(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 109682392 A
(43)申请公布日 2019.04.26

(21)申请号 201811622803.5

(22)申请日 2018.12.28

(71)申请人 山东大学
地址 250061 山东省济南市历下区经十路17923号

(72)发明人 张伟 饶振环 吴悦晨 宋柯 鲁威志

(74)专利代理机构 济南圣达知识产权代理有限公司 37221
代理人 黄海丽

(51)Int.Cl.
G01C 21/36(2006.01)
G01C 21/34(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

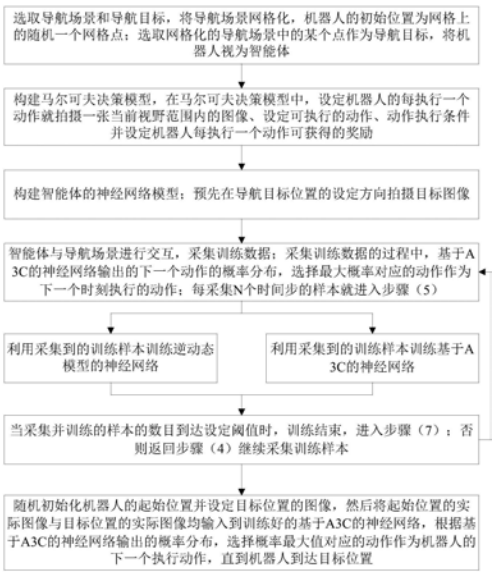
权利要求书3页 说明书9页 附图4页

(54)发明名称

基于深度强化学习的视觉导航方法及系统

(57)摘要

本公开公开了基于深度强化学习的视觉导航方法及系统，包括：随机初始化机器人的起始位置并设定目标位置的图像，然后将起始位置的实际图像与目标位置的实际图像均输入到训练好的基于A3C算法的神经网络，根据基于A3C算法的神经网络输出的概率分布，选择概率最大值对应的动作作为机器人的下一个执行动作，直到机器人到达目标位置。



CN 109682392 A

1.基于深度强化学习的视觉导航方法,其特征是,包括:

随机初始化机器人的起始位置并设定目标位置的图像,然后将起始位置的实际图像与目标位置的实际图像均输入到训练好的基于A3C算法的神经网络,根据基于A3C算法的神经网络输出的概率分布,选择概率最大值对应的动作作为机器人的下一个执行动作,直到机器人到达目标位置。

2.如权利要求1所述的方法,其特征是,基于A3C算法的神经网络的训练过程为:

步骤(1):选取导航场景和导航目标,将导航场景网格化,机器人的初始位置为网格上的随机一个网格点;选取网格化的导航场景中的某个点作为导航目标,将机器人视为智能体;

步骤(2):设定视觉导航任务为寻找机器人由初始位置到导航目标位置的导航路径;

预先在导航目标位置的设定方向拍摄目标图像;

构建视觉导航任务的马尔可夫决策过程模型,在马尔可夫决策过程模型中,设定机器人的每执行一个动作就拍摄一张当前视野范围内的图像、设定可执行的动作、动作所对应的执行条件并设定机器人每执行一个动作获得的奖励;

步骤(3):构建智能体的神经网络模型;所述智能体的神经网络模型,包括:相互交叉的基于A3C算法的神经网络和基于逆动态模型的神经网络;

步骤(4):智能体从导航场景中采集训练数据;采集训练数据的过程中,基于A3C算法的神经网络输出的下一个动作的概率分布,选择最大概率对应的动作作为智能体下一个时刻执行的动作;每采集N个时间步的样本就进入步骤(5);

步骤(5):利用步骤(4)采集到的训练样本训练智能体的神经网络;包括步骤(51)和步骤(52);所述步骤(51)和步骤(52)是同时进行,且同时结束并进入步骤(6)的;

步骤(51):利用采集到的训练样本训练基于逆动态模型的神经网络,进入步骤(6);

步骤(52):利用采集到的训练样本训练基于A3C算法的神经网络,进入步骤(6);

步骤(6):当采集并训练的样本的数目均到达设定阈值时,训练结束,得到训练好的基于A3C算法的神经网络;否则,返回步骤(4)继续采集训练样本。

3.如权利要求2所述的方法,其特征是,

所述步骤(2)中构建视觉导航任务的马尔可夫决策过程模型:M(状态,动作,奖励);其中,

状态是指机器人视野范围内的图像,机器人当前视野范围内的图像被称之为当前状态;在当前状态下,机器人执行一个动作后的视野范围内的图像,称之为下一时刻的状态;目标图像是指机器人在导航目标位置所拍摄的图像,目标图像被称之为目标状态;

动作是指机器人在每个时间间隔内选取的动作,所述动作,包括:前进一步、左转90度或右转90度;前进一步的步长为单个网格的长度;机器人在当前状态下采取的动作为当前动作,在上一时刻采取的动作为上一时刻的动作;

奖励是指机器人采取某个动作后,若到达导航目标位置且拍摄的视野范围内的图像与目标图像一致,则获得的奖励值为1;若未到达目标状态,则获得的奖励值为0;

时间步:在当前状态下,机器人采取动作后,获得下一时刻的状态,将这个过程所用时间长度称之为一个时间步长,简称时间步。

4.如权利要求2所述的方法,其特征是,

所述智能体的神经网络模型的结构包括:两条并发的通道,通道之间互有交叉;

其中,第一个通道包括:依次连接的第一卷积层、第二卷积层、第一全连接层、第二全连接层、第三全连接层和第一输出层;

第二个通道包括:依次连接的第三卷积层、第四卷积层、第四全连接层、第一长短期记忆单元层和第二输出层;

所述第一全连接层和第四全连接层的输出端均与第二全连接层的输入端连接;

所述第二全连接层的输出端与第二输出层的输入端连接;

基于A3C算法的神经网络由两个通道中除第一个通道中的第三全连接层和输出层外的其他网络组成;逆动态模型的神经网络由两个通道中除第二个通道中的第一长短期记忆单元层和输出层外的其他网络组成。

5.如权利要求2所述的方法,其特征是,步骤(4)的具体步骤为:

在当前的导航场景下,智能体采集当前图像X_t和目标图像X_g,智能体将目标图像X_g输入基于A3C算法的神经网络模型的第一卷积层,智能体将当前图像X_t输入基于A3C算法的神经网络模型的第三卷积层,基于A3C算法的神经网络模型输出设定的可执行动作的概率分布,获取最大概率对应的动作a_t,智能体执行动作a_t后,采集到新图像X_{t+1},获得奖励r,进而完成一次数据采集过程;如果奖励r=1,即智能体到达导航目标位置;如果奖励r=0,即智能体未到达导航目标位置,智能体根据概率分布选择的动作,完成动作的执行,继续拍摄新的图像。

6.如权利要求5所述的方法,其特征是,

将数据采集过程每执行N次,就暂停数据采集,开始利用采集的N次数据对网络进行训练;同时在数据采集的过程中,保存每一次的状态、每一次的执行动作和每一次执行动作的奖励r,每一次的状态、每一次的执行动作和每一次执行动作的奖励r被称之为训练样本;每一次的状态,包括:智能体上一时刻的图像X_{t-1}、当前图像X_t以及目标图像X_g;每一次动作包括:上一时刻的动作a_{t-1}和当前动作a_t。

7.如权利要求2所述的方法,其特征是,

在步骤(51)所述训练逆动态模型的神经网络的过程中,

第一卷积层用于输入上一时刻的状态图像X_{t-1},提取视觉特征并输出;

第二卷积层用于输入第一卷积层输出的视觉特征,进一步提取视觉特征并输出;

第一全连接层用于输入第二卷积层输出的视觉特征,将其映射到256维的特征并输出;

第三卷积层用于输入当前状态图像X_t,提取视觉特征并输出;

第四卷积层用于输入第三卷积层输出的视觉特征,进一步处理提取视觉特征并输出;

第四全连接层用于输入第四卷积层输出的视觉特征,将其映射到256维的特征并输出给第二全连接层;

第二全连接层用于输入第一全连接层和第四全连接层输出的视觉特征,将两个256维的视觉特征映射到256维的特征并输出;

第三全连接层用于输入第二全连接层输出的256维特征,将其映射到64维特征并输出;

第一输出层用于输入第三全连接层输出的64维特征,将其映射为对上一时刻动作的预测值的概率分布。

8.如权利要求2所述的方法,其特征是,

在步骤(52)所述训练基于A3C算法的神经网络的过程中，

第一卷积层用于输入目标状态图像X_g，提取视觉特征并输出；

第二卷积层用于输入第一卷积层输出的视觉特征，进一步处理提取视觉特征并输出；

第一全连接层用于输入第二卷积层输出的视觉特征，将其映射到256维的特征并输出；

第二全连接层用于第一全连接层和第四全连接层输出的视觉特征，将两个256维的视觉特征映射到一个256维的特征并输出；

第三卷积层用于输入当前状态图像X_t，提取视觉特征并输出；

第四卷积层用于输入第四卷积层输出的视觉特征，进一步处理提取视觉特征并输出；

第四全连接层用于输入第四卷积层输出的视觉特征，将其映射到256维特征并输出；

第一长短期记忆单元层用于输入第四全连接层输出的256维特征，提取其在时间序列上的信息，映射到256维特征并输出；

第二输出层用于输入第二全连接层和第一长短期记忆单元层输出的特征，将其映射到一个状态值函数V及对当前动作的预测值的概率分布。

9.如权利要求1所述的方法，其特征是，

随机初始化机器人的起始位置并设定目标位置的图像，然后将起始位置的实际图像与目标位置的实际图像均输入到训练好的基于A3C算法的神经网络，根据基于A3C算法的神经网络输出的概率分布，选择概率最大值对应的动作作为机器人的下一个执行动作，直到机器人到达目标位置，具体包括：

步骤(71)：随机初始化机器人的起始位置并设定目标位置的图像X_g，设定当前时刻t，进入步骤(72)；

步骤(72)：机器人获取当前视野范围内的图像X_t；若机器人处于目标位置，即图像X_t与X_g为同一幅图像，则完成机器人从起始位置到目标位置的导航；

若机器人不处于目标位置，即图像X_t与图像X_g不是同一幅图像，则将当前视野范围内的图像和设定的目标位置的图像均输入到训练好的基于A3C算法的神经网络，获得可执行动作的概率分布；进入步骤(73)；

步骤(73)：机器人对可执行动作的概率分布按概率进行采样，获取动作并执行，进入下一时刻t+1，令t＝t+1，返回步骤(72)。

10.基于深度强化学习的视觉导航系统，其特征是，包括：

训练模块，用于训练好的基于A3C算法的神经网络；

导航模块，用于随机初始化机器人的起始位置并设定目标位置的图像，然后将起始位置的实际图像与目标位置的实际图像均输入到训练好的基于A3C算法的神经网络，根据基于A3C算法的神经网络输出的概率分布，选择概率最大值对应的动作作为机器人的下一个执行动作，直到机器人到达目标位置。

基于深度强化学习的视觉导航方法及系统

技术领域

[0001] 本公开涉及基于深度强化学习的视觉导航方法及系统。

背景技术

[0002] 本部分的陈述仅仅是提到了与本公开相关的背景技术，并不必然构成现有技术。

[0003] 视觉导航是一项用于确定当前位置，然后根据图像或者视频输入规划朝向某些目标位置的路径的智能。由于相机视角的局限性，每次只能观察到环境的一部分，这使得仅依靠视觉输入来导航非常困难。近年来深度强化学习在诸如Atari游戏，电脑围棋和目标定位等领域取得了巨大成功，受此鼓舞，学界开始使用深度强化学习来训练智能体学会导航到特定目标。

[0004] 目标驱动的导航任务要求智能体经过训练后，能从任意的位置和朝向导航到一系列不同的目标，因此，对于不同的目标，不必重新训练模型。目前存在一些基于深度强化学习的目标驱动导航问题的开创性工作。

[0005] 据发明人了解，在实施本发明的过程中，需要解决的技术问题如下：

[0006] 首先，由于智能体必须学会从随机状态导航到不同的目标，智能体需要学习当前状态和目标、动作之间的关联。其次，智能体与环境进行交互，然后生成与每一个目标有关的样本。然而，为一个目标采集的样本只能用来训练智能体导航到这个目标，这是一种样本效率低下的方式。

发明内容

[0007] 为了解决现有技术的不足，本公开提供了基于深度强化学习的视觉导航方法及系统；

[0008] 第一方面，本公开提供了基于深度强化学习的视觉导航方法；

[0009] 基于深度强化学习的视觉导航方法，包括：

[0010] 随机初始化机器人的起始位置并设定目标位置的图像，然后将起始位置的实际图像与目标位置的实际图像均输入到训练好的基于A3C算法的神经网络，根据基于A3C算法的神经网络输出的概率分布，选择概率最大值对应的动作作为机器人的下一个执行动作，直到机器人到达目标位置。

[0011] 作为一种可能的实现方式，基于A3C算法的神经网络的训练过程为：

[0012] 步骤(1)：选取导航场景和导航目标，将导航场景网格化，机器人的初始位置为网格上的随机一个网格点；选取网格化的导航场景中的某个点作为导航目标，将机器人视为智能体；

[0013] 步骤(2)：设定视觉导航任务为寻找机器人由初始位置到导航目标位置的导航路径；

[0014] 预先在导航目标位置的设定方向拍摄目标图像；

[0015] 构建视觉导航任务的马尔可夫决策过程模型，在马尔可夫决策过程模型中，设定

机器人的每执行一个动作就拍摄一张当前视野范围内的图像、设定可执行的动作、动作所对应的执行条件并设定机器人每执行一个动作获得的奖励；

[0016] 步骤(3)：构建智能体的神经网络模型；所述智能体的神经网络模型，包括：相互交叉的基于A3C算法的神经网络和基于逆动态模型的神经网络；

[0017] 步骤(4)：智能体从导航场景中采集训练数据；采集训练数据的过程中，基于A3C算法的神经网络输出的下一个动作的概率分布，选择最大概率对应的动作作为智能体下一个时刻执行的动作；每采集N个时间步的样本就进入步骤(5)；

[0018] 步骤(5)：利用步骤(4)采集到的训练样本训练智能体的神经网络；包括步骤(51)和步骤(52)；所述步骤(51)和步骤(52)是同时进行，且同时结束并进入步骤(6)的；

[0019] 步骤(51)：利用采集到的训练样本训练基于逆动态模型的神经网络，进入步骤(6)；

[0020] 步骤(52)：利用采集到的训练样本训练基于A3C算法的神经网络，进入步骤(6)；

[0021] 步骤(6)：当采集并训练的样本的数目均到达设定阈值时，训练结束，得到训练好的基于A3C算法的神经网络；否则，返回步骤(4)继续采集训练样本。

[0022] 作为一种可能的实现方式，所述选取导航场景是指：高仿真框架AI2-THOR；

[0023] 作为一种可能的实现方式，所述将导航场景进行网格化处理，是指：将场景中的位置网格化，网格化的场景包括若干个网格点，相邻网格点之间间距相同；使得机器人只能到达场景中的若干个网格点，降低场景的导航复杂度。

[0024] 作为一种可能的实现方式，所述步骤(2)中构建视觉导航任务的马尔可夫决策过程模型：M(状态，动作，奖励)；其中，

[0025] 状态是指机器人视野范围内的图像，机器人当前视野范围内的图像被称之为当前状态；在当前状态下，机器人执行一个动作后的视野范围内的图像，称之为下一时刻的状态；目标图像是指机器人在导航目标位置所拍摄的图像，目标图像被称之为目标状态；

[0026] 动作是指机器人在每个时间间隔内选取的动作，所述动作，包括：前进一步、左转90度或右转90度；前进一步的步长为单个网格的长度；机器人在当前状态下采取的动作为当前动作，在上一时刻采取的动作为上一时刻的动作；

[0027] 奖励是指机器人采取某个动作后，若到达导航目标位置且拍摄的视野范围内的图像与目标图像一致，则获得的奖励值为1；若未到达目标状态，则获得的奖励值为0；

[0028] 时间步：在当前状态下，机器人采取动作后，获得下一时刻的状态，将这个过程所用时间长度称之为一个时间步长，简称时间步。

[0029] 作为一种可能的实现方式，所述智能体的神经网络模型的结构包括：两条并发的通道，通道之间互有交叉；

[0030] 其中，第一个通道包括：依次连接的第一卷积层、第二卷积层、第一全连接层、第二全连接层、第三全连接层和第一输出层；

[0031] 第二个通道包括：依次连接的第三卷积层、第四卷积层、第四全连接层、第一长短期记忆单元层和第二输出层；

[0032] 所述第一全连接层和第四全连接层的输出端均与第二全连接层的输入端连接；

[0033] 所述第二全连接层的输出端与第二输出层的输入端连接；

[0034] 基于A3C算法的神经网络由两个通道中除第一个通道中的第三全连接层和输出层

外的其他网络组成;逆动态模型的神经网络由两个通道中除第二个通道中的第一长短期记忆单元层和输出层外的其他网络组成。

[0035] 作为一种可能的实现方式,

[0036] 第一卷积层,卷积核大小为8x8,步长为4x4,输出16个特征图;激活函数为线性整流单元ReLU;

[0037] 第二卷积层,卷积核大小为4x4,步长为2x2,输出32个特征图;激活函数为线性整流单元ReLU;

[0038] 第一全连接层,包括256个隐藏单元,激活函数为线性整流单元ReLU;

[0039] 第二全连接层,包括256个隐藏单元,激活函数为线性整流单元ReLU;

[0040] 第三全连接层,包括64个隐藏单元,激活函数为线性整流单元ReLU;

[0041] 第三卷积层,与第一卷积层共享参数;

[0042] 第四卷积层,与第二卷积层共享参数;

[0043] 第四全连接层,与第一全连接层共享参数;

[0044] 第一长短期记忆单元层,包括256个隐藏单元;第一长短期记忆单元层为长短期记忆网络。

[0045] 作为一种可能的实现方式,对智能体的神经网络模型进行训练,训练过程包含两个部分:一是训练数据的采集,即步骤(4);二是利用采集到的训练数据训练神经网络的参数,即步骤(5);步骤(4)和步骤(5)两个过程是交替进行的。

[0046] 作为一种可能的实现方式,步骤(4)的具体步骤为:

[0047] 在当前的导航场景下,智能体采集当前图像X_t和目标图像X_g,智能体将目标图像X_g输入基于A3C算法的神经网络模型的第一卷积层,智能体将当前图像X_t输入基于A3C算法的神经网络模型的第三卷积层,基于A3C算法的神经网络模型输出设定的可执行动作的概率分布,获取最大概率对应的动作a_t,智能体执行动作a_t后,采集到新图像X_{t+1},获得奖励r,进而完成一次数据采集过程。

[0048] 如果奖励r=1,即智能体到达导航目标位置;

[0049] 如果奖励r=0,即智能体未到达导航目标位置,智能体根据概率分布选择的动作,完成动作的执行,继续拍摄新的图像。

[0050] 将数据采集过程每执行N次,就暂停数据采集,开始利用采集的N次数据对网络进行训练;同时在数据采集的过程中,保存每一次的状态、每一次的执行动作和每一次执行动作的奖励r,每一次的状态、每一次的执行动作和每一次执行动作的奖励r被称之为训练样本;每一次的状态,包括:智能体上一时刻的图像X_{t-1}、当前图像X_t以及目标图像X_g;每一次动作包括:上一时刻的动作a_{t-1}和当前动作a_t。

[0051] 在步骤(51)所述训练逆动态模型的神经网络的过程中,

[0052] 第一卷积层用于输入上一时刻的状态图像X_{t-1},提取视觉特征并输出;

[0053] 第二卷积层用于输入第一卷积层输出的视觉特征,进一步提取视觉特征并输出;

[0054] 第一全连接层用于输入第二卷积层输出的视觉特征,将其映射到256维的特征并输出;

[0055] 第三卷积层用于输入当前状态图像X_t,提取视觉特征并输出;

[0056] 第四卷积层用于输入第三卷积层输出的视觉特征,进一步处理提取视觉特征并输

出；

[0057] 第四全连接层用于输入第四卷积层输出的视觉特征，将其映射到256维的特征并输出给第二全连接层；

[0058] 第二全连接层用于输入第一全连接层和第四全连接层输出的视觉特征，将两个256维的视觉特征映射到256维的特征并输出；

[0059] 第三全连接层用于输入第二全连接层输出的256维特征，将其映射到64维特征并输出；

[0060] 第一输出层用于输入第三全连接层输出的64维特征，将其映射为对上一时刻动作的预测值的概率分布。

[0061] 在步骤(52)所述训练基于A3C算法的神经网络的过程中，

[0062] 第一卷积层用于输入目标状态图像X_g，提取视觉特征并输出；

[0063] 第二卷积层用于输入第一卷积层输出的视觉特征，进一步处理提取视觉特征并输出；

[0064] 第一全连接层用于输入第二卷积层输出的视觉特征，将其映射到256维的特征并输出；

[0065] 第二全连接层用于第一全连接层和第四全连接层输出的视觉特征，将两个256维的视觉特征映射到一个256维的特征并输出；

[0066] 第三卷积层用于输入当前状态图像X_t，提取视觉特征并输出；

[0067] 第四卷积层用于输入第四卷积层输出的视觉特征，进一步处理提取视觉特征并输出；

[0068] 第四全连接层用于输入第四卷积层输出的视觉特征，将其映射到256维特征并输出；

[0069] 第一长短期记忆单元层用于输入第四全连接层输出的256维特征，提取其在时间序列上的信息，映射到256维特征并输出；

[0070] 第二输出层用于输入第二全连接层和第一长短期记忆单元层输出的特征，将其映射到一个状态值函数V及对当前动作的预测值的概率分布。

[0071] 作为一种可能的实现方式，步骤(51)：利用采集到的训练样本来训练逆动态模型的神经网络；训练时逆动态模型的神经网络的输入值是样本中的上一时刻的状态X_{t-1}和当前状态X_t，采用监督学习的方式进行训练，损失函数设置为交叉熵分类损失函数，标签为上一时刻的动作a_{t-1}。

[0072] 作为一种可能的实现方式，步骤(52)：利用采集到的训练样本来训练基于A3C算法的神经网络；训练时网络的输入值是样本中的当前图像X_t和目标图像X_g，采用强化学习的方式进行训练，用到样本中的当前动作a_t和奖励r。

[0073] 作为一种可能的实现方式，随机初始化机器人的起始位置并设定目标位置的图像，然后将起始位置的实际图像与目标位置的实际图像均输入到训练好的基于A3C算法的神经网络，根据基于A3C算法的神经网络输出的概率分布，选择概率最大值对应的动作作为机器人的下一个执行动作，直到机器人到达目标位置，具体包括：

[0074] 步骤(71)：随机初始化机器人的起始位置并设定目标位置的图像X_g，设定当前时刻t，进入步骤(72)；

[0075] 步骤(72):机器人获取当前视野范围内的图像X_t;若机器人处于目标位置,即图像X_t与X_g为同一幅图像,则完成机器人从起始位置到目标位置的导航;

[0076] 若机器人不处于目标位置,即图像X_t与图像X_g不是同一幅图像,则将当前视野范围内的图像和设定的目标位置的图像均输入到训练好的基于A3C算法的神经网络,获得可执行动作的概率分布;进入步骤(73);

[0077] 步骤(73):机器人对可执行动作的概率分布按概率进行采样,获取动作并执行,进入下一时刻t+1,令t=t+1,返回步骤(72)。

[0078] 第二方面,本公开还提供了基于深度强化学习的视觉导航系统;

[0079] 基于深度强化学习的视觉导航系统,包括:

[0080] 训练模块,用于训练好的基于A3C算法的神经网络;

[0081] 导航模块,用于随机初始化机器人的起始位置并设定目标位置的图像,然后将起始位置的实际图像与目标位置的实际图像均输入到训练好的基于A3C算法的神经网络,根据基于A3C算法的神经网络输出的概率分布,选择概率最大值对应的动作作为机器人的下一个执行动作,直到机器人到达目标位置。

[0082] 与现有技术相比,本公开的有益效果是:

[0083] 由于智能体的神经网络模型既包括基于A3C算法的神经网络,又包括逆动态模型的神经网络,二者这两个神经网网络有部分交叉,所以在训练的过程中,可以实现训练速度的提升,在使用该模型的时候,由于该模型在训练的过程中考虑了上一时刻跟当前时刻之间的状态关系,所以该模型能够对目标图像给出精准的导航路线,即使一个目标结束后,再给出第二个目标也不需要重新对模型进行训练。

附图说明

[0084] 构成本申请的一部分的说明书附图用来提供对本申请的进一步理解,本申请的示意性实施例及其说明用于解释本申请,并不构成对本申请的不当限定。

[0085] 图1为本发明的流程图;

[0086] 图2(a)图2(b)为目标驱动的视觉导航任务示例;

[0087] 图3为模型网络架构及训练流程;

[0088] 图4(a)-图4(d)AI2-THOR平台中的一些典型场景示例(包括浴室,卧室,客厅,厨房)。

具体实施方式

[0089] 应该指出,以下详细说明都是示例性的,旨在对本申请提供进一步的说明。除非另有指明,本文使用的所有技术和科学术语具有与本申请所属技术领域的普通技术人员通常理解的相同含义。

[0090] 需要注意的是,这里所使用的术语仅是为了描述具体实施方式,而非意图限制根据本申请的示例性实施方式。如在这里所使用的,除非上下文另外明确指出,否则单数形式也意图包括复数形式,此外,还应当理解的是,当在本说明书中使用术语“包含”和/或“包括”时,其指明存在特征、步骤、操作、器件、组件和/或它们的组合。

[0091] 英文缩写介绍:基于演员评论家算法(Asynchronous advantage actor-critic,

简称A3C)

[0092] 视觉导航是计算机视觉和机器人应用中的一个基本问题。本发明提出了一个新的模型嵌入式actor-critic方案,使得智能体学会仅依赖视觉状态输入,就能从任意位置导航到多个不同目标。本发明提出的方案的关键设计是逆动态模型(inverse dynamics model,简称InvDM)。逆动态模型的作用是捕捉当前状态与目标状态之间在导航上的联系,同时提供密集的训练信号,以此来缓解奖励稀疏的问题。在The House OfinteRactions(AI2-THOR)平台上进行验证,结果表明本发明提出的算法比传统的强化学习方法收敛得更快,同时能达到更好的导航表现。

[0093] 与常规的导航任务相比,目标驱动的导航任务需要智能体学会一系列不同的目标,这使其更具有挑战性。如图2(a)和图2(b)所示,目标驱动的导航任务要求智能体经过训练后,能从任意的位置和朝向导航到一系列不同的目标,因此,对于不同的目标,不必重新训练模型。

[0094] 本发明提出了一个新的模型嵌入式actor-critic方案,使得智能体仅依赖视觉状态输入就能同时学会导航到多个目标。首先,如图2(a)和图2(b)所示,在本发明的actor-critic框架中,本发明引入了一个逆动态模型(inverse dynamics model,InvDM)。逆动态模型以一个辅助任务的形式进行训练。这个辅助任务基于智能体当前状态和上一刻状态来预测其上一刻的动作。

[0095] 逆动态模型有三个优点:

[0096] 1)动作可以被视为区分状态序列的合适标准。经过训练,逆动态模型使得智能体更好的预测当前状态和目标之间的差异,例如,当前状态与目标之间在导航上的关联。

[0097] 2)由于预测上一刻的动作的辅助任务是通过自监督学习的方式来训练的,这可以用来引导智能体更有效地探索。因此尽管没有增加显式的奖励,仍能促进智能体的训练。换句话说,这个辅助任务能够提供密集的训练信号来解决奖励稀疏这一强化学习方法经常遇到的问题。

[0098] 3)由于不同的目标只具有不同的奖励函数,而拥有相同的马尔可夫决策过程(MDP)的转移结构,当要训练的导航目标们处在同一个场景中时,可以共同训练逆动态模型。因此,不同的导航目标在训练时可能会相互促进。换句话说,训练智能体导航到一个目标可能会帮助其训练导航到其他目标。

[0099] 如图1所示,基于深度强化学习的视觉导航方法,包括:

[0100] 步骤(1):选取导航场景和导航目标,将导航场景网格化,机器人的初始位置为网格上的随机一个网格点;选取网格化的导航场景中的某个点作为导航目标,将机器人视为智能体;

[0101] 步骤(2):设定视觉导航任务为寻找机器人由初始位置到导航目标位置的导航路径;

[0102] 预先在导航目标位置的设定方向拍摄目标图像;

[0103] 构建视觉导航任务的马尔可夫决策过程模型,在马尔可夫决策过程模型中,设定机器人的每执行一个动作就拍摄一张当前视野范围内的图像、设定可执行的动作、动作所对应的执行条件并设定机器人每执行一个动作获得的奖励;

[0104] 步骤(3):构建智能体的神经网络模型;所述智能体的神经网络模型,包括:相互交

叉的基于A3C算法的神经网络和基于逆动态模型的神经网络；

[0105] 步骤(4)：智能体从导航场景中采集训练数据；采集训练数据的过程中，基于A3C算法的神经网络输出的下一个动作的概率分布，选择最大概率对应的动作作为智能体下一个时刻执行的动作；每采集N个时间步的样本就进入步骤(5)；

[0106] 步骤(5)：利用步骤(4)采集到的训练样本训练智能体的神经网络；包括步骤(51)和步骤(52)；所述步骤(51)和步骤(52)是同时进行，且同时结束并进入步骤(6)的；

[0107] 步骤(51)：利用采集到的训练样本训练基于逆动态模型的神经网络，进入步骤(6)；

[0108] 步骤(52)：利用采集到的训练样本训练基于A3C算法的神经网络，进入步骤(6)；

[0109] 步骤(6)：当采集并训练的样本的数目均到达设定阈值时，训练结束，得到训练好的基于A3C算法的神经网络；否则，返回步骤(4)继续采集训练样本；

[0110] 步骤(7)：随机初始化机器人的起始位置并设定目标位置的图像，然后将起始位置的实际图像与目标位置的实际图像均输入到训练好的基于A3C算法的神经网络，根据基于A3C算法的神经网络输出的概率分布，选择概率最大值对应的动作作为机器人的下一个执行动作，直到机器人到达目标位置。

[0111] 本发明在The House Of inteRactions(AI2-THOR)，一个接近真实场景的3D室内导航平台上验证本发明提出的方法。本发明使用异步优势演员-评论家算法(Asynchronous advantage actor-critic，简称A3C)作为本发明的方法的基础框架。实验结果表明，所提出的方法可以加速智能体在目标驱动的视觉导航任务上的学习速率，并且随着目标数量的增加，方法具有鲁棒性。不仅如此，本发明还使得智能体仅依靠二元奖励就能同时学习多个环境中的多个目标。

[0112] 本发明提出了一个自监督的逆动态模型(InvDM)来更好地预测当前状态和目标状态之间的差异，强化学习的目的是训练智能体与环境交互进而最大化未来累计奖励的期望值。这关系到马尔可夫决策过程(MDP)中的策略优化。在目标驱动的视觉导航任务中，马尔可夫决策过程可用公式元组表示为$M(s,g,a,r,\gamma)$，其中$s\in S$表示一个确定的状态空间，$g\in G$表示一系列可能的目标，$a\in A$表示动作空间，r表示状态奖励函数，$\gamma\in(0,1]$是一个折扣因子。奖励函数$r_g(s,a,s')$取决于当前的目标和状态。一个随机策略不$(a|s,g)$将每一个状态-目标对映射到一个动作，同时定义智能体的行为。

[0113] 在每一个离散的时刻t下，智能体观察到状态s_t，然后根据策略$\pi(a_t|s_t,g_t)$选择一个动作a_t。一个时间步之后，智能体获得一个数值奖励r_t，然后智能体便到达了一个新的状态s_{t+1}。这一过程一直持续直到智能体到达指定的目标。R_t表示从时间步t开始直到智能体到达目标的累计奖励。智能体的目的是学到一个最优策略π，这个策略能最大化上述的累计奖励R_t的期望。A3C算法能够使用n步的累计奖励R_t同时更新策略函数$\pi(a_t|s_t,g_t;\theta_\pi)$以及状态值函数$V(s_t,g_t;\theta_v)$。每次经过$t_{max}$步或者智能体到达指定目标时更新策略和状态值函数。从t时刻开始的累计奖励R_t定义如下：

[0114] $$R_t=\sum_{i=0}^{k-1}\gamma^i r_{t+i}+\gamma^k V(s_{k+t},g_{k+t};\theta_v) \qquad (1)$$

[0115] 在公式(1)中k值大小随状态的改变而改变，且不大于t_{max}。

[0116] 为了防止过早地收敛到局部最优，强化学习方法通常将策略π的熵H加入到待优化的目标函数中。最终的目标函数的梯度如下：

[0117] $$\nabla_{\theta_\pi}\log\pi(a_t|s_t,g_t;\theta_\pi)\big(R_t-V(s_t,g_t;\theta_\pi)\big)+\beta\nabla_{\theta_\pi}H\big(\pi(s_t,g_t;\theta_\pi)\big) \quad (2)$$

[0118] 公式(2)中β用于控制熵正则项的强度。因此，最终的梯度更新规则如下所示：

[0119] $$\theta_\pi\leftarrow\theta_\pi+\eta\nabla_{\theta_\pi}\log\pi(a_t|s_t,g_t;\theta_\pi)\big(R_t-V(s_t,g_t;\theta_v)\big)+\beta\nabla_{\theta}H\big(\pi(s_t,g_t;\theta_\pi)\big) \quad (3)$$

[0120] $$\theta_v\leftarrow\theta_v+\eta\nabla_{\theta_\pi}\big(R_t-V(s_t,g_t;\theta_v)\big)^2 \quad (4)$$

[0121] 公式(3)和公式(4)中η代表学习速率。

[0122] 如图3所示，面对目标驱动的视觉导航任务，本发明基于A3C算法设计了一个新的模型嵌入式深度神经网络模型。这个模型将目标当作状态输入的一部分，使智能体同时学习导航到一系列不同的目标。与此同时，模型的双通道工作方式使得智能体可能学会两种不同的特征表达:通用的特征和专属的特征。通用的特征表达仅依赖于当前的状态，能够为智能体例如场景理解等感知处理的功能。而专属的特征表达依赖当前的状态和目标，能够帮助智能体进行长期的路径规划。本发明提出的模型的输入由当前观察到的状态x_t和目标的状态x_g组成，模型输出一个在动作空间内的概率分布和一个值函数。值函数可以表示智能体从任何一个状态s到达一个给定目标g的效用。本发明通过端到端的强化学习结合辅助的辅助目标来训练本发明提出的模型。训练的目的是用actor-critic方法最大化累积奖励的同时最小化由预测的上一刻的动作和真实的上一刻的动作所定义的辅助损失函数。

[0123] 模型的细节如图3所示。首先，模型的特征提取部分由两层卷积网络和一层全连接网络组成。特征提取部分处理当前状态和目标状态的图片，通过共享网络参数的形式分别产生视觉特征f_s和f_g。第一层卷积网络的卷积核大小为8x8，步长为4x4，输出16个特征图。第二层卷积网络的卷积核大小为4x4，步长为2x2，输出32个特征图。之后的全连接层包含256个隐藏单元。上述三层网络的激活函数均为线性整流单元(ReLU)。其次，状态的视觉特征$f_s(X_t)$被级联到目标的视觉$f_g(X_g)$，经过一层包含256个隐藏单元以及ReLU激活函数的全连接层后输出隐藏激活单元$h_a(f_s,f_g)$。动作预测模块$g_a(h_a)$由一层包含64个隐藏单元的全连接层以及一个柔性最大(softmax)输出层组成，它用来预测上一步的动作a_t。最后，在网络的另一端，状态的视觉特征$f_s(X_t)$经过一层包含256个隐藏单元的长短期记忆单元(LSTM)输出隐藏激活单元$h_s(f_s)$。将隐藏激活单元h_a级联到h_s，然后连接一个柔性最大层(softmax)输出策略π，连接一个全连接层输出值函数V。

[0124] 对于视觉导航任务来说，如果能捕捉到当前状态与目标之间的联系，智能体就能很好的处理规划与实时动作选择之间的关系。为此，如图3所示，本发明引入了一个逆动态模型(InvDM)。在本发明的结构中，逆动态模型的以辅助任务的方式来训练。这个辅助任务的内容是根据当前状态与上一步的状态来预测上一步的动作。动作预测可以用来衡量连续状态之间的差异。因此，通过训练之后，逆动态模型可以预测当前状态与目标之间在导航上的差异与联系，进而为智能体的长期规划提供帮助。

[0125] 在具体实现上，辅助任务以自监督的方式进行训练并且可以产生额外的连续梯度。由于这样的辅助任务可以提供额外的密集训练信号，因此奖励稀疏这一强化学习领域常见的问题就能得到解决。此外，改变智能体的目标，在整个框架中只会带来奖励函数的改变，而不会造成马尔可夫决策过程中的转移模型的变化，因此，在不同的导航目标，智能体

均可以共同训练逆动态模型，这意味着在不同的导航目标下训练可以相互提升。

[0126] 逆动态模型的训练流程如图3所示。逆动态模型的输入包括智能体观察到的当前状态x_t和上一步的状态x_{t-1}，输出对上一步动作的预测在动作空间的概率分布。这个动作预测作为一个额外的优化项，由交叉熵分类损失函数来定义：

[0127] $$L_a = -\sum_i a_i \cdot \log \bar{a}_i \qquad (5)$$

[0128] 公式(5)中的i表示动作的索引，a和$\bar{a}$分别表示实际采取的动作和预测的动作。

[0129] AI2-THOR是Unity3D游戏引擎中的一个开源集，提供了在一组近乎真实的室内场景中进行导航模拟的功能。选择了四个不同的场景来进行导航性能验证：浴室，卧室，厨房和客厅。一个可供智能体进行导航和交互的卧室场景。实验中用到的环境的具体细节，如图4(a)-图4(d)所示。

[0130] • 动作空间：智能每一步有三种动作可供选择：前进，左转和右转。前进的步长固定(0.5米)，转向动作的转角固定(90度)。固定的步长和转角将环境离散化为一个网格状的空间表示。

[0131] • 状态和目标：状态和目标都是智能体所观察到的第一视角图像。实际输入到智能体的时被降采样到大小为84x84的RGB图片。使用图像作为目标描述的好处是可以灵活地指定新目标。给定一张目标图片，任务目标是导航到拍摄目标图像的位置和视角。

[0132] • 奖励设置：环境仅在任务完成时提供奖励回报(值为1)。

[0133] 本发明以每2000帧图像(状态)里智能体完成的轨迹的次数来评价智能体的表现。每给定一个目标，本发明随机初始化智能体的起始位置。

[0134] 模型的训练参数如下：折扣因子$\gamma=0.99$，RMSProp优化器的衰减因子和探索率分别为$\alpha=0.99$，$\in=0.1$，熵正则项的系数为$\beta=0.01$。训练中本发明使用了16个线程，在每个线程中智能体每采取五次动作后更新一次网络参数($t_{max}=5$)。为了防止智能体的性能偏置到某个目标下，在每个线程里轮流训练智能体到达各个环境里的各个目标。

[0135] 本发明使用A3C作为基线算法来评估逆动态模型(InvDM)的效果。本发明在上面提到的浴室，卧室，厨房和客厅等四个场景里比较A3C和A3C+InvDM在目标个数变化时的性能，如一个目标，两个目标和四个目标。从图3中本发明可以看出四个场景的具体情况，浴室的尺寸最小而厨房的尺寸最大。

[0136] 本发明提出了一个模型嵌入式的actor-critic方案，使得智能体能够学会同时导航到多个目标。在本发明的架构中中包含了一个特别设计的逆动态模型(InvDM)逆动态模型以一个辅助任务的形式进行训练，帮助智能体捕捉当前状态与目标之间在导航上的联系，并且提供额外的密集训练信号来解决奖励稀疏的问题。在AI2-THOR平台上的实验结果说明本发明提出的模型不仅使智能体能够同时学习导航到多个不同的目标，还使智能体的样本效率得到显著提升。

[0137] 以上所述仅为本申请的优选实施例而已，并不用于限制本申请，对于本领域的技术人员来说，本申请可以有各种更改和变化。凡在本申请的精神和原则之内，所作的任何修改、等同替换、改进等，均应包含在本申请的保护范围之内。

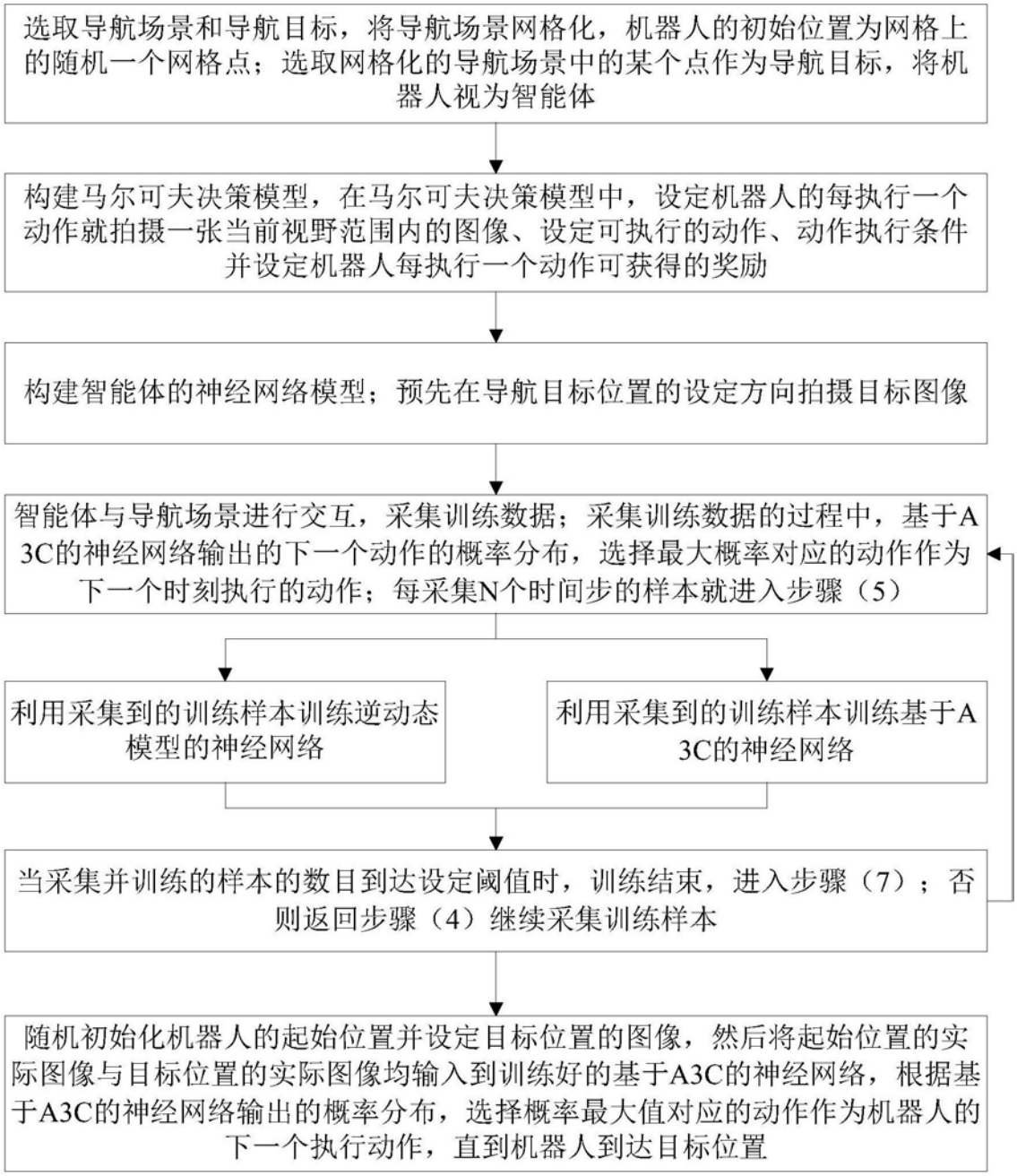

选取导航场景和导航目标，将导航场景网格化，机器人的初始位置为网格上的随机一个网格点；选取网格化的导航场景中的某个点作为导航目标，将机器人视为智能体
构建马尔可夫决策模型，在马尔可夫决策模型中，设定机器人的每执行一个动作就拍摄一张当前视野范围内的图像、设定可执行的动作、动作执行条件并设定机器人每执行一个动作可获得的奖励
构建智能体的神经网络模型；预先在导航目标位置的设定方向拍摄目标图像
智能体与导航场景进行交互，采集训练数据；采集训练数据的过程中，基于A3C的神经网络输出的下一个动作的概率分布，选择最大概率对应的动作作为下一个时刻执行的动作；每采集N个时间步的样本就进入步骤（5）
利用采集到的训练样本训练逆动态模型的神经网络
利用采集到的训练样本训练基于A3C的神经网络
当采集并训练的样本的数目到达设定阈值时，训练结束，进入步骤（7）；否则返回步骤（4）继续采集训练样本
随机初始化机器人的起始位置并设定目标位置的图像，然后将起始位置的实际图像与目标位置的实际图像均输入到训练好的基于A3C的神经网络，根据基于A3C的神经网络输出的概率分布，选择概率最大值对应的动作作为机器人的下一个执行动作，直到机器人到达目标位置

图1


Bedroom
Goal C
Goal D
Goal B
Goal A

图2(a)

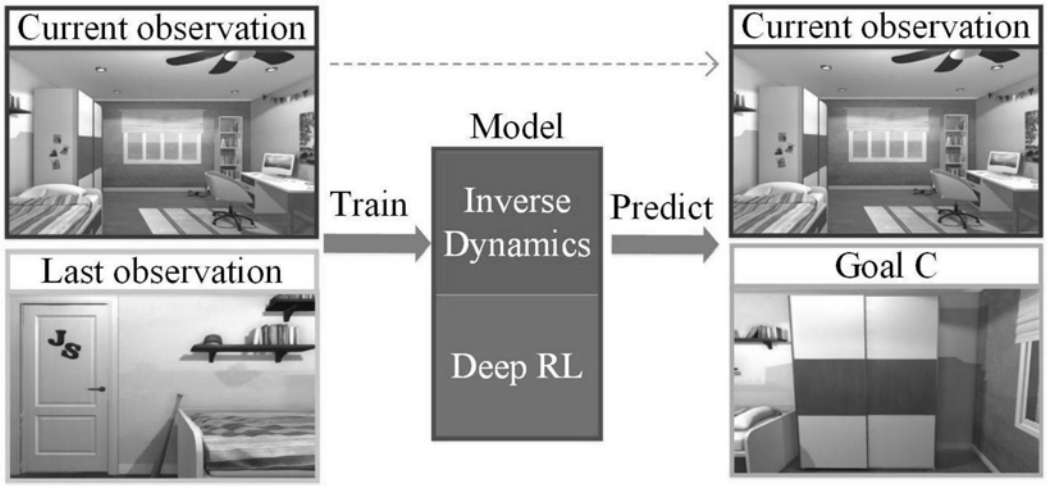

Current observation
Last observation
Train
Model
Inverse Dynamics
Deep RL
Predict
Current observation
Goal C

图2(b)

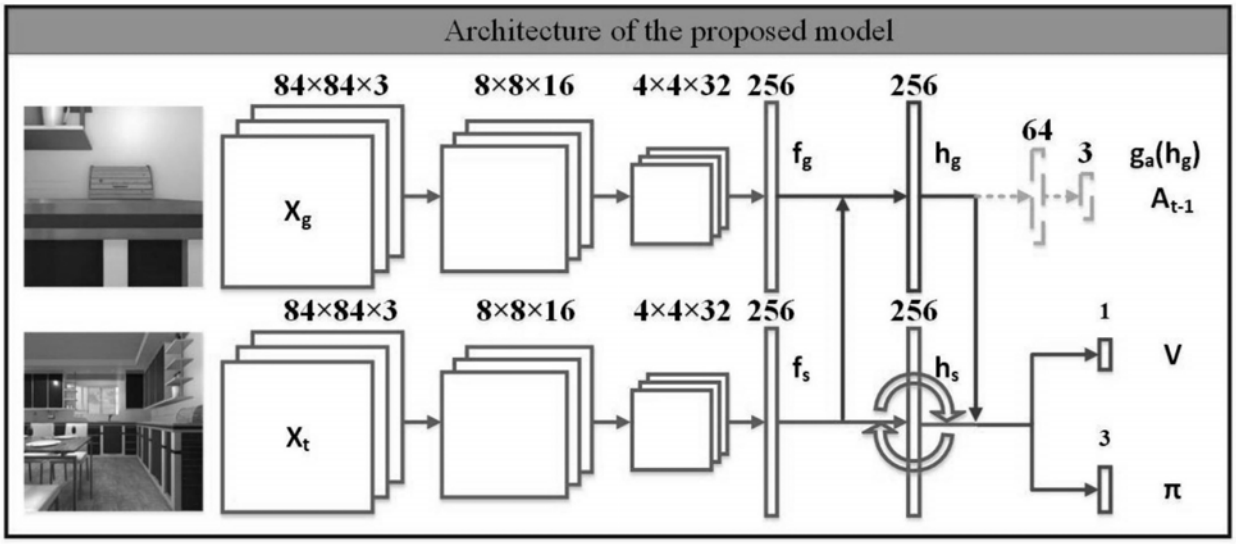


图3

图4(a)

图4(b)

图4(c)

图4(d)