Probing Representations of Numbers in Vision and Language Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

The ability to represent and reason about numbers in different contexts is an important aspect of human and animal cognition. Literature in numerical cognition posits the existence of two number representation systems: one for representing small, exact numbers, which is largely based on visual processing, and another system for representing larger, approximate quantities. In this work, we investigate number sense in vision and language models by examining learned representations and asking: What is the structure of the space representing numbers? Which modality contributes mostly to the representation of a number? While our analyses reveal that small numbers are processed differently from large numbers, as in biological systems, we also found a strong linguistic contribution in the structure of number representations in vision and language models, highlighting a difference between representations in biology and artificial systems.

1 Introduction

Whether it is foraging for food in novel environments, or counting slices of cake at a birthday party, humans and animals daily demonstrate various aspects of numerical competence. Reasoning about quantities is a fundamental cognitive feature, and it plays an important role in survival and successful reproduction. Literature in numerical cognition points to two systems of number representation; an approximate system that supports intuitive reasoning about numerical magnitudes, often been referred to as *number sense* (Dehaene, 1997; Lipton & Spelke, 2003), as well as the system in charge of rapid, confident and accurate discrimination of small numerosities known as *subitizing* (Kaufman et al., 1949). Thus, natural intelligence is endowed with neural mechanisms that support emergence of number competence as a by-product of exposure to natural visual stimuli, without any explicit training for numerosity estimation (Nieder, 2020).

In contrast, the majority of existing work in artificial intelligence has focused on more advanced aspects of numerical competence, such as counting (Zhang et al., 2018; Mandler & Shebo, 1982; Trott et al., 2018), arithmetic or quantitative reasoning (Geva et al., 2020; Drori et al., 2022; Lewkowycz et al., 2022). While these are important components of analytical skill, they are either gradually acquired during development or via formal education, and are thus associated with various higher-level cognitive functions. Fewer studies have investigated how more fundamental notions of numerical competence, such as those that have been documented in naïve animals or prelinguistic infants, develop in artificial systems (Creatore et al., 2021; Testolin et al., 2020). In this work, we ask whether contemporary deep neural networks trained on large amounts of static image-text data learn representations that are functionally comparable to those underlying number processing in biological cognition. We study whether distinct number representations can be observed for small versus large numerosities and how this depends on the specific modality. Our findings suggest that number

Submitted to 4th Workshop on Shared Visual Representations in Human and Machine Visual Intelligence (SVRHM) at NeurIPS 2022. Do not distribute.

representations in vision and language models are structured in a way that is consistent with the two-system theory in numerical cognition. When comparing modalities, we found that language contributes more than vision towards accurate number representation. We argue that biases in the data used to train these models might explain some of our findings.

2 Related Work

Numeracy-related research in artificial intelligence and machine learning spans a spectrum of motivations. On the one hand, there is a strong application-driven incentive to improve performance and the quality of representations on tasks requiring numerical skills (*e.g.,* arithmetic, magnitude comparison or numerical common sense, among others) which has been lagging behind other NLP benchmarks (Thawani et al., 2021; Wallace et al., 2019; Parcalabescu et al., 2021; Lin et al., 2020). Helpful approaches improving performance on such tasks include using curated synthetic data (Geva et al., 2020; Zhang et al., 2015), or heuristics such as counting-specific model components (Zhang et al., 2018; Trott et al., 2018).

On the other hand, understanding how the more abstract concept of a number is related to visually perceived numerosity is associated with the grounding problem studied in artificial intelligence and cognitive science (Harnad, 2003). Understanding the representation of numbers, and factors that affect it, offers the potential of enriching such representations in artificial systems by informed inductive biases. Creatore et al. (2021) show that basic neural networks can develop internal representations that support qualitatively different numerosity perceptions systems, akin to number sense and subitizing, though with some differences compared to human processes. Wallace et al. (2019) find that token embeddings learned from text can accurately encode magnitudes for numbers within the training range, while failing to extrapolate to numbers outside the range. While affirming that linguistic data contains a significant amount of information about numbers, it is unclear how these representations relate to biological ones, where small numerosities are perceived differently from larger ones.

As well, the existing work in ML related to numeracy has predominantly been focused on specific benchmark performance (Parcalabescu et al., 2021; Zhang et al., 2015; Wallace et al., 2019), and less so on fine-grained analysis of learned number representation and its relation to biological representations. In this work, we scrutinize these representations in the context of what is known about number representation in biological systems.

3 Methods: Models and Datasets

Transformer-based neural network models (Vaswani et al., 2017) have become a de facto standard modelling choice in NLP since their inception, and the extension to support the visual modality has opened doors to study a new space of problems and multimodal interactions (Du et al., 2022; Hendricks et al., 2021; Bugliarello et al., 2021). Here, to characterize representations of numbers, we use the VILBERT (Lu et al., 2019) architecture, a vision and language model that extends the BERT (Devlin et al., 2018) architecture to support processing of visual and text inputs. It processes inputs via two separate, parallel streams, which are subsequently combined via co-attentional transformer layers. Text is presented as a sequence of tokens, while image is serialized as a sequence of region-of-interest features extracted from a separate convolutional neural network (Ren et al., 2015).

Most vision and language models are pre-trained on pretext tasks with multimodal data, analogous to the training process used with text-based transformer models. Then, models are further fine-tuned on a specific transfer task such as visual question answering (VQA), visual commonsense reasoning, referring expressions, or caption-based image retrieval by learning a task-specific decoding head. The lingustic stream of the VILBERT model is initialized with BERT weights, and the model is pretrained on 3.3M image-text pairs from Conceptual Captions (CC; Sharma et al., 2018).

Here, we also study the model that has been fine-tuned on VQA datasets which include explicit number-related questions. In VQA, the task is to answer a question about a given image. We investigate representations in models that have been fine-tuned on two widely used VQA datasets: VQAv2 (Goyal et al., 2017), and Visual Genome (VG; Krishna et al., 2017). While the original VILBERT implementation uses a classifier as a decoding head, we use auto-regeressive token decoder, which is more flexible as it does not require a priori specification of the number of output classes. In VQAv2 there are 10 human responses for each image-question pair, while there is just one response in VG. In our analyses we focus on number-related questions (*i.e.,* those that start with *How many...* or *What number...*), and use the existing data splits. In total, such questions represent about 11.5% and 8% of all questions in VQAv2 and VG, respectively.

4 Results

4.1 Analysis: Number Representation Similarity

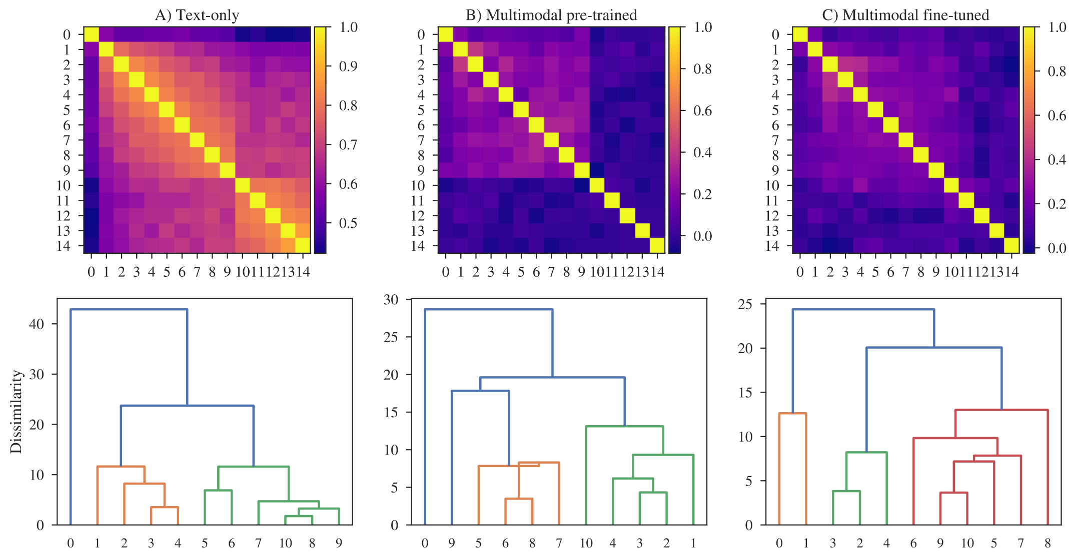

Figure 1: Visualizing similarities between number embeddings. Top: Pairwise cosine similarities between token embeddings for number tokens 0 to 14. Bottom: Dendrograms showing the hierarchy of number clusters based on similarities between embeddings.

First, we look at differences in number representations between text-only models and multimodal models. Rather than looking at aggregate benchmark performance on number-related tasks (Parcalabescu et al., 2021; Wallace et al., 2019), we examine the relationship between learned number representations. Specifically, we are interested in whether learned representations capture any structure reflecting the order of numbers at the qualitative level, and the change in structure during the pre-train/fine-tune process.

We extract learned token embeddings for numbers 0 to 14 from three models: BERT (text-only), VILBERT pre-trained on CC (multimodal), and pre-trained VILBERT fine-tuned on VG (multimodal). For two-digit numbers, we only consider a single token (*i.e.,* "14" instead of "1" and "4"). Figure 1 (top) shows pairwise cosine similarities between extracted token embeddings. Text-only embeddings, shown in Figure 1 (A), display a pattern of 3 visually distinct clusters: one for the token 0, one for numbers 1-9, and one for numbers 10 and larger. Multimodal pre-training appears to distort that pattern, especially as numbers from 10 to 14 become less similar to any other numbers. This occurs because captions in CC contain only numbers from 0 to 9. As well, CC has a peculiar distribution of numbers, with 3 and 4 being the most frequent, possibly due to common occurrences of "3D" and "4K" tokens in the dataset.

To highlight the similarity between individual representations, we plot results of hierarchical clustering of representations of numbers from 0 to 10 in the bottom of Figure 1.[1] Clustering was performed on 2D PCA-projected representations of token embeddings, using the centroid method and the Euclidean metric for calculating the distance between clusters. The algorithm starts by treating each token as an individual cluster, and proceeds to iteratively merge least dissimilar clusters. In all cases, we observe that some of the first clusters formed from singletons are those for subsequent numbers (*e.g.,* some of them are (3, 4) and (5, 6) for text-only representations; (2, 3) for multimodal pre-trained

[1]This is the range present in all conditions we examined.

representations; and (0, 1) and (2, 3) for fine-tuned multimodal representations). We observe further interpretable groupings in the case of BERT, as 2 is merged with (3, 4), 1 with (2, (3, 4)), 7 with (9, (10, 8))) etc. For fine-tuned representation a cluster is formed for (2, 3) and 4, and another one for all larger numbers. In general, we also observe anti-patterns (*i.e.,* 10 merging with 8 for BERT; or 10 with small numbers for multimodal pre-trained).

Restricting the analysis to the following three clusters: 0, small numbers within subitizing range (*i.e.,* 1–4), and larger numbers outside the subitizing range (*i.e.,* 5–10), which we take to be the gold standard reflective of number representation structure in humans and animals, allows us to evaluate cluster assignments observed in hierarchical clustering. Specifically, we consider cluster assignments at points where the distance cut-off value defines three clusters for each dendrogram in Figure 1. To compute the F_1 score, we follow the approach outlined in Schütze et al. (2008, Section 16.3). The highest F_1 score is observed for text-only representations ($F_1 = 1.00$), followed by multimodal fine-tuned on VG ($F_1 = 0.90$), multimodal-pretrained ($F_1 = 0.78$), and multimodal fine-tuned on VQAv2 ($F_1 = 0.53$). This leads us to conclude that text-only number representations are structured in a way that is most similar to the structure of number representations in humans and animals.

4.2 Influence of Modality and Numerosity in Number Representation

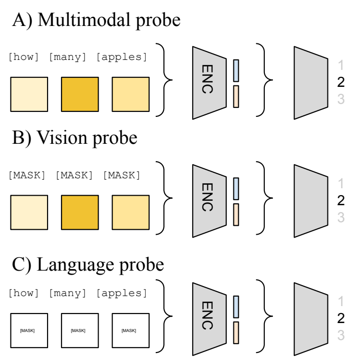

Figure 2: Probes for multimodal, visual and language number representations.

Dataset	Probe	(1, 4)	(5, 10)	(1, 10)
VQAv2	Vis	44.22	23.71	37.36
VQAv2	Lang	46.59	31.26	38.75
VQAv2	[Lang, Vis]	55.09	29.04	46.39
VG	Vis	56.72	31.81	52.66
VG	Lang	58.57	37.35	54.46
VG	[Lang, Vis]	69.10	31.61	64.17
Random	n/a	25.27	16.12	10.46

Table 1: Classification accuracy on test sets for number labels from pooled feature representations.

In biological systems, the processing of items within the subitizing range is attributed to the visual system, while larger numbers are assumed to be processed in a different way (Kaufman et al., 1949). In the previous analysis we found text-based number representations to be the most interpretable. However, it is unclear which modality contributes the most to the representation of a number in multimodal models, and whether this depends on the number range. In this section, we design a probe to answer that question, inspired by similar work in the domain (Lin et al., 2020; Wallace et al., 2019; Parcalabescu et al., 2021).

We train probes to predict numerals based on features extracted from different modalities: multimodal (concatenated visual and linguistic features), visual, and linguistic. Features are extracted from a fine-tuned model as pooled representations of an input question (text) or image (vision) from the 'CLS' (for text) or 'IMG' (for images) tokens at the encoder output during the forward pass on a dataset. In other words, for each (question, image) tuple from a VQA dataset we get two feature vectors. For the vision probe, used to examine the contribution of visual modality in representing numbers, we entirely mask the input question; for the language probe, we entirely mask the visual input to the model. Figure 2 illustrates the probing process. The probe is trained to minimize the cross-entropy loss when predicting the corresponding numeric label—the answer associated with the (question, image) tuple. By ablating one modality in this way, we can study the contribution of the other modality in predicting numbers. Each probe is trained on features from a train split of a corresponding dataset, and tested on the val split of the same dataset. As well, separate probes are trained for different number ranges. Further details on training and evaluation are provided in Appendix A.1.

The probing results are shown in Table 1. In most cases, multimodal features are best at encoding numbers compared to features from a single modality. They are better at encoding smaller numbers (*i.e.,* 1–4) than the larger numbers (*i.e.,* 5–10). In fact, training on the full range of 10 numbers reduces the accuracy. As for individual modalities, linguistic features appear not only to be better at encoding numbers than the visual features, but also better than multimodal features for larger numbers. The fact that language is more informative of small numbers than vision is a remarkable difference between number representation in humans and deep neural networks, as animals and pre-linguistic infants are able to subitize without having developed or acquired language. The reason why it might be somewhat easier to predict the number given a masked image, than to predict the number given masked text is that the question type *How many...* or *What number...* is more informative of the space of possible answers than the image itself. Given an image, without any text, the space of possible answers is more diverse (*i.e.,* yes/no answers, number, color, nouns, verbs etc).

5 General Discussion

The ability to represent and reason about numerical quantities has been extensively studied in human and animal cognition. Human brains, as well as brains of other animals, are equipped with a form of rudimentary number sense essential for survival and reproduction (Dehaene, 1997; Nieder, 2020). In this work, we investigated whether contemporary neural networks processing visual and linguistic inputs develop a notion of a number that is comparable to that observed in biological systems. Namely, we investigated how are numbers represented, and whether small numerosities in the subitizing range (*i.e.,* 1–4 items) are processed differently from large numerosities.

First, we found interpretable structure among number representations—in some instances, representations between subsequent numbers were more similar compared to representations between non-subsequent numbers. When we coarsely clustered number representations into groups based on how numbers are represented and processed in biological systems (small numbers in the subitizing range vs. numbers outside that range), we surprisingly observed a perfect score for number representations coming from a model that has only been trained on text (BERT). We speculate that number ordering, as observed during cluster merge process, as well as grouping of small vs larger numbers could be due to the statistical distribution of numbers in the training text. That is, pairs of numbers such as (1, 2) or (3, 4) are more likely to occur than (1, 5) or (2, 4). In addition, the Newcomb-Benford law (Newcomb, 1881; Benford, 1938), stating that leading digits are likely to be small, might imply better representation of numbers within the subitizing range in real-world data, which could explain some of the patterns we observe. It is worth noting that the distribution of digits in multi-modal data did not adhere to that law.

Second, we examined to what extent individual modalities in vision and language models contribute to the representation of a number. We designed a probe that ablated one modality and learned to predict numbers based on the inputs to the other, non-ablated modality. While multimodal features were best at predicting the number overall, and especially small numbers in the subitizing range, linguistic features were better at this task than visual features. We found this result surprising in light of the fact that humans and non-human animals develop numerical competence through exposure to natural visual stimuli. The higher accuracy on smaller number ranges for all modalities is also likely to be explained by better representation of small numbers in training data, which is the case for both datasets. Numbers 1–4 account for 83.6% (VQAv2) and 93.1% (VG) of probe training data (with remaining numbers being 5–10). As a reference, the Newcom-Benford law predicts that numbers 1–4 would account for approximately 70% of data. We consider it an open question as to why linguistic features appear better than multimodal features for the representation of larger numbers (*i.e.,* 5–10).

In future work, we would like to better examine the role of pre-training, robustness and generality of learned number representations. Since vision and language models are known to latch onto surface-level correlations in the data in VQA (Goyal et al., 2017; Agrawal et al., 2016), it is unclear how transferable learned number representations are. As well, due to distributional statistics of numerical data in these datasets, it is difficult to discern whether subitizing-like patterns we observe are simply due to better representation of small numbers, or are indicative of an emergent phenomenon with distinctive cognitive and behavioural characteristics.[2] We postulate that systematic assessments, similar to those used in cognitive science and psychology, might help to accurately characterize the role of different factors contributing to number sense in artificial systems.

[2]However, data biases can play a fundamental role in emergence of higher-level skills: For example, Chan et al. (2022) find that few-shot learning emerges only with certain distributional statistics and only for some architectures.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. Analyzing the behavior of visual question answering models. *CoRR*, abs/1606.07356, 2016. URL `http://arxiv.org/abs/1606.07356`.

Frank Benford. The law of anomalous numbers. *Proceedings of the American philosophical society*, pp. 551–572, 1938.

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

Stephanie CY Chan, Adam Santoro, Andrew K Lampinen, Jane X Wang, Aaditya Singh, Pierre H Richemond, Jay McClelland, and Felix Hill. Data distributional properties drive emergent few-shot learning in transformers. *arXiv preprint arXiv:2205.05055*, 2022.

Celestino Creatore, Silvester Sabathiel, and Trygve Solstad. Learning exact enumeration and approximate estimation in deep neural network models. *Cognition*, 215:104815, 2021. ISSN 0010-0277. doi: https://doi.org/10.1016/j.cognition.2021.104815. URL `https://www.sciencedirect.com/science/article/pii/S0010027721002341`.

Stanislas Dehaene. *The number sense: How the mind creates mathematics*. Oxford University Press New York, 1997. ISBN 0195110048.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL `http://arxiv.org/abs/1810.04805`.

Iddo Drori, Sarah Zhang, Reece Shuttleworth, Leonard Tang, Albert Lu, Elizabeth Ke, Kevin Liu, Linda Chen, Sunny Tran, Newman Cheng, et al. A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level. *Proceedings of the National Academy of Sciences*, 119(32):e2123433119, 2022.

Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. In *A Survey of Vision-Language Pre-Trained Models*. IJCAI-2022, 2022.

Mor Geva, Ankit Gupta, and Jonathan Berant. Injecting numerical reasoning skills into language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 946–958, 2020.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Stevan Harnad. Symbol-grounding problem. In *Encyclopedia of Cognitive Science*, volume 67. MacMillan: Nature Publishing Group, 2003.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021.

Edna L Kaufman, Miles W Lord, Thomas Whelan Reese, and John Volkmann. The discrimination of visual number. *The American journal of psychology*, 62(4):498–525, 1949.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*, 2022.

Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. Birds have four legs?! numersense: Probing numerical commonsense knowledge of pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6862–6868, 2020.

Jennifer S Lipton and Elizabeth S Spelke. Origins of number sense: Large-number discrimination in human infants. *Psychological science*, 14(5):396–401, 2003.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pp. 13–23, 2019.

George Mandler and Billie J Shebo. Subitizing: an analysis of its component processes. *Journal of experimental psychology: general*, 111(1):1, 1982.

Simon Newcomb. Note on the frequency of use of the different digits in natural numbers. *American Journal of mathematics*, 4(1):39–40, 1881.

Andreas Nieder. The adaptive value of numerical competence. *Trends in Ecology & Evolution*, 35(7): 605–617, 2020.

Letitia Parcalabescu, Albert Gatt, Anette Frank, and Iacer Calixto. Seeing past words: Testing the cross-modal capabilities of pretrained v&l models on counting tasks. In *Proceedings of the 1st Workshop on Multimodal Semantic Representations (MMSR)*, pp. 32–44, 2021.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL `https://aclanthology.org/P18-1238`.

Alberto Testolin, Serena Dolfi, Mathijs Rochus, and Marco Zorzi. Visual sense of number vs. sense of magnitude in humans and machines. *Scientific reports*, 10(1):1–13, 2020.

Avijit Thawani, Jay Pujara, Filip Ilievski, and Pedro Szekely. Representing numbers in nlp: a survey and a vision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–656, 2021.

Alexander Trott, Caiming Xiong, and Richard Socher. Interpretable counting for visual question answering. In *International Conference on Learning Representations*, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. Do nlp models know numbers? probing numeracy in embeddings. 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)*, pp. 5307–5315, 2019.

Jianming Zhang, Shugao Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Salient object subitizing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4045–4054, 2015.

Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. In *International Conference on Learning Representations*, 2018.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] We perform an analysis of existing models to interpret and characterize their behaviours.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The code is proprietary, but we included details needed to reproduce experiments.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Appendix

A.1 Technical Details for Probing Experiments

Our probe is an MLP with 2 hidden layers, with 100 units in each, and a linear layer at the output (10 units). Output of each unit in each layer is passed through a ReLU non-linearity. Training labels are numbers from 1–10, encoded as one-hot vectors. We use cross-entropy loss at the output, which is minimized using Adam optimizer (Kingma & Ba, 2014) with a learning rate of 0.001. We train the probe for 50K steps using feature vectors (*i.e.,* ‘CLS‘ or ‘IMG‘ tokens) extracted from the forward pass of a training split of VQA dataset through the VILBERT model, and evaluate on the val split of the same dataset. We only extract features from (image, question) pairs where the question starts with “How many“ or “What number“. We normalize answers so that “1.“ and “1” is the same answer.

For VILBERT pre-training and fine-tuning we use 16 TPUv3s, while for evaluation (collecting pooled features) we use 1 GPU. GPUs are either Tesla V100 or P100.