Communication breakdown: On the low mutual intelligibility between human and neural captioning

Roberto Dessì
Meta AI (FAIR)
Universitat Pompeu Fabra
rdessi@meta.com

Eleonora Gualdoni and **Francesca Franzon**
Universitat Pompeu Fabra
{name.lastname}@upf.edu

Gemma Boleda and **Marco Baroni**
ICREA
Universitat Pompeu Fabra
{name.lastname}@upf.edu

Abstract

We compare the 0-shot performance of a neural caption-based image retriever when given as input either human-produced captions or captions generated by a neural captioner. We conduct this comparison on the recently introduced IMAGECODE data-set (Krojer et al., 2022), which contains hard distractors nearly identical to the images to be retrieved. We find that the neural retriever has much higher performance when fed neural rather than human captions, despite the fact that the former, unlike the latter, were generated without awareness of the distractors that make the task hard. Even more remarkably, when the same neural captions are given to human subjects, their retrieval performance is almost at chance level. Our results thus add to the growing body of evidence that, even when the "language" of neural models resembles English, this superficial resemblance might be deeply misleading.

1 Introduction

Neural vision-and-language models have achieved impressive results in tasks such as visual commonsense reasoning and question answering (e.g., Chen et al., 2019; Lu et al., 2019). However, Krojer et al. (2022) recently showed, in the context of caption-based image retrieval, that state-of-the-art multimodal models still perform poorly when the candidate pool contains very similar distractor images (such as close frames from the same video).

Here, we show that, when the best pre-trained image retrieval system of Krojer et al. (2022) is fed captions produced by an out-of-the box neural caption generator, its performance makes a big jump forward. 0-shot image retrieval accuracy improves by almost 6% compared to the highest previously reported human-caption-based performance by the same model, with fine-tuning and various *ad-hoc* architectural adaptations. This is remarkable, because the off-the-shelf caption generator we use (unlike the humans who wrote the original captions in the data-set) is *not* taking the set of distractor images into account. Even more remarkably, we show that, when human subjects are tasked with retrieving the right image using the same neural captions that help the model so much, their performance is only marginally above chance level.

2 Setup

Data We use the more challenging *video* section of the IMAGECODE data-set (Krojer et al., 2022). Since we do not fine-tune our model, we only use the validation set, including 1,872 data points.[1] Henceforth, when when we employ the term IMAGECODE, we are referring to this subset. Each data-point consists of a target image and 9 distractors, where the target and the distractors are frames from the same (automatically segmented) scene in a video. We also use the human captions in the data-set, that were produced by subjects that had access to the distractors while annotating each target (they were instructed to take distractors into account, without explicitly referring to them). Having access to this "common ground" (Brennan and Clark, 1996), annotators produced highly context-dependent descriptions (see example human captions in Fig. 1). The data-set contains one single caption per image.

Neural caption generation We use the ClipCap caption generation system (Mokady et al., 2021) without fine-tuning. For details and hyperparameters of the generation process see Appendix A. In short, ClipCap processes an image with a CLIP visual encoder (Radford et al., 2021) and learns a mapping from the resulting visual embedding to a sequence of embeddings in GPT-2 space (Radford et al., 2019), that are used to kickstart the generation of a sequence of tokens. We report experiments with the ClipCap variant fine-tuned on the COCO

[1] We use the validation set because IMAGECODE test set annotations are not publicly available.

setup	acc
neural captions, 0-shot	27.9
human captions, 0-shot	17.4
human captions, Krojer et al's best	22.3

Table 1: Percentage IMAGECODE accuracy of 0-shot image retriever when given neural vs. human captions as input. Last row reports accuracy of best fine-tuned, architecturally-adjusted model from Krojer et al. (2022) (featuring a context module, temporal embeddings and a ViT-B/16 backbone).

data-set (Lin et al., 2014), where the weights of the multimodal mapper were updated and those of the language model (GPT-2) were kept frozen. We obtained very similar results with the other publicly available ClipCap variants. We generate a single *neural caption* for each IMAGECODE target image by passing it through ClipCap. Note that, as there is no way to make this out-of-the-box architecture distractor-aware, the neural captions do not take distractors into account.

Image retrieval We use the simplest CLIP-based retrieval system of Krojer et al. (2022) (the one without context module and temporal embeddings), which corresponds to a standard CLIP architecture from Radford et al. (2021). The caption and each image in the set are passed through a transformer-based text encoder and a transformer-based visual encoder, respectively. Retrieval is successful if the dot product between the resulting caption and target image representations is larger than that of the embedded caption with any distractor representation. We use the ResNet-based CLIP retriever (He et al., 2015), whereas Krojer et al. (2022) used the ViT-B/16 architecture, since we found the former having a higher retrieval accuracy compared to what they used (17.4% in Table 1 here vs. 14.9% in their paper).

3 Results and analysis

Neural vs. human caption performance As shown in Table 1, the out-of-the-box neural image retrieval model has a clear preference for neural captions. It reaches 27.9% IMAGECODE accuracy when taking neural captions as input, vs. 17.4% with human captions (chance level is at 10%). For comparison, the best fine-tuned, architecture-adjusted model of Krojer et al. (2022) reached 22.3% performance with human captions.

A concrete sense of the differences between the two types of captions is given by the examples in Fig. 1. The examples in this figure are picked randomly. Based on manual inspection of a larger set, we are confident they are representative of the full data. Clearly, neural captions are shorter (avg. length at 11.4 tokens vs. 23.2 for human captions) and more plainly descriptive (although the description is mostly only vaguely related to what's actually depicted). Since there is no way to make the out-of-the-box ClipCap system distractor-aware, the neural captions are not highlighting discriminative aspects of a target image compared to the distractors. Human captions, on the other hand, use very articulated language to highlight what is unique about the target compared to the closest distractors (often focusing on rather marginal aspects of the image, because of their discriminativeness, e.g., for the first example in the figure, the fact that the blue backpack is hardly visible). It is not surprising that a generic image retriever, that was not trained to handle this highly context-based linguistic style, would not get much useful information out of the human captions. It is interesting, however, that this generic system performs relatively well with the neural captions, given how off-the-mark and non-discriminative the latter typically are.

As more quantitative cues of the differences between caption types, we observe that human captions are making more use of both rare lemmas and function words (see frequency plots in Appendix B).[2] Extracting the lemmas that are statistically most strongly associated to the human caption set (see Appendix C for method and full top list), we observe "meta-visual" words such as *visible* and *see*, pronouns and determiners cuing anaphoric structure (*the*, *her*, *his*), and function words signaling a more complex sentence structure, such as auxiliaries, negation and connectives. Among the most typical neural lemmas, we find instead general terms for concrete entities such as *people*, *woman*, *table* and *food*.

Are neural captions really discriminative? By looking at Figure 1, we see that neural captions might be (very noisily) descriptive of the target, but they seem hardly discriminative with respect to the nearest distractors. Recall that each IMAGECODE set contains a sequence of 10 frames from the same scene. In general, the frames that are farther away

[2]Code to reproduce our analysis with human and model-generated captions is available at https://github.com/franfranz/emecomm_context

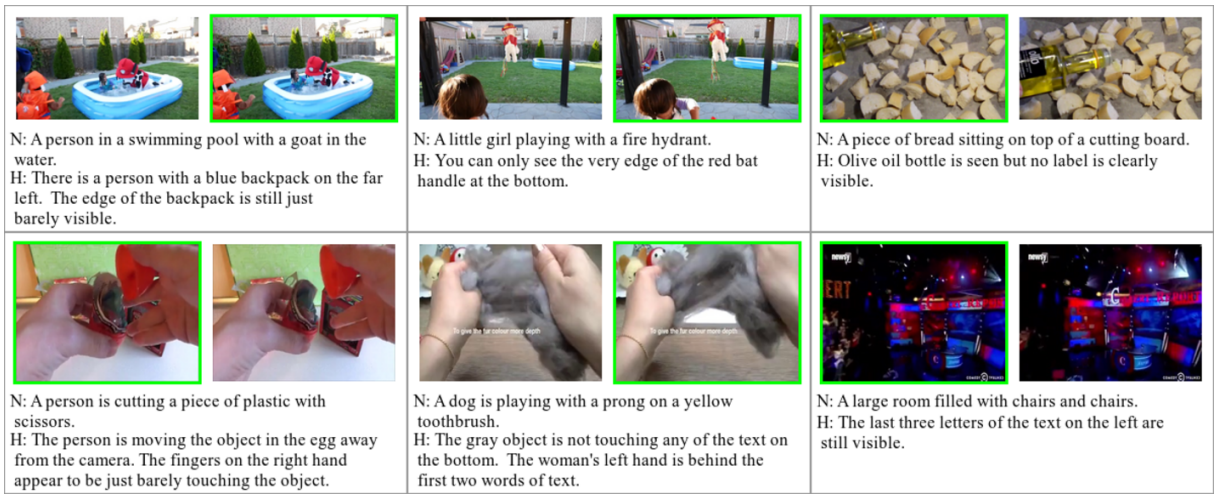


Figure 1: Randomly selected data points from IMAGECODE with neural (N) and human (H) captions, where, given the neural caption, the neural retriever guessed the target and the human retriever failed. For each candidate set, we show the target (marked in green) and (randomly) either the immediately preceding or following distractor frame. Appendix E reports the whole candidate sets for each data-point (10 images per set), as well as providing details on the random selection process.

in time might be easier to discriminate than the closest ones (consider the full candidate set examples in Appendix E: it is in general much easier to tell apart the first and last frames than two adjacent images). It could be, then, that the non-random but still low performance of the image retriever with neural captions is due to the combination of two factors. On the one hand, the neural captions might suffice for the retriever to exclude the farthest distractors. On the other, its performance at telling the closest frames apart is actually random.

To rule out this explanation, we repeated the retrieval experiment in the most challenging setup, in which we excluded all but the distractors immediately preceding and following the target frame (if the target is the first/last image, we pick the two frames following/preceding it, respectively). The retriever using neural captions still reaches 48.7% accuracy, well above chance level (33.3%). We must thus conclude that neural captions such as those in Fig. 1 do carry a non-negligible degree of discriminative power for a neural image retriever.

On a related point, neural captions such as those in the figure seem so generic that one could imagine the neural caption generator would produce the same caption for close-by frames. This is not the case. We consider the 678 IMAGECODE cases in which a candidate set is repeated across multiple data points, with only the target changing, and in which the targets are adjacent or one-frame apart. In 93.7% of such cases, ClipCap generated at least two non-identical captions. For example, the frame on the left of the center-bottom pair in Fig. 1 is also used as a target, and for it ClipCap generates the caption "A person holding a water bottle with a dog in it." Looking at the right-bottom pair, the frame on the right is also a target, and ClipCap generated the following caption for it: "A picture of a kitchen with a bunch of televisions on it." Future research should ascertain to what extent these intuitively uninformative variations in frame description actually contain cues that are systematically discriminative for the retrieval system.

Human performance on neural captions Looking again at examples such as those in Fig. 1 (all cases in which the image retriever correctly identified the target), we might conjecture that the neural captions are more informative for the neural retriever model than they are for humans. To verify this hypothesis, we organized a crowd-sourcing experiment in which human subjects had to perform the same 10-image-set target discrimination task we submitted to the neural image retriever.

More precisely, we selected a subset of IMAGECODE that is balanced in terms of image retriever performance as follows. We used all 522 sets where the retriever guessed the target, and we randomly added the same number of sets where the retriever got it wrong (so that its accuracy, on this subset, is at 50% by construction). We collected human discrimination decisions for this subset of 1,044 items, when either human or neural captions

retriever \ captions	human	neural
human	54.3	12.8
neural	16.3	50.0

Table 2: Percentage accuracy on an IMAGECODE subset (balanced to get 50% accuracy of the 0-shot neural retriever with neural captions): human vs. neural retrievers tested with neural vs. human captions as inputs.

are given as input. We collected one rating per item-caption combination from a total of 36 Amazon Mechanical Turk[3] participants, that each provided a total of 58 ratings. Experimental details are given in Appendix D.

On the relevant IMAGECODE subset, humans clearly outperform the neural retriever when human captions are given: 54.3% human discrimination accuracy vs. 16.3% for the neural retriever.[4] Strikingly, the pattern sharply reverses with neural captions: 50.0% for the neural image retriever vs. 12.8% for humans (not much above the 10% random baseline).

We thus confirm that neural captions such as those presented in Fig. 1, despite being apparently vague and inaccurate descriptions of the target image "in plain English", carry significantly more discriminative value for the neural retriever than they do for humans (recall that the examples in this figure were selected among the cases where the neural caption allowed the neural retriever to guess the right target, while human subjects failed the task).

4 Conclusion

Previous research has shown that neural caption generators occasionally produce highly counterintuitive or irrelevant image descriptions (e.g., Lake et al., 2017; Rohrbach et al., 2018). We provide here evidence that such descriptions might only be misleading or uninformative for humans, while still being relatively "understandable" to neural models. We discuss below the **Limitations** that delimit the scope of our finding. Still, we can tentatively conclude that, even when they are trained on English, deep nets might pack and retrieve information from token sequences that are different from those an English speaker would encode in and extract from them.

A better understanding of this behaviour could help design higher-performance systems. For example, we could implement a module translating human captions to the "machine code" that neural models prefer, leading to better caption-based retrieval; or, from a model-to-model communication perspective (Zeng et al., 2022), optimize caption generation directly for neural model understanding, instead of imitating human captions. Similar ideas have recently proposed in the context of textual information retrieval (e.g., Haviv et al., 2021; Shin et al., 2020).

From a less optimistic perspective, our results can be interpreted as another cautionary tale about the degree to which neural models truly "understand language" (Webson and Pavlick, 2022), and suggest that a good grasp of their counter-intuitive behaviour should be a priority of current research, or else malicious agents could rely on the models' opaque behaviour for adversarial attacks (Wallace et al., 2019) and other types of model misuse.

Limitations

The results we presented are limited to one specific data-set tested with a single caption generator and image retriever pair (with both systems relying on the CLIP image encoder). Future work should verify whether they generalize to other neural model combinations and data-sets.

We observe a considerable increase in accuracy when the neural image retriever is fed machine-generated captions instead of human ones. However, accuracy is still at 27.9%, suggesting that the retrieval system has only a very partial understanding of captions, whether machine- or human-produced. How to improve its performance remains a question for future work. In the current setup, the caption generation system, unlike human annotators, only receives the target image as input, and it is unaware of the distractors. Making the caption generation system distractor-aware (perhaps taking inspiration from work on "image difference captioning", e.g., Guo et al., 2022) might improve the performance of the image retriever. Distractor-aware neural captions would also be more fairly comparable to the distractor-aware human captions we got from the IMAGECODE data-set.

Last but not least, we provided evidence that captions that carry virtually no discriminative infor-

[3] https://www.mturk.com/

[4] See Appendix D for discussions of why our human-to-human discrimination accuracy is considerably lower than that reported by Krojer et al. (2022) on the whole IMAGECODE data-set.

mation for humans are instead helping the neural retriever identify target images well above chance level. We still lack, however, an understanding of *how* the neural retriever accomplishes this surprising feat: developing such an understanding is perhaps the most important direction for future work.

Ethics Statement

We rely on existing, publicly available data-sets (Das et al., 2013; Krojer et al., 2022; Li et al., 2019; Xu et al., 2016; Lin et al., 2014) and pre-trained models (Mokady et al., 2021; Radford et al., 2021).

We re-normed a subset of the data used by Krojer et al. (2022) using crowdsourcing. The experiment was approved by the ethical board of Universitat Pompeu Fabra in the context of the AMORE project (grant agreement No. 715154). Participants had to agree to an informed consent form before doing the experiment, and they were allowed to leave it at any time. No personal data were collected, except for the participants' AMT worker IDs, needed for their payment. They were paid 12.5$ for completing the task (that took about 20 minutes). The crowdsourcing experiment procedure is described in more detail in Appendix D.

As we only run zero-shot experiments with pre-trained models, compute usage is negligible.

Our research contributes to an expanding body of evidence showing that, while pre-trained deep models are apparently responding to natural language prompts, their "language" might differ from human language (e.g., Lu et al., 2022; Shin et al., 2020; Webson and Pavlick, 2022). Understanding this gap between human and machine language is important, in order to improve human-machine interaction, but also because it can be exploited for harmful purposes, such as adversarial attacks (Wallace et al., 2019).

Acknowledgements

We would like to thank Michele Bevilacqua for comments on an earlier draft and the FAIR conference participants for fruitful discussion. We thank the EMNLP area chair and anonymous reviewers for feedback.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreements No. 715154 and No. 101019291) and the Spanish Research Agency (ref. PID2020-112602GB-I00). This paper reflects the authors' view only, and the funding agencies are not responsible for any use that may be made of the information it contains.

References

Harald Baayen. 2001. *Word Frequency Distributions*. Kluwer, Dordrecht, The Netherlands.

Susan Brennan and Herbert Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. UNITER: UNiversal Image-TExt Representation learning. In *Proceedings of ECCV*, pages 104–120, virtual conference.

Pradipto Das, Chenliang Xu, Richard Doell, and Jason Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of CVPR*, pages 2634–2641, Portland, Oregon.

Stefan Evert. 2005. *The Statistics of Word Cooccurrences*. Ph.D dissertation, Stuttgart University.

Zixin Guo, Tzu-Jui, Julius Wang, and Jorma Laaksonen. 2022. CLIP4IDC: CLIP for image difference captioning. https://arxiv.org/abs/2206.00629.

Adi Haviv, Jonathan Berant, and Amir Globerson. 2021. BERTese: Learning to speak to BERT. In *Proceedings of EACL*, pages 3618–3623, Online.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition.

Benno Krojer, Vaibhav Adlakha, Vibhav Vineet, Yash Goyal, Edoardo Ponti, and Siva Reddy. 2022. Image retrieval from contextual descriptions. In *Proceedings of ACL*, pages 3426–3440, Dublin, Ireland.

Brenden Lake, Tomer Ullman, Joshua Tenenbaum, and Samuel Gershman. 2017. Building machines that learn and think like people. *Behavorial and Brain Sciences*, 40:1–72.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan Kankanhalli. 2019. Video storytelling: Textual summaries for events. *IEEE Transactions on Multimedia*, 22(2):554–565.

Tsung-Yi Lin, Michael Maire, Serge 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*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of NeurIPS*, Vancouver, Canada. Published online: https://papers.nips.cc/paper/2019.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of ACL*, pages 8086–8098, Dublin, Ireland.

Ron Mokady, Amir Hertz, and Amit Bermano. 2021. ClipCap: CLIP prefix for image captioning. https://arxiv.org/abs/2111.09734.

Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. 2019. PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, 51(1):195–203.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, pages 8748–8763, virtual conference.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of EMNLP*, pages 4035–4045, Brussels, Belgium.

Taylor Shin, Yasaman Razeghi, Robert Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of EMNLP*, pages 4222–4235, virtual conference.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of EMNLP*, pages 2153–2162, Hong Kong, China.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of NAACL*, pages 2300–2344, Seattle, WA.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of CVPR*, pages 5288–5296, Sydney, Australia.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. https://arxiv.org/abs/2204.00598.

A Neural caption generation details

In our experiments, we use the pre-trained ClipCap caption generation model from Mokady et al. (2021), which employs a Transformer mapper trained on the COCO data-set (Lin et al., 2014) while the CLIP image encoder (Radford et al., 2021) and the GPT-2 language model (Radford et al., 2019) are frozen. In the ClipCap architecture, the mapper projects a CLIP-extracted embedding into the multidimensional space of GPT-2 word embeddings to trigger image-conditioned text generation. ClipCap has two architectural variants, one that uses a Transformer-based mapper and one that employs an MLP-based mapper. We refer to Mokady et al. (2021) for a detailed description of the MLP variant. The model that uses a Transformer mapper extracts a visual embedding from a pre-trained CLIP image encoder and feeds such representation together with a set of learned constant embeddings into GPT-2.

To produce a caption, we generate text using beam search with 5 beams, without tuning this value, and retaining the single maximum likelihood sequence. We set a maximum caption length of 67 tokens. Given that neural captions have an average length of around 11 tokens, it is unlikely that this limit is of any practical import. Additionally, the pre-trained CLIP text encoder from Radford et al. (2021), which both Krojer et al. (2022) and we use, cannot process contexts larger than 75 tokens, and thus extra tokens would be ignored in any case.

B Caption frequency distribution analysis

We tokenize, part-of-speech tag and lemmatize human and neural captions with Spacy.[5] We use the resulting part-of-speech and lemma sequences to compute the statistics reported in this Appendix and in Appendix C.

We counted the occurrences of the different parts of speech, normalized over the total amount of produced tokens, in both human and neural captions (Fig. 2). Their distribution reveals that, unsurprisingly, both types of caption mostly use nouns to

[5] https://spacy.io/

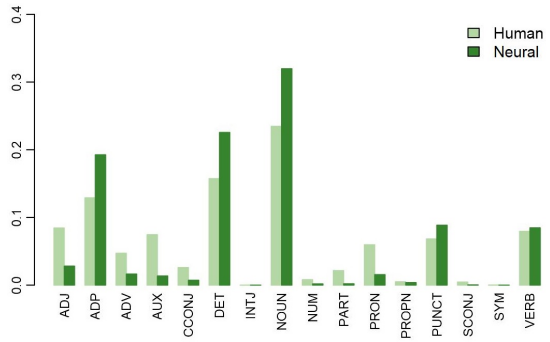


Figure 2: Part of Speech frequency distribution in human and neural captions.

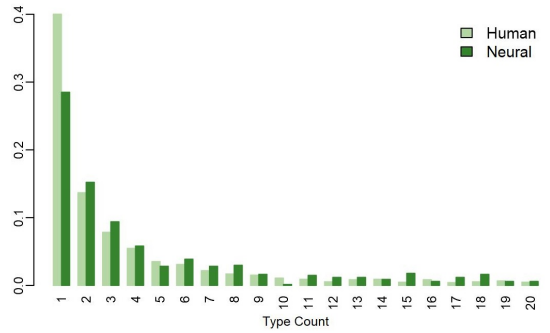


Figure 3: Lemma frequency spectrum in human and neural captions (only frequency of first 20 counts shown). The x-axis represents an occurrence count, the y-axis the number of distinct lemmas with that count in the captions, normalized over the total amount of distinct lemma types.

denote the entities presented in the images, but humans tend to modify them with remarkably higher adjective usage. Human-generated captions also display more functional words, pointing to the higher syntactic complexity already suggested by sentence length.

We computed the frequency spectrum (Baayen, 2001) of lemma types occurring in the two sets of captions. The normalized count of distinct lemmas with caption occurrence from 1 to 20 are plotted in Fig. 3. Human captions make a larger use of lemmas occurring only once, displaying a clear Zipfian trend. This trend is also present, but much less pronounced, in the neural captions.

C Neural vs. human caption lemma analysis

We use the Local Mutual Information score (Evert, 2005) to extract lemmas that are most significantly associated with neural vs. human captions, based on their relative frequency of occurrence in the two sets.

Top 20 most typical lemmas of neural caption set (min LMI: 124.2): *a, stand, of, in, next, hold, people, woman, group, front, on, sit, man, person, table, with, food, couple, cell, phone.*

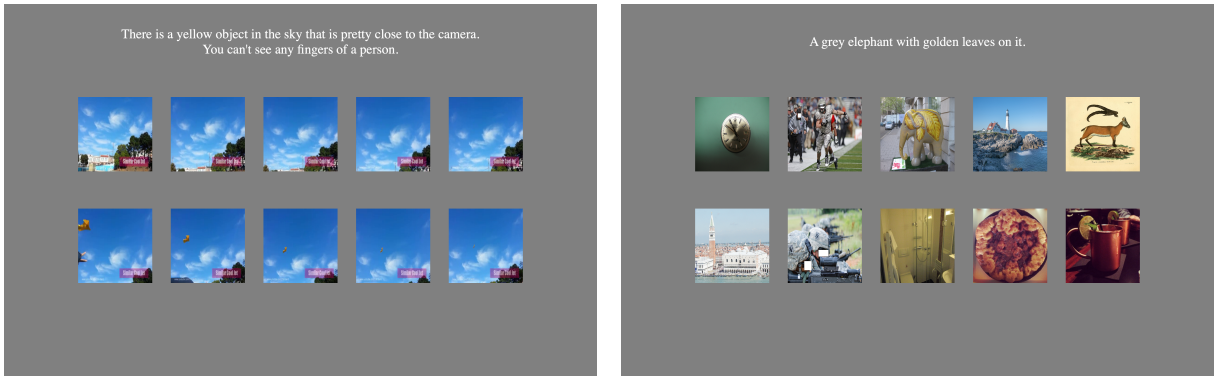
Top 20 most typical lemmas of human caption set (min LMI: 89.7): *the, be, right, see, left, can, and, 's, visible, you, her, have, hand, his, not, there, at, face, but, just.*

Besides looking at lemmas most typical of each set, we explore whether there is some non-trivial degree of overlap between the words occurring in the neural vs. human captions for each target (excluding stop words and punctuation, that would artificially increase overlap). We find that the average lemma overlap, measured as intersection-over-union (IOU), is at 5.2% (st. dev.: 6.6%). This might look non-negligible, but it does not significantly differ from random overlap according to a permutation test.

D Crowdsourcing experiment details

We populated the IMAGECODE stimulus subset for the human retrieval experiment as follows. We took all 522 candidate sets where the neural retriever guessed the target from the IMAGECODE *video* section. We further sampled without replacement the same amount of cases from the sets that the retriever got wrong (thus obtaining a balanced sample where the neural retriever accuracy is at 50%). We presented to subjects these 1,044 sets with both the human captions from Krojer et al. (2022) and the captions produced by our neural caption generation system. This resulted in 2,088 questions posed to subjects.

We randomly divided the entire set into 36 blocks of 58 questions (always containing both neural and human captions in similar amounts). In each screen, the 10 images from a set were presented at the center, arranged in two arrays of 5 images, with the caption written above–see Fig. 4a. Participants were asked to click on the image that matched the caption best. They were shown one example before starting the task. They were also warned that some cases could be more challenging than others. We asked them to always reply with the answer they found most plausible. Finally, they were warned that the experiment contained some control items, used to ensure annotation quality.



(a) (b)

Figure 4: Examples of screens shown to the participants. In panel (a), a set with a human caption; In panel (b), an attention check.

Each subject was presented with one block of questions, plus 5 randomly placed controls, designed to ensure that annotators were paying attention to the task. These cases were made intentionally very simple: targets were surrounded by random distractors, i.e., images that were neither contextually relevant nor very similar to the target. Targets and distractors for the attention checks were extracted from the less challenging *static* section of the IMAGECODE data-set (Krojer et al., 2022). We made sure internally that these sets could be easily processed with 100% retrieval accuracy. See Fig. 4b for an example.

The data collection routine was written in Psychopy (Peirce et al., 2019) and launched through Pavlovia.[6] There was no time limit for completing the study.

We recruited participants via Amazon Mechanical Turk.[7] We only accepted annotators from the US, with HIT approval rate higher than 89% and number of approved HITs higher than 1,000. We informed them that we would not collect any personal data (except for their workerID, that we would not make public), and that the goal of the experiment was to study how well people identify images based on descriptions. Before being able to access the link of the experiment, participants had to complete an informed consent form. They were able to quit the experiment at any time. We paid them 12.5$ for completing the task. The experiment was approved by the ethical board of Universitat Pompeu Fabra in the context of the AMORE project (grant agreement No. 715154).

[6] https://pavlovia.org/

[7] https://www.mturk.com/

We excluded the data of participants that made more than one mistake when scoring the controls, suggesting that they were not paying enough attention to the task. After a first round of data collection, we computed mean accuracy and standard deviation on human captions (without looking at neural caption performance). To further filter out low-quality trials, we removed participants with human caption accuracy more than one standard deviation below the mean, again suggesting scarce attention to the task. This resulted in 6 participants being removed, with the corresponding data being collected again. The boxplot in Fig. 5 shows the distribution of accuracy on human and neural captions for our final 36 participants. All participants reached well-above-chance accuracy on human captions, with a clear contrast with respect to their neural caption performance (the worst performance on human captions is comparable to the best performance on neural captions).

Our final cumulative human accuracy on human captions is considerably lower than the one reported by Krojer et al. (2022) for the whole IMAGECODE collection (54.3% vs. 90%). We conjecture that this is due in part to the fact that our items only come from the more challenging IMAGECODE *video* subset, and in part to the fact that their two-stage data-collection setup allowed a subject white-listing procedure we could not implement. Still, three authors performed the caption retrieval task, with resulting accuracies at 52%, 56% and 76%, respectively. The performance distribution of this supposedly "high-quality" annotators is comparable to the one of the 36 crowd-sourced participants, suggesting that the low overall accu-

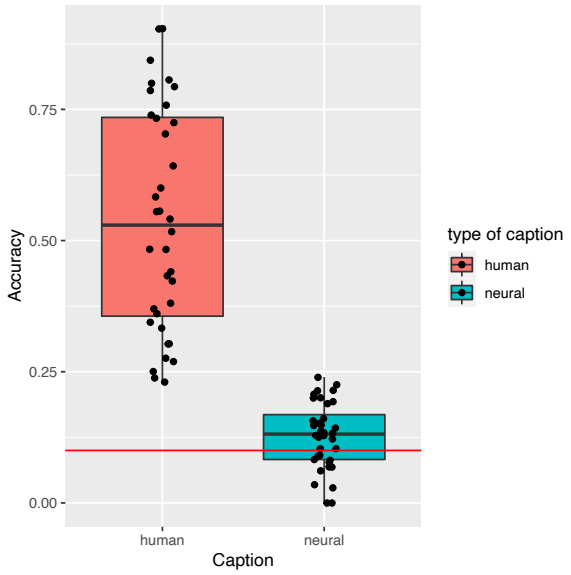


Figure 5: Accuracy of our 36 participants on human vs. neural captions. Each point represents one participant. The red line represents chance level.

racy is genuinely due to the difficulty of the task, and not to poor quality control.

E Fig. 1 example selection method and full candidate sets

The examples in Fig. 1 were randomly selected among trials in which, given the neural caption, the neural model guessed the target and the human annotators missed it. We avoided re-sampling the same candidate set more than once. We also discarded images displaying identifiable persons or large portions of text.

In each example, the target image is presented with a distractor, which can be the frame immediately preceding the target or the frame following it in the original sequence. The choice to show the preceding vs. following distractor frame was random.

In Fig. 6, we report the full sequences of distractors of each selected set, with the target marked in green, and the corresponding captions produced by the neural model (N), and by humans (H).

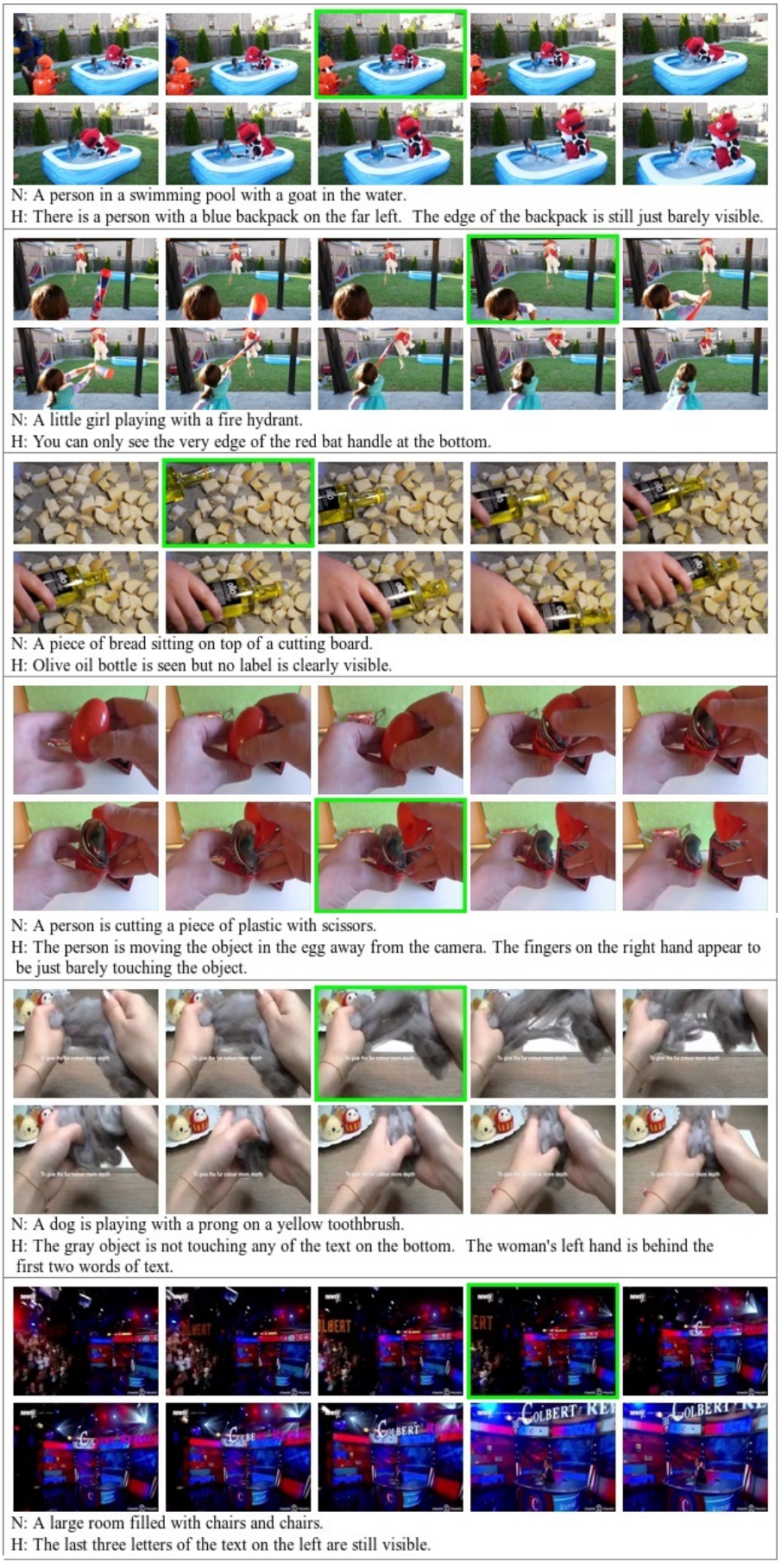


Figure 6: Whole candidate sets for each example in Fig. 1. The target image is marked in green.