arXiv:2303.05052v1 [cs.RO] 9 Mar 2023

经遗传算法优化的基于 VQA 的机器人状态识别

Kento Kawaharazuka[1], Yoshiki Obinata[1],Naoaki Kanazawa[1],Kei Okada[1],和 Masayuki Inaba[1]

摘要—机器人对物体和环境的状态识别已通过多种方式实现。在大多数情况下，这一过程通过处理点云、利用标注图像进行学习以及使用专用传感器来完成。与此不同，本研究提出了一种状态识别方法，将视觉问答（VQA）应用于在大规模数据集上训练的预训练视觉语言模型（PTVLM）。借助 VQA，可以用口语直观地描述机器人状态识别。另一方面，同一事件可以通过多种方式提问，而状态识别的性能会因问题而异。因此，为了提升基于 VQA 的状态识别性能，我们使用遗传算法搜索合适的问题组合。我们展示了该系统不仅能够识别冰箱门的开/关状态和显示屏的开/关状态，还能识别透明门的开/关状态以及水的状态，而这些状态此前一直难以识别。

I. 引言

机器人在空间中移动并执行任务时，识别周围的物体、工具和环境必不可少。例如，机器人在 [1], [2] 周围移动时，需要识别房间门和电梯门的开/关状态、灯的开/关状态等。此外，机器人巡查区域时，还需要识别冰箱门和柜门的开/关状态、显示屏的开/关状态，以及水龙头是否仍在流水等。迄今为止，这些状态识别主要采用以下方法：通过人工编程从原始图像或点云中提取特征 [3], [4], ；人工标注图像并使用神经网络进行训练 [5], ；或者根据待识别的状态安装相应的传感器 [6]。换言之，我们针对每种需要识别的状态分别构建了识别器。另一方面，采用这些方法时，随着待识别状态数量的增加，识别器的数量和需要收集的数据也会随之增加，这不仅需要耗费大量时间构建系统，也使资源管理变得困难。此外，人们在四处走动时，会不断识别各种

难以轻易编程的状态。这些状态包括透明门的开/关、水的状态等。机器人不太容易识别这些状态，因此人们针对识别本身开展了各种研究[7]。

因此，在本研究中，我们提出了一种使用口语进行状态识别的方法。我们应用视觉

[1] 作者来自日本东京都文京区本乡 7-3-1，邮编 113-8656，东京大学信息科学与技术研究生院机械信息学系。 [kawaharazuka, obinata, kanazawa, k-okada, inaba]@jsk.t.u-tokyo.ac.jp

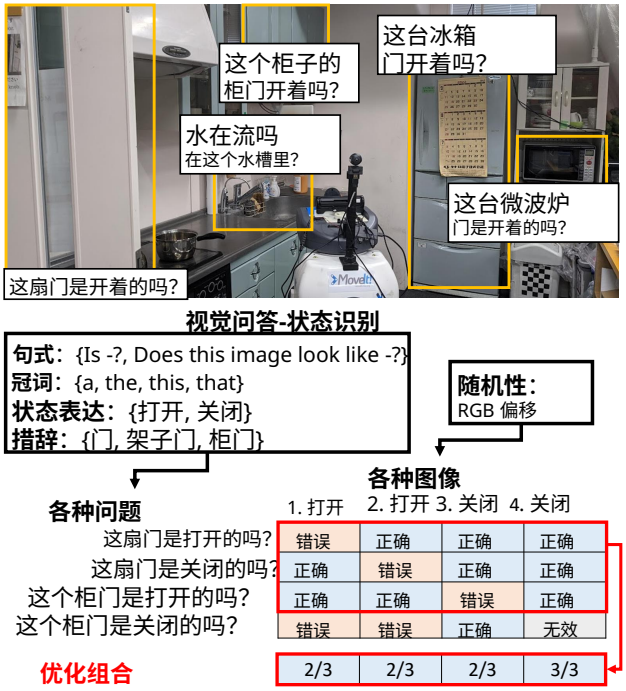

图 1. 本研究的概念。通过遗传算法优化适当问题的组合，以使用视觉问答进行二元状态识别。

基于视觉问答（VQA） [8]，使用在大规模数据集上训练的预训练视觉语言模型（PTVLM） [9], [10] 。通过询问有关当前图像的问题并获取句子形式的答案来进行状态识别。借助口语，即使是难以用程序轻易描述的状态也能得到识别。另一方面，初步实验表明，识别性能很大程度上取决于问题，尤其是所用的句式、冠词、状态表达和措辞。由于不同问题各自容易识别和难以识别的状态并不相同，因此仅凭一个问题只能准确识别部分状态。因此，本研究开发了一种利用遗传算法搜索合适问题组合的方法，并构建了高性能状态识别器。我们展示了该系统不仅能够识别冰箱门的开/关状态和显示屏的开/关状态，还能够识别以往难以识别的透明门开/关状态以及水的状态。本研究表明，只需为视觉问答提供候选问题，即可自动生成合适且高性能的识别器；并且，状态

识别可以通过口语和单个视觉语言模型简单、快速地完成，从而显著提升机器人的识别表现。

II. 通过遗传算法优化的基于 VQA 的机器人状态识别

A. 使用预训练视觉语言模型的基于 *VQA* 的机器人状态识别

首先，我们介绍一种使用视觉问答的状态识别方法。视觉问答是一项通过向图像 V提出问题Q 来获取答案 A 的任务。具体而言，我们处理二元状态识别，这种识别方式可轻松用于机器人的动作分支。换言之，该任务是通过提出诸如“是否是 -？ ”或“是否会 -？ ”之类的问题，获取“是”或“否”的回答。另一方面，这类问题并非总能得到“是”或“否”的回答，往往会得到其他短语。因此，获得的回答分为三类：正确、错误和无效。在此，通过增强图像或更改问题，可以针对同一图像获得多个回答。获得的正确、错误和无效响应数量表示为 $N_{\{correct,wrong,invalid\}}$，并将正确率 $R_{correct}$ 定义为 $N_{correct}/（N_{correct}+N_{wrong}）$，将无效率 $R_{invalid}$ 定义为 $N_{nvalid}/（N_{correct}+N_{wrong}+N_{invalid}）$（计算$R_{correct}$时不包括$N_{invalid}$ ）。如果对于某张图像，$R_{correct}$ 大于 0.5，则该图像被正确识别。

接下来，我们介绍如何对图像进行增强，以及如何针对图像生成各种问题。首先，我们使用 RGB-Shift 进行图像增强：从均匀分布中采样一个随机值，并将其加到每个 RGB 值上，取值范围为 [-0.1, 0.1] 。这样，我们便能针对同一图像和问题获得多个回答，并对这些回答取平均值（本研究中，我们为每个图像-问题对准备了六张增强图像）。当然，也可以采用其他增强方法，例如高斯噪声。接下来，我们通过组合四种句式变化来生成各种类型的问题：疑问形式、冠词、状态表达和措辞。句式指提问的方式；本研究使用了“Is -?”（例如“Is this door open?”）和“Does -?”（例如“Does this image look like this door is open?”）等形式。冠词指用于目标对象的“a”“the”“this”“that”。状态表达指使用反义词，例如用“closed”表示“open”，用“is not”表示“is”，等等。措辞指针对显示器使用“display”“monitor”“TV”等词；本研究为每个状态识别器准备了最多四种措辞。因此，我们最多会准备$2\times4\times2\times4$，也就是 64 个问题。这里最多有$2^{64}-1$，即约 10^{19} 种问题组合。当然，问题类型并不限于这些，还可以通过单复数等方式进一步增加。

在本研究中，我们使用 OFA [11]作为预训练视觉语言模型，可用于视觉问答。OFA 是一种通过同时学习图像描述、视觉定位和文本到图像生成等多项任务而具备强大泛化能力的模型。

B. 使用遗传算法优化合适的问题组合

对于第II-A节中提到的大量问题，我们搜索能够实现最佳准确率的问题组合。在本研究中，我们使用遗传算法来完成这一任务。问题组合由二进制向量 $\boldsymbol{s}$ ($\in \{0,1\}^{N_q}$ 表示，其中N_q 表示所有问题的数量，并对其进行优化。值为 1 的问题将被使用，而值为 0 的问题不使用。作为优化数据集，我们收集了 $N_{data}^{train}=20$ 幅图像。例如，在进行门的开闭状态识别时，我们准备了 10 幅门处于打开状态的图像和 10 幅门处于关闭状态的图像。

给定 $\boldsymbol{s}$ 的评估值是以下三个值（a）–（c）之和。首先，尽可能正确地识别数据集中图像的状态是最重要的。也就是说，对于数据集中的每幅图像 i ($1\leq i\leq N_{data}^{train}$)，我们使用 Correct Rate $R_{correct}^{i}$ 超过 0.5 的图像数量 $N_{correct}^{img}$。该$R_{correct}^{i}$ 是所有值为 $\boldsymbol{s}$ 的问题以及经过 RGB 偏移增强的图像 i 的正确回答率。在实际应用中，我们使用正确识别的图像数量占所有图像的比例（a） $R_{correct}^{img}=N_{correct}^{img}/N_{data}^{train}$ 作为评估值。其次，如果相同，则 Correct Rate 应更高。因此，我们将 Correct Rate 的期望值，即数据集中所有图像的平均值（b） $R_{correct}^{ave}$ of $R_{correct}^{i}$作为评估值。对于（a）和（b），我们还定义 $R'_{correct}=N_{correct}/(N_{correct}+N_{wrong}+N_{invalid})$，并使用 $R_{correct}^{img'}$ 和 $R_{correct}^{ave'}$ 作为评估值进行比较。预计这样可以将无效回答视为与错误回答等价，从而避免选择包含大量无效回答的问题。第三，如果 $R_{correct}^{img}$ 相同，则使用较少的问题可以降低计算复杂度。另一方面，问题数量越多，越有可能基于更多回答实现稳定的状态识别。因此，我们使用组合中所使用问题数占所有问题数的比例（c） $R_q=N_q^s/N_q$（其中 N_q^s是 $\boldsymbol{s}$中 1 的数量）作为评估值。在本研究中，我们使用以下四种评估值 E比较优化结果，

$$E_+=\alpha R_{correct}^{img}+R_{correct}^{ave}+\beta R_q \tag{1}$$
$$E'_+=\alpha R_{correct}^{img'}+R_{correct}^{ave'}+\beta R_q \tag{2}$$
$$E_-=\alpha R_{correct}^{img}+R_{correct}^{ave}-\beta R_q \tag{3}$$
$$E'_-=\alpha R_{correct}^{img'}+R_{correct}^{ave'}-\beta R_q \tag{4}$$

其中， α 和 β 为权重系数；在本研究中，$\alpha=100$ 和 $\beta=0.1$ 用于按照 (a)、(b) 和 (c) 的顺序优先考虑评估值。同样，我们将使用这些评估值获得的问题组合分别记作$\boldsymbol{s}_+$、 $\boldsymbol{s}'_+$、 $\boldsymbol{s}_-$、$\boldsymbol{s}'_-$。

遗传算法的详细设置如下。使用 DEAP 库 [12]，并使用函数 cxTwoPoint

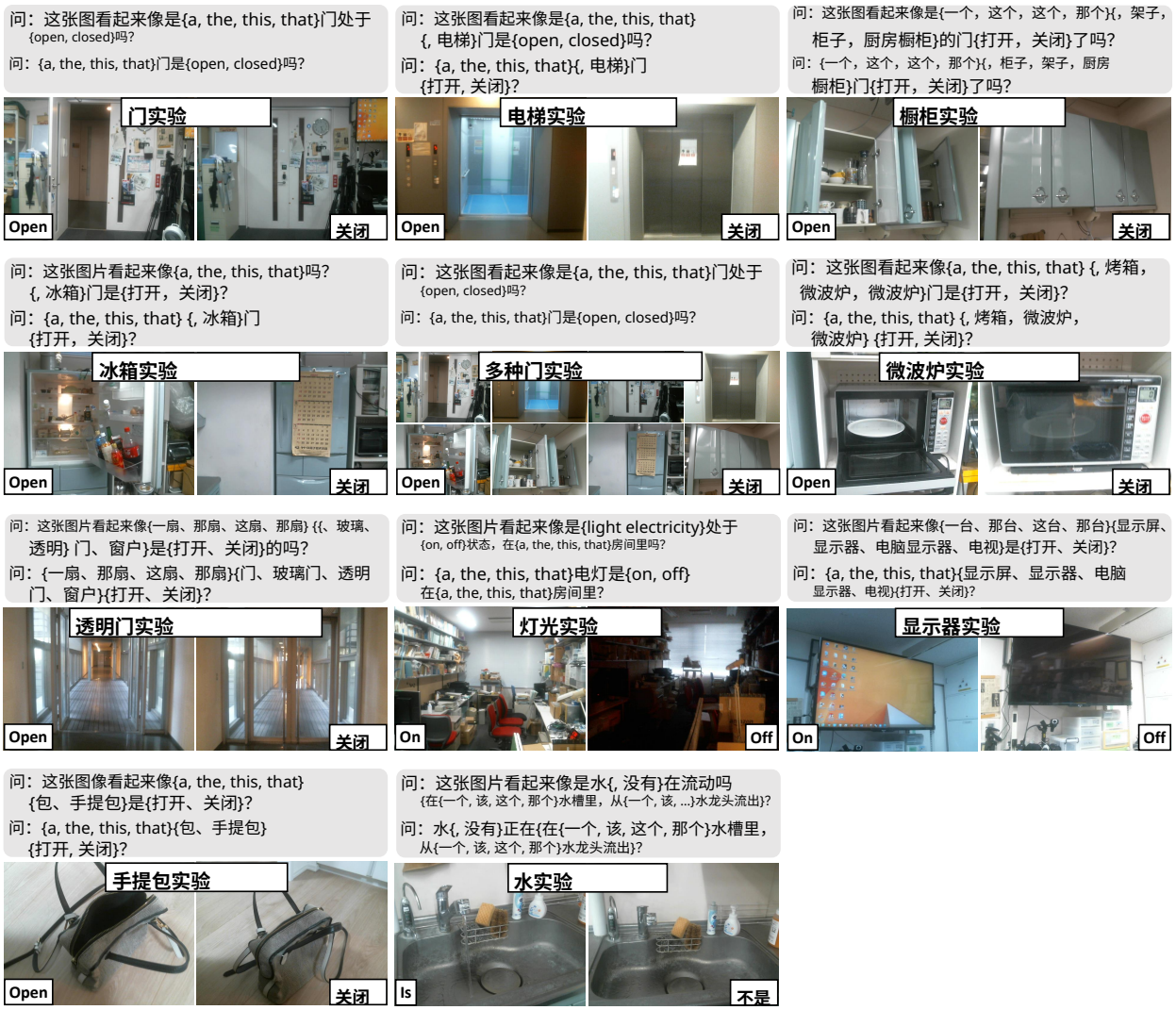

图 2。 问题集及门、电梯、柜子、冰箱、多种门、微波炉、透明门、灯的代表性图像显示屏、手提包和水实验

ht, ent.

进行交叉，概率为 50%；使用函数 mutFlipBit 进行变异，概率为 20%。使用函数 selTournament 选择个体，锦标赛规模设为 5。个体数量和迭代代数分别设为 1000 和 200。

此外，在本研究中，我们准备了 $s_{\{does,is\}}$，其中使用了所有“Does -?”或“Is -?”形式的问题。我们还将 s的所有值均为 1 的 s 记为 s_{all}。使用训练数据集优化问题组合后，我们准备了一个不同于训练数据集的测试数据集，并比较 $N^{img}_{correct}$、$R^{ave}_{correct}$、$R_{invalid}$和 N^{s}_{q}（$R_{invalid}$是所有响应中无效响应的比例）。测试数据集中的图像数量与训练数据集相同。

III. 实验

实验中使用的问题集和代表性图像如图 2 所示。实验旨在识别普通门、电梯门、柜子门、冰箱门、各种门，以及微波炉门和透明门是打开还是关闭，灯或显示屏是开启还是关闭，手提包是打开还是关闭，以及水是否在流动。

A. 门实验

表 I 展示了标准门开闭状态的识别结果。尽管 s_{does} 中只有 15/20 被正确识别，但 s_{is}、s_{all} 和优化结果表明，所有图像均被正确识别。s_{+}和 s'_{+} 使用了大量问题，而 s_{-} 和 s'_{-} 仅使用最佳问题。比较 s_{+} 和 s'_{+}时，s'_{+} 的无效响应更少，且使用的问题更少。这被认为是因为 s'_{+} 将无效视为与错误相同，因此选项仅限于不太可能产生无效响应的选项。

B. 电梯实验

表 II 展示了电梯门开闭状态的识别结果。这里使用了“门”和“电梯门”两种措辞。虽然$s_{\{does,is,all\}}$ 在部分图像上识别出错，但$s_{\{+,-\}}$ 和 $s'_{\{\{+,-\}}$ 能够正确识别所有图像。

表 I 门实验结果

	$N_{correct}^{img}$	$R_{correct}^{ave}$	$R_{invalid}$	N_q^s
$\boldsymbol{s}_+$	20 / 20	0.981	0.035	11 / 16
$\boldsymbol{s}'_+$	20 / 20	0.977	0.001	8 / 16
$\boldsymbol{s}_-, \boldsymbol{s}'_-$	20 / 20	1.000	0.000	1 / 16
$\boldsymbol{s}_{does}$	15 / 20	0.800	0.000	8 / 16
$\boldsymbol{s}_{is}$	20 / 20	0.904	0.056	8 / 16
$\boldsymbol{s}_{all}$	20 / 20	0.852	0.028	16 / 16

此外，$\boldsymbol{s}_{\{+,-\}}$ 和 $\boldsymbol{s}'_{\{+,-\}}$ 的优化结果完全一致，这表明仅使用最佳问题的准确度显著高于使用其他问题。最佳问题是“这张图像看起来像是门打开了吗？”。此外，比较 $\boldsymbol{s}_{does}$ 和 $\boldsymbol{s}_{is}$时，$\boldsymbol{s}_{does}$的 $R_{correct}^{ave}$ 更低，且 $R_{invalid}$ 比 $\boldsymbol{s}_{is}$更小，这与第 III-A 节中的趋势相同。

表 II 电梯实验结果

	$N_{correct}^{img}$	$R_{correct}^{ave}$	$R_{invalid}$	N_q^s
$\boldsymbol{s}_{\{+,-\}}, \boldsymbol{s}'_{\{+,-\}}$	20 / 20	0.992	0.000	1 / 32
$\boldsymbol{s}_{does}$	17 / 20	0.750	0.000	16 / 32
$\boldsymbol{s}_{is}$	17 / 20	0.796	0.048	16 / 32
$\boldsymbol{s}_{all}$	18 / 20	0.773	0.024	32 / 32

C. 橱柜实验

表 III 展示了柜门开闭状态的识别结果。这里使用了四种措辞：“门”、“柜门”、“厨房橱柜门”和“架子门”。对于所有问题组合，所有状态均成功识别。其特征与上述实验相同，$\boldsymbol{s}_+$ 的 $R_{invalid}$ 和 N_q^s 大于 $\boldsymbol{s}'_+$。此外，$\boldsymbol{s}_{does}$ 的 $R_{correct}^{ave}$ 较低，且 $R_{invalid}$ 小于 $\boldsymbol{s}_{is}$。

表 III 橱柜实验结果

	$N_{correct}^{img}$	$R_{correct}^{ave}$	$R_{invalid}$	N_q^s
$\boldsymbol{s}_+$	20 / 20	0.963	0.082	26 / 64
$\boldsymbol{s}'_+$	20 / 20	0.971	0.000	17 / 64
$\boldsymbol{s}_-, \boldsymbol{s}'_-$	20 / 20	1.000	0.000	1 / 64
$\boldsymbol{s}_{does}$	20 / 20	0.865	0.000	32 / 64
$\boldsymbol{s}_{is}$	20 / 20	0.878	0.125	32 / 64
$\boldsymbol{s}_{all}$	20 / 20	0.873	0.062	64 / 64

D. 冰箱实验

表 IV 展示了冰箱门开闭状态的识别结果。优化后的问题组合在准确率方面优于 $\boldsymbol{s}_{\{does,is,all\}}$。就所有优化结果而言，$N_{correct}^{img}$ 在用于优化的训练数据集上为 20/20，而在某些情况下，测试数据集上的结果为 19/20。其他特征与上述实验相似。

E. 多种门实验

表 V 展示了迄今描述的全部四种门的开闭状态识别结果。措辞仅使用“门”

表 IV 冰箱实验结果

	$N_{correct}^{img}$	$R_{correct}^{ave}$	$R_{invalid}$	N_q^s
$\boldsymbol{s}_+, \boldsymbol{s}'_+$	19 / 20	0.964	0.000	3 / 32
$\boldsymbol{s}_-$	19 / 20	0.975	0.000	1 / 32
$\boldsymbol{s}'_-$	20 / 20	0.983	0.000	1 / 32
$\boldsymbol{s}_{does}$	15 / 20	0.727	0.000	16 / 32
$\boldsymbol{s}_{is}$	17 / 20	0.737	0.025	16 / 32
$\boldsymbol{s}_{all}$	17 / 20	0.732	0.013	32 / 32

，但所给数据集包含标准门、电梯门、柜门和冰箱门。训练数据集和测试数据集分别使用了 80 张门的图像。优化后的问题组合在准确率方面优于 $\boldsymbol{s}_{\{does,is,all\}}$。换言之，如果待识别的状态具有相似特征，就可以在不更改问题的情况下进行状态识别，无需分别构建各个识别器。此外，仅用一个问题即可识别的图像数量为 72/80，低于优化结果中的 76/80，这表明采用恰当的问题组合能够实现更高的识别准确率。

表 V 多种门实验结果

	$N_{correct}^{img}$	$R_{correct}^{ave}$	$R_{invalid}$	N_q^s
$\boldsymbol{s}_+$	76 / 80	0.884	0.075	7 / 16
$\boldsymbol{s}_-$	76 / 80	0.890	0.041	3 / 16
$\boldsymbol{s}'_{\{+,-\}}$	76 / 80	0.890	0.004	4 / 16
$\boldsymbol{s}_{does}$	64 / 80	0.789	0.000	8 / 16
$\boldsymbol{s}_{is}$	73 / 80	0.857	0.092	8 / 16
$\boldsymbol{s}_{all}$	72 / 80	0.823	0.046	16 / 16

F. 微波炉实验

表 VI 展示了微波炉门开闭状态的识别结果。此处使用“门”“微波炉门”“烤箱门”和“微波炉门”这四种措辞。优化后的$\boldsymbol{s}_{\{+,-\}}$ 尤其在准确率方面优于$\boldsymbol{s}_{\{does,is,all\}}$。另一方面，$N_{correct}^{img}$ 在 $\boldsymbol{s}'_{\{+,-\}}$ 中的数值小于其在$\boldsymbol{s}_{\{+,-\}}$中的数值。这被认为是因为 $\boldsymbol{s}'_{\{+,-\}}$ 将无效响应视为错误，因此会把许多无效响应但状态识别准确的问题排除在选项之外。$R_{invalid}$ 在$\boldsymbol{s}_{\{+,-\}}$ 中的数值大于其在$\boldsymbol{s}_{\{does,is,all\}}$中的数值，这表明$\boldsymbol{s}_{\{+,-\}}$ 会主动使用会产生无效响应的问题。请注意，多种门实验不包括微波炉门，因为简单的“门”这一措辞使得识别微波炉门的状态变得困难。

G. 透明门实验

表 VII 展示了透明门开闭状态的识别结果。这里使用了四种措辞：“门”、“透明门”、“玻璃门”和“窗户”。优化的问题组合在准确率方面显著优于 $\boldsymbol{s}_{\{does,is,all\}}$。准确率

表 VI 微波炉实验结果

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^{s}_{q}
$\boldsymbol{s}_{+}$	17 / 20	0.703	0.315	6 / 64
$\boldsymbol{s}_{-}$	16 / 20	0.689	0.210	5 / 64
$\boldsymbol{s}'_{+}$	12 / 20	0.546	0.008	47 / 64
$\boldsymbol{s}'_{-}$	12 / 20	0.596	0.015	8 / 64
$\boldsymbol{s}_{does}$	12 / 20	0.506	0.000	32 / 64
$\boldsymbol{s}_{is}$	15 / 20	0.621	0.084	32 / 64
$\boldsymbol{s}_{all}$	14 / 20	0.561	0.042	64 / 64

$\boldsymbol{s}_{\{does,is,all\}}$ 的准确率对于 $N^{img}_{correct}$ 为 10/20，这是因为门是透明的，导致系统错误地识别为门始终处于打开状态。另一方面，即使门是透明的，有些问题也能正确识别其状态，例如用于 $\boldsymbol{s}'_{-}$ 的以下两个问题："这张图像看起来像一扇关闭的窗户吗？"和"这扇窗户是打开的吗？"。其他优化结果得到的问题也表明，"窗户"这一措辞非常重要。

表 VII 透明门实验结果

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^{s}_{q}
$\boldsymbol{s}_{\{+,-\}}, \boldsymbol{s}'_{+}$	15 / 20	0.692	0.000	5 / 64
$\boldsymbol{s}'_{-}$	16 / 20	0.729	0.000	2 / 64
$\boldsymbol{s}_{does}$	10 / 20	0.574	0.000	32 / 64
$\boldsymbol{s}_{is}$	10 / 20	0.465	0.116	32 / 64
$\boldsymbol{s}_{all}$	10 / 20	0.519	0.058	64 / 64

H. 灯光实验

接下来，我们将说明，除了像目前为止所见的门的开/关状态之外，还可以识别各种其他状态。表 VIII 展示了房间灯光开关状态的识别结果。此处使用"灯"和"电力"作为措辞。优化的问题组合在准确率方面显著优于 $\boldsymbol{s}_{\{does,is,all\}}$ 。尽管 $R^{ave}_{correct}$ 较小，但$\boldsymbol{s}_{\{+,-\}}$ 在$N^{img}_{correct}$ 方面优于 $\boldsymbol{s}'_{\{+,-\}}$ 。这被认为是因为 $\boldsymbol{s}'_{\{+,-\}}$ 将无效响应视为错误，因此，具有许多无效响应但状态识别准确的问题会作为选项被淘汰。请注意，所有优化的问题组合中使用的词都是"电力"。其他特征与上述实验相同。

表 VIII 灯光实验结果

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^{s}_{q}
$\boldsymbol{s}_{\{+,-\}}$	20 / 20	0.799	0.094	3 / 32
$\boldsymbol{s}'_{\{+,-\}}$	17 / 20	0.856	0.000	3 / 32
$\boldsymbol{s}_{does}$	10 / 20	0.598	0.000	16 / 32
$\boldsymbol{s}_{is}$	10 / 20	0.573	0.069	16 / 32
$\boldsymbol{s}_{all}$	11 / 20	0.586	0.034	32 / 32

I. 显示器实验

表IX展示了显示屏开关状态的识别结果。这里使用了四个词作为措辞："显示屏"、"显示器"、"电脑显示器"和"电视"。优化的问题组合在准确率上显著优于 $\boldsymbol{s}_{\{does,is,all\}}$。 $\boldsymbol{s}_{-}$ 和 $\boldsymbol{s}'_{-}$ 完全相同，只使用了一个问题："这张图像看起来像是这个显示屏处于关闭状态吗？"另一方面，像 $\boldsymbol{s}_{+}$ 或$\boldsymbol{s}'_{+}$ 这样包含许多问题的组合，其准确率略高于 $\boldsymbol{s}_{-}$ 或 $\boldsymbol{s}'_{-}$ ，后者的 $N^{img}_{correct}$为20/20。

表IX 显示屏实验结果

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^{s}_{q}
$\boldsymbol{s}_{+}$	20 / 20	0.950	0.557	20 / 64
$\boldsymbol{s}'_{+}$	20 / 20	0.957	0.000	6 / 64
$\boldsymbol{s}_{-}, \boldsymbol{s}'_{-}$	19 / 20	0.925	0.000	1 / 64
$\boldsymbol{s}_{does}$	12 / 20	0.716	0.000	32 / 64
$\boldsymbol{s}_{is}$	14 / 20	0.693	0.491	32 / 64
$\boldsymbol{s}_{all}$	14 / 20	0.705	0.245	64 / 64

J. 手提包实验

表X显示了手提包开/关状态的识别结果。此处，"包"和"手提包"均作为措辞使用。与$\boldsymbol{s}_{\{does,is,all\}}$相比，优化后的问题组合准确得多。其他特征与之前的实验相同。

表X 手提包实验结果

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^{s}_{q}
$\boldsymbol{s}_{+}$	18 / 20	0.815	0.497	8 / 32
$\boldsymbol{s}_{-}$	18 / 20	0.815	0.195	5 / 32
$\boldsymbol{s}'_{\{+,-\}}$	18 / 20	0.750	0.000	3 / 32
$\boldsymbol{s}_{does}$	15 / 20	0.670	0.002	16 / 32
$\boldsymbol{s}_{is}$	11 / 20	0.530	0.461	16 / 32
$\boldsymbol{s}_{all}$	14 / 20	0.622	0.232	32 / 32

K. 水实验

表 XI 展示了水是否正在流动的识别结果。此处，在"水正在流动"之后添加了两种修改："在一个、该、这个、那个水槽中"或"从一个、该、这个、那个水龙头流出"。优化的问题组合在准确率上显著优于 $\boldsymbol{s}_{\{does,is,all\}}$ 。与第 III-H 节类似，$\boldsymbol{s}_{\{+,-\}}$ 在 $N^{img}_{correct}$方面优于 $\boldsymbol{s}'_{\{+,-\}}$ ，这表明包含大量无效响应的问题十分重要。请注意，对于优化中使用的所有数据， $N^{img}_{correct}$ 均为 19/20，这表明测试数据集上的准确率略低。此外，优化结果中使用的所有问题都采用了"从一个、该、这个、那个水龙头流出"这一修改。

表 XI 水实验结果

	$N^{img}_{correct}$	$R^{ave}_{correct}$	$R_{invalid}$	N^{s}_{q}
$\boldsymbol{s}_{+}$	17 / 20	0.797	0.179	6 / 32
$\boldsymbol{s}_{-}$	17 / 20	0.796	0.033	5 / 32
$\boldsymbol{s}'_{\{+,-\}}$	15 / 20	0.735	0.033	5 / 32
$\boldsymbol{s}_{does}$	12 / 20	0.580	0.000	16 / 32
$\boldsymbol{s}_{is}$	13 / 20	0.667	0.114	16 / 32
$\boldsymbol{s}_{all}$	12 / 20	0.622	0.057	32 / 32

IV. 讨论

以下是对所得实验结果中重要内容的总结。

- 同时使用许多问题（例如$\boldsymbol{s}_{\{does,is,all\}}$）的准确率并不高，因此有必要只使用合适的问题。
- 状态识别越容易，即使只有一个问题，准确率也越高。
- 在许多情况下，仅靠一个问题的准确性不足；通过适当组合多个问题，可以构建准确率更高的识别器。
- 可以将无效视为错误，以减少无效回答的数量，但这不一定会提高准确率。相反，一些重要问题往往会被排除在选项之外，准确率也常常因此下降。
- 对于某些状态识别，即使使用与优化时不同的数据集，问题数量较多也可能比问题数量较少实现更高的准确率。

综合考虑上述所有因素，我们基本上应使用通过$\boldsymbol{s}_{\{+,-\}}$获得的问题。即使耗时较长，为实现准确识别，也建议使用 $\boldsymbol{s}_{+}$；若希望在较短时间内实现准确识别，则建议使用$\boldsymbol{s}_{-}$。

其他一些值得注意的特性包括：

- 与 $\boldsymbol{s}_{does}$中的问题相比， $\boldsymbol{s}_{is}$中的问题会产生更多无效响应。
- 与 $\boldsymbol{s}_{does}$中的问题相比， $\boldsymbol{s}_{is}$中的问题具有更高的识别准确率。
- 通过优化获得的所有问题都采用相同的措辞。

这些特性被认为会因所使用的视觉问答模型而异，并非普遍适用的知识。然而，借助本研究，只需提供一些数据集，就能获得合适的问题组合，而无需担心这些因模型而异的特性。

本研究的重要贡献在于，无需重新训练网络或进行单独编程，仅凭口语就能直观地构建识别器。此外，只需一个预训练模型，资源也非常易于管理。另一方面，由于问题句的选择依赖启发式方法，我们加入了优化机制来解决这一问题。这样就能非常轻松地构建状态识别器，而人们唯一需要考虑的就是增加问题的多样性。未来，有必要将本研究应用于实际任务，并找出可能在现实世界中出现的问题。尤其是，我们希望研究在更加杂乱的场景中的状态识别，以及图像中出现人类、机器人本体等其他各种物体时的识别性能。此外，我们认为，异常检测和目标检测也将成为可能与本研究采用相同的方式，并且我们希望在不久的将来处理范围更广的状态识别问题。

五、结论

本研究提出了一种利用预训练视觉语言模型（PTVLM）中的视觉问答（VQA）实现机器人二元状态识别的方法。在视觉问答中，对于多张增强图像，我们整合了针对各种问题的答案，这些问题在句式、冠词、状态表达和措辞方面各不相同。对于每个问题，有些状态容易识别，有些状态则难以识别；因此，通过采用适当的问题组合，可以构建能够准确识别任意状态的识别器。通过遗传算法优化这一组合，现在可以识别此前难以识别的透明门、水等状态。我们认为，该方法将革新机器人的识别策略，因为它无需重新训练网络或编写程序，只需向单个模型提供一组问题，即可轻松构建能够识别复杂状态的识别器。

参考文献

[1] M. Saito、H. Chen、K. Okada、M. Inaba、L. Kunze 和 M. Beetz，“大规模室内环境中的语义目标搜索”，载于*IEEE/RSJ*智能机器人与系统国际会议，现实世界中的主动语义感知与目标搜索研讨会，2011年。[2] K. Okada、M. Kojima、Y. Sagawa、T. Ichino、K. Sato 和 M. Inaba，“面向日常环境任务的人形机器人视觉行为验证系统”，载于*2006*年*IEEE-RAS*人形机器人国际会议论文集，2006年，第7–12页。[3] R. T. Chin 和 C. R. Dyer，“机器人视觉中的基于模型识别”，*ACM*计算调查，第18卷，第1期，第67–108页，1986年。[4] S. M. Z. Borgsen、M. Sch¨opfer、L. Ziegler 和 S. Wachsmuth，“基于三维传感器的自动门检测”，载于*2014*年加拿大计算机与机器人视觉会议论文集，2014年，第276–282页。[5] X. Li、M. Tian、S. Kong、L. Wu 和 J. Yu，“面向基于视觉的水面垃圾捕捞机器人的改进YOLOv3检测方法”，《先进机器人系统国际期刊》，第17卷，第3期，2020年。[6] T. Takahata，“采用同心硅玻璃混合透镜的热成像与彩色成像同轴成像系统的同轴度评估”，《传感器》，第20卷，2020年。[7] M. P. Khaing 和 M. Masayuki，“基于卷积神经网络的透明物体检测”，载于大数据分析与深度学习应用国际会议论文集，2018年。[8] S. Antol、A. Agrawal、J. Lu、M. Mitchell、D. Batra、C. L. Zitnick 和 D. Parikh，“VQA：视觉问答”，载于*2015*年*IEEE/CVF*计算机视觉国际会议论文集，2015年，第2425–2433页。[9] A. Radford、J. W. Kim、C. Hallacy、A. Ramesh、G. Goh、S. Agarwal、G. Sastry、A. Askell、P. Mishkin、J. Clark、G. Krueger 和 I. Sutskever，“通过自然语言监督学习可迁移的视觉模型”，arXiv预印本 arXiv:2103.00020，2021年。[10] F. Li、H. Zhang、Y. Zhang、S. Liu、J. Guo、L. M. Ni、P. Zhang 和 L. Zhang，“视觉语言智能：任务、表征学习与大型模型”，arXiv预印本 arXiv:2203.01922，2022年。[11] P. Wang、A. Yang、R. Men、J. Lin、S. Bai、Z. Li、J. Ma、C. Zhou、J. Zhou 和 H. Yang，“OFA：通过简单的序列到序列学习框架统一架构、任务与模态”，arXiv预印本 arXiv:2202.03052，2022年。[12] F. Fortin、F. D. Rainville、M. Gardner、M. Parizeau 和 C. Gagn´e，“DEAP：轻松实现进化算法”，《机器学习研究期刊》，第13卷，第2171–2175页，2012年。