Visual Commonsense in Pretrained Unimodal and Multimodal Models

Anonymous ACL submission

Abstract

Our commonsense knowledge about objects includes their typical visual attributes; we know that bananas are typically yellow or green, and not purple. Text and image corpora, being subject to reporting bias, represent this world-knowledge to varying degrees of faithfulness. In this paper, we investigate to what degree unimodal (language-only) and multimodal (image and language) models capture a broad range of visually salient attributes. To that end, we automatically extract a visually-grounded commonsense dataset covering 5 property types (color, shape, material, size, and visual co-occurrence) for over 5000 subjects. We validate this dataset by showing that our grounded color data correlates much better than ungrounded text-only data with crowdsourced color judgments provided by Paik et al. (2021). We then use our dataset to evaluate pretrained unimodal models and multimodal models. Our results show that multimodal models better reconstruct attribute distributions, but are still subject to reporting bias. Moreover, increasing model size does not enhance performance, suggesting that the key to visual commonsense lies in the data.

1 Introduction

The observation that human language understanding happens in a rich multimodal environment has led to an increased focus on visual grounding in natural language processing (NLP) (Bisk et al., 2020; Baltrusaitis et al., 2019), driving comparisons between traditional unimodal text-only models and multimodal models which take both text and image inputs. In this work, we explore to what extent unimodal and multimodal models are able to capture commonsense visual concepts across five types of relations: color, shape, material, size, and visual co-occurrence (see Fig. 1). We further explore how such an ability is influenced by the reporting bias (defined in Section 2.3) in training data. We measure visual commonsense, defined as knowledge about visual properties which humans are implicitly aware of even without explicit visual cues, through frequency distributions. A visually aware language model should be able to capture such properties upon elicitation. The color, shape, material, and co-occurrence data are mined from Visual Genome (Krishna et al., 2016), and the size data are created from object lists.

Paik et al. (2021) evaluate language models' color perception using a human-annotated color dataset (CoDa), finding that reporting bias negatively influences model performance and that multimodal training can mitigate those effects. In this work, we confirm those findings while extending the evaluation to a broader range of visually salient properties, resulting in a more comprehensive metric for visual commonsense. In order to elicit visual commonsense from language models, we utilize soft prompt tuning (Qin and Eisner, 2021), which trains optimal templates by gradient descent for each model and relation type that we explore. We also utilize knowledge distillation to enhance a text-only model's visual commonsense ability, where the vision-language model serves as the teacher.

The major contributions of this work are: (1) we design a comprehensive analytic dataset for probing English visual commonsense that is applicable to any language model; (2) apply the dataset to study the capacity of unimodal language models and multimodal vision-language (VL) models to capture empirical distributions of visually salient properties; and (3) train a knowledge-distilled version of a VL model that achieves improved performance on our task. The dataset and code will be made available at `http://anonymous_url`.

2 Related Work

2.1 Vision-Language Modeling

Recent advances in vision-language modeling have achieved great success. Most of them learn joint

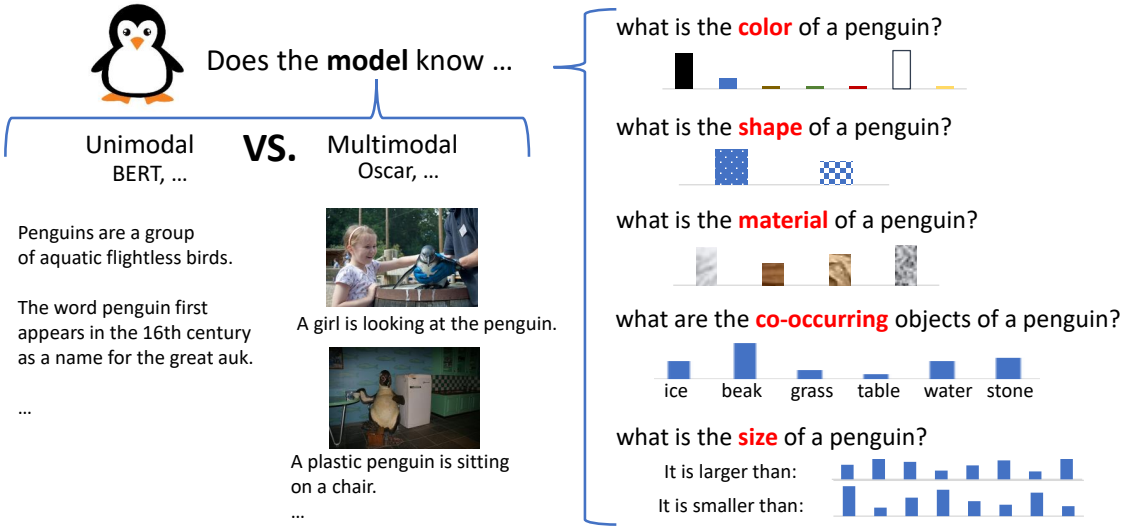

Figure 1: Illustration of our main idea with input "penguin" as an example. We compare unimodal and multimodal models in terms of their ability to capture commonsense knowledge. The commonsense knowledge is evaluated by five relation types: color, shape, material, size, and visual co-occurrence. To do evaluation, we compare the model outputs with the gold distribution, which is mined from Visual Genome.

image and text representations from cross-modal training of transformers with self-attention, including LXMERT (Tan and Bansal, 2019), ViLBERT (Lu et al., 2019), UNITER (Chen et al., 2020), etc. Oscar (Li et al., 2020) additionally uses object tags in images as anchor points to ease the learning of image-text alignments and VinVL (Zhang et al., 2021) presents an improved object detection model. CLIP (Radford et al., 2021) learns by predicting caption-image alignment from a large internet corpus of (image, text) pairs.

While our work uses textual prompt tuning techniques, there have also been work on visual prompt engineering to enhance the performance of pretrained vision-language models. Zhou et al. (2021) model context in prompts as continuous representations and learn to optimize that context. Yao et al. (2021) develop a cross-modal prompt tuning framework that reformulates visual grounding as a fill-in-the-blank problem for both image and text.

2.2 Visual Commonsense

In one of the early attempts at learning visual commonsense, Vedantam et al. (2015) measure the plausibility of a commonsense assertion in the form of (obj1, relation, obj2) based on how similar it is to known plausible assertions, using both visual scenes and accompanying text. Zellers et al. (2021) learn physical commonsense through interaction, and uses this knowledge to ground language. Frank et al. (2021) probe whether vision-language models have learned to construct cross-modal representations using both modalities via cross-modal input ablation.

Note that our definition of visual commonsense differs from that of Zellers et al. (2019), where the model is required to perform commonsense reasoning based on an image. Our idea of visual commonsense is more similar to the idea of stereotypic tacit assumptions (Prince, 1978) – the propositional beliefs that humans hold about generic concepts, such as "dogs have to be walked." Weir et al. (2020) probe neural language models for such human tacit assumptions and demonstrate the models' success. We extend this intuition to visual concepts and explore how visual information may help language models to capture such assumptions.

Zhu et al. (2020) investigate the "language prior" problem in Visual Question Answering models, where models tend to answer questions based on word frequencies in the data and ignore the image contents. In this work, we explore to what extent such language prior is correct when there is no image input.

2.3 Reporting Bias

Pretrained language models such as BERT (Devlin et al., 2019) are trained on millions of tokens of text, capturing statistical regularities present in the training corpora. However, their textual training data can suffer from reporting bias, where the frequency distribution of specific events and properties in text

may not reflect the real-world distribution of such properties (Gordon and Van Durme, 2013). For example, while grass is most commonly green, this may not be reported as much in web corpora, and while motorcycle crashes may be more common in the real world, plane crashes may be mentioned far more in news text.

Misra et al. (2016) shows that "human-centric" image annotations contain reporting bias as well and that the noise in annotations exhibits structure and can be modeled.

3 Datasets

3.1 Dataset Mining

Our data take the form of (subject, object) tuples for each relation, with the goal being to predict the object (and its distribution) from the subject and relation. Relations include color, shape, material, size, and object co-occurrence. Table 1 summarizes the number of classes and subject-object pairs for each relation. [1]

Color, Shape, Material For color, shape, and material, the subject is a noun and the object is the color, shape, or material property of the noun, mined from Visual Genome (VG) (Krishna et al., 2016) [2] attributes. We manually create a list of single-word attributes for each relation, and only VG subjects that are matched with a specific attribute for more than a threshold number of times are recorded, in order to avoid noise in the dataset. The thresholds for color, material, and shape are 5, 2, and 1, respectively, chosen based on the availability of attributes of each relation in VG. VG attributes are filtered with the following steps: (1) attribute "Y colored / made / shaped" is treated as "Y"; (2) select only the last word for compound attributes (e.x. treat "forest green" as "green"); (3) similar attributes are merged into one of the main attribute classes (e.x. "maroon" and "crimson" are merged into "red").

The above procedure produces a distribution over the set of attributes for each subject noun. From that distribution, a (subject, object) data instance is generated for each subject where the object is the attribute that associates with it the most. See the first three rows of Table 1 for examples.

[1] See Appendix A.1 for more information on the object classes.

[2] Licensed under CC-BY 4.0.

Size Size is separated into `size_smaller` and `size_larger`, where the subject is a noun and the object is another noun that is smaller or larger, respectively, than the subject. To form the size dataset, we obtain a set of concrete nouns which we classify into 5 size categories (`tiny`, `small`, `medium`, `large`, and `huge`). We randomly pick two nouns from different categories to form a (subject, object) pair.

Visual Co-occurrence The visual co-occurrence dataset is generated in a way similar to that of the color, shape, and material datasets. The (subject, object) tuple here contains two nouns corresponding to objects that may appear in the same context. Co-occurrence distribution is extracted from Visual Genome where two objects that occur in the same scene graph together for more than 8 times are recorded.

3.2 Data Grouping

Following Paik et al. (2021), we split the color, shape, and material datasets each into three groups: Single, Multi, and Any. The Single group is for subjects whose most common attribute covers more than 80% of the probability, e.g., the color of *snow* is almost always white. The Multi group is defined as subjects not in the Single group where more than 90% of the probability falls in the top 4 attribute classes, e.g., the color of a penguin in Fig. 1. The rest of the subjects are in the Any group. Lower model performance for the Single group would indicate the influence of reporting bias.

3.3 Templates

In order to elicit model response and extract target objects and distributions from text, we manually design a set of templates for each relation. There are 7 templates for color, shape, and material each, 8 for size, and 4 for visual co-occurrence. See Table 1 for example templates.

3.4 Wikipedia Data

In order to compare the text-only and multimodal datasets, we mine the color, shape, and material datasets from Wikipedia data, which is typically used in model pretraining. To mine these text-based datasets, we combine the sets of subjects in VG, take the manual list of attributes as objects again, and extract (subject, object) pairs if the pair matches any of the pre-defined templates. In Section 3.5 we will show the advantages of the VG

Relation	# Classes	# (subj, obj) Pairs	Ex Template	Ex (subj, obj) Pair
color	12	2877	[subj] *can be of color* [obj]	(*sky*, *blue*)
shape	12	706	[subj] *has shape* [obj] .	(*egg*, *oval*)
material	18	1423	[subj] *is made of* [obj] .	(*sofa*, *cloth*)
size (smaller)	107	2000	[subj] *is smaller than* [obj] .	(*book*, *elephant*)
size (larger)	107	2000	[subj] *is larger than* [obj] .	(*face*, *spoon*)
co-occurrence	5939	2108	[subj] *co-occurs with* [obj] .	(*fence*, *horse*)

Table 1: Summary of the dataset mined from Visual Genome and manual templates, including the number of classes, (subject, object) pairs, and examples for each relation.

Source	Group	Spearman ρ	# Subjs	Avg # Occ	Top5 # Occ	Btm5 # Occ	Acc@1
VG	All	64.3 ± 23.9	355	1252.6	64.6	308.6	
	Single	62.2 ± 24.0	131	494.9	64.6	1181.6	80.2
	Multi	69.3 ± 20.7	136	1156.1	2062.2	347.0	
	Any	58.4 ± 27.1	88	2529.6	8452.4	1213.4	
Wikipedia	All	33.4 ± 30.6	302	543.6	1758.0	49.8	
	Single	29.6 ± 29.9	110	352.2	345.8	35.0	35.5
	Multi	33.9 ± 30.9	119	500.8	1242.0	27.6	
	Any	38.2 ± 30.4	73	902.0	3000.2	161.2	

Table 2: Evaluation of the VG-mined and Wikipedia-mined color datasets by comparing with the human-annotated dataset CoDa. Reported are the average Spearman correlation ($\times 100$), number of common subjects, average number of occurrences of the common subjects, average number of occurrences of subjects with top- and bottom-5 Spearman correlations, and the percentage of top-1 attributes being matched for the single group. VG has higher correlations with human annotations.

dataset over this text-based data.

3.5 Dataset Evaluation

To ensure the validity of the datasets mined from Visual Genome, we compare our color dataset with the human annotated CoDa dataset (Paik et al., 2021), which we assume is close to real-world color distributions and has minimal reporting bias. We see a reasonably strong correlation with human annotations, indicating that our dataset is a good and cost-effective approximation to human annotations.

Metrics We report the Spearman's rank-order correlation between the two distributions in comparison, averaged across all subjects. The Spearman correlation is used instead of the Pearson correlation since we care more about the rank of the object distributions than the exact values, which may be variable due to data variability. The top-1 accuracy (Acc@1) is measured by the percentage of the objects with the highest probability in the source distribution matching those in the target distribution. Those two metrics are also used in later sections when evaluating model distributions.

Analysis Table 2 shows the detailed results of the evaluation of the VG and Wikipedia color datasets by comparing with the human-annotated dataset, CoDa. We can see that the VG dataset has much higher Spearman correlation with CoDa, as well as substantially higher top-1 accuracy for the Single group. The VG correlation is expected to be low for the Any group, because objects in the Any group can have many possible colors.

Reporting bias is present in both datasets, as the average number of occurrences of Single group subjects are much fewer than that of the Multi and Any group subjects. Counter-intuitively, for VG, the highly-correlated Single group subjects have fewer average occurrences than the ones with low correlations. This is contrary to our expectation that more frequent objects would better reflect the human-perceived distribution and can be explained by Single subjects being easier to represent even without a large amount of data.

One example where the Wikipedia distribution diverges from the CoDa distribution is "penguin," whose most likely color in CoDa is black, followed by white and gray; however, its top color in Wikipedia is blue, because "blue penguin" is a specific species with an entry in Wikipedia, even if it is not as common as black and white penguins. One example where the VG distributions diverge from CoDa is "mouse," because in VG, most occurrences of "mouse" are computer mice, which are most commonly black, whereas when asked about the word "mouse", human annotators typically think about the animal, meaning that the most likely colors in CoDa are white and gray.[3]

[3] Additional examples are provided in Appendix A.3.

4 Probing Visual Commonsense

4.1 Models

Following Paik et al. (2021), we apply zero-shot probes to models that are trained on a language modeling objective, and conduct representation probes for those that are not. We report the prediction accuracy and the Spearman correlation of the output distribution with the true distribution.

We examine 6 transformer-based models, trained on a variety of data. BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), and RoBERTa (Liu et al., 2019) are trained on text only using a masked language modeling objective (MLM). Oscar (Li et al., 2020) is a vision-language model based on the BERT architecture, trained with an combined MLM and contrastive loss on text-image pairs. As our experiments involve exclusively textual inputs, we later describe a knowledge-distilled version of Oscar ("Distilled") which corrects for the lack of image input in our task. Finally, we use representations from CLIP (ViT-B/32) (Radford et al., 2021), which is trained with a contrastive image-caption matching loss.

We use models trained with an MLM objective (BERT, Distilled, etc) directly for zero-shot prediction of masked tokens[4]. For Oscar we add a word-prediction head on top of it. The results across templates are aggregated in two modes. In the "best template" mode, for each example, the highest Spearman correlation among all templates is reported, and the top-1 result is regarded as correct if the true target object is the same as the top-1 result of any of the templates. In the "average template" mode, the output distribution is the mean of the distributions across all templates.

Since CLIP is not trained on a token-prediction objective, we implement logistic regression on top of the frozen encoder output, to predict the target attribute or object. The input is each of the templates with the subject [X] filled with an input in the dataset. Like Paik et al. (2021), to give the model ample chance of success, we take the template that results in the best test accuracy score, report that accuracy and the Spearman correlation associated with that template.

Vokenization Tan and Bansal (2020) introduce the "vokenization" method, which aligns language tokens to their related images, mitigating the shortcomings of models trained on visually-grounded datasets in text-only tasks. Since our task is purely text-based, we also experiment with a pretrained vokenization model (BERT + VLM on Wiki).

[4]For the target words that contain more than one subword tokens, we use the first token as the target.

4.2 Elicitation Methods

We compare the visual commonsense abilities of pretrained unimodal and multimodal models. Given a list of prompts and a subject word, each model outputs the distribution of the target word.

Soft prompt tuning In order to overcome the limitation of self-designed prompts, we incorporate prompt tuning techniques in Qin and Eisner (2021), which learns soft prompts by gradient descent. The algorithm minimizes the log loss:

$$\sum_{(x,y)\in E_r} -\log \sum_{\mathbf{t}\in T_r} p(y|\mathbf{t}, x)$$

for a set of example pairs E_r and template set T_r.

Knowledge distillation Through preliminary experiments, we notice, as expected, that pretrained Oscar, even without visual input, achieved better results than BERT. This led us to consider knowledge distillation (Hinton et al., 2015; Sanh et al., 2019). We use Oscar as the teacher and BERT as the student, and the weights of the student are adjusted to simulate the output distribution of the teacher. The training data is part of the Oscar pretraining corpus: COCO (Lin et al., 2014), Flickr30k (Young et al., 2014), and GQA (Hudson and Manning, 2019).

4.3 Size Evaluation

We use two evaluation strategies for size, because the size dataset differs from the other datasets in that we use relative sizes (X is larger/smaller than Y), as absolute size information is hard to obtain.

Rank partition First, similar to the previous prediction task, given a template such as "[X] is larger than [Y]" and an object [X], we ask the model to predict the distribution of [Y], taking only the distribution D of nouns in the size dataset. For the current object [X], we take the nouns in size categories that are smaller than the category of [X] (N_{sm}), and those that are in larger categories (N_{lg}). Let the length of N_{sm} be m and the length of N_{lg} be n. Then for the "larger" templates, we compute the average percentage of overlap between the top n objects in D and N_{lg} and that between the bottom m objects in D and and N_{sm}. For the "smaller" templates, the "top" and "bottom" are reversed.

		Color		Shape		Material		Cooccur
Tune	**Model**	**Spearman ρ**	**Acc@1**	**Spearman ρ**	**Acc@1**	**Spearman ρ**	**Acc@1**	**Spearman ρ**
N	BERT_b	26.1 ± 31.0*	11.7	38.7 ± 15.1	6.7	33.7 ± 19.6	30.0	4.7 ± 3.5
	Oscar_b	26.4 ± 30.7*	24.0	45.9 ± 14.1	**53.0**	38.6 ± 17.5	**43.3**	9.8 ± 6.9
	Distilled	34.8 ± 27.3	27.5	$\mathbf{46.2 \pm 14.2}$	37.3	36.1 ± 20.2	37.7	$\mathbf{10.1 \pm 7.5}$
	BERT_l	$\mathbf{37.6 \pm 30.3}$	**30.3**	42.7 ± 17.1	28.4	36.6 ± 19.1	35.7	5.2 ± 3.8
	Oscar_l	31.8 ± 28.3	17.1	40.0 ± 16.9	38.1	$\mathbf{39.2 \pm 17.1}$	40.5	9.7 ± 6.7
Y	BERT_b	48.0 ± 22.9	47.4	49.2 ± 12.7*	76.1	41.2 ± 15.3	45.2	11.3 ± 7.9
	Oscar_b	$\mathbf{58.1 \pm 21.1}$	**67.9**	50.4 ± 11.5*	81.3	45.3 ± 14.3	**66.2**	12.7 ± 9.3
	Distilled	57.1 ± 21.9	64.6	$\mathbf{50.5 \pm 12.3}$	**82.8**	$\mathbf{45.4 \pm 14.8}$	**66.2**	$\mathbf{13.0 \pm 10.1}$
	BERT_l	37.6 ± 30.3	30.3	49.2 ± 12.6	78.4	43.7 ± 15.1	53.3	11.4 ± 8.0
	Oscar_l	57.6 ± 21.6	65.3	50.1 ± 12.2	81.3	45.2 ± 15.2	65.8	12.8 ± 9.6

Table 3: Spearman correlation and top-1 accuracy (both $\times$ 100) of zero shot probing, before and after soft prompt tuning ("N" and "Y" for the "Tune" column). This is the "average template" case where the output distribution is the mean of distributions across all templates. The Spearman correlation reported is the mean across all subjects $\pm$ standard deviation, comparing the output distribution and the Visual Genome distribution. The subscripts b and l indicate the size of the model, and Distilled is the BERT model after distilling from Oscar. Asterisk indicates where there is no significant difference between BERT_b and Oscar_b (t-test p-value > 0.05).

Adjective projection The second approach follows that of van Paridon et al. (2021), which projects the word to be evaluated onto an adjective scale. In this case, we compute the word embeddings of the adjectives "small" and "large" and the nouns from models, so the scale is $\overrightarrow{\text{large}} - \overrightarrow{\text{small}}$ and the projection is calculated by cosine similarity. For instance, for the example noun "bear", the projection score is given by:

$$cos\_sim(\overrightarrow{\text{large}} - \overrightarrow{\text{small}}, \overrightarrow{\text{bear}})$$

With good word embeddings, larger nouns are expected to have higher projection scores. The validity of the adjective scales from word representations is shown by Kim and de Marneffe (2013).

5 Experiments

5.1 Implementation Details

Dataset splits Each of the color, shape, material, size, and co-occurrence datasets is split into 80% training data and 20% test data. All evaluation metrics are reported on the test set. The training set is used for the logistic regression and the soft prompt tuning algorithm.

Model training For the classification head, we use the sklearn implementation of Logistic Regression (random_state=0, C=0.316, max_iter=2000). For soft prompt tuning, we use the implementation from Qin and Eisner (2021) [5]. For knowledge distillation, we use the Kullback-Leibler loss to measure the divergence between the output distributions of BERT and Oscar, and optimize the pretrained BERT on that loss to match the outputs of Oscar. Configurable parameters are set the same as for Oscar pretraining.

[5] https://github.com/hiaoxui/soft-prompts

5.2 Results

The experimental results show that multimodal models outperform text-only models in capturing visual commonsense. However, all models are subject to the influence of reporting bias, as they correlate better with the distributions from Wikipedia than those from VG. Prompt tuning and knowledge distillation substantially enhance model performance, while increasing model size does not.

Color, Shape, Material The resulting model performance for the "average template" mode is shown in Table 3. Prompt tuning is done in this mode only. Note that because the top-1 accuracy is taken among all possible classes of each relation, it should be interpreted together with the number of classes (Table 1).

We can see from Table 3 that Oscar does better than BERT in almost all cases. Significant difference between Oscar (base) and BERT (base) is seen in most cases. Also, after soft prompt tuning, both the Spearman correlation and the accuracy substantially improved. Although the standard deviations of the Spearman correlations are large, we find consistent improvement per example with both prompt tuning and multimodal pretraining (Appendix A.2).

Table 3 also shows that knowledge distillation helps improve the performance of BERT in all cases, and the distilled model can sometimes even outperform the teacher model, Oscar. Moreover, the large version of each model does not necessarily perform better than their base counterparts,

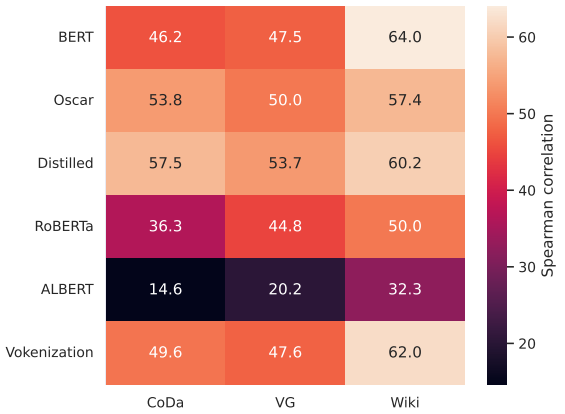

Figure 2: Spearman correlations ($\times 100$) for color, under the "best template" case, for base models on CoDa, VG, and Wikipedia. While all models correlate the best with Wikipedia, BERT is the most biased.

suggesting that increasing the size of the model would not enhance the model's ability to understand visual commonsense. Instead, training with visually grounded data would.

Fig. 2 illustrates the Spearman correlations of different models with the color distributions from CoDa, VG and Wikipedia, under the "best template" mode.[6] We assume that CoDa contains no reporting bias, in which case we can interpret Table 2 as showing that VG contains a relatively small amount of it, and Wikipedia contains a relatively large amount. All models correlate moderately with all three datasets, with the highest correlations to Wikipedia, indicating text-based reporting bias in all model types. BERT has the largest correlation gap between Wikipedia and CoDa.

Visual Co-occurrence Table 3 also contains the results on visual co-occurrence before and after prompt tuning. Only the Spearman correlations are reported, because the top-1 accuracy is meaningless due to the large number of possible co-occurring objects with any noun.

Before prompt tuning, BERT has small Spearman correlations, suggesting that it may contain little knowledge about the visual co-occurrence relationship. Oscar demonstrates more such knowledge under the zero-shot setting. After prompt tuning, all model performances improve.

Size Table 4 shows results of the rank partition method (Section 4.3), before and after prompt tuning. Surprisingly, prompt tuning does not help in this case. Moreover, the performance for the

[6] Appendix A.2 contains further details.

Tune	Model	Larger	Smaller
N	BERT_b	80.0	67.1
	Oscar_b	79.5	67.7
	Distilled	**84.6**	60.7
	BERT_l	80.9	66.1
	Oscar_l	79.4	**70.7**
Y	BERT_b	69.9	55.7
	Oscar_b	70.6	57.3
	Distilled	70.6	57.3
	BERT_l	70.0	55.7
	Oscar_l	70.6	57.3

Table 4: Percent correct for size relation, for "larger" and "smaller" templates, before and after soft prompt tuning. Interestingly, tuning does not help with size.

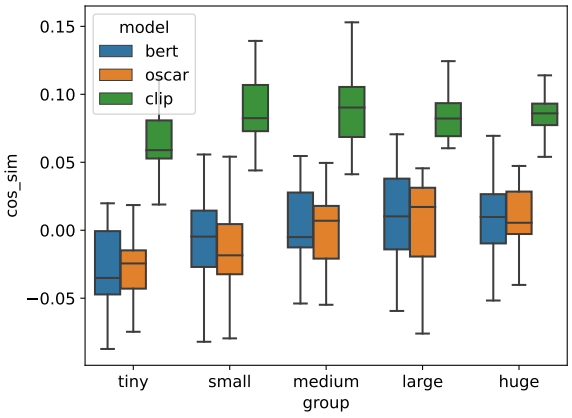

Figure 3: The size projection scores, where the x-axis indicates the object groups. Outliers are omitted. All three models perform reasonably well, as larger objects have higher cosine similarities in general.

"larger" templates is higher than that of the "smaller" templates, suggesting that the models contain inherent preference towards the "larger" templates.

Fig. 3 shows the results of the adjective projection method.[7] For BERT and Oscar, we use the average embedding of the subword tokens of the nouns projected onto that of the adjectives "large" and "small". For CLIP, we take the textual encoder outputs as the embeddings, resulting in a different score range from that of BERT and Oscar. The results show the following trend: larger objects are projected onto the "large" end of the spectrum, although the trend is sometimes broken towards the "huge" end. This may be due to the "huge" group including nouns such as "pool" and "house" which can be modified by a relative size indicator "small".

5.3 Results with Classification Head

Table 5 shows the results of BERT, CLIP, and Oscar when topped with a classification head. We observe that Oscar and CLIP achieve similar per-

[7] Appendix A.2 contains per-object plot for BERT vs Oscar.

	Color		Shape		Material		Co-occur
Model	**Spearman ρ**	**Acc@1**	**Spearman ρ**	**Acc@1**	**Spearman ρ**	**Acc@1**	**Spearman ρ**
BERT_b	48.0 ± 21.6	51.4	53.2 ± 13.4	78.4	41.3 ± 15.6	51.1	30.2 ± 11.7
Oscar_b	$\mathbf{52.5 \pm 20.8}$	63.1	54.4 ± 14.8	**80.6**	$\mathbf{43.2 \pm 14.4}$	**63.0**	31.2 ± 12.1
CLIP	51.9 ± 20.8	**63.8**	$\mathbf{54.5 \pm 13.9}$	79.9	42.9 ± 15.0	**63.0**	$\mathbf{31.3 \pm 11.6}$

Table 5: Spearman correlation and top-1 accuracy (both $\times$ 100) with a logistic regression head on model encoder outputs. Oscar and CLIP have comparable performance, both slightly better than BERT.

		Color		Shape		Material	
Group	**Model**	**Spearman ρ**	**Acc@1**	**Spearman ρ**	**Acc@1**	**Spearman ρ**	**Acc@1**
Single	BERT_b	36.8 ± 19.0	54.8	48.3 ± 12.3	83.0	35.9 ± 14.3	51.6
	Oscar_b	39.9 ± 15.3	60.3	$\mathbf{49.3 \pm 11.6}$	87.0	$\mathbf{38.5 \pm 12.8}$	**65.1**
	CLIP	$\mathbf{41.0 \pm 15.2}$	**66.3**	49.2 ± 14.5	**90.0**	38.1 ± 12.8	64.1
Multi	BERT_b	49.7 ± 21.2	42.3	$\mathbf{65.9 \pm 16.9}$	59.5	53.8 ± 16.2	51.3
	Oscar_b	$\mathbf{51.2 \pm 19.9}$	50.6	65.2 ± 17.4	64.9	$\mathbf{56.2 \pm 13.0}$	53.9
	CLIP	50.5 ± 21.1	**55.4**	64.6 ± 18.9	**67.6**	56.2 ± 14.3	**59.2**
Any	BERT_b	56.5 ± 19.5	46.1	100.0 ± 0	–	58.7 ± 15.2	**35.7**
	Oscar_b	$\mathbf{62.5 \pm 18.9}$	**58.4**	100.0 ± 0	–	60.4 ± 17.1	**35.7**
	CLIP	60.3 ± 18.2	55.8	100.0 ± 0	–	$\mathbf{63.5 \pm 20.5}$	21.4

Table 6: Per-group Spearman correlation and top-1 accuracy (both $\times$ 100) with a logistic regression head on model encoder outputs. Note that the Any group for shape only has one example, so the accuracy is less meaningful and is omitted. All models have higher correlations in the Multi and Any groups than the Single group, which is a sign of reporting bias.

formance and are both better than BERT. Note that, while Visual Genome is part of Oscar's pretraining corpus and one might suspect that that gives it an advantage, CLIP is trained on a large corpus from web search that is unrelated to Visual Genome. Therefore, we can conclude that multimodal models pretrained on both images and text outperform text-only models.

Table 6 breaks down the results in Table 5 into three subject groups. Oscar and CLIP outperform BERT in almost all cases. The top-1 accuracy is higher for the Single group than for the Multi and Any groups, perhaps because the Single group subjects have only one most likely target attribute, which may be easier to predict. Note that the Spearman correlations for all three models become higher from group Single to Multi to Any. Paik et al. (2021) argue that higher correlation for the Any and Multi groups is a sign of model reporting bias, as objects in those two groups are more often reported. Thus, the results here indicate that reporting bias is still present in multimodal models.

5.4 Analysis and Limitations

In Table 3, the accuracy of BERT for shape is particularly low (only 6.7%), despite that shape has only 12 classes. We hypothesize that this is due to reporting bias on shape in the text corpora that BERT is trained on. This hypothesis is supported by mining sentences from Wikipedia that contain (noun, attribute) pairs, where we see that the relation shape has fewer number of occurrences than material and color (see Appendix A.3).

Finally, although multimodal models show improvement on the task, the improvement is sometimes not significant and the resulting correlations are still weak. Further work is needed to enhance the visual commonsense abilities of the models and mitigate reporting bias, and our datasets can serve as an evaluation method.

6 Conclusion

In this paper, we probe knowledge about visually salient properties from pretrained neural networks. We automatically extract dataset of five visual relations–color, shape, material, size, and co-occurrence, and show that it has much higher correlation with human perception for color than data mined from text copora. Then, we apply various types of probing techniques and discover that visually-supervised models can better capture such visual properties than pure language models. Knowledge distillation can sometimes further enhance model performance. Despite their higher performance, visually-supervised models are still subject to the influence of reporting bias, as shown by the per-group analysis, where both types of models perform better for the Multi group than the Single group.

References

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

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.

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)*.

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.

Stella Frank, Emanuele Bugliarello, and Desmond Elliott. 2021. Vision-and-language or vision-for-language? on cross-modal influence in multimodal transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for compositional question answering over real-world images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (ICCV)*, page 6700–6709.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625–1630, Seattle, Washington, USA. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting language and vision using crowdsourced dense image annotations.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations (ICLR)*.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision (ECCV)*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, 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 (ECCV)*.

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.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Ishan Misra, C. Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2930–2939.

Cory Paik, Stéphane Aroca-Ouellette, Alessandro Roncone, and Katharina Kann. 2021. The world of an octopus: How reporting bias influences a language model's perception of color. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 823–835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jeroen van Paridon, Qiawen Liu, and Gary Lupyan. 2021. How do blind people know that blue is cold? distributional semantics encode color-adjective associations. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Ellen F. Prince. 1978. On the function of existential presupposition in discourse. In *Chicago Linguistic Society (Vol. 14, pp. 362–376)*.

Guanghui Qin and Jason Eisner. 2021. Learning how to ask: Querying lms with mixtures of soft prompts. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

page 5203–5212, Online. Association for Computational Linguistics.

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. In *International Conference on Machine Learning (ICML)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

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 (EMNLP)*.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2066–2080, Online. Association for Computational Linguistics.

Ramakrishna Vedantam, Xiaoyu Lin, Tanmay Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Learning common sense through visual abstraction. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2542–2550.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. On the existence of tacit assumptions in contextualized language models. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Yuan Yao, Ao Zhang, Zhengyan Zhang, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2021. CPT: colorful prompt tuning for pre-trained vision-language models. *ArXiv*, abs/2109.11797.

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. In *Transactions of the Association for Computational Linguistics*, volume 2, pages 67–78.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rowan Zellers, Ari Holtzman, Matthew Peters, Roozbeh Mottaghi, Aniruddha Kembhavi, Ali Farhadi, and Yejin Choi. 2021. PIGLeT: Language grounding through neuro-symbolic interaction in a 3D world. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2040–2050.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Making visual representations matter in vision-language models. In *2021 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2021. Learning to prompt for vision-language models. *arXiv preprint arXiv:2109.01134*.

Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. 2020. Overcoming language priors with self-supervised learning for visual question answering. In *International Joint Conference on Artificial Intelligence (IJCAI)*, page 1083–1089.

A Appendix

A.1 List of Objects

Table 7 shows the list of all possible attributes for relations color, shape, and material. Table 8 shows the list of objects in the five categories of relation size. Visual co-ocurrence has a large number of objects that are not listed here for space reasons.

Relation	Classes
Color	**black**, **blue** (aqua, azure, cyan, indigo, navy), **brown** (khaki, tan), **gray** (grey), **green** (turquoise), **orange** (amber), **pink** (magenta), **purple** (lavender, violet), **red** (burgundy, crimson, maroon, scarlet), **silver**, **white** (beige), **yellow** (blond, gold, golden)
Shape	**cross**, **heart**, **octagon**, **oval**, **polygon** (heptagon, hexagon, pentagon), **rectangle**, **rhombus** (diamond), **round** (circle), **semicircle**, **square**, **star**, **triangle**
Material	**bronze** (copper), **ceramic**, **cloth**, **concrete**, **cotton**, **denim**, **glass**, **gold**, **iron**, **jade**, **leather**, **metal**, **paper**, **plastic**, **rubber**, **stone** (cobblestone, slate), **tin** (pewter), **wood** (wooden)

Table 7: List of all objects for relation color, shape, and material. Inside the parentheses are the attributes that are grouped into the object class.

Size	Objects
Tiny	ant, leaf, earring, candle, lip, ear, eye, nose, pebble, shrimp, pendant, spoon, dirt, pill, bee
Small	bird, tomato, pizza, purse, bowl, cup, mug, tape, plate, potato, bottle, faucet, pot, knob, dish, book, laptop, menu, flower, pillow, clock, teapot, lobster, duck, balloon, helmet, hand, face, lemon, microphone, foot, towel, shoe
Medium	human, door, dog, cat, window, lamp, chair, tire, tv, table, desk, sink, guitar, bicycle, umbrella, printer, scooter, pumpkin, monitor, bag, coat, vase, deer, horse, kite
Large	elephant, car, tree, suv, pillar, stairway, bed, minivan, fireplace, bus, boat, cheetah, wall, balcony, bear, lion
Huge	building, airplane, plane, clocktower, tower, earth, pool, mountain, sky, road, house, hotel, tank, town, city, dinasour, whale, school

Table 8: List of objects in five size categories.

A.2 Additional Probing

Best template mode Table 9 contains zero-shot results under the "best template" mode, for BERT (base), Oscar (base), BERT distilled from Oscar, RoBERTa (base), ALBERT (base), and Vokenization. These results demonstrate similar trends as the ones in the "average template" mode.

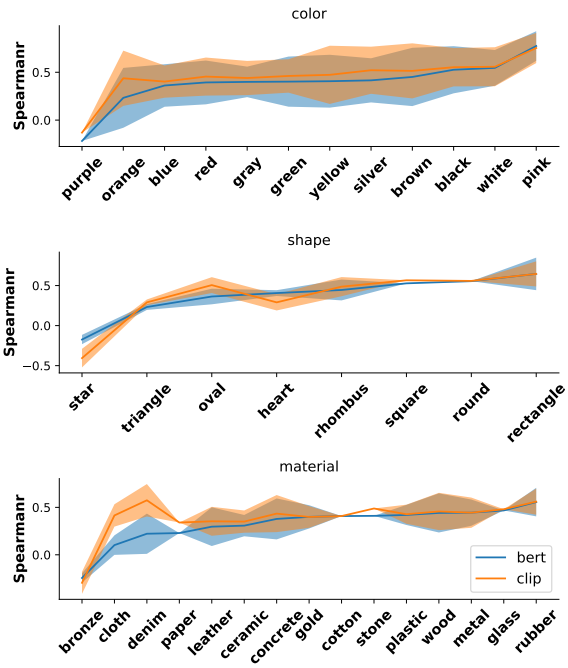

Figure 4: Spearman correlation per object class for BERT and CLIP with the logistic regression head, for color, shape, and material. The error margins are the standard deviations.

Per-object analysis Fig. 4 illustrates the fine-grained Spearman correlation ± standard deviation per object group for BERT and CLIP.

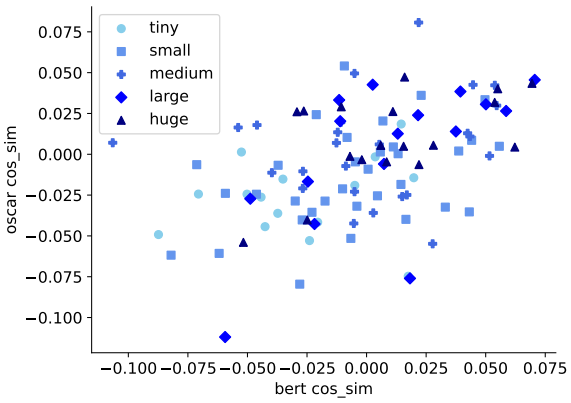

Figure 5: The size projection scores from BERT and Oscar, where each point is one object. Cosine similarities are correlated between Oscar and BERT.

Size per-object Fig. 5 shows how the per-object projection scores on the size spectrum from BERT and Oscar are correlated.

Per-Subject Comparison Fig. 6 and Fig. 7 show how the Spearman correlations of 10 individual subjects improve after soft prompt tuning and after multimodal pretraining. Consistent improvement

	Color		Shape		Material		Cooccur
Model	**Spearman ρ**	**Acc@1**	**Spearman ρ**	**Acc@1**	**Spearman ρ**	**Acc@1**	**Spearman ρ**
BERT$_b$	47.5 ± 21.6	41.8	48.2 ± 12.0	64.3	41.9 ± 15.4	55.3	6.1 ± 4.0
Oscar$_b$	50.0 ± 19.8	**59.8**	**52.7 ± 10.0**	**89.3**	**46.5 ± 13.7**	**74.6**	10.1 ± 7.2
Distilled	**53.7 ± 21.3**	57.7	51.4 ± 11.1	74.3	46.0 ± 13.6	**74.6**	**10.4 ± 7.8**
RoBERTa$_b$	44.8 ± 19.8	41.6	45.4 ± 12.4	69.3	33.0 ± 15.5	39.1	1.1 ± 1.4
ALBERT$_b$	20.2 ± 24.8	13.4	29.8 ± 15.7	13.6	25.0 ± 17.9	27.8	6.6 ± 5.1
Vokenization	47.6 ± 20.9	51.6	49.8 ± 13.1	72.9	39.4 ± 16.0	52.5	6.0 ± 3.7

Table 9: Spearman correlation and top-1 accuracy (both × 100) of zero shot probing. This is the "best template" case discussed in Section 4.1.

can be seen in color, material, and cooccurrence. Although we report average Spearman correlations in Table 3 and there are large standard deviations, here we show that when improvement is observed collectively, it is also consistent across subjects. With shape, the improvement is less obvious (45.9 to 50.4 for prompt tuning and 49.2 to 50.4 for multimodal pretraining).

A.3 Error Analysis

Data The three subjects with the highest and lowest Spearman correlation are shown in Fig. 8 and Fig. 9.

Wikipedia Table 10 shows the number of (noun, attribute) pairs of the three relation types in Wikipedia. Shape has fewer occurrences than material and color.

	Color	Shape	Material
Total	331480	195921	307879
Avg 12	27623.3	16326.8	24634.7

Table 10: First row is the total number of occurrences of (noun, attribute) pairs for relations shape, material, and color in Wikipedia. Second row is the average number of occurrences across the top 12 attributes for each relation. Shape has the fewest number of occurrences.

Model Table 11 shows the errors made by BERT and Oscar in the "average template" mode before prompt tuning. Overall, subjects with low correlation are those that are less often reported in Visual Genome as well as in textual data.

A.4 Resources

BERT, RoBERTa, ALBERT We use the Huggingface implementations of BERT, RoBERTa, and ALBERT.

Oscar See the GitHub repository for the code and pretrained Oscar: https://github.com/microsoft/Oscar.

CLIP We use the CLIP model released by OpenAI: https://github.com/openai/CLIP.

Vokenization See the GitHub repository for the pretrained model: https://github.com/airsplay/vokenization.

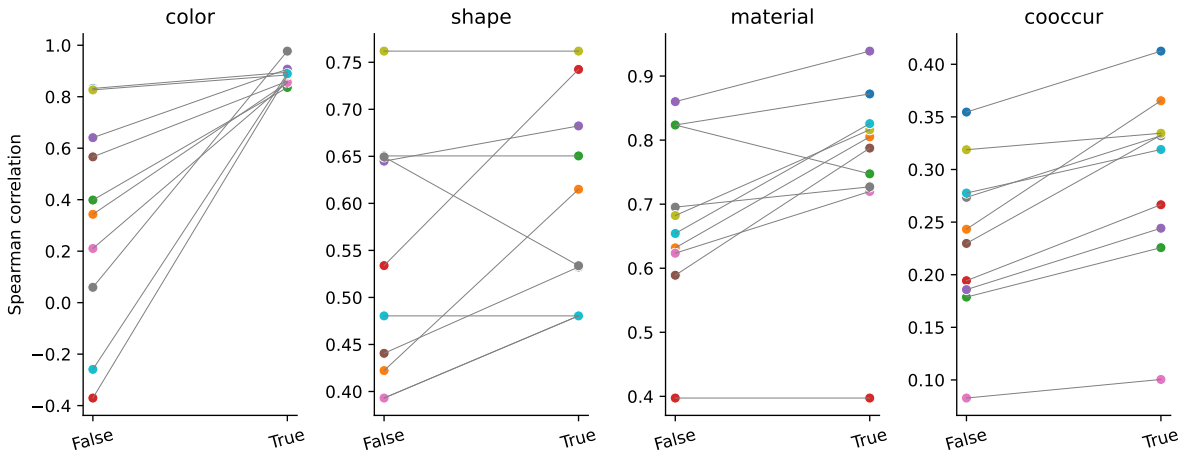

Figure 6: Spearman correlation of 10 subjects for each relation type before and after soft prompt tuning, with Oscar (base). Almost all individual subject has increased correlation after prompt tuning, except in relation shape.

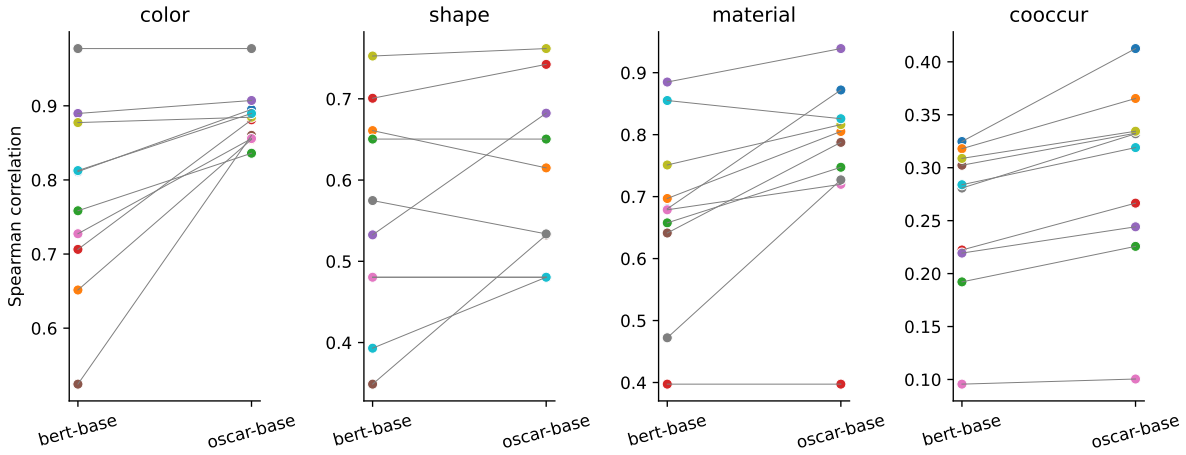

Figure 7: Spearman correlation of 10 subjects for each relation type with BERT (base) and Oscar (base), after soft prompt tuning. Almost all individual subject has higher correlation with Oscar than with BERT, except in relation shape.

	High Corr Subjs		Low Corr Subjs	
Relation	**BERT**$_b$	**Oscar**$_b$	**BERT**$_b$	**Oscar**$_b$
Color	lace, jacket, design	balloon, jacket, apple	flush, water faucet, muffler	hinge, leg, slack
Shape	mirror, vase, container	chair, pizza, vase	connector, log, knot	banana, toast, phone
Material	wall, tray, board	fence, wall, shelf	sheep, fabric, patch	elephant, rug, patch

Table 11: Three subjects each with high and low correlations for relations color, shape, and material.

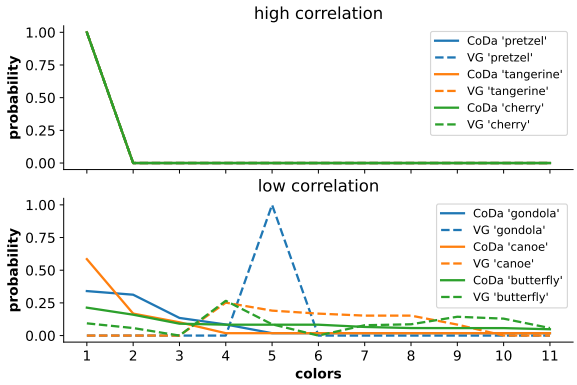

Figure 8: VG vs. CoDa distribution of 3 subjects with the lowest and highest correlation, ordered by probability of colors in CoDa.

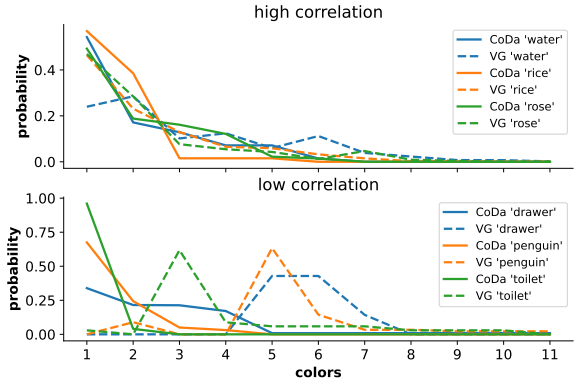

Figure 9: Wikipedia vs. CoDa distribution of 3 subjects with the lowest and highest correlation, ordered by probability of colors in CoDa.