Visual Question Answering Using BLIP

Shreyas Kabra
2021563
shreyas21563@iiitd.ac.in

Ritwik Harit
2021557
ritwik21557@iiitd.ac.in

Vasan Vohra
2021572
vasan21572@iiitd.ac.in

Abstract

Our study evaluates the BLIP model's performance in Visual Question Answering (VQA) across three datasets. Using WUPS Score, BERT Score, VQA Score, etc, we assess its ability to generalise and predict answers accurately. We provide insights into metric selection, rationale, and advantages/disadvantages. This analysis informs on the BLIP model's capabilities in VQA, guiding future research.

1. Introduction

In the realm of Computer Vision and Natural Language Processing, Visual Question Answering (VQA) presents a formidable challenge, demanding a deep comprehension of both linguistic nuances and visual semantics. This task involves an AI system analysing an image alongside a corresponding natural language query and then generating a coherent textual response. Our project endeavours to assess the efficacy of the BLIP model across three distinct datasets (VQA v2.0 dataset (training), VQA v2.0 dataset (validation) and DAQUAR - DAtaset for QUestion Answering on Real-world images), aiming to ascertain its ability to generalise and accurately predict answers. To comprehensively evaluate the model's performance, we adopt a diverse set of evaluation metrics, including WUPS Score, BERT Score, and VQA Score, etc. Each metric offers unique insights into the model's capabilities, and we delve into the rationale behind their selection, along with their respective advantages and disadvantages, to provide a comprehensive evaluation framework.

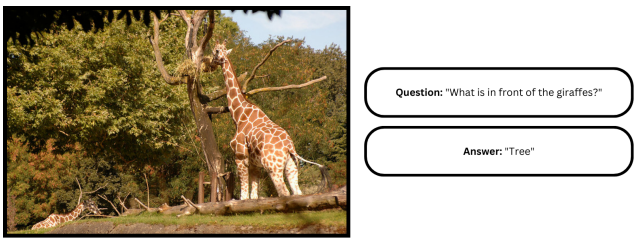


Figure 1. Random examples from the VQA v2 training dataset

2. Literature Review

2.1. Bootstrapping Language-Image Pre-training

The BLIP model architecture (see Fig. 2) proposed by Li et. al [2], is a new Vision-Language Pretraining framework that achieves state-of-the-art performance on various vision-language tasks by addressing the limitations of existing methods. BLIP utilises a new dataset bootstrapping technique called CapFilt, which generates synthetic captions and filters out noisy captions to improve the quality of the dataset. The proposed framework introduces a multimodal mixture of encoder-decoder (MED) model architecture and leverages pre-training objectives such as image-text contrastive learning, image-text matching, and image-conditioned language modelling to achieve flexible transfer learning and effective multi-task pre-training.

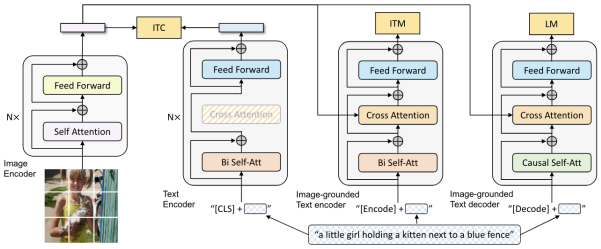


Figure 2. BLIP Architecture

2.2. Wu-Palmer Similarity Score

In paper by Malinowsk et. al [3] introduces a performance measure called the WUPS score for evaluating the quality of system-generated answers. It draws inspiration from Fuzzy Sets theory and utilises the Wu-Palmer Similarity (WUPS) score to account for semantic fuzziness between classes. WUPS score penalises both underestimation and overestimation of answers. The formula considers the intersection of system and ground-truth answers, employing a soft membership measure. Empirical findings suggest a WUP score of approximately 0.9 for precise answers, prompting down-weighting for scores below a threshold. A curve over thresholds illustrates the trade-off between precision and

forgiveness, with WUPS at 0 being the most lenient measure and WUPS at 1.0 equating to standard accuracy. Further details about the evaluation metric will be discussed in the subsequent sections.

3. Dataset Description

We have utilized three distinct datasets to evaluate the performance of the BLIP model: VQA v2.0 Training, VQA v2.0 Validation, and the DAQUAR Dataset. Refer to Table 1 for further details.

Dataset	#Images	#Questions	#Answers
VQA v2.0 Training	82783	443757	4437570
VQA v2.0 Validation	40504	214354	2143540
DAQUAR	1449	5674	5674

Table 1. Comparison of #Image, #Question, and #Answer Across Datasets

For the VQA v2.0 Training and Validation datasets, we were provided with 10 answers for each question. The majority is considered as the final ground truth answer.

4. Exploratory Data Analysis

4.1. Question Length Distribution Across Datasets

The graph in Fig. 3 indicates that the majority of questions in both the training and validation sets of the VQA v2 dataset consist of 4 to 7 words, with the distributions almost overlapping, implying the validation set's representativeness of the training set. Conversely, questions in the DAQUAR dataset tend to be longer, primarily falling within the range of 6 to 11 words.

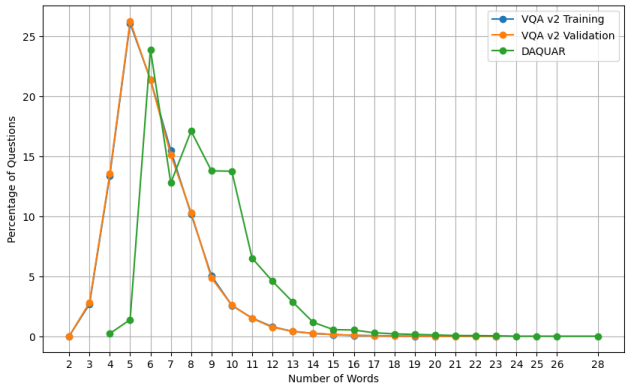


Figure 3

4.2. Answer Length Distribution Across Datasets

The graph in Fig. 4 reveals that the majority of answers in both the VQA v2 datasets (training and validation sets) and the DAQUAR dataset consist of 1 or 2 words. This could suggest that the questions are designed to elicit simple or direct information from the images, rather than requiring complex or elaborate explanations.

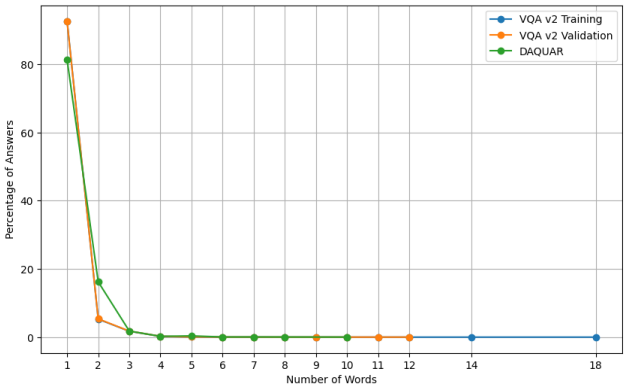


Figure 4

4.3. Question Type Distribution Across Datasets

The graph in Fig. 5 provides insights into the design of questions within the datasets. In the VQA v2 datasets (both training and validation sets), questions starting with "how many," "is the," and "what" are predominant. Conversely, in the DAQUAR dataset, questions frequently begin with "what is on the," "what is the," and "what is." This suggests a focus on different types of queries in each dataset, with VQA v2 emphasising queries about quantity, state, and general attributes, while DAQUAR leans towards inquiries about object identification and description.

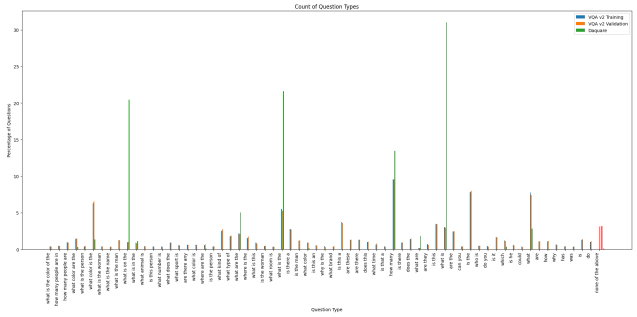


Figure 5

5. Results

5.1. Accuracy

Accuracy is the proportion of correctly predicted instances (True Positive and True Negative) out of all instances. In this an answer is correctly predicted only if it is exactly

same as the ground truth answer i.e. strict matching.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Dataset	Accuracy
VQA v2.0 Training	0.769
VQA v2.0 Validation	0.766
DAQUAR	0.230

Table 2. Accuracy Scores

5.2. BLEU Score

It is the geometric average of the modified n-gram precisions, p_n, using n-grams up to length N and positive weights w_n summing to one. Let c be the length of the predicted sentence and r be the ground truth sentence length. The brevity penalty BP is calculated as

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases}$$

Then the BLEU score is

$$BLEU = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

.

We utilize BLEU-1, BLEU-2, BLEU-3, and BLEU-4 by respectively adjusting N and applying uniform weights w_n = 1/N.

Dataset	BLEU1	BLEU2	BLEU3	BLEU4
VQA v2.0 Training	0.763	0.552	0.438	0.349
VQA v2.0 Validation	0.760	0.551	0.438	0.354
DAQUAR	0.183	0.081	0.037	0.0

Table 3. BLEU Scores

5.3. BERT Scores

BERTScore leverages the pre-trained contextual embeddings from BERT [1] and matches words in candidate and reference sentences by cosine similarity. It has been shown to correlate with human judgment on sentence-level and system-level evaluation. Moreover, BERTScore computes precision, recall, and F1 measure, which can be useful for evaluating different language generation tasks.

Dataset	BERT Pre	BERT Rec	BERT F1
VQA v2.0 Training	0.985	0.986	0.985
VQA v2.0 Validation	0.985	0.985	0.985
DAQUAR	0.945	0.935	0.939

Table 4. BERT Precision, Recall and F1 Scores

5.4. WUPS Score

The WUPS calculates the similarity between two words based on their longest common subsequence in the taxonomy tree. If the similarity between two words is less than a threshold then a score of zero will be given to the candidate answer. We have used the Wordnet [4] database to measure the similarity between the words. The WUPS Score is calculated using the below formula

$$\text{WUPS}(A,T) = \frac{1}{N}\sum_{i=1}^{N} \min\{\prod_{a \in A^i} \max_{t \in T^i} \mu(a,t), \prod_{t \in T^i} \max_{a \in A^i} \mu(a,t)\}$$

Dataset	WUPS 0.0	WUPS 0.9
VQA v2.0 Training	86.573	79.484
VQA v2.0 Validation	86.223	79.203
DAQUAR	58.122	30.680

Table 5. WUPS Score at 0.0 and 0.9 Threshold

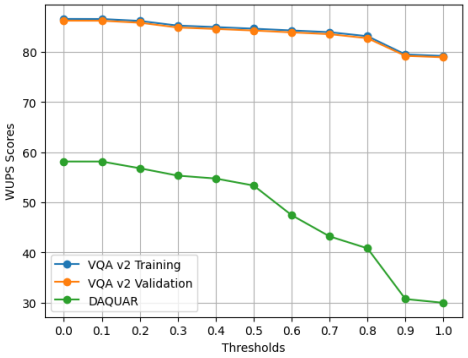


Figure 6. WUPS Score at different threshold

5.5. VQA Score

The VQA accuracy metric assesses the correspondence between the model's response and all potential answers provided for a given question. If the model's response matches

exactly with a minimum of three among the available answers, the VQA accuracy is rated as 1; otherwise, it falls below 1. To evaluate this metric, multiple answers to a question are required. However, since we don't have that in the DAQUAR dataset, we haven't calculated it for this.

$$Acc(ans) = \min\left\{\frac{\text{\#humans that said ans}}{3}, 1\right\}$$

Dataset	VQA Score
VQA v2.0 Training	84.89
VQA v2.0 Validation	84.73
DAQUAR	–

Table 6. VQA Scores

6. Experimental Setup

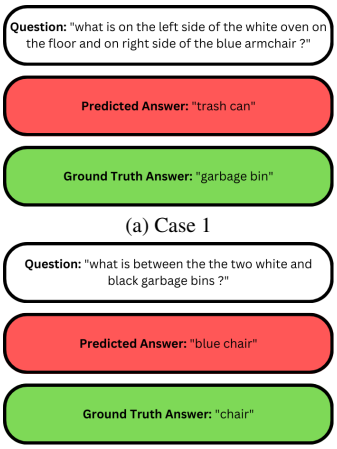
Since the VQA v2.0 dataset was very large (approximately 82,000 in the training dataset), we divided it into 10 parts and split these parts among the 3 members to achieve parallelisation. We used the BLIP model for each part and then combined the outputs for all the parts to evaluate the results. For the validation set of VQA v2 we divided it into 2 parts and combined the outputs for both parts to evaluate the results. DAQUAR dataset was not large (approximately 1500 images), so we did not divide it into parts.

7. Analysis

In this section we will go through the drawbacks and advantages of the evaluation metric with the help of two question-answer pairs projected on the same image.

Figure 7

Fig. 7 shows an image instance from DAQUAR dataset to which the BLIP question-answer model is applied. Fig. 8 presents two cases which includes the question asked, the prediction we get and the ground truth on the image.



(a) Case 1

(b) Case 2

7.1. Accuracy

As evident from the results, the BLIP model exhibits relatively low accuracy. However, relying solely on accuracy metrics may not comprehensively assess the model's performance. As shown in the example above, strict matching between predicted and best ground truth results may overlook semantically similar answers, as explained below :

- **Case 1**: The model predicted "trash can" while the ground truth was "garbage bin", resulting in a 0 accuracy score despite the similarity in meaning.
- **Case 2**: The model predicted "blue chair" while the ground truth was "chair", resulting in a 0 accuracy score despite the more detailed answer.

Hence, accuracy alone may not adequately reflect the model's effectiveness in capturing nuanced semantic relationships.

7.2. BLEU Score

The BLIP model's subpar accuracy underscores the importance of adopting nuanced evaluation metrics such as the BLEU score. Unlike traditional accuracy, BLEU considers n-gram overlaps, providing a more comprehensive understanding of the model's performance across diverse datasets.

- **Case 1**: The model predicted "trash can" while the ground truth was "garbage bin", resulting in a 0 BLUE score despite the similarity in meaning because it checks for the n-gram overlaps.
- **Case 2**: The model predicted "blue chair" while the ground truth was "chair", resulting in a 0.5 BLUE score due to a small degree of overlap in the answer.

Thus, the BLUE score is better than the standard accuracy but is still insufficient because it does not take the semantic similarity into account.

7.3. BERT Score

- **Case 1**: The model achieved a BERT Precision score of 0.90, indicating high semantic similarity between the model's output ("trash can") and the ground truth value ("garbage bin").
- **Case 2**: The model predicted "blue chair" while the ground truth was "chair", resulting in a BERT Precision score of 0.83 due to high similarity.

Thus, as shown by the results, the BERT score is a highly valuable metric for assessing the performance of the model.

7.4. WUPS Score

- **Case 1**: The WUPS score is good in case 1. The score is 0.76 at 0 threshold and 0.0076 at 0.9 threshold.
- **Case 2**: The WUPS does not provide a good score in case 2. The score is 0.11 at 0 threshold and is 0.011 at 0.9 threshold.

WUPS score provides us insight into the semantic similarity between our model's output and the ground truth values in some cases, however, it is unable to capture the semantic similarity of the words in all the cases.

7.5. VQA Accuracy

- **Case 1**: VQA accuracy is 0.67 because the model's output matched 2 answers out of the 10 available answers.
- **Case 2**: VQA accuracy is 0.33 because the model's output matched exactly 1 answer of the 10 available answers.

VQA score provides a more accurate depiction of the model's performance as compared to simple accuracy by comparing the predictions with all the available answers. However, it still does not take the semantic similarity into account, which is not ideal.

8. Commentary

8.1. Ritwik Harit

Through our study, we have used a comprehensive framework to evaluate the performance of the BLIP model across different datasets. We have conducted a detailed exploratory data analysis to showcase the datasets and their characteristics. Moreover, the Results and Analysis section highlights the areas where our model performs well and where it needs improvement. As shown by poor accuracy on the DAQUAR dataset, the BLIP model is unable to predict the exact answers in many cases. However, the high BERT scores show that there is a high degree of semantic similarity between the predictions of the BLIP model and the ground truth answers. Such specialised metrics for evaluation provide us with better insights into the performance of the model, which have not been used in the original paper.

8.2. Shreyas Kabra

Our task was to evaluate the performance of the BLIP model across different datasets including the one it was trained on. As shown by the results, the model performs exceptionally well for the VQA v2 dataset, however, its performance decreases sharply on the DAQUAR dataset. However, the value of the BERT score is still very good on the DAQUAR dataset, implying that the model is able to produce results which are semantically similar to the ground truth. Moreover, good results on the VQA datasets suggest that the model is trained well and gives precise results for these datasets, as shown by high accuracy. However, such precision is lost when the DAQUAR dataset is used, as shown by the low accuracy score.

8.3. Vasan Vohra

In this project, we looked at how to utilize BLIP on different datasets to bring out the best outcomes in the VQA, i.e., Visual Question Answering. The task is to extract an answer from a given image based on the question provided. The first step was an extensive dataset analysis to understand the features. The next step is to understand the different metrics that could be utilized to evaluate the results. We have provided an explanation of why a particular metric is better than another. The low accuracy and BLEU score in DAQUAR make it clear that the model cannot give exact answers. The high BERT score in the same means that the meaning of the outcomes of the BLIP is closer to the result we need to reach. The overall results on the VQA dataset were better than DAQUAR. Using better and more explanatory evaluations in the model outputs makes this project different from the already defined papers.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[3] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27, 2014.

[4] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.