Be Different to Be Better!
A Benchmark to Leverage the Complementarity of Language and Vision

Sandro Pezzelle[1], **Claudio Greco**[2], **Greta Gandolfi**[2], **Eleonora Gualdoni**[2], **Raffaella Bernardi**[2,3]
[1]Institute for Logic, Language and Computation, University of Amsterdam
[2]CIMeC, [3]DISI, University of Trento
s.pezzelle@uva.nl,
{greta.gandolfi|eleonora.gualdoni}@studenti.unitn.it,
{claudio.greco|raffaella.bernardi}@unitn.it

Abstract

This paper introduces BD2BB, a novel language and vision benchmark that requires multimodal models combine *complementary* information from the two modalities. Recently, impressive progress has been made to develop universal multimodal encoders suitable for virtually any language and vision tasks. However, current approaches often require them to combine *redundant* information provided by language and vision. Inspired by real-life communicative contexts, we propose a novel task where either modality is necessary but not sufficient to make a correct prediction. To do so, we first build a dataset of images and corresponding sentences provided by human participants. Second, we evaluate state-of-the-art models and compare their performance against human speakers. We show that, while the task is relatively easy for humans, best-performing models struggle to achieve similar results.

1 Introduction

Human communication, in real-life situations, is multimodal (Kress, 2010): To convey and understand a message uttered in natural language, people build on what is present in the multimodal context surrounding them. As such, speakers do not need to "repeat" something that is already provided by the environment; similarly, listeners leverage information from various modalities, such as vision, to interpret the linguistic message. Integrating information from multiple modalities is indeed crucial for attention and perception (Partan and Marler, 1999) since combined information from concurrent modalities can give rise to different messages (McGurk and MacDonald, 1976).

The argument that language and vision convey different, possibly complementary aspects of meaning has been largely made to motivate the need for multimodal semantic representations of words (Baroni, 2016; Beinborn et al., 2018). However, computational approaches to language and vision typically do not fully explore this complementarity. To illustrate, given an image (e.g., the one depicted in Figure 1), popular tasks involve describing it in natural language, e.g., *"A tennis player about to hit the ball"* (Image Captioning; see Bernardi et al., 2016); answering questions that are grounded in it, e.g., Q: *"What sport is he playing?"*, A: *"Tennis"* (Visual Question Answering; see Antol et al., 2015); having a dialogue on its entities, e.g., Q: *"Is the person holding a racket?"*, A: *"Yes."* (visually-grounded dialogue; see De Vries et al., 2017; Das et al., 2017). While all these tasks challenge models to perform visual grounding, i.e., an effective *alignment* of language and vision, none of them require a genuine *combination* of complementary information provided by the two modalities. All the information is fully available in the visual scene, and language is used to describe or retrieve it.

In this work, we propose a novel benchmark, *Be Different to Be Better* (in short, **BD2BB**), where the *different*, complementary information provided by the two modalities should push models develop a *better*, richer multimodal representation. As illustrated in Figure 1, models are asked to choose, among a set of **candidate actions**, the one a person who sees the visual context depicted by the **image** would do based on a certain **intention** (i.e., their goal, attitude or feeling). Crucially, the resulting multimodal input (the sum of the image and the intention) will be richer compared to that conveyed by either modality in isolation; in fact, the two modalities convey complementary or *non-redundant* information (Partan and Marler, 1999).

To illustrate, a model that only relies on the (non-grounded) linguistic information conveyed by the intention, i.e., *"If I have tons of energy"*, might consider as equally plausible any actions that have to do with playing a sport, e.g., *"I will play base-*

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2751–2767

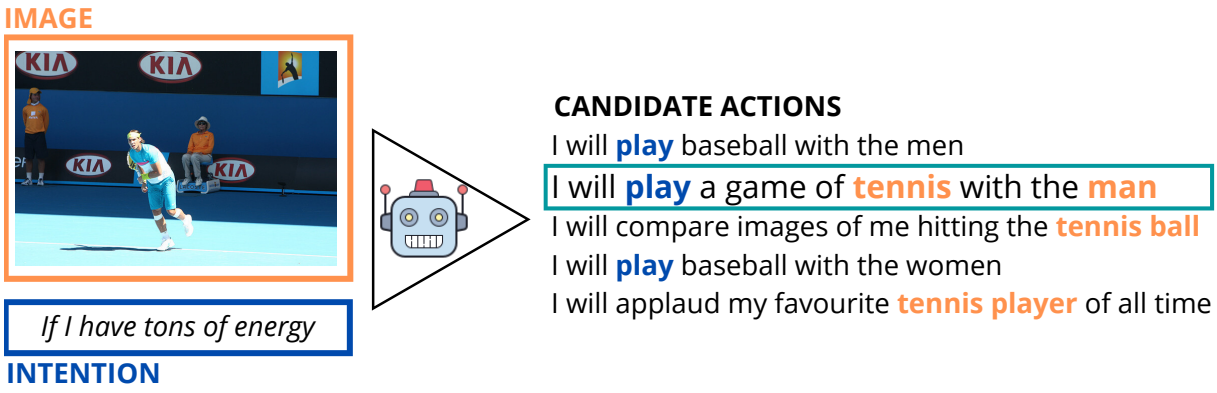

Figure 1: One real sample of our proposed task. Given an **image** depicting, e.g., a tennis player during a match and the **intention** *"If I have tons of energy"*, the task involves choosing, from a list of 5 **candidate actions**, the target action that unequivocally applies to the combined multimodal input: *"I will play a game of tennis with the man"*. The task is challenging: a model exploiting a language or vision bias could fall into the trap of decoy actions containing words highlighted in blue or orange, respectively. Therefore, selecting the *target* action requires models perform a genuine integration of the two modalities, whose information is complementary. Best viewed in color.

ball with the men" or *"I will play a game of tennis with the man"*. Similarly, a model that only relies on the visual information conveyed by the image—a tennis player during a match—might consider as equally plausible any actions that have to do with 'tennis' and/or 'player', e.g., *"I will applaud my favourite tennis player of all time"* or *"I will play a game of tennis with the man"*. In contrast, a model that genuinely combines information conveyed by both modalities should be able to select the *target* action, namely the only one that is both consistent with the intention and grounded in the image, i.e., *"I will play a game of tennis with the man"*. Moreover, similarly to real-life communicative scenarios, in our approach different language inputs *modulate* differently the same visual context, and this gives rise to various multimodal messages. To illustrate, if the image in Figure 1 is paired with the intention *"If I am tired watching"*, the target action *"I will play a game of tennis with the man"* is no longer valid. Indeed, the target action in this context is *"I will leave the tennis court"* (see Figure 3).

Our work has the following key contributions:

- We introduce a novel multimodal benchmark: the set of $\sim$ 10K $\langle image, intention, action \rangle$ datapoints collected via crowdsourcing and enriched with meta-annotation; the multiple choice task, **BD2BB**, which requires proper integration of language and vision and is specifically aimed at testing SoA pretrained multimodal models. The benchmark, together with the code and trained models, is available at: https://sites.google.com/view/bd2bb

- We test various models (including the SoA multimodal, transformer-based LXMERT; Tan and Bansal, 2019) and show that, while **BD2BB** is a relatively easy task for humans ($\sim$ 80% acc.), best systems struggle to achieve a similar performance ($\sim$ 60% acc.).

- We extensively analyze the results and show the advantage of exploiting multimodal pretrained representations. This confirms they are effective, but not enough to solve the task.

2 Related Work

Since the introduction of the earliest multimodal tasks, such as Image Captioning (IC; see Bernardi et al., 2016) and Visual Question Answering (VQA; Antol et al., 2015), a plethora of tasks dealing with language and vision have been proposed. In parallel, baseline models have been replaced by more powerful attention-based systems (Anderson et al., 2018) and, more recently, by transformer-based architectures pretrained on several tasks (Tan and Bansal, 2019; Lu et al., 2019; Chen et al., 2019). These latter models build on multimodal representations that are meant to be task-agnostic; as such, they can be transferred to virtually any other multimodal task with minimal fine-tuning. Our work contribute to these two lines of research by (1) introducing a novel multimodal task, and (2) by evaluating a SoA multimodal encoder on it.

Multimodal tasks VQA was originally proposed to overcome the challenge of quantitatively evaluate IC models. The task (and its evaluation)

is straightforward: given an image and a question about its visible *objects*, systems have to provide the correct answer by aligning information from the two modalities (Antol et al., 2015). Driven by VQA, several datasets have been proposed to minimize the bias observed in natural images (Goyal et al., 2017; Ray et al., 2019); to force models to "reason" over a joint visual and linguistic input (Johnson et al., 2017; Suhr et al., 2019); to deal with objects' attributes and relations (Krishna et al., 2017); to encompass more diverse (Zhu et al., 2016) and goal-oriented questions and answers (Gurari et al., 2018). At the same time, some work proposed higher-level evaluations of VQA models and showed their limitations (Hodosh and Hockenmaier, 2016; Shekhar et al., 2017); similarly, recent attention has been paid to understand what makes a question "difficult" for a model (Bhattacharya et al., 2019; Terao et al., 2020). Despite impressive progress, current approaches to VQA do not tackle one crucial limitation of the task: the answer to a question is given by the *alignment* of language and vision rather than their *complementary* integration.

Moving from objects to *actions*, several tasks have been proposed to mimic more realistic settings where a higher degree of integration between modalities is required. One is visual storytelling (Huang et al., 2016; Gonzalez-Rico and Pineda, 2018; Lukin et al., 2018), where models have to understand the action depicted in each photo and their relations to generate a story. Similar abilities are required in the task of generating non-grounded, human-like questions about an image (Mostafazadeh et al., 2016; Jain et al., 2017), and in that of asking discriminative questions over pairs of similar scenes (Li et al., 2017). Related tasks are also those of predicting motivations of visually-grounded actions (Vondrick et al., 2016) or generating explanations for a given answer (Park et al., 2018; Hendricks et al., 2018).

An even higher level of understanding of vision and language is required in the tasks of filling the blank with the correct answer (Yu et al., 2015); answering questions from videos and subtitles (Lei et al., 2018); having a dialogue on objects (De Vries et al., 2017; Das et al., 2017) or events (Mostafazadeh et al., 2017); answering and justifying commonsense questions (Zellers et al., 2019). However, all these tasks require making *commonsense* inferences over the two modalities rather than integrating their complementary information to answer a *grounded* question.

More akin to ours are the approaches by Iyyer et al. (2017), which aims to predict the subsequent scene and dialogue in a comic strip, and Kruk et al. (2019), where the goal is to compute the communicative intent of a social media post. Though they both require a challenging integration of language and vision, these tasks (as well as the type of data they use) are crucially different from **BD2BB**, where the task is to predict the action that is consequent to a given intention based on the image.

Transformer-based multimodal models Developing universal multimodal encoders whose pretrained representations are suitable for virtually any multimodal task is a crucial challenge. Inspired by the success of BERT, a pretrained transformer-based language encoder (Devlin et al., 2019), similar architectures have been recently proposed in the domain of language and vision (Lu et al., 2019; Tan and Bansal, 2019; Chen et al., 2019; Su et al., 2020; amd Nan Duan et al., 2020). While these architectures achieve state-of-the-art performance in many tasks, their novelty and complexity leave several questions open, and further work is needed to better understand, e.g., which layers are more suitable for transferability (Tamkin et al., 2020), or what is the relation between pretraining and downstream tasks (Zamir et al., 2018; Singh et al., 2020). Moreover, to prove they are readily applicable to novel multimodal benchmarks, pretrained universal encoders should be ideally effective with only minimal fine-tuning on the target tasks.

In this light, we believe that more efforts should be put in developing datasets that are challenging and yet relatively small, in line with the 'diagnostic' datasets proposed for VQA (Johnson et al., 2017) and the easy vs. hard subsets introduced by Akula et al. (2020) for visual referring expression recognition. Our contribution follows this line of thought.

3 Data

In this section, we describe how we collected intentions and actions through crowdsourcing, and the subsequent phase of data meta-annotation. Consistently with our purposes, we needed images that elicit goals and feelings (the intentions) in the annotators, as well as consequent actions. To this end, we used the partition of the MS-COCO dataset (Lin et al., 2014) provided by Vondrick et al. (2016),[1]

[1] http://visiond1.cs.umbc.edu/webpage/codedata/intention/motivations_clean.zip

Figure 2: Data collection. Examples of good (top) and bad (bottom) annotations provided to participants in the task instructions. Errors and corresponding warnings are shown to make participants familiarize with the tool.

where each of the $10,191$ images depicts at least one person. This choice was aimed to make the participants' task more natural: indeed, the presence of people in the image allows more possibilities of interaction, and therefore guarantees that some actions can be performed in that situation.

3.1 Data Collection

We set up an annotation tool on Figure-Eight[2] (see Figure 2) where annotators were shown an image and asked to imagine themselves being in that situation, as ideal observers not represented in the picture. We instructed them to carefully look at the image and think about 1) an intention, i.e., *how they might feel/behave if they were in that situation*; 2) an action, i.e., *what they would do based on that feeling/behavior*. Intentions and actions were typed in free form by participants in two separate text boxes; by instructions, their sentences had to complete the provided opening words *If I...* and *I will...*, respectively. To ensure that intentions conveyed information that was complementary (non-redundant) to that by the image, participants were instructed not to mention *any* of the entities (people, objects, etc.) shown in the image. In contrast, to ensure that actions contained information that was grounded in the image, participants were asked to mention at least one visible entity when writing their action (see errors and warnings in Figure 2).[3]

We randomly selected $\sim$ 3.6K images from the split by Vondrick et al. (2016) and, for each of them, we collected on average 5 $\langle intention, action \rangle$ tuples by 5 participants. In total, $\sim$ 18K unique $\langle image, intention, action \rangle$ datapoints were collected. Participants were recruited from native-English countries only. Overall, 477 annotators (based on the IP) took part in the data collection; on average, each of them provided 38 annotations. Participants were paid 0.04$ per tuple.[4] In total, the data collection costed $\sim$ 900$.

INTENTIONS	ACTIONS
1. *If I* want to be on a spotlight	1. *I will* stay behind the player
2. *If I* want to give encouragement	2. *I will* applaud the player
3. *If I* want to make my dream come true	3. *I will* have to win the tennis match
4. *If I* have tons of energy	4. *I will* play a game of tennis with the man
5. *If I* get tired of watching	5. *I will* leave the tennis court

Figure 3: Five $\langle intention, action \rangle$ tuples provided by 5 unique participants for the image in Figure 1.

A few filtering steps were needed to get rid of datapoints with invalid annotations. First, we discarded those datapoints where intentions and/or actions were either not in English (e.g., bot-generated *Lorem Ipsum* sequences) or nonsense strings (e.g., random sequences of characters). This step was done semi-manually and filtered out $\sim$ 3K datapoints. Second, we removed datapoints where the action did not contain any noun nor pronoun. After this, we were left with $12,457$ valid datapoints.

To illustrate the type of data collected, Figure 3 reports the 5 $\langle intention, action \rangle$ tuples provided by 5 annotators for the image in Figure 1. As can be noted, the same visual context elicits different intentions, which in turn give rise to different possible actions. Crucially, no intentions refer to

[2] https://www.figure-eight.com/

[3] Further details on data collection and meta-annotation, dataset and models are given in Appendix A.

[4] This corresponds to a hourly wage of around 8$/hour.

anything that is visible in the image, which makes them suitable for virtually any visual context. As for the actions, in contrast, they all 1) mention at least one entity that is *grounded* in the given scene, e.g., "player" or "tennis court", which makes them plausible only for sports contexts, particularly 'tennis'; 2) match their corresponding intention, but not (or to a much lesser extent) the others; i.e., different intentions trigger different actions, and the verb in the action is a proxy for such diversity. Below, we describe the meta-annotation process we performed to categorize each datapoint with respect to: 1) the topic of its action, e.g., 'tennis'; and 2) the argument structure of the verbs in its action.

3.2 Meta-Annotation

Topic For each of the $12,457$ datapoints, we built a 512-d semantic representation of its action using the off-the-shelf Universal Sentence Encoder (USE; Cer et al., 2018). We then run a k-means clustering algorithm over these vectors and obtained 60 *topic* clusters.[5] By manual inspection, 54 clusters were found to consistently group together actions revolving on the same topic, e.g., 'tennis' or 'birthday', in a way that it was easy to label them using such terms. Since for the remaining 6 clusters this was not straightforward due to the presence of rather disconnected actions, we filtered these clusters out. We further polished the 54 clusters (a) by manually moving actions to clusters that fit them better, and (b) by removing actions that were not in line with the cluster topic. Moreover, we removed actions that did not comply with the instructions provided to annotators during the data collection. After these steps, we were left with $10,287$ $\langle image, intention, action \rangle$ datapoints.

Argument structure Using the Stanford NLP Parser (Chen and Manning, 2014), we annotated the actions in each of the $10,287$ topic-categorized datapoints by means of a 4-code annotation schema. In particular, from each parsed action we extracted its main verb (code1) and its direct or indirect object (code2). Moreover, when present, the verb of the coordinate or subordinated sentence was also extracted (code3), as well as other nouns in any complement position of the main or secondary verb (code4).[6] All the outputs by the parser were manually checked and fixed where needed. Given the action *"I will swing the racket to hit the ball"*, for example, we thus obtained the following argument structure annotation: $\langle swing \rangle$ (code1), $\langle racket \rangle$ (code2), $\langle hit \rangle$ (code3), $\langle ball \rangle$ (code4). As can be seen, this simplified representation of the action provides information on both its verbs (that are *consequent* to the intention) and nouns (*grounded* in the image). The $10,287$ annotated datapoints were used to build the dataset for our task.

[5]The best number of clusters was chosen based on the Elbow method, which relies on cluster consistency.

[6]While verbs were lemmatized, we did not do so for nouns due to the visual difference between, e.g., *player* and *players*.

4 Task

We introduce the *Be Different to Be Better* (**BD2BB**) task, where the *different*, i.e., complementary information provided by the two modalities should push models develop a *better*, i.e., richer multimodal representation. To evaluate these abilities, we frame our task as a multiple-choice problem (similar to Antol et al., 2015; Yu et al., 2015; Zhu et al., 2016) where either modality is necessary but not sufficient to perform a correct prediction. The task is the following (see Figure 1): given an image and a corresponding intention, the model has to choose the correct action over a set of 5 candidate actions. We refer to the correct action as the *target* action; to the wrong actions as the *decoy* actions. Similarly to Chao et al. (2018), decoy actions are carefully selected to be as plausible as possible when evaluated against either the intention (2 decoys) or the image (the other 2) only. Below, we explain how language-based and image-based decoys were selected based on the meta-annotation.

Language-based decoys For each of the $10,287$ $\langle image, intention, action \rangle$ datapoints, we randomly selected a number of datapoints from the entire data that had the following criteria: 1) their action belonged to a different topic cluster than the one including the target action; 2) their action did not share any noun with the target action, i.e., their $\langle code2 \rangle$ and $\langle code4 \rangle$ were different. We then computed a similarity score between the target action and each of these selected actions by means of the cosine of their USE representations. We ranked these scores and selected as our language decoys the two with the highest similarity. This way, we obtained language-based decoys that are semantically very similar to the target action, but are on a different topic and do not share any noun with it.

Vision-based decoys For each datapoint, we randomly selected a number of datapoints from the

I:	*If I want to protect myself, I will...*	*If I want to enjoy the sun, I will...*	*If I want to get the blood pumping, I will...*	*If I want to be noticed, I will...*
L:	sit on my skateboard instead of actually riding it	take a huge bite out of my sandwich	take a ride on the aerial tramway	put on a costume and join the parade
L:	wear jeans when racing on a skateboard	take a bite of the burger	ride a horse in the rodeo	join the men on the street
T:	**wear a helmet while riding my motor bike**	**eat my food on the roof patio**	**ride a motorcycle**	**wear a sign**
V:	look at the motorcycle display	use my phone to order from a take out menu	seat next to a bike and read a book	at least match my colors to look fancy
V:	challenge the people to a race	assist the group with cutting food	help the person who has fallen off their bike	teach him how to tie a tie

Figure 4: Four samples from our dataset. **I**: Intention; **T**: Target action; **L/V**: Language-/Vision-based decoys.

entire data that had the following criteria: 1) their action belonged to the same topic cluster of the target one; 2) their action did not share any verb with the target action, i.e., their $\langle code1 \rangle$ and $\langle code3 \rangle$ were different. We then ranked these actions with respect to their USE similarity with the target one, and selected as our vision-based decoys the two with the lowest score. This way, we obtained vision-based decoys that are about the same topic of the target action; at the same time, they do not share any verbs with it and are semantically different.

4.1 Dataset

Our final dataset includes $10,265$ samples[7] as the ones depicted in Figure 4: each sample consists of a unique $\langle image, intention, action \rangle$ datapoint paired with 4 carefully-selected decoy actions. Consistently with out purpose of making **BD2BB** a challenging benchmark for pretrained multimodal architectures (see Section 1), we split the dataset into "unusual" train/val/test partitions; i.e., we selected 20% samples for training; the remaining for validation (40%) and test (40%). We propose having small training data and larger validation and test sets should become a standard, as pretrained models already build on a massive amount of data.

Table 1 reports the descriptive statistics of the dataset, including the number of unique images, intentions and actions per split, and the average length of the sentences. All the experiments reported in the paper are performed on these splits.

5 Experiments

To test the importance of combining information from the two modalities and the independent contribution of either modality, we experiment with 3 settings of the **BD2BB** task: *L*, where the target action among the 5 candidates has to be guessed based on the intention only; *V*, where only the image is provided; *LV*, where both the image and the intention are provided. For each setting of the task, we evaluate the performance of (1) a simple baseline trained from scratch on the task; (2) a state-of-art transformer-based pretrained model fine-tuned on the task; (3) the same transformer-based model trained from scratch on the task. Moreover, results by models are compared to (4) human performance.

5.1 Models

Baseline For each $\langle image, intention, action \rangle$ datapoint in the sample, baseline$_{LV}$ builds a multimodal representation by concatenating the 2048-d visual features of the image (extracted from a pretrained ResNet-101; He et al., 2016) with the 300-d embedding of the intention and the 300-d embedding of the action. Embeddings for both the intention and the action are obtained by summing the GloVe embeddings (Pennington et al., 2014) of the words in them. The concatenated features are linearly projected into a vector (8192-d), passed through ReLU, and linearly projected into a single value. Softmax probabilities are computed over the 5 sample's candidate values. The baseline$_{L}$ only concatenates intention and action embeddings (600-d representation); baseline$_{V}$ concatenates the visual features with the action embedding (2348-d). Finally, to account for any bias due to unavoidable association and repetition patterns among the actions, we test a version of the baseline which only encodes the actions. We refer to it as *actions-only*.

RoBERTa In setting *L*, we employ the robustly optimized version of BERT, RoBERTa (Liu et al., 2019); this model is a universal language encoder pretrained on the task of masked language modeling, which achieves best-performing performance in the challenging multiple-choice

[7] For 22 datapoints it was not possible to find all the decoys, hence they were discarded during the creation of the dataset.

	#samples (%)	#img	#int	#act	#t-act	#d-act	avg int len	avg act len
train	2102 (20%)	1517	1683	5063	2102	4228	22.15	35.34
val	4082 (40%)	2447	2772	6082	3567	4133	20.76	36.20
test	4081 (40%)	2425	2720	6108	3561	4138	20.49	36.00
total	10265 (100%)	3215	6192	8751	8738	6339	20.94	35.94

Table 1: Descriptive statistics of the dataset including, from left to right: 1) # (and %) of unique samples; 2) # of unique images; 3) # of unique intentions; 4) # of unique actions; 5) # of unique target actions; 6) # of unique decoy actions; 7) average number of tokens in intentions; 8) average number of tokens in actions.

SWAG task (Zellers et al., 2018). We adapt RoBERTa$_{BASE}$ to our task as following: for each of the 5 $\langle image, intention, action\rangle$ datapoints in the sample, RoBERTa encodes the input as a sequence composed by $\langle CLS\rangle$, the intention, $\langle SEP\rangle$, the action, and $\langle EOS\rangle$. The encoding corresponding to the $\langle CLS\rangle$ token (768-d) is passed through Tanh, linearly projected into a vector (768-d), passed to Dropout (Srivastava et al., 2014), and linearly projected into a single value. Softmax probabilities are computed over the 5 sample's candidate values. As mentioned above, we evaluate two model versions: RoBERTa$_L$, pretrained and fine-tuned on our task, and RoBERTa^s_L, trained from scratch on BD2BB.

LXMERT In settings *LV* and *V*, we employ LXMERT (Learning Cross-Modality Encoder Representations from Transformers; Tan and Bansal, 2019), a universal multimodal encoder pretrained on five language and vision tasks which is state-of-art on VQA2.0 (Goyal et al., 2017). This model represents an image by the set of position-aware object embeddings for the 36 most salient regions detected by Faster R-CNN (Ren et al., 2015) and processes the textual input by position-aware randomly-initialized word embeddings. Like RoBERTa, LXMERT uses the special tokens $\langle CLS\rangle$ and $\langle SEP\rangle$ but, differently from RoBERTa, here $\langle SEP\rangle$ is used both to separate sequences and to denote the end of the textual input. Hence, we take this into account when adapting LXMERT to our task. Similar to RoBERTa, we use the encoding corresponding to $\langle CLS\rangle$ (768-d) to obtain a probability distribution over the 5 sample's candidate values. For each task setting, we evaluate each model in two versions, i.e., pretrained model fine-tuned on our task (LXMERT$_{LV}$ and LXMERT$_V$); trained from scratch (LXMERT$^s_{LV}$ and LXMERT^s_V).

Experimental setup For baseline models, we perform hyperparameter search on learning rate, Dropout, and hidden size; as for transformer-based models, we use the best configurations reported in the source papers (reproducibility details in Appendix B). All models are trained with 3 random seeds for 50 epochs with Adam (Kingma and Ba, 2015) minimizing a Cross Entropy Loss between the probability distribution over the 5 sample's candidate actions and the ground-truth action. For each of the 3 runs, we consider the model with the highest validation accuracy. Average accuracy and standard deviation over 3 runs is computed.

5.2 Human Evaluation

We randomly extracted 300 unique samples from the dataset and split them into 3 partitions including 100 samples each. For each partition, we collected judgments by 3 participants in each setting of the task: *L*, *V*, and *LV*. Crucially, participants did the task only once per partition; i.e., they judged each sample only in one of the 3 task settings. Using Quiz Maker,[8] we collected 2,700 unique responses from 11 subjects who participated on a voluntary basis. For each setting of the task, we counted as 'correctly predicted' the samples where at least 2 out of 3 annotators converged on the target action. Moreover, for each task setting we computed the 'best' accuracy, i.e., the average of the 3 participants who achieved the highest accuracy in each split.

6 Results

Results by both models and humans are reported in Table 2. Several key observations can be made.

Multimodal integration is the key. The overall best-performing model in **BD2BB** is LXMERT$_{LV}$ (62.2%), which outperforms the other pretrained models, i.e., RoBERTa$_L$ (56.2%) and LXMERT$_V$ (59.2%). On the one hand, this shows that disposing of both modalities is beneficial to perform the

[8] https://www.quiz-maker.com

	model	accuracy	
		val ± std	test ± std
SCRATCH	actions-only	44.0 ± 0.4	44.6 ± 0.8
	baseline$_L$	45.3 ± 0.9	45.9 ± 0.9
	baseline$_V$	45.8 ± 0.8	46.1 ± 0.8
	baseline$_{LV}$	48.6 ± 0.9	49.0 ± 0.9
SCRATCH	RoBERTa^s_L	47.0 ± 0.2	47.2 ± 0.1
	LXMERT^s_V	30.9 ± 0.9	31.8 ± 0.4
	LXMERT$^s_{LV}$	50.4 ± 0.3	51.3 ± 0.4
PRETRAIN	RoBERTa$_L$	55.9 ± 0.9	56.2 ± 1.3
	LXMERT$_V$	59.1 ± 0.2	59.2 ± 0.6
	LXMERT$_{LV}$	**62.8** ± 2.3	**62.2** ± 2.2
	humans$_L$	50.0 (best 54.0)	
	humans$_V$	72.3 (best 73.7)	
	humans$_{LV}$	79.0 (best 82.3)	
	chance	20.0	20.0

Table 2: Results for the 3 settings: *L*, *V*, and *LV*. ^{s} refers to transformer-based models trained from scratch. For each model, we report average accuracy and std over 3 runs. Human accuracy is computed over 300 samples (we report values based on both majority vote, i.e., 2 out of 3, and average of best participants; see 5.2).

task. This is in line with the results by human participants, who achieve the highest accuracy in the multimodal setting (79% vs. 50% of *L* and 72.3% of *V*). On the other hand, the finding that LXMERT$_V$ surpasses RoBERTa$_L$ (+3%) confirms that the image provides more information compared to the intention. This, again, is consistent with human results, where the gap between *V* and *LV* (−7%) is much smaller compared to that between *L* and *LV* (−29%). For humans, this visual advantage is likely due to (MS-COCO) images depicting complex events that elicit a broad range of aspects related to people's experience of the world. As for the models, it confirms that LXMERT, thanks to its massive pretraining, is effective in extracting fine-grained information from images.

Models are far from humans. Humans achieve around 80% accuracy ('best' 82%) on the multimodal version of the task. This is a high result, in line with previous work with a similar setup (consider, e.g., SWAG, where 'expert' human accuracy is around 85% with 4 choices, i.e., chance level at 25%; Zellers et al., 2018). At the same time, the non-perfect human accuracy reveals that the benchmark is challenging due to the careful selection of plausible decoys. Compared to humans, the best-performing LXMERT$_{LV}$ achieves much lower results (−17%), which indicates that **BD2BB** is challenging and far from being solved. Since the gap between best-performing models and human participants in unimodal settings is smaller (−13% in *V* and −6% in *L*), the biggest computational challenge lies in the integration of complementary information from different modalities.

Pretrained is better. Pretrained models neatly outperform the baseline in all the versions of the task[9] and, more interestingly, also all their counterparts trained from scratch. As can be seen in Table 2, indeed, transformer-based models trained from scratch achieve results that are only slightly better than those by the baseline in both *LV* and *L*; as for *V*, LXMERT^s_V turns out to perform worse than the baseline^s_V (and even worse than the actions-only baseline). This clearly shows that these architectures are very effective when building on their pretraining, but suffer when challenged to learn a task from scratch with relatively few samples.

7 Analysis

Best models' errors We perform an analysis on the errors made by the 3 pretrained models to check whether they fall more often into the language-based or vision-based decoys. To do so, we focus on each model's best run, and compute the proportion of wrong predictions in the test set that belong to one or the other decoy type. For comparison, a model that makes modality-balanced wrong predictions should fall into language-/vision-based decoys 50% of the times. Quite surprisingly, RoBERTa$_L$ has only a moderate bias toward language-based decoys: in fact, only 60.2% of its errors are of this type. As for LXMERT$_V$, no bias at all is observed toward the vision-based decoys (48.6%). Finally, the best-performing LXMERT$_{LV}$ is shown to be halfway between these models, with only a slight preference for language-based (55.1%) over vision-based decoys (44.9%).

In Figure 5, we report two cherrypicked examples where LXMERT$_{LV}$ either correctly predicts the target action (left) or choses a wrong one, in this case a vision-based decoy (right). It is worth mentioning that these two cases are challenging: for

[9] It should be noted that the baselines are only slightly better than *actions-only*; this suggests that these models are only marginally capable of extracting (and combining) relevant information for the task from the image and the intention.

I: *If I am in the mood to act silly, I will…* | *If I don't like this, I will…*

L: attend a dinner like this man holding a gift | sit next to the woman on the bench
L: buy him a cake and invite his friends to party | get my face painted
T: **act silly with this man and eat cake** | **avert my eyes from the man who looks silly**
V: help my child cut their cake | teach him how to tie a tie
V: have cake with soldiers | **wear a costume and march in a parade**

Figure 5: Two samples where humans give the correct answer in the *LV* setting—but neither in *L* nor in *V*. LXMERT$_{LV}$ picks the correct answer (**blue**) in the left sample, a wrong one (**red**) in the right sample. **I**: Intention; **T**: Target action; **L/V**: Language-/Vision-based decoys. Best viewed in color.

both of them, human annotators were able to pick the correct action only in the multimodal version of the task—but neither in *L* nor in *V*. As can be seen, in the leftmost example the model does a good job in combining complementary information from language and vision. In the rigthmost one, instead, it picks an action that is very much plausible based on the image, but not in presence of the given intention containing a negation (*don't*). Taken together, these analyses indicate that no simple strategies can be exploited by models to detect and rule out decoy types. Language- and vision-based decoys are equally challenging, and combining complementary information is needed to solve the task.

Hard test To explore the robustness of the pretrained models, we check how well they perform on a subset of the test set where several features of the samples were *unseen* in training. In particular, neither the image nor the intention were seen in training; moreover, the target action could be seen as a decoy but never as the target. In Table 3 we report the results by the 3 pretrained models on this subset ($1,505$ samples); we refer to it as the *hard* test. As can be seen, all models experience a small decrease in accuracy compared to the whole test set—while humans do not. This indicates that the hard test is indeed more challenging. However, pretrained models are overall robust to unseen features. In line with the standard test set, LXMERT$_{LV}$ still outperforms the unimodal models, though its drop in performance (-4%) is more pronounced compared to them ($-1/2\%$). This suggests that part of the advantage of the multimodal system over the unimodal ones is due to its fine-tuning. Indeed, pretraining on its own is not enough to properly combine complementary information from the intention and the image. Finally, since humans do not perform worse in these samples, the performance gap with LXMERT$_{LV}$ increases to $\sim 20\%$.

model	accuracy	humans
	hard test $\pm$ std	
RoBERTa$_L$	55.1 ± 1.6	56.5
LXMERT$_V$	56.9 ± 0.8	73.9
LXMERT$_{LV}$	**58.3** ± 2.7	78.3

Table 3: Accuracy of the pretrained transformer-based models on the *hard* samples of the test set. Human accuracy is computed over 92 samples.

8 Conclusion

Inspired by real-life communicative contexts where language and vision are *non-redundant*, we proposed a novel benchmark to challenge models combine complementary multimodal information. This is a crucial ability that, we believe, our benchmark will contribute push further. In particular, recently proposed universal multimodal encoders can greatly benefit from relatively small but challenging resources as is **BD2BB**, which can be used to shed light on model abilities and help developing architectures which exhibit more human-like skills.

Here, we evaluated LXMERT and showed that it struggles to achieve results that are comparable to those by humans. In the future, we plan to evaluate other multimodal encoders on it, and to contribute to the development of better multimodal systems.

Acknowledgments

The authors kindly acknowledge SAP for sponsoring the work. We are grateful to Moin Nabi and Tassilo Klein (SAP AI Research) for the valuable discussion in the early stages of the project. We thank all the participants who voluntarily took part in the human evaluation, and the attendees of the SiVL workshop co-located with ECCV 2018 for their feedback on a preliminary version of the task and data collection pipeline. We kindly acknowledge the support of NVIDIA Corporation with the donation of the GPUs used in our research. The first author is funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455 awarded to Raquel Fernández).

References

Arjun R. Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020. Words aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *The 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

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

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.

Marco Baroni. 2016. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13.

Lisa Beinborn, Teresa Botschen, and Iryna Gurevych. 2018. Multimodal grounding for language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Nilavra Bhattacharya, Qing Li, and Danna Gurari. 2019. Why does a visual question have different answers? In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4271–4280.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174.

Wei-Lun Chao, Hexiang Hu, and Fei Sha. 2018. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. 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 431–441.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: learning universal image-text representations. *CoRR*, abs/1909.11740.

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

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Diana Gonzalez-Rico and Gibran Fuentes Pineda. 2018. Contextualize, show and tell: A neural visual storyteller. *CoRR*, abs/1806.00738.

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.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Generating counterfactual explanations with natural language. In *ICML Workshop on Human Interpretability in Machine Learning*, pages 95–98.

Micah Hodosh and Julia Hockenmaier. 2016. Focused evaluation for image description with binary forced-choice tasks. In *Proceedings of the 5th Workshop on Vision and Language*, pages 19–28.

Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet

Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239.

Mohit Iyyer, Varun Manjunatha, Anupam Guha, Yogarshi Vyas, Jordan L Boyd-Graber, Hal Daumé III, and Larry S Davis. 2017. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6478–6487.

Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. 2016. Revisiting visual question answering baselines. In *European Conference on Computer Vision*, pages 727–739. Springer.

Unnat Jain, Ziyu Zhang, and Alexander G Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6485–6494.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Gunther R Kress. 2010. *Multimodality: A social semiotic approach to contemporary communication*. Taylor & Francis.

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

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in Instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4614–4624.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379.

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

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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

Stephanie Lukin, Reginald Hobbs, and Clare Voss. 2018. A pipeline for creative visual storytelling. In *Proceedings of the First Workshop on Storytelling*, pages 20–32.

Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature*, 264(5588):746–748.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472.

Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1802–1813.

Gen Li amd Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. 2020. Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of AAAI*.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *31st IEEE Conference on Computer Vision and Pattern Recognition*.

Sarah Partan and Peter Marler. 1999. Communication goes multimodal. *Science*, 283(5406):1272–1273.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. 2019. Sunny and dark outside?! Improving answer consistency in VQA through entailed question generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5860–5865, Hong Kong, China. Association for Computational Linguistics.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto, and Raffaella Bernardi. 2017. FOIL it! Find One mismatch between Image and Language caption. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 255–265.

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are we pretraining it right? Digging deeper into visio-linguistic pretraining. *CoRR*, abs/2004.08744.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*.

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.

Alex Tamkin, Trisha Singh, Davide Giovanardi, and Noah Goodman. 2020. Investigating transferability in pretrained language models. *CoRR*, abs/2004.14975.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114.

Kento Terao, Toru Tamaki, Bisser Raytchev, Kazufumi Kaneda, and Shun'ichi Satoh. 2020. Which visual questions are difficult to answer? Analysis with entropy of answer distributions. *CoRR*, abs/2004.05595.

Carl Vondrick, Deniz Oktay, Hamed Pirsiavash, and Antonio Torralba. 2016. Predicting motivations of actions by leveraging text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2997–3005.

Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. 2015. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2461–2469.

Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. 2018. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3712–3722.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004.

Appendices

A Further Details on Data

A.1 Data Collection

Crowdsourcers are presented with detailed instructions and examples before starting with the annotation task. First, we introduce the task and provide them with some details to familiarize with the annotation tool. Then, we give them instructions regarding the constraints to be observed, i.e., for intentions: (1) to use the present tense and (2) do not mention any of the entities depicted in the image; for actions: (1) to use the present tense and (2) do mention entities that are visible in the image. To make instructions and constraints clearer, we show them several examples of good/wrong annotations

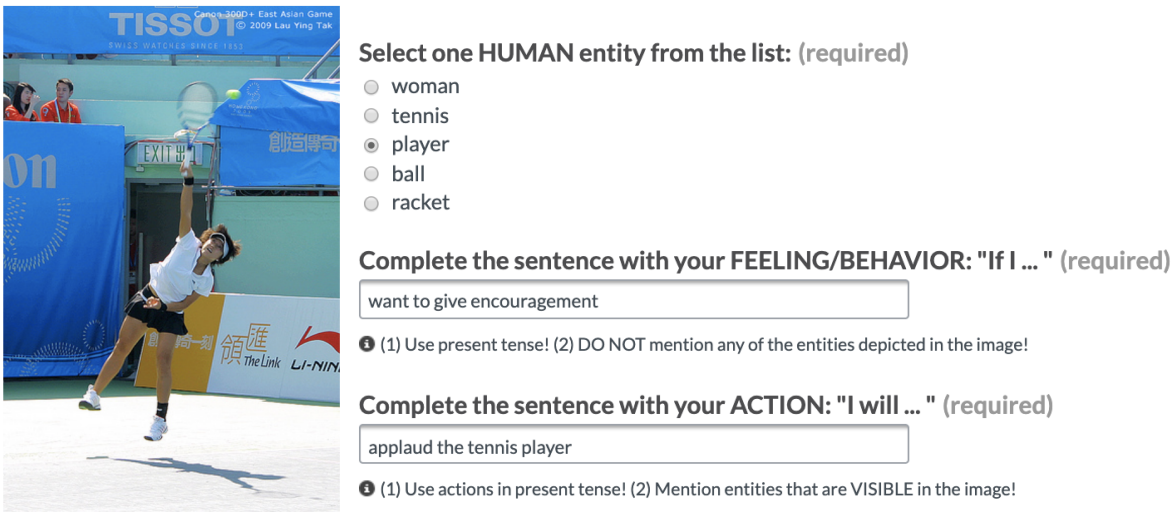

Figure 6: Data collection. One annotation sample presented to participants. Given an image, participants are asked to provide an intention and an action. To ensure they are doing the task properly, a verification question is asked preliminarly. Answering the question correctly (multiple correct answers) leads to the proper annotation phase.

(see Figure 2). Moreover, to make sure participants are performing the task properly (and, crucially, to avoid collecting fake data from automatic bots), a verification question is asked at the beginning of each image's annotation phase. The verification question has multiple correct answers, and only by picking one of these answers participants can proceed with the annotation phase (see Figure 6).

In addition, we add two sanity checks to the collected intentions. We check that (1) they have a length of at least 5 tokens; if this is not the case, participants are shown a warning and asked to fix their sentence; (2) they do not contain any noun referring to an entity that is grounded in the image; this is checked by means of a simple heuristic which extracts all the nouns from a given image's MS-COCO captions. Nouns with frequency > 1 are not allowed, and when typing them turkers are warned to modify their sentence.

A.2 BD2BB Dataset Statistics

As described in Section 4, the final BD2BB dataset includes $10,265$ samples, where each sample includes a $\langle image, intention, target\_action \rangle$ triple associated with 4 selected decoy actions. These triples were provided by 430 unique annotators. In particular, 253 were from the USA, 111 from the United Kingdom, 53 from Canada, 6 from Ireland, 5 from New Zealand, 2 from Australia. Each of them provided, on average, 23.87 $\langle image, intention, target\_action \rangle$ tuples contained in the dataset (min 1, max 192).

Each sample contains 5 actions. On average, these actions were provided by 4.90 unique annotators (min 3, max 5); moreover, they were collected for 4.96 (min 3, max 5) unique images, i.e., the decoy actions in each sample refer to different images than the target one in most of the cases.

A.3 Meta-Annotation

Topics We manually inspected the 60 clusters obtained through k-means clustering and removed 6 clusters for which we could not identify a coherent topic. Examples of the actions for each of the remainining 54 clusters, and the corresponding labels we assigned to them, are provided in Table 4. The 60 clusters were reviewed by two of the authors. We kept only clusters for which full agreement was met.

Numeric 4-Code Annotation We organize our data through a two-step system of *wordcodes* using codes to mark the syntatic class and the word-type. With the Stanford NLP parser (Chen and Manning, 2014), we extract from each action syntatic information and mark: 1) the main verb: "code1"; 2) the direct or indirect object of the main verb, as well as other complements related to the main verb: "code2"; 3) the second verb – if present (i.e., the verb of the coordinated or subordinated sentence): "code3"; 4) the object of the second verb – if present: "code4". In this case, we considered not only the direct object of the second verb, but also all the words referring to an object grounded

labels	action example	code1	code2	code3	code4
tennis	grab my tennis raquet firmly and hit the ball	grab	racket	hit	ball
food	grab some delicious food	grab	food		
cake	cut the cake	cut	cake		
snacks	purchase a hot dog	purchase	hotdog		
actions with ball	hit the ball as hard as i can	hit	ball		
skateboard 1	go skateboarding	go	skateboard		
bikes and motos	take a ride on the motorbike	ride	motorbike		
skateboard 4	pull off this skateboard trick	pull off	trick		
surf	grab my surfboard and join the woman	grab	surfboard	join	woman
phone	call someone for a chat	call	someone		
interact with people	join these people and talk	join	people	talk	
baseball 2	yell at the batter to distract him	yell	batter	distract	batter
sport audience	watch this game	watch	game		
approaching women	try to get the woman's attention	get	attention		
pizza	order a slice of pizza	order	pizza		
ski	use my ski poles judiciously	use	ski poles		
drink	i will drink my drink and watch people walk by	drink	drink	watch	people
kids	move the baby so i can use the computer	move	baby	use	computer
cooking	help those women to cook	help	women	cook	
videogames	grab an extra remote and join the game	grab	remote	join	game
pets	take a piece of cake and give it to the dog	take	cake	give	dog
clothing	wear my sun glasses	wear	glasses		
relax	i would look for a seat to rest	look for	seat		
umbrella	use the pink umbrella	use	umbrella		
urban activities	try to cross the street to investigate the trams	cross	street	investigate	trams
laptop	i will use that laptop the best way	use	laptop		
baseball 3	i will play as batter in a game of baseball	play	game		
baseball 1	watch a baseball game	watch	baseball game		
team sports	i play a soccer game	play	soccer		
frisbee 2	join a frisbee team	join	team		
birthday	i will sing happy birthday to the girl	sing	happy birthday		girl
water sports	grab my board and ride the waves	grab	board	ride	wave
photo	to go to the bathroom to get a selfie	go to	bathroom	get	selfie
zoo animals	ride an elephant	ride	elephant		
public transports	i will get on the bus and take a trip	get on	bus	take	trip
skateboard 2	will sit on the wall and watch the skateboarder	sit	wall	watch	skateboarder
frisbee 1	i will leave these men to play their little frisbee game	leave	men	play	frisbee
wii	play a wii game	play	wii		
bedtime	instead go into my room and lay down	go	room	lay	
manual work / hobbies	use the scissors to make oragmi	use	scissors	make	origami
animals farm	watch the man shear the sheep	watch	man	shear	sheep
good intentions	get the right job	get	job		
kite	enjoy watching the people fly their kites	enjoy		watch	people
horse riding	ride a horse	ride	horse		
toilet things	brush my teeth	brush	teeth		
skateboard 3	i will go to skate park	go	skatepark		
street scenes	stealthily unzip his backpack and take his possessions	unzip	backpack	take	possession
ski and snow	take off my shirt and do a big ski jump in front of her	take off	shirt	do jump	woman
snowboard	go snowboarding	go	snowboard		
airport	board that ancient plane	board	plane		
fruit	buy and eat a banana	buy	banana	eat	banana
haircut	use the hairdryer	use	hairdryer		
women and food	tell the girl i hope she enjoys her pizza	tell	girl	enjoy	pizza
reading	read the newspaper	read	newspaper		

Table 4: We report the label assigned to each of the 54 clusters (which summarizes its main topic), and one example of the actions included in it. Each action was annotated with codes to mark the verb (code1) and the complement object (code2) of the main sentence, and the verb (code3) and complements (code4) of the secondary sentence. Clusters are listed by their size: in descending order, from biggest to smallest.

labels	#actions	#code1	#code2	#code3	# code4
tennis	**580**	**90**	50	79	41
food	408	76	63	**81**	57
cake	334	60	37	65	**74**
snacks	316	68	**82**	26	50
actions with ball	298	71	27	54	34
skateboard 1	270	61	48	51	43
bikes and motos	269	86	55	59	51
skateboard 4	267	54	25	38	33
surf	262	66	50	52	22
phone	261	72	48	60	49
interact with people	261	66	58	62	22
baseball 2	259	82	42	69	30
sport audience	250	70	40	32	46
approaching women	227	84	54	49	70
pizza	226	43	23	37	42
ski	223	53	35	26	34
drink	222	53	46	50	39
kids	213	78	47	41	73
cooking	213	68	70	45	45
videogames	212	47	34	42	40
pets	202	80	47	44	32
clothing	202	54	61	48	47
relax	192	33	14	46	61
umbrella	186	56	24	32	26
urban activities	181	75	56	55	59
laptop	180	69	34	43	45
baseball 3	177	33	30	27	6
baseball 1	177	42	32	60	44
team sports	172	38	31	27	50
frisbee 2	172	25	25	29	22
birthday	170	62	71	46	59
water sports	165	87	60	38	41
photo	163	39	21	30	44
zoo animals	161	57	25	32	39
public transports	159	46	28	23	22
skateboard 2	158	45	36	35	25
frisbee 1	154	39	11	31	27
wii	149	36	22	35	22
bedtime	144	53	38	51	29
manual work / hobbies	139	69	75	44	60
animals farm	139	69	41	32	26
good intentions	132	66	64	44	32
kite	125	28	18	31	17
horse riding	118	49	22	22	29
toilet things	105	43	38	29	24
skateboard 3	98	22	16	18	14
street scenes	96	56	37	26	35
ski and snow	95	48	26	31	23
snowboard 1	94	27	26	21	17
airport	93	48	30	35	12
fruit	89	33	18	24	20
haircut	54	31	21	19	15
women and food	43	24	18	22	14
reading	32	11	11	11	7

Table 5: Statistics on the meta-annotation of the data. For each cluster, we report the number of actions, the number of verbs in the main (code1) and in the secondary sentence (code3), the number of nouns occuring as complements in the main (code2) and in the secondary sentence (code4).

in the corresponding image that specify the action expressed by the sentence. This way, for each action in which this was possible, we have a word that underlines the link between the linguistic and the visual aspect of the annotation. All the outputs by the parser were manually checked and fixed were needed. This was done by two of the authors: First, a subset of the data was annotated by the two auhtors together; then, each of the authors annotated a different subset. Only doubtful cases were discussed. In Table 4, for each action given as an example of the cluster, we highlight the words cor-

cluster	action	code1	code2	code3	code4
food	join the people in the restaurant to enjoy a meal	**join 1**	**people 77**	enjoy 15	meal 28
food	get some food with the people	get 107	food 6	0	**people 666**
frisbee	join this man playing frisbee	**join 9**	man 11	play 13	**frisbee 14**
frisbee	catch the frisbee and throw it again	catch 777	**frisbee 777**	throw 8	**frisbee 14**

Table 6: Examples of actions and corresponding word-type codes. Note that: (1) a given verb, e.g., *join*, is assigned different codes in different clusters (lines 1 and 3); (2) a given object within the same cluster, e.g., *frisbee* at line 4, is assigned different codes in different syntactic positions; (3) a given object, e.g., *frisbee* at lines 3 and 4, is assigned the same code if belonging to the same cluster and in the same syntactic position.

Model	**Number of parameters**
baseline$_L$	4931585
baseline$_V$	19251201
baseline$_{LV}$	21708801
RoBERTa$_L$	124646401
LXMERT$_V$	194352385
LXMERT$_{LV}$	194352385

Table 7: Number of parameters of each model. The number of parameters is the same both in models trained from scratch and in pre-trained ones.

responding to each of the four codes. Statistics about this meta-annotation are reported in Table 5.

Furthermore, for each topic cluster, we assign a numeric *wordcode* to each unique word-type in the 4 *syntactic classes* described above. In other words, each sentence is translated into a code composed of 4 numbers, each one representing a unique word in the corresponding *syntactic class*.[10] Illustrative examples are given in Table 6.[11]

B Further Details on Experiments

B.1 Models

The number of parameters of each model is reported in Table 7. The number of parameters is the same both in models trained from scratch and in pre-trained ones. The validation accuracy and epoch of the best models for each one of the three runs are reported in Table 8. For each of the three runs, we consider the model obtaining the best validation accuracy. For each model, we report mean and standard deviation of the test accuracies obtained across the three runs.

Baseline Our baseline is inspired by Jabri et al. (2016), but we use Softmax instead of Sigmoid as the final activation function to compute a probability distribution over all the candidates and choose the best one. We consider a version receiving image, intention and actions (**baseline**$_{LV}$), a version receiving image and actions (**baseline**$_V$), and a version receiving intention and actions (**baseline**$_L$). We used PyTorch 1.4.0. Baseline models were run on a CPU and their training took 33 seconds per epoch on average. We used a batch size equal to 32. We performed a grid search over two hyperparameters: the size of the hidden layer receiving concatenated figures (we tried values 8192 and 2048) and the dropout probability of zeroing elements of the input tensor right after the ReLU activation function (we tried values 0.0 and 0.5). The combination of parameters which leaded to the best validation accuracy was a hidden layer having size 8192 and a dropout probability of 0.0 corresponding to not having any dropout.

RoBERTa The RoBERTa$_{BASE}$ model we used has 12 self-attention layers with 12 heads each. It uses three special tokens, namely CLS, which is taken to be the representation of the given sequence, SEP, which separates sequences, and EOS, which denotes the end of the input. For each of the 5 $\langle image, intention, action \rangle$ datapoints in the sample, RoBERTa encodes the input as a sequence composed by CLS, the intention, SEP, the action, and EOS. As in the original work, we use the representation corresponding to the CLS token to use the encoder in the downstream task. For RoBERTa we used PyTorch 1.0.1 and we started from the source code available at `https://github.com/huggingface/transformers`. Both when fine-tuning the pre-trained model and when training the model from scratch, we used a batch size equal to 32 with 8 gradient accumulation steps, thereby having a batch size equal to 256, a weight decay equal to 0.01, gradient clipping equal to 5, and a learning rate which is warmed up over the

[10] When we choose to consider more than one object, we create a compositional code, using the '+' mark

[11] Here numbers are assigned randomly, just to provide a concrete example of our meta-annotation.

Model	Run 1		Run 2		Run 3	
	Epoch	**Valid. acc.**	**Epoch**	**Valid. acc.**	**Epoch**	**Valid. acc.**
baseline$_L$	19	0.449	28	0.446	41	0.462
baseline$_V$	25	0.453	21	0.467	23	0.453
baseline$_{LV}$	22	0.481	34	0.496	36	0.480
RoBERTa^s_L	3	47.1	2	46.8	2	47.1
LXMERT^s_V	8	32.0	8	29.9	48	30.7
LXMERT$^s_{LV}$	35	50.2	9	50.8	28	50.2
RoBERTa$_L$	12	0.571	36	0.557	38	0.550
LXMERT$_V$	38	0.593	49	0.588	31	0.592
LXMERT$_{LV}$	44	0.643	36	0.647	18	0.595

Table 8: Epoch and validation accuracy of the best models for each run.

first 10% steps to a peak value of 0.00005 and then linearly decayed.

LXMERT The LXMERT model we used has a Object-Relationship Encoder and a Language Encoder which encode relationships between regions and relationships words, respectively, through a self-attention mechanism, and a Cross-Modality Encoder which encode relationships between regions and words and vice-versa through a cross-modal attention mechanism followed by a self-attention mechanism. The number of layers in the Language Encoder, Object-Relationship Encoder, and Cross-Modality Encoder are 9, 5, and 5, respectively. As in RoBERTa, LXMERT uses the special tokens CLS and SEP. Differently from RoBERTa, LXMERT uses the special token SEP both to separate sequences and to denote the end of the textual input. As in the original work, we use the representation corresponding to the CLS token to use the encoder in the downstream task. For RoBERTa we used PyTorch 1.0.1 and we started from the source code available at https://github.com/airsplay/lxmert. As with RoBERTa, both when fine-tuning the pre-trained model and when training the model from scratch, we used a batch size equal to 32 with 8 gradient accumulation steps, thereby having a batch size equal to 256, a weight decay equal to 0.01, gradient clipping equal to 5, and a learning rate which is warmed up over the first 10% steps to a peak value of 0.00005 and then linearly decayed.