

Premise-based Multimodal Reasoning: Conditional Inference on Joint Textual and Visual Clues

Qingxiu Dong^{1*}, Ziwei Qin^{1*}, Heming Xia¹, Tian Feng¹,
Shoujie Tong¹, Haoran Meng¹, Lin Xu¹, Zhongyu Wei²,

Weidong Zhan¹, Baobao Chang¹, Sujian Li¹, Tianyu Liu³, Zuifang Sui¹

¹ Key Laboratory of Computational Linguistics, Peking University, MOE, China

² School of Data Science, Fudan University ³ Tencent Cloud Xiaowei

{dqx, qinziwei}@stu.pku.edu.cn

Abstract

It is a common practice for recent works in vision language cross-modal reasoning to adopt a binary or multi-choice classification formulation taking as input a set of source image(s) and textual query. In this work, we take a sober look at such an “unconditional” formulation in the sense that no prior knowledge is specified with respect to the source image(s). Inspired by the designs of both visual commonsense reasoning and natural language inference tasks, we propose a new task termed “**P**remise-based **M**ulti-modal **R**easoning” (PMR) where a textual premise is the background presumption on each source image. The PMR dataset contains 15,360 manually annotated samples which are created by a multi-phase crowd-sourcing process. With selected high-quality movie screenshots and human-curated premise templates from 6 pre-defined categories, we ask crowd-source workers to write one true hypothesis and three distractors (4 choices) given the premise and image through a cross-check procedure. Besides, we generate adversarial samples to alleviate the annotation artifacts and double the size of PMR. We benchmark various state-of-the-art (pretrained) multi-modal inference models on PMR and conduct comprehensive experimental analyses to showcase the utility of our dataset.

1 Introduction

Cross-modal reasoning between image and text has been recognized as a fundamental and long-standing task in both academia and industry, which has recently attracted intensive attention from both natural language processing and computer vision communities (Su et al., 2019; Yu et al., 2020; Chen et al., 2020). Researchers try to teach machines to perceive, understand and reason with both visual and textual clues, which mimics human cognitive process (Lake et al., 2016).

The canonical form for the cross-modal reasoning tasks usually take the source image(s) as the input and request the inference model to perform a multi-choice classification according to the specified textual query, including visual question answering (Antol et al., 2015; Goyal et al., 2017), visual commonsense reasoning (Zellers et al., 2018), visual entailment (Xie et al., 2019) and image-text grounding (Suhr et al., 2017, 2019), etc. We argue the default cross-modal setting is “static” or “unconditional” in the sense that no prior presumption or belief is attached to the input images (Ren et al., 2021). The lack of prior presumption would be insignificant while answering factoid questions according to the images, e.g. “how many dogs are lying on the grass?” (Suhr et al., 2019) or “is the bowl to the right of the green apple?” (Hudson and Manning, 2019). However it may not be the case for more sophisticated cross-modal reasoning that involves human-like cognition and commonsense, e.g. “what may happen to [personA]?”, “what will [personA] say to [personB]” or “what will [personA] do to [personB]”. The answers to these questions are ambiguous by merely looking at the source images, instead we offer an extra textual premise, which serves as the prior belief to the source images, such as “to [personA], nothing could be more somber”, “[personA] is the son of [personB]”. We name the proposed task as “**P**remise-based **M**ulti-modal **R**easoning” (PMR)¹. In PMR, the inference model should be able to reason with both textual (from the premise) and visual (from images) clues.

The motivations for PMR are two-folds. 1) From the social psychology perspective, the design of PMR is inspired by the ABC Theory (Ellis, 1995), which represents a widely accepted framework for how one’s feelings and behavioral patterns are connected. As illustrated in Figure 1, the theory

*Equal contribution.

¹The dataset and baseline models can be found in <https://2030nlp.github.io/PMR/>.



Figure 1: Connections between the ABC theory (Ellis, 1995) and the proposed PMR.

claims that human emotions and resulting behavior do not come directly from the events, but from the interpretations (textual premise) we make of those events (depicted in the source images). 2) From the methodology perspective, PMR gains insights from both visual commonsense reasoning (Zellers et al., 2018; Lei et al., 2020; Park et al., 2020) and natural language inference (Dagan et al., 2006; Bowman and Zhu, 2019; Xie et al., 2019) tasks. From the view of commonsense reasoning, the textual premise serves as the prior belief to the image as mentioned before. Meanwhile in the world of natural language inference, the input image can be viewed as the supplementary evidence that supports the textual premise-hypothesis classification. The proposed PMR could be readily seen as a meaningful extension on the joint cross-modal entailment and commonsense reasoning with both visual and textual clues.

Given the input premise-image pair, the practitioners are requested to choose the only true hypothetical action from four candidates in PMR. All four hypothetical actions are written by crowd-sourced workers in a multi-phase cross-check annotation pipeline. We encourage the annotators to write challenging distracting actions that are logically correct with respect to either the premise or the image, but are contradicted with the joint premise-image pair while combining the visual and textual clues. In order to succeed on PMR, the reasoning model should excel at both language grounding among premise, image and candidate actions, and cross-modal understanding on commonsense and logical inference. We establish multiple competitive cross-modal pretrained models as baselines. We hope the proposed PMR could pave the way for the “conditional” cross-modal commonsense and logical reasoning that requires a human-like cognitive process.

2 Dataset and Task Overview

We describe the proposed PMR task with an example in Figure 2 I&II. Given a source image and a textual premise, the inference model should perceive and understand the image in combination

with the premise so as to choose the exclusive correct action among the four hypothetical candidates. The premise would serve as the background knowledge or domain-specific commonsense for the given image. In the running example, the model should be able to recognize what [person2] wears from the image and infer whether [person4] would give his seat to [person2] under the premise “[person4] is very friendly”. The corrected answer is ‘C’ according to the visual and textual clues. In total, we collect about 15k instances for PMR. We list the statistics for PMR in Table 2.

3 Data Collection

We collect PMR dataset in a multi-step crowd-sourcing pipeline, including 1) image and premise creation, 2) annotator recruitment and instruction, 3) cross-checking annotation.

3.1 Image and Premise Creation

The source images of PMR are selected from the image pool² in the VCR (Zellers et al., 2018) dataset. The entire image pool contains 110k high-quality movie screenshots. To make the images fit better with PMR, we screened out those which have low brightness, more than five people, or more than 15 tags, and finally got 29,987 images. We also kept the object anchor information in VCR along with the source images, which was identified using Mask-RCNN (He et al., 2017).

As there are so many possible choices while writing premise for an image, to facilitate analysis of model performance by premise type, we constrained the premise to six categories and manually wrote templates for each category (see Appendix D). As Figure 4 shows, the six categories are personality, identity, emotion, relationship, environment and antecedent. To complete templates, we create multiple word sets and fill corresponding words into the slots randomly. And finally, we obtained 30,759 predefined premises, which would be presented to annotators for selection.

²The images originate from Fandango MovieClips (<https://youtube.com/user/movieclips>) and Large Scale Movie Description Challenge (Rohrbach et al., 2017).



Premise: [person4] is very friendly.

Actions:

- A. [person2] with a cap wants to sit by the window but [person4] refuses him without any hesitance.
- B. [person2] in a white coat wants to sit by the window but [person4] refuses him without any hesitance.
- C. [person2] with a cap wants to sit by the window and [person4] exchanges his seat with him generously.
- D. [person2] in a white coat wants to sit by the window and [person4] exchanges his seat with him generously.

Figure 2: We demonstrate a source image with object anchors (e.g. ‘[person1]’, ‘[person2]’) in I, and the corresponding premise and hypothetical actions in II.

	Ori.			Adv.			Total
	Train	Val	Test	Train	Val	Test	
#samples	12,080	1,538	1,742	12,080	1,538	1,742	30,720
#unique 1-gram	9,882	3,819	4,101	8,046	3,071	3,359	11,041
#unique 2-gram	72,048	17,678	19,292	50,526	12,236	13,453	84,365
Avg premise length	9.48	9.47	9.54	9.48	9.47	9.54	9.49
Avg action text length	14.38	14.41	14.45	14.20	14.42	14.31	14.31
Avg #objects mentioned	1.92	1.91	1.94	2.42	2.43	2.38	2.17
#images	9,536	1,213	1,370	9,536	1,213	1,370	12,119
#movies covered	1,353	209	170	1,353	209	170	1,732

Table 1: The statistics of PMR dataset. Ori. stands for the manually annotated part of PMR, and Adv. represents the adversarial samples generated automatically.

3.2 Annotator Recruitment and Instruction

Through job descriptions posted on the online forums, we invite more than a hundred applicants who hold a bachelor of arts or higher degree to attend the online pre-annotating instruction and qualification test. In the pre-annotating instruction, we organized a two-hour training session, which covers the basics and goals of PMR, to better instruct the annotators, we extracted the text from VCR as the reference actions. Specifically, we filtered out questions about the following behavior of persons and retrieved the correct answers as the reference actions.

After the pre-annotating instruction, we conducted a qualification test by asking each participant to write the true and false hypothetical actions for 10 sampled image-premise pairs, the authors manually assess the annotation quality and eventually select 61 annotators whose educational background covers literature, linguistics, sociology, etc. Among them, 19 annotators who got higher scores in the qualification test were recognized as advanced workers and assigned to the distractor collection phase (phase 2 in Figure 3), while the

other annotators were assigned to the hypothetical action creation phase (phase 1 in Figure 3).

3.3 Cross-check Annotation

We divide the annotation process into 3 phases. In phase 1, workers are presented with an image with bounding boxes on it, six predefined premises of different categories and six reference actions. Firstly, they are supposed to choose an appropriate premise. Workers can adjust the person tags to meet with the given image, but they have to assure that the modified premise still belongs to one of the six categories. Secondly, they ought to write two hypothetical actions, which describe what will happen next. Among them, **Action-True** contains image information and meets the chosen premise. In contrast, **Action-False** contains image information but does not meet the chosen premise. Thirdly, to make it easier for replacing certain words in phase 2, they need to enclose the words with curly brackets that mention the information of the image in both written actions.

In phase 2, advanced workers, serving as examiners, are responsible for checking whether the

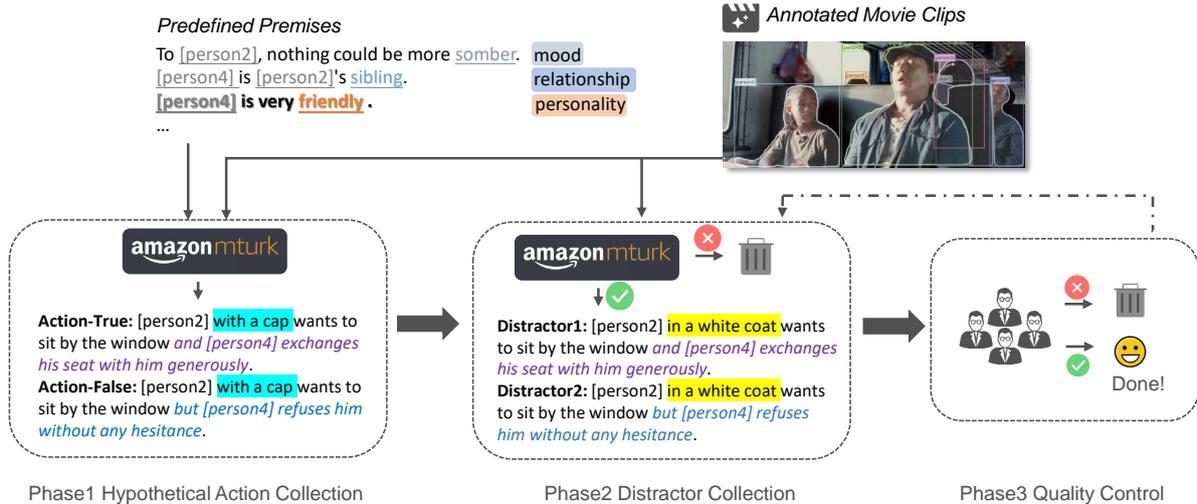


Figure 3: The overview for the cross-check annotation for PMR. In the textual premises, the underlined tokens denote the slots in the pre-defined templates. In the hypothetical action collection, the words related with visual information are **highlighted in blue**, while the correspondent substitution in the distractors is **highlighted in yellow**.

annotation in phase 1 conforms to labeling instruction. They can drop it while meeting poor-quality one, and once they accept it, they need to write another two distractors corresponding to Action-True and Action-False respectively. We instruct the annotators to write challenging distractors that are logically correct with respect to either the premise or the image, but contradict with the joint premise-image pair while combining the visual and textual clues.

In phase 3, to ensure the quality of the examiners' work, we sample 10% HITs in phase 2 to check if the annotation meets our standard. This work is performed by the authors of this paper, and we feedback to examiners timely.

Post-Processing In order to ground objects from images with entities in text, we follow the VCR which substitutes all the tags both in premise and answers with the index in objects list. Besides, since crowd workers can conditionally modify the given premises, which results in the missing of labels of category, we calculate the BLEU (Papineni et al., 2002) score between premise and each template from six categories and labeled it with the most likely category with the highest BLEU score.

Annotation Cost We drop all the annotations that are rejected in phase3 and obtained 15,360 items after 60-day work. (See Appendix E for detailed pricing strategy) We split the total dataset into train, valid and test set in 8:1:1.

Adversarial Samples Generation To alleviate the possible bias introduced by annotation artifacts, inspired by Zellers et al., for each sample in the crowd-sourcing set of PMR, we pick three actions from other samples but similar to the premise as negatives and generate adversarial samples with the same scale. See Appendix C for detailed methods.

4 Experiments

To set up the benchmark, we introduce multiple text-only and pretrained multi-modal baseline models as well as human performance on PMR.

4.1 Baselines

Text-only Baselines We begin with a “blind” setting, where only the text is given without access to the image or bounding boxes annotations. In this setting, the models have to choose the correct answers with only textual clues. Specifically we finetune the BERT (Devlin et al., 2018) model in the action-only and premise+action scenarios. (See Appendix 6 for training details)

Multi-modal Baselines Throughout the visual reasoning tasks, cross-modal pretrained models have achieved state-of-the-art performance (Zellers et al., 2018; Goyal et al., 2017; Hudson and Manning, 2019). We introduce three powerful pretrained models and test their performance on PMR:

- **VL-BERT**(Su et al., 2019) A dual-stream pretrained model, is extended from BERT by ap-

pending visual feature embedding along with the subsequent sentences.

- **ERNIE-VL**(Yu et al., 2020) is a knowledge-enhanced approach to learning joint representations of vision and language, which introduced structure knowledge with scene graph prediction tasks while pretraining.
- **UNITER**(Chen et al., 2020) is also an extension of BERT to the visual domain. It is trained with a conditional masking strategy that allows the model to learn an informative representation of one modality conditioned on the other.

4.2 Detailed Experiment Settings

Before feeding into BERT, we concatenate the question and each answer as a sequence and replace detection tags with object names in it. As for person detection tags, we substitute them with gender-neutral names to bridge the gap between PMR and pretraining corpus (Zellers et al., 2018).

In terms of multimodal baselines, while an annotation is processed as above, we regard region features as visual tokens to be concatenated along with text sequence. Region features are extracted with Faster RCNN. Different from ERNIE-VL and UNITER which freeze Faster RCNN all the time, VL-BERT updates it while pretraining and finetuning.

4.3 Results

Are Premises Critical for Models to Predict?

As is shown in Table 2, E-L achieves the highest accuracy while training and testing on the original split, but if trained with only actions as text inputs, the accuracy of it decreases sharply by 23.3%. More details can be found in Table 3. Premises help E-L correct the prediction of Action-False with the ratio from 30.5% to 8.2%.

Is Vision Modality Useful? As for the performance of text-only models and multi-modal models in Table 2, considering the information from images improves the accuracy by 10%-14.7%. Besides, Distractor1, as generated by substituting objects in Action-True, contradicts with corresponding images, which can be hard negatives while the visual signal is imperceptible. In Table 3, trained on the original split, E-L predicts Distractor1 with a ratio of 10.7%, nearly half of the prediction of

	Ori. Train		Adv. Train		Mix Train		
	OT	AT	OT	AT	OT	AT	MT
Random	25.0	25.0	25.0	25.0	25.0	25.0	25.0
B-B [†]	39.3	23.1	22.9	41.9	25.6	25.0	25.3
B-B	65.2	21.8	27.2	23.2	25.8	26.9	26.4
V-B	75.4	37.4	22.8	80.2	70.7	66.4	68.6
E-B	79.0	46.2	33.7	82.9	72.7	76.0	74.4
U-B	77.4	50.7	35.8	80.3	72.7	70.1	73.2
V-L	79.3	47.0	25.3	82.5	77.3	75.4	76.4
E-L [†]	56.6	51.1	40.7	75.5	50.2	70.9	60.6
E-L	79.9	52.1	33.4	83.6	77.1	78.0	77.6
U-L	77.0	57.9	35.7	81.9	74.6	72.0	73.2

Table 2: The performance of baselines trained and tested on different dataset split. Ori. Train, Adv. Train and Mix Train respectively stand for the models being trained with the manually annotated part of PMR, adversarial samples and the mixture of the above two. Accordingly, OT, AT and MT represent testing on original, adversarial and mixed test set. For baselines, B, V, E and U are short for BERT, VL-BERT, ERNIE-VIL and UNITER. “-B” and “-L” stand for the two different sizes of models, “Base” and “Large”. Besides, we trained and tested models without premises as input, which indicated with sign [†].

	Ori. Train				Mix Train			
	AT	D1	AF	D2	AT	D1	AF	D2
BERT [†]	39.3	15.2	31.2	14.4	25.6	23.5	26.1	24.8
BERT	65.2	19.8	10.6	4.5	25.8	25.8	24.3	24.1
E-L [†]	56.6	8.0	30.5	4.9	50.2	6.5	37.4	6.0
E-L	79.9	10.7	8.2	1.2	77.1	9.6	11.1	2.1

Table 3: Detailed performance on Ori. Test. AT, D1, AF and D2 represent the ratio of the four-type prediction of models, Action-True, Distractor1, Action-False and Distractor2 respectively.

BERT, suggesting that vision modality is beneficial to distinguish distractors.

Does Crowd-sourcing Introduce Annotation Artifacts to Dataset?

Table 2 shows that BERT with only actions outperforms random baseline by 14.3 on accuracy, indicating that textual bias is introduced by manual annotations, which is inevitable due to fixed patterns of thinking during labeling but is tolerable. Adding premises as background presumption, BERT can discover the relation between premises and actions with the performance improved a lot.

We also conduct a quantitative analysis of bias in data. To check whether labels are significantly associated with certain words, we compute the point-

Word	green	red	n't	fight	throw	hit
No. (Ori.)	1	2	7	11	18	23
PMI (Ori.)	1.620	1.562	1.322	1.318	1.305	1.301
No. (Mix)	1	2	26	16	52	155
PMI (Mix)	1.488	1.408	1.231	1.242	1.212	1.172

Table 4: PMI calculated for the words in Ori. and Mix datasets. No. is the reverse order in vocabulary according to PMI.

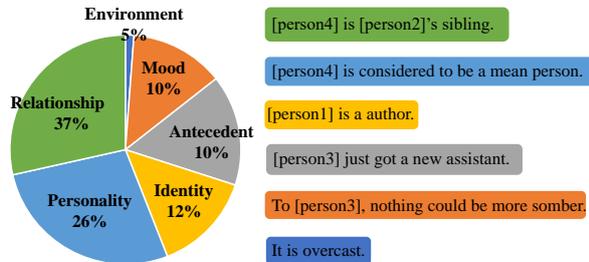


Figure 4: An overview of the premises from the samples in validation set.

wise mutual information (PMI) (Gururangan et al., 2018). As each sample in PMR has four choices, we consider the premise concatenated respectively with four answers as four binary-class items with label *False* if the answer is incorrect, and compute PMI between words and *False* label.

As Table 4 shows, among the top 30 words, green and red are the top two with the highest PMI, they are commonly used by crowd workers to modify the objects and generate distractors mismatched with images. Other words such as leave, fight and hit, crowd workers tend to write universal negative containing these words so as to pass the qualification check, which also results in high PMI of these words.

Does Training with Adversarial Samples Help to Relieve the Above Issue? Yes, firstly in Table 4, we can find that the PMIs of words decrease, indicating that adversarial samples help to balance the correlation of words and labels.

Secondly, Table 2 shows that there is a great gap in the performance of multi-modal models between testing on OT and AT while training on the original set, which also testifies the existence of bias in the original set. Nevertheless, after training on mixed data, multi-modal models are able to achieve high accuracy both on OT and AT, suggesting that adversarial samples are helpful to improve models' robustness. Besides, We can notice that E-L may be misled by the fewer incorrect samples in the adversarial split, and it results in a decline of 2.8%

on OT, which can be seen as a compromise between generalization and performance.

Thirdly, we find that another huge gap of 39.4% on OT between BERT trained on the original set and mixed set, and it indicates that the adversarial samples help to change the data distribution and alleviate the bias from the text.

5 Analysis

5.1 Premise Distribution

From Figure 4, we can find that Relationship and Personality account for nearly two thirds of the samples, while the other four categories hold the rest one third, which indicates that the understanding of interpersonal relationships and the correlation between character's personalities and behaviors are test points of PMR. In terms of the reason why there is such a distribution, we maintain that this is because Relationship and Personality are more likely to conform with the instruction of annotation. As presented in Section 2, we expect premises as a supplement of images, so the information implied in premises should not be repeated or contradictory with the content of the image. Choosing the other four premises is more likely to violate the above requirements. For instance, people's mood tends to explicitly show through expression; the dressing of the characters can reflect their identity and occupation.

5.2 Substitution Strategy

We manually classified 100 samples from the validation set into nine types, which is presented in Table 5. As we can see, the types of substitution are varied, and most of them are related to PERSON supertype, which reflects the emphasis on the characters' behavior.

In terms of the understanding of images, the levels of difficulty vary among different types. For instance, in APPEARANCE, CLOTHING, OBJECTS and COLOR, the words involved in the replacement process are relatively simple, such as long->short, white->red, knives->guns, which are likely to appear as labels in the datasets of object detection. Thus, systems with pre-trained objects detection models are expected to have an excellent performance on the four types. On the contrary, as for the types like EMOTION, BODY STATE and LOCATION, the texts substituted tend to be abstract and mostly are phrases consisting of more than one word. Consequently, these types are more

Supertype	Type	Action-True	Distractor1
PERSON	Appearance (8.0%)	[person2] with long hair will fall on the [couch1] and have a sleep immediately .	[person2] with short hair will fall on the [couch1] and have a sleep immediately .
	Emotion (3.0%)	[person1] weeps and tells [person2] his sad story , and he listens attentively .	[person1] smiles and tells [person2] his sad story , and he listens attentively .
	Clothing (22.0%)	[person2] will button up his shirt , because now it does n’t look neat .	[person2] will button up his coat , because now it does n’t look neat .
	Body State (10.0%)	As [person2] is ill in bed , [person1] will take care of [person2] .	As [person2] is ill sitting in the chair , [person1] will take care of [person2] .
	Location (18.0%)	[person4] who is in a car will ask [person2] to live with him because [person2] ’s roommate has just died of an accident .	[person4] hiding behind the column will ask [person2] to live with him because [person2] ’s roommate has just died of an accident .
GENERAL	Objects (23.0%)	[person1] and [person2] are trying to kill each other with knives .	[person1] and [person2] are trying to kill each other with guns .
	Color (1.0%)	[person3] with a red tie will play a beautiful tune with his guitar .	[person3] with a white tie will play a beautiful tune with his guitar .
	Texture (8.0%)	[person1] will have a rest on the chair .	[person1] will have a rest on the stone chair .
ENVIRONMENT	Scenery (7.0%)	[person1] and [person2] are walking down the street together .	[person1] and [person2] are walking down the park together .

Table 5: Nine types of substitution in Phase 2, the percentage of which in the original validation subset and the examples. For each one, the words before and after the replacement are highlighted in red and blue respectively.

challenging, and in order to succeed on them, the model is supposed to excel at in-depth image understanding and fine-grained grounding.

Comparing the difference before and after the substitution, we may notice that the substitution, within the scope of several consecutive words, can be regarded as a method of disturbance, and correspondingly, the distractors served as adversarial choices are beneficial to counter the possibility of models exploiting unimodal priors.

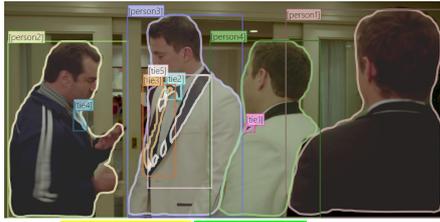
5.3 Case Study

We conduct a qualitative analysis shown in Figure 5, where the first two samples are correctly predicted by ERNIE-VIL. For each of them, premise 1 is the original. To explore the influence of different premises on the model predictions with the same image as background, we write premise 2 for the test, with which the model is expected to choose Action-False as the correct answer. The result shows that ERNIE-VIL predicted correctly both with premise 1 and 2, which indicates the model can distinguish the possible person’s behaviors resulting from the two opposite premises. But for case 3, the model makes an incorrect prediction with Action-False chosen, suggesting that the

performance is still unsatisfactory due to limited commonsense knowledge. Besides, even if the simple objects substitution can be easily detected by comparing the text with the images directly, case 4 shows that the model is not able to determine the relative position of people, which indicates that a powerful understanding capacity of images is needed to be added.

6 Related Work

Multimodal Commonsense Reasoning A series of tasks and datasets have been proposed for cross-modal commonsense reasoning. VCR (Zellers et al., 2018) is the most related work to the proposed PMR. It requires machines to understand a image and answer a multi-choice question. Specifically, questions are like ‘what is going to happen next’, ‘infer the relationship between [personA] and [personB]’ and ‘why is [personA] smiling’, as well as the rationale why the answer is true. In the video understanding regime, VLEP (Lei et al., 2020) is a dataset for future event prediction from videos. Given a video with aligned dialogue, and two possible future events, the AI system is required to choose the more likely event from two provided options. In the world of cross-modal com-



Premises: 1. [person3] is indifferent. 2. [person3] is helpful.

Actions:

- ⊙ AT: [person2] raises his hand , which is hurt badly , but [person3] just looks at it coldly and says nothing.
- ⊙ D1: [person2] raises his left hand , which is hurt badly , but [person3] just looks at it coldly and says nothing.
- ⊙ AF: When [person2] raises his hand , [person3] sees that it is hurt and offers to bandage it.
- D2: When [person2] raises his right hand , [person3] sees that it is hurt and offers to bandage it.



Premises: 1. The situation is critical, but people are lucky. 2. this is telling a nightmarish story.

Actions:

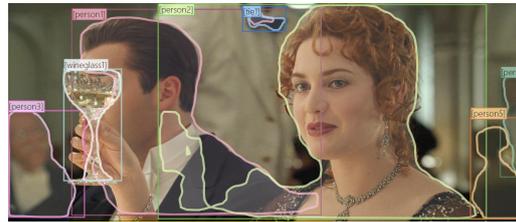
- ⊙ AT: the water fills the room , but [person2] who has black hair will be saved by a passing boat.
- ⊙ D1: the water fills the room , but [person2] who has golden hair will be saved by a passing boat.
- ⊙ AF: the water fills the room and [person2] who has black hair will drown.
- D2: the water fills the room and [person2] who has golden hair will drown.



Premises: [person1] just got a new spouse.

Actions:

- AT: [person1] who is laughing happily will announce his engagement.
- D1: [person1] who is sobbing bitterly will announce his engagement.
- ⊙ AF: [person1] in a suit is sharing his happiness of being single with others.
- D2: [person1] in a t-shirt is sharing his happiness of being single with others.



Premise: [person1] is thought to be a/an outgoing man.

Actions:

- AT: [person1] will chat with the strange girl on his left and ask her to have a drink.
- ⊙ D1: [person1] will chat with the strange girl in front of him and ask her to have a drink.
- AF: [person1] will keep silent at the wine party.
- D2: [person1] will keep silent at the dock.

Figure 5: There are four prediction samples from ERNIE-VIL. For each sample, We list actions in the order of AT, D1, AF and D2, which represent Action-True, Distractor1, Action-False, Distractor2 respectively. Model predictions are highlighted in different colors, which correspond to different premises, and the predictions are tagged with ⊙ if correct, and ⊗ otherwise.

monsense graph, VisualCOMET (Park et al., 2020) is a repository of Visual Commonsense Graphs that consists of 1.4 million textual descriptions of visual commonsense inferences. The proposed PMR is different from the above-mentioned dataset as we request the models to perceive and understand the source image with the supervision from a specified textual premise, and reason conditionally with joint visual and textual clues.

Natural Language Inference Early methods for textual NLI mainly relied on conventional, feature-based methods trained from small-scale datasets (Dagan et al., 2006). The release of larger datasets, such as SNLI (Bowman et al., 2015), MultiNLI (Williams et al., 2018), made neural network methods feasible. In the field of cross-modal entailment, the SNLI-VE (Xie et al., 2019) dataset casts the source image as the visual premise, and asks the inference model judge whether the specified textual hypothesis entails or contradicts with the visual premise. The proposed PMR can also be viewed as

a cross-modal entailment task, however difference from SNLI-VE, both premise and hypothetical actions are textual, and the corresponding source images can be viewed as the supplementary evidence while judging the entailment labels. Besides, the existing textual and visual-related NLI tasks are formulated in a three-way classification, while the proposed PMR is model as a 4-choice classification task.

7 Conclusion

We propose a premise-based cross-modal reasoning (PMR) task, in which the textual premise serves as the presumptions to the source images. With selected images and human-curated premise templates, the collected dataset is formulated as a 4-choice classification task, in which we ask human annotators to write one true hypothetical action and three distracting actions. The PMR task gains insights from both the visual commonsense reasoning and the natural language inference tasks. We hope the proposed dataset and task can pave the

way for human-like cognitive reasoning from both visual and textual clues.

Acknowledgements

This paper is supported by the National Key Research and Development Program of China 2020AAA0106700 and NSFC project U19A2065.

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.
- Samuel Bowman and Xiaodan Zhu. 2019. Deep learning for natural language inference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 6–8.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, pages 177–190, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- Albert Ellis. 1995. Changing rational-emotive therapy (RET) to rational emotive behavior therapy (REBT). *Journal of Rational-Emotive and Cognitive-Behavior Therapy*, 13(2):85–89.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. *CoRR*, abs/1803.02324.
- Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.
- Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.
- Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. 2016. Building Machines That Learn and Think Like People. *arXiv*.
- 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*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Jae Sung Park, Chandra Bhagavatula, Roozbeh Motlaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision*, pages 508–524. Springer.
- Shuhuai Ren, Junyang Lin, Guangxiang Zhao, Rui Men, An Yang, Jingren Zhou, Xu Sun, and Hongxia Yang. 2021. Learning relation alignment for calibrated cross-modal retrieval. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 514–524, Online. Association for Computational Linguistics.
- Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision*, 123(1):94–120.
- Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vi-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*.
- Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.
- Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy. Association for Computational Linguistics.

- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.
- Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.
- Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2020. [Ernie-vil: Knowledge enhanced vision-language representations through scene graph](#). *CoRR*, abs/2006.16934.
- Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2018. [From Recognition to Cognition: Visual Commonsense Reasoning](#). *arXiv*.

Appendix

A Author Statement

Hereby we confirm that we bear all responsibility in case of violation of rights, etc., and confirmation of the data license. This work is licensed under a CC BY-NC license, and both images and annotations can be accessed at <https://2030nlp.github.io/PMR/>.

B Additional Data Analysis

We explore the features of the language in PMR. As is shown in Table 1, the number of 2-grams has reached 84,365, the length of premises average to over 9, and additionally, our action texts average at more than 14 words, which is higher than the common question answering datasets and indicates PMR has high language complexity and diversity. Furthermore, we calculate the average number of objects mentioned in each sample. Since PMR aims to test the model’s capability to infer the action of people, the objects mentioned are mainly persons, but due to our proposed generation strategy of distractors, the text also mention a variety of other objects, such as ‘bowl’ and ‘chair’. Figure 10 shows the objects distribution both in text(reference) and images(total).

C Details for Adversarial Samples Generation

For each example, manually writing three distractors is expensive and unscalable, which costs nearly 1.5 times as much as a single question-answer pair. Therefore, Zellers et al. proposes Adversarial Matching to obtain high-quality negatives automatically. For a bunch of question-answer pairs, it aims to take the answer from other question as a negative for a question, with the constraint that the negative is supposed to be relevant to the question, but not overly similar to its true positive at the same time.

Inspired by it, we regarded the 15k premises and their Action-Trues as pairs and took a similar method to double the scale of PMR. Specifically, we chose the most related three actions from other pairs as negatives for a premise according to the relevance scores. Firstly, for a premise p_i , in order not to conflict with o_i , the objects list of p_i ’s corresponding image, we performed remapping³, that’s

³<https://groups.google.com/group/visualcommonsense/attach/500d18f416f1b/dataloader.py?part=0.1>

substituting the objects mentioned in candidate actions with objects in o_i . Secondly, we calculated the relevance scores for p_i with each remapped action by a bert-base model, which was trained with randomly sampled premise action pairs. Then, we picked the top3 actions with the highest confidence as negatives.

This method is simple but effective and efficient compared to the original adversarial matching. Due to the remapping strategy, the objects in actions are going to be randomly substituted, which may cause the contradiction with the premise and thus avoid them being false negatives. To figure out the effectiveness of our methods, we performed a human test on the 50 items from the adversarial samples generated by it, and the result showed that the human got an accuracy of 0.86 which demonstrated that our methods can generate high-quality negatives. What’s more, overemphasizing the difference with the true positive tends to lead the distractors to be more distinguishable and lower the difficulty of datasets, which was testified by the performance of VL-BERT with an accuracy of 1.0 on the test set of adversarial samples generated in the same way of VCR.

Therefore, we maintain that the remapping strategy is enough to balance the difficulty with the correctness of adversarial samples.

D Manually Written Templates for Premises

We list all the templates for premises by category. "[]" indicates the slots to be filled with words of different types.

1. Relationship

- They are [N].
- [person1] and [person2] are [N].
- The two speakers are [N].
- [person1] is [person2]’s [n].
- [person1] is talking with his/her [n].
- Because of what happened before, they have a [adj] relationship now.
- The relationship between [person1] and [person2] is very [adj].

2. Personality

- [person1] is [adj].
- [person1] has a [adj] temper.
- Both [person1] and [person2] are [adj].
- [person1]’s personality is very [adj].

- [person1] is thought to be a/an [adj] man.
- [person1] gives people a [adj] feeling.
- They all know that [person1] is a/an [adj] person.
- [person1] is considered to be a/an [adj] person.
- [person1] is quite [adj] about most things.

3. Identity

- [person1] is a [job1].
- [person1]'s job is a [job1].
- The occupation of [person1] is a [job1].
- [person1] works in [place1].
- [person1] works for a/an [place1].
- The line of work that the [person1] is in is [place1]

4. Antecedent

- There was a/an [incident1].
- [incident1] happened.
- [person1] and [person2] had a fight.
- [person1] had no idea who [person2] was supposed to be.
- [person1] and [person2] were very familiar with each other.
- [person1] broke things on accident.
- Something unfortunate had just happened.
- Something fortunate had just happened.
- [person1] just lost [relative1].
- [person1] just got a new [relative1].

5. Mood

- [person1] is [adj].
- [person1] feels [adj].
- To [person1], nothing could be more [adj].
- Having a conversation with [person2], [person1] is very [adj].
- [person1] has a [adj] time with [person2].
- Because of [person2]'s behavior, [person1] feels very [adj].
- [person1] is so [adj] to talk with [person2].
- people have [adj] looks on their faces.
- [person1] looks [adj] today.
- Recently, [person1] is becoming more and more [adj].
- This scene makes them very [adj].
- [person1] makes [person2] [adj].
- [person1] is feeling a bit [adj].
- [person1] is in a [adj] mood today.
- [person1] is filled with [n].

6. Environment

- It is [weather1].
- The atmosphere is [emotion1].
- There is a/an [emotion1] in the air.
- This is telling a [emotion1] story.
- The scene is [emotion1].

E Payment for Worker

Crowd workers performed annotations mainly in phase1 and phase2, and we paid them different prices according to the workload and the quality. In Phase 1, workers are responsible for selecting the premise and writing two hypothetical actions. If accepted in Phase 2, they would be paid 0.15\$ for each sample. And they would only obtain 1.5 cents per sample if rejected. In Phase 2, the total salary was calculated by multiplying the basic salary by qualification rate. For basic salary, if anyone accepted the annotation in Phase 1 and generate another two distractors, he or she would get 7.5 cents, otherwise, they only got 1.5 cents (No distracter needs to be generated in such a case certainly). In terms of qualification rate, it was obtained by checking the quality of the result in Phase 2, and we would calculate it for every worker. On average, the annotators were paid about 3 times as much as the prevailing local minimum wage per hour.

F Data Availability and Copyright

According to Section 107 of the Copyright Law⁴, and 28A and 30 of the Copyright Acts⁵, there is one exception to copyright infringement which is fair use (or fair dealing). Fair use is appropriate for public benefit purposes, like research. Our use is not of commercial nature. Besides, we only use texts that are publicly available, and the source will be stated according to law. Users can download the images directly from the original source.

G Experimental Details

We list all the hyperparameters training different models in Table 6.

H Annotation Interface

We present instruction for annotation in Figure 6, and the interfaces⁶ used in 3 phases in Figure 7, Figure 8 and Figure 9.

⁴<https://www.copyright.gov/title17/92chap1.html#107>

⁵<https://www.gov.uk/government/publications/copyright-acts-and-related-laws>

⁶The interfaces is constructed referring to the code from <https://visualcommonsense.com/explore>

	BS	CPU/GPU	LR	Optimizer	Warmup Steps	Epochs/Steps
BERT	16	CPU	5e-5	Adam	500	10 epochs
VL-BERT-Base	4	4 2080T GPUs	7e-5	SGD	1000	20 epochs
ERNIE-VIL-Base	4	1 2080T GPUs	2e-5	Adam	8000	22500 steps
UNITER-Base	16	1 2080T GPUs	3e-5	AdamW	1084	20 epochs
VL-BERT-Large	4	4 A40 GPUs	7e-5	SGD	1000	20 epochs
ERNIE-VIL-Large	4	1 A40 GPUs	2e-5	Adam	8000	22500 steps
UNITER-Large	16	1 A40 GPUs	3e-5	AdamW	1084	20 epochs

Table 6: Hyperparameters for training.

Instructions

- According to the image and its label information (click "show/hide" below), choose one of the six optional premises that can be used as the supplementary information of the image (if the options given are not suitable, you can modify or rewrite one to fill in the blank freely).
- Write two hypothetical actions, which describe what will happen next. Among them, action-True containing image information and meets the chosen premise. In contrast, action-False contains image information but **does not** meet the chosen premise.
- Please enclose the words with "[]" that mention the information of image in both written actions.

Tip: If there is no idea, you can click "Unfold Action" to get the reference items.

Note:

- For high-quality results, please read [More Instructions](#) conscientiously before annotating.
- Text** is supposed to be filled in the blank instead of the indexes of reference premises or actions.
- No other language except **English** should be used.
- PERSON should be enclosed with [] if you want to refer to someone from the image in the premise and actions. That's, to write [person1] instead of person1 or person 1.

Figure 6: Annotation Instruction.

Fold Premise

- [person2] just lost friend.
- [person3] is conscientious.
- There is a/an boring in the air.
- [person1] is a drummer.
- The relationship between [person1] and [person3] is very bumpy.
- [person3] makes [person2] vibrant.

Unfold Action

hide all [person1] [person2] [person3] [car1] [pottedplant1]

show all [pottedplant2]

Premise

Hypothetical Action-True

Hypothetical Action-False

Submit

Figure 7: Annotation interface for phase 1.

hide all [person1] [person2] [person3] [person4] [person5]

show all

Fold Premise
The atmosphere is cheerful.

Fold Action
True: {[person4] in an orange vest} finds something interesting and he will share that with his friends.
False: {[person4] in an orange vest} finds something horrible and he will run away.

Premise
The atmosphere is cheerful.

Action-true
[person4] in (an orange vest) finds something interesting and he will share that with his friends.

Distractor-true
[person4] in (an orange vest) finds something horrible and he will run away.

Action-false
[person4] in (a white gown) finds something interesting and he will share that with his friends.

Distractor-false
[person4] in (a white gown) finds something horrible and he will run away.

Reject
 Accept

Submit

Figure 8: Annotation interface for phase 2.

hide all [person1] [person2] [person3] [tie1]

show all

Fold Premise
context-o: [person1] just got a new rival.
context-e: [person1] just got a new rival.

Fold Action
Action-True-o: {[person1]} is talking to (the person in front of her) about her new competitor.
Action-True-e: {[person1]} is talking to (the person in front of her) about her new competitor.
Action-False-o: {[person1]} is talking to (the person in front of her) about her new friends.
Action-False-e: {[person1]} is talking to (the person in front of her) about her new friends.

Distractor-True: {[person1]} is talking to (the policeman) about her new competitor.
Distractor-False: {[person1]} is talking to (the doctor) about her new friends.

Examer-Accept: True
Examer-Reject: False

Accept
 Reject

Submit

Figure 9: Interface for quality control.

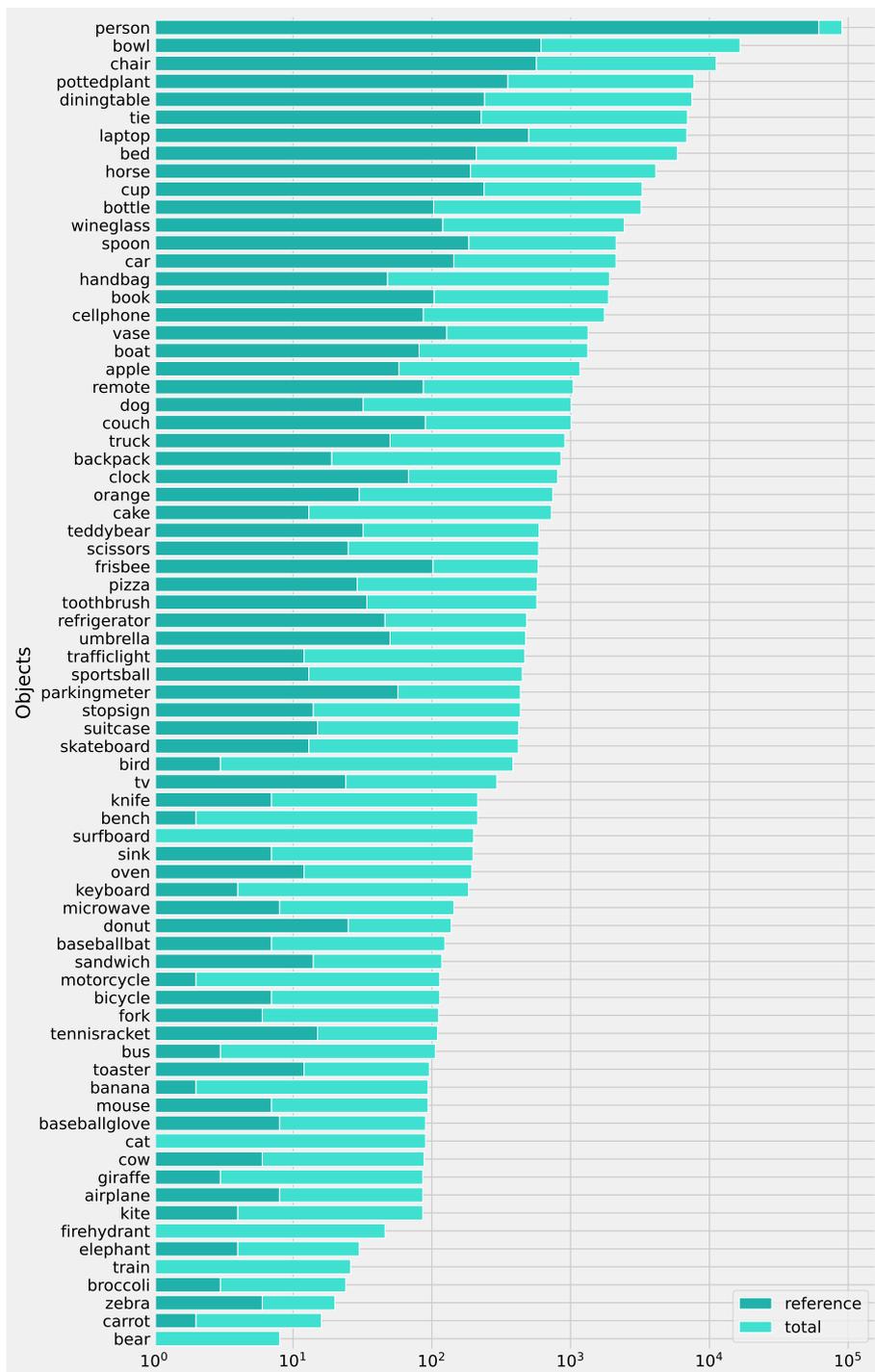


Figure 10: Objects distribution.