

批归一化数据的处理方法及装置、电子设备和存储介质

本公开要求在2019年07月19日提交中国专利局、申请号为201910656284.2、申请名称为“批归一化数据的处理方法及装置、电子设备和存储介质”的中国专利申请的优先权，其全部内容通过引用结合在本公开中。

5

技术领域

本公开涉及数据处理技术领域，尤其涉及一种针对批归一化数据的处理方法及装置、电子设备和存储介质。

10 背景技术

在深度神经网络中采取批归一化（BN，Batch Normalization），可以使得该神经网络即使采用最大学习率也不会发散，同时还能增加该神经网络的泛化性能。在BN层后可以连接有激励层，在激励层使用的激励函数可以为线性整流函数（ReLU，Rectified linear unit）。需要对BN+ReLU构成的该神经网络的性能提升进行完善。

15

发明内容

本公开提出了一种针对批归一化数据处理的技术方案。

根据本公开的一方面，提供了一种针对批归一化数据的处理方法，所述方法包括：

20 将多个样本数据输入待训练目标网络中的批归一化BN层进行归一化处理，得到BN层的处理结果，所述多个样本数据通过对多个图像数据进行特征提取得到；

将所述BN层的处理结果根据指定的常数偏移量进行初始BN的偏移调整，得到偏移BN层的处理结果；

将所述偏移BN层的处理结果通过激励层的线性整流函数ReLU进行非线性映射，逐级得到损失函数后再反向传播，得到第一目标网络。

25 采用本公开，通过设置常数偏移量对初始BN进行偏移调整，得到偏移BN层的处理结果，使进入待训练目标网络中不可训练区域的网络参数，通过该偏移BN层的处理结果重新迁移到可训练区域，或者将进入待训练目标网络中不可训练区域的网络参数，通过该偏移BN层的处理结果进行网络剪枝，从而改善了网络的性能。

可能的实现方式中，所述将多个样本数据输入待训练目标网络中的BN层进行归一化处理，得到BN层的处理结果，包括：

30 根据所述多个样本数据对应的均值和方差，对所述多个样本数据进行归一化处理，得到归一化处理结果；

根据所述BN层的缩放系数和偏移系数，对所述归一化处理结果进行线性变换，得到所述BN层的处理结果。

35 采用本公开，可以对多个样本数据进行归一化处理，根据所述BN层的缩放系数和偏移系数，对所述归一化处理结果进行线性变换，得到所述BN层的处理结果，从而减少样本数据分散程度，有利于加速网络的训练。

可能的实现方式中，所述将所述BN层的处理结果根据指定的常数偏移量进行初始BN的偏移调整，得到偏移BN层的处理结果，包括：

40 设置所述常数偏移量为正数，通过所述常数偏移量进行初始BN的偏移调整，得到所述偏移BN层的处理结果。

采用本公开，设置常数偏移量的取值为正数，根据常数偏移量对初始BN进行偏移调整，得到偏移BN层的处理结果后，使进入待训练目标网络中不可训练区域的网络参数，通过该偏移BN层的处理结果重新迁移到可训练区域。

可能的实现方式中, 所述将所述BN层的处理结果根据指定的常数偏移量进行初始BN的偏移调整, 得到偏移BN层的处理结果, 包括:

设置所述常数偏移量为负数, 通过所述常数偏移量进行初始BN的偏移调整, 得到所述偏移BN层的处理结果。

- 5 采用本公开, 设置常数偏移量的取值为负数, 根据常数偏移量对初始BN进行偏移调整, 得到偏移BN层的处理结果后, 将进入待训练目标网络中不可训练区域的网络参数, 通过该偏移BN层的处理结果进行网络剪枝, 从而得到保证网络稀疏性的通用剪枝网络, 使用该剪枝网络可以降低数据运算量。

- 10 可能的实现方式中, 所述将所述偏移BN层的处理结果通过激励层的ReLU进行非线性映射, 逐级得到损失函数后再反向传播, 得到第一目标网络, 包括:

将所述偏移BN层的处理结果通过所述ReLU进行非线性映射后, 进入下一层计算, 最终得到损失函数;

根据损失函数的反向传播, 得到所述第一目标网络。

- 15 采用本公开, 通过ReLU进行非线性映射, 再利用损失函数反向传播, 使求导得到梯度的运算量降低且ReLU会使得神经网络中的部分输出为零, 从而有助于形成网络的稀疏性。

可能的实现方式中, 所述常数偏移量的取值范围在[0.01, 0.1]之间。

采用本公开, 常数偏移量的取值范围在[0.01, 0.1]区间时, 可以抑制网络参数进入不可训练区域, 从而改善网络的性能, 同时兼容BN层的表达能力。

- 20 可能的实现方式中, 所述常数偏移量的取值范围在[-0.1, -0.01]之间。

采用本公开, 常数偏移量的取值范围在[-0.1, -0.01]区间时, 可以促进网络剪枝, 从而增加网络训练或模型推理的速度。

根据本公开的一方面, 提供了一种图像分类方法, 所述方法包括:

获取图像数据;

- 25 采用上述针对批归一化数据的处理方法得到的第一目标网络, 对所述图像数据进行图像分类, 得到图像分类处理结果。

采用本公开, 通过第一目标网络进行图像分类, 不仅数据运算量低且提高了图像分类的精度。

根据本公开的一方面, 提供了一种图像检测方法, 所述方法包括:

- 30 获取图像数据;

采用上述针对批归一化数据的处理方法得到的第一目标网络, 对所述图像数据中的目标区域进行图像检测, 得到图像检测结果。

采用本公开, 通过第一目标网络进行图像检测, 不仅数据运算量低且提高了图像检测的精度。

- 35 根据本公开的一方面, 提供了一种视频处理方法, 所述方法包括:

获取视频图像;

采用上述针对批归一化数据的处理方法得到的第一目标网络, 对所述视频图像按照预设处理策略进行编码、解码及播放处理中的至少一种视频处理, 得到视频处理结果。

- 40 采用本公开, 通过第一目标网络进行视频处理, 不仅数据运算量低且提高了视频处理的精度。

根据本公开的一方面, 提供了一种针对批归一化数据的处理装置, 所述装置包括:

归一化单元, 用于将多个样本数据输入待训练目标网络中的批归一化BN层进行归一化处理, 得到BN层的处理结果, 所述多个样本数据通过对多个图像数据进行特征提取所得到;

偏移单元，用于将所述BN层的处理结果根据指定的常数偏移量进行初始BN的偏移调整，得到偏移BN层的处理结果；

处理单元，用于将所述偏移BN层的处理结果通过激励层的线性整流函数ReLU进行非线性映射，逐级得到损失函数后再反向传播，得到第一目标网络。

5 可能的实现方式中，所述归一化单元，用于：

根据所述多个样本数据对应的均值和方差，对所述多个样本数据进行归一化处理，得到归一化处理结果；

根据所述BN层的缩放系数和偏移系数，对所述归一化处理结果进行线性变换，得到所述BN层的处理结果。

10 可能的实现方式中，所述偏移单元，用于：

设置所述常数偏移量为正数，通过所述常数偏移量进行初始BN的偏移调整，得到所述偏移BN层的处理结果。

可能的实现方式中，所述偏移单元，用于：

15 设置所述常数偏移量为负数，通过所述常数偏移量进行初始BN的偏移调整，得到所述偏移BN层的处理结果。

可能的实现方式中，所述处理单元，用于：

将所述偏移BN层的处理结果通过所述ReLU进行非线性映射后，进入下一层计算，最终得到损失函数；

根据损失函数的反向传播，得到所述第一目标网络。

20 可能的实现方式中，所述常数偏移量的取值范围在[0.01, 0.1]之间。

可能的实现方式中，所述常数偏移量的取值范围在[-0.1, -0.01]之间。

根据本公开的一方面，提供了一种图像分类装置，所述装置包括：

第一获取器，用于获取图像数据；

25 第一处理器，用于采用上述针对批归一化数据的处理方法得到的第一目标网络，对所述图像数据进行图像分类，得到图像分类处理结果。

根据本公开的一方面，提供了一种图像检测装置，所述装置包括：

第二获取器，用于获取图像数据；

30 第二处理器，用于采用上述针对批归一化数据的处理方法得到的第一目标网络，对所述图像数据中的目标区域进行图像检测，得到图像检测结果。

根据本公开的一方面，提供了一种视频处理装置，所述装置包括：

第三获取器，用于获取视频图像；

35 第三处理器，用于采用上述针对批归一化数据的处理方法得到的第一目标网络，对所述视频图像按照预设处理策略进行编码、解码及播放处理中的至少一种视频处理，得到视频处理结果。

根据本公开的一方面，提供了一种电子设备，包括：

处理器；

用于存储处理器可执行指令的存储器；

其中，所述处理器被配置为：执行上述针对批归一化数据处理的方法。

40 根据本公开的一方面，提供了一种计算机可读存储介质，其上存储有计算机程序指令，所述计算机程序指令被处理器执行时实现上述针对批归一化数据处理的方法。

根据本公开的一方面，提供了一种计算机程序，其中，所述计算机程序包括计算机可读代码，当所述计算机可读代码在电子设备中运行时，所述电子设备中的处理器执行用于实现上述针对所述批归一化数据处理的方法。

在本公开实施例中，将多个样本数据输入待训练目标网络中的BN层进行归一化处

理,得到BN层的处理结果,所述多个样本数据通过对多个图像数据进行特征提取所得到;将所述BN层的处理结果根据指定的常数偏移量进行初始BN的偏移调整,得到偏移BN层的处理结果;将所述偏移BN层的处理结果通过激励层的ReLU进行非线性映射,逐级得到损失函数后再反向传播,得到第一目标网络。采用本公开,在对BN层偏移处理后再接入激励层,将偏移BN层的处理结果通过ReLU进行非线性映射后进行损失函数的反向传播,得到第一目标网络(该第一目标网络是待训练目标网络进行训练后得到的目标网络),具有偏移BN+ReLU的该第一目标网络,通过设置常数偏移量对初始BN进行偏移调整,得到偏移BN层的处理结果,使进入待训练目标网络中不可训练区域的网络参数,通过该偏移BN层的处理结果重新迁移到可训练区域,或者将进入待训练目标网络中不可训练区域的网络参数,通过该偏移BN层的处理结果进行网络剪枝,从而改善了网络的性能。

应当理解的是,以上的一般描述和后文的细节描述仅是示例性和解释性的,而非限制本公开。

根据下面参考附图对示例性实施例的详细说明,本公开的其它特征及方面将变得清楚。

附图说明

此处的附图被并入说明书中并构成本说明书的一部分,这些附图示出了符合本公开的实施例,并与说明书一起用于说明本公开的技术方案。

图1示出根据本公开实施例的针对批归一化数据的处理方法的流程图。

图2示出根据本公开实施例的应用于图像分类场景的偏移处理效果示意图。

图3示出根据本公开实施例的应用于迁移学习场景的偏移处理效果示意图。

图4示出根据本公开实施例的针对批归一化数据的处理装置的框图。

图5示出根据本公开实施例的电子设备的框图。

图6示出根据本公开实施例的电子设备的框图。

具体实施方式

以下将参考附图详细说明本公开的各种示例性实施例、特征和方面。附图中相同的附图标记表示功能相同或相似的元件。尽管在附图中示出了实施例的各种方面,但是除非特别指出,不必按比例绘制附图。

在这里专用的词“示例性”意为“用作例子、实施例或说明性”。这里作为“示例性”所说明的任何实施例不必解释为优于或好于其它实施例。

本文中术语“和/或”,仅仅是一种描述关联对象的关联关系,表示可以存在三种关系,例如,A和/或B,可以表示:单独存在A,同时存在A和B,单独存在B这三种情况。

另外,本文中术语“至少一种”表示多种中的任意一种或多种中的至少两种的任意组合,例如,包括A、B、C中的至少一种,可以表示包括从A、B和C构成的集合中选择的任意一个或多个元素。

另外,为了更好的说明本公开,在下文的具体实施方式中给出了众多的具体细节。本领域技术人员应当理解,没有某些具体细节,本公开同样可以实施。在一些实例中,对于本领域技术人员熟知的方法、手段、元件和电路未作详细描述,以便于凸显本公开的主旨。

深度神经网络中,BN往往是不可或缺的一种归一化方式。BN可以使得该神经网络能用最大学习率而不发散同时增加模型的泛化性能。ReLU是该神经网络中的一种非线性激活函数。相对其他非线性激活函数(如Sigmoid、Tanh等),ReLU在负值输入时激活值

是恒0，因此，其能表达特征的稀疏属性，从而使网络的训练能更快收敛。

对于稀疏属性而言，ReLU会使该神经网络中一部分神经元的输出为0，也可以说用于该神经网络中参数运算的权重为0（从全局视角来看，是去掉了部分权重），这样就造成了网络具备稀疏性，减少了参数的相互依存关系，缓解了过拟合问题的发生，且用于该神经网络中参数运算的权重为0（从全局视角来看，是去掉了部分权重），使得运算速度更快，网络的训练能更快收敛。一个例子中，参数运算的权重有10万个，如果将这个神经网络部署在手机或车载等运算负载不能过大的终端时，会使得运算量非常大，而当使得部分权重为0（即从运算中去掉部分权重），网络具备稀疏性，既不会影响神经网络太多的网络性能，又能提高该神经网络部署在手机或车载等终端的运行效率，使得运算负载不超出预期，那么，这种网络稀疏性是用户期望出现的稀疏性结果，可以称为良性的稀疏性。

针对稀疏性而言，如果神经网络中权重为0的网络通道（神经网络中至少一个对应的输入输出构成的网络通道）过多，会出现不良的稀疏性，这是不利的，需要去掉或者说抑制这种不良的稀疏性。

由于网络具备稀疏性可以降低数据运算量，因此，考虑到网络具备稀疏性的这一好处，如果在神经网络中存在权重为0的网络通道（神经网络中至少一个对应的输入输出构成的网络通道），那么网络参数的变少，可以提高运行效率，所以说，将该神经网络中参数运算的部分权重设置为0（从全局视角来看，是去掉了部分权重），可以使得运算速度更快。采用本公开，将偏移BN层的处理结果通过ReLU进行非线性映射后进行损失函数的反向传播，得到的第一目标网络，可以对这两方面都进行完善，在将多个样本数据输入待训练目标网络中的BN层进行归一化处理，得到BN层的处理结果，再将所述BN层的处理结果根据指定的常数偏移量进行初始BN的偏移调整，通过对常数偏移量取不同的值，得到不同偏移BN层的处理结果。比如，当常数偏移量取正数时，通过BN层偏移处理，可以抑制第一目标网络的网络稀疏性；当常数偏移量取负数时，通过BN层偏移处理，可以促进第一目标网络的网络稀疏性，得到剪枝网络。就剪枝网络而言，通过剪枝网络可以减少深层网络的繁重计算。一个典型的剪枝网络分步描述为：首先训练一个大的网络模型，进行剪枝处理，最后进行网络模型的微调。在剪枝过程中，根据用户预期的标准，裁剪冗余权重（去掉部分权重），只有重要的权重被保留以保证网络模型准确度及性能。剪枝处理是一种模型压缩方法，对深度神经网络的稠密连接引入稀疏性，通过将“不重要”的权值直接置零来减少非零权值数量，以达到提高网络模型运行效率的目的。

由于BN层中的归一化处理，导致激励层（包括用于执行非线性映射的ReLU）接在BN层后面时，会使得该神经网络在网络初始阶段或学习率较大情况下使得BN层的参数出现一段稳定的不可训练区域，当参数进入到该区域后无法从样本数据中得到梯度并更新，因此只能在L2损失函数作用下逐渐趋于0，造成该网络通道被剪枝处理。

就不可训练区域而言，所谓不可训练区域是指进入激励层ReLU的输入参数为负数时导致ReLU输入为恒等于0，无梯度回传。一种出现该不可训练区域的原因是：BN层的两个参数在 γ 取值为较小值如0.1，而 β 取值为负数如-100时，会导致BN层输出结果经过ReLU

的非线性映射后恒等于0，如果恒等于0，则无法进行梯度求导，也就是说无梯度回传，从而后续损失函数的反向传输中梯度下降无法进行，导致参数无法进行更新。

综上所述，在实际应用中，发明人发现，在BN+ReLU的网络中参数进入不可训练区域的概率在训练初期以及学习率较大时是比较随机的，但在训练过程中仍表现出来部分选择性，即对损失影响较小的参数会更加可能进入到不可训练区域而被剪枝。因此这种

现象表现出上述描述的两面性，一方面其会作为一种剪枝方法能在网络性能基本不变的情况下使网络的参数量减小，需要促进这种稀疏性；另一方面也反而会使得网络的表达能力下降，进而使网络的性能变差，需要抑制这种稀疏性。

5 由于在深度神经网络中采用BN+ReLU的网络组合方式，会导致部分网络通道（如BN参数的通道）无法训练而坍塌（进一步导致其前一层的卷积运算也无法训练），因此，一方面，本公开对BN的形式加以改进，具体是增加指定的常数偏移量（此时取正数），以进行初始BN的偏移调整。根据调整后得到的偏移BN层的处理结果在网络训练的初始阶段或由于学习率较大而进入不可训练区域时可在L2损失函数的作用下重新回到可训练区域，从而保证了网络的表达能力，抑制稀疏性。该方法可以弥补上述BN+ReLU网络组合方式会导致部分网络通道无法训练而坍塌的问题。该方案针对每个BN在其原有形式上加

10 上一个的指定的正常数偏移量（如常数 α ），便可使得网络具有剪枝效果也可使在训练过程中处于不可训练区的网络参数重新回到可训练区域，从而提升网络的性能。另一方面，本公开对BN的形式加以改进，具体是增加指定的常数偏移量（此时取负数），以进行初始BN的偏移调整。根据调整后得到的偏移BN层的处理结果在完全兼容原始BN的表达能力的15 基础上，通过调整BN偏置项额外的偏移使得网络可直接被训练得到一个剪枝网络。由于是对原始BN的形式做微小调整，所以该方案称为偏移BN(psBN, post shifted Batch Normalization)，用户可以根据自己的需求（如想进一步提升网络的性能或增加网络的通道稀疏性）选择相应偏移常数 α 的符号，即根据用户的需求来选择 α 的取值为正数或负数。

需要指出的是，一个网络中可能会存在多个BN层，本公开中，对于每个BN层，可以都根据增加的常数偏移量对BN层进行偏移调整，得到偏移BN层的处理结果。且在多个BN层中采用的常数偏移量可以是统一的偏移量，也就是说，在同一个网络的至少一个BN层均增加该常数偏移量，并设置相同值，具体取值根据用户需求进行设置，该常数偏移量可以为正数，也可以为负数。

20 对于每个BN层而言，当常数偏移量的取值为正数，根据常数偏移量对初始BN进行偏移调整，得到偏移BN层的处理结果后，使进入待训练目标网络中不可训练区域的网络参数，通过该偏移BN层的处理结果重新迁移到可训练区域。

对于每个BN层而言，当常数偏移量的取值为正数，根据常数偏移量对初始BN进行偏移调整，得到偏移BN层的处理结果后，将进入待训练目标网络中不可训练区域的网络参数，通过该偏移BN层的处理结果进行网络剪枝，从而得到保证网络稀疏性的通用剪枝网络，使用该剪枝网络可以降低数据运算量。

图1示出根据本公开实施例的针对批归一化数据的处理方法的流程图，该方法应用于针对批归一化数据的处理装置，例如，该处理装置部署于终端设备或服务器或其它处理设备执行的情况下，可以执行图像分类、图像检测和视频处理等等。其中，终端设备可以为用户设备（UE, User Equipment）、移动设备、蜂窝电话、无绳电话、个人数字处理35 （PDA, Personal Digital Assistant）、手持设备、计算设备、车载设备、可穿戴设备等。在一些可能的实现方式中，该处理方法可以通过处理器调用存储器中存储的计算机可读指令的方式来实现。如图1所示，该流程包括：

步骤S101、将多个样本数据输入待训练目标网络中的BN层进行归一化处理，得到BN层的处理结果，所述多个样本数据通过对多个图像数据进行特征提取得到。

40 一个示例中，待训练目标网络可以为针对图像处理的图卷积网络（如CNN卷积神经网络），包括：1）输入层：用于样本数据的输入；2）卷积层：使用卷积核进行特征提取和特征映射；3）激励层：由于卷积也是一种线性运算，因此需要增加非线性映射，要

接入激励层，激励层中包括用于非线性映射的ReLU，以进行非线性映射，因为卷积层的计算还是一种线性计算，而所谓激励层，可以对卷积层的输出结果做一次非线性映射；4) 池化层：进行下采样，对特征图稀疏处理，减少数据运算量；5) 全连接FC层：在CNN的尾部进行重新拟合，减少特征信息的损失；6) 输出层：用于输出结果。其中，中间还可以使用一些其他的功能层，如用于在CNN卷积神经网络中对特征进行归一化的BN层；对某些（图片）数据的进行分区域的单独学习的切分层；对独立进行特征学习的分支进行融合的融合层等等。

一些可能的实现方式中，可以将卷积层和激励层合并在一起称为卷积层，BN层可以位于输入层以进行特征预处理，也可以位于卷积层中，本公开所采用神经网络的具体结构不限于上述的描述。

步骤S102、将所述BN层的处理结果根据指定的常数偏移量（如 α ）进行初始BN的偏移调整，得到偏移BN层的处理结果。

一个示例中，偏移调整的计算公式如公式（1）所示：

$$y = \text{BN}(\hat{x}) + \alpha = \gamma \cdot \frac{\hat{x} - \mu_{\beta}}{\sqrt{\sigma_{\beta}^2 + \epsilon}} + \beta + \alpha \quad (1)$$

其中， $\text{BN}(\hat{x})$ 为步骤S101中得到的BN层的处理结果（或称原有BN层的处理结果）；

$\hat{x}$ 为BN层的输入特征， γ 为BN层的缩放系数； β 为BN层的偏移系数； μ_{β} 为样本数据的均值； σ_{β} 为样本数据的标准差； ϵ 为固定常数，可以等于 10^{-5} 。ReLU 保持不变，例如

$\text{ReLU}(y) = \max(0, y)$ 。 y 为偏移BN层的处理结果，可表示为偏移BN（psBN），其具有与BN相同的表现能力，当特征参数在训练期间进入不可训练区域时可再次训练。可以根据偏移BN（psBN）提高网络模型的性能，比如，可将其作为CIFAR-10的分类和MS-COCO2017上的物体检测。

步骤S103、将所述偏移BN层的处理结果通过激励层的激活函数ReLU进行非线性映射，逐级得到损失函数后再反向传播，得到第一目标网络。

一个示例中，待训练目标网络可以为BN+ReLU构成的神经网络，经步骤S101-步骤S103训练得到的第一目标网络为BN（psBN）+ReLU构成的神经网络。

采用本公开的一个完整示例中，可以将多个样本数据输入待训练目标网络中的批归一化BN层进行归一化处理，得到BN层（普通BN或称原始BN）的处理结果。该处理结果具体为归一化及对归一化进一步做线性变换后得到的处理结果。所述多个样本数据通过对多个图像数据进行特征提取所得（获取多个图像数据，根据从所述多个图像数据中提取的多个特征参数得到样本数据集，样本数据集中包括多个样本数据）。就该归一化处理而言，在批处理BN中从一批样本数据（特征参数）中得到其均值和方差，根据均值和方差对样本数据做归一化，对归一化的特征参数做线性变换（BN乘以一个缩放系数和偏移系数），得到BN层（普通BN或称原始BN）的处理结果。将所述BN层的处理结果根据指定的常数偏移量进行初始BN的偏移调整，得到偏移BN层的处理结果，即对普通BN或称原始BN的输出，加上微小常数偏移量（视乎用户的需要可选择偏移量的符号），得到

偏移BN层的处理结果（新BN层输出结果），将所述偏移BN层的处理结果通过激励层的激活函数ReLU进行非线性映射后进行损失函数的反向传播，迭代训练得到上述第一目标网络。

5 采用本公开，通过设置常数偏移量对初始BN进行偏移调整，得到偏移BN层的处理结果，使进入待训练目标网络中不可训练区域的网络参数，通过该偏移BN层的处理结果重新迁移到可训练区域，或者将进入待训练目标网络中不可训练区域的网络参数，通过该偏移BN层的处理结果进行网络剪枝，从而改善了网络的性能。

可能的实现方式中，将多个样本数据输入待训练目标网络中的BN层进行归一化处理，得到BN层的处理结果，包括：根据所述多个样本数据对应的均值（ μ_g ）和方差（ σ_g^2 ），
10 对所述多个样本数据进行归一化处理，得到归一化处理结果。根据所述BN层的缩放系数（ γ ）和偏移系数（ β ），对所述归一化处理结果进行线性变换，得到所述BN层的处理结果。

采用本公开，可以对多个样本数据进行归一化处理，根据所述BN层的缩放系数和偏移系数，对所述归一化处理结果进行线性变换，得到所述BN层的处理结果，从而减少样
15 本数据分散程度，有利于加速网络的训练。

可能的实现方式中，所述将所述BN层的处理结果根据指定的常数偏移量进行初始BN的偏移调整，得到偏移BN层的处理结果，包括：设置所述常数偏移量为正数，通过所述常数偏移量进行初始BN的偏移调整，得到偏移BN层的处理结果。

采用本公开，设置常数偏移量的取值为正数，根据常数偏移量对初始BN进行偏移调
20 整，得到偏移BN层的处理结果后，使进入待训练目标网络中不可训练区域的网络参数，通过该偏移BN层的处理结果重新迁移到可训练区域。

一示例中， α 为正数，例如 α 的值在[0.01, 0.1]之间，可以兼容BN层表达能力，即不改变BN层参数的先验，不对网络造成不良影响的同时起到抑制参数进入不可训练区域的效果。样本数据为初始BN层中的特征参数，在网络训练的初始阶段或由于学习率较大时
25 该特征参数进入不可训练区域，根据偏移BN层的处理结果可以使得该特征参数重新回到可训练区域，由于抑制了参数进入不可训练区域，从而保证了网络的表达能力，改善了网络的性能。具体来说， $\alpha > 0$ ，即取值为正数时，由于当BN层的参数进入不可训练区域之后，BN层的参数 γ 与 β 会在权值衰减的作用下以相同的速度衰减到0，但由于偏置项有正常数 α 的存在，使得偏置项最终会大于0，则ReLU进入到线性区（即梯度可经ReLU回
30 传），从而使得神经网络中的神经元被再次激活（即BN层的参数重新进入可训练区域），因此，在 α 为正数时可以达到抑制稀疏性的目的。

采用本公开，通过训练目标网络（如神经网络中用于对视频数据处理，如针对图像处理的图卷积网络），以提升其性能。主要是针对BN+ReLU的网络，ReLU不动，通过指定的常数偏移量调整BN产生偏移后得到psBN，以得到训练后的目标网络为psBN+ ReLU
35 的网络，从而优化网络性能，其中， α 为正数的取值时，是为了抑制作用，即迁移到可训练区域，以在具备网络稀疏性时，去掉不良的稀疏性结果。

可能的实现方式中，所述将所述BN层的处理结果根据指定的常数偏移量进行初始BN的偏移调整，得到偏移BN层的处理结果，包括：设置所述常数偏移量为负数，通过所述常数偏移量进行初始BN的偏移调整，得到偏移BN层的处理结果。

- 5 采用本公开，设置常数偏移量的取值为负数，根据常数偏移量对初始BN进行偏移调整，得到偏移BN层的处理结果后，将进入待训练目标网络中不可训练区域的网络参数，通过该偏移BN层的处理结果进行网络剪枝，从而得到保证网络稀疏性的通用剪枝网络，使用该剪枝网络可以降低数据运算量。

- 一示例中， α 为负数，例如 α 的值在 $[-0.1, -0.01]$ 之间，可以兼容BN层表达能力，即不改变BN层参数的先验，不对网络造成不良影响的同时使得网络具有更少参数。样本数据为初始BN层中的特征参数，这时会使得更多的BN参数处于不可训练区域，从而使得该部分通道在训练过程中被剪枝。由于促进了网络剪枝，从而增加了网络训练或模型推理的速度，使得网络具有更少的参数同时，较小地影响网络的性能。具体来说， $\alpha < 0$ 时的

- 原理则与上述 $\alpha > 0$ 的情况相反，偏置项加上负常数 α 后可诱导进入ReLU的输入参数小于0，梯度无法经过ReLU回传，从而使得BN层的参数在权值衰减的作用下衰减到0，实现网络剪枝的功能，因此，在 α 为负数时可以达到促进稀疏性的目的。

- 采用本公开，通过训练目标网络（如神经网络中用于对视频数据处理，如针对图像处理的图卷积网络），以提升其性能。主要是针对BN+ReLU的网络，ReLU不动，通过指定的常数偏移量调整BN产生偏移后得到psBN，以得到训练后的目标网络为psBN+ ReLU的网络，从而优化网络性能。其中， α 为负数的取值时，是为了促进作用，即得到剪枝网络。

- 可能的实现方式中，所述将所述偏移BN层的处理结果通过激励层的ReLU进行非线性映射，逐级得到损失函数后再反向传播，得到第一目标网络，包括：将所述偏移BN层的处理结果通过所述ReLU进行非线性映射后，进入下一层计算，最终得到损失函数；根据损失函数的反向传播，得到所述第一目标网络。需要指出的是：由于神经网络是多层结构，这里描述的该偏移BN+ReLU只是神经网络其中一层的结构，因此，该层的输出需要逐层传导后，最终才会得到损失函数。

- 采用本公开，通过ReLU进行非线性映射，再利用损失函数反向传播，使求得得到梯度的运算量降低且ReLU会使得神经网络中的部分输出为零，从而有助于形成网络的稀疏性。

- 30 针对上述训练得到的第一目标网络，对应的应用场景包括：

本公开的图像分类方法，所述方法包括：获取图像数据；采用本公开上述方法得到的第一目标网络，对所述图像数据进行图像分类，得到图像分类处理结果。

本公开的图像检测方法，所述方法包括：获取图像数据；采用本公开上述方法得到的第一目标网络，对所述图像数据中的目标区域进行图像检测，得到图像检测结果。

- 35 本公开的视频处理方法，所述方法包括：获取视频图像；采用本公开上述方法得到的第一目标网络，对所述视频图像按照预设处理策略进行编码、解码及播放处理中的至少一种视频处理，得到视频处理结果。

图2示出根据本公开实施例的应用于图像分类场景的偏移处理效果示意图，其中，BN+ReLU这一行为采用待训练的网络进行图像分类得到的处理结果，BN+LeakyReLU这

一行为采用一般优化训练网络进行图像分类得到的处理结果，psBN+ReLU这一行为采用本公开对网络进行训练所得到的第一目标网络进行图像分类的处理结果（如多次训练的平均准确率），网络使用ResNet-20和VGG16-BN两个网络为例。从图2中可见：采用本公开得到的处理结果在多个结果中最优越。采用本公开，针对BN+ReLU的网络，ReLU不动，通过指定的常数偏移量调整BN产生偏移后得到psBN，得到第一目标网络为psBN+ReLU的网络，从而优化网络性能。其中，带泄露线性整流函数（Leaky ReLU，Leaky Rectified linear unit）与ReLU同为激活函数，它是ReLU的变体，该Leaky ReLU的输出对负值输入有很小的坡度，由于导数总是不为零，这能减少神经网络中静默神经元的出现，允许基于梯度的学习（虽然会很慢），解决了Relu函数进入负区间后，导致神经元不学习的问题。

图3示出根据本公开实施例的应用于迁移学习场景的偏移处理效果示意图。对于图像尺寸为500或800的图像数据，采用RetinaNet网络所得到的图像检测效果 $AP^{bbox}(\text{RetinaNet})$ （即检测的平均准确率）中，括号中的数值是采用相关技术得到的准确率，括号外的数值是采用发明人复现RetinaNet网络进行图像检测的结果，而 $AP^{bbox}(\text{RetinaNet+psBN})$ 是使用本公开的方案将RetinaNet网络修改为带有偏移BN的RetinaNet网络进行图像检测所得到的检测准确率。从图3中直观可以看出：采用本公开所得到的数值更高，也就是说准确率相比采用以往相关技术更加准确，显然采用本公开 $AP^{bbox}(\text{RetinaNet+psBN})$ 所得到的图像检测效果更好。

本领域技术人员可以理解，在具体实施方式的上述方法中，各步骤的撰写顺序并不意味着严格的执行顺序而对实施过程构成任何限定，各步骤的具体执行顺序应当以其功能和可能的内在逻辑确定。

本公开提及的上述各个方法实施例，在不违背原理逻辑的情况下，均可以彼此相互结合形成结合后的实施例，限于篇幅，本公开不再赘述。

此外，本公开还提供了针对批归一化数据的处理装置、电子设备、计算机可读存储介质、程序，上述均可用来实现本公开提供的任一种针对批归一化数据的处理方法，相应技术方案和描述和参见方法部分的相应记载，不再赘述。

图4示出根据本公开实施例的针对批归一化数据的处理装置的框图，如图4所示，该处理装置，包括：归一化单元31，用于将多个样本数据输入待训练目标网络中的BN层进行归一化处理，得到BN层的处理结果，所述多个样本数据通过对多个图像数据进行特征提取所得；偏移单元32，用于将所述BN层的处理结果根据指定的常数偏移量进行初始BN的偏移调整，得到偏移BN层的处理结果；处理单元33，用于将所述偏移BN层的处理结果通过激励层的ReLU进行非线性映射，逐级得到损失函数后再反向传播，得到第一目标网络。

可能的实现方式中，所述归一化单元，用于：根据所述多个样本数据对应的均值和方差，对所述多个样本数据进行归一化处理，得到归一化处理结果；根据所述BN层的缩放系数和偏移系数，对所述归一化处理结果进行线性变换，得到所述BN层的处理结果。

可能的实现方式中，所述偏移单元，用于：设置所述常数偏移量为正数，通过所述常数偏移量进行初始BN的偏移调整，得到所述偏移BN层的处理结果。从而，将进入待训练目标网络中不可训练区域的网络参数，通过所述偏移BN层的处理结果重新迁移到可训练区域。

可能的实现方式中，所述偏移单元，用于：设置所述常数偏移量为负数，通过所述常数偏移量进行初始BN的偏移调整，得到所述偏移BN层的处理结果。从而，将进入待训练目标网络中不可训练区域的网络参数，通过所述偏移BN层的处理结果进行网络剪

枝，得到剪枝网络。

可能的实现方式中，所述处理单元，用于：将所述偏移BN层的处理结果通过所述ReLU进行非线性映射后，进入下一层计算，最终得到损失函数；根据损失函数的反向传播，得到所述第一目标网络。

5 可能的实现方式中，所述常数偏移量的取值范围在[0.01, 0.1]之间。

可能的实现方式中，所述常数偏移量的取值范围在[-0.1, -0.01]之间。

本公开的图像分类装置，所述装置包括：第一获取器，用于获取图像数据；第一处理器，用于采用本公开上述方法得到的第一目标网络，对所述图像数据进行图像分类，得到图像分类处理结果。

10 本公开的一种图像检测装置，所述装置包括：第二获取器，用于采集视频图像；第二处理器，用于采用本公开上述方法得到的第一目标网络，对所述图像数据中的目标区域进行图像检测，得到图像检测结果。

本公开的一种视频处理装置，所述装置包括：第三获取器，用于获取视频图像；第三处理器，用于采用本公开上述方法得到的第一目标网络，对所述视频图像按照预设处理策略进行编码、解码及播放处理中的至少一种视频处理，得到视频处理结果。

15 需要指出的是，上述第一获取器、第二获取器及第三获取器所执行的获取操作，不限于获取的方式，比如，可以是第一获取器、第二获取器及第三获取器自身执行获取操作（如对图像数据或视频图像等的采集操作），得到操作结果；又如，可以是第一获取器、第二获取器及第三获取器通过无线或有线通信方式与其他可执行获取操作的处理设备进行通信，并获得由该处理设备执行获取操作（如对图像数据或视频图像等的采集操作）所得到的操作结果。其中，有线通信方式的接口不限于串行通信接口、总线接口及其他类型接口。

25 在一些实施例中，本公开实施例提供的装置具有的功能或包含的模块可以用于执行上文方法实施例描述的方法，其具体实现可以参照上文方法实施例的描述，为了简洁，这里不再赘述。

本公开实施例还提出一种计算机可读存储介质，其上存储有计算机程序指令，所述计算机程序指令被处理器执行时实现上述方法。计算机可读存储介质可以是易失性计算机可读存储介质或非易失性计算机可读存储介质。

30 本公开实施例还提出一种电子设备，包括：处理器；用于存储处理器可执行指令的存储器；其中，所述处理器被配置为实现上述方法。

电子设备可以被提供为终端、服务器或其它形态的设备。

35 本公开实施例还提出一种计算机程序，其中，所述计算机程序包括计算机可读代码，当所述计算机可读代码在电子设备中运行时，所述电子设备中的处理器执行用于实现上述任一实施例所述的方法。

图5是根据一示例性实施例示出的一种电子设备800的框图。例如，电子设备800可以是移动电话，计算机，数字广播终端，消息收发设备，游戏控制台，平板设备，医疗设备，健身设备，个人数字助理等终端。

40 参照图5，电子设备800可以包括以下一个或多个组件：处理组件802，存储器804，电源组件806，多媒体组件808，音频组件810，输入/输出（I/O）的接口812，传感器组件814，以及通信组件816。

处理组件802通常控制电子设备800的整体操作，诸如与显示，电话呼叫，数据通信，相机操作和记录操作相关联的操作。处理组件802可以包括一个或多个处理器820来执行

指令，以完成上述的方法的全部或部分步骤。此外，处理组件802可以包括一个或多个模块，便于处理组件802和其他组件之间的交互。例如，处理组件802可以包括多媒体模块，以方便多媒体组件808和处理组件802之间的交互。

5 存储器804被配置为存储各种类型的数据以支持在电子设备800的操作。这些数据的示例包括用于在电子设备800上操作的任何应用程序或方法的指令，联系人数据，电话簿数据，消息，图片，视频等。存储器804可以由任何类型的易失性或非易失性存储设备或者它们的组合实现，如静态随机存取存储器（SRAM），电可擦除可编程只读存储器（EEPROM），可擦除可编程只读存储器（EPROM），可编程只读存储器（PROM），只读存储器（ROM），磁存储器，快闪存储器，磁盘或光盘。

10 电源组件806为电子设备800的各种组件提供电力。电源组件806可以包括电源管理系统，一个或多个电源，及其他与为电子设备800生成、管理和分配电力相关联的组件。

多媒体组件808包括在所述电子设备800和用户之间的提供一个输出接口的屏幕。在一些实施例中，屏幕可以包括液晶显示器（LCD）和触摸面板（TP）。如果屏幕包括触摸面板，屏幕可以被实现为触摸屏，以接收来自用户的输入信号。触摸面板包括一个或多个触摸传感器以感测触摸、滑动和触摸面板上的手势。所述触摸传感器可以不仅感测触摸或滑动动作的边界，而且还检测与所述触摸或滑动操作相关的持续时间和压力。在
15 一些实施例中，多媒体组件808包括一个前置摄像头和/或后置摄像头。当电子设备800处于操作模式，如拍摄模式或视频模式时，前置摄像头和/或后置摄像头可以接收外部的多媒体数据。每个前置摄像头和后置摄像头可以是一个固定的光学透镜系统或具有焦距和光学变焦能力。
20

音频组件810被配置为输出和/或输入音频信号。例如，音频组件810包括一个麦克风（MIC），当电子设备800处于操作模式，如呼叫模式、记录模式和语音识别模式时，麦克风被配置为接收外部音频信号。所接收的音频信号可以被进一步存储在存储器804或经由通信组件816发送。在一些实施例中，音频组件810还包括一个扬声器，用于输出音频
25 信号。

I/O接口812为处理组件802和外围接口模块之间提供接口，上述外围接口模块可以是键盘，点击轮，按钮等。这些按钮可包括但不限于：主页按钮、音量按钮、启动按钮和锁定按钮。

传感器组件814包括一个或多个传感器，用于为电子设备800提供各个方面的状态评估。例如，传感器组件814可以检测到电子设备800的打开/关闭状态，组件的相对定位，
30 例如所述组件为电子设备800的显示器和小键盘，传感器组件814还可以检测到电子设备800或电子设备800一个组件的位置改变，用户与电子设备800接触的存在或不存在，电子设备800方位或加速/减速和电子设备800的温度变化。传感器组件814可以包括接近传感器，被配置用来在没有任何的物理接触时检测附近物体的存在。传感器组件814还可以包括光
35 传感器，如CMOS或CCD图像传感器，用于在成像应用中使用。在一些实施例中，该传感器组件814还可以包括加速度传感器，陀螺仪传感器，磁传感器，压力传感器或温度传感器。

通信组件816被配置为便于电子设备800和其他设备之间有线或无线方式的通信。电子设备800可以接入基于通信标准的无线网络，如WiFi，2G或3G，或它们的组合。在一个
40 示例性实施例中，通信组件816经由广播信道接收来自外部广播管理系统的广播信号或广播相关信息。在一个示例性实施例中，所述通信组件816还包括近场通信（NFC）模块，以促进短程通信。例如，在NFC模块可基于射频识别（RFID）技术，红外数据协会（IrDA）技术，超宽带（UWB）技术，蓝牙（BT）技术和其他技术来实现。

在示例性实施例中，电子设备800可以被一个或多个应用专用集成电路（ASIC）、数

字信号处理器（DSP）、数字信号处理设备（DSPD）、可编程逻辑器件（PLD）、现场可编程门阵列（FPGA）、控制器、微控制器、微处理器或其他电子元件实现，用于执行上述方法。

在示例性实施例中，还提供了一种计算机可读存储介质，例如包括计算机程序指令的存储器804，上述计算机程序指令可由电子设备800的处理器820执行以完成上述方法。

图6是根据一示例性实施例示出的一种电子设备900的框图。例如，电子设备900可以被提供为一服务器。参照图6，电子设备900包括处理组件922，其进一步包括一个或多个处理器，以及由存储器932所代表的存储器资源，用于存储可由处理组件922的执行的指令，例如应用程序。存储器932中存储的应用程序可以包括一个或一个以上的每一个对应于一组指令的模块。此外，处理组件922被配置为执行指令，以执行上述方法。

电子设备900还可以包括一个电源组件926被配置为执行电子设备900的电源管理，一个有线或无线网络接口950被配置为将电子设备900连接到网络，和一个输入输出（I/O）接口958。电子设备900可以操作基于存储在存储器932的操作系统，例如Windows Server™，Mac OS X™，Unix™，Linux™，FreeBSD™或类似。

在示例性实施例中，还提供了一种计算机可读存储介质，例如包括计算机程序指令的存储器932，上述计算机程序指令可由电子设备900的处理组件922执行以完成上述方法。

本公开可以是系统、方法和/或计算机程序产品。计算机程序产品可以包括计算机可读存储介质，其上载有用于使处理器实现本公开的各个方面的计算机可读程序指令。

计算机可读存储介质可以是可以保持和存储由指令执行设备使用的指令的有形设备。计算机可读存储介质例如可以是一——但不限于——电存储设备、磁存储设备、光存储设备、电磁存储设备、半导体存储设备或者上述的任意合适的组合。计算机可读存储介质的更具体的例子（非穷举的列表）包括：便携式计算机盘、硬盘、随机存取存储器（RAM）、只读存储器（ROM）、可擦式可编程只读存储器（EPROM或闪存）、静态随机存取存储器（SRAM）、便携式压缩盘只读存储器（CD-ROM）、数字多功能盘（DVD）、记忆棒、软盘、机械编码设备、例如其上存储有指令的打孔卡或凹槽内凸起结构、以及上述的任意合适的组合。这里所使用的计算机可读存储介质不被解释为瞬时信号本身，诸如无线电波或者其他自由传播的电磁波、通过波导或其他传输媒介传播的电磁波（例如，通过光纤电缆的光脉冲）、或者通过电线传输的电信号。

这里所描述的计算机可读程序指令可以从计算机可读存储介质下载到各个计算/处理设备，或者通过网络、例如因特网、局域网、广域网和/或无线网下载到外部计算机或外部存储设备。网络可以包括铜传输电缆、光纤传输、无线传输、路由器、防火墙、交换机、网关计算机和/或边缘服务器。每个计算/处理设备中的网络适配卡或者网络接口从网络接收计算机可读程序指令，并转发该计算机可读程序指令，以供存储在各个计算/处理设备中的计算机可读存储介质中。

用于执行本公开操作的计算机程序指令可以是汇编指令、指令集架构（ISA）指令、机器指令、机器相关指令、微代码、固件指令、状态设置数据、或者以一种或多种编程语言的任意组合编写的源代码或目标代码，所述编程语言包括面向对象的编程语言——诸如Smalltalk、C++等，以及常规的过程式编程语言——诸如“C”语言或类似的编程语言。

计算机可读程序指令可以完全地在用户计算机上执行、部分地在用户计算机上执行、作为一个独立的软件包执行、部分在用户计算机上部分在远程计算机上执行、或者完全在远程计算机或服务器上执行。在涉及远程计算机的情形中，远程计算机可以通过任意种类的网络——包括局域网（LAN）或广域网（WAN）——连接到用户计算机，或者，可以连接到外部计算机（例如利用因特网服务提供商来通过因特网连接）。在一些实施例中，通过利用

计算机可读程序指令的状态信息来个性化定制电子电路，例如可编程逻辑电路、现场可编程门阵列（FPGA）或可编程逻辑阵列（PLA），该电子电路可以执行计算机可读程序指令，从而实现本公开的各个方面。

5 这里参照根据本公开实施例的方法、装置（系统）和计算机程序产品的流程图和/或框图描述了本公开的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

这些计算机可读程序指令可以提供给通用计算机、专用计算机或其它可编程数据处理装置的处理器，从而生产出一种机器，使得这些指令在通过计算机或其它可编程数据处理装置的处理器执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各个方面的指令。

10 也可以把计算机可读程序指令加载到计算机、其它可编程数据处理装置、或其它设备上，使得在计算机、其它可编程数据处理装置或其它设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得在计算机、其它可编程数据处理装置、或其它设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

附图中的流程图和框图显示了根据本公开的多个实施例的系统、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、程序段或指令的一部分，所述模块、程序段或指令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。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，两个连续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，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动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。

在不违背逻辑的情况下，本申请不同实施例之间可以相互结合，不同实施例描述有所侧重，为侧重描述的部分可以参见其他实施例的记载。

15 以上已经描述了本公开的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实施例的原理、实际应用或对市场中技术的技术改进，或者使本技术领域的其它普通技术人员能理解本文披露的各实施例。