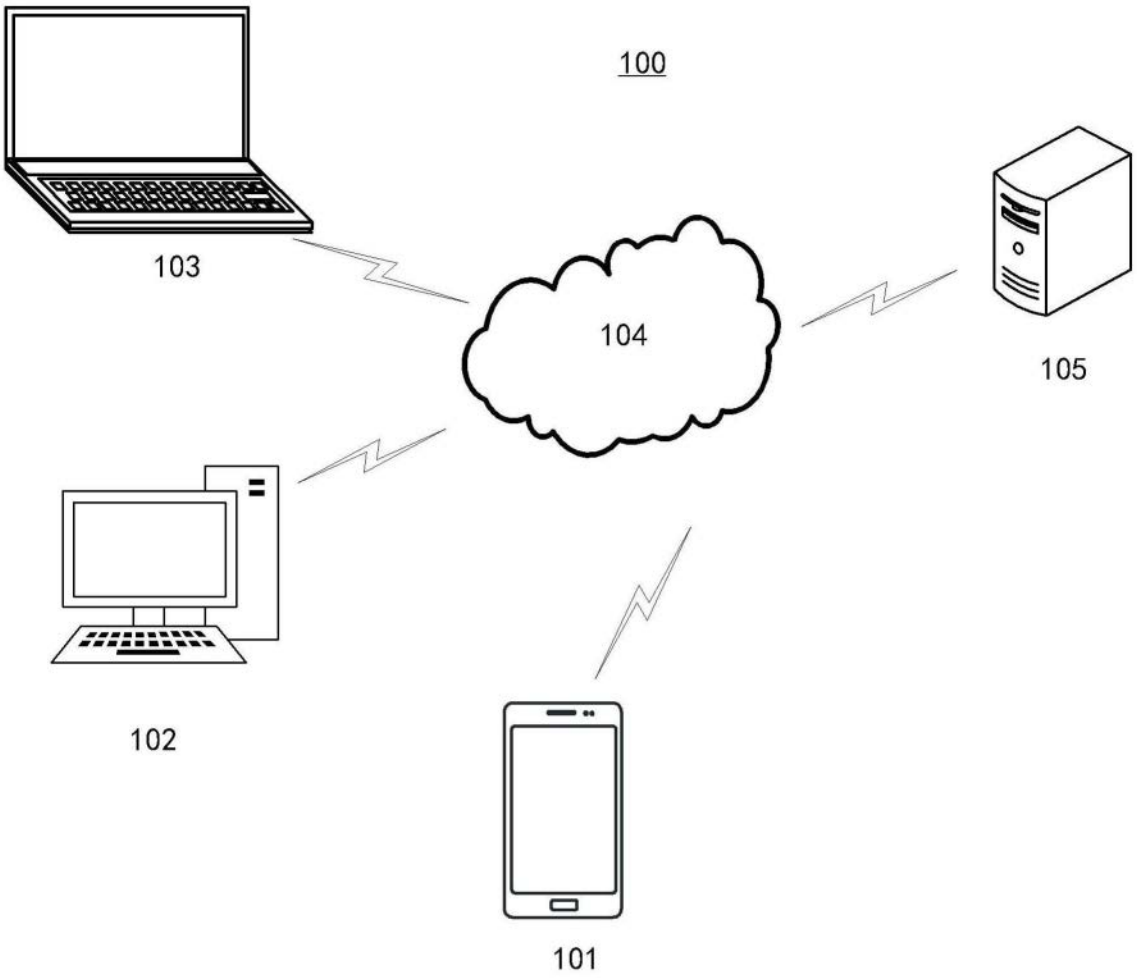



图1

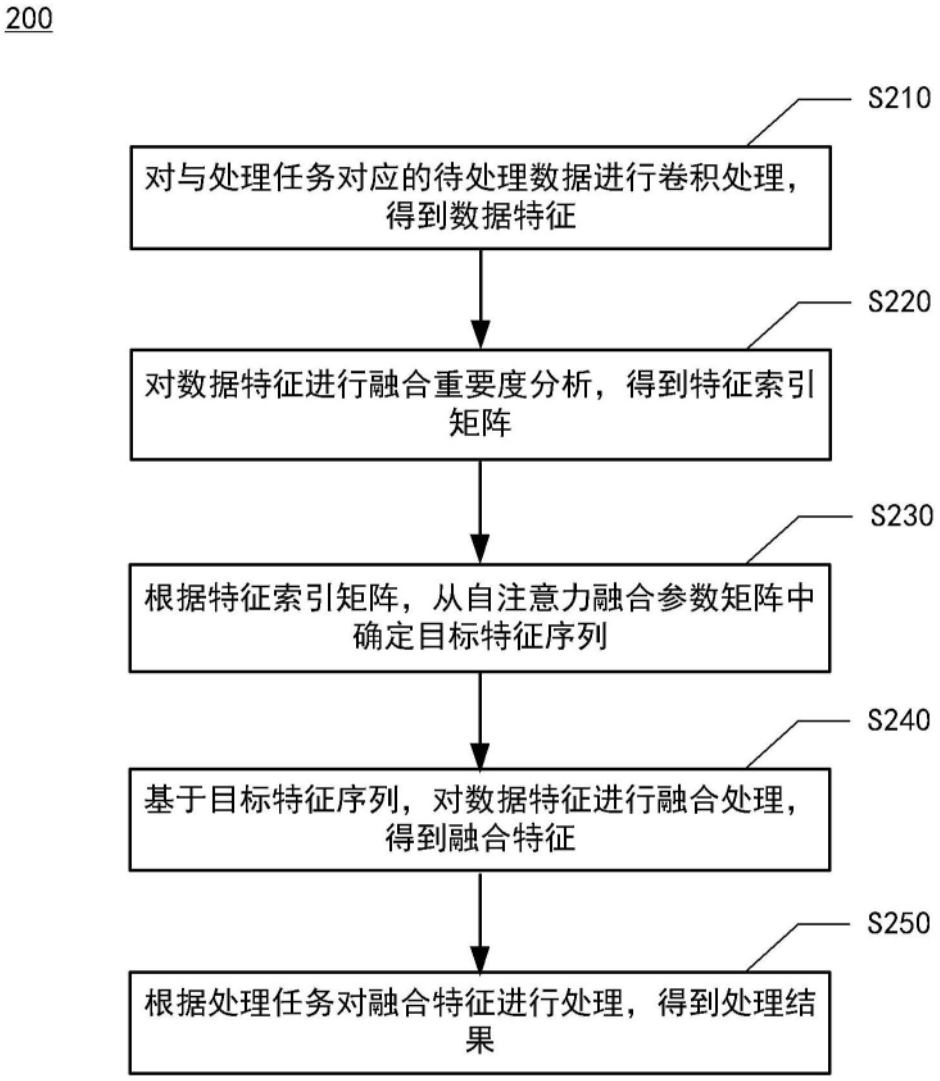

200
S210
对与处理任务对应的待处理数据进行卷积处理，得到数据特征
S220
对数据特征进行融合重要度分析，得到特征索引矩阵
S230
根据特征索引矩阵，从自注意力融合参数矩阵中确定目标特征序列
S240
基于目标特征序列，对数据特征进行融合处理，得到融合特征
S250
根据处理任务对融合特征进行处理，得到处理结果

图2

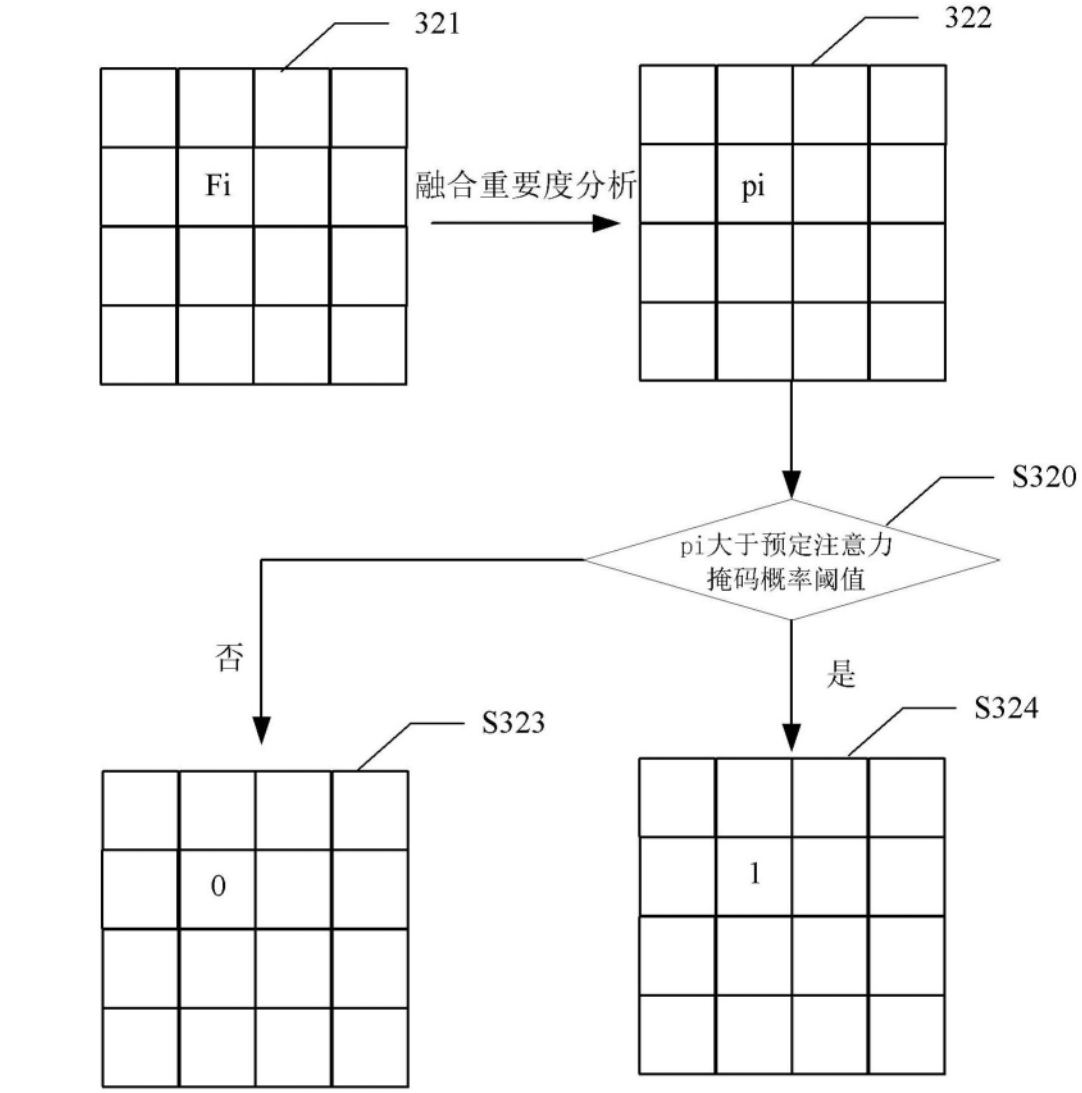

300
321
322
Fi
融合重要度分析
pi
S320
pi大于预定注意力
掩码概率阈值
否
是
S323
S324
0
1

图3

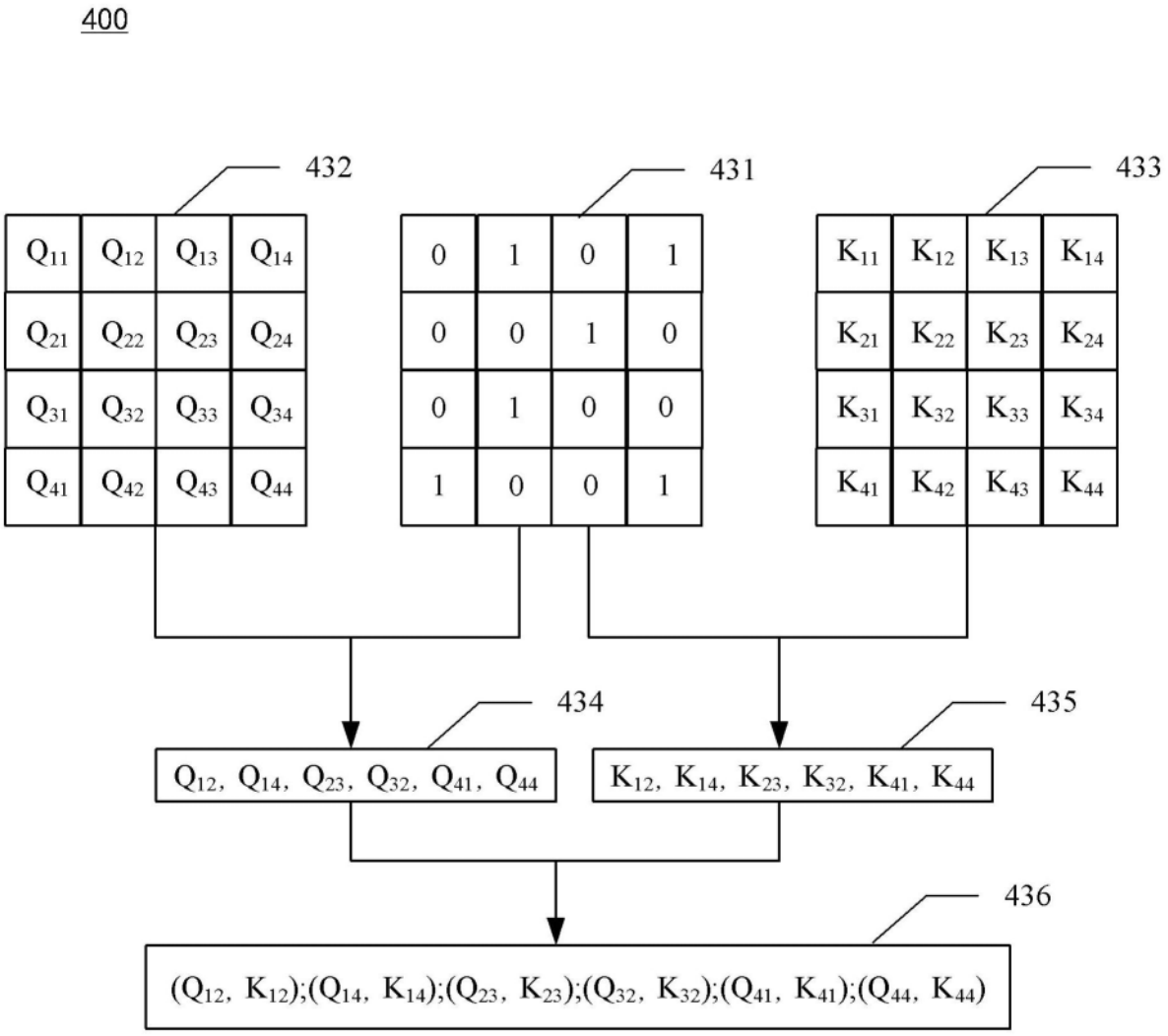

400
432
431
433
$Q_{11}$ $Q_{12}$ $Q_{13}$ $Q_{14}$
$Q_{21}$ $Q_{22}$ $Q_{23}$ $Q_{24}$
$Q_{31}$ $Q_{32}$ $Q_{33}$ $Q_{34}$
$Q_{41}$ $Q_{42}$ $Q_{43}$ $Q_{44}$
0 1 0 1
0 0 1 0
0 1 0 0
1 0 0 1
$K_{11}$ $K_{12}$ $K_{13}$ $K_{14}$
$K_{21}$ $K_{22}$ $K_{23}$ $K_{24}$
$K_{31}$ $K_{32}$ $K_{33}$ $K_{34}$
$K_{41}$ $K_{42}$ $K_{43}$ $K_{44}$
434
435
$Q_{12}$, $Q_{14}$, $Q_{23}$, $Q_{32}$, $Q_{41}$, $Q_{44}$
$K_{12}$, $K_{14}$, $K_{23}$, $K_{32}$, $K_{41}$, $K_{44}$
436
($Q_{12}$, $K_{12}$);($Q_{14}$, $K_{14}$);($Q_{23}$, $K_{23}$);($Q_{32}$, $K_{32}$);($Q_{41}$, $K_{41}$);($Q_{44}$, $K_{44}$)

图4

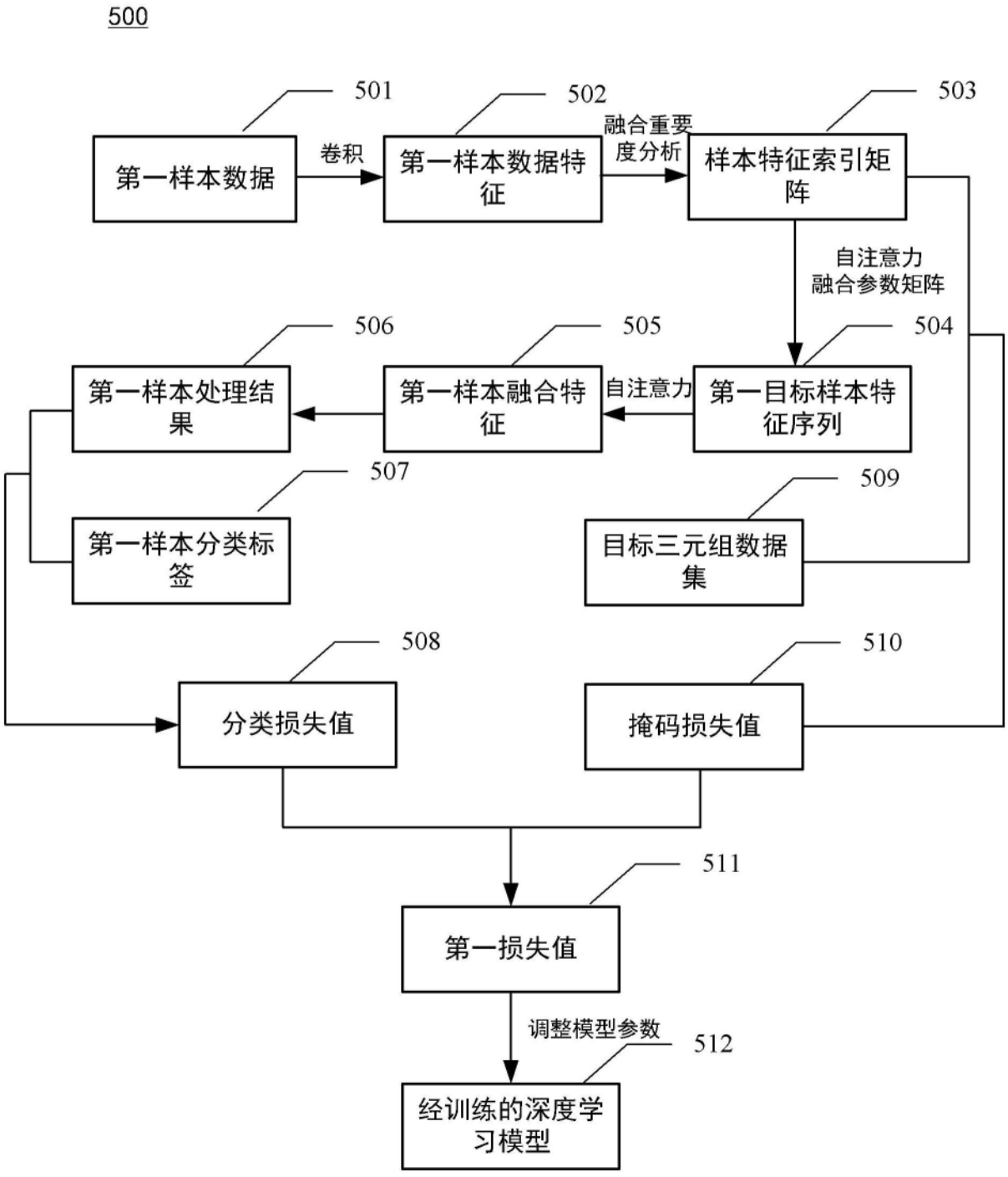

500
501
第一样本数据
卷积
502
第一样本数据特征
融合重要度分析
503
样本特征索引矩阵
自注意力融合参数矩阵
504
第一目标样本特征序列
自注意力
505
第一样本融合特征
506
第一样本处理结果
507
第一样本分类标签
509
目标三元组数据集
508
分类损失值
510
掩码损失值
511
第一损失值
调整模型参数
512
经训练的深度学习模型

图5

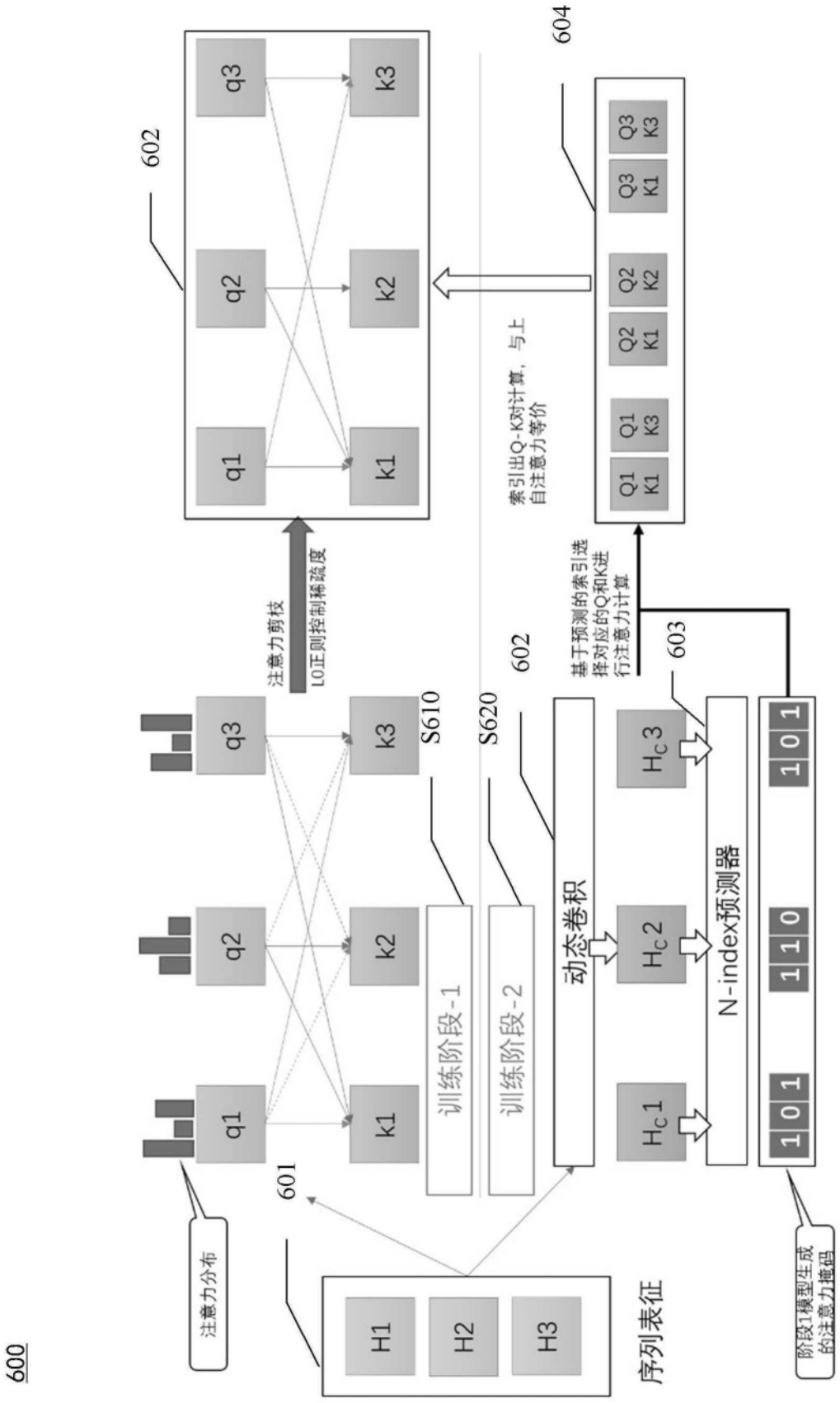

600
注意力分布
601
序列表征
H1
H2
H3
q1
q2
q3
k1
k2
k3
训练阶段-1
训练阶段-2
S610
S620
602
注意力剪枝
L0正则控制稀疏度
动态卷积
Hc1
Hc2
Hc3
603
基于预测的索引选择对应的Q和K进行注意力计算
N-index预测器
1 0 1
1 1 0
1 0 1
阶段1模型生成的注意力掩码
604
索引出Q-K对计算，与上自注意力等价
Q1 K1
Q1 K3
Q2 K1
Q2 K2
Q3 K1
Q3 K3

图6

700

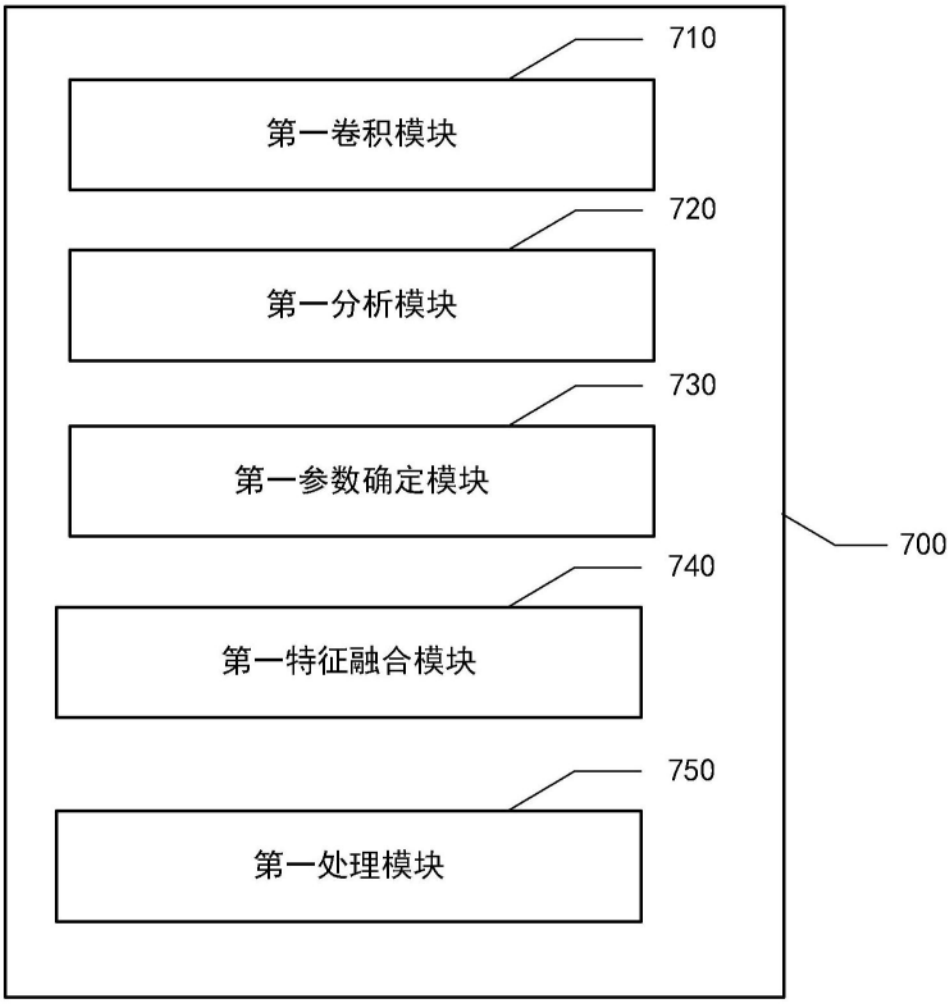

710
第一卷积模块
720
第一分析模块
730
第一参数确定模块
740
第一特征融合模块
750
第一处理模块
700

图7

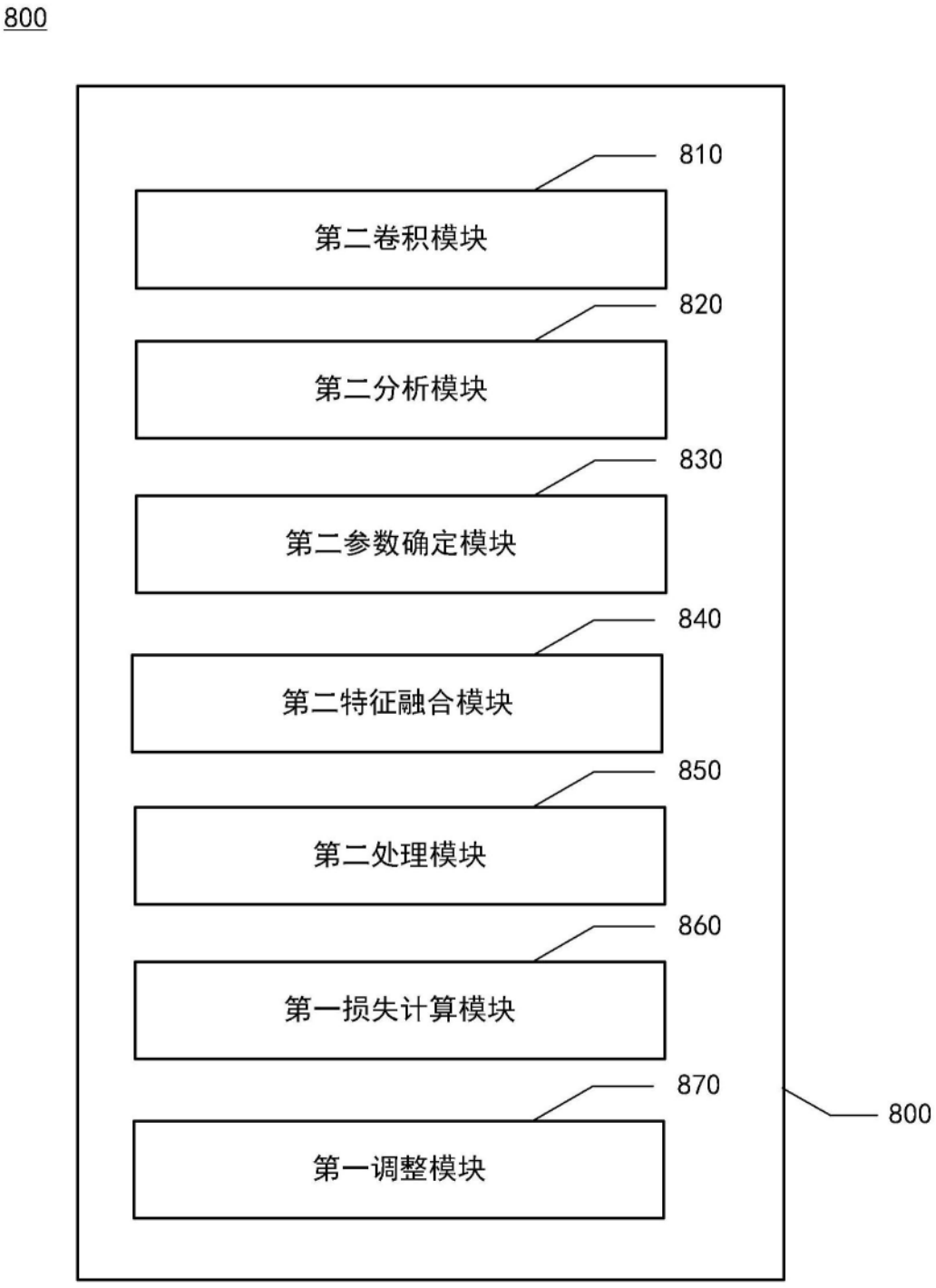

800
810
第二卷积模块
820
第二分析模块
830
第二参数确定模块
840
第二特征融合模块
850
第二处理模块
860
第一损失计算模块
870
第一调整模块
800

图8

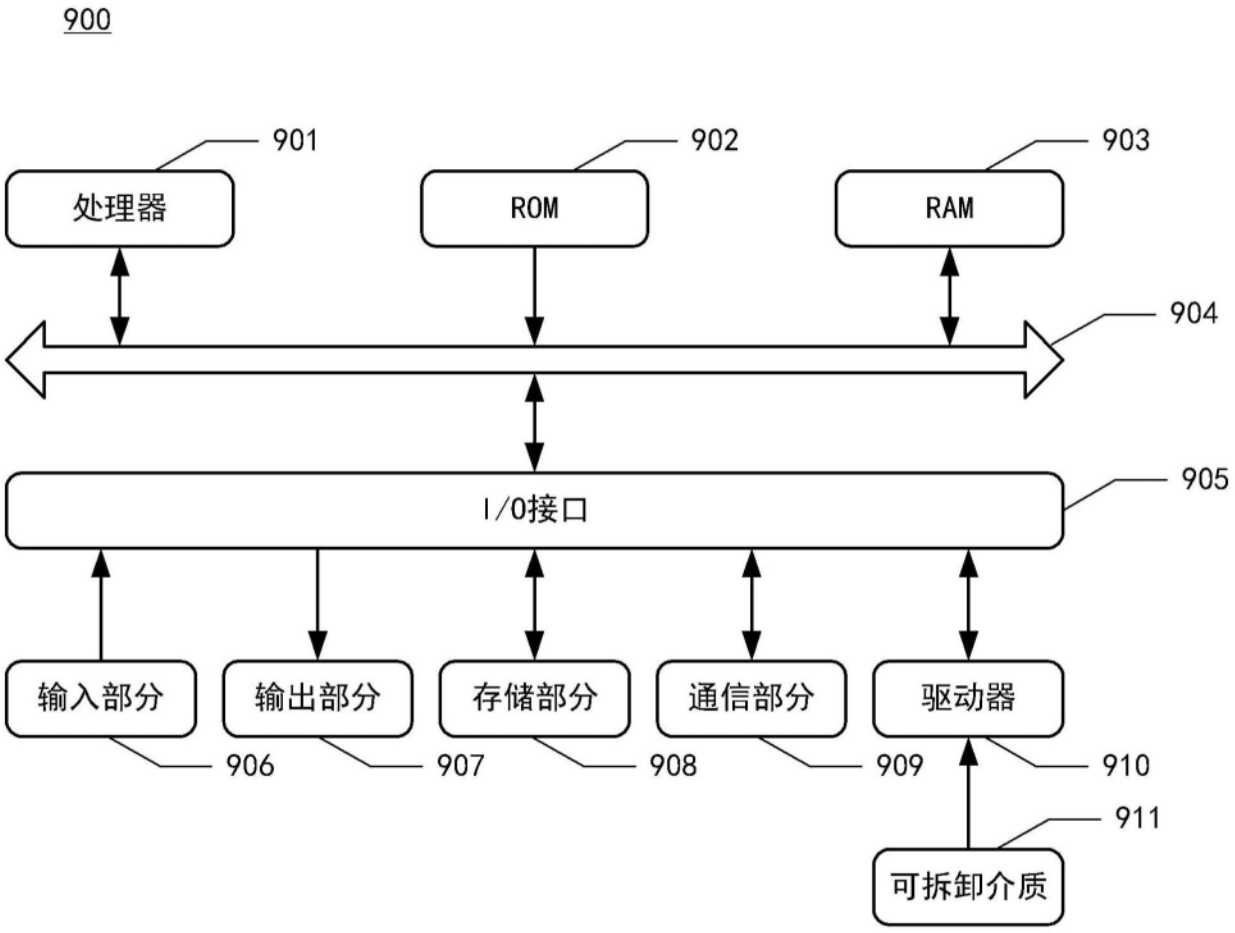

900
901
处理器
902
ROM
903
RAM
904
905
I/O接口
输入部分
输出部分
存储部分
通信部分
驱动器
906
907
908
909
910
911
可拆卸介质

图9