



(12) 发明专利

(10) 授权公告号 CN 114443899 B

(45) 授权公告日 2024.12.31

(21) 申请号 202210108236.1

G06F 16/78 (2019.01)

(22) 申请日 2022.01.28

G06V 10/764 (2022.01)

G06V 10/80 (2022.01)

(65) 同一申请的已公布的文献号

申请公布号 CN 114443899 A

(43) 申请公布日 2022.05.06

(73) 专利权人 腾讯科技(深圳)有限公司

地址 518057 广东省深圳市南山区高新区

科技中一路腾讯大厦35层

(72) 发明人 刘刚

(74) 专利代理机构 北京三高永信知识产权代理

有限责任公司 11138

专利代理师 李文静

(56) 对比文件

CN 110674350 A, 2020.01.10

CN 113010736 A, 2021.06.22

审查员 陈竹心

(51) Int. Cl.

G06F 16/75 (2019.01)

G06F 16/783 (2019.01)

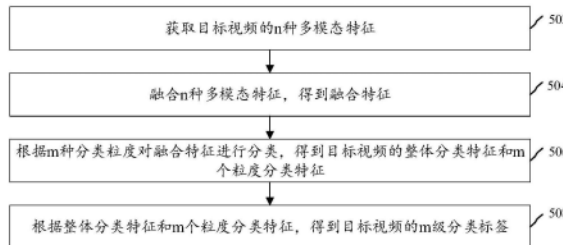
权利要求书4页 说明书22页 附图10页

(54) 发明名称

视频分类方法、装置、设备及介质

(57) 摘要

本申请公开了一种视频分类方法、装置、设备及介质,涉及机器学习领域。该方法包括:获取目标视频的 n 种多模态特征, n 种多模态特征包括图像特征、音频特征和文本特征中的至少两种, n 为大于1的正整数;融合 n 种多模态特征,得到融合特征;根据 m 种分类粒度对融合特征进行分类,得到目标视频的整体分类特征和 m 个粒度分类特征, m 种分类粒度用于表示在目标维度下 m 种相互关联的用于分类的粒度, m 个粒度分类特征用于表示与 m 种分类粒度对应的分类特征;根据整体分类特征和 m 个粒度分类特征,得到目标视频的 m 级分类标签, m 级分类标签是按照 m 种分类粒度排列的视频标签。本申请可以融合多模态的信息,使得视频标签更为准确。



1. 一种视频分类方法,其特征在于,所述方法包括:

获取目标视频的 n 种多模态特征,所述 n 种多模态特征包括图像特征、音频特征和文本特征中的至少两种, n 为大于1的正整数;

融合所述 n 种多模态特征,得到融合特征;

根据 m 种分类粒度对所述融合特征进行分类,得到所述目标视频的整体分类特征和 m 个粒度分类特征,所述 m 种分类粒度用于表示在目标维度下 m 种相互关联的用于分类的粒度,所述 m 个粒度分类特征用于表示与所述 m 种分类粒度对应的分类特征, m 为正整数,所述整体分类特征是由所述 m 个粒度分类特征中的各个特征均取出的一部分特征融合得到的;

将所述整体分类特征分别和所述 m 个粒度分类特征进行结合,得到 m 个目标分类特征;

根据所述 m 个目标分类特征确定所述目标视频的 m 级分类标签,所述 m 级分类标签是按照所述 m 种分类粒度排列的视频标签。

2. 根据权利要求1所述的方法,其特征在于,所述将所述整体分类特征分别和所述 m 个粒度分类特征进行结合,得到 m 个目标分类特征,包括:

对于所述 m 个粒度分类特征中的第 i 个粒度分类特征,确定与所述第 i 个粒度分类特征相邻的邻居粒度分类特征, i 为小于 $m+1$ 的正整数, i 的初始值为1;

调用与所述第 i 个粒度分类特征对应的标签分类器,对所述整体分类特征和所述邻居粒度分类特征进行数据处理,输出与所述第 i 个粒度分类特征对应的第 i 个目标分类特征;

将所述 i 更新为 $i+1$ 后,重复上述两个步骤,直至得到所述 m 个目标分类特征。

3. 根据权利要求1或2所述的方法,其特征在于,所述根据 m 种分类粒度对所述融合特征进行分类,得到所述目标视频的整体分类特征和 m 个粒度分类特征,包括:

调用级联的 m 个标签分类器,对所述融合特征进行数据处理,得到所述 m 个粒度分类特征;

按照所述 m 个标签分类器的级联顺序,融合所述 m 个标签分类器的隐藏层特征,得到所述整体分类特征。

4. 根据权利要求1或2所述的方法,其特征在于,所述融合所述 n 种多模态特征,得到融合特征,包括:

基于注意力机制,确定与所述 n 种多模态特征对应的 n 种权重;

根据所述 n 种权重,对所述 n 种多模态特征进行加权计算,得到所述融合特征。

5. 根据权利要求4所述的方法,其特征在于,所述 n 种多模态特征包括图像特征、音频特征和文本特征;

所述基于注意力机制,确定与所述 n 种多模态特征对应的 n 种权重,包括:

基于注意力机制,确定所述图像特征的图像权重,以及确定所述音频特征的音频权重,以及确定所述文本特征的文本权重;

所述根据所述 n 种权重,对所述 n 种多模态特征进行加权计算,得到所述融合特征,包括:

根据所述图像权重、所述音频权重和所述文本权重,对所述图像特征、所述音频特征和所述文本特征进行加权计算,输出所述融合特征。

6. 根据权利要求1或2所述的方法,其特征在于,所述方法还包括:

对所述融合特征进行压缩处理和激活处理,输出中间融合特征;

对所述中间融合特征进行校准处理,输出优化后的融合特征。

7.根据权利要求1或2所述的方法,其特征在于,所述n种多模态特征包括所述图像特征;

所述方法还包括:

从所述目标视频中抽取出目标视频帧;

调用图像特征提取网络,对所述目标视频帧进行数据处理,输出所述目标视频帧的视频帧特征;

融合所述视频帧特征,得到所述目标视频的所述图像特征。

8.根据权利要求1或2所述的方法,其特征在于,所述n种多模态特征包括所述音频特征;

所述方法还包括:

获取所述目标视频的梅尔频谱图;

调用音频特征提取网络,对所述梅尔频谱图进行数据处理,输出所述目标视频的所述音频特征。

9.根据权利要求1或2所述的方法,其特征在于,所述n种多模态特征包括所述文本特征;

所述方法还包括:

提取所述目标视频中的文本内容;

调用文本特征提取网络,对所述文本内容和所述目标视频的标题进行数据处理,输出所述目标视频的所述文本特征。

10.一种视频分类模型的训练方法,其特征在于,所述视频分类模型包括特征提取网络层、特征融合网络层和分类网络层,所述方法包括:

获取样本训练集,所述样本训练集包括样本视频和所述样本视频对应的真实标注;

调用所述特征提取网络层,对所述样本视频进行数据处理,输出n种样本多模态特征,所述n种样本多模态特征包括样本图像特征、样本音频特征和样本文本特征中的至少两种,n为大于1的正整数;

调用所述特征融合网络层,对所述n种样本多模态特征进行融合处理,输出样本融合特征;

调用所述分类网络层,分别对所述n种样本多模态特征和样本融合特征进行数据处理,分别输出n种样本粒度分类标签和样本融合分类标签;

分别计算所述n种样本粒度分类标签与所述真实标注之间的交叉熵,得到n种粒度交叉熵;计算所述样本融合分类标签与所述真实标注之间的交叉熵,得到融合交叉熵;

根据所述n种粒度交叉熵和所述融合交叉熵,对所述视频分类模型进行训练。

11.根据权利要求10所述的方法,其特征在于,所述根据所述n种粒度交叉熵和所述融合交叉熵,对所述视频分类模型进行训练,包括:

分别计算所述n种样本多模态特征和所述样本融合特征之间的相对熵,得到n种粒度相对熵;

计算所述n种粒度相对熵的和,得到相对熵损失;

以最小化所述相对熵损失为训练目标,根据所述n种粒度交叉熵和所述融合交叉熵,对

所述视频分类模型进行训练。

12. 根据权利要求11所述的方法, 其特征在于, 所述 n 种样本多模态特征包括所述样本图像特征;

所述方法还包括:

计算所述样本图像特征的概率分布和所述样本融合特征的概率分布之间的相对熵, 得到所述 n 种粒度相对熵中的图像相对熵。

13. 根据权利要求11所述的方法, 其特征在于, 所述 n 种样本多模态特征包括所述样本音频特征;

所述方法还包括:

计算所述样本音频特征的概率分布和所述样本融合特征的概率分布之间的相对熵, 得到所述 n 种粒度相对熵中的音频相对熵。

14. 根据权利要求11所述的方法, 其特征在于, 所述 n 种样本多模态特征包括所述样本文本特征;

所述方法还包括:

计算所述样本文本特征的概率分布和所述样本融合特征的概率分布之间的相对熵, 得到所述 n 种粒度相对熵中的文本相对熵。

15. 一种视频分类装置, 其特征在于, 所述装置包括:

特征提取模块, 用于获取目标视频的 n 种多模态特征, 所述 n 种多模态特征包括图像特征、音频特征和文本特征中的至少两种, n 为大于1的正整数;

融合模块, 用于融合所述 n 种多模态特征, 得到融合特征;

分类模块, 用于根据 m 种分类粒度对所述融合特征进行分类, 得到所述目标视频的整体分类特征和 m 个粒度分类特征, 所述 m 种分类粒度用于表示在目标维度下 m 种相互关联的用于分类的粒度, 所述 m 个粒度分类特征用于表示与 m 种分类粒度对应的分类特征, m 为正整数, 所述整体分类特征是由所述 m 个粒度分类特征中的各个特征均取出的一部分特征融合得到的;

所述分类模块, 还用于将所述整体分类特征分别和所述 m 个粒度分类特征进行结合, 得到 m 个目标分类特征; 根据所述 m 个目标分类特征确定所述目标视频的 m 级分类标签, 所述 m 级分类标签是按照所述 m 种分类粒度排列的视频标签。

16. 一种视频分类模型的训练装置, 其特征在于, 所述视频分类模型包括特征提取网络层、特征融合网络层和分类网络层, 所述装置包括:

样本获取模块, 用于获取样本训练集, 所述样本训练集包括样本视频和所述样本视频对应的真实标注;

样本特征提取模块, 用于调用所述特征提取网络层, 对所述样本视频进行数据处理, 输出 n 种样本多模态特征, 所述 n 种样本多模态特征包括样本图像特征、样本音频特征和样本文本特征中的至少两种, n 为大于1的正整数;

样本融合模块, 用于调用所述特征融合网络层, 对所述 n 种样本多模态特征进行融合处理, 输出样本融合特征;

样本分类模块, 用于调用所述分类网络层, 分别对所述 n 种样本多模态特征和样本融合特征进行数据处理, 分别输出 n 种样本粒度分类标签和样本融合分类标签;

训练模块,用于分别计算所述n种样本粒度分类标签与所述真实标注之间的交叉熵,得到n种粒度交叉熵;计算所述样本融合分类标签与所述真实标注之间的交叉熵,得到融合交叉熵;

所述训练模块,还用于根据所述n种粒度交叉熵和所述融合交叉熵,对所述视频分类模型进行训练。

17.一种计算机设备,其特征在于,所述计算机设备包括:处理器和存储器,所述存储器中存储有至少一段程序,所述至少一段程序由所述处理器加载并执行以实现如权利要求1至9中任一项所述的视频分类方法,或,如权利要求10至14中任一项所述的视频分类模型的训练方法。

18.一种计算机可读存储介质,其特征在于,所述计算机可读存储介质中存储有至少一条程序代码,所述程序代码由处理器加载并执行以实现如权利要求1至9中任一项所述的视频分类方法,或,如权利要求10至14中任一项所述的视频分类模型的训练方法。

19.一种计算机程序产品,包括计算机程序或指令,其特征在于,所述计算机程序或指令被处理器执行时实现权利要求1至9中任一项所述的视频分类方法,或,如权利要求10至14中任一项所述的视频分类模型的训练方法。

视频分类方法、装置、设备及介质

技术领域

[0001] 本申请涉及机器学习领域,特别涉及一种视频分类方法、装置、设备及介质。

背景技术

[0002] 在视频被上传到视频平台后,会给视频加上分类标签,比如,一个视频的分类标签是科技-智能手机-国产手机。

[0003] 相关技术会预先训练一个用于视频分类的机器学习模型。从目标视频中抽取图像特征,将图像特征输入到机器学习模型中,机器学习模型会对图像特征进行数据处理,输出目标视频的分类标签。

[0004] 但是相关技术用到了目标视频中与图像特征相关的信息,最终输出的分类标签不准确。

发明内容

[0005] 本申请实施例提供了一种视频分类方法、装置、设备及介质,该方法可以结合目标视频中的多模态特征,从多个模态来确定目标视频的分类标签,所述技术方案如下:

[0006] 根据本申请的一个方面,提供了一种视频分类方法,该方法包括:

[0007] 获取目标视频的 n 种多模态特征,所述 n 种多模态特征包括图像特征、音频特征和文本特征中的至少两种, n 为大于1的正整数;

[0008] 融合所述 n 种多模态特征,得到融合特征;

[0009] 根据 m 种分类粒度对所述融合特征进行分类,得到所述目标视频的整体分类特征和 m 个粒度分类特征,所述 m 种分类粒度用于表示在目标维度下 m 种相互关联的用于分类的粒度,所述 m 个粒度分类特征用于表示与 m 种分类粒度对应的分类特征;

[0010] 根据所述整体分类特征和所述 m 个粒度分类特征,得到所述目标视频的 m 级分类标签,所述 m 级分类标签是按照所述 m 种分类粒度排列的视频标签。

[0011] 根据本申请的另一个方面,提供了一种视频分类装置,该装置包括:

[0012] 特征提取模块,用于获取目标视频的 n 种多模态特征,所述 n 种多模态特征包括图像特征、音频特征和文本特征中的至少两种, n 为大于1的正整数;

[0013] 融合模块,用于融合所述 n 种多模态特征,得到融合特征;

[0014] 分类模块,用于根据 m 种分类粒度对所述融合特征进行分类,得到所述目标视频的整体分类特征和 m 个粒度分类特征,所述 m 种分类粒度用于表示在目标维度下 m 种相互关联的用于分类的粒度,所述 m 个粒度分类特征用于表示与 m 种分类粒度对应的分类特征, m 为正整数;

[0015] 所述分类模块,还用于根据所述整体分类特征和所述 m 个粒度分类特征,得到所述目标视频的 m 级分类标签,所述 m 级分类标签是按照所述 m 种分类粒度排列的视频标签。

[0016] 根据本申请的一个方面,提供了一种视频分类模型的训练方法,该视频分类模型包括特征提取网络层、特征融合网络层和分类网络层,该方法包括:

[0017] 获取样本训练集,所述样本训练集包括样本视频和所述样本视频对应的真实标注;

[0018] 调用所述特征提取网络层,对所述样本视频进行数据处理,输出n个样本多模态特征,所述n种样本多模态特征包括样本图像特征、样本音频特征和样本文本特征中的至少两种,n为大于1的正整数;

[0019] 调用所述特征融合网络层,对所述n个样本多模态特征进行融合处理,输出样本融合特征;

[0020] 调用所述分类网络层,分别对所述n个样本多模态特征和样本融合特征进行数据处理,分别输出n个样本粒度分类标签和样本融合分类标签;

[0021] 分别计算所述n个样本粒度分类标签与所述真实标注之间的交叉熵,得到n个粒度交叉熵;计算所述样本融合分类标签与所述真实标注之间的交叉熵,得到融合交叉熵;

[0022] 根据所述n个粒度交叉熵和所述融合交叉熵,对所述视频分类模型进行训练。

[0023] 根据本申请的一个方面,提供了一种视频分类模型的训练装置,该视频分类模型包括特征提取网络层、特征融合网络层和分类网络层,该装置包括:

[0024] 样本获取模块,用于获取样本训练集,所述样本训练集包括样本视频和所述样本视频对应的真实标注;

[0025] 样本特征提取模块,用于调用所述特征提取网络层,对所述样本视频进行数据处理,输出n个样本多模态特征,所述n种样本多模态特征包括样本图像特征、样本音频特征和样本文本特征中的至少两种,n为大于1的正整数;

[0026] 样本融合模块,用于调用所述特征融合网络层,对所述n个样本多模态特征进行融合处理,输出样本融合特征;

[0027] 样本分类模块,用于调用所述分类网络层,分别对所述n个样本多模态特征和样本融合特征进行数据处理,分别输出n个样本粒度分类标签和样本融合分类标签;

[0028] 训练模块,用于分别计算所述n个样本粒度分类标签与所述真实标注之间的交叉熵,得到n个粒度交叉熵;计算所述样本融合分类标签与所述真实标注之间的交叉熵,得到融合交叉熵;

[0029] 所述训练模块,还用于根据所述n个粒度交叉熵和所述融合交叉熵,对所述视频分类模型进行训练。

[0030] 根据本申请的另一方面,提供了一种计算机设备,该计算机设备包括:处理器和存储器,存储器中存储有至少一段程序由处理器加载并执行以实现如上方面所述的视频分类方法,或,视频分类模型的训练方法。

[0031] 根据本申请的另一方面,提供了一种计算机存储介质,计算机可读存储介质中存储有至少一条程序代码,程序代码由处理器加载并执行以实现如上方面所述的视频分类方法,或,视频分类模型的训练方法。

[0032] 根据本申请的另一方面,提供了一种计算机程序产品或计算机程序,上述计算机程序产品或计算机程序包括计算机指令,上述计算机指令存储在计算机可读存储介质中。计算机设备的处理器从上述计算机可读存储介质读取上述计算机指令,上述处理器执行上述计算机指令,使得上述计算机设备执行如上方面所述的视频分类方法,或,视频分类模型的训练方法。

[0033] 本申请实施例提供的技术方案带来的有益效果至少包括：

[0034] 该方法融合目标视频的多模态特征,得到融合特征,根据融合特征确定目标视频的分类标签。该方法能够通过多模态特征,建模挖掘更细的粒度信息并加深对视频内容的理解和提取。在对融合特征进行分类时,从整体的角度充分利用不同模态间的依赖关系和约束信息,提高视频分类结果的准确性。

附图说明

[0035] 为了更清楚地说明本申请实施例中的技术方案,下面将对实施例描述中所需要使用的附图作简单地介绍,显而易见地,下面描述中的附图仅仅是本申请的一些实施例,对于本领域普通技术人员来讲,在不付出创造性劳动的前提下,还可以根据这些附图获得其他的附图。

[0036] 图1是本申请一个示例性实施例提供的计算机系统的结构示意图；

[0037] 图2是本申请一个示例性实施例提供的计算机系统的结构示意图；

[0038] 图3是本申请一个示例性实施例提供的视频分类模型的示意图；

[0039] 图4是本申请一个示例性实施例提供的视频分类模型的示意图；

[0040] 图5是本申请一个示例性实施例提供的视频分类方法的流程示意图；

[0041] 图6是本申请一个示例性实施例提供的视频分类方法的流程示意图；

[0042] 图7是本申请一个示例性实施例提供的分类模块的结构示意图；

[0043] 图8是本申请一个示例性实施例提供的视频分类方法的流程示意图；

[0044] 图9是本申请一个示例性实施例提供的融合模块的结构示意图；

[0045] 图10是本申请一个示例性实施例提供的特征融合方法的流程示意图；

[0046] 图11是本申请一个示例性实施例提供的特征融合方法的流程示意图；

[0047] 图12是本申请一个示例性实施例提供的视频分类模型的训练方法的流程示意图；

[0048] 图13是本申请一个示例性实施例提供的计算相对熵的示意图；

[0049] 图14是本申请一个示例性实施例提供的视频分类模型的训练方法的流程示意图；

[0050] 图15是本申请一个示例性实施例提供的视频推荐方法的流程示意图；

[0051] 图16是本申请一个示例性实施例提供的视频推荐方法的流程示意图；

[0052] 图17是本申请一个示例性实施例提供的视频分类装置的结构示意图；

[0053] 图18是本申请一个示例性实施例提供的视频分类模型的训练装置的结构示意图；

[0054] 图19是本申请一个示例性实施例提供的计算机设备的结构示意图。

具体实施方式

[0055] 为使本申请的目的、技术方案和优点更加清楚,下面将结合附图对本申请实施方式作进一步地详细描述。

[0056] 首先,对本申请实施例中涉及的名词进行介绍:

[0057] 人工智能(Artificial Intelligence, AI):是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说,人工智能是计算机科学的一个综合技术,它企图了解智能的实质,并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能

也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。

[0058] 人工智能技术是一门综合学科,涉及领域广泛,既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0059] 计算机视觉技术(Computer Vision,CV):计算机视觉是一门研究如何使机器“看”的科学,更进一步的说,就是指用摄影机和电脑代替人眼,使电脑处理成为更适合人眼观察或传送给仪器检测的图像。作为一个科学学科,计算机视觉研究相关的理论和技术,试图建立能够从图像或者多维数据中获取信息的人工智能系统。计算机视觉技术通常包括图像处理、图像识别、图像语义理解、图像检索、OCR(Optical Character Recognition,光学字符识别)、视频处理、视频语义理解、视频内容/行为识别、三维物体重建、3D技术、虚拟现实、增强现实、同步定位与地图构建等技术,还包括常见的人脸识别、指纹识别等生物特征识别技术。

[0060] 自然语言处理(nature Language processing,NLP):是计算机科学领域与人工智能领域中的一个重要方向。它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。自然语言处理是一门融语言学、计算机科学、数学于一体的科学。因此,这一领域的研究将涉及自然语言,即人们日常使用的语言,所以它与语言学的研究有着密切的联系。自然语言处理技术通常包括文本处理、语义理解、机器翻译、机器人问答、知识图谱等技术。

[0061] 机器学习(Machine Learning,ML):是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。

[0062] 相对熵(Relative Entropy):又被称为Kullback-Leibler(KL)散度或信息散度(information divergence),是两个概率分布间差异的非对称性度量。相对熵等价于两个概率分布的信息熵(Shannon entropy)的差值。

[0063] 交叉熵(Cross Entropy):用于度量两个概率分布间的差异性信息。交叉熵可在机器学习中作为损失函数,p表示真实标记的分布,q则为训练后的模型的预测标记分布,交叉熵可以衡量p与q的相似性。

[0064] 梅尔频谱图:将普通频谱图的频率标度(Hz,赫兹)转化为梅尔频率标度得到的频谱图。映射关系如下:梅尔频率标度 $\text{mel}(f) = 2595 * \log_{10}(1 + f/700)$,f表示以赫兹为单位的频率标度。

[0065] 随着人工智能技术研究和进步,人工智能技术在多个领域展开研究和应用,例如常见的智能家居、智能穿戴设备、虚拟助理、智能音箱、智能营销、无人驾驶、自动驾驶、无人机、机器人、智能医疗、智能客服等,相信随着技术的发展,人工智能技术将在更多的领域得到应用,并发挥越来越重要的价值。

[0066] 图1示出了本申请一个示例性实施例提供的计算机系统的结构示意图。计算机系

统100包括:终端120和服务器140。

[0067] 终端120上安装有与视频播放相关的应用程序。该应用程序可以是app(application,应用程序)中的小程序,也可以是专门的应用程序,也可以是网页客户端。终端120包括内容提供端和内容消费端中的至少一种,其中,内容提供端指提供视频的终端,内容消费端指播放视频的终端。示例性的,用户在终端120上播放视频。终端120是智能手机、平板电脑、电子书阅读器、MP3播放器、MP4播放器、膝上型便携计算机和台式计算机中的至少一种。

[0068] 终端120通过无线网络或有线网络与服务器140相连。

[0069] 服务器140可以是独立的物理服务器,也可以是多个物理服务器构成的服务器集群或者分布式系统,还可以是提供云服务、云数据库、云计算、云函数、云存储、网络服务、云通信、中间件服务、域名服务、安全服务、CDN(Content Delivery Network,内容分发网络)、以及大数据和人工智能平台等基础云计算服务的云服务器。服务器140用于为视频播放的应用程序提供后台服务,并将视频和与视频相关的信息发送到终端120上。可选地,服务器140承担主要计算工作,终端120承担次要计算工作;或者,服务器140承担次要计算工作,终端120承担主要计算工作;或者,服务器140和终端120两者采用分布式计算架构进行协同计算。

[0070] 如图2所示,在本申请的一种可选设计中,计算机系统包括内容生产端201、上下行内容接口服务器202、调度中心服务器203、内容数据库204、视频分类模型205、视频分类服务206、内容排重服务207、内容分发出服务208、内容消费端209、人工审核系统210、内容存储服务211、下载文件系统212和视频内容抽帧和音频分离213。

[0071] 内容生产端201提供的内容包括PGC(Professional Generated Content,专业生成的内容)、UGC(User Generated Content,用户原创内容)、MCN(Multi-Channel Network,多频道网络)和PUGC(Professional generated content+User Generated Content,专业用户生产内容)中的至少一种。内容生产端201通过移动端接口API(Application Programming Interface,应用程序接口)系统或后台接口API系统向上下行内容接口服务器202提供视频。可选地,内容生产端201通过与上下行内容接口服务器202的通讯,获取上下行内容接口服务器202的接口地址,通过前述的接口地址向上下行内容接口服务器202上传视频。

[0072] 上下行内容接口服务器202与内容生产端201直接通讯,将内容生产端201提供的视频的元信息写入内容数据库204,以及将视频内容提交给调度中心服务器203。视频的元信息包括视频的文件大小、封面图链接、码率、文件格式、标题、发布时间、作者中的至少一种。

[0073] 调度中心服务器203用于负责视频流转的整个调度过程。示例性的,调度中心服务器203用于通过上下行内容接口服务器202接收入库的内容,从内容数据库204中获取视频的元信息。示例性的,调度中心服务器203还用于调度人工审核系统210和其它处理系统,控制调度的顺序和优先级。示例性的,调度中心服务器203还用于调用内容排重服务207,过滤掉不必要的重复相似内容,对于没有达到重复过滤的内容,本系统会输出内容相似度和相似关系链,供系统打散使用。示例性的,调度中心服务器203还用通调用人工审核系统210的内容启用通过内容出口分发服务208(比如,推荐引擎或者搜索引擎或者运营商)将视频推

荐给内容消费者209。示例性的,调度中心服务器203还用于负责和视频分类服务通讯,完成视频的多级分类和标记。

[0074] 内容数据库204存储有视频的元信息。示例性的,上下行内容接口服务器202在接收到视频后,会对视频进行转码操作,转码完成后,将视频的元信息保存到内容数据库204中。可选地,内容数据库204还存储有人工审核系统210的人工审核结果。可选地,内容数据库204还存储有内容排重服务207的排重结果。

[0075] 视频分类模型205用于利用视频的多模态特征生成不同粒度下的视频标签。示例性的,视频的标签为“科技-智能手机-国产手机”。

[0076] 视频分类服务206用于将视频分类模型205服务化,以便实现和调度中心服务器203之间的通讯。

[0077] 内容排重服务207用于排除相似的上传视频。可选地,内容排重服务207将上传视频向量化,得到上传视频向量;通过向量索引来比较上传视频向量和已有视频向量之间的相似程度。内容排重服务207可以实现去重服务的工程化和并行化,避免重复的视频被启用。

[0078] 内容分发出服务208用于向内容消费端209提供视频的索引信息,以便内容消费端208下载视频的流媒体信息,通过视频播放器播放视频。

[0079] 内容消费端209用于从内容分发出服务208处获取视频。可选地,内容消费端209根据内容分发出服务208获取不同类别的视频和不同视频频道的展示内容。可选的,内容消费端209直接从内容存储服务211中下载视频。

[0080] 人工审核系统210用于对视频的内容进行二次审核。示例性的,人工审核系统210是一个基于web(网络)数据库开发的系统。由于视频分类模型205可能会出现错误,通过进行二次的人工审核处理,提升视频本身分类标签的准确性和效率。

[0081] 内容存储服务211用于存储内容生产端201上传的视频。可选地,内容存储服务211包括对内服务的数据源和对外服务的数据源,对内服务的数据源和对外服务的数据源是分开部署的,避免相互影响。可选地,内容存储服务211通过CDN(全称Content Delivery Network,内容分发网络)加速服务器进行分布式缓存来实现加速。

[0082] 下载文件系统212用于控制视频下载的速度和进度。可选地,下载文件系统是一组并行的服务器,包括相关的任务调度集群和分发集群构成。

[0083] 视频内容抽帧和音频分离213用于对视频进行采样抽帧以及音频信息的采用和分离。示例性的,当下载文件系统212下载的视频提供给视频分类模型205时,视频内容抽帧和音频分离213获取视频的关键帧或者调用TSN(Time Sensitive Network,时间敏感型网络)网络进行采样。

[0084] 图3示出了本申请一个示例性实施例提供的视频分类模型的结构示意图。视频分类模型300包括特征提取模块301、特征融合模块302和分类模块303。

[0085] 特征提取模块301用于提取目标视频的多模态特征。特征提取模块301包括第一特征特征提取网络、第二特征特征提取网络、第三特征特征提取网络、……、第n特征特征提取网络,共n个特征提取网络层,每个特征提取网络层可以提取目标视频的不同模态的特征。例如,第一特征特征提取网络用于提取目标视频的图像特征,第二特征特征提取网络用于提取目标视频的音频特征。

[0086] 特征融合模块302用于融合目标视频的多模态特征。特征融合模块302的输入是特征提取模块301输出的多模态特征,输出是融合特征305。在融合目标视频的多模态特征,得到融合特征305后。可选地,调用优化模块304,对融合特征304依次进行压缩处理、激活处理和上下文选通处理,得到融合特征305。

[0087] 分类模块303用于根据融合特征305确定目标视频在不同粒度下的标签。分类模块303的输入是融合特征305,输出是目标视频的不同粒度下的标签。可选地,以分类标签输出三种粒度的标签为例进行说明,分类模块303包括级联的一级标签分类器306、二级标签分类器307和三级标签分类器308,其中,一级标签分类器306、二级标签分类器307和三级标签分类器308输出的分类标签按照粒度大小从大到小排列。调用一级标签分类器306对融合特征305进行数据处理,输出一级类分类特征,根据一级类分类特征可确定出一级类标签。调用二级标签分类器307对融合特征305和上述一级类分类特征进行数据处理,输出二级类分类特征,根据二级类分类特征可确定出二级类标签。调用三级标签分类器308对融合特征305、二级类分类特征和二级标签分类器307的隐藏层特征进行数据处理,输出三级类分类特征,根据三级类分类特征可确定出三级类标签。示例性的,分类模块205输出三种粒度的分类标签是“科技-手机-国产手机”。

[0088] 在接下来的实施例中,以提取图像特征、音频特征和文本特征为例进行说明,图4示出了本申请一个示例性实施例提供的视频分类模型的结构示意图。视频分类模型400包括图像特征提取模块401、音频特征提取模块402、文本特征提取模块403、特征融合模块302和分类模块303。

[0089] 图像特征提取模块401用于提取整个目标视频的图像特征。图像特征提取模块的输入是目标视频404,输出是图像特征407。在一种可选的设计中,通过稀疏采样操作,从目标视频404中抽取出目标视频帧;调用图像特征提取网络层405,对目标视频帧进行数据处理,输出与目标视频帧对应的视频帧特征;调用图像特征融合网络层406,从图像特征的角度,融合视频帧特征,得到目标视频的图像特征407。示例性的,通过TSN网络,从目标视频404中抽取出目标视频帧。示例性的,图像特征提取网络层405是ResNet(残差网络)或者Xception(Inception网络的一种)。示例性的,图像特征融合网络层308是NeXtVLAD网络(一种改良后的VLAD网络,增加了VLAD网络中VLAD层的非线性参数,VLAD指Vector of Local Aggregated Descriptors,局部聚合描述向量),NeXtVLAD网络可以将高维的视频帧特征分解为一组低维的视频帧特征,将低维的视频帧特征聚合低维的图像特征。

[0090] 音频特征提取模块402用于提取整个目标视频的音频特征。音频特征提取模块的输入是目标视频的音频的频谱图408,输出是音频特征412。在一种可选的设计中,将目标视频的频谱图408转换为梅尔频谱图409;调用音频特征提取网络,对梅尔频谱图409进行数据处理,输出目标视频的音频特征412,由于梅尔频谱图409包括互不重叠的不同频带的频谱,因此,这里输出的音频特征是包括多个频带的音频特征,所以还需要调用音频特征融合网络411,融合多个频带的音频特征,得到最终的音频特征412。示例性的,音频特征提取网络312是Vggish网络。示例性的,音频特征融合网络411是NeXtVLAD网络,类似的,在音频特征融合网络411是NeXtVLAD网络的情况下,音频特征提取模块402中的音频特征融合网络层411也会将不同频带下高维的音频特征分解为一组低维的特征,将上述低维的特征聚合永远忽代表整个目标视频的低维的音频特征。可选地,选取目标视频前10分钟的16kHz的音

频,使用25ms的汉明时窗,10ms的帧移对音频进行短时傅里叶变换得到频谱图408,随后将频谱图映射到64阶梅尔滤波器组中,计算得到梅尔频谱图409。梅尔频谱图409以960ms的时长组帧,帧间没有重叠,每一帧时长10ms,每一帧包含64个梅尔频带。

[0091] 文本特征提取模块403用于提取整个目标视频的文本特征。文本特征提取模块的输入是标题和文本内容413,输出是文本特征416。在一种可选的设计中,提取目标视频中的文本内容;调用文本特征提取网络层415,对文本内容和目标视频的标题进行数据处理,输出目标视频的文本特征416。示例性的,文本特征提取网络层415是基于信息流大规模文本语料的语料训练BERT(Bidirectional Encoder Representation from Transformers,双向编码器表示)模型。在实践中,由于一些视频没有标题或标题传达的信息不足,因此需要提供目标视频的文本内容。可选地,通过OCR(Optical Character Recognition,光学字符识别)对目标视频的内容进行文本识别,得到文本内容。但是通过OCR得到的文本内容也存在着一些问题,比如:画面切换过程中OCR识别不准、固定位置OCR需要去重、口述类OCR需保留以及新闻滚屏类OCR需删除。可选地,调用文本预处理网络层414对文本内容进行预处理,这里的预处理包括但不限于是过滤单字类/纯数字/纯字母OCR、过滤相邻两帧文本位置偏移小且文本重复率高的OCR、过滤文本在屏幕底端且较小的OCR。

[0092] 需要说明的是,上述图像特征提取模块4014、音频特征提取模块402、文本特征提取模块403均属于特征提取模块,在本申请实施例中,特征提取模块的数量至少为两个,特征提取模块的数量可以更多。

[0093] 特征融合模块304用于融合目标视频的特征。特征融合模块304可以融合图像特征407、音频特征412和文本特征416,得到融合特征303。

[0094] 分类模块303用于根据融合特征303确定目标视频在不同粒度下的标签。

[0095] 图5示出了本申请一个示例性实施例提供的视频分类方法的流程示意图。该方法可由图1所示的终端120或服务器140或其他计算机设备执行,该方法包括:

[0096] 步骤502:获取目标视频的n种多模态特征。

[0097] 目标视频可以是任意视频。目标视频可以是计算机设备存储的本地视频,也可以是从网络上下载的视频,也可以是其它计算机设备提供的视频,本申请实施例对此不作限定。

[0098] 多模态特征指目标视频在不同模态下对应的特征。多模态特征包括至少两种特征。在本申请实施例中,n种多模态特征包括图像特征、音频特征和文本特征中的至少两种,n为大于1的正整数。

[0099] 在本申请实施例的一种可选设计中,使用特征提取网络层来获取多模态特征。示例性的,使用图像特征提取网络层来获取图像特征;使用音频特征提取网络层来获取音频特征;使用文本特征提取网络层来获取文本特征。

[0100] 在n种多模态特征包括图像特征的情况下,从目标视频中抽取出目标视频帧;调用图像特征提取网络,对目标视频帧进行数据处理,输出目标视频帧的视频帧特征;融合视频帧特征,得到目标视频的图像特征。目标视频帧是按照预设的抽帧策略,从目标视频中提取的视频帧。示例性的,调用TSN网络,从目标视频306中抽取出目标视频帧。可选地,图像特征提取网络是ResNet或者Xception。

[0101] 在n种多模态特征包括音频特征的情况下,获取目标视频的梅尔频谱图;调用音频

特征提取网络,对梅尔频谱图进行数据处理,输出目标视频的音频特征。可选地,调用Vggish网络,对梅尔频谱图进行数据处理,输出目标视频的音频特征。

[0102] 在n种多模态特征包括文本特征的情况下,提取目标视频中的文本内容;调用文本特征提取网络,对文本内容和目标视频的标题进行数据处理,输出目标视频的文本特征。在一种可选设计中,通过OCR对目标视频的内容进行文本识别,得到文本内容。在一种可选设计中,调用基于信息流大规模文本语料的预料训练BERT模型,对文本内容和目标视频的标题进行数据处理,输出目标视频的文本特征。

[0103] 步骤504:融合n种多模态特征,得到融合特征。

[0104] 融合特征是指包含n种多模态特征的特征。

[0105] 可选地,采用向量按位乘法的方法,来融合n种多模态特征。

[0106] 步骤506:根据m种分类粒度对融合特征进行分类,得到目标视频的整体分类特征和m个粒度分类特征。

[0107] m种分类粒度用于表示在目标维度下m种相互关联的用于分类的粒度,m为正整数。目标维度指用于对融合特征进行分类的维度,得到的m种粒度分类特征都是对应目标维度的特征。比如,得到3个粒度分类特征,3个粒度分类特征都是对应科技维度的。可选地,目标维度是根据目标视频的视频内容确定的。在一种可选地设计中,目标维度是m种分类粒度中最大的分类粒度。

[0108] m个粒度分类特征用于表示与m种分类粒度对应的分类特征。示例性的,假设有3种分类粒度,对目标视频进行分类,得到3种分类粒度特征。其中,分类粒度特征1对应“科技”,是粒度最大的特征,分类粒度特征2对应“手机”,是粒度居中的特征,分类粒度特征3对应“国产手机”,是粒度最小的特征。

[0109] 整体分类特征是包括m个粒度分类特征的特征。整体分类特征可以包括完整的m个粒度分类特征,也就可以包括m个粒度分类特征的部分特征。示例性的,整体分类特征从m个粒度分类特征的各个粒度分类特征中均取出一部分特征,融合取出的特征,得到整体分类特征。

[0110] 在一种可选设计中,调用级联的m个标签分类器,对融合特征进行数据处理,得到m个粒度分类特征。融合级联的m个标签分类器的隐藏层特征,得到整体分类特征。可选地,m个标签分类器属于全连接前馈网络。

[0111] 步骤508:根据整体分类特征和m个粒度分类特征,得到目标视频的m级分类标签。

[0112] 申请人认为,最终得到的某一粒度的分类标签应该综合考虑整个分类体系的整体分类特征,以整体分类特征为参考来确定目标视频的m级分类标签。

[0113] 在一种可选地设计中,将整体分类特征分别和m个粒度分类特征进行结合,得到m个目标分类特征;根据m个目标分类特征确定目标视频的m级分类标签。m个目标分类特征用于确定m个分类粒度下的分类标签。

[0114] 在一种可选设计中,目标分类特征包括多个候选标签的概率值,将目标分类特征中最大概率值对应的候选标签确定为分类标签。

[0115] m级分类标签是按照m种分类粒度排列的视频标签。示例性的,目标视频的3级分类标签是“生活-食品制作-川菜制作”。

[0116] 综上所述,本实施例融合目标视频的多模态特征,得到融合特征,根据融合特征确

定目标视频的分类标签。该方法能够通过多模态特征,建模挖掘更细的粒度信息并加深对视频内容的理解和提取。在对融合特征进行分类时,从整体的角度充分利用不同模态间的依赖关系和约束信息,提高视频分类结果的准确性。

[0117] 图6示出了本申请一个示例性实施例提供的视频分类方法的流程示意图。该方法可由图1所示的终端120或服务器140或其他计算机设备执行,该方法包括:

[0118] 步骤601:获取目标视频的 n 种多模态特征。

[0119] 目标视频可以是任意视频。目标视频可以是计算机设备存储的本地视频,也可以是从网络上下载的视频,也可以是其它计算机设备提供的视频,本申请实施例对此不作限定。

[0120] 多模态特征指目标视频在不同模态下对应的特征。多模态特征包括至少两种特征。

[0121] 步骤602:基于注意力机制,确定与 n 种多模态特征对应的 n 种权重。

[0122] 不同模态的特征会从不同角度刻画目标视频,不同模态对分类结果的贡献是不一致的。比如,图像特征提供的信息量较大,而音频特征提供的信息量较小。所以,需要通过注意力机制为 n 种多模态特征赋予权重,以便后续得到的融合特征更为准确。

[0123] 示例性的,基于门控注意力机制,确定与 n 种多模态特征对应的 n 种权重。该方法基于门控机制动态生成的维度特定标量多模态特征的权重。

[0124] 步骤603:根据 n 种权重,对 n 种多模态特征进行加权计算,得到融合特征。

[0125] 可选地,加权计算的方法是向量按位相乘。

[0126] 步骤604:调用级联的 m 个标签分类器,对融合特征进行数据处理,得到 m 个粒度分类特征。

[0127] m 个标签分类器用于从 m 种分类粒度对融合特征进行分类,得到代表 m 种分类粒度的 m 个粒度分类特征。在一种可选设计中, m 个标签分类器之间连接方式可参考图3或图4所示的分类模块303中标签分类器的连接方式。

[0128] m 个粒度分类特征用于表示 m 种分类粒度对应的特征。

[0129] 步骤605:按照 m 个标签分类器的级联顺序,融合 m 个标签分类器的隐藏层特征,得到整体分类特征。

[0130] 在一种可选设计中, m 个标签分类器属于全连接前馈网络。

[0131] 示例性的,3个标签分类器的级联顺序是“标签分类器1-标签分类器2-标签分类器3”,则在融合标签分类器的隐藏层特征时,依次将标签分类器1、标签分类器2和标签分类器3的隐藏层特征进行融合。

[0132] 需要说明的是,步骤604和步骤605不分先后顺序,可以先执行步骤604后执行步骤605,也可以先执行步骤605后执行步骤604,本申请实施例对此不做限定。

[0133] 步骤606:将整体分类特征分别和 m 个粒度分类特征进行结合,得到 m 个目标分类特征。

[0134] 可选地,将 m 个粒度分类特征分别拼接到整体分类特征后,得到 m 个目标分类特征。或者,将 m 个粒度分类特征分别拼接到整体分类特征前,得到 m 个目标分类特征。

[0135] 步骤607:根据 m 个目标分类特征确定目标视频的 m 级分类标签。

[0136] 在一种可选设计中,目标分类特征包括多个候选标签的概率值,将目标分类特征

中最大概率值对应的候选标签确定为分类标签。

[0137] 综上所述,本实施例融合目标视频的多模态特征,得到融合特征,根据融合特征确定目标视频的分类标签。该方法能够通过多模态特征,建模挖掘更细的粒度信息并加深对视频内容的理解和提取。在对融合特征进行分类时,从整体的角度充分利用不同模态间的依赖关系和约束信息,提高视频分类结果的准确性。

[0138] 而且,在对融合特征进行分类时,从不同粒度得到粒度分类特征,能够通过多模态层次建模挖掘更细粒度的信息并加深对视频内容的理解,充分利用整体分类特征作为约束信息,提高模型预测准确率。

[0139] 图7示出了本申请一个示例性实施例提供的分类模块的结构示意图。

[0140] 分类模块305包括标签分类器501、标签分类器502和标签分类器503,标签分类器501、标签分类器502和标签分类器503是一种三级层级由粗到细的层级体系,由标签分类器501约束标签分类器502,由标签分类器502约束标签分类器503。标签分类器501的输入是融合特征,输出是一级类标签。标签分类器501会将一级类标签提供给标签分类器502。标签分类器502的输入是一级类标签和融合特征,输出是二级类标签。标签分类器502会将自身的隐藏层特征和二级类标签提供给标签分类器503。标签分类器503的输入是标签分类器502的隐藏层特征、二级类标签和融合特征,输出是三级类标签。综上,分类模块305通过上述层次结构在标签分类器之间传递信息,以生成不同粒度的分类标签。

[0141] 但是,分类模块305存在一个问题,二级类标签受一级类标签的影响,三级类标签受一级类标签和二级类标签的影响。而一级类标签和二级类标签属于标签嵌入向量(label embedding),有明显的约束作用。在一级类标签错误的情况下,二级类标签和三级类标签必然出错。因此,对分类模块305进行优化和展开,得到分类模块500。

[0142] 分类模块500包括标签分类器501、标签分类器502和标签分类器503。但是,分类模块500内的标签分类器的连接方式与分类模块305不同。其中,将所有标签分类器的隐藏层特征进行融合,得到整体分类特征504,整体分类特征504会作为输入部分输入到每个标签分类器中,另外相邻标签分类器输出的目标分类特征也会输入到标签分类器中。而每个标签分类器的输入包括两部分:(1)整体分类特征504;(2)相邻标签分类器输出的目标分类特征。比如,标签分类器501的输入是整体分类特征504和标签分类器502输出的目标分类特征,标签分类器502的输入是整体分类特征504、标签分类器501和标签分类器503输出的目标分类特征,标签分类器503的输入是整体分类特征504和标签分类器502输出的目标分类特征。通过上述方式得到的分类标签都会包括整体分类特征504和相邻标签分类器输出的分类标签的信息,可以有效改善和提升分类标签的准确性。而且,某一级分类标签同时受整体分类特征和相邻分类标签的影响,而不是只受上一级分类标签的影响,减弱了对分类标签的约束,即使一级类标签出错,二级类标签和三级类标签也不一定会出错。

[0143] 图8示出了本申请一个示例性实施例提供的视频分类方法的流程示意图。该方法可由图1所示的终端120或服务器140或其他计算机设备执行,该方法包括:

[0144] 步骤801:对于 m 个粒度分类特征中的第 i 个粒度分类特征,确定与第 i 个粒度分类特征相邻的邻居粒度分类特征。

[0145] 其中, i 为小于 $m+1$ 的正整数, i 的初始值为1。

[0146] 示例性的,共有4个按顺序排列的粒度分类特征,那么第1个粒度分类特征对应的

邻居粒度分类特征是第2个粒度分类特征,第2个粒度分类特征对应的邻居粒度分类特征是第1个粒度分类特征和第3个粒度分类特征,第3个粒度分类特征对应的邻居粒度分类特征是第2个粒度分类特征和第3个粒度分类特征,第4个粒度分类特征对应的邻居粒度分类特征是第3个粒度分类特征。

[0147] 步骤802:调用与第i个粒度分类特征对应的标签分类器,对整体分类特征和邻居粒度分类特征进行数据处理,输出与第i个粒度分类特征对应的第i个目标分类特征。

[0148] 可选地,标签分类器是全连接前馈网络。标签分类器用于结合整体分类特征和第i个粒度分类特征的邻居粒度分类特征,得到第i个目标分类特征。

[0149] 步骤803:将i更新为i+1后,重复上述两个步骤,直至得到m个目标分类特征。

[0150] 上述步骤801至步骤802需要迭代m次,每次迭代得到一个目标分类特征,直至得到m个目标分类特征。

[0151] 综上所述,本实施例提供了一种确定目标分类特征的方法,该方法可以通过整体分类特征以及和粒度分类特征相邻的邻居粒度分类特征,得到用于分类的目标分类特征。可以有效整合不同特征之间的信息,改善和提升分类标签的准确性。减弱分类标签的约束作用,比如,在一级类标签出错的情况下,二级类标签和三级类标签也不一定会出错。

[0152] 图9示出了本申请一个示例性实施例提供的融合模块的结构示意图。以融合图像特征和音频特征这两种多模态特征为例进行说明:

[0153] 一方面,将目标视频的图像901输入到翻转白化模块903中,实现图像901的翻转白化,以便后续的特征提取流程。将翻转白化后的图像输入到图像特征提取网络层904中,得到目标视频的图像特征。另一方面,将目标视频的音频902输入在音频特征提取网络层905中,得到目标视频的音频特征。

[0154] 将目标视频的图像特征和音频特征输入到融合模块906中,以实现图像特征和音频特征的融合。可选地,融合图像特征和音频特征的方式是基于门控的融合方法、基于注意力机制的融合方法和基于张量的融合方法中的至少一种。以基于门控的融合方法为例,该融合方法会为图像特征和音频特征分配权重,计算图像特征和音频特征的加权和,得到融合特征。

[0155] 选通模块907用于减少神经网络过拟合的结构。选通模块907应用在网络的训练阶段中,可以防止网络的过拟合。在网络的应用阶段,选通模块907是非必要的。在一种可选设计中,选通模块907是dropout层。

[0156] 全连接层908和SE模块909(S表示Squeeze,压缩。E表示Excitation,激活)用于校准融合特征。在一种可选设计中,SE模块909是SE Context Gating (Squeeze&Excitation Context Gating,用于上下选通的压缩和激活)。

[0157] 分类模块910用于对校准后的融合特征进行分类,得到目标视频的分类标签。

[0158] 图10示出了本申请一个示例性实施例提供的特征融合方法的流程示意图。该方法可由图1所示的终端120或服务器140或其他计算机设备执行,该方法包括:

[0159] 步骤1001:基于注意力机制,确定与n种多模态特征对应的n种权重。

[0160] 不同模态的特征会从不同角度刻画目标视频,不同模态对分类结果的贡献是不一致的。比如,图像特征提供的信息量较大,而音频特征提供的信息量较小。所以,需要通过注意力机制为n种多模态特征赋予权重,以便后续得到的融合特征更为准确。

[0161] 可选地,基于门控注意力机制,确定与n种多模态特征对应的n种权重。该方法基于门控机制动态生成的维度特定标量多模态特征的权重。

[0162] 步骤1002:根据n种权重,对n种多模态特征进行加权计算,得到融合特征。

[0163] 可选地,加权计算的方法是向量按位相乘。

[0164] 步骤1003:对融合特征进行压缩处理和激活处理,输出中间融合特征。

[0165] 压缩处理用于对融合特征进行全局平均池化,将 $C \times H \times W$ 的融合特征转化为 $1 \times 1 \times C$ 。

[0166] 激活处理用于对压缩处理之后的融合特征做一个非线性变换。

[0167] 可选地,压缩处理和激活处理通过SE模块实现。

[0168] 步骤1004:对中间融合特征进行校准处理,输出优化后的融合特征。

[0169] 可选地,校准处理可表示为以下公式: $Y = \sigma(WX + b) * X$,其中,X是输入的中间融合特征,Y是优化后的融合特征, σ 是激活函数(比如,Sigmoid函数,S型生长曲线),W和b通过训练得到的参数。 $\sigma(WX + b)$ 为一串01向量,来对中间融合特征的激活或者抑制。这种做法的目的有两点:一是通过这种激活机制体现中间融合特征中各个模态特征的内部关联,二是能对中间融合特征做进一步校准。

[0170] 需要说明的是,步骤1003和步骤1004是可选步骤,可以执行步骤1003和步骤1004,也可以不执行步骤1003和步骤1004。

[0171] 综上所述,本实施例提供了特征融合的方法,该方法可以融合目标视频的多模态特征,并对融合特征进行优化。使得融合特征能够从不同模态刻画目标视频,使融合特征能够较为准确地代表目标视频。

[0172] 在接下来的实施例中,以目标视频的多模态特征包括图像特征、音频特征和文本特征为例进行说明,图11示出了本申请一个示例性实施例提供的视频分类模型的训练方法的流程示意图。该方法可由图1所示的终端120或服务器140或其他计算机设备执行,该方法包括:

[0173] 步骤1101:基于注意力机制,确定与图像特征对应的图像权重。

[0174] 可选地,基于门控注意力机制,确定与图像特征对应的图像权重。

[0175] 在另一种可选实施方式中,基于张量的方法,确定与图像特征对应的图像权重。

[0176] 步骤1102:基于注意力机制,确定与音频特征对应的音频权重。

[0177] 可选地,基于门控注意力机制,确定与音频特征对应的音频权重。

[0178] 在另一种可选实施方式中,基于张量的方法,确定与音频特征对应的音频权重。

[0179] 步骤1103:基于注意力机制,确定与文本特征对应的文本权重。

[0180] 可选地,基于门控注意力机制,确定与文本特征对应的文本权重。

[0181] 在另一种可选实施方式中,基于张量的方法,确定与文本特征对应的文本权重。

[0182] 需要说明的是,步骤1101、步骤1102和步骤1103在实际执行时,不分先后顺序。例如,按照步骤1101、步骤1102和步骤1103的顺序执行,或者,按照步骤1102、步骤1103和步骤1101。本申请实施例对此不作限定。

[0183] 步骤1104:根据图像权重、音频权重和文本权重,对图像特征、音频特征和文本特征进行加权计算,得到融合特征。

[0184] 可选地,加权计算的方法是向量按位相乘。

[0185] 示例性的,将图像特征记为 a ,将音频特征记为 b ,将文本特征记为 c ,图像权重为 k_1 ,音频权重记为 k_2 ,文本权重记为 k_3 ,其中, $k_1+k_2+k_3=1$,则融合特征 $d=a*k_1+b*k_2+c*k_3$ 。

[0186] 步骤1105:对融合特征进行压缩处理和激活处理,输出中间融合特征。

[0187] 压缩处理用于对融合特征进行全局平均池化,将 $C \times H \times W$ 的融合特征转化为 $1 \times 1 \times C$ 。

[0188] 激活处理用于对压缩处理之后的融合特征做一个非线性变换。

[0189] 可选地,压缩处理和激活处理通过SE模块实现。

[0190] 步骤1106:对中间融合特征进行校准处理,输出优化后的融合特征。

[0191] 需要说明的是,步骤1105和步骤1106是可选步骤,可以执行步骤1105和步骤1106,也可以不执行步骤1105和步骤1106。

[0192] 综上所述,本实施例提供了融合图像特征、音频特征和文本特征的方法,该方法可以融合目标视频的图像特征、音频特征和文本特征,并对融合特征进行优化。使得融合特征能够从图像、音频和文本三种模态刻画目标视频,使融合特征能够较为准确地代表目标视频。

[0193] 图12示出了本申请一个示例性实施例提供的视频分类模型的训练方法的流程示意图。视频分类模型包括特征提取网络层、特征融合网络层和分类网络层,该方法可由图1所示的终端120或服务器140或其他计算机设备执行,该方法包括:

[0194] 步骤1202:获取样本训练集。

[0195] 其中,样本训练集包括样本视频和样本视频对应的真实标注;

[0196] 样本训练集包括一个或多个样本视频。

[0197] 步骤1204:调用特征提取网络层,对样本视频进行数据处理,输出 n 个样本多模态特征。

[0198] n 个样本多模态特征包括至少两种样本多模态特征。示例性的, n 种样本多模态特征包括样本图像特征、样本音频特征和样本文本特征中的至少两种, n 为大于1的正整数。

[0199] 步骤1206:调用特征融合网络层,对 n 个样本多模态特征进行融合处理,输出样本融合特征。

[0200] 样本融合特征是指包含 n 种多模态特征的特征。

[0201] 可选地,采用向量按位乘法的方法,来融合 n 种样本多模态特征。

[0202] 步骤1208:调用分类网络层,分别对 n 个样本多模态特征和样本融合特征进行数据处理,分别输出 n 个样本粒度分类标签和样本融合分类标签。

[0203] 可选地,分类网络层包括 $n+1$ 个级联的标签分类器, $n+1$ 个级联的标签分类器与 n 个样本多模态特征、样本融合特征一一对应。

[0204] 步骤1210:分别计算 n 个样本粒度分类标签与真实标注之间的交叉熵,得到 n 个粒度交叉熵。

[0205] n 个粒度交叉熵用于衡量 n 个样本粒度分类标签与真实标注之间的差异性信息。

[0206] 可选地, n 个粒度交叉熵包括图像交叉熵、音频交叉熵和文本交叉熵。图像交叉熵是样本图像特征对应的交叉熵,音频交叉熵是样本音频特征对应的交叉熵,文本交叉熵是样本文本特征对应的交叉熵。

[0207] 示例性的,如图13所示,当n个样本多模态特征包括样本图像特征1301、样本音频特征1302和样本文本特征1303时,调用标签分类器1305,对样本图像特征1301进行数据处理,得到样本图像分类标签;根据样本图像分类标签与真实标注之间的交叉熵,得到图像交叉熵。调用标签分类器1306,对样本音频特征1301进行数据处理,得到样本音频分类标签;根据样本音频分类标签与真实标注之间的交叉熵,得到音频交叉熵1310。调用标签分类器1307,对样本文本特征1301进行数据处理,得到样本文本分类标签;根据样本文本分类标签与真实标注之间的交叉熵,得到文本交叉熵1311。

[0208] 步骤1212:计算样本融合分类标签与真实标注之间的交叉熵,得到融合交叉熵。

[0209] 示例性的,如图13所示,调用标签分类器1308,对样本融合特征1304进行数据处理,得到样本融合分类标签;根据样本文本分类标签与真实标注之间的交叉熵,得到融合交叉熵。

[0210] 需要说明的是,步骤1210和步骤1212在执行时,不分先后顺序。可以先执行步骤1210,后执行步骤1212。也可以先执行步骤1212,后执行步骤1210。

[0211] 步骤1214:根据n个粒度交叉熵和融合交叉熵,对视频分类模型进行训练。

[0212] 在训练的过程中,如果只看一个模态的训练,则视频分类模型可能拟合到容易收敛的模态,从而影响网络参数的学习。为了避免这种情况,引入相对熵来监督视频分类模型的学习。此时,该步骤包括以下子步骤:

[0213] 1、分别计算n个样本多模态特征和样本融合特征的相对熵,得到n个粒度相对熵。

[0214] 可选地,n个粒度相对熵包括图像交叉熵、音频交叉熵和文本交叉熵中的至少两种。n个粒度相对熵用于描述n个粒度交叉熵和融合交叉熵之间的信息熵差值。

[0215] 示例性的,当n个样本多模态特征包括样本图像特征1301、样本音频特征1302和样本文本特征1303时,计算样本图像特征的概率分布和样本融合特征的概率分布之间的相对熵,得到n个粒度相对熵中的图像相对熵1312。以及计算样本音频特征的概率分布和样本融合特征的概率分布之间的相对熵,得到n个粒度相对熵中的音频相对熵1313。以及计算样本文本特征的概率分布和样本融合特征的概率分布之间的相对熵,得到n个粒度相对熵中的文本相对熵1314。

[0216] 2、计算n个粒度相对熵的和,得到相对熵损失。

[0217] 可选地,直接计算n个粒度相对熵的和,得到相对熵损失。或者,为n个粒度相对熵赋予权重;加权计算n个粒度相对熵的和,得到相对熵损失。

[0218] 3、以最小化相对熵损失为训练目标,根据n个粒度交叉熵和融合交叉熵,对视频分类模型进行训练。

[0219] 可选地,以最小化相对熵损失为训练目标,根据n个粒度交叉熵和融合交叉熵,按照误差反向传播算法调整视频分类模型的模型参数,对视频分类模型进行训练。

[0220] 综上所述,本实施例对视频分类模型进行训练。可以有效提高模型准确率,加快模型收敛速度。而且引入相对熵可以加强多个模态的协同作用。

[0221] 在接下来的实施例中,以n个样本多模态特征包括样本图像特征、样本音频特征和样本文本特征为例进行说明,图14示出了本申请一个示例性实施例提供的视频分类模型的训练方法的流程示意图。视频分类模型包括特征提取网络层、特征融合网络层和分类网络层,该方法可由图1所示的终端120或服务器140或其他计算机设备执行,该方法包括:

- [0222] 步骤1401:获取样本训练集。
- [0223] 其中,样本训练集包括样本视频和样本视频对应的真实标注;
- [0224] 样本训练集包括一个或多个样本视频。
- [0225] 步骤1402:调用图像特征提取网络层,对样本视频进行数据处理,输出样本图像特征。
- [0226] 可选地,从样本视频中抽取出样本目标视频帧;调用图像特征提取网络,对样本目标视频帧进行数据处理,输出样本样本视频帧的样本视频帧特征;融合样本视频帧特征,得到样本视频的样本图像特征。
- [0227] 步骤1403:调用音频特征提取网络层,对样本视频进行数据处理,输出样本音频特征。
- [0228] 可选地,获取样本视频的梅尔频谱图;调用音频特征提取网络,对梅尔频谱图进行数据处理,输出样本视频的样音频特征。
- [0229] 步骤1404:调用文本特征提取网络层,对样本视频进行数据处理,输出样本文本特征。
- [0230] 可选地,提取样本视频中的文本内容;调用文本特征提取网络,对文本内容和样本视频的标题进行数据处理,输出样本视频的样本文本特征。
- [0231] 需要说明的是,步骤1402、步骤1403和步骤1404在执行时,不分先后顺序。
- [0232] 步骤1405:调用特征融合网络层,对样本图像特征、样本音频特征和样本文本特征进行融合处理,输出样本融合特征。
- [0233] 样本融合特征是指包含样本图像特征、样本音频特征和样本文本特征的特征。
- [0234] 可选地,采用向量按位乘法的方法,来融合样本图像特征、样本音频特征和样本文本特征。
- [0235] 步骤1406:调用分类网络层,分别对样本图像特征、样本音频特征、样本文本特征和样本融合特征进行数据处理,分别输出样本图像分类标签、样本音频分类标签、样本文本分类标签和样本融合分类标签。
- [0236] 可选地,分类网络层包括4个级联的标签分类器,4个级联的标签分类器与样本图像特征、样本音频特征、样本文本特征和样本融合特征一一对应。
- [0237] 步骤1407:计算样本图像分类标签与真实标注之间的交叉熵,得到图像交叉熵。
- [0238] 示例性的,如图13所示,调用标签分类器1305,对样本图像特征1301进行数据处理,得到样本图像分类标签;根据样本图像分类标签与真实标注之间的交叉熵,得到图像交叉熵。
- [0239] 步骤1408:计算样本音频分类标签与真实标注之间的交叉熵,得到音频交叉熵。
- [0240] 示例性的,如图13所示,调用标签分类器1306,对样本音频特征1301进行数据处理,得到样本音频分类标签;根据样本音频分类标签与真实标注之间的交叉熵,得到音频交叉熵1310。
- [0241] 步骤1409:计算样本文本分类标签与真实标注之间的交叉熵,得到文本交叉熵。
- [0242] 示例性的,如图13所示,调用标签分类器1307,对样本文本特征1301进行数据处理,得到样本文本分类标签;根据样本文本分类标签与真实标注之间的交叉熵,得到文本交叉熵1311。

[0243] 步骤1410:计算样本融合分类标签与真实标注之间的交叉熵,得到融合交叉熵。

[0244] 示例性的,如图13所示,调用标签分类器1308,对样本融合特征1304进行数据处理,得到样本融合分类标签;根据样本文本分类标签与真实标注之间的交叉熵,得到融合交叉熵。

[0245] 需要说明的是,步骤1407、步骤1408、步骤1409和步骤1410在执行时,不分先后顺序。

[0246] 步骤1411:根据图像交叉熵、音频交叉熵、文本交叉熵和融合交叉熵,对视频分类模型进行训练。

[0247] 在本申请的一种可选设计中,引入相对熵来监督视频分类模型的学习:计算样本图像特征的概率分布和样本融合特征的概率分布之间的相对熵,得到 n 个粒度相对熵中的图像相对熵;计算样本音频特征的概率分布和样本融合特征的概率分布之间的相对熵,得到 n 个粒度相对熵中的音频相对熵;计算样本文本特征的概率分布和样本融合特征的概率分布之间的相对熵,得到 n 个粒度相对熵中的文本相对熵。计算图像相对熵、音频相对熵和文本相对熵的和,得到相对熵损失。以最小化相对熵损失为训练目标,根据图像交叉熵、音频交叉熵、文本交叉熵和融合交叉熵,对视频分类模型进行训练。

[0248] 综上所述,本实施例对视频分类模型进行训练,将样本图像特征、样本音频特征和样本文本特征引入视频分类模型的训练。可以有效提高模型准确率,加快模型收敛速度。而且引入相对熵可以加强图像、音频和文本模态的协同作用。

[0249] 在接下来的实施例中,根据视频的分类标签向用户推荐目标视频,图15示出了本申请一个示例性实施例提供的视频推荐方法的流程示意图。该方法包括:

[0250] 步骤1501:获取待推荐视频的 n 种多模态特征。

[0251] 待推荐视频可以是任意视频。待推荐视频可以是计算机设备存储的本地视频,也可以是从网络上下载的视频,也可以是其它计算机设备提供的视频,本申请实施例对此不作限定。

[0252] 在本申请实施例中, n 种多模态特征包括图像特征、音频特征和文本特征中的至少两种, n 为大于1的正整数。

[0253] 步骤1502:融合 n 种多模态特征,得到融合特征。

[0254] 可选地,采用向量按位乘法的方法,来融合 n 种多模态特征。

[0255] 步骤1503:根据 m 种分类粒度对融合特征进行分类,得到待推荐视频的整体分类特征和 m 个粒度分类特征。

[0256] m 种分类粒度用于表示在目标维度下 m 种相互关联的用于分类的粒度, m 为正整数。目标维度指用于对融合特征进行分类的维度,得到的 m 种粒度分类特征都是对应目标维度的特征。

[0257] 整体分类特征是包括 m 个粒度分类特征的特征。整体分类特征可以包括完整的 m 个粒度分类特征,也就可以包括 m 个粒度分类特征的部分特征。

[0258] 步骤1504:根据整体分类特征和 m 个粒度分类特征,得到待推荐视频的 m 级分类标签。

[0259] 在一种可选地设计中,将整体分类特征分别和 m 个粒度分类特征进行结合,得到 m 个目标分类特征;根据 m 个目标分类特征确定目标视频的 m 级分类标签。 m 个目标分类特征用

于确定 m 个分类粒度下的分类标签。

[0260] 步骤1505:在目标用户帐号的推荐标签与 m 级分类标签匹配的情况下,向目标用户帐号推荐待推荐视频。

[0261] 可选地,目标用户帐号的推荐标签是根据目标用户帐号的历史浏览记录确定的。或者,目标用户帐号的推荐标签是由用户执行设置的。

[0262] 可选地,在目标用户帐号的推荐标签与 m 级分类标签完全相同的情况下,向用户帐号推荐待推荐视频。可选地,在目标用户帐号的推荐标签与 m 级分类标签部分相同的情况下,向用户帐号推荐待推荐视频。

[0263] 示例性的,目标用户帐号的推荐标签是“科技-手机-国产手机”,待推荐视频1的分类标签是“科技-手机-外国手机”,待推荐视频2的分类标签是“生活-美食-中华美食”,由于待推荐视频2的分类标签与目标用户帐号的推荐标签完全不同,而待推荐视频1的分类标签与目标用户帐号的推荐标签存在部分相同,因此,向目标用户帐号推荐待推荐视频1。

[0264] 综上所述,本实施例提供了向用户帐号推荐视频的方法,通过获取用户帐号的推荐,有针对性地向用户帐号推荐视频,确保推荐的视频符合用户的喜好。

[0265] 在接下来的实施例中,根据视频的分类标签向用户推荐目标视频,图16示出了本申请一个示例性实施例提供的视频推荐方法的流程示意图。

[0266] 步骤1601:获取待推荐视频的图像特征。

[0267] 待推荐视频可以是任意视频。待推荐视频可以是计算机设备存储的本地视频,也可以是从网络上下载的视频,也可以是其它计算机设备提供的视频,本申请实施例对此不作限定。

[0268] 可选地,从待推荐视频中抽取出目标视频帧;调用图像特征提取网络,对目标视频帧进行数据处理,输出目标视频帧的视频帧特征;融合视频帧特征,得到待推荐视频的图像特征。

[0269] 步骤1602:获取待推荐视频的音频特征。

[0270] 可选地,获取待推荐视频的梅尔频谱图;调用音频特征提取网络,对梅尔频谱图进行数据处理,输出待推荐视频的音频特征。可选地,调用Vggish网络,对梅尔频谱图进行数据处理,输出待推荐视频的音频特征。

[0271] 步骤1603:获取待推荐视频的文本特征。

[0272] 可选地,提取待推荐视频中的文本内容;调用文本特征提取网络,对文本内容和待推荐视频的标题进行数据处理,输出待推荐视频的文本特征。

[0273] 步骤1604:融合图像特征、音频特征和文本特征,得到融合特征。

[0274] 可选地,采用向量按位乘法的方法,来融合 n 种多模态特征。

[0275] 步骤1605:根据 m 种分类粒度对融合特征进行分类,得到待推荐视频的整体分类特征和 m 个粒度分类特征。

[0276] m 种分类粒度用于表示在目标维度下 m 种相互关联的用于分类的粒度, m 为正整数。目标维度指用于对融合特征进行分类的维度,得到的 m 种粒度分类特征都是对应目标维度的特征。

[0277] 整体分类特征是包括 m 个粒度分类特征的特征。整体分类特征可以包括完整的 m 个粒度分类特征,也就可以包括 m 个粒度分类特征的部分特征。

[0278] 步骤1606:根据整体分类特征和 m 个粒度分类特征,得到待推荐视频的 m 级分类标签。

[0279] 在一种可选地设计中,将整体分类特征分别和 m 个粒度分类特征进行结合,得到 m 个目标分类特征;根据 m 个目标分类特征确定目标视频的 m 级分类标签。 m 个目标分类特征用于确定 m 个分类粒度下的分类标签。

[0280] 步骤1605:在目标用户帐号的推荐标签与 m 级分类标签匹配的情况下,向目标用户帐号推荐待推荐视频。

[0281] 可选地,目标用户帐号的推荐标签是根据目标用户帐号的历史浏览记录确定的。或者,目标用户帐号的推荐标签是由用户执行设置的。

[0282] 可选地,在目标用户帐号的推荐标签与 m 级分类标签完全相同的情况下,向用户帐号推荐待推荐视频。可选地,在目标用户帐号的推荐标签与 m 级分类标签部分相同的情况下,向用户帐号推荐待推荐视频。

[0283] 综上所述,本实施例提供了向用户帐号推荐视频的方法,通过获取用户帐号的推荐,有针对性地向用户帐号推荐视频,确保推荐的视频符合用户的喜好。

[0284] 下述为本申请装置实施例,可以用于执行本申请方法实施例。对于本申请装置实施例中未披露的细节,请参照本申请方法实施例。

[0285] 请参考图17,其示出了本申请一个实施例提供的视频分类装置的框图。上述功能可以由硬件实现,也可以由硬件执行相应的软件实现。该装置1700包括:

[0286] 特征提取模块1701,用于获取目标视频的 n 种多模态特征,所述 n 种多模态特征包括图像特征、音频特征和文本特征中的至少两种, n 为大于1的正整数;

[0287] 融合模块1702,用于融合所述 n 种多模态特征,得到融合特征;

[0288] 分类模块1703,用于根据 m 种分类粒度对所述融合特征进行分类,得到所述目标视频的整体分类特征和 m 个粒度分类特征,所述 m 种分类粒度用于表示在目标维度下 m 种相互关联的用于分类的粒度,所述 m 个粒度分类特征用于表示与 m 种分类粒度对应的分类特征;

[0289] 所述分类模块1703,还用于根据所述整体分类特征和所述 m 个粒度分类特征,得到所述目标视频的 m 级分类标签,所述 m 级分类标签是按照所述 m 种分类粒度排列的视频标签。

[0290] 在一种可选设计中,所述分类模块1703,还用于将所述整体分类特征分别和所述 m 个粒度分类特征进行结合,得到 m 个目标分类特征;根据所述 m 个目标分类特征确定所述目标视频的所述 m 级分类标签。

[0291] 在一种可选设计中,所述分类模块1703,还用于对于所述 m 个粒度分类特征中的第 i 个粒度分类特征,确定与所述第 i 个粒度分类特征相邻的邻居粒度分类特征, i 为小于 $m+1$ 的正整数, i 的初始值为1;调用与所述第 i 个粒度分类特征对应的标签分类器,对所述整体分类特征和所述邻居粒度分类特征进行数据处理,输出与所述第 i 个粒度分类特征对应的第 i 个目标分类特征;将所述 i 更新为 $i+1$ 后,重复上述两个步骤,直至得到所述 m 个目标分类特征。

[0292] 在一种可选设计中,所述融合模块1702,还用于调用级联的 m 个标签分类器,对所述融合特征进行数据处理,得到所述 m 个粒度分类特征;按照所述 m 个标签分类器的级联顺序,融合所述 m 个标签分类器的隐藏层特征,得到所述整体分类特征。

[0293] 在一种可选设计中,所述融合模块1702,还用于基于注意力机制,确定与所述 n 种

多模态特征对应的 n 种权重;根据所述 n 种权重,对所述 n 种多模态特征进行加权计算,得到所述融合特征。

[0294] 在一种可选设计中,所述分类模块1703,还用于基于注意力机制,确定所述图像特征的图像权重,以及确定所述音频特征的音频权重,以及确定所述文本特征的文本权重;所述分类模块1703,还用于根据所述图像权重、所述音频权重和所述文本权重,对所述图像特征、所述音频特征和所述文本特征进行加权计算,输出所述融合特征。

[0295] 在一种可选设计中,所述融合模块1702,还用于对所述融合特征进行压缩处理和激活处理,输出中间融合特征;对所述中间融合特征进行校准处理,输出优化后的融合特征。

[0296] 在一种可选设计中,所述特征提取模块1701,还用于从所述目标视频中抽取出目标视频帧;调用图像特征提取网络,对所述目标视频帧进行数据处理,输出所述目标视频帧的视频帧特征;融合所述视频帧特征,得到所述目标视频的所述图像特征。

[0297] 在一种可选设计中,所述特征提取模块1701,还用于获取所述目标视频的梅尔频谱图;调用音频特征提取网络,对所述梅尔频谱图进行数据处理,输出所述目标视频的所述音频特征。

[0298] 在一种可选设计中,所述特征提取模块1701,还用于提取所述目标视频中的文本内容;调用文本特征提取网络,对所述文本内容和所述目标视频的标题进行数据处理,输出所述目标视频的所述文本特征。

[0299] 综上所述,本实施例融合目标视频的多模态特征,得到融合特征,根据融合特征确定目标视频的分类标签。该方法能够充分利用不同模态间的依赖关系和约束信息,提高视频分类结果的准确性。得到更多对视频内容进行刻画的多模态内容理解特征,包括视频的多级细粒度分类信息,助力推荐系统的内容分发;对内容的理解能够充分利用视频内容的标题文本,音频内容,视频内容作为基础,刻画更加全面和准确,提升内容分发的效率;能够通过多模态层次建模挖掘更细粒度信息加深对视频内容的理解,充分利用类别层次依赖关系和约束信息,提高模型的准确率。

[0300] 请参考图18,其示出了本申请一个实施例提供的视频分类模型的训练装置的框图。上述功能可以由硬件实现,也可以由硬件执行相应的软件实现。该装置1800包括:

[0301] 样本获取模块1801,用于获取样本训练集,所述样本训练集包括样本视频和所述样本视频对应的真实标注;

[0302] 样本特征提取模块1802,用于调用所述特征提取网络层,对所述样本视频进行数据处理,输出 n 个样本多模态特征,所述 n 种样本多模态特征包括样本图像特征、样本音频特征和样本文本特征中的至少两种, n 为大于1的正整数;

[0303] 样本融合模块1803,用于调用所述特征融合网络层,对所述 n 个样本多模态特征进行融合处理,输出样本融合特征;

[0304] 样本分类模块1804,用于调用所述分类网络层,分别对所述 n 个样本多模态特征和样本融合特征进行数据处理,分别输出 n 个样本粒度分类标签和样本融合分类标签;

[0305] 训练模块1805,用于分别计算所述 n 个样本粒度分类标签与所述真实标注之间的交叉熵,得到 n 个粒度交叉熵;计算所述样本融合分类标签与所述真实标注之间的交叉熵,得到融合交叉熵;

[0306] 所述训练模块1805,还用于根据所述n个粒度交叉熵和所述融合交叉熵,对所述视频分类模型进行训练。

[0307] 在一种可选设计中,所述训练模块1805,还用于分别计算所述n个粒度交叉熵和所述融合交叉熵之间的相对熵,得到n个粒度相对熵;计算所述n个粒度相对熵的和,得到相对熵损失;以最小化所述相对熵损失为训练目标,根据所述n个粒度交叉熵和所述融合交叉熵,对所述视频分类模型进行训练。

[0308] 在一种可选设计中,所述n个样本多模态特征包括所述样本图像特征;训练模块1805,还用于计算所述样本图像特征的概率分布和所述样本融合特征的概率分布之间的相对熵,得到所述n个粒度相对熵中的图像相对熵。

[0309] 在一种可选设计中,所述n个样本多模态特征包括所述样本音频特征;训练模块1805,还用于计算所述样本音频特征的概率分布和所述样本融合特征的概率分布之间的相对熵,得到所述n个粒度相对熵中的音频相对熵。

[0310] 在一种可选设计中,所述n个样本多模态特征包括所述样本文本特征;训练模块1805,还用于计算所述样本文本特征的概率分布和所述样本融合特征的概率分布之间的相对熵,得到所述n个粒度相对熵中的文本相对熵。

[0311] 综上所述,本实施例对视频分类模型进行训练。可以有效提高模型准确率,加快模型收敛速度。而且引入相对熵可以加强多个模态的协同作用。

[0312] 图19是根据一示例性实施例示出的一种计算机设备的结构示意图。所述计算机设备1900包括中央处理单元(Central Processing Unit,CPU)1901、包括随机存取存储器(Random Access Memory,RAM)1902和只读存储器(Read-Only Memory,ROM)1903的系统存储器1904,以及连接系统存储器1904和中央处理单元1901的系统总线1905。所述计算机设备1900还包括帮助计算机设备内的各个器件之间传输信息的基本输入/输出系统(Input/Output,I/O系统)1906,和用于存储操作系统1913、应用程序1914和其他程序模块1915的大容量存储设备1907。

[0313] 所述基本输入/输出系统1906包括有用于显示信息的显示器1908和用于用户输入信息的诸如鼠标、键盘之类的输入设备1909。其中所述显示器1908和输入设备1909都通过连接到系统总线1905的输入输出控制器1910连接到中央处理单元1901。所述基本输入/输出系统1906还可以包括输入输出控制器1910以用于接收和处理来自键盘、鼠标、或电子触控笔等多个其他设备的输入。类似地,输入输出控制器1910还提供输出到显示屏、打印机或其他类型的输出设备。

[0314] 所述大容量存储设备1907通过连接到系统总线1905的大容量存储控制器(未示出)连接到中央处理单元1901。所述大容量存储设备1907及其相关联的计算机设备可读介质为计算机设备1900提供非易失性存储。也就是说,所述大容量存储设备1907可以包括诸如硬盘或者只读光盘(Compact Disc Read-Only Memory,CD-ROM)驱动器之类的计算机设备可读介质(未示出)。

[0315] 不失一般性,所述计算机设备可读介质可以包括计算机设备存储介质和通信介质。计算机设备存储介质包括以用于存储诸如计算机设备可读指令、数据结构、程序模块或其他数据等信息的任何方法或技术实现的易失性和非易失性、可移动和不可移动介质。计算机设备存储介质包括RAM、ROM、可擦除可编程只读存储器(Erasable Programmable Read

Only Memory, EPROM)、带电可擦可编程只读存储器 (Electrically Erasable Programmable Read-Only Memory, EEPROM)、CD-ROM、数字视频光盘 (Digital Video Disc, DVD) 或其他光学存储、磁带盒、磁带、磁盘存储或其他磁性存储设备。当然,本领域技术人员可知所述计算机设备存储介质不局限于上述几种。上述的系统存储器1904和大容量存储设备1907可以统称为存储器。

[0316] 根据本公开的各种实施例,所述计算机设备1900还可以通过诸如因特网等网络连接到网络上的远程计算机设备运行。也即计算机设备1900可以通过连接在所述系统总线1905上的网络接口单元1912连接到网络1911,或者说,也可以使用网络接口单元1912来连接到其他类型的网络或远程计算机设备系统(未示出)。

[0317] 所述存储器还包括一个或者一个以上的程序,所述一个或者一个以上程序存储于存储器中,中央处理器1901通过执行该一个或一个以上程序来实现上述视频分类方法,或,视频分类模型的训练方法的全部或者部分步骤。

[0318] 在示例性实施例中,还提供了一种计算机可读存储介质,所述计算机可读存储介质中存储有至少一条指令、至少一段程序、代码集或指令集,所述至少一条指令、所述至少一段程序、所述代码集或指令集由处理器加载并执行以实现上述各个方法实施例提供的视频分类方法,或,视频分类模型的训练方法。

[0319] 本申请还提供一种计算机可读存储介质,所述存储介质中存储有至少一条指令、至少一段程序、代码集或指令集,所述至少一条指令、所述至少一段程序、所述代码集或指令集由所述处理器加载并执行以实现上述方法实施例提供的视频分类方法,或,视频分类模型的训练方法。

[0320] 本申请还提供一种计算机程序产品或计算机程序,上述计算机程序产品或计算机程序包括计算机指令,上述计算机指令存储在计算机可读存储介质中。计算机设备的处理器从上述计算机可读存储介质读取上述计算机指令,上述处理器执行上述计算机指令,使得上述计算机设备执行如上方面实施例提供的视频分类方法,或,视频分类模型的训练方法。

[0321] 上述本申请实施例序号仅仅为了描述,不代表实施例的优劣。

[0322] 本领域普通技术人员可以理解实现上述实施例的全部或部分步骤可以通过硬件来完成,也可以通过程序来指令相关的硬件完成,所述的程序可以存储于一种计算机可读存储介质中,上述提到的存储介质可以是只读存储器,磁盘或光盘等。

[0323] 以上所述仅为本申请的可选实施例,并不用以限制本申请,凡在本申请的精神和原则之内,所作的任何修改、等同替换、改进等,均应包含在本申请的保护范围之内。

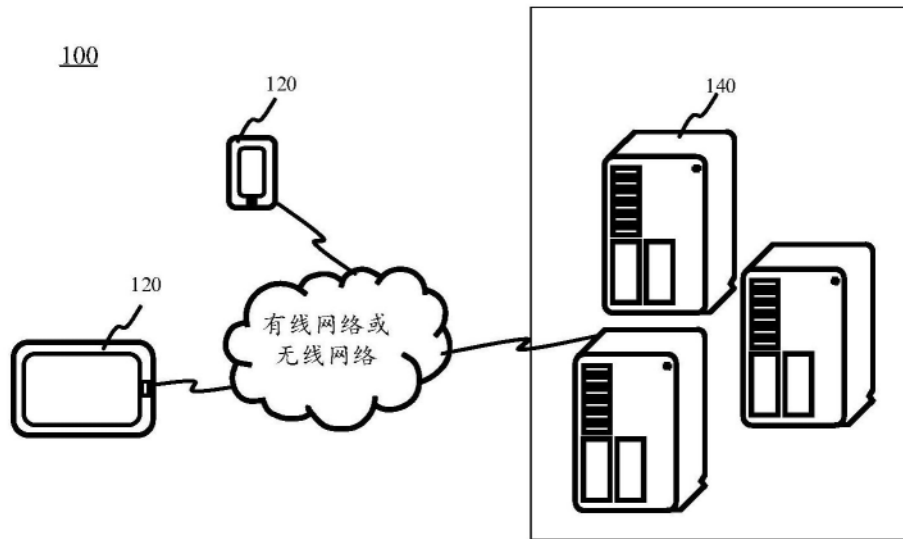


图1

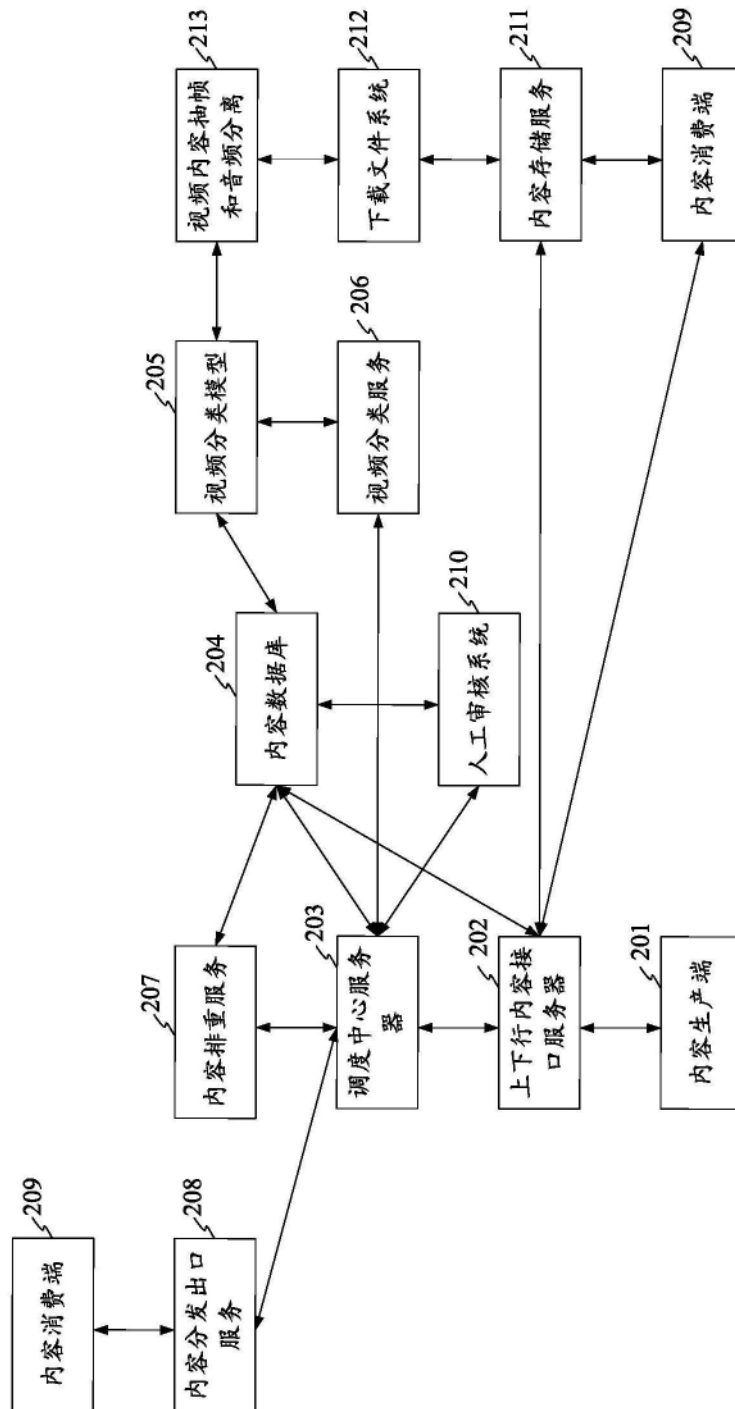


图2

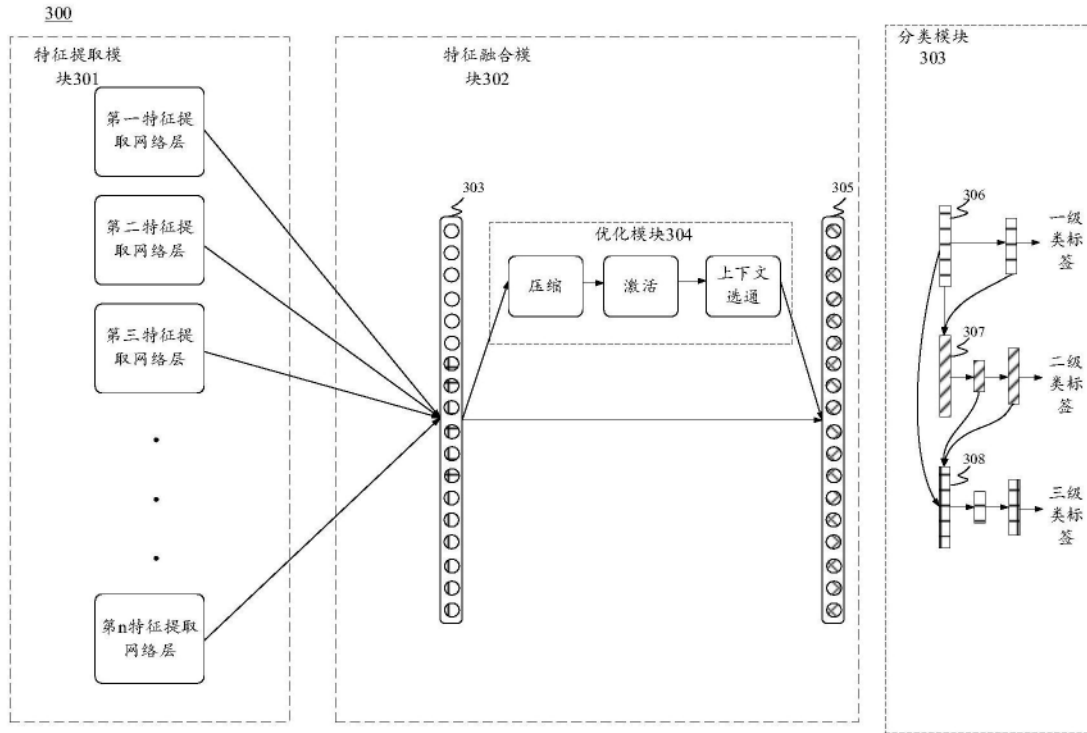


图3

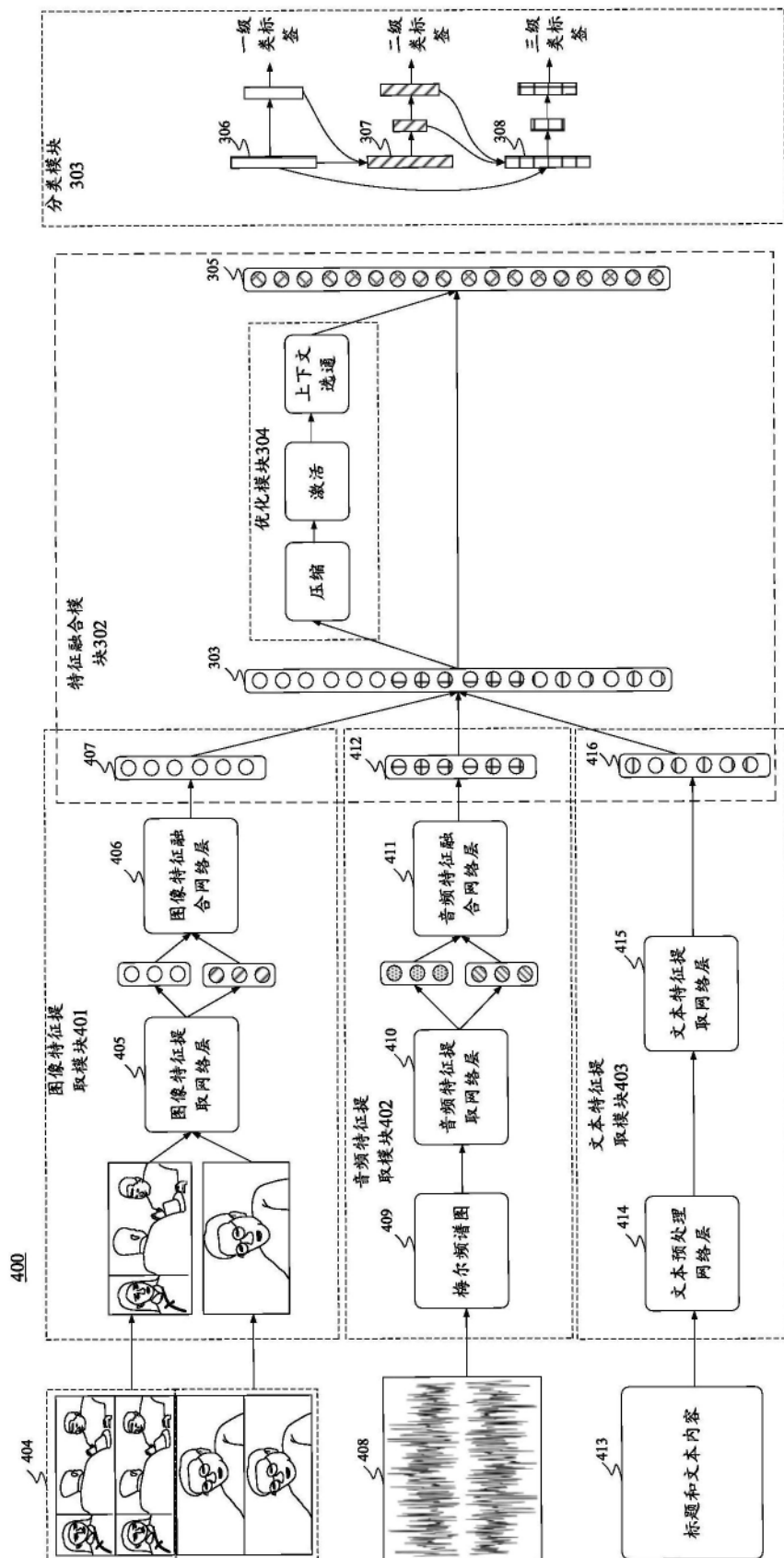


图4

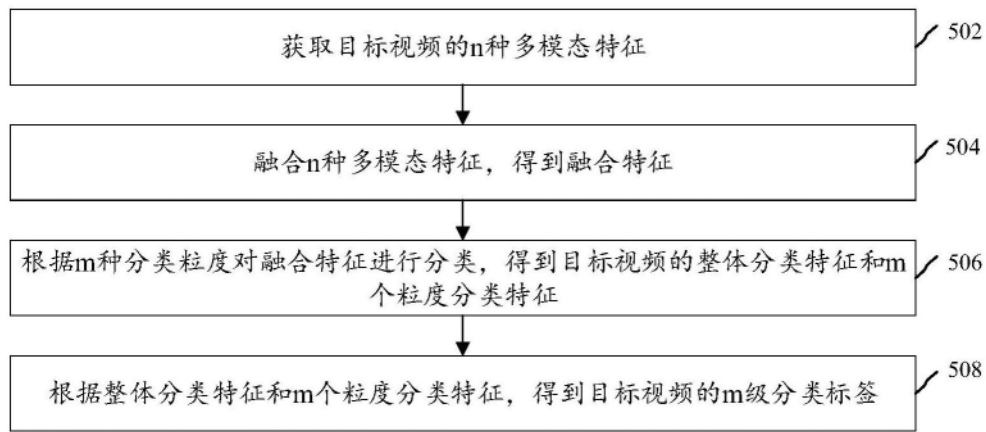


图5

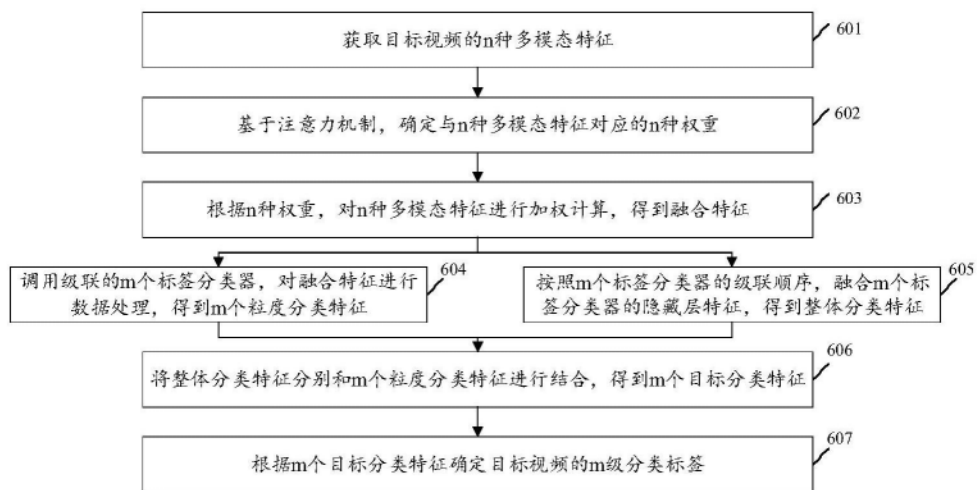


图6

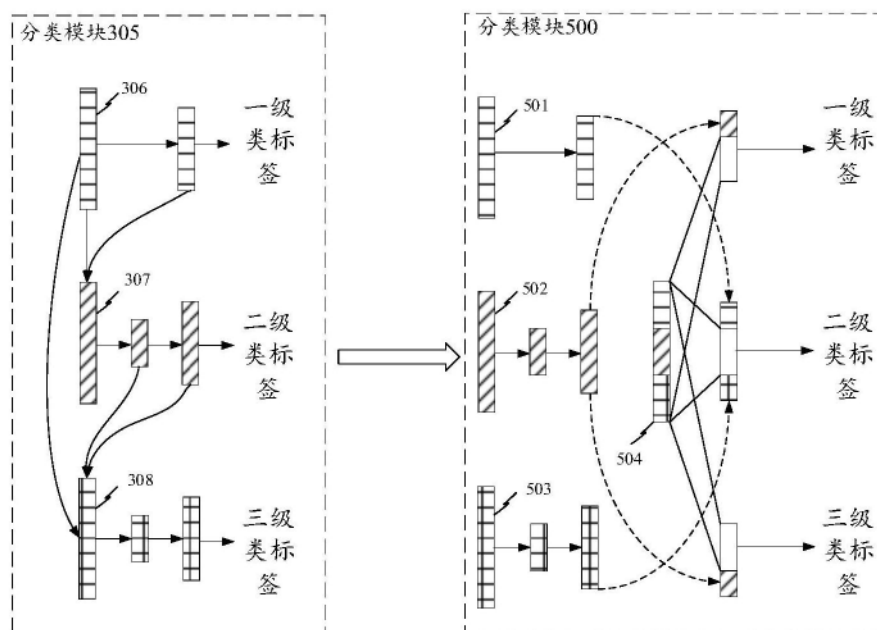


图7

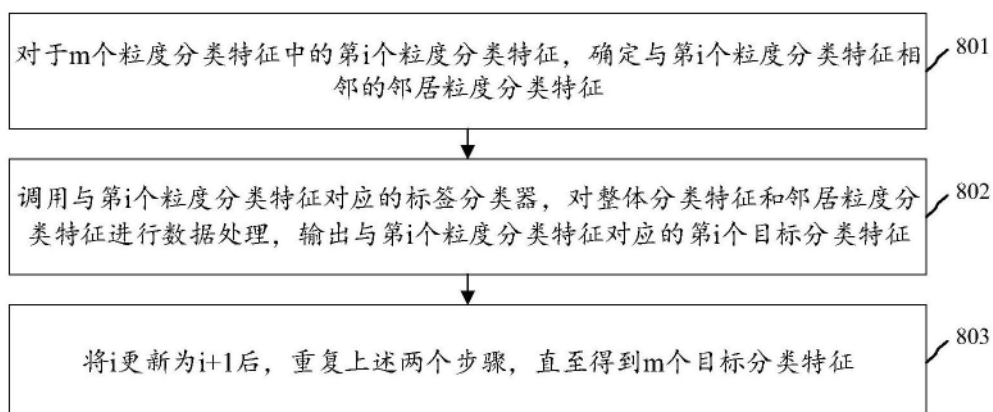


图8

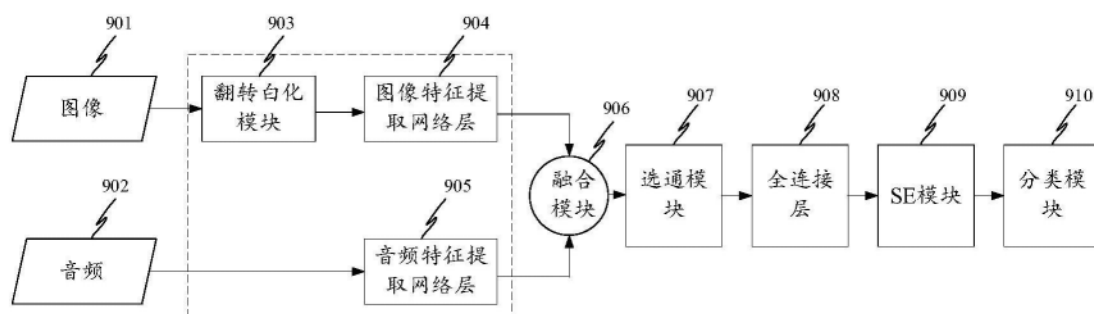


图9

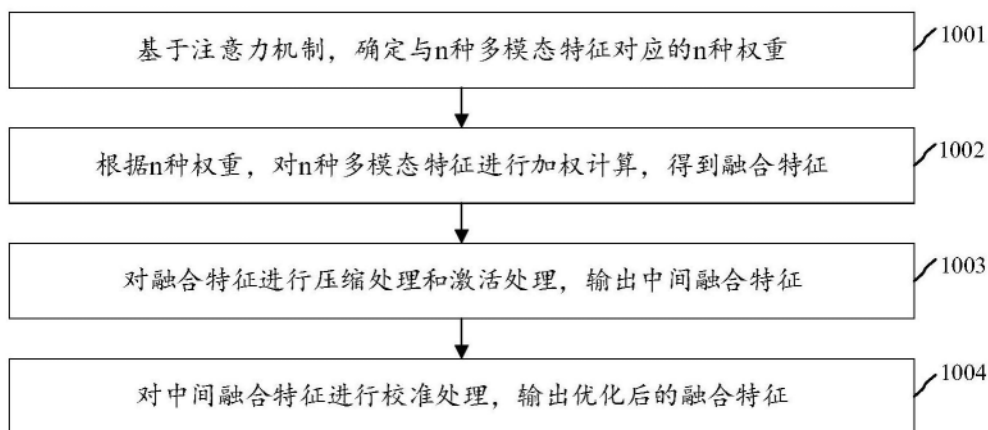


图10

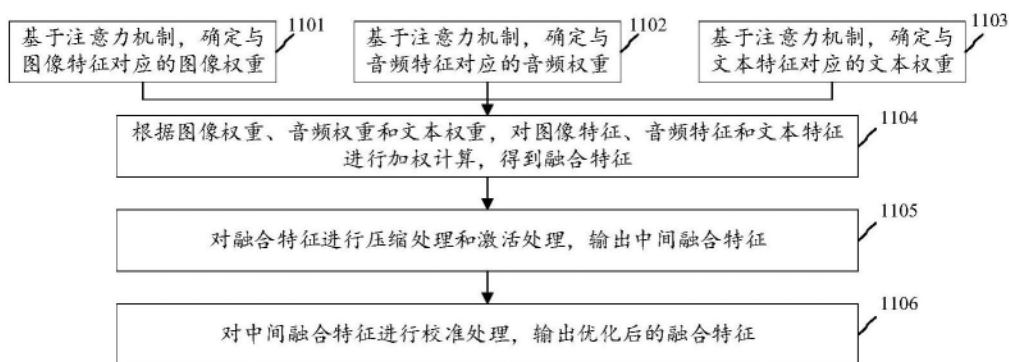


图11

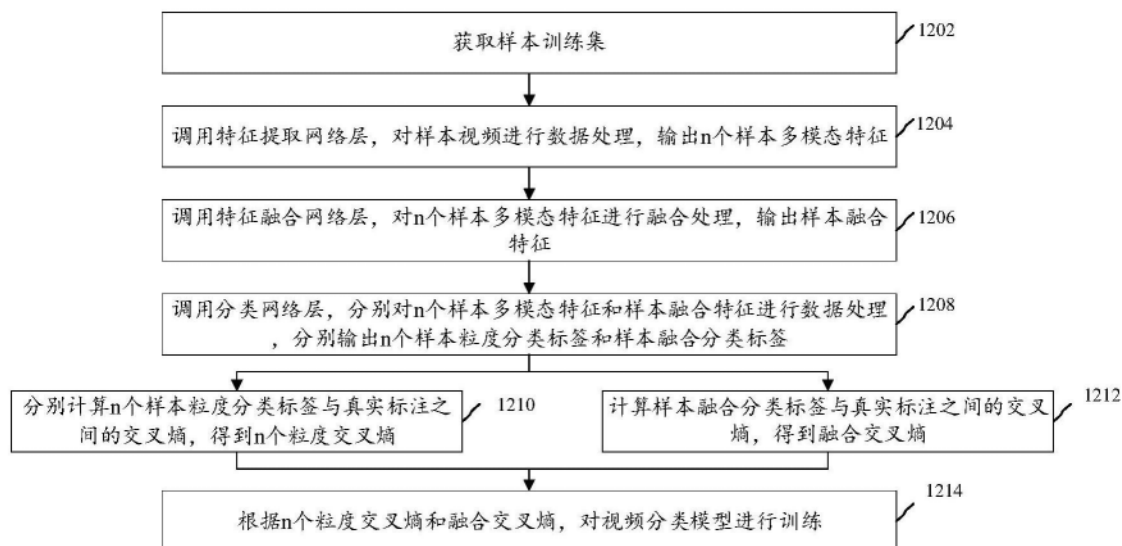


图12

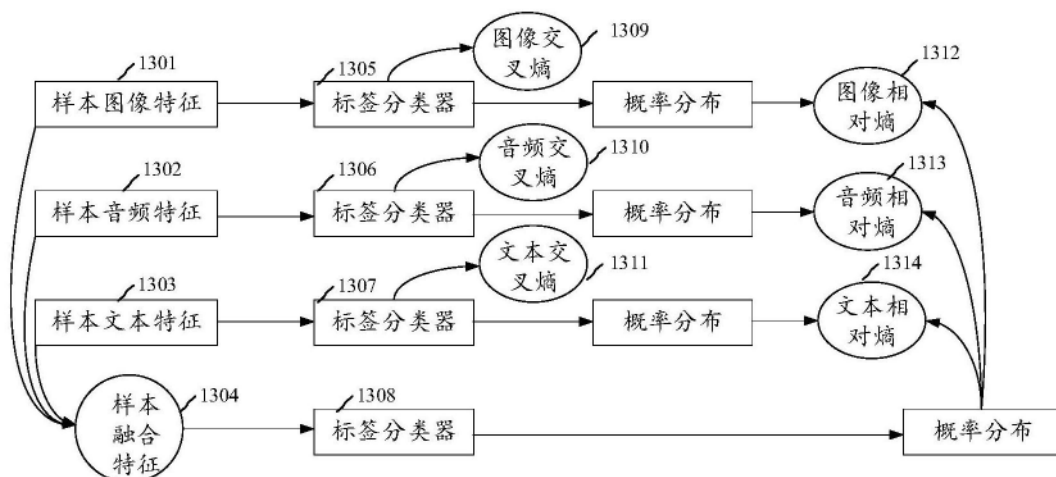


图13

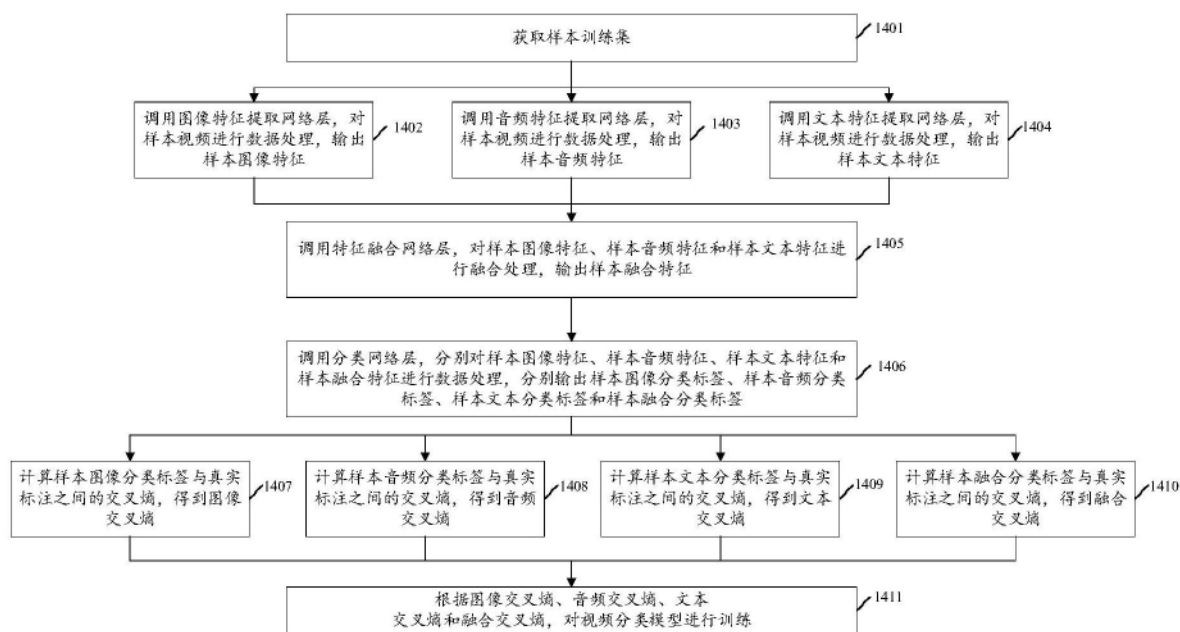


图14

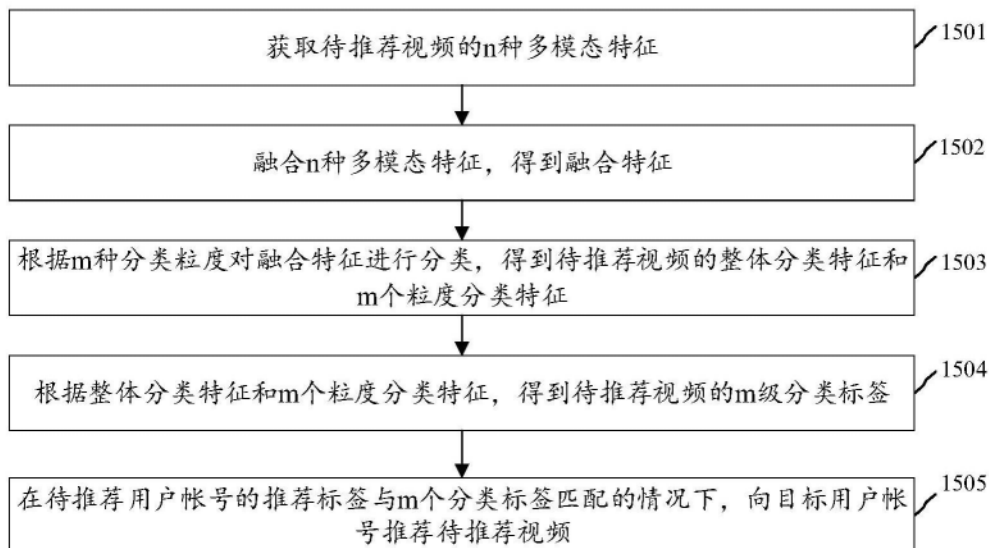


图15

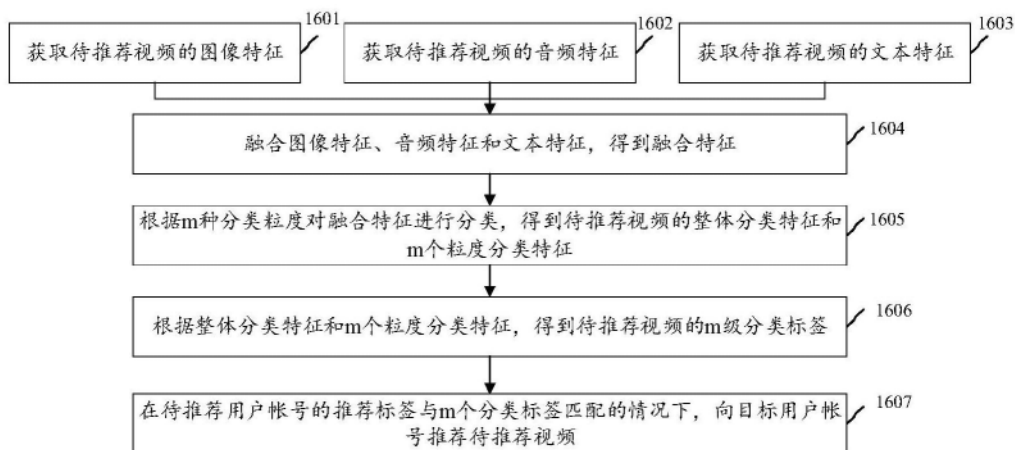


图16

1700



图17

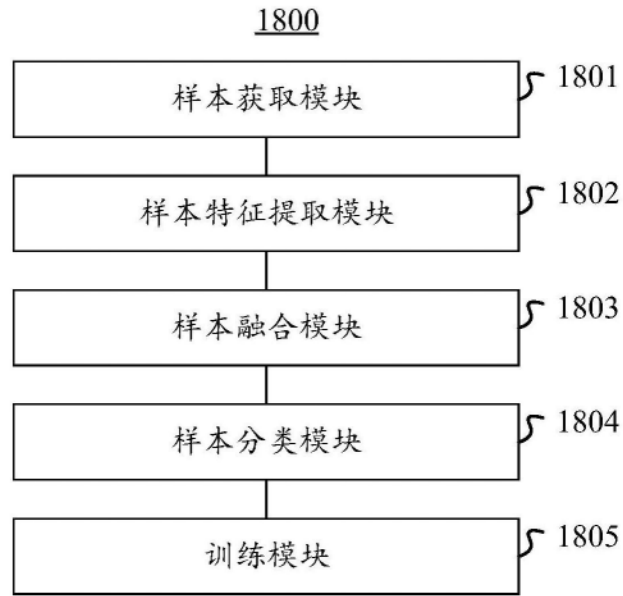


图18

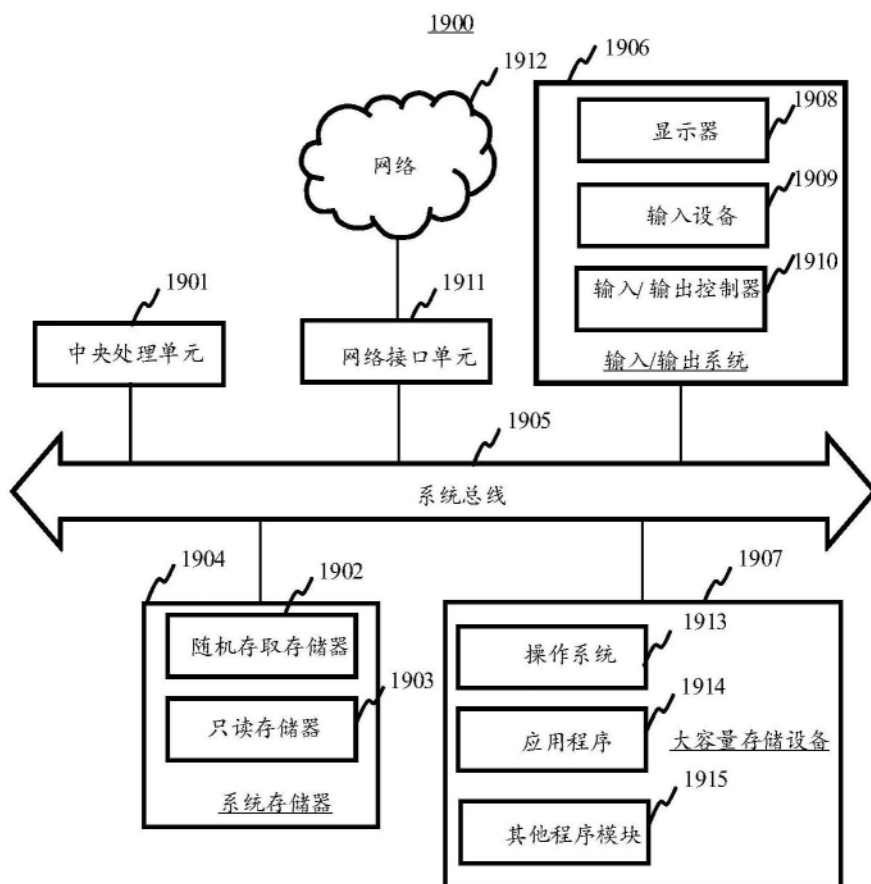


图19