From Recognition to Cognition: Visual Commonsense Reasoning

Rowan Zellers[♠] Yonatan Bisk[♠] Ali Farhadi[♠♡] Yejin Choi[♠♡]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[♡]Allen Institute for Artificial Intelligence
visualcommonsense.com

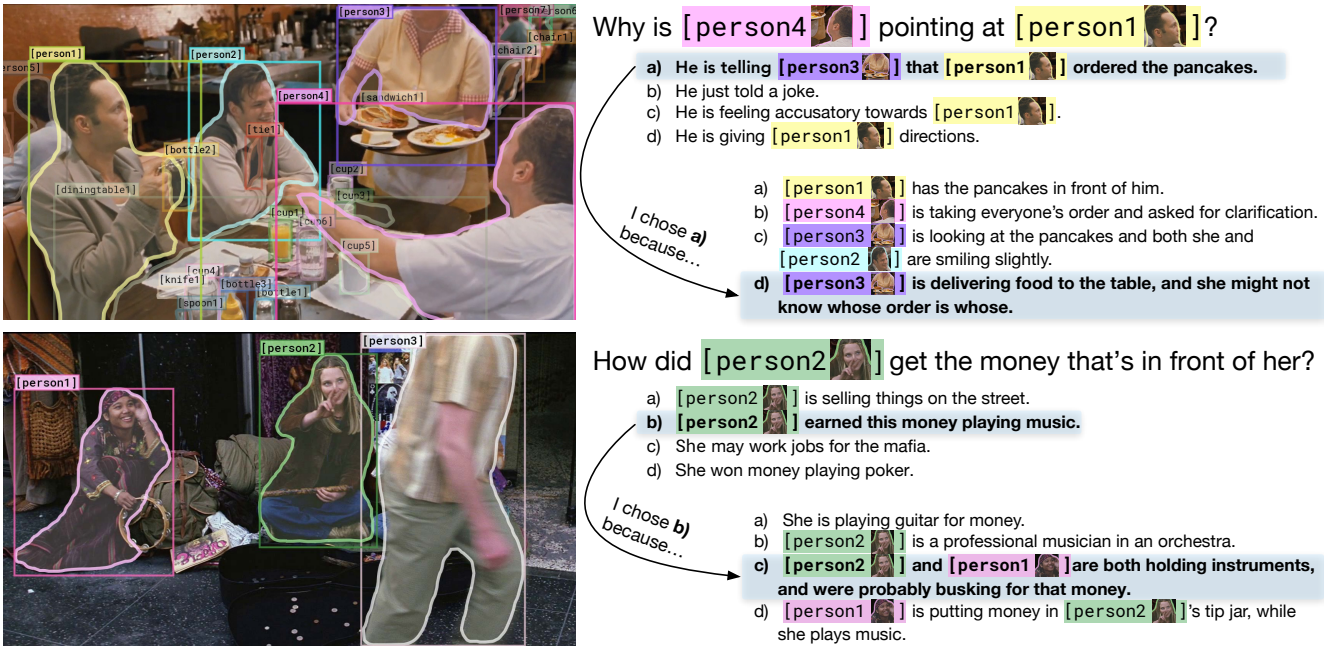


Figure 1: VCR: Given an image, a list of regions, and a question, a model must answer the question and provide a *rationale* explaining why its answer is right. Our questions challenge computer vision systems to go beyond recognition-level understanding, towards a higher-order cognitive and commonsense understanding of the world depicted by the image.

Abstract

*Visual understanding goes well beyond object recognition. With one glance at an image, we can effortlessly imagine the world beyond the pixels: for instance, we can infer people's actions, goals, and mental states. While this task is easy for humans, it is tremendously difficult for today's vision systems, requiring higher-order cognition and commonsense reasoning about the world. We formalize this task as **Visual Commonsense Reasoning**. Given a challenging question about an image, a machine must answer correctly and then provide a rationale justifying its answer.*

*Next, we introduce a new dataset, VCR, consisting of 290k multiple choice QA problems derived from 110k movie scenes. The key recipe for generating non-trivial and high-quality problems at scale is **Adversarial Matching**, a new approach to transform rich annotations into multiple choice questions with minimal bias. Experimental results show that while humans find VCR easy (over 90% accuracy), state-of-the-art vision models struggle (~45%).*

*To move towards cognition-level understanding, we present a new reasoning engine, Recognition to Cognition Networks (**R2C**), that models the necessary layered inferences for grounding, contextualization, and reasoning. **R2C** helps narrow the gap between humans and machines (~65%); still, the challenge is far from solved, and we provide analysis that suggests avenues for future work.*

1. Introduction

With one glance at an image, we can immediately infer what is happening in the scene beyond what is visually obvious. For example, in the top image of Figure 1, not only do we see several objects (people, plates, and cups), we can also reason about the entire situation: three people are dining together, they have already ordered their food before

the photo has been taken, [person3] is serving and not eating with them, and what [person1] ordered are the pancakes and bacon (as opposed to the cheesecake), because [person4] is pointing to [person1] while looking at the server, [person3].

Visual understanding requires seamless integration between *recognition* and *cognition*: beyond recognition-level perception (e.g., detecting objects and their attributes), one must perform cognition-level reasoning (e.g., inferring the likely intents, goals, and social dynamics of people) [13]. State-of-the-art vision systems can reliably perform *recognition-level* image understanding, but struggle with complex inferences, like those in Figure 1. We argue that as the field has made significant progress on recognition-level building blocks, such as object detection, pose estimation, and segmentation, now is the right time to tackle cognition-level reasoning at scale.

As a critical step toward complete visual understanding, we present the task of **Visual Commonsense Reasoning**. Given an image, a machine must answer a question that requires a thorough understanding of the visual world evoked by the image. Moreover, the machine must provide a rationale justifying why that answer is true, referring to the details of the scene, as well as background knowledge about how the world works. These questions, answers, and rationales are expressed using a mixture of rich natural language as well as explicit references to image regions. To support clean-cut evaluation, all our tasks are framed as multiple choice QA.

Our new dataset for this task, VCR, is the first of its kind and is large-scale — 290k pairs of questions, answers, and rationales, over 110k unique movie scenes. A crucial challenge in constructing a dataset of this complexity at this scale is how to avoid annotation artifacts. A recurring challenge in most recent QA datasets has been that human-written answers contain unexpected but distinct biases that models can easily exploit. Often these biases are so prominent so that models can select the right answers without even looking at the questions [28, 61, 72].

Thus, we present **Adversarial Matching**, a novel QA assignment algorithm that allows for robust multiple-choice dataset creation at scale. The key idea is to recycle each correct answer for a question exactly three times — as a negative answer for three other questions. Each answer thus has the same probability (25%) of being correct: this resolves the issue of answer-only biases, and disincentivizes machines from always selecting the most generic answer. We formulate the answer recycling problem as a constrained optimization based on the relevance and entailment scores between each candidate negative answer and the gold answer, as measured by state-of-the-art natural language inference models [10, 57, 15]. A neat feature of our recycling algorithm is a knob that can control the tradeoff between human and machine difficulty: we want the problems to be hard for machines while easy for humans.

Narrowing the gap between recognition- and cognition-level image understanding requires grounding the meaning of the natural language passage in the visual data, understanding the answer in the context of the question, and reasoning over the shared and grounded understanding of the question, the answer, the rationale and the image. In this paper we introduce a new model, **Recognition to Cognition Networks** (**R2C**). Our model performs three inference steps. First, it *grounds* the meaning of a natural language passage with respect to the image regions (objects) that are directly referred to. It then *contextualizes* the meaning of an answer with respect to the question that was asked, as well as the global objects not mentioned. Finally, it *reasons* over this shared representation to arrive at an answer.

Experiments on VCR show that **R2C** greatly outperforms state-of-the-art visual question-answering systems: obtaining 65% accuracy at question answering, 67% at answer justification, and 44% at staged answering and justification. Still, the task and dataset is far from solved: humans score roughly 90% on each. We provide detailed insights and an ablation study to point to avenues for future research.

In sum, our major contributions are fourfold: (1) we formalize a new task, Visual Commonsense Reasoning, and (2) present a large-scale multiple-choice QA dataset, VCR, (3) that is automatically assigned using Adversarial Matching, a new algorithm for robust multiple-choice dataset creation. (4) We also propose a new model, **R2C**, that aims to mimic the layered inferences from recognition to cognition; this also establishes baseline performance on our new challenge. The dataset is available to download, along with code for our model, at visualcommonsense.com.

2. Task Overview

We present VCR, a new task that challenges vision systems to holistically and cognitively understand the content of an image. For instance, in Figure 1, we need to understand the activities ([person3] is delivering food), the roles of people ([person1] is a customer who previously ordered food), the mental states of people ([person1] wants to eat), and the likely events before and after the scene ([person3] will serve the pancakes next). Our task covers these categories and more: a distribution of the inferences required is in Figure 2.

Visual understanding requires not only answering questions correctly, but doing so *for the right reasons*. We thus require a model to give a *rationale* that explains why its answer is true. Our questions, answers, and rationales are written in a mixture of rich natural language as well as detection tags, like '[person2]': this helps to provide an unambiguous link between the textual description of an

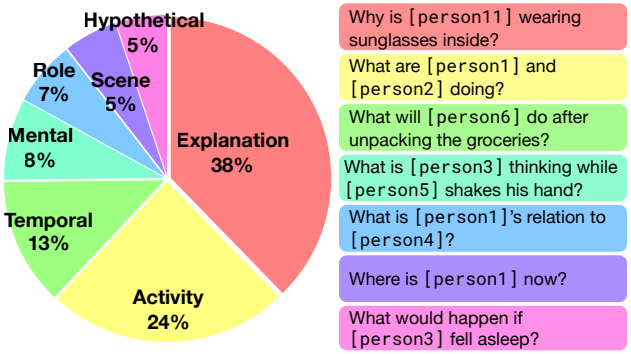


Figure 2: Overview of the types of inference required by questions in VCR. Of note, 38% of the questions are explanatory 'why' or 'how' questions, 24% involve cognition-level activities, and 13% require temporal reasoning (i.e., what might come next). These categories are not mutually exclusive; an answer might require several hops of different types of inferences (see appendix Sec A).

object ('the man on the left in the white shirt') and the corresponding image region.

To make evaluation straightforward, we frame our ultimate task – of staged answering and justification – in a multiple-choice setting. Given a question along with four answer choices, a model must first select the right answer. If its answer was correct, then it is provided four rationale choices (that could purportedly justify its correct answer), and it must select the correct rationale. We call this $Q \rightarrow AR$ as for the model prediction to be correct requires *both the chosen answer and then the chosen rationale* to be correct.

Our task can be decomposed into two multiple-choice sub-tasks, that correspond to answering ($Q \rightarrow A$) and justification ($QA \rightarrow R$) respectively:

Definition VCR *subtask.* A single example of a VCR subtask consists of an image $\boldsymbol{I}$, and:

- A sequence $\boldsymbol{o}$ of object detections. Each object detection o_i consists of a *bounding box* $\mathbf{b}$, a segmentation mask $\mathbf{m}$[1], and a class label $\ell_i \in \mathcal{L}$.
- A *query* $\boldsymbol{q}$, posed using a mix of natural language and pointing. Each word $\boldsymbol{q}_i$ in the query is either a word in a vocabulary $\mathcal{V}$, or is a tag referring to an object in $\boldsymbol{o}$.
- A set of N *responses*, where each response $\boldsymbol{r}^{(i)}$ is written in the same manner as the query: with natural language and pointing. Exactly one response is correct.

The model chooses a single (best) response.

In question-answering ($Q \rightarrow A$), the query is the question and the responses are answer choices. In answer justification ($QA \rightarrow R$), the query is the concatenated question and correct answer, while the responses are rationale choices.

[1]The task is agnostic to the representation of the mask, but it could be thought of as a list of polygons $\boldsymbol{p}$, with each polygon consisting of a sequence of 2d vertices inside the box $\boldsymbol{p}_j = \{x_t, y_t\}_t$.

In this paper, we evaluate models in terms of accuracy and use N=4 responses. Baseline accuracy on each subtask is then 25% ($1/N$). In the holistic setting ($Q \rightarrow AR$), baseline accuracy is 6.25% ($1/N^2$) as there are two subtasks.

3. Data Collection

In this section, we describe how we collect the questions, *correct answers* and *correct rationales* for VCR. Our key insight – towards collecting commonsense visual reasoning problems at scale – is to carefully select interesting situations. We thus extract still images from movie clips. The images from these clips describe complex situations that humans can decipher without additional context: for instance, in Figure 1, we know that [person3] will serve [person1] pancakes, whereas a machine might not understand this unless it sees the entire clip.

Interesting and Diverse Situations To ensure diversity, we make no limiting assumptions about the predefined set of actions. Rather than searching for predefined labels, which can introduce search engine bias [76, 16, 20], we collect images from movie scenes. The underlying scenes come from the Large Scale Movie Description Challenge [67] and YouTube movie clips.[2] To avoid simple images, we train and apply an 'interestingness filter' (e.g. a closeup of a syringe in Figure 3).[3]

We center our task around challenging questions requiring cognition-level reasoning. To make these cognition-level questions simple to ask, and to avoid the clunkiness of referring expressions, VCR's language integrates object tags ([person2]) and explicitly excludes referring expressions ('the woman on the right.') These object tags are detected from Mask-RCNN [29, 24], and the images are filtered so as to have at least three high-confidence tags.

Crowdsourcing Quality Annotations Workers on Amazon Mechanical Turk were given an image with detections, along with additional context in the form of video captions.[4] They then ask one to three questions about the image; for each question, they provide a reasonable answer and a rationale. To ensure top-tier work, we used a system of quality checks and paid our workers well.[5]

The result is an underlying dataset with high agreement and diversity of reasoning. Our dataset contains a myriad of interesting commonsense phenomena (Figure 2) and a great diversity in terms of unique examples (Supp Section A); almost every answer and rationale is unique.

[2]Namely, Fandango MovieClips: youtube.com/user/movieclips.

[3]We annotated images for 'interestingness' and trained a classifier using CNN features and detection statistics, details in the appendix, Sec B.

[4]This additional clip-level context helps workers ask and answer about what will happen next.

[5]More details in the appendix, Sec B.

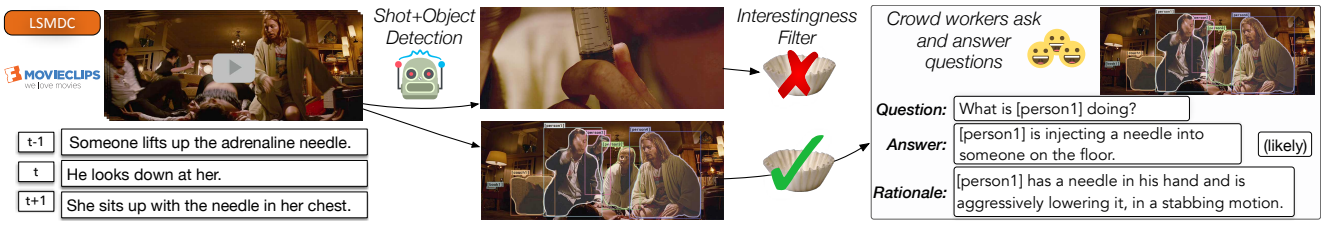


Figure 3: An overview of the construction of VCR. Using a state-of-the-art object detector [29, 24], we identify the objects in each image. The most interesting images are passed to crowd workers, along with scene-level context in the form of scene descriptions (MovieClips) and video captions (LSMDC, [67]). The crowd workers use a combination of natural language and detection tags to ask and answer challenging visual questions, also providing a rationale justifying their answer.

4. Adversarial Matching

We cast VCR as a four-way multiple choice task, to avoid the evaluation difficulties of language generation or captioning tasks where current metrics often prefer incorrect machine-written text over correct human-written text [49]. However, it is not obvious how to obtain high-quality incorrect choices, or counterfactuals, at scale. While past work has asked humans to write several counterfactual choices for each correct answer [75, 46], this process is expensive. Moreover, it has the potential of introducing annotation artifacts: subtle patterns that are by themselves highly predictive of the 'correct' or 'incorrect' label [72, 28, 61].

In this work, we propose Adversarial Matching: a new method that allows for any 'language generation' dataset to be turned into a multiple choice test, while requiring minimal human involvement. An overview is shown in Figure 4. Our key insight is that the problem of obtaining good counterfactuals can be broken up into two subtasks: the counterfactuals must be as **relevant** as possible to the context (so that they appeal to machines), while they cannot be overly **similar** to the correct response (so that they don't become correct answers incidentally). We balance between these two objectives to create a dataset that is challenging for machines, yet easy for humans.

Formally, our procedure requires two models: one to compute the relevance between a query and a response, P_{rel}, and another to compute the similarity between two response choices, P_{sim}. Here, we employ state-of-the-art models for Natural Language Inference: BERT [15] and ESIM+ELMo [10, 57], respectively.[6] Then, given dataset examples $(\boldsymbol{q}_i, \boldsymbol{r}_i)_{1\leq i\leq N}$, we obtain a counterfactual for each $\boldsymbol{q}_i$ by performing maximum-weight bipartite matching [55, 40] on a weight matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$, given by

$$\mathbf{W}_{i,j} = \log(P_{rel}(\boldsymbol{q}_i, \boldsymbol{r}_j)) + \lambda \log(1 - P_{sim}(\boldsymbol{r}_i, \boldsymbol{r}_j)). \quad (1)$$

Here, λ>0 controls the tradeoff between similarity and relevance.[7] To obtain multiple counterfactuals, we perform several bipartite matchings. To ensure that the negatives are diverse, during each iteration we replace the similarity term with the maximum similarity between a candidate response $\boldsymbol{r}_j$ and all responses currently assigned to $\boldsymbol{q}_i$.

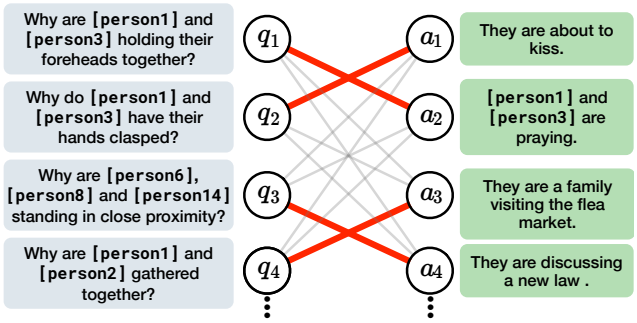


Figure 4: Overview of Adversarial Matching. Incorrect choices are obtained via maximum-weight bipartite matching between queries and responses; the weights are scores from state-of-the-art natural language inference models. Assigned responses are highly relevant to the query, while they differ in meaning versus the correct responses.

Ensuring dataset integrity To guarantee that there is no question/answer overlap between the training and test sets, we split our full dataset (by movie) into 11 folds. We match the answers and rationales invidually for each fold. Two folds are pulled aside for validation and testing.

5. Recognition to Cognition Networks

We introduce Recognition to Cognition Networks (**R2C**), a new model for visual commonsense reasoning. To perform well on this task requires a deep understanding of language, vision, and the world. For example, in Figure 5, answering '*Why is* [person4] *pointing at* [person1]*?*' requires multiple inference steps. First, we **ground** the meaning of the query and each response, which involves referring to the image for the

[6] We finetune P_{rel} (BERT), on the annotated data (taking steps to avoid data leakage), whereas P_{sim} (ESIM+ELMo) is trained on entailment and paraphrase data - details in appendix Sec C.

[7] We tuned this hyperparameter by asking crowd workers to answer multiple-choice questions at several thresholds, and chose the value for which human performance is above 90% - details in appendix Sec C.

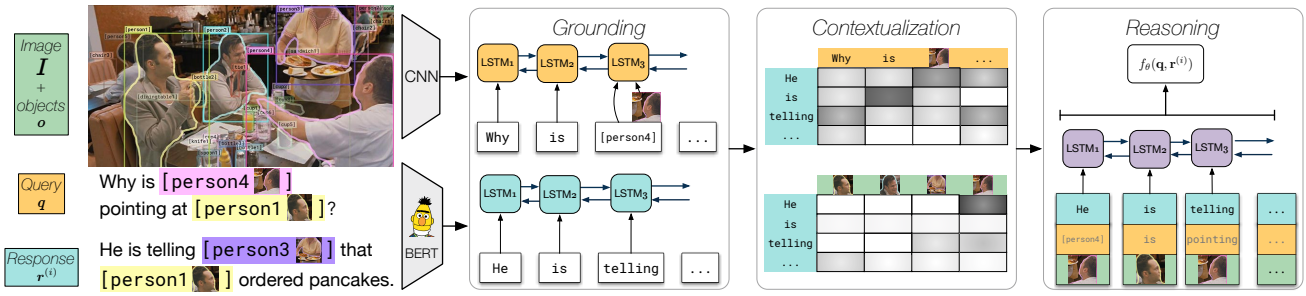


Figure 5: High-level overview of our model, **R2C**. We break the challenge of Visual Commonsense Reasoning into three components: grounding the query and response, contextualizing the response within the context of the query and the entire image, and performing additional reasoning steps on top of this rich representation.

two people. Second, we **contextualize** the meaning of the query, response, and image together. This step includes resolving the referent 'he,' and why one might be pointing in a diner. Third, we **reason** about the interplay of relevant image regions, the query, and the response. In this example, the model must determine the social dynamics between [person1] and [person4]. We formulate our model as three high-level stages: grounding, contextualization, and reasoning, and use standard neural building blocks to implement each component.

In more detail, recall that a model is given an image, a set of objects $\boldsymbol{o}$, a query $\boldsymbol{q}$, and a set of responses $\boldsymbol{r}^{(i)}$ (of which exactly one is correct). The query $\boldsymbol{q}$ and response choices $\boldsymbol{r}^{(i)}$ are all expressed in terms of a mixture of natural language and pointing to image regions: notation-wise, we will represent the object tagged by a word w as o_w. If w isn't a detection tag, o_w refers to the entire image boundary. Our model will then consider each response $\boldsymbol{r}$ separately, using the following three components:

Grounding The grounding module will learn a joint image-language representation for each token in a sequence. Because both the query and the response contain a mixture of tags and natural language words, we apply the same grounding module for each (allowing it to share parameters). At the core of our grounding module is a bidirectional LSTM [34] which at each position is passed as input a word representation for w_i, as well as visual features for o_{w_i}. We use a CNN to learn object-level features: the visual representation for each region o is Roi-Aligned from its bounding region [63, 29]. To additionally encode information about the object's class label ℓ_o, we project an embedding of ℓ_o (along with the object's visual features) into a shared hidden representation. Let the output of the LSTM over all positions be $\mathbf{r}$, for the response and $\mathbf{q}$ for the query.

Contextualization Given a grounded representation of the query and response, we use attention mechanisms to contextualize these sentences with respect to each other and the image context. For each position i in the response, we will define the attended query representation as $\hat{\mathbf{q}}_i$ using the following equation:

$$\alpha_{i,j} = \underset{j}{\text{softmax}}(\mathbf{r}_i \mathbf{W} \mathbf{q}_j) \qquad \hat{\mathbf{q}}_i = \sum_j \alpha_{i,j} \mathbf{q}_j. \tag{2}$$

To contextualize an answer with the image, including implicitly relevant objects that have not been picked up from the grounding stage, we perform another bilinear attention between the response $\mathbf{r}$ and each object o's image features. Let the result of the object attention be $\hat{\mathbf{o}}_i$.

Reasoning Last, we allow the model to *reason* over the response, attended query and objects. We accomplish this using a bidirectional LSTM that is given as context $\hat{\mathbf{q}}_i$, $\mathbf{r}_i$, and $\hat{\mathbf{o}}_i$ for each position i. For better gradient flow through the network, we concatenate the output of the reasoning LSTM along with the question and answer representations for each timestep: the resulting sequence is max-pooled and passed through a multilayer perceptron, which predicts a logit for the query-response compatibility.

Neural architecture and training details For our image features, we use ResNet50 [30]. To obtain strong representations for language, we used BERT representations [15]. BERT is applied over the entire question and answer choice, and we extract a feature vector from the second-to-last layer for each word. We train **R2C** by minimizing the multi-class cross entropy between the prediction for each response $\boldsymbol{r}^{(i)}$, and the gold label. See the appendix (Sec E) for detailed training information and hyperparameters.[8]

6. Results

In this section, we evaluate the performance of various models on VCR. Recall that our main evaluation mode is the staged setting ($Q \rightarrow AR$). Here, a model must choose the right answer for a question (given four answer choices), and then choose the right rationale for that question and answer (given four rationale choices). If it gets either the answer or the rationale wrong, the entire prediction will be wrong. This holistic task decomposes into two sub-tasks wherein we can train individual models: question answering ($Q \rightarrow A$)

[8] Our code is also available online at visualcommonsense.com.

	Model	$Q \rightarrow A$ Val	$Q \rightarrow A$ Test	$QA \rightarrow R$ Val	$QA \rightarrow R$ Test	$Q \rightarrow AR$ Val	$Q \rightarrow AR$ Test
	Chance	25.0	25.0	25.0	25.0	6.2	6.2
Text Only	BERT	53.8	53.9	64.1	64.5	34.8	35.0
Text Only	BERT (response only)	27.6	27.7	26.3	26.2	7.6	7.3
Text Only	ESIM+ELMo	45.8	45.9	55.0	55.1	25.3	25.6
Text Only	LSTM+ELMo	28.1	28.3	28.7	28.5	8.3	8.4
VQA	RevisitedVQA [38]	39.4	40.5	34.0	33.7	13.5	13.8
VQA	BottomUpTopDown[4]	42.8	44.1	25.1	25.1	10.7	11.0
VQA	MLB [42]	45.5	46.2	36.1	36.8	17.0	17.2
VQA	MUTAN [6]	44.4	45.5	32.0	32.2	14.6	14.6
	R2C	**63.8**	**65.1**	**67.2**	**67.3**	**43.1**	**44.0**
	Human		91.0		93.0		85.0

Table 1: Experimental results on VCR. VQA models struggle on both question-answering ($Q \rightarrow A$) as well as answer justification ($Q \rightarrow AR$), possibly due to the complex language and diversity of examples in the dataset. While language-only models perform well, our model **R2C** obtains a significant performance boost. Still, all models underperform human accuracy at this task. For more up-to-date results, see the leaderboard at visualcommonsense.com/leaderboard.

as well as answer justification ($QA \rightarrow R$). Thus, in addition to reporting combined $Q \rightarrow AR$ performance, we will also report $Q \rightarrow A$ and $QA \rightarrow R$.

Task setup A model is presented with a query $\boldsymbol{q}$, and four response choices $\boldsymbol{r}^{(i)}$. Like our model, we train the baselines using multi-class cross entropy between the set of responses and the label. Each model is trained separately for question answering and answer justification.[9]

6.1. Baselines

We compare our **R2C** to several strong language and vision baselines.

Text-only baselines We evaluate the level of visual reasoning needed for the dataset by also evaluating purely text-only models. For each model, we represent $\boldsymbol{q}$ and $\boldsymbol{r}^{(i)}$ as streams of tokens, with the detection tags replaced by the object name (e.g. `chair5` → `chair`). To minimize the discrepancy between our task and pretrained models, we replace person detection tags with gender-neutral names.

a. **BERT** [15]: BERT is a recently released NLP model that achieves state-of-the-art performance on many NLP tasks.

b. **BERT (response only)** We use the same BERT model, however, during fine-tuning and testing the model is only given the response choices $\boldsymbol{r}^{(i)}$.

c. **ESIM+ELMo** [10]: ESIM is another high performing model for sentence-pair classification tasks, particularly when used with ELMo embeddings [57].

[9] We follow the standard train, val and test splits.

Model	$Q \rightarrow A$	$QA \rightarrow R$	$Q \rightarrow AR$
R2C	**63.8**	**67.2**	**43.1**
No query	48.3	43.5	21.5
No reasoning module	63.6	65.7	42.2
No vision representation	53.1	63.2	33.8
GloVe representations	46.4	38.3	18.3

Table 2: Ablations for **R2C**, over the validation set. 'No query' tests the importance of integrating the query during contextualization; removing this reduces $Q \rightarrow AR$ performance by 20%. In 'no reasoning', the LSTM in the reasoning stage is removed; this hurts performance by roughly 1%. Removing the visual features during grounding, or using GloVe embeddings rather than BERT, lowers performance significantly, by 10% and 25% respectively.

d. **LSTM+ELMo**: Here an LSTM with ELMo embeddings is used to score responses $\boldsymbol{r}^{(i)}$.

VQA Baselines Additionally we compare our approach to models developed on the VQA dataset [5]. All models use the same visual backbone as **R2C** (ResNet 50) as well as text representations (GloVe; [56]) that match the original implementations.

e. **RevisitedVQA** [38]: This model takes as input a query, response, and image features for the entire image, and passes the result through a multilayer perceptron, which has to classify 'yes' or 'no'.[10]

f. **Bottom-up and Top-down attention** (BottomUpTopDown) [4]: This model attends over region proposals given by an object detector. To adapt to VCR, we pass this model object regions referenced by the query and response.

g. **Multimodal Low-rank Bilinear Attention** (MLB) [42]: This model uses Hadamard products to merge the vision and language representations given by a query and each region in the image.

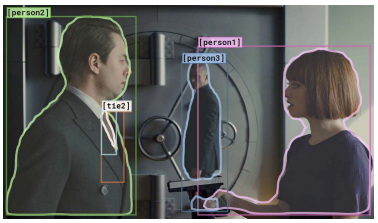
h. **Multimodal Tucker Fusion** (MUTAN) [6]: This model expresses joint vision-language context in terms of a tensor decomposition, allowing for more expressivity.

We note that BottomUpTopDown, MLB, and MUTAN all treat VQA as a multilabel classification over the top 1000 answers [4, 50]. Because VCR is highly diverse (Supp A), for these models we represent each response $\boldsymbol{r}^{(i)}$ using a GRU [11].[11] The output logit for response i is given by the dot product between the final hidden state of the GRU encoding $\boldsymbol{r}^{(i)}$, and the final representation from the model.

Human performance We asked five different workers on Amazon Mechanical Turk to answer 200 dataset questions from the test set. A different set of five workers were asked to choose rationales for those questions and answers. Predictions were combined using a majority vote.

[10] For VQA, the model is trained by sampling positive or negative answers for a given question; for our dataset, we simply use the result of the perceptron (for response $\boldsymbol{r}^{(i)}$) as the i-th logit.

[11] To match the other GRUs used in [4, 42, 6] which encode $\boldsymbol{q}$.



Why is [person1] pointing a gun at [person2]?

a) [person1] wants to kill [person2].(1%)

b) [person1] and [person3] are robbing the bank and [person2] is the bank manager. (71%)

c) [person2] has done something to upset [person1]. (18%)

d) Because [person2] is [person1]'s daughter. [person1] wants to protect [person2]. (8%)

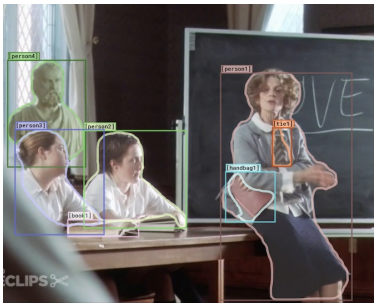
b) is right because...

a) [person1] is chasing [person1] and [person3] because they just robbed a bank. (33%)

b) Robbers will sometimes hold their gun in the air to get everyone's attention. (5%)

c) The vault in the background is similar to a bank vault. [person3] is waiting by the vault for someone to open it. (49%)

d) A room with barred windows and a counter usually resembles a bank. (11%)



What would [person1] do if she caught [person2] and [person3] whispering?

a) [person1] would look to her left. (7%)

b) She would play with [book1]. (7%)

c) She would look concerned and ask what was funny. (39%)

d) She would switch their seats. (45%)

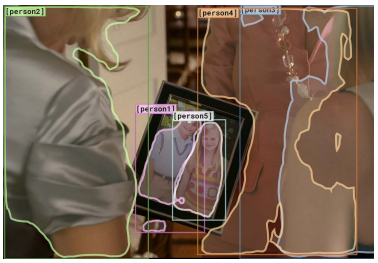
d) is right because...

a) When students are talking in class they're supposed to be listening - the teacher separates them. (64%)

b) Plane seats are very cramped and narrow, and it requires cooperation from your seat mates to help get through. (15%)

c) It's not unusual for people to want to get the closest seats to a stage. (14%)

d) That's one of the only visible seats I can see that's still open, the plane is mostly full. (6%)



What's going to happen next?

a) [person2] is going to walk up and punch [person4] in the face. (10%)

b) Someone is going to read [person4] a bedtime story. (15%)

c) [person2] is going to fall down. (5%)

d) [person2] is going to say how cute [person4]'s children are. (68%)

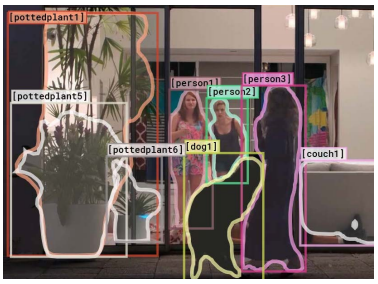
d) is right because...

a) They are the right age to be father and son and [person5] is hugging [person3] like they are his son. (1%)

b) It looks like [person4] is showing the photo to [person2], and [person2] will want to be polite. (31%)

c) [person2] is smirking and looking down at [person4]. (6%)

d) You can see [person4] smiling and facing the crib and decor in the room (60%)



Why can't [person3] go in the house with [person1] and [person2]?

a) She does not want to be there. (12%)

b) [person3] has [dog1] with her. (14%)

c) She needs the light. (45%)

d) She is too freaked out (26%)

b) is right because...

a) [person1] is going away by himself. (60%)

b) [dog1] is small enough to carry. [person3] appears to own him. (33%)

c) If [dog1] was in the house, he would likely knock over [pottedplant6] and likely scratch [couch1]. (4%)

d) [person1] looks like he may have lead [person2] into the room to see [dog1].(1%)

Figure 6: Qualitative examples from **R2C**. Correct predictions are **highlighted in blue**. Incorrect predictions are in red with the correct choices **bolded**. For more predictions, see see visualcommonsense.com/explore.

6.2. Results and Ablations

We present our results in Table 1. Of note, standard VQA models struggle on our task. The best model, in terms of $Q \rightarrow AR$ accuracy, is MLB, with 17.2% accuracy. Deep text-only models perform much better: most notably, BERT [15] obtains 35.0% accuracy. One possible justification for this gap in performance is a bottlenecking effect: whereas VQA models are often built around multilabel classification of the top 1000 answers, VCR requires reasoning over two (often long) text spans. Our model, **R2C** obtains an additional boost over BERT by 9% accuracy, reaching a final performance of 44%. Still, this figure is nowhere near human performance: 85% on the staged task, so there is significant headroom remaining.

Ablations We evaluated our model under several ablations to determine which components are most important. Removing the query representation (and query-response contextualization entirely) results in a drop of 21.6% ac-

curacy points in terms of $Q \rightarrow AR$ performance. Interestingly, this setting allows it to leverage its image representation more heavily: the text based response-only models (BERT response only, and LSTM+ELMo) perform barely better than chance. Taking the reasoning module lowers performance by 1.9%, which suggests that it is beneficial, but not critical for performance. The model suffers most when using GloVe representations instead of BERT: a loss of 24%. This suggests that strong textual representations are crucial to VCR performance.

Qualitative results Last, we present qualitative examples in Figure 6. **R2C** works well for many images: for instance, in the first row, it correctly infers that a bank robbery is happening. Moreover, it picks the right rationale: even though all of the options have something to do with 'banks' and 'robbery,' only **c)** makes sense. Similarly, analyzing the examples for which **R2C** chooses the right answer but the wrong rationale allows us to gain more insight into its understanding of the world. In the third row, the model incorrectly believes there is a crib while assigning less probability mass on the correct rationale - that [person2] is being shown a photo of [person4]'s children, which is why [person2] might say how cute they are.

7. Related Work

Question Answering Visual Question Answering [5] was one of the first large-scale datasets that framed visual understanding as a QA task, with questions about COCO images [49] typically answered with a short phrase. This line of work also includes 'pointing' questions [45, 93] and templated questions with open ended answers [86]. Recent datasets also focus on knowledge-base style content [80, 83]. On the other hand, the answers in VCR are entire sentences, and the knowledge required by our dataset is largely background knowledge about how the world works.

Recent work also includes movie or TV-clip based QA [75, 51, 46]. In these settings, a model is given a video clip, often alongside additional language context such as subtitles, a movie script, or a plot summary.[12] In contrast, VCR features no extra language context besides the question. Moreover, the use of explicit detection tags means that there is no need to perform person identification [66] or linkage with subtitles.

An orthogonal line of work has been on referring expressions: asking to what image region a natural language sentence refers to [60, 52, 65, 87, 88, 59, 36, 33]. We explicitly avoid referring expression-style questions by using indexed detection tags (like [person1]).

Last, some work focuses on commonsense phenomena, such as 'what if' and 'why' questions [79, 58]. However, the space of commonsense inferences is often limited by the underlying dataset chosen (synthetic [79] or COCO [58] scenes). In our work, we ask commonsense questions in the context of rich images from movies.

Explainability AI models are often right, but for questionable or vague reasons [7]. This has motivated work in having models provide explanations for their behavior, in the form of a natural language sentence [31, 9, 41] or an attention map [32, 35, 37]. Our rationales combine the best of both of these approaches, as they involve both natural language text as well as references to image regions. Additionally, while it is hard to evaluate the quality of generated model explanations, choosing the right rationale in VCR is a multiple choice task, making evaluation straightforward.

Commonsense Reasoning Our task unifies work involving reasoning about commonsense phenomena, such as physics [54, 84], social interactions [2, 77, 12, 27], procedure understanding [91, 3] and predicting what might happen next in a video [74, 17, 92, 78, 18, 64, 85].

Adversarial Datasets Past work has proposed the idea of creating adversarial datasets, whether by balancing the dataset with respect to priors [25, 28, 62] or switching them at test time [1]. Most relevant to our dataset construction methodology is the idea of Adversarial Filtering [89].[13] Correct answers are human-written, while wrong answers are chosen from a pool of machine-generated text that is further validated by humans. However, the correct and wrong answers come from fundamentally different sources, which raises the concern that models can cheat by performing authorship identification rather than reasoning over the image. In contrast, in Adversarial Matching, the wrong choices come from the exact same distribution as the right choices, and no human validation is needed.

8. Conclusion

In this paper, we introduced Visual Commonsense Reasoning, along with a large dataset VCR for the task that was built using Adversarial Matching. We presented **R2C**, a model for this task, but the challenge – of cognition-level visual undertanding – is far from solved.

Acknowledgements

We thank the Mechanical Turk workers for doing such an outstanding job with dataset creation - this dataset and paper would not exist without them. Thanks also to Michael Schmitz for helping with the dataset split and Jen Dumas for legal advice. This work was supported by the National Science Foundation through a Graduate Research Fellowship (DGE-1256082) and NSF grants (IIS-1524371, 1637479, 165205, 1703166), the DARPA CwC program through ARO (W911NF-15-1-0543), the IARPA DIVA program through D17PC00343, the Sloan Research Foundation through a Sloan Fellowship, the Allen Institute for Artificial Intelligence, the NVIDIA Artificial Intelligence Lab, and gifts by Google and Facebook. The views and conclusions contained herein are those of the authors and should not be interpreted as representing endorsements of IARPA, DOI/IBC, or the U.S. Government.

[12] As we find in Appendix D, including additional language context tends to boost model performance.

[13] This was used to create the SWAG dataset, a multiple choice NLP dataset for natural language inference.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Dont just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018. 8

[2] Alexandre Alahi, Kratarth Goel, Vignesh Ramanathan, Alexandre Robicquet, Li Fei-Fei, and Silvio Savarese. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–971, 2016. 8

[3] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4575–4583, 2016. 8

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018. 6

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 6, 8, 13, 19, 21

[6] Hedi Ben-younes, Remi Cadene, Matthieu Cord, and Nicolas Thome. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 6

[7] Or Biran and Courtenay Cotton. Explanation and justification in machine learning: A survey. In *IJCAI-17 Workshop on Explainable AI (XAI)*, page 8, 2017. 8

[8] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642, 2015. 18, 19

[9] Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. Do explanations make vqa models more predictable to a human? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1036–1042, 2018. 8

[10] Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1657–1668, 2017. 2, 4, 6, 18

[11] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, 2014. 6

[12] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *CVPR*, 2018. 8

[13] Ernest Davis and Gary Marcus. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58:92–103, 2015. 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. Ieee, 2009. 21

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 4, 5, 6, 7, 17, 19, 20

[16] Jacob Devlin, Saurabh Gupta, Ross B. Girshick, Margaret Mitchell, and C. Lawrence Zitnick. Exploring nearest neighbor approaches for image captioning. *CoRR*, abs/1505.04467, 2015. 3

[17] Kiana Ehsani, Hessam Bagherinezhad, Joseph Redmon, Roozbeh Mottaghi, and Ali Farhadi. Who let the dogs out? modeling dog behavior from visual data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 8

[18] Panna Felsen, Pulkit Agrawal, and Jitendra Malik. What will happen next? forecasting player moves in sports videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3342–3351, 2017. 8

[19] Andrew Flowers. The Most Common Unisex Names In America: Is Yours One Of Them?, June 2015. 17, 20

[20] David F. Fouhey, Weicheng Kuo, Alexei A. Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *CVPR*, 2018. 3

[21] Yarin Gal and Zoubin Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027, 2016. 21

[22] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. 2017. 18

[23] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018. 21

[24] Ross Girshick, Ilija Radosavovic, Georgia Gkioxari, Piotr Dollár, and Kaiming He. Detectron. https://github.com/facebookresearch/detectron, 2018. 3, 4, 14, 15, 22

[25] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, volume 1, page 9, 2017. 8

[26] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 19

[27] Agrim Gupta, Justin Johnson, Li Fei-Fei, Silvio Savarese, and Alexandre Alahi. Social gan: Socially acceptable trajectories with generative adversarial networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 8

[28] Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proc. of NAACL*, 2018. 2, 4, 8, 19

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 3, 4, 5, 14, 15, 22

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 15, 21

[31] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016. 8

[32] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. *European Conference on Computer Vision (ECCV)*, 2018. 8

[33] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017. 8

[34] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, Nov. 1997. 5

[35] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. Explainable neural computation via stack neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 53–69, 2018. 8

[36] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4418–4427. IEEE, 2017. 8

[37] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 8

[38] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016. 6

[39] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, Hawaii*, pages 2680–8, 2017. 19

[40] Roy Jonker and Anton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987. 4

[41] Jinkyu Kim, Anna Rohrbach, Trevor Darrell, John Canny, and Zeynep Akata. Textual explanations for self-driving vehicles. In *15th European Conference on Computer Vision*, pages 577–593. Springer, 2018. 8

[42] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard Product for Low-rank Bilinear Pooling. In *The 5th International Conference on Learning Representations*, 2017. 6

[43] K Kim, C Nan, MO Heo, SH Choi, and BT Zhang. Pororoqa: Cartoon video series dataset for story understanding. In *Proceedings of NIPS 2016 Workshop on Large Scale Computer Vision System*, 2016. 19

[44] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 17, 21

[45] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. 8

[46] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 4, 8, 13, 19

[47] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650, 2016. 19

[48] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 21

[49] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 4, 8, 13, 14, 22

[50] Xiao Lin and Devi Parikh. Leveraging visual question answering for image-caption ranking. In *European Conference on Computer Vision*, pages 261–277. Springer, 2016. 6

[51] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron C Courville, and Christopher Joseph Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[52] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 8

[53] Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*, 2016. 19

[54] Roozbeh Mottaghi, Mohammad Rastegari, Abhinav Gupta, and Ali Farhadi. what happens if... learning to predict the effect of forces in images. In *European Conference on Computer Vision*, pages 269–285. Springer, 2016. 8

[55] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957. 4

[56] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 6, 17

[57] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 2227–2237, 2018. 2, 4, 6, 18

[58] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Inferring the why in images. *arXiv preprint arXiv:1406.5472*, 2014. 8

[59] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *Proc. ICCV*, 2017. 8

[60] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 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, 2015. 8

[61] Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis Only Baselines in Natural Language Inference. *arXiv:1805.01042 [cs]*, May 2018. arXiv: 1805.01042. 2, 4

[62] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, 2018. 8

[63] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 5

[64] Nicholas Rhinehart and Kris M Kitani. First-person activity forecasting with online inverse reinforcement learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3696–3705, 2017. 8

[65] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016. 8

[66] Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. Generating descriptions with grounded and co-referenced people. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*, Piscataway, NJ, USA, July 2017. IEEE. 8

[67] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie Description. *International Journal of Computer Vision*, 123(1):94–120, May 2017. 3, 4, 14, 15, 16

[68] Rachel Rudinger, Chandler May, and Benjamin Van Durme. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79, 2017. 19, 22

[69] Maarten Sap, Marcella Cindy Prasettio, Ari Holtzman, Hannah Rashkin, and Yejin Choi. Connotation frames of power and agency in modern films. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2329–2334, 2017. 19, 22

[70] Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013. 21

[71] Alexandra Schofield and Leo Mehr. Gender-distinguishing features in film dialogue. In *Proceedings of the Fifth Workshop on Computational Linguistics for Literature*, pages 32–39, 2016. 19, 22

[72] Roy Schwartz, Maarten Sap, Ioannis Konstas, Li Zilles, Yejin Choi, and Noah A. Smith. The effect of different writing tasks on linguistic style: A case study of the ROC story cloze task. In *Proc. of CoNLL*, 2017. 2, 4, 19

[73] Rico Sennrich, Orhan Firat, Kyunghyun Cho, Alexandra Birch, Barry Haddow, Julian Hitschler, Marcin Junczys-Dowmunt, Samuel Läubli, Antonio Valerio Miceli Barone, Jozef Mokry, and Maria Nadejde. Nematus: a toolkit for neural machine translation. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 65–68, Valencia, Spain, April 2017. Association for Computational Linguistics. 18

[74] Krishna Kumar Singh, Kayvon Fatahalian, and Alexei A Efros. Krishnacam: Using a longitudinal, single-person, egocentric dataset for scene understanding tasks. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016. 8

[75] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 4, 8, 13, 19

[76] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1521–1528. IEEE, 2011. 3, 19

[77] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 8

[78] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Anticipating visual representations from unlabeled video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 98–106, 2016. 8

[79] Misha Wagner, Hector Basevi, Rakshith Shetty, Wenbin Li, Mateusz Malinowski, Mario Fritz, and Ales Leonardis. Answering visual what-if questions: From actions to predicted scene descriptions. In *Visual Learning and Embodied Agents*

in Simulation Environments Workshop at European Conference on Computer Vision, 2018. 8

[80] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony R. Dick. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 2017. 8

[81] John Wieting, Jonathan Mallinson, and Kevin Gimpel. Learning paraphrastic sentence embeddings from back-translated bitext. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 274–285, 2017. 18

[82] Adina Williams, Nikita Nangia, and Samuel Bowman. 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, 2018. 18

[83] Qi Wu, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4622–4630, 2016. 8

[84] Tian Ye, Xiaolong Wang, James Davidson, and Abhinav Gupta. Interpretable intuitive physics model. In *European Conference on Computer Vision*, pages 89–105. Springer, 2018. 8

[85] Yuya Yoshikawa, Jiaqing Lin, and Akikazu Takeuchi. Stair actions: A video dataset of everyday home actions. *arXiv preprint arXiv:1804.04326*, 2018. 8

[86] Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. Visual Madlibs: Fill in the blank Image Generation and Question Answering. *arXiv:1506.00278 [cs]*, May 2015. arXiv: 1506.00278. 8

[87] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 8

[88] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speakerlistener-reinforcer model for referring expressions. In *Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017. 8

[89] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 8

[90] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017. 19

[91] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 8

[92] Yipin Zhou and Tamara L Berg. Temporal perception and prediction in ego-centric video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4498–4506, 2015. 8

[93] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7W: Grounded Question Answering in Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 8, 13

[94] Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *arXiv preprint arXiv:1506.06724*, 2015. 20