论文推荐

Who Let The Dogs Out?

Modeling Dog Behavior From Visual Data

21821155 张玚

Abstract

作者研究了如何直接建模一个视觉智能体。计算机视觉通常专注于解决各种与视觉智能相关的子任务。作者偏离了处理计算机视觉任务的标准方法，直接对视觉智能体进行建模。作者的模型将视觉信息作为输入并直接预测视觉智能体的动作。为了达成这一目标，作者引入了 DECADE，一个包含以狗为第一人称视角的视频以及相应动作的数据集。利用这样的数据集，作者可以建模狗的行为方式和动作规划方式。在多种度量方式下，对于给定视觉输入，作者能成功地在各种环境下建模智能体。此外，相比用图像分类训练出的表征学习，作者的模型学得的表征能编码不同的信息，还可以泛化至其他的领域。特别是，通过将这种对狗的建模用于表征学习，作者在可行走表面预测和场景分类任务中得到了非常好的结果。

1、Introduction

作者重视代表性任务（包括图像分类、目标识别、目标检测、图像分割等等）在计算机视觉研究中的影响，主张继续对这些基本问题进行研究。然而，这些代表性任务的理想结果与视觉智能系统的期望功能之间存在差距。在这篇论文中，受近期关于行为与互动在视觉理解中作用的研究的启发，作者将视觉智能问题定义为“理解视觉数据，使得智能体能够在视觉世界中执行动作并解决问题”。根据这个定义，作者建议学习在视觉世界中扮演一个视觉智能体来处理问题。

一般来说，学习像视觉智能体一样行事是一个极具挑战性且难以定义的问题。动作对应的动作范围广，语义复杂。作者别具一格，只考虑最基本的、无语义的动作、简单的动作。

作者将对狗建模，作为视觉智能体。狗相对人来说，有着更简单的动作空间，使研究变得相对简单。同时，它们能很好地展示视觉智能的特性，例如它们可以分辨食物、障碍、别的动物以及人类，并作出相应的反应输出。然而，它们的目的和动机通常是事先不知道的。因此作者可以说是在建模一个黑箱。关于这个黑

箱系统，作者只知道它的输入和输出。

本论文研究如何基于视觉输入学习模仿狗的行为和动作规划方式。作者编写了一个以狗为第一人称视角的动作数据集 DECADE，包括以狗为第一人称视角的视频及其对应的运动。为了记录相关的运动，作者在狗的身体和关节处安装了惯性测量单元 IMU。作者记录了这些装置的绝对位置，然后计算狗的四肢与身体之间的相对角度。

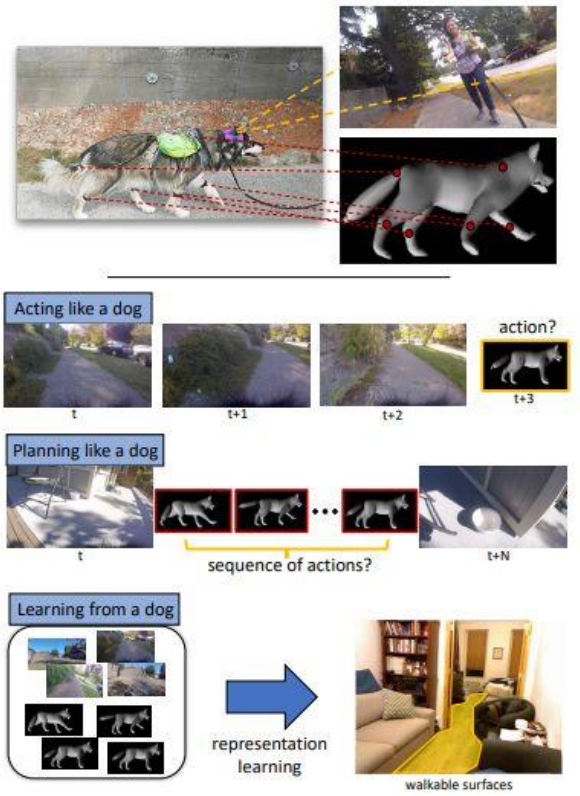

图 1. 作者解决了三个问题：(1) 模仿狗的行为：根据给出的一系列狗之前的相关行为照片，预测狗接下来的行为动作。(2) 模仿狗的动作规划方式：目的是找出一组动作使狗能从一个给定位置移动到另一给定位置。(3) 利用关于狗的数据来学习：利用学得的知识解决这一问题（例如：预测一个可供行走的地面区域）。

使用 DECADE 数据集，如图 1，作者探索了上面提到的三个主要问题。

在学习模仿狗的行为时，作者通过观察狗到目前为止的观察结果来预测狗在未来可能的动作（关节屈伸）。在模仿狗的动作规划方式时，作者解决了预测狗的系列运动动作的问题，这些动作将狗的状态从一个特定状态转变为目标状态。在利用狗作监督时，作者发现将狗的动作用于表征学习的潜力。

结果是令人欣喜的。作者的模型可以预测狗在各种场景下的运动（模仿狗的行为），也可以预测狗如何决定从一个状态转化为另一状态（模仿狗的动作规划

方式)。除此之外，作者还展示了根据狗的行为构建的模型也可以泛化至其他的一些任务。更重要的是，在使用狗行为模型为可行走表面预测以及场景识别等任务作预训练之后，这些任务的结果准确率都得到了提高。

2、Related Work

几乎没有和建模狗行为相关的工作，只有 Visual prediction、Sequence to sequence models、Ego-centric vision、Ego-motion estimation、Action inference & Planning、Inverse Reinforcement Learning、Self-supervision，这些都是最相关的，但是实际上也和作者的工作很不同。

3、Dataset

作者在狗的头上安装了 GoPro 相机来拍摄以自我为中心的视频。作者以每秒 5 帧的速度对帧进行子采样。。作者使用惯性测量单元 IMUs 来测量身体的位置和运动。四个 IMU 测量狗的四肢位置，一个测量狗的尾巴，一个测量狗的身体位置。IMU 使作者能够捕捉到角位移的运动。

对于每一帧，作者都有六个 IMU 的绝对角位移。每个角位移都表示为一个四维四元数向量。两个连续帧之间角位移的差值(时间上为 0.2s)表示狗在该时间步长的动作。

狗背上的 Arduino 连接到 IMUs 并记录位置信息。它还通过安装在狗背上的麦克风收集音频数据。作者收集了各种场景的数据：客厅、楼梯、阳台、街道和狗公园都是这些场景的例子，有 50 多个不同的地点。作者记录了狗狗在进行活动时的行为，如散步、跟随、抓取、与其他狗狗互动以及跟踪物体。没有为视频帧提供注释，作者使用原始数据进行实验。

4、Acting like a dog

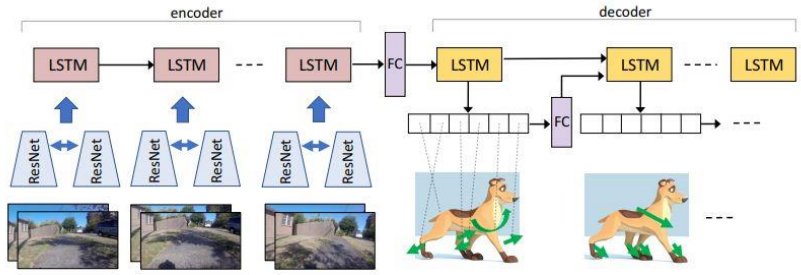

图 2. 模仿狗行为的模型架构。该模型是一个编码器-解码器神经网络。编码器接收一系列图像对，解码器输出各个关节的预测动作。编码器和解码器之间有一个全连接层（FC），

以更好地捕捉相关域中的变化（从图像变为动作）。在解码器中，每一个时间步的动作输出概率会被传输至下一个时间步。两个 ResNet 塔共享权重。

网络如图 2。

作者的运动预测模型是基于 encoder-decoder 架构，其目标是找到输入图像和未来动作之间的映射。例如，如果狗看到其主人拿着一袋食物，很有可能狗会坐下来等待食物，或者如果狗看到其主人扔一个球，狗可能会跟踪球并朝它跑去。

5、Planning like a dog

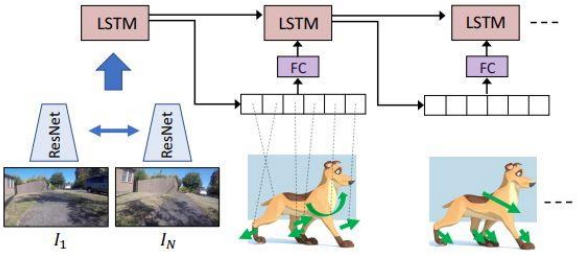

图 3. 用于模仿狗动作规划方式的模型架构。这个模型结合了 CNN 和 LSTM。模型的输入是两个图像 I_1 和 I_N，它们在视频中相差 N-1 个时间步。LSTM 接收来自 CNN 的特征数据作为输入，然后输出一组能使狗从 I_1 的状态转化为 I_N 的动作（关节屈伸）。

网络如图 3。

另一个目标是模拟狗如何计划行动来完成任务。为此，作者设计了一个任务:给定一对非连续的图像帧，计划狗从第一帧(开始状态)到第二帧(结束状态)的一系列关节运动。

狗所做的每一个动作都改变了世界的状态，因此也就为下一步做了计划。因此，作者设计了一个递归神经网络。此外，在当前时间步长的低概率动作可能会导致在序列中更远处的高概率轨迹。使用动作概率可以防止早期剪枝以保留未来动作的所有可能性。

6、Learning from a dog

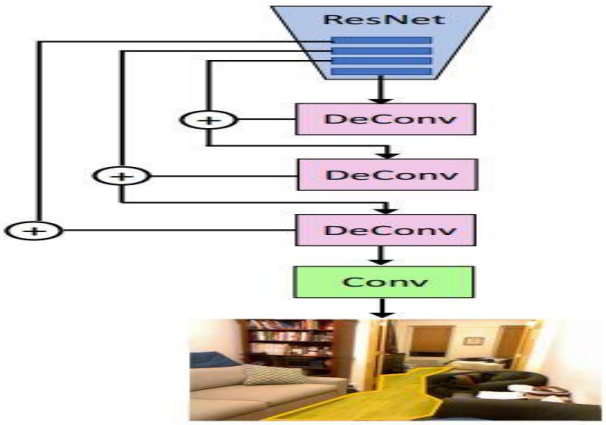

图 4. 用于预测可行走表面的模型架构。作者使用解卷积和卷积层来增强 ResNet 的最后四层，得出可供行走的表面。

当作者从狗狗观察到的图像中学习预测狗狗关节的运动时，作者得到了一个编码不同类型信息的图像表示。为了学习表示，通过观察狗观察的在时间 t-1 至 t 的图像，作者训练了 ResNet-18 模型来估计当前狗运动(IMU 的变化从时间 t - 1 到 t)。然后作者测试这种表示方法，使用分离的数据在不同的任务与 ImageNet 训练的 ResNet-18 模型比较。在实验中，作者选择了利用 SUN397 数据集进行可步行表面估计和场景分类的任务。图 4 描述了从图像中估计可行走表面的模型。

7、Experiments & Results

Learning to Act Like a dog：

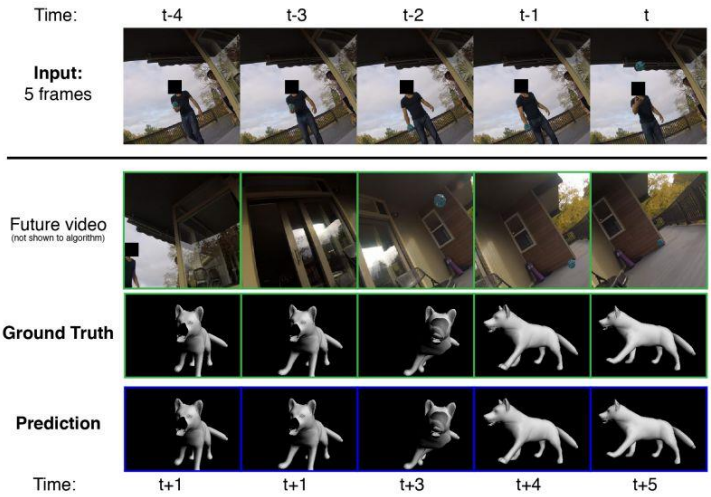

图 5. 定性结果：模型学会了如何执行动作。作者向模型输入了一个视频的五帧，这五帧中一个男人开始向一只狗扔球。在视频中，这个球撞到墙反弹，而狗转向右边来追这个球。仅仅是使用了视频一开始的五帧，该模型就能精确地预测出狗在球飞过时如何转向右侧的。

Model	Test Accuracy	Perplexity
Nearest Neighbor	12.64	N/A
CNN	19.84	0.2171
Our Model–1 tower	18.04	0.2023
Our Model–1 frame/timestep	19.28	0.242
Our Model	**21.62**	**0.2514**

表 2.模仿动作模型的输出结果。作者输入了视频的前五帧然后预测接下来的五个动作。

实验举例如图 5。获取 5 帧的视频序列并预测接下来 5 个的工作，准确性结果如表 2，可见相较于基线都有很大提升。

Learning to plan like a dog：

Model	Test Accuracy	Perplexity
Nearest Neighbor	14.09	N/A
CNN	14.61	0.1419
Our Model	**19.77**	**0.2362**

表 3. 模仿规划方式模型的输出结果。预测了从开始帧到结束帧之间的动作组。作者认为从开始的图像转化成结束的图像需要五步（每步间隔 0.2s）。

表 3 显示了作者对规划任务的实验结果。结果表明，作者的模型在像狗一样进行有挑战性的规划任务时，在 Accuracy 和 Perplexity 方面都优于这些基线。

Model	Angular metric	All joints
Random	131.70	4e-4
CNN-acting	63.42	8.67
Our model-acting	**59.61**	**9.49**
CNN-planning	76.18	0.14
Our model-planning	**63.14**	**3.66**

表 4. 对模型效果的评估。第一列（Angular metric）当中的数值越小越好。第二列（All joints）当中数值越大越好。

可以看出，在 Learning to act like a dog 和 Learning to plan like a dog 上，作者的结果都有较大的提升。

Learning from a dog：

Model	Pre-training task	IOU
ResNet-18	ImageNet Classification	42.88
ResNet-18	Acting like a dog	**45.60**

表 5. 评估可行走表面的预测。评价指数是 IOU。作者模型从 acting 任务中学习到的有更好的效果。

这里分别选择了可行走表面预测和场景分类两项任务进行测试。

可行走表面预测任务的目标是标记与图像中可行走区域（例如，地板、地毯和地毯区域）相对应的像素。在作者的数据集中，作者有一些在室内和室外场景中遛狗的序列。有各种各样的障碍（例如，家具，人，或墙壁）在场景中，狗狗会避开。作者推测，这项学习可为估计可行走表面提供强有力的启发。

人类和狗的可行走表面的定义是不一样的。有些区域不适合人类行走和适合狗行走。然而，由于作者实验用的狗是体型较大的狗，所以人类和狗对行走能力的定义大致相同。

表 5 显示了结果。与 ImageNet 相比，作者的特性提供了 3%的显著改进。作者使用 IOU 作为评价指标。

对于场景分类任务，作者使用 SUN 397 数据集进行了一个额外的场景识别实验。作者学习到的表征的准确率为 4.48。这很有趣，因为作者的数据集不包含许多场景类型（加油站、商店等）。

8、Conclusion

作者研究的任务是直接建模一个视觉智能体。作者的模型从以自我为中心的视频和运动信息中学习，像狗在相同的情况下一样行动和计划。作者在定量和定性结果上都取得了一些成功。作者的实验表明，作者的模型可以预测狗狗未来的动作，并能设计出类似于狗狗的动作。

这是视觉智能体端到端建模的第一步。这种方法不需要手动标记数据等代表性任务任务或目标的详细语义信息。尽管缺少语义标签，但是作者可以在各种各样的代表性场景中使用这个模型，并学习有用的信息。

对于这项工作，作者只考虑了视觉数据。然而，视觉智能体在与世界交互时使用多种输入模式，包括声音、触摸、气味等。作者有兴趣扩展当前模型，以包含更多的输入模式在一个组合的端到端模型。作者的工作也仅限于为一只特定的狗做模型。从多只狗身上收集数据并评估狗的综合能力将会很有趣。作者希望这项工作为更好地理解视觉智能和居住在作者世界上的其他智能生物铺平了道路。

9、Reasons for recommendation

很少见以狗为第一视角建模视觉智能体，十分有趣和新颖，引入了大量的测试数据，并且不需要标记数据。此外，还分别建模了不同的网络，并且都取得了良好的效果和实验结果。