RT-1: Robotics Transformer for Real-World Control at Scale

1. 摘要

通过从庞大的、多样性的以及与任务不相关的数据集中提取到相关信息，现代机器学习模型可以解决一些没遇到过的或者是只包含在小型的特定任务数据集中的下游任务，而且性能也达到了一个较高的水准。在计算机视觉、自然语言处理和语音识别等领域中，现代机器学习模型已经表现得相当不错，但在机器人领域仍没有展现应有的效果。另外，在机器人领域中，收集真实世界的机器人训练数据比较困难，因此阻碍了机器学习模型泛化能力的提升。因此，我们认为获得高泛化能力模型的关键在于一个可以容纳大量机器人数据的高容量的模型以及持续进行各种未知任务的训练。在这篇文章里，我们提出一类称为“Robotics Transformer”的模型，这类模型展现了令人信服的拓展能力。我们收集了大量机器在真实世界执行任务的数据，在这个基础上，通过控制训练数据规模、训练数据的类别和模型大小，研究了不同种类的模型和它们的泛化能力。我们把这个项目的资料上传到 robotics-transformer.github.io。

2. 介绍

端到端的机器人训练基本上是模仿或者是强化学习这两种方式。这两种方式都需要收集任务导向的训练数据，包括单任务和多任务场景设定。通过这样

的训练方式，模型只能在这些任务场景中有不错的表现。这种工作流程类似其它领域的有监督学习，比如计算机视觉和自然语言处理；在流程中，需要收集任务导向的数据并对这些数据打标签，然后部署模型去处理特定的任务场景，而且不同的任务场景间几乎没有关联。近几年，计算机视觉、自然语言和其它领域都从单一小规模向大型通用模型转变，这些模型提前使用大规模的数据集进行训练。这些模型成功的关键在于一个可以容纳大量训练数据的高容量的架构以及持续进行各种未知任务的训练。在语言处理或者感知方面，如果模型可以通过吸收大量经验信息学习到普遍性原理，模型可以运用这些原理高效地完成那些单一的任务。在有监督的训练方式中，会尽量避免使用大量的任务导向数据，原因是收集这样的数据成本不低；这机器人领域里，收集任务导向数据会更加困难，可能需要大量的人工操作或者是需要人工设定的自动化操作。

因此我们提出一个疑问：是否可以在宽范围的机器人数据集里训练一个单一的、性能强大的、适应多任务的模型；如果可以，这个模型是否还具备与其它领域模型相同的泛化能力。

在机器人领域，训练这样的模型有很多困难。近年来，机器人领域发表的文章提出了一些适应多任务的策略（Reed et al.,2022;Jang et al.,2021），有的模型在面对真实世界的任务时表现非常有限，例如 Gato（Reed et al.,2022）；有的模型只针对训练任务的表现而非泛化到新任务，例如最近的指令响应型模型（Shridhar et al.,2021;2022）；能泛化到新任务的模型也只表现出较低的泛化水平（Jang et al.,2021）。

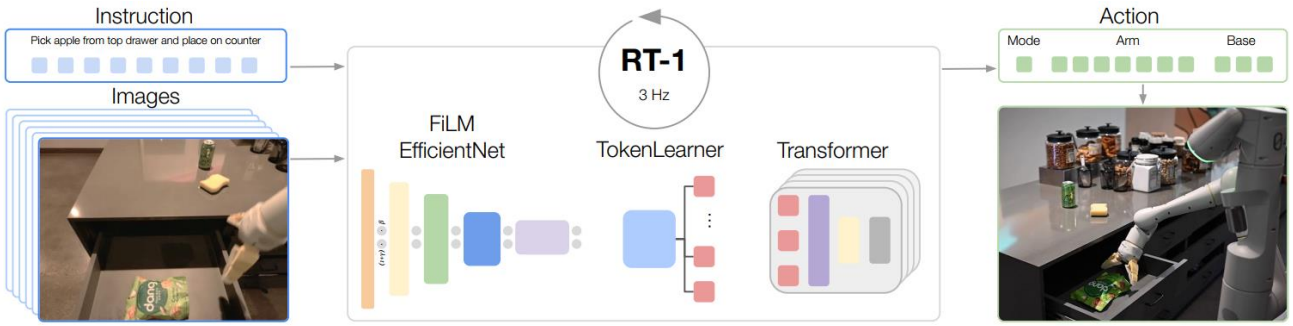


图 0-1（a）RT-1 接受图片和自然语言指令输入并输出独立的基座和机械臂动作指令。尽管 RT-1 的模型尺寸非常大（35M 个参数），它运行的频率可以达到 3 赫兹，这是源于它拥有高效的模型架构：FiLM（Perez et al.,2018）conditioned EfficientNet（Tan & Le,2019），TokenLearner（Ryoo et al.,2021）,和 Transformer（Vaswani et al.,2017）。

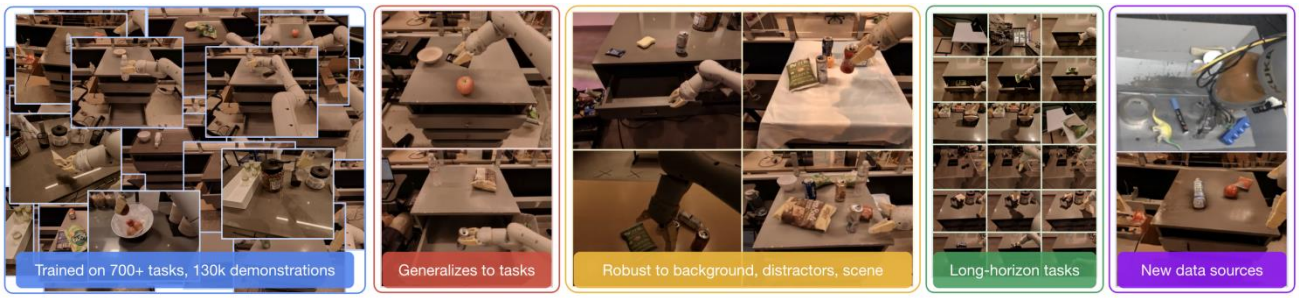


图 0-2（b）基于大规模、真实世界的训练（130k 训练规模）和验证（3000 次真实世界的试验）,RT-1 的展示了令人印象深刻的泛化能力、鲁棒性以及从宽范围数据学习的能力。

制作合适的数据集和设计合适的模型是达成最终目标的两大挑战。在大规模机器学习项目中，训练数据的记录方式和采集策略通常是无名英雄（Radford et al.,2021; Ramesh et al.,2021）；在以人工采集和机型适配为通常情况的机器人领域尤其突显（Dasari et al.,2019;Ebert et al.,2021）。在文章中，我们会指出好的泛化能力需要训练数据具有规模和覆盖范围，包含了各种各样的任务和场景设定。同时，训练数据中相似的任务内容需要有良好的联系以促进模型的泛化能力，从而模型可以在训练中提取相似任务中的共同特征并在面对由这些特征重新组成的新任务时可以胜任。我们在训练中所使用的数据包含了约 130k 回合和超过 700 个任务。这些数据由一组包含 13 台机器人的队伍采集 17 个月。在以下的验证中，我们将数据集分成了多个方面进行验证。

模型架构设计是第二个挑战。高效的多任务的机器人学习需要一个高容量的模型，而 Transformer（Vaswani et al.,2017）模型很好的满足这个要求，特别

是，像我们所提出的，所训练的这些多任务场景都是基于自然语言指令下的情况。机器人的控制器同样需要足够快地实时地执行指令，这也是 Transformer 模型的主要的挑战。综上，我们提出一个新的模型架构，称为 RT-1（Robotics Transformer 1），它可以高效地进行推理从而实时控制机器人。新架构将高维的输入输出，包括图像、自然语言指令输入和执行器指令，编码成紧凑的代号供 Transformer 模型使用。

我们的主贡献是提出 RT-1 模型以及在大量宽泛的真实世界的机器人任务数据中测试这个模型。对比以往文章所提出的方案，我们的测试表明 RT-1 模型在泛化能力和鲁棒性中具有重大的提升；而且验证和排除了很多模型架构和训练数据集的组成方案。我的实验结果表明，RT-1 模型可以执行超过 700 种训练指令，成功率约为 97%，并且在新任务、新干扰和新场景背景下的泛化能力相较第二名高出 25%、36%和 18%。这个性能可以执行 SayCan（Ahn et al.,2022）框架里至多 50 回这么长的的任务序列。我还表明 RT-1 可以结合仿真数据或者其它型号的机器人数据，可以保持在原有任务场景下的性能还提高对新场景的泛化能力。图 1b 展示了 RT-1 模型的性能。

3. 相关文献

最近的文献提出基于 Transformer 架构的机器人控制策略。一些文章同样使用了 Transformer 来处理自然语言指令，从而组成一个鲁棒性高的框架来执行和泛化到新任务（Zhang & Chai,2021;Pashevich et al.,2021;Silva et al.,2021;Jang et al.,2021;Ahn et al.,2022;Nair et al.,2022）。在本文中，我们将自然语言指令和视觉输入、机器人控制指令序列输出定义为一个映射问题并使用 Transformer 来学习这个映射。我们的灵感来源于游戏控制（Chen et al.,2021;Lee et al.,2022a）、仿真

机器人导航（Fang et al.,2019）、移动控制（Janner et al.,2021;Gupta et al.,2022）和运动控制（Jiang et al.,2022）。我们也注意到提及的文章中探索了除自然语言外的其它指令下发方式（e.g.,Jang et al.,2021;Jiang et al.,2022）并将 Transformer 运用到其它机器人的平台上（e.g.,Gupta et al.,2022）。这也是 RT-1 往后的探索方向。

除了基于 Transformer 的策略外，本文还专注于在真实世界里机器人控制可扩展的泛化能力和鲁棒性。现有的论文中专注于在真实世界里和基于 Transformer 的机器人控制主要解决如何高效地通过几个实例来学习执行每个任务（Shridhar et al.,2022）。行为 Transformer（Shafiullah et al.,2022）和 Gato（Reed et al.,2022）提倡在大规模机器人和非机器人数据集上训练单一模型。上述的文章都定位在真实世界的机器人任务中；e.g.,Gato 只高效地学习执行一个任务（堆叠彩色方块）而没有验证在其它新任务的泛化情况或者其它真实世界的环境设定。从技术角度来看，本文验证了如何构建一个基于 Transformer 策略的模型架构以达到高容量和具有泛化性，并且具有足够的推理效率来实现在真实世界里控制机器人。

使用高容量的 Transformer 模型来训练机器人控制策略还是相当新的方法。面向多任务和基于自然语言指令下发的机器人控制策略训练则有相当长的历史了。RT-1 模型建立在这些研究上。相当体量的研究主要解决机器人抓取任务的训练策略和推理模型设计以获得抓取取其物品的泛化性（Saxena et al.,2006;Lenz et al.,2015;Pinto & Gupta,2016;Gupta et al.,2018;Viereck et al.,2017）。以往的研究工作已经尝试通过将自然语言、视觉、机器人控制和端到端的训练方式组成串行的流程使机器人能响应自然语言的指令。面向多任务的机器人训练也尝试了通过强化学习来训练（Chung et al.,2015;Raffin et al.,2019;Jurgenson et al,.2020;Huang et al.,2020）,(Deisenroth et al.,2014;Devin et al.,2017;Fox et

al,.2019;Kalashnikov et al,.2021a)。还有部分研究是专注于收集展示数据和试验数据的（Sharma et al.,2018;Dasari et al.,2019;Yu et al.,2020;Singh et al.,2020;James et al.,2020）。本文的工作验证了基于多任务和自然语言指令下发的训练机器人控制模型的能力，展示了模型在更大范围内、更多动作种类、更多的物品种类和更多的环境背景的推理结果，提出了新的模型架构和设计方向，使得训练能扩展到更大的范围。

4. 预备知识

4.1. 机器人学习

我们的目的是在视觉作为输入的基础上，训练机器人控制策略来响应自然语言下发的指令。理论上，我们将问题设定为生成序列决策指令。在 t=0 时刻，决策策略π接受一个自然语言指令 i 和一张图片输入x_0。π生成一个动作指令的分布$\pi\left(\cdot|i,\{x_j\}_{j=0}^{t}\right)$，从这个分布上采样得到动作指令$a_0$并发送到机器人控制器上。这个过程不断重复直到满足停止条件。从$t=0$到停止时刻 T，整个交互阶段$i,\{(x_j,a_j)\}_{j=0}^{T}$称为一个回合。每个回合结束后会生成一个才二值的得分$r\in\{0,1\}$来表示机器人是否完成了下发的指令。训练目的是在给定的这些指令中，从状态x_0开始，经过中间过渡阶段，将平均得分最大化。

4.2. Transformers

RT-1 使用 Transformer 模型结构（Vaswani et al.,2017）来建模决策策略π。Transformer 结构是一个序列式的模型，由 self-attention 层和全连接层组成的神

经网络；将一个输入序列$\{\xi_h\}_{h=0}^{H}$映射到一个输出序列$\{y_h\}_{k=0}^{K}$上。最初，transformer 是为文本序列设计的，每个输入ξ_i和输出y_k都表示一个文本符号；后来也被应用到图像作为输入的领域（Parmar et al.,2018）和其它输入形式（Lee et al.,2022a;Reed et al.,2022）。将决策策略π参数化为：先将输入序列$i,\{x_j\}_{j=0}^{t}$映射到序列$\{\xi_h\}_{h=0}^{H}$上以及将期望的动作序列a_t映射到序列$\{y_h\}_{k=0}^{K}$，然后训练 Transformers 模型将$\{\xi_h\}_{h=0}^{H} \rightarrow \{y_h\}_{k=0}^{K}$上。在下一小节我们会详细描述这个过程。

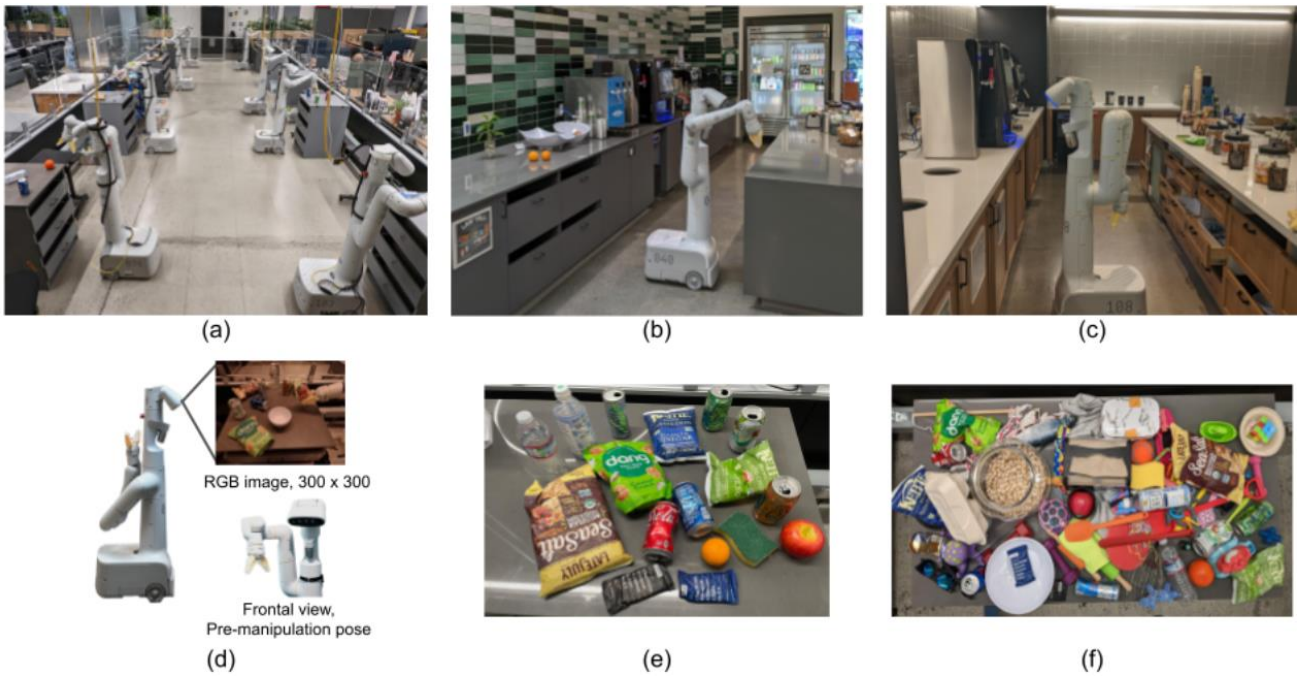
4.3. 模仿学习

模仿学习是在一个数据集 D 上训练一个策略π（Pomerleau,1988;Zhang et al.,2018;Jang et al.,2021）。需要指出的是，我们假设有一个数据集$D = \left\{\left(i^{(n)},\left\{\left(x_t^{(n)},a_t^{(n)}\right)\right\}_{t=0}^{T^{(n)}}\right)\right\}_{n=0}^{N}$，里面包含了每个动作的真实序列。我们运用 behavioral cloning（Pomerleau,1988），通过优化输入图像序列和自然语言指令输出动作序列与真实序列的 negative log-likelihood 获得决策策略π。

5. 系统概况

本文的目的是搭建和展示一个通用机器人学习系统。这个系统可以从大量的训练数据中提取有用的信息并能高效地将模型泛化。我们的机器人平台是 Everyday Robot。它的机械臂拥有 7 个自由度和一个两指抓取器。上装安装在一个可移动的底座上，如所示。我们在三个厨房的环境里收集训练数据和验证模型，其中包含两个真实的办公室厨房和一个其它样式的仿真厨房。这个仿真厨

房是为了收集大量数据而建造的，里面有一个操作台，如所示。两个真实的厨房有类似的操作台，但照明、背景墙和还有其它厨房布置都不同。我们在这三个厨房场景里验证模型，评估模型的性能和泛化性。



6. 模型

我们的模型建立在 Transformer 框架上（Vaswani et al.,2017），接受图像序列和任务描述作为输入，输出代号化的控制指令，如所示。我将自上而下的介绍模型架构。详细的模型细节请参考附录 C.3。

6.1. 任务指令和图像的代号化

RT-1 建立在高效的数据处理和紧凑的代号化的图像和自然语言指令。RT-1 采用一个图像特征提取层来处理按时间排列的包含 6 张图像的序列。这个图像特征提取层在 ImageNet 上做预训练 （即 EfficientNet-B3）（Tan & Le，2019），

输入的图像尺寸是 300*300，输出尺寸为 9*9*512 的特征。与 Reed et al.（2022）不同，我们不提前将这些特征组成视觉代号喂给 Transfomrer，而是展开成 81 个视觉代号传给下一层网络。

为了增加自然语言指令到视觉代号中，我们将自然语言指令通过一个预训练的语言编码层附加到视觉代号上，这样可以提前提取出与任务相关的视觉特征从而提高模型的性能。这些自然语言指令首先经过 Universal Sentence Encoder（Cer et al.，2018）编码，然后输出给 identity-initialized FiLM 层（Perez et al.，2018），再输出到 EfficientNet。正常来说，在预训练网络中直接插入 FiLM 层会破坏其中的连接关系影响预训练网络的效果。为了消除这个问题，我们将密集层（fc 和 hc）初始化为 0，使得 FiLM 在起始阶段不起作用从而保留预训练参数的效果。另外，我们也发现这样的初始化操作与非预训练的 EfficientNet 结合也获得较好的效果，但没有比使用预训练的效果好。

RT-1 的 FiLM+EfficientNet-B3 共有 16M 的参数，包含了 26 层 MBConv 块和 FiLM 层，最后输出 81 个视觉-自然语言代号。

6.2. TokenLearner

RT-1 使用 TokenLearner（Ryoo et al .,2021）来进一步压缩代号的数量以便在后面的 Transformer 中提高效率。TokenLearner 以代号为单位，将输入的代号集合映射到较小的代号集合中。这使得我们可以根据所需要的信息，通过 soft-select 的方式筛选图像代号，只把关键的代号传递给后续的 Transformer 层。TokenLearner 将来自预训练层的 81 个代号重新采样至只有 8 个并传递给 Transformer 层。

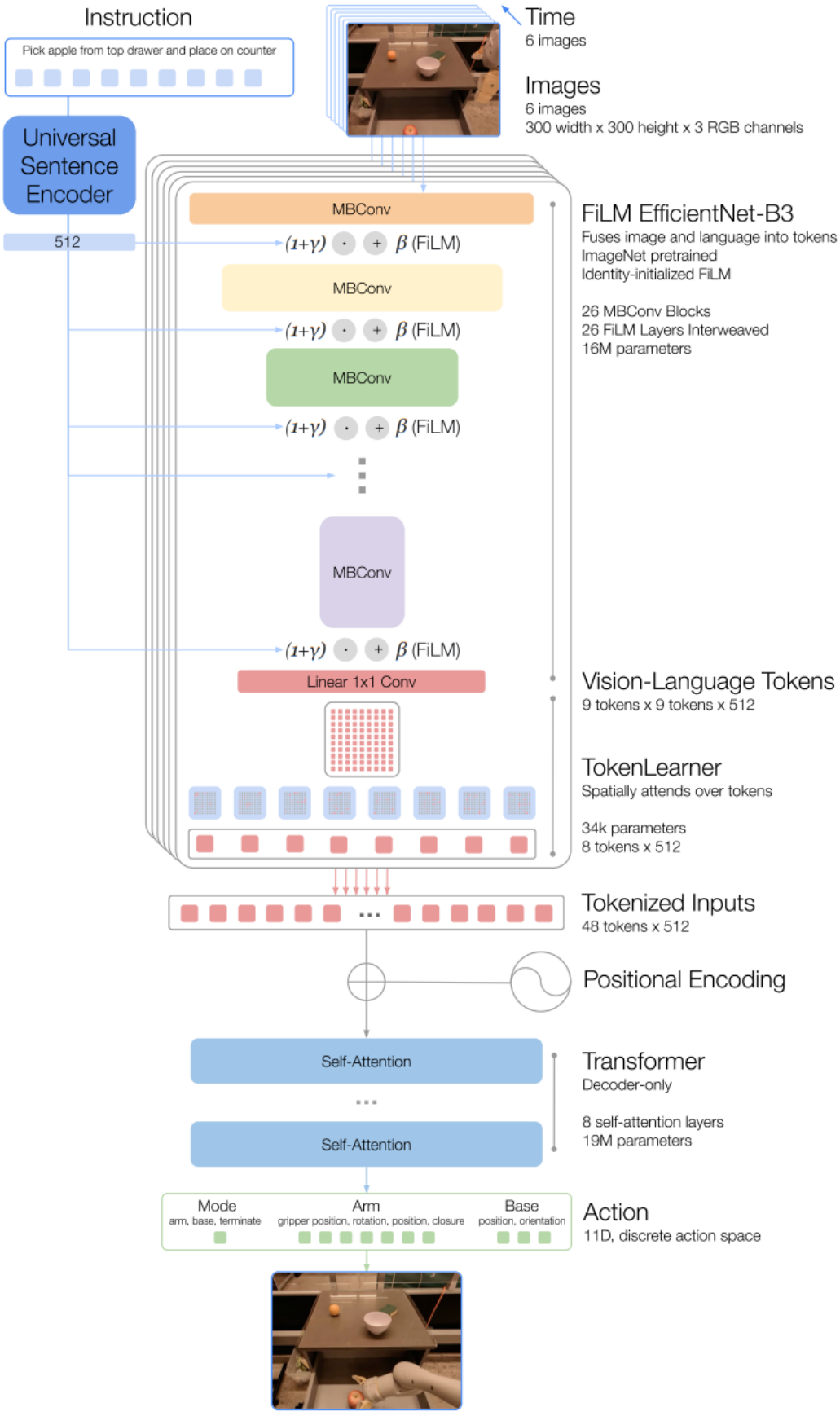


图 0-3 RT-1 的模型架构。自然语言指令以 USE 编码通过 FiLM 层附加到预训练的 EfficientNet 的输出上。输

出的视觉-自然语言代号再经过 TokenLearner 重新采样并传递到解码器式的 Transformer，最终输出代号化的操指令。

6.3. Transformer

将 6 张图像的代号组合起来构成 48 个代号，附加每张图像的位置编码信息，输入到 Transformer 中。Transformer 是解码器式的序列模型结构，带有 8 个 self-attention 层共 19M 参数，最终输出控制机器人所需要的代号指令。

6.4. Action Tokenization

在 RT-1 中，每个控制指令分解成 256 位并编码成代号。前面提到过，我们将每个动作指令分解成机械臂运动的 7 个自由度（x，y，z，roll，pitch，yaw，抓取器开关），底座的运动（x，y，yaw），以及一个单独的模式来在机械臂控制、底座控制和停止动作回合中切换。每个自由度都分解成 256 位，每一位都均匀分布在整个有效范围内。

6.5. Loss

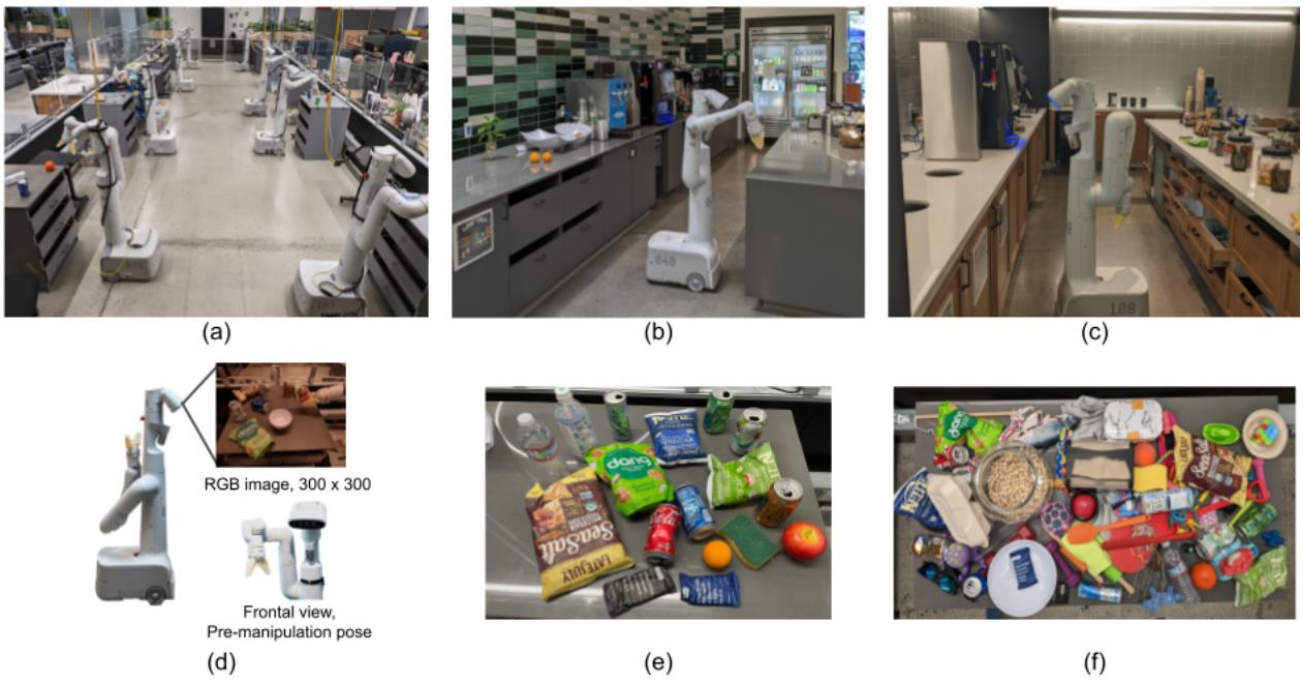
根据先前的基于 Transformer 的控制器（Reed et al.,2022；Lee et al.,2022a），本文模型中使用了标准的 categorical cross-entropy 和 casual masking 作为优化目标。

6.6. Inference Speed

区别于运用在其它领域的大模型，比如说自然语言处理和图像处理，在实时控制机器人的场景中，需要保证模型的推理速度足够快和稳定。人类在执行

本文中动作指令时大概需要 2-4 秒，因此我们希望模型推理速度不能比人类慢太多。我们从实验结果推算，满足这样的执行速度至少需要整个 3hz 的控制频率，考虑系统其它模块的耗时，留给模型推理的时间不到 100ms。而且这样的要求也限制了模型的大小。我们在实验中进一步探索了模型的大小对推理速度的影响。为此我采用了两个方法来提高推理速度：（i）使用 Tokenlearner 将前面的代号重新采样到较小的数量（Ryoo et al.,2021），（ii）在指令执行过程中，连续的相似的操作环境中使用相同的代号。以上两个方法将推理速度分别提高了 2.4 和 1.7 倍。详细推理过程请参考附录 C.1。

本文的目标是搭建一个具有高性能、好的泛化性至新任务以及面对干扰因素和不同的环境背景时有好的鲁棒性的机器人控制系统。因此我们要采集大规模、多样性的机器人轨迹数据，包含了多任务、多个抓取物体和多个操作环境。我的基础训练集包含了约 130k 个回合的机器人执行动作，这些数据是由 13 个机器人在 17 个月内采集的。这些机器人在几个办公室厨房的场景中采集数据，我们也将这些场景称为“机器人课堂”。训练数据的详细说请参考附录 C.2。



7. 数据集

7.1. Skills and Instructions

在本文中，对机器人任务的定义没有固定的格式，因此我们只计算系统能执行的任务数量。自然语言指令的格式是一个动词+一个或多个名词，例如“把水瓶放到右上角”，“把可乐瓶移到绿色薯片袋旁边”，“打开抽屉”。RT-1 可以在多个真实的办公室厨房环境中执行超过 700 个自然语言指令，我们在实验章节详细说明验证过程。为了可以定性的分析系统的性能，我们将指令按照动作的类型分类，也称为技能类型。详细的指令类型请看表。

目前技能类型包含了“抓取”、“放置”、“打开/关闭抽屉”、“拿出抽屉里的物品/把物品放进抽屉”、“把长条形的物品放到右上角/碰倒”、“拿手帕”和打开罐头。选择这些技能的原因是想展示 RT-1 处理多个动作与不同物品组合的能力，评估系统的能力，比如泛化到新指令、是否能执行很多任务等。我们还把“抓取”动作的对象扩大了，评估泛化到其它物品的情况。其它验证过程请看附录 C.4。

Skill	Count	Description	Example Instruction
Pick `Object`	130	Lift the object off the surface	pick iced tea can
Move `Object` Near `Object`	337	Move the first object near the second	move pepsi can near rxbar blueberry
Place `Object` Upright	8	Place an elongated object upright	place water bottle upright
Knock `Object` Over	8	Knock an elongated object over	knock redbull can over
Open `Drawer`	3	Open any of the cabinet drawers	open the top drawer
Close `Drawer`	3	Close any of the cabinet drawers	close the middle drawer
Place `Object` into `Receptacle`	84	Place an object into a receptacle	place brown chip bag into white bowl
Pick `Object` from `Receptacle` and Place on the Counter	162	Pick an object up from a location and then place it on the counter	pick green jalapeno chip bag from paper bowl and place on counter
Section 6.3 and 6.4 tasks	9	Skills trained for realistic, long instructions	open the large glass jar of pistachios pull napkin out of dispenser grab scooper
Total	744		