Visual Navigation with Spatial Attention 复现

龚威龙

摘要

这项工作侧重于对象目标视觉导航，旨在从给定类中找到对象的位置，在每个步骤中，代理都会获得场景的以自我为中心的 RGB 图像。我们建议使用强化学习算法来学习代理的策略。我们的主要贡献是用于视觉导航任务的新颖注意力概率模型。这种注意力编码了关于观察到的物体的语义信息，以及关于它们位置的空间信息。“什么”和“哪里”的这种组合允许代理有效地导航到寻找的对象。注意力模型被证明可以改进代理的策略并在常用数据集上实现最先进的结果

关键词：注意概率分布；视觉导航

1 引言

人类和动物可以相对较好地适应新环境。这种对新环境的适应虽然很自然，但并非微不足道。它需要在新的观察结果和我们过去的经验之间找到相似之处。这在很大程度上是可能的，因为我们有能力对新的视觉信息进行分类，并智能地关注最相关的语义线索。例如，当在以前未访问过的厨房中寻找烤面包机时，我们的直觉是寻找冰箱，而忽略其他“不相关”的信息，因为我们过去的经验表明烤面包机通常位于离冰箱不远的地方。而智能体显然不具备此能力，如何改进智能体使之具备此能力，是我们所追求的。对象目标视觉导航任务包括两个基本组成部分：场景的语义理解和路径规划。

2 相关工作

导航是移动机器人技术中最基本的问题之一。传统的导航方法将问题分解为两个独立的阶段：映射周围环境和规划通往目标的路径[1]强化学习 (RL) 方法被应用于学习机器人任务的策略[2-3]。虽然 RL 方法能够以端到端的方式学习复杂任务，但它们在视觉导航任务中的主要挑战是理解视觉线索和导航计划。我们的空间信息依赖于图像区域上的注意力概率分布。该组件在我们参与的嵌入中充当重要的构建块，它结合了图像的空间和语义信息。我们在 AI2THOR[4]上验证了我们的视觉导航方法，这是一个由接近照片般逼真的 3D 室内场景[5]组成的环境。我们增加了 Wortsman 等人[6]的自适应视觉导航 (SAVN) 工作，并通过参与观察为其模型不可知元学习器 (MAML)[7]提供输入。我们的工作开发了一个嵌入式注意力模块，它结合了语义和空间信息[8]。空间信息由图像区域上的注意力概率分布编码，这些区域的语义信息由卷积网络编码。视觉任务中的注意力主要用于语言增强任务.

2.1 SAVN 自适应视觉导航框架

SAVN[6]依赖于自适应导航，因此它的策略受益于适应相关的导航子任务，例如，进入走廊、接近冰箱等。为了处理如此复杂的任务，SAVN 应用了模型不可知元学习 (MAML)，它在智能体与场景交互时改变策略参数。这种参数的转变允许智能体在与场景交互时适应场景。SAVN 目标是使 agent 在与环境交互时不断学习。与 MAML 一样，我们将 SGD 更新用于这种 adaptation。这些 SGD 更新会在 agent 与场景交互时修改 agent 的策略网络，使 agent 能够适应场景。本文建议这些更新应针对 loss 进行，将其称为交互损失。最小化交互损失应该有助于 agent 完成其导航任务。SAVN 通过使用交互损

失 $\mathcal{L}_{\text{int}}^{\tau}(\theta,\alpha)$ 实现此行为，该损失应用于动作 a 的 $\hat{k}$-prefix α, 即 $\alpha=(a_1,\ldots,a_{\hat{k}})$ 。因此对于一系列动作 a 及其前缀 α 的任务 τ 学习策略 $\pi_\theta(\cdot\mid s)$ 的损失函数是

$$\min_{\theta}\sum_{\tau\in\mathcal{T}_{\text{rain}}}\mathbb{E}_{a\sim\pi_\theta}\left[\mathcal{L}_{\text{nav}}^{\tau}\left(\theta-\nabla_\theta\mathcal{L}_{\text{int}}^{\tau}(\theta,\alpha),a\right)\right] \tag{1}$$

3 本文方法

下文讲介绍复现论文的主要创新内容以及其主要思想和奖励设计等。本文主要创新内容是提出了一种新颖的注意力机制，该方法能够使用卷积网络对有关观察到的对象的语义信息进行编码，并使用注意力概率模型对有关其位置的空间信息进行编码。通过在输入图像的 $n_v\times n_v$ 个子窗口上构造一个注意力概率分布。这种概率分布将高概率分配给图像中具有相关信息的子窗口，并将低概率分配给不具有相关信息的子窗口。通过这样做，注意力概率分布将空间信息引入到过程中。我们的注意力概率分布由三个注意力单元组成：目标注意力单元，它结合了图像中的目标信息；动作注意单元，它考虑了代理的最后一个动作；记忆注意单元，它从场景中先前看到的图像中“记住”相关信息。然后将 $n_v\times n_v$ 子窗口上的这三个分布融合成图像子窗口上的单个注意概率分布。我们用 $p^t(i,j)$ 表示时间 t 在 $n_v\times n_v$ 个子窗口上的融合概率分布。第 t 个图像的空间参与嵌入 $\hat{v}_{i,j}^t$ 结合了图像中的语义信息以及关于不同对象位置的空间信息。语义信息由向量 $v_{i,j}^t\in\mathbb{R}^{d_v}$ 表示，而空间信息由注意力概率分布 $p^t(i,j)$ 表示。这种嵌入允许代理根据图像的语义和空间信息选择下一步，因为它被作为导航策略的输入。

3.1 输入特征

本文的输入是由三个注意力单元组成。分别是目标注意力单元，动作注意单元，记忆注意单元。它们的详细信息如下：目标注意力单元：它以第 t 步的图像和目标（由单词给出）作为输入，旨在关注图像中与目标相关的信息。我们用 u_g 表示目标的词向量，它是由 GloVe 嵌入[9]，$v_{i,j}^t$ 表示第 t 个时间步的 $n_v\times n_v$ 图像向量，W_v 为在 d 维空间中嵌入子窗口嵌入 $v_{i,j}^t$ 的可训练参数，W_g 为在同一空间中嵌入目标嵌入 u_g 的可训练参数。对于每个子窗口索 $\text{i,j}\in\{1,\ldots,\text{nv}\}$，时间 t 的目标 u_g 和图像子窗口 $v_{i,j}^t$ 之间的潜在交互 φtg(・) 采取以下形式：

$$\phi_g^t(i,j)=\left\langle\frac{W_v v_{i,j}^t}{\left\|W_v v_{i,j}^t\right\|},\frac{W_g u_g}{\|W_g u_g\|}\right\rangle \tag{2}$$

通过运用 softmax 操作获得相应的注意力概率分布：

$$p_g^t(i,j)=\frac{e^{\phi_g^t(i,j)}}{\sum_{s,t=1}^{n_v}e^{\phi_g^t(s,t)}} \tag{3}$$

动作注意单元：它将图像和最后一步的动作分布作为输入。我们用 $u_a^{(t-1)}$ 表示策略 πθ(・|s) 利用时间 t-1 的动作分布，W_a 为在同一空间中嵌入目标嵌入 $u_a^{(t-1)}$ 的可训练参数。时间 t 的目标 $u_a^{(t-1)}$ 和图像子窗口 $v_{i,j}^t$ 之间的潜在交互 φtg(・) 采取以下形式：

$$\phi_a^t(i,j)=\left\langle\frac{W_v v_{i,j}^t}{\left\|W_v v_{i,j}^t\right\|},\frac{W_a u_a^{(t-1)}}{\left\|W_a u_a^{(t-1)}\right\|}\right\rangle \tag{4}$$

通过运用 softmax 操作获得相应的注意力概率分布：

$$p_a^t(i,j)=\frac{e^{\phi_a^t(i,j)}}{\sum_{s,t=1}^{n_v}e^{\phi_a^t(s,t)}} \tag{5}$$

记忆注意单元：它总结了智能体的经验，旨在根据情节中已经收集到的信息来关注图像的各个部分。我们用 $u_m^{(t-1)}$ 表示从 LSTM 单元的隐藏状态中提取的记忆，W_m 为在同一空间中嵌入目标嵌入 $u_m^{(t-1)}$ 的可训练参数。时间 t 的目标 $u_m^{(t-1)}$ 和图像子窗口 $v_{i,j}^t$ 之间的潜在交互 φtg(•) 采取以下形式：

$$\phi_m^t(i,j) = \left\langle \frac{W_v v_{i,j}^t}{\left\|W_v v_{i,j}^t\right\|}, \frac{W_m u_m^{(t-1)}}{\left\|W_m u_m^{(t-1)}\right\|} \right\rangle \tag{6}$$

通过运用 softmax 操作获得相应的注意力概率分布：

$$p_m^t(i,j) = \frac{e^{\phi_m^t(i,j)}}{\sum_{s,t=1}^{n_v} e^{\phi_m^t(s,t)}} \tag{7}$$

最后，我们学习实值权重函数来融合不同的注意力概率分布时间 t，得到的融合注意力概率分布：

$$p^t(i,j) \propto \left(p_g^t(i,j)^{\beta_g} p_a^t(i,j)^{\beta_a} p_m^t(i,j)^{\beta_m}\right) \tag{8}$$

3.2 整体模型框架

SAVN 的网络模型优化了两个目标函数，自监督交互损失 $\mathcal{L}_{\text{int}}^{\tau}(\theta,\alpha)$ 和导航损失 $\mathcal{L}_{\text{nav}}^{\tau}(\theta,a)$，如图2所示。在训练过程中，交互和导航梯度通过网络反向传播，自监督损失的参数在每一集结束时使用导航梯度更新。在测试时，交互损失的参数保持固定，而网络的其余部分使用交互梯度进行更新。

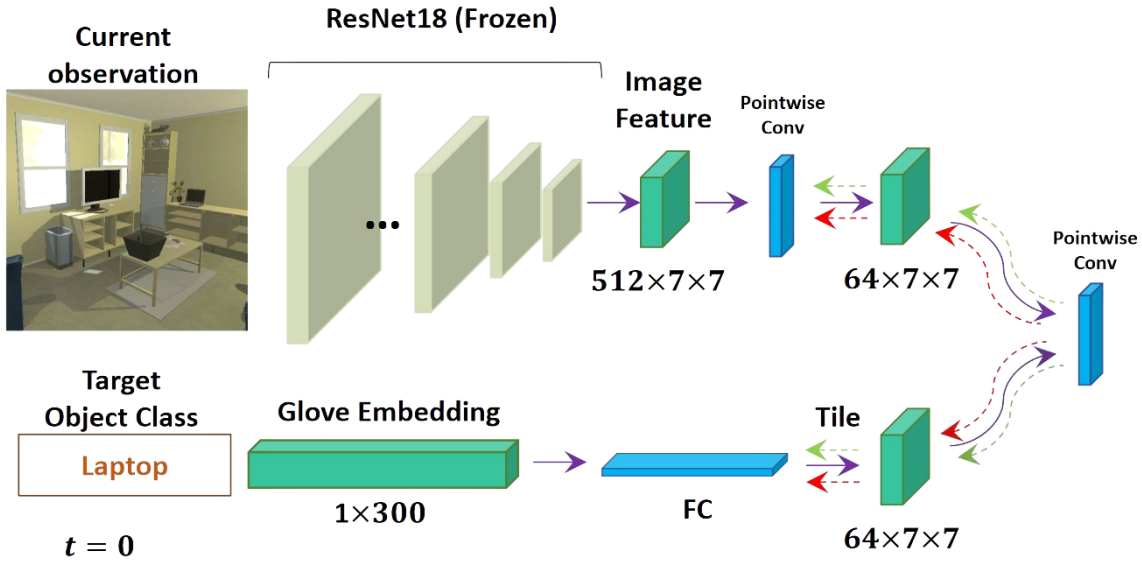

图 1: 方法示意图

本文在 SAVN 的基础上增加了嵌入式注意力模块，其对应网络模型如图 2所示。

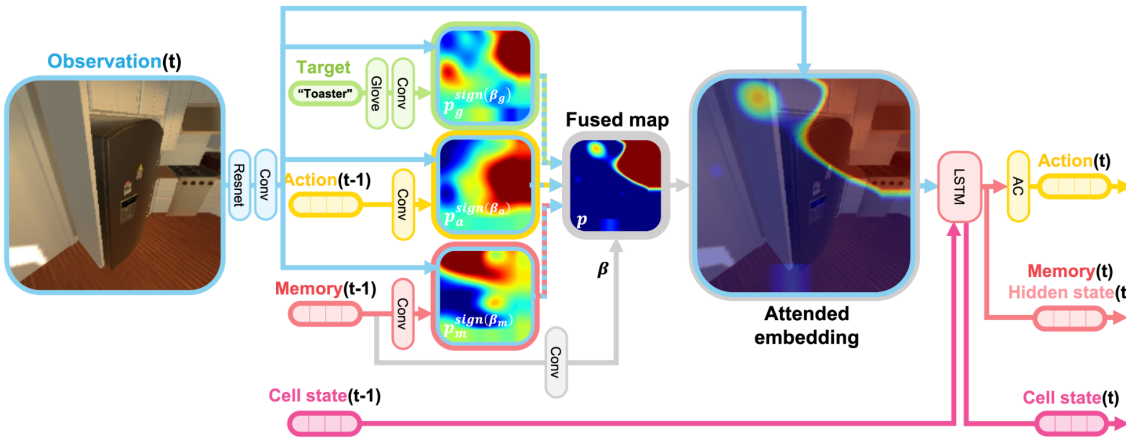

图 2: 方法示意图

在 LSTM 单元中，此时刻智能体的经验，记忆等会被存储在隐藏状态中，供下一时刻使用。

3.3 损失函数定义

我们学习了一个策略 $\pi_\theta(\cdot \mid s)$，它在场景 S 中选择一个给定的以自我为中心的 RGB 图像 s 的动作。我们使用 gradient decent（策略梯度）来改进策略的参数 θ 以在每一集中导航. 学习这些参数是为了最大化给定情节中一系列动作的预期奖励 $\mathbb{E}\left[\mathcal{R}^\tau\right]$。在我们的实验评估中，我们使用 SAVN 导航奖励 $\mathcal{R}^\tau_{\text{nav}}$，它为除了完成之外的任何步骤减去 0.01，并为成功导航添加 5。我们还使用 actor-critic 系列算法来最小化其导航损失 $\mathcal{L}^\tau_{\text{nav}}(\theta, a)$，它由服务于 actor 的负预期奖励和服务于 critic 的学习价值函数组成。

4 复现细节

4.1 与已有开源代码对比

在复现的实现过程中，我大多数参考的原论文源代码。

Procedure 1 SAVN-Training(Ttrain, α, β1, β2, k)

Input: Randomly initialize θ, φ
Output: θ, φ

while *not converged* **do**
  for *mini-batch of tasks* $\tau_i \in \mathcal{T}_{train}$ **do**
    $\theta_i \leftarrow \theta$
    $t \leftarrow 0$
    while *termination action is not issued* **do**
      Take action a sampled from $\pi_{\theta_i}(s_t)$
      t $\leftarrow t+1$
      if *t is divisible by k* **then**
        $\theta_i \leftarrow \theta_i - \alpha \nabla_{\theta_i} \mathcal{L}^\phi_{\text{int}}\left(\theta_i, \mathcal{D}^{(t,k)}_\tau\right)$
      end
    end
    $\theta \leftarrow \theta - \beta_1 \sum_i \nabla_\theta \mathcal{L}_{\text{nav}}(\theta_i, \mathcal{D}_\tau)$
    $\phi \leftarrow \phi - \beta_2 \sum_i \nabla_\phi \mathcal{L}_{\text{nav}}(\theta_i, \mathcal{D}_\tau)$
  end
end

上面的伪代码学习了一个策略网络 πθ 和一个以 φ 为参数的损失网络，其步长超参数为 α，β1，β2。

4.2 实验环境搭建

实验环境采用的是 Ubuntu，机器信息为 RTX3070，cuda9，pytorch 版本不同可能对于一些方法的实现有所改动。

4.3 创新点

尝试使用 GA3C 替换 A3C，但是没有取得成功。

5 实验结果分析

实验结果如下表所示，表中使用成功率和按路径长度 (SPL) 加权的成功率对这些方法进行评估，将我们的结果与最先进的结果进行了比较，并显示了在成功率和路径长度 (SPL) 方面对短路径和长路径的改进。在我们的实验验证过程中，A2C 算法最终效果比 A3C(asynchronous actor-critic) 算法都要好。

表 1: 定量结果

Architecture	SPL	Success	SPL L≥5	Success L≥5
SAVN	16.15	40.86	13.91	28.70
Ours (A3C)	16.99	43.20	15.51	31.71
Ours (A2C)	**17.88**	**46.20**	**15.94**	**32.63**

6 总结与展望

原文提出了一种用于视觉导航的端到端强化学习。原文的框架基于适合视觉导航的新颖注意力概率模型，因为它编码了有关观察对象的语义信息和有关其位置的空间信息。具体来说，注意力模型由三部分组成：目标、动作和记忆。原文采用了 A2C(Advantage Actor Critic) 算法取得了 SOTA, 我们在实现的时候尝试参考 GA3C(混合 CPU/GPU 版的 A3C), 理论上可以提升网络的训练速度，但是最终没有成功。在实现过程中，本文在处理数据上有很多地方有不足之处，参考了网上代码，这是需要之后去提升的地方。

参考文献

[1] BLÖSCH M, WEISS S, SCARAMUZZA D, et al. Vision based MAV navigation in unknown and unstructured environments[J]., 2010: 21-28.

[2] KOHL N, STONE P. Policy gradient reinforcement learning for fast quadrupedal locomotion[C]//IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004: vol. 3. 2004: 2619-2624.

[3] PETERS J, SCHAAL S. Reinforcement learning of motor skills with policy gradients[J]. Neural networks, 2008, 21(4): 682-697.

[4] ZHU Y, MOTTAGHI R, KOLVE E, et al. Target-driven visual navigation in indoor scenes using deep reinforcement learning[J]., 2017: 3357-3364.

[5] KOLVE E, MOTTAGHI R, HAN W, et al. Ai2-thor: An interactive 3d environment for visual ai[J]. arXiv preprint arXiv:1712.05474, 2017.

[6] WORTSMAN M, EHSANI K, RASTEGARI M, et al. Learning to learn how to learn: Self-adaptive visual navigation using meta-learning[J]., 2019: 6750-6759.

[7] FINN C, ABBEEL P, LEVINE S. Model-agnostic meta-learning for fast adaptation of deep networks[C]//International conference on machine learning. 2017: 1126-1135.

[8] SCHWARTZ I, YU S, HAZAN T, et al. Factor graph attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 2039-2048.

[9] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]//Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014: 1532-1543.