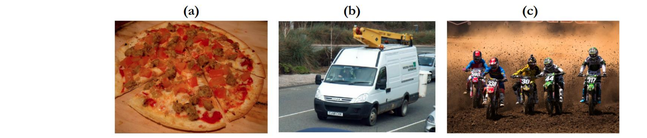
Making the V in Text-VQA Matter

Shamanthak Hegde, Soumya Jahagirdar, Shankar Gangisetty

KLE Technological University — Creating Value, Leveraging Knowledge
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction



- Recent text based VQA models have been found to answer questions without using the information in the image. This diminishes the meaning of VQA resulting in incorrect answers in different scenarios.
- There exists a bias in the TextVQA[1] dataset, that allows the models to predict biased answers to the questions with just OCR tokens given as input, which shows the lack of understanding of the visual features.
- Although image is given as an input to the model, the model predicts correct answers due to the bias in the dataset and not focusing on the visual features.
- On a random set of around 100 QA pairs, our human volunteers classified them as, the questions being answered using visual features, textual features or a combination of both, as shown in Fig 2.

Fig 1: Distribution of Union Dataset

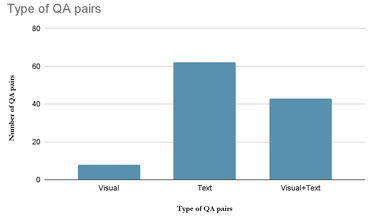
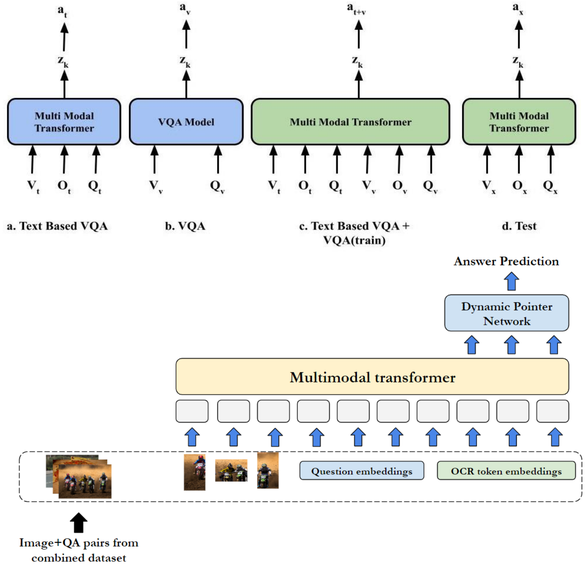


Fig 2: Distribution in the type of QA pairs

Approach



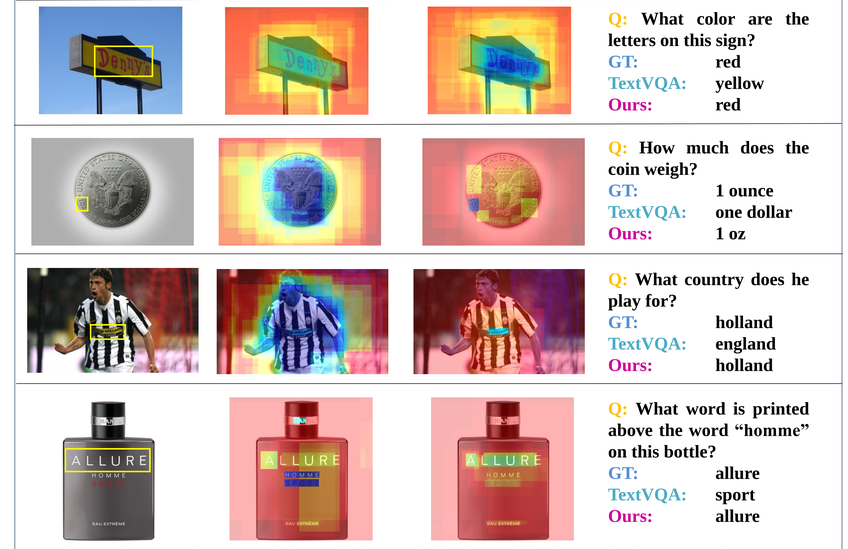
We train a multi modal transformer on our Union dataset consisting of TextVQA, VQA and STVQA. VQA dataset consists of variety of rich object features and text based VQA assists in moderating reasoning in both modality.

Quantitative Results

Method	Data for pretraining	Data for finetuning	Test Acc.
M4C	-	TextVQA	39.01
M4C (Ours)	-	TextVQA + VQA + STVQA	39.16
TAP	TextVQA	-	49.71
TAP (Ours)	TextVQA + VQA + STVQA	TextVQA	47.75

Table 1 in the paper. Accuracy of standard models on TextVQA test set.

Qualitative Results



We show using attention map of the prediction made by the model when trained on our Union Dataset, predicts correct answer by attending to the appropriate regions of the image to answer the question.

Conclusion

We proposed a method to use our Union dataset, a combination of text based VQA and VQA datasets which helped to focus on visual features along with the text present in the image and evaluate on state of the art models. The attention maps generated by our method shows improvement in the reasoning process compared to the original methods. An unbiased dataset, various self supervised tasks can result in creating a well designed VQA model capable of proper reasoning.

References

[1] Singh et. al. Towards VQA models that can read, *CVPR 2019*
[2] Hu et. al. Iterative answer prediction with pointer augmented multimodal transformers for TextVQA, *CVPR 2020*