TidyBot: personalized robot assistance with large language models

汪俊扬 12111028　邓皓文 12110510 毛新科11910412
季亦冰12110501　周靖东 1211102
Ng Wooi Cheng 12111128　Daniel Tan Sioa Hen 12111127

2024.4.7

Group 4

Motivation and Main Problem

Explores a method for personalized household robotic assistance utilizing large language models (LLMs)

Long-standing Goal

Robotics aims to develop personalized household assistance robots, focusing on tidying rooms by moving objects to proper places, reflecting a persistent aspiration in robotics research.

Challenges in Personalization:

Determining proper receptacles for objects poses a challenge due to diverse individual preferences and cultural norms, necessitating personalized solutions beyond generic rules.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Motivation and Main Problem

Explores a method for personalized household robotic assistance utilizing large language models (LLMs)

Existing Approaches

Prior methods either rely on tedious user specifications or generic rules derived from aggregated data, lacking efficiency and personalization required for autonomous tasks.

Proposed Solution

Leveraging large language models (LLMs), the paper suggests summarizing user-provided examples into generalized rules, aiming for efficient personalization without extensive datasets.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Motivation and Main Problem

Explores a method for personalized household robotic assistance utilizing large language models (LLMs)

Implementation and Evaluation:

The proposed approach is implemented in the TidyBot system, enabling users to provide example placements, which are then summarized by an LLM and utilized for object manipulation, achieving high accuracy in real-world scenarios.

Contributions and Validation:

Contributions include the utilization of LLMs for robotic generalization, a benchmark dataset, and real-world system implementation. Validation through quantitative evaluations confirms the effectiveness of the proposed approach in personalized household assistance.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Problem Setting

The paper addresses the task of 6D object pose estimation in cluttered scenes, where the goal is to estimate the 3D translation and rotation of known objects in a scene. This involves localizing the object's center in the image and predicting its distance from the camera to estimate translation, and regressing convolutional features to a quaternion representation to estimate rotation. The challenge arises due to the variety of objects and complexities caused by clutter and occlusions between objects.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Related Work & Limitations of Prior Work

1. Ask a person to specify a target location for every object

FROM: Rasch, R., Sprute, D., Pörtner, A., Battermann, S., & König, M. (2019).Tidy up my room: Multi-agent cooperation for service tasks in smart environments. Journal of Ambient Intelligence and Smart Environments, 11(3), 261–275.

2. Learn generic (non-personalized) rules about where objects typically go inside a house by averaging over many users

FROM: Kant, Y., Ramachandran, A., Yenamandra, S., Gilitschenski, I., Batra,D., Szot, A., & Agrawal, H. (2022). Housekeep: Tidying vir_x0002_tual households using commonsense reasoning. arXiv preprint arXiv:2205.10712

3.Works that focus on personalization aim to extrapolate from a few userexamples given similar choices made by other users, using methods such as collaborative filtering

FROM:Abdo, N., Stachniss, C., Spinello, L., & Burgard, W. (2015). Robot,organize my shelves! tidying up objects by predicting user pref_x0002_erences. In 2015 IEEE international conference on robotics and automation (ICRA).

limitation: all of these approaches require collecting large datasets with user preferences or generating datasets from manually constructed, simulated scenarios. Such datasets can be expensive to acquire and may not generalize well if they are too small.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Theory

LLM (Large Language Model)

What are Large Language Models?

Large Language Models (LLM) are deep learning models that are designed to understand and generate human-like text based on vast amounts of data they have been trained on. These models use deep learning techniques, particularly transformer architectures, to process and generate text.

The underlying transformer is a set of neural networks that consist of an encoder and a decoder, which extract meanings from a sequence of text and understand the relationships between words and phrases in it, in another word, understand basic grammar.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

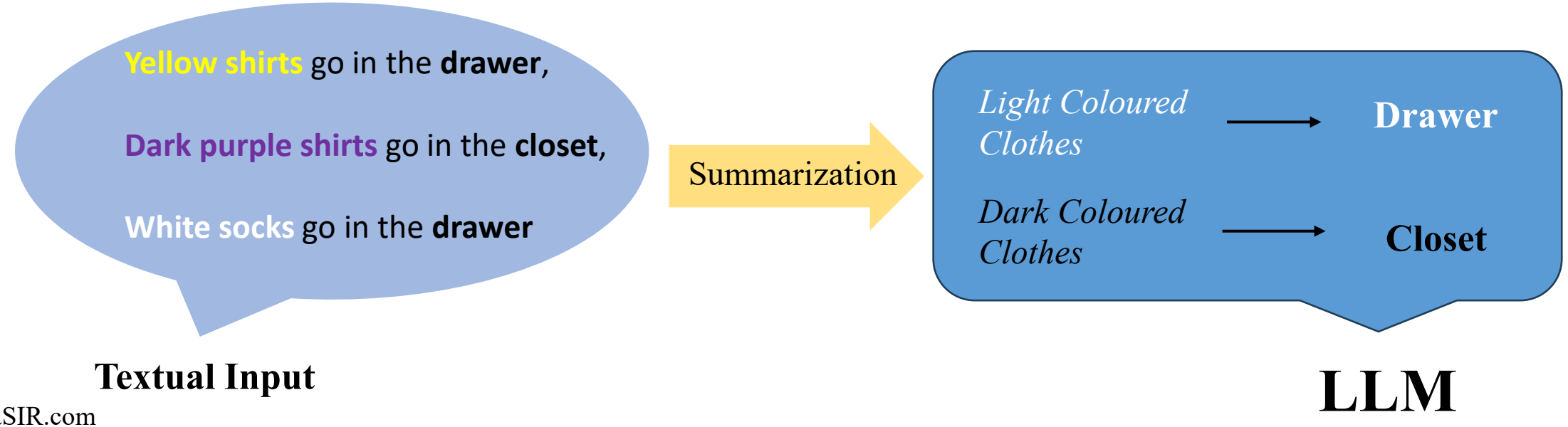
Theory

LLM (Large Language Model)

LLMs have a wide range of applications, including natural language understanding, text generation, translation, summarization, question answering, and more.

LLM's Summarization Capability.

In this approach, the summarization capability of LLM has been highlighted as the generalization of example preferences can be provided by using the capability.



AncoraSIR.com

Theory

LLM (Large Language Model)

Users provide examples, which LLM summarizes into personalized rules (mapping object categories to receptacles).

Robot executes cleanup task by identifying objects and moving them to target receptacles based on rules.

Robot carry out the cleanup task by repeatedly picking up objects, identifying them, and moving them to their target receptacles

AncoraSIR.com

Proposed Approach / Algorithm / Method

The approach is to utilize the summarization capabilitiesnof large language models (LLMs) to provide generalizationfrom a small number of example preferences.

By using the summarization provided by LLMs for generalization in robotics, it can produce generalized rules from a small number of examples, in a form that is human interpretable (text) and is expressed in nouns that can be grounded in images using open-vocabulary image classifiers.

The authors want to set a publicly released benchmark dataset for evaluating generalization of receptacle selection preferences

AncoraSIR.com

SUSTech
Southern University of Science and Technology

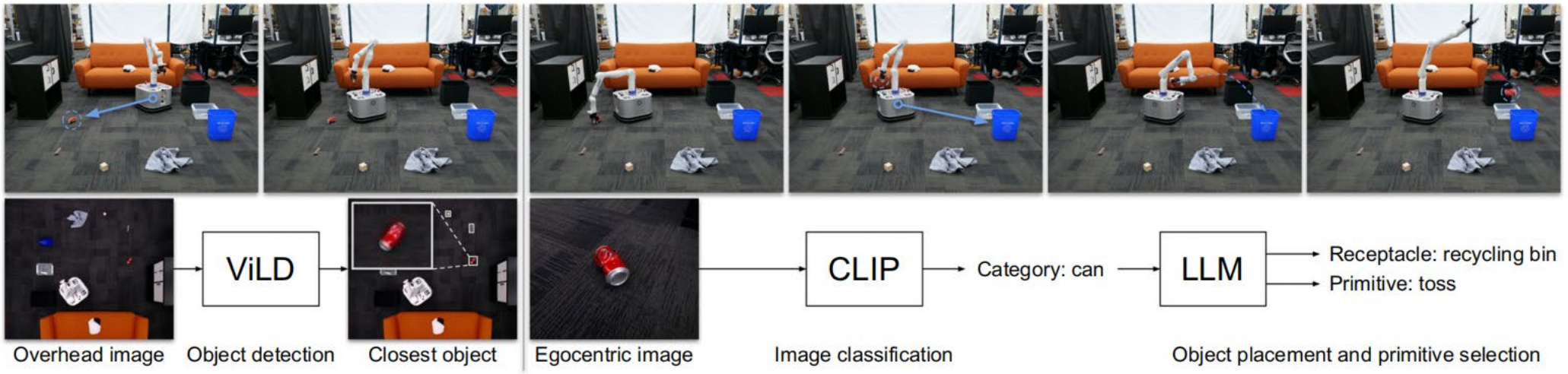
Proposed Approach / Algorithm / Method

1.Personalized receptacle selection

2.Personalized primitive selection

3.Real-world robotic system

```
objects = ["yellow shirt", "dark purple
shirt", "white socks", "black shirt"]
receptacles = ["drawer", "closet"]
pick_and_place("yellow shirt", "drawer")
pick_and_place("dark purple shirt", "closet")
pick_and_place("white socks", "drawer")
pick_and_place("black shirt", "closet")
# Summary: Put light-colored clothes in the
drawer and dark-colored clothes in the closet.
```



AncoraSIR.com

Experimental Results

1.Benchmark dataset

Success on this benchmark is measured by the object placement accuracy: the number of objects placed in the correct receptacle divided by the total number of objects. Evaluating accuracy separately for seen and unseen objects, to tease apart memorization versus generalization

Category	Attribute	Function	Subcategory	Multiple
86/96	27/96	24/96	31/96	17/96

2. Baseline comparisons

Method	Accuracy (unseen) (%)
Examples only	78.5
WordNet taxonomy	67.5
RoBERTa embeddings	77.8
CLIP embeddings	83.7
Summarization (ours)	**91.2**

Method	Category (%)	Attribute (%)	Function (%)	Subcategory (%)	Multiple (%)
Examples only	80.1	72.7	75.7	77.0	81.5
WordNet taxonomy	69.1	59.8	61.4	71.3	74.1
RoBERTa embeddings	78.6	75.5	71.8	71.7	87.5
CLIP embeddings	84.6	79.8	85.5	84.7	87.9
Summarization (ours)	**91.0**	**85.6**	**93.9**	**90.1**	**93.5**

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Experimental Results

3. Ablation studies

The goal of these experiments is to compare its performance to alternatives with far less information (using only common sense, without preferences) or far more information (using human-generated summarizations).

Model	Commonsense		Summarization	
	Seen (%)	Unseen (%)	Seen (%)	Unseen (%)
text-davinci-003	45.0	45.6	**91.8**	**91.2**
text-davinci-002	41.8	37.5	84.1	75.7
code-davinci-002	41.4	39.4	88.6	83.2
PaLM 540B	**45.5**	**49.6**	84.6	75.7

4. Human evaluation

1. **Evaluate** whether humans prefer the object placements generated by our LLM summarization method over those of the CLIP embeddings baseline.
2. **Evaluate** whether human-preferred object placements align with the ground truth placements in our benchmark.

Method	Category (%)	Attribute (%)	Function (%)	Subcategory (%)	Multiple (%)	Overall (%)
CLIP embeddings	19.7	23.7	11.2	22.6	21.2	19.1
Summarization (ours)	**47.4**	**41.9**	**60.0**	**46.1**	**40.6**	**46.9**
Equally preferred	32.9	34.4	28.8	31.3	38.2	34.1

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Experimental Results

5. Real-world experiments

The robot is placed inside a room with various objects and receptacles on the floor and is then tasked with picking up all the objects and putting them into the correct receptacles according to user preferences.

	CLIP (%)	ViLD (%)	OWL-ViT (%)
Summarized categories	**95.5**	76.1	45.9
Scenario object names	**70.7**	59.9	24.8
All object names	**52.3**	36.5	18.5

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Experimental Setup

Hardware and Environment

- Robot Platform: The experiments were conducted using a Franka Emika Panda robotic arm, which is commonly used in research for its precision and versatility.
- Environment: The real-world tests were conducted in an environment set up to mimic a typical household, complete with various commonly found items and furniture. This setup was designed to assess the robot's ability to navigate and manipulate objects in a space resembling its intended use case.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Experimental Setup

Software and Models

- Large Language Models (LLMs): The core of their experimental setup involved integrating Large Language Models for understanding user preferences and guiding the robot's actions. These models were trained to generate summaries of short descriptions, provided by users, about where objects should be placed.
- Perception and Planning: The robot utilized advanced perception algorithms to identify and locate objects in its environment. Planning algorithms then determined the most efficient sequence of actions to tidy up the space according to inferred user preferences.

Experimental Tasks

- Tidying Task: The primary task involved the robot picking up objects from a cluttered environment and placing them in their designated locations based on the user's preferences.
- Preference Learning: A significant aspect of the experiment was learning and generalizing user preferences to unseen objects. This involved the robot interpreting instructions or preferences expressed in natural language and applying these preferences to similar objects.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Experimental Setup

Data Collection and User Interaction

- User Preferences: To train and evaluate the model, the researchers collected data on user preferences for object placement. This involved short descriptions or instructions provided by users on where objects should be placed.
- Evaluation on Unseen Objects: A critical part of the experimental setup was evaluating the robot's performance on objects that were not part of the training dataset. This tested the robot's ability to generalize learned preferences to new situations.

AncoraSIR.com

Discussion of Results

1.The summarization capabilities of large language models (LLMs) can be used to generalize user preferences for personalized robotics.

2.Given a handful of example preferences for a particular person, LLM summarization can infer a generalized set of rulesto manipulate objects according to the user's preferences.

3.The summarization approach outperforms several strong baselines on our benchmark. And human responses were aligned with benchmark ground truth.

4.We also evaluate the approach on a real-world mobile manipulator called TidyBot, which can successfully clean up test scenarios with a success rate of 85.0%.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Critique / Limitations / Open Issues

Further explaination of the title with supporting evidence

LLM summarization

In instances where LLMs are tasked with summarizing preferences, there are occasional shortcomings observed. One prevalent issue occurs when the generated summary merely itemizes seen objects without effectively categorizing them. Such summaries tend to be overly specific and lack the ability to generalize to unseen objects.

Another common failure arises when the LLM aggregates receptacles into broad groups (e.g., grouping "top drawer" and "bottom drawer" as "drawers"), leading to subpar performance when utilizing the summary for receptacle selection.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Critique / Limitations / Open Issues

Further explaination of the title with supporting evidence

Real-world system

The real-world system implementation simplifies tasks with hand-written manipulation instructions, top-down grasping, and assuming known receptacle locations. Addressing these limitations involves integrating advanced manipulation techniques and enhancing perception capabilities.

Priority in excessively cluttered environments

Moreover, mobile robots' inability to traverse obstacles limits their functionality in cluttered environments. Incorporating advanced high-level planning could enable the robot to prioritize clearing paths over simply picking up the nearest object.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Future Work for Paper / Reading

Further explaination of the title with supporting evidence

Special case sorting plan:

In cluttered environments where mobility is challenging, Tidybot needs to first clear a passable path before proceeding with sorting according to regular rules.

Special Item Retrieval:

When facing items like books, improper grasping mode may cause significant deformation, leading to grip failure. It's necessary to enhance the adaptability of the grip and expand the range of graspable items to achieve optimal performance.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Extended Readings

Further explaination of the title with supporting evidence

Robot, organize my shelves! Tidying up objects by predicting user preferences

N. Abdo, C. Stachniss, Luciano Spinello, Wolfram Burgard

Published in IEEE International Conference. 26 May 2015. Computer Science

This paper predicts pairwise object preferences of the user, and subdivides the objects in containers by modeling a spectral clustering problem, which is easy to update, does not require complex modeling, and improves with the amount of user data.

AncoraSIR.com

Extended Readings

Further explaination of the title with supporting evidence

Rearrangement: A Challenge for Embodied AI

Dhruv Batra, Angel X. Chang, S. Chernova, A. Davison, Jia Deng, V. Koltun, S. Levine, J. Malik, Igor Mordatch, Roozbeh Mottaghi, M. Savva, Hao Su

Published in arXiv.org. 3 November 2020. Computer Science

This paper predicts pairwise object preferences of the user, and subdivides the objects in containers by modeling a spectral clustering problem, which is easy to update, does not require complex modeling, and improves with the amount of user data.

AncoraSIR.com

Summary

Further explaination of the title with supporting evidence

- Problem the reading is discussing

  Personalized robot assistance with larg e language models

- Why is it important and hard

  Make robots smarter; Personalization and accuracy.

- What is the key limitation of prior work

  LLM summarization, advanced manipulation techniques and perception capabilities

- What is the key insight of the proposed work

  Combine language based planning and perception with the few-shot summarization capabilities of large language models

- What did they demonstrate by this insight?

  This approach enables fast adaptation and achieves 91.2%accuracy on unseen objects in our benchmark dataset.

AncoraSIR.com

SUSTech
Southern University of Science and Technology

Q&A

Team4