WSDM Cup 2023 Challenge on Visual Question Answering

Dmitry Ustalov[1,*], Nikita Pavlichenko[1], Daniil Likhobaba[1] and Alisa Smirnova[2]

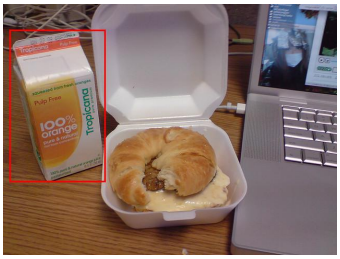
[1]Toloka, Belgrade, 11000, Serbia
[2]Toloka, Lucerne, 6005, Switzerland

Abstract

We challenge you with a visual question answering task!! Given an image and a textual question, draw the bounding box around the object correctly responding to that question. Every image-question pair contains the response, with only one correct response per image. In this paper, we describe the setup, timeline, and results of our competition at WSDM Cup '23, which attracted 48 participants worldwide.

Keywords

visual question answering, visual grounding, crowdsourcing, machine learning competition, WSDM Cup



(a) What do we use to support the immune system and get vitamin C?

(b) What do people use for cutting?

(c) What do you use to hit the ball?

Figure 1: Given an image and a textual question, draw a bounding box containing the correct answer to the question. Above is a sample of three image-question pairs from the training subset of our dataset. Every image contains the response, with only one correct response per image. Bounding boxes are drawn for illustrative purposes only; they are not parts of images in our dataset but are available as ground truth. All images are from the MS COCO dataset [1].

WSDM 2023 Crowd Science Workshop on Collaboration of Humans and Learning Algorithms for Data Labeling, March 3, 2023, Singapore

*Corresponding author.
dustalov@toloka.ai (D. Ustalov); pavlichenko@toloka.ai (N. Pavlichenko); likhobaba-dp@toloka.ai (D. Likhobaba); zero@toloka.ai (A. Smirnova)
https://linkedin.com/in/ustalov/ (D. Ustalov); https://linkedin.com/in/nikita-pavlichenko/ (N. Pavlichenko); https://linkedin.com/in/daniil-likhobaba/ (D. Likhobaba); https://linkedin.com/in/alisa-smirnova-2b862029/ (A. Smirnova)
0000-0002-9979-2188 (D. Ustalov); 0000-0002-7330-393X (N. Pavlichenko); 0000-0002-0322-3774 (D. Likhobaba); 0000-0002-7108-9917 (A. Smirnova)

1. Introduction

Recently featured multi-modal deep learning models like CLIP [2] and DALL-E [3] demonstrate very impressive results in such difficult tasks as text-image similarity measurement and text-to-image generation, respectively. At the same time, modern machine learning methods achieve superhuman results on such challenging multi-task benchmarks as SuperGLUE [4]. We would like to increase the level of difficulty for these methods and set up a new benchmark, *Toloka Visual Question Answering Challenge*.

Our task is formulated as follows. Given an image and a textual question in English, draw a bounding box selecting the object that correctly responds to the question (Figure 1). For example, a bathroom photo might have a question like "Where do I wash my hands?" with the sink selected as the answer. To solve such a task successfully, one has to combine visual, textual, and commonsense information non-trivially.

Most of the previous work is focused on two setups. First, the visual question answering setup, VQA [5], that assumes for each image-question pair the textual response involving commonsense knowledge, e.g., for question "Where is the child sitting?" with an image of a kitchen, the answer could be "fridge", which can be obtained by enumerating the objects detected in the image and utilizing knowledge of how the world works. Second, the TextVQA task [6], e.g., "What is the speed limit of this road?" $\rightarrow$ "20", which can be approached by optical character recognition methods like in the recently introduced VTVQA dataset.[1] We believe that our setup using free-form, open-ended textual questions to the images with bounding boxes as the answers offers a fair challenge for today's multi-modal models.

2. Data Description

Our dataset is comprised of the images associated with textual questions (Figure 1). One entry (instance) in our dataset is a question-image pair labeled with the ground truth coordinates of a bounding box containing the visual answer to the given question. We guarantee that each image contains one and only one correct response to the given question. The images were obtained from a subset of the Microsoft Common Objects in Context, MS COCO,[2] dataset [1] that was licensed under the Creative Commons Attribution (CC BY) license.

To collect the dataset, we ran an annotation campaign on the Toloka crowdsourcing platform.[3] The workers were asked to select the images containing the objects they found subjectively interesting. Then, they had to compose questions about these objects. Finally, for each question-image pair, we asked workers to select the answer on the image using a bounding box, allowing us to exclude unanswerable questions. Although it was possible to facilitate the question composition using such models as DH-GAN [7], we decided to stick to the pure crowdsourcing approach. This also allowed us to avoid synthetic data in our task and stick to the more natural formulations made by real humans.

Our dataset had 45,199 instances split among three subsets: *train* (38,990 instances), *public*

[1] https://github.com/bytedance/VTVQA
[2] https://cocodataset.org/
[3] https://toloka.ai/

Table 1
Overview of our competition dataset

Column	Type	Description
`image`	`string`	URL of an image on a public content delivery network
`question`	`string`	question in English
`width`	`integer`	image width
`height`	`integer`	image height
`left`	`integer`	bounding box coordinate: left
`top`	`integer`	bounding box coordinate: top
`right`	`integer`	bounding box coordinate: right
`bottom`	`integer`	bounding box coordinate: bottom

test (1,705 instances), and *private test* (4,504 instances). The entire *train* dataset was available for everyone since the start of the challenge. The *public test* dataset was available since the evaluation phase of the competition (see Section 5), but without any ground truth labels. The *private test* dataset was not available until the challenge ended. The data were provided as files in the comma-separated values (CSV) format as described in Table 1.

Since we used images from MS COCO, we have explicitly checked the overlap between rectangles in our dataset and the original dataset. About 20% of bounding boxes had non-empty overlap, and we put all such instances into the train dataset. All the annotation, including bounding boxes and questions, was done on Toloka from scratch using only the CC BY-licensed images from MS COCO.

After the competition ended, we released our complete dataset with the ground truth under the same CC BY license as the subset of MS COCO to foster research and development of multi-modal question answering models:

- Zenodo: https://doi.org/10.5281/zenodo.7057740
- Hugging Face Hub: https://huggingface.co/datasets/toloka/WSDMCup2023
- Kaggle: https://www.kaggle.com/datasets/dustalov/toloka-wsdm-cup-2023-vqa
- GitHub: https://github.com/Toloka/WSDMCup2023

3. Metrics and Evaluation Methods

Since the answers in our task are bounding box coordinates and there is only one bounding box per image, we use the *intersection over union* (IoU) aka *Jaccard index* evaluation criterion. For the i-th image, we define it as

$$\mathrm{IoU}_i = \frac{I_i}{U_i},$$

where I_i is the intersection of the ground truth bounding box area and the predicted bounding box area, and U_i is the union of these boxes. Thus, for the entire dataset of N images, the

evaluation criterion is *average intersection over union*, AIoU:

$$\mathrm{AIoU} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{IoU}_i \,.$$

4. Baselines

YOLOR + CLIP Baseline. Shortly before starting the challenge, we released a starter kit that included Python code for a simple zero-shot prediction baseline. First, it used a detection model, YOLOR [8], to generate candidate rectangles. Then, it applied CLIP [2] to measure the similarity between the question and a part of the image bounded by each candidate rectangle. To make a prediction, it used the candidate with the highest similarity. This baseline method achieved $\mathrm{IoU} = 0.21$ on both public and private test subsets that we expected to be surpassed by the participating teams.

Crowdsourcing Baseline. We evaluated how well non-expert human annotators can solve our task by running a dedicated round of crowdsourcing annotations on Toloka. We found them to tackle this task successfully without knowing the ground truth. On all three subsets of our data, the average IoU value is 0.87 ± 0.01, which we consider as a *strong human baseline* for our task; see Section 3 for more information on the evaluation criteria. Krippendorff's α coefficients for the public test is 0.68 and for the private test is 0.66, showing the decent agreement between the responses; we used $1 - \mathrm{IoU}$ as the distance metric when calculating the α coefficient [9].

5. Platform and Timeline

We hosted our competition on the CodaLab platform: https://codalab.lisn.upsaclay.fr/competitions/7434. Before the start of the competition, all the parts of our dataset were frozen and did not change during the competition.

Our competition had three key phases: the practice phase, the evaluation phase, and the reproduction phase (Table 2). In September, we started the *practice* phase to let the contestants get used to the task and training data, including the ground truth data. Then, we started the *evaluation* phase using the public test dataset without ground truth labels. The contestants had to submit their predictions to the competition platform, which resulted in leaderboard updates. Finally, for the sake of reproducibility, soon after the end of the evaluation phase, we started the *reproduction* phase. In this phase, we asked the contestants to provide their solution as a container image. We ran their code to obtain answers for the private test dataset to determine the winners.

6. Results

We had 48 overall participants in our competition, 9 of whom submitted their code during the reproduction stage. Given the width of the gap between the simple zero-shot and human baselines (Section 4), the contestants invented creative ways to address this task. As we used

Table 2
Complete timeline of our competition at WSDM Cup '23

Event	Date
Practice Starts	September 16, 2022
Evaluation Starts	September 30, 2022
Evaluation Ends	December 16, 2022
Reproduction Starts	December 19, 2022
Reproduction Ends	January 16, 2023
Post-Competition Starts	January 16, 2023
WSDM Cup Workshops	March 3, 2023

Table 3
Baselines and final team standings on the *private test* subset, obtained at the reproduction phase of our competition; for visual convenience, we multiplied the IoU values by 100; out of 48 participants, only 9 submitted their code during the reproduction phase

Rank	CodaLab Login(s)	IoU
—	Crowdsourcing Baseline	87.154
1	wztxy89	**76.347**
2	jinx, Zhouyang_Chi	**76.342**
3	komleva.ep	**75.591**
4	xexanoth	74.667
5	Man_of_the_year	72.768
6	Haoyu_Zhang, KhylonWong	71.998
7	nika-li	70.525
8	blinoff	62.037
9	Ndhuynh	61.247
—	YOLOR + CLIP Baseline	21.292

images from the well-known MS COCO dataset, the contestants were welcome to use pre-trained computer vision, language, and multi-modal models trained on this and other datasets.

Table 3 shows the competition results. Even though the participants managed to improve dramatically upon our zero-shot YOLOR + CLIP approach from the starter kit, none of the participating systems outperformed our crowdsourcing baseline.

In the following three paragraphs, we briefly describe the methods reported by the three winning teams according to the reproduction phase on the private test subset of our dataset.

3rd Place. This team fine-tuned the pre-trained multi-modal OFA model [10] on the competition dataset. In order to increase the prediction quality, they additionally used data from the pre-processed GQA dataset [11].

2nd Place. This team devised a three-step pipeline solution. First, at the *coarse tuning* step, they generated textual pseudo answers for the questions and tuned the OFA model to produce textual answers. Then, at the *fine tuning* step, they used prompt engineering of the coarse-tuned

OFA model to draw the bounding boxes. Finally, at the *post-processing* step, they ran an ensemble of these coarse- and fine-tuned models to propose and select the best bounding box candidate.

1st Place. This team created a variant detector using Uni-Perceiver as the multi-modal backbone network [12], with ViT-Adapter for cross-modal localization [13], and DINO as the prediction head [14]. They also included an auxiliary loss [15] and a test-time augmentation module for improved performance, which helped them win the challenge.

7. Conclusion

In our visual question answering task, the inputs were an image and a question, and the output was the bounding box. We had 48 participants in our competition, 9 of whom submitted their code to the final reproduction stage. Even though the participating systems offered near-human-like performance, none of them outperformed non-expert annotators by a significant margin. We believe that it makes our benchmark relevant for the near future until larger multi-modal models are made available. The entire dataset, except images, was created using crowdsourcing on the Toloka platform, making it a valuable tool for creating challenging benchmarks.

Acknowledgments

We are grateful to our colleagues, Natalia Fedorova, Sergey Koshelev, Evgenia Sukhodolskaya, Mikhail Potalitsin, Oleg Pavlov, and Ekaterina Fedorenko, for their contributions to the competition organization and dataset. We would like to thank the CodaLab and the WSDM Cup teams, especially Hady W. Lauw, and the competition participants for making it a big success.

References

[1] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C. L. Zitnick, Microsoft COCO: Common Objects in Context, in: Computer Vision – ECCV 2014, Springer International Publishing, Cham, Switzerland, 2014, pp. 740–755. doi:10.1007/978-3-319-10602-1_48.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning Transferable Visual Models From Natural Language Supervision, in: Proceedings of the 38th International Conference on Machine Learning, volume 139 of *Proceedings of Machine Learning Research*, PMLR, Virtual, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.

[3] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-Shot Text-to-Image Generation, CoRR abs/2102.12092 (2021). URL: https://arxiv.org/abs/2102.12092.

[4] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman, SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems, in: Advances in Neural Information Processing Systems, volume 32, Curran Associates,

Inc., Vancouver, BC, Canada, 2019, pp. 3266–3280. URL: https://proceedings.neurips.cc/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.

[5] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, D. Parikh, VQA: Visual Question Answering, in: 2015 IEEE International Conference on Computer Vision (ICCV), IEEE, Santiago, Chile, 2015, pp. 2425–2433. doi:10.1109/ICCV.2015.279.

[6] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, M. Rohrbach, Towards VQA Models That Can Read, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 2019, pp. 8309–8318. doi:10.1109/CVPR.2019.00851.

[7] S. Kai, L. Wu, S. Tang, Y. Zhuang, Z. He, Z. Ding, Y. Xiao, B. Long, Learning to Generate Visual Questions with Noisy Supervision, in: Advances in Neural Information Processing Systems, volume 34, Curran Associates, Inc., Virtual, 2021, pp. 11604–11617. URL: https://proceedings.neurips.cc/paper/2021/file/60792d855cd8a912a97711f91a1f155c-Paper.pdf.

[8] C. Wang, I. Yeh, H. M. Liao, You Only Learn One Representation: Unified Network for Multiple Tasks, CoRR abs/2105.04206 (2021). URL: https://arxiv.org/abs/2105.04206. arXiv:2105.04206.

[9] K. Krippendorff, Content Analysis: An Introduction to Its Methodology, 4th ed., SAGE Publications, Inc, Thousand Oaks, CA, USA, 2018.

[10] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, H. Yang, OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework, in: Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, Baltimore, MD, USA, 2022, pp. 23318–23340. URL: https://proceedings.mlr.press/v162/wang22al.html.

[11] D. A. Hudson, C. D. Manning, GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 2019, pp. 6693–6702. doi:10.1109/CVPR.2019.00686.

[12] X. Zhu, J. Zhu, H. Li, X. Wu, H. Li, X. Wang, J. Dai, Uni-Perceiver: Pre-training Unified Architecture for Generic Perception for Zero-shot and Few-shot Tasks, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 2022, pp. 16783–16794. doi:10.1109/CVPR52688.2022.01630.

[13] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, Y. Qiao, Vision Transformer Adapter for Dense Predictions, The Eleventh International Conference on Learning Representations (2023). URL: https://openreview.net/forum?id=plKu2GByCNW.

[14] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. Ni, H.-Y. Shum, DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection, The Eleventh International Conference on Learning Representations (2023). URL: https://openreview.net/forum?id=3mRwyG5one.

[15] A. Kirillov, R. Girshick, K. He, P. Dollár, Panoptic Feature Pyramid Networks, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Long Beach, CA, USA, 2019, pp. 6392–6401. doi:10.1109/CVPR.2019.00656.