

Supplementary Materials: Towards Flexible Evaluation for Generative Visual Question Answering

Anonymous Authors

1 TRADITIONAL VQA EVALUATION METRICS

Traditional VQA evaluation metrics contain Exact Match [9] and VQA Score [1]. They apply for different settings in VQA datasets. For datasets where each sample contains only one correct answer, like DAQUAR [9], TDIUC [8], GQA [7], the Exact Match is used. If each sample contains ten candidate answers, like VQA v2 [4], OKVQA [10], VizWiz [6], VQA Score is commonly used.

Exact Match. Exact Match calculates by judging whether the response is identical to the annotated ground-truth answer, and if matches, the score is 1, otherwise 0.

VQA Score. VQA Score evaluates how many times the response appear in the ten candidate answers, and is computed as follows:

$$accuracy = \min(\frac{\# \text{ correct hits}}{3}, 1)$$

As there are ten candidate answers, # correct hits represents numbers of matched answers, which means as long as there are three or more candidates are the same with the predicted answer, the answer will be considered fully correct, and gets a score of 1.

2 PROMPT FOR DECODER MODELS

The following is the similarity calculation prompt provided to the LLMs and ChatGPT in the experiments:

Sentence similarity evaluation here refers to the task of measuring the semantic similarity score between two sentences. For example, "what a good day" and "how nice the weather is" are almost the same, your output shall be {"score":0.91}. Now please evaluate the similarity score between the following two sentences: sentence1: sample["sentence1"]. sentence2: sample["sentence2"]. The score shall range continuously from 0-1. DO NOT output anything else but the .json-style dictionary, like {"score":x}, where x is your predicted score.

The sample["sentence1"] and sample["sentence2"] indicate a pair of texts for similarity calculation. The question and answer are concatenated with the prompt "Question: {question} Answer: {answer}" before similarity calculation, importing contextual information, just the same as other models in experiments.

ChatGPT prompt for converting each question-answer pair into a description:

Concatenate the question with the answer and form assertions. For example, Question:What kind of dog is in the photo? Answer:golden retriever. Assertion: The dog in the photo is a golden retriever. Infer for the following: Question: {question} Answer: {answer}. Please think of three different forms of naturally-sounded assertions for this question-answer pair with small disturbance but do not output them. Choose the two assertions that are closest in meaning to the original question-answer for output. Output shall be in .json style so that I can directly save them in a .txt and open by json. Do not output anything else including explanation, reasoning or instructions.

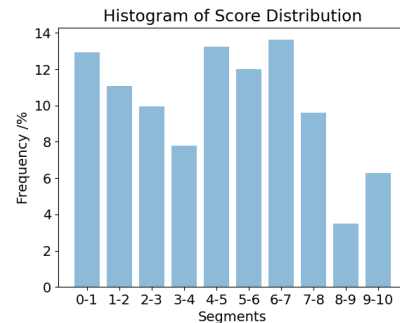


Figure 1: The annotated scores distribution of AVE.

3 ANNOTATION SCORE DISTRIBUTION

The score distribution of the dataset Assessing VQA Evaluators (AVE) is shown in Figure 1. The annotation covers all scores.

4 SELECTION OF ANNOTATING ASPECT

For the evaluation of the quality of responses, we considered adopting multiple aspects for analyzing. The text summary task [2, 3, 5] adopts four aspects, i.e., *relevance*, *consistency*, *fluency* and *coherence* for analyzing the generated summary based on the reference text. As VQA responses are generally short, the *fluency* and *coherence* that measure the fluency of the text is less necessary. The aspect of *consistency* measures whether the generated summary contains hallucination that generates untrue information. In the scene of VQA response evaluation, such measure corresponds to the overlap in semantics from the response towards the ground-truth answer, which is similar to *relevance* that measures how well the generated text captures the key points. Therefore, we decide to use semantic similarity as the only score of annotation.

5 ANNOTATION RULES

5.1 Scores Annotation

All of the following in this section is the annotation rules provided to the three annotators during the annotation of scores.

Scoring Format: Discrete integer scoring from 0 to 10 (e.g., 0, 1, 2, 3).

Scoring Rules:

Note that it is not about judging whether the response to the question is correct, but whether the response and the standard answer are semantically the same under the question (For example, even if the standard answer is obviously unreasonable, as long as the answer and response are semantically similar, a high score shall be given).

Semantically similar (the answer is fairly correct in meaning) but different in specific form (for example, the meaning expressed

is similar but different in word choice, tense, number), score 6-10 based on the degree of semantic similarity. Examples:

(1) Question: What is the last letter on the license plate? Standard Answer: letter j Response: j Scoring (this is a reasonable range, just mark a specific score when actually annotating): 9-10 Reason (no need to mark, this is to help understand the rules): Under this question, the response and the standard answer are semantically the same, only the form is different. Similarly, synonyms should also be scored highly.

(2) Question: The young man above the swimming pool is wearing what? Standard Answer: swimsuit Response: trunks Scoring: 7-8 Reason: The question asks what is being worn, and swimsuit (swimwear) includes trunks (swim shorts), so the answer is quite correct. This kind of inclusive or included relationship should be scored based on the semantic similarity of the two words. In addition, trunks as swim shorts is a less common meaning and requires more attention to the different meanings of words, not entirely based on experience.

(3) Question: How many trees are there? Standard Answer: 3 Response: three Scoring: 10 Reason: The meanings are exactly the same.

Semantically dissimilar, the answer is incorrect, but the answer is a possible answer for that type of question, score 1-5 based on the degree of semantic similarity. Examples:

(1) Question: What color do you think the trousers the boy is wearing have? Standard Answer: white Response: blue Scoring: 2-3 Reason: The question asks about color, and although the answer blue is different from the standard answer white, both are common answers under the category of color questions. Furthermore, if it's white and black, the difference between the two is greater than between white and blue, so the scoring range should be further reduced to 1-2.

Semantically dissimilar, but the answer and the standard answer mean the same under the question, then score 4-7 based on the correctness. Examples:

(1) Question: What is lit? Standard Answer: cake Response: candle Scoring: 5-7 Reason: Although cake and candle are very different semantically, in this question, they actually mean the same thing. Therefore, one should not only look at the answer but also focus on the question.

For numerical type answers, score 1-8 based on how much the number in the standard answer and the number in the response differ. Examples:

(1) Question: How many trees are there? Standard Answer: 4 Response: 5 Scoring: 3-5 Reason: Although 4 and 5 are different, the difference between them is not particularly large. If the standard answer is still 4, but the response becomes 1, then the scoring should be appropriately lowered to 1-3. If the standard answer is 70, and the response is 75, then it can be considered quite correct, scoring 5-7. If the standard answer is 1 and the response is 0, then score 1-2. Judge the score based on whether the numbers are relatively close to each other.

(2) Question: When did this accident happen? Standard Answer: 1945 Response: 1940 Scoring: 5-7 Reason: The two years are quite close. But if the answer becomes the 1940s (i.e., 1941-1949), it includes 1945, and the range is not particularly large, so it is quite correct, scoring 6-8.

For answers with significantly different meanings, score based on semantic similarity without range restrictions. Examples:

(1) Question: What color bathing suit is the woman wearing? Standard Answer: no woman Response: red Scoring: 0-1 Reason: The meanings are very different.

5.2 Manual Filter

The following is the annotation rules provided to the three annotators during the last stage, manual filter, in the construction of our AVE dataset.

Read the question, answer, response and all augmentation results carefully, and decide whether the augmentation has changed the original meaning. Labels shall be in yes, no, unsure.

6 ANSWER TEMPLATES

In the fourth step of constructing the proposed dataset, i.e., description generation, beside collecting ChatGPT-transformed results, we augment each short response with manual written templates: (1) *Answer: {response}.*, (2) *The answer to this question is {response}.*, (3) *As shown in the image and question, the answer is {response}.*, (4) *The answer you are asking for is {response}.*, (5) *As can be deduced from the image, the answer to this question is {response}.*, (6) *The answer to your question appears to be {response}, as shown in the image.*

REFERENCES

- [1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- [2] Manik Bhandari, Pranav Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100* (2020).
- [3] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* 9 (2021), 391–409.
- [4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. 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*. 6904–6913.
- [5] Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283* (2018).
- [6] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3608–3617.
- [7] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6700–6709.
- [8] Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*. 1965–1973.
- [9] Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems* 27 (2014).
- [10] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 3195–3204.