Making the V in Text-VQA Matter

Shamanthak Hegde
KLE Technological University
Hubballi, India
01fe19bcs233@kletech.ac.in

Soumya Jahagirdar
CVIT, IIIT Hyderabad
Hyderabad, India
soumya.jahagirdar@research.iiit.ac.in

Shankar Gangisetty
IIIT Hyderabad
Hyderabad, India
shankar.gangisetty@ihub-data.iiit.ac.in

Abstract

Text-based VQA aims at answering questions by reading the text present in the images. It requires a large amount of scene-text relationship understanding compared to the VQA task. Recent studies have shown that the question-answer pairs in the dataset are more focused on the text present in the image but less importance is given to visual features and some questions do not require understanding the image. The models trained on this dataset predict biased answers due to the lack of understanding of visual context. For example, in questions like "What is written on the signboard?", the answer predicted by the model is always "STOP" which makes the model to ignore the image. To address these issues, we propose a method to learn visual features (making V matter in TextVQA) along with the OCR features and question features using VQA dataset as external knowledge for Text-based VQA. Specifically, we combine the TextVQA dataset and VQA dataset and train the model on this combined dataset. Such a simple, yet effective approach increases the understanding and correlation between the image features and text present in the image, which helps in the better answering of questions. We further test the model on different datasets and compare their qualitative and quantitative results.

1. Introduction

In recent years, deep learning models that require an understanding of visual scenes by answering questions about everyday scenes have become important. Towards this, many works [3, 27] have introduced datasets and methods that present varied types of questions over different scenes. A few works [4, 10, 16, 17, 22] have focused on the datasets that require models to read the text present in the images.

Figure 1. (a) Visual cues can provide additional context and help clarify meaning, (b) while textual cues can provide more detailed information to support understanding. (c) Combined cues that integrate both visual and textual information can be powerful.

Recently a few works have introduced works on text-based video question answering [11, 28]. These datasets and works provide methods with the ability to learn to answer questions belonging to a certain domain. In Fig. 1, we can see that dataset are designed by keeping the domain into consideration. Question-answer pairs in the VQA dataset are framed based on the visual scene only. One such example is seen in Fig. 1(a) *"How many slices of pizza are there?"* .

On other hand, datasets like TextVQA [22] have questions majorly that require textual content in the image to answer the questions and little to no visual information is needed to obtain the answer. An example is shown in Fig. 1(b) where the question is *"What is the license plate number?"* and answers can be obtained just by using the OCR information. Ideally, a good VQA system should be able to **look** and **read** as shown in Fig. 1 (c) where the question is *"What is the number on the middle of the bike?"*. To answer this question, the model should first look at the im-

age and find the region "middle" as instructed by the question. Then, the model should read the number written on the region of interest. Current methods and datasets suffer from bias of domain-specific questions that result in these methods to learn shortcuts to obtain the answers. In [8], the authors introduce a technique of augmenting dataset to remove these language priors existing in the VQA dataset. Complementary images are added to the existing VQA dataset such that language priors are removed. On the other hand, authors of [7, 19], show that there is an obvious gap in the TextVQA models to learn to look at the images while answering the questions and VQA models to learn to read. Though the models trained specifically on domain specific dataset perform very well, they tend to fail when questions from other domains are asked. The language priors in these domain-specific datasets make the existing methods under exploit combining information from multiple modalities but only use these language priors to obtain higher accuracy. To subside this effect of language priors for the task of Text-based VQA, we propose a method to **Make the V in TextVQA Matter**.

In this work, to dissolve this bias, we propose a new method of multimodal training on the union of Visual Question Answering (VQA) [3] and Text-based Visual Question Answering (TextVQA [22] + ST-VQA [4]) datasets. Specifically, we balance the Text-based VQA dataset by adding the images in the VQA dataset which contain text in them. We call this merged dataset as **Union Dataset**. Our dataset is more balanced compared to TextVQA only and VQA only in terms of types of questions that require both looking and reading the image to answer. We train the state-of-the-art models of TextVQA task on the union dataset and perform exhaustive experiments. These models include iterative answer prediction with pointer-augmented multimodal transformers for TextVQA [9] and text-aware pre-training for Text-VQA and Text-Caption [26]. We provide attention maps for a better understanding of the proposed method and compare them with attention maps obtained from existing methods. We also show the generalization of such a method to new datasets like [20] by directly testing on the new dataset and also by fine-tuning on it.
Our main contributions are as follows:

1. We balance the current Text-based VQA datasets by combining (union) images from VQA dataset such that the images should contain textual information. This results in the dataset twice as only Text-based VQA dataset with questions that will make the methods learn to look and read.

2. We evaluate state-of-the-art TextVQA models on the proposed union dataset and show that the models trained on existing out of balance datasets exploit the language prior to obtain answer. This observation
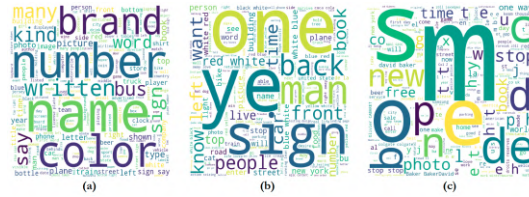

Figure 2. (a) Wordcloud for words in question in our union dataset. (b) Wordcloud for words in answers. (c) Wordcloud for words in OCR tokens.

helps our premise that combining the datasets can help in making the visual information matter in TextVQA. In addition to this, we test our hypothesis and show that it generalizes well on new out of domain test set. We show this by evaluating the performance of the model trained with our union dataset on KVQA [20] dataset.

2. Related Work

2.1. Debiasing in Visual Question Answering

In VQA bias is ubiquitous, existing VQA [3] dataset have biases between the questions and answers. For example, (i) strong correlations between questions and answers, i.e., *language prior* [2, 8] such as answering *"green"* for the question *"What color is the grass?"*, *"tennis"* for the question *"What sports ... "* will obtain 40% accuracy [18], (ii) questioner tends to ask about the objects seen in the image, i.e., *visual priming bias* [3, 8] such as answering *"yes"* to all the questions *"Do you see a ..."* achieves nearly 90% accuracy because the model is trained and tested on the quite different scenarios. Recently, many methods have been proposed to overcome the biases in VQA. These methods can be classified as (i) non-augmentation-based methods [6, 12, 13, 18, 21] seek to reduce the language biases explicitly or improve attention on the image (ii) augmentation-based methods [1, 14, 23, 29] seek to balance the biased dataset for unbiased training.

In [12], the authors use a dual masking strategy, wherein they train a VQA model by masking the most relevant image region or the question words and they use a negative answer assignment mechanism for providing the answers to the counterfactual samples synthesized which exploits the probability distribution of the answers based on their frequency in the original training set. In CF-VQA [18], the authors make use of both question and image, but use the two modalities individually without combining them. They subtract the pure language bias effect from the multi-modal knowledge of standard VQA models. In [29], the authors have proposed a self-supervised learning framework for VQA to automatically balance the biased data. They make use of an auxiliary task named question-image cor-
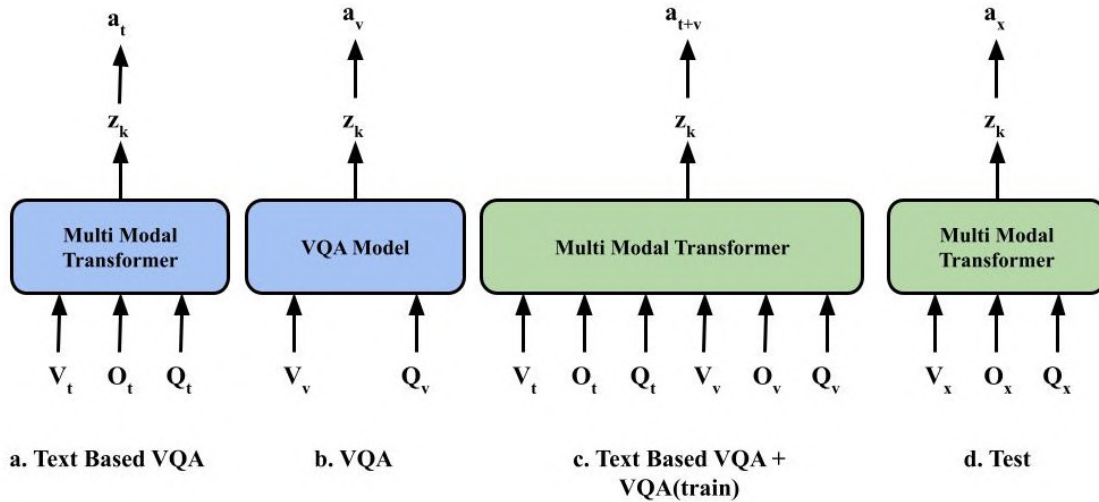
Figure 3. Different models used for TextVQA and VQA and combined tasks.(a) The existing method for Text-based VQA using Multi-Modal Transformer. (b) Existing VQA models for VQA tasks. (c) Our method where we pass combined dataset of Text-based VQA and VQA datasets for training. (d) Testing our method on different datasets.

relation estimation (QICE) to estimate the relevance between questions and images and generate a set of balanced question-image pairs with binary labels as either relevant or irrelevant, which are then used by the self-supervised auxiliary task to assist the VQA model to overcome language priors. In [23], the authors propose a general method to improve OOD generalization. The model is discouraged from using spurious correlations that only appear in subsets of the training data, and rather ensure that it uses reliable ones that are more likely to generalize at test time. More precisely, data is partitioned into multiple training environments such that spurious correlations vary across environments while reliable ones remain stable. by using unsupervised clustering, prior knowledge, and auxiliary annotations in existing datasets. Then, multiple copies of a neural network, one per environment are trained. Some of their weights are shared across environments, while others are subject to a variance regularizer in parameter space. This leads the model to extract features that are stable across environments since they are optimized to be predictive under a classifier common to all environments.

2.2. Biases in Text-based Visual Question Answering

In text-based VQA we expect a model to answer truthfully based on the visual evidence contained in the image, scene text and the correct intention of the question. Unfortunately, this is not always the case even for state-of-the-art methods. Instead of exploiting the image and scene text to find the correct answer, most models frequently rely on spurious correlations and follow the bias that naturally exists within the training data. This severely limits the generalization of Text-based VQA models in real-world scenarios, where the test distribution of facts (e.g., colors, counts, objects, position of objects, etc.) is often different from the training distribution. Few works such as [15, 19] make use of an M4C [9] like multimodal transformer while additionally having to train a separate decoder to ground the answer with bounding boxes in LOGOS [15] and a segmentation network to output segmentation maps of the answer region in MTXNet [19]. These work provide a visual analysis on the region of interest of the model while answering the question. However, as per our knowledge, we are the first to propose solution for debiasing text-based VQA.

3. Benchmarking Text-based VQA Models

In this section, we explain our method, as shown in Fig. 3. Fig. 3a and 3b are models used for text-based VQA (TextVQA and ST-VQA) and VQA datasets. Then, we train a Multi Modal Transformer on the combined dataset wherein we specifically make use of two models M4C [9] and TAP [26]. M4C [9] is a multimodal transformer encoder with a dynamic pointer network decoder to select the answer from either vocabulary or detected OCR tokens. TAP [26] is an extended version of M4C which is pretrained on a large corpus of data, performing tasks such as masked language modelling(MLM), relative position prediction (RPP) and Image-Text Matching (ITM).

3.1. Union of Visual and Text Based Datasets

We combine Text-based VQA: TextVQA + ST-VQA, $Y = (V_t, O_t, Q_t)$ where V_t, O_t, Q_t are objects, OCR, ques-
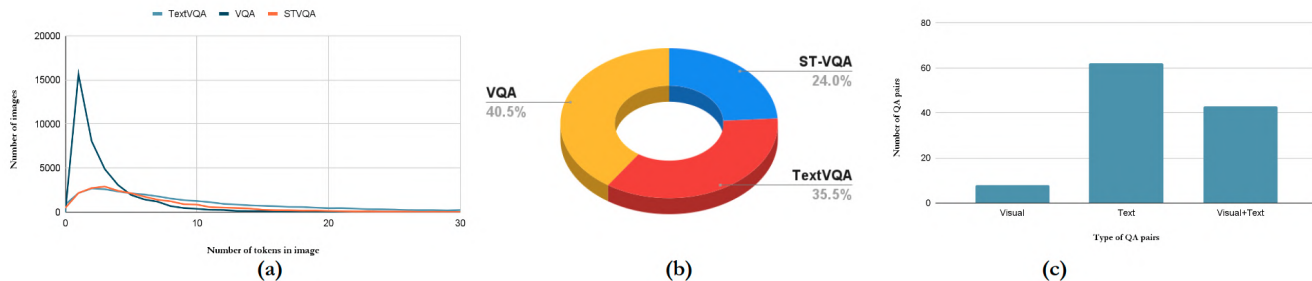
Figure 4. (a) Graph showing the distribution of length of OCR tokens in images of Union dataset. (b) Distribution of our Union Dataset. Our Union Dataset contains 35.5% of question-answer pairs from TextVQA [22] dataset, 24.0% question-answer pairs from ST-VQA [4] dataset, and 40.5% question-answer pairs from VQA [3] dataset. (c) Bar chart showing ablation study when random of 100 QA pairs were given to human volunteers to classify each QA pair based on the answer based on Visual, Textual, or Visual+Textual.

tions and VQA $Z = (V_v, O_v, Q_v)$ where V_v, O_v, Q_v are objects, OCR, questions of VQA and call the combined dataset as **Union Dataset**. Fig. 2 shows the word clouds for (a) words in the question of the Union dataset. (b) words in answers of Union dataset, and (c) words in OCR tokens of Union dataset. In Fig. 4 (a) shows the distribution of length of OCR tokens in images of Union dataset (TextVQA, VQA, STVQA). (b) Shows the % distribution of TextVQA, ST-VQA and VQA dataset in Union dataset. It can be seen that, the Union dataset has balanced distribution.

We extract their corresponding object, OCR features along with the question feature to obtain dataset W. We consider the images from VQA [3] dataset that contains OCR tokens. Fig. 4(b) shows the distribution of the number of question-answer pairs in each dataset. This union or merging of dataset results in balanced question-answer pairs which enable the Text-based VQA models to look (question-answer pairs from VQA dataset) and read (question-answer pairs from TextVQA and ST-VQA dataset). This results in current state-of-the-art Text-based VQA methods to learn to look and read.

$$W = Y \cup Z \qquad (1)$$

3.2. Multi-modal Transformer

We use a multimodal transformer architecture containing three modalities – objects detected V, OCR tokens O and question words Q. We pass the feature embeddings to the model by projecting them in d-dimensional common embedding space with the following steps:

Embedding of detected objects. Given image I, we obtain N visual objects V (generally N is 100) and their corresponding location using a pretrained object detector (MaskRCNN). We consider the location of the jth object where $j = 1, 2, ..., N$ by obtaining the relative bounding box x_j^b. We combine object feature x_j^{fr} and bounding box

x_j^b to get the final object embedding x_j^{obj} of the corresponding object V_j.

$$x_j^{obj} = x_j^{fr} + x_j^b \qquad (2)$$

OCR embedding. We consider the M OCR tokens O (generally M is 50) extracted using EasyOCR (for VQA images) and Rosetta [5] (for TextVQA and STVQA images). We extract the FastText word embedding feature x_i^{ft} of the ith OCR token where $(i = 1, 2, ..., M)$ along with the appearance feature x_i^{ap}, and the bounding box of each token, we sum it to get the OCR embedding x_i^{ocr} of corresponding OCR token O_i.

$$x_i^{ocr} = x_i^{ft} + x_i^b \qquad (3)$$

Question words embedding. We embed the Q question words (generally Q is 20) to a feature vector x^{ques} using a pretrained BERT. We only use the first three layers of BERT to extract features of question words.

After embedding all entities from each modality as vectors in the d-dimensional joint embedding space, we apply a stack of L transformer layers [24] with a hidden dimension of d over the list of all entities. Through the multihead self-attention mechanism in transformers, each entity is allowed to freely attend to all other entities. Using the same transformer layers as a decoder, we predict the answer word by word in an autoregressive manner for a total of T steps.

4. Experiments

In this section, we experiment and validate the performance of the proposed method of combining VQA and Text-based VQA datasets for better generalization of VQA systems that can see, read and reason. We first discuss the datasets in Sec. 4.1. The quantitative and qualitative results are presented in Sec. 4.4 and 4.5 respectively. We also provide several ablation studies in Sec. 4.6.
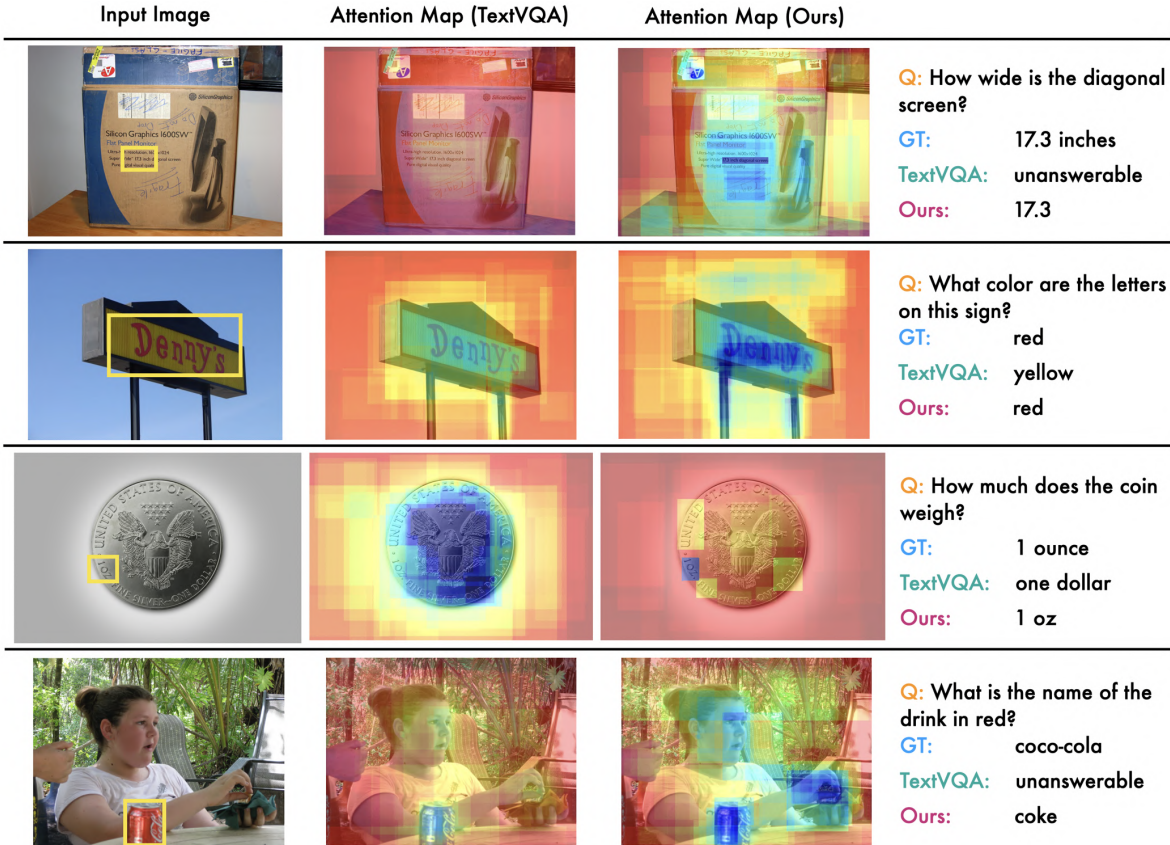
Figure 5. Qualitative results of TAP: Comparison on TextVQA and Ours with attention maps.

Method	Data for pretraining	Data for Finetuning data	Test Acc.
M4C	-	TextVQA	39.01
M4C	-	TextVQA + VQA + STVQA	39.16
TAP	TextVQA	-	49.71
TAP	TextVQA + VQA + STVQA	TextVQA	47.75

Table 1. Evaluation of TextVQA [22] test data on text-based VQA models trained on our Union dataset. It can be seen that combining data from multiple sources helps the models that can only read to also look at the images thereby answering questions that require both textual and visual reasoning.

4.1. Datasets

To showcase the effectiveness of the proposed method, we make use of three popular datasets, namely, VQA [3], TextVQA [22] and ST-VQA [4]. We obtain the images in the VQA dataset that contains text and combine them with TextVQA+STVQA dataset to obtain a union dataset: **VQA+TextVQA+STVQA**. We organize the test set of TextVQA as a test dataset for the models trained on the union dataset. We also evaluate our method on KVQA [20] to show the generalization and domain transfer of the knowledge learned from the union dataset to a specific domain dataset such as KVQA. The training set of

union datasets comprises 97,578 question-answer pairs and the test set contains 5,734 question-answer pairs.

4.2. Performance metrics.

We use Accuracy as the evaluation metric as it measures the percentage of questions for which the predicted answer matches exactly with atleast three of the target answers for the question.

$$Acc(\text{ans}) = min\left(\frac{\text{No of humans that said ans}}{3}, 1\right) \quad (4)$$

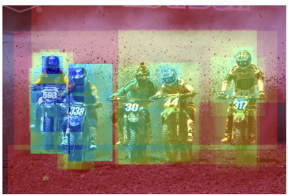We also show attention maps obtained from M4C [9] and TAP [26] trained with the original configurations and

Figure 6. Qualitative results of M4C: Comparison on TextVQA and Ours with attention maps. It can be seen that M4C trained on **Union dataset** performs better for the questions that require models that need both visual and textual explanations to answer questions.

trained on the union dataset.

4.3. Implementation details.

The proposed method can be applied to different and newer approaches proposed in Visual Question Answering. The key idea is, making the visual cues important in text-based visual question answering. We used AdamW as the optimizer. The learning rate for the union dataset is set to $1e-4$. We train M4C [9] and TAP [26] for 24,000 iterations with a batch size of 64. We recognize a maximum of 50 OCR tokens in the union dataset and detect a maximum of 100 objects from the image. We set the maximum decoding steps to 12 and use the answer vocabulary from the union dataset. In the case of experiments on test set of TextVQA, we train M4C on union dataset (TextVQA+STVQA+VQA) and directly evaluate on test set of TextVQA. To evaluate the generalization of the proposed method, we evaluate it on test set of KVQA.

4.4. Quantitative results

We evaluate the performance of the text-based models trained on our Union dataset and compare it against the state-of-the-art models, namely M4C and TAP which are usually trained on TextVQA and STVQA datasets. The comparative results are shown in Table. 1 using accuracy as the evaluation metric, although we cannot quantitatively display the reduction in bias. Among the two main baselines, M4C trained on our Union dataset slightly outperforms the existing M4C trained on TextVQA however, as shown in Sec. 4.5, our model predicts better answers that are unbiased and predicted by looking at the appropriate visual features of the images. In the case of TAP, our model slightly under-performs compared to the TAP model originally trained on just the TextVQA dataset. The main reason for the decrease in accuracy in case of TAP or the small increase in accuracy in case of M4C is because the models now rely lesser on the bias to answer the questions. With the reduced bias, using more relevant and appropriate data would assist the model to predict correct answers and thus

| Input Image | Attention Map (Ours) | Input Image | Attention Map (Ours) |

Q: How many people in this image were born in Russia?

Knowledge: Grand Duchess Olga of Russia reads a book as her sister Grand Duchess Anastasia looks on.

GT: 2

Ours (M4C): 2

Q: How many political parties has the person in the image been a part of?

Knowledge: Tun Abdul Razak Memorial in Kuala Lumpur.

GT: 1

Ours (M4C): 1

Q: How many people in this image were born in United States of America?

Knowledge: Crow at The Grove of Los Angeles, California in 2002, with guitarist Peter Stroud.

GT: 2

Ours (TAP): 2

Q: How many people in this image were born in United States of America?

Knowledge: Secretary of Defense Donald H. Rumsfeld (left) and the Commander of U.S. Central....

GT: 2

Ours (TAP): 2

Figure 7. Qualitative results of M4C and TAP: Answering of KVQA questions based on knowledge with attention maps.

Method	Pre-training data	Fine-tuning data	Test Acc.
MemNet [20]	-	KVQA	50.2
UNITER [25]	-	KVQA	**69.3**
M4C	-	-	22.89
M4C	-	KVQA	47.38
TAP	TextVQA + VQA + STVQA	-	15.68
TAP	TextVQA + VQA + STVQA	KVQA	47.49

Table 2. Evaluation of KVQA [20] test data on text-based VQA models trained on our Union dataset.

increase the accuracy.

4.5. Qualitative results

We showcase a detailed qualitative analysis of our proposed method. In Fig. 6, we show the predictions of M4C model trained on the union dataset and predictions of M4C model trained only on TextVQA dataset. It can be seen that our method predicts correct answers for the questions that require both visual and textual understanding to answer the question. Example: *"What number is on the middle bike?"*, the answer to this question predicted by our method is *"30"*, whereas the M4C trained only on TextVQA has predicted *"598"*. A model trained only on TextVQA suffers from the bias by the type of question-answer pairs present in the dataset. M4C trained only on TextVQA fails to comprehend the meaning of the word "middle", whereas M4C trained by our proposed method can answer the question. This is because of the questions in the VQA dataset, which

Q: How many km's are on the sad sign?
GT: 1,2,3
TextVQA: 2
Ours (M4C):   3

Q: What is the first letter on the woman's back?
GT: trinity
TextVQA: b
Ours (TAP):   w

Figure 8. Ground Truth Limitations: In a few cases the dataset contains ambiguous answers such as the first case where the number of km's on the sad sign is asked, but the ground truth answer provided is *1, 2, 3* which is incorrect.

require the models to look at the image to answer a question, thereby understanding the spatial positions and image features while reasoning over the image to answer the question rather than to just read the text. Fig. 5 shows the qualitative results for predictions made by TAP on the question-answer pairs in TextVQA dataset that require both visual and text reasoning to obtain the answers. It can be seen that, TAP trained on Union dataset can look as well as read the content in the image, whereas original TAP fails to do so. Attention maps for all the examples shown in Fig. 6 and 5 show that the proposed method indeed looks at the image based on the question asked and can also read the required textual content to obtain accurate answers.

4.6. Ablation study

We perform two ablation studies to demonstrate the performance of our model in the case (i) of an external knowledge based VQA dataset named, KVQA - Knowledge aware Visual Question Answering [20] and (ii) wherein the ground truth of a given question in the TextVQA dataset [22] is wrong. It can be seen in the Fig. 7 that the text-based models trained on our Union dataset looks into the appropriate image regions based on the given question and external knowledge provided. In Table. 2, we can see the performance of our text-based models performing well on such datasets as well while giving results that are comparable to the existing models used for the KVQA dataset [20]. UNITER [25] is one of the VQA models generally used for datasets like VQA. It achieves a state-of-the-art accuracy of 69.3 on the KVQA dataset due to its larger pretraining data. Our models get less accuracy (15.68 and 22.89) when it is tested on it without finetuning on the training dataset. However once it has been finetuned, the model is also able to answer such questions that focuses more on the visual features and it achieves an accuracy of approxi-

mately 47.49. Further in Fig. 8, we also show the performance of our models trained on the Union dataset predicting correct answers for the questions with wrong ground truth by looking at the image features and predicting answers by reasoning over the images. As shown in Fig. 8(a), *"How many km's are on the sad sign?"* our model can differentiate between parts of the image by localizing to the 'sad sign' and predicting the number of km's to be *'3'*. On the other side in Fig.8(b), *"What is the first letter on the woman's back?"* is the question to which our model predicts 'w' which is correct when compared with the wrong ground truth to be *"trinity"*.

5. Conclusion

In this work, we address the problem of focusing on the text present in the image compared to visual features and proposed a method to focus on visual features along with the text present in the image. We use a Union dataset, a combination of both Text-based VQA and VQA datasets. We evaluate our method on the state-of-the-art models. We show that our method attends to the corresponding visual features while answering a question. The qualitative result of the samples with the wrong ground truth show that our method outperforms the existing state-of-the-art models in terms of reasoning over the image. Our exhaustive quantitative and qualitative analysis suggests that having an unbiased dataset can result in better-comprehending models thereby taking a step towards well-designed VQA models that are capable of reasoning over multiple modalities. With more appropriate and unbiased data, we could achieve better results and answering through proper reasoning. Self-supervised training on various captioning datasets would help in better understanding of the image and can act as a substitute for the lack of proper scene-text data.

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *CVPR*, 2020. 2

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *CVPR*, 2018. 2

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, 2015. 1, 2, 4, 5

[4] Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Lluís Gómez i Bigorda, Marçal Rusiñol, C. V. Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, 2019. 1, 2, 4, 5

[5] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018. 4

[6] Rémi Cadène, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases for visual question answering. In *NeurIPS*, 2019. 2

[7] Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton van den Hengel, and Qi Wu. Structured multimodal attentions for textvqa. *CoRR*, abs/2006.00753, 2020. 2

[8] 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. *International Journal of Computer Vision*, 2016. 2

[9] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*, 2020. 2, 3, 5, 6

[10] Soumya Jahagirdar, Shankar Gangisetty, and Anand Mishra. Look, read and ask: Learning to ask questions by reading text in images. In Josep Lladós, Daniel Lopresti, and Seiichi Uchida, editors, *ICDAR*, 2021. 1

[11] Soumya Jahagirdar, Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Watching the news: Towards videoqa models that can read. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023*. IEEE, 2023. 1

[12] Camila Kolling, Martin D. More, Nathan Gavenski, Eduardo H. P. Pooch, Otávio Parraga, and Rodrigo C. Barros. Efficient counterfactual debiasing for visual question answering. In *WACV*. IEEE, 2022. 2

[13] Gouthaman KV and Anurag Mittal. Reducing language biases in visual question answering with visually-grounded question encoder. In *ECCV (13)*, 2020. 2

[14] Zujie Liang, Weitao Jiang, Haifeng Hu, and Jiaying Zhu. Learning to contrast the counterfactual samples for robust visual question answering. In *EMNLP (1)*. Association for Computational Linguistics, 2020. 2

[15] XiaoPeng Lu, Zhenhua Fan, Yansen Wang, Jean Oh, and Carolyn Penstein Rosé. Localize, group, and select: Boosting text-vqa by scene text modeling. *2021 IEEE/CVF International Conference on Computer Vision Workshops (IC-CVW)*, 2021. 3

[16] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for VQA on document images. In *WACV*, 2021. 1

[17] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: visual question answering by reading text in images. In *ICDAR*, 2019. 1

[18] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual VQA: A cause-effect look at language bias. In *CVPR*, 2021. 2

[19] Varun Nagaraj Rao, Xingjian Zhen, Karen Hovsepian, and Mingwei Shen. A first look: Towards explainable textvqa models via visual and textual explanations. *CoRR*, abs/2105.02626, 2021. 2, 3

[20] Naganand Yadati Sanket Shah, Anand Mishra and Partha Pratim Talukdar. Kvqa: Knowledge-aware visual question answering. In *AAAI*, 2019. 2, 5, 7, 8

[21] Ramprasaath Ramasamy Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry P. Heck, Dhruv Batra, and Devi Parikh. Taking a HINT: leveraging explanations to make vision and language models more grounded. In *ICCV*, 2019. 2

[22] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards VQA models that can read. In *CVPR*, 2019. 1, 2, 4, 5, 8

[23] Damien Teney, Ehsan Abbasnejad, and Anton van den Hengel. Unshuffling data for improved generalization in visual question answering. In *ICCV*. IEEE, 2021. 2, 3

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, 2017. 4

[25] Peter Vickers, Nikolaos Aletras, Emilio Monti, and Loïc Barrault. In factuality: Efficient integration of relevant facts for visual question answering. In *ACL/IJCNLP (2)*, 2021. 7, 8

[26] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florêncio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. TAP: text-aware pre-training for text-vqa and text-caption. In *CVPR*, 2021. 2, 3, 5, 6

[27] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *CVPR*, 2019. 1

[28] Minyi Zhao, Bingjia Li, Jie Wang, Wanqing Li, Wenjing Zhou, Lan Zhang, Shijie Xuyang, Zhihang Yu, Xinkun Yu, Guangze Li, Aobotao Dai, and Shuigeng Zhou. Towards video text visual question answering: Benchmark and baseline. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 1

[29] Xi Zhu, Zhendong Mao, Chunxiao Liu, Peng Zhang, Bin Wang, and Yongdong Zhang. Overcoming language priors with self-supervised learning for visual question answering. In *IJCAI*, 2020. 2