Proceedings of Machine Learning Research 176:232–245, 2022 NeurIPS 2021 Competition and Demonstration Track

WebQA: A Multimodal Multihop NeurIPS Challenge

Yingshan Chang YINGSHAC@CS.CMU.EDU
Yonatan Bisk YBISK@CS.CMU.EDU
Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA USA

Editors: Douwe Kiela, Marco Ciccone, Barbara Caputo

Abstract

Scaling the current QA formulation to the open-domain and multi-hop nature of web searches requires fundamental advances in visual representation learning, multimodal reasoning and language generation. To facilitate research at this intersection, we propose WebQA challenge that mirrors the way humans use the web: 1) Ask a question, 2) Choose sources to aggregate, and 3) Produce a fluent language response. Our challenge for the community is to create unified multimodal reasoning models that can answer questions regardless of the source modality, moving us closer to digital assistants that search through not only text-based knowledge, but also the richer visual trove of information.

Keywords: Vision-and-Language, Multimodal reasoning, Multimodal information retrieval, Knowledge aggregation, Multi-hop question answering

1. Introduction

Amazing strides have been made in the Question-Answering (QA) field. A myriad of tasks have emerged such as open-domain QA Rajpurkar et al. (2016); Dhingra et al. (2017), VQA Goyal et al. (2017); Antol et al. (2015); Marino et al. (2019); Hudson and Manning (2019), multi-hop QA Yang et al. (2018), and multi-modal QA Hannan et al. (2020); Singh et al. (2021); Talmor et al. (2021). However, at the very core of existing QA systems still lies keyword matching, span-based extraction and classification over a pre-defined answer vocabulary. While the wild success achieved by existing QA systems is undeniable, those models still largely fall short of human performance because they lack language groundable visual representations for novel objects and the ability to reason over multiple pieces of knowledge. Take for example a question which requires combining multimodal information dispersed in different knowledge pieces to reveal the whole picture (below): *Are the Golden Lion, Golden Unicorn and Golden Eagle statues on the same side of the Old State House?*

Even if a question answering system successfully figured out that the lion and unicorn statues are on the east side while the eagle statue is on the west side, it would struggle without the ability to reason about union or intersection. However, humans are comfortable with this kind of reasoning process because we can switch between modalities with ease as well as combining parts into the whole.

To encourage such multi-modal, multi-hop question answering, we introduced WebQA (Chang et al., 2021),[1] a novel dataset consisting of open-domain, factoid QA pairs that requires information retrieval and aggregation given a list of external snippets and images. Apart from questions that require reasoning over images, WebQA also includes text-only questions to ensure improved visual reasoning performance does not come at the cost of textual reasoning. The challenge motivates the move to build more intelligent search engines which are able to bear the bulk of information filtering and aggregation work for downstream users. Our formulation is specifically constructed to center on core scientific questions while allowing for rapid deployment of resulting technologies for shared interest with industry.

2. Current Challenges

We briefly discuss three important aspects of Question-Answering research and their associated challenges. These factors are independent of one another and if simultaneously achieved would accurately reflect question answering tasks we engage in our daily life.

Multi-hop QA requires systems to gather information from different sources of evidence. The subsequent processing of gathered pieces should involve consolidation and reasoning beyond simple concatenation. Based on the logical relationships between different sources, such consolidation may either process the sources in a chain or in parallel. More complex scenarios may also benefit from revisiting an evidence at multiple processing steps. The primary challenge for multi-hop QA lies in how multiple sources can be effectively combined such that they collectively inform the derivation of the final answer.

Multi-modal QA requires systems to take input from different modalities. This is necessitated by the fact that knowledge can exist not only in a text-only landscape, but also in visual or audio forms. The main difference between multi-modal QA and VQA is that, the multi-modal property lies on the knowledge side for multi-modal QA, versus on the question side for VQA. Also, multi-modal QA is more demanding than VQA because a query is usually agnostic to the answer modality, although it is possible to infer a preferred answer modality from the content of the query. In contrast, a query in VQA is, by task definition, paired with a given image.

Text and Image are the two dominant modalities at interest, since they are the major sources humans acquire knowledge from, though, some recent work (Chen et al., 2020; Talmor et al., 2021) has considered tables as a modality different from text, given their more structured organization of information. Speech is another important knowledge-carrying modality, especially targeting at low-literacy or visually-impaired users, which grants future exploration. Since different modalities vary a lot in the raw input format, (e.g. language comes in discrete symbols, images in 2D pixel-matrices, speech in waveforms), research has focused on jointly representing multi-modal information in a single unified space before

1. All data and leaderboard are available at https://webqna.github.io/

further consolidation. Such joint representations highlight the main difficulty of multimodal QA. Existing unification approaches include joint embedding or translation. Joint embeddings aim to represent text and images in the same embedding space so subsequent processing can operate agnostic to the modality. Translation-based approaches focus on translating information in one modality (typically images) into the other modality (typically text), so that they can be analyzed in one representational space. This often leverages image captioning (Li et al., 2019) and scene-graph generation (Shi et al., 2019).

Open-domain QA In the literature, the "domain" can both refer to the answer domain and the knowledge domain. An open answer domain means answers are produced by free-form generation, as opposed to classification over pre-defined candidates. This requires a generative model, whereas most current models focus on discriminative tasks. An open knowledge domain means that, in order to answer the query, one is expected to first perform knowledge retrieval in an unrestricted knowledge base (e.g. the Web). However, most existing open-domain QA datasets provide a restricted set of knowledge candidates among which to select supporting evidence, for the ease of prototype building and model comparison. In addition, a more unrestricted retrieval space is often approximated by Wikipedia, the collection of all knowledge candidates across a particular dataset, or a large-scale knowledge graph. To extract knowledge from a large, heterogeneous space, one has to overcome challenges on efficient filtering, indexing and ranking. The authentic "retrieval in-the-wild" setting could also involve extra difficulties such as noise and misinformation handling.

WebQA Challenge Highlights WebQA is interdisciplinary and reflects challenges in all three aspects. WebQA requires a system to 1) represent both text and images (**multimodal**), 2) identify relevant knowledge in either modality (**open knowledge domain**), 3) aggregate information from multiple sources via logical or numerical reasoning (**multi-hop**), and 4) generate answers in free-form natural language (**open answer domain**).

3. Problem Definition

Given a question Q, and a set of candidate sources s_1, s_2, ..., s_n, a system is expected to answer the question with a natural language sentence and indicate which sources were used to support the answer. A source can be either a snippet or an (image, description) pair. Under this task setup, systems are required to reason over a heterogeneous space without knowing in advance which modality the target knowledge comes from. In total, WebQA has over 34K QA training pairs, with an additional 5K and 7.5K held out for development and testing. See Chang et al. (2021) for a detailed breakdown of dataset statistics.

4. Related Datasets

Recent work in question answering has transitioned from multiple-choice (Clark et al., 2019; Marino et al., 2019; Hannan et al., 2020) and span prediction (Talmor and Berant, 2018; Yang et al., 2018; Tu et al., 2019; Welbl et al., 2018; Choi et al., 2018; Chen et al., 2017; Hannan et al., 2020; Joshi et al., 2017) to the more general open-domain paradigm. In addition, novelty of newly released datasets have centered around multimodal knowledge retrieval (Talmor et al., 2021; Singh et al., 2021), answer modality disambiguation (Hannan

	#Train	#Dev	#Test	#Img	Len Q	Len A
VQA v2 (Goyal et al., 2017)	443K	214K	453K	200K	6.1	1.2
OKVQA(Marino et al., 2019)	9.0K	0	5.0K	14.0K	8.1	1.3
MultiModalQA(Talmor et al., 2021)	23.8K	2.4K	3.6K	57.7K	18.2	2.1
ManyModalQA(Hannan et al., 2020)	2.0K	3.0K	5.1K	2.9K	–	1.0
MIMOQA(Singh et al., 2021)	52.4K	0.7K	3.5K	400.0K	–	–
WebQA (ours)	34.2K	5K	7.5K	390.0K	17.5	12.5

Table 1: Comparison of relevant benchmarks by size and average question/answer lengths.

	Eval Metrics	Answer Schema
VQA v2 OK-VQA	$min\{\frac{\#\text{human agreement}}{3}, 1\}$	Top training answers
MultimodalQA	Exact Match F1	Txt: span/Y/N Img: Fixed vocab Table: Y/N, cell selection, or logical operation.
ManymodalQA	Classification Accuracy	Word selection from context or vocab
MIMOQA	Txt: ROUGE-1/-2/-L or BLEU Img: Precision@1/@2/@3	Span prediction + Image retrieval
WebQA (ours)	Fluency: BARTScore Keyword Acc: Recall/F1	Complete NL sentence

Table 2: Comparison of evaluation metrics and answer schemas

et al., 2020) as well as information consolidation with logical operations (Talmor et al., 2021) or message passing along a reasoning chain (Jhamtani and Clark, 2020; Yang et al., 2018; Welbl et al., 2018; Chen et al., 2020). Table 1 and 2 compare WebQA with recent related datasets in this field in terms of basic statistics, evaluation metrics and answer schema. Specifically, we compare WebQA with MultiModalQA (Talmor et al., 2021), ManyModalQA (Hannan et al., 2020) and MINOQA (Singh et al., 2021) in detail, since they share the most high-level commonalities, while each highlights different sub-challenges.

MultimodalQA focuses on heterogeneous knowledge extraction across snippets, tables and images. However, MultiModalQA requires different answer schemas for TextQA, ImageQA and TableQA. TextQA answers are either a span, "yes" or "no", while ImageQA selects from a fixed answer set. TableQA can select a table cell, combine several cells, or produce "yes"/"no". By contrast, we state that modality unification has to be done either explicitly or implicitly before generating the final answer.

ManyModalQA features ambiguity in the choice of answer modality. We focus more on building multimodal knowledge space and less on distinguishing the answer modality. Importantly, ManyModalQA restricts all answers to be a single word chosen from a predefined vocabulary and given context. While such simplification facilitates easy evaluation, it is unnatural when considering the unlimited coverage of the Web, and constantly shifting

domain knowledge, since the finite answer space imposes a hard limit on the flexibility of an answering system. To handle the more general real-world use case, we formulate WebQA as a free-form generation task.

MIMOQA highlights accompanying a textual answer with an image to improve the completeness and informativeness of an answer presented to the user. Their approach only requires span prediction and source selection, both under the classification banner. We differ by our additional requirement of summarizing selected snippets and images into a natural language sentence, suggesting our advantage over MIMOQA in reasoning depth.

To summarize our unique contributions, previous work does not require the answers to be complete, free-form natural language sentences, as opposed to extractive spans, or elements from a finite set. Also, no existing multi-modal QA challenges has supported both natural language generation (NLG) evaluation and accuracy-style evaluation as we do.

5. Dataset Collection

We collected data via crowdsourcing with a five-stage pipeline: qualification, interestingness filter, QA-pair creation, validation and multi-human-reference generation. Annotator pay averaged $13/hr overall, including the qualification HITs which paid less than annotation. We closely spot-check quality and generously bonus out-of-the-box thinking to incentivize workers for producing high-quality and diverse samples.

Qualification We restricted to crowdworkers located in the US or Canada, who has at least 1,000 approved annotations with an approval rate $\geq$ 95%. The pool of annotators were selected with a tutorial and a qualification exam. The exam included 15 annotation examples, some of which obviously violated the annotation guidelines. Annotators had to point out problematic examples as well as the specific instruction violated. One had to score at least 80% on the qualification exam in order to be granted access to our main task interface. Considering that workers who had patience to complete the qualification exam were generally more coachable, we gave workers who achieved 60% - 80% at their first attempt a second chance.

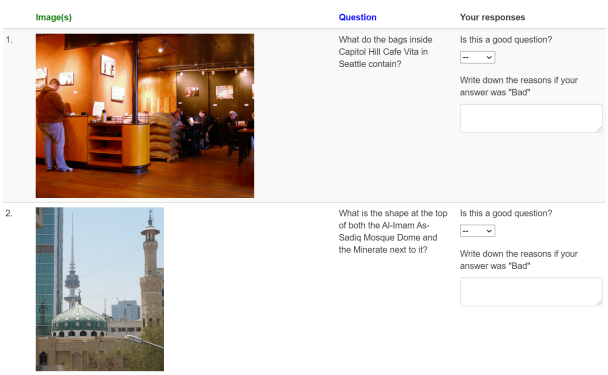
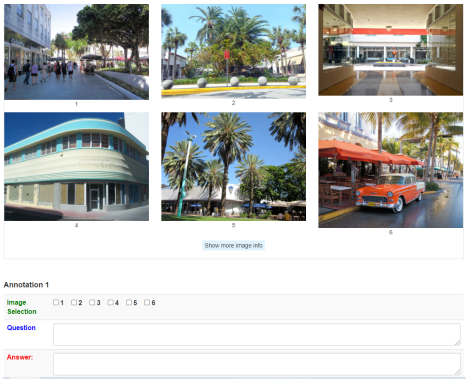
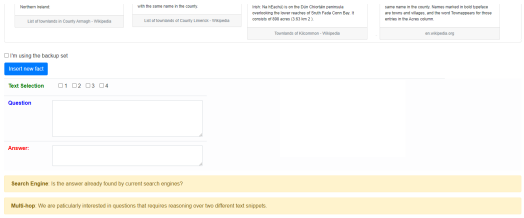


Image Interesting-ness Filter We aimed to include visually interesting images as the evidence for our knowledge-seeking queries. Most categories in the topic list on Wikimedia Commons do not satisfy this criterion. Therefore, we seeded our image pool with natural scenes and designed a filter HIT to obtain image groups that are both semantically relevant and visually interesting. In the filter HIT, we presented 10 images at a time, which are returned by an Image Search API call using a seed term. Tasks for the annotators included 1) selecting 3 out of the 10 images that are distinct, but related, and 2) assigning a label that would best summarize the shared semantics. We further paired up image triples obtained

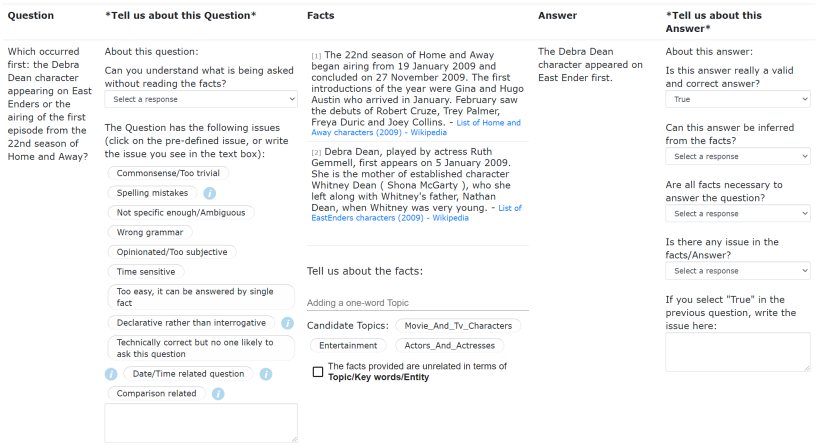
from this stage according to the cosine similarity between their semantic labels, resulting in groups of 6 images which were to appear as prompts in the QA-pair creation stage.

QA Pair Creation We crowdsourced question-answer pairs with either text- or image-based sources. All sources were crawled from Wikipedia and grouped according to their topics. In each HIT, an annotator was presented with a group of 6 candidate knowledge sources, among which one had to select one or two and write down a question attending to information present in the selected sources. Annotation tasks were released batch-wise to facilitate expert-feedback-in-the-loop for quality maintenance.

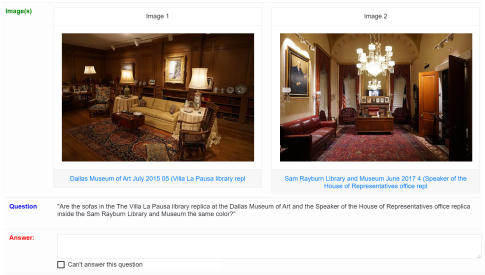




Validation All data was run through validation HITs to ensure the correctness of source selection, the necessity of all sources, and the correctness of the answer. Samples were only kept upon agreement of 3 different validators.



Multi-human-reference generation On the test set, each question received answers from multiple (3-6) humans for improved evaluation and to compute human performance.



6. Evaluation Metrics

The goal of WebQA is a system that answers a question, by aggregating knowledge to produce an accurate and fluent natural language response. To assess progress, on this overall goal we evaluate model performance with respect to both source retrieval and question answering separately. While source retrieval is easily evaluated via F1, question answering quality is factored into two scores: fluency (**FL**) and accuracy (**Acc**). **FL** is based on BARTScore (Yuan et al., 2021), which measures the grammaticality and general semantic relevance between an output and a reference. However, we cannot rely solely on embedding-similarity-based metrics (Yuan et al., 2021; Zhang et al., 2019) as they cannot precisely distinguish between entities within an answer domain (e.g. all color words). Therefore, we also compute an accuracy measure, **Acc**, with precision and recall within a corresponding answer domain, which not only penalizes mindless guessing, but also remains lenient with superfluous words outside the answer domain.

7. Participating and Comparison Systems

Below we detail several of the most important components of the natural baselines and strongest performing models in this space.

Baseline: VLP+x101fpn This system finetuned VLP (Zhou et al., 2020) with a ResNeXt-101 FPN image feature extraction backbone. Separate models were finetuned for source retrieval and question answering respectively on top of VLP's released checkpoints.

Baseline: VLP+VinVL This system replaced the image feature extraction backbone in the previous system with the latest state-of-the-art visual representation model VinVL (Zhang et al., 2021). Other training techniques remained the same.

PICa To explore to what extent models can answer web search queries by leveraging implicit knowledge stored in parameters, this system prompted GPT-3 (Brown et al., 2020) (davinci) with engineered prefixes containing a few examples selected from the training set (Yang et al., 2021). All images were converted to textual descriptions by the Oscar (Li et al., 2020) image captioning model to pass to GPT-3. This system required oracle sources.

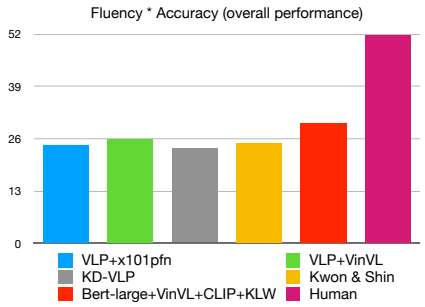
KD-VLP This system used image grid features from CNNs in place of the region features proposed by an object detector. This would overcome the hard limit imposed by a fixed vocabulary of 1600 object classes on which an object detector was trained. The multimodal transformer was fine-tuned from KD-VLP's released weights. KD-VLP (Liu et al., 2021) was pre-trained to gain object awareness with knowledge distillation from external object detectors. Please refer to Liu et al. (2021) for more details.

BERT-Large+VinVL+CLIP+KLW In this system, the multimodal transformer was initialized from BERT-Large (Devlin et al., 2018) and visual representations were augmented with CLIP (Radford et al., 2021) features. Additionally, the training objective for QA was modified to put greater weight (keywords loss weight, KLW) on keywords that describe shapes, colors and binary judgments (i.e. yes/no). More balanced data sampling strategies were also proved to be helpful. Specifically, for the retrieval task, instances corresponding to different ground-truth sources were grouped into one batch. For the QA task, all QA

instances are fully shuffled to guarantee that both image- and text-based instances are present in each batch. During inference, the threshold for source retrieval was dynamically determined with multiple stages to reduce both false negatives and empty yields.

Kwon & Shin For retrieval, this system initialized transformers with Zhou et al. (2020) and Sharma et al. (2018) and performed knowledge distillation from the provided baselines (Section 7). This resulted in two distilled models $\mathcal{M}_{vqa}^{dist}$ and $\mathcal{M}_{cc}^{dist}$, which were ensembled into $\mathcal{M}^{ens}$. The retrieval results were produced by binary classification via lightGBM, where the input features were the modality information of each source and variant statistics of scores (e.g. mean, median, quantiles, top2, ...) predicted by $\mathcal{M}^{ens}$. This provides the final prediction model (lightGBM) a global view of all candidate sources, rather than making independent decisions for each source. For the QA task, two transformers were trained on Visual Question Answering (VQA) (Zhou et al., 2020) and Google's Conceptual Captions (Sharma et al., 2018) initializations, which were ensembled to produce the final answer.

8. Results



During the competition, retrieval F1 improved dramatically, by ten absolute points, and QA performance (**FL∗Acc**) by five points. The strongest entry came from aggregating representations from multiple sources. Specifically, **Bert-large** with **+VinVL** (Zhang et al., 2021) **+CLIP** (Radford et al., 2021) **+KLW** (Dong et al., 2019). When combined, this system won first place in the challenge, with an integrated 29.92 **FL∗Acc** score. A very different approach was to leverage **PICa** (Yang et al., 2021) to achieve a 40.1 **FL ∗ Acc** score. Since **PICa** is not capable of source retrieval, it was not included in the final ranking for fair comparison, but serves as an upper-bound for the best possible performance of the strongest known NLG models given perfect source information. Though important insights have been shared among the Vision and Language community, current models still lag far behind human performance, making WebQA an important playground for further improvements.

9. Lessons Learned and Open Challenges

9.1. Extending retrieval to full-scale

Given the great interest in open-domain, multi-modal retrieval as a prerequisite for the QA task, we added a full retrieval setting where a retriever is supposed to select from the entire collection of sources (∼900K) across the dataset. We benchmarked full-scale retrieval on WebQA with both sparse (BM25 Robertson and Zaragoza (2009)) and dense (CLIP Radford et al. (2021)) retrieval methods and observed a huge drop in F1 score at this large scale (see Chang et al. (2021)). What holds promise is that having VLP re-rank the top 20 sources predicted by CLIP doubles F1, suggesting ample room for a future of large-scale coarse-to-fine filtering that better deals with the efficiency-accuracy trade-off.

	Retrieval	Fluency	Accuracy	FL*Acc
VLP+x101pfn	68.86	44.56	40.35	24.53
VLP+VinVL	70.87	45.73	42.16	25.86
KD-VLP	74.55	42.7	38.51	23.57
Kwon & Shin	**78.45**	46.13	40.25	24.96
Bert-large+VinVL+CLIP+KLW	76.59	**48.47**	**47.10**	**29.92**
Human	90.54	55.09	94.33	52.35

Table 3: Comparison of top performing submissions against the baselines on both individual tasks and overall performance (grey background). Specifically, models performed retrieval (showing the largest gains) in order to then answer questions (Accuracy) in natural language (Fluency). The combined result is evaluated as FL*Acc.

9.2. Multi-modal QA Modelling

Be cautious when using pre-trained object detectors for open-domain QA First, the category list on which object detectors are trained is impoverished. This leads to the problem that the vision-and-language transformer cannot establish a meaningful association between the image regions and text tokens, even when the classification tags of region proposals are included in the input. Apart from failing to provide meaningful associative links, a finite vocabulary also restricts the image features into representing only those entities that are deemed important in object detection benchmarks, leading to defective image representations for downstream QA. For example, the most popular object category returned by a detector is "person". Nonetheless random figures detected in an image rarely become the supporting evidence for a knowledge-seeking query.

Second, object detectors are not optimized towards distinguishing visual properties that an object classifier should be invariant to. For example, varying the color, texture or brightness on an object should not change its category. However, distinguishing between those low-level features and verbalizing the results into language are the core skills required by WebQA and web queries more generally.

Third, the object detection task does not equip the model with the ability to associate descriptive language with visual properties. Queries in WebQA target a multitude of knowledge types that can be gathered from visual cues, including textures, shapes, colors, cardinality and object affordances. Unfortunately, an image encoder is unaware of how to verbalize all these interesting aspects. And it is unclear whether this ability should emerge during fine-tuning since the autoregressive MLM loss only serves as distant supervision for the associative behavior between visual cues and descriptive language.

Multi-modal fusion does NOT naturally emerge Despite the rapidly advancing architectures, objectives and pre-training tasks in multi-modal modelling, cross-modal fusion still remains largely unsolved. This is a bitter lesson informed by the finding that initial modeling attempts to solve WebQA do not know how to "properly" read images, even though images are explicitly given as knowledge sources. This failing explains why a) ablating out the visual input does not lead to dramatic performance drop, b) more performance

gain can be obtained with a large-scale pre-trained language model even if it does not take visual input, c) directly fine-tuning from the BERT checkpoint demonstrates an advantage over the VLP checkpoint. Importantly, simply matching the dimensions between visual and textual token embeddings so the model can easily integrate components, is far from a working solution for multi-modal fusion.

We posit that insufficient (visual) attention guidance during training constitutes the reason why the supposed visual understanding component in vision-and-language models failed to play a constructive role. We do not deny that pre-trained vision-and-language models are good at image-text matching when each image is treated as a single component to score. However, in the WebQA task setup, it is more helpful to view an image as a cluttered collection of distinct components which may or may not provide useful clues. This demands proactive decision on where to attend to within an image. Nevertheless, all approaches implemented for WebQA at this point do not incorporate instance-level supervision that would teach the model "where to look at" within an image.

Most of the time, holistic image encodings concatenated with the textual embeddings are fed to the transformer in a brute-force manner, with an unrealistic hope that stacked attention layers should somehow figure out an association between the input and the expected answer. But it is doubtful whether the current supervision signals provide enough discriminative power for the model to learn a robust association between what it was fed with and what it was supposed to produce. We expect that multi-modal QA should benefit from fine-grained supervision guiding the model to "look" at a targeted image area where the answer can be derived from.

10. Conclusion

We have documented here the community's valiant and informative first attempts at tackling the challenge of WebQA (Chang et al., 2021), which was designed to simulate the heterogeneous information landscape one might expect when performing web search. WebQA covers a wide range of open-domain factoid queries that require attending to images for supporting knowledge, while also forcing systems to reason about text and generate language. To recap, WebQA evaluates the following capabilities with respect to open-domain multimodal QA: 1) Perform information retrieval from both text and image collections with hard distractors, 2) Aggregate language-groundable knowledge from multiple resources, as opposed to extracting an existing text span or image patch, 3) Perform logical or numeric reasoning in a multimodal space, and 4) Generate answers in natural languages, as opposed to a classification label.

Various representation methods, model architectures, training techniques, objective functions and data augmentation approaches have been explored to solve the task, but there is still a considerable performance gap between models and humans. We hope WebQA will continue to be a playground for the community to benchmark important aspects of web search, around reasoning, knowledge aggregation, rich visual understanding and the use of a unified model across modalities.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*, 2015.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. *arXiv preprint arXiv:2109.00590*, 2021.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. Hybridqa: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, 2020.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, 2018.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Bhuwan Dhingra, Kathryn Mazaitis, and William W. Cohen. Quasar: Datasets for question answering by search and reading. *CoRR*, abs/1707.03904, 2017. URL http://arxiv.org/abs/1707.03904.

Jiarong Dong, Ke Gao, Xiaokai Chen, Junbo Guo, Juan Cao, and Yongdong Zhang. Not all words are equal: Video-specific information loss for video captioning. *arXiv preprint arXiv:1901.00097*, 2019.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Darryl Hannan, Akshay Jain, and Mohit Bansal. Manymodalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7879–7886, 2020.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 6700–6709. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00686. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Harsh Jhamtani and Peter Clark. Learning to explain: Datasets and models for identifying valid reasoning chains in multihop question-answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 137–150, 2020.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, 2017.

Hui Li, Peng Wang, Chunhua Shen, and Anton van den Hengel. Visual question answering as reading comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6319–6328, 2019.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

Yongfei Liu, Chenfei Wu, Shao-yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. Kd-vlp: Improving end-to-end vision-and-language pretraining with object knowledge distillation. *arXiv preprint arXiv:2109.10504*, 2021.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00331. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Marino_OK-VQA_A_Visual_Question_Answering_Benchmark_Requiring_External_Knowledge_CVPR_2019_paper.html.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100, 000+ questions for machine comprehension of text. In Jian Su, Xavier Carreras, and Kevin Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics, 2016. doi: 10.18653/v1/d16-1264. URL https://doi.org/10.18653/v1/d16-1264.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Hrituraj Singh, Anshul Nasery, Denil Mehta, Aishwarya Agarwal, Jatin Lamba, and Balaji Vasan Srinivasan. Mimoqa: Multimodal input multimodal output question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5317–5332, 2021.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651, 2018.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021.

Ming Tu, Guangtao Wang, Jing Huang, Yun Tang, Xiaodong He, and Bowen Zhou. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2704–2713, 2019.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302, 2018.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. An empirical study of gpt-3 for few-shot knowledge-based vqa. *arXiv preprint arXiv:2109.05014*, 2021.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi

Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1259. URL https://doi.org/10.18653/v1/d18-1259.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. Bartscore: Evaluating generated text as text generation. *arXiv preprint arXiv:2106.11520*, 2021.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049, 2020.