Seeing things or seeing scenes: Investigating the capabilities of V&L models to align scene descriptions to images

Anonymous ACL submission

Abstract

Images can be described in terms of the objects they contain, or in terms of the types of scene or place that they instantiate. In this paper we address to what extent pretrained Vision and Language models can learn to align descriptions of both types with images. We compare 3 state-of-the-art models, VisualBERT, LXMERT and CLIP. We find that (i) V&L models are susceptible to stylistic biases acquired during pretraining; (ii) only CLIP performs consistently well on both object- and scene-level descriptions. A follow-up ablation study shows that CLIP uses object-level information in the visual modality to align with scene-level textual descriptions.

1 Introduction

Grounding symbols in perception (Harnad, 1990) is a crucial step towards achieving full understanding of natural language (Bender and Koller, 2020; Bisk et al., 2020). This endeavour has received new impetus through the development of pretrained Vision and Language (V&L) models (e.g. Lu et al., 2019; Tan and Bansal, 2019; Li et al., 2019; Chen et al., 2020; Li et al., 2020a; Su et al., 2020; Wang et al., 2020; Luo et al., 2020; Li et al., 2021; Huang et al., 2021; Radford et al., 2021). Similarly to unimodal language models such as BERT (Devlin et al., 2019), V&L models are intended to be task-agnostic and are extensively pretrained on paired image-text data, achieving good performance on several tasks after finetuning (e.g. Lu et al., 2020; Li et al., 2020c; Kim et al., 2021). Pretraining usually includes an image-text alignment task to discover implicit cross-modal relationships. Although the importance of this task is widely recognized and adopted during model pretraining, it is unclear how the models perform on it, since they are usually evaluated on downstream tasks.

The data used for V&L pretraining usually contains highly descriptive text which mentions objects and their spatial relationships. For instance, the COCO (Chen et al., 2015) and Localized Narratives (LN; Pont-Tuset et al., 2019) captions for Figure 1 are of this type, though they differ stylistically. By contrast, the third caption in the figure, from the novel HL1K dataset introduced in Section 3 below, is what we refer to as 'scene-level', focusing on *what type* of scene or location is depicted.

LN: This is the picture of a stadium. In the foreground there is a person [. . .] At the back there are group of people sitting [. . .].
COCO: a baseball player getting ready to swing at a baseball game in a stadium packed with people.
HL1K: the picture is shot in a baseball field

Figure 1: An example of scene with COCO and Localized Narrative (LN) object-level captions, versus HL1K scene-level description (Section 3)

Note that both the object- and scene-level descriptions in the Figure describe the picture, albeit in different ways. Indeed, it would be expected that, for a V&L model to display true grounding capabilities, it should be able to match both types of descriptions with the image. For models which do display this cabability, a natural follow-up question is whether their representations capture interesting connections between scenes on the one hand, and the objects within them on the other.

Research on human perception suggests that humans do not perceive scenes exclusively in terms of the objects they contain, and that visual salience is not exclusively determined by bottom-up features such as colour and texture. Rather, visual stimuli

are considered 'scenes' because their elements constitute a meaningful whole, both in terms of their contents (e.g. one expects an oven in a kitchen, but not in a living room) and in terms of their spatial arrangement (e.g. ovens do not typically hang from the ceiling) (Malcolm et al., 2016).

These observations have provided the impetus to work showing that violations of scene 'semantics' (content) and 'syntax' (spatial arrangement) exact a cognitive cost during perception (e.g. Biederman et al., 1982; Võ and Wolfe, 2013). A related strand of modeling research in computer vision has also shown that scene-level priors generate expectations about objects and their configurations, impacting the salience of objects in a way that classical, feature-based models of attention (e.g. Itti and Koch, 2001) would not predict (Torralba et al., 2006; Oliva and Torralba, 2007). Indeed, the problem of linking low-level features with high-level semantic information is an instance of the problem referred to as the 'semantic gap' in computer vision (Ma et al., 2010).

In this paper we investigate whether V&L models are able to handle object-level and scene-level descriptions equally well. A positive answer to this question would suggest that such models are learning useful associations between the elements of a scene and the overall scene type, as captured in textual descriptions.

We perform an analysis in a zero-shot setting on three state-of-the-art pretrained V&L models. To our knowledge, this is the first systematic comparison of model capabilities on object- versus scene-level grounding. The goal of this study is therefore not to establish new SOTA results, but to further our understanding of what V&L models learn, as a function of the data they are pretrained on and the model architecture. Therefore we choose three models differing in many settings (including training set size, architecture, number of parameters and model size). All of the models are however optimized on the image-sentence alignment task.

We find that only one of the models under comparison, CLIP (Radford et al., 2021), performs consistently well on both object- and scene-level image-text matching. We then investigate this model's abilities in depth, using an ablation method to identify the elements of a text and/or an image which contribute to these abilities.

	Training size (# image-sentence pairs)	Model size (# parameters)	Pretraining Objectives
CLIP	400M	151M	ISA
VisualBERT	330k	112M	ISA, MLM
LXMert	9.18M	228M	ISA, MLM MOP, VQA

Table 1: Comparison of training settings for the three models (**ISA**: Image-Sentence Alignment, **MLM**: Masked Language Modeling, **MOP**: Masked Object Prediction, **VQA**: Visual Question Answering)

2 Models

Current V&L models typically combine textual and visual features in a single or a dual-stream architecture. Though the two architectures have been found to perform roughly at par when trained on the same data in comparable settings Bugliarello et al. (2020), in this paper we include widely-used representatives of both at the time of writing, as we are interested in their zero-shot grounding capabilities in their original settings. We also include a third model which differs in structure and is trained on a much larger and more varied dataset. Table 1 gives an overview of some of the properties of the models we consider.

LXMERT (Tan and Bansal, 2019) is a dual-stream model, which encodes text and visual features in parallel, combining them using cross-modal layers. LXMERT is trained on COCO captions (Chen et al., 2015) as well as a variety of VQA datasets, with an image-text alignment objective, among others. We use the implementation of LXMERT in the `transformers`[1] library.

VisualBERT (Li et al., 2019) is a single-stream, multimodal version of BERT (Devlin et al., 2019), with a Transformer stack to encode image regions and linguistic features and align them via self-attention. It is pretrained on COCO captions (Chen et al., 2015). Image-text alignment is conceived as an extension of the next-sentence prediction task in unimodal BERT. Thus, VisualBERT expects an image i and a correct caption c_1, together with a second caption c_2, with the goal of determining whether c_2 matches $\langle i, c_1 \rangle$. We use the publicly available implementation of the model.[2]

CLIP (Radford et al., 2021) combines a transformer encoder for text with an image encoder based on Visual Transformer (Dosovitskiy et al.,

[1] github.com/huggingface/transformers
[2] https://github.com/uclanlp/visualbert

	LXMERT		VisualBERT	
	I	C	I	C
COCO	✓	✓	✓	✓
LN	✓	✗	✓	✗
HL1K	✓	✗	✓	✗
ADE20K	✗	✗	✗	✗

Table 2: Presence of the (I)mages and (C)aptions of the dataset used for the experiments in the training data of VisualBERT and LXMERT. The composition of CLIP's training data is not known.

2020), jointly trained using contrastive learning to maximise scores for aligned image-text pairs. CLIP is trained on around 400m pairs sourced from the Internet, a strategy similar to the web-scale training approach used for unimodal models such as GPT-3 (Brown et al., 2020). We note that the visual backbone for this model differs from that of LXMERT and VisualBERT, both of which use Faster-RCNN (Ren et al., 2015).

For all experiments, we truncate textual captions to a maximum length of 50 tokens, following standard practice for such models, including CLIP.

3 Data

We use four different datasets for our experiments, which overlap to different degrees with the data that LXMERT and VisualBERT were trained on.[3] The extent of overlap is shown in Table 2.

Localized Narratives Localized Narratives (LN) Pont-Tuset et al. (2019) is a V&L dataset created by transcribing speech from annotators who were instructed to give object-by-object descriptions as they moved a mouse over image regions. LN captions tend to be highly detailed and stylistically similar to speech. We use LN as a source of object-level captions. The images in LN come from pre-existing datasets; this allows us to align LN captions with images and captions from datasets such as COCO and ADE20K.

ADE20K ADE20K (Zhou et al., 2017) is a computer vision dataset containing 20k images comprehensively annotated with objects, parts and scene labels. We use ADE20K as a source of scene-level captions. For our experiments, we filter out images with scenes which in the dataset are labelled as `unknown`. We produce captions for each image using a simple template-based generation method, whereby a scene label is inserted into one of the templates below:

- *it is a* SCENE
- *this is a* SCENE
- *it is located in* SCENE

We align the resulting scene-level descriptions and the corresponding ADE20K images to the corresponding object-level captions in LN.

COCO COCO (Lin et al., 2014a) consists of images paired with captions and object annotations. LN captions are also available for the same images. We use images and captions from the 2017 COCO validation split, as well as the corresponding LN captions.

HL1K **H**igh **L**evel Scenes - 1k (HL1K) is a new dataset collected for the purposes of the present study. HL1K is composed of 1k images, each depicting at least one person, sampled from the 2014 COCO `train` split. We crowd-sourced three annotations per image on Amazon Mechanical Turk, showing crowd workers the image and asking them to write a description in response to the question *Where is the picture taken?* Crowd workers were asked to respond using full sentences and it was made clear to them that their answer to this question should bring to bear their knowledge of typical, or common, scenes. Figure 2 shows an image with three different scene descriptions.

Where is the picture taken?

- in a bedroom
- the picture is taken in a bedroom
- this is the bedroom

Figure 2: COCO image with three HL1K scene descriptions.

Descriptions were corrected for typos using the Neuspell Toolkit (Jayanthi et al., 2020). Finally,

[3] CLIP was trained on a web-harvested dataset.

we paired our scene-level HL1K captions with the previously available COCO and LN object-level captions. Figure 1 provides an example.

	ADE20k	HL1K	COCO
images	19733	1000	5000
COCO captions	19733	3000	5000
Localized Narratives	19733	1000	5000

Table 3: Dataset statistics.

Dataset statistics are shown in Table 3. For ADE20k, the numbers are for images which are not labelled as having an *unknown* scene. In COCO, there are five captions associated with each image. For the purposes of the present study, a single caption is randomly selected in each case.

4 Image-sentence alignment experiments

We first test models in the image-sentence alignment task on both object- and scene-level descriptions. Since we are interested in the capabilities of the pretrained models, and since pretraining included alignment for all models we use (see Table 1), we do not finetune them. Rather, we use the models' pretrained alignment head to predict whether a scene-level or object-level caption correctly describes an image, or not.[4]

Table 4 shows that LXMERT and VisualBERT perform adequately on object-level COCO Captions, though performance is lower than would be expected, given that they were pretrained on this dataset. In the case of LXMERT, one possible explanation is catastrophic forgetting, arising from the fact that this model is pretrained for its final ten epochs on VQA (similar observations are made by Parcabalescu et al., 2021). For both models, performance drops dramatically on LN captions. This is likely due to a stylistic difference: compared to COCO captions, LN captions are longer, more discursive and contain disfluencies.

In contrast, CLIP performs close to ceiling on all three datasets, possibly reflecting the benefits accrued from the size and diversity of its pretraining data.

On scene-level captions, performance is somewhat above chance for LXMERT on ADE20k template-based descriptions, and for VisualBERT on HL1K. Otherwise, performance is below 50% for both models. Once again, CLIP performs above 90%, though there is a drop in performance from the template-based ADE20k descriptions to human-authored HL1K scene-level captions, possibly reflecting the more predictable nature of the former.

[4]Note that this setting is the same used by the models is their pretraining.

		LXMERT	CLIP	VisualBERT
Object	ADE20k + LN	28.4	**96.8**	39.0
	COCO + LN	59.1	**98.7**	65.2
	COCO Cap.	79.3	**99.1**	64.4
Scene	ADE20k	58.0	**97.6**	17.3
	HL1K	45.5	**91.5**	55.3

Table 4: Image-sentence alignment accuracies on object-level and scene-level captions. Chance performance is at 50%. (*LN* = Localized Narratives)

5 Ablation experiments on CLIP

Since CLIP is the only one of the three models which is successful at matching scene-level and object-level captions to images, we probe its capabilities further, paying particular attention to the question whether CLIP links *scene types* (e.g. kitchen) to *scene contents* (e.g. oven, pizza) in image-text matching.

Whereas a standard image-text alignment setup compares the model's success at identifying actual versus random captions, here we directly compare the preference of the model for scene- versus object-level descriptions, as a function of (i) the entities mentioned in the object-level caption; (ii) the entities visible in an image. To this end, we use textual and visual ablation on captions and images; an example is shown in Figure 3.

5.1 Textual ablation

Given an object-level caption, we identify all the NPs in the caption and create new versions by removing each possible subset of the set of NPs, with the restriction that the resulting caption must always contain at least one NP. When NP removal results in dangling predicates, we remove them to preserve grammaticality. NPs are detected with Spacy v.3, using the pipeline for English with the `en_core_web_md` pretrained models. The right panel of Figure 3 shows the original caption and examples of ablated captions.

For a given image i with object-level caption o and scene-level caption s, we compare how $P(o|i)$ – CLIP's estimate of the probability that o matches i – changes as NPs are removed from o, and to what extent this causes CLIP to assign higher prob-

Original Image

Occluded Image

COCO: A man rides a motorcycle on a road through a grassy, hilly area.

Ablated Captions:

- a grassy, hilly area *(A man, a motorcycle, a road)*
- a road *(A man, a motorcycle, a grassy, hilly area)*
- a road through a grassy, hilly area *(A man, a motorcycle)*
- A man rides a road *(a motorcycle, a grassy, hilly area)*
- a motorcycle a grassy, hilly area *(A man, a road)*
- A man rides a grassy, hilly area *(a motorcycle, a road)*
- A man rides a motorcycle *(a road, a grassy, hilly area)*
- A man rides a road through a grassy, hilly area *(a motorcycle)*
- a motorcycle on a road *(A man, a grassy, hilly area)*
- A man rides a motorcycle on a road *(a grassy, hilly area)*
- A man rides a motorcycle a grassy, hilly area *(a road)*
- a motorcycle on a road through a grassy, hilly area. *(A man)*

Figure 3: Example of visual and textual ablation. *Left*: Original image and image with occluded object. *Right*: Original caption and different ablated captions. NPs removed are shown in parentheses.

ability $P(s|i)$, to s as the match for i. We report two comparisons, one on LN captions versus ADE20K template-based scene descriptions; and one on COCO captions against HL1K scene-level descriptions.

To control for possible loss of grammaticality after ablation, we score ablated captions with GRUEN (Zhu and Bhat, 2020), a BERT-based model which has been shown to yield scores that correlate highly with human judgments.[5] CLIP probabilities for ablated textual captions yielded a significant, but very low correlation with grammaticality (Pearson's $r = 0.1, p < .01$) suggesting that grammaticality did not affect the scores.

5.2 Visual ablation

Given an object-level caption and an image, we extract all nouns from the caption and extract the embedding vector for each noun using pretrained FastText embeddings.[6] We pass the image through the Faster-RCNN object detector[7] to detect entities. We extract embeddings for each entity label. Then, we identify regions to be masked by comparing embeddings for entity labels l_e against embeddings for nouns n_e in the caption, considering them a match if cosine$(l_e, n_e) \geq 0.7$. Bounding box regions corresponding to matched entities are occluded with a greyscale mask. The left panel of Figure 3 compares the original and masked image.

Once again, we are interested in whether CLIP's estimate of the alignment probability of object- versus scene-level captions, changes as elements of the visual input are masked.

	ADE20k	HL1K
T	205k	10027
V	10788	625
V+T	1078	625

Table 5: Total number of ablations generated per dataset, across all the ablations experiments using T(extual) ablation, V(isual) ablation, or both (V+T).

Table 5 provides the number of ablations anal-

[5] GRUEN returns a combined score consisting of a linear combinaton of Grammaticality, Focus and Coherence. Here, we use only the Grammaticality scores.

[6] We use the model with 2m word vectors trained with subword information from common Crawl `https://fasttext.cc/docs/en/english-vectors.html`

[7] Faster R-CNN ResNet-50 FPN pre-trained on COCO, available from the `torchvision` module in `Pytorch`

	ADE20k		HL1K	
	LN	Scene	COCO	Scene
No ablation	55.6	44.4	95.7	4.3
T	22.0	78.0	67.2	32.8
V	74.9	25.1	71.2	28.8
V+T	68.4	31.6	63.3	36.7

Table 6: CLIP preferences for object-level versus scene-level captions. Results shown are percentages of times the model assigns higher alignment probability to object-level versus scene-level captions (sub-columns), when there is no ablation, T(extual) ablation, V(isual) ablation, or both (V+T) for each dataset (columns).

ysed in the study. For both ADE20k and HL1k we obtain a number of ablated captions which is greater than the respective dataset sizes in Table 3, because for each example we generate all the possible combinations of noun phrases. For the Visual and Visual+Textual ablations, the number of ablated instances is lower than the dataset size, because we omit all the images where no object is detected.

5.3 Results

The results of image-sentence alignment using CLIP, after ablation, are shown in Table 6.

Without ablation, the model assigns higher probability to object-level descriptions, suggesting that CLIP has higher confidence in aligning an image-text pair when the text focuses on objects rather than scenes. This preference is far more marked for COCO/HL1K, in line with the observation (Table 4) that HL1K scene descriptions are somewhat more challenging for this model.

As entity-level information is removed from the object-level caption (row T in Table 6), the model's assigns higher probability to the scene-level caption, suggesting that the model leverages the visual information to align with the scene description.

In contrast, visual ablation (row V) results in the opposite tendency: when entities are occluded in the image, the model assigns higher probability to object-level captions compared to scene-level descriptions.

These results suggest that CLIP aligns images to scene-level descriptions based on the entities visible in the images. As these are masked in the image, entity-level captions are aligned with higher probability. On the other hand, when both sources of information are ablated, CLIP once again assigns higher probability to object-level captions.

5.4 Scenes vs. entities

Our findings suggest that CLIP reasons about scenes on the basis of salient objects within them. If this is the case, then the probability assigned by clip to an image-scene caption pair should diminish, as more salient entities are visually ablated in the image.

To investigate this further, we use scene labels extracted from the HL1K captions and the object detections produced for the visual ablation (Section 5.2). For a scene label s and entity label e, we compute $P(s|e)$ as follows. Let e be an entity detected n_e times in the dataset, of which $n_{e,s}$ times in images depicting scene s. We compute:

$$P(s|e) = \frac{n_{e,s}}{n_e}$$

Figure 4 shows visualisations for entities detected in four example scene types found in the HL1K dataset.

For all images with at least three detected entities, we consider the image-sentence alignment probability assigned by CLIP to the *scene*-level description, when the top 1, 2 or 3 most likely entities in the scene are masked. We therefore average over those images containing at least three detected entities (53/174 total scenes).

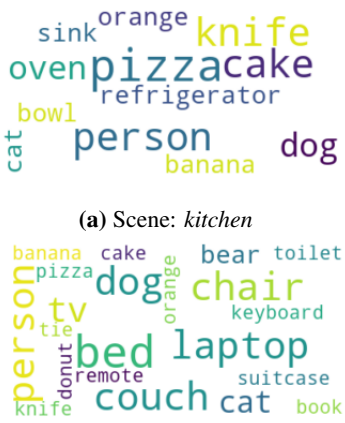
Figure 5 displays the average alignment probability assigned by CLIP to images and scene-level captions, as entities are progressively masked in the image. The figure displays a linear trend, with the probability dropping as more likely entities are removed. A one-way ANOVA comparing the change in log probability as 1, 2 or 3 entities are removed showed that the difference is significant ($F(2, 156) = 4.25, p < 0.05$).

Thus, when CLIP aligns images with scenes, it is relying on object-level information in the visual modality. This explains why the removal of object mentions in text results in higher preference for scene-level descriptions, since the objects are detectable in the image. By the same token, masking objects in images causes the model to rely more on the entity-level information in the text.

5.5 Effect of length and informativeness

So far, our analysis suggests that CLIP reasons about scenes based on object-level information.

However, the length of the caption might be a possible confounding factor. Some of our results might simply be due to the model assigning a higher alignment probability to a caption which



(a) Scene: *kitchen* **(b)** Scene: *road*

(c) Scene: *room* **(d)** Scene: *park*

Figure 4: Visualisations of entities (e) in four different scene types (s). Font size is proportional to $P(s|e)$

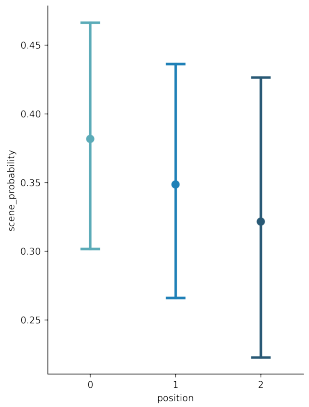


Figure 5: CLIP scene-level description probabilities after masking top 1-3 entities. Error bars represent standard deviations.

resort: 3% *person*: 96%

resort: 99% *snowboard*: 1%

Figure 6: Scene vs entity one-to-one comparison. In the top image, there are many people in the foreground and the entity *person* is preferred over the scene label *resort*. At the bottom, people are snowboarding in the background and the scene label is preferred over the the entity label *snowboard*.

is longer or more informative. This could provide an alternative explanation for the changes observed above in the alignment probabilities after textual ablation.

To account for this, we replicate the alignment experiment using single words. Once again, we use the scene labels extracted from HL1K scene-level descriptions and identify the top three most likely entities in a given scene, as in the previous experiment (see Figure 5).

Given an image, we compare image-text alignment probabilities in CLIP for single-word object labels (e.g. *motorbike*) and single-word scene labels (e.g. *road*).

In this setting, CLIP displays a moderate preference for scene labels (63%), suggesting that such labels are more informative than object-level labels, for the one-word alignment task.

We performed a qualitative analysis, inspecting 5 cases where the model has a clear preference for scene label or object label. A representative example is shown in Figure 6. CLIP assigns higher probability to object labels when images have salient, foregrounded entities. When entities are less salient or in the background, the model prefers scene labels.

6 Related work

V&L models have been extensively evaluated on tasks such as Visual Question Answering (Goyal

et al., 2017) or image retrieval (Lin et al., 2014b). More recently, there has been increased interest in understanding the nature of the representations and capabilities learned by large, pretrained models, for example via probe tasks or investigation of their attention heads (see Belinkov and Glass, 2019, for a survey). This has also been done for V&L models. For example, Li et al. (2020b) consider VisualBERT's attention heads in a manner similar to Clark et al. (2019), showing that it is able to ground entities and syntactic relations (see also Ilharco et al., 2020; Dahlgren Lindström et al., 2020). Hendricks and Nematzadeh (2021) similarly seek to obtain an in-depth understanding of the representations learned by V&L models, finding that they have difficulty with grounding verbs in visual data, compared to other morphosyntactic categories. The present work has a similar motivation, but focuses on models' ability to reason in a grounded way about the relationship between entities and scenes. Our method of ablation in the textual and visual modalities was developed concurrently with similar methods by Frank et al. (2021), who use it to uncover asymmetries in the extent to which V&L models rely on textual or visual modalities.

More generally, a number of tasks have been developed to test the ability of V&L models to reason with a combination of linguistic and visual cues, including VCR (Zellers et al., 2019), SWAG (Zellers et al., 2018) and NLVR (Suhr et al., 2017, 2019). Pezzelle et al. (2020), in work complementary to our own, address the relationship between visual and textual modalities, exploring a task in which the text does not provide an object-level description of an image.

Scene recognition is a central task in computer vision, with extensive work on scene categorisation systems (e.g. Anderson et al., 2021) and several datasets in addition to the ones used in this paper, including ImageNET (Deng et al., 2009), Places (Zhou et al., 2014) and SUN (Xiao et al., 2010). However, there has been little work at the V&L interface, exploring the capabilities of models to link scene- and object-level representations. Some precedent for the concerns addressed in this paper are found in the image captioning literature. For example, an influential proposal by (Anderson et al., 2018) combines top-down and bottom-up attention to combine local and global features. CapWAP (Fisch et al., 2020) conditions image captioning on questions that determine which information is relevant to current communicative needs, going beyond object-level description. Closer to the scope of the work presented here, a recent pretrained V&L model, SemVLP (Li et al., 2021), combines single- and dual-streams for feature-level and high-level semantic alignment. We plan to investigate this model further in future work.

7 Conclusions

In order to address the symbol grounding problem, V&L models should be able to capture the relationship between an 'object-level' view of an image, focusing on objects and their configuration, and the higher-level scene it corresponds to. This paper has found that when models do this, they rely on object-level information in the *visual* modality, to link images to scene descriptions in the *textual* modality; this is influenced by the probability of entities occurring in particular scene types.

Of the models tested, we find that LXMERT and VisualBERT perform poorly on this task, and also suffer when captions deviate stylistically from their pretraining data. For these models, testing on ADE20k, amounts to a full zero-shot setting, whereas for Localized Narratives and HL1K, this only applies to the textual input, as the images are included in their training data. With the exception of HL1K, a new dataset, it is an open question whether testing for CLIP was zero-shot, since this model was trained on web-scale data, which is often unfathomable (Bender et al., 2021). On the other hand, model size is clearly not the determining factor; CLIP has fewer parameters than LXMERT, for example (cf. Table 1).

We believe that two additional factors contribute to the success of CLIP. First, its contrastive learning objective may result in greater sensitivity to fine-grained distinctions between captions for image-sentence alignment. A second feature is its visual backbone, which (in the version used in this paper) is based on Visual Transformer (ViT Dosovitskiy et al., 2020). Recently, BERT-inspired architectures have achieved notable success on computer vision tasks (see also Bao et al., 2021). Tuli et al. (2021) have shown that ViT is more consistent with characteristics of human vision than a convolutional network, extracting image features which are not strictly local. This could partially underlie the model's ability to use object-level information in an image to align to scene-level captions.

8 Ethical considerations

For the studies presented here, we used a new dataset, HL1K, collected using the Amazon Mechanical Turk crowdsourcing platform. For the data collection, participants were shown images and asked to answer questions such as *Where is the picture taken?* Answers took the form of short statements. Workers were paid at the rate of €0.03 per item, an amount we consider equitable for the work involved, and in line with rates for similar tasks. No sensitive or identifying information was collected. All other data and models used are publicly available. The HL1K dataset will be made available upon publication.

References

Matt D Anderson, Erich W Graf, James H Elder, Krista A Ehinger, and Wendy J Adams. 2021. Category systems for real-world scenes. *Journal of Vision*, 21(2)(8):1–31.

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.

Hangbo Bao, Li Dong, and Furu Wei. 2021. BEiT: BERT Pre-Training of Image Transformers. *arXiv*, 2106.08254:1–16.

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the fourth ACM Conference on Fairness, Accountability, and Transparency (FAccT'21)*, Online. Association for Computing Machinery.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. 1982. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *ArXiv*, 2005.14165.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2020. Multimodal pretraining unmasked: Unifying the vision and language berts. *arXiv preprint arXiv:2011.15124*.

Xinlei Chen, Hao Fang, Tsung-yi Lin, Ramakrishna Vedantam, C Lawrence Zitnick, Saurabh Gupta, and Piotr Doll. 2015. Microsoft COCO Captions : Data Collection and Evaluation Server. *arXiv*, 1504.00325:1–7.

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 *ECCV*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? An analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Adam Dahlgren Lindström, Johanna Björklund, Suna Bensch, and Frank Drewes. 2020. Probing multimodal embeddings for linguistic properties: the visual-semantic case. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 730–744, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, pages 248–255.

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.

Alexey Dosovitskiy, Lucas Beye, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognitoin at scale. *arXiv*, 2010.11929.

Adam Fisch, Kenton Lee, Ming-Wei Chang, Jonathan Clark, and Regina Barzilay. 2020. CapWAP: Image captioning with a purpose. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8755–8768, Online. Association for Computational Linguistics.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-Language or Vision-for-Language? On Cross-Modal Influence in Multimodal Transformers. *ArXiv*, 2109.04448.

Yash Goyal, Tejas Khot, Aishwarya Agrawal, 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 International Conference on Computer Vision and Pattern Recognition (CVPR'17)*, pages 6904–6913.

Stevan Harnad. 1990. The symbol grounding problem. *Physica*, D42(1990):335–346.

Lisa Anne Hendricks and Aida Nematzadeh. 2021. Probing Image-Language Transformers for Verb Understanding. *ArXiv*, 2106.09141.

Zhicheng Huang, Zhaoyang Zeng, Yupan Huang, Bei Liu, Dongmei Fu, and Jianlong Fu. 2021. Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning. *ArXiv*, 2104.03135.

Gabriel Ilharco, Rowan Zellers, Ali Farhadi, and Hannaneh Hajishirzi. 2020. Probing Contextual Language Models for Common Ground with Visual Representations. *arXiv*, 2005.00619.

L Itti and C Koch. 2001. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203.

Sai Muralidhar Jayanthi, Danish Pruthi, and Graham Neubig. 2020. Neuspell: A neural spelling correction toolkit. *arXiv preprint arXiv:2010.11085*.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*.

Chenliang Li, Ming Yan, Haiyang Xu, Fuli Luo, Wei Wang, Bin Bi, and Songfang Huang. 2021. SemVLP: Vision-Language Pre-training by Aligning Semantics at Multiple Levels. *arXiv preprint*, 2103.07829.

Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 11336–11344. AAAI Press.

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. In *Arxiv*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020b. What Does BERT with Vision Look At? In *Proceedings ofthe 58th Annual Meeting ofthe Association for Computational Linguistics (ACL'20)*, pages 5265–5275, Online. Association for Computational Linguistics.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020c. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

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

Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014b. Microsoft COCO: Common objects in context. In *Proceedings of the 2014 European Conference on Comupter Vision (ECCV'14)*, volume 8693 LNCS, pages 740–755, Berlin and Heidelberg. 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, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. 2020. UniVL: A unified video and language pre-training model for multimodal understanding and generation. *arXiv*.

Hao Ma, Jianke Zhu, Michael Rung Tsong Lyu, and Irwin King. 2010. Bridging the semantic gap between image contents and tags. *IEEE Transactions on Multimedia*, 12(5):462–473.

George L. Malcolm, Iris I.A. Groen, and Chris I. Baker. 2016. Making Sense of Real-World Scenes. *Trends in Cognitive Sciences*, 20(11):843–856.

Aude Oliva and Antonio Torralba. 2007. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527.

Letitia Parcabalescu, Albert Gatt, Annette Frank, and Iacer Calixto. 2021. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. In *Proceedings of the Workshop Beyond Language: Multimodal Semantic Representations (MMSR'21)*, Groningen, The Netherlands.

Sandro Pezzelle, Claudio Greco, Greta Gandolfi, Eleonora Gualdoni, and Raffaella Bernardi. 2020. Be Different to Be Better! A Benchmark to Leverage the Complementarity of Language and Vision. In *Findings ofthe Association for Computational Linguistics: EMNLP 2020*, pages 2751–2767, Online. Association for Computational Linguistics.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2019. Connecting Vision and Language with Localized Narratives. *arXiv*, 1912.03098.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint*, 2103.00020.

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 28 (NeurIPS 2015)*, Montreal, Canada.

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 *International Conference on Learning Representations*.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A Corpus of Natural Language for Visual Reasoning. In *Proceedings ofthe 55th Annual Meeting ofthe Association for Computational Linguistics (ACL'17)*, pages 217–223, Vancouver, BC. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. *arXiv*, 1811.00491:6418–6428.

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 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. 2006. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4):766–786.

Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. 2021. Are Convolutional Neural Networks or Transformers more like human vision? *arXiv*, 2105.07197.

Melissa L.H. Võ and Jeremy M. Wolfe. 2013. Differential Electrophysiological Signatures of Semantic and Syntactic Scene Processing. *Psychological Science*, 24(9):1816–1823.

Jianfeng Wang, Xiaowei Hu, Pengchuan Zhang, Xiujun Li, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. 2020. MiniVLM: A Smaller and Faster Vision-Language Model. *arXiv*, 2012.06946.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'10)*, pages 3485–3492.

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

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. *arXiv*, 1808.05326.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. 2017. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641.

Wanzheng Zhu and Suma Bhat. 2020. GRUEN for evaluating linguistic quality of generated text. *arXiv*, 2010.02498.