
# What do Models Learn From Training on More Than Text? Measuring Visual Commonsense Knowledge

# What do Models Learn From Training on More Than Text? Measuring Visual Commonsense Knowledge

**Lovisa Hagström**
Chalmers University of Technology
lovhag@chalmers.se

**Richard Johansson**
University of Gothenburg
richard.johansson@gu.se

## Abstract

There are limitations in learning language from text alone. Therefore, recent focus has been on developing multimodal models. However, few benchmarks exist that can measure what language models learn about language from multimodal training. We hypothesize that training on a visual modality should improve on the visual commonsense knowledge in language models. Therefore, we introduce two evaluation tasks for measuring visual commonsense knowledge in language models[1] and use them to evaluate different multimodal models and unimodal baselines. Primarily, we find that the visual commonsense knowledge is not significantly different between the multimodal models and unimodal baseline models trained on visual text data.

## 1 Introduction

Language models (LMs) trained on large amounts of textual data have shown great performance on several textual tasks (Devlin et al., 2019; Brown et al., 2020). However, recent work has illuminated limitations with text-only training of LMs. These limitations arise from a lack of meaning (Bender and Koller, 2020) and experience (Bisk et al., 2020), together with the problem of reporting bias (Gordon and Van Durme, 2013). Multimodal training has been identified as one way to create models that do not suffer from the aforementioned limitations (Paik et al., 2021; Huang et al., 2021). While several multimodal models have been developed (Tan and Bansal, 2019; Li et al., 2019, 2020), few evaluation methods exist that can tell us whether multimodal training mitigates text-only training limits.

If we wish to successfully create multimodal LMs that learn from more than text, we need a way to evaluate them for what we expect them to have learned from their multimodal training.

[1]Code publicly available at: github.com/lovhag/measure-visual-commonsense-knowledge

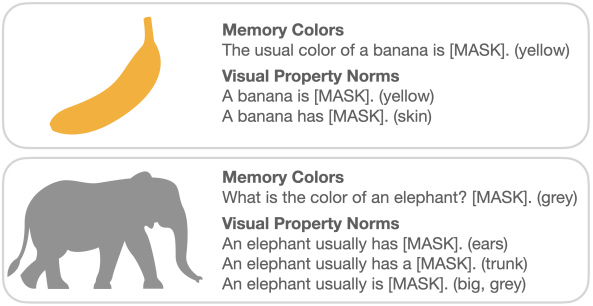


Figure 1: We introduce the two evaluation tasks Memory Colors and Visual Property Norms for measuring visual commonsense knowledge in a LM.

One hypothesis is that multimodal training should aid LMs in learning commonsense knowledge (Zhang et al., 2021). There are several text-only evaluation tasks that aim to measure the commonsense knowledge in LMs (Zellers et al., 2019b; Zhou et al., 2020), but none of them focus explicitly on the commonsense knowledge learned through training on more than text.

In this work, we focus on models trained on images and text, denoted *vision-and-language models*. We reason that if there is any additional information to be learnt from a visual modality it should firstly be basic visual commonsense knowledge. That is, visual conceptual knowledge that is viewed as commonsense by humans, and thus not attainable from text alone due to reporting bias.

We propose a simple method for measuring the visual commonsense knowledge of a model using two zero-shot masked language text-only tasks, depicted in Figure 1. The first task is the Memory Colors evaluation task (Norlund et al., 2021) and the second we create based on the visual features in the Centre for Speech, Language and the Brain (CSLB) concept property norms dataset (Devereux et al., 2014). We refer to the latter task as the Visual Property Norms evaluation task. We complement our work with the results of four vision-and-language models and four baselines on these two tasks.



*Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*
*Student Research Workshop*, pages 252 - 261

## 2 Evaluation Tasks

Our aim is to evaluate models for visual commonsense knowledge. To do this we make use of the existing Memory Colors evaluation task described in section 2.1, and introduce a new evaluation task, Visual Property Norms in section 2.2. Memory Colors is smaller than Visual Property Norms and specifically focuses on visual information related to the color of different concepts, so it is potentially easier. We include both tasks to get a performance curve over increasing difficulty.

Common for both tasks is that they contain queries in English relating to visual properties of tangible concepts and that these queries are based on the knowledge of multiple human participants. Therefore, the tasks can be considered to evaluate a basic aspect of visual commonsense knowledge.

Also common for both tasks is that they use textual templates containing a `[MASK]` token to be predicted by a model in a cloze-style fashion, similarly to the method used by Kassner and Schütze (2020) and Petroni et al. (2019). The advantages with querying the models in this fashion is that most LMs[2] already have been exposed to this type of query format, including most multimodal models. We can then evaluate any model in a masked language modelling fashion on these tasks without additional training or having to make model-specific adaptations, enabling easy evaluation for researchers who wish to use these evaluation tasks.

This form of cloze-style evaluation is also referred to as *prompt-based retrieval*. The reliability of this method has recently been questioned by Jiang et al. (2020) and Cao et al. (2021) due to the query format sensitivity of LMs. To alleviate this issue, we evaluate the models using several different prompts for each of the two tasks.

### 2.1 Memory Colors

The Memory Colors evaluation task is a text-only zero-shot cloze test in English that evaluates a model for its knowledge of memory colors. It queries a model for the color of 109 typical objects using 13 different query templates. The task has been created with the help of 11 human participants, so to some extent it encodes human visual commonsense knowledge limited to colors. Some examples of queries can be seen in Figure 1.

We use the same evaluation metric as specified by Norlund et al. (2021), i.e. the accuracy score after masking the model output for the 11 possible colors black, blue, brown, green, grey, orange, pink, purple, red, white and yellow.

[2]Excluding autoregressive LMs.

### 2.2 Visual Property Norms

We also introduce a new cloze task in English to evaluate for visual commonsense knowledge, denoted Visual Property Norms. It is the largest query-based pure-language evaluation task capable of evaluating LMs for visual commonsense knowledge, containing 6,541 visual conceptual features produced by human participants.

We base it on the CSLB concept property norms dataset (Devereux et al., 2014) that contains the conceptual knowledge of 30 human participants for each of 541 concrete objects, with 123 participants in total. This knowledge is represented as a set of features per object, for which each feature is specified with a production frequency (PF). The PF describes how many of 30 participants produced that feature, so a feature with a high PF can be considered to be more apparent to the participants, since more came to think of it. All features are also categorized as either *encyclopaedic*, *functional*, *other perceptual*, *taxonomic* or *visual perceptual*. Table 1 contains some examples of visual perceptual features in the dataset.

| Concept | Relation | Feature | PF |
|---|---|---|---|
| Cherry | has a | stalk | 17 |
| Fern | is | green | 29 |
| Hair | is | thin | 22 |
| Plum | has | flesh | 9 |

Table 1: Some concepts and their visual perceptual features in the concept property norms dataset.

We create our evaluation task from the concept property norms dataset in a set of steps. Firstly, since our goal is to measure visual commonsense knowledge, we only make use of the *visual perceptual* features. Since we wish to perform cloze tests through masked language modelling, only feature alternatives describable by one wordpiece from the BERT base uncased tokenizer are included.

Furthermore, we only include the four most common feature relations in the task. These are *has*, *has a*, *made of* and *is*. We then part the data into five different segments based on production frequency. This is done by thresholding the features for each concept such that only features with a PF above the set threshold for a certain data segment

are included as gold labels in that segment. The segments and their PF thresholds are listed in the appendix.

Lastly, we create queries from the concepts in each data segment using 8 different query templates, seen in the appendix. Some examples of Visual Property Norms queries can be seen in Figure 1.

Similarly to Weir et al. (2020) we use the mean average precision (mAP) as our evaluation metric, since there may be multiple correct answers for each query in our evaluation data. We calculate this score for each concept and relation, per query template and production frequency segment. We then get a final score for each production frequency segment by taking the average score over all query templates and concepts per segment. This metric is measured over a vocabulary that has been masked to only contain the 614 possible answer alternatives in the Visual Property Norms evaluation data.

## 3 Models

We evaluate four multimodal pre-trained models for their visual commonsense knowledge. These are CLIP-BERT both with and without imagination[3](Norlund et al., 2021), a LXMERT base uncased (Tan and Bansal, 2019) and VisualBERT (Li et al., 2019). We also evaluate four unimodal baseline models. These are a BERT base uncased pretrained on English Wikipedia and BookCorpus, a BERT base uncased further trained on the pure-text part of the CLIP-BERT training data (BERT-CLIP-BERT-train) and two BERT base uncased models trained on the pure-text part of the LXMERT training data, one from scratch and one initialized from pre-trained BERT weights (BERT-LXMERT-train-scratch and BERT-LXMERT-train).

All models are to some extent based on the BERT base architecture and consequently share the same vocabulary and tokenizer. They are also of similar sizes with $\sim$ 110M trainable weights, the exception being LXMERT with $\sim$ 230M trainable weights. Additional information about the models can be found in the appendix.

**Adapting the models for pure-text queries** The majority of current multimodal models have not been developed to be queried only with text. In this case, both CLIP-BERT and VisualBERT should work well with only removing their visual features input, since they are single-stream models. However, LXMERT is a dual-stream model that requires a visual feature input. We handle the removal of visual information by simply removing the visual processing chain in LXMERT, making the language input the only input given to the Cross-Modality Encoder in the model. This would not work if we still wanted to use the model in a multimodal fashion, but we can make this adaption since we are only interested in querying the model for visual commonsense knowledge via language.

[3]The explicit version has the ability to "imagine" visual features when queried with text.

## 4 Results

The results of the models on our two evaluation tasks can be seen in Figure 2. We format the analysis of the results around a set of questions.

**Do the multimodal models display more memory colors knowledge?** The multimodal CLIP-BERT-explicit model has the best performance on this task. So to some extent, yes. But it is worth noting that the unimodal BERT model trained on LXMERT training data is second best on the task, outperforming both LXMERT and VisualBERT, indicating a small multimodal advantage.

**Is performance on Memory Colors indicative of performance on Visual Property Norms?** The ranking visible in Figure 2a does not entirely differ from that in Figure 2b. The main exception being CLIP-BERT-explicit, which has the best performance on Memory Colors, but is outperformed by most other models on Visual Property Norms. We perform a closer analysis of how these results compare by extracting Visual Property Norm results for colors in the appendix.

**Do the models perform better when evaluated on more apparent concept features?** We can observe how the model performance unanimously increases with increased production frequency threshold in Figure 2b. Thus, it appears as though the models agree more with concept features that can be regarded as more apparent.

**Do the multimodal models contain more visual commonsense knowledge?** The results in Figure 2b do not really indicate clear advantage of either unimodal or multimodal models. The multimodal model CLIP-BERT-implicit may generally have the best performance on the task, but the unimodal models trained on visual text data do not differ much in performance. For example, the unimodal BERT-LXMERT-train performs almost on par with CLIP-BERT-implicit.

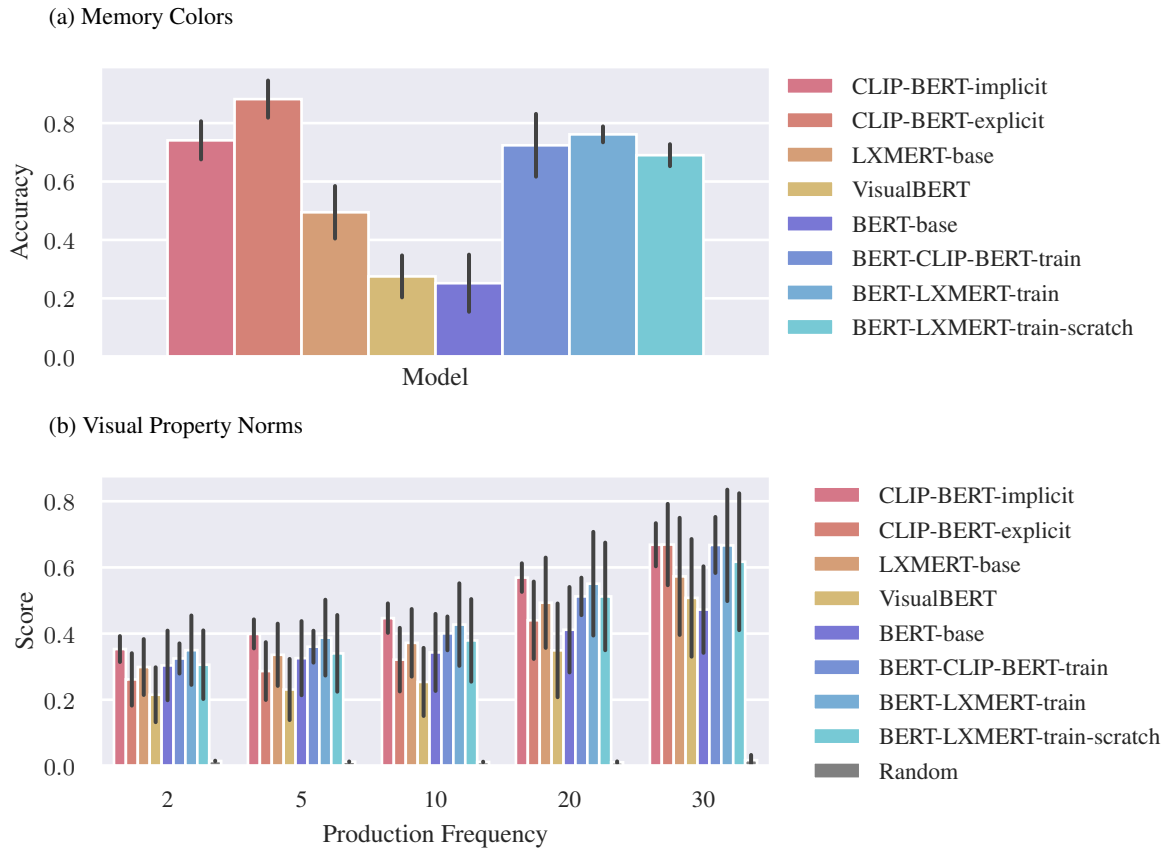


Figure 2: The model accuracy on Memory Colors and model scores on Visual Property Norms per production frequency segment. The multimodal model results are depicted with warmer colors, and the unimodal model results are depicted in cooler colors. The error bars indicate the standard deviation of the model performance over the different query templates. The score has been calculated by masking the vocabulary of the models to only contain the possible answers of the task.

This conclusion is similar to that of Yun et al. (2021), who also compared vision-and-language models to text-only models trained on captions. They found that the models have similar performance with respect to their internal linguistic representations for general tasks.

These results do not mean that the idea of having models learn language from more than text has failed. They do however indicate that there is more work to be done on developing models that use multimodal pretraining to improve on their natural language understanding.

However, we cannot exclude the possibility in our work that the multimodal models suffer in performance due to a lack of visual feature input. Future work investigating this would be valuable.

**Are the models sensitive to how they are queried?** Prevalent for all models is that their performance varies greatly with how they are queried. BERT-LXMERT-train may have the best performance on Visual Property Norms if queried differently. We evaluate the model performances depending on query template in the appendix. This highlights the importance of querying the models with different prompts, since the models may perform dissimilarily depending on prompt due to the degree of prompt-dataset fitness, as reported by Cao et al. (2021).

**Does fine-tuning on visual language develop visual commonsense knowledge?** In both Figures 2a and 2b it is visible that unimodal model performance greatly improves with fine-tuning on visual text corpora. Potential explanations for this are that the models become more attuned to the task with fine-tuning, or that corpora from VQA and image captioning do not suffer as much from reporting bias compared to more common corpora. Thus, text that has been curated to explicitly contain visual information may suffice as a replacement for images.

## 5 Related Work

Weir et al. (2020) also use the CSLB concept property norms to probe LMs for commonsense knowledge. Our work differs from theirs in that we focus on visual commonsense knowledge and evaluate several multimodal models for whether their multimodal training grants them additional visual commonsense knowledge.

Norlund et al. (2021) also query a multimodal model for visual commonsense knowledge but with a focus on memory colors. Paik et al. (2021) present similar work but with more focus on probing and reporting bias. In our work, we include general visual commonsense knowledge concepts and evaluate several multimodal models.

Additionally, Iki and Aizawa (2021) evaluate several vision-and-language models on GLUE, to investigate the effect of an additional visual modality on the general linguistic capabilities of a model. Our work differs in that we evaluate the models specifically for visual commonsense knowledge.

Other tasks that have been developed to evaluate the performance of vision-and-language models are Visual Question Answering (VQA) tasks and Visual Commonsense Reasning (VCR) tasks (Goyal et al., 2017; Hudson and Manning, 2019; Zellers et al., 2019a). Our work differs from these in that we evaluate for visual knowledge in models without conditioning on an image, to investigate whether the linguistic capabilities of a model improve from training on more than text. In the aforementioned tasks, the text prompts are always conditioned on an image provided with the prompt, obstructing equal comparisons with text-only models.

## 6 Limitations

Our work is limited to a subset of vision-and-language models, so the results found may not translate to all such model types. Also, since our evaluation utilizes prompt-based retrieval, its measurement accuracy depends on how well this method works for LMs. Additionally, as previously mentioned, we do not investigate how well the multimodal models adapt to a unimodal input. Thus, our results depend on whether the models were functioning adequately with our method of adapting them to a unimodal input.

## 7 Ethical Considerations

Our work should not have any direct ethical implications, since we mainly introduce evaluation tasks and evaluate different models on them. We do however investigate visual conceptual perceptions based on data from a potentially small group of people whose world-view may be culturally different from that of other individuals. This means that we may encourage knowledge that benefits some people more than others. Similar issues are discussed by Liu et al. (2021). Our investigation is limited to English-language models and datasets, limiting the generality of our conclusions.

## 8 Conclusions

We introduce new evaluation methods for measuring the visual commonsense knowledge in LMs and evaluate a number of multimodal LMs on these benchmarks. We find that there are no significant differences in performance between models trained on pure text and models trained on images and text. Most prominently, we find that a unimodal LM trained on image captions and VQA queries can attain a visual commonsense knowledge on par with that of a multimodal model.

We also confirm the results by Jiang et al. (2020) and Cao et al. (2021), that LMs are sensitive to query format even when querying for commonsense knowledge. This casts some doubts on what is really measured in a model for a cloze task and whether we can reason about LMs as having knowledge. An interesting future step would be to investigate this further and see if it would be more applicable to use e.g. probing or some other evaluation method.

Nonetheless, this is a first step towards measuring the visual commonsense knowledge in multimodal as well as unimodal LMs. We hope that the evaluation tasks introduced here may aid other researchers in their aim to create models that learn language from more than text.

## Acknowledgements

We thank the anonymous reviewers for their valuable feedback and knowledge sharing.

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

Additionally, the computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973.

## References


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.

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 Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Boxi Cao, Hongyu Lin, Xianpei Han, Le Sun, Lingyong Yan, Meng Liao, Tong Xue, and Jin Xu. 2021. Knowledgeable or educated guess? revisiting language models as knowledge bases. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1860–1874, Online. Association for Computational Linguistics.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46(4):1119–1127.

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.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30.

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

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. 2021. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Taichi Iki and Akiko Aizawa. 2021. Effect of visual extensions on natural language understanding in vision-and-language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2189–2196, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

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

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, 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 *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tobias Norlund, Lovisa Hagström, and Richard Johansson. 2021. Transfering knowledge from vision to language: How to achieve it and how to measure it? In *Proceedings of the Fourth BlackboxNLP*

*Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–162, Punta Cana, Dominican Republic.

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*, pages 823–835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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 2463–2473, Hong Kong, China. Association for Computational Linguistics.

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.

Nathaniel Weir, Adam Poliak, and Benjamin Van Durme. 2020. Probing neural language models for human tacit assumptions. In *42nd Annual Virtual Meeting of the Cognitive Science Society (CogSci)*.

Tian Yun, Chen Sun, and Ellie Pavlick. 2021. Does vision-and-language pretraining improve lexical grounding? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4357–4366, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019b. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Yian Zhang, Alex Warstadt, Xiaocheng Li, and Samuel R. Bowman. 2021. When do you need billions of words of pretraining data? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1112–1125, Online. Association for Computational Linguistics.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9733–9740.

## A Additional model information

Additional information about the models used in our work and their training datasets can be found in Tables 2 and 3. We can observe that VisualBERT has been trained on a data amount that is quite small compared to those of CLIP-BERT and LXMERT.

It is also worth noting on the different backbones of the models. CLIP-BERT is a single-stream multimodal model with a CLIP backbone for visual processing. LXMERT is a dual-stream multimodal model with a Faster R-CNN detector backbone. While VisualBERT is a single-stream model that also utilizes Faster R-CNN detector backbone. Since CLIP has been trained on the immense WIT dataset, the backbone data sizes differ greatly between CLIP-BERT and the other multimodal models.

## B Additional information on Visual Property Norms

Information about the different segments and number of entries per segment in the Visual Property Norms can be seen in Table 4.

## C Additional results on Visual Property Norms

Additional model results on the Visual Property Norms can be found here.

Figure 3 indicates model performance per feature relation across the production frequency segments. We can observe how the models show the best performance for the *is made of* relation, which arguably can be associated more with visual perceptual properties.

Figure 4 shows model score per query template across all production frequency segments, indicating that CLIP-BERT-implicit benefits from being more robust to different query templates. Additionally, these results indicate that BERT-LXMERT-train would have the best overall score on Visual Property Norms if the queries containing "q: a" were to be removed.

Lastly, Figure 5 contains the results of the models on the color part of Visual Property Norms which has been filtered to only contain queries with

| Model | Text | Visual text | Images+Text | Backbone | Training objectives |
|---|---|---|---|---|---|
| BERT | 80M | | | | MLM, NSP |
| CLIP-BERT-implicit | 80M | | 4.7M | 400M | MLM |
| CLIP-BERT-explicit | 80M | | 4.7M | 400M | MLM |
| BERT-CLIP-BERT-train | 80M | 4.7M | | | MLM |
| LXMERT | | | 9.2M | 0.1M | MLM, RFR, DLC, ITM, IQA |
| BERT-LXMERT-train | 80M | 9.0M | | | MLM |
| BERT-LXMERT-train-scratch | | 9.0M | | | MLM |
| VisualBERT | 80M | | 1.7M | 0.1M | MLM, ITM |

Table 2: An overview of the pre-trained models, the sizes of their training datasets and their pre-training objectives. The sizes are measured in number of training samples. The backbone column indicates the training data sizes for the image processing backbones of the models. For the training objectives, ITM refers to Image-Text Matching, RFR to RoI-Feature Regression, DLC to Detected Label Classification, MVM to Masked Visual Modeling and IQA to image QA.

| Dataset | Data sources | # of text | # of images |
|---|---|---|---|
| CLIP-BERT V+L | MS COCO, SBU Captions, VG-QA, CC | 4.72M | 2.91M |
| LXMERT V+L | MS COCO, VG, VQA, GQA, VG-QA | 9.18M | 0.18M |
| VisualBERT V+L | MS COCO, VQA | 1.27M | 0.12M |

Table 3: The vision-language datasets on which the multimodal models originally were trained. More information about the datasets can be found in the articles that introduced the models.

| PF | entries | *has* | *has a* | *made of* | *is* |
|---|---|---|---|---|---|
| 2 | 6,541 | 1,675 | 1,190 | 1,176 | 2,500 |
| 5 | 3641 | 1,016 | 642 | 760 | 1,223 |
| 10 | 2001 | 583 | 347 | 509 | 562 |
| 20 | 613 | 169 | 88 | 209 | 147 |
| 30 | 27 | 5 | 2 | 10 | 10 |

Table 4: The data segments segmented based on production frequencies together with their number of entries. The entries are calculated as the number of feature-concept-label entries, where there can be several features belonging to the same feature and concept. The PF column indicates the production frequency threshold for each segment, all features with a production frequency higher or equal to this threshold are included in the segment. We also list the number of labels per feature relation type.

gold labels describing colors. Here, we see some indications of a better performance of CLIP-BERT-explicit for colors. Potentially, the imagination capacity of this model is more helpful for queries with answers relating to more basic visual properties, such as color.

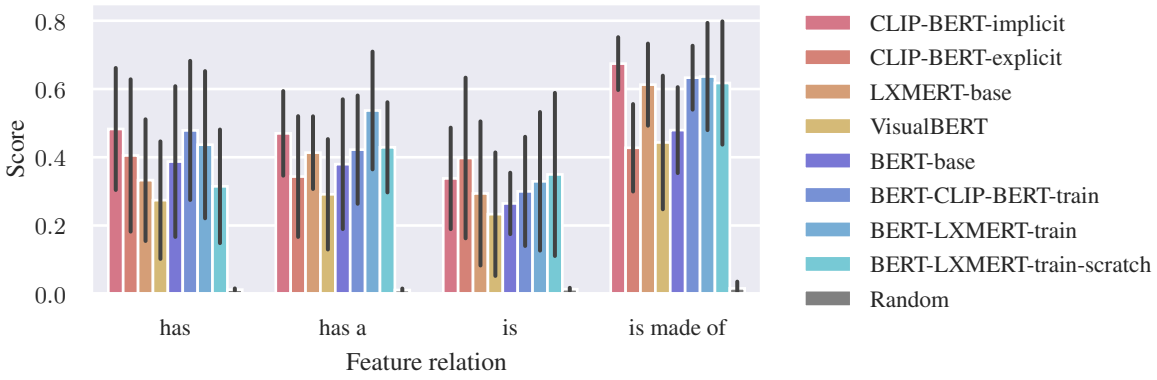


Figure 3: The model scores on Visual Property Norms per feature relation. The error bars indicate the standard deviation of the model performance over the different query templates. The score has been calculated by masking the vocabulary of the models to only contain the possible answers of the task.

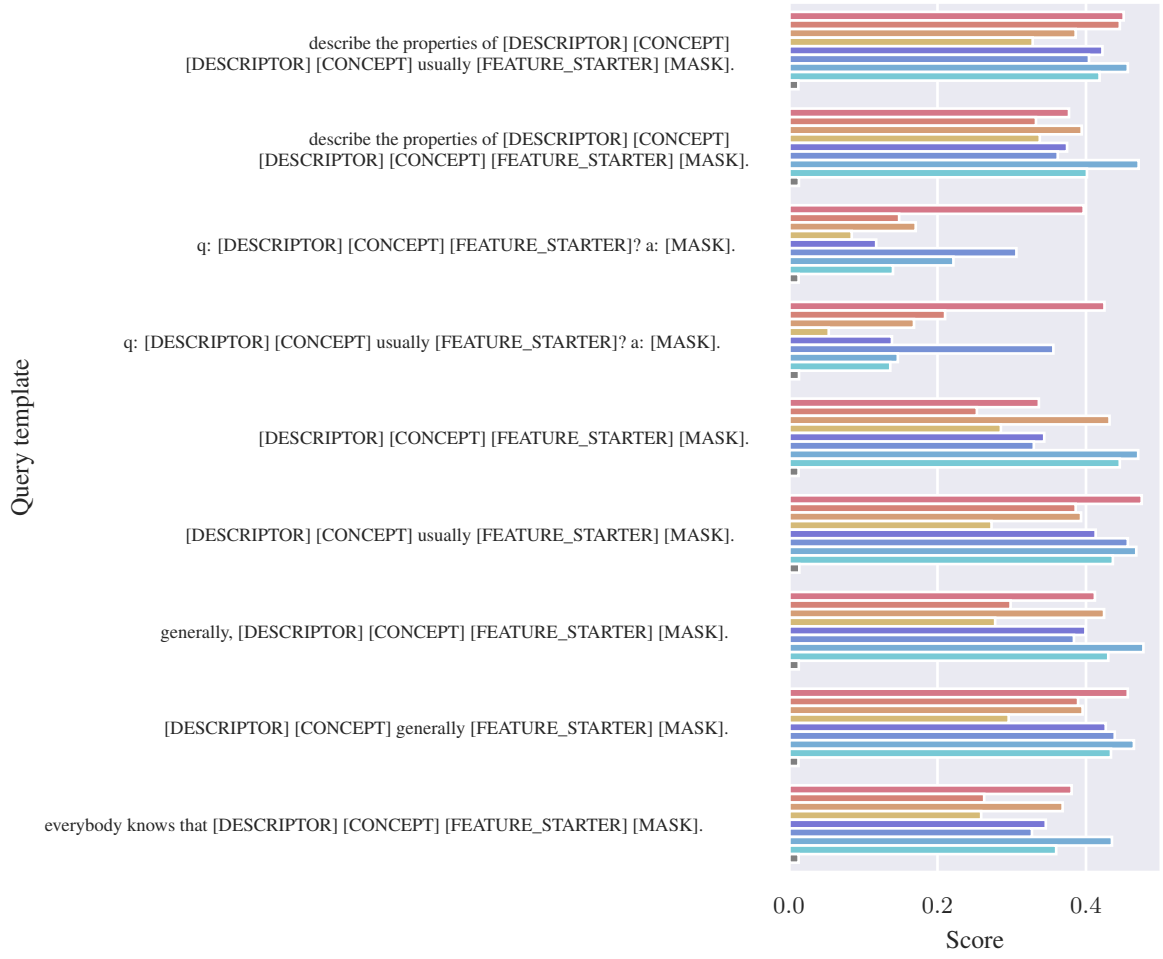


Figure 4: The score for each model on Visual Property Norms per query template. The score has been calculated by masking the vocabulary of the models to only contain the possible answers of the task.

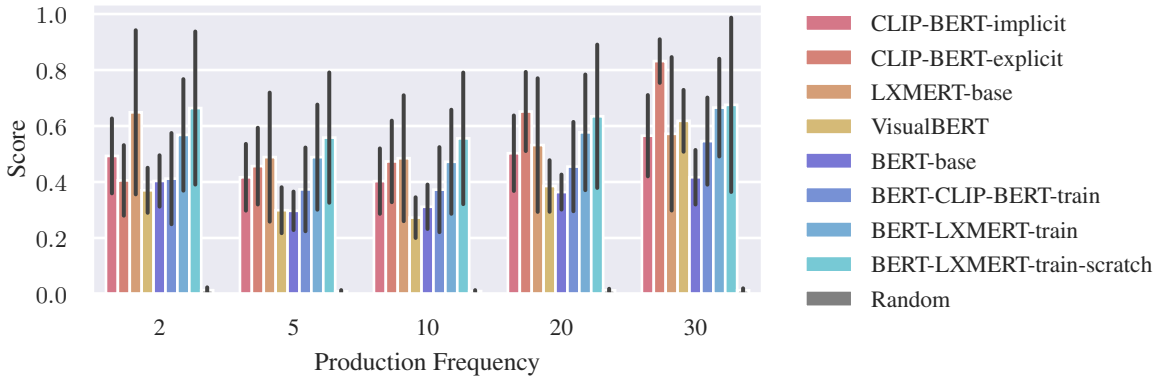


Figure 5: The score for each model per production frequency segment on Visual Property Norms that has been filtered to only contain samples for which the correct answer is one or more out of 11 possible colors. The score has been calculated by masking the vocabulary of the models to only contain the possible answers of the task.