Criteria for Human-Compatible AI in Two-Player Vision-Language Tasks

Cheolho Han[1,†], **Sang-Woo Lee**[1,†], **Yujung Heo**[1], **Wooyoung Kang**[1], **Jaehyun Jun**[2], **Byoung-Tak Zhang**[1,2]
[1]School of Computer Science and Engineering, Seoul National University
[2]Interdisciplinary Program in Neuroscience, Seoul National University
{chhan, slee, yjheo, wykang, jhjun, btzhang}@bi.snu.ac.kr

Abstract

We propose rule-based search systems that outperform not only the state-of-the-art but the human performance, measured in accuracy, in GuessWhat?!, a vision-language game where either of two players can be a human. Although those systems achieve the high accuracy, they do not meet other requirements to be considered as an AI system that communicates effectively with humans. To clarify what they lack, we suggest the use of three criteria to enable effective communication with humans in vision-language tasks. These criteria also apply to other two-player vision-language tasks that require communication with humans, e.g., ReferIt.

1 Introduction

Recent advances in computer vision and natural language processing have led researchers' attentions to the intersection of these two areas, vision-language tasks. An initiative to this kind of task was image description [Kiros *et al.*, 2014; Vinyals *et al.*, 2015; Xu *et al.*, 2015; Johnson *et al.*, 2016b; Mao *et al.*, 2016]. In the image description task, an image is given, and the model is supposed to generate descriptions or captions on the image. However, generated descriptions have been difficult to evaluate, and results have not been directly related to how well the model understand the image. To show the comprehension of the model, visual question answering (VQA) was introduced [Antol *et al.*, 2015; Johnson *et al.*, 2016a; Agrawal *et al.*, 2017; Fukui *et al.*, 2016; Kim *et al.*, 2016]. In VQA, an image and a question about the image is given, and the model is supposed to answer the question. However, the communication occurs one-way, and the model has the passive role only to answer questions.

Some vision-language tasks require active bidirectional communication between two (or possibly more) agents. To interactively communicate over an image, visual dialogues were introduced [Lazaridou *et al.*, 2016; Das *et al.*, 2016; Mao *et al.*, 2016; de Vries *et al.*, 2017; Strub *et al.*, 2017; Das *et al.*, 2017]. In visual dialogues, an image is given, and two (or possibly more) agents communicate over the image. Visual dialogues involving agents with specific roles or tasks were mainly studied up to the present.

Figure 1: An example of the GuessWhat?! game. The correct object is highlighted by a green mask.

ReferIt game [Kazemzadeh *et al.*, 2014] is an example of visual dialogue. It is a two player game referring to objects in an image of natural scenes. One player is shown an image with a target object and has to explain it to distinguish from the other, where what it says is called the referring expression. The other player is shown the same image and the referring expression written by the other player and guesses the target object.

GuessWhat?! game is also an example of visual dialogue (Fig. 1). It contains dialogue on question answering about the given image over two players. Its goal is to locate an unknown object in a rich image scene by asking a sequence of questions. One player is randomly assigned an object in the image and the other player is to locate the hidden object with a series of Yes/No questions.

On the two tasks mentioned above, each player also can be human or agent. If one player is human and the other is agent, the agent must generate meaningful dialog with humans in natural and conversational language about a visual image. This aspect is crucial to solve the task successfully. Such tasks have been evaluated in terms of task-specific performance metrics such as accuracy or success rate of the task. However, they are not the only criteria to enable efficient bidirectional communication with humans.

In this paper, we pose the problem that we need more criteria to measure and analyze the bidirectional communication between human and agent other than metrics. To demonstrate that, we first try to tackle GuessWhat?! game which we mentioned above and show that our proposed rule-based search

† authors contributed equally.

systems outperformed not only state-of-the-art, but the human performance measure, the success rate of the task. Then, we suggest the use of some criteria to enable efficient bidirectional communication between human and agent in vision-language tasks such as GuessWhat?! game.

The rest of the paper is organized as follows. First we review related works to vision-language tasks on Section 2 and we proposed rule-based search systems which outperformed state-of-the-art performance on Section 3. Then, we suggest the use of some criteria for measure and analyze the bidirectional communication between human and agent on Section 4. Finally, we discuss conclusions and future work on Section 5.

2 Related Works

2.1 Image Description

Automatic image description is a challenging problem that involves analyzing an image, reasoning contextual information between existing objects in the image and generating textual descriptions. It has been first stage on research about vision-language grounding. [Vinyals *et al.*, 2015] proposed neural image caption (NIC) generator inspired by advances in machine translation. They replaced encoding step which extracts abstract representations of source language using RNN to using CNN fed into given image. Encouraged by advances in employing attention in machine translation and object recognition, attention mechanism is introduced by [Xu *et al.*, 2015]. The mechanism can attend to salient part of given image while generating its caption so that demonstrated the learned alignments correspond very well to human intuition.

Many previous papers on image description have focused on describing the entire image. On the other hand, [Johnson *et al.*, 2016b] address a new task, dense captioning, which requires a model to predict a set of descriptions of regions of given image. It is important to understand each object or part not an entire image for high-level scene understanding. [Mao *et al.*, 2016] also focused on generating an unambiguous description of a specific object or region in an image. They considered both description generation and description comprehension and jointly modeled both tasks combining CNN with RNN.

These models are just passive roles only to generate description about given image on the task, not bidirectional communication. And, generated descriptions have been difficult to evaluate, and results have not been directly related to how well the model understand the image. Therefore, the task could be extended for bidirectional communication task such as ReferIt or GuessWhat?! game and it needs further consideration for evaluation and analysis.

2.2 ReferIt

ReferIt [Kazemzadeh *et al.*, 2014; Lazaridou *et al.*, 2016] is a two player game referring an object in an image of natural scenes. One player is shown an image with a target object and has to explain it to distinguish from the other, where what it says is called the referring expression. The other player is shown the same image and the referring expression written by the other player and guesses the target object. To accomplish this game, both agents should cooperate and learn the relation between vision and language. [Lazaridou *et al.*, 2016] designed referIt game between two agents, constitute the referring expression by a binary vector between two agents, formulated the game as classification, and solved the classification problem by neural networks. On the task, these agents develop their own artificial language from the need to communicate in order to succeed at the game. It showed some correlation with human language, but also showed some mismatches. If one player is agent and the other is human, the referring expression may not carry the exact meaning and cause the confusion between the players. In terms of meaninful vision-language integration between human and agent, we argue that we need to analyze details of these referring expressions other than metrics such as success rate of game and accuracy.

2.3 GuessWhat?!

GuessWhat?! is a cooperative two-player guessing game proposed by [de Vries *et al.*, 2017]. The goal of the game is to locate an unknown object in a rich image scene by asking a sequence of questions. One player who called Oracle is randomly assigned an object in the image and the other player who called the Questioner does not know the object assigned to the Oracle. The goal of the Questioner is to locate the hidden object with a series of Yes/No questions which are answered by the Oracle. If the Questioner selects a right object, we consider the game successful.

[de Vries *et al.*, 2017] collect a large-scale human-played GuessWhat?! game dataset consisting of 800K visual question-answering pairs on 66K images and propose baseline deep learning model. To solve the proposed task successfully, [de Vries *et al.*, 2017] suggested that an agent is required higher-level image understanding, like spatial reasoning, visual properties, object taxonomy, and interaction. The authors also proposed that the agent should understand the relationships between objects and how they are expressed in natural language. The baseline model consists of 3 parts: Oracle, Guesser and Question generator. Oracle is a model-based a simple neural network, which fed embedded inputs and classifies answer among Yes/No or N/A. The role of guesser is to predict one hidden object. It compares dot-products of embedded vectors of image, dialogue, and information of candidate objects and classifies the most probable object among the candidates. Question generator is to generate questions reflecting the context of the previous question answering pairs based on the Hierarchical Recurrent Encoder Decoder(HRED) model.

As a follow-up research, [Strub *et al.*, 2017] present end-to-end reinforcement learning optimization for question generation task to find the correct object efficiently. They define GuessWhat?! game as a Markov Decision Process: A state x_t is the tokens generated on the dialogue until time t and an action u_t is to select a new word with zero-one reward depending on the Questioner's choice. They train the question generator with policy gradient and obtain about 17% improvement of accuracy as compared with the baseline model.

GuessWhat?! game have been measured only by the success rate of the game. However, as following Section 3, we

show rule-based search system which attains not only state-of-the-art but also human performance. At the point, we underline that it is not enough to measure the bidirectional communications only by the metrics such as success rate of the game or accuracy and propose criteria for meaningful evaluation on Section 4.

3 Rule-Based Search Systems

3.1 Methods

We constructed rule-based search systems using only the spatial information of the target object for GuessWhat?!. We can divide an image evenly into three parts by two vertical lines and divide each part continuously by horizontal lines or vertical lines in turns (Fig. 2). Then in a current region of interests divided by two vertical lines, we say the left of the left-side region, the right of the right-side region, and N/A of the middle region. Similarly, we say the top, the bottom, and N/A when a region is divided by horizontal lines. Through this protocol or language, we can ask and answer for the location of the center of the target object. Rule-based search systems use this simple language to locate the target object. We may also utilize statistics or the distribution of the spatial information. For the first turn, we may divide an image not evenly. We found that the range of the middle side of 0.18 covers 1/3 target objects, and the other sides of 0.82 covers 2/3 target objects, which means the distribution of target objects was denser in the evenly-divided middle side. Therefore, we set the vertical lines to locate at 0.41 and 0.59.

Given a segmentation model, we can further improve our system. To explore this case, we stole a look at the segmentation information of the candidates, which is supposed to be known when the guesser selects a candidate after a series of question-answer pairs. Then we can implement the binary search based on the spatial information of the candidates. If a segmentation model gives the segmentation similar with that of candidates given in the dataset, we can get an algorithm close to the binary search, which is optimal.

Proposed rule-based search systems do not break the rule of GuessWhat?! game. The spatial information is commonly used by humans as appears in the dataset. Moreover, we can substitute the spatial information with other features or properties. We can choose a real-valued feature without the point mass like the area and the color. Then the feature gives another rule-based search algorithm. This search algorithm is optimal (for the brightness of the center of the target object) or near-optimal (given an optimal segmentation model for the area.) We can also use real-valued features with the point mass or unordered (nominal or categorical) features by choosing a feature that evenly divide the candidates, which gives a near-optimal algorithm.

3.2 Results

[de Vries *et al.*, 2017] and [Strub *et al.*, 2017] constructed the question generator by the hierarchical recurrent encoder-encoder (HRED) and recurrent neural network (RNN) with reinforcement learning, respectively (Table 1). The oracle and the guesser was trained before interacting with the question generator. Therefore, the oracle and the guesser do not benefit from interaction. In contrast, in proposed rule-based search systems, the oracle and the questioner share a set of strict rules. Proposed systems outperformed the state-of-the-art accuracy in two turns and the human accuracy in four or five turns. This result is remarkable improvement of accuracy considering that [de Vries *et al.*, 2017] and [Strub *et al.*, 2017] generated 5 and 8 questions respectively to get their accuracy and that the minimum, mode, mean, and maximum of the number of questions by humans are 1, 3, 5.2, and 24, respectively. Furthermore, we improved the system by tuning the first division of an image, utilizing statistics on spatial information (denoted Fine-Tune). To explore the effect of segmentation, we stole a look at the segmentation information of the candidates (denoted Segment Info).

Table 1: Test accuracy on the GuessWhat?! dataset. Even if the image information is not used, the proposed method outperforms the state-of-the-art deep learning methods in two turns and exceeds human performance in four or five turns. We improved the system by tuning the first division of an image, utilizing statistics on the spatial information (denoted Fine-Tune). To explore the effect of segmentation, we stole a look at the segmentation information of the candidates (denoted Segment Info). The deep learning methods constructed the question generator by the hierarchical recurrent encoder-decoder (HRED) or the recurrent neural network (RNN) with reinforcement learning (RL).

Model	Accuracy
Baseline	16.04
1 Question	38.96
2 Questions	56.25
3 Questions	76.61
4 Questions	85.85
5 Questions	**94.34**
1 Question w/ Fine-Tune	*39.82*
2 Questions w/ Fine-Tune	*59.40*
1 Question w/ Segment Info	*48.12*
2 Questions w/ Segment Info	*87.67*
HRED [de Vries *et al.*, 2017]	46.8
RNN w/ RL [Strub *et al.*, 2017]	52.3
Human [de Vries *et al.*, 2017]	90.8
Human [Strub *et al.*, 2017]	84.4

4 More Criteria

The search system exploits low-level features and strictly follows a predefined set of rules without considering the uncertainty or ambiguity. If either the oracle or the questioner is a human, the search system may not successfully communicate with the human. If the search system takes the part of the questioner, then it may explain how it works at the beginning. If the search system plays the role of the oracle, the search system is unlikely to take the initiative, so it may not have a chance to show how it works. To develop more satisfying systems, we investigate characteristics of the effective AI system for bidirectional communication with humans in

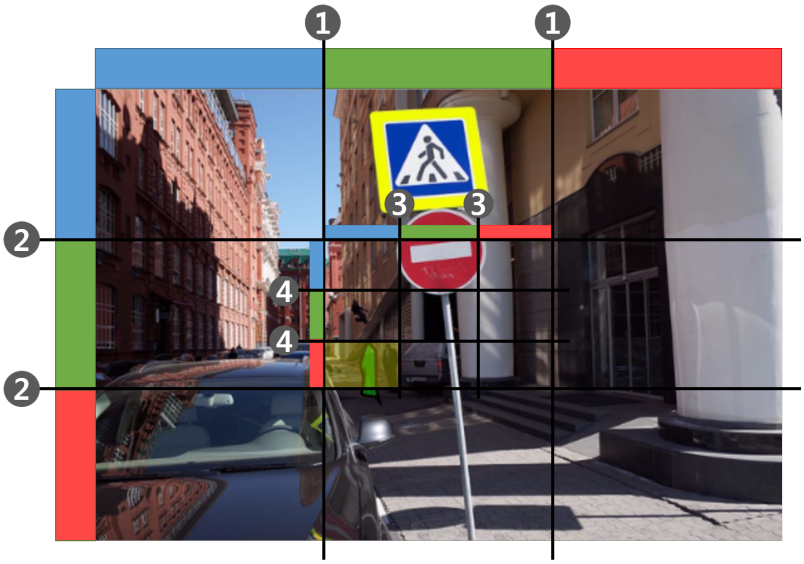

Figure 2: A sequence of divisions of an image by rule-based search systems

vision-language tasks. We will present some considerations in designing such systems. Then we will review some criteria in vision-language tasks and suggest the use of some criteria to be adopted in GuessWhat?!.

To develop effective AI systems for bidirectional communication with humans in vision-language tasks, we need to formulate a problem or task first. Many vision-language tasks have been proposed for this purpose. A task naturally gives a main set of criteria, called objective functions and constraints in optimization. However, this main set may not be enough, then we need to add more criteria. Several criteria in vision-language tasks have been made. We may group these criteria into subjective, task-specific, or similarity criteria.

The subjective evaluation by humans is widely performed in many AI researches as well as vision-language areas. People observe or interact with systems and then evaluate how well or alike humans systems behave. The subjective evaluation is not objective so may not be considered scientific, but the subjective evaluation is crucial because it tells how humans actually feel about the system. However, the subjective evaluation is costly in general.

Task-specific criteria are essential to show how well the system performs in each task. In vision-language tasks, some cross-modal classification or retrieval metrics have been used including accuracy, median rank (mRank), and precision / recall at k (P/R@k) [Kent *et al.*, 1955]. These criteria are measured on data that was collected before constructing the system, so they cost less than subjective criteria, which require additional human efforts.

Similarity criteria evaluate how human-like systems behave. In language generations like machine translations and text summarizations, some language similarity metrics have been used including bilingual evaluation understudy (BLEU) [Papineni *et al.*, 2002], metric for evaluation of translation with explicit ordering (METEOR) [Banerjee and Lavie, 2005], recall-oriented understudy for gisting evaluation (ROUGE) [Lin, 2004], and consensus-based image description evaluation (CIDEr) [Vedantam *et al.*, 2015]. These similarity criteria are calculated systematically, so they cost less than subjective criteria.

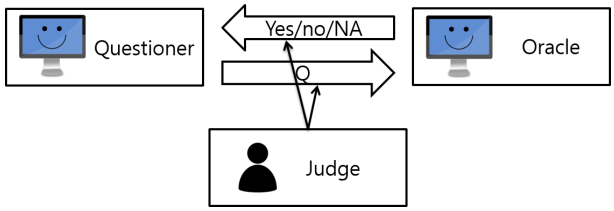

Figure 3: GuessWhat?! setting with a judge for complement or substitution with human evaluation

Adversarial evaluation through neural networks [Bowman *et al.*, 2015; Kannan and Vinyals, 2017; Li *et al.*, 2017] was suggested as a similarity criterion. Unlike previous similarity criteria, the adversarial evaluation is not fixed. Instead, it changes when a neural network called a discriminator learns whether the speaker is a human or not. After learning, the discriminator tells the human from the other for the test data. Since the discriminator is a neural network, it may catch various complex patterns which could not be found by other similarity metrics. However, training the discriminator is necessary unlike other similarity metrics.

Beyond the accuracy, we need to adopt more criteria in GuessWhat?!. Basically, more criteria are considered the better in evaluating vision-language tasks, which is partially because we do not have a unique criterion that everyone agrees with. However, we have limited resources, so we have to choose some criteria among them. The accuracy is the obvious task-specific criterion, but people would not be satisfied only with the accuracy achieved by two AI players. We need to solve the following constrained optimization prob-

lem to develop systems that can communicate with humans in GuessWhat?!. Given an objective functional f (the accuracy in GuessWhat?!), human oracle O_h, and human questioner Q_h, the optimal oracle O^* and questioner Q^* are given by solving

$$\begin{aligned} &\max_{O,Q} f(O,Q) \\ &\text{s.t. } O \text{ is compatible with } Q_h \\ &\quad\ Q \text{ is compatible with } O_h \end{aligned} \tag{1}$$

where, for a threshold $t > 0$,

$$\begin{aligned} &O \text{ is compatible with } Q_h \\ &\quad \text{if } f(O,Q_h) > (1-t)\cdot \max_{O} f(O,Q_h) \\ &Q \text{ is compatible with } O_h \\ &\quad \text{if } f(O_h,Q) > (1-t)\cdot \max_{Q} f(O_h,Q) \end{aligned} \tag{2}$$

This problem is a joint optimization problem and is difficult to solve because it involves two optimization functions and the interaction with human oracles as well as the observation of human questioners. Instead, a two-phase greedy optimization was commonly used in previous works. It involves the observation of human questioners like the previous joint optimization problem, but it involves only one optimization function at each phase and the interaction with an oracle model instead of human oracles.

$$\hat{O} = \operatorname*{argmax}_{O}\ f(O,Q_h) \tag{3}$$

$$\hat{Q} = \operatorname*{argmax}_{Q}\ f(\hat{O},Q) \tag{4}$$

However, $\hat{O}$ and $\hat{Q}$ are marginally optimal, and $\hat{Q}$ may not be compatible with O_h. We may employ human oracles to determine whether $\hat{Q}$ is compatible with O_h even if it costs large. Under the assumption that a questioner Q similar with the human questioner Q_h is compatible with O_h, similarity criteria may complement or substitute with human evaluation. Criteria that distinguish the AI system from humans are necessary for this purpose. The adversarial metric is a promising criterion among them because it involves the neural network, which can learn complex patterns, as a discriminator or judge who determines whether two players are human-like (Fig. 3).

We reviewed some criteria in vision-language tasks and suggested the use of some criteria to be adopted in GuessWhat?!. We grouped the criteria into subjective, task-specific, or similarity criteria and gave some examples. In GuessWhat?!, we need more criteria other than the accuracy. If we are affordable, then we can employ human evaluation. Otherwise, we may choose similarity criteria. We recommended the adversarial evaluation as a promising similarity criterion.

5 Conclusion

We proposed rule-based search systems which used only the spatial information of the target object for GuessWhat?! game. It can be regarded as the artificial language between agents on the GuessWhat?! task. Our rule-based search system outperformed the state-of-the-art accuracy in two turns and the human accuracy in four or five turns. In the view of the results, we argue that we need to measure the performance of the system and analyze details of the results with more concrete criteria not just task-specific metrics such as the success rate of game and accuracy. We suggested the use of criteria for bidirectional communication between humans and agents in vision-language tasks. The adversarial evaluation can be considered as a promising similarity criterion.

Acknowledgments

This work was supported by the Institute for Information & Communications Technology Promotion (2015-0-00310-SW.StarLab) and Korea Evaluation Institute of Industrial Technology (10044009-HRI.MESSI, 10060086-RISF)

References

[Agrawal *et al.*, 2017] Aishwarya Agrawal, Aniruddha Kembhavi, Dhruv Batra, and Devi Parikh. C-vqa: A compositional split of the visual question answering (vqa) v1. 0 dataset. *arXiv preprint arXiv:1704.08243*, 2017.

[Antol *et al.*, 2015] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[Banerjee and Lavie, 2005] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, volume 29, pages 65–72, 2005.

[Bowman *et al.*, 2015] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*, 2015.

[Das *et al.*, 2016] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. *arXiv preprint arXiv:1611.08669*, 2016.

[Das *et al.*, 2017] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. *arXiv preprint arXiv:1703.06585*, 2017.

[de Vries *et al.*, 2017] Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[Fukui *et al.*, 2016] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[Johnson *et al.*, 2016a] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick,

and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. *arXiv preprint arXiv:1612.06890*, 2016.

[Johnson *et al.*, 2016b] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

[Kannan and Vinyals, 2017] Anjuli Kannan and Oriol Vinyals. Adversarial evaluation of dialogue models. *arXiv preprint arXiv:1701.08198*, 2017.

[Kazemzadeh *et al.*, 2014] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014.

[Kent *et al.*, 1955] Allen Kent, Madeline M Berry, Fred U Luehrs, and James W Perry. Machine literature searching viii. operational criteria for designing information retrieval systems. *Journal of the Association for Information Science and Technology*, 6(2):93–101, 1955.

[Kim *et al.*, 2016] Jin-Hwa Kim, Sang-Woo Lee, Donghyun Kwak, Min-Oh Heo, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Multimodal residual learning for visual qa. In *Advances in Neural Information Processing Systems*, pages 361–369, 2016.

[Kiros *et al.*, 2014] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Multimodal neural language models. In *Icml*, volume 14, pages 595–603, 2014.

[Lazaridou *et al.*, 2016] Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. Towards multi-agent communication-based language learning. *arXiv preprint arXiv:1605.07133*, 2016.

[Li *et al.*, 2017] Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*, 2017.

[Lin, 2004] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out: Proceedings of the ACL-04 workshop*, volume 8. Barcelona, Spain, 2004.

[Mao *et al.*, 2016] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20, 2016.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[Strub *et al.*, 2017] Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems. *arXiv preprint arXiv:1703.05423*, 2017.

[Vedantam *et al.*, 2015] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015.

[Vinyals *et al.*, 2015] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[Xu *et al.*, 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, volume 14, pages 77–81, 2015.