Probability Distributions as a Litmus Test to Inspect NNs Grounding Skills

Alex J. Lucassen[1], Alberto Testoni[2] and Raffaella Bernardi[1,2]

[1]CIMeC, University of Trento, Palazzo Fedrigotti - corso Bettini 31, 38068 Rovereto (TN), Italy
[2]DISI, University of Trento, Via Sommarive, 9 I-38123 Povo (TN), Italy

Abstract

Today AI systems are trained by ultimately using a classifier to perform a down-streaming task and are mostly evaluated on the task-success they reach. Not enough attention is given to how the classifier distributes the probabilities among the candidates out of which the target with the highest probability is selected. We propose to take the probability distribution as a litmus test to inspect models' grounding skills. We take a visually grounded referential guessing game as test-bed and use the probability distribution as a way to evaluate whether question answer pairs are well grounded by the model. To this end, we propose a method to obtain such soft-labels automatically and show they correlate well with human uncertainty about the grounded interpretation of the QA pair. Our result shows that higher task accuracy does not necessarily correspond to a more meaningful probability distribution; we do not consider trustworthy the models which do not pass our litmus test.

Keywords

Referential Guessing Games, Soft-labels, Interpretable and Trustworthy Agents.

1. Introduction

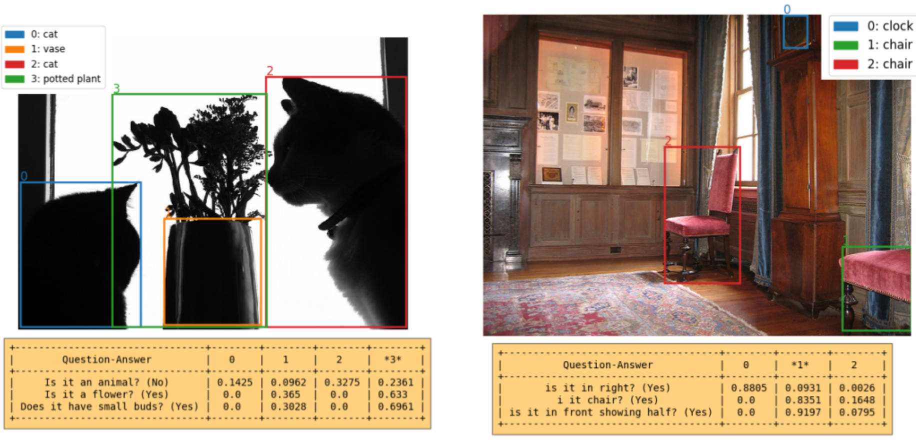
Classification tasks represent the backbone of most AI systems. These systems are trained to assign probabilities to a set of labels, while the underneath model learns to reach a suitable representation of the given input. The evaluation usually consists of comparing the probability distribution of the model against the ground-truth label. Instead, we conjecture it is important to look into the probability distribution over the whole set of possible candidate labels and not only the one that receives the highest probability. In our work, we will bring evidence that modelling uncertainty is a crucial step for building trustworthy AI systems.

Probabilities are crucial during problem-solving tasks. In Cognitive Science, a typical test to study humans' problem-solving strategies is the 20Q game in which a Questioner has to ask a sequence of questions to guess which is the target object the other player, the Oracle, has been assigned. The studies focus on how the questioner' conjectures about the target drive the sequence of questions he/she asks. Models have been evaluated on tasks such as the 20Q game but again the evaluation has mostly focused on task success or on the linguistic quality of the generated sequence of questions. We believe models should be compared also on how

NL4AI 2022: Sixth Workshop on Natural Language for Artificial Intelligence, November 30, 2022, Udine, Italy [1]

alex.lucassen@studenti.unitn.it (A. J. Lucassen); alberto.testoni@unitn.it (A. Testoni); raffaella.bernardi@unitn.it (R. Bernardi)
0000-0002-1388-0444 (A. J. Lucassen); 0000-0002-4130-0056 (A. Testoni); 0000-0002-3423-1208 (R. Bernardi)

CEUR Workshop Proceedings (CEUR-WS.org)



```
+--------------------------------+---------+---------+---------+---------+
|        Question-Answer         |    0    |    1    |    2    |   *3*   |
+--------------------------------+---------+---------+---------+---------+
|     Is it an animal? (No)      | 0.1425  | 0.0962  | 0.3275  | 0.2361  |
|     Is it a flower? (Yes)      |  0.0    | 0.365   |  0.0    | 0.633   |
| Does it have small buds? (Yes) |  0.0    | 0.3028  |  0.0    | 0.6961  |
+--------------------------------+---------+---------+---------+---------+
```

```
+--------------------------------------+---------+---------+---------+
|           Question-Answer            |    0    |   *1*   |    2    |
+--------------------------------------+---------+---------+---------+
|        is it in right? (Yes)         | 0.8805  | 0.0931  | 0.0026  |
|          i it chair? (Yes)           |  0.0    | 0.8351  | 0.1648  |
| is it in front showing half? (Yes)   |  0.0    | 0.9197  | 0.0795  |
+--------------------------------------+---------+---------+---------+
```

Figure 1: Although in both games the target object is guessed correctly (receives the highest probability) at the last turn, the trust on the model decreases when looking at the probability distribution at first turn.

sound and interpretable the probability is that they assign to the various candidates. When challenged with tasks in which the agent aims to incrementally gain confidence in its hypothesis and therefore arrives to take a well-thought decision, the probability assigned to the candidates at a certain step should impact the reasoning process and guide the natural language generation. Therefore, assigning meaningful probabilities is crucial also for the efficacy and coherence of the dialogue structure. We take GuessWhat?! [2], a grounded version of the 20Q game, as a test-bed and shed light on how the probability distribution of models serves as a litmus test for evaluating the extent to which their task success can be trusted.

Figure 1 illustrates two cases where a state-of-the-art (SOTA) model correctly predicts the target object based on the full dialogue generated by humans playing the game. However, if we look at the probabilities assigned after the first turn we see that they are *uninterpretable*. In the example on the left, one would expect the first Question Answer (QA) pair (the category question: 'Is it an animal?' 'No') to lead to low probabilities for objects #0 and #2, which are cats, and higher for the other two candidates, the vase and the potted plant. Contrary to the expectations, the cats receive higher probabilities than the other two objects, and one cat seems more probable than the other one, clearly just for some data bias. In the example on the right, after the first QA pair (a spatial question: 'is it in right? 'Yes'), object #0 is assigned a very high probability, much higher than object #1, which is actually further to the right in the image. In this paper, we call attention to such a crucial aspect of neural network models and propose a method to evaluate their probability distribution success by taking human disagreement as a proxy of human uncertainty, and designing rule-based systems that simulate such disagreement by leveraging on the human annotation and which can then be used to automatically annotate larger data.

Given the nature of the GuessWhat?! dialogues, at each turn we can identify the set of objects that, given the information received till that point, could be the target and those which have been instead excluded. We call these two sets, the **reference set** and the **complement set**.

Each turn brings an update of the reference set: when the question is positively (vs. negatively) answered the reference set maintains all the candidates that have (vs. don't have) the property asked in the question. The candidates that are not maintained in the reference set move to the complement set. We would expect the probabilities assigned to members of these two sets to differ for the QA pair to be well grounded, and those of the complement set to be close to zero. Obviously, the cut between the two sets is not always crystal clear and the members in the sets might have a different status: for some members of the referent (vs. complement) set, the model could be more confident than for others, that is, it should assign a higher (vs. lower) probability; they should receive different **soft-labels**. The difficulty is how to obtain such labels so as to properly evaluate the model's grounding and reasoning skills.

We contribute to this challenge in the following way. First of all, we propose a method for obtaining soft-label distribution automatically. It consists of two phases: 1) a data collection step with subjects on a small sample, and 2) an automatic annotation of the full dataset through a rule-based system, after having verified its correlation with the human data on the sample set. Concretely, we take the first turns of GuessWhat?! games as case-study and collect human annotations. Our data collection experiment shows that humans highly agree on grounding category and object questions ("Is it an animal?"/ "Is it a dog?") but less so in grounding spatial questions, in particular absolute questions ("Is it on the left?"). We take humans' disagreement as a proxy for humans' uncertainty and obtain human-based soft-labels for a sample of the GuessWhat?! dataset. We implement rule-based systems that simulate such disagreement/uncertainty and use the human annotation data to verify that the resulting soft-labels correlate well with those derived from human disagreement on the sample dataset. We use such systems to automatically annotate the first turns in the GuessWhat?! test set containing category and spatial questions. Secondly, we report how much models' probability distribution correlates with the soft-labels obtained automatically for the full test set. Finally, we propose metrics that quantify to which extent the models' probability distribution reflects the intuition that candidates in the complement set should have a probability as close as possible to zero or at least lower than a certain threshold while all the candidates in the referential set should receive a probability not lower than a given threshold. Since grounding negatively answered questions has been shown to be harder than grounding the positively answered ones both for models and for humans [3], we report our results by comparing turns containing a negative vs. positive answer.

2. Related Work

GuessWhat?! Over the years, a number of different approaches have been proposed for the GuessWhat?! task [4]. [2] proposed a baseline model for the Oracle, the Questioner, and the Guesser, which is the model that would guess the target object after the dialogue has been concluded. [5] introduced a Deep Reinforcement Learning model for building a multi-modal goal direction dialogue system. [6] sought to mitigate the issue of dialogue agents focusing on simple utterances, and introduced a class of temperature-based extensions for policy gradient methods called Tempered Policy Gradients (TPGs), which would mitigate the problem. This method was used to create an improved Guesser model based on Deep Reinforcement Learning, TPGs and Memory-Attention (tow-hop attention) [6, 4]. [7] proposed a grounded dialogue state encoder (GDSE), which combines training the Question generator and the Guesser by learning both in a multi-task fashion. More recently, [8] proposed new models for the GuessWhat?! task using VilBert, thereby allowing the models to take advantage of the pretrained vision-linguistic model. The models for the Oracle, Guesser, and Questioner all outperformed state-of-the-art models. [9] adapted LXMERT [10], a multimodal universal encoder, to act as the Oracle, reaching an important improvement over the baseline. [11] used GuessWhat?! as a way to evaluate the performance of two pre-trained Transformers, LXMERT and RoBERTa [12], and [3] used Guesser models with LXMERT and RoBERTa as encoders. In our work, we take GDSE as a baseline model and LXMERT as a representative instance of pre-trained transformer-based models. None of these works, however, has studied how to assign probabilities to different candidates. Even [8], in which for the first time the Guesser has been trained incrementally turn by turn neglected this analysis of the probability distribution. [13] proposed Confirm-it, a cognitively-inspired beam search re-ranking strategy for the GuessWhat Questioner agent that exploits the probabilities assigned by the Guesser to guide the Question Generator. However, the authors rely on the probabilities of the Guesser without investigating them in detail. So far, there has been no careful study of how to distribute the probability for the Guesser.

Evaluation of Visually Grounded Models Through the study of probability distributions generated by the model, our work also relates to research on natural language understanding in vision and language (V&L) models. While V&L models are often evaluated by looking at performance on V&L tasks, recent studies have focused on models' understanding of specific linguistic structures [14, 15]. [14] proposed a new framework for investigating a model's fine-grained understanding of linguistic structures in spatial expressions. They used the visual dialogue dataset OneCommon and evaluated the model's understanding through reference resolution. [15] propose VALSE (Vision and Language Structured Evaluation), a benchmark for evaluating the grounding capabilities of pretrained V&L models on specific linguistic phenomena. Models have to distinguish real captions from foils (captions altered to differ on some linguistic phenomenon). Our work differs from these studies by proposing a metric to evaluate the grounding of the reference set for a single QA pair, rather than focusing on specific linguistic structures or phenomena.

Soft labels There has been a growing interest in soft labels, defined as probability distributions over annotator labels. Soft labels have been shown to help build a more robust representation. [16] show that models trained with soft labeling are better at generalizing on unseen data. [17] use soft labels as an auxiliary task in a multi-task learning setting with a main task based on hard labels, and show that this improves on the main task. These works build on earlier lines of research that replaced the training objective moving from the most likely label to full distribution over labels [16]. Alternatives to one-hot encodings have been proposed also for image classification [18, 19, 16]. Most previous works on soft labels aim at improving the task accuracy. Instead, we propose to use soft-labels as a test-bed to evaluate the trustworthiness of computational models. We conjecture that generating probability distributions inspired by soft labels should come naturally as a by-product of an effective model training, without the need to

Table 1
Composition of the human annotation dataset

Question type	n
Category	25
Object	25
Color	40
Absolute Spatial	85

explicitly provide this information at training time. As an upper-bound, we also train a model to replicate soft-labels distributions using KL divergence loss [20].

3. Dataset and soft-label annotation

In this Section, we are going to describe the data collection procedure on a sample dataset with human annotators. Then, we present how we design the automatic annotation on the GuessWhat?! test set. The human annotation allows us to study human disagreement and verify the reliability of the automatic annotation.

3.1. Human annotation collected for a sample set

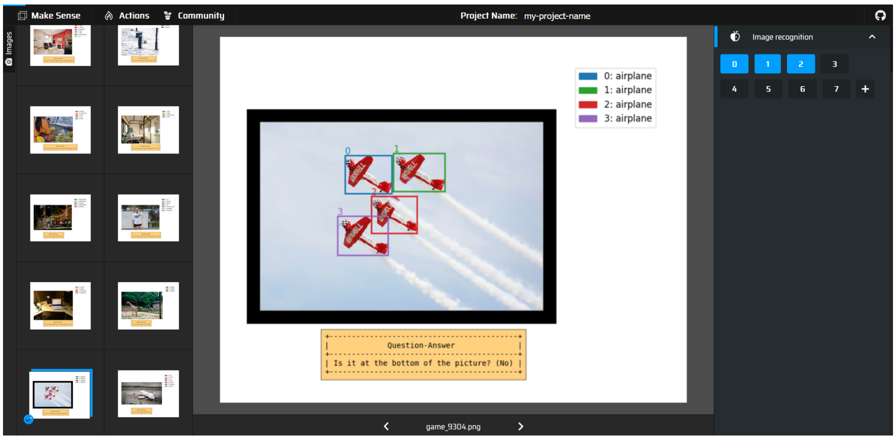
From the training set of GuessWhat?! human-human dialogues, we selected a subset of first questions (i.e., questions asked at the first turn of the dialogue) focusing on the most frequent question types, namely category, object, color, and absolute spatial questions. The distribution of such a subset is given in Table 1.[1]

In order to collect humans' judgment, participants were asked to use the website makesense.ai. As illustrated in Figure 2, they were given an image with colored bounding boxes for each candidate object, a list of MS-COCO labels corresponding to such objects, and a QA pair ('Is it at the bottom of the picture? No'). The participants had to select the set of possible target objects for the given QA pair (the reference set, i.e., all the objects that are not considered to be at the bottom of the picture). Participants were also informed that there were no correct or wrong answers. We collected human annotation data for 175 QA pairs, each annotated by three participants. In total, we recruited nine people, most of whom were university students.

3.2. Automatic soft-label annotation for a large-scale dataset

Starting from the sample of human-annotated data, we designed methods to annotate the first QA pairs containing category and absolute questions automatically. We focus on these two question types because the GuessWhat?! dataset contains, for each candidate object in the image, its category label and its spatial coordinates. We can thus easily exploit these two sources of information for the automatic annotation. We believe other strategies could be found for other question types so as to apply the rule-based system leveraging on human annotations.

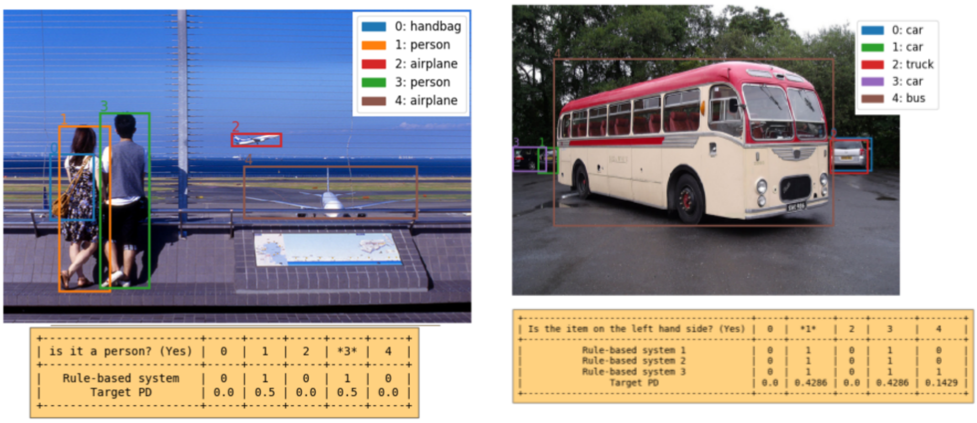
[1] We initially started with 25 questions per question type and then we expanded the two question types with less agreement, namely color questions and absolute spatial questions.



```
+--------------------------------------+
|            Question-Answer           |
+--------------------------------------+
| Is it at the bottom of the picture? (No) |
+--------------------------------------+
```

Figure 2: An example of the set up seen by human annotators in makesense.ai.

Automatic annotation methods For the category questions, we simply rely on the MS-COCO labels and divide the probability uniformly among the objects in the referent set, and leave 0 to the probability of the candidates in the complement set. An example of the resulting probability distribution is given in Figure 3 (left). Here, given the QA pair 'Is it a person? Yes', the probability is uniformly distributed among the two candidate objects (#1 and #3) which are associated with the label 'person'.



```
+--------------------------+-----+-----+-----+-----+-----+
| is it a person? (Yes)    |  0  |  1  |  2  | *3* |  4  |
+--------------------------+-----+-----+-----+-----+-----+
|   Rule-based system      |  0  |  1  |  0  |  1  |  0  |
|       Target PD          | 0.0 | 0.5 | 0.0 | 0.5 | 0.0 |
+--------------------------+-----+-----+-----+-----+-----+
```

```
+---------------------------------------------+-----+--------+-----+--------+--------+
| Is the item on the left hand side? (Yes)    |  0  |  *1*   |  2  |   3    |   4    |
+---------------------------------------------+-----+--------+-----+--------+--------+
|              Rule-based system 1             |  0  |   1    |  0  |   1    |   0    |
|              Rule-based system 2             |  0  |   1    |  0  |   1    |   0    |
|              Rule-based system 3             |  0  |   1    |  0  |   1    |   1    |
|                  Target PD                   | 0.0 | 0.4286 | 0.0 | 0.4286 | 0.1429 |
+---------------------------------------------+-----+--------+-----+--------+--------+
```

Figure 3: Examples of how the rule based system assigns soft-labels for category (left) and spatial (right) questions. Category: The only candidates that are labelled as "person" are #1 and #3, hence the probability of 1 is divided equally between them. Spatial: The three systems agree in considering candidates #0 and #2 as not being on the left, hence they receive a 0 probability score. The probability of 1 is divided among the other candidates proportionally to the number of rules that consider the object as being on the left, resulting into 0.4286 assigned to #1 and #3 (3/3 rules considered them being on the left) and 0.1429 assigned to #4 (only 1/3 rule).

For the absolute spatial questions, we designed a rule-based method. We implemented three systems that simulate three different ways of grounding spatial questions: two of them based on the coordinates of the bounding boxes of the candidate objects, and one on the centroid of the square bounding box containing a given candidate. These three systems are meant to simulate different humans' interpretations of the same question and are explained in detail in Appendix A. We then combine the three systems so as to generate soft-label probabilities over the candidates. In particular, for each rule system, the objects in the reference set get assigned a 1 and the objects in the complement set are assigned a 0. We then sum the result for the three rules for each candidate object, so that an object that is in the reference set for all three rules receives a 3. We now have a list of numbers from 0 to 3, reflecting how often each candidate object is in the reference set. Finally, we apply a Softmax function to this list to turn it into a probability distribution. Figure 3 (right) shows the soft-label probabilities generated out of the answers given by our three rule-based systems. The three rules agree in considering #0 and #2 not to be on the left, hence their probability is scored 0 and the probability of 1 is divided among the other candidates proportionally to the number of rules that consider them to be on the left, namely #1 and #3 receive 0.4286 while #4 receives 0.1429.

Training and Testing data set We extracted the first-turn QA pair containing either a category or an absolute spatial question out of the training and testing set. From the training set, we obtained 59,518 QA containing a category question and 1274 containing an absolute spatial question. From the test set, we obtained 9421 QA-image pairs (4774 Yes and 4647 No) and 235 (104 Yes and 131 No) for category and absolute spatial questions, respectively. We have automatically annotated these 1st QA pairs in the training and test sets with soft-labels using the methods described above.

4. Models and Metrics

Below we describe the models we will evaluate through the paper; in the setting we consider, models receive as input the dialogues or QA pair generated by the Amazon Mechanical Turks who played the GuessWhat?! games, i.e. the dialogues/QA in the dataset released by [2].

4.1. Models

Baseline We use the model provided by [13] for the GuessWhat?! task, which is based on the GDSE architecture [7]. The images are encoded using a ResNet-152 network [21], with an LSTM being used to encode the dialogue history. From this a multi-modal shared representation is generated, which is used to train both the Question Generator and the Guesser module in a joint multi-task learning method. The original model uses the cross-entropy loss (CE) against the ground-truth target object, we experiment with both CE and Kullback-Leibler (KL) divergence loss [20] against a probability distribution over candidate objects.

Upper-bound We train the GDSE model in a multi-task setting. For the main task, we train the model to predict a probability distribution that is a one-hot encoding for the target object, so that the Guesser is still trained to guess the target object. For the auxiliary task, we train the Guesser to predict a probability distribution for a QA pair by using soft labels as a target probability distribution. In both of these cases, KL-divergence loss is used as the loss function to calculate the distance between the predicted probability distribution and the target probability distribution (one-hot encoding for the target object for complete dialogues and soft labels for single QA). We derived the soft labels from the rule-based systems described above. We will refer to this model as `MTL-KL`.

LXMERT We compare the models described above with the Guesser model trained by [3] which is based on a multimodal Transformer-based model, LXMERT [10].

4.2. Metrics

Besides looking at the task-accuracy of the model in identifying the target object at the end of the dialogue, we propose the following metrics to evaluate to which extent the models are trustworthy. They focus on the probability distributions generated by the models and serve as a **litmus test** of the model's grounding skills. When a model performs poorly on one or more of these metrics, it calls into question the reliability of the model. The metrics are all based on the automatic soft-labels we obtain on the test set with the rule-based systems which we use to identify the members of the reference and the complement set. As a preliminary step, we calculate the Pearson's correlation coefficient between the soft-labels obtained from human disagreement and those assigned by the rule-based system on the sample set. Given the high correlation we obtain (See Section 5), we can use the metrics below.

- **Correlation with rule-based soft-labels** We calculate the Pearson's correlation coefficient between the models' probability distribution and the soft-labels assigned by rule-based systems.
- **Percentage of well-grounded QA pairs:** We define a QA pair to be well-grounded when all candidates in the referent (vs. complement) set have been assigned a probability higher (vs. below) a certain threshold θ. Given an image I, a question-answer pair QA, a set of candidate objects $O = \{o_1 ... o_n\}$ and corresponding probabilities $P(o)$, a reference set $RS \subset O$ and a complement set $CS \subset O$ such that $RS \cup CS = O$ and $RS \cap CS = \emptyset$, we define two criteria for a question-answer pair to be considered well-grounded. *Well-grounded against RS:* given a threshold θ, we consider a game to be well-grounded if $\forall x \in RS, P(x) > \theta$. *Well-grounded against CS:* given a threshold θ, we consider a game to be well-grounded if $\forall x \in CS, P(x) < \theta$.
- **Average Probability of the complement set:** We consider a model to be *trustworthy* if all candidates in the complement set receive a probability close to zero. To this end, for each game, we calculate the average probability of the candidate entities in the complement set and the standard deviation. The lower this probability is, the higher is our trust in the model.

5. Experiments

5.1. From human disagreement to automatic soft-labels

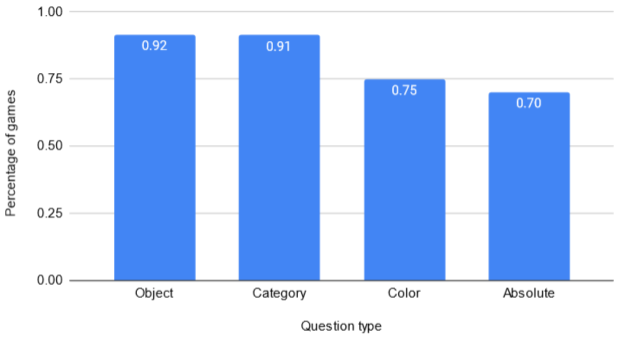
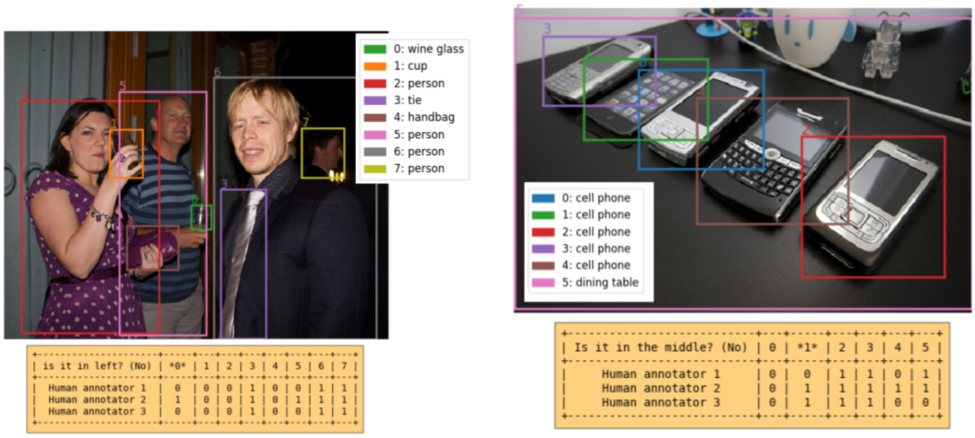


Figure 4: Human almost fully agree on object and category questions, whereas there is some disagreement on color and absolute questions.

Human disagreement on the sample set Figure 4 shows the percentage of games on which human annotators fully agree in identifying the reference set. The comparison of human annotation of the reference sets reveals that humans highly agree with each other on grounding category and objects questions (91% and 92%, respectively); whereas they disagree on 25% of the color questions and 30% of the absolute spatial questions. Figure 5 illustrates examples of spatial questions on which humans disagree. On the left, we can see that human participant 2 also included objects #0 and #5 in the reference set, showing the ambiguity of this kind of spatial questions when dealing with objects located towards the centre of the image. On the right image, instead, there is another source of ambiguity related to the word 'middle'. In fact, annotators seem to interpret this question as either 'middle' of the image or 'middle' of the group of the most salient objects appearing in it. Our aim is to model the disagreement between annotators to obtain reliable probability distributions to model uncertainty.



```
+----------------------+-----+---+---+---+---+---+---+---+
| is it in left? (No)  | *0* | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
+----------------------+-----+---+---+---+---+---+---+---+
|  Human annotator 1   |  0  | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
|  Human annotator 2   |  1  | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
|  Human annotator 3   |  0  | 0 | 0 | 1 | 0 | 0 | 1 | 1 |
+----------------------+-----+---+---+---+---+---+---+---+
```

```
+---------------------------+---+-----+---+---+---+---+
| Is it in the middle? (No) | 0 | *1* | 2 | 3 | 4 | 5 |
+---------------------------+---+-----+---+---+---+---+
|     Human annotator 1     | 0 |  0  | 1 | 1 | 0 | 1 |
|     Human annotator 2     | 0 |  1  | 1 | 1 | 1 | 1 |
|     Human annotator 3     | 0 |  1  | 1 | 1 | 0 | 0 |
+---------------------------+---+-----+---+---+---+---+
```

Figure 5: Human disagreement. Left: the objects that the annotators disagree about the objects located near the middle of the image. Right: human disagreement appears to come from a relational reading of the question.

Quality of the automatic soft-labels annotation To evaluate the quality of the rule-based system annotation, we compute the Pearson's correlation between the soft-labels obtained out of

Table 2

Test set: Task accuracy (TA) obtained by the models when receiving the full humans generated dialogues, the Pearson's correlation (r) between the soft-labels assigned by the rule-based systems and the models for category (CQ) and spatial (SQ) questions in the first turn of the test set games, and the average probability (prob) assigned to entities in the complement set after the first turn containing a category question for positively-answered (Y) and negatively-answered (N) questions.

	TA	r-CQ	r-SQ	prob-Y (SD)	prob-N (SD)
STL-CE	61.2%	0.80	0.68	0.25% (3.41)	0.82% (4.55)
STL-KL	60.6%	0.80	0.64	0.28% (4.06)	0.69% (4.63)
LXMERT	70.2%	0.73	0.66	0.19% (2.20)	0.08% (0.88)
MTL-KL	60.0%	0.97	0.68	0.19% (3.06)	0.34% (3.99)

human disagreement and those obtained out of the rule-based systems. For category questions, the correlation is almost perfect. This is due to the fact that for these questions (appearing in the first turn of GuessWhat dialogues) there is no uncertainty about their interpretation. This is also revealed by the high agreement between human annotators and between the automatic annotation based on MS-COCO labels and human annotation. The Pearson's correlation shows that the ensemble of rule-based systems successfully captures humans' disagreement on spatial absolute questions: its soft-labels have a Pearson's correlation of 0.94 with those derived from humans data. These results bring the first evidence on the feasibility of the method we propose, namely to implement rule-based systems to simulate human disagreement on a small sample of data and then use such systems to automatically annotate large-scale datasets. Of course, it remains to be seen whether the method can be extended to other question types, and whether having such soft-labels for first questions could help models in properly assigning probabilities through the dialogue, incrementally.

5.2. Soft-labels evaluation on the test set

In this experiment, we compute the metrics introduced above to evaluate the probability distributions assigned by the models in the test set. In particular, we are interested in assessing the ability of the models to generate probability distributions over candidate objects that effectively mirror the uncertainty of GuessWhat?! games with incomplete dialogue history.

Pearson's correlation Given the reliability of our proposed rule-based systems to simulate humans' uncertainty / disagreement, we claim they can be used to automatically annotate the GuessWhat?! test set focusing on the first turns of the games which contain category questions. Table 2 shows the task accuracy of different models together with their Pearson's correlation (r) with the soft-labels described above. We can observe some interesting properties emerging from the interplay between these two sets of metrics. First of all, we can observe that LXMERT outperforms the other models to a large extent in accurately identifying the target object at the end of the dialogue. The other models (STL and MTL, regardless of the loss function used) show a similar accuracy. However, if we look at the ability of the models to generate probability distributions that mirror human uncertainty, we can see that the pre-trained transformer-based

Table 3
Percentage of well-grounded QA pairs in the first turn when containing a category question or an absolute spatial question, when looking at the **complement set** (Left) with $theta = 0.4\%$ and at the **reference set** (Right) with $theta = 0.1\%$.

	Category		Absolute Spatial	
	YES	NO	YES	NO
STL-CE	96.6	78.2	31.7	12.0
STL-KL	97.5	80.9	30.8	13.0
LXMERT	97.9	95.1	57.7	30.5
MTL-KL	95.6	95.5	-	-

	Category		Absolute Spatial	
	YES	NO	YES	NO
STL-CE	95.88	80.00	76.72	87.50
STL-KL	93.62	77.26	82.76	78.47
LXMERT	57.02	47.72	59.48	57.64
MTL-KL	99.98	99.98	-	-

model LXMERT is far away from the upper-bound MTL-KL model (explicitly trained to replicate these probabilities) but, surprisingly, also lower than the baseline model. This result shows the importance of our evaluation criterion to shed light on the behaviour of computational models.

Well-grounded QA pairs We compare models based on how well they ground the QA pairs via analysing the probability assigned to the members of the referential set vs. those assigned to the members of the complement set. To verify whether negatively answered questions are harder to be grounded as attested in the literature [3], we report results for the Yes- and the No-first turns containing category or absolute spatial questions. Table 3 reports the percentage of well-grounded QA pairs for the complement set and the reference set . Given the small dataset size of spatial questions, for MTL-KL we focus only on category questions. Remember that this metric captures the ability of the model to assign a probability below a given threshold (vs. above) for all objects belonging to the complement set (vs. reference set). We distinguish between questions that receive positive or negative answers. Overall, we can see that the proposed MTL-KL model effectively manages to include and exclude objects from the reference/complement set and assign soft-label probabilities when dealing with uncertainty. Looking at the complement set (Table 3, left), for category questions we can see that all models except STL-CE effectively assign a low probability to objects belonging to the complement set when dealing with positive answers. However, if we look at questions that receive negative answers, only LXMERT and MTL-KL do not show a degradation in the model's performance. The advantage of LXMERT is even more prominent when looking at spatial questions, both for positive and negative answers. Combining these results with the ones reported in Table 2, we can conclude that LXMERT effectively assigns low probabilities to candidate objects belonging to the complement set but it fails at generating probability distributions that mimic human uncertainty.

Table 2 also shows the average probability assigned to entities in the complement set for category questions. We can see that LXMERT and MTL-KL effectively handle category questions with both positive and negative answers, while the single-task models struggle with the latter, assigning a much higher probability to entities in the complement set.[2] Moving to the reference set, Table 3 (right) also reports the percentage of well-grounded games in the reference set.

[2] We have evaluated models changing the θ value and noticed that it only impacts results on the negatively answered questions.

LXMERT is by far the worst model in assigning a probability over a threshold as low as 0.1% to all objects belonging to the reference set while MTL-KL reaches an almost perfect performance on category questions. Despite the high task accuracy reached by the model, LXMERT is shown to lack the ability to generate trustworthy probabilities.

6. Conclusion

Classification tasks represent the core component of most AI systems. Assigning reliable probability distributions across labels represents a crucial step toward building trustworthy systems. In our work, we take a referential visual dialogue task, GuessWhat?!, as a test bed and we scrutinize the probability distribution assigned by different models to the set of possible referents. We experimented with just the QA in the first turn of the dialogue. We focus our attention on category and spatial questions and designed a set of heuristics that mimic the human annotation we collected for these question types. While category questions show a high agreement between human annotators, spatial questions show an intrinsic uncertainty and lead to human disagreement. Our approach of combining three different rule-based systems effectively takes human annotators' disagreement into account when generating probability distributions over candidate objects. The *well-grounded* metric we propose shows that computational models struggle in generating reliable probability distributions regardless of their architecture or pre-training regime. In line with previous work, we show that models perform worse when dealing with negative answers to polar questions. We used as an upper-bound a model trained to replicate these probabilities using KL divergence loss. We show that the model that performs best in the task accuracy (LXMERT, a pre-trained transformer-based model) does not generate trustworthy probabilities over candidate objects. Our work shows the importance of scrutinizing the ability of computational models to assign reliable probability distributions at inference time.

Acknowledgments

This work was supported by the University of Trento. The first author is a student of the Erasmus Mundus European Masters Program in Language and Communication Technologies. We also thank the anonymous reviewers for their valuable comments.

References

[1] D. Nozza, L. Passaro, M. Polignano, Preface to the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI), in: D. Nozza, L. C. Passaro, M. Polignano (Eds.), Proceedings of the Sixth Workshop on Natural Language for Artificial Intelligence (NL4AI 2022) co-located with 21th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2022), November 30, 2022, CEUR-WS.org, 2022.

[2] H. De Vries, F. Strub, S. Chandar, O. Pietquin, H. Larochelle, A. Courville, Guesswhat?! visual object discovery through multi-modal dialogue, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5503–5512.

[3] A. Testoni, C. Greco, R. Bernardi, Artificial intelligence models do not ground negation, humans do. guesswhat?! dialogues as a case study, Frontiers in big Data 4 (2021).

[4] G. M. Elshamy, M. Alfonse, M. M. Aref, A guesswhat?! game for goal-oriented visual dialog: A survey, in: 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), IEEE, 2021, pp. 116–123.

[5] F. Strub, H. De Vries, J. Mary, B. Piot, A. Courville, O. Pietquin, End-to-end optimization of goal-driven and visually grounded dialogue systems, arXiv preprint arXiv:1703.05423 (2017).

[6] R. Zhao, V. Tresp, Learning goal-oriented visual dialog via tempered policy gradient, in: 2018 IEEE Spoken Language Technology Workshop (SLT), IEEE, 2018, pp. 868–875.

[7] R. Shekhar, A. Venkatesh, T. Baumgärtner, E. Bruni, B. Plank, R. Bernardi, R. Fernández, Beyond task success: A closer look at jointly learning to see, ask, and guesswhat, arXiv preprint arXiv:1809.03408 (2018).

[8] T. Tu, Q. Ping, G. Thattai, G. Tur, P. Natarajan, Learning better visual dialog agents with pretrained visual-linguistic representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5622–5631.

[9] A. Testoni, C. Greco, T. Bianchi, M. Mazuecos, A. Marcante, L. Benotti, R. Bernardi, They are not all alike: Answering different spatial questions requires different grounding strategies, in: Proceedings of the Third International Workshop on Spatial Language Understanding, 2020, pp. 29–38.

[10] H. Tan, M. Bansal, Lxmert: Learning cross-modality encoder representations from transformers, arXiv preprint arXiv:1908.07490 (2019).

[11] C. Greco, A. Testoni, R. Bernardi, Grounding dialogue history: Strengths and weaknesses of pre-trained transformers, in: International Conference of the Italian Association for Artificial Intelligence, Springer, 2020, pp. 263–279.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[13] A. Testoni, R. Bernardi, Looking for confirmations: An effective and human-like visual dialogue strategy, arXiv preprint arXiv:2109.05312 (2021).

[14] T. Udagawa, T. Yamazaki, A. Aizawa, A linguistic analysis of visually grounded dialogues based on spatial expressions, arXiv preprint arXiv:2010.03127 (2020).

[15] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, A. Gatt, Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena, arXiv preprint arXiv:2112.07566 (2021).

[16] J. C. Peterson, R. M. Battleday, T. L. Griffiths, O. Russakovsky, Human uncertainty makes classification more robust, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9617–9626.

[17] T. Fornaciari, A. Uma, S. Paun, B. Plank, D. Hovy, M. Poesio, Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning, in: 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2021.

[18] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern

recognition, 2016, pp. 2818–2826.

[19] R. A. Krishna, K. Hata, S. Chen, J. Kravitz, D. A. Shamma, L. Fei-Fei, M. S. Bernstein, Embracing error to enable rapid crowdsourcing, in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 3167–3179.

[20] S. Kullback, R. A. Leibler, On information and sufficiency, The annals of mathematical statistics 22 (1951) 79–86.

[21] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, IEEE Computer Society, 2016, pp. 770–778. URL: https://doi.org/10.1109/CVPR.2016.90. doi:10.1109/CVPR.2016.90.

A. Rule-Based Systems

We designed a rule-based method for automatic annotation of absolute spatial questions. We implemented three systems that simulate three different ways of grounding spatial questions: two of them based on the coordinates of the bounding boxes of the candidate objects, and one on their centroid, i.e., the centroid of the squared bounding box containing a given candidate.

Each system defines rules for different types absolute spatial questions, specifically, those related to left/right, top/bottom, the middle, and quadrants of the image (e.g. top right). The systems differ in how they assign labels for areas of the image which showed a larger human disagreement so that when combined they reflect different humans' interpretations of the same question.

Rule-based System 1 (R1) The first rule-based system is relatively simplistic and strict. It is based on the intuition that (nearly) the entire object should be in a region of space for it to be considered in that region. The rules and their implications for the candidate answers are illustrated in Figure 6

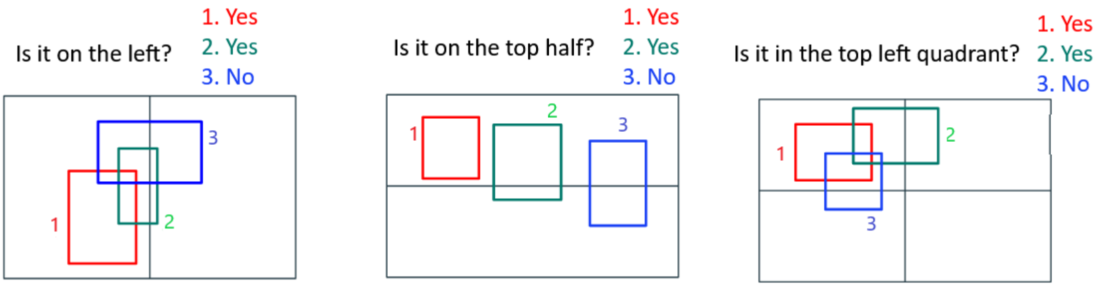


Figure 6: Illustration of the R1 rules for left/right, top/bottom, and quadrant questions respectively.

Left and right questions A candidate object is considered to be on the left of the image when more than 80% of the bounding box is on the left 50% of the image. The inverse is true for the right.

Top and bottom questions A candidate object is considered to be on the top (half) of the image when more than 80% of the bounding box is on the top 50% of the image. The inverse is true for the bottom half.

Middle and center questions A candidate object is considered to be in the middle or center when in both a horizontal and a vertical direction, at least 50% of the bounding box is inside the central 50% of the image.

Quadrants All candidate objects where the entire bounding box is in the top left quadrant of the image are considered to be in the top left of the image. The same holds for other quadrants.

Rule-Based System 2 (R2) The second rule-based system for assigning the candidate objects is more fine-grained than the first one and again uses bounding boxes. It is simultaneously more lenient when it comes to large objects sticking out into other regions of space, and stricter when it comes to small objects near the middle of the image. This is because humans disagree more on objects near the middle for left/right/top/bottom questions. Hypothetical examples of the rules and their implications can be found in Figure 7 and Figure 8.

Left and right questions Unlike in R1, we make a distinction between the 'left half' or 'left picture' and the more generic 'left'. These more specific questions make up 9.6% and 5.2% of the left and right questions respectively, and there is some indication in the human disagreement that they are not always treated the same as generic questions. The intuition is that when the question specifically asks about the left half, small objects close to the middle may still count as 'left' as long as they are on the left 50%, while for the generic 'is it on the left?', the middle may already be considered 'not left'.

To determine whether an object is on the left or left half of an image, we consider the following two conditions:

1. Two-thirds of the bounding box must be on the left 50% of the image.
2. Some part of the bounding box must be on the leftmost 40% of the image

A candidate object is on the left half of the image when the first condition is met. It is considered to be on the left when both conditions are met. The same rule applies in the inverse for questions about the right.

Top and bottom questions A candidate object is considered to be on the top half if at least 75% of the bounding box is on the top half. The inverse is true for bottom questions.

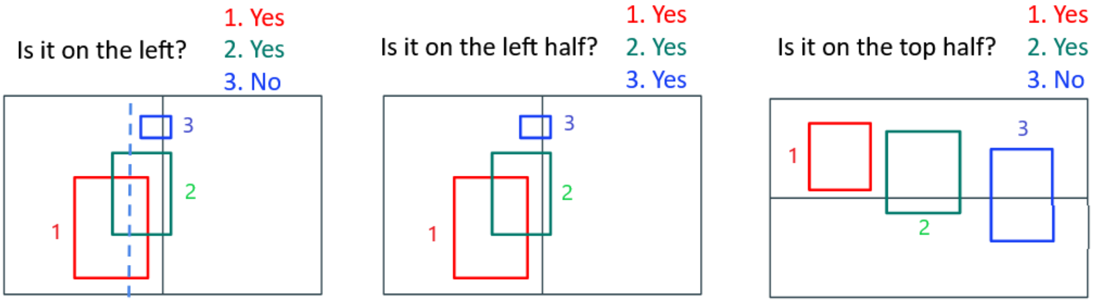


Figure 7: Illustration of the R2 rules for left/right and top/bottom questions.

Middle and center questions A candidate object is considered to be in the middle of the image if the entire bounding box is either in the middle 50% horizontally or vertically.

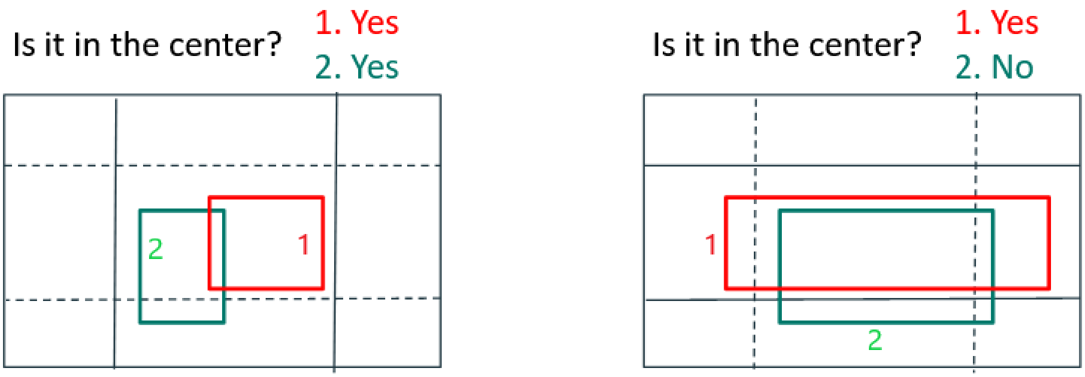


Figure 8: Illustration of the R2 rules for middle and center questions.

Quadrants Lastly, the rule for quadrants such as “top left” is the same as in the first rule-based system.

- All candidate objects where the entire bounding box is in the top left quadrant of the image are considered to be in the top left of the image.

Rule-Based System 3 (R3) The third rule-based system for assigning candidate answers is based entirely on the centroid of the bounding box. The rules are visualized using hypothetical examples in Figures 9 and 10.

Left and right questions A candidate object is considered to be on the left when the centroid of the bounding box is on the left 50% of the image. The inverse holds for right questions.

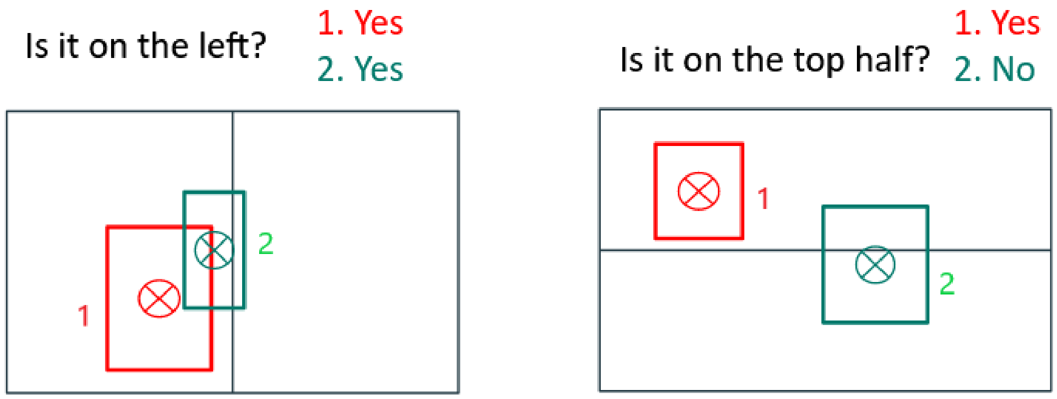


Figure 9: Illustration of the R3 rules for left/right and top/bottom questions respectively.

Top and bottom questions A candidate object is considered to be on the top part when the centroid of the bounding box is on the upper 50% of the image. The inverse holds for bottom questions.

Middle and center questions A candidate object is considered to be in the middle of the image when the centroid of the bounding box is in the middle 50% of the image both horizontally and vertically.

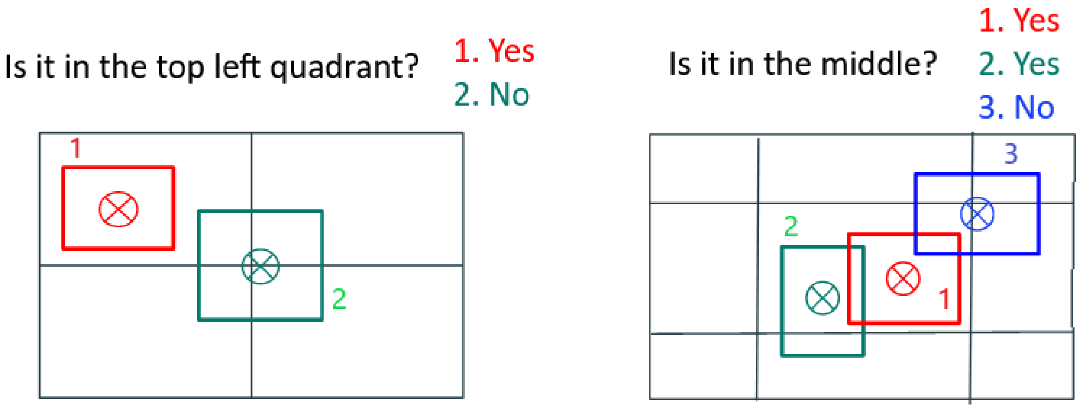


Figure 10: Illustration of the R3 rules for quadrant and middle/center questions respectively.

Quadrants A candidate object is considered to be in the top left when the centroid of the bounding box is in the left 50% horizontally and the top 50% vertically. The same holds for the other quadrants.

Accuracy of the rules In order to assess the accuracy of the annotation assigned by the rules on a larger dataset containing unseen data, we calculate the accuracy in assigning a candidate answer to the target object (available from the GuessWhat?! dataset). For rule-based system 1, the rules lead to correct candidate answers for the target object in 90.5% of the absolute spatial questions in the training set. The accuracy for R2 and R3 is 92.6% and 89.7% respectively.

As we designed these systems to simulate the variation in human interpretation of questions, we also investigate the union of the systems. The union of R1 and R2 for assigning the target object is 94.9%, that of R1 and R3 is 93.9% and that of R2 and R3 is 92.2%.