Learning Better Visual Dialog Agents with Pretrained Visual-Linguistic Representation

Tao Tu[1], Qing Ping[2,*], Govindarajan Thattai[2], Gokhan Tur[2], Prem Natarajan[2]
[1]National Taiwan University
[2]Amazon Alexa AI
ttaoREtw@gmail.com, {pingqing,thattg,gokhatur,premknat}@amazon.com

Abstract

GuessWhat?! is a visual dialog guessing game which incorporates a Questioner agent that generates a sequence of questions, while an Oracle agent answers the respective questions about a target object in an image. Based on this dialog history between the Questioner and the Oracle, a Guesser agent makes a final guess of the target object. While previous work has focused on dialogue policy optimization and visual-linguistic information fusion, most work learns the vision-linguistic encoding for the three agents solely on the GuessWhat?! dataset without shared and prior knowledge of vision-linguistic representation. To bridge these gaps, this paper proposes new Oracle, Guesser and Questioner models that take advantage of a pretrained vision-linguistic model, VilBERT. For Oracle model, we introduce a two-way background/target fusion mechanism to understand both intra and inter-object questions. For Guesser model, we introduce a state-estimator that best utilizes VilBERT's strength in single-turn referring expression comprehension. For the Questioner, we share the state-estimator from pretrained Guesser with Questioner to guide the question generator. Experimental results show that our proposed models outperform state-of-the-art models significantly by 7%, 10%, 12% for Oracle, Guesser and End-to-End Questioner respectively.

1. Introduction

Multi-modal dialog tasks have gained increasing popularity in recent years such as *GuessWhat?!* [8], *GuessWhich?!* [5], VisDial [7], VDQG [16], vision-and-language navigation R2R [3], ImageChat [27], Alfred [25] and so on. Multi-modal dialog tasks are challenging as the models need to perform high-level image understanding and visual grounding, and such visual grounding should be properly combined with understanding and tracking of multi-turn dialogues in the meanwhile.

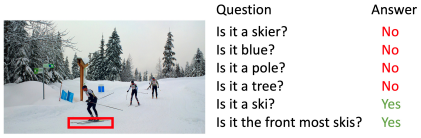

Figure 1: Example of the *GuessWhat?!* dataset

The *GuessWhat?!* dataset is a challenging dataset for a two-player game, where one player will ask a sequence of binary questions and make a final guess for an object in an image designated by another player. The first player performs two sub-tasks, namely as a Questioner to ask questions, and as a Guesser to make the final guess. The second player serves as the Oracle to give Yes/No answer to first player's questions. An example of the *GuessWhat?!* game can be seen in Figure-1. The *GuessWhat?!* game is a good test-bed for such multi-modal tasks such as VQA, referring expression comprehension and generation, and it is also organized in a multi-turn multi-agent dialog. This paper focuses on the three agents for the two players in the *GuessWhat?!* dataset, namely the Oracle model, Guesser model and the Questioner model.

The Oracle task can be considered as an object-aware Visual Question Answering task (VQA), where the inputs are an image, a question, and a pre-defined target object, and the output is an answer of Yes/No/NA depending on whether the question matches the target object. The baseline Oracle model [8] encodes the target object with only category and spatial information but no visual information, which may be insufficient for answering more complex questions about color, shape, relation, actions of an object. To bridge this gap, we introduce VilBERT-Oracle, which takes advantage of the VilBERT model's ability to achieve state-of-the-art performance on VQA tasks [17, 18]. We also introduce a two-way background/target fusion mechanism on top of the VilBERT encoder to learn how to predict

*Corresponding author

correct binary answers with respective to a target object.

The Guesser model can be considered as a special case of referring expression comprehension problem. Given an image and an entire dialog history of questions (referring expression) and its corresponding answers, the Guesser model has to look at the entire dialog and make a final guess. One intuitive solution is to simply concatenate the entire dialog and feed them to the model [9, 29, 26, 18]. However, this might be inadequate if the dialog history is not properly dissected in a way to promote/demote objects according to question and answer in each turn. Recent work [21] introduces object state tracking mechanism, where the belief of all objects is dynamically updated after each turn. Another issue is that almost all existing Guesser models learn the vision-linguistic associations between object and question from scratch on the *GuessWhat?!* dataset, which may be sparse in coverage of referring expressions for new objects. To bridge this gap, we propose VilBERT-Guesser, which is built on top of VilBERT's strength in single-turn referring expression comprehension, and introduces the object state tracking mechanism into VilBERT encoder to learn to update the belief of object states throughout the dialog. To our best knowledge, this is the first work that brings dialog state tracking to large-scale pre-trained vision-linguistic model, which is meant to work only on single-turn text descriptions.

The Questioner model can be considered as a special case of referring expression generation problem. Previous works for Questioner model intuitively encode the image feature and dialog history information to a fused representation, and utilize a language decoder to generate the question [29, 24, 39, 37, 1, 2]. The multi-modal fusion modules are mostly learned from scratch on the *GuessWhat?!* dataset, which may be insufficient for similar reasons as the Guesser models. Moreover, the encoding of the dialog history as a whole, poses challenges for language generator which tends to forget long-term history and generates repeated questions. Recent work introduces state-tracking to Questioner, which dynamically feeds the updated beliefs over objects into Questioner, so that the language generator could generate more targeted questions in each turn[21]. Inspired by this work, we introduce object state estimation mechanism to our VilBERT-Questioner. Moreover, once the VilBERT-Guesser is trained, we load its weights to the state-estimator of VilBERT-Questioner, so that the later could take advantage of the Guesser's ability to make reliable predictions for estimating object states.

Our major contribution of the paper are as follows. First, we propose novel Oracle, Guesser and Questioner models that are built on top of a state-of-the-art vision-linguistic pre-trained model. The proposed models outperform existing state-of-the-art models with significant margins. Second, we propose a unified framework for Guesser and Questioner so that Questioner can take advantage of the robust state-estimator learned from VilBERT-Guesser. Third, we conduct thorough ablation-study and analysis and find that a shared vision-linguistic representation cross the three agents may be beneficial for mutual-understanding and end-game success. Our code is made publicly available. [1]

2. Related Work

2.1. Oracle

The original work for *GuessWhat?!* proposes a baseline Oracle [8] that concatenates question encoding, along with the spatial and category information of the target object, and feeds it into a MLP layer to predict the final answer. However, the baseline Oracle may be insufficient to deal with more challenging visual questions such as colors, shapes, relations, actions and so on, without visual information encoding. The Oracle task can be considered as a special case of Visual Question Answering (VQA) problem with an extra input of object identifier. Methods proposed in [34, 13, 12, 36] achieve competitive performance on VQA tasks. However these models cannot be readily used in this task, unless they are adapted to the extra input of object identifier.

2.2. Guesser

The Guesser model plays an important role in the *GuessWhat?!* game, which should perform both referring expression comprehension on the dialog to describe the visual objects, and perform multi-turn dialog reasoning. Earlier work proposes Guesser models that fuse the encoding of entire dialog with the object category and spatial embedding [8, 29] to predict the target object. Later work in [26, 18, 9] adopted similar approaches and treated entire dialog history as a whole. This might be problematic for two reasons. First, reasoning over such multi-turn dialog is challenging without turn-by-turn explicit dialog state tracking. Second, the lack of turn-level visual grounding can also confuse the Guesser model as to which object the question is referring to in each turn. On the multi-modal representation, original Guesser model encodes no visual information. Some approaches have used image features such as VGG features [28] and Faster-RCNN features[33, 17, 20] into Guesser models, which have shown improvement in accuracy. Recent work in [20, 33] proposes to break down the dialog into turn-level question/answer, and update the final guess with soft state tracking [20], which shows good performance gains.

2.3. Questioner

Questioner plays a key role in the GuessWhat game, since it has to both ask visually meaningful questions and

[1] https://github.com/amazon-research/read-up

guide the dialog towards goal-oriented end-game success rate. [8] proposed the first Questioner model with an encoder-decoder structure where dialog history is encoded with the hierarchical recurrent encoder decoder (HRED) [23] and conditioned on the image which is encoded as fixed-length VGG features [28]. Later work have introduced a shared dialog state encoder for both Guesser and Questioner, where the visual encoder is based on ResNet [10] and an LSTM-based language encoder [11]. More recent work in [21] has incorporated turn-level object state tracking into Questioner, and has shown some improvement in a supervised learning setting. All the approaches mentioned above learn visual grounding and object state-tracking from scratch on the *GuessWhat?!* dataset, which may be insufficient to generalize to new objects/games due to the sparse semantic coverage of objects.

While this paper focuses on using supervised methods for the three models, its worth mentioning there are methods as in [29, 38, 2, 39, 21] that use Reinforcement Learning (RL) approaches to learn Questioner/Guesser model with different variants of end-game success reward.

3. Vision-Linguistic Pretrained Model: Vilbert

Before introducing our VilBERT-based models, we briefly review the model structure of VilBERT [17]. VilBERT is a model for learning task-agnostic joint representation of image content and natural language. Similar as the BERT architecture, VilBERT processes both visual and textual inputs in separate streams then interacts them through co-attention transformer layers [17]. Given an image I represented as a set of object/region features $o_1, o_2, ..., o_M$ and a text input $w_1, w_2, ..., w_L$, the VilBERT model outputs final representations $h_{o1}, h_{o2}, ..., h_{oM}$ for vision information and $h_{w1}, h_{w2}, ..., h_{wL}$ for text information. For more details of VilBERT, please refer to the original work [17]. There are also concurrent work such as VL-Bert [31], Lxmert [32], Oscar [15], UNITER [6] and so on.

4. Proposed Method

In this section, we describe the three models built upon VilBERT, namely VilBERT-Oracle, VilBERT-Guesser and VilBERT-Questioner.

4.1. VilBERT-Oracle Model

The Oracle task can be considered as an object-aware Visual Question Answering task (VQA), where the inputs are an image, a question, and a pre-defined target object, and the output is an answer of (Yes, No, NA) depending on whether the question matches the target object.

The Oracle model structure is illustrated in Figure-2. The model is composed a multi-modal encoder (VilBERT) and a background/target fusion predictor. The multi-modal encoder includes language encoding for a question $\mathbf{q} = \{\texttt{[CLS]}, \mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_L\}$ and vision encoding, which in turn involves both target object encoding $\mathbf{o}_{\text{tgt}}$ and image/all objects encoding $\mathbf{O}_I = \{\texttt{[IMG]}, \mathbf{o}_{\text{tgt}}, \mathbf{o}_1^{\text{pred}}, \mathbf{o}_2^{\text{pred}}, ..., \mathbf{o}_M^{\text{pred}}\}$. Features of $\mathbf{q}$ are word embeddings pretrained by Vilbert model. Features $\mathbf{o}_{\text{tgt}}, \mathbf{o}_1^{\text{pred}}, \mathbf{o}_2^{\text{pred}}, ..., \mathbf{o}_M^{\text{pred}}$ are visual features of target object and all M regions/objects predicted by the object detection model such as Faster-RCNN [22]. The input features are then fed into Vilbert to obtain final hidden states for visual information $\mathbf{H}_O = \{\mathbf{h}_{[IMG]}, \mathbf{h}_{\text{tgt}}, \mathbf{h}_{o_1}, ..., \mathbf{h}_{o_M}\}$ and text information $\mathbf{H}_q = \{\mathbf{h}_{[CLS]}, \mathbf{h}_{w_1}, ..., \mathbf{h}_{w_L}\}$.

For our two-way background/target fusion, we fuse the background image hidden states $\mathbf{h}_{\texttt{[IMG]}}$ and language output $\mathbf{h}_{\texttt{[CLS]}}$, and target object hidden states $\mathbf{h}_{\texttt{tgt}}$ and language output $\mathbf{h}_{\texttt{[CLS]}}$ respectively by taking element-wise multiplication between each pair, and concatenate them with the target object category embedding as fusion result: $\mathbf{x}_{fusion} = (\mathbf{h}_{\texttt{[IMG]}} \odot \mathbf{h}_{\texttt{[CLS]}}) \oplus (\mathbf{h}_{\texttt{tgt}} \odot \mathbf{h}_{\texttt{[CLS]}}) \oplus \mathbf{c}_{cat}$. The final fusion vector $\mathbf{x}_{fusion}$ is fed into multi-layer perceptron followed by softmax to predict the answer p_i. Finally, the loss of VilBERT-Oracle is defined with cross-entropy loss on the three answer classes.

$$L_{VilBERT-Oracle} = -\sum_{i=1}^{N}\sum_{j=1}^{K} y_{i,j} \cdot log(p_{i,j}) \quad (1)$$

Where N is total number of questions, and K is the size of answer classes. Our intuition is that if the question matches the target object, then the fusion of $\mathbf{h}_{\texttt{tgt}}$ and $\mathbf{h}_{\texttt{[CLS]}}$ would give stronger signals. The fusion between $\mathbf{h}_{\texttt{[IMG]}}$ and $\mathbf{h}_{\texttt{[CLS]}}$ is expected to learn to understand object relations that goes beyond what $\mathbf{h}_{\texttt{tgt}}$ can represent.

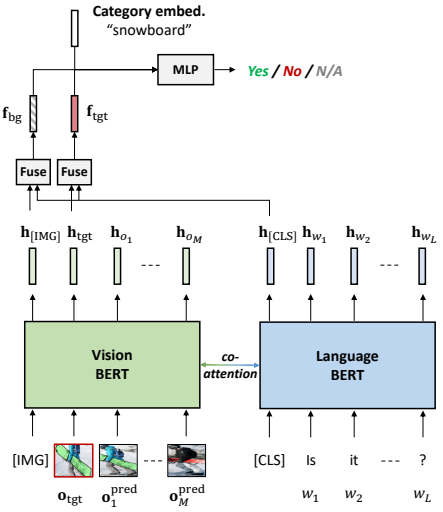

Figure 2: Illustration of Oracle-Vilbert model.

4.2. Vilbert-Guesser Model

The Guesser model can be considered as a special case of referring expression comprehension problem. Given an image I with a set of regions/objects $\{\mathbf{o}_1, ..., \mathbf{o}_N\}$, and an entire dialog history of question (referring expression) and corresponding answers $\{(q_1, a_1), ..., (q_T, a_T)\}$, a Guesser model predicts the likelihood of all objects in the image to be the target object.

The model structure of VilBERT-Guesser is illustrated in Figure-3. The model is composed of a multi-modal encoder (VilBERT), a global image/text fusion layer, a state-weighting layer, and answer-updating layer. Specifically, during each turn, we first feed the visual and language features ($\{\mathbf{o}_1, ..., \mathbf{o}_N\}$, (q_t)) into the model, where (q_t) are the word embeddings of current-turn question. After the VilBERT layers we obtain the final visual hidden states $\{\mathbf{h}_{<\texttt{IMG}>}, \mathbf{h}_{o_1}, ..., \mathbf{h}_{o_N}\}$ and take element-wise multiplication with the sentence-level VilBERT language output $\mathbf{h}_{<\texttt{CLS}>}$ for each visual state to get fused visual output $\mathbf{f}_{o_i}$: $\mathbf{f}_{o_i} = \mathbf{h}_{o_i} \odot \mathbf{h}_{<\texttt{CLS}>}$. Our intuition is that the fused visual output of the object $\mathbf{f}_{o_i}$ that matches the question description encoded by $\mathbf{h}_{<\texttt{CLS}>}$ will have stronger signals compared to irrelevant objects. Next, the fused output for each object is weighted by previous-turn object state belief $\mathbf{p}_t$ to derive $\mathbf{f'}_{o_i}$: $\mathbf{f'}_{o_i} = \mathbf{f}_{o_i} \odot p_{t_i}$, where p_{t_i} is the belief of the ith object in previous turn-t. Next we further update the weighted output of each object by adding the answer embedding of this turn to it: $\mathbf{v}_{o_i} = \mathbf{f'}_{o_i} + \mathbf{a}_t$. Now the final visual output of each object $\mathbf{v}_{o_i}$ should ideally satisfy all the following: (1) object(s) matching the current-turn question should have stronger signals; (2) if the objects have higher likelihood indicated in previous turn, that belief should carry over to the current turn; (3) if the answer is positive/negative, then the belief should be updated to reflect the increased/decreased belief of certain objects. The fundamental basis of all of the above is the robust referring expression comprehension that Vilbert has been pre-trained for.

Eventually the final visual output of each object $\mathbf{v}_{o_i}$ is fed into MLP layer followed by softmax to derive the updated state belief for this turn $\mathbf{p}_{t+1}$, which will be used to re-weight $\mathbf{f}_{o_i}$ for next-turn. We also accumulate the belief states cross-turns for faster convergence: $\mathbf{p}_{t+1} = \alpha \cdot \mathbf{p}'_{t+1} + (1 - \alpha) \cdot \mathbf{p}_t$, where $\alpha \in [0, 1]$ is the state accumulation coefficient. Finally, at the final turn T, Guesser-VilBERT makes a guess by picking the object with highest probability. Therefore the loss of VilBERT-Guesser can be defined as cross-entropy loss over all objects in an image:

$$L_{VilBERT-Guesser} = -\sum_{i=1}^{|D|}\sum_{j=1}^{M} y_{i,j} \cdot log(p_{i,j}) \qquad (2)$$

Where $|D|$ is the number of dialogues, and M is the number of objects in each image. To view a concrete example of how object states are updated turn-by-turn, please refer to our supplementary materials (Figure **??**-**??**).

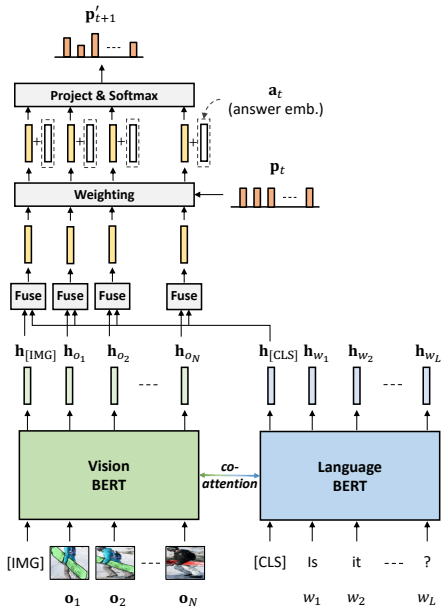

Figure 3: Illustration of Guesser-Vilbert model.

4.3. VilBERT-Questioner Model

The Questioner model can be considered as a special case of referring expression generation problem. Given an image I with K regions/objects $\{\mathbf{o}_1^{\text{pred}}, ..., \mathbf{o}_K^{\text{pred}}\}$ and a dialog history $\{(q_1, a_1), ..., (q_{t-1}, a_{t-1})\}$, the Questioner model is expected to generate a new question q_t that seeks useful information about the target object strategically. The model structure of VilBERT-Questioner is depicted in Figure-4.

The VilBERT-Questioner model is composed of a state-estimator, a state-reweighting layer, a vis-diff layer, and a question generator. In each turn, starting from uniformly distributed object states p_0 , we first re-weight visual features of all objects $\{\mathbf{o}_1^{\text{pred}}, ..., \mathbf{o}_K^{\text{pred}}\}$ with the last-turn object states p_{t-1} . Then we feed the re-weighted object visual features into vis-diff module to derive the most distinctive feature of each object relative to others and merge the representation to $\mathbf{v}_t$ [35, 21]. Then the language decoder (LSTM [11]) generates question conditioned on the encoder output $\mathbf{v}_t$.

The generated question together with its corresponding answer, is fed into the state-estimator, which is the pre-trained VilBERT-Guesser, to get updated state belief p_{t+1} as input to next-turn Vilbert-Questioner encoder. The loss function of the VilBERT-Questioner model is defined as fol-

lows:

$$L_{VilBERT-Questioner} = -\sum_{i=1}^{T}\sum_{j=1}^{L}\sum_{k=1}^{|V|} y_{i,j,k} \cdot log(p_{i,j,k}) \tag{3}$$

Where T is the number of turns in the dialog, L is the length of questions, and $|V|$ is the size of vocabulary for the language decoder.

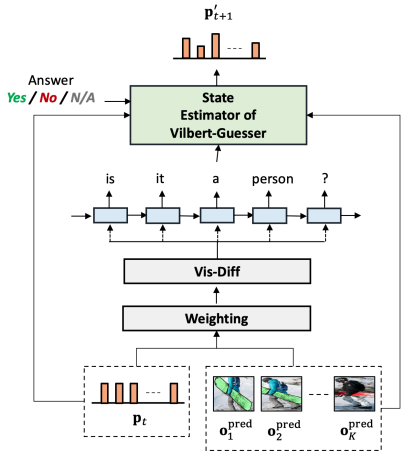

Figure 4: Illustration of Vilbert-Questioner model.

5. Experiments

5.1. Dataset

The *GuessWhat?!* dataset [8] contains 155k dialogues with 821k question-answer pairs on 66k unique images with 134k unique objects. The answers are 52.2%, 45.6% and 2.2% for (Yes,No,N/A) respectively. 84.6% of the dialogues are successful games. We use the partitioned datasets with training (70%), validation (15%) and test (15%) in all of our experiments, as specified in the original *GuessWhat?!* dataset [8].

5.2. Evaluation Metrics

Independent Accuracy Independent accuracy refers to the percentage of correct predictions one agent achieves by isolating it on the ground-truth data without interacting with other agents. The Oracle and Guesser models can be evaluated independently.

End-to-end Success Rate. The Questioner models cannot be evaluated independently due to its dependency on dynamically generated dialog history. The Questioner models can only be evaluated jointly by having all three agents play the *GuessWhat?!* game together and measuring the success rate at the end of the game, which is the percentage of games where the Guesser model makes correct guesses based on generated dialogues.

Semantic Diversity and Rate of Games with Repeated Questions. One common problem for Questioner models is the generation of repeated questions within a game. Repeated questions reduces the opportunities to ask more meaningful questions. Therefore, we measure the percentage of games with at least one repeated question, as in previous work [24].

5.3. Experiment Settings

The backbone of the VilBERT encoder [17] in all our models is adapted from the official VilBERT implementation [2]. We use Faster R-CNN model [22] for feature extraction, pre-trained on Visual Genome dataset [14] with ResNet-101 backbone [10]. Relative coordinates and area are concatenated with feature vectors before feeding them into VilBERT. For the VilBERT-Oracle, the category embedding size is 512, and the number of bounding boxes M is 100. For the VilBERT-Guesser, the answer embedding size is 128 and the state accumulation coefficient α is 0.9. For the VilBERT-Questioner decoder, the word embedding size is 512 and the number of bounding boxes K is 100.

5.4. Results

5.4.1 The Oracle Model

To the best of our knowledge, all existing work use the same baseline Oracle [8] except [30]. We compare the performance of the baseline oracles with the proposed VilBERT-Oracle. Further, we also modify the baseline Oracles by introducing image-level and object-level Faster-RCNN features as extra input for predicting answer.

From the results in Table-1, we observe the following. First, introducing visual features increases the accuracy of baseline Oracle. This is intuitive since the baseline Oracle only relies on category/spatial information and not any visual information to predict any answer, it is prone to errors on challenging questions that need robust visual grounding. Second, VilBERT-Oracle further outperforms baseline Oracle, RCNN-Oracle and MultiHop Oracle. We attribute this to the two-way global/target fusion on top of VilBERT encoder, which not only supports matching of the target object with descriptive question, but also helps with contextually capturing the underlying relationships between objects in the image, therefore making it easier to answer more complex questions such as *is it to the left of the women in red?*.

To corroborate to this observation, we further break down Oracle models' performances across different types of questions same as previous work [24], as in column 2 and 3 in Table-4. From the table we see that VilBERT-Oracle

[2] https://github.com/facebookresearch/vilbert-multi-task

Oracle Models	Accuracy
Baseline Oracle [8]	78.5%
RCNN-Oracle (Ours)	81.7%
Multi-hop FiLM Oracle[30]	83.1%
VilBERT-Oracle (Ours)	**85.0%**

Table 1: Comparison of Oracle Models (Independent Accuracy Evaluation)

indeed performs much better on all types of questions other than object type compared to baseline Oracle by 8% - 19%.

5.4.2 The Guesser Model

For Guesser models, we compare proposed VilBERT-Guesser with a comprehensive set of baselines and SOTAs under an independent evaluation setting. The input is an image along with the entire dialog history of question/answer pairs, and the output is the target object, all from ground-truth.

From the results in Table-2, we observe the following. First, Guesser models that utilize image features (VilBERT [18], GST [20], ATT-R4 (w2v) [9], and HACAN [33]) achieve slightly higher accuracy compared to models that include no visual information (LSTM [9], Guesser [29], and RIG [26]). Second, Guesser models that encode text at turn-level instead of dialog-level shows slightly better performance (GST[20] and HACAN [33]). Third, our VilBERT-Guesser outperforms all baseline models by an absolute margin of 10%. Intuitively, the VilBERT-Guesser model encodes visual information of objects through pre-trained vision-linguistic layers of VilBERT; the turn-level state tracking and state accumulation mechanism, update the belief state with information from the VilBERT output, as it was used for referring expression comprehension. Both of these factors contribute to the improvement over state-of-the-art Guesser model. Please see supplementary material for example of object state update process.

5.4.3 The Questioner Model

End-to-end success rate. Table-3 compares different Questioner models in end-to-end self-play games based on success rate. The dialog sessions were generated by making Questioner and Oracle talking to each other, and having Guesser to make a final guess about the target object. Rows 1-5 correspond to baseline and SOTA Questioner model results, rows 6-10 refer to different combinations of proposed models, and rows 11-12 are two variants of the VilBERT-Questioner model.

From the results (row 1-10), we observe the following. First, the state-of-the-art Questioner (VDST) only

Guesser Models	Accuracy
Mask-RCNN (no gt bbox) [4]	57.9%
LSTM [8]	61.3%
PLAN [40]	63.4%
Guesser [29]	63.8%
RIG [26]	64.2%
12-in-1 Vilbert [18]	65.7%
GST [20]	65.7%
ATT-R4 (w2v) [9]	65.8%
HACAN [33]	66.8%
Multi-hop FiLM Guesser [30]	69.5%
Vilbert-Guesser (Ours)	**76.5%**

Table 2: Comparison of Guesser Models (Independent Accuracy Evaluation)

achieves slightly improvement over baseline Questioner (45.9% over 44.6%) when collaborating with baseline Oracle and Guesser. On the other hand, when combined with state-of-the-art Guesser (GST), VDST achieves much higher performance (50.6%). Our speculation is that GST and VDST share very similar model structures, which promotes mutual understanding in self-play games [21, 20]. Similarly for GDSE-SL, the Guesser and Questioner share the same encoder that encodes visually grounded dialog states. This might promote mutual understanding of the two models in end-to-end games. Second, our VilBERT-Questioner, combined with baseline Oracle and Guesser (row-8), also achieves higher end-to-end success rate over VDST (52.5% over 45.9%). Third, for row 9 and 10, when we introduce two or three proposed models into the game, the performance continuously improves (55.7% and 62.8% respectively). As contrast, as in row 6 and 7, when only VilBERT-Oracle or VilBERT-Guesser is introduced in the game, the improvement is minimal. We will discuss details in Ablation Study.

Please note that we did not include RL-based models for comparison of success rate, since this work focuses on supervised learning. Previous work also show evidence that RL-based Questioner models may generate unnatural questions as indicated by skewed distribution across different visual attributes [19, 24]. As can be seen in Table-3, RL [29] and VDST [21], which are both RL-based models tend to generate more questions about object and location, while BL [8], SL [24], CL[24], and our VilBERT-Questioner have less skewed distributions across different question types.

Rate of Games with Repeated Questions. One common issue for Questioner models is question repetition [24]. We compare the rate of games with repeated questions, for the baseline Questioner models and our VilBERT-Questioner. The results are reported in Table-5. From the

results we can see that our VilBERT-Questioner has the lowest rate of games with repeated questions.

6. Ablation Study

6.1. Variants of Guesser Model

In this paper, we decompose the dialog history to turns, and feed question representation to the VilBERT-Guesser encoder, instead of concatenating the entire dialogue history together [9, 29, 26, 18]. We compare two methods, namely injecting answer information to the post-layers of the state-estimator (**Post-Fusion**), versus concatenating a pair of question and answer as input to the model (**Pre-Concatenation**).

From Table-6 we observe that both variants outperforms the previous work [18] significantly by 9%-11%. Second, the Post-Fusion method further outperforms Pre-Concatenation method by 2%. We hypothesize that pretrained vision-linguistic models like VilBERT may be best used to its strength if we isolate the input to a way similar to how the model has been pretrained, in this case, on single-turn descriptive questions.

6.2. Variants of Questioner Model

For Vilbert-Questioner, the weights of state-estimator is loaded from pretrained VilBERT-Guesser. We compare two variants: fine-tuning the state-estimator together with question generator (**w/ fine-tune**), versus freezing the weights of state-estimator while training the rest of the Vilbert-Questioner (**w/o fine-tune**). These results are reported in row 11-12 of Table-3. The results show that the VilBERT-Questioner **w/ fine-tune** performs poorer than the VilBERT-Questioner **w/o fine-tune**. This is intuitive since the state-estimator learned by VilBERT-Guesser is already good at inferring object states turn-by-turn, fine-tuning it with the question generation objective may confuse the state-estimator to go astray from predicting the correct states and thus make the end-to-end performance worse.

6.3. Variants of Combinations of the Three Models

As is shown in the ablation study (Table-3 row 6,7 and 9), when only the VilBERT-Oracle or VilBERT-Guesser is introduced to the end-to-end game, the overall accuracy has shown minor improvement over the baselines. Whereas when both are introduced, the end-to-end accuracy has improved significantly (55.7% over 50.6%). We investigate the possible causes as follows.

When Oracle model's performance is very poor, even a highly accurate Guesser may not achieve a high end-to-end success rate. To simulate this effect, we keep a fixed ground-truth dialog and add random errors to the answer data with varying levels (10%-90%) by toggling the yes/no answer. We run both the baseline guesser and our VilBERT-Guesser on this corrupted dataset to study model performance deterioration. The results are reported in Figure 5. When the corruption ratio is low (10%-40%), VilBERT-Guesser consistently outperforms the baseline Guesser. When the Oracle data is corrupted to a large ratio ($\geq$ 50%), the accuracy of VilBERT-Guesser drops faster than baseline model, suggesting that it is more sensitive to the correctness of the Oracle output. This partially explains why row-6 shows little improvement over row-2 in Table-3, since in both settings the baseline Oracle is expected to have lower accuracy.

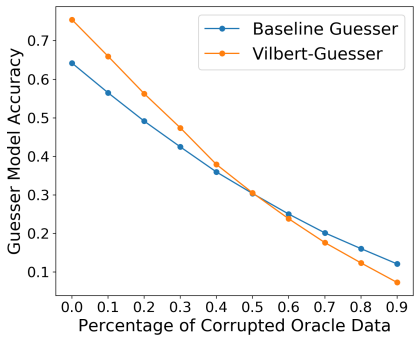

Figure 5: Performance of Two Guesser Models on Corrupted Oracle Data with Varying Corruption Ratios

Second, only introducing a better Oracle, while keeping a Guesser that is less sensitive to the correct/wrong answers, may also not help end-to-end success rate, as indicated in row-7 versus row-2 in Table-3. To demonstrate this, we compute a confusion matrix (Table-7) between baseline Guesser and VilBERT-Guesser, both run on dialogs generated with VilBERT-Oracle and VDST [21]. From the table, it is clear that given a better Oracle (VilBERT-Oracle), VilBERT-Guesser is able to make more correct guess compared to a baseline Guesser. For more details of baseline Guesser and VilBERT-Guesser model behavior, please refer to examples in our supplementary materials.

Examining row 6-10 together, we argue that sharing a similar visual-linguistic encoder across the three agents may be beneficial for the end-to-end game.

7. Conclusion

In this paper, we propose three novel models, VilBERT-Oracle, VilBERT-Questioner and VilBERT Guesser for the *GuessWhat?!* game. The proposed models take advantage of a pretrained visual-linguistic encoder (VilBERT[17]) that has shown state-of-the-art performance in multiple vision-language tasks especially in VQA and referring expression comprehension. A state-estimator is introduced to the Guesser and Questioner model to handle object state update turn-by-turn. Experimental results show

	Oracle	Guesser	Questioner	Success Rate
1	Baseline Oracle [8]	Baseline Guesser [29]	Baseline Questioner [29]	44.6%
2	Baseline Oracle [8]	Baseline Guesser [29]	VDST [21]	45.9%
3	Baseline Oracle [8]	GDSE-SL [24]	GDSE-SL [24]	47.8%
4	Baseline Oracle [8]	Guesser (MN) [39]	TPG [39]	48.8%
5	Baseline Oracle [8]	GST [20]	VDST [21]	50.6%
6	Baseline Oracle [8]	**Vilbert-Guesser (Ours)**	VDST [21]	47.5%
7	**Vilbert-Oracle (Ours)**	Baseline Guesser [29]	VDST [21]	47.8%
8	Baseline Oracle [8]	Baseline Guesser [29]	**VilBert-Questioner (Ours)**	52.5%
9	**Vilbert-Oracle (Ours)**	**Vilbert-Guesser (Ours)**	VDST [21]	55.7%
10	**Vilbert-Oracle (Ours)**	**Vilbert-Guesser (Ours)**	**Vilbert-Questioner (Ours)**	**62.8%**
11	**Vilbert-Oracle (Ours)**	**Vilbert-Guesser (Ours)**	**Vilbert-Questioner (w/ fine-tune)**	57.0%
12	**Vilbert-Oracle (Ours)**	**Vilbert-Guesser (Ours)**	**Vilbert-Questioner (w/o fine-tune)**	**62.8%**

Table 3: Comparison of Different Questioner Models in End-to-End Evaluation

Type	Baseline Oracle [8]	VilBERT-Oracle	BL[8]	SL [24]	CL [24]	RL [29]	VDST [21]	VilBERT-Questioner	Human
Object	94%	94%	49.00	48.08	46.40	24.00	36.44	65.23	38.12
Color	63%	**82%**	2.75	13.00	12.51	0.12	0.01	9.1	15.50
Shape	67%	**75%**	0.00	0.01	0.02	0.00	0.00	0.00	0.30
Size	60%	**77%**	0.02	0.33	0.39	0.02	0.01	0.01	1.38
Texture	70%	**83%**	0.00	0.33	0.15	0.01	0.00	0.00	0.89
Location	67%	**77%**	47.25	37.09	38.54	74.80	64.80	25.60	40.00
Action	65%	**81%**	1.34	7.97	7.60	0.66	0.30	5.04	7.59
Other	75%	**82%**	1.12	5.28	5.90	0.49	0.03	1.95	8.60

Table 4: Oracle Accuracy by Types of Question and Question Distribution for Models.

Questioners	% Games with Repeated Q's
GDSE-BL [8]	93.50
GDSE-SL [24]	55.80
CL [24]	52.19
RL [29]	96.47
VDST [21]	40.05
VilBERT-Questioner	**32.56**
Human	N/A

Table 5: Rate of Games with Repeated Questions of Different Questioner Models. Note: * the VDST Model Used Here is the Model Trained in Supervised Learning Setting.

Models	Accuracy
12-in-1 VilBERT: Concatenate Entire Dialog	65.7%
Answer Pre-Concatenation (Ours)	74.3%
Answer Post-Fusion (Ours)	**76.5%**

Table 6: Different Variants of Answer Fusion

		VilBERT-Guesser		
		Correct	Wrong	Total
Baseline Guesser	Correct	7565	1993	47.8%
	Wrong	3573	6864	52.2%
	Total	55.7%	44.3%	

Table 7: Confusion Matrix of End-to-End Dialogue Success Rate Generated from Baseline Guesser and VilBERT-Guesser Using VilBERT-Oracle.

that feeding the VilBERT model with turn-level text description is better than feeding a long dialog history, in accordance with how the VilBERT model has been pretrained. Ablation study suggests that a shared vision-linguistic encoder may be beneficial for such three-agent games. For future work, we plan to explore different reinforcement learning approaches with Questioner and Guesser to further improve end-to-end success rates.

References

[1] Ehsan Abbasnejad, Qi Wu, Iman Abbasnejad, Javen Shi, and Anton van den Hengel. An active information seeking model for goal-oriented vision-and-language tasks. *arXiv preprint arXiv:1812.06398*, 2018.

[2] Ehsan Abbasnejad, Qi Wu, Qinfeng Shi, and Anton van den Hengel. What's to know? uncertainty as a guide to asking goal-oriented questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4155–4164, 2019.

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

[4] Gabriele Bani, Davide Belli, Gautier Dagan, Alexander Geenen, Andrii Skliar, Aashish Venkatesh, Tim Baumgartner, Elia Bruni, and Raquel Fernández. Adding object detection skills to visual dialogue agents. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[5] Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. Evaluating visual conversational agents via cooperative human-ai games. *arXiv preprint arXiv:1708.05122*, 2017.

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

[7] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

[8] 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*, pages 5503–5512, 2017.

[9] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7746–7755, 2018.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[11] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[12] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[13] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018.

[14] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[15] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[16] Yining Li, Chen Huang, Xiaoou Tang, and Chen Change Loy. Learning to disambiguate by asking discriminative questions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3419–3428, 2017.

[17] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[18] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[19] Mauricio Mazuecos, Alberto Testoni, Raffaella Bernardi, and Luciana Benotti. On the role of effective and referring questions in guesswhat?! In *Proceedings of the First Workshop on Advances in Language and Vision Research*, pages 19–25, 2020.

[20] Wei Pang and Xiaojie Wang. Guessing state tracking for visual dialogue. *arXiv preprint arXiv:2002.10340*, 2020.

[21] Wei Pang and Xiaojie Wang. Visual dialogue state tracking for question generation. In *AAAI*, pages 11831–11838, 2020.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[23] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*, 7(8):434–441, 2015.

[24] Ravi Shekhar, Aashish Venkatesh, Tim Baumgärtner, Elia Bruni, Barbara Plank, Raffaella Bernardi, and Raquel Fernández. Beyond task success: A closer look at jointly learning to see, ask, and guesswhat. *arXiv preprint arXiv:1809.03408*, 2018.

[25] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10740–10749, 2020.

[26] Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, Matthew Turk, and William Yang Wang. What should i ask? using conversationally informative rewards for goal-oriented visual dialog. *arXiv preprint arXiv:1907.12021*, 2019.

[27] Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. Image-chat: Engaging grounded conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2414–2429, 2020.

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[29] Florian Strub, Harm de Vries, Jeremie Mary, Bilal Piot, Aaron Courvile, and Olivier Pietquin. End-to-end optimization of goal-driven and visually grounded dialogue systems.

[30] Florian Strub, Mathieu Seurin, Ethan Perez, Harm De Vries, Jérémie Mary, Philippe Preux, and Aaron CourvilleOlivier Pietquin. Visual reasoning with multi-hop feature modulation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 784–800, 2018.

[31] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[32] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[33] Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. Making history matter: History-advantage sequence training for visual dialog. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2561–2569, 2019.

[34] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016.

[35] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[36] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6281–6290, 2019.

[37] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton van den Hengel. Asking the difficult questions: Goal-oriented visual question generation via intermediate rewards. *arXiv preprint arXiv:1711.07614*, 2017.

[38] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, Jianfeng Lu, and Anton Van Den Hengel. Goal-oriented visual question generation via intermediate rewards. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 186–201, 2018.

[39] Rui Zhao and Volker Tresp. Improving goal-oriented visual dialog agents via advanced recurrent nets with tempered policy gradient. In *LaCATODA IJCAI*, 2018.

[40] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4252–4261, 2018.