On Agreements in Visual Understanding

Yassine Mrabet **Dina Demner-Fushman**

Lister Hill National Center for Biomedical Communications
U.S. National Library of Medicine
National Institutes of Health
8600 Rockville Pike, Bethesda, MD, 20814, USA
mrabety@mail.nih.gov,ddemner@nih.gov

Abstract

Grounding linguistic symbols with digitized images requires a reliable representation of visual concepts. Whether from a cognitive perspective, or a computational perspective, those *iconic representations* are meant to be reused in different linguistic contexts. For visual grounding applications, this poses an important premise question: what would be the levels of agreement on purely visual content when no linguistic descriptions are involved? In particular, (i) the agreement between human assessors on purely visual content can be a relevant indicator of the scalability of visual grounding as a computational approach to natural language understanding, and (ii) the agreement between computer models and human assessors can give us some insights on the difficulty of bringing added value from grounding. In this paper, we study these agreements through the design of a new image similarity collection. In particular, we study inter-human, inter-model, and human-model agreements both on open-domain images and on medical images involving a more challenging context. Our experiments show that coarse-grained agreement between human assessors can reach 90.1%, at different expertise levels, even when no linguistic descriptions are associated with the images. However, a detailed analysis of deep learning search results on our collection showed that different interpretations of the same neural layer have highly different perspectives on visual content, with average correlation ratios ranging between 0.1 and 0.4 for the top 50 results. Although these findings confirm that there is a sufficiently common ground in cognitive iconic representations to build relevant references for visually-grounded language models, they also show that relying on one single model (or layer) for image representation is not suited for grounding applications, and that ensemble representations might be a more viable option.

1 Background

Today's state-of-the-art natural language understanding models rely on symbolic contexts such as sequences of words in a sentence, paragraph, or document, to build relevant neural encoders from large textual corpora [6, 16, 11]. The benefits of this approach have been demonstrated on several benchmarks related to text-classification, named entity recognition, and answer extraction [23]. However, the reuse of these models for different domains, or with texts having substantial structural and syntactic differences remain challenging, and often requires transfer learning networks with dedicated datasets.

Visually-Grounded Language Models (VGLM) follow a different approach by trying to enrich text-based representations with information from associated images [22, 24, 18]. These models

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

are particularly useful for emerging applications such as Visual Question Answering [1, 14, 2, 13], Image-based dialogue systems [4], or image captioning[17, 10].

Reliable representations of visual concepts are a crucial prerequisite of VGLM. From the perspective of the cognitive theory of connectionism [7], this purely visual component is referred to as an *iconic representation* that is separate from the symbolic component, even if they have only one shared learning process. This theoretical separation raises two fundamental questions:

- To what extent would we agree on visual contents when no linguistic descriptions are involved?
- How difficult would it be to build relevant image representations that can be used as reference for visually-grounded language models?

In this paper, we investigate both aspects through the design of a new image similarity collection. To answer the first question, we designed dedicated manual annotation guidelines to study human agreement on purely visual content. To answer the second question, we studied the correlations in search results when different computer-models are used, and the accuracy of these computer models when averaging the raters point of view as a gold standard.

Our results show that inter-rater agreement can reach 90.1% overall, with slight variations across different levels of expertise. On another hand, different readings of the same neural layers for image representation showed low levels of correlation on the top 50 results, with the best model achieving a precision of 75% on the top 5 results. These two factors combined highlight a potentially strong bottleneck for future endeavors in visual grounding.

We present and discuss these results in more detail in section 3. We present our image annotation approach, source data, and experimental setting in the next section.

2 Approach

Image collection. We collected all accessible images from the PubMed Central Open Access subset of biomedical articles[1], the chest X-ray collection from the Indiana University, images from the History of Medicine section of the National Library of Medicine, MedPix radiology images[2], and the USC Othopedic Surgical Anatomy association, leading to a total of 5,362,166 images. Both PubMed Central and History of Medicine contain large subsets of open-domain images such as portraits, maps, and animal species. We filtered out abstract illustrations (e.g., charts) using a supervised SVM classifier similar to Simpson et al.'s approach [21]. Many images have also multiple sub-figures which can bring a substantial ambiguity to the manual assessment task. We applied a panel segmentation classifier adapted from [3] and filtered out the images having two or more sub-figures. This process led to the final collection size of 678,347.

Queries. We defined four query categories to build assessor groups with similar expertise: i.e., Medical (Expert), Medical (Intermediate), Medical (Open), and Open Domain. The test images (queries) were selected by a medical doctor with research experience in information retrieval. Two hundred total queries were selected manually using an online search. The query categories were assigned to each test image by one medical expert and a computer scientist with no medical background, then manually reconciled. We selected 98 open-domain images (49%), and assigned the relevant subsets to different assessor groups (cf. 1). The specific numbers of assessors per category were computed automatically to ensure a balanced workload.

Deep Learning Features. We considered the convolutional model from the Visual Geometry Group (VGG16) [20] and the residual (RESNET) models that obtained the state-of-the-art performance in the ILSVRC 2015 challenge [8]. Both models are pre-trained on ImageNet which was shown to be relevant for several medical tasks [15]. In order to pick the best layers for the collection we conducted preliminary experiments on the ImageCLEF 2011 dataset for medical image retrieval [9]. The best performing feature space was found to be the sum and max aggregations of the convolutional layer 5A of RESNET50, with a top ten precision of 16% and a Mean Average Precision of 3.52% on the first thousand results. In order to take into account these observations and the unaltered embeddings vectors, we derived three feature spaces from the RESNET embeddings:

[1] `https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/`

[2] `https://medpix.nlm.nih.gov`

	Queries		Assessors			
Category	Number	Ratio	MD	ML	CS	Total
Medical (Expert)	24	12%	2	0	0	(2)
Medical (Intermediate)	32	16%	2	2	0	(4)
Medical (Open)	46	23%	0	2	0	(2)
Open Domain	98	49%	0	0	7	(7)
Total	200	100%	2	2	7	11 (15*)

Table 1: Query statistics and assignments per category. MD (Medical Doctor); ML (Medical Librarian); CS (Computer Scientist). *: 4 assessors contributed to two different categories.

- The full feature vector including 25,088 dimensions (R_{full}).
- A sum based aggregation of the convolutional layer (512 dimensions) ($R_{\sum}$).
- A max pooling of the same layer (512 dimensions) (R_{max}).

For all three variants, we use a brute-force search method based on the euclidean distance between the query images and the images in our collection.

Global features. Global features bring a distinct perspective that might enrich the annotation pools with good candidates for both positive and negative matches. For our experiments we reused features from the OpenI image retrieval system, which obtained the best results at the medical ImageCLEF challenge [5]. OpenI uses global features, such as the Fuzzy Color Texture Histogram, the Color Layout Descriptor, the Edge Histogram Descriptor. We created two feature spaces for this global feature set: (1) a clustering-based space, called CW, where each global feature is used to generate cluster words (cluster assignments) for the collection images, which are then used for retrieval with boolean queries, and (2) a concatenation of all global features vectors, called GF, which is then used with the same brute force search method that was applied to the deep learning features.

Table 2 presents a summary of each feature space and the associated search method. We *select the top 50* results from all listed methods to form the final result pools for manual assessment.

Feature	Dimensions	Search method
RESNET Full (R_{full})	25,033	BFS
RESNET SUM ($R_{\sum}$)	512	BFS
RESNET MAX (R_{max})	512	BFS
GF	492	BFS
CW	Word vocabulary	Boolean queries

Table 2: List of methods selected for result pooling. (BFS stands for brute-force search).

It is important to note that this list of features, combinations, and search methods is far from being exhaustive. Restrictions had to be made to obtain a manageable assessment workload (e.g. RESNET vs other architectures) and a good trade-off between the type of a feature space and its computational cost (e.g., GF vs SIFT[12] and ORB[19]). To make these choices we relied either on preliminary experiments or on results reported in related works.

Assessment Guidelines. One of the goals of our study is to observe inter-rater agreements on image similarity when no linguistic bias is introduced (e.g., image captions or labels). Such approach requires a minimal frame to give an orientation to the assessors without interfering with their own reasoning process. We establish such a frame by using the words *content* and *thing*. I.e.: two images are considered to have a similar content if they describe the same thing, or things. We leave it up to the assessors of the different groups to decide what are the important things in an image.

We use three labels to rate a candidate result that is similar to the query image: (5) Very Similar; (4) Similar; (3) Somewhat Similar. We use two labels to assess a candidate result that is not similar to the query image: (2) Graphically Similar (i.e., close in shape and/or color, but not content-wise) and (1) Very Different. At the end of the process, our annotators rated 45,360 results for 200 queries.

Aggregation	Open Domain	Med. (Open)	Med. (Intermediate)	Med. (Expert)	(Micro) Total
(1) (2,3,4,5)	86.10 (0.14)	**91.96** (**0.04**)	79.77 (0.15)	**91.60** (**0.04**)	85.76 (0.10)
(1,2) (3,4,5)	**91.31** (**0.09**)	90.91 (**0.04**)	81.43 (0.15)	81.58 (0.09)	**90.41** (**0.09**)
(1,2,3) (4,5)	77.56 (0.21)	89.43 (0.05)	77.03 (0.18)	62.70 (0.18)	77.65 (0.16)
(1,2,3,4) (5)	82.37 (0.18)	82.88 (0.08)	**92.12** (**0.06**)	83.44 (0.08)	83.18 (0.12)

Table 3: Inter-Rater Agreements. Computed as micro-average accuracy % based on different binary aggregations of the 5 similarity labels. Numbers in () indicate the standard deviation.

3 Results & Discussion

Inter-Rater Agreements. We selected a 10% random subset of queries from each query category to compute inter-rater agreements. All assessors in a given category rated their common subset, leading to either double or multiple ratings according to the number of assessors in the group. We study inter-rater agreements with the Mirco-average accuracy (MAA) according to different binary aggregations of the 5 ratings (cf. table 3). The MAA computation relies on exact counts of agreeing rating pairs, A, and disagreeing rating pairs, D:

$$MAA = \frac{A}{A + D} \tag{1}$$

The binary agreement results are relatively high, with a best accuracy of 90.41% on the second aggregation. This means that, on average, most assessors agreed on what constituted an important content in the query image, and differentiated clearly between the similar-content labels (3, 4, 5) and the different-content labels (1 and 2). We can also notice a weaker agreement on this distinction for the Medical (Expert), and Medical (Intermediate) categories with only 81.5% and 81.4% accuracy.

This observation corresponds to the qualitative feedback reported by our expert assessors. For instance, a radiologist noted that he considered images showing similar conditions in the same organ as similar or very similar, images showing an abnormality in the same organ as somewhat similar, and images showing the same organ in a healthy condition as graphically similar. However, an MD with specialization in immunology relied mostly on the abnormal or healthy distinction in her assessments without using the graphically similar category when the image described the same organs. Similar observations can also be made for the Medical (Intermediate) category which was assessed by two medical librarians.

These relative divergences are a good example of the challenges facing visual grounding approaches with contextual and domain-specific images.

Model agreements. In order to study the agreement between the considered models, we computed the Common Result Ratio at n (CR_n) between several candidate methods. CR_n is defined as:

$$CR_n(m_1, m_2) = \frac{|m_1(n) \cap m_2(n)|}{n} \tag{2}$$

where m_1, and m_2 are two given image search methods and $m_1(n)$, and $m_2(n)$ are their respective result sets at rank n. CR_n is a direct correlation measure that shows exactly how many distinct (or redundant) results are found by a given pair of methods. The values were computed by applying the search methods to the full collection of 678K images using our 200 test queries. The CR_n analysis led to the following observations:

- Despite the fact that the full vector, and the sum and max aggregations are extracted from the same convolutional layer, their disagreement in terms of similarity rankings is substantial. The correlation between the full vector and both aggregations do not exceed 25% in the top 10 result and up to the top 500. The correlations between the sum and max aggregations have only 40% of results in common in the top 10 retrieved images, a value that increases only slightly for the top 500.
- The correlations between the GF variants and the deep learning features are very low (below 0.05), and almost equivalent to the synthetic baseline of two methods agreeing only on the first result. The correlation between GF and CW is also close to the minimal baseline.

Feature Space	Dimensions	P@5	P@10	P@20	P@50
R_{full}	25K	71.0	65.4	60.9	54.8
$R_{\sum}$	512	<u>73.3</u>	<u>67.4</u>	<u>64.21</u>	**58.5**
R_{max}	512	**75.7**	**69.6**	**64.28**	<u>58.0</u>
GF	492	49.5	40.4	35.0	30.5
CW	word index/search	32.7	27.0	23.6	20.2

Table 4: Micro-Average Precision (P@N) of methods

Human-Model Agreements. We study the level of agreement between the manually produced ratings and the automatic search methods through the precision of the top N results when considering the best binary aggregation, i.e., (1,2) (3,4,5), according to the inter-rater agreements (cf. table 3). When multiple annotations exist for an image pair, we computed their final similarity value by rounding the average rating scores. Table 4 presents the Micro-Average Precision values at different ranks, noted (P@N).

The best performing approach was the Max-based aggregation of the convolutional layer from RESNET with a micro average precision of 64.2% for the top 50 results and 75.7% for the top 5 results. The full vector from the convolutional layer had a weaker performance than the Max and Sum aggregations, showing that the additional dimensions brought more noise than relevant information. Global features (GF) had a substantially lower performance than the RESNET approaches, with the cluster words variant causing an additional loss of relevant information.

By studying further the behaviour of each method for specific query categories, we observed that the open domain and medical (open) categories played an important role in decreasing the overall performance. For instance, the best approach reached 89.1% P@5 and 78.2% P@50 for the medical (expert) category, but only 68.3% P@5 and 43.9% P@50 for the open domain category. This open-domain performance can be explained by the very strong distractors that face some queries and the limited number of good matches for a few others. For instance, one of the open-domain queries is a distant picture of the moon which is very similar to many kinds of tissue samples and microbial cultures. Another example is a query image of a postage stamp which had only 5 similar and very similar matches, with the vast majority of results consisting of brochures for different drugs, medical procedures, and health campaigns, which have been rated mostly as graphically similar.

The high levels of agreement obtained on the binary aggregations (up to 90.1%) are very encouraging for future visual grounding efforts, especially considering that our guidelines were designed to have a very low level of control. The baseline RESNET embeddings also achieved up to 75% precision on the top 5 similarity results on average for the 200 test queries. However, the disagreements between close computer models (i.e., different aggregations of the same neural layers) were substantial, with correlation values ranging between 0.1 and 0.4 for the top 50 results. This highlights that relying on the perspective of only one neural layer (or one aggregation) might not be suitable to build visually-grounded language models. Further research efforts are therefore needed to asses whether representations based on an ensemble of models would provide a more stable perspective on visual contents.

4 Conclusion

We studied agreement levels on visual similarity when no linguistic descriptions are involved. We found that a high level of agreement (90.1%) can be reached between human assessors on whether or not two images describe a similar content, with a moderate decrease in agreement on domain-specific images, and that baseline RESNET embeddings trained on ImageNet are able to reach a precision of 75% on the top 5 most similar results. Our experiments also showed that different readings of the same neural layers lead to substantially different similarity results. These findings suggest that convolutional embeddings of images can be a good reference point for visually grounded language models, but that more sophisticated image representation methods are needed to bypass the closed-world assumption that is implicitly made by any individual neural layer and its aggregations. A new image similarity collection was created to support this study and will be shared publicly to promote further research efforts.

Acknowledgments

We would like to thank Soumya Gayen for extracting the deep learning features and conducting the preliminary tests on the ImageCLEF collection, and Dr. Russell Loane for implementing and optimizing the brute-force search. We would also like to express our gratitude to the collection annotators: Dr. Asma Ben Abacha, Prerak Dalal, Soumya Gayen, Dr. Travis Goodwin, Michael Kushnir, Le Lan, Mark Sharp, Sonya E. Shoushan, Dr. James Smirniotopoulos for his valuable expertise and feedback, and Sabir Tehseen. This work was supported by the intramural research program at the U.S. National Library of Medicine, National Institutes of Health.

References

[1] Asma Ben Abacha, Sadid A Hasan, Vivek V Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. Vqa-med: Overview of the medical visual question answering task at imageclef 2019. In *CLEF2019 Working Notes. CEUR Workshop Proceedings*, pages 09–12, 2019.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[3] Emilia Apostolova, Daekeun You, Zhiyun Xue, Sameer Antani, Dina Demner-Fushman, and George R Thoma. Image retrieval from scientific publications: Text and image content processing to separate multipanel figures. *Journal of the American Society for Information Science and Technology*, 64(5):893–908, 2013.

[4] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512, 2017.

[5] Dina Demner-Fushman, Sameer Antani, Matthew Simpson, and George R Thoma. Design and development of a multimodal biomedical information retrieval system. *Journal of Computing Science and Engineering*, 6(2):168–177, 2012.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, 2019.

[7] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335 – 346, 1990.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] Jayashree Kalpathy-Cramer, Henning Müller, Steven Bedrick, Ivan Eggel, Alba García Seco de Herrera, and Theodora Tsikrika. The CLEF 2011 medical image retrieval and classification tasks. In *Working Notes of CLEF 2011 (Cross Language Evaluation Forum)*, September 2011.

[10] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[11] Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*, 2019.

[12] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[13] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690, 2014.

[14] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015.

[15] Obioma Pelka, Felix Nensa, and Christoph M Friedrich. Annotation of enhanced radiographs for medical image retrieval with deep convolutional neural networks. *PloS one*, 13(11):e0206229, 2018.

[16] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

[17] Anna Rohrbach, Marcus Rohrbach, Siyu Tang, Seong Joon Oh, and Bernt Schiele. Generating descriptions with grounded and co-referenced people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4979–4989, 2017.

[18] Deb K Roy. Learning visually grounded words and syntax for a scene description task. *Computer speech & language*, 16(3-4):353–385, 2002.

[19] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE international conference on*, pages 2564–2571. IEEE, 2011.

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[21] Matthew S Simpson, Daekeun You, Md Mahmudur Rahman, Zhiyun Xue, Dina Demner-Fushman, Sameer Antani, and George Thoma. Literature-based biomedical image classification and retrieval. *Computerized Medical Imaging and Graphics*, 39:3–13, 2015.

[22] Akira Utsumi. A distributional semantic model of visually indirect grounding for abstract words. *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2018 Workshop, Montreal, Canada*, 2018.

[23] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

[24] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded language learning.