Approximation sacrifices variability: estimating the closeness between human and machine language in image descriptions

Anonymous ACL submission

Abstract

We examine language generated by two state-of-the-art image description systems, BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023). We illustrate that while these systems generate accurate image descriptions, they lack the complexity and variability inherent in human-authored descriptions. Our study provides motivation for a closer exploration of how generic model-generated texts and languages as spread of such visibly accurate language might be misleading. Our study motivates a closer exploration of how generic model-generated image descriptions can be as ubiquitous use of the models might lead to a less variable language.

1 Introduction

Human language is naturally variable and it is hard to reproduce such variability computationally as shown by studies on natural language inference (Plank et al., 2014; Pavlick and Kwiatkowski, 2019; Plank, 2022), object naming (Silberer et al., 2020) and image captioning (Bernardi et al., 2016). One major problem preventing neural generative models of language-and-vision from producing image descriptions with human-level of variability is the training objective of maximum likelihood estimation (MLE). This method tends to prioritise certain combinations of words present in the model's training data during testing (Devlin et al., 2015; Dai et al., 2017) and a few approaches to mitigate this issue have been introduced (Welleck et al., 2020). Still, current language-and-vision models built on top of the transformer architecture (Vaswani et al., 2017) and trained with MLE have shown tremendous improvements in the quality of generated image descriptions (Alayrac et al., 2022; Li et al., 2023) due to their massive pre-training data, size and sophisticated model designs. It is unclear how much the language that such models acquire through MLE mirrors not only statistical patterns and structures, but also variability of human image descriptions (Castro Ferreira et al., 2016; Li et al., 2016; van Miltenburg et al., 2018).

In this study we present experiments on the analysis of the language that current state-of-the-art pre-trained image description models generate when applied. We use "language" to describe all tokens that a speaker (a multi-modal model) generates for a specific corpus of images. Both models and humans can be considered speakers, albeit with their own individual languages. We offer insights into how close these languages are by answering the following questions:

- How predictable (surprising) each language is and are machine-generated texts as complex as human ones?
- Do changes in the surprisal levels indicate better or worse quality of image descriptions?
- To what extent do the languages generated by the models align semantically with human language?

We use the measure of surprisal (Hale, 2001; Levy, 2008) to evaluate the differences in variability between languages of humans and models in the multi-modal domain. We use surprisal interchangeably with complexity as, for example, the more surprising the human text is for the model, the more complex it is. We also evaluate the quality of image descriptions and their similarity with human ground-truth texts.Our results highlight issues with current image description models and the metrics used to evaluate them as we show that machine-generated texts might seem accurate, but they are still far from being truly human-like. Our study aims to advocate for robust assessment of cutting-edge image description systems and raise awareness about the dangers of producing generic, non-variable language when these models are used in real-world applications.

2 Experiments

We use BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023) as machine speakers for generating image descriptions. These models, recognised as state-of-the-art image description systems trained on extensive data and multiple tasks, are expected to yield high-quality image descriptions[1]. Rather than focusing solely on the models themselves, evaluating the descriptions generated by these models offers insights into whether their language exhibits similarities with human descriptions at syntactic and semantic levels. This evaluation is crucial as it is predictive of how such descriptions might be perceived by humans in practical applications.

2.1 Models of language

We train two types of n-gram models: one on human-generated language and another on machine-generated language. Each model is a simple bi-gram language model with Laplace Smoothing as implemented in NLTK toolkit (Bird et al., 2009)[2]. The models are expected to capture the n-gram distribution in languages of different speakers and by testing these models on the language of a chosen speaker, we capture its complexity.

2.2 Data

We use Karpathy test split of MSCOCO dataset (Lin et al., 2014) – a standard benchmark for evaluation of models such as BLIP or BLIP-2. With 5000 scenes, each described by at least 5 humans, we gather 25010 human-generated descriptions. We split the dataset into LM_{train} (4500) and LM_{test} (500) for training and testing our n-gram models. In LM_{train} we obtain 22508 human-generated and 4500 machine-generated descriptions. The LM_{test} includes 2502 human-generated and 500 machine-generated descriptions. We generate four different sets texts by building machine speakers with different decoding algorithms such as greedy search, beam search, contrastive search (Su et al., 2022) and nucleus sampling (Holtzman et al., 2020). Such diversity in machine speakers allows for a nuanced evaluation of their quality and similarity with the human texts, as the decoding choice has a direct effect on the speaker's language.

[1]We load pre-trained models from HuggingFace and run them on NVIDIA GeForce RTX 3090 Ti.

[2]We trained and tested a variety of n-gram models, but bi-gram models showed the best performance, e.g. lowest perplexity on the test set.

2.3 Evaluation

In psycholinguistic studies of human language the notion of surprisal has been typically defined as the negative log-likelihood of the conditional probability of the target word given a specific context (Hale, 2001; Levy, 2008). This formulation aligns with the task of language modelling, wherein a model learns to predict the next word based on previous context and generates probabilities for the subsequent word. Negative log probabilities of an event can be intuitively interpreted as the information entropy of an event. Therefore, we compute entropy $\mathbf{H}$ of a n-gram model on predicting words in a specific sentence as follows:

$$\mathbf{H} = -\frac{1}{\mathbf{N}} \log P\left(\mathbf{w}_t \mid \mathbf{w}_{t-n}, \ldots, \mathbf{w}_{t-1}\right), \quad (1)$$

where $\mathbf{N}$ is the sentence's length and n is an n-gram window. To understand how well our model predicts each sentence, we compute perplexity $\mathbf{P} = 2^{\mathbf{H}}$. Looking at perplexity we can assess the acceptability of the language used by each machine speaker, an aspect typically overlooked by the automatic metrics. As comparing perplexities of n-gram models can be tricky if models have different vocabularies (Buck et al., 2014), we report vocabulary size of each n-gram model and human speakers as $|\mathcal{V}|$. The models are expected to share similar vocabularies as they are trained and tested on the data within the same domain and dataset.

3 Estimating the language complexity

The evaluation of complexity of languages from different speakers is shown in Table 1. We first evaluate and compare language complexity of speakers trained and tested on the same data. Human language complexity is determined by the score of 538.4. In comparison, machine language shows smaller complexity for itself (diagonal rows), except the model trained on M_{nu} and tested on M_{be}. This is the first indication that machine speakers generate simpler descriptions, which are highly predictive of each other, while humans generate much less predictable language. Are the differences in perplexities between different speakers significant? We conduct a Student's t-test (Gosset, 1908) to compare the perplexities of the models trained and tested on the same data. The model generates one description per image and there are N human descriptions for an image. Thus, we replicate the model's perplexity score N times, resulting in N

		Human	BLIP				BLIP-2			
			M_{gr}	M_{be}	M_{co}	M_{nu}	M_{gr}	M_{be}	M_{co}	M_{nu}
	$\lvert\mathcal{V}\rvert$	6919	1222	1291	1300	2100	1476	1616	1711	2383
P	Human	***538.4***	**206.3**	**240.6**	**1248.5**	**791.5**	**188.2**	**183.5**	**419.9**	**614.0**
	M_{gr}	*357.0*	90.3	119.9	593.6	439.5	106.1	121.3	204.5	423.4
	M_{be}	*342.4*	90.8	93.6	563.8	401.6	109.3	110.2	203.5	397.3
	M_{co}	*448.6*	194.5	224.8	269.0	497.5	149.2	155.0	183.7	513.2
	M_{nu}	*438.0*	177.9	187.3	602.1	422.3	178.1	179.3	289.0	395.3
BERTSCORE ↑			0.843	0.847	0.787	0.797	0.864	0.868	0.844	0.818
MOVERSCORE ↑			0.609	0.615	0.567	0.583	0.623	0.629	0.613	0.597

Table 1: Performance of bi-gram language models trained on different speakers (Human, BLIP, BLIP-2) in terms of averaged perplexity **P**, BERTSCORE and MOVERSCORE. The columns and rows denote what the model has been trained and tested on, respectively. Models trained on machine-generated language were tested with different decoding algorithms: gr (greedy), be (beam), co (contrastive) or nu (nucleus)). $|\mathcal{V}|$ is the vocabulary size of the corresponding model. Numbers in **bold** are comparisons across models tested on human data, numbers in *italics* are perplexities of the model trained on human data and red cells denote perplexities of a model trained and tested on the same data. Numbers in blue are best-performing models.

pairs of human-model perplexity scores. We observe a significant difference[3] between the means of perplexities: 34.010*, 33.735*, 18.883*, 7.510* for BLIP and 32.646*, 32.440*, 24.675*, 9.146* for BLIP-2 for greedy, beam, contrastive and nucleus search respectively. The languages of humans and machines employing more deterministic algorithms (greedy, beam) differ in surprisal levels when compared. In contrast, stochastic algorithms result in smaller differences, leading to similar surprisal levels between the two. Now we turn to the cross-language comparison between speakers. For a human speaker, model-generated data is easy to predict as when tested on such data, perplexity decreases. The texts generated by contrastive and nucleus search are harder to predict for a human speaker. How complex is human language for machine speakers? The picture varies as some perplexities drop down (greedy and beam), while others climb up (contrastive and nucleus) compared to the perplexity level of the human speaker. Interestingly, the scores exhibit different behaviour between BLIP and BLIP-2 as the same search does not necessarily yield the same level and direction of change in perplexity, e.g., $M_{co} > M_{nu}$ for BLIP, but the opposite holds for BLIP-2. Overall, the bi-gram model trained on texts produced by BLIP-2 and generated with nucleus search achieves perplexity closest to the one of the human speaker on human language.

3.1 Evaluating accuracy and similarity

Machine speakers produce less variable language, but are the descriptions they produce true to the image? There are two possibilities: either the machine generates accurate yet different types of descriptions compared to humans, or it produces inaccurate ones. To better understand the relationship between description accuracy and the deviation of the model's perplexity from human perplexity, we compute BERTSCORE (Zhang et al., 2020). BERTSCORE uses contextualised embeddings to measure how accurate the text is compared to the ground truth. We also compute MOVERSCORE (Zhao et al., 2019), a monolingual metric which evaluates the similarity between a pair of sentences. This metric is particularly suited to measure the semantic similarity between image captions. The last two rows in Table 1 show the results for BERTSCORE and MOVERSCORE. The highest scores for both metrics are achieved by greedy and beam search with M_{gr} for BLIP-2 being the highest. Less deterministic decoding algorithms achieve lower scores. However, in terms of perplexity, the bi-gram model that matches human perplexity the closest is the one that uses nucleus sampling, while greedy and beam have a much lower perplexity when tested on human data. Thus, perplexity and both automatic metrics are inversely related. This result demonstrates that such metrics as BERTSCORE might not be good measures of generation. Such metrics favour models that produce language of lower complexity than human language, but with higher similarity to it, hence the generated texts are correct, but very general and not specific in details. If we keep selecting the best model based on the automatic measures alone, we generate languages which are generic and do not replicate variability in human descriptions.

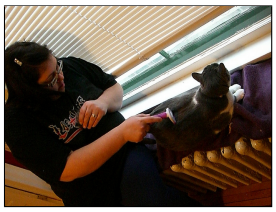
3.2 Qualitative analysis

In Figure 1 we examine descriptions of two models that have demonstrated the closest ability to gen-

[3] * means that the p-value is smaller or equal 0.05.

Nucleus search, MOVERSCORE ↑: A dog is standing on a small boat in water. **Humans**: (i) A dog standing on top of a boat in a body of water. (ii) Dog in flat boat barking at what he sees. (iii) A dog standing in a small boat in the water. (iv) Two dogs are pictured but only one of them is in the boat. (v) A white dog is standing in a small boat.

Contrastive search, MOVERSCORE ↑: A mouse sitting on a desk next to a keyboard. **Humans**: (i) Closeup of a white computer mouse on a table. (ii) A small white mouse sitting next to a keyboard. (iii) A computer mouse next to a computer keyboard on a wooden table. (iv) An apple mouse sits on a desk next to a keyboard. (v) An apple mouse sitting next to a computer.



Nucleus search, MOVERSCORE ↓: Window has a blind. **Humans**: (i) A person brushing a cat with a brush. (ii) A woman brushing her cat that is laying on a radiator. (iii) A woman is brushing her large gray cat. (iv) Cage sitting atop a heater looking out windowpane. (v) A woman that is sitting down near a cat.

Contrastive search, MOVERSCORE ↓: Portrait of Joseph Gilbert, c1910-1911. **Humans**: (i) A vintage photo of a man in a suit and tie. (ii) A portrait of a person in a picture frame. (iii) A man in a suit is not smiling in an old photograph. (iv) An old black and white picture of a man in a suit and tie. (v) There is an image of a man wearing a suit.

Figure 1: Examples of texts generated by BLIP-2 and humans for the two models whose language is the closest to human in terms of complexity (Table 1).

erate language with similar levels of complexity to human language, e.g., M_{nu} and M_{co} for BLIP-2. More examples of descriptions can be found in Appendix A. The description generated with nucleus search and having a higher MOVERSCORE (upper left) appears to be correct for the image, but there are much fewer details mentioned in text compared to human descriptions. The machine-generated description is shorter and more straightforward. On the contrary, humans are using a variety of referring expressions, mentions of attributes, and different object names. In general, the model produces a good description, but metrics such as BERTSCORE or MOVERSCORE seem to give higher scores for "vanilla" descriptions, which have a general fit for all possible human descriptions for this image. When the similarity score drops (bottom left), the description is still correct but more specific to the image and not necessarily in the same way that human descriptions are. For example, human texts mention a woman, a cat, a brush, and describe the process happening in the image, while the model-generated text describes blinds and windows, which are not the main focus of the image. A different search, contrastive search, can also generate descriptions that are very similar to human texts (top right). However, the situation itself might be easy to describe for both humans and models e.g., how much can one say about a computer mouse on a table given the task of describing an image in eight words (MSCOCO task)? This "simplicity" of the visual situation matters because it is likely to prompt less variability in human descriptions and, consequently, in machine-generated descriptions. The model can also generate untrue personalised descriptions (bottom right), while humans, possibly because of the uncertain context, generate a much more general description.

Overall, it appears that machine-generated language that is the closest to human-generated language in terms of complexity might differ from human language on a very fine-grained level. These differences might tell us more about how generic the machine language is. As models are trained to approximate a single image description from human text, they tend to either go very specific or very general. Human language, on the contrary, incorporates a wider range of nuanced details specific to various contexts.

4 Conclusion

In this study we examined the languages generated by two state-of-the-art image descriptions systems, BLIP (Li et al., 2022) and BLIP-2 (Li et al., 2023). We specifically looked at the differences in complexity between these languages. Machine-generated texts employ more predictable words and a smaller vocabulary, while human language is more complex, harder to predict, and richer in variability. N-gram models, trained on data from machine speakers demonstrate significant surprisal levels when encountering human-generated texts, despite the fact that the texts they generate themselves appear to be correct. Automatic metrics find texts which are similar to humans in terms of complexity to be less accurate, thus, they do not take into account the variability in the generated language. Our results indicate that if image captioning models are applied in real life, the language they produce will be more general and generic. Future work should explore if a better training objective and datasets can be used in order to generate descriptions with human level of variability.

5 Limitations

Our evaluation can be expanded if additional models and datasets are used. However, from preliminary experiments (not reported in this short paper) we have found that similar observations would hold for many other image captioning models and datasets. Also, directly comparing n-gram models with varying vocabularies may pose challenges, requiring alternative methods.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 23716–23736. Curran Associates, Inc.

Raffaella Bernardi, Ruket Çakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.*, 55:409–442.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly.

Christian Buck, Kenneth Heafield, and Bas van Ooyen. 2014. N-gram counts and language models from the common crawl. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Thiago Castro Ferreira, Emiel Krahmer, and Sander Wubben. 2016. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–577, Berlin, Germany. Association for Computational Linguistics.

Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional GAN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2989–2998. IEEE Computer Society.

Jacob Devlin, Saurabh Gupta, Ross B. Girshick, Margaret Mitchell, and C. Lawrence Zitnick. 2015. Exploring nearest neighbor approaches for image captioning. *ArXiv*, abs/1505.04467.

William Sealy Gosset. 1908. The probable error of a mean. *Biometrika*, 6(1):1–25. Originally published under the pseudonym "Student".

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. 2022. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 12888–12900. PMLR.

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.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference*

on Empirical Methods in Natural Language Processing, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Carina Silberer, Sina Zarrieß, Matthijs Westera, and Gemma Boleda. 2020. Humans meet models on object naming: A new dataset and analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1893–1905, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *NeurIPS*.

Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. 2018. Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1730–1741, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. 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 563–578, Hong Kong, China. Association for Computational Linguistics.

A Appendix A

Beam search, MOVERSCORE ↑ A man riding a wave on a surfboard in the ocean. **Humans**: (i) A man is riding a wave on a surfboard. (ii) Man in blue swim trunks surfing a small wave. (iii) A man catches a wave as he surfs in the ocean. (iv) A man riding a surfboard on a wave in the ocean. (v) A man riding the waves on his surfboard.

Greedy search, MOVERSCORE ↑ A bird is perched on the back of a giraffe. **Humans**: (i) A bird sitting on the back of a giraffe next to a tree. (ii) A little bird sitting on the back of a giraffe. (iii) The rear end of a giraffe with a bird sitting on it. (iv) A bird is sitting on the behind of a giraffe. (v) A bird sitting on the back of a giraffe.

Beam search, MOVERSCORE ↓ Woman wearing a pink shirt. **Humans**: (i) Two people smiling and using cellular phones in a group of people. (ii) Two people sitting close to one another talking on cell phones. (iii) Two women on cell-phones laughing with trees in the background. (iv) Young women are making faces holding their cell phones. (v) Two women wearing bracelets are talking on their mobile phones.

Greedy search, MOVERSCORE ↓ A group of people on horses. **Humans**: (i) A group of horse-mounted police standing in front of a crowd. (ii) Riot police form a line to control a crowd. (iii) Police on horseback and bicycles holding a crowd back on a street. (iv) Police on foot, horse-back, and bicycle engage in crowd control on a city street. (v) Police sitting on a street blocking people in a certain section.

Figure 2: Examples of model-generated and human-generated texts for a subset of images.