

系统增强学习方法和装置、电子设备、计算机存储介质

本公开要求在 2018 年 5 月 7 日提交中国专利局、申请号为 CN201810428099.3、发明名称为“系统增强学习方法和装置、电子设备、计算机存储介质”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。

技术领域

本公开涉及计算机视觉技术，尤其是一种系统增强学习方法和装置、电子设备、计算机存储介质。

背景技术

当前，多层级系统取得了很好的效果，如在多人姿态检测任务上，基于多层级系统设计的算法（物体检测器+单人姿态检测器）在很多数据库上都取得了最好的结果。同时，多层级结构的算法也允许调整和替换不同的模块，模型的灵活性很强。因此具有非常强的潜力和广大的前景。

发明内容

本公开实施例提供一种系统增强学习技术。

根据本公开实施例的一个方面，提供一种系统增强学习方法，包括：

基于系统的第一网络对输入图像进行处理，得到第一结果；

将所述第一结果输入所述系统的第二网络，得到第二结果；

经增强网络基于所述第二结果获得增强操作，基于所述增强操作对所述第一结果进行调整，获得目标结果。

可选地，在本公开上述任一方法实施例中，所述经增强网络基于所述第二结果获得增强操作，基于所述增强操作对所述第一结果进行调整，基于所述调整后的第一结果获得目标结果，包括：

经所述增强网络基于所述第二网络输出的所述第二结果获得增强操作，基于所述增强操作对所述第一结果进行调整，得到第一中间结果；

将所述第一中间结果输入所述第二网络，基于所述第一中间结果得到第二结果，将所述第二结果输入所述增强网络；

响应于达到预设条件，输出所述第二结果作为所述目标结果。

可选地，在本公开上述任一方法实施例中，所述增强操作包括至少一个调整动作；

所述经增强网络基于所述第二网络输出的第二结果获得增强操作，基于所述增强操作对所述第一结果进行调整，得到第一中间结果，包括：

基于所述第二结果经增强网络获得调整动作概率，基于所述调整动作概率确定至少一个调整动作；

基于所述至少一个调整动作对所述第一结果进行调整，得到第一中间结果。

可选地，在本公开上述任一方法实施例中，所述经增强网络基于所述第二网络输出的第二结果对所述第一结果进行调整，得到第一中间结果，包括：

基于所述第二结果和所述第一结果，经增强网络获得调整动作概率，基于所述调整动作概率确定至少一个调整动作；

基于所述至少一个调整动作对所述第一结果进行调整，得到第一中间结果。

可选地，在本公开上述任一方法实施例中，所述基于所述第二结果和所述第一结果，经增强网络获得调整动作概率，基于所述调整动作概率确定至少一个调整动作，包括：

基于所述第二结果和所述第一结果得到连接结果；

基于所述连接结果经增强网络获得调整动作概率，基于所述调整动作概率确定至少一个调整动作。

可选地，在本公开上述任一方法实施例中，所述第一网络包括以下一个或多个网络：

检测神经网络、识别神经网络、分类神经网络；

所述第二网络包括以下一个或多个网络：

检测神经网络、识别神经网络、分类神经网络。

可选地，在本公开上述任一方法实施例中，还包括：基于样本图像训练所述系统，所述样本图像已知样本目标结果。

可选地，在本公开上述任一方法实施例中，所述基于样本图像训练所述系统，包括：

基于所述系统的第一网络对所述样本图像进行处理，得到第一样本结果；

将所述第一样本结果输入所述系统的第二网络，得到第二样本结果；

经增强网络基于所述第二样本结果获得增强操作，基于所述增强操作对所述第一样本结果进行调整，基于调整后的第一样本结果获得预测目标结果；

基于所述预测目标结果和所述样本目标结果分别调整所述第二网络和所述增强网络的参数。

可选地，在本公开上述任一方法实施例中，所述经增强网络基于所述第二样本结果获得增强操作，基于所述增强操作对所述第一样本结果进行调整，基于调整后的第一样本结果获得预测目标结果，包括：

经增强网络基于所述第二样本结果获得增强操作，基于所述增强操作对所述第一样本结果进行调整，得到第二中间样本结果；

基于所述第二中间样本结果通过所述第二网络得到中间预测结果，将所述中间预测结果输入所述增强网络；

响应于达到预设条件，输出所述中间预测结果作为所述预测目标结果。

可选地，在本公开上述任一方法实施例中，所述基于所述预测目标结果和所述样本目标结果调整所述增强网络的参数，包括：

基于所述中间预测结果和样本目标结果确定第一损失；

基于至少一个所述第一损失调整所述增强网络的参数。

可选地，在本公开上述任一方法实施例中，所述基于至少一个所述第一损失调整所述增强网络的参数，包括：

按照获得至少一个所述第一损失的顺序，按序基于至少一个所述第一损失调整所述增强网络的参数。

可选地，在本公开上述任一方法实施例中，所述基于所述中间预测结果和样本目标结果确定第一损失，包括：

基于所述中间预测结果和样本目标结果确定当前中间奖励；

基于所述中间预测结果经得分网络确定当前预测奖励；所述得分网络与所述增强网络共享部分网络结构；

基于所述中间预测结果经所述增强网络和所述第二网络确定下一个中间预测结果；基于所述下一个中间预测结果经所述得分网络确定下一个预测奖励；

基于所述当前中间奖励、所述当前预测奖励和所述下一个预测奖励确定所述第一损失。

可选地，在本公开上述任一方法实施例中，所述基于所述当前中间奖励、所述当前预测奖励和所述下一个预测奖励确定所述第一损失，包括：

基于所述当前中间奖励加所述下一个预测奖励，减去所述当前预测奖励，得所述第一损失。

可选地，在本公开上述任一方法实施例中，所述基于所述预测目标结果和所述样本目标结果调整所述第二网络的参数，包括：

基于所述预测目标结果和所述样本目标结果确定第二损失，基于所述第二损失调整所述第二网络的参数。

根据本公开实施例的另一个方面，提供一种系统增强学习装置，包括：

第一网络模块，用于对输入图像进行处理，得到第一结果；

第二网络模块，用于对所述第一结果进行处理，得到第二结果；

增强网络模块，用于基于所述第二结果获得增强操作；

执行器，用于基于所述增强操作对所述第一结果进行调整，基于调整后的第一结果获得目标结果。

可选地，在本公开上述任一装置实施例中，所述增强网络模块，包括：

调整模块，用于基于所述第二网络模块输出的第二结果对所述第一结果进行调整，得到第一中间结果；

结果输出模块，与将所述第一中间结果输入所述第二网络模块，基于所述第一中间结果得到第二结果，将所述第二结果输入所述调整模块，响应于达到预设条件，输出所述第二结果作为所述目标结果。

可选地，在本公开上述任一装置实施例中，所述调整模块，用于基于所述第二结果经增强网络模块获得调整动作概率，基于所述调整动作概率确定至少一个调整动作；基于所述至少一个调整动作对所述第一结果进行调整，得到第一中间结果。

可选地，在本公开上述任一装置实施例中，所述增强网络模块，用于基于所述第二结果和所述第一结果，获得调整动作概率，基于所述调整动作概率确定至少一个调整动作；

所述执行器，用于基于所述至少一个调整动作对所述第一结果进行调整，得到第一中间结果。

可选地，在本公开上述任一装置实施例中，所述增强网络模块，用于基于所述第二结果和所述第一结果得到连接结果；基于所述连接结果经增强网络获得调整动作概率，基于所述调整动作概率确定至少一个调整动作。

可选地，在本公开上述任一装置实施例中，还包括：训练模块，用于基于样本图像训练所述包括第一网络、第二网络和增强网络的系统，所述样本图像已知样本目标结果。

可选地，在本公开上述任一装置实施例中，所述第一网络包括以下一个或多个网络：

检测神经网络、识别神经网络、分类神经网络；

所述第二网络包括以下一个或多个网络：

检测神经网络、识别神经网络、分类神经网络。

可选地，在本公开上述任一装置实施例中，所述训练模块，用于将样本图像输入所述第一网络模块、所述第二网络模块、所述增强网络模块和所述执行器，获得预测目标结果；基于所述预测目标结果和所述样本目标结果分别调整所述第二网络和所述增强网络的参数。

可选地，在本公开上述任一装置实施例中，所述增强模块，用于基于所述第二网络模块得到的第二样本结果获得增强操作；

所述执行器，用于基于所述增强操作对所述第一网络模块得到的第一样本结果进行调整，得到第二中间样本结果；

所述第二网络模块，用于基于所述第二中间样本结果得到中间预测结果，将中间预测结果输入增强模块，响应于达到预设条件，输出所述中间预测结果作为所述预测目标结果。

可选地，在本公开上述任一装置实施例中，所述训练模块，包括：

第一损失模块，用于基于所述中间预测结果和样本目标结果确定第一损失；

第一参数调整模块，用于基于至少一个所述第一损失调整所述增强网络的参数。

可选地，在本公开上述任一装置实施例中，所述第一参数调整模块，用于按照获得至少一个所述第一损失的顺序，按序基于至少一个所述第一损失调整所述增强网络的参数。

可选地，在本公开上述任一装置实施例中，所述第一损失模块，包括：

奖励确定模块，用于基于所述中间预测结果和样本目标结果确定当前中间奖励；基于所述中间预测结果经得分网络确定当前预测奖励；所述得分网络与所述增强网络共享部分网络结构；基于所述中间预测结果经所述增强网络和所述第二网络确定下一个中间预测结果；基于所述下一个中间预测结果经所述得分网络确定下一个预测奖励；

损失确定模块，用于基于所述当前中间奖励、所述当前预测奖励和所述下一个预测奖励确定所述第一损失。

可选地，在本公开上述任一装置实施例中，所述损失确定模块，用于基于所述当前中间奖励加所述下一个预测奖励，减去所述当前预测奖励，得所述第一损失。

可选地，在本公开上述任一装置实施例中，所述第二网络调整模块，用于基于所述预测目标结果和所述样本目标结果确定第二损失，基于所述第二损失调整所述第二网络的参数。

根据本公开实施例的另一个方面，提供一种电子设备，包括：存储器，用于存储可执行指令；

以及处理器，用于与所述存储器通信以执行所述可执行指令从而完成如上任意一项所

述系统增强学习方法的操作。

根据本公开实施例的另一个方面，提供一种计算机存储介质，用于存储计算机可读的指令，所述指令被执行时执行如上任意一项所述系统增强学习方法的操作。

根据本公开实施例的另一个方面，提供一种计算机程序产品，包括计算机可读代码，当所述计算机可读代码在设备上运行时，所述设备中的处理器执行用于实现如上任意一项所述系统增强学习方法的指令。

基于本公开上述实施例提供的一种系统增强学习方法和装置、电子设备、计算机存储介质，基于系统的第一网络对输入图像进行处理，得到第一结果；将第一结果输入系统的第二网络，得到第二结果；基于第二结果获得增强操作，基于增强操作对第一结果进行调整，获得目标结果，通过增强网络实现信息从下游层级反馈回上游层级，优化了系统的输出结果。

下面通过附图和实施例，对本公开的技术方案做进一步的详细描述。

附图说明

构成说明书的一部分的附图描述了本公开的实施例，并且连同描述一起用于解释本公开的原理。

参照附图，根据下面的详细描述，可以更加清楚地理解本公开，其中：

图1为本公开实施例提供的系统增强学习方法的一个流程示意图。

图2为本公开实施例提供的系统增强学习方法的另一流程示意图。

图3为本公开实施例提供的系统增强学习方法的一个训练示例图。

图4为本公开实施例提供的系统增强学习装置的一个结构示意图。

图5为适于用来实现本公开实施例的终端设备或服务器的电子设备的结构示意图。

具体实施方式

现在将参照附图来详细描述本公开的各种示例性实施例。应注意到：除非另外具体说明，否则在这些实施例中阐述的部件和步骤的相对布置、数字表达式和数值不限制本公开的范围。

同时，应当明白，为了便于描述，附图中所示出的各个部分的尺寸并不是按照实际的比例关系绘制的。

以下对至少一个示例性实施例的描述实际上仅仅是说明性的，决不作为对本公开及其应用或使用的任何限制。

对于相关领域普通技术人员已知的技术、方法和设备可能不作详细讨论，但在适当情况下，所述技术、方法和设备应当被视为说明书的一部分。

应注意：相似的标号和字母在下面的附图中表示类似项，因此，一旦某一项在一个附图中被定义，则在随后的附图中不需要对其进行进一步讨论。

本公开实施例可以应用于计算机系统/服务器，其可与众多其它通用或专用计算系统环境或配置一起操作。适于与计算机系统/服务器一起使用的众所周知的计算系统、环境和/或配置的例子包括但不限于：个人计算机系统、服务器计算机系统、瘦客户机、厚客户机、手持或膝上设备、基于微处理器的系统、机顶盒、可编程消费电子产品、网络个人电脑、小型计算机系统、大型计算机系统和包括上述任何系统的分布式云计算技术环境，等等。

计算机系统/服务器可以在由计算机系统执行的计算机系统可执行指令（诸如程序模块）的一般语境下描述。通常，程序模块可以包括例程、程序、目标程序、组件、逻辑、数据结构等等，它们执行特定的任务或者实现特定的抽象数据类型。计算机系统/服务器可以在分布式云计算环境中实施，分布式云计算环境中，任务是由通过通信网络链接的远程处理设备执行的。在分布式云计算环境中，程序模块可以位于包括存储设备的本地或远程计算系统存储介质上。

多层次系统在不同层级的连接上经常是不可微的，这样整个系统就无法联合优化，无法达到最好的效果。其中，不可微指无法求梯度。例如：涉及到包围框的系统，在进行裁剪包围框内的图片的操作时，无法计算这个操作对包围框坐标的梯度。

图1为本公开实施例提供的系统增强学习方法的一个流程示意图。如图1所示，该实施例方法包括：

步骤110，基于系统的第一网络对输入图像进行处理，得到第一结果。

可选地，系统的第一网络可以是实现一个具体功能的神经网络，或神经网络的部分网络层，还可以包括多个神经网络；可选地，系统中的第一网络与第二网络存在关联关系，例如：第一网络是人脸检测网络，第二网络是人脸识别网络，此时第二网络以第一网络输出的检测结果作为输入，第二网络以第一网络的输出为基础得到识别结果；通常情况下，第一网络与第二网络的关联只有正向传播的检测结果，不同层级之间的输出具有强相关性，错误的检测结果（第一网络的输出）将导致错误的识别结果（第二网络的输出）。

例如，第一网络可以包括但不限于以下一个或多个网络：检测神经网络、识别神经网络、分类神经网络；

第二网络可以包括但不限于以下一个或多个网络：检测神经网络、识别神经网络、分类神经网络。

在一个可选示例中，该步骤 S110 可以由处理器调用存储器存储的相应指令执行，也可以由被处理器运行的第一网络模块 41 执行。

步骤 120，将第一结果输入系统的第二网络，得到第二结果。

在一个可选示例中，该步骤 S120 可以由处理器调用存储器存储的相应指令执行，也可以由被处理器运行的第二网络模块 42 执行。

步骤 130，经增强网络基于第二结果获得增强操作，基于增强操作对第一结果进行调整，基于调整后的第一结果获得目标结果。

为了得到更好的第二结果，可选地，在系统中添加增强网络，该增强网络以第二网络输出的第二结果为输入，以输出的增强操作对第一结果进行调整，实现下游层级对上游层级的信息反馈，经过信息反馈获得目标结果。

在一个可选示例中，该步骤 S130 可以由处理器调用存储器存储的相应指令执行，也可以由被处理器运行的增强网络模块 43 执行。

基于本公开上述实施例提供的一种系统增强学习方法，基于系统的第一网络对输入图像进行处理，得到第一结果；将第一结果输入系统的第二网络，得到第二结果；经增强网络基于第二结果获得增强操作，基于增强操作对第一结果进行调整，获得目标结果，通过增强网络实现信息从下游层级反馈回上游层级，优化了系统的输出结果。

在一个或多个可选的实施例中，操作 130 可以包括：

经增强网络基于第二网络输出的第二结果获得增强操作，基于增强操作对第一结果进行调整，得到第一中间结果；

将第一中间结果输入第二网络，基于第一中间结果得到第二结果，将第二结果输入增强网络；

迭代执行上述过程，响应于达到预设条件，输出第二结果作为目标结果。

可选地，预设条件可以包括：迭代调整次数达到预设次数，或第二结果的效果达到预期等，本公开实施例不限制预设条件的具体形式。

可选地，增强网络基于第二结果获得对第一结果的调整动作，而一个或较少的调整动作调整后得到的第一中间结果可能不能获得较好的目标结果，因此，通过迭代执行，每次以上次调整后的第一中间结果获得的第二结果为增强网络的输入，得到新的调整动作，直到达到预设条件。

可选地，增强操作可以包括至少一个调整动作，

经增强网络基于第二网络输出的第二结果获得增强操作，基于增强操作对第一结果进行调整，得到第一中间结果，包括：

基于第二结果经增强网络获得调整动作概率，基于调整动作概率确定至少一个调整动作；

基于至少一个调整动作对第一结果进行调整，得到第一中间结果。

可选地，增强网络可以是一个分类神经网络，基于第二结果从多个调整动作中确定一个或多个概率较大（如，概率最大）的调整动作，对第一结果进行调整。可选地，在一个示例中，系统输入的为 RGB 图像，需要输出的目标结果是图像识别结果（如：语义分割或者人体姿态检测），系统处理过程可以包括：

a. 利用第一网络，对输入的 RGB 图像进行处理，得到第一结果 Ω ； Ω 是一个抽象的参数，对于不同的识别任务， Ω 可以不同。为方便理解，以下针对该示例的讨论以多层级多人姿态预测系统为例，第一个阶段的网络（对应第一网络）是一个物体检测网络，输出的 Ω 即为物体的包围框坐标。

b. 以第一结果 Ω 作为第二网络的输入或根据第一结果 Ω 计算出第二网络的输入，经过第二网络输出第二结果；通常来讲，第二网络的输入与第一网络的输出相关。若 Ω 为包围框的坐标，则第二网络的输入即为原图片在包围框内的部分。第二网络的输出则是物体识别的结果，如：人体的各个关键点在图中分布的热力图。

c. 增强网络（例如：增强网络采用一个分类网络，一个 18 层的残差网络 (ResNet-18)）根据第二网络的输出，选择一个动作调整第一结果 Ω ，并返回到步骤 b，迭代执行步骤 c 和步骤 b，直到达到预设条件，本示例中可以设置预设条件为迭代次数，即达到预设迭代次数，其预设迭代次数根据需要进行设置（例如：设置预设迭代次数为 20 次）。增强网络以第二网络的输出作为输入，产生一个针对多动作的概率分布，并选出概率较大的至少一个作为决策的结果；其中多个动作可以包括：以第一网络的输出为包围框为例，动作集为一系列可以调整包围框的动作，例如：放大或缩小包围框的高度和/或宽度，上下左右平移，以及一个不调整的动作等多个动作。

可选地，利用执行器根据所选的动作调整第一结果 Ω ，执行器是一种通用的说法，因为增强网络确定的只是一个动作的编号（分类网络的分类结果），需要一个功能将这个编号转化为动作（如调整包围框的具体动作）。可以通过一个函数来直接实现，通常将这个实现的函数称为执行器。

为了防止第二网络输出的结果偏差过大，需要将第一网络输出的第一结果作为增强网络的参考输入，可选地，经增强网络基于第二网络输出的第二结果对第一结果进行调整，得到第一中间结果，包括：

基于第二结果和第一结果，经增强网络获得调整动作概率，基于调整动作概率确定至少一个调整动作；

可选地，可以基于第二结果和第一结果得到连接结果，基于连接结果经增强网络获得调整动作概率，或者，当第一结果和第二结果都是图像时，基于图像通道级联第一结果和第二结果，基于级联后的第一结果和第二结果获得调整动作概率，本公开实施例不限制第一结果和第二结果之间的具体关联方法，只要在获得调整动作概率时参考第一结果即可。

基于至少一个调整动作对第一结果进行调整，得到第一中间结果。

在一个或多个可选的实施例中，还包括：基于样本图像训练系统，样本图像已知样本目标结果。

为了实现更好的增强学习效果，需要对第一结果调整的调整动作是较为准确的，因此，在增强学习之前，还可以基于已知样本目标结果的样本图像对系统中的至少一个网络进行训练；可选地，仅对第二网络和增强网络进行训练，增强网络经过训练可以得到更准确的调整动作，而第二网络经过训练可以得到更好的目标结果。

可选地，基于样本图像训练系统，包括：

基于系统的第一网络对样本图像进行处理，得到第一样本结果；

将第一样本结果输入系统的第二网络，得到第二样本结果；

经增强网络基于第二样本结果获得增强操作，基于增强操作对第一样本结果进行调整，基于调整后的第一样本结果获得预测目标结果；

基于预测目标结果和样本目标结果分别调整第二网络和增强网络的参数。

作为一个可选示例，图2为本公开实施例提供的系统增强学习方法的另一流程示意图。如图2所示，第二网络与增强网络以相互的结果互为输入，因此，训练过程中，需要第二网络和增强网络共同训练；训练过程中需要系统中包括的网络与预测过程一样，对样本图像进行处理，得到预测目标结果，基于预测目标结果和样本目标结果调整第二网络和增强网络的参数。可选地，经增强网络基于第二样本结果获得增强操作，基于增强操作对第一样本结果进行调整，基于调整后的第一样本结果获得预测目标结果，包括：经增强网络基于第二样本结果获得增强操作，基于增强操作对第一样本结果进行调整，得到第二中间样本结果；

基于第二中间样本结果通过第二网络得到中间预测结果，将中间预测结果输入增强网络；

迭代执行上述过程，响应于达到预设条件，输出中间预测结果作为预测目标结果。

可选地，预设条件可以包括：迭代次数达到预设次数，或第二结果的效果达到预期等，本公开实施例不限制预设条件的具体形式。

本公开实施例中，将增强网络和第二网络平等处理，分别以最后输出的预测目标结果和样本目标结果对增强网络和第二网络的参数进行调整。但是，由于增强学习（增强网络）要求的数据量远远大于监督学习，并且增强学习的梯度方差也远比监督学习（第二网络）的梯度方差大，因此如果直接把增强网络和第二网络放在一起联合训练的话，会导致监督学习的网络受到极大的影响，降低准确率。为了解决这个问题，本公开实施例提出了平衡增强学习和监督学习不同训练要求的方法。

可选地，基于预测目标结果和样本目标结果调整增强网络的参数，包括：

基于中间预测结果和样本目标结果确定第一损失；

基于至少一个第一损失调整增强网络的参数。

可选地，在训练过程中，对于输入，增强网络会不断调整第一结果，多次调整对应多个动作，迭代完成后，得到一个动作序列（包括多个动作）。在调整一次第一样本结果时，第二网络会产生一个新的对应的预测中间样本结果（如热力图）。对于增强网络，本实施例通过整个动作序列的预测中间样本结果去训练它，而对于第二网络的有监督学习，可以只选择用最终的预测目标结果去训练它。通过不同的训练动作，可以实现平衡监督学习和增强学习的不同训练要求。

可选地，基于至少一个第一损失调整增强网络的参数，包括：

按照获得至少一个第一损失的顺序，按序基于至少一个第一损失调整增强网络的参数。

基于得到的中间预测结果与样本目标结果可确定损失，得到至少一个损失，基于得到的至少一个损失调整增强网络的参数。

可选地，基于中间预测结果和样本目标结果确定第一损失，包括：

基于中间预测结果和样本目标结果确定当前中间奖励；

基于中间预测结果经得分网络确定当前预测奖励；

其中，得分网络与增强网络共享部分网络结构；可选地，得分网络与增强网络只有最后一层不同，增强网络最后一层为全连接层，得到一维特征向量，基于该一维特征向量确定选择动作的概率；得分网络最后一层为全连接层，得到一个特征值，该特征值即为得分。

基于中间预测结果经增强网络和第二网络确定下一个中间预测结果；基于下一个中间预测结果经得分网络确定下一个预测奖励；

基于当前中间奖励、当前预测奖励和下一个预测奖励确定第一损失。

基于中间预测结果得到至少一个第一损失，例如，每个中间预测结果对应一个第一损失。

本实施例中，图3为本公开实施例提供的系统增强学习方法的一个训练示例图。如图3所示，增加一个得分网络，增强网络和得分网络构成的增强学习代理模型采用了执行者-评分者(增强网络结合得分网络，Actor Critic)的算法，使用了时序不同的学习方法(例如：TD-Learning)；其中，Actor Critic 方法：结合了政策渐变(Policy Gradient，执行者)和函数逼近(Function Approximation，评分者)的方法，Actor 实现增强网络基于概率确定动作，Critic 基于 Actor 的行为评判基于概率确定动作的得分，Actor 根据 Critic 的评分修改确定动作的概率，Actor Critic 方法的优势在于，可以进行单步更新，比传统的方法更快(回合结束更新)。TD-learning 这个算法也是已知的算法。增强学习的反向传播(back-propagation)算法大概有两种思路，一种是一次性把整个序列都执行完，然后算出整个序列的奖励(reward)，再一步步往回优化，执行反向传播(back-propagation)算法。另一种是，每次只执行一次动作，然后就马上优化一次。这个就是 TD-learning 算法的思想。TD-learning 算法的好处在于，网络收敛得会更快，能把训练时间缩短而不会对性能造成很大影响。

可选地，基于当前中间奖励、当前预测奖励和下一个预测奖励确定第一损失，包括：基于当前中间奖励加下一个预测奖励，减去当前预测奖励，得第一损失。

通过当前预测奖励和下一个预测奖励的加入，可减小损失函数的方差，让训练变得更加稳定。

可选地，基于预测目标结果和样本目标结果调整第二网络的参数，包括：

基于预测目标结果和样本目标结果确定第二损失，基于第二损失调整第二网络的参数。

本领域普通技术人员可以理解：实现上述方法实施例的全部或部分步骤可以通过程序指令相关的硬件来完成，前述的程序可以存储于一计算机可读取存储介质中，该程序在执行时，执行包括上述方法实施例的步骤；而前述的存储介质包括：ROM、RAM、磁碟或者光盘等各种可以存储程序代码的介质。

图4为本公开实施例提供的系统增强学习装置的一个结构示意图。该实施例的装置可用于实现本公开上述各方法实施例。如图4所示，该实施例的装置包括：

第一网络模块 41，用于对输入图像进行处理，得到第一结果。

第二网络模块 42，用于对第一结果进行处理，得到第二结果。

增强网络模块 43，用于基于第二结果获得增强动作。

执行器 44，用于基于增强操作对第一结果进行调整，基于调整后的第一结果获得目标结果。

基于本公开上述实施例提供的一种系统增强学习装置，基于系统的第一网络对输入图像进行处理，得到第一结果；将第一结果输入系统的第二网络，得到第二结果；经增强网络基于第二结果获得增强操作，基于增强操作对第一结果进行调整，获得目标结果，通过增强网络实现信息从下游层级反馈回上游层级，优化了系统的输出结果。

在一个或多个可选的实施例中，增强网络模块 43，包括：

调整模块，用于基于第二网络模块输出的第二结果对第一结果进行调整，得到第一中间结果；

结果输出模块，与将第一中间结果输入第二网络模块，基于第一中间结果得到第二结果，将第二结果输入调整模块；响应于达到预设条件，输出第二结果作为目标结果。

可选地，预设条件可以包括：调整次数达到预设次数，或第二结果的效果达到预期等，本公开实施例不限制预设条件的具体形式。

可选地，增强网络基于第二结果获得对第一结果的调整动作，而一个或较少的调整动作调整后得到的第一中间结果可能不能获得较好的目标结果，因此，通过迭代执行，每次以上次调整后的第一中间结果获得的第二结果为增强网络的输入，得到新的调整动作，直到达到预设条件。

可选地，调整模块，用于基于第二结果经增强网络模块获得调整动作概率，基于调整动作概率确定至少一个调整动作；基于至少一个调整动作对第一结果进行调整，得到第一中间结果。

可选地，增强网络模块，用于基于第二结果和第一结果，获得调整动作概率，基于调整动作概率确定至少一个调整动作；

上述执行器，用于基于至少一个调整动作对第一结果进行调整，得到第一中间结果。

可选地，增强网络模块，用于基于第二结果和第一结果得到连接结果；

基于连接结果经增强网络获得调整动作概率，基于调整动作概率确定至少一个调整动作。

在一个或多个可选的实施例中，还包括：训练模块，用于基于样本图像训练包括第一

网络、第二网络和增强网络的系统。

其中，样本图像已知样本目标结果。

为了实现更好的增强学习效果，需要对第一结果调整的调整动作是较为准确的，因此，在增强学习之前，还可以基于已知样本目标结果的样本图像对系统中的至少一个网络进行训练；可选地，仅对第二网络和增强网络进行训练，增强网络经过训练可以得到更准确的调整动作，而第二网络经过训练可以得到更好的目标结果。

可选地，第一网络包括以下一个或多个网络：检测神经网络、识别神经网络、分类神经网络；

第二网络包括以下一个或多个网络：检测神经网络、识别神经网络、分类神经网络。

可选地，训练模块，用于将样本图像输入第一网络模块、第二网络模块、增强网络模块和执行器，获得预测目标结果；基于预测目标结果和样本目标结果分别调整第二网络和增强网络的参数。

可选地，增强模块，用于基于第二网络模块得到的第二样本结果获得增强操作；

执行器，用于基于增强操作对第一网络模块得到的第一样本结果进行调整，得到第二中间样本结果；

第二网络模块，用于基于第二中间样本结果得到中间预测结果，将中间预测结果输入增强模块，响应于达到预设条件，输出中间预测结果作为预测目标结果。

可选地，训练模块，包括：

第一损失模块，用于基于中间预测结果和样本目标结果确定第一损失；

第一参数调整模块，用于基于至少一个第一损失调整增强网络的参数。

可选地，第一参数调整模块，用于按照获得至少一个第一损失的顺序，按序基于至少一个第一损失调整增强网络的参数。

可选地，第一损失模块，包括：

奖励确定模块，用于基于中间预测结果和样本目标结果确定当前中间奖励；基于中间预测结果经得分网络确定当前预测奖励；得分网络与增强网络共享部分网络结构；基于中间预测结果经增强网络和第二网络确定下一个中间预测结果；基于下一个中间预测结果经得分网络确定下一个预测奖励；

损失确定模块，用于基于当前中间奖励、当前预测奖励和下一个预测奖励确定第一损失。

可选地，损失确定模块，用于基于当前中间奖励加下一个预测奖励，减去当前预测奖

励，得第一损失。

可选地，第二网络调整模块，用于基于预测目标结果和样本目标结果确定第二损失，基于第二损失调整第二网络的参数。

根据本公开实施例的另一个方面，提供一种电子设备，包括：存储器，用于存储可执行指令；

以及处理器，用于与存储器通信以执行可执行指令从而完成本公开上述任一实施例中所述的系统增强学习方法。

根据本公开实施例的另一个方面，提供一种计算机存储介质，用于存储计算机可读的指令，指令被处理器执行时，该处理器执行本公开上述任一实施例提供的系统增强学习方法。

根据本公开实施例的另一个方面，提供一种计算机程序产品，包括计算机可读代码，当计算机可读代码在设备上运行时，设备中的处理器执行本公开上述任一实施例提供的系统增强学习方法的操作。

根据本公开实施例的再一个方面，提供的另一种计算机程序产品，用于存储计算机可读指令，所述指令被执行时使得计算机执行上述任一可能的实现方式提供的系统增强学习方法的操作。

在一个或多个可选实施方式中，本公开实施例还提供了一种计算机程序产品，用于存储计算机可读指令，所述指令被执行时使得计算机执行上述任一实施例中所述的系统增强学习方法的操作。

该计算机程序产品可以具体通过硬件、软件或其结合的方式实现。在一个可选例子中，所述计算机程序产品具体体现为计算机存储介质，在另一个可选例子中，所述计算机程序产品具体体现为软件产品，例如软件开发包（Software Development Kit, SDK）等等。

根据本公开实施例还提供了另一种系统增强学习方法及其对应的装置和电子设备、计算机存储介质以及计算机程序产品，其中，该方法包括：基于系统的第一网络对输入图像进行处理，得到第一结果；将第一结果输入系统的第二网络，得到第二结果；经增强网络基于第二结果对第一结果进行调整，获得目标结果。

在一些实施例中，该目标跟踪指示可以具体为调用指令，第一装置可以通过调用的方式指示第二装置执行目标跟踪，相应地，响应于接收到调用指令，第二装置可以执行上述目标跟踪方法中的任意实施例中的步骤和/或流程。

应理解，本公开实施例中的“第一”、“第二”等术语仅仅是为了区分，而不应理解成

对本公开实施例的限定。

还应理解，在本公开中，“多个”可以指两个或两个以上，“至少一个”可以指一个、两个或两个以上。

还应理解，对于本公开中提及的任一部件、数据或结构，在没有明确限定或者在前后文给出相反启示的情况下，一般可以理解为一个或多个。

还应理解，本公开对各个实施例的描述着重强调各个实施例之间的不同之处，其相同或相似之处可以相互参考，为了简洁，不再一一赘述。

本公开实施例还提供了一种电子设备，例如可以是移动终端、个人计算机（PC）、平板电脑、服务器等。下面参考图 5，其示出了适于用来实现本公开实施例的终端设备或服务器的电子设备 500 的结构示意图：如图 5 所示，电子设备 500 包括一个或多个处理器、通信部等，所述一个或多个处理器例如：一个或多个中央处理单元（CPU）501，和/或一个或多个图像处理器（GPU）513 等，处理器可以根据存储在只读存储器（ROM）502 中的可执行指令或者从存储部分 508 加载到随机访问存储器（RAM）503 中的可执行指令而执行各种适当的动作和处理。通信部 512 可包括但不限于网卡，所述网卡可包括但不限于 IB（Infiniband）网卡。

处理器可与只读存储器 502 和/或随机访问存储器 503 中通信以执行可执行指令，通过总线 504 与通信部 512 相连、并经通信部 512 与其他目标设备通信，从而完成本公开实施例提供的任一项方法对应的操作，例如，基于系统的第一网络对输入图像进行处理，得到第一结果；将第一结果输入系统的第二网络，得到第二结果；经增强网络基于第二结果获得增强操作，基于增强操作对第一结果进行调整，获得目标结果。

此外，在 RAM 503 中，还可存储有装置操作所需的各种程序和数据。CPU501、ROM502 以及 RAM503 通过总线 504 彼此相连。在有 RAM503 的情况下，ROM502 为可选模块。RAM503 存储可执行指令，或在运行时向 ROM502 中写入可执行指令，可执行指令使中央处理单元 501 执行上述通信方法对应的操作。输入/输出（I/O）接口 505 也连接至总线 504。通信部 512 可以集成设置，也可以设置为具有多个子模块（例如多个 IB 网卡），并在总线链接上。

以下部件连接至 I/O 接口 505：包括键盘、鼠标等的输入部分 506；包括诸如阴极射线管（CRT）、液晶显示器（LCD）等以及扬声器等的输出部分 507；包括硬盘等的存储部分 508；以及包括诸如 LAN 卡、调制解调器等网络接口卡的通信部分 509。通信部分 509 经由诸如因特网的网络执行通信处理。驱动器 510 也根据需要连接至 I/O 接口 505。可拆卸

介质 511, 诸如磁盘、光盘、磁光盘、半导体存储器等等, 根据需要安装在驱动器 510 上, 以便于从其上读出的计算机程序根据需要被安装入存储部分 508。

需要说明的, 如图 5 所示的架构仅为一种可选实现方式, 在具体实践过程中, 可根据实际需要对上述图 5 的部件数量和类型进行选择、删减、增加或替换; 在不同功能部件设置上, 也可采用分离设置或集成设置等实现方式, 例如 GPU513 和 CPU501 可分离设置或者可将 GPU513 集成在 CPU501 上, 通信部可分离设置, 也可集成设置在 CPU501 或 GPU513 上, 等等。这些可替换的实施方式均落入本公开公开的保护范围。

特别地, 根据本公开的实施例, 上文参考流程图描述的过程可以被实现为计算机软件程序。例如, 本公开的实施例包括一种计算机程序产品, 其包括有形地包含在机器可读介质上的计算机程序, 计算机程序包含用于执行流程图所示的方法的程序代码, 程序代码可包括对应执行本公开实施例提供的方法步骤对应的指令, 例如, 基于系统的第一网络对输入图像进行处理, 得到第一结果; 将第一结果输入系统的第二网络, 得到第二结果; 经增强网络基于第二结果获得增强操作, 基于增强操作对第一结果进行调整, 获得目标结果。在这样的实施例中, 该计算机程序可以通过通信部分 509 从网络上被下载和安装, 和/或从可拆卸介质 511 被安装。在该计算机程序被中央处理单元 (CPU) 501 执行时, 执行本公开的方法中限定的上述功能的操作。

本说明书中各个实施例均采用递进的方式描述, 每个实施例重点说明的都是与其它实施例的不同之处, 各个实施例之间相同或相似的部分相互参见即可。对于系统实施例而言, 由于其与方法实施例基本对应, 所以描述的比较简单, 相关之处参见方法实施例的部分说明即可。

可能以许多方式来实现本公开的方法和装置。例如, 可通过软件、硬件、固件或者软件、硬件、固件的任何组合来实现本公开的方法和装置。用于所述方法的步骤的上述顺序仅是为了进行说明, 本公开的方法的步骤不限于以上具体描述的顺序, 除非以其它方式特别说明。此外, 在一些实施例中, 还可将本公开实施为记录在记录介质中的程序, 这些程序包括用于实现根据本公开的方法的机器可读指令。因而, 本公开还覆盖存储用于执行根据本公开的方法的程序的记录介质。

本公开的描述是为了示例和描述起见而给出的, 而并不是无遗漏的或者将本公开限于所公开的形式。很多修改和变化对于本领域的普通技术人员而言是显然的。选择和描述实施例是为了更好说明本公开的原理和实际应用, 并且使本领域的普通技术人员能够理解本公开从而设计适于特定用途的带有各种修改的各种实施例。

权利要求

1、一种系统增强学习方法，其特征在于，包括：

基于系统的第一网络对输入图像进行处理，得到第一结果；

将所述第一结果输入所述系统的第二网络，得到第二结果；

经增强网络基于所述第二结果获得增强操作，基于所述增强操作对所述第一结果进行调整，基于调整后的第一结果获得目标结果。

2、根据权利要求 1 所述的方法，其特征在于，所述经增强网络基于所述第二结果获得增强操作，基于所述增强操作对所述第一结果进行调整，基于所述调整后的第一结果获得目标结果，包括：

经所述增强网络基于所述第二网络输出的所述第二结果获得增强操作，基于所述增强操作对所述第一结果进行调整，得到第一中间结果；

将所述第一中间结果输入所述第二网络，基于所述第一中间结果得到第二结果，将所述第二结果输入所述增强网络；

响应于达到预设条件，输出所述第二结果作为所述目标结果。

3、根据权利要求 2 所述的方法，其特征在于，所述增强操作包括至少一个调整动作；

所述经增强网络基于所述第二网络输出的第二结果获得增强操作，基于所述增强操作对所述第一结果进行调整，得到第一中间结果，包括：

基于所述第二结果经增强网络获得调整动作概率，基于所述调整动作概率确定至少一个调整动作；

基于所述至少一个调整动作对所述第一结果进行调整，得到第一中间结果。

4、根据权利要求 2 所述的方法，其特征在于，所述经增强网络基于所述第二网络输出的第二结果对所述第一结果进行调整，得到第一中间结果，包括：

基于所述第二结果和所述第一结果，经增强网络获得调整动作概率，基于所述调整动作概率确定至少一个调整动作；

基于所述至少一个调整动作对所述第一结果进行调整，得到第一中间结果。

5、根据权利要求 4 所述的方法，其特征在于，所述基于所述第二结果和所述第一结果经增强网络获得调整动作概率，基于所述调整动作概率确定至少一个调整动作，包括：

基于所述第二结果和所述第一结果得到连接结果；

基于所述连接结果经增强网络获得调整动作概率，基于所述调整动作概率确定至少一个调整动作。

6、根据权利要求 1-5 任一所述的方法，其特征在于，所述第一网络包括以下一个或多个网络：

检测神经网络、识别神经网络、分类神经网络；

所述第二网络包括以下一个或多个网络：

检测神经网络、识别神经网络、分类神经网络。

7、根据权利要求 1-6 任一所述的方法，其特征在于，还包括：基于样本图像训练所述系统，所述样本图像已知样本目标结果。

8、根据权利要求 7 所述的方法，其特征在于，所述基于样本图像训练所述系统，包括：

基于所述系统的第一网络对所述样本图像进行处理，得到第一样本结果；

将所述第一样本结果输入所述系统的第二网络，得到第二样本结果；

经增强网络基于所述第二样本结果获得增强操作，基于所述增强操作对所述第一样本结果进行调整，基于调整后的第一样本结果获得预测目标结果；

基于所述预测目标结果和所述样本目标结果分别调整所述第二网络和所述增强网络的参数。

9、根据权利要求 8 所述的方法，其特征在于，所述经增强网络基于所述第二样本结果获得增强操作，基于所述增强操作对所述第一样本结果进行调整，基于调整后的第一样本结果获得预测目标结果，包括：

经增强网络基于所述第二样本结果获得增强操作，基于所述增强操作对所述第一样本结果进行调整，得到第二中间样本结果；

基于所述第二中间样本结果通过所述第二网络得到中间预测结果，将所述中间预测结果输入所述增强网络；

响应于达到预设条件，输出所述中间预测结果作为所述预测目标结果。

10、根据权利要求 9 所述的方法，其特征在于，所述基于所述预测目标结果和所述样本目标结果调整所述增强网络的参数，包括：

基于所述中间预测结果和样本目标结果确定第一损失；

基于至少一个所述第一损失调整所述增强网络的参数。

11、根据权利要求 10 所述的方法，其特征在于，所述基于至少一个所述第一损失调整所述增强网络的参数，包括：

按照获得至少一个所述第一损失的顺序，按序基于至少一个所述第一损失调整所述增

强网络的参数。

12、根据权利要求 10 或 11 所述的方法，其特征在于，所述基于所述中间预测结果和样本目标结果确定第一损失，包括：

基于所述中间预测结果和样本目标结果确定当前中间奖励；

基于所述中间预测结果经得分网络确定当前预测奖励；所述得分网络与所述增强网络共享部分网络结构；

基于所述中间预测结果经所述增强网络和所述第二网络确定下一个中间预测结果；基于所述下一个中间预测结果经所述得分网络确定下一个预测奖励；

基于所述当前中间奖励、所述当前预测奖励和所述下一个预测奖励确定所述第一损失。

13、根据权利要求 12 所述的方法，其特征在于，所述基于所述当前中间奖励、所述当前预测奖励和所述下一个预测奖励确定所述第一损失，包括：

基于所述当前中间奖励加所述下一个预测奖励，减去所述当前预测奖励，得所述第一损失。

14、根据权利要求 8-13 任一所述的方法，其特征在于，所述基于所述预测目标结果和所述样本目标结果调整所述第二网络的参数，包括：

基于所述预测目标结果和所述样本目标结果确定第二损失，基于所述第二损失调整所述第二网络的参数。

15、一种系统增强学习装置，其特征在于，包括：

第一网络模块，用于对输入图像进行处理，得到第一结果；

第二网络模块，用于对所述第一结果进行处理，得到第二结果；

增强网络模块，用于基于所述第二结果获得增强操作；

执行器，用于基于所述增强操作对所述第一结果进行调整，基于调整后的第一结果获得目标结果。

16、根据权利要求 15 所述的装置，其特征在于，所述增强网络模块，包括：

调整模块，用于基于所述第二网络模块输出的第二结果对所述第一结果进行调整，得到第一中间结果；

结果输出模块，与将所述第一中间结果输入所述第二网络模块，基于所述第一中间结果得到第二结果，将所述第二结果输入所述调整模块，响应于达到预设条件，输出所述第二结果作为所述目标结果。

17、根据权利要求 16 所述的装置，其特征在于，所述调整模块，用于基于所述第二

结果经增强网络模块获得调整动作概率，基于所述调整动作概率确定至少一个调整动作；基于所述至少一个调整动作对所述第一结果进行调整，得到第一中间结果。

18、根据权利要求 16 所述的装置，其特征在于，所述增强网络模块，用于基于所述第二结果和所述第一结果，获得调整动作概率，基于所述调整动作概率确定至少一个调整动作；

所述执行器，用于基于所述至少一个调整动作对所述第一结果进行调整，得到第一中间结果。

19、根据权利要求 18 所述的装置，其特征在于，所述增强网络模块，用于基于所述第二结果和所述第一结果得到连接结果；基于所述连接结果经增强网络获得调整动作概率，基于所述调整动作概率确定至少一个调整动作。

20、根据权利要求 15-19 任一所述的装置，其特征在于，还包括：训练模块，用于基于样本图像训练所述包括第一网络、第二网络和增强网络的系统，所述样本图像已知样本目标结果。

21、根据权利要求 20 所述的装置，其特征在于，所述第一网络包括以下一个或多个网络：

检测神经网络、识别神经网络、分类神经网络；

所述第二网络包括以下一个或多个网络：

检测神经网络、识别神经网络、分类神经网络。

22、根据权利要求 20 或 21 所述的装置，其特征在于，所述训练模块，用于将样本图像输入所述第一网络模块、所述第二网络模块、所述增强网络模块和所述执行器，获得预测目标结果；基于所述预测目标结果和所述样本目标结果分别调整所述第二网络和所述增强网络的参数。

23、根据权利要求 22 所述的装置，其特征在于，所述增强模块，用于基于所述第二网络模块得到的第二样本结果获得增强操作；

所述执行器，用于基于所述增强操作对所述第一网络模块得到的第一样本结果进行调整，得到第二中间样本结果；

所述第二网络模块，用于基于所述第二中间样本结果得到中间预测结果，将中间预测结果输入增强模块，响应于达到预设条件，输出所述中间预测结果作为所述预测目标结果。

24、根据权利要求 23 所述的装置，其特征在于，所述训练模块，包括：

第一损失模块，用于基于所述中间预测结果和样本目标结果确定第一损失；

第一参数调整模块，用于基于至少一个所述第一损失调整所述增强网络的参数。

25、根据权利要求 24 所述的装置，其特征在于，所述第一参数调整模块，用于按照获得至少一个所述第一损失的顺序，按序基于至少一个所述第一损失调整所述增强网络的参数。

26、根据权利要求 24 或 25 所述的装置，其特征在于，所述第一损失模块，包括：

奖励确定模块，用于基于所述中间预测结果和样本目标结果确定当前中间奖励；基于所述中间预测结果经得分网络确定当前预测奖励；所述得分网络与所述增强网络共享部分网络结构；基于所述中间预测结果经所述增强网络和所述第二网络确定下一个中间预测结果；基于所述下一个中间预测结果经所述得分网络确定下一个预测奖励；

损失确定模块，用于基于所述当前中间奖励、所述当前预测奖励和所述下一个预测奖励确定所述第一损失。

27、根据权利要求 26 所述的装置，其特征在于，所述损失确定模块，用于基于所述当前中间奖励加所述下一个预测奖励，减去所述当前预测奖励，得所述第一损失。

28、根据权利要求 22-27 任一所述的装置，其特征在于，所述第二网络调整模块，用于基于所述预测目标结果和所述样本目标结果确定第二损失，基于所述第二损失调整所述第二网络的参数。

29、一种电子设备，其特征在于，包括：存储器，用于存储可执行指令；

以及处理器，用于与所述存储器通信以执行所述可执行指令从而完成权利要求 1 至 14 任意一项所述系统增强学习方法的操作。

30、一种计算机存储介质，用于存储计算机可读取的指令，其特征在于，所述指令被执行时执行权利要求 1 至 14 任意一项所述系统增强学习方法的操作。

31、一种计算机程序，包括计算机可读代码，其特征在于，当所述计算机可读代码在设备上运行时，所述设备中的处理器执行用于实现权利要求 1 至 14 任意一项所述系统增强学习方法的指令。