



(12) 发明专利申请

(10) 申请公布号 CN 116095183 A

(43) 申请公布日 2023.05.09

(21) 申请号 202310077949.0

H04N 19/91 (2014.01)

(22) 申请日 2023.01.13

H04N 19/70 (2014.01)

(66) 本国优先权数据

202210249906.1 2022.03.14 CN

(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 张琛 汤姆·莱德 康宁 张世枫

(74) 专利代理机构 深圳市深佳知识产权代理事务所(普通合伙) 44285

专利代理师 夏欢

(51) Int. Cl.

H04L 69/04 (2022.01)

G06N 3/0464 (2023.01)

G06N 3/0455 (2023.01)

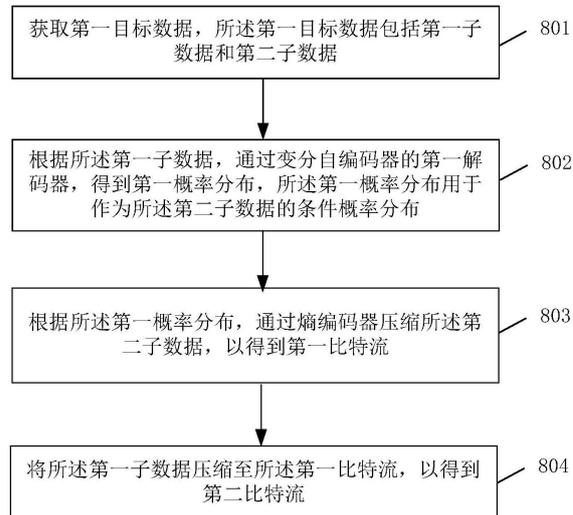
权利要求书5页 说明书28页 附图11页

(54) 发明名称

一种数据压缩方法以及相关设备

(57) 摘要

本申请涉及人工智能领域,公开了一种数据压缩方法,包括:获取第一目标数据,第一目标数据包括第一子数据和第二子数据;根据第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,第一概率分布用于作为第二子数据的条件概率分布;根据第一概率分布,通过熵编码器压缩第二子数据,以得到第一比特流;将所述第一子数据压缩至所述第一比特流,以得到第二比特流。本申请相比于现有技术中反编码机制所需的额外设置的初始比特,本申请实施例中无需额外设置的初始比特,可以实现单数据点的压缩,且降低了并行压缩时的压缩比。



1. 一种数据压缩方法,其特征在于,包括:

获取第一目标数据,所述第一目标数据包括第一子数据和第二子数据;

根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为所述第二子数据的条件概率分布;

根据所述第一概率分布,通过熵编码器压缩所述第二子数据,以得到第一比特流;

将所述第一子数据压缩至所述第一比特流,以得到第二比特流。

2. 根据权利要求1所述的方法,其特征在于,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的;或者,

所述第一目标数据为文字序列,所述第一子数据和所述第二子数据为对所述文字序列进行数据切分后得到的;或者,

所述第一目标数据为二进制流,所述第一子数据和所述第二子数据为对所述二进制流进行数据切分后得到的;或者,

所述第一目标数据为视频,所述第一子数据和所述第二子数据为对所述视频的多个图像帧进行数据切分后得到的。

3. 根据权利要求1或2所述的方法,其特征在于,所述第一子数据和所述第二子数据为对所述图像块在空间维度或者通道维度上进行数据切分后得到的。

4. 根据权利要求1至3任一所述的方法,其特征在于,所述将所述第一子数据压缩至所述第一比特流,包括:

根据所述第一子数据,通过所述变分自编码器中的变分编码器,得到隐变量的近似后验分布;

根据所述近似后验分布,从所述第一比特流中通过所述熵编码器解码出所述隐变量,得到第三比特流;

根据所述隐变量,通过所述变分自编码器的第二解码器,得到第二概率分布;所述第二概率分布用于作为所述第一子数据的条件概率分布;

根据所述第二概率分布,通过所述熵编码器将所述第一子数据压缩至所述第三比特流,以得到第四比特流;

根据所述隐变量的先验分布,通过所述熵编码器将所述隐变量压缩至所述第四比特流,得到第二比特流。

5. 根据权利要求1至4任一所述的方法,其特征在于,所述第一解码器包括第一卷积神经网络和第二卷积神经网络,所述根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,包括:

对包括所述第二子数据的第二目标数据进行空间维度到通道维度的像素重置操作,以得到第三子数据,所述第二目标数据和所述第一目标数据的尺寸大小一致,所述第三子数据和所述第一子数据在空间维度的尺寸相同;

根据所述第一子数据,通过所述第一卷积神经网络,得到第四子数据,所述第四子数据和所述第三子数据在通道维度的尺寸相同;

将所述第三子数据和所述第四子数据进行融合,以得到融合后的子数据;

根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

6. 根据权利要求5所述的方法,其特征在于,所述将所述第三子数据和所述第四子数据

进行融合,包括:

将所述第四子数据中部分通道的数据替换为所述第三子数据中对应通道的数据,以得到融合后的子数据。

7. 根据权利要求5或6所述的方法,其特征在于,所述方法还包括:

将所述融合后的子数据和所述第一子数据沿着通道维度进行拼接操作,以得到拼接后的子数据;

所述根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布,包括:根据所述拼接后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

8. 一种数据解压缩方法,其特征在于,包括:

获取第二比特流;

从所述第二比特流中解码出第一子数据,以得到第一比特流;

根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为第二子数据的条件概率分布;

根据所述第一概率分布,通过熵编码器从所述第一比特流中解压出第二子数据;所述第一子数据和所述第二子数据用于还原得到第一目标数据。

9. 根据权利要求8所述的方法,其特征在于,所述从所述第二比特流中解码出第一子数据,以得到第一比特流,包括:

获取隐变量的先验分布;

根据所述先验分布,通过熵编码器从所述第二比特流中解压出所述隐变量,得到第四比特流;

根据所述隐变量,通过所述变分自编码器的第二解码器,得到第二概率分布;所述第二概率分布用于作为第一子数据的条件概率分布;

根据所述第二概率分布,通过所述熵编码器从所述第四比特流中解压出所述第一子数据,得到第三比特流;

根据所述第一子数据,通过所述变分自编码器中的变分编码器,得到隐变量的近似后验分布;

根据所述近似后验分布,通过所述熵编码器将所述隐变量压缩至所述第三比特流,得到第一比特流。

10. 根据权利要求8或9所述的方法,其特征在于,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的;或者,

所述第一目标数据为文字序列,所述第一子数据和所述第二子数据为对所述文字序列进行数据切分后得到的;或者,

所述第一目标数据为二进制流,所述第一子数据和所述第二子数据为对所述二进制流进行数据切分后得到的;或者

所述第一目标数据为视频,所述第一子数据和所述第二子数据为对所述视频的多个图像帧进行数据切分后得到的。

11. 根据权利要求8至10任一所述的方法,其特征在于,所述第一子数据和所述第二子数据为对所述图像块在空间维度或者通道维度上进行数据切分后得到的。

12. 根据权利要求8至11任一所述的方法,其特征在于,所述第一解码器包括第一卷积

神经网络和第二卷积神经网络,所述根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,包括:

对包括所述第二子数据的第二目标数据进行空间维度到通道维度的像素重置操作,以得到第三子数据,所述第二目标数据和所述第一目标数据的尺寸大小一致,所述第三子数据和所述第一子数据在空间维度的尺寸相同;

根据所述第一子数据,通过所述第一卷积神经网络,得到第四子数据,所述第四子数据和所述第三子数据在通道维度的尺寸相同;

将所述第三子数据和所述第四子数据进行融合,以得到融合后的子数据;

根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

13. 根据权利要求12所述的方法,其特征在于,所述将所述第三子数据和所述第四子数据进行融合,包括:

将所述第四子数据中部分通道的数据替换为所述第三子数据中对应通道的数据,以得到融合后的子数据。

14. 根据权利要求12或13所述的方法,其特征在于,所述方法还包括:

将所述融合后的子数据和所述第一子数据沿着通道维度进行拼接操作,以得到拼接后的子数据;

所述根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布,包括:根据所述拼接后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

15. 一种数据压缩装置,其特征在于,包括:

获取模块,用于获取第一目标数据,所述第一目标数据包括第一子数据和第二子数据;

压缩模块,用于根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为所述第二子数据的条件概率分布;

根据所述第一概率分布,通过熵编码器压缩所述第二子数据,以得到第一比特流;

将所述第一子数据压缩至所述第一比特流,以得到第二比特流。

16. 根据权利要求15所述的装置,其特征在于,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的;或者,

所述第一目标数据为文字序列,所述第一子数据和所述第二子数据为对所述文字序列进行数据切分后得到的;或者,

所述第一目标数据为二进制流,所述第一子数据和所述第二子数据为对所述二进制流进行数据切分后得到的;或者,

所述第一目标数据为视频,所述第一子数据和所述第二子数据为对所述视频的多个图像帧进行数据切分后得到的。

17. 根据权利要求15或16所述的装置,其特征在于,所述压缩模块,具体用于:

根据所述第一子数据,通过所述变分自编码器中的变分编码器,得到隐变量的近似后验分布;

根据所述近似后验分布,从所述第一比特流中通过所述熵编码器解码出所述隐变量,得到第三比特流;

根据所述隐变量,通过所述变分自编码器的第二解码器,得到第二概率分布;所述第二概率分布用于作为所述第一子数据的条件概率分布;

根据所述第二概率分布,通过所述熵编码器将所述第一子数据压缩至所述第三比特流,以得到第四比特流;

根据所述隐变量的先验分布,通过所述熵编码器将所述隐变量压缩至所述第四比特流,得到第二比特流。

18. 根据权利要求15至17任一所述的装置,其特征在于,所述第一解码器包括第一卷积神经网络和第二卷积神经网络,所述压缩模块,具体用于:

对包括所述第二子数据的第二目标数据进行空间维度到通道维度的像素重置操作,以得到第三子数据,所述第二目标数据和所述第一目标数据的尺寸大小一致,所述第三子数据和所述第一子数据在空间维度的尺寸相同;

根据所述第一子数据,通过所述第一卷积神经网络,得到第四子数据,所述第四子数据和所述第三子数据在通道维度的尺寸相同;

将所述第三子数据和所述第四子数据进行融合,以得到融合后的子数据;

根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

19. 根据权利要求18所述的装置,其特征在于,所述将所述第三子数据和所述第四子数据进行融合,包括:

将所述第四子数据中部分通道的数据替换为所述第三子数据中对应通道的数据,以得到融合后的子数据。

20. 根据权利要求18或19所述的装置,其特征在于,所述装置还包括:

拼接模块,用于将所述融合后的子数据和所述第一子数据沿着通道维度进行拼接操作,以得到拼接后的子数据;

所述根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布,包括:根据所述拼接后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

21. 一种数据解压缩装置,其特征在于,包括:

获取模块,用于获取第二比特流;

解压模块,用于从所述第二比特流中解码出第一子数据,以得到第一比特流;

根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为第二子数据的条件概率分布;

根据所述第一概率分布,通过熵编码器从所述第一比特流中解压出第二子数据;所述第一子数据和所述第二子数据用于还原得到第一目标数据。

22. 根据权利要求21所述的装置,其特征在于,所述接码模块,具体用于:

获取隐变量的先验分布;

根据所述先验分布,通过熵编码器从所述第二比特流中解压出所述隐变量,得到第四比特流;

根据所述隐变量,通过所述变分自编码器的第二解码器,得到第二概率分布;所述第二概率分布用于作为第一子数据的条件概率分布;

根据所述第二概率分布,通过所述熵编码器从所述第四比特流中解压出所述第一子数据,得到第三比特流;

根据所述第一子数据,通过所述变分自编码器中的变分编码器,得到隐变量的近似后验分布;

根据所述近似后验分布,通过所述熵编码器将所述隐变量压缩至所述第三比特流,得到第一比特流。

23. 根据权利要求21或22所述的装置,其特征在于,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的;或者,

所述第一目标数据为文字序列,所述第一子数据和所述第二子数据为对所述文字序列进行数据切分后得到的;或者,

所述第一目标数据为二进制流,所述第一子数据和所述第二子数据为对所述二进制流进行数据切分后得到的;或者

所述第一目标数据为视频,所述第一子数据和所述第二子数据为对所述视频的多个图像帧进行数据切分后得到的。

24. 根据权利要求21至23任一所述的装置,其特征在于,所述第一解压器包括第一卷积神经网络和第二卷积神经网络,所述解压模块,具体用于:

对包括所述第二子数据的第二目标数据进行空间维度到通道维度的像素重置操作,以得到第三子数据,所述第二目标数据和所述第一目标数据的尺寸大小一致,所述第三子数据和所述第一子数据在空间维度的尺寸相同;

根据所述第一子数据,通过所述第一卷积神经网络,得到第四子数据,所述第四子数据和所述第三子数据在通道维度的尺寸相同;

将所述第三子数据和所述第四子数据进行融合,以得到融合后的子数据;

根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

25. 根据权利要求24所述的装置,其特征在于,所述将所述第三子数据和所述第四子数据进行融合,包括:

将所述第四子数据中部分通道的数据替换为所述第三子数据中对应通道的数据,以得到融合后的子数据。

26. 根据权利要求24或25所述的装置,其特征在于,所述装置还包括:

拼接模块,用于将所述融合后的子数据和所述第一子数据沿着通道维度进行拼接操作,以得到拼接后的子数据;

所述根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布,包括:根据所述拼接后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

27. 一种数据压缩装置,其特征在于,包括存储介质、处理电路以及总线系统;其中,所述存储介质用于存储指令,所述处理电路用于执行存储器中的指令,以执行所述权利要求1至14中任一项所述的方法的步骤。

28. 一种计算机可读存储介质,其上存储有计算机程序,其特征在于,该程序被处理器执行时实现权利要求1至14中任一项所述的方法的步骤。

29. 一种计算机程序产品,其特征在于,所述计算机程序产品包括代码,当所述代码被执行时,用于实现权利要求1至14任一项所述的方法的步骤。

## 一种数据压缩方法以及相关设备

[0001] 本申请要求于2022年3月14日提交中国专利局、申请号202210249906.1、发明名称为“一种数据压缩方法以及相关设备”的中国专利申请的优先权,其全部内容通过引用结合在本申请中。

### 技术领域

[0002] 本申请涉及人工智能领域,尤其涉及一种数据压缩方法以及相关设备。

### 背景技术

[0003] 如今多媒体数据占据了互联网的绝大部分流量。对于图像数据的压缩对于多媒体数据的存储和高效传输有着至关重要的作用。所以图像编码是一项具有重大实用价值的技术。

[0004] 对于图像编码的研究已经有较长的历史了,研究人员提出了大量的方法,并制定了多种国际标准,比如JPEG, JPEG2000, WebP, BPG等图像编码标准。这些编码方法虽然在目前都得到了广泛应用,但是针对现在不断增长的图像数据量及不断出现的新媒体类型,这些传统方法显示出了某些局限性。

[0005] 基于人工智能的无损压缩方案利用了深度生成模型能够比传统的方案更准确地估计数据的概率分布这一特性,得到了远优于传统无损压缩方案的压缩比。在基于人工智能的无损压缩方案中,被广泛使用的深度生成模型包括自回归模型 (autoregressive model), 变分自编码器 (variational auto-encoder, VAE), 流模型 (normalizing flows) 等。一般来讲,自回归模型可较好地兼容算术编码器和霍夫曼编码;变分自编码器结合使用反编码 (bits-back) 机制可较好地兼容非对称数字系统;流模型可以兼容上述三种不同的熵编码器。除了压缩比以外,评价无损压缩解决方案的还有吞吐率这一指标。对于基于人工智能的无损压缩解决方案来说,由于模型规模远大于传统方案,因此整体吞吐率低于传统方案。另外,综合压缩比和吞吐率两个指标来说,基于不同生成模型的无损压缩解决方案目前没有绝对的先后之分。目前的研究尚处于对不同生成模型的压缩方案探索其帕累托前沿的阶段。

[0006] 其中,区别于全观测模型(如自回归模型),变分自编码器模型是一种隐变量模型。该类模型并非对数据数据本身直接建模,而是额外引入了一个(或者多个)隐变量,然后对先验分布,似然函数以及近似后验分布进行建模。由于从变分自编码器中无法直接获得数据数据的边际分布,传统的熵编码方式无法直接被沿用。为了能够使用变分自编码器进行数据的无损压缩,基于反编码机制的变分自编码无损压缩方案被提出。bits-back ANS是该方案的原始形式,适用于只包含一个隐变量的变分自编码器模型,并且可以推广适用到包含多个隐变量的变分自编码器模型。

[0007] 现行的基于反编码机制的变分自编码器无损压缩方案均需要额外的初始比特用以解压出隐变量的样本。额外的初始比特为随机生成的数据,该数据的大小需要考虑进压缩成本中,且在待串行压缩的数据数量较少时,额外的平均成本较高;且,由于所需的额外

初始比特与待压缩数据点的个数成正比,因此无法实现高效的并行压缩。

## 发明内容

[0008] 本申请提供一种数据压缩方法,相比于现有技术中反编码机制所需的额外设置的初始比特,本申请实施例中无需额外设置的初始比特,可以实现单数据点的压缩,且大大降低了并行压缩时的压缩比。

[0009] 第一方面,本申请提供一种数据压缩方法,包括:获取第一目标数据,所述第一目标数据包括第一子数据和第二子数据;

[0010] 在一种可能的实现中,第一目标数据可以为供压缩的图像数据或者是其他数据(例如文本、视频、二进制流等)。

[0011] 在一种可能的实现中,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的;或者,

[0012] 所述第一目标数据为文字序列,所述第一子数据和所述第二子数据为对所述文字序列进行数据切分后得到的;或者,

[0013] 所述第一目标数据为二进制流,所述第一子数据和所述第二子数据为对所述二进制流进行数据切分后得到的;或者,

[0014] 所述第一目标数据为视频,所述第一子数据和所述第二子数据为对所述视频的多个图像帧进行数据切分后得到的。

[0015] 在一种可能的实现中,所述第一子数据和所述第二子数据为对所述图像块在空间维度或者通道维度上进行数据切分后得到的。其中,对于图像数据来说,包含了一个通道维度(C)和两个空间维度(宽W和高H)。

[0016] 根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为所述第二子数据的条件概率分布;根据所述第一概率分布,通过熵编码器压缩所述第二子数据,以得到第一比特流;将所述第一比特流作为初始比特流,对所述第一子数据进行压缩(也就是将所述第一子数据压缩至所述第一比特流),以得到第二比特流。

[0017] 相比于现有技术中反编码机制所需的额外设置的初始比特,本申请实施例中无需额外设置的初始比特,可以实现单数据点的压缩,且大大降低了并行压缩时的压缩比。

[0018] 在一种可能的实现中,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的。

[0019] 在一种可能的实现中,所述第一子数据和所述第二子数据为对所述图像块在空间维度或者通道维度上进行数据切分后得到的。

[0020] 在一种可能的实现中,变分自编码器可以包括变分编码器,解码器(例如本申请实施例中的第一解码器和第二解码器)和隐变量的先验分布。

[0021] 在一种可能的实现中,解码器可以由解码器层(例如本申请实施例中的第一卷积神经网络以及第二卷积神经网络)构成,且包含解码器层的个数与变分自编码器中隐变量的个数相同。解码器层的作用是输入更深层的隐变量,输出当前层数据的条件概率分布(当前层数据可以是更浅层的隐变量或者数据数据)。

[0022] 在现有的变分自编码器模型中,变分编码器需要输入整个数据数据以预测隐变量

的近似后验分布,解码器中输入隐变量直接预测整个数据数据的条件概率分布。在本申请实施例中,将待压缩的数据分成至少两部分,即:第一子数据和第二子数据。和现有将全部数据输入到变分编码器不同的是,本申请实施例中仅将数据的一部分(第一子数据)输入到变分编码器,来预测隐变量的近似后验分布,且隐变量输入第一解码器后预测第一子数据的条件概率分布;第二子数据的条件概率分布依赖于第一子数据,具体可以由将第一子数据输入第一解码器来确定。

[0023] 在一种可能的实现中,解码器可以实现像素重置操作。

[0024] 在一种可能的实现中,所述第一解码器可以包括第一卷积神经网络和第二卷积神经网络,所述根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,具体可以包括:对包括所述第二子数据的第二目标数据进行空间维度到通道维度的像素重置操作,以得到第三子数据,所述第二目标数据和所述第一目标数据的尺寸大小一致,所述第三子数据和所述第一子数据在空间维度的尺寸相同;

[0025] 其中,包括所述第二子数据的第二目标数据可以为和第一目标数据尺寸相同的数据,其中,在第一目标数据中,除了第二子数据之外的元素可以被置零(或者其他预设数值),以得到第二目标数据,对第二目标数据进行像素重置操作后,可以将其转化为和第一子数据在空间维度的尺寸相同的第三子数据。

[0026] 本申请实施例通过使用基于通道优先像素重置定义的回归结构的编码器层,充分利用图片像素间的相关关系,从而在获得更低的编码长度的前提下大幅降低模型所需参数量,进而提高了压缩的吞吐率以及降低了模型存储的空间成本。

[0027] 在一种可能的实现中,可以根据所述第一子数据,通过所述第一卷积神经网络,得到第四子数据,所述第四子数据和所述第三子数据在通道维度的尺寸相同。也就是说,可以通过第一卷积神经网络,对第一子数据进行特征提取以及尺寸的变换,以便得到一个赫尔第三子数据在通道维度的尺寸相同的第四子数据。

[0028] 在一种可能的实现中,可以将所述第三子数据和所述第四子数据进行融合,以得到融合后的子数据。可选的,融合方式可以为对应通道的数据替换。

[0029] 在一种可能的实现中,所述将所述第三子数据和所述第四子数据进行融合,具体可以包括:将所述第四子数据中部分通道的数据替换为所述第三子数据中对应通道的数据,以得到融合后的子数据。

[0030] 在一种可能的实现中,可以根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

[0031] 在一种可能的实现中,还可以将所述融合后的子数据和所述第一子数据沿着通道维度进行拼接操作(concat),以得到拼接后的子数据;进而,所述根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布,具体可以包括:根据所述拼接后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

[0032] 第二方面,本申请提供了一种数据解压缩方法,包括:

[0033] 获取第二比特流;

[0034] 从所述第二比特流中解码出第一子数据,以得到第一比特流;

[0035] 根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为第二子数据的条件概率分布;

- [0036] 根据所述第一概率分布,通过熵编码器从所述第一比特流中解压出第二子数据;所述第一子数据和所述第二子数据用于还原得到第一目标数据。
- [0037] 在一种可能的实现中,所述从所述第二比特流中解码出第一子数据,以得到第一比特流,包括:
- [0038] 获取隐变量的先验分布;
- [0039] 根据所述先验分布,通过熵编码器从所述第二比特流中解压出所述隐变量,得到第四比特流;
- [0040] 根据所述隐变量,通过所述变分自编码器的第二解码器,得到第二概率分布;所述第二概率分布用于作为第一子数据的条件概率分布;
- [0041] 根据所述第二概率分布,通过所述熵编码器从所述第四比特流中解压出所述第一子数据,得到第三比特流;
- [0042] 根据所述第一子数据,通过所述变分自编码器中的变分编码器,得到隐变量的近似后验分布;
- [0043] 根据所述近似后验分布,通过所述熵编码器将所述隐变量压缩至所述第三比特流,得到第一比特流。
- [0044] 在一种可能的实现中,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的;或者,
- [0045] 所述第一目标数据为文字序列,所述第一子数据和所述第二子数据为对所述文字序列进行数据切分后得到的;或者,
- [0046] 所述第一目标数据为二进制流,所述第一子数据和所述第二子数据为对所述二进制流进行数据切分后得到的;或者
- [0047] 所述第一目标数据为视频,所述第一子数据和所述第二子数据为对所述视频的多个图像帧进行数据切分后得到的。
- [0048] 在一种可能的实现中,所述第一子数据和所述第二子数据为对所述图像块在空间维度或者通道维度上进行数据切分后得到的。
- [0049] 在一种可能的实现中,所述第一解码器包括第一卷积神经网络和第二卷积神经网络,所述根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,包括:
- [0050] 对包括所述第二子数据的第二目标数据进行空间维度到通道维度的像素重置操作,以得到第三子数据,所述第二目标数据和所述第一目标数据的尺寸大小一致,所述第三子数据和所述第一子数据在空间维度的尺寸相同;
- [0051] 根据所述第一子数据,通过所述第一卷积神经网络,得到第四子数据,所述第四子数据和所述第三子数据在通道维度的尺寸相同;
- [0052] 将所述第三子数据和所述第四子数据进行融合,以得到融合后的子数据;
- [0053] 根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。
- [0054] 在一种可能的实现中,所述将所述第三子数据和所述第四子数据进行融合,包括:
- [0055] 将所述第四子数据中部分通道的数据替换为所述第三子数据中对应通道的数据,以得到融合后的子数据。
- [0056] 在一种可能的实现中,所述方法还包括:
- [0057] 将所述融合后的子数据和所述第一子数据沿着通道维度进行拼接操作,以得到拼

接后的子数据；

[0058] 所述根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布,包括:根据所述拼接后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

[0059] 第三方面,本申请提供了一种数据压缩装置,包括:

[0060] 获取模块,用于获取第一目标数据,所述第一目标数据包括第一子数据和第二子数据;

[0061] 压缩模块,用于根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为所述第二子数据的条件概率分布;

[0062] 根据所述第一概率分布,通过熵编码器压缩所述第二子数据,以得到第一比特流;

[0063] 将所述第一子数据压缩至所述第一比特流,以得到第二比特流。

[0064] 在一种可能的实现中,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的;或者,

[0065] 所述第一目标数据为文字序列,所述第一子数据和所述第二子数据为对所述文字序列进行数据切分后得到的;或者,

[0066] 所述第一目标数据为二进制流,所述第一子数据和所述第二子数据为对所述二进制流进行数据切分后得到的;或者

[0067] 所述第一目标数据为视频,所述第一子数据和所述第二子数据为对所述视频的多个图像帧进行数据切分后得到的。

[0068] 在一种可能的实现中,所述第一子数据和所述第二子数据为对所述图像块在空间维度或者通道维度上进行数据切分后得到的。

[0069] 在一种可能的实现中,所述压缩模块,具体用于:

[0070] 根据所述第一子数据,通过所述变分自编码器中的变分编码器,得到隐变量的近似后验分布;

[0071] 根据所述近似后验分布,从所述第一比特流中通过所述熵编码器解码出所述隐变量,得到第三比特流;

[0072] 根据所述隐变量,通过所述变分自编码器的第二解码器,得到第二概率分布;所述第二概率分布用于作为所述第一子数据的条件概率分布;

[0073] 根据所述第二概率分布,通过所述熵编码器将所述第一子数据压缩至所述第三比特流,以得到第四比特流;

[0074] 根据所述隐变量的先验分布,通过所述熵编码器将所述隐变量压缩至所述第四比特流,得到第二比特流。

[0075] 在一种可能的实现中,所述第一解码器包括第一卷积神经网络和第二卷积神经网络,所述压缩模块,具体用于:

[0076] 对包括所述第二子数据的第二目标数据进行空间维度到通道维度的像素重置操作,以得到第三子数据,所述第二目标数据和所述第一目标数据的尺寸大小一致,所述第三子数据和所述第一子数据在空间维度的尺寸相同;

[0077] 根据所述第一子数据,通过所述第一卷积神经网络,得到第四子数据,所述第四子数据和所述第三子数据在通道维度的尺寸相同;

- [0078] 将所述第三子数据和所述第四子数据进行融合,以得到融合后的子数据;
- [0079] 根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。
- [0080] 在一种可能的实现中,所述将所述第三子数据和所述第四子数据进行融合,包括:
- [0081] 将所述第四子数据中部分通道的数据替换为所述第三子数据中对应通道的数据,以得到融合后的子数据。
- [0082] 在一种可能的实现中,所述装置还包括:
- [0083] 拼接模块,用于将所述融合后的子数据和所述第一子数据沿着通道维度进行拼接操作,以得到拼接后的子数据;
- [0084] 所述根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布,包括:根据所述拼接后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。
- [0085] 第四方面,本申请提供了一种数据解压缩装置,包括:
- [0086] 获取模块,用于获取第二比特流;
- [0087] 解压模块,用于从所述第二比特流中解码出第一子数据,以得到第一比特流;
- [0088] 根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为第二子数据的条件概率分布;
- [0089] 根据所述第一概率分布,通过熵编码器从所述第一比特流中解压出第二子数据;所述第一子数据和所述第二子数据用于还原得到第一目标数据。
- [0090] 在一种可能的实现中,所述接码模块,具体用于:
- [0091] 获取隐变量的先验分布;
- [0092] 根据所述先验分布,通过熵编码器从所述第二比特流中解压出所述隐变量,得到第四比特流;
- [0093] 根据所述隐变量,通过所述变分自编码器的第二解码器,得到第二概率分布;所述第二概率分布用于作为第一子数据的条件概率分布;
- [0094] 根据所述第二概率分布,通过所述熵编码器从所述第四比特流中解压出所述第一子数据,得到第三比特流;
- [0095] 根据所述第一子数据,通过所述变分自编码器中的变分编码器,得到隐变量的近似后验分布;
- [0096] 根据所述近似后验分布,通过所述熵编码器将所述隐变量压缩至所述第三比特流,得到第一比特流。
- [0097] 在一种可能的实现中,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的;或者,
- [0098] 所述第一目标数据为文字序列,所述第一子数据和所述第二子数据为对所述文字序列进行数据切分后得到的;或者,
- [0099] 所述第一目标数据为二进制流,所述第一子数据和所述第二子数据为对所述二进制流进行数据切分后得到的;或者
- [0100] 所述第一目标数据为视频,所述第一子数据和所述第二子数据为对所述视频的多个图像帧进行数据切分后得到的。
- [0101] 在一种可能的实现中,所述第一子数据和所述第二子数据为对所述图像块在空间

维度或者通道维度上进行数据切分后得到的。

[0102] 在一种可能的实现中,所述第一解码器包括第一卷积神经网络和第二卷积神经网络,所述解压模块,具体用于:

[0103] 对包括所述第二子数据的第二目标数据进行空间维度到通道维度的像素重置操作,以得到第三子数据,所述第二目标数据和所述第一目标数据的尺寸大小一致,所述第三子数据和所述第一子数据在空间维度的尺寸相同;

[0104] 根据所述第一子数据,通过所述第一卷积神经网络,得到第四子数据,所述第四子数据和所述第三子数据在通道维度的尺寸相同;

[0105] 将所述第三子数据和所述第四子数据进行融合,以得到融合后的子数据;

[0106] 根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

[0107] 在一种可能的实现中,所述将所述第三子数据和所述第四子数据进行融合,包括:

[0108] 将所述第四子数据中部分通道的数据替换为所述第三子数据中对应通道的数据,以得到融合后的子数据。

[0109] 在一种可能的实现中,所述装置还包括:

[0110] 拼接模块,用于将所述融合后的子数据和所述第一子数据沿着通道维度进行拼接操作,以得到拼接后的子数据;

[0111] 所述根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布,包括:根据所述拼接后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

[0112] 第五方面,本申请提供了一种数据压缩装置,包括存储介质、处理电路以及总线系统;其中,所述存储介质用于存储指令,所述处理电路用于执行存储器中的指令,以执行上述第一方面任一所述的数据压缩方法。

[0113] 第六方面,本申请提供了一种数据压缩装置,包括存储介质、处理电路以及总线系统;其中,所述存储介质用于存储指令,所述处理电路用于执行存储器中的指令,以执行上述第二方面任一所述的数据压缩方法。

[0114] 第七方面,本申请实施例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计算机程序,当其在计算机上运行时,使得计算机执行上述第一方面任一所述的数据压缩方法。

[0115] 第八方面,本申请实施例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计算机程序,当其在计算机上运行时,使得计算机执行上述第二方面任一所述的数据压缩方法。

[0116] 第九方面,本申请实施例提供了一种计算机程序,当其在计算机上运行时,使得计算机执行上述第一方面任一所述的数据压缩方法。

[0117] 第十方面,本申请实施例提供了一种计算机程序,当其在计算机上运行时,使得计算机执行上述第二方面任一所述的数据压缩方法。

[0118] 第十一方面,本申请提供了一种芯片系统,该芯片系统包括处理器,用于支持执行设备(例如数据压缩装置或者数据解压缩装置)或训练设备实现上述方面中所涉及的功能,例如,发送或处理上述方法中所涉及的数据和/或信息。在一种可能的设计中,所述芯片系统还包括存储器,所述存储器,用于保存执行设备或训练设备必要的程序指令和数据。该芯

片系统,可以由芯片构成,也可以包括芯片和其他分立器件。

[0119] 本申请实施例提供了一种数据压缩方法,包括:获取第一目标数据,所述第一目标数据包括第一子数据和第二子数据;根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为所述第二子数据的条件概率分布;根据所述第一概率分布,通过熵编码器压缩所述第二子数据,以得到第一比特流;将所述第一子数据压缩至所述第一比特流,以得到第二比特流。相比于现有技术中反编码机制所需的额外设置的初始比特,本申请实施例中无需额外设置的初始比特,可以实现单数据点的压缩,且降低了并行压缩时的压缩比。

## 附图说明

- [0120] 图1为人工智能主体框架的一种结构示意图;
- [0121] 图2为本申请实施例的应用场景示意;
- [0122] 图3为本申请实施例的应用场景示意;
- [0123] 图4为一种基于CNN的数据处理过程示意;
- [0124] 图5为一种基于CNN的数据处理过程示意;
- [0125] 图6为本申请实施例提供的一种系统架构法的实施例示意;
- [0126] 图7为本申请实施例提供的一种芯片的结构示意;
- [0127] 图8为本申请实施例提供的一种数据压缩方法的流程示意;
- [0128] 图9为本申请实施例提供的一种像素置换操作的实施例示意;
- [0129] 图10为本申请实施例提供的一种解码器的处理流程示意;
- [0130] 图11为本申请实施例提供的解码器的结构示意;
- [0131] 图12为本申请实施例提供的一种数据压缩方法的流程示意;
- [0132] 图13为本申请实施例提供的一种数据压缩方法的流程示意;
- [0133] 图14为本申请实施例提供的一种数据解压缩方法的流程示意;
- [0134] 图15为本申请实施例提供的数据压缩装置的一种结构示意图;
- [0135] 图16为本申请实施例提供的数据解压缩装置的一种结构示意图;
- [0136] 图17为本申请实施例提供的执行设备的一种结构示意图。

## 具体实施方式

[0137] 下面结合本发明实施例中的附图对本发明实施例进行描述。本发明的实施方式部分使用的术语仅用于对本发明的具体实施例进行解释,而非旨在限定本发明。

[0138] 下面结合附图,对本申请的实施例进行描述。本领域普通技术人员可知,随着技术的发展和场景的出现,本申请实施例提供的技术方案对于类似的技术问题,同样适用。

[0139] 本申请的说明书和权利要求书及上述附图中的术语“第一”、“第二”等是用于区别类似的对象,而不必用于描述特定的顺序或先后次序。应该理解这样使用的术语在适当情况下可以互换,这仅仅是描述本申请的实施例中对相同属性的对象在描述时所采用的区分方式。此外,术语“包括”和“具有”以及他们的任何变形,意图在于覆盖不排他的包含,以便包含一系列单元的过程、方法、系统、产品或设备不必限于那些单元,而是可包括没有清楚地列出的或对于这些过程、方法、产品或设备固有的其它单元。

[0140] 首先对人工智能系统总体工作流程进行描述,请参见图1,图1示出的为人工智能主体框架的一种结构示意图,下面从“智能信息链”(水平轴)和“IT价值链”(垂直轴)两个维度对上述人工智能主题框架进行阐述。其中,“智能信息链”反映从数据的获取到处理的一系列过程。举例来说,可以是智能信息感知、智能信息表示与形成、智能推理、智能决策、智能执行与输出的一般过程。在这个过程中,数据经历了“数据—信息—知识—智慧”的凝练过程。“IT价值链”从人工智能的底层基础设施、信息(提供和处理技术实现)到系统的产业生态过程,反映人工智能为信息技术产业带来的价值。

[0141] (1) 基础设施

[0142] 基础设施为人工智能系统提供计算能力支持,实现与外部世界的沟通,并通过基础平台实现支撑。通过传感器与外部沟通;计算能力由智能芯片(CPU、NPU、GPU、ASIC、FPGA等硬件加速芯片)提供;基础平台包括分布式计算框架及网络等相关的平台保障和支持,可以包括云存储和计算、互联互通网络等。举例来说,传感器和外部沟通获取数据,这些数据提供给基础平台提供的分布式计算系统中的智能芯片进行计算。

[0143] (2) 数据

[0144] 基础设施的上一层的数据用于表示人工智能领域的数据来源。数据涉及到图形、图像、语音、文本,还涉及到传统设备的物联网数据,包括已有系统的业务数据以及力、位移、液位、温度、湿度等感知数据。

[0145] (3) 数据处理

[0146] 数据处理通常包括数据训练,机器学习,深度学习,搜索,推理,决策等方式。

[0147] 其中,机器学习和深度学习可以对数据进行符号化和形式化的智能信息建模、抽取、预处理、训练等。

[0148] 推理是指在计算机或智能系统中,模拟人类的智能推理方式,依据推理控制策略,利用形式化的信息进行机器思维和求解问题的过程,典型的功能是搜索与匹配。

[0149] 决策是指智能信息经过推理后进行决策的过程,通常提供分类、排序、预测等功能。

[0150] (4) 通用能力

[0151] 对数据经过上面提到的数据处理后,进一步基于数据处理的结果可以形成一些通用的能力,比如可以是算法或者一个通用系统,例如,翻译,文本的分析,计算机视觉的处理,语音识别,图像的识别等等。

[0152] (5) 智能产品及行业应用

[0153] 智能产品及行业应用指人工智能系统在各领域的产品和应用,是对人工智能整体解决方案的封装,将智能信息决策产品化、实现落地应用,其应用领域主要包括:智能终端、智能交通、智能医疗、自动驾驶、智慧城市等。

[0154] 本申请可以应用于人工智能领域的图像压缩领域中,下面将对多个落地到产品的多个应用场景进行介绍。

[0155] 一、应用于终端设备中的图像压缩过程

[0156] 本申请实施例提供的图像压缩方法可以应用于终端设备中的图像压缩过程,具体的,可以应用于终端设备上的相册、视频监控等。具体的,可以参照图2,图2为本申请实施例的应用场景示意,如图2中示出的那样,终端设备可以获取到待压缩图片,其中待压缩图片

可以是相机拍摄的照片或是从视频中截取的一帧画面。终端设备可以通过嵌入式神经网络(neural-network processing unit,NPU)中的人工智能(artificial intelligence,AI)编码单元对获取到的待压缩图片进行特征提取,将图像数据变换成冗余度更低的输出特征,且产生输出特征中各点的概率估计,中央处理器(central processing unit,CPU)通过输出特征中各点的概率估计对提取获得的输出特征进行算术编码,降低输出特征的编码冗余,进一步降低图像压缩过程中的数据传输量,并将编码得到的编码数据以数据文件的形式保存在对应的存储位置。当用户需要获取上述存储位置中保存的文件时,CPU可以在相应的存储位置获取并加载上述保存的文件,并基于算数解码获取到解码得到的特征图,通过NPU中的AI解码单元对特征图进行重构,得到重构的图像。

[0157] 二、应用于云侧的图像压缩过程

[0158] 本申请实施例提供的图像压缩方法可以应用于云侧的图像压缩过程,具体的,可以应用于云侧服务器上的云相册等功能。具体的,可以参照图3,图3为本申请实施例的应用场景示意,如图3中示出的那样,终端设备可以获得待压缩图片,其中待压缩图片可以是相机拍摄的照片或是从视频中截取的一帧画面。终端设备可以通过CPU对待压缩图片进行无损编码压缩,得到编码数据,例如但不限于基于现有技术中的任意一种无损压缩方法,终端设备可以将编码数据传输至云侧的服务器,服务器可以对接收到的编码数据进行相应的无损解码,得到待压缩图像,服务器可以通过图形处理器(graphics processing unit,GPU)中的AI编码单元对获取到的待压缩图片进行特征提取,将图像数据变换成冗余度更低的输出特征,且产生输出特征中各点的概率估计,CPU通过输出特征中各点的概率估计对提取获得的输出特征进行算术编码,降低输出特征的编码冗余,进一步降低图像压缩过程中的数据传输量,并将编码得到的编码数据以数据文件的形式保存在对应的存储位置。当用户需要获取上述存储位置中保存的文件时,CPU可以在相应的存储位置获取并加载上述保存的文件,并基于算数解码获取到解码得到的特征图,通过NPU中的AI解码单元对特征图进行重构,得到重构的图像,服务器可以通过CPU对待压缩图片进行无损编码压缩,得到编码数据,例如但不限于基于现有技术中的任意一种无损压缩方法,服务器可以将编码数据传输至终端设备,终端设备可以对接收到的编码数据进行相应的无损解码,得到解码后的图像。

[0159] 由于本申请实施例涉及大量神经网络的应用,为了便于理解,下面先对本申请实施例可能涉及的神经网络的相关术语和概念进行介绍。

[0160] (1) 神经网络

[0161] 神经网络可以是由神经单元组成的,神经单元可以是指以 $x_s$ 和截距1为输入的运算单元,该运算单元的输出可以为:

$$[0162] \quad h_{W,b}(x) = f(W^T x) = f\left(\sum_{s=1}^n W_s x_s + b\right);$$

[0163] 其中, $s=1,2,\dots,n$ , $n$ 为大于1的自然数, $W_s$ 为 $X_s$ 的权重, $b$ 为神经单元的偏置。 $f$ 为神经单元的激活函数(activation functions),用于将非线性特性引入神经网络中,来将神经单元中的输入信号转换为输出信号。该激活函数的输出信号可以作为下一层卷积层的输入,激活函数可以是sigmoid函数。神经网络是将多个上述单一的神经单元联结在一起形成的网络,即一个神经单元的输出可以是另一个神经单元的输入。每个神经单元的输入

可以与前一层的局部接受域相连,来提取局部接受域的特征,局部接受域可以是由若干个神经元组成的区域。

[0164] (2) 深度神经网络

[0165] 深度神经网络(deep neural network,DNN),也称多层神经网络,可以理解为具有多层隐含层的神经网络。按照不同层的位置对DNN进行划分,DNN内部的神经网络可以分为三类:输入层,隐含层,输出层。一般来说第一层是输入层,最后一层是输出层,中间的层数都是隐含层。层与层之间是全连接的,也就是说,第*i*层的任意一个神经元一定与第*i*+1层的任意一个神经元相连。

[0166] 虽然DNN看起来很复杂,但是就每一层的工作来说,其实并不复杂,简单来说就是如下线性关系表达式: $\vec{y} = \alpha(\mathbf{W} \cdot \vec{x} + \vec{b})$ ,其中, $\vec{x}$ 是输入向量, $\vec{y}$ 是输出向量, $\vec{b}$ 是偏移向量, $\mathbf{W}$ 是权重矩阵(也称系数), $\alpha(\cdot)$ 是激活函数。每一层仅仅是对输入向量 $\vec{x}$ 经过如此简单的操作得到输出向量 $\vec{y}$ 。由于DNN层数多,系数 $\mathbf{W}$ 和偏移向量 $\vec{b}$ 的数量也比较多。这些参数在DNN中的定义如下所述:以系数 $\mathbf{W}$ 为例:假设在一个三层的DNN中,第二层的第4个神经元到第三层的第2个神经元的线性系数定义为 $\mathbf{W}_{24}^3$ 。上标3代表系数 $\mathbf{W}$ 所在的层数,而下标对应的是输出的第三层索引2和输入的第二层索引4。

[0167] 综上,第*L*-1层的第*k*个神经元到第*L*层的第*j*个神经元的系数定义为 $\mathbf{W}_{jk}^L$ 。

[0168] 需要注意的是,输入层是没有 $\mathbf{W}$ 参数的。在深度神经网络中,更多的隐含层让网络更能够刻画现实世界中的复杂情形。理论上而言,参数越多的模型复杂度越高,“容量”也就越大,也就意味着它能完成更复杂的学习任务。训练深度神经网络的也就是学习权重矩阵的过程,其最终目的是得到训练好的深度神经网络的所有层的权重矩阵(由很多层的向量 $\mathbf{W}$ 形成的权重矩阵)。

[0169] (3) 卷积神经网络(convolutional neuron network,CNN)是一种带有卷积结构的深度神经网络。卷积神经网络包含了一个由卷积层和子采样层构成的特征抽取器,该特征抽取器可以看作是滤波器。卷积层是指卷积神经网络中对输入信号进行卷积处理的神经元层。在卷积神经网络的卷积层中,一个神经元可以只与部分邻层神经元连接。一个卷积层中,通常包含若干个特征平面,每个特征平面可以由一些矩形排列的神经单元组成。同一特征平面的神经单元共享权重,这里共享的权重就是卷积核。共享权重可以理解为提取特征的方式与位置无关。卷积核可以以随机大小的矩阵的形式化,在卷积神经网络的训练过程中卷积核可以通过学习得到合理的权重。另外,共享权重带来的直接好处是减少卷积神经网络各层之间的连接,同时又降低了过拟合的风险。

[0170] CNN是一种非常常见的神经网络,下面结合图4重点对CNN的结构进行详细的介绍。如前文的基础概念介绍所述,卷积神经网络是一种带有卷积结构的深度神经网络,是一种深度学习(deep learning)架构,深度学习架构是指通过机器学习的算法,在不同的抽象层级上进行多个层次的学习。作为一种深度学习架构,CNN是一种前馈(feed-forward)人工神经网络,该前馈人工神经网络中的各个神经元可以对输入其中的图像作出响应。

[0171] 如图4所示,卷积神经网络(CNN) 200可以包括输入层210,卷积层/池化层220(其中池化层为可选的),以及全连接层(fully connected layer) 230。

[0172] 卷积层/池化层220:

[0173] 卷积层：

[0174] 如图4所示卷积层/池化层220可以包括如示例221-226层，举例来说：在一种实现中，221层为卷积层，222层为池化层，223层为卷积层，224层为池化层，225为卷积层，226为池化层；在另一种实现方式中，221、222为卷积层，223为池化层，224、225为卷积层，226为池化层。即卷积层的输出可以作为随后的池化层的输入，也可以作为另一个卷积层的输入以继续进行卷积操作。

[0175] 下面将以卷积层221为例，介绍一层卷积层的内部工作原理。

[0176] 卷积层221可以包括很多个卷积算子，卷积算子也称为核，其在图像处理中的作用相当于一个从输入图像矩阵中提取特定信息的过滤器，卷积算子本质上可以是一个权重矩阵，这个权重矩阵通常被预先定义，在对图像进行卷积操作的过程中，权重矩阵通常在输入图像上沿着水平方向一个像素接着一个像素(或两个像素接着两个像素……这取决于步长stride的取值)的进行处理，从而完成从图像中提取特定特征的工作。该权重矩阵的大小应该与图像的大小相关，需要注意的是，权重矩阵的纵深维度(depth dimension)和输入图像的纵深维度是相同的，在进行卷积运算的过程中，权重矩阵会延伸到输入图像的整个深度。因此，和一个单一的权重矩阵进行卷积会产生一个单一纵深维度的卷积化输出，但是大多数情况下不使用单一权重矩阵，而是应用多个尺寸(行×列)相同的权重矩阵，即多个同型矩阵。每个权重矩阵的输出被堆叠起来形成卷积图像的纵深维度，这里的维度可以理解为由上面所述的“多个”来决定。不同的权重矩阵可以用来提取图像中不同的特征，例如一个权重矩阵用来提取图像边缘信息，另一个权重矩阵用来提取图像的特定颜色，又一个权重矩阵用来对图像中不需要的噪点进行模糊化等。该多个权重矩阵尺寸(行×列)相同，经过该多个尺寸相同的权重矩阵提取后的特征图的尺寸也相同，再将提取到的多个尺寸相同的特征图合并形成卷积运算的输出。

[0177] 这些权重矩阵中的权重值在实际应用中需要经过大量的训练得到，通过训练得到的权重值形成的各个权重矩阵可以用来从输入图像中提取信息，从而使得卷积神经网络200进行正确的预测。

[0178] 当卷积神经网络200有多个卷积层的时候，初始的卷积层(例如221)往往提取较多的一般特征，该一般特征也可以称之为低级别的特征；随着卷积神经网络200深度的加深，越往后的卷积层(例如226)提取到的特征越来越复杂，比如高级别的语义之类的特征，语义越高的特征越适用于待解决的问题。

[0179] 池化层：

[0180] 由于常常需要减少训练参数的数量，因此卷积层之后常常需要周期性的引入池化层，在如图4中220所示例的221-226各层，可以是一层卷积层后面跟一层池化层，也可以是多层卷积层后面接一层或多层池化层。在图像处理过程中，池化层的唯一目的就是减少图像的空间大小。池化层可以包括平均池化算子和/或最大池化算子，以用于对输入图像进行采样得到较小尺寸的图像。平均池化算子可以在特定范围内对图像中的像素值进行计算产生平均值作为平均池化的结果。最大池化算子可以在特定范围内取该范围内值最大的像素作为最大池化的结果。另外，就像卷积层中用权重矩阵的大小应该与图像尺寸相关一样，池化层中的运算符也应该与图像的大小相关。通过池化层处理后输出的图像尺寸可以小于输入池化层的图像的尺寸，池化层输出的图像中每个像素点表示输入池化层的图像的对应子

区域的平均值或最大值。

[0181] 全连接层230:

[0182] 在经过卷积层/池化层220的处理后,卷积神经网络200还不足以输出所需要的输出信息。因为如前所述,卷积层/池化层220只会提取特征,并减少输入图像带来的参数。然而为了生成最终的输出信息(所需要的类信息或其他相关信息),卷积神经网络200需要利用全连接层230来生成一个或者一组所需要的类的数量的输出。因此,在全连接层230中可以包括多层隐含层(如图4所示的231、232至23n),该多层隐含层中所包含的参数可以根据具体的任务类型的相关训练数据进行预先训练得到,例如该任务类型可以包括图像识别,图像分类,图像超分辨率重建等等……

[0183] 在全连接层230中的多层隐含层之后,也就是整个卷积神经网络200的最后层为输出层240,该输出层240具有类似分类交叉熵的损失函数,具体用于计算预测误差,一旦整个卷积神经网络200的前向传播(如图4由210至240方向的传播为前向传播)完成,反向传播(如图4由240至210方向的传播为反向传播)就会开始更新前面提到的各层的权重值以及偏差,以减少卷积神经网络200的损失,及卷积神经网络200通过输出层输出的结果和理想结果之间的误差。

[0184] 需要说明的是,如图4所示的卷积神经网络200仅作为一种卷积神经网络的示例,在具体的应用中,卷积神经网络还可以以其他网络模型的形式存在,例如,仅包括图4中所示的网络结构的一部分,比如,本申请实施例中所采用的卷积神经网络可以仅包括输入层210、卷积层/池化层220和输出层240。

[0185] 需要说明的是,如图4所示的卷积神经网络100仅作为一种卷积神经网络的示例,在具体的应用中,卷积神经网络还可以以其他网络模型的形式存在,例如,如图5所示的多个卷积层/池化层并行,将分别提取的特征均输入给全连接层230进行处理。

[0186] (4) 损失函数

[0187] 在训练深度神经网络的过程中,因为希望深度神经网络的输出尽可能的接近真正想要预测的值,所以可以通过比较当前网络的预测值和真正想要的值,再根据两者之间的差异情况来更新每一层神经网络的权重向量(当然,在第一次更新之前通常会有初始化的过程,即为深度神经网络中的各层预先配置参数),比如,如果网络的预测值高了,就调整权重向量让它预测低一些,不断地调整,直到深度神经网络能够预测出真正想要的值或与真正想要的值非常接近的值。因此,就需要预先定义“如何比较预测值和值之间的差异”,这便是损失函数(loss function)或函数(objective function),它们是用于衡量预测值和值的差异的重要方程。其中,以损失函数举例,损失函数的输出值(loss)越高表示差异越大,那么深度神经网络的训练就变成了尽可能缩小这个loss的过程。

[0188] (5) 反向传播算法

[0189] 神经网络可以采用误差反向传播(back propagation, BP)算法在训练过程中修正初始的神经网络模型中参数的大小,使得神经网络模型的重建误差损失越来越小。具体地,前向传递输入信号直至输出会产生误差损失,通过反向传播误差损失信息来更新初始的神经网络模型中参数,从而使误差损失收敛。反向传播算法是以误差损失为主导的反向传播运动,旨在得到最优的神经网络模型的参数,例如权重矩阵。

[0190] (6) 无损压缩:对数据进行压缩的技术,压缩后的数据长度小于原始数据长度。压

缩后的数据通过解压,恢复的数据必须与原始数据完全相同。

[0191] (7) 压缩长度:压缩后的数据所占的存储空间。

[0192] (8) 压缩率:原始数据长度和压缩后数据长度的比值。如果没有压缩,值为1。该值越大越好。

[0193] (9) 每维比特数:压缩后的数据每个维度(字节)的平均比特长度。计算公式为:8/压缩率。如果没有压缩,该值为8。该值越小越好。

[0194] (10) 吞吐率:平均每秒处理的数据量大小。

[0195] (11) 隐变量:一种具有特定概率分布的数据,通过建立这些数据与原始数据的条件概率,能够得到原始数据的概率分布。

[0196] (12) 编码/解码:对数据压缩的过程是编码,解压的过程是解码。

[0197] (13) 反编码:一种特殊的编码技术,利用系统中存储的额外二进制数据用解码生成特定的数据。

[0198] 下面结合图6对本申请实施例提供的系统架构进行详细的介绍。图6为本申请一实施例提供的系统架构示意图。如图6所示,系统架构500包括执行设备510、训练设备520、数据库530、客户设备540、数据存储系统550以及数据采集系统560。

[0199] 执行设备510包括计算模块511、I/O接口512、预处理模块513和预处理模块514。计算模块511中可以包括目标模型/规则501,预处理模块513和预处理模块514是可选的。

[0200] 作为一种示例,所述执行设备510可以为手机、平板、笔记本电脑、智能穿戴设备等,终端设备可以对获取到的图片进行压缩处理。作为另一示例,所述终端设备可以为虚拟现实(virtual reality,VR)设备。作为另一示例,本申请实施例也可以应用于智能监控中,可以在所述智能监控中配置相机,则智能监控可以通过相机获取待压缩图片等,应当理解,本申请实施例还可以应用于其他需要进行图像压缩的场景中,此处不再对其他应用场景进行一一列举。

[0201] 数据采集设备560用于采集训练数据。在采集到训练数据之后,数据采集设备560将这些训练数据存入数据库530,训练设备520基于数据库530中维护的训练数据训练得到目标模型/规则501。

[0202] 上述目标模型/规则501(例如本申请实施例中的变分自编码器、熵编码器等)能够用于实现数据的压缩以及解压缩任务,即,将待处理数据(例如本申请实施例中的第一目标数据)输入该目标模型/规则501,即可得到压缩后的数据(例如本申请实施例中的第二比特流)。需要说明的是,在实际应用中,数据库530中维护的训练数据不一定都来自于数据采集设备560的采集,也有可能是从其他设备接收得到的。另外需要说明的是,训练设备520也不一定完全基于数据库530维护的训练数据进行目标模型/规则501的训练,也有可能从云端或其他地方获取训练数据进行模型训练,上述描述不应该作为对本申请实施例的限定。

[0203] 根据训练设备520训练得到的目标模型/规则501可以应用于不同的系统或设备中,如应用于图6所示的执行设备510,所述执行设备510可以是终端,如手机终端,平板电脑,笔记本电脑,增强现实(augmented reality,AR)/虚拟现实(virtual reality,VR)设备,车载终端等,还可以是服务器或者云端等。在图6中,执行设备510配置输入/输出(input/output,I/O)接口512,用于与外部设备进行数据交互,用户可以通过客户设备540向I/O接口512输入数据。

[0204] 预处理模块513和预处理模块514用于根据I/O接口512接收到的输入数据进行预处理。应理解,可以没有预处理模块513和预处理模块514或者仅有的一个预处理模块。当不存在预处理模块513和预处理模块514时,可以直接采用计算模块511对输入数据进行处理。

[0205] 在执行设备510对输入数据进行预处理,或者在执行设备510的计算模块511执行计算等相关的处理过程中,执行设备510可以调用数据存储系统550中的数据、代码等以用于相应的处理,也可以将相应处理得到的数据、指令等存入数据存储系统550中。

[0206] 最后,I/O接口512将处理结果,呈现给客户设备540,从而提供给用户。

[0207] 在图6所示情况下,用户可以手动给定输入数据,该“手动给定输入数据”可以通过I/O接口512提供的界面进行操作。另一种情况下,客户设备540可以自动地向I/O接口512发送输入数据,如果要求客户设备540自动发送输入数据需要获得用户的授权,则用户可以在客户设备540中设置相应权限。用户可以在客户设备540查看执行设备510输出的结果,具体的呈现形式可以是显示、声音、动作等具体方式。客户设备540也可以作为数据采集端,采集如图所示输入I/O接口512的输入数据及输出I/O接口512的输出结果作为新的样本数据,并存入数据库530。当然,也可以不经过客户设备540进行采集,而是由I/O接口512直接将如图所示输入I/O接口512的输入数据及输出I/O接口512的输出结果,作为新的样本数据存入数据库530。

[0208] 值得注意的是,图6仅是本申请实施例提供的一种系统架构的示意图,图中所示设备、器件、模块等之间的位置关系不构成任何限制,例如,在图6中,数据存储系统550相对执行设备510是外部存储器,在其它情况下,也可以将数据存储系统550置于执行设备510中。

[0209] 下面介绍本申请实施例提供的一种芯片硬件结构。

[0210] 图7为本申请一实施例提供的芯片硬件结构图,该芯片包括神经网络处理器700。该芯片可以被设置在如图6所示的执行设备510中,用以完成计算模块511的计算工作。该芯片也可以被设置在如图6所示的训练设备520中,用以完成训练设备520的训练工作并输出目标模型/规则501。如图6所示的图像处理模型中各层的算法均可在如图7所示的芯片中得以实现。

[0211] 神经网络处理器(neural processing unit,NPU)700作为协处理器挂载到主中央处理单元(host central processing unit,host CPU)上,由主CPU分配任务。NPU的核心部分为运算电路703,控制器704控制运算电路703提取存储器(权重存储器702或输入存储器701)中的数据并进行运算。

[0212] 在一些实现中,运算电路703内部包括多个处理单元(process engine,PE)。在一些实现中,运算电路703是二维脉动阵列。运算电路703还可以是一维脉动阵列或者能够执行例如乘法和加法这样的数学运算的其它电子线路。在一些实现中,运算电路703是通用的矩阵处理器。

[0213] 举例来说,假设有输入矩阵A,权重矩阵B,输出矩阵C。运算电路703从权重存储器702中取矩阵B相应的数据,并缓存在运算电路703中每一个PE上。运算电路703从输入存储器701中取矩阵A数据与矩阵B进行矩阵运算,得到的矩阵的部分结果或最终结果,保存在累加器(accumulator)708中。

[0214] 向量计算单元707可以对运算电路703的输出做进一步处理,如向量乘,向量加,指数运算,对数运算,大小比较等等。例如,向量计算单元707可以用于神经网络中非卷积/非

FC层的网络计算,如池化(pooling),批归一化(batch normalization),局部响应归一化(local response normalization)等。

[0215] 在一些实现中,向量计算单元707能将经处理的输出的向量存储到统一存储器706。例如,向量计算单元707可以将非线性函数应用到运算电路703的输出,例如累加值的向量,用以生成激活值。在一些实现中,向量计算单元707生成归一化的值、合并值,或二者均有。在一些实现中,处理过的输出的向量能够用作到运算电路703的激活输入,例如用于在神经网络中的后续层中的使用。

[0216] 统一存储器706用于存放输入数据以及输出数据。

[0217] 权重数据直接通过存储单元访问控制器(direct memory access controller, DMAC) 705将外部存储器中的输入数据搬运到输入存储器701和/或统一存储器706、将外部存储器中的权重数据存入权重存储器702,以及将统一存储器706中的数据存入外部存储器。

[0218] 总线接口单元(bus interface unit, BIU) 710,用于通过总线实现主CPU、DMAC和取指存储器709之间进行交互。

[0219] 与控制器704连接的取指存储器(instruction fetch buffer) 709,用于存储控制器704使用的指令。

[0220] 控制器704,用于调用取指存储器709中缓存的指令,实现控制该运算加速器的工作过程。

[0221] 一般地,统一存储器706、输入存储器701、权重存储器702以及取指存储器709均为片上(on-chip)存储器,外部存储器为该NPU外部的存储器,该外部存储器可以为双倍数据率同步动态随机存储器(double data rate synchronous dynamic random access memory, DDR SDRAM)、高带宽存储器(high bandwidth memory, HBM)或其他可读可写的存储器。

[0222] 数据的无损压缩是信息技术领域的重要基础方向之一。其目的为建立原数据空间到编码空间的双射,使得出现频率较高且较长的数据被较短的编码表示,从而在平均意义上获得更短的数据表示长度,并且能根据该双射在原数据空间和编码空间之间实现一对一的转换。根据香农信源编码定理,数据的最优无损压缩长度由数据概率分布的香农信息熵决定;并且对数据概率分布估计地越准确,越能得到接近最优无损压缩长度。

[0223] 基于人工智能的无损压缩方案利用了深度生成模型能够比传统的方案更准确地估计数据的概率分布这一特性,得到了远优于传统无损压缩方案的压缩比。在基于人工智能的无损压缩方案中,被广泛使用的深度生成模型包括自回归模型(autoregressive Models),变分自编码器(variational auto-encoder, VAE),流模型(normalizing flows)等。一般来讲,自回归模型可较好地兼容算术编码器和霍夫曼编码;变分自编码器结合使用反编码(bits-back)机制可较好地兼容非对称数字系统;流模型可以兼容上述三种不同的熵编码器。除了压缩比以外,评价无损压缩解决方案的还有吞吐率这一指标。对于基于人工智能的无损压缩解决方案来说,由于模型规模远大于传统方案,因此整体吞吐率低于传统方案。另外,综合压缩比和吞吐率两个指标来说,基于不同生成模型的无损压缩解决方案目前没有绝对的先后之分。目前的研究尚处于对不同生成模型的压缩方案探索其帕累托前沿的阶段。

[0224] 其中,区别于全观测模型(如自回归模型),变分自编码器模型是一种隐变量模型。该类模型并非对数据数据本身直接建模,而是额外引入了一个(或者多个)隐变量,然后对先验分布,似然函数以及近似后验分布进行建模。由于从变分自编码器中无法直接获得数据数据的边际分布,传统的熵编码方式无法直接被沿用。为了能够使用变分自编码器进行数据的无损压缩,基于反编码机制的变分自编码无损压缩方案被提出。bits-back ANS是该方案的原始形式,适用于只包含一个隐变量的变分自编码器模型,并且可以推广适用到包含多个隐变量的变分自编码器模型。

[0225] 以Bits-Back ANS和包含一个隐变量的变分自编码器为例,在包含一个隐变量的变分自编码器中,模型可以分为三个模块,即:先验模块、变分编码器模块和解码器模块。以上三个模块可以用来分别确定以下三个分布的参数,即:隐变量的先验分布,隐变量的似然函数(数据的条件概率分布)和隐变量的近似后验分布。

[0226] 该技术方案中数据的压缩步骤为:

[0227] 1. 获取待压缩数据;

[0228] 2. 获取额外的初始比特数据(比特流1);

[0229] 3. 将待压缩数据输入变分编码器,从而获得隐变量的近似后验分布;根据近似后验分布从比特流1中使用熵编码器解码出一个隐变量的样本,并获得比特流2;

[0230] 4. 将解压出的隐变量样本输入解码器,从而获得数据的条件概率分布;根据数据的条件概率分布使用熵编码器将待压缩数据压缩进比特流2,从而获得比特流3;

[0231] 5. 从先验模块中获取隐变量的先验分布;根据隐变量的先验分布使用熵编码器将上述隐变量的样本压缩进比特流3,从而获得比特流4;

[0232] 6. 输出比特流4作为最终压缩后比特数据。

[0233] 该技术方案中数据的解压步骤为:

[0234] 1. 获取待解压的比特数据(比特流4);

[0235] 2. 从先验模块中获取隐变量的先验分布;根据隐变量的先验分布使用熵编码器从第四比特流中解压出压缩阶段使用的隐变量样本,从而获得比特流3;

[0236] 3. 将上述隐变量样本输入解码器,从而获得数据的条件概率分布;根据数据的条件概率分布使用熵编码器从比特流3中解压出被压缩的数据,并获得比特流2;

[0237] 4. 将解压出的数据输入变分编码器,从而获得隐变量的近似后验分布;根据近似后验分布使用熵编码器压缩上述隐变量样本进入比特流2,从而获得比特流1;

[0238] 5. 输出比特流1作为还原的额外初始比特;

[0239] 6. 输出解压缩出的数据。

[0240] 现行的基于反编码机制的变分自编码器无损压缩方案均需要额外的初始比特用以解压出隐变量的样本。额外的初始比特为随机生成的数据,该数据的大小需要考虑进压缩成本中,且在待串行压缩的数据数量较少时,额外的平均成本较高;且,由于所需的额外初始比特与待压缩数据点的个数成正比,因此无法实现高效的并行压缩。

[0241] 基于以上技术背景,本发明是针对基于变分自编码器的人工智能无损压缩方案进行的改良。本发明改良了该细分领域的两大痛点问题:一是通过引入一种特殊的自回归结构降低了变分自编码器达到相同压缩比所需的参数量,从而提升了吞吐率;二是通过引入一种特殊的的变分编码器和解码器结构和提出新的反编码算法,移除了基于变分自编码器

无损压缩方案之前所必须的随机初始比特,从而实现了该方案的单数据点压缩解压以及高效并行压缩解压。

[0242] 参照图8,图8为本申请实施例提供的一种数据压缩方法的实施示意图,如图8示出的那样,本申请实施例提供的一种数据压缩方法包括:

[0243] 801、获取第一目标数据,所述第一目标数据包括第一子数据和第二子数据。

[0244] 在一种可能的实现中,第一目标数据可以为供压缩的图像数据或者是其他数据(例如文本、视频等),其中,第一目标数据可以是上述终端设备通过摄像头拍摄到的图像(或者是图像的一部分),或者,该第一目标数据还可以是从终端设备内部获得的图像(例如,终端设备的相册中存储的图像,或者,终端设备从云端获取的图片)。应理解,上述第一目标数据可以是具有图像压缩需求的数据,本申请并不对第一目标图像的来源作任何限定。

[0245] 在一种可能的实现中,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的。

[0246] 在一种可能的实现中,所述第一子数据和所述第二子数据为对所述图像块在空间维度或者通道维度上进行数据切分后得到的。其中,对于图像数据来说,包含了一个通道维度(C)和两个空间维度(宽W和高H)。

[0247] 例如,第一目标数据可以包括6个通道,第一子数据可以为第一目标数据中的前三个通道的数据,第二子数据可以为第一目标数据中的后三个通道数据。

[0248] 例如,第一目标数据在空间维度上的尺寸为 $N*N$ ,第一子数据可以为第一目标数据中空间维度在 $(0至N/2)*N$ 的数据,第二子数据可以为第一目标数据中空间维度在 $(N/2至N)*N$ 的数据。

[0249] 应理解,本申请并不限定在对第一目标数据进行数据切分的切分规则。

[0250] 802、根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为所述第二子数据的条件概率分布;

[0251] 在一种可能的实现中,变分自编码器可以包括变分编码器,解码器(例如本申请实施例中的第一解码器和第二解码器)和隐变量的先验分布。

[0252] 在一种可能的实现中,解码器可以由解码器层(例如本申请实施例中的第一卷积神经网络以及第二卷积神经网络)构成,且包含解码器层的个数与变分自编码器中隐变量的个数相同。解码器层的作用是输入更深层的隐变量,输出当前层数据的条件概率分布(当前层数据可以是更浅层的隐变量或者数据数据)。

[0253] 在现有的变分自编码器模型中,变分编码器需要输入整个数据数据以预测隐变量的近似后验分布,解码器中输入隐变量直接预测整个数据数据的条件概率分布。在本申请实施例中,将待压缩的数据分成至少两部分,即:第一子数据和第二子数据。和现有将全部数据输入到变分编码器不同的是,本申请实施例中仅将数据的一部分(第一子数据)输入到变分编码器,来预测隐变量的近似后验分布,且隐变量输入第一解码器后预测第一子数据的条件概率分布;第二子数据的条件概率分布依赖于第一子数据,具体可以由将第一子数据输入第一解码器来确定。

[0254] 接下来介绍本申请实施例中的解码器的结构:

[0255] 在一种可能的实现中,解码器可以实现像素重置操作。

[0256] 接下来介绍本申请实施例中的空间维度到通道维度的像素重置操作：

[0257] 在一种可能的实现中，可以配置参数（记为 $k$ ）和两个在通道维度和空间维度进行像素重置的可逆操作（记为空间转通道操作和通道转空间操作）。其中，参数 $k$ 取值正整数，决定了上述两个可逆操作中输入和输出张量空间维度尺寸变化的比例。对于同一个 $k$ ，上述空间转通道操作和通道转空间操作互逆。

[0258] 对于图像来说，图像数据可以表示为向量，其包含了一个通道维度（ $C$ ）和两个空间维度（宽 $W$ 和高 $H$ ）。由于深度学习技术中数据批处理的特性，其相应的张量表征则多出一个批维度（ $N$ ），即图片数据张量包含四个维度（ $NCHW$ 或者 $NHWC$ ）。以 $HCHW$ 为例，一个大小为 $n_1 * c_1 * h_1 * w_1$ 的张量经过参数为 $k$ 的空间转通道操作可以变为大小为 $n_1 * k^2 * c_1 * h_1 / k * w_1 / k$ 的张量。此处要求 $h_1$ 和 $w_1$ 均可以整除 $k$ 。一个大小为 $n_2 * c_2 * h_2 * w_2$ 的张量经过参数为 $k$ 的通道转空间操作变为大小为 $n_2 * c_2 / k^2 * k * h_2 * k * w_2$ 的张量。可以看出，上述两个操作不改变张量中元素的总个数，只改变元素在张量中的位置。对于不同的像素重置的规则，可以得到不同的像素重置装置。本申请实施例使用的像素重置操作采用通道优先的方式。由于空间转通道和通道转空间两个操作在固定 $k$ 时互逆，图9示出了当 $n$ 取值为1， $h$ 和 $w$ 取值为4， $c$ 取值为3， $k$ 取值为2时，空间转通道的操作效果。

[0259] 以第一解码器为例，在一种可能的实现中，所述第一解码器可以包括第一卷积神经网络和第二卷积神经网络，所述根据所述第一子数据，通过变分自编码器的第一解码器，得到第一概率分布，具体可以包括：对包括所述第二子数据的第二目标数据进行空间维度到通道维度的像素重置操作，以得到第三子数据，所述第二目标数据和所述第一目标数据的尺寸大小一致，所述第三子数据和所述第一子数据在空间维度的尺寸相同；

[0260] 其中，包括所述第二子数据的第二目标数据可以为和第一目标数据尺寸相同的数据，其中，在第一目标数据中，除了第二子数据之外的元素可以被置零（或者其他预设数值），以得到第二目标数据，对第二目标数据进行像素重置操作后，可以将其转化为和第一子数据在空间维度的尺寸相同的第三子数据。例如可以参照图10，其中，当前层变量可以为上述第二目标数据，可以对当前层变量进行像素重置操作。

[0261] 在一种可能的实现中，可以根据所述第一子数据，通过所述第一卷积神经网络，得到第四子数据，所述第四子数据和所述第三子数据在通道维度的尺寸相同。也就是说，可以通过第一卷积神经网络，对第一子数据进行特征提取以及尺寸的变换，以便得到一个赫尔第三子数据在通道维度的尺寸相同的第四子数据。

[0262] 在一种可能的实现中，可以将所述第三子数据和所述第四子数据进行融合，以得到融合后的子数据。可选的，融合方式可以为对应通道的数据替换。

[0263] 在一种可能的实现中，所述将所述第三子数据和所述第四子数据进行融合，具体可以包括：将所述第四子数据中部分通道的数据替换为所述第三子数据中对应通道的数据，以得到融合后的子数据。

[0264] 其中，在计算第二子数据的第 $i+1$ 个通道像素的概率分布时，可以将第三子数据 $z'_{i-1}$ 前 $i$ 个通道替换第四子数据 $z''_i$ 的前 $i$ 个通道。

[0265] 在一种可能的实现中，可以根据所述融合后的子数据，通过所述第二卷积神经网络，得到所述第一概率分布。

[0266] 在一种可能的实现中，还可以将所述融合后的子数据和所述第一子数据沿着通道

维度进行拼接操作 (concat), 以得到拼接后的子数据; 进而, 所述根据所述融合后的子数据, 通过所述第二卷积神经网络, 得到所述第一概率分布, 具体可以包括: 根据所述拼接后的子数据, 通过所述第二卷积神经网络, 得到所述第一概率分布。

[0267] 示例性的, 参照图10, 更深层隐变量为 $z_i$ , 输出为当前层变量 $z_{i-1}$ 的条件概率分布。神经网络一的输入为 $z_i$ , 输出为与 $z_{i-1}$ 大小相同 (包括的元素数量相同) 的张量 $z'_i$ 。张量 $z'_i$ 经过参数为 $k$ 的空间到通道像素重置操作变为和 $z_i$ 空间维度大小相同的张量 $z''_i$ 。由于张量 $z_i$ 与张量 $z''_i$ 空间维度大小相同, 故可以按通道维度进行拼接操作, 得到拼接后的张量 $z'''_i$ 。记张量 $z_{i-1}$ 经过参数为 $k$ 的空间到通道像素重置操作变为和 $z_i$ 空间维度大小相同的张量 $z'_{i-1}$ 。该解码器层一引入自回归结构的方式为: 将张量 $z'''_i$ 输入神经网络二, 得到张量 $z'_{i-1}$ 第一个通道像素的概率分布参数; 将张量 $z'_{i-1}$ 前 $i$ 个通道替换 $z'''_i$ 中来自 $z''_i$ 的前 $i$ 个通道并输入神经网络二, 得到张量 $z'_{i-1}$ 第 $i+1$ 个通道像素的概率分布参数。

[0268] 参照图11, 图11为本申请实施例中的一个变分自编码器的示意, 其中, 解码器一相当于本申请实施例中的第二解码器, 解码器二相当于本申请实施例中的第一解码器。

[0269] 示例性, 如下表1所述, 表1示出了包含一个隐变量的变分自编码器为例的示例性流程。其流程的压缩流程如表1左所示, 解压流程如图1右所示。其隐变量的近似后验分布 $q(z^1|x)$ 的参数由变分编码器输入待压数据后给出。其最深层隐变量的先验分布 $p(z^1_1)$ 的参数由模型中最深层隐变量先验分布模块的参数直接给出。其余的条件概率分布的参数均由相应的解码器层, 通过输入条件数据的值输出。每一个涉及的解码器层结构可以参照上述实施例中关于解码器的描述。表1中 $x_1, \dots, x_{12}$ 为数据 $x$  (包含3个通道) 输入通过通道优先的像素重置 (参数 $k$ 为2) 进行空间维度到通道维度的变换所得的12个通道。对于 $z^1_1, \dots, z^1_{12}$ 同理。

[0270] 表1

	Algorithm 1 SHVC Encoding	Algorithm 2 SHVC Decoding
	<b>Input:</b> data to compress $x$	<b>Input:</b> bit stream $c$
	<b>Step 0:</b> Get auxiliary initial bits $c_0$	<b>Step 1:</b> Decode $z^{(1)}$ with $p(z^{(1)})$
	<b>Step 1:</b> Decode $z^{(1)}$ with $q(z^{(1)} x)$	Decode $z^{(1)}_1$ with $p(z^{(1)}_1)$
	<b>Step 2:</b> Encode $x$ with $p(x z^{(1)})$	Decode $z^{(1)}_2$ with $p(z^{(1)}_2 z^{(1)}_1)$
	Encode $x_{12}$ with $p(x_{12} x_{1:11}, z^{(1)})$	...
	Encode $x_{11}$ with $p(x_{11} x_{1:10}, z^{(1)})$	Decode $z^{(1)}_{10}$ with $p(z^{(1)}_{10} z^{(1)}_{1:9})$
	Encode $x_{10}$ with $p(x_{10} x_{1:9}, z^{(1)})$	Decode $z^{(1)}_{11}$ with $p(z^{(1)}_{11} z^{(1)}_{1:10})$
	...	Decode $z^{(1)}_{12}$ with $p(z^{(1)}_{12} z^{(1)}_{1:11})$
[0271]	Encode $x_2$ with $p(x_2 x_1, z^{(1)})$	<b>Step 2:</b> Decode $x$ with $p(x z^{(1)})$
	Encode $x_1$ with $p(x_1 z^{(1)})$	Decode $x_1$ with $p(x_1 z^{(1)})$
	<b>Step 3:</b> Encode $z^{(1)}$ with $p(z^{(1)})$	Decode $x_2$ with $p(x_2 x_1, z^{(1)})$
	Encode $z^{(1)}_{12}$ with $p(z^{(1)}_{12} z^{(1)}_{1:11})$	...
	Encode $z^{(1)}_{11}$ with $p(z^{(1)}_{11} z^{(1)}_{1:10})$	Decode $x_{10}$ with $p(x_{10} x_{1:9}, z^{(1)})$
	Encode $z^{(1)}_{10}$ with $p(z^{(1)}_{10} z^{(1)}_{1:9})$	Decode $x_{11}$ with $p(x_{11} x_{1:10}, z^{(1)})$
	...	Decode $x_{12}$ with $p(x_{12} x_{1:11}, z^{(1)})$
	Encode $z^{(1)}_2$ with $p(z^{(1)}_2 z^{(1)}_1)$	<b>Step 3:</b> Encode $z^{(1)}$ with $q(z^{(1)} x)$
	Encode $z^{(1)}_1$ with $p(z^{(1)}_1)$	<b>Output:</b> auxiliary initial bit stream $c_0$ , data to decompress $x$
	<b>Output:</b> final bit stream $c$	

[0272] 本申请实施例通过使用基于通道优先像素重置定义自回归结构的编码器层, 充分利用图片像素间的相关关系, 从而在获得更低的编码长度的前提下大幅降低模型所需参数量, 进而提高了压缩的吞吐率以及降低了模型存储的空间成本。

[0273] 803、根据所述第一概率分布, 通过熵编码器压缩所述第二子数据, 以得到第一比特流;

[0274] 在一种可能的实现中,可以根据所述第一概率分布,通过熵编码器压缩所述第二子数据,以得到第一比特流。第一比特流可以作为初始比特流,并对所述第一子数据进行压缩。相比于现有技术中反编码机制所需的额外设置的初始比特,本申请实施例中无需额外设置的初始比特,可以实现单数据点的压缩,且大大降低了并行压缩时的压缩比。

[0275] 804、将所述第一子数据压缩至所述第一比特流。

[0276] 在一种可能的实现中,所述将所述第一比特流作为初始比特流,对所述第一子数据进行压缩,具体可以包括:根据所述第一子数据,通过所述变分编码器中的变分编码器,得到隐变量的近似后验分布;根据所述近似后验分布,从所述第一比特流中通过所述熵编码器得到所述隐变量,得到第三比特流;根据所述隐变量,通过所述变分编码器的第二解码器,得到第二概率分布;所述第二概率分布用于作为所述第一子数据的条件概率分布;根据所述第二概率分布,通过所述熵编码器将所述第一子数据压缩至所述第三比特流,以得到第四比特流;根据所述隐变量的先验分布,通过所述熵编码器将所述隐变量压缩至所述第四比特流,得到第二比特流。

[0277] 示例性的,在编码侧,可以执行如下步骤:

[0278] 1. 获取第一目标数据;

[0279] 2. 将第一目标数据分为第一子数据和第二子数据,将第一子数据输入第一解码器获取第二子数据的条件概率分布;使用第二子数据的条件概率分布压缩第二子数据从而获得初始比特数据(第一比特流);

[0280] 3. 将第一子数据输入变分编码器,从而获得隐变量的近似后验分布;根据近似后验分布从第一比特流中使用熵编码器解码出一个隐变量的样本,并获得第三比特流;

[0281] 4. 将解压出的隐变量样本输入第二解码器,从而获得数据的条件概率分布;根据数据的条件概率分布使用熵编码器将第一子数据压缩进第三比特流,从而获得第四比特流;

[0282] 5. 从先验模块中获取隐变量的先验分布;根据隐变量的先验分布使用熵编码器将上述隐变量的样本压缩进第三比特流,从而获得第二比特流;

[0283] 6. 输出第二比特流作为最终压缩后的比特数据。

[0284] 相应的,在解码段,解压步骤可以为:

[0285] 1. 获取待解压的比特数据(第二比特流);

[0286] 2. 从先验模块中获取隐变量的先验分布;根据隐变量的先验分布使用熵编码器从第二比特流中解压出压缩阶段使用的隐变量样本,从而获得第四比特流;

[0287] 3. 将上述隐变量样本输入解码器一,从而获得子数据一的条件概率分布;根据子数据一的条件概率分布使用熵编码器从第四比特流中解压出被压缩的子数据一,并获得第三比特流;

[0288] 4. 将解压出的子数据一输入变分编码器,从而获得隐变量的近似后验分布;根据近似后验分布使用熵编码器压缩上述隐变量样本进入第三比特流,从而获得第一比特流;

[0289] 5. 将子数据一输入解码器二,从而获得子数据二的条件概率分布;根据子数据二的条件概率分布从第一比特流中解压出子数据二,此时比特流被消耗完毕;

[0290] 6. 将子数据一和子数据二按照压缩时的分离方式逆转,从而得到原数据,输出解压出的原数据(第一目标数据)。

[0291] 参照图12,图12为隐变量数量为1时的数据压缩过程的一个示意,其中S为第一目标数据,S1为第一子数据,S2为第二子数据。

[0292] 为了比较本申请实施例与现有方案的区别,将其流程分别展示于图13。图13左示出了现有方案的压缩解压核心流程;图13右示出了本申请实施例无需额外初始比特的压缩解压核心流程。

[0293] 参照表2,表2为基于通道优先像素重置定义的回归结构的解码器层和无需额外初始比特的变分自编码器无损压缩解决方案的核心方法流程。

[0294] 表2

Algorithm 3 SHVC-ArIB Encoding	Algorithm 4 SHVC-ArIB Decoding
<b>Input:</b> data to compress $x$ <b>Step 0:</b> Get autoregressive initial bits by encoding $x_{7:12}$ Encode $x_{12}$ with $p(x_{12} x_{1:11})$ ... Encode $x_7$ with $p(x_7 x_{1:6})$ <b>Step 1:</b> Decode $z^{(1)}$ with $q(z^{(1)} x_{1:6})$ <b>Step 2:</b> Encode $x_{1:6}$ with $p(x_{1:6} z^{(1)})$ Encode $x_6$ with $p(x_6 x_{1:5},z^{(1)})$ ... Encode $x_1$ with $p(x_1 z^{(1)})$ <b>Step 3:</b> Encode $z^{(1)}$ with $p(z^{(1)})$ Encode $z_{12}^{(1)}$ with $p(z_{12}^{(1)} z_{1:11}^{(1)})$ Encode $z_{11}^{(1)}$ with $p(z_{11}^{(1)} z_{1:10}^{(1)})$ Encode $z_{10}^{(1)}$ with $p(z_{10}^{(1)} z_{1:9}^{(1)})$ ... Encode $z_2^{(1)}$ with $p(z_2^{(1)} z_1^{(1)})$ Encode $z_1^{(1)}$ with $p(z_1^{(1)})$ <b>Output:</b> final bit stream $c$	<b>Input:</b> bit stream $c$ <b>Step 1:</b> Decode $z^{(1)}$ with $p(z^{(1)})$ Decode $z_1^{(1)}$ with $p(z_1^{(1)})$ Decode $z_2^{(1)}$ with $p(z_2^{(1)} z_1^{(1)})$ ... Decode $z_{10}^{(1)}$ with $p(z_{10}^{(1)} z_{1:9}^{(1)})$ Decode $z_{11}^{(1)}$ with $p(z_{11}^{(1)} z_{1:10}^{(1)})$ Decode $z_{12}^{(1)}$ with $p(z_{12}^{(1)} z_{1:11}^{(1)})$ <b>Step 2:</b> Decode $x_{1:6}$ with $p(x_{1:6} z^{(1)})$ Decode $x_1$ with $p(x_1 z^{(1)})$ ... Decode $x_6$ with $p(x_6 x_{1:5},z^{(1)})$ <b>Step 3:</b> Decode $z^{(1)}$ with $q(z^{(1)} x_{1:6})$ <b>Step 4:</b> Decode $x_{7:12}$ with $p(x_{7:12} x_{1:6})$ Decode $x_7$ with $p(x_7 x_{1:6})$ ... Decode $x_{12}$ with $p(x_{12} x_{1:11})$ <b>Output:</b> data to decompress $x$

[0296] 接下来结合实验结果描述本申请实施例的有益效果。

[0297] 在本申请实施例中,由于使用的编码器较好地利用了图片数据的像素间的相关性,能够在比同类型模型无损压缩方案给出更优编码长度的同时,将模型大小减小100倍。

[0298] 表3展示了本方案 (SHVC) 与其他业界最优方案在公开数据集上的平均每维度编码比特数 (bpd)。可以看出本方案效果在所有对比方案中 (包括传统方案,VAE模型方案和流模型方案) 均为最优或者接近最优。在同类型方案中 (VAE based) 为最优。

[0299] 表4展示了本方案除了编码长度优势外,由于模型参数数量的较少,其推理时间大大降低,从而提升了压缩和解压的吞吐率。其数据统计设定为10000张CIFAR10图片,批尺寸为100,硬件为一张V100显卡。

[0300] 表3

	Compression Model	CIFAR10	ImageNet32	ImageNet64	CLIC.mobile	CLIC.pro	DIV2K
[0301]	<i>Generic</i>						
	PNG [4]	5.71	5.87	6.39	3.90	4.00	3.09
	FLIF [35]	4.19	4.19	4.52	2.49	2.78	2.91
	JPEG-XL [1]	5.74	5.89	6.39	2.36	2.63	2.79
[0301]	<i>VAE-Based</i>						
	L3C [22]	-	4.76	4.42	2.64	2.94	3.09
	Bit-Swap [18]	3.82	4.50	-	-	-	-
	HiLLoC [37]	3.56 <sup>‡</sup>	4.20 <sup>‡</sup>	3.90 <sup>‡</sup>	-	-	-
	SHVC	<b>3.16/3.41<sup>‡</sup></b>	<b>3.98</b>	<b>3.68/3.71<sup>‡</sup></b>	<b>1.96*</b>	<b>2.02*</b>	<b>2.57*</b>
	SHVC Lite	<b>3.76</b>	<b>4.49</b>	<b>4.16</b>	-	-	-
[0301]	<i>Flow-Based</i>						
	IDF [14]	3.34/3.60 <sup>‡</sup>	4.18	3.90/3.94 <sup>‡</sup>	-	-	-
	IDF++ [3]	3.26	4.12	3.81	-	-	-
	LBB [13]	3.12	3.88	3.70	-	-	-
	iVPF [44]	3.20/ 3.49 <sup>‡</sup>	4.03	3.75/3.79 <sup>‡</sup>	2.39*	2.54*	2.68*
	iFlow [43]	3.12/3.36 <sup>‡</sup>	3.88	3.70/3.65 <sup>‡</sup>	2.26*	2.44*	2.57*

[0302] 表4

	SHVC	Bit-Swap	HiLLoC	IDF
[0303] BPD	<b>3.18</b>	3.82	3.32	3.34
Time (s)	<b>4.63</b>	5.86	10.20	20.58

[0304] 此外,本申请实施例可以在避免现行反编码机制所需的额外初始比特,实现单数据点的压缩和高效并行压缩。表5中展示了本申请实施例 (SHVC)、使用 (SHVC-ARIB) 以及使用确定性后验 (本质上为自编码器模型) 和无反编码机制的方案 (SHVC-Det) 三种情况下考虑进额外初始比特时的平均每维度编码长度。从表5可以看出,使用本方案可以比现行的反编码压缩算法将额外空间成本降低高达30倍。

[0305] 表5

	SHVC	SHVC-ARIB	SHVC-Det
[0306] CIFAR10	4.18	3.19	3.37
ImageNet32	5.03	4.00	4.17
ImageNet64	4.57	3.71	3.90

[0307] 应理解,本申请实施例也可以用到数据的有损压缩中。

[0308] 本申请实施例提供了一种数据压缩方法,包括:获取第一目标数据,所述第一目标数据包括第一子数据和第二子数据;根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为所述第二子数据的条件概率分布;根据所述第一概率分布,通过熵编码器压缩所述第二子数据,以得到第一比特流;将所述第一比特流作为初始比特流,对所述第一子数据进行压缩,以得到第二比特流。相比于现有技术中反编码机制所需的额外设置的初始比特,本申请实施例中无需额外设置的初始比特,可以实现单数据点的压缩,且大大降低了并行压缩时的压缩比。

[0309] 参照图14,图14为本申请实施例提供的一种数据解压缩方法的流程示意,如图14所示,本申请实施例提供的数据解压缩方法,包括:

[0310] 1401、获取第二比特流以及隐变量的先验分布;

[0311] 1402、根据所述先验分布,通过熵编码器从所述第二比特流中解压出所述隐变量,得到第四比特流;

[0312] 1403、根据所述隐变量,通过所述变分自编码器的第二解码器,得到第二概率分布;所述第二概率分布用于作为第一子数据的条件概率分布;

[0313] 1404、根据所述第二概率分布,通过所述熵编码器从所述第四比特流中解压出所述第一子数据,得到第三比特流;

[0314] 1405、根据所述第一子数据,通过所述变分自编码器中的变分编码器,得到隐变量的近似后验分布;

[0315] 1406、根据所述近似后验分布,通过所述熵编码器将所述隐变量压缩至所述第三比特流,得到第一比特流;

[0316] 1407、根据所述第一子数据,通过所述变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为所述第二子数据的条件概率分布;

[0317] 1408、根据所述第一概率分布,通过所述熵编码器从所述第一比特流中解压出第二子数据;所述第一子数据和所述第二子数据用于确定第一目标数据。

[0318] 在一种可能的实现中,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的。

[0319] 在一种可能的实现中,所述第一子数据和所述第二子数据为对所述图像块在空间维度或者通道维度上进行数据切分后得到的。

[0320] 在一种可能的实现中,所述第一解码器包括第一卷积神经网络和第二卷积神经网络,所述根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,包括:

[0321] 对包括所述第二子数据的第二目标数据进行空间维度到通道维度的像素重置操作,以得到第三子数据,所述第二目标数据和所述第一目标数据的尺寸大小一致,所述第三子数据和所述第一子数据在空间维度的尺寸相同;

[0322] 根据所述第一子数据,通过所述第一卷积神经网络,得到第四子数据,所述第四子数据和所述第三子数据在通道维度的尺寸相同;

[0323] 将所述第三子数据和所述第四子数据进行融合,以得到融合后的子数据;

[0324] 根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

[0325] 在一种可能的实现中,所述将所述第三子数据和所述第四子数据进行融合,包括:

[0326] 将所述第四子数据中部分通道的数据替换为所述第三子数据中对应通道的数据,以得到融合后的子数据。

[0327] 在一种可能的实现中,所述方法还包括:

[0328] 将所述融合后的子数据和所述第一子数据沿着通道维度进行拼接操作,以得到拼接后的子数据;

[0329] 所述根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布,包括:根据所述拼接后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

[0330] 关于数据解压缩方法的描述,可以参照图8以及对应的实施例中关于数据解压缩相关的描述,这里不再赘述。

[0331] 在图1至图14所对应的实施例的基础上,为了更好的实施本申请实施例的上述方案,下面还提供用于实施上述方案的相关设备。具体参阅图15,图15为本申请实施例提供的数据压缩装置1500的一种结构示意图,数据压缩装置1500可以是终端设备或服务器,数据压缩装置1500包括:

[0332] 获取模块1501,用于获取第一目标数据,所述第一目标数据包括第一子数据和第

二子数据；

[0333] 关于获取模块1501的具体描述可以参照上述实施例中步骤801的描述,这里不再赘述。

[0334] 压缩模块1502,用于根据所述第一子数据,通过变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为所述第二子数据的条件概率分布;

[0335] 根据所述第一概率分布,通过熵编码器压缩所述第二子数据,以得到第一比特流;

[0336] 将所述第一比特流作为初始比特流,对所述第一子数据进行压缩,以得到第二比特流。

[0337] 关于压缩模块1502的具体描述可以参照上述实施例中步骤802至步骤804的描述,这里不再赘述。

[0338] 在一种可能的实现中,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的。

[0339] 在一种可能的实现中,所述第一子数据和所述第二子数据为对所述图像块在空间维度或者通道维度上进行数据切分后得到的。

[0340] 在一种可能的实现中,所述压缩模块,具体用于:

[0341] 根据所述第一子数据,通过所述变分自编码器中的变分编码器,得到隐变量的近似后验分布;

[0342] 根据所述近似后验分布,从所述第一比特流中通过所述熵编码器得到所述隐变量,得到第三比特流;

[0343] 根据所述隐变量,通过所述变分自编码器的第二解码器,得到第二概率分布;所述第二概率分布用于作为所述第一子数据的条件概率分布;

[0344] 根据所述第二概率分布,通过所述熵编码器将所述第一子数据压缩至所述第三比特流,以得到第四比特流;

[0345] 根据所述隐变量的先验分布,通过所述熵编码器将所述隐变量压缩至所述第四比特流,得到第二比特流。

[0346] 在一种可能的实现中,所述第一解码器包括第一卷积神经网络和第二卷积神经网络,所述压缩模块,具体用于:

[0347] 对包括所述第二子数据的第二目标数据进行空间维度到通道维度的像素重置操作,以得到第三子数据,所述第二目标数据和所述第一目标数据的尺寸大小一致,所述第三子数据和所述第一子数据在空间维度的尺寸相同;

[0348] 根据所述第一子数据,通过所述第一卷积神经网络,得到第四子数据,所述第四子数据和所述第三子数据在通道维度的尺寸相同;

[0349] 将所述第三子数据和所述第四子数据进行融合,以得到融合后的子数据;

[0350] 根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

[0351] 在一种可能的实现中,所述将所述第三子数据和所述第四子数据进行融合,包括:

[0352] 将所述第四子数据中部分通道的数据替换为所述第三子数据中对应通道的数据,以得到融合后的子数据。

[0353] 在一种可能的实现中,所述装置还包括:

[0354] 拼接模块,用于将所述融合后的子数据和所述第一子数据沿着通道维度进行拼接

操作,以得到拼接后的子数据;

[0355] 所述根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布,包括:根据所述拼接后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

[0356] 参阅图16,图16为本申请实施例提供的数据解压缩装置1600的一种结构示意图,数据解压缩装置1600可以是终端设备或服务器,数据解压缩装置1600可以包括:

[0357] 获取模块1601,用于获取第二比特流以及隐变量的先验分布;

[0358] 关于获取模块1601的具体描述可以参照上述实施例中步骤1401的描述,这里不再赘述。

[0359] 解压模块1602,用于根据所述先验分布,通过熵编码器从所述第二比特流中解压出所述隐变量,得到第四比特流;

[0360] 根据所述隐变量,通过所述变分自编码器的第二解码器,得到第二概率分布;所述第二概率分布用于作为第一子数据的条件概率分布;

[0361] 根据所述第二概率分布,通过所述熵编码器从所述第四比特流中解压出所述第一子数据,得到第三比特流;

[0362] 根据所述第一子数据,通过所述变分自编码器中的变分编码器,得到隐变量的近似后验分布;

[0363] 根据所述近似后验分布,通过所述熵编码器将所述隐变量压缩至所述第三比特流,得到第一比特流;

[0364] 根据所述第一子数据,通过所述变分自编码器的第一解码器,得到第一概率分布,所述第一概率分布用于作为所述第二子数据的条件概率分布;

[0365] 根据所述第一概率分布,通过所述熵编码器从所述第一比特流中解压出第二子数据;所述第一子数据和所述第二子数据用于确定第一目标数据。

[0366] 关于解压模块1602的具体描述可以参照上述实施例中步骤1402至步骤1408的描述,这里不再赘述。

[0367] 在一种可能的实现中,所述第一目标数据为图像块,所述第一子数据和所述第二子数据为对所述图像块进行数据切分后得到的。

[0368] 在一种可能的实现中,所述第一子数据和所述第二子数据为对所述图像块在空间维度或者通道维度上进行数据切分后得到的。

[0369] 在一种可能的实现中,所述第一解码器包括第一卷积神经网络和第二卷积神经网络,所述解压模块,具体用于:

[0370] 对包括所述第二子数据的第二目标数据进行空间维度到通道维度的像素重置操作,以得到第三子数据,所述第二目标数据和所述第一目标数据的尺寸大小一致,所述第三子数据和所述第一子数据在空间维度的尺寸相同;

[0371] 根据所述第一子数据,通过所述第一卷积神经网络,得到第四子数据,所述第四子数据和所述第三子数据在通道维度的尺寸相同;

[0372] 将所述第三子数据和所述第四子数据进行融合,以得到融合后的子数据;

[0373] 根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

[0374] 在一种可能的实现中,所述将所述第三子数据和所述第四子数据进行融合,包括:

[0375] 将所述第四子数据中部分通道的数据替换为所述第三子数据中对应通道的数据,以得到融合后的子数据。

[0376] 在一种可能的实现中,所述装置还包括:

[0377] 拼接模块,用于将所述融合后的子数据和所述第一子数据沿着通道维度进行拼接操作,以得到拼接后的子数据;

[0378] 所述根据所述融合后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布,包括:根据所述拼接后的子数据,通过所述第二卷积神经网络,得到所述第一概率分布。

[0379] 接下来介绍本申请实施例提供的一种执行设备,请参阅图17,图17为本申请实施例提供的执行设备的一种结构示意图,执行设备1700具体可以表现为虚拟现实VR设备、手机、平板、笔记本电脑、智能穿戴设备、监控数据处理设备等,此处不做限定。其中,执行设备1700上可以部署有图15对应实施例中所描述的数据压缩装置、或者图16对应实施例中所描述的数据解压缩装置。具体的,执行设备1700可以包括:接收器1701、发射器1702、处理器1703和存储器1704(其中执行设备1700中的处理器1703的数量可以一个或多个,图15中以一个处理器为例),其中,处理器1703可以包括应用处理器17031和通信处理器17032。在本申请的一些实施例中,接收器1701、发射器1702、处理器1703和存储器1704可通过总线或其它方式连接。

[0380] 存储器1704可以包括只读存储器和随机存取存储器,并向处理器1703提供指令和数据。存储器1704的一部分还可以包括非易失性随机存取存储器(non-volatile random access memory,NVRAM)。存储器1704存储有处理器和操作指令、可执行模块或者数据结构,或者它们的子集,或者它们的扩展集,其中,操作指令可包括各种操作指令,用于实现各种操作。

[0381] 处理器1703控制执行设备的操作。具体的应用中,执行设备的各个组件通过总线系统耦合在一起,其中总线系统除包括数据总线之外,还可以包括电源总线、控制总线和状态信号总线等。但是为了清楚说明起见,在图中将各种总线都称为总线系统。

[0382] 上述本申请实施例揭示的方法可以应用于处理器1703中,或者由处理器1703实现。处理器1703可以是一种集成电路芯片,具有信号的处理能力。在实现过程中,上述方法的各步骤可以通过处理器1703中的硬件的集成逻辑电路或者软件形式的指令完成。上述的处理器1703可以是通用处理器、数字信号处理器(digital signal processing,DSP)、微处理器或微控制器,还可进一步包括专用集成电路(application specific integrated circuit,ASIC)、现场可编程门阵列(field-programmable gate array,FPGA)或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件。该处理器1703可以实现或者执行本申请实施例中的公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者该处理器也可以是任何常规的处理器等。结合本申请实施例所公开的方法的步骤可以直接体现为硬件译码处理器执行完成,或者用译码处理器中的硬件及软件模块组合执行完成。软件模块可以位于随机存储器,闪存、只读存储器,可编程只读存储器或者电可擦写可编程存储器、寄存器等本领域成熟的存储介质中。该存储介质位于存储器1704,处理器1703读取存储器1704中的信息,结合其硬件完成上述方法的步骤。

[0383] 接收器1701可用于接收输入的数字或字符信息,以及产生与执行设备的相关设置

以及功能控制有关的信号输入。发射器1702可用于通过第一接口输出数字或字符信息；发射器1702还可用于通过第一接口向磁盘组发送指令，以修改磁盘组中的数据；发射器1702还可以包括显示屏等显示设备。

[0384] 本申请实施例中还提供一种包括计算机程序产品，当其在计算机上运行时，使得计算机执行如前述图8所示实施例描述的方法所执行的步骤，或者，使得计算机执行如前述图14所示实施例描述的方法所执行的步骤。

[0385] 本申请实施例中还提供一种计算机可读存储介质，该计算机可读存储介质中存储有用于进行信号处理的程序，当其在计算机上运行时，使得计算机执行如前述图8所示实施例描述的方法所执行的步骤，或者，使得计算机执行如前述图14所示实施例描述的方法所执行的步骤。

[0386] 另外需说明的是，以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的单元可以是或者也可以不是物理上分开的，作为单元显示的部件可以是或者也可以不是物理单元，即可以位于一个地方，或者也可以分布到多个网络单元上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。另外，本申请提供的装置实施例附图中，模块之间的连接关系表示它们之间具有通信连接，具体可以实现为一条或多条通信总线或信号线。

[0387] 通过以上的实施方式的描述，所属领域的技术人员可以清楚地了解到本申请可借助软件加必需的通用硬件的方式来实现，当然也可以通过专用硬件包括专用集成电路、专用CPU、专用存储器、专用元器件等来实现。一般情况下，凡由计算机程序完成的功能都可以很容易地用相应的硬件来实现，而且，用来实现同一功能的具体硬件结构也可以是多种多样的，例如模拟电路、数字电路或专用电路等。但是，对本申请而言更多情况下软件程序实现是更佳实施方式。基于这样的理解，本申请的技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品存储在可读取的存储介质中，如计算机的软盘、U盘、移动硬盘、ROM、RAM、磁碟或者光盘等，包括若干指令用以使得一台计算机设备（可以是个人计算机，训练设备，或者网络设备等）执行本申请各个实施例所述的方法。

[0388] 在上述实施例中，可以全部或部分地通过软件、硬件、固件或者其任意组合来实现。当使用软件实现时，可以全部或部分地以计算机程序产品的形式实现。

[0389] 所述计算机程序产品包括一个或多个计算机指令。在计算机上加载和执行所述计算机程序指令时，全部或部分地产生按照本申请实施例所述的流程或功能。所述计算机可以是通用计算机、专用计算机、计算机网络、或者其他可编程装置。所述计算机指令可以存储在计算机可读存储介质中，或者从一个计算机可读存储介质向另一计算机可读存储介质传输，例如，所述计算机指令可以从一个网站站点、计算机、训练设备或数据中心通过有线（例如同轴电缆、光纤、数字用户线（DSL））或无线（例如红外、无线、微波等）方式向另一个网站站点、计算机、训练设备或数据中心进行传输。所述计算机可读存储介质可以是计算机能够存储的任何可用介质或者是包含一个或多个可用介质集成的训练设备、数据中心等数据存储设备。所述可用介质可以是磁性介质，（例如，软盘、硬盘、磁带）、光介质（例如，DVD）、或者半导体介质（例如固态硬盘（Solid State Disk,SSD））等。

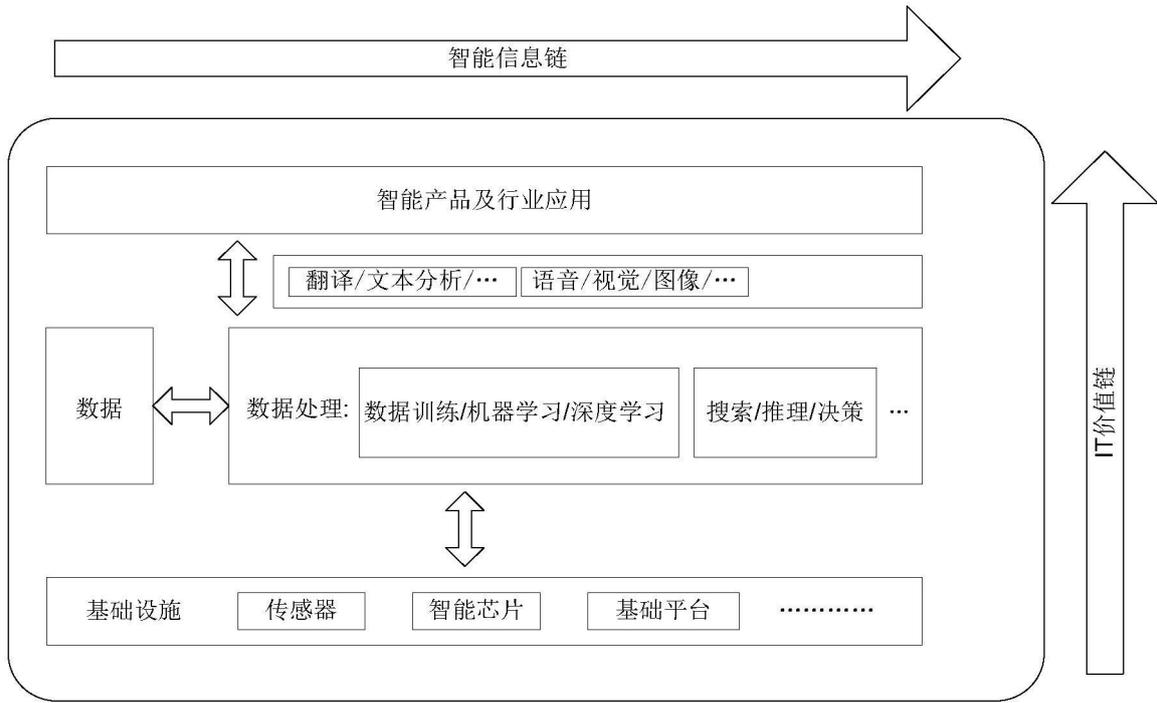


图1

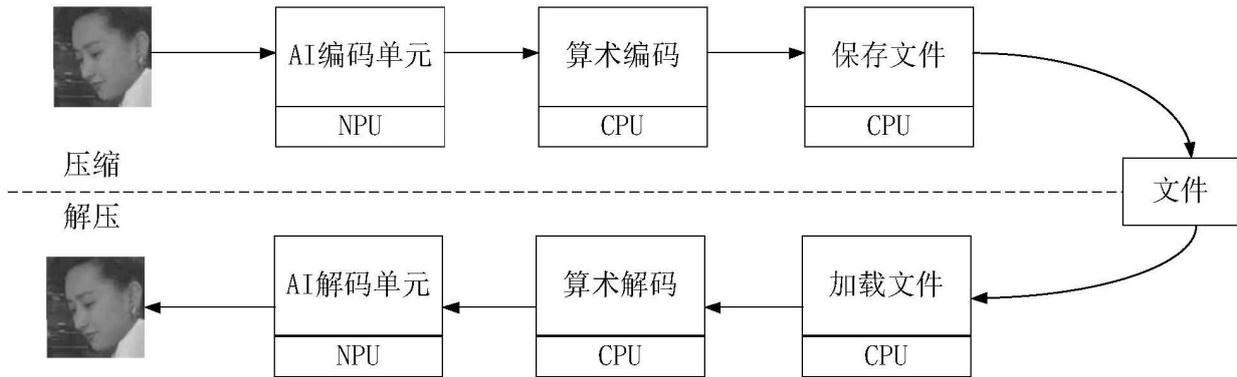


图2

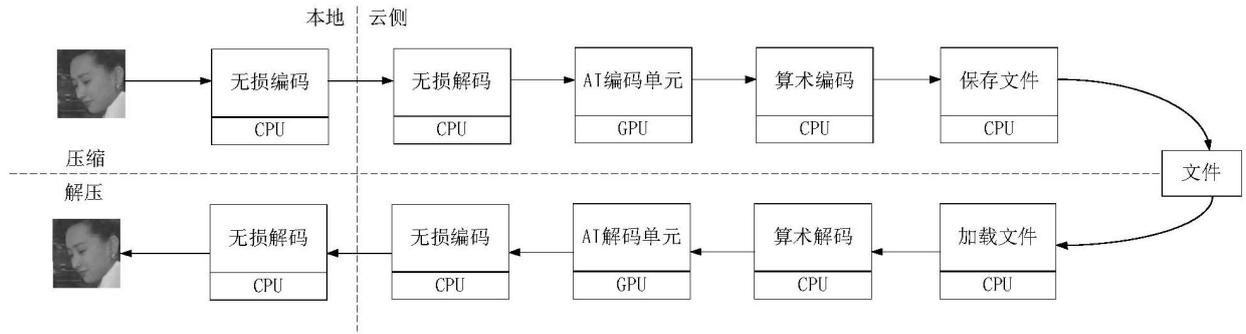


图3

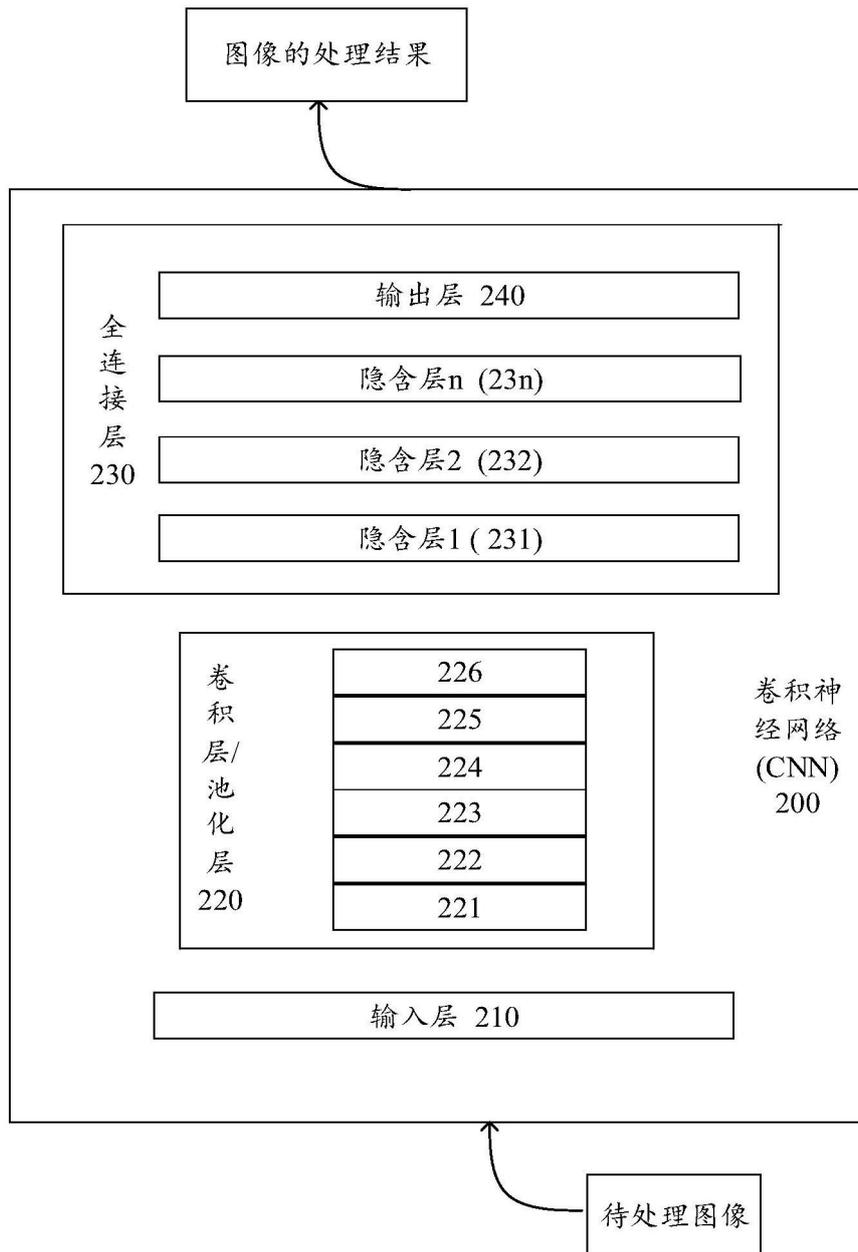


图4

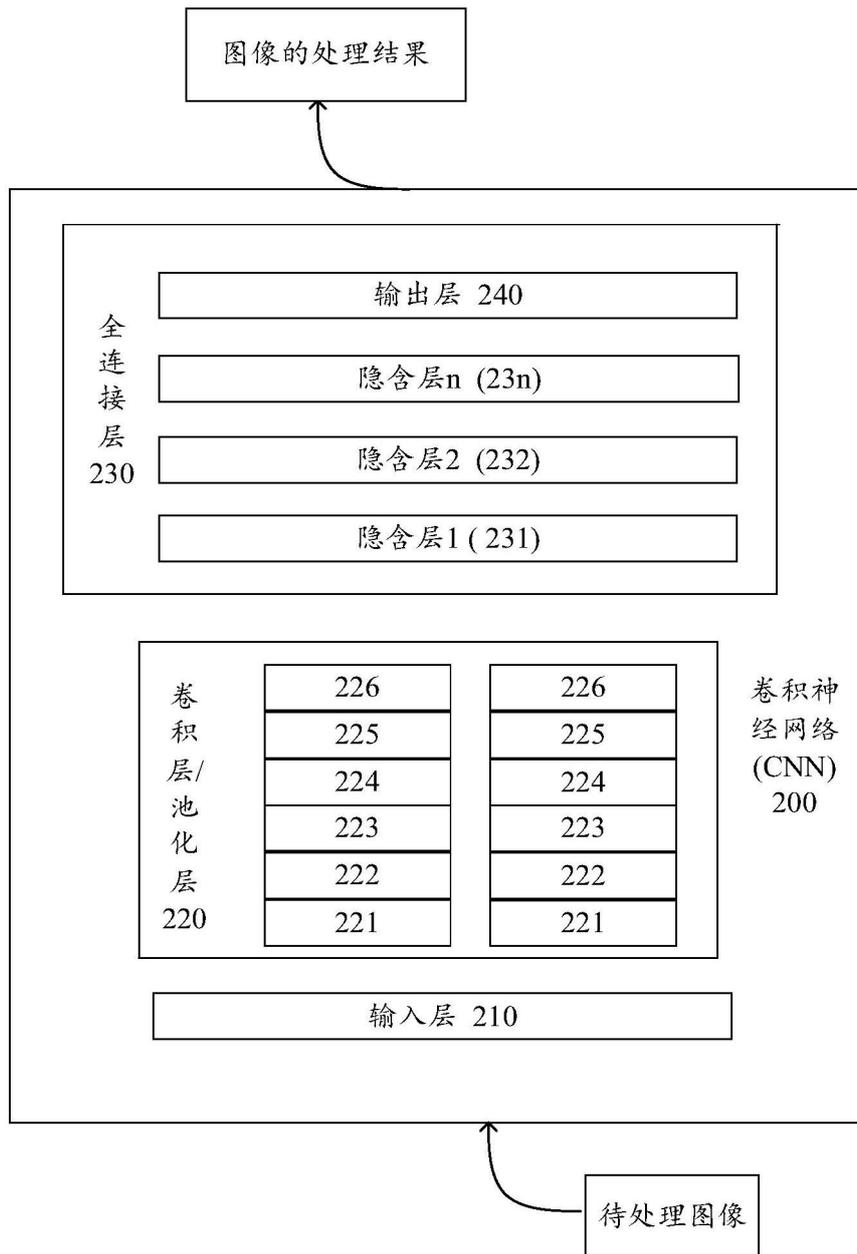


图5

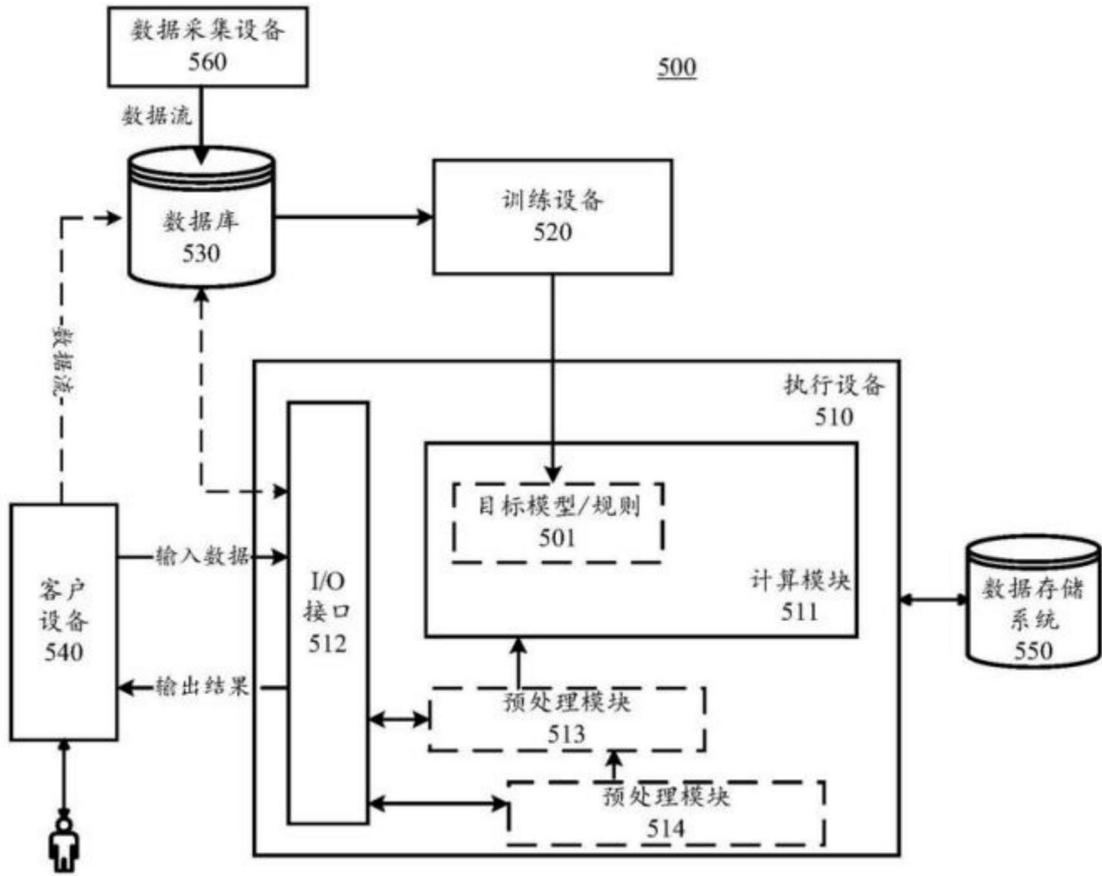


图6

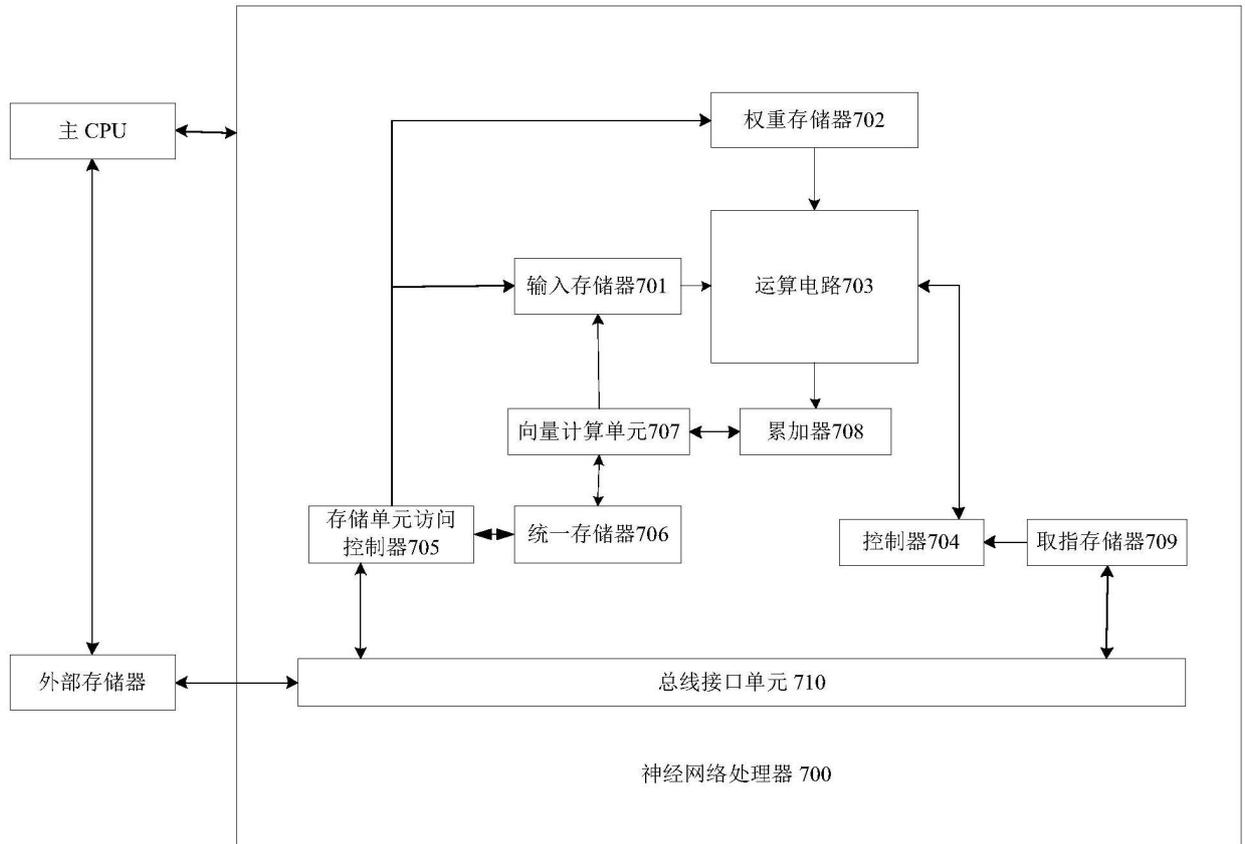


图7

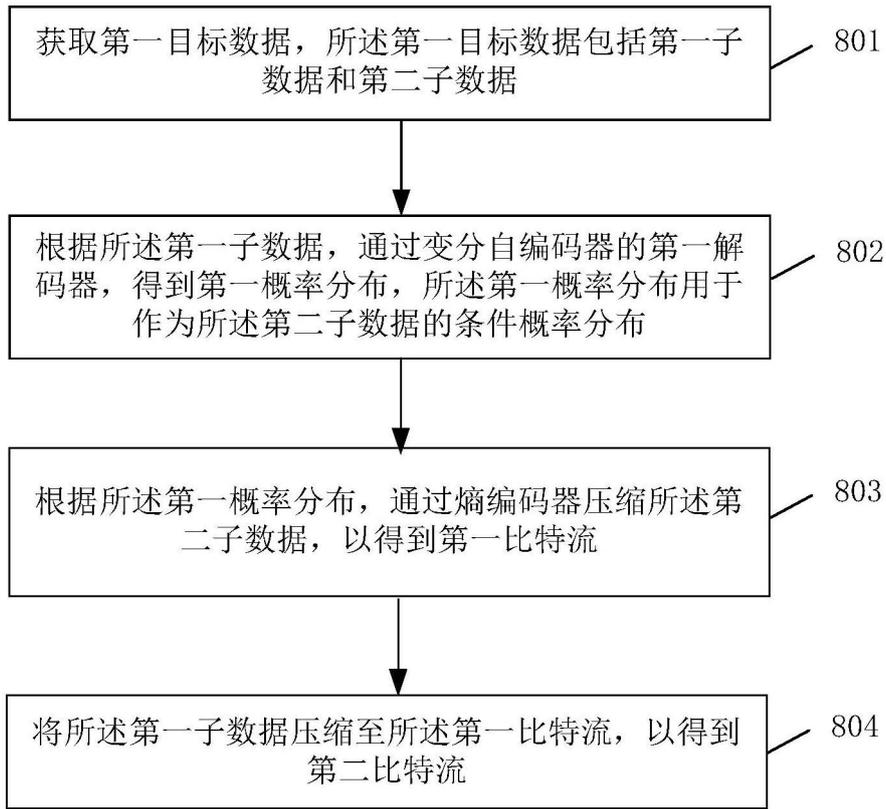


图8

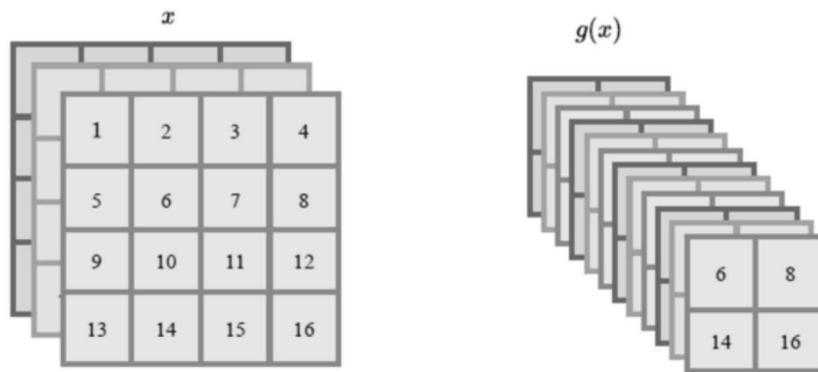


图9

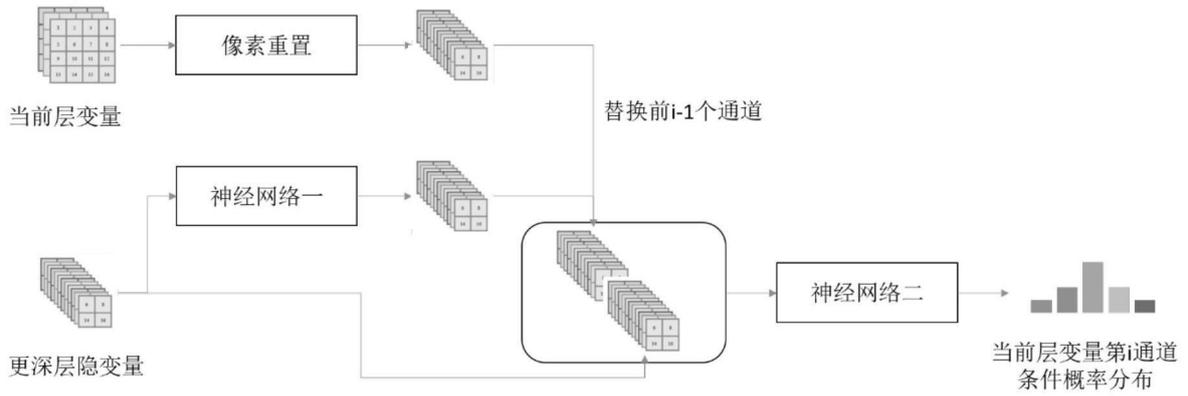


图10

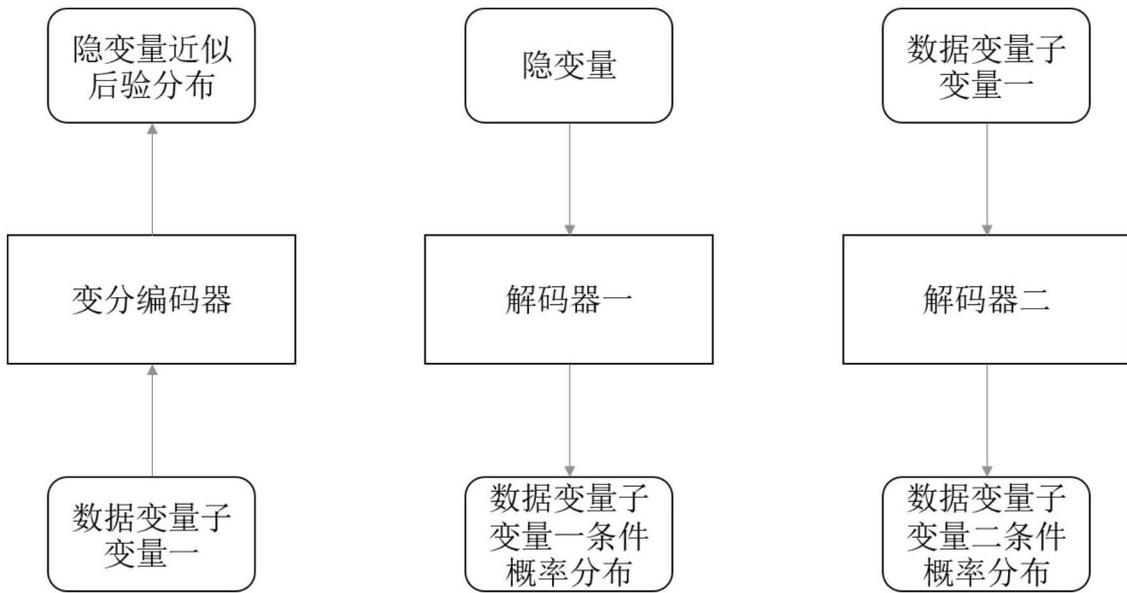


图11



图12

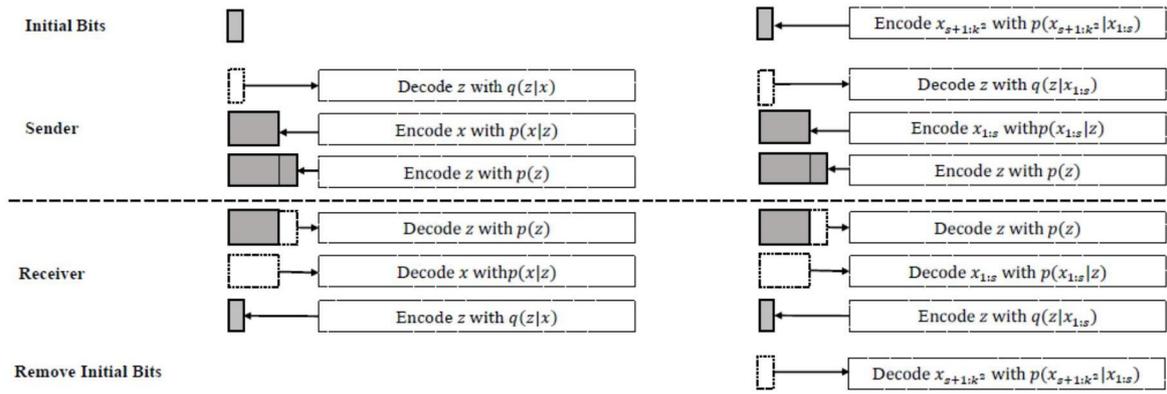


图13

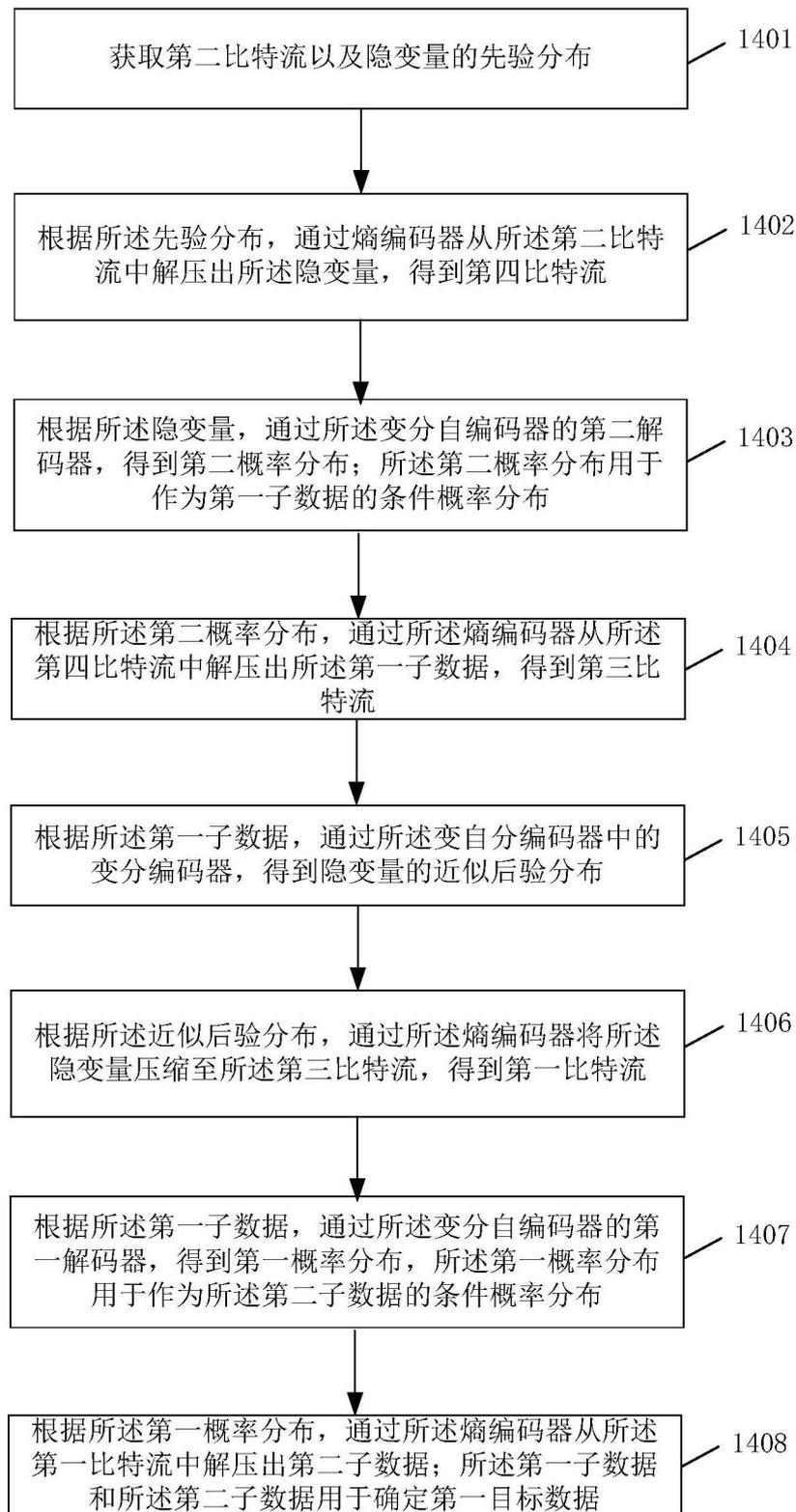


图14

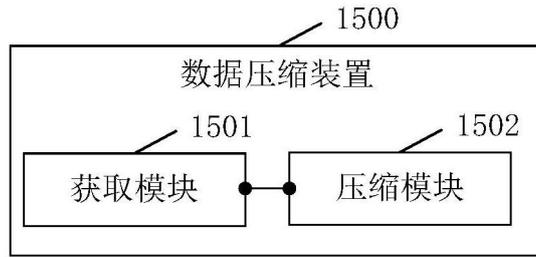


图15

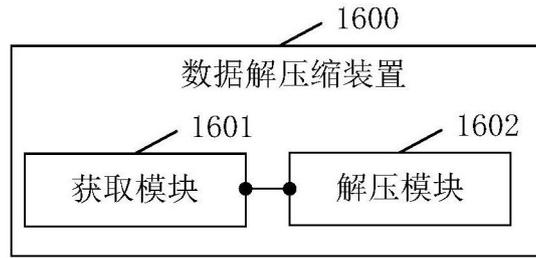


图16

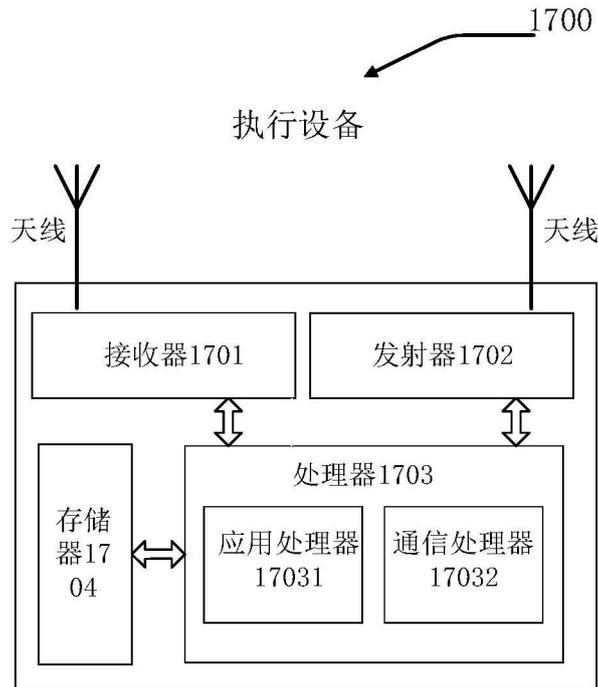


图17