Find Someone Who: Visual Commonsense Understanding in Human-Centric Grounding

Haoxuan You[1], Rui Sun[1], Zhecan Wang[1], Kai-Wei Chang[2], Shih-Fu Chang[1]
[1] Columbia University, New York
[2] University of California, Los Angeles
{hy2612, rs4110, zw2627, sc250}@columbia.edu, kwchang@cs.ucla.edu

Abstract

From a visual scene containing multiple people, human is able to distinguish each individual given the context descriptions about what happened before, their mental/physical states or intentions, *etc*. Above ability heavily relies on human-centric commonsense knowledge and reasoning. For example, if asked to identify the "person who needs healing" in an image, we need to first know that they usually have injuries or suffering expressions, then find the corresponding visual clues before finally grounding the person.

We present a new commonsense task, *Human-centric Commonsense Grounding*, that tests the models' ability to ground individuals given the context descriptions about what happened before, and their mental/physical states or intentions. We further create a benchmark, *HumanCog*, a dataset with 130k grounded commonsensical descriptions annotated on 67k images, covering diverse types of commonsense and visual scenes. We set up a context-object-aware method as a strong baseline that outperforms previous pre-trained and non-pretrained models. Further analysis demonstrates that rich visual commonsense and powerful integration of multi-modal commonsense are essential, which sheds light on future works. Data and code will be available at https://github.com/Hxyou/HumanCog.

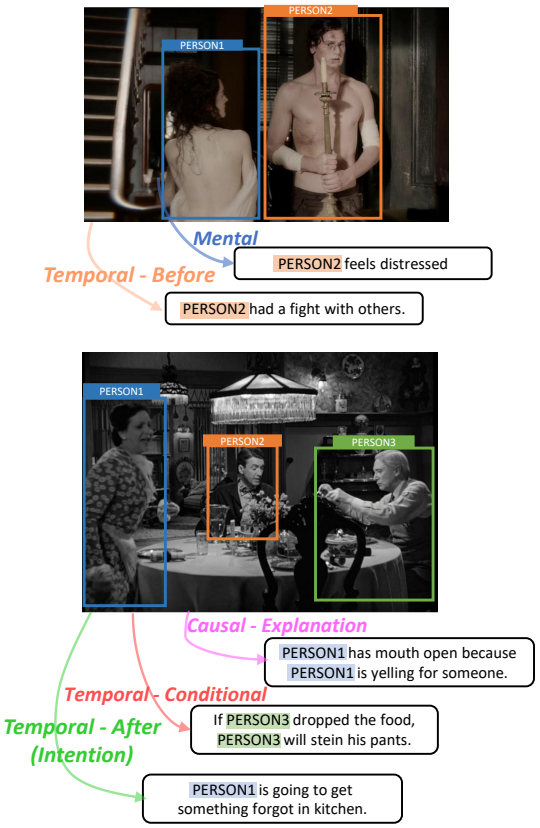

Figure 1: *Human-centric Commonsense Grounding*: given an image, a set of candidate person boxes and a human-centric commonsensical description, a model must ground the persons in description to correct person boxes in image.

1 Introduction

Visual scenes often involve multiple people. For instance, as in movies, a frame can involve multiple characters. Complex human interaction happens because different people may have different intentions, roles, and emotions. When observing such scenes, humans can understand the scene and differentiate the characters from each other according to the context description, such as what will happen/happened to them, attributes, mental/physical states, and intentions.

Take the bottom image in Fig. 1 as an example. When the context description talks about the figure who is "going to get something forgot in the kitchen", humans can match that description to PERSON1 because PERSON1 is standing up, looking around, and speaking. We can also connect dots via the understanding that PERSON2 and PERSON3 are sitting and focusing on eating food without any sign of leaving. We can achieve this because we understand the commonsense, such as causal, temporal, mental, *etc*., behind the human

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 5444–5454
December 7-11, 2022

interactions and the subtle visual clues can serve as hints to identify them.

Understanding human-centric commonsense relations is important in widely broad fields. *e.g.* in human-robot communication, it's crucial for medical-aid robots to identify "person who needs healing" and take actions to help. Despite the importance and challenges, lack of development exists for this task. Existing works are limited to conventional visual object grounding. The state-of-the-art grounding models can ground objects by the description of their geometric/spatial relation and appearance (Mao et al., 2016; Lin et al., 2014; Plummer et al., 2015), and ground people by actions (Cui et al., 2021), but struggle with complex scenes requiring human-centric commonsense knowledge and sophisticated reasoning ability. Meanwhile, tasks that focus on evaluating commonsense reasoning often take the form of multi-choice QA (Zellers et al., 2019) or free-form generation (Park et al., 2020). Such formulation tends to offer less interpretability and could contain easy shortcuts. We formulate our task as a grounding task, with a simple output format (i.e., finding the alignment between humans and bounding boxes) while covering a wide range of commonsense understanding.

In this paper, we formulate the task as **Human-centric Commonsense Grounding**. Given an image containing multiple candidate persons and a commonsensical description (including temporal, causal, mental , *etc.* human-interaction), a machine must ground the persons mentioned in the description to correct person boxes in the image. To back the study of this task, we introduce **HumanCog**, a new dataset with 130k "commonsensical" descriptions where context descriptions are constituted with human-centric commonsense relations like mental states, intentions, *etc.*. Those relations with associated pronouns can all be grounded to 230k persons in 67k images, covering diverse visual scenes. *HumanCog* is automatically collected by transferring the questions and correct answers in Visual Commonsense Reasoning (VCR) to grounded statements through a set of pre-defined rules, which also preserves the paired person-box groundings (*i.e.*, co-reference links). Since the questions in VCR require commonsense to answer, our transferred grounded statements also cover various types of commonsense needed to ground the persons. Further we employ NLP specialists to iteratively refine the rules until reaching a acceptably low error rate. Moreover, the validation and test sets are verified by Amazon Mechanical Turk, and the result testifies the preciseness of our annotation.

We introduce a context-object-aware method as a strong baseline on this task based on pre-trained vision and language Transformer architecture (Chen et al., 2019). We take the candidate person region features as weights in classifier and classify the person tokens in text by cross-entropy loss. To facilitate the interaction between people and visual scenes, detected context objects are also input to the model. Further, at the feature level, we draw the person tokens in text and neighbor context objects pertaining to corresponding persons, while push them away from other persons, through a proposed context contrastive loss.

Comprehensive experiments are conducted with a wide range of methods, from heuristic methods to pre-trained models. We further present both qualitative and quantitative analysis and find that rich contextualized visual representation, effective usage of context objects, and better integrating vision and text by vision-language pre-training, are the keys to improve the performance.

In summary, our main contribution is threefold. (1) We introduce a new task, *human-centric commonsense grounding*, to ground the persons mentioned in commonsensical descriptions. (2) A large-scale dataset, *HumanCog*, containing 130k commonsensical descriptions on 67k images. (3) A context-object-aware model to facilitate the visual commonsense learning, establishing a strong baseline on our new challenge.

2 Related Work

Visual Grounding Dataset Ground the corresponding regions in images given text information is an essential task to bridge image and text modalities. There are in general two conventional settings in existing grounding datasets. In the first setting, the entire sentence is mainly describing one object and its environment/attribute, and only refers to one box in image, which is termed as Referring expression comprehension (REC). RefCOCO (Yu et al., 2016), RefCOCO+ (Yu et al., 2016) and RefCOCOg (Mao et al., 2016) are commonly used REC datasets annotated on top of MSCOCO (Lin et al., 2014). RefCOCO and RefCOCO+ are annotated using ReferIt Game (Kazemzadeh et al., 2014), but RefCOCO+ focuses more on appearance description since location words are not al-

lowed. RefCOCOg is collected by Mechanical Turks in a non-interactive setting. CLEVR-Ref+ (Liu et al., 2019) is collected on synthesized images where objects of different attributes are put on plane. KB-REF (Wang et al., 2020) enriches the sentence description by injecting knowledge retrieved from the external knowledge base. Flick30k Entity (Plummer et al., 2015) is the pioneer to establish the second setting: multiple phrases inside one caption can be grounded to different boxes. Who's Waldo(Cui et al., 2021) further studies grounding the persons in sentence. Our work differs from Who's Waldo and KB-REF in that the descriptions contain rich human-centric commonsense such as temporal, causal, mental, *etc.*

Visual Grounding Approaches Current methods can be divided into two categories: one-stage and two-stage(Qiao et al., 2020). In two-stage methods, a set of candidate object regions are first detected by object detection models, then multimodal models are used to predict the links between detected boxes and text. LSTM-based models (Luo and Shakhnarovich, 2017; Hu et al., 2017), attention-based models(Yu et al., 2018; Kim et al., 2018; Fukui et al., 2016), graph-based models (Liu et al., 2020; Yang et al., 2019) and pre-trained models (Li et al., 2019; Lu et al., 2019; Chen et al., 2019) are explored. In one-stage models, the coordinates of grounded object box are directly predicted by a single model (Liao et al., 2020; Deng et al., 2021; Kamath et al., 2021).More can be found in a survey (Qiao et al., 2020). Since our dataset already gives ground-truth person candidate boxes, *i.e.*, the first stage results in two-stage schema are provided, we mainly focuses on building better model for image-text understanding (second stage in two-stage schema).

Multimodal Commonsense Reasoning Multimodal commonsense reasoning has attracted wide research interest in recent years. VCR (Zellers et al., 2019) introduces commonsense question that requires a deep understanding of both image and text, and is formulated as a multi-choice answering task. VisualCOMET (Park et al., 2020) focuses on inferring the temporal and causal information given current image and description, regarded as a generation task. VLEP (Lei et al., 2020) also requires machine to predict future event but is in multi-choice answering format. We are similar to VCR in that we collect images from VCR and transform the questions&answers in VCR to statements. However, we differs from above works in that we target at the human-centric commonsense grounding ability of machines.

3 Task: Human-centric Commonsense Grounding

We present a challenging task, *human-centric commonsense grounding*, to mimic the inference ability of humans to distinguish wanted persons in image by corresponding commonsensical description. The input of one sample in this task includes: (1) An image I. (2) A set of N ($N \geq 2$) candidate person boxes $\boldsymbol{r}$, covering all the persons in the image. (3) A commonsensical description $\boldsymbol{t} = \{t_i\}_{i=1}^{n}$ of the image, where n is the token number, *e.g.*, "PERSONX feels distressed." t_i is either a token in vocabulary or a person link (PERSONX in above example) that remains to be grounded/referred to ground-truth person in image. At least one person link exists.

Given above input, the goal of the task is to ground/refer the person links to corresponding correct person boxes out of all candidate person boxes, *i.e.*, $\arg\max_{\boldsymbol{r}} f(t_i|\boldsymbol{t}, I, \boldsymbol{r}) = r_j, \{t_i, r_j\} \in \mathcal{L}$, where $\mathcal{L}$ is the set of ground-truth reference pairs and f is the desired model. We evaluate the accuracy of correct prediction among candidate person boxes.

Take Fig. 1 as an example. For the top picture, a commonsensial description is "PERSONX feels distressed", where "PERSONX" is an person link, and its corresponding ground-truth person in image should be "PERSON2". It's noted that there might be more than one person links referring to the same person in image (see the bottom picture in Fig. 1).

4 Dataset Collection and Analysis

The *HumanCog* dataset contains 130k commonsensical descriptions on 67k images, where in total 230k persons are grounded. In the following, we describe how *HumanCog* is constructed and annotated, and provide detailed analysis of the dataset.

4.1 Data Collection

To support the research of human-centric commonsense grounding task, we hope the samples in dataset should have two properties: (1) cover a wide range of visual scenes, (2) have rich and practical commonsense in the description. Although employing annotators to annotate from scratch is a

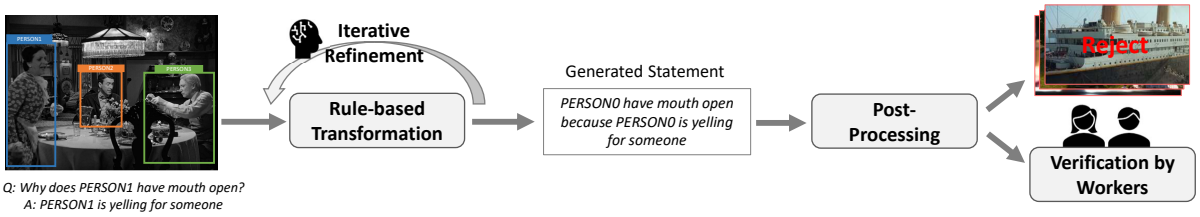

Figure 2: Diagram of data collection.

Question: What will PERSON3 do if PERSON3 dropped the food?
Answer: He will stein his pants

Step1: ***Detect If-condition in Q; Delete will-support.***
Generated: What ~~will~~ PERSON3 ~~do~~ if PERSON3 dropped the food?

Step2: ***Reverse wh-movement.***
Generated: What PERSON3 if PERSON3 dropped the food?

Step3: ***Delete wh-movement.***
Generated: PERSON3 ~~What~~ if PERSON3 dropped the food?

Step4: ***Concatenate Answer.***
Generated: PERSON3 will stein his pants if PERSON3 dropped the food?

Step5: ***Switch Order; Delete/Add marks.***
Generated: If PERSON3 dropped the food, PERSON3 will stein his pants.

Figure 3: An example of applying one of the rules to a sample.

feasible way, it might be costly to build such a large-scale dataset. Instead, we find VCR (Zellers et al., 2019) a perfect base for our usage, because VCR's images are from movie clips depicting complex and diverse situations, and its questions and answers are carefully annotated by turkers in free-form focusing on various commonsense. To build *HumanCog* dataset, we extend and tailor VCR dataset by following steps: (1) Rule-based Transformation with Iterative Refinement (2) Post-processing (3) Validation with Amazon Turker.

Transformation via rules with Iterative Refinement Since the task of VCR is question answering, the text part of each sample in VCR contains one question and four answer choices. To extend VCR for our task, following (Demszky et al., 2018), we transform the questions and answers to statements/descriptions via a set of rules. More specifically, in each example, among four answer choices, we take the correct answer a_i^{gt} and question q_i as input $\{q_i, a_i^{gt}\}$, since the other answer choices are semantically wrong or irrelevant. Assume we have a set of Q pre-defined rules $\{T_i\}_{i=1}^{Q}$. During transformation, we first examine whether each sample $\{q_i, a_i^{gt}\}$ can match certain rule out of all rules. We discard the unmatched examples while keeping the matched ones, which preserves 93.3% samples in train and validation set of VCR (223k out of 239k). Then we transform each $\{q_i, a_i^{gt}\}$ to a statement via the matched rule.

To design different rules, we start from question types and find there are mainly 7 question types in VCR, which begin with *what*, *whose*, *how*, *where*, *who* and *which*. We first define a basic rule for each question type. Then we employ NLP specialists to iteratively refine those basic rules to cover as many as possible various scenarios under each question type. To be more concrete, in every iteration, 20 question-answer pairs per question types are sampled, the NLP specialist have to examine whether current rules can perfectly transform them. If not, they can revise current rules or create new rules for unseen scenarios. Tens of iterations are conducted until current rules can correctly transform all samples in last 5 iterations. In Fig. 3, We show one example of applying a rule to a question-answer pair in our data. At the end, 15 rules are summarized to do the transformation. The accuracy of our rules are validated by further Amazon Turker Annotation, which will be introduced later.

Post-processing After obtaining the statements, several steps are applied in post-processing. (1) VCR contains all person bounding boxes and object bounding boxes in image, which are already verified by Amazon Turker, and annotated person/object-region co-reference links. Among verified bounding boxes, we only keep the person bounding boxes as candidate person boxes in our task. As for the co-reference links, we keep person-box links as ground-truth grounding labels, and replace the object-region links mentioned in statement with their object names. In that way, each token in statements is either a word in vocabulary or a person link. (2) We remove the samples that have no person links in statements or no person candidate box in images. (3) We remove samples that have only one candidate person box in image, in that the accuracy would be 100% for those sam-

Dataset	V. Src.	#Description	#Image	#Target	Avg. Word Len.	Human-centric	Knowledge Reqiured
RefCOCO	MSCOCO	142k	20k	142k	3.61	✗	Spatial/Appearance/Action
RefCOCO+	MSCOCO	141k	20k	141k	3.53	✗	Appearance
RefCOCOg	MSCOCO	104k	27k	104k	8.43	✗	Spatial/Appearance/Action
Flickr 30k entities	Flickr	159k	32k	276k	-	✗	Spatial/Appearance/Action
Who's Waldo	News	193k	193k	215k	-	✓	Event/Activity
HumanCog(ours)	Movies	130k	67k	230k	10.32	✓	Commonsense-Temporal/Causal/Mental

Table 1: Comparison with other grounding datasets. '-' denotes data not provided in their paper. Additionally, the average number of people in the image (sentence) in our dataset is 4.11 (1.85).

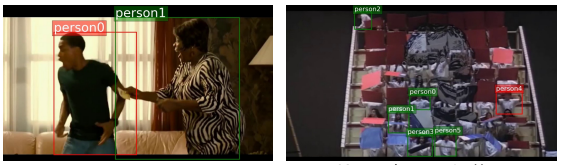

Figure 4: Ambiguous and unreadable examples.

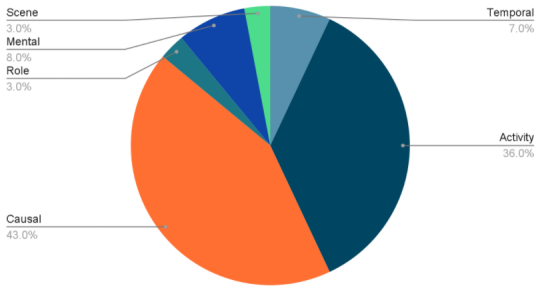

Figure 5: Commonsense Type Analysis

ples. (4) Some samples contain too many persons in image, where the persons tend to be blurry and incomplete. To reduce such noise, we remove samples that have more than 10 persons in image. (5) In some cases, there will be two or more person links tied together in description, *e.g.*, "PERSON1 and PERSON2". As a result, the person links can be exchanged, which causes the ambiguity. We simply remove those samples.

In summary, after above post-processing, we keep 134k samples (out of 223k), which are split into 120k training, 7k validation and 7k test set.

Validation with Amazon Turker Workers on Amazon Turker are employed to verify the validation and test set of our data. Now that the person-box grounding links and candidate person boxes have been checked by workers in VCR, we assume the correctness of them are guaranteed. The two focuses of our verification are ambiguity of grounding links and grammar mistakes or typos in transformed statement. The ambiguity means that in some samples the person mentioned in statement can refer to multiple person boxes in the image. For example, in Fig. 4 (a), the description is "PERSONX" is quite shocked". However, both "PERSON0" and "PERSON1" are quite shocked in the image. So the ground-truth "PERSON1" is not complete and this sample is ambiguous. Even though it's acceptable that the noise brought by incomplete links exists in the training, we hope to remove it from validation and test set to make sure a precise evaluation. By annotation of workers, only 29% data are ambiguous to certain degree. For grammar mistakes or typos, it's because questions and answers in VCR are in free-form, and finite rules may not cover all cases. As shown in Fig. 4 (b), "PERSONX are ' s arms raised" is wrong due to the lack of consideration of genitive cases in rule design. Through annotation, we find only 3% data having grammar mistakes or typos.

We remove the ambiguous and unreadable samples. Then, the validation/test set shrinks to 4.9k/4.9k samples, while the training set remains 120k. We pay the workers 0.05$ per sample to make their wage 12$ to 15$ per hour. In order to obtain high quality data, only workers that have finished more than 400 HITs with a decent approval rate of 96% are allowed for our annotation, which gives us around 90% agreement in identifying the most likely referred person.

4.2 Data Analysis

Commonsense Types Our dataset covers plentiful daily scenarios with an enormous diversity in commonsense types. We classify the commonsense types according to the templates. As shown in Fig. 5, 43% samples involve causal commonsense, 36% samples are related to highly semantic activity commonsense. For some categories, such as causal and temporal, we can further find sub-categories.

Causal commonsense can be either causal inference or causal explanation. Temporal commonsense includes before, after and conditional commonsense. Some examples are shown in Fig. 1 and more can be found in experiment section.

Comparison with Other Grounding Datasets Tab. 1 exhibits a comparison of our dataset to previous visual grounding datasets. *HumanCog* is the only one specializing on human-centric commonsense grounding. *Who's Waldo* (Cui et al., 2021) is most similar to us, in that both focus on grounding persons. Nevertheless, their samples are crawled from news, where the descriptions are mostly about low-level human actions that are visible straight from images, seldom requiring extra hop of inference and commonsense knowledge.

5 Method

In this section, we introduce a context-object-aware method as a strong baseline to solve the task. To jointly encode vision and text input, our architecture is built on a pre-trained vision-language Transformer, UNITER (Chen et al., 2019) for its generality, which will be covered in Sec. 5.1. In Sec. 5.2, we introduce the classification loss and the proposed context contrastive loss that can facilitate the visual commonsense learning between the human-object interaction. The diagram of method is shown in Fig. 6

5.1 Architecture

Visual Input Given the image I and candidate person boxes $\boldsymbol{r}$, we follow UNITER (Chen et al., 2019) to use the Faster R-CNN (Ren et al., 2015) to extract the pooled ROI features of each region r_i as visual features. Location features are encoded by a 7-dimensional vector, $[x_1, y_1, x_2, y_2, w, h, w * h]$[1]. Visual and location features are transformed into the same dimension through two FC layers, and are then summed up and normalized by a LN, as the input features of each region. Additionally, to complement person representation and enrich the visual scene understanding, we take extra detected object proposals[2] $\boldsymbol{r}'$ by Faster R-CNN and append their features together with candidate person boxes as input. It's validated in experiments that the additional proposals are essentially helpful to this task.

[1][normalized top/left/bottom/right coordinates, width, height, area.]

[2]Objectness Threshold is 0.2, Max. No. of object is 100.

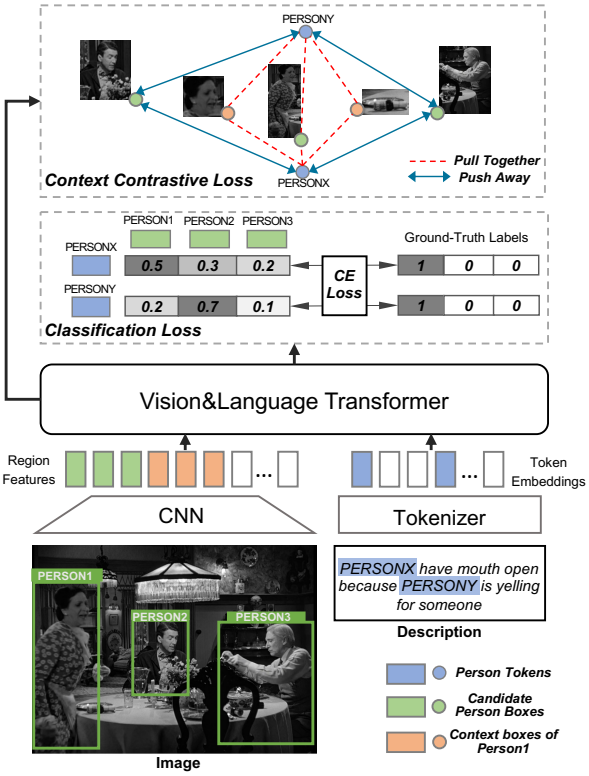

Figure 6: Diagram of our context-object-aware method

Textual Input We tokenize the input description into WordPieces (Wu et al., 2016). The word embeddings and position embeddings are summed up and normalized by a LN, as input text features. As for person links in description, following VL-BERT (Su et al., 2019), we replace them with random neutral names, *e.g.,* James or Mary. Compared with initializing new embeddings, it can better utilize the pre-training knowledge.

Transformer Encoder The visual and textual features are input into the Transformer (Chen et al., 2019), pre-trained with 9.5M image-caption pairs. The self-attention layers inside enable the contextualization of the two modalities. We take the hidden layers' features for loss calculation.

5.2 Loss Function

Classification Loss We treat the task as a classification problem, where each person link t_i in description $\boldsymbol{t}$ should be classified into the ground-truth person box r_j out of N candidate person boxes $\boldsymbol{r}$. In that way, we transform the features of candidate person boxes into classifier weights and apply

a cross-entropy loss $\mathcal{L}_{cls}$:

$$Q(i,j) = f(t_i)W_1 \times (f(r_j)W_2)^T$$
$$\mathcal{L}_{cls} = -\frac{1}{k}\sum_{i=1}^{k}\log(Softmax(Q(i,:))),$$

where $f(\cdot)$ denotes the final layer's feature output, W_1 and W_2 are linear weight matrices, and k is the number of person links in description.

Context Contrastive Loss Although the classification loss is straightforward to model human-human interaction, the relationship between detected object proposals and persons is not fully exploited, *i.e.*, the human-object interaction. The surrounding objects can provide plentiful and distinctive semantics to the persons, which is essential to diversify persons and identify ground-truth (GT) persons from other persons. In response to that, we propose a context contrastive loss, where context objects pertaining to GT persons are regarded as positive instances and their features are aligned with corresponding person embeddings in text. More specifically, we pull the person links in description closer to context objects pertaining to corresponding GT persons and push them away from other negative persons in feature space. At first, for person link t_i in text, whose GT person box is r_j in image, we define the pertaining context objects $C(i)$ as those detected boxes that have higher IoU scores with GT person and lower IoU scores with other persons:

$$C(i) = \{r_c^{'} | r_c^{'} \in \boldsymbol{r}^{'} \text{ and } IoU(r_c^{'}, r_j) > T_1$$
$$\text{and } max(IoU(r_c^{'}, \boldsymbol{r} \backslash r_j)) < T_2, \{t_i, r_j\} \in \mathcal{L}\},$$

where T_1 and T_2 are two thresholds as hyper-parameters. Then we further include GT person box r_j also in the positive instances, $P(i) = C(i) \vee r_j$. The negative instances are other person boxes $N(i) = \boldsymbol{r} \backslash r_j$. Contrastive loss has been widely studied in recent works, *e.g.*, InfoNCE (Oord et al., 2018) and its related extensions (He et al., 2020; Khosla et al., 2020; Radford et al., 2021; Jia et al., 2021). To realize contrastive learning in our scenario, considering the pertaining objects that have more overlap with GT person r_j tend to be semantically more aligned, we further utilize their IoU scores as the weights in proposed context contrastive loss:

$$\mathcal{L}_{con} = -\frac{1}{k}\sum_{i=1}^{k}\mathcal{L}_{con}^{i}$$

$$\mathcal{L}_{con}^{i} = \sum_{r_p \in P(i)} \frac{IoU(r_p, r_j)}{|P(i)|} \cdot \log \frac{\exp(g_l(t_i) \cdot g_l(r_p)/\tau)}{\sum_{r_n \in \mathcal{N}(i) \vee P(i)} \exp(g_l(t_i) \cdot g_l(r_n)/\tau)},$$

where τ is the temperature and $g_l(\cdot)$ denotes the encoded feature of l-th hidden layer from the last (l=3 in our case). Through context contrastive loss, we encourage the person link representations more similar to the correct persons and contextual neighbors of them, and more distinguished from other persons.

In training, two losses are ensembled together with a coefficient λ to adjust the significance of contrastive loss, which is shown as follows. In experiments, we find $\lambda = 1$ already gives satisfactory performance. In inference, we take the classification logits and select the box with the highest score as prediction, regardless of the contrastive part.

$$\mathcal{L}_{train} = \mathcal{L}_{cls} + \lambda\mathcal{L}_{con}.$$

6 Results

For evaluation, we comprehensively study the effect of various methods on this task. We first introduce previous state-of-the-art (SOTA) methods and compare them with the proposed context-object-aware model (Sec. 6.2). And we conduct ablation experiments to quantify the importance of different components in this task (Sec. 6.3). Further, we provide qualitative results for analysis (Sec. 6.4).

6.1 Experimental Setup

We compute accuracy of predicted person boxes for all mentioned persons in descriptions as the evaluation metric. For the visual features, if not specified, we use an off-the-shelf pre-trained Faster-RCNN (Anderson et al., 2018) to extract the region features of both person candidate boxes and detected boxes from image. As for the training of proposed context-object-aware model, pre-trained weight of UNITER is loaded as the initialization. AdamW (Loshchilov and Hutter, 2017) optimizer is used with a learning rate of 6e-5. Following UNITER, we use dynamic sequence length to batch the samples by their number of input tokens, so that padding is reduced and training is speeded up. We set the batch size to 4000 and train 4k steps. Our model is implemented in Pytorch and trained with 4 TITAN RTX GPUs. (Paszke et al., 2019)

Model	Model-size	# Pre-training Pairs	Acc.
Random	-	-	30.9
B→ S	-	-	39.2
L→ R	-	-	31.0
L→ R (Biggest)	-	-	39.6
BAN(LSTM)	-	-	56.2 ±0.36
BAN(BERT)	-	-	62.8 ±0.32
VL-BERT	base	3.3M	67.4 ±0.27
	large	3.3M	68.2 ±0.22
VILLA	base	9.5M	68.1 ±0.56
	large	9.5M	68.5 ±0.52
UNITER	base	9.5M	67.9 ±0.29
	large	9.5M	68.9 ±0.31
Ours	large	9.5M	**69.8** ±0.23
Human	-	-	92.3

Table 2: Comparison against previous methods

Model	Acc.(%)
Ours	69.8
– Context Contrastive Loss	68.9
– Detected Context Objects	66.9
– Pre-training on Image-Text Pairs	64.5

Table 3: Ablation of different components in our method

6.2 Compared with Previous Methods

Several previous visual grounding methods are implemented to be compared with our context-object-aware model, including heuristic methods, models w/o image-text pre-training and models w/ image-text pre-training.

Heuristic Methods Similar to (Cui et al., 2021), we probe the biases in dataset by several hand-crafted heuristics. To be more specific, we assign the persons mentioned in description from left to right to person candidate boxes in image that are sorted based on following heuristics: (1) big to small with decreasing areas; (2) left to right according to the upper-left coordinates (3) left to right with only top-k biggest boxes, where k is the number of mentioned persons in description.

Human Evaluation We go through the test set and obtained a 92.3% accuracy with human evaluation. It can be treated as an reasonable upperbound of current machine models.

Methods w/o Image-Text Pre-training In previous works of non-pretrained vision&Language models, we choose BAN (Kim et al., 2018) for its superior performance on visual grounding and other downstream tasks. BAN extracts visual features and text features, and then fuse two modalities by a bilinear attention network. A classification loss as in Sec. 5.2 is applied afterward to do the classification. In our implementation, text feature can be extracted either by a LSTM (Hochreiter and Schmidhuber, 1997) module, as in the original paper, or a pre-trained BERT (Devlin et al., 2018) module. We name those two BAN(LSTM) and BAN(BERT).

Methods w/ Image-Text Pre-training Recently, there have been a lot research interests in vision&language pre-training models for their effectiveness and generalizability. We implement widely-used VL-BERT (Su et al., 2019), UNITER (Chen et al., 2019) and VILLA (Gan et al., 2020), whose numbers of pre-training image-caption pairs range from 3.3M to 9.5M. Similarly, we also apply the classification loss for training.

The full experimental result, including random guessing, is shown in Tab. 2. By comparing the heuristic methods with random guessing, it's found that the strongest heuristic can only improve 9%, which indicates the spatial bias and area bias is not severe in our dataset. From the comparison of models w/ and w/o image-text pre-training, we find that, in general, larger pre-training data can bring higher performance. Last but not least, proposed context-object-aware method can further outperform UNITER by 0.9%, reaching a final performance of 69.8% and establishing a strong baseline.

6.3 Ablation Study

Components We further validate the effectiveness of different components in our method. In Tab. 3, we present the results. If the multi-modal Transformer weight is initialized from a BERT model without image-caption pre-training, the performance drops by 5.3%, which is in line with the finding in previous section that image-text pre-training is beneficial. Removing the context contrastive loss and the detected context objects also bring a performance degrade of 0.9% and 2.9% respectively. It highlights the importance of incorporating contextual objects and enhancing human-object interactions. In summary, both strong image-text fusion and effective human-object visual commonsense modeling are crucial for this task, which suggests the avenues to future works.

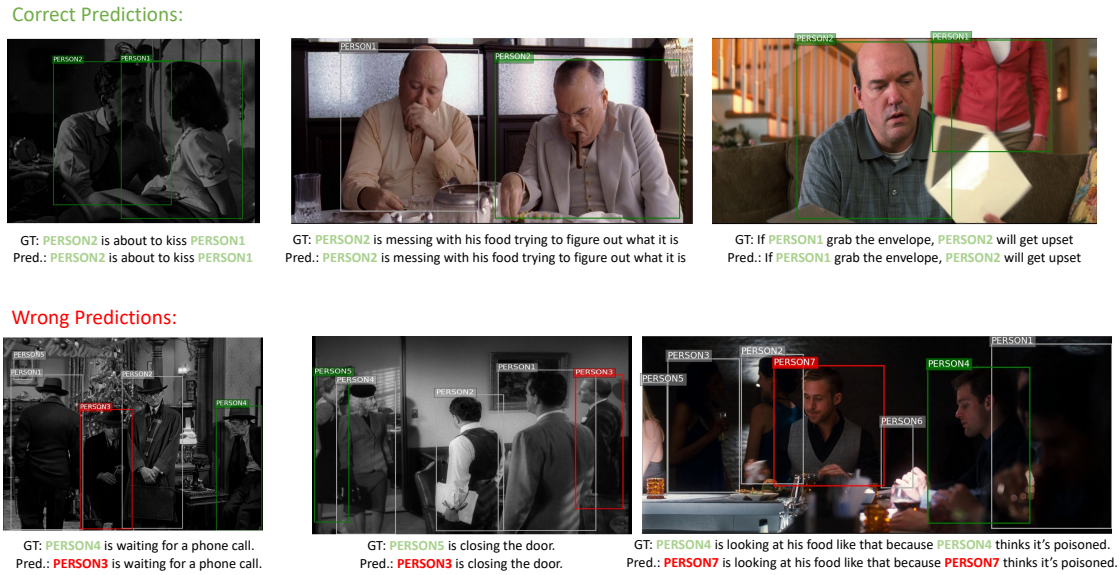

Figure 7: Qualitative Results. GT denotes ground-truth links in description. Pred. denotes predicted links.

Hyperparameters We ablated where to insert the context contrastive loss. We found taking the output of 3rd Transformer layer from the last to compute context contrastive loss performs the best. Others bring at most 0.6% performance drop. We infer that contrastive and classification loss are not fully complementary when added to the same layer. Putting the auxiliary loss earlier helps the last several layers specialize in the classification target.

6.4 Qualitative Results

In Fig.7, we present qualitative results of our method. Ours can correctly ground to the GT person boxes in many cases. In the second correct samples, our model can predict it's PERSON2 instead of PERSON1 probably by the facial expression and the food in front of him. Our model can also work well under the complex causal scenarios, such as the third correct sample. Sometimes, we fail on those samples where the visual information might be incomplete, blurry or misleading. For example, in the first wrong sample, the object PERSON3 is holding a cell phone from appearance, which might mislead the model. And in the 2nd wrong sample, the door is unseen from the image. Some too complex descriptions that require detailed and subtle observation may also cause failure. In the 3rd wrong sample, the PERSON7 is actually hiding something while the PERSON4 is stirring his food, making models hard to tell.

7 Conclusion

In this work, we present a new task: *human-centric commonsense grounding*, where machines are required to ground the persons mentioned in a commonsensical description. Correspondingly, we collect a new dataset for training and evaluation. Also, we proposed a context-object-aware method that exploits background context objects via context contrastive loss for a strong vision-language understanding. Through detailed analysis, we find there is still an ample room for improvement and we point out the potential directions for further works.

Acknowledgement

This work is supported by DARPA MCS program under Cooperative Agreement N66001-19-2-4032.

Limitations

One limitation of our work is that the training data might be noisy compared with validation and test data. But according to the statistics summarized from validation and test data, we only have around 3% unreadable samples and 29% ambiguous samples in the training set, which is acceptable. It's also a design choice to include ambiguous samples as we would like to test if a model can generalize well by learning from noisy data. The noisy setting simulates the process that humans learn from ambiguous examples rather than heavily curated data. And our experimental results prove that trained with noisy data, machines can still greatly outperform random guess.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

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

Yuqing Cui, Apoorv Khandelwal, Yoav Artzi, Noah Snavely, and Hadar Averbuch-Elor. 2021. Who's waldo? linking people across text and images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1374–1384.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1769–1779.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. *Advances in Neural Information Processing Systems*, 33:6616–6628.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

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

Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1115–1124.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. 2021. Mdetr-modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referit game: Referring to objects in photographs of natural scenes. In *EMNLP*.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear attention networks. *Advances in Neural Information Processing Systems*, 31.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. What is more likely to happen next? video-and-language future event prediction. *arXiv preprint arXiv:2010.07999*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. 2020. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan Yuille. 2019. Clevr-ref+: Diagnosing visual reasoning with referring expressions. *arXiv preprint arXiv:1901.00850*.

Yongfei Liu, Bo Wan, Xiaodan Zhu, and Xuming He. 2020. Learning cross-modal context graph for visual grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11645–11652.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Ruotian Luo and Gregory Shakhnarovich. 2017. Comprehension-guided referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7102–7111.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *In Proceedings of the European Conference on Computer Vision (ECCV)*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2020. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

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

Peng Wang, Dongyang Liu, Hui Li, and Qi Wu. 2020. Give me something to eat: referring expression comprehension with commonsense knowledge. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 28–36.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4145–4154.

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. 2018. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315.

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

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.