

# Beyond task success: A closer look at jointly learning to see, ask, and GuessWhat

## Supplementary Material

Ravi Shekhar<sup>†</sup>, Aashish Venkatesh<sup>\*</sup>, Tim Baumgärtner<sup>\*</sup>, Elia Bruni<sup>\*</sup>,

Barbara Plank<sup>♥</sup>, Raffaella Bernardi<sup>†</sup> and Raquel Fernández<sup>\*</sup>

<sup>†</sup>University of Trento, <sup>\*</sup>University of Amsterdam, <sup>♥</sup>IT University of Copenhagen  
bplank@itu.dk raffaella.bernardi@unitn.it raquel.fernandez@uva.nl

### 1 Details on the *GuessWhat?!* dataset

The *GuessWhat?!* dataset contains 77,973 images with 609,543 objects. It consists of around 155K dialogues containing around 821K question/answer pairs composed out of 4900 words (min 5 times occurrences) on 66,537 unique images and 134,073 objects. Answers are Yes (52.2%), No (45.6%) and NA (not applicable, 2.2%); dialogues contain on average 5.2 questions and there are on average 2.3 dialogues per image. There are successful (84.6%), unsuccessful (8.4%) and not completed (7.0%) dialogues.

### 2 Models and experimental setting

#### 2.1 Oracle

Figure 1 illustrates the architecture of the Oracle model by de Vries et al. (2017), which we re-implemented for our study.

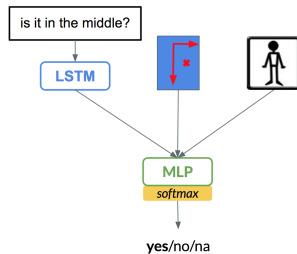


Figure 1: Oracle model.

#### 2.2 Details on our GDSE model

As explained in section 4.2, our architecture consists of a visually-grounded dialogue state Encoder, a Question Generator (QGen) module and Guesser module. Both QGen and Guesser modules are trained using ground truth data. The Encoder is updated during the training of both QGen and Guesser through backpropagation.

In the Encoder, the dialogue history is represented by word embeddings of dimension 512 which is processed by a LSTM with a hidden layer of 1024 dimension. This is then concatenated with visual features pre-computed from the second last layer of ResNet-152. The pre-computed visual features have a dimension of 2048. For pre-computing the visual features all the images are

resized to 224x224 and then passed through the ResNet-152. The combined features (1024+2048 = 3072) are then passed through a linear layer to scale down the features to the size of 512. This is then passed through a tanh layer to get a final representation from the encoder which is going to be the input to the QGen and Guesser.

The QGen acts as the decoder in a seq2seq model (Sutskever et al., 2014). QGen is a LSTM with hidden layer of 512 dimension. Similar to seq2seq models, the Encoder representation is taken to be the starting hidden state of the decoder. The QGen module gets word embeddings concatenated with scaled-down image features as the input. The image features are scaled down from 2048 to 512 and the size word embedding 512 dimensions. For each object, the Guesser receives the representation of the object category, viz., a dense category embedding obtained from its one-hot class vector using a learned look-up table, and its spatial representation, viz., an 8-dimensional vector.

The QGen and the Guesser are optimized while training by minimizing the negative log-likelihood for generated question words and the correct object selection respectively. They are trained using ADAM optimizer with a learning rate of 0.0001. All the parameters are tuned on the validation set. Model is trained for 100 epochs and the best model is selected based on the validation game accuracy.

### 3 Analysis

We provide further details and examples related to the analyses carried out in Section 7.2 of the main paper.

#### 3.1 Question classification

The analysis by question type is based on a classification on a set of keywords. These keywords have been annotated using information in the MS-COCO dataset plus manual annotation. First, we created the possible question categories by inspecting the human dialogues. As explained in the paper, the resulting categories are ENTITY, subdivided into SUPER-CATEGORY and OBJECT, and ATTRIBUTE, sub-divided into COLOR, LOCATION, SHAPE, SIZE, TEXTURE and ACTION. We exploited the super-category and object annotations from MS-COCO. To further enrich these annotations, we manually annotated the words in the human dialogues that occur at least

---

**Algorithm 1: Question Classification.**


---

```

Input : Question and annotated words (from Table 1).
Output Question Classification
:
1 Let  $Q = \{w_1 \dots w_l \dots w_n\}$  denotes all the words for the given Question ;
2 Let Color, Shape, Size, Texture, Location, Action, Object, Super
   denotes all words present in the 'Color', 'Shape', 'Size', 'Texture', 'Location',
   'Action', 'Object' and 'Super-category' respectively ;
3 Let  $Q_{cat}$  a Empty List ;
4 for  $\forall w_k \in Q$  do
5   if  $w_k \in Color$  then
6      $Q_{cat} \leftarrow Q_{cat} + color$  ;      // Append color to  $Q_{cat}$ 
7     break
8   end
9 end
10 for  $\forall w_k \in Q$  do
11   if  $w_k \in Shape$  then
12      $Q_{cat} \leftarrow Q_{cat} + shape$  ;      // Append shape to  $Q_{cat}$ 
13     break
14   end
15 end
16 for  $\forall w_k \in Q$  do
17   if  $w_k \in Size$  then
18      $Q_{cat} \leftarrow Q_{cat} + size$  ;      // Append size to  $Q_{cat}$ 
19     break
20   end
21 end
22 for  $\forall w_k \in Q$  do
23   if  $w_k \in Texture$  then
24      $Q_{cat} \leftarrow Q_{cat} + texture$  ;    // Append texture to  $Q_{cat}$ 
25     break
26   end
27 end
28 for  $\forall w_k \in Q$  do
29   if  $w_k \in Location$  then
30      $Q_{cat} \leftarrow Q_{cat} + location$  ; // Append location to  $Q_{cat}$ 
31     break
32   end
33 end
34 for  $\forall w_k \in Q$  do
35   if  $w_k \in Action$  then
36      $Q_{cat} \leftarrow Q_{cat} + action$  ;    // Append action to  $Q_{cat}$ 
37     break
38   end
39 end
40 if  $Q_{cat}$  is EMPTY then
41   for  $\forall w_k \in Q$  do
42     if  $w_k \in Object$  then
43        $Q_{cat} \leftarrow object$  ;          // Assign object to  $Q_{cat}$ 
44       break
45     end
46   end
47 end
48 if  $Q_{cat}$  is EMPTY then
49   for  $\forall w_k \in Q$  do
50     if  $w_k \in Super$  then
51        $Q_{cat} \leftarrow super$  ; // Assign super-category to  $Q_{cat}$ 
52       break
53     end
54   end
55 end
56 if  $Q_{cat}$  is EMPTY then
57    $Q_{cat} \leftarrow not-classified$  ; // Assign not-classified to  $Q_{cat}$ 
58 end
59 return  $Q_{cat}$ 

```

---

Question	Question type
is it a <i>basket</i> ?	OBJECT
is it a <i>human</i> ?	SUPER-CATEGORY
is it the person in the <i>middle</i> ?	LOCATION
is the person wearing a <i>white</i> shirt?	COLOR
is it the <i>round</i> table?	SHAPE
is it the <i>little</i> plate?	SIZE
is he wearing a <i>striped</i> shirt?	TEXTURE

Table 1: Examples of questions from the human dialogues with keywords in italics and the types assigned through our classification procedure.

40 times in the training and testing sets. In Table 6, we report the complete list of keywords highlighting those obtained from COCO. Algorithm 1 provides the pseudo-code of the question classification heuristics we used. Table 1 provides some examples of the resulting classification.

### 3.2 Dialogue strategy

Table 2 and 3 provides the details of Table 3 and 4 of the main paper, respectively including standard deviation for CL & RL.

Table 4 shows how dialogues start, i.e., the percentages of the type of questions used right at the beginning of a game. We can observe that all the models start with ENTITY questions more than 95% of the time. Sometimes GDSE-CL starts with ATTRIBUTE question (around 4%), while RL start with ATTRIBUTE question negligible number of times (0.003%).

### 3.3 Question repetition

In Table 5, we look more closely at the repetitions of the questions. Repetition is performed by string matching. We see that in the game played by the baseline (BL) more than 93% of games have some form of repetitions. In case of GDSE, the games played in the supervised learning setting have slightly more repetition than in the cooperative learning one (55.80% vs. 52.19%). RL has even more repetitive (96%) questions. For RL and BL, majority of repetition is for the cases where same question is repeated consecutively.

We further looked at where repetition is happening i.e. at the start or towards the end of dialogue. For this, we looked at the dialogues and computed if there is any repetition in the dialogue after the 5<sup>th</sup> question. For the baseline, around 67% of the dialogues are having repetitions toward the end of dialogue. This also explains why for the baseline 5 is the optimal number of questions. Even for GDSE based models, around 47% of repetitions happens towards the end of dialogue. Furthermore, for RL the majority of the repetitions are in terms of consecutive question repetitions.

### 3.4 Quality of the dialogues

Figures 2 to 5 provide further examples of the games played by the models.


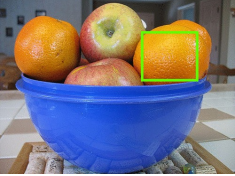
	<b>Humans</b>	[success]	<b>GDSE-CL</b>	[success]	<b>RL</b>	[failure]
	1. is it a person?	no	1. is it a person?	no	1. is it a person?	no
	2. is it a birthday hat?	no	2. is it a cake?	no	2. is it a food?	no
	3. is it the candles?	no	3. is it a cup?	no	3. is it a table?	no
	4. is it a donut?	no	4. is it a table?	no	4. is it in left?	yes
	5. is it the cake?	no	5. is it a chair?	yes	5. is it in left?	yes
	6. is it a piece of clothing?	no	6. is it the one on the left side of the table?	yes	6. is it in front?	yes
	7. is it a chair?	yes	7. is it the one which is sitting?	yes	7. is it in top?	no
	8. is it the one the baby is sitting in?	yes	8. is it just behind the girl?	yes	8. is it in top?	yes
	<b>Humans</b>	[success]	<b>GDSE-CL</b>	[success]	<b>RL</b>	[failure]
	1. is it a fruit?	yo	1. is it an orange?	yes	1. is it food?	yes
	2. is it the apple?	no	2. is it in the front?	yes	2. is it one of the oranges?	yes
	3. is it the left orange?	no	3. is it an apple?	no	3. is it in left?	no
	4. is it the smaller orange on the right?	no	4. is it all the orange?	yes	4. is it in middle?	no
	5. is it the bigger orange on the right?	yes	5. can you see most of it?	yes	5. is it in top?	yes
			6. is it on the left side of the picture?	no	6. is it in top?	yes
			7. is it the top right corner?	yes	7. is it in top?	yes
			8. so it is an apple touching the apple?	yes	8. is it in top?	yes

Figure 2: Game examples in which GDSE-CL succeeds and RL fails at guessing the target object (in the green box). In both examples, GDSE-CL asks further attribute-specific questions. In the top example, *chair* is not that visible so the RL model does not ask a question about it, after Q3 it just asks location questions. In contrast, GDSE-CL after getting *chair* (Q5) it tries to ask questions to distinguish amongst chairs. Similarly, in the bottom example, after getting *orange* (Q2), RL keeps asking simplistic location questions. The GDSE-CL model, after getting *orange* (Q1), tries to ask attribute questions along with location questions.

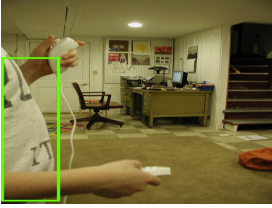
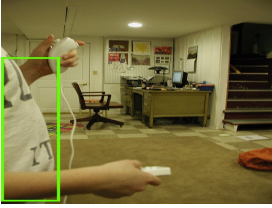
	<b>GDSE-CL</b>	[success]	<b>RL</b>	[success]
	1. is it a person?	yes	1. is it a person?	yes
	2. is it the man in the white shirt?	yes	2. is it in foreground?	yes
	3. is it the whole man?	yes	3. is it in left?	yes
	4. does he have a yellow shirt?	yes	4. is it in right?	no
	5. is he on the right side?	no	5. is it in middle?	yes
	6. is he on the left side?	yes	6. is it in right?	no
	7. his face is not visible?	yes	7. is it in middle?	yes
	8. is his face visible?	yes	8. is it in right?	no
	<b>Humans</b>	[success]		
	1. is it in the background?	no		
	2. is it a person?	yes		

Figure 3: Game example in which both GDSE-CL and RL succeed at guessing the target object (green box), using a different strategy. Here, we have only one instance of the target object super-category *person*, and both models are able to guess the object correctly. While RL keeps asking location questions (Q3-8), GDSE-CL asks questions about color, location, visibility. Even after getting wrong Oracle answer to a question (Q8), it is able to guess the correct object based on previous evidence.



	<b>GDSE-CL</b>	[failure]	<b>RL</b>	[success]
	1. is it a person?	yes	1. is it a person?	yes
	2. is it the man on the motorcycle?	yes	2. is it in foreground?	yes
	3. is it the whole person?	yes	3. is it in left?	no
	4. is it a part of the man?	no	4. is it in right?	yes
	5. is the whole body shown?	yes	5. is it in right?	yes
	6. can i see his whole body?	yes	6. is it in right?	yes
	7. is it the man on top of the bike?	yes	7. is it in right?	yes
	8. is it the whole person?	yes	8. is it in right?	yes
	<b>Humans</b>	[success]		
	1. is it a person?	yes		
	2. are they wearing a helmet?	yes		

Figure 4: Game example in which GDSE-CL fails and RL succeeds at guessing the target object (green box). Here, there are only two instances of the target object super-category *person*. The RL strategy of asking location-based questions works well here.


	<b>Humans</b>	[success]	<b>GDSE-CL</b>	[failure]	<b>RL</b>	[failure]
	1. is it a sofa?	no	1. is it furniture?	yes	1. is it a person?	no
	2. is it green in color?	no	2. is it a couch?	no	2. is it a computer?	no
	3. is it a clock?	no	3. a chair then?	yes	3. is it a couch?	no
	4. is it a tv?	no	4. the blue one?	yes	4. is it in left?	yes
	5. is it blue color?	no	5. the one on the bottom?	no	5. is it in front?	no
	6. is it a pillow?	no	6. the left one?	yes	6. in top	yes
	7. it is a chair?	yes	7. the whole thing?	yes	7. is it in top?	yes
	8. it is black color?	yes	8. just the backrest?	yes	8. is it in top?	yes

Figure 5: Game examples in which both GDSE-CL and RL fail. The presence of multiple instances of similar target objects (category *chair*) in a similar context makes the task very difficult to the models.

Question type	Example	BL	SL	CL	RL	H
<b>ENTITY</b>		<b>49.00</b>	<b>48.07</b>	<b>46.51(±1.33)</b>	<b>23.99(±0.04)</b>	<b>38.11</b>
-SUPER-CATEGORY	<i>Is it a vehicle?</i>	19.6	12.38	12.58(±0.37)	14.00(±0.03)	14.51
-OBJECT	<i>Is it a skateboard?</i>	29.4	35.70	33.92(±1.03)	9.99(±.02)	23.61
<b>ATTRIBUTE</b>		<b>49.88</b>	<b>46.64</b>	<b>47.60(±0.61)</b>	<b>75.52(±0.05)</b>	<b>53.29</b>
-COLOR	<i>Is he wearing blue?</i>	2.75	13.00	12.51(±1.27)	0.12(±0.004)	15.50
-SHAPE	<i>Is it square?</i>	0.00	0.01	0.02(±0.012)	0.003(±0.001)	0.30
-SIZE	<i>The bigger one?</i>	0.02	0.33	0.39(±0.074)	0.024(±0.003)	1.38
-TEXTURE	<i>Is it wood?</i>	0.00	0.13	0.15(±0.094)	0.013(±0.004)	0.89
-LOCATION	<i>The one from the left?</i>	47.25	37.09	38.54(±2.6)	74.80(±0.03)	40.00
-ACTION	<i>Are they standing?</i>	1.34	7.97	7.60(±2.52)	0.66(±0.12)	7.59
<b>Not classified</b>		<b>1.12</b>	<b>5.28</b>	<b>5.90(±0.78)</b>	<b>0.49(±0.009)</b>	<b>8.60</b>
KL divergence wrt Human distribution		0.953	0.042	0.038(±0.004)	0.396(±0.0001)	—

Table 2: 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.

Question type shift	BL	SL	CL	RL	H
SUPER-CAT $\rightarrow$ OBJ/ATT	89.05	92.61	89.75(±3.61)	95.63(±0.163)	89.56
OBJECT $\rightarrow$ ATTRIBUTE	67.87	60.92	65.06(±1.92)	99.46(±0.215)	88.70

Table 3: Proportion of question type shift vs. no type shift in consecutive questions  $Q_t \rightarrow Q_{t+1}$  where  $Q_t$  has received a Yes answer.

	BL	SL	CL	RL	H
ENTITY	97.26	97.26	94.67 (±1.83)	98.77 (±0.06)	78.48
SUPER-CAT	75.24	68.74	69.29 (±2.79)	84.41(±0.08)	52.32
OBJECT	22.03	28.52	25.37(±1.73)	14.37(±0.03)	26.16
ATTRIBUTE	1.72	1.39	3.89 (±1.64)	0.003(±0.001)	13.95

Table 4: Percentages of question type of the first question in the dialogues.

## References

- Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.
- 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)*.

	% game having at least 1 repetition	% game having at least 1 consecutive repetition	% game having at least 1 repetition after 5 <sup>th</sup> Que	% game having at least 1 consecutive repetition after 5 <sup>th</sup> Que
BL	93.5	65.42	67.36	46.89
GDSE-SL	55.80	34.24	47.61	27.52
GDSE-CL	52.19( $\pm 4.7$ )	25.26( $\pm 6.43$ )	36.03( $\pm 5.65$ )	20.66( $\pm 5.2$ )
RL	96.47 ( $\pm 0.04$ )	73.72( $\pm 0.27$ )	69.43( $\pm 0.2$ )	50.75( $\pm 0.27$ )

Table 5: Percentages of repeated questions in all games for different models.

ENTITY	
SUPER-CATEGORY	<i>'person', 'vehicle', 'outdoor', 'animal', 'accessory', 'sports', 'kitchen', 'food', 'furniture', 'electronic', 'appliance', 'indoor', 'utensil', 'human', 'cloth', 'cloths', 'clothing', 'people', 'persons'</i>
OBJECT	<i>'bicycle', 'car', 'motorcycle', 'airplane', 'bus', 'train', 'truck', 'boat', 'traffic light', 'fire hydrant', 'stop sign', 'parking meter', 'bench', 'bird', 'cat', 'dog', 'horse', 'sheep', 'cow', 'elephant', 'bear', 'zebra', 'giraffe', 'backpack', 'umbrella', 'handbag', 'tie', 'suitcase', 'frisbee', 'skis', 'snowboard', 'sports ball', 'kite', 'baseball bat', 'baseball glove', 'skateboard', 'surfboard', 'tennis racket', 'bottle', 'wine glass', 'cup', 'fork', 'knife', 'spoon', 'bowl', 'banana', 'apple', 'sandwich', 'orange', 'broccoli', 'carrot', 'hot dog', 'pizza', 'donut', 'cake', 'chair', 'couch', 'potted plant', 'bed', 'dining table', 'toilet', 'tv', 'laptop', 'mouse', 'remote', 'keyboard', 'cell phone', 'microwave', 'oven', 'toaster', 'sink', 'refrigerator', 'book', 'clock', 'vase', 'scissors', 'teddy bear', 'hair drier', 'toothbrush', 'meter', 'bear', 'cell', 'phone', 'wine', 'glass', 'racket', 'baseball', 'glove', 'hydrant', 'drier', 'kite', 'sofa', 'fork', 'adult', 'arms', 'baby', 'bag', 'ball', 'bananas', 'basket', 'bat', 'batter', 'bike', 'birds', 'board', 'body', 'books', 'bottles', 'box', 'boy', 'bread', 'brush', 'building', 'bunch', 'cabinet', 'camera', 'candle', 'cap', 'carrots', 'cars', 'cart', 'case', 'catcher', 'cell phone', 'chairs', 'child', 'chocolate', 'coat', 'coffee', 'computer', 'controller', 'counter', 'cows', 'cupboard', 'cups', 'curtain', 'cycle', 'desk', 'device', 'dining table', 'dish', 'doll', 'door', 'dress', 'driver', 'equipment', 'eyes', 'fan', 'feet', 'female', 'fence', 'fire', 'flag', 'flower', 'flowers', 'foot', 'frame', 'fridge', 'fruit', 'girl', 'girls', 'glasses', 'guy', 'guys', 'hair drier', 'handle', 'hands', 'hat', 'helmet', 'house', 'jacket', 'jar', 'jeans', 'kid', 'kids', 'lady', 'lamp', 'leg', 'legs', 'luggage', 'machine', 'male', 'man', 'meat', 'men', 'mirror', 'mobile', 'monitor', 'mouth', 'mug', 'napkin', 'pan', 'pants', 'paper', 'pen', 'picture', 'pillow', 'plant', 'plate', 'player', 'players', 'pole', 'pot', 'purse', 'rack', 'racket', 'road', 'roof', 'screen', 'shelf', 'shelves', 'shirt', 'shoe', 'shoes', 'short', 'shorts', 'shoulder', 'signal', 'sign', 'silverware', 'skate', 'ski', 'sky', 'snow', 'soap', 'speaker', 'stairs', 'statue', 'stick', 'stool', 'stove', 'street', 'suit', 'sunglasses', 'suv', 'teddy', 'tennis', 'tent', 'tomato', 'towel', 'tower', 'toy', 'traffic', 'tray', 'tree', 'trees', 't-shirt', 'tshirt', 'vegetable', 'vest', 'wall', 'watch', 'wheel', 'wheels', 'window', 'windows', 'woman', 'women'</i>
ATTRIBUTES	
COLOR	<i>'white', 'red', 'black', 'blue', 'green', 'yellow', 'orange', 'brown', 'pink', 'grey', 'gray', 'dark', 'purple', 'color', 'colored', 'colour', 'blond', 'beige', 'bright'</i>
SIZE	<i>'small', 'little', 'long', 'large', 'largest', 'big', 'tall', 'smaller', 'bigger', 'biggest', 'tallest'</i>
TEXTURE	<i>'metal', 'silver', 'wood', 'wooden', 'plastic', 'striped', 'liquid'</i>
SHAPE	<i>'circle', 'rectangle', 'round', 'shape', 'square', 'triangle'</i>
LOCATION	<i>'1st', '2nd', 'third', '3', '3rd', 'four', '4th', 'fourth', '5', '5th', 'five', 'first', 'second', 'last', 'above', 'across', 'after', 'around', 'at', 'away', 'back', 'background', 'before', 'behind', 'below', 'beside', 'between', 'bottom', 'center', 'close', 'closer', 'closest', 'corner', 'directly', 'down', 'edge', 'end', 'entire', 'facing', 'far', 'farthest', 'floor', 'foreground', 'from', 'front', 'furthest', 'ground', 'hidden', 'in', 'inside', 'left', 'leftmost', 'middle', 'near', 'nearest', 'next', 'next to', 'off', 'on', 'out', 'outside', 'over', 'part', 'right', 'rightmost', 'row', 'side', 'smaller', 'top', 'towards', 'under', 'up', 'upper', 'with'</i>

Table 6: Lists of keywords used to classify questions with the corresponding class according to Algorithm 1. Words in *italics* come from COCO object category/super-category.