In: *Proceedings of the 31st International Conference on Artificial Neural Networks (ICANN 2022)*, LNCS, Vol 13530, pp. 258-269, Bristol, UK, Sep 6–9, 2022

Learning Visually Grounded Human-Robot Dialog in a Hybrid Neural Architecture

Xiaowen Sun, Cornelius Weber, Matthias Kerzel, Tom Weber, Mengdi Li, and Stefan Wermter

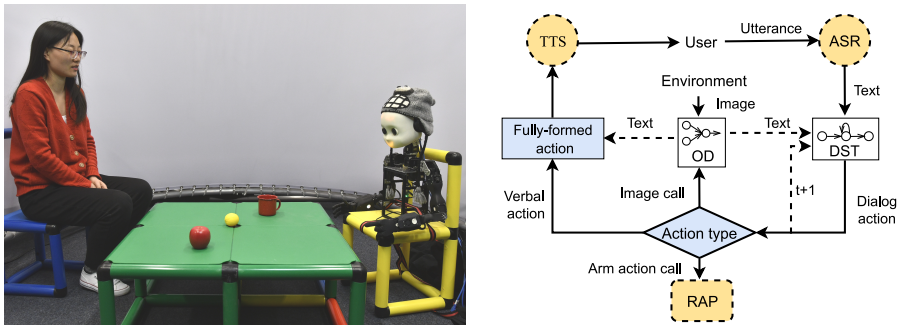
Knowledge Technology, Department of Informatics, University of Hamburg
{xiaowen.sun, cornelius.weber, matthias.kerzel, tom.weber, stefan.wermter}@uni-hamburg.de, mli@informatik.uni-hamburg.de
www.knowledge-technology.info

Abstract. Conducting a dialog in human-robot interaction (HRI) involves complexities that are hard to reconcile by individual research or engineering works. Towards the development of a robotic dialog agent, we develop a verbal and visual instruction scenario in which a robot needs to enter into a dialog to resolve ambiguities. We propose a novel hybrid neural architecture to learn the robotic part of the interaction. A neural dialog state tracker learns to process the user input depending on visual inputs and dialog instances. It uses variables to allow certain generality to generate the robot's physical or verbal actions. We train it on a new visual dialog dataset, test different forms of input representations, and validate the robot agent on unseen examples. We evaluate our hybrid neural network approach in handling an HRI conversation scenario that is extendable to a real robot. Furthermore, we demonstrate that the hybrid approach allows generalization to a large range of unseen visual inputs and verbal instructions.

Keywords: Human-robot interaction · Visual dialog generation · Natural language processing · Computer vision · Recurrent neural networks.

1 Introduction

Human-robot interaction (HRI) utilizing dialog is a challenge for neural network research. Open-domain dialog agents produce dialog actions of reasonable quality, but they do not pursue any specific goal [27]. In contrast, for real-world applications, dialog is usually domain-specific and goal-oriented. Task-oriented dialog systems show good performance in specific tasks helping in our daily lives, such as for making a restaurant reservation, receiving orders, getting technical support, or giving smart-home commands [2]. Beyond these language-based systems, visually grounded human-robot interaction is usually situated in an environment that can be perceived visually and with other sensors, and in which a robot can interact, or execute commands. Verbal interaction usually involves multi-turn dialog. Due to ambiguities in language or complex scenarios, if one person wants to instruct another person to perform a certain task, this often requires multiple turns to unambiguously determine the goal. Conducting a dialog



(a) Conversational scenario. (b) Hybrid neural architecture.

Fig. 1: (a) Visually grounded human-robot conversational scenario. The user is talking to the NICO robot about objects on the table. (b) Hybrid neural architecture. White boxes: Object Detection (OD) and Dialog State Tracker (DST) are both trained components. Yellow boxes: Automatic Speech Recognition (ASR), Text to Speech (TTS) and Robotic Arm Planner (RAP) are pre-trained components. Blue boxes represent Human-Robot Interaction Policy (HRIP). The dashed arrows denote conditional input.

that requires disambiguation remains a scientific challenge in visually grounded human-robot interaction.

The development of visually grounded dialog requires suitable multimodal datasets. Related areas like Visual Question Answering [10], Visual Dialog [6], Visual Dialog Generation [7], Image-Grounded Conversations [25], and CLEVR-dialog [18] are mainly focussed on understanding the image information, not the interaction within an environment. A seller-buyer interaction in a virtual shop is covered in the Situated and Interactive MultiModal Conversations (SIMMC) dataset [24], where the focus is however on fashion and furniture, but not HRI.

To address this gap, we designed the conversational scenario shown in Figure 1a, presenting a new task and dataset in the area of robotic manipulation research. The robot, Neuro-Inspired COmpanion (NICO) [16], sits at a table on which there are some common household objects. The user uses natural language to instruct NICO to pick or point to one of those objects. We assume there are always three different objects on the table and the robot cannot point to two objects at once. Sometimes, the user's command may be ambiguous. For example, the user instruction could contain two targets (e.g. `Show me the lemon and apple`). NICO can understand that this is an ambiguous instruction and give feedback. Another situation is that the user gives an unambiguous instruction (e.g. `Point to the red object`), but multiple objects have the same color. In such situations, the robot needs to use visual information to understand the ambiguity of the command and request the user for additional input. Once, the user and NICO reach a consensus, NICO will execute the appropriate action.

To solve this task, our agent needs to recognize the objects and relate the visual information to the objects' names, colors, and positions as communicated by language. Our contributions in the HRI domain are: 1) we propose a new task of

intermediate complexity for visually grounded human-robot conversation, which deals with the ambiguity between human instructions and the environment. 2) We provide a multimodal dataset including images and dialog instances related to the images. 3) We propose a new architecture, a hybrid neural model that tracks the dialog state for the visually grounded human-robot conversation and evaluates the model's performance.

2 Background and related work

The Situated and Interactive Multimodal Conversations (SIMMC) incorporates multimodal inputs (e.g., vision, memories of previous interactions, and users' utterances) for multimodal actions (e.g., representing the search results while generating the agent's next utterance) [24]. The next generation of SIMMC 2.0 is still focussed on a shopping scenario [17], which is devised of four main benchmark tasks: Multimodal Disambiguation, Multimodal Coreference Resolution, Multimodal Dialog State Tracking, and Response Generation [17]. The primary task of SIMMC is dialog state tracking. To solve this task, many studies focus on transformer architecture, such as Bidirectional Encoder Representations from Transformers (BERT) [8], Bidirectional Auto Regressive Transformers (BART) [19] and Generative Pre-trained Transformer (GPT) [29] to solve this task recently [13, 14, 28].

Visual Question Answering (VQA) [10] and Visual Dialog (VisDial) [6] are used for common-sense learning of visual-language representations, which are both based on Microsoft Common Objects datasets [21]. In contrast, GuessWhat? [7] is a two-player guessing game, which aims to find an unknown object in a rich image scene by question-answering strategies based on reinforcement learning. The CLEVR-dialog [18] dataset focuses on multi-round reasoning learning in visual dialog, which constructs a dialog grammar that is grounded in the scene graphs of the images from the CLEVR dataset. Lu et al. [22] present Vision-and-Language BERT (VilBERT) which extends BERT [8] to process visual and linguistic input. Murahari et al. [26] pretrained the VilBERT on the VQA [10] dataset and fine-tuned it on the VisDial dataset, then created the VisDial-BERT for multi-turn visually grounded conversations. The augmented extended train robots dataset [15], which expands the extended train robots dataset [1], offers tasks for a robotic agent to reach for objects in three-dimensional space based on augmented reality and a simulation environment. However, in all above approaches, no dialog studies focus on the visually grounded human-robot interaction domain.

A traditional dialog pipeline usually contains natural language understanding, a dialog manager, and natural language generation. The core part of a dialog system is the dialog manager, including a dialog state tracker and dialog policy [4, 27]. Recurrent neural networks (RNNs) are usually trained in an end-to-end fashion to match an observable dialog history to output sentences [11]. A hybrid approach is also attractive for task-oriented dialog modelling, since it can combinine multiple approaches, such as rule-based and data-driven [9].

Table 1: List of instruction utterances. The parts left and right of the slash can be substituted for each other. An ambiguous instruction refers to multiple objects. An unambiguous instruction refers to one specific object.
Show me: SM, Where is: WI, Point to: PT, Where are: WA.

Unambiguous instructions	`SM/ WI/ PT the [name].`
	`SM/ WI/ PT the [color] object/ [name].`
	`SM/ PT the [position] object/ [name].`
	`SM/ PT the [color] object/ [name] on the [position].`
Ambiguous instructions	`SM/ PT the object on the table.`
	`SM/ WA the [color1] and [color2] objects.`
	`SM/ WA the [name1] and [name2].`

Hybrid-code networks combine an RNN with domain-specific knowledge, which perform well on the bAbI dataset [3], and are applied to a real customer support domain [31].

Inspired by the lack of robotic visually grounded datasets, we generated an artificial multimodal dataset for our HRI domain, which mainly focuses on human use of language, including naturally occurring ambiguities, to instruct the humanoid robot to point to an object in the environment. Learning from the principle of the Hybrid-code networks, we train an RNN for dialog state tracking and define an HRI policy for our scenario.

3 Multimodal dataset for human-robot interaction

We propose a dataset consisting of two modalities, visual scenes and conversations in text form. The visual scenes were generated with Blender[1] and CoppeliaSim[2]. The user and NICO use language to talk about objects' characteristics in the scene, such as an object's position, color, and name.

3.1 Human-robot conversation task definition

The task is set in the context of robot manipulation with human instruction, which requires understanding user utterances, using symbols and recognizing objects, and using the acquired knowledge. The subtasks are:

Subtask 1: Opening greetings. The greeting is the start of the conversation.

Subtask 2: Receiving user requests. We assume that the user request is always related to the objects on the table. Nevertheless, there are still two types of ambiguities: ambiguous instructions from user utterances and ambiguous scenes that

[1] https://www.blender.org/
[2] https://www.coppeliarobotics.com/

Table 2: List of dialog actions.

Tasks	No.	Content
Subtask 1	$a_1 \sim a_4$	`[Greeting], what can I do for you.`
Subtask 2	a_5, (UOI)	`Please give me a specific target instruction.`
	a_6, (AIR)	`I cannot point to multiple things at once.`
	a_7, (UIR)	`Ok, let me check.`
	a_8, (IC)	`image_call`
Subtask 3	a_9, (NC)	`Do you mean the [name]?`
	a_{10}, (PC)	`Do you mean the one on the [position]?`
	a_{11}, (OI)	`Ok, let me show you the [name].`
	a_{12}, (SAR)	`There is more than one object you ask for.`
	a_{13}, (VRF)	`I cannot see anything.`
Subtask 4	a_{14}	`arm_action_call`
	a_{15}	`Am I wrong, or do you want to change your mind?`
	a_{16}	`It is fine, You can try again by saying hello.`
	a_{17}	`Sorry, I am still learning.`
	a_{18}	`Here it is.`
	a_{19}	`Do you want to try again, start by saying hello.`
	a_{20}	`You are welcome, See you.`

contain multiple objects with the same characteristic (e.g. color). In this subtask, the agent clarifies ambiguous instructions from the user. If an instruction does not contain enough information to identify any object on the table.NICO will, therefore, ask the user to specify and include more detailed information, which we call Unspecified Object Instruction (UOI, a_5). If the user's instructions contain multiple names or colors, NICO will announce his inability to point to multiple objects at once. We term this action Ambiguous Instruction Recognition (AIR, a_6). When the robot receives unambiguous instructions, it will confirm the valid instruction. We term this action Unambiguous Instruction Recognition (UIR, a_7). After this, NICO will detect the relevant objects using the camera command Image Call (IC, a_8). Table 1 shows all ambiguous and unambiguous instructions in this subtask.

Subtask 3: Confirming user requests. Upon receiving results from OD, the robot confirms the user's request. The main challenge of this subtask is matching an unambiguous instruction with the environment, so the model can find a corresponding dialog action (refer to Table 2 for a list of all dialog actions). NICO tackles this matching problem with one of the following approaches. If the user asks for the position, he replies with the name. We term this action Name Confirmation (NC, a_9). If the user's instructions contain a name and/or a color, NICO replies with the position. We term this action Position Confirmation (PC, a_{10}). If the user requests an object by its color, name, and position, NICO replies with the name. We term this action Object Identification (OI, a_{11}). If the user's request contains ambiguous information in relation to the environment, so that the matching task cannot be fulfilled, the robot should recognize this ambigu-

ity. An action we term Scene Ambiguity Recognition (SAR, a_{12}). The robot can also automatically detect if an image capture failure within the camera occurred, henceforth called Visual Recognition Failure (VRF, a_{13}).

Subtask 4: Issuing *Arm_action_calls*. After getting the confirmed information, the user either gives a negative or an affirmative answer. For the negative response, the robot will ask for the reason and guide the user to restart the conversation. When receiving an affirmative response, the robot will call for the arm motor to execute the specific action.

3.2 Visual scenes

We selected a subset of objects from the Yale-CMU-Berkeley (YCB) objects [5], representing objects that often occur in everyday situations.We use 28 objects each for both, the training and test set, where each set contains 9 colors (red (8 objects), blue (3 objects), yellow (5 objects), brown (2 objects), green (1 objects), orange (3 objects), black (2 objects), white (3 objects), purple (1 objects)). We selected three different objects out of 28 objects for every scene. Following the combination formula, we generate 3276 scenes for every set. An ambiguous scene means objects have the same color. An unambiguous scene means the objects have three different colors. For each set there are 1136 ambiguous scenes and 2140 unambiguous scenes. The test and training set contain the same number of objects and the same color balance to let the percentage of ambiguous scenes be the same (34.7%).

3.3 Conversations

Based on the task that we define in Section 3.1 and the visual scenes, we generate a dialog instance dataset, which is inspired by Bordes and Weston's work [3, 30]. We use 15 unambiguous instruction utterances (see Table 1) and create 15 templates for every visual scene. Every template also includes some ambiguous instructions. Overall, we have 147420 dialog instances for both, the training and test set. For the training set, there are 117936 dialog instances for training and 29484 dialog instances for validation.

4 Approach

An overview of our hybrid neural architecture is shown in Fig. 1b. Its six components are Automatic Speech Recognition (ASR), Text-to-Speech (TTS), Object Detection (OD), Robotic Arm Planner (RAP), Human-robot interaction policy (HRIP), and Dialog State Tracker (DST). The cycle begins when the user starts with a greeting. The very first action of the architecture is to process the user input with its ASR. The ASR result is then fed into the DST, which classifies the dialog action (Table 2 states all the dialog actions used in this study). In addition to the user's utterance, the DST potentially has two other input modalities,

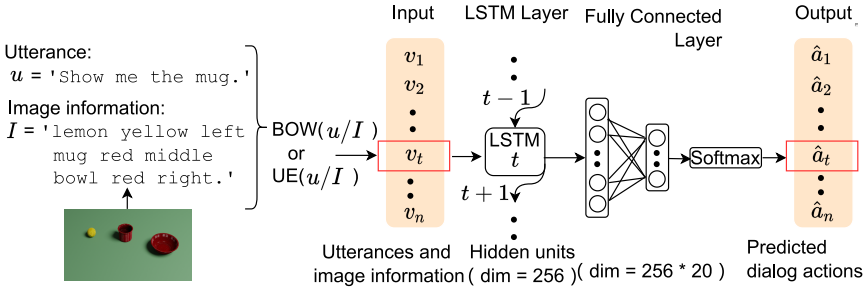


Fig. 2: Dialog State Tracker (DST). We use bag of words (BOW) or utterance embeddings (UE) to represent utterances and image information. These are fed as input v_t at time step t into the network. The output $\hat{a}_t$ at time step t denote the predicted dialog action. n is the number of the conversational turns of the person as well as of the agent.

depending on the situation. One is from the OD containing the scene's visual information. The other is its previous dialog action. Based on the classified action, the HRIP determines NICO's behavior. The behavior entails detecting the object with its camera (OD), pointing to a target (RAP) or forming appropriate verbal responses and using TTS to generate a vocal response. The main contribution of this work is that the DST can handle the ambiguity in a user's language input and in the environment the input pertains to. The OD and DST are both neural models trained on our datasets. In this paper, the focus is mainly on OD, DST, and HRIP. Inspired by [31], we call our combination of learning-based DST and a code-based HRIP a hybrid-code dialog manager.

4.1 Object detection (OD)

Our object detection is realized with the single-stage object-detector RetinaNet [20]. The neural architecture backbone of RetinaNet is formed by a Feature Pyramid Network (FPN) and a connected deep residual network. In essence, it constructs a semantic feature map at different scales and thus compensates for the CNN's low resolution on high-level feature maps. The FPN regression and classification subnetworks perform the actual object detection. In this work, the RetinaNet is trained on a corpus of one thousand images like the one shown in Fig. 2. The images are generated using Blender and automatically annotated with bounding boxes. During the dialog, the OD is invoked by the robot and supplied with an image of the scene. The RetinaNet subsequently detects the objects and returns the object names, positions and colors. These attributes can then be used to formulate appropriate responses through the DST.

4.2 Dialog state tracker (DST)

Multimodal feature representation As shown in Fig. 1b, the whole visually grounded human-robot dialog is based on multimodal, audio-visual input,

being the user's utterance and an image of the environment. However, the DST receives pure text input (see Fig. 2, image I and utterances $U = [u_1, \ldots, u_n]$). The two modalities, therefore, have to be processed into text format, so that the DST can make use of it. Representing natural language can be done in manifold ways. We decided to use the bag of words (BOW) for its simplicity and utterance embeddings (UE) for its successful applications in other natural language tasks. In order to create a vector of inputs, we build a vocabulary dictionary for the training set. For the BOW, we create a vector from the sum of the individual one-hot vectors of this dictionary. For the UE, we employ a word2vec model [23] that has been pre-trained on the Google News corpus to obtain word vector representations, and then, we calculate the mean vector of these representations. These representations are applied to both, utterances and image descriptions. Moreover, each dialog instance d_j combines user utterances, image information and labeled dialog actions ($d_j = [(u_1, a_1), \ldots, (I, a_i), \ldots, (u_n, a_n)]$, i is the position of an image reading within a dialog instance, n is the number of conversational turns). There is a total of m dialog instances ($D = [d_1, \ldots, d_m]$).

LSTM + FCL Besides the user's utterances U and image I, which both are represented by BOW or UE, the previous action (PA) is an additional input to the DST. Every utterance u_i of the dialog instance, its image information I and its previous action a_{i-1} are concatenated to form a feature vector. They are fed into an RNN, specifically, a long short-term memory (LSTM) network [12]. The output of the LSTM is passed to a fully connected layer (FCL), after which the softmax function is applied. The output of the model are predicted dialog actions $(\hat{a}_1, \ldots, \hat{a}_n)$.

4.3 Human-robot interaction policy (HRIP)

A dialog manager usually contains a dialog state tracker and a dialog policy. Here, we train a neural network for the state tracking. Additionally, rules for knowledge extraction and decision-making are needed. When the DST recognizes that the user gives a specific instruction, HRIP extracts the user's target and matches it with the output of the OD. With the matched information, a decision will be made. Based on the predicted dialog action of the DST, rules determine the robot's behavior. Possible robot behavior includes calling for the camera to get image information, calling for the robot's arm to execute the specific action (e.g. point to the object), using the OD result to formulate correct verbal answers and generating speech to respond to the user.

5 Experiments

During training, each dialog instance constituted its own minibatch, and updates were computed on full rollouts (i.e., non-truncated backpropagation through time). Because it is a multi-class classification task, categorical cross-entropy (CCE) was used to calculate the error terms. We selected the AdaDelta optimizer

Table 3: Average test accuracy of labeled actions in subtask 2 and subtask 3. Unspecified Object Instruction (UOI), Ambiguous Instruction Recognition (AIR), Unambiguous Instruction Recognition (UIR), Image Call (IC), Name Confirmation (NC), Position Confirmation (PC), Object Identification (OI), Scene Ambiguity Recognition (SAR), Visual Recognition Failure (VRF).

Labeled	Average accuracy(%)±Standard deviation					
actions	BOW	(BOW, PA)	UE	(UE, PA)	(BOW, UE)	(BOW, UE, PA)
UOI	100±0	100±0	100±0	100±0	100±0	100±0
AIR	100±0	100±0	73.98±12.19	77.50±10.45	100±0	100±0
UIR	100±0	100±0	98.33±0.85	99.58±0.63	100±0	100±0
IC	100±0	100±0	98.72±1.94	99.91±0.26	100±0	100±0
NC	78.24±5.84	78.96±6.08	98.09±2.18	99.99±0.02	95.51±3.93	97.96±3.29
PC	100±0	99.94±0.18	97.76±1.54	99.48±1.52	100±0	96.64±9.44
OI	99.56±0.75	100±0	89.24±12.99	98.15±2.64	99.81±0.55	97.63±4.40
SAR	91.65±6.3	90.98±7.84	55.79±22.32	42.70±15.85	74.17±7.47	67.40±9.97
VRF	100±0	100±0	97.38±6.32	99.83±0.48	100±0	100±0

[32] to minimize the loss function. We evaluate six different variants of inputs to the RNN of the DST: bag of words only BOW, utterance embeddings UE only, bag of words and utterance embeddings (BOW, UE) together, and the previous action added to all those combinations. Each combination was trained 9 times with different sampling order for 30 epochs to reduce noise and avoid biases.

5.1 Results

Subtask 1 (*Opening greetings*) and subtask 4 (*Issuing Arm_action_calls*) are both essential parts of our visual grounded human-robot dialog. Since they are not the focus of our research, we define them in a simple fashion, using the same data in training and test sets. The DST successfully predicts the correct dialog action with an accuracy of 100% for all action classes belonging to these subtasks. Table 3 only shows the results for subtask 2 (*Issuing user requests*) and subtask 3 (*confirming user requests*). Mean and standard deviation of the accuracy are computed from 9 runs. The challenge of subtask 2 is to deal with an ambiguous instruction from the user, and the challenge of subtask 3 is to correctly combine the user's instruction and image input.

Action AIR responds to a type of ambiguous instructions (see table 1). Compared with variants BOW and (BOW, UE) that can predict the dialog action with 100% of accuracy, variants UE cannot properly react to a user's unspecific instruction, achieving only accuracy of 73.98%. Action SAR responds to an ambiguous scene. The variant BOW performs better than UE and (BOW, UE) in situations when SAR is the expected dialog action. In unambiguous situations when action NC is expected, utterance embedding cannot help the model handle ambiguous situations, but it helps the model perform well in unambiguous

situations. Comparing variant BOW with (BOW, UE), a striking difference is that the addition of UE helps in NC but has a negative effect in SAR.

5.2 Discussion

The results indicate that using BOW to represent the inputs for the DST model can be a worthwhile option to tackle visually grounded human-robot dialog state tracking. For the BOW, we found that one problem is the multiple meanings of the word *orange*, which can be either a color or an object name, but is represented the same via BOW. For the UE in subtask 2, DST may predict a wrong dialog action, e.g. for *Show me the [name1] and [name2]*, it incorrectly predicts UIR. The reason is that the UE uses word2vec that does not have a representation of stop words like *and*. Furthermore, the pretraining of the UE on a non-related corpus could explain some issues. For the action SAR in subtask 3, the predicted dialog action is NC, which means the model misunderstood the user's instruction for a position. These limitations of our model are related to the simple utterance representations, which might be remedied in future work by more sophisticated sentence embeddings.

6 Conclusion and future work

Integrating visual information into a dialog is an essential necessity for robots to interact and cooperate with humans naturally. To this end, we propose a visually grounded human-robot dialog task along with a dataset. Moreover, we designed a novel hybrid neural architecture to solve this task, entailing a neural model to track the dialog state and a knowledge-based policy for the robot behavior. The hybrid nature of the architecture allows integrating state-of-the-art neural modules for vision and language processing with symbolic reasoning mechanisms. We explored how to represent user utterance and visual scene inputs to let the dialog model learn interaction skills. The results show that the simple bag of words (BOW) method can solve this task better than utterance embeddings (UE) based on a pre-trained word2vec model. Notably, the model generalizes to objects in the test set that were never shown in training, indicating that the model can generalize to any unseen object, provided the OD can recognize it. Moreover, although the number of dialog actions are fixed, they are dependent on the scenario and not on the model architecture, thereby allowing adaptation and application to a multitude of different scenarios and domains.

In future work, we plan to use ill-formed utterances where the user language does not strictly follow a correct grammar, to train the model and to improve the model's robustness. Also, we plan to deploy the model on the real NICO robot, making the model cooperate with its camera and robot arm (RAP), and to evaluate our architecture in the real world, including ASR and TTS. Furthermore, more variations in the table scene, like a random number of objects, might increase the robustness and applicability of the model in the real world.

Acknowledgment

The authors gratefully acknowledge support from the China Scholarship Council (CSC) and the German Research Foundation DFG under project CML (TRR 169). We thank Alexander Sutherland for his advice on the experimental design.

References

1. Alomari, M., Dukes, K.: Extended train robots. (2016). https://doi.org/10.5518/32
2. Bagaskara, A., Naufal, A.R., Dhojopatmo, I.E., Abdurrab, A., Budiharto, W.: Development of smart restaurant application for dine-in. In: Conference on Computer Science and Artificial Intelligence. vol. 1, pp. 230–235 (2021)
3. Bordes, A., Boureau, Y.L., Weston, J.: Learning end-to-end goal-oriented dialog. Preprint arXiv:1605.07683 (2016)
4. Brabra, H., Báez, M., Benatallah, B., Gaaloul, W., Bouguelia, S., Zamanirad, S.: Dialogue management in conversational systems: A review of approaches, challenges, and opportunities. IEEE Transactions on Cognitive and Developmental Systems (2021)
5. Calli, B., Walsman, A., Singh, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: Benchmarking in manipulation research: The YCB object and model set and benchmarking protocols. Preprint arXiv:1502.03143 (2015)
6. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J.M., Parikh, D., Batra, D.: Visual Dialog. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
7. De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.: GuessWhat?! Visual object discovery through multi-modal dialogue. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for language understanding. arXiv:1810.04805 (2018)
9. Goel, R., Paul, S., Hakkani-Tür, D.: HyST: A hybrid approach for flexible and accurate dialogue state tracking. Preprint arXiv:1907.00883 (2019)
10. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
11. Henderson, M., Thomson, B., Young, S.: Word-based dialog state tracking with recurrent neural networks. In: 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue. pp. 292–299 (2014)
12. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation. **9**(8), 1735–1780 (1997)
13. Huang, X., Tan, C.S., Ng, Y.B., Shi, W., Yeo, K.H., Jiang, R., Kim, J.j.: Joint generation and Bi-Encoder for situated interactive multimodal conversations. In: AAAI 2021 DSTC9 Workshop (2021)
14. Jeong, Y., Lee, S.J., Ko, Y., Seo, J.: TOM: End-to-end task-oriented multimodal dialog system with GPT-2. In: AAAI 2021 DSTC9 Workshop (2021)
15. Kerzel, M., Abawi, F., Eppe, M., Wermter, S.: Enhancing a neurocognitive shared visuomotor model for object identification, localization, and grasping with learning from auxiliary tasks. IEEE Transactions on Cognitive and Developmental Systems pp. 1–13 (2020)

16. Kerzel, M., Strahl, E., Magg, S., Navarro-Guerrero, N., Heinrich, S., Wermter, S.: NICO—Neuro-Inspired COmpanion: A developmental humanoid robot platform for multimodal interaction. In: IEEE International Symposium on Robot and Human Interactive Communication. pp. 113–120 (2017)
17. Kottur, S., Moon, S., Geramifard, A., Damavandi, B.: SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. arXiv:2104.08667 (2021)
18. Kottur, S., Moura, J.M., Parikh, D., Batra, D., Rohrbach, M.: CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. Preprint arXiv:1903.03166 (2019)
19. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. Preprint arXiv:1910.13461 (2019)
20. Lin, T.Y., Goyal, P., Girshick, R.B., He, K., Dollár, P.: Focal loss for dense object detection. IEEE International Conference on Computer Vision (2017)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
22. Lu, J., Batra, D., Parikh, D., Lee, S.: ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: Advances in Neural Information Processing Systems. vol. 32 (2019)
23. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)
24. Moon, S., Kottur, S., Crook, P.A., De, A., Poddar, S., Levin, T., Whitney, D., Difranco, D., Beirami, A., Cho, E., et al.: Situated and interactive multimodal conversations. Preprint arXiv:2006.01460 (2020)
25. Mostafazadeh, N., Brockett, C., Dolan, B., Galley, M., Gao, J., Spithourakis, G.P., Vanderwende, L.: Image-Grounded Conversations: Multimodal context for natural question and response generation. Preprint arXiv:1701.08251 (2017)
26. Murahari, V., Batra, D., Parikh, D., Das, A.: Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In: European Conference on Computer Vision. pp. 336–352. Springer (2020)
27. Ni, J., Young, T., Pandelea, V., Xue, F., Adiga, V., Cambria, E.: Recent advances in deep learning based dialogue systems: A systematic survey. Preprint arXiv:2105.04387 (2021)
28. Qian, K., Beirami, A., Kottur, S., Shayandeh, S., Crook, P., Geramifard, A., Yu, Z., Sankar, C.: Database search results disambiguation for task-oriented dialog systems. Preprint arXiv:2112.08351 (2021)
29. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding by generative pre-training (2018)
30. Weston, J., Bordes, A., Chopra, S., Rush, A.M., Van Merriënboer, B., Joulin, A., Mikolov, T.: Towards AI-complete question answering: A set of prerequisite toy tasks. Preprint arXiv:1502.05698 (2015)
31. Williams, J.D., Asadi, K., Zweig, G.: Hybrid Code Networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. Preprint arXiv:1702.03274 (2017)
32. Zeiler, M.D.: ADADELTA: An adaptive learning rate method. Preprint arXiv:1212.5701 (2012)