Hallucinate or ground: how general or specific are objects descriptions generated by a vision-and-language transformer?

Nikolai Ilinykh **Simon Dobnik**
Centre for Linguistic Theory and Studies in Probability (CLASP)
Department of Philosophy, Linguistics and Theory of Science (FLoV)
University of Gothenburg, Sweden
{nikolai.ilinykh,simon.dobnik}@gu.se

Abstract

We examine a pre-trained bi-directional multi-modal transformer how it produces descriptions of objects in images. We find that such a model still inclines towards generating a generally more common noun. We believe this is an essential caveat for building robust models and better multi-modal datasets.

1 Introduction

One of the most prominent tasks of representation learning is the multi-modal one (Bisk et al., 2020; Bender and Koller, 2020). Although recent work shows that text-only models can capture meaning by looking at the form only (Merrill et al., 2022), the advances made by large multi-modal architectures are not easily denied. Many language-and-vision transformers (Tan and Bansal, 2019; Lu et al., 2019a), datasets, tasks (Antol et al., 2015; Das et al., 2018) and test suits targeted towards specific phenomena (Park et al., 2020) have moved the NLP field further. The representations captured in these setups are different from uni-modal ones. Multi-modal representations have also been shown to benefit tasks such as machine translation (Elliott et al., 2015). However, it is still hard to learn a model that provides optimal fusion of modalities (Elliott, 2018; Frank et al., 2021).

What makes multi-modal representations useful? For example, Ilinykh and Dobnik (2021) have shown that multi-modal representations positively bias the model towards more global attention patterns (unlike uni-modal representations), allowing for higher quality generation. On the contrary, Hagström and Johansson (2022) have found no significant value in using language-and-vision representations compared to text-only features when identifying common-sense visual knowledge. However, this might not be globally the case as the task and the dataset are also important to the extend they

Figure 1: MSCOCO image-text example: "a child holding a flowered umbrella and petting a yak". The model takes a description with the words in red replaced with the MASK token and generates "a woman holding a flowered plant and petting a bull goat".

focus on *general* vs *context-specific* knowledge which is captured differently by different models (Dobnik et al., 2018; Storks et al., 2021). In multi-modal models visual representations are known to be often dismissed by (Agrawal et al., 2018; Thomason et al., 2019). In this paper we therefore test the ability of multi-modal language models to establish an image-specific understanding of words during description generation.

We examine a multi-modal bi-directional transformer which is expected to have both (i) a general knowledge from large-scale pre-training and (ii) a specific understanding of image-text situations after being fine-tuned on several downstream tasks. In this respect, we discover insights about the ability of such models to learn situation-specific contextual representations of visual scenes. We test the model with the *masked language modelling* task in the context of object descriptions. Specifically, we mask a subset of words (nouns), which typically describe objects in the original caption and measure the model's perplexity. The example is shown in Figure 1. Our results demonstrate that the model

struggles to understand specific image-text situations and focuses on common-sense knowledge predicted from a language model.

2 Data and models

We use ViLBERT (Lu et al., 2019b) and its public implementation in Volta (Bugliarello et al., 2021). ViLBERT takes a set of object features $\mathbf{V} = \{\mathrm{v}_1, \ldots, \mathrm{v}_N\}$ and a sequence of words $\mathbf{W} = \{\mathrm{w}_1, \ldots, \mathrm{w}_M\}$ as its input. Each input modality has its own CLS token. The training objectives are identical to those of BERT (Devlin et al., 2019). In particular, the first objective is a standard masked word prediction: part of words in $\mathbf{W}$ are replaced with a special MASK token and the model learns to predict the correct token from masked features. Similarly, the model is also trained for masked region prediction: to predict object labels from a distribution given some masked visual features. Finally, both CLS tokens are combined in the image-text matching objective: given a caption, predict whether it describes the image. Captions of other images are used as negative examples. We use a publicly available model[1] that has been pre-trained on the Conceptual Captions dataset (Sharma et al., 2018) and fine-tuned for several tasks, including the image retrieval task on the train set of the MSCOCO dataset (Lin et al., 2014). The benefit of the Volta model is that it has a large *general* knowledge about the multi-modal world. We keep all parameters of the model fixed as reported in Bugliarello et al. (2021) and run it on the validation set of MSCOCO with 5,000 images where each image has five human-produced captions. Visual features of images were pre-extracted by the authors of the Volta framework were 36 objects per image were detected[2].

3 Experimental setup

We identify nouns in image descriptions, mask them and retrieve the model's predictions for masks. ViLBERT is using the BERT-based WordPiece tokeniser (Devlin et al., 2019) which may split a noun into two or more sub-words. It is hard to manually build a set of rules that correctly map nouns to original descriptions with sub-words produced by BERT since the number of words changes after the tokenisation step. To overcome this problem, we employ a BERT tokeniser fine-tuned for part-of-speech tagging[3]. This tokeniser produces a POS tag per each output item.

The Volta model takes a caption $\mathbf{W}$, consisting of M words $\{\mathrm{w}_1, \ldots, \mathrm{w}_M\}$, where some of the words are replaced with the MASK token. We replace only words that have NOUN as the POS tag. On the vision side, we keep the original input to the model which includes random masking scheme for object features. We set $M = 20$, each word $\mathrm{w}_m \in \mathbb{R}^{1\times D}$, where $D = 512$ and every object feature $\mathrm{v}_n \in \mathbb{R}^{1\times F}$ where $F = 2048$. The model outputs a probability distribution over the entire vocabulary $\mathcal{V} = 30522$ on the language side. At each step m we sample the most probable word and reconstruct the caption. We evaluate the model only on the first caption from a set of 5.

4 Evaluation

As our evaluation measure we use perplexity. Perplexity allows us to examine the model's confidence in predicting tokens of the text, and it is defined either either a normalised inverse probability of the test set or an exponential of cross-entropy. We use the second definition and first compute a cross-entropy loss $\mathcal{L}$ over a full dataset of size T as follows:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{|M|} \mathrm{y}_{t,m}\ \log_b(\hat{\mathrm{y}}_{t,m}), \qquad (1)$$

where b is the base of the log (either 2 or exp[4]), y and $\hat{\mathrm{y}}$ are the two probability distributions, the ground-truth and the predicted probability distribution respectively. The perplexity is computed as a base of the logarithm used in the loss function computation to the power of the loss, $\mathrm{PPL} = b^{\mathcal{L}}$; we set $b = \exp$. We calculate loss and perplexity over the masked tokens only.

5 Results

Figure 2 shows some examples where perplexity of the model is the highest and when it is the lowest. The mean perplexity of the model (on the word level) across all examples is 371.78, while the standard deviation is 2660.35 and the variance is

[1] https://github.com/e-bug/volta/blob/main/MODELS.md

[2] https://github.com/e-bug/volta/blob/main/data/README.md

[3] https://huggingface.co/vblagoje/bert-english-uncased-finetuned-pos

[4] The choice of the log base is not essential because differences are scaled by a constant factor.

(a) **G**: a young boy holding a set of nintendo wii game controllers
H: a young boy holding a bag of water in a gym

(b) **G**: a fire hydrant is painted to look like a dalmation
H: a red street that is painted to look like a lighthouse

(c) **G**: two brown grizzly bears are rearing up and fighting
H: two brown grizzly bears are rearing up and fighting

(d) **G**: a surfer is riding on a big wave
H: a surfer is riding on a big wave

Figure 2: Examples of images, ground-truth (**G**) texts and texts with predicted nouns (**H**) for cases with high perplexity (a, b) and low perplexity (c, d).

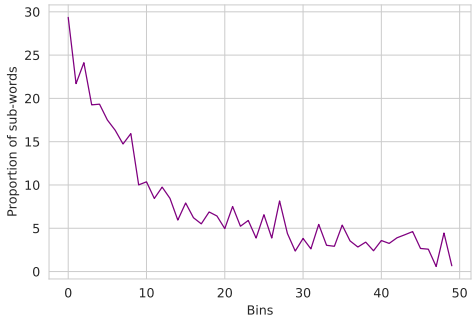

Figure 3: Summed proportions of sub-words per item per bin. The examples and bins are sorted from the ones with the highest perplexity scores (left) to the ones with the smallest perplexity score (right).

7077511.33. One interpretation of these numbers is that the model performs very differently on many examples in the test set and its performance also largely varies from item to item. For example, the highest perplexity we observe is 97078.98, while the lowest is 1.01. This indicates that the model is confused by a large variety of visual situations and descriptions in the data. This is a positive characteristic of the dataset but also a very challenging one for the model.

Sub-word tokenisation increases perplexity One reason for the model's confusion could be the *tokenisation scheme*: input words are split into sub-words by a WordPiece tokeniser which increases the complexity of mapping between texts and images as more words may correspond to a scene. The model is now required to establish a link between a visual representation of the object and multiple sub-words which constitute a noun that describes an object. Unlike nouns its sub-words do not immediately correspond to visual semantic information and extra reasoning over sub-words is necessary to make a correct prediction.

We evaluate whether there is a correlation between the number of sub-words in tokenised input texts and perplexity scores. We focus on sub-words extracted from nouns and use the method to identify nouns described in Section 3. We calculate the proportion of sub-words per item by dividing the number of nouns with (any number of) sub-words by the total number of detected nouns. For example, if a caption has two nouns and one has been split into multiple sub-words the proportion is 0.5. Next, we sort image-text pairs by their perplexity score from the highest to the lowest and split all 5,000 examples to 50 bins. We sum the scores in each bin. The results in Figure 3 show that bins with high perplexity also have a high proportion of sub-words and bins with low perplexity have a low properties of sub-words. There is a significant correlation between perplexity and the proportion of sub-words (Pearson's, $r = 0.64, p = 4.4e-07$). The result demonstrates how a design choice impacts the model's confusion which directly affects image-specific noun prediction.

Perplexity, general and specific knowledge Next, we ask if the model predicts more frequent nouns across the data when its perplexity is high. We believe that the commonality of the noun is related to its frequency in the data: for example, it is simpler to predict "wall" rather than "window"

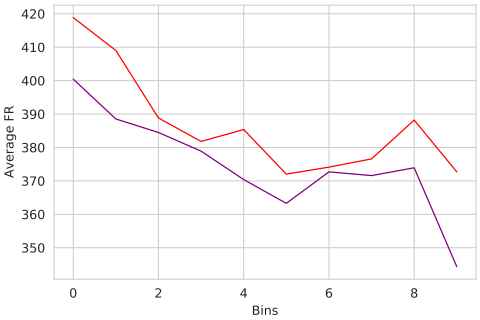

Figure 4: Relation between the commonality ranking of nouns in ground-truth descriptions and perplexity for humans (red) and model (purple). Average FR stands for average frequency rank (high value = low rank!) and bins represent items grouped by perplexity scores from high to low. For better visualisation we divide the frequency ranks by 1,000.

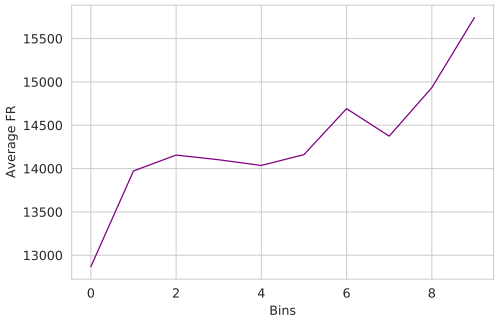

Figure 5: Average commonality/frequency rank (low value = high rank) when generating 100 most common nouns distributed in bins with decreasing perplexity from left to right.

in "a person is standing next to the MASK" when (i) the dataset contains more images of people standing next to something, and (ii) such images are described with the mention of the wall. This is a data-related bias that the model captures. To get the frequency ranks of nouns we count them in the *ground-truth* test set and sort them in a descending order which gives us a noun commonality ranking $\mathcal{C}$. Words that are common are assigned high ranks. Next, for each predicted noun we look-up its rank from $\mathcal{C}$ and compute an average frequency rank of nouns in the predicted caption. The 5,000 items are sorted in a descending order from high to low perplexity. For visualisation purposes we group items into 10 bins. We also calculate an average frequency ranking of nouns in the ground-truth descriptions to compare the model with the human performance. The results in Figure 4 demonstrate the lower the perplexity (the later bins), the more likely the model is to pick common words (low value = high rank, closer to the top of $\mathcal{C}$). This indicates that a lot can be predicted from commonsense knowledge captured in a language model. Humans follow the same trend. This may be connected to a low "surprisal" of visual scenes, their typicality and the way they can be described with common, general nouns instead of specific ones. The model is less confused by picking a more common noun which guarantees a better fit with the ground-truth produced by a human.

To discern this question further we examine the perplexity of the model when predicting only most common words, thus in situations when the model is taking a safer choice and therefore most likely ignoring visual information. We consider 100 nouns with highest ranks (low value = high rank). The results are shown in (Figure 5) and demonstrate that when perplexity is high (the lower bins), the model indeed tends to choose among the most common nouns (low values). Bins with lower perplexity contain less common nouns. This result demonstrates that the model may be biased towards more common nouns when it struggles to predict image-specific nouns describing objects.

6 Conclusion

We have examined the ability of the multi-modal bi-directional transformer to predict image-specific object descriptions represented by nouns. Although our experiments touch only at a surface of the problem, our results indicate that the model is biased to predict more common (frequent) nouns for image-description pairs which are hard to ground. The vision and language transformer appears to lack specific, contextual knowledge about the visual situation and the ability to balance between a general language-based and vision-specific knowledge. Future work should focus on further evaluation of the visual transformer models which would lead to model improvements in terms of information fusion and grounding different modalities.

Acknowledgements

The research reported in this paper was supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4971–4980. Computer Vision Foundation / IEEE Computer Society.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433.

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.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. 2018. Embodied question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 1–10. Computer Vision Foundation / IEEE Computer Society.

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.

Simon Dobnik, Mehdi Ghanimifard, and John Kelleher. 2018. Exploring the functional and geometric bias of spatial relations using neural language models. In *Proceedings of the First International Workshop on Spatial Language Understanding*, pages 1–11, New Orleans. Association for Computational Linguistics.

Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, and Eva Hasler. 2015. Multi-language image description with neural sequence models. *CoRR*, abs/1510.04709.

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*, pages 9847–9857, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Lovisa Hagström and Richard Johansson. 2022. What do models learn from training on more than text? measuring visual commonsense knowledge. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 252–261, Dublin, Ireland. Association for Computational Linguistics.

Nikolai Ilinykh and Simon Dobnik. 2021. How vision affects language: Comparing masked self-attention in uni-modal and multi-modal transformer. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pages 45–55, Groningen, Netherlands (Online). Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019a. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019b. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

William Merrill, Alex Warstadt, and Tal Linzen. 2022. Entailment semantics can be extracted from an ideal language model. In *Proceedings of the 26th Conference on Computational Natural Language Learning*. arXiv.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the dynamic context of a

still image. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V*, volume 12350 of *Lecture Notes in Computer Science*, pages 508–524. Springer.

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

Shane Storks, Qiaozi Gao, Yichi Zhang, and Joyce Chai. 2021. Tiered reasoning for intuitive physics: Toward verifiable commonsense language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4902–4918, Punta Cana, Dominican Republic. 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.

Jesse Thomason, Daniel Gordon, and Yonatan Bisk. 2019. Shifting the baseline: Single modality performance on visual navigation & QA. 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 1977–1983, Minneapolis, Minnesota. Association for Computational Linguistics.