WebQA Dataset 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.

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, text-based annotated samples are all classified as *text* and have 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*, *Others*. They are demonstrated in Figure. 1

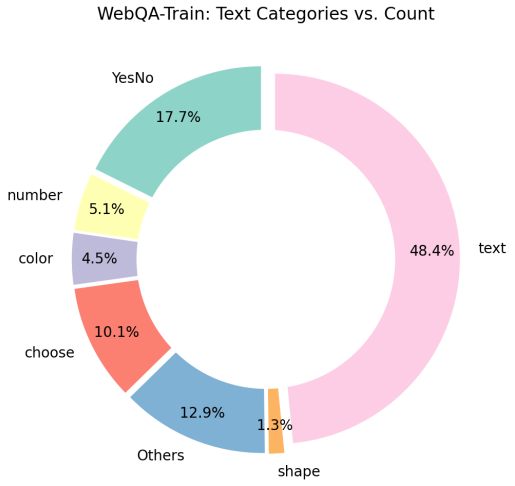

Figure 1. Statistical information of Question Categories in Training Set

More information with regard to question-answer pairs is 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. Example topics are streets, plaza, american museum of natural history, etc. The distribution of topics in the training dataset is plotted in Figure. 2. Most of the topics, which include 348 topics, are overlapped between the 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 are shared between training and validation. It means that there is no out-of-distribution problem with regard to training and evaluation.

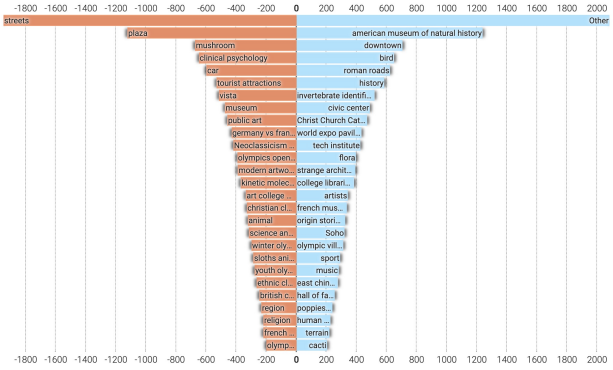

Figure 2. Distribution of topics of the training dataset

Another part of the 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 the sense that they have a 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 question-answering pairs. Table 1 shows how many average positives and negatives belong to each sample in the 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 1. 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 that one sample has.

Split	shape(%)	YesNo(%)	color(%)
Train	92.67	66.54	99.33
Dev	95.95	68.24	97.21

Table 2. Domain set overlap information about *YesNo*, *color*, *shape* types of questions and discuss whether ground-truth answers include corresponding domain set tokens which are used in the WEBQA final testing process.(domain set are specially designed token sets including {Yes, No} for *YesNo* type, {square, circle, ...} for *shape* type, {golden, red, ...} for *color* type)

2. 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 an answer or question overlap between the different splits of the dataset.

As mentioned in Section1, WEBQA has different types of questions and their evaluation metrics are different based on their types. We consider their overlap analysis separately. For the first type, including question types of *YesNo*, *number*, *color*, *shape*, WEBQA uses the F1 results calculated from intersected keywords and generated tokens as evaluation metrics. The intersected keywords are made from domain set keywords and ground-truth test keywords in the test set. Since this type of question has typical answering scopes, we can use the provided domain set and calculate its coverage on the training dataset and development dataset. Since *number* type keywords is not provided, we leave this part of question-answering pairs to be future work. Table 2 shows that there are still relatively some samples that have no words related in the domain sets.

The evaluation script provided by the WebQA dataset calculates the accuracy and F1 score based on the number of intersections between true answers and the domain set and the number of intersections between predicted answers and the domain set. Take an example of the *YesNo* question type, where the domain set is just {Yes, No}. There are more than 30% answers of the training and development dataset that do not contain the keywords "yes" or "no". Given the official evaluation script, all the predicted answers under such cases are marked as wrong. To further investigate the real accuracy of those cases in the *YesNo* question type where the true answers are not related to the domain set, we manually labeled the correctness of the answers produced by the baseline models with respect to the development dataset.

The number of out-of-domain answers in the development is 263, and we randomly sampled 50 pairs from the validation outputs from both the vinvl and x101fpn baseline models that are marked as incorrect. Table 3 shows that the real accuracy in both baseline models is fairly high, indicating that those cases falling out of the domain set are dragging down the actual accuracy of the model. For future work, we need to find a better way to evaluate the accuracy of the out-of-domain cases instead of just finding the intersections of the domain set.

Baseline Model	vinvl	x101fpn
Accuracy	82%	82%
# Samples	50	50

Table 3. Manually calculated accuracy of 50 randomly selected samples of the vinvl and x101fpn baseline models where the real answers have no intersection with the *YesNo* domain set.

For the second type, including question types of *choice*, *Others*, and *text*, WEBQA uses the recall calculated from test keywords and generated tokens as evaluation metrics. Since the test keywords are not publicly available, we utilize *noun_chunks* API in the *spaCy* to analyze the noun overlap between the development dataset and training dataset. Table 4 shows that more than 30% of nouns mentioned in the questions or answers in the dev/test dataset actually have appeared in the training dataset before. As a result, it is one of the potential points for model improvement.

	#train	#dev	#shared	overlap(%)
Questions	53,469	9,793	3,736	38.14
Answers	31,224	5,367	1,895	35.31
	#train	#test	#shared	overlap(%)
Questions	53,469	20,787	6,290	30.26

Table 4. Noun overlap information in answers and questions between training dataset and development dataset. Each number in the table represents the unique noun number in the splits of dataset. Only *choice*, *Others* , and *text* samples are taken into consideration in this table. Overlap ratio is calculated by dividing number of shared nouns with total number of nouns in the dev/test dataset.

Positiveness	#PosFacts	#NegFacts
Train	2,658	3,673
Dev	561	1,343
Test	1,216	1,732

Question Types	#YesNo	#number	#color	#shape	#choose	#Others
Train	369	95	75	28	168	265
Dev	142	57	40	18	111	132
Test	597	227	160	49	438	529

Table 5. Statistics of the image classification subset in terms of positiveness and question types

3. Fact Analysis

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 (Ren et al., 2015) 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.

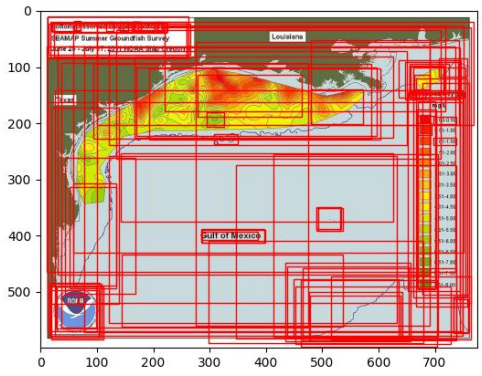

Figure 3. The bounding boxes produced by RCNN

Figure 3 shows an example image and the bounding boxes produced by the RCNN model. Note that there are a large number of bounding boxes, some of them even overlapping, because the RCNN model was trained on visual genome dataset which contains detailed descriptions for objects, their attributes, and relationships (Krishna et al., 2017). This is ideal for WebQA because we want the VLP model to be aware of as many details about the images as possible.

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 the positives for the VLP 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 4 and Figure 5.

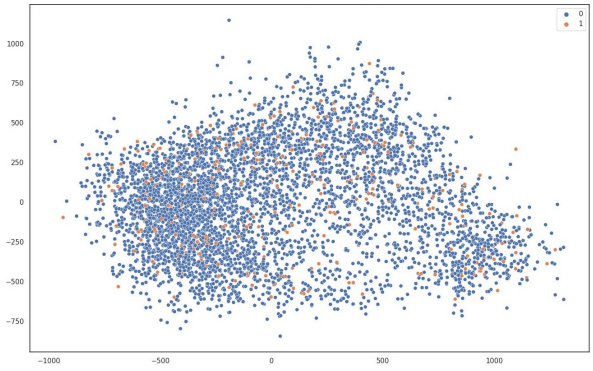

Figure 4. The PCA of RCNN embeddings vs. positiveness

3.2. Text Embedding Visualization

In (Chang et al., 2021), the authors used Bert-base-cased model as tokenizer to obtain the textual embeddings of textual sources and image captions. Rather that having multiple embeddings in each sentence, in this analysis, we used sentence transformers (Reimers & Gurevych, 2019) directly to extract utterance-level embeddings for simplicity and intuitiveness. Each token in a sentence will be represented by a

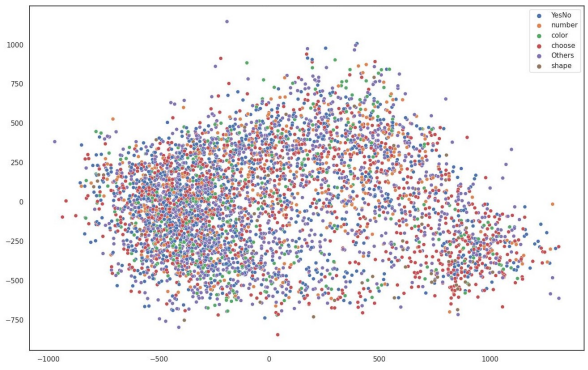

Figure 5. The PCA of RCNN embeddings vs. question types

384-dimension vector, and the embedding of each sentence is by taking average of all token embeddings. The samples choices are aligned with the selected images. Following the same PCA settings, we visualize the reduced textual embeddings in Figure. 6 and Figure. 7, hoping to verify our assumptions.

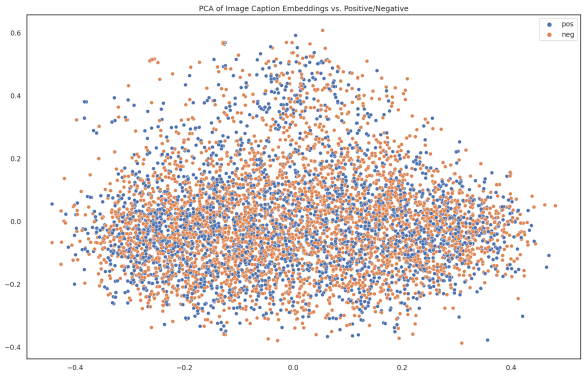

Figure 6. The PCA of Image Caption Embeddings vs. positiveness

As shown from Figure. 4 to Figure. 7, there's no apparent pattern or clusters in regard to question types or fact positiveness. This gives us a rough impression that the images or image captions alone 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.

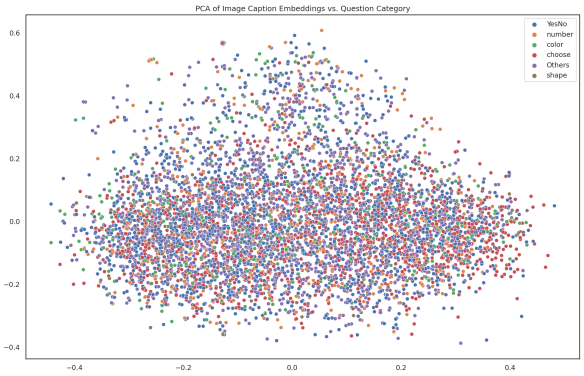

Figure 7. The PCA of Image Caption Embeddings vs. Question Categories

3.3. Image Classification Experiments

We extracted a subset from WebQA that has a balanced number of positive and negative facts. See Table 5 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 more 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 classifiers share the same structure, a ResNet50 (He et al., 2015) trained on ImageNet dataset (Russakovsky et al., 2014) followed by a linear output layer. They are finetuned on the image subset for 15 epochs.

The results shown in Table 6 indeed meet our expectation. The topic classifier shows some classification capability in terms of accuracy and F1 score on classifying 74 topics. However, the question type and positiveness classifier failed to even converge. And their classification performance is no better than random guesses.

	Accuracy	F1 Score	Precision	Recall
Topic	0.32	0.35	0.35	0.40
Question Type	0.33	0.19	0.20	0.21
Positiveness	0.55	0.50	0.51	0.51

Table 6. Image classification model performance on test set

In conclusion, we believe that the image negatives cannot be distinguished from the positives without seeing the questions.

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.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D. A., Bernstein, M. S., and Fei-Fei, L. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. doi: 10.1007/s11263-016-0981-7. URL https://doi.org/10.1007/s11263-016-0981-7.

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.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL https://arxiv.org/abs/1908.10084.

Ren, S., He, K., Girshick, R. B., and Sun, J. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015. URL http://arxiv.org/abs/1506.01497.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. S., Berg, A. C., and Fei-Fei, L. Imagenet large scale visual recognition challenge. *CoRR*, abs/1409.0575, 2014. URL http://arxiv.org/abs/1409.0575.