Pragmatic Natural Language Generation with Neural Language Models

---

Evaluating Pragmatic Adequacy in GPT-4v Image Descriptions

Author
Christian Gerber, Ayodeji Olupinla
christian.gerber@student.uni-tuebingen.de
ayodeji.olupinla@student.uni-tuebingen.de

Supervisor
Prof. Dr. Michael Franke
michael.franke@uni-tuebingen.de

Seminar für Sprachwissenschaft
Eberhard Karls Universität Tübingen

January 2024

Abstract

This report aims to evaluate pragmatic adequacy of image descriptions generated by GPT-4v. The goal was to answer the question: How good this powerful neural image captioner is in pragmatically describing pictures? By providing a single picture, which is a combination of two similar pictures, our objective was to assess the model's ability to distinguish one image from the other solely based on description. The leverage metrics were based on the paper "Evaluating pragmatic abilities of image captioners on A3DS" by Tsvilodub and Michael Franke (2023).

Contents

List of Tables

List of Figures

1 Introduction

In the field of neural-based natural language generation (npNLG) advanced language models like GPT-4v present new opportunities to create automated image captions. Due to the background that humans are highly flexible and pragmatic in their use of language, receiving a simple true statement like" the ball is red" would be too plain. Therefore newer neural models like GPT-4v offer the opportunity to extend mere factual descriptions. Due to this new offer, we test if it's pragmatically informative and efficient in communication. By applying established metrics for pragmatic adequacy, we aim to contribute the descriptive power of neural image captioners and answer the question: "How well can GPT-4v predict descriptions of pictures that are not only true but pragmatically adequate?".

2 Background

The field of npNLG emphasizes the need for cooperative and effective communication. Its roots are cognitive science and computational linguistics. For this paper, the challenge lies not only in recognizing and naming objects within the given image, but also in doing it with contextual awareness. In other words, the description must be informative in a specific context, in order for a distinction between similar objects and understanding the unique aspects. Searle (1969) and Grice (1975) set the ground for pragmatics and communicative tasks. Pioneers like Vinyals et al. (2016) and Karpathy et al. (2014) managed to achieve image caption tasks, that were generally accurate, but they usually lacked of the ability of task-specific adaptability seen in human communication. Our research focuses on pragmatic abilities of these models. The abilities of not only labeling images by giving an generic description but also doing so in human-like task-oriented manners. Especially tasks that require identification of an item, where humans tend to distinguish features, such as the surroundings. The challenge of this study is to fulfill these grounded models with the ability to focus further on the pragmatic use of language. While the previous research established a framework for evaluating pragmatic adequacy, they also show the necessity to generate captions, that communicate with the users. This study builds upon the existing research by evaluating GPT-4v's performance in generating pragmatically adequate image descriptions. By applying the metrics from the "Evaluating Pragmatic Abilities of Image Captioners on A3DS" paper to produce captions, that are not only generic, but also informative to distinguish between similar pictures for its users.

3 Data

The data used for this study came from the dataset 3DShapes (Burgess and Kim, 2018) (introduced in Kim and Mnih (2018)) in the "Annotated 3D Shapes" (A3DS) dataset.

4 Method

We make use of 6 different parameters taken from the "Evaluating pragmatic abilities of image captioners on A3DS" by Tsvilodub and Michael Franke (2023).In order to test how pragmatically adequate GPT-4v is as an image captioner.These parameters are as follows:

- Shape Type
- Shape colour
- Shape scale
- Shape orientation relative to the background
- Wall colour
- Floor colour

4.1 Image Selection

As previously mentioned, we selected our images from the A3DS dataset. A total of 50 images were chosen and saved to our device to proceed with the experiment. Cropping the images was a necessary step, as the generated images from the A3DS dataset included axes. To ensure consistency in the dimensions of the images, we developed a Python script specifically for cropping out the axes. This approach allowed us to maintain uniform dimensions across all images. Furthermore we added the clean sentences containing the ground truths of the images into a text file.

4.2 Image Pair Settings

After generating the images, we paired them according to three categories. These categories are based on the number of features the images have in common with each other:

- Category I: One Feature - In this category, images are paired if they share one similar feature.
- Category II: Two Features - In this category, images are paired if they share two similar features.
- Category III: Three Features - In this category, images are paired if they have at least three features in common.

To sort the images into these categories, we employed Python to achieve accurate and efficient results. After sorting, we obtained a total of 2000 different pairings.

4.3 Image Visualization

Once the pairings were provided by Python, we needed to visualize them to make them compatible for analysis by the GPT-4 model. We supplied Python with the names of the images and the pairings and wrote a script to visualize the pairings. This process involved merging two different images into one, with a small white line in between for clarity, specifically designed for the GPT model's analysis.

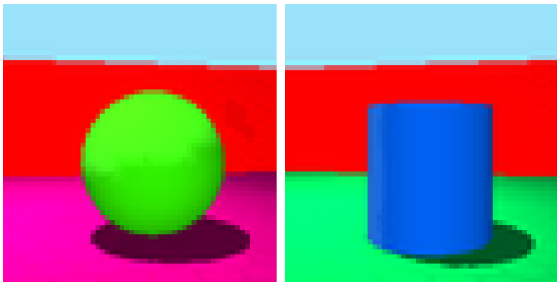

Figure 1: Example image pair matching on three features

4.4 Application of Metrics

The generated descriptions were evaluated using the same four metrics from the referenced paper.We used the following metrics:

- y : This represents the generated caption
- c : This represents the number of contrastive features in a generated caption y
- k: This represents the total number of features mentioned in the generated caption y
- z: This represents the ground truth number of contrastive features between the images.

We then combined these metrics to create categories to evaluate our descriptions:

- *Discriminativity d:* d = 1 if c > 0 else 0, indicating if the caption successfully identifies the target, thus a binary measure of task success.
- *Contrastive efficiency e* (applies only to discriminative captions, i.e., for $d = 1$): $e = 1$ if $k = c = 1$, else: $e = 1 - \frac{c-1}{k-1}$, indicating whether the description avoids overmodification with contrastive features. This notion captures the extent to which the caption is economic and observes the communicative Maxim of Quantity, i.e., includes necessary details for the task but not more (Grice, 1975).
- *Relevance r*: $r = 1 - \frac{k-c}{6-z}$, indicates the propensity to avoid producing redundant non-contrastive features. This is formalized via the proportion of mentioned non-contrastive features $(k - c)$ compared to all non-contrastive features $(6 - z)$. It represents the communicative Maxim of Relevance (Grice, 1975) by measuring the degree to which details unnecessary for the task are excluded.
- *Optimal discriminativity od*: $od = 1$ if $c = 1$ else 0. It is a binary indicator summarizing d and e, by binarizing the observance of the Maxim of Quantity for contrastive captions only (Grice, 1975).

5 Experiment

We devised a prompt in order to produce a caption by the model that would pragmatically distinguish the images provided. This prompt was made to generate descriptions for each image and then recorded for subsequent analysis. We added some instructions so the GPT-4v model was concise and straight to the point. To asses the capability to differentiate between the images, we implemented the prompt in two sessions. One session excluding the ground truths (wgt), so as to let the model distinguish between the images based on its perception, and another time with the ground truths (gt).
The Prompt was formulated as follows:
"Based on this image, describe one of the images in such a way that it can be uniquely identified from the other, without mentioning which one you describe. The description should always include the features :object color, object shape, object orientation, object scale , wall color, floor color, except if both images have the feature in common, then do not mention it at all. Make the description short."
We wanted the model to give us distinguishable features based on the parameters we defined in the methods section. This approach intend to produce direct responses that help with the application of our evaluation metrics.

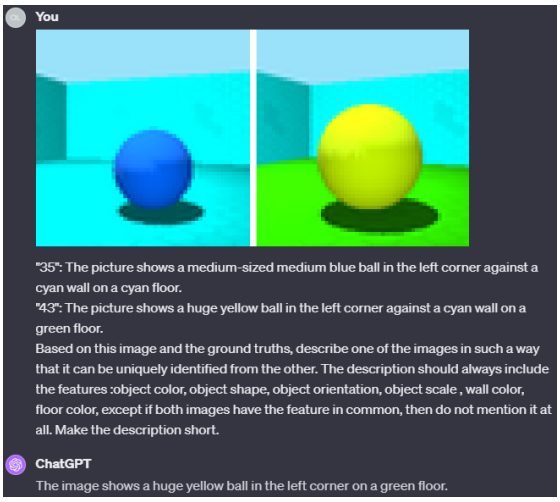

Figure 2: Example of using the prompt in GPT-4v

6 Results

We provided the model with images that shared a number of similar features, more specifically groups of images with one feature, two features and three or more features. By adding these images, we were able to generate concrete results for our evaluation metrics. GPT-4v was first tested on ten paired images without their ground truths falling into category one, two or three and afterwards tested on the same pictures with their ground truths added. This allowed us to see in which conditions it's more difficult for the OpenAI to produce appropriate captions. The results were merely satisfying. As Table 1 shows, it managed to find at least one contrastive feature ($c > 0$, $d = 1$) for all the test sets and categories. In terms of Contrastive Efficiency (e) the score increased for each category set with its best score of 0.717. This was achieved by applying the prompt together with the image to GPT-4v, that shared at least three features (wgt, three-features). The best relevance (r) scores were achieved in the second category with 0.800 in the wgt setting. Proving, that a lot of times, GPT-4v would give back unrelevant features in terms of distinguishing. Only two results for the optimal discriminativity (od) were calculated. As these were the only sets containing captions, where the contrastive features found (c) were only one. With its best score of 0.400 from the wgt-three features pairing. Interestingly to see, we thought the results would improve based on providing the ground truths to the model, but actually it was the other way around. In general, the OpenAI had less difficulties in distinguishing the images, based on their features, the more features they had in common. Especially features like shape, size and orientation was difficult for the model to distinguish without their ground truths. Nevertheless, most of the time it still failed to exclude these three features, when the images both had it in common. Only after the model is told that it did something wrong, it would check again and give back the correct caption.

7 Discussion

- **Impact of Ground Truth Data:** The impact of ground truth data during the experiment was interesting. Contrary to our expectations, providing the ground truth data did not improve the ability of the model to generate captions that are pragmatically adequate. This suggests that the GPT 4 model cannot solely depend on explicit data input, due to the complex nature of AI models when handling tasks that require a high degree of contextual awareness. A lot of times, the AI model would give back captions including features both had in common.

Score	one-feature		two-features		three-features	
	wgt	gt	wgt	gt	wgt	gt
Discriminativity (d)	1.000	1.000	1.000	1.000	1.000	1.000
Contrastive Efficiency (e)	0.200	0.100	0.350	0.317	**0.717**	0.458
Relevance (r)	0.400	0.500	**0.800**	0.600	0.608	0.517
Optimal Discriminativity (od)	0.000	0.000	0.300	0.000	**0.400**	0.000

Table 1: Pragmatic evaluation results by test set category for each model (wgt: without ground truths, gt: with ground truths), averaged across test sets within category. Bold numbers indicate best performance across models and test sets.

- **Pragmatic Adequacy vs Accuracy:** The findings suggest that although the GPT-4v model is generally reliable in producing correct information, its capability to mimic human-like qualities such as subtlety varies. This indicates a distinction between the model's accuracy in data processing and its ability to replicate subtle human communication.

7.1 Challenges and Limitations

- **Contextual Understanding Limitation:** The model's performance in pragmatically adequate caption generation suggests limitations in its ability to fully understand and interpret context as humans do.
- **Lack of Consistency with ground truth data:** The surprising result, where providing ground truth data did not consistently improve performance indicates a complexity in the way the model processes and uses additional information.
- **Instability in replicating human communication:** The model's unstable ability to replicate human-like communication such as subtlety and contextual relevance reveals a gap in its linguistic capabilities compared to human communication.

8 Conclusion

The GPT-4v model demonstrates a significant capability in generating factually correct image captions, but its performance in creating captions that are pragmatically adequate and contextually nuanced is less consistent. This highlights a gap in the model's ability to fully replicate human-like subtlety and contextual understanding in language use and calls for the need for continued research and development to enhance AI models' proficiency in nuanced communication tasks. For future research we would suggest using different prompts as well as more data, in order to find a more precise results.

References

Tsvilodub, P., & Franke, M. (2023). Evaluating pragmatic abilities of image captioners on A3DS.

In A. Rogers, J. Boyd-Graber, & N. Okazaki (Hrsg.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (S. 1277–1285). Association for Computational Linguistics.

John R Searle. 1969. Speech acts: An essay in the philosophy of language, volume 626. Cambridge university press.

Herbert P Grice. 1975. Logic and conversation. In Speech acts, pages 41–58. Brill

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2016. Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge.

IEEE transactions on pattern analysis and machine intelligence, 39(4):652–663.

Andrej Karpathy, Armand Joulin, and Li F Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. Advances in neural information processing systems, 27.