Probing Image–Language Transformers for Verb Understanding

Lisa Anne Hendricks Aida Nematzadeh
DeepMind
{lmh, nematzadeh} @google.com

Abstract

Multimodal image–language transformers have achieved impressive results on a variety of tasks that rely on fine-tuning (*e.g.*, visual question answering and image retrieval). We are interested in shedding light on the quality of their pretrained representations – in particular, if these models can distinguish different types of verbs or if they rely solely on nouns in a given sentence. To do so, we collect a dataset of image–sentence pairs (in English) consisting of 421 verbs that are either visual or commonly found in the pretraining data (*i.e.*, the Conceptual Captions dataset). We use this dataset to evaluate pretrained image–language transformers and find that they fail more in situations that require verb understanding compared to other parts of speech. We also investigate what category of verbs are particularly challenging.

1 Evaluating Verb Understanding

The success of image–language models in real-world applications relies on their ability to relate different aspects of language (such as verbs or objects) to images, which we refer to as multimodal understanding. For example, an image-retrieval model needs to distinguish between "eating an apple" and "cutting an apple" and a captioning model must accurately describe the actions in a scene.

Previous work shows that image–language benchmarks do not always fully measure such multimodal understanding: object retrieval models fail to account for linguistic structure (Akula et al., 2020), visual question answering (VQA) models overly rely on language priors (Goyal et al., 2017; Agrawal et al., 2018), and captioning metrics do not always measure if captions "hallucinate" objects in an image (Rohrbach et al., 2018). Inspired by this, prior work introduced tasks to specifically examine whether models can relate objects to images (Shekhar et al., 2017) or classify frequent interactions associated with objects (Chao et al., 2015). However, both these datasets are limited to the 80 objects in the MSCOCO detection challenge (Lin et al., 2014).

To address this gap, we design a benchmark focused on verbs called SVO-Probes for examining **s**ubject, **v**erb, **o**bject triplets; more specifically, we collect a set of image–sentence pairs (in English) where each pair is annotated with whether the sentence corresponds to the image or not. As shown in Fig. 1, for a given sentence, in addition to a *positive* image that matches the sentence, our dataset includes controlled *negative* images that do not correspond to specific aspects of the sentence (*i.e.*, subject, verb, and object). These controlled examples enable us to probe models for their understanding of verbs as well as subjects and objects. Our dataset consists of 421 verbs and includes over $48,000$ image–sentence pairs.

We use our benchmark to evaluate the recent family of multimodal (image–language) transformers that have shown impressive results on benchmarks like VQA and image retrieval (Lu et al., 2019; Chen et al., 2020; Tan and Bansal, 2019; Li et al., 2020b,a; Huang et al., 2020). Our goal is to investigate if the good performance of these models is due to learned representations that successfully relate different aspects of language to images. More specifically, we evaluate a few architectural variations of these models in a zero-shot way by using the pretrained models to classify if image–sentence pairs from SVO-Probes match.

Our results show that the performance of all evaluated models is worst on verbs, with subjects being easier than verbs but harder than objects. We find that this observation does not depend on the frequency of test examples in pretraining data. Moreover, it is considerably harder for all models to correctly classify image–sentence pairs that do not

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 3635–3644
August 1–6, 2021. ©2021 Association for Computational Linguistics

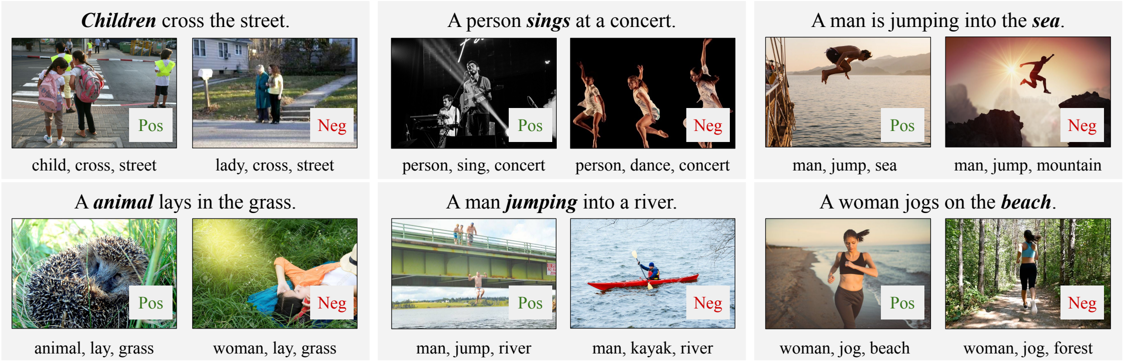

Figure 1: Examples from SVO-Probes. Images on the left and right show positive and negative image examples for each sentence. Below each image is the ⟨subject, verb, object⟩ triplet corresponding to the image.

match; the image–language transformers overpredict that sentences corresponds to images.

Additionally, we compare an image–language transformer pretrained on a large automatically-curated dataset (*i.e.*, Conceptual Captions, Sharma et al., 2018) with one pretrained on the smaller but manually-annotated MSCOCO (Chen et al., 2015). Conceptual Captions is more noisy than MSCOCO in that its sentences do not necessarily correspond to its images. Interestingly, we observe that the model pretrained on MSCOCO performs better. This result shows that the image–language transformers are not robust to dataset noise as they learn to predict that somewhat-related image–sentence pairs correspond to each other.

Despite their good performance on downstream tasks, image–language transformers fail on our task that requires multimodal understanding since they cannot distinguish between finer-grained differences between images. Our results highlight that there is still considerable progress to be made when training multimodal representations, and that verbs in particular are an interesting challenge in image–language representation learning.

2 Related Work

Image–language transformers build on the transformer architecture (Vaswani et al., 2017) by incorporating additional loss functions (to learn image features and align image and language modalities), using self-attention to combine modalities, and training on paired image–text data (Lu et al., 2019; Chen et al., 2020; Tan and Bansal, 2019; Li et al., 2020b,a; Huang et al., 2020). The impressive performance of these models on many image–language benchmarks has inspired recent work that studies different architectural choices made in these models (Cao et al., 2020; Hendricks et al., 2021).

Compared to previous image–language models, multimodal transformers both use a new architecture and are frequently trained on a much larger dataset – the Conceptual Captions dataset consisting of $3m$ image–text pairs (Sharma et al., 2018). Singh et al. (2020) show that on fine-tuned tasks, the performance of multimodal transformers (*i.e.*, Lu et al., 2019; Li et al., 2019) are less sensitive to dataset size; the domain match between pretraining and fine-tuning datasets is more important.

Datasets. Our proposed dataset is most similar to the FOIL benchmark (Shekhar et al., 2017) which tests if image–language models can differentiate between sentences that vary with respect to only one noun. FOIL consists of $64,300$ images from MSCOCO (Chen et al., 2015); each image is paired with a corresponding sentence that describes the image (*i.e.*, a *positive* example) and one that does not (*i.e.*, a *negative* example). Negative sentences are collected by replacing object words in the positive sentences with a similar object (*e.g.*, changing the word "dog" to "cat" in "The dog ran."). Shekhar et al. (2017) use the FOIL dataset in a few tasks including a classification task where the model is asked to classify if a sentence matches the image or not. We use the same task setup because it allows us to probe image–language transformers in a zero-shot setting as these models are generally trained to classify whether an image–text pair match. Our work is different than FOIL in that we focus on verb understanding as opposed to noun understanding; moreover, our dataset provides different negative types (by replacing subjects, verbs, or objects).

Other datasets focus on relationship or interaction detection (*e.g.*, HICO and VRD; Chao et al., 2015; Lu et al., 2016a). These datasets are evalu-

dataset	Ims	Subjs	Verbs	Objs	Sents	Negs
FOIL	32k	n/a	0	70	✓	✓
HICO	10k	n/a	117	80	✗	✓
VRD	1k	100	70	100	✗	✗
V-COCO	5k	n/a	26	48	✗	✗
ImSitu	25k	950	504	1840	✗	✗
SVO-Probes	14k	100	421	275	✓	✓

Table 1: **Im**ages, **Subj**ects, Verbs, **Obj**ects, **Sent**ences, and **Neg**atives in other datasets and SVO-Probes. Image numbers are for the *evaluation* set.

ated in a classification setting in which the input is an image and the output is a detected relationship (for HICO, an object and interaction, for VRD two objects and their relationship) and have a limited number of verbs and objects. V-COCO (Gupta and Malik, 2015) and ImSitu (Yatskar et al., 2016) both includes verbs but do not provide negatives for a controlled evaluation of verb (or noun) understanding. Finally, other work has explored how creating hard negatives (*e.g.*, by substituting words in train examples) leads to better test performance (Gupta et al., 2020; Hendricks et al., 2018; Faghri et al., 2017). In contrast, our work focuses on creating hard evaluation examples to probe learned representations.

In summary, SVO-Probes is unique as it tests understanding of a broad range of verbs as well as subjects and objects in a controlled way. Furthermore, our dataset includes image–sentence pairs; thus, it can be used to evaluate image–language transformers that process image–sentence pairs. Finally, SVO-Probes is designed as a zero-shot task to evaluate pretrained image–language transformers and is collected to have a similar distribution to Conceptual Captions which is commonly used in pretraining these models. See Table 1 for a comparison between SVO-Probes and other datasets.

3 Task Setup and Dataset Collection

Our goal is to examine *verb-understanding* in *pretrained* multimodal transformers. To do so, we need a task that requires an understanding of a given verb in a sentence, *e.g.*, a model cannot succeed at the task by relying on nouns. We also need to include a diverse set of verbs, and examine each verb in at least a few situations. To test the pretrained representations, we need to examine the models in a zero-shot setting (without fine-tuning).

Inspired by the FOIL setup (Shekhar et al., 2017), we use a zero-shot classification task where a model is asked to identify if a sentence and an image correspond to each other. As a result, we need a dataset that provides "match" or "not match" labels between images and sentences. We collect a dataset of image–sentence pairs (SVO-Probes) that given a sentence, provides such labels for at least two images.[1] Some of these images are *positive* examples, *i.e.*, the sentence correctly describes them. Others are *negative* examples where some aspect of the sentence (*e.g.*, verb) does not match the image. Figure 1 shows some examples from our dataset.

We systematically collect negative examples such that they only differ from the positive image with respect to the subject, verb, or object of the sentence. Finally, we consider sentences whose subjects, verbs, and objects are frequent in the Conceptual Captions (CC) training dataset. Since CC is the dataset most frequently used for pretraining multimodal transformers, we can examine what the pretrained representations capture (in contrast to examining these models' generalization ability). We next describe our pipeline to create SVO-Probes.

Creating a verb list. To ensure that we have a large number of verbs in our dataset, we first created a verb list by considering a subset of verbs that occur in the train split of the Conceptual Captions dataset (CC-train). More specifically, we consider verbs that are visually recognizable in the images; to identify the visual verbs, we use the *imSitu* dataset (Yatskar et al., 2016) that includes verbs that annotators marked as reliably recognizable. Moreover, we include verbs that occur at least 50 times in CC-train.

Curating triplets. Given a positive example, we need to systematically generate negatives by replacing the subject, verb, or the object. As a result, we collect a set of ⟨subject, verb, object⟩ (SVO) triplets from CC-train sentences for our verbs. We extract the subject, verb, and direct object from the dependency parse trees. and remove triplets where subjects or objects are pronouns or have less than two characters. Finally, we discard SVO triplets with frequency smaller than five.

We consider three negative types for a given triplet: a subject-, verb-, or object-negative where respectively, the subject, verb, or object in the triplet are replaced by a different word. For example, given the triplet ⟨girl, lie, grass⟩, examples of subject-negative, verb-negative, and object-

[1] We note that our dataset is limited to English sentences; we simply use "sentences" to refer to English sentences.

negative are ⟨puppy, lie, grass⟩, ⟨girl, sit, grass⟩, and ⟨girl, lie, beach⟩.

Since our goal is to examine verb understanding, we only keep the triplets that have at least one verb negative. This enables us to evaluate a model's capacity in distinguishing images that mainly differ with respect to the verb; for example, ⟨girl, lie, grass⟩ vs. ⟨girl, sit, grass⟩. Adding this constraint results in 11230 SVO triplets and 421 verbs. In this set, 1840 SVO triplets (and 53 verbs) have at least two verb and object negatives.

Collecting images. The next step is collecting images that match the curated SVO triplets. We query for SVO triplets using the Google Image Search API. We retrieve 5 images for each triplet, then remove any images with urls in Conceptual Captions. To make sure that these automatically-retrieved images certainly match the triplets, we set up an annotation task where we ask workers on Amazon Mechanical Turk (AMT) to verify if the subject, verb, and object are present in the image. We ask three people to annotate each image, and only keep images where at least two annotators agree that the subject, verb, and object are depicted in the image. Moreover, we discard images marked as a cartoon by annotators. We find that 58% of our images pass this initial annotation process. We pay workers $0.04 per HIT for all tasks.

Collecting sentences. Multimodal transformer models are trained on pairs of images and *sentences*; to evaluate them, we require image–sentence pairs as opposed to image–SVO pairs. Given an image and an SVO triplet, we next ask annotators to write a sentence that uses all the words in the triplet and describes the image. For example, as shown in Figure 1 top right, given the triplet ⟨man, jump, sea⟩, an annotator might write "A man is jumping into the sea.". We ask annotators to refrain from writing additional information to ensure that a collected sentence examines the words in the SVO (as opposed to words that we are not controlling for). Annotators are given the option to not write a sentence if they do not think the subject, verb, and object can be combined into a grammatical sentence that describes the image. 86% of our images pass this phase of our pipeline.

We observe that for a given SVO, different images elicit slightly different sentences. For example, the triplet ⟨person, jog, beach⟩ resulted in the sentences "A person jogging along the beach." and "A person jogs at the beach.". Additionally, annotators pluralize nouns to ensure the sentence describes the image (*e.g.*, Figure 1 top left, the subject "child" is written as "children" in the sentence).

Confirming the negative image. Finally, given a *positive* triplet (*e.g.*, ⟨girl, lie, grass⟩) and its negative (*e.g.*, ⟨girl, sit, grass⟩), we need to confirm that the positive's sentence does not match the image retrieved for the negative triplet. To do so, we ask three annotators to select which images (positive, negative, neither, or both) match a given sentence. Image–sentence pairs where two out of three annotators agree are accepted into our dataset; 68% of the pairs pass this final annotation stage.

4 Experimental Setup and Results

We investigate if current image–language transformers can relate different aspects of language (and in particular verbs) to images by evaluating these models against both FOIL and SVO-Probes. More specifically, we evaluate a few architectural variations of image–language transformer models (based on the implementation of the models by Hendricks et al., 2021) that differ in their choice of multimodal attention and loss functions; this way we can examine whether our findings are sensitive to these slight differences. The base multimodal transformer (**MMT**) closely replicates the ViLBERT architecture (Lu et al., 2019): this model includes three loss functions, masked language modeling (MLM) and masked region modeling (MRM) losses on the language and image inputs and an image–text matching (ITM) loss that classifies if an image–sentence pair match. Importantly, the multimodal attention of MMT is similar to the hierarchical co-attention in Lu et al. (2016b) where each modality (*i.e.*, image or language) attends *only* to the other modality. More specifically, in the multimodal self-attention layer of transformer (Vaswani et al., 2017), for queries on the language input, keys and values are taken from images and vice versa.

Different interactions of image (language) queries, keys, and values in multimodal self-attention results in variations of image–language transformers. We describe the model variations we study in Table 2. We also consider models that either lack the MLM or MRM loss. Models are pretrained on Conceptual Captions (CC) unless stated otherwise. For reference, we report the Recall@1 performance on the zero-shot image-retrieval task

Name	Multimodal Attention	Similar Model	MLM	MRM	*ZS Flickr*
MMT	Queries from L (I) take values and keys from *only* I (L)	ViLBERT; LXMERT	✓	✓	**41.9**
Merged–MMT	Queries from L (I) take values and keys from *both* L and I	UNITER	✓	✓	**40.0**
Lang–MMT	Queries are *only* from L (Hendricks et al., 2021)		✓	✓	33.6
Image–MMT	Queries are *only* from I (Hendricks et al., 2021)		✓	✓	31.6
SMT	Single-Modality Transformers without multimodal attention		✓	✓	16.9
No-MRM–MMT	The same as MMT		✓	✗	**41.1**
No-MLM–MMT	The same as MMT		✗	✓	20.2

Table 2: Different variants of the image–language transformer architecture we test. L and I stand for language and image, respectively. We note that models with Merged attention (like UNITER) are also referred to as single-stream models. ViLBERT: Lu et al. (2019); LXMERT: Tan and Bansal (2019); UNITER: Chen et al. (2020)

on Flickr (*ZS Flickr*), where a model must retrieve an image from the Flickr dataset (Young et al., 2014) that matches an input sentence. Since MMT performs best on *ZS Flickr*, we do most of our experiments on this model unless stated otherwise.

We first evaluate our image–language transformers on FOIL to examine their noun understanding and then test them on SVO-Probes which probes for subject, verb, and object understanding in learned representations. Following FOIL, we report the accuracy on positive and negative pairs. All our models have an image-text classification output used in pretraining to align images and sentences. We calculate accuracy by passing images through our models and labeling an image–sentence pair as negative if the classifier output is < 0.5 and positive otherwise. We report the average over the two pairs (see Avg columns in Tables 3 and 4) by weighting them equally, since we expect models to perform well on both positive and negative pairs. In FOIL, there are equal positive and negative pairs.

Another possible way to set-up our evaluations is as image-retrieval (reporting recall@1 as a metric). However, the retrieval setting does not highlight the difference in performance between positive and negative pairs. For example, a model might rank the pairs correctly even when their scores are very close (positive score is 0.91 and negative one is 0.9). In this example, the model is wrong about the negative pair (it is assigned a high score) but the retrieval setting does not capture this. However, the classification metric will penalize the model for assigning a high score to a negative pair. As a result, the classification metric better differentiates between the models by examining if they correctly label both the positive and negative pairs.

4.1 Evaluating Nouns with FOIL

We examine noun understanding in image–language transformers with the FOIL dataset (Shekhar et al., 2017). Given image–sentence pairs from FOIL, we evaluate the MMT model in a zero-shot setting by using it to classify if the image and sentence match. Table 3 compares MMT with the best model from the FOIL paper (HieCoAtt Shekhar et al., 2017) and, to our knowledge, the best-performing model on the task without using ground-truth annotations (Freq+MM-LSTM from Madhyastha et al., 2018). Note that these models are trained specifically for the FOIL task (*i.e.*, on the train split of FOIL), whereas the MMT model (pretrained on CC) is tested in a zero-shot setting.

MMT achieves an accuracy considerably worse than the best models on FOIL (Shekhar et al., 2017; Madhyastha et al., 2018) on all pairs; this is surprising given that image–language transformers achieve state-of-the-art results on zero-shot image retrieval tasks based on Flickr (Young et al., 2014) and MSCOCO (Chen et al., 2015). In particular, MMT overpredicts that image–sentence pairs match, resulting in the highest accuracy on the positive pairs (99.0) but the lowest on negative pairs (11.8). Thus MMT cannot distinguish between sentences that only differ with respect to nouns.

We investigate whether this poor performance of MMT is due to mismatch between the pretraining (*i.e.*, CC) and FOIL test (*i.e.*, MSCOCO) datasets. Thus, we compare our MMT model pretrained on Conceptual Captions with one pretrained on MSCOCO (MMT-COCO). As expected, MMT-COCO has considerably higher performance on all pairs (compare to MMT); however, the accuracy is still significantly higher on positive pairs than negative ones, showing that the model overpredicts that image–sentence pairs match. Our result shows that despite their impressive performance on downstream tasks, image–language transformer models perform poorly in distinguishing between semantically similar sentences. Next we examine how well these models perform on our proposed

Model	Avg	Pos.	Neg.
HieCoAtt*	64.1	91.9	36.4
Freq + MM-LSTM †	**87.9**	86.7	**89.0**
MMT	55.4	**99.0**	11.8
MMT-COCO	72.0	95.0	49.0

Table 3: Performance on FOIL averaged over all (Avg), positive (Pos.), and negative (Neg.) pairs. *Shekhar et al. (2017); †Madhyastha et al. (2018).

probing dataset which is designed to have a similar vocabulary to the CC pretraining dataset.

4.2 Comparing Models on SVO-Probes

We evaluate all models (see Table 2) on SVO-Probes and report overall accuracy and accuracy for subject, verb, and object negatives in Table 4.

The MMT model (with the best performance on *ZS Flickr*) performs poorly on SVO-Probes, achieving an overall average accuracy of 64.3. The best overall average accuracy (No-MRM–MMT; 69.5) shows that SVO-Probes is challenging for image–language transformers. In particular, models struggle with classifying negative pairs; Lang–MMT achieves the highest accuracy over negative pairs (56) which is slightly higher than chance at 50. [2]

Though No-MRM–MMT and MMT perform similarly on *ZS Flickr*, No-MRM–MMT performs better on SVO-Probes. This suggests that the masked region modelling loss is not needed for good performance on *ZS Flickr*; also, it impedes the model from learning fine-grained representations needed to perform well on SVO-Probes. More surprisingly, Lang–MMT, which performs worse on *ZS Flickr* than MMT, outperforms MMT on SVO-Probes. The image representations in Lang–MMT are not updated with an attention mechanism. In Sec. 4.4, we explore if the stronger attention mechanism in MMT leads to overfitting of the training images and thus weaker performance.

We crafted SVO-Probes such that it includes words from the pretraining dataset of image–language transformers (*i.e.*, CC), whereas FOIL is collected from MSCOCO. Comparing the performance of MMT (with CC pretraining) on FOIL and SVO-Probes (55.4 in Table 3 vs. 64.3 in Table 4), we see that the domain mismatch between pretraining and test data plays a role in MMT's performance. Interestingly, comparing the performance of MMT-COCO (MMT with COCO pretraining) on FOIL to MMT (with CC pretraining) on SVO-Probes, we find that SVO-Probes is more challenging than FOIL when there is no domain mismatch (72.0 in Table 3 vs. 64.3 in Table 4).

When comparing different negative types across *all* models, we observe that verbs are harder than subjects and objects; compare average accuracy for Subj., Verb, and Obj. Negative columns in Table 4. For example, in MMT, the subject and object negative average accuracies (67.0 and 73.4) are considerably higher than the average accuracy for verb negatives (60.8). Moreover, when breaking down the accuracies for positive and negative pairs (Pos. and Neg. columns in Table 4), we observe that the accuracies of positive pairs are similar (ranging between 80.2 and 94.4) across all models except SMT (which performs close to chance); however, for negative pairs, there is more variation in accuracy across models especially for verb negatives (ranging between 22.4 and 54.6, Neg. columns under "Verb Negative"). These results show that negative pairs are better than positive ones in distinguishing between different model architectures.

We also find that subjects are harder than objects across all models (when comparing average accuracies of subject and object negatives). To better understand this result, we examined 21 nouns that occur both as subjects and objects in SVO-Probes' sentences. Interestingly, over these 21 nouns, for our MMT model, the accuracies of negative pairs are 42.9 and 56.4 for subject and object negatives, respectively. This suggests that the subject position might be more challenging than the object one which we further explore in Sec. 4.3.

4.3 Accuracy and Frequency at Training

Our overall results on SVO-Probes (Table 4) show that for image–language transformers, verb negatives are more challenging than subject and object ones, and also subject negatives are harder than object ones. We examine if this observation is due to properties of SVO-Probes as opposed to differences specific to subjects, verbs, and objects. First, we explore whether the frequency of SVO triplets in pretraining data impacts the accuracy of negative pairs in our MMT model. We focus on negative

[2] We focus on image–language transformers, but we also tested a baseline model where image features are embedded with the detector used in our transformers and language features with BERT. Features are pooled using element-wise multiplication. This baseline achieves 66.3% accuracy overall with 75.4% and 57.3% accuracy on positives and negatives. Similar to transformers, performance on verbs is the worst.

	Overall			Subj. Negative			Verb Negative			Obj. Negative		
	Avg	Pos.	Neg.	Avg	Pos.	Neg.	Avg	Pos.	Neg.	Avg	Pos.	Neg.
# Examples	48k	12k	36k	8k	3k	5k	34k	11k	23k	11k	3k	8k
MMT	64.3	*93.8*	34.8	67.0	*94.4*	39.5	60.8	*93.8*	27.8	73.4	*94.4*	52.4
Merged–MMT	64.7	**94.4**	35.0	69.1	**94.9**	43.2	60.7	**94.4**	27.0	74.1	**94.9**	53.3
Lang–MMT	*68.1*	80.2	**56.0**	*71.5*	82.1	**60.9**	*64.5*	80.2	*48.9*	*77.7*	81.4	**74.1**
Image–MMT	64.3	91.6	37.0	68.2	92.1	44.2	59.7	91.6	27.8	75.6	91.5	59.6
SMT	52.4	49.1	*55.6*	52.6	47.7	57.5	51.8	49.1	**54.6**	53.9	50.7	57.0
No-MRM–MMT	**69.5**	85.4	53.7	**73.5**	87.4	*59.7*	**65.5**	85.6	45.5	**80.1**	86.2	**74.1**
No-MLM–MMT	60.8	92.3	29.3	64.8	93.9	35.8	57.4	92.5	22.4	69.5	93.6	45.5

Table 4: Results on SVO-Probes on different models for subject, verb, and object negatives. Best results are shown in bold; second best results are italicized.

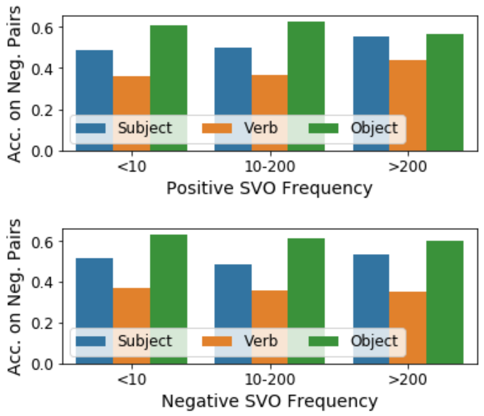

Figure 2: Accuracy of negative pairs for subject, verb, and object negatives given SVO frequencies in CC.

pairs as there is more variation in negative-pair accuracies across both models as well as subject, verb, and object negatives. We consider the frequency of positive and negative SVO triplets: a positive SVO corresponds to a positive image matching a given sentence, but a negative SVO and its extracted negative image do not match the sentence.

We group SVOs based on their frequency in CC-train into low (less than 10), medium (between 10-200), and high (greater than 200) frequency bins. Fig. 2 plots the negative-pair accuracy for subject, verb, and objects across these different frequency bins over positive and negative SVO frequencies. We confirm that our result on the difficulty of negative types does not depend on the frequency of positive or negative SVOs in pretraining data. In both plots of Fig. 2, the negative types in order of difficulty (lower accuracy) are verbs, subjects, and objects independent of the frequency bin.

Similarity between SVOs. We examine if the similarity between the SVO triplets corresponding to the negative and positive images can explain the difference in performance of subject, verb, and object negatives. For example, we expect that distinguishing ⟨child, cross, street⟩ and ⟨adult, cross, street⟩ to be harder than differentiating one of them from ⟨dog, cross, street⟩: “child” and “adult” are more similar to each other than to “dog”. To test this, we measure the similarity between subjects, verbs, and objects in their corresponding negative types using the cosine similarity between word2vec (Mikolov et al., 2013) embeddings.

The average similarities between subjects, verbs, and objects are 0.49, 0.29, 0.27, respectively. Thus, subject words in negative examples tend to be more similar than object words. Furthermore, we find that there is a small positive correlation (as measured by Spearman rank correlation) between SVO similarity and classifier scores for negative pairs – .264 and .277 for subjects and objects respectively – suggesting that when SVOs corresponding to the image and sentence are similar, the classifier tends to assign a higher score (more positive) to the pair. This partially explains why accuracy on subjects is lower than on objects in Table 4. Even though verb negatives are harder for our model, the similarity for verb negatives is similar to that of object negatives. The correlation coefficient between similarity and classifier score is weaker (.145) for verbs, suggesting that word similarity factors less in how well the model classifies verb negatives.

4.4 Similarity to Pretraining Data

We next consider the similarity between images in SVO-Probes and CC. To measure the similarity between images, we sample 1 million images from CC. For each image in SVO-Probes, we find the 10 nearest neighbors in the feature embedding space

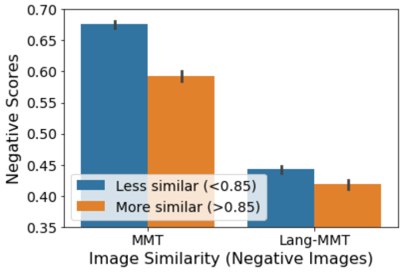

Figure 3: Comparing negative scores on MMT and Lang–MMT for images less or more similar to CC.

of CC, and average the distance to compute a similarity score for the image. Figure 3 plots the average score from our classifier for negative pairs with images that are less or more similar to the pretraining data (since we are classifying negative pairs, the lower score the better). We compare the MMT and Lang–MMT models since they have considerably different performance on SVO-Probes and *ZS Flickr*. The difference in average scores between less similar and more similar examples for MMT is 0.083. This is noticeably greater than the difference in average scores between less and more similar examples for Lang–MMT (0.024), suggesting that the image similarity influences Lang–MMT less than MMT. One hypothesis is that the stronger attention mechanism in MMT overfits to the training images which makes the MMT model less robust.

4.5 The Choice of Pretraining Dataset

In Sec. 4.2, we observe that models perform particularly poorly in classifying negative pairs. We investigate whether the choice of pretraining dataset impacts this observation. Conceptual Captions (CC), the most-common pretraining dataset for image–language transformers, is curated by scraping images and alt-text captions from the web. As a result, compared to manually-annotated datasets such as MSCOCO, CC is noisy – it contains examples where the sentence and its corresponding image do not completely align. For example, a sentence can mention objects that are not in the image or, in extreme cases, does not describe the image at all.

We hypothesize that image–language transformers treat correspondences due to dataset noise as "real" relations; in other words, they learn that if a image–sentence pair is somewhat semantically related, it should be classified as a positive match, even if some aspects of the sentence do not describe the image. At the same time, we can think of negatives in SVO-Probes as examples with noisy correspondences where a specific aspect of a sentence (*e.g.*, the verb) does not match the image. We compare our MMT model (with CC pretraining) to one trained on a manually-annotated and less noisy dataset, MSCOCO (referred to as MMT-COCO).

	Overall			Neg. Acc.		
Train	Avg.	Pos.	Neg.	S	V	O
CC	64.3	93.8	34.8	39.5	27.8	52.4
COCO	68.0	75.2	60.9	66.0	55.5	73.4

Table 5: Comparing performance when training our MMT model on COCO and CC.

Table 5 reports the overall accuracy of the two models on SVO-Probes as well as a breakdown over subject, verb, and object negatives for negative-pair accuracies. MMT-COCO performs better than MMT pretrained on CC (avg. accuracy of 68 vs 64.3). This is surprising since MMT-COCO has a different image and language distribution in its pretraining dataset. The accuracy of positive pairs in MMT-COCO is considerably lower than MMT while it performs noticeably better for negative pairs: unlike MMT, the MMT-COCO model does not overpredict that image–sentence pairs match. Our results show the image–language transformers are not robust to dataset noise. Less-noisy datasets (such as MSCOCO), despite their small size and domain mismatch, are more suitable for learning representations that are sensitive to finer-grained differences in images. Alternatively, models which are more robust to noise in datasets could be beneficial for tasks like ours.

4.6 Which Verbs Are the Hardest?

We investigate which verbs are hardest for MMT. We consider verbs with many examples in SVO-Probes: we keep SVO triplets with at least 30 negative images, resulting in a set of 147 verbs and 887 SVO triplets across $4,843$ images. Table 6 lists the easiest and hardest verbs (with highest and lowest accuracy for negative pairs) for the MMT model. Easy and hard verbs have a diverse set of properties; for example, easy verbs include sporting activities like "tackle" as well as verbs like "lead" that occurs in a variety of contexts. We also examine the 20 most difficult and easiest verbs for *all* our models (described in Table 2). Most difficult verbs for all models include: "cut", "argue", and"break" and the easiest ones include: "direct", "battle", "surround", "skate", and "participate".

Easy	Hard
tackle, reach, arrive, pitch, accept, congratulate, lead, present, celebrate, attend	argue, beat, break, burn, buy, cast, comb, crash, cut, decorate

Table 6: Hard and easy verbs for our MMT model

We test if verbs that occur in both SVO-Probes and imSitu are easier for our model to classify. Verbs in imSitu are considered visual as the dataset collection pipeline for imSitu includes an explicit annotation step to determine if verbs are visual. Surprisingly, we find verbs in imSitu are harder for our MMT model. On closer inspection, some verbs in our dataset but *not* in imSitu (*e.g.*, "swim") are clearly visual. An interesting future direction is to investigate which visual properties of a verb make it harder or easier for image–language models to learn.

5 Conclusions

Although image–language transformers achieve impressive results on downstream tasks, previous work suggests performance on these tasks can be confounded by factors such as over-reliance on language priors (Goyal et al., 2017). We collect a dataset of image–sentence pairs to examine multimodal understanding by testing the ability of models to distinguish images that differ with respect to subjects, verbs, and objects.

Our results show that image–language transformers fail at identifying such fine-grained differences; they incorrectly classify image–sentence pairs that do not match. Surprisingly, a model trained on a manually-annotated and smaller dataset does better on our task, suggesting that models have trouble ignoring noise in larger but automatically-curated pretraining datasets. Additionally, verb understanding is harder than subject or object understanding across all models we study. This motivates the need for researchers to not only examines models on objects, but develop datasets and architectures which allow for better verb understanding as well.

Acknowledgements

We would like to thank Antoine Miech for detailed comments on our paper. Also, thanks to Phil Blunsom, Laura Rimell, Stephen Clark, Andrew Zisserman, Jean-Baptiste Alayrac and the anonymous reviewers for their helpful feedback. We also thank Susie Young, Zhitao Gong, and Tyler Liechty for supporting our data collection.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't 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.

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. *arXiv preprint arXiv:2005.01655*.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. *arXiv preprint arXiv:2005.07310*.

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision (ECCV)*.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

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.

Saurabh Gupta and Jitendra Malik. 2015. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*.

Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. 2020. Contrastive learning for weakly supervised phrase grounding. *ECCV*.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 264–279.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. 2021. Decoupling the role of data, attention, and losses in multimodal transformers. *arXiv preprint arXiv:2102.00529*.

Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pretraining. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*.

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

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

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.

Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2016a. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer.

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.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016b. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297.

Pranava Madhyastha, Josiah Wang, and Lucia Specia. 2018. Defoiling foiled image captions. *NAACL*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. *NeurIPS*.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *EMNLP*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

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)*.

Amanpreet Singh, Vedanuj Goswami, and Devi Parikh. 2020. Are we pretraining it right? digging deeper into visio-linguistic pretraining. *arXiv preprint arXiv:2004.08744*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Empirical Methods in Natural Language Processing*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Conference on Computer Vision and Pattern Recognition*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.