WebQA Data Analysis

Haofei Yu Jiyang Tang Ruiyi Wang Ziang Zhou

1. Statistical Analysis

WEBQA (Chang et al., 2021) is a novel and challenging benchmark for multi-modal, multi-hop, open-domain question answering. Generally speaking, this dataset includes 36,766 training samples, 4,966 validation samples, and 7,540 test samples.Based on different annotation standards, the whole dataset can be classified into two types of data: image-based annotated samples and text-based annotated samples. Table 1 shows that the image-based annotated samples are 51.24% of the total number while the text-based annotated samples are 48.76% of the total number.

The text-based questions are spread out across a large territory and it makes them hard to find a categorization method that would nicely summarize them into discrete buckets. The linguistic complexity of text-based annotated questions is also greater than that of image-based questions. As a result, tet-based annotated samples are all classified as $text$ and has no fine-grained sub-classes. When it comes to the image-based questions, since these question-answer pairs are more straightforward and easy to answer, the dataset are classified into 6 different types: $YesNo$, $choice$, $number$, $color$, $shape$, $Other$.

More information with regard to question-answer pairs are provided in the WEBQA dataset. The first part of the additional information is the topic of the question-answer pairs. The second part of that is the negative text and image pairs. Concerning topic information, there are 433 different topics in the training dataset and 356 topics in the development dataset. Most of the topics, which include 348 topics, are overlapped between training and development dataset. Only 401 samples in the training dataset own unique topics and only 9 samples in the development dataset have unique topics. Therefore, we can conclude that most topics for the data is shared between training and validation. It means that there is no out-of-distribution problem with regard to training and evaluation.

Another part of additional information mentioned in the WEBQA dataset is both positive and negative image-based and text-based facts. The negatives in the dataset are hard in a sense that they have large lexical overlap with the question but are not helpful for answering the questions. The positives are carefully checked during the initial annotation process and are ensured that false positives are not included. It is worth noticing that for image-based question-answering pairs, text positives are not provided while for text-based question-answering pairs, image positives are not provided. As a result, we calculate statistical information about positive samples and negative samples separately on image-based question-answering pairs and text-based quetion-answering pairs. Table 2 shows how many average positives and negatives belong to each sample in the dataset.

WEBQA	Train(%)	Dev(%)	Total(%)
# text	42.68	5.88	48.56
# YesNo	15.56	1.98	17.54
# choice	8.90	1.21	10.11
# number	4.45	0.61	5.06
# color	3.96	0.43	4.39
# shape	1.18	0.17	1.35
# Other	11.37	1.60	12.97
WEBQA	#Train	#Dev	#Test
# sample	36,766	4,966	7,540

Table 1. Statistical information of WEBQA dataset (Total(%) column excludes test samples and only considers train and dev samples since the question categorization is not provided in the test dataset).

Train	#Img(+)	#Img(-)	#Txt(+)	#Txt(-)
Image-based QA	1.44	15.85	0	15.35
Text-based QA	0	11.61	2.03	14.62
Dev	#Img(+)	#Img(-)	#Txt(+)	#Txt(-)
Image-based QA	1.44	15.88	0	15.29
Text-based QA	0	11.78	2.04	14.63

Table 2. Statistical information about both image-based and text-based negatives and positives. Each number in this table represents the average number of text-based and image-based positives and negatives for each sample.

2. Train-Test Overlap Analysis

Lewis et al. (2020) mentions that 60-70% of test-time answers are also present somewhere in the training sets in widely used open-domain question answering datasets. Moreover, they find that 30% of test-set questions have a near-duplicate paraphrase in their corresponding training sets. In this section, we follow this idea and analyze whether there is answer and question overlap between training dataset and test dataset.

Mentioned in Section1, since WEBQA has different types of questions and their evaluation metrics are different based on their types, we separately consider their overlap.

For the first type, including question types of $YesNo$, $number$, $color$, $shape$, WEBQA uses the F1 results calculated from domain keywords and generated tokens as evaluation metrics. For the second type, including question types of $choice$, $Other$, and $text$, WEBQA uses the recall calculated from test keywords and generated tokens as evaluation metrics.

3. Image Fact Aanalysis

As mentioned in (Chang et al., 2021), questions can contain one or two positive image facts and several negative ones. The input of the VLP model is the prediction result of a pretrained faster RCNN on such images. More specifically, the output of the faster RCNN model's last layer, the bounding box coordinates, the object class labels, and the confidence scores are used.

The negative image facts are obtained using hard negative mining. However, (Chang et al., 2021) didn't mention the detailed procedure of this process. We obtained more details from the authors. During the initial annotation process, each annotator was provided with a set of evidences and they selected some of them to create a question that can be answered by these evidences. The unselected evidences naturally became negative facts. In addition, the authors have a retrieval baseline that selects evidences based on their captions' lexical overlap with the question, and the samples that deceived this baseline became additional hard negatives. Finally, the authors manually validated these hard negatives to ensure there is no false positive.

Intuitively, human first read the question and then select which images to use as positive samples based on the question content. That means the hard negatives must be difficult to be distinguished from each other for the VQA model without seeing the question. Therefore, we want to verify this assumption by analyzing the image facts without the questions.

3.1. Image Embedding Visualization

We treat the output of the RCNN model's final layer as the image embeddings, and visualized them with regard to positiveness and question type using PCA dimension reduction, as shown in Figure 1 and Figure 2.

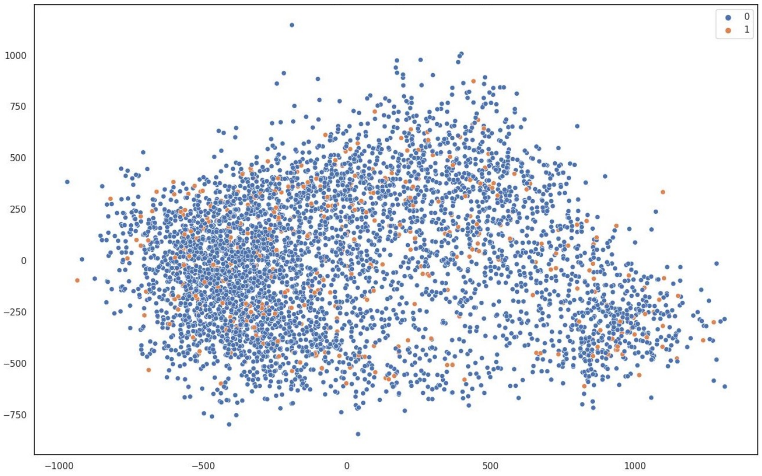

Figure 1. The PCA of RCNN embeddings vs. positiveness

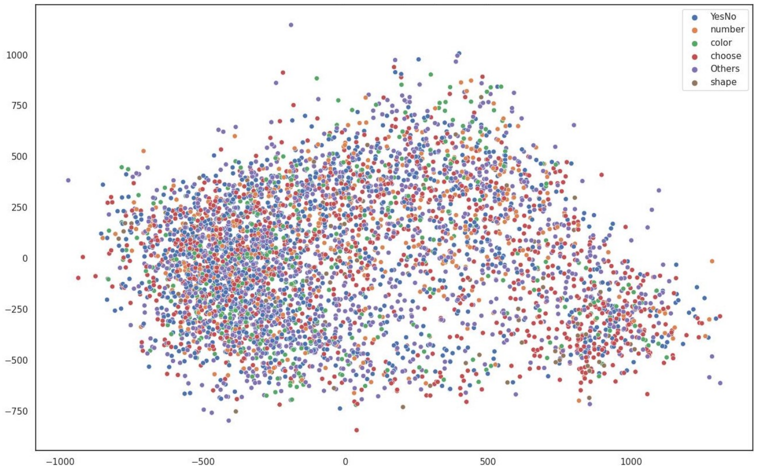

Figure 2. The PCA of RCNN embeddings vs. question types

As shown in both figures, there's no apparent pattern or clusters in regard to question types or fact positiveness. This gives us a rough impression that the images themselves contain insufficient information to infer their question types and fact positiveness. However, this doesn't necessarily verify our assumption, because

- there could be some patterns in data that cannot be seen from PCA visualization,
- and the RCNN embeddings do not comprehensively represent the images.

Therefore, to properly verify our assumption, we need to perform some classification experiments on raw image input.

Positiveness	Positive facts	Negative facts				
Train	2658	3673				
Test	1216	1732				
Question Types	YesNo	Number	Color	Shape	Choose	Other
Train	369	95	75	28	168	265
Test	597	227	160	49	438	529

Table 3. Statistics of the image classification subset in terms of positiveness and question types

3.2. Image Classification Experiments

We extracted a subset from WebQA that has a balanced number of positive and negative facts. See Table 3 for the statistics of the subset. In addition, there are 74 question topics in this subset.

We first need a baseline experiment to prove that our classifier indeed has the capability of distinguishing different types of images. Therefore, a topic classifier is trained. Then, two classifiers are trained, one for predicting question types and the other for predicting fact positiveness.

We expect the topic classifier to show some classification capability and the other two two classifiers to diverge during the training process or to perform badly on the test data.

The results shown in Table 4 indeed meet our expectation. Therefore, we have confirmed the quality of hard negative mining of image facts.

	Accuracy	F1 Score
Topic Classifier		
Question Type Classifier		
Positiveness Classifier		

Table 4. Image classification experiment results

4. Future Work

References

Chang, Y., Narang, M., Suzuki, H., Cao, G., Gao, J., and Bisk, Y. Webqa: Multihop and multimodal QA. *CoRR*, abs/2109.00590, 2021. URL https://arxiv.org/abs/2109.00590.

Lewis, P. S. H., Stenetorp, P., and Riedel, S. Question and answer test-train overlap in open-domain question answering datasets. *CoRR*, abs/2008.02637, 2020. URL https://arxiv.org/abs/2008.02637.