Evaluating Pragmatic Abilities of Image Captioners on A3DS

Polina Tsvilodub and **Michael Franke**
Department of Linguistics
University of Tübingen
{polina.tsvilodub, michael.franke}@uni-tuebingen.de

Abstract

Evaluating grounded neural language model performance with respect to pragmatic qualities like the trade off between truthfulness, contrastivity and overinformativity of generated utterances remains a challenge in absence of data collected from humans. To enable such evaluation, we present a novel open source image-text dataset "Annotated 3D Shapes" (A3DS) comprising over nine million *exhaustive* natural language annotations and over 12 million variable-granularity captions for the 480,000 images provided by Burgess and Kim (2018). We showcase the evaluation of pragmatic abilities developed by a task-neutral image captioner fine-tuned in a multi-agent communication setting to produce *contrastive* captions. The evaluation is enabled by the dataset because the exhaustive annotations allow to quantify the presence of contrastive features in the model's generations. We show that the model develops human-like patterns (informativity, brevity, over-informativity for specific features (e.g., shape, color biases)).

1 Introduction and Related Work

In human communication, language is rarely used as a unimodal channel; rather, language is mostly used in reference to the surroundings, i.e., it is *grounded* in the physical world. Thus, in order to build artificial agents that could be potentially employed in scenarios requiring natural communication with humans, it is crucial to develop approaches for training such agents to communicate about the world in a human-like way (Lake et al., 2017). However, automatically evaluating the human-likeness of a trained system without costly human feedback is a recurring problem in NLP.

In this paper, we set out to provide tools for evaluating human-like pragmatic abilities of grounded models and evaluate a model trained interactively via reinforcement learning, which is commonly suggested to give rise to task-oriented behavior (Lazaridou and Baroni, 2020).

Grounding of neural language models has been advanced greatly in recent years through *image captioning models*. Starting with the work by Vinyals et al. (2016) and Karpathy et al. (2014), neural encoder-decoder architectures have been dominating the field, recently extending to unified architectures (Zhou et al., 2020). However, these approaches are *task neutral*, i.e., the models are trained to produce generally true image captions.

In contrast, humans are highly flexible and *pragmatic* in their use of language and, e.g., adapt the granularity of their utterances to the requirements of the communicative task (Searle, 1969). It is generally guided by conversational maxims, suggesting that cooperative speakers should only provide as much information as required in a given context, be truthful, relevant, and brief (Grice, 1975). Therefore, faced with a simple referential task of picking out a target item among an array of distractors, humans tend to mention *contrastive* features of the target (e.g., Kramer and van Deemter, 2012), i.e., the ones setting it apart from distractors. On the other hand, *biases* towards producing shape and color descriptions even when these aren't contrastive have been identified (e.g., Degen et al., 2020). For grounded language models, the underlying pragmatic reasoning formalized as nested Bayesian inference about the behavior of speakers and listeners (Goodman and Frank, 2016) inspired decoding schemes applied on top of standardly trained models (e.g., Cohn-Gordon et al., 2018; Zarrieß et al., 2021; Shen et al., 2019; Vedantam et al., 2017; Andreas and Klein, 2016).

However, evaluating the pragmatic qualities of models' predictions when they are applied to specific tasks (e.g., referential tasks) remains a challenge. Currently standard metrics like BLEU-n, ROUGE, CIDEr and METEOR (Papineni et al., 2002; Banerjee and Lavie, 2005; Vedantam et al.,

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1277–1285
July 9-14, 2023 ©2023 Association for Computational Linguistics

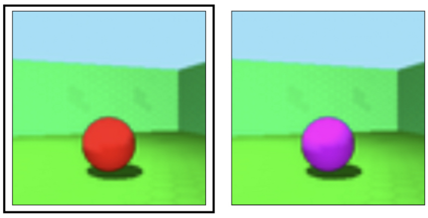

Figure 1: Example image pair matching on five features (left: red, right: purple ball), left image is target. Example exhaustive ground truth caption for target: "A tiny red ball near the right corner in front of a light green wall on green floor." Example short ground truth caption: "A ball on green floor." Contrastive caption predicted by RP model: "A tiny **red** ball green near the floor in green of".[1]

2015; Lin, 2004) for evaluating models' generations make reference to the surface form of ground truth image annotations. They cannot provide insight into models' mechanics and possible biases based on *context-dependent functional aspects* like mentioning contrastive features or being overinformative. Given that model predictions might not always be syntactically well-formed and yet still count as functionally expedient for a human (e.g., see Fig. 1), evaluating pragmatic aspects of natural language image captions is important. We propose a new dataset and metrics facilitating such evaluation in the next sections.

2 Methods

2.1 A3DS

To enable such evaluation, we provide novel annotations for the dataset 3DShapes (Burgess and Kim, 2018) (introduced in Kim and Mnih (2018)) in the "Annotated 3D Shapes" (A3DS) dataset. The image dataset consists of 480,000 unique images of 3D geometric objects, constructed by varying six features ($\times$number of distinct feature values): shape type ($\times$4), shape color ($\times$10), shape scale ($\times$8), shape orientation relative to the background ($\times$15), wall color ($\times$10) and floor color ($\times$10). For each image, two sets of ground truth captions were generated: *exhaustive* captions mentioning all six features and their values, and *short* captions, mentioning two or three features of the image only (see example annotation in Fig. 1). The captions were constructed with a hand-written grammar from the numeric labels shipped with the original dataset.

[1]The last token was predicted nine times. This shows how the caption can be contrastive for the given task inspite of surface form artefacts.

For each distinct feature value, different natural language descriptions were created. In total, over nine million exhaustive captions and 12 million short captions are released as part of this work.[2] The important advantage of this synthetic dataset for investigating referential language use of models trained on it is that the numeric labels allow to easily identify *contrastive* versus *redundant* features of the target image in any given context of distractor images. Furthermore, training with fully exhaustive captions allows to focus evaluations on models' contrastive abilities, excluding insufficient granularity of training data as a potential reason for a system's failure to be contrastive.

Because all natural language expressions for each label are known, it is possible to comprehensively evaluate model predictions by-feature. Predictions of fine-tuned models which may deviate from ground truth captions in their surface form (e.g., due to language drift; see, e.g., Lazaridou et al. (2020)) can also be evaluated. We consider a caption contrastive if at least one of the known contrastive features for a given context (target and distractors) is mentioned in the target's description. For contrastive color features, a caption is considered contrastive if it mentions the respective color irrespective of other mentioned aspects, if the color is unique for the target. If several features in the target image have the same color, the description is considered contrastive only if the color name occurs together with the correct head noun (e.g., "floor", "wall", object shape). For other contrastive features like shape, the respective expression (e.g., "ball", "in the left corner") has to literally occur in the generated caption. For the example, in Fig. 1, we were able to identify that the caption is contrastive because the contrastive feature is the red color of the ball in the target image (left), there is only one red feature in the target image, and the generated caption contains the term "red".

We suggest informative metrics for evaluating pragmatic abilities of models on this dataset in the next section.

2.2 Evaluation Metrics

The metrics are informed by notions that are considered important in the cognitive science literature for cooperative and efficient pragmatic communi-

[2]`https://tinyurl.com/2p8w6rct`. The repository also contains endpoints for running model evaluations described in the next section and a sandboxed version of the dataset and the pretrained model for easy exploration.

cation (e.g., Grice, 1975) and used commonly in the literature on computational generation of referring expressions (e.g., Kramer and van Deemter, 2012). In the context of a reference task, we define pragmatically relevant categories of features a model might mention. Given a target and distractor image, each feature falls in one of the following three categories:

- *Contrastive* feature: true of target and false of distractor.
- *Non-contrastive* feature: true of both the target and the distractor, and, therefore, redundant for the purpose of reference.
- *False* feature: false of the target.

From these categories, we derive the following metrics (higher values are better), where c is the number of contrastive features mentioned in a generated caption y, k is the total number of features mentioned in y, and z is the ground truth number of contrastive features between the images:

- *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).

In the next section, we showcase how these metrics can be applied in order to evaluate the development of pragmatic abilities of an image captioner through fine-tuning in an interactive setting.

2.3 Experiment

The multi-agent communication setting wherein the image captioner is trained as the sender agent together with an artificial receiver agent to complete a communicative task (e.g., reference game) allows to fine-tune the sender's captioning behavior based directly on task performance, e.g., via *deep reinforcement learning* (e.g., Lazaridou et al., 2020; Lazaridou and Baroni, 2020; Lazaridou et al., 2016; Havrylov and Titov, 2017), without making use of a supervised task specific dataset. Applied to the reference task, the idea is that the sender agent will learn to produce more contrastive descriptions which are helpful for the receiver to complete the task. Lazaridou et al. (2020) compare sender agent architectures in terms of their task-specific improvement, but they do not investigate properties like overinformativity that might have emerged during the multi-agent training.

To investigate these potenential effects, following the "multi-task learning" training regime from Lazaridou et al. (2020) we pretrained a *baseline* image captioner (B) on 150,000 image-exhaustive caption pairs constructed from 30,000 images sampled from A3DS. It was then fine-tuned on another 150,0000 pairs on a reference game together with a listener agent. In the reference game, both agents received concatenated pairs of images $i = [i_1; i_2]$, where $i_t, t \in \{1, 2\}$ was the target known only to the sender. The sender was trained to produce a description of the target, so that the listener guesses the target correctly, given the same images in randomized order. The sender received the reward $r = 1$ if the guess was correct, and $r = -1$ otherwise. Both the sender and the listener consisted of a pretrained ResNet-50 image encoder which was not fine-tuned during the reference game, and a trainable linear layer projecting the ResNet image features to 512-dimensional features. These were input into one-layer LSTM language modules with the hidden layer size $h = 512$. Further architectural and training details followed Lazaridou et al. (2020).[3]

We trained two sender-agent pairs in the reference game setting: in the *random pairs* setting (RP),

[3]The weight λ_s for the speaker loss was set to 0.75.

Score	one feature			two features			three features		
	B	**RP**	**SP**	**B**	**RP**	**SP**	**B**	**RP**	**SP**
Discriminativity	**0.999**	0.822	0.824	0.997	0.576	0.586	0.984	0.527	0.541
Contrastive efficiency	0.198	**0.879**	0.875	0.203	0.963	0.955	0.251	0.856	0.875
Relevance	0.150	0.668	0.640	0.162	0.522	0.521	0.149	**0.684**	0.665
Optimal contrastivity	0.014	**0.457**	0.452	0.039	**0.485**	0.476	0.148	0.335	0.367
Mentioned features #	5.880	2.944	3.125	5.871	2.950	3.133	5.876	2.955	3.135
Listener accuracy	—	0.919	0.895	—	0.887	0.900	—	0.862	0.860

Table 1: Pragmatic evaluation results by test set category for each model (B: pretrained baseline, RP: random pairs fine-tuning, SP: similar pairs fine-tuning), averaged across test sets within category. Bold numbers indicate best performance across models and test sets.

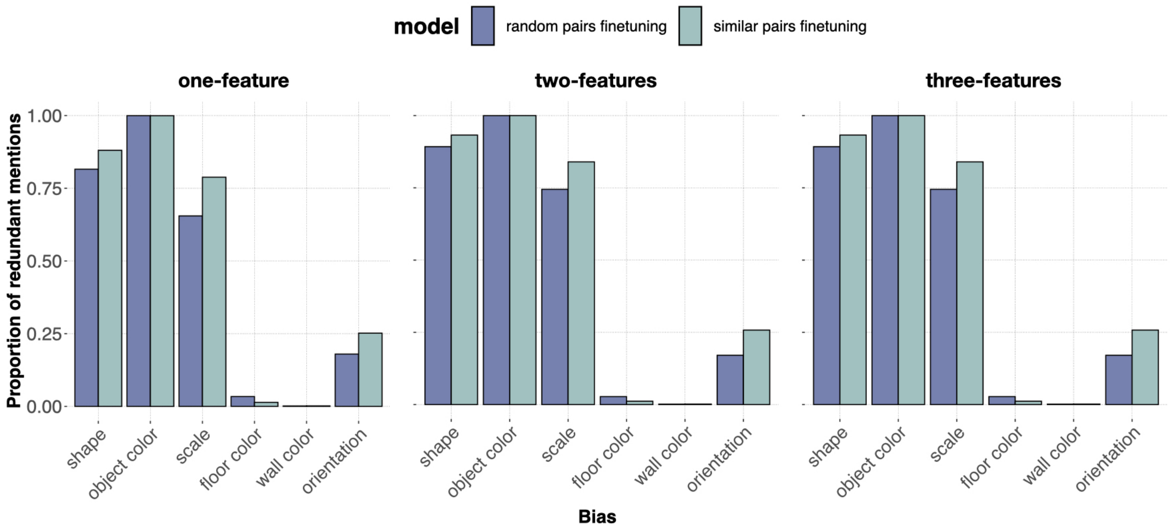

Figure 2: Generation proportions of each feature (x-axis) when it was *non-contrastive* for each model (color) by test category (facets). Generation proportions of all features for the baseline (not shown) are at ceiling on all test sets, except for the scale category being at around 0.9 due to a tokenization glitch.

the agents saw pairs of (distinct) images selected at random. In the *similar* pairs setting (SP), they received images which had at least three overlapping features (e.g., target and distractor depicted the same shape of the same color with background of the same color).[4]

3 Results

The agents were evaluated on three categories of test sets, each set containing 7500 image pairs. In the *one-feature* category, six sets were constructed where test pairs matched at least on one of each possible features. The *two-features* category included three sets of pairs matched on at least two object features and a set with two random matching features. The *three-features* category included sets where at least all object features, all background features, or three randomly sampled features matched. These sets allowed to evaluate in which conditions it was more difficult for the sender to produce appropriate captions. In the following, the fine-tuned sender models (**RP** and **SP**) are compared to the baseline model (**B**), which is the pretrained task-neutral image captioner. The average number of falsely named features was 0.720 for baseline, 0.139 (RP) and 0.316 (SP). Table 1 shows listener test accuracies on all test splits, showing that the agents successfully learned the reference task (0.5 is chance). In terms of discriminativity d, it was more difficult for the fine-tuned models to identify the correct feature when two or three features were identical across the pair (Table 1). These average difficulties were driven by the fail-

[4]The speaker included ≈5.3M, the listener ≈2.15M trainable parameters. Pretraining for 10 epochs took around 10h, fine-tuning for 5 epochs—20h/model on NVIDIA A40 GPU.

ure on test sets where the non-contrastive features included shape (e.g., a pair showing a red vs. a blue block), indicating that the shape was easiest to pick up on for the models, although all features were mentioned in all training captions. For instance, d was 0.750 for SP on the object color-scale matched test set, and 0.724 on the random two-feature test set, but 0.501 on the shape-object color matched set. The discriminativity on random and background feature matched three-feature test sets was 0.618 | 0.875 (RP) and 0.854 | 0.605 (SP), while it was only 0.087 (RP) and 0.164 (SP) on the object feature matched test set.

The better contrastive performance of the baseline came at a cost of generally overmodifying the messages with contrastive features (see low contrastive efficiency, Table 1). Low relevance scores also show that the baseline did not identify functionally appropriate features well. In contrast, both fine-tuned models showed higher contrastive efficiency and relevance, indicating that the task based fine-tuning might have helped the models to learn contrastiveness. The fine-tuned models also showed higher optimal constrastivity which is, however, still far from perfect. In general, no qualitative differences between the two- and three-feature datasets or RP and SP settings are apparent.

Figure 2 shows how frequently the models' predictions mentioned a specific feature when it was contrastively *irrelevant* (i.e., it zooms in on predictions where $r < 1$). For the fine-tuned models, it suggests potential biases towards redundantly producing object-related features (shape, scale, color of object), matching human biases (see Section 1), as opposed to background descriptions. The proportions slightly increase for object color and scale in the two- and three-feature test sets, potentially hinting at overmodification as the model's loophole behavior in a more complex setting. The SP model has a stronger redundancy propensity than RP. The apparent trend towards mentioning shape is in line with the pattern of discriminativity results described above where models relied on the shape being the discriminative feature between target and distractor.

4 Conclusion

We provide the A3DS dataset alongside evaluation metrics for investigating referential pragmatic abilities acquired by grounded language models on this dataset. With this dataset, we identify that an image captioner fine-tuned interactively via reinforcement learning developed a strikingly human-like shape bias, while being less overinformative than a task-neutral model. Future research could expand such evaluations by including metrics which investigate additional aspects that might matter to human referential expression generation (e.g., the current metrics are agnostic to the surface order of discriminative features, while humans have preferences towards certain adjective ordering; Scontras et al. (2017)). Although these results are specific to the given architecture, with this work we hope to inspire research opening up black box language models—an important task in the age of LLMs.

Limitations

The identified tendencies towards mentioning object-related features and the reliance on the shape as a contrastive feature might be driven by the grammatical structure of the annotations, mostly presenting object features in sentence-initial subject position, although 40% of exhaustive captions mention either the scale or the object color as the last word in the sentence. Therefore, these results call for investigating the biases of model architectures less sensitive to sentence length than LSTMs, as well as extending the annotations with additional grammars. Further, this evaluation provides descriptive results of the models' pragmatic abilities, leaving the question of whether it is indeed a pragmatic inductive bias or, e.g., structural language drift (Lazaridou et al., 2020) causing the observed patterns, unanswered. Finally, since the evaluation pertains to the surface form of the predictions, applying decoding schemes other than greedy decoding used in this work might provide different patterns, indicating to which degree potential biases are due to model mechanics in opposition to sampling parameters.

Acknowledgements

We would like to thank Elia Bruni for his support of the work which led to this paper, and Xenia Ohmer and Leon Schmid for helpful discussions. We also acknowledge support by the state of Baden-Württemberg through the computing resources provided by bwHPC and the German Research Foundation (DFG) through grant INST 35/1597-1 FUGG. Michael Franke is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 39072764.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Chris Burgess and Hyunjik Kim. 2018. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/.

Reuben Cohn-Gordon, Noah Goodman, and Christopher Potts. 2018. Pragmatically informative image captioning with character-level inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 439–443, New Orleans, Louisiana. Association for Computational Linguistics.

Judith Degen, Robert D Hawkins, Caroline Graf, Elisa Kreiss, and Noah D Goodman. 2020. When redundancy is useful: A Bayesian approach to "overinformative" referring expressions. *Psychological Review*, 127(4):591.

Noah D Goodman and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11):818–829.

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

Serhii Havrylov and Ivan Titov. 2017. Emergence of language with multi-agent games: Learning to communicate with sequences of symbols. *Advances in neural information processing systems*, 30.

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.

Hyunjik Kim and Andriy Mnih. 2018. Disentangling by factorising. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR.

Emiel Kramer and Kees van Deemter. 2012. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218.

Brenden M. Lake, Tomer D. Ullman, Joshua B. Tenenbaum, and Samuel J. Gershman. 2017. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253.

Angeliki Lazaridou and Marco Baroni. 2020. Emergent multi-agent communication in the deep learning era. *ArXiv*, abs/2006.02419.

Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. 2016. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*.

Angeliki Lazaridou, Anna Potapenko, and Olivier Tieleman. 2020. Multi-agent communication meets natural language: Synergies between functional and structural language learning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7663–7674, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Gregory Scontras, Judith Degen, and Noah D Goodman. 2017. Subjectivity predicts adjective ordering preferences. *Open Mind*, 1(1):53–66.

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

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically informative text generation. 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 4060–4067, Minneapolis, Minnesota. Association for Computational Linguistics.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 251–260.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

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.

Sina Zarrieß, Hendrik Buschmeier, Ting Han, and Simeon Schüz. 2021. Decoding, fast and slow: A case study on balancing trade-offs in incremental, character-level pragmatic reasoning. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 371–376, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☒ A2. Did you discuss any potential risks of your work?
The work is confined to a synthetic dataset depicting and describing geometric objects, such that the used tools cannot be directly applied to socially relevant scenarios which might pose risks. The used architectures were light-weight, such that training theoretically doesn't require resources beyond modern laptop hardware.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
abstract, section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

2.1

☑ B1. Did you cite the creators of artifacts you used?
abstract, 2.1

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
The license and terms are provided in the online repository released with the paper. The original data source distributed the data under Apache 2.0 allowing reuse.

☒ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Given the original Apache 2.0 license, the original data allows both research, private and commercial use, therefore not imposing any limitations. The submitted online repository providing the data provides the same conditions for the derivatives.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
The data only contains descriptions of abstract synthetically generated geometric shapes.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
2.1

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Sections 2.3., 3 provide train/test split set sizes and construction statistics.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ **Did you run computational experiments?**

2.3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
2.3

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No hyperparameter search was conducted. Since the computational experiment architecture replicates existing cited work, parameters reported there or single selected parameters were used.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
3

☒ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
The used Spacy model is reported in the supplementary online repository documentation exposing the newly created resource.

D ☒ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.