

Jointly Learning to See, Ask, *Decide when to Stop*, and then GuessWhat

Ravi Shekhar[†], Alberto Testoni[†], Raquel Fernández* and Raffaella Bernardi[†]

[†]University of Trento, *University of Amsterdam

ravi.shekhar@unitn.it alberto.testoni@unitn.it

raquel.fernandez@uva.nl raffaella.bernardi@unitn.it

Abstract

We augment a task-oriented visual dialogue model with a decision-making module that decides which action needs to be performed next given the current dialogue state, i.e. whether to ask a follow-up question or stop the dialogue. We show that, on the *GuessWhat?!* game, the new module enables the agent to succeed at the game with shorter and hence less error-prone dialogues, despite a slightly decrease in task accuracy. We argue that both dialogue quality and task accuracy are essential features to evaluate dialogue systems.¹

1 Introduction

The development of conversational agents that ground language in visual information is a challenging problem that requires the integration of dialogue management skills with multimodal understanding. A common test-bed to make progress in this area are guessing tasks where two dialogue participants interact with the goal of letting one of them guess a visual target (Das et al., 2017a; de Vries et al., 2017; Das et al., 2017b). We focus on the *GuessWhat?!* game, which consists in guessing a target object within an image which is visible to both participants. One participant (the Questioner) is tasked with identifying the target object by asking yes-no questions to the other participant (the Oracle), who is the only one who knows the target. Participants are free to go on with the task for as many turns as required.

Most models of the Questioner agent in the *GuessWhat?!* game consist of two disconnected modules, a Question Generator and a Guesser, which are trained independently with Supervised

Learning or Reinforcement Learning (de Vries et al., 2017; Strub et al., 2017). In contrast, Shekhar et al. (2019) model these two modules jointly. They show that thanks to its joint architecture, their Questioner model leads to dialogues with higher linguistic quality in terms of richness of the vocabulary and variability of the questions, while reaching a performance similar to the state of the art with Reinforcement Learning. They argue that achieving high task success is not the only criterion by which a visually-grounded conversational agent should be judged. Crucially, the dialogue should be coherent, with no unnatural repetitions nor irrelevant questions. We claim that to achieve this, a conversational agent needs to learn a strategy to decide how to respond at each dialogue turn, based on the dialogue history and the current context. In particular, the Questioner model has to learn when it has gathered enough information and it is therefore ready to guess the target.

In this work, we extend the joint Questioner architecture proposed by Shekhar et al. (2019) with a decision-making component that decides whether to ask a follow-up question to identify the target referent, or to stop the conversation to make a guess. Shekhar et al. (2018) had added a similar module to the baseline architecture by de Vries et al. (2017). Here we show that the novel joint architecture by Shekhar et al. (2019) can also be augmented with a decision-making component and that this addition leads to further improvements in the quality of the dialogues. Our extended Questioner agent reaches a task success comparable to Shekhar et al. (2019), but it asks fewer questions, thus significantly reducing the number of games with repetitions.

¹Copyright © 2019 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

2 Task and Models

2.1 Task

The *GuessWhat?!* dataset² was collected via Amazon Mechanical Turk by de Vries et al. (2017). The task involves two human participants who see a real-world image, taken from the MS-COCO dataset (Lin et al., 2014). One of the participants (the Oracle) is assigned a target object in the image and the other participant (the Questioner) has to guess it by asking Yes/No questions to the Oracle. There are no time constraints to play the game. Once the Questioner is ready to make a guess, the list of candidate objects is provided and the game is considered successful if the Questioner picks the target object. The dataset consists of around 155k English dialogues about approximately 66k different images. Dialogues contain on average 5.2 questions-answer pairs.

We use the same train (70%), validation (15%), and test (15%) splits as de Vries et al. (2017). The test set contains new images not seen during training. Following Shekhar et al. (2019), we use two experimental setups for the number of questions to be asked by the Questioner, motivated by prior work: 5 questions (5Q) as de Vries et al. (2017), and 8 questions (8Q) as Strub et al. (2017).

2.2 Models

We focus on developing a Questioner agent able to decide when it has asked enough information to identify the target object. We first describe the baseline model proposed by de Vries et al. (2017). Then we describe the model proposed by Shekhar et al. (2019) and extend it with a decision making module.

Baseline de Vries et al. (2017) model the Questioner agent of the *GuessWhat?!* game as two disjoint models a Question Generator (QGen) and a Guesser trained independently. After a fixed number of questions by QGen, the Guesser selects a candidate object.

QGen is implemented as a Recurrent Neural Network (RNN) with a transition function handled with Long-Short-Term Memory (LSTM), on which a probabilistic sequence model is built with a Softmax classifier. Given the overall image (encoded by extracting its VGG features) and the current dialogue history (i.e., the previous sequence

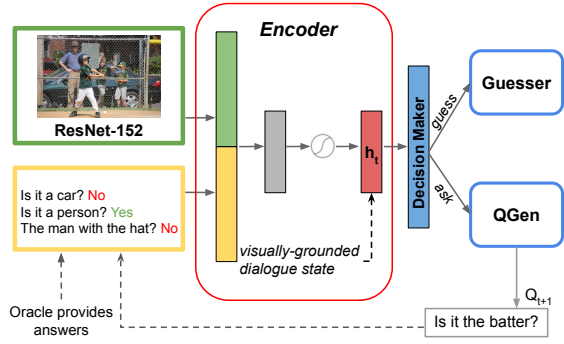


Figure 1: Proposed visually-grounded dialogue state encoder with a decision-making component.

of questions and answers), QGen produces a representation of the visually grounded dialogue (the RNN’s hidden state QH_{t-1} at time $t - 1$ in the dialogue) that encodes information useful to generate the next question q_t . The best performing model of the Guesser by de Vries et al. (2017) represents candidate objects by their object category and spatial coordinates. These features are passed through a Multi-Layer Perceptron (MLP) to get an embedding for each object. The Guesser also takes as input the dialogue history processed by an LSTM, whose hidden state GH_{t-1} is of the same size as the MLP output. A dot product between both returns a score for each candidate object in the image.

Shekhar et al. (2018) extend the baseline architecture of de Vries et al. (2017) with a third model, a decision-making component that determines, after each question/answer pair, whether the QGen model should ask another question or whether the Guesser model should guess the target object.

Grounded Dialogue State Encoder (GDSE)

Shekhar et al. (2019) address one of the fundamental weakness of the Questioner model by de Vries et al. (2017), i.e., having two disconnected QGen and Guesser modules. They tackle this issue with a multi-task approach, where a common visually-grounded dialogue state encoder (GDSE) is used to generate questions and guess the target object. Two learning paradigms are explored: supervised learning (SL) and co-operative learning (CL). In SL, the Questioner model is trained using human data. While in CL, the Questioner model is trained on both generated and human data. First, the Guesser is trained on the generated questions and answers and then the QGen is “readapted” using the human data. Their results show that

²Available at <https://guesswhat.ai/download>.

training these two modules jointly improves the performance of the Questioner model, reaching a task success comparable to RL-based approaches (Strub et al., 2017).

Adding a Decision Making module (GDSE-DM) We extend the GDSE model of Shekhar et al. (2019) with a decision-making component (DM). The DM determines whether QGen should ask a follow-up question or the Guesser should guess the target object, based on the image and dialogue history. As shown in Figure 1, the DM component is modelled as a binary classifier that uses the visually-grounded dialogue state h_t to decide whether to *ask* or *guess*. It is implemented by a Multi Layer Perceptron (MLP_d) trained together with the encoder with negative log-likelihood loss:

$$\mathcal{L}_D = -\log p(\text{dec}_{label}) \quad (1)$$

where dec_{label} is the decision label, i.e., ‘ask’ or ‘guess’. The MLP_d consists of three hidden layers whose dimensions are 256, 64, and 16, respectively; after each hidden layer a ReLU non-linearity is applied.

To train the DM, we need decision labels. For the SL setting, we follow the label generation procedure introduced by Shekhar et al. (2018): decision labels are generated by annotating all the last question-answer pairs in the games with *guess* and all other question-answer pairs as *ask*. For the CL setting, we label the question/answer pairs based on whether the Guesser module is able to correctly predict the target object given the current dialogue. If the Guesser module is able to make a correct prediction after a given question/answer pair, we label that dialogue state with *guess* and otherwise with *ask*. This process results in an unbalanced dataset for the DM where the guess label makes up for only 20% of states. We address this class imbalance by adding a weighing factor, α , to the loss. The balanced loss is given by

$$\mathcal{L}_D = \alpha_{label} \cdot (-\log p(\text{dec}_{label})) \quad (2)$$

where $\alpha_{guess} = 0.8$ and $\alpha_{ask} = 0.2$.

The DM, for both SL and CL, is trained with Cross Entropy loss in a supervised manner using decision labels after each question/answer pair. During inference, the model continues to ask questions unless the DM chooses to end the conversation or the maximum number of questions has been reached. The GDSE-DM model trained with

Model	5Q	8Q
Baseline	41.2	40.7
GDSE-SL	47.8	49.7
GDSE-CL	53.7 (± 0.83)	58.4 (± 0.12)
GDSE-SL-DM	46.78	49.12
GDSE-CL-DM	49.77(± 1.16)	53.89(± 0.24)

Table 1: Test set accuracy for each model (for setups with 5 and 8 questions).

SL and CL will be referred to as SL-DM and CL-DM, respectively. It has to be highlighted that the tasks of generating a question and guessing the target object are not equally challenging: while the Guesser has to learn the probability distribution of the set of possible objects in the image, QGen needs to fit the distribution of natural language words, which is a much harder task. As in Shekhar et al. (2019), we address this issue by making the learning schedule task-dependent using a *modulo-n* training setup. In the SL setting, n indicates after how many epochs of QGen training the Guesser is updated together with QGen; for CL, QGen is updated at every n^{th} epoch, while the Guesser is updated at all other epochs. We found the optimal value of n to be equal to 5 for both the SL and the CL setting. The models are trained for 100 epochs with Adam optimizer and a learning rate of 0.0001 and we select the Questioner module with the best performance on the validation set.

3 Results

In this section, we report the task success accuracy of our GDSE-DM model, which extends the joint GDSE architecture with a decision-making component. Following Shekhar et al. (2019), to neutralize the effect of random sampling in CL training, we use 3 runs and report mean and standard deviation.

Table 1 gives an overview of the accuracy results obtained by the models. Our main goal is to show the effect of adding a DM module to the joint GDSE architecture. We therefore do not compare to other approaches that use RL.³ As we can see, adding a DM to the GDSE model decreases its accuracy by 0.5-1% in the supervised learning setting and by 4-5% in the cooperative learning set-

³For completeness, the RL model by Strub et al. (2017) has accuracy 56.2(± 0.24) and 56.3(± 0.05) for the 5Q and 8Q settings, respectively.

Model	5Q	8Q
GDSE-SL-DM	3.83	5.49
GDSE-CL-DM	4.02(± 0.10)	5.46(± 0.10)

Table 2: Average number of questions asked by the GDSE-DM models when the maximum number of questions is set to 5 or 8.

ting. We believe that the higher drop in accuracy of the CL-DM model can be attributed to the decision labels used by this model. In the SL-DM setting, the model is trained on human data, which leads to a more reliable decision label. In contrast, in the CL-DM setting, the model is trained on automatically generated data, which includes possible errors by both the QGen and the Oracle. Overall, this results in more noisy dialogues. We think that, due to the accumulation of these errors, the decision labels of the generated dialogue deviate significantly from the human data and thus the DM fails to capture them.

Despite the drop in task success, the DM agent seems to be more efficient. Table 2 shows that the average number of questions asked by the DM-based models is lower: the GDSE model without a DM always asks the maximum number of questions allowed (either 5 or 8 questions); while, on average, the GDSE-DM agent asks around 3.8 to 5.5 questions, even when it is allowed to ask up to 8. As we shall see in the next section, this leads to dialogues that are more natural and less repetitive.

4 Analysis

In this section, we look into the advantage brought about by the DM in terms of the quality of the dialogues produced by the model.

Following Shekhar et al. (2019), we report statistics about the dialogue produced by the models with respect to lexical diversity (measured as type/token ratio over all games), question diversity (measured as the percentage of unique questions over all games), and percentage of games with at least one repeated question (see Table 3). The main drawback of the models asking a fixed number of questions is that they repeat questions within the same dialogue. While the introduction of the joint GDSE architecture by Shekhar et al. (2019) substantially reduced the percentage of games with repeated questions with respect to the baseline model (from 93.5% to 52.16%), more

	Lexical diversity	Question diversity	% Games with repeated Q's
Baseline	0.030	1.60	93.50
GDSE-SL	0.101	13.61	55.80
GDSE-CL	0.115 (± 0.02)	14.15 (± 3.0)	52.19 (± 4.7)
GDSE-SL-DM	0.047	1.62	42.47
GDSE-CL-DM	0.135(± 0.02)	10.25(± 2.46)	32.51(± 6.45)
Humans	0.731	47.89	—

Table 3: Statistics of the linguistic output of all models with the 8Q setting compared to human dialogues in all test games.

than 50% of dialogues included repetitions, which make them unnatural. We can see how adding a DM component to GDSE addresses this important problem: with the CL-DM setting, the percentage of games with repeated questions goes down to 32.51% (-19.68%, from 52.19 to 32.51). The reduction is also substantial for the SL-DM model (-13.33%, from 55.80 to 42.47) albeit less impressive.

Given that the number of questions asked by the DM-based models is lower (as shown in Table 2), it is to be expected that the lexical and question diversity of the dialogues produced by these models will also be somewhat lower. Indeed, we observe a rather significant drop in diversity for the SL-DM setting. The CL-DM model, on the other hand, is rather robust to diversity loss: in fact, lexical diversity increases slightly with respect to GDSE-CL (0.135 vs. 0.115 on average), while question diversity decreases by a couple of points only, remaining much higher than that of the baseline model.

Following Shekhar et al. (2019), we also looked into the distribution of the types of questions asked by the models. Questions are divided into two broad categories; ENTITY (about the target object category, e.g., *'is it a animal?'*) and ATTRIBUTE (about the target object property, e.g., *'is it green one?'*). ENTITY questions are further sub-divided into 'object category' and 'super-category'. ATTRIBUTE questions are divided into 'color', 'shape', 'size', 'texture', 'location', and 'action' questions. Table 4 provides distribution of questions by different models. Compared with their counterparts, the DM-based models ask more object questions. The SL-DM also lowers significantly the number of location questions (from 37.09 to 21.70), which are the type of question most commonly repeated by the various models, as shown by Shekhar et al. (2019). We also computed the Kullback-Leibler


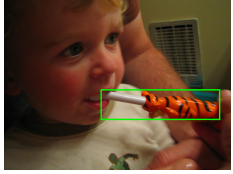
	Baseline [failure]	GDSE-SL [success]	GDSE-CL [success]
	1. is it a person? no	1. is it a person? no	1. is it a person? no
	2. is it a tree? no	2. is it a ski? yes	2. is it a tree? no
	3. is it a tree? no	3. is it on the person in the front? yes	3. is it a ski? yes
	4. is it a tree? no	DM stops asking to guess	4. is it both skis of the person in front? yes
	5. is it a tree? no	4. is it the left most person? yes	DM stops asking to guess
	6. is it a tree? no	5. both skis? yes	5. is it on the left? yes
	7. is it a tree? no	6. both skis? yes	6. is it on the left? yes
	8. is it a tree? no	7. the whole skis? yes	7. it is second to the right? no
	8. the whole skis? yes	8. is it the one on the left? yes	
	Baseline [success]	GDSE-SL [success]	GDSE-CL [success]
	1. is it the baby? no	1. is it a person? no	1. is it a person? no
	2. is it the toothbrush? yes	2. is it a toothbrush? yes	2. is it a toothbrush? yes
	3. is it the toothbrush? yes	3. is it the whole thing? yes	DM stops asking to guess
	4. is it the toothbrush? yes	DM stops asking to guess	3. is it the one kid holding? yes
	5. is it the toothbrush? yes	4. the one in the hand? yes	4. is it the one left of the baby? no
	6. is it the toothbrush? yes	5. the one in the front? yes	5. is it next to the one with the red handle? yes
	7. is it the toothbrush? yes	6. the whole thing? yes	6. is it next to the baby? no
	8. is it the entire toothbrush? yes	7. the whole of it? yes	7. is it one next to the baby? no
	8. all of it? yes	8. next to the whole object? yes	

Figure 2: Game examples where most models succeed at guessing the target object (framed). In red, the point in the dialogue where the DM component decides to stop asking questions and guess the target. Many of the questions asked after the decision point by the GDSE model without a DM are repeated, and thus do not add any extra information.

(KL) divergence to measure how the output of each model differs from the human distribution of fine-grained question classes. We can see that GDSE-DM models have comparatively higher degree of divergence than GDSE, in particular the SL-DM model, which asks a substantially larger proportion of ENTITY questions.

The sample dialogues in Figure 2 provide a qualitative illustration of the output of our models, showing how the DM-based Questioner stops asking questions when it has enough information to guess the target object.

Question type	BL	SL	CL	SL-DM	CL-DM	H
ENTITY	49.00	48.07	46.51	71.03	51.36	38.11
SUPER-CAT	19.6	12.38	12.58	15.35	15.40	14.51
OBJECT	29.4	35.70	33.92	55.68	35.97	23.61
ATTRIBUTE	49.88	46.64	47.60	27.27	45.21	53.29
COLOR	2.75	13.00	12.51	10.57	8.41	15.50
SHAPE	0.00	0.01	0.02	0.0	0.07	0.30
SIZE	0.02	0.33	0.39	0.01	0.67	1.38
TEXTURE	0.00	0.13	0.15	0.01	0.25	0.89
LOCATION	47.25	37.09	38.54	21.70	39.92	40.00
ACTION	1.34	7.97	7.60	3.96	8.01	7.59
Not classified	1.12	5.28	5.90	1.70	3.43	8.60
KL wrt Human	0.953	0.042	0.038	1.48	0.055	—

Table 4: Percentage of questions per question type in all the test set games played by humans (H) and the models with the 8Q setting, and KL divergence from human distribution of fine-grained question types.

5 Conclusion

We have enriched the Questioner agent in the goal-oriented dialogue game *GuessWhat?!* with a Decision Making (DM) component. Based on the visually grounded dialogue state, our Questioner model learns whether to ask a follow-up question or to stop the conversation to guess the target object. We show that the dialogue produced by our model has less repetitions and less unnecessary questions, thus potentially leading to more efficient and less unnatural interactions – a well known limitation of current visual dialogue systems. As in Shekhar et al. (2018), where a simple baseline model was extended with a DM component, task accuracy slightly decreases while the quality of the dialogues increases.

A first attempt to partially tackle the issue within the *GuessWhat?!* game was made by Strub et al. (2017), who added a <stop> token to the vocabulary of the question generator module to learn when to stop asking questions using Reinforcement Learning. This is a problematic approach as it requires the QGen to generate probabilities over a non-linguistic token; further, the decision to ask more questions or guess is a binary decision and thus it is not desirable to incorporate it within the large softmax output of the QGen.

Jiaping et al. (2018) propose a hierarchical RL-based Questioner model for the *GuessWhich* image-guessing game introduced by Chattopad-

hyay et al. (2017) using the *VisDial* dataset (Das et al., 2017a). The first RL layer is a module that learns to decide when to stop asking questions. We believe that a decision making component for the *GuessWhich* game is an ill-posed problem. In this game, the Questioner does not see the pool of candidate images while carrying out the dialogue. Hence, it will never know when it has gathered enough information to distinguish the target image from the distractors. In any case, our work shows that a simple approach can be used to augment visually-grounded dialogue systems with a DM without having to use the high complexity of RL paradigms.

Task accuracy and dialogue quality are equally important aspects of visually-grounded dialogue systems. It remains to be seen how such systems can reach higher task accuracy while profiting from the better quality that DM-based models produce.

References

- Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.
- Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- Abhishek Das, Satwik Kottur, José M.F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. In *International Conference on Computer Vision (ICCV)*.
- Zhang Jiaping, Zhao Tiancheng, and Yu Zhou. 2018. Multimodal hierarchical reinforcement learning policy for task-oriented visual dialog. In *Proceeding of the SigDial Conference*, pages 140–150. Association for Computational Linguistics.
- Sang-Woo Lee, Yu-Jung Heo, and Byoung-Tak Zhang. 2017. Answerer in questioner’s mind for goal-oriented visual dialogue. In *NIPS Workshop on Visually-Grounded Interaction and Language (ViGIL)*. ArXiv:1802.03881. Last version Feb. 2018.
- T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, Dollár, P., and C. L. Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of ECCV (European Conference on Computer Vision)*.
- Ravi Shekhar, Tim Baumgärtner, Aashish Venkatesh, Elia Bruni, Raffaella Bernardi, and Raquel Fernández. 2018. Ask no more: Deciding when to guess in referential visual dialogue. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 1218–1233.
- Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. 2019. Beyond task success: A closer look at jointly learning to see, ask, and guesswhat. In *NAACL*.
- Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.
- Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville. 2017. Guesswhat?! Visual object discovery through multi-modal dialogue. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton van den Hengel. 2018. Goal-oriented visual question generation via intermediate rewards. In *Proceedings of the European Conference of Computer Vision (ECCV)*.
- Yan Zhu, Shaoting Zhang, and Dimitris Metaxas. 2017. Interactive reinforcement learning for object grounding via self-talking. In *NIPS Workshop on Visually-Grounded Interaction and Language (ViGIL)*.