A First Look: Towards Explainable TextVQA Models via Visual and Textual Explanations

Varun Nagaraj Rao [*] [†], **Xingjian Zhen** [*] [‡] [§] , **Karen Hovsepian** [†] , **Mingwei Shen** [†]
[†] PARS [¶], Amazon.com, Seattle [‡] University of Wisconsin-Madison
varao@amazon.com, xzhen3@wisc.edu, {khhovsep, mingweis}@amazon.com
[*] Equal Contribution [§] Work done while an intern at Amazon

Abstract

Explainable deep learning models are advantageous in many situations. Prior work mostly provide unimodal explanations through post-hoc approaches not part of the original system design. Explanation mechanisms also ignore useful textual information present in images. In this paper, we propose MTXNet, an end-to-end trainable multimodal architecture to generate multimodal explanations, which focuses on the text in the image. We curate a novel dataset TextVQA-X, containing ground truth visual and multi-reference textual explanations that can be leveraged during both training and evaluation. We then quantitatively show that training with multimodal explanations complements model performance and surpasses unimodal baselines by up to 7% in CIDEr scores and 2% in IoU. More importantly, we demonstrate that the multimodal explanations are consistent with human interpretations, help justify the models' decision, and provide useful insights to help diagnose an incorrect prediction. Finally, we describe a real-world e-commerce application for using the generated multimodal explanations.

1 Introduction

The ability to explain decisions through voice, text and visual pointing, is inherently human. Deep learning models on the other hand, are rather opaque black boxes that don't reveal very much about how they arrived at a specific prediction. Recent research effort, aided by regulatory provisions such as GDPRs "right to explanation" (Goodman and Flaxman, 2017), have focused on peeking beneath the hood of these black boxes and designing systems that inherently enable explanation. Explainable multimodal architectures can also be used to reduce the effort required for manual compliance checks of products sold by online retailers. Further, explanations can be provided as evidence to justify decisions and help improve customer and seller partner experiences.

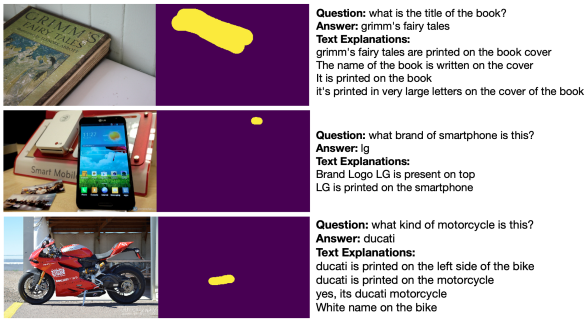

Figure 1: Sample Ground Truth Labels

We choose the TextVQA task proposed by Singh et al. (2019) for realizing the system, motivated by two reasons. First, the task is multimodal and is naturally suited for generating multimodal explanations. Second, the task specifically focuses on the text in the image, known to encode essential information for scene understanding and reasoning (Hu et al., 2020), and allows for better quality of explanations including the text recognized. Several approaches have been proposed for the TextVQA task (Singh et al., 2019; Hu et al., 2020; Mishra et al., 2019; Biten et al., 2019; Kant et al., 2020), but they do not include a means for explaining the model decision. In addition to allowing humans to interpret the model's decision, we believe the explanations can also provide valuable insight into what component could be improved.

Most prior explanation approaches (Hendricks et al., 2016, 2018; Li et al., 2018) have been unimodal and do not focus on the text in the image. Only recently, Huk Park et al. (2018) and Wu and Mooney (2019) generated multimodal explanations for the VQA and Activity Recognition tasks. They curated datasets (VQA-X, ACT-X) consisting of single reference ground truth textual explanations and relied on implicit attention-based visual expla-

[¶] Product Assurance, Risk, and Security https://www.amazon.jobs/en/teams/product-assurance-risk-security

Proceedings of the Third Workshop on Multimodal Artificial Intelligence, pages 19–29

nations without any access to labeled visual ground truth. However, their models cannot read and incorporate text in the image into the explanations. In addition, it is debatable whether attention mechanisms are indeed explanations (Wiegreffe and Pinter, 2019; Jain and Wallace, 2019). Moreover, other works (Das et al., 2017) have shown that current VQA attention models do not seem to look at the same regions as humans, resulting in inconsistent explanations.

The goal of our work is two-fold. First, to collect a multimodal explanations dataset (TextVQA-X) thereby highlighting the need to curate datasets where explanations are not post-hoc but part of the initial interpretable model design. Non post-hoc explanations which may not be faithful to the model decision but are in line with human explanations are still beneficial to end users. Figure 1 provides a representative example. Second, to implement a multimodal explanation system that has the ability to not only read and reason about the text in the image, but more importantly justify its decision with natural language and visually highlight the evidence, useful to even non-experts (Miller et al., 2017). The explanations and model decision must be tightly coupled and mutually influence each other through an end-to-end trainable architecture. In summary, our contributions are as follows:

- We present TextVQA-X, a novel dataset of human-annotated multimodal explanations that includes ground truth segmentation maps and multi-reference textual explanations containing text in the image. The raw dataset is available publicly [1]. (Section 3)
- We propose the first end-to-end trainable MTXNet architecture that produces high quality textual and visual explanations, focusing on the text in the image. (Section 4)
- Qualitative and quantitative results show that textual and visual explanations help justify a model's decision and help diagnose the reasons for an incorrect prediction. (Section 5)
- We describe a real-world e-commerce system that can leverage the multimodal explanations and also highlight its challenges. (Section 6)

2 Related Work

VQA / TextVQA. The VQA task (Antol et al., 2015) has received a lot of research attention in terms of both datasets (Antol et al., 2015; Johnson et al., 2017; Hudson and Manning, 2019) and methods (Anderson et al., 2018; Ben-Younes et al., 2017; Lu et al., 2019). Oftentimes however, these models predict an answer without completely understanding the question and do not change answers across images (Agrawal et al., 2016). Further, they ignore the text in the image and tend to focus on visual components such as objects. To address this limitation, the TextVQA task was proposed by Singh et al. (2019) and has received recent research attention (Kant et al., 2020; Hu et al., 2020; Biten et al., 2019; Mishra et al., 2019). However, not having reliable explanation mechanisms that focus on the text in the image, as part of the system design makes it difficult to diagnose prediction failures. Our work, thus allows for better diagnosis of model failures through explanations in line with human interpretations and focus on the text in the image.

Explanations. Prior explanation approaches (Shortliffe and Buchanan, 1975; Van Lent et al., 2004; Zeiler and Fergus, 2014; Goyal et al., 2016; Ribeiro et al., 2016; Selvaraju et al., 2017; Das et al., 2017) focus on parts of the input that is relevant to the model's decision, but not on explicitly generating explanations as model predictions. Hendricks et al. (2016, 2018) were the first to generate natural language justifications for image classifiers. Unlike our model however, explanations are unimodal and there are no reference human explanations. Closer to our objective Huk Park et al. (2018) generate multimodal explanations and curate a new VQA-X dataset. Wu and Mooney (2019) extend their work to ensure explanations can be traced back to an object ensuring local faithfulness. However, their explanations do not contain the text in the image. They use implicit attention for visual explanations and have no access to visual ground truth during training. Further, they use a single textual explanation reference during training. In contrast, our work incorporates multimodal explanations which focuses on the text in the image.

3 TextVQA-X Dataset

To train and evaluate multimodal explanation models that focus on the text in the image, we collect the TextVQA-X dataset by human annotation of a subset of samples from the TextVQA dataset (Singh et al., 2019).

[1] https://github.com/amzn/explainable-text-vqa

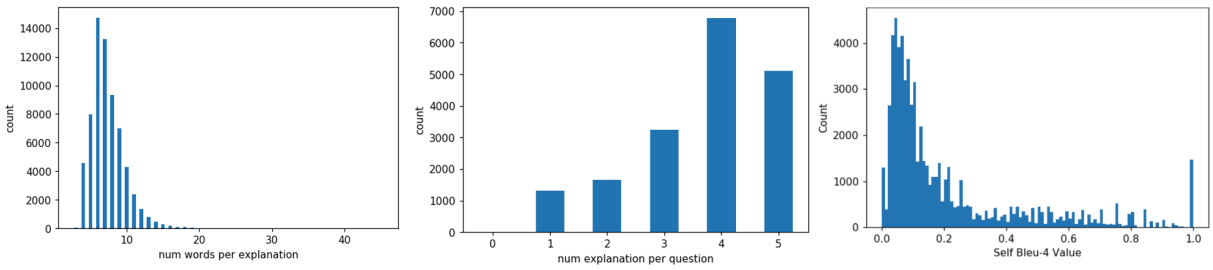

Figure 2: TextVQA-X Dataset Statistics

3.1 Ground Truth Label Collection

We used the Sagemaker Ground Truth (Amazon-AWS, 2018) platform to create a labeling task for gathering visual and textual explanations. Human annotators were asked to provide a single textual explanation that answers the question "Why do you think <answer> is the correct answer for the given question and image pair?". Specific instructions added that annotators should try to incorporate the answer and/or the text in the image as part of their explanation. The annotators were also asked to make use of a brush to segment image regions relevant to both the answer and written explanation. Sample annotations are shown in Figure 1. Each image and question pair can have up to 5 distinct human annotators allowing for multi-reference training and evaluation (Zheng et al., 2018). A single segmentation map is obtained by using a threshold of 0.5 obtained as an average over all annotations. Bad actors were identified and most were removed through a combination of heuristics and manual checks. Overall, we collected more than 67K explanations among over 800 unique workers.

3.2 TextVQA Explanation Dataset (TextVQA-X).

Dataset Statistic	Value
Num. Unique Images	11681
Num. Questions	18096
Num. Unique Questions	15374
Num. Visual Explanations	67055
Num. Textual Explanations	67055
Num. Unique Textual Explanations	61999
Avg. Num Textual Explanations per Question	3.71
Avg. Words per Textual Explanation	7.36
Avg. Characters per Textual Explanation	36.92
Textual Explanation Vocab Size	17910

Table 1: TextVQA-X Dataset Summary

In order to obtain a measure of the quality of explanations and to help filter out bad actors, we make use of the Self-BLEU-4 metric (Zhu et al., 2018). The Self-BLEU score is used to measure how one sentence resembles the rest in a generated collection by regarding one sentence as the hypothesis and the rest as references. A higher Self-BLEU score implies higher similarity of the hypothesis with all the references. A lower Self-BLEU implies higher diversity and lesser overlap. Although we would like to have several diverse textual explanations, we noticed that most good textual explanation annotations have overlap with others. The average Self-BLEU-4 across all annotations was 0.21 indicating consistent overlap and quality.

Comparison with VQA-X and VQA-HAT datasets. With respect to textual explanations, the TextVQA-X includes multi-references with an average of 3.71 explanations for each QA pair that can be utilized for both training and testing. In contrast, VQA-X (Huk Park et al., 2018) contains an average of 1.27 explanations with a single textual explanation for QA pairs in the training set and three textual explanations for test/val QA pairs. VQA-HAT (Das et al., 2017) does not include textual explanations. As far as visual explanations are concerned, there are a number of distinctions among these datasets. First, both VQA-X and VQA-HAT are defined on the VQA task, which does not require reading text in the. In contrast, the TextVQA-X is specifically designed to focus on the text in the image. Second, TextVQA-X includes one ground truth visual explanation for both training and testing (total 67K), whereas VQA-X includes explanations only as part of testing for a small random subset (total 6K). And third, similar to VQA-X, TextVQA-X annotators were asked to directly segment the relevant image region. On the contrary, VQA-HAT annotations were collected by having humans unblur the images and are more likely to introduce noise when irrelevant regions are uncovered.

4 Multimodal Text-in-Image Explanation Network (MTXNet)

We design our Multimodal Text-in-Image Explanation Network (MTXNet) to allow for end-to-end multitask training of answer prediction, text generation and semantic segmentation extending the M4C model proposed in (Hu et al., 2020). In the subsequent subsections we describe each of the individual components in more detail.

4.1 Graph Attention Network (GAT)

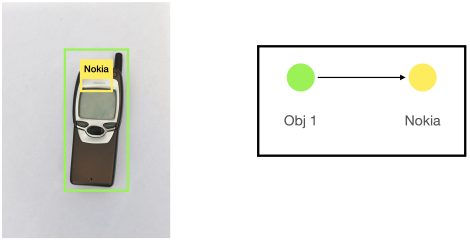

Figure 4: An example of how to build the graph

Many questions in the TextVQA dataset require the model to acknowledge the spatial relationship between objects and OCR tokens. To better encode the relationship between objects and OCR tokens and subsequently generate better quality explanations, we leverage graph neural networks. The ideal way to build the graph is to link together relevant components such as question words, OCR tokens and object labels. However, there are two limitations in the existing TextVQA dataset that prevent us from adopting this approach. First, the OCR tokens may be misspelled due to an inaccurate OCR system. And second, the object labels are not included and only the bounding box coordinates are present. Thus, for our model we build the graph using only the visual inputs (object and OCR region bounding boxes). Each object location and OCR token is treated as a node in the graph. Whenever the bounding box associated with node *i* is contained in node *j*, we add an edge from node *j* to node *i*. An example is presented in Figure 4. We then make use of the Graph Attention Network (GAT) (Veličković et al., 2017) to operate on the structured data. Unlike Graph Convolutional Networks (GCN) (Kipf and Welling, 2016) that treat each adjacent node equally, GATs incorporate attention into the layer-wise propagation rule and allows the model to variably weigh adjacent nodes based on relevancy.

4.2 Multimodal Transformer (MMT)

The multimodal transformer operates on three modalities - question words, visual objects and OCR tokens. The feature definitions are identical to that proposed in M4C (Hu et al., 2020) with the addition of textual explanation embeddings whose embedding process resembles that of the question words. The object embedding is obtained as a combination of the 2048-dim Faster R-CNN detector output and 4-dimensional relative location feature $[x_{min}/W_{im}, y_{min}/H_{im}, x_{max}/W_{im}, y_{max}/H_{im}]$. The OCR token embedding is obtained as a combination of 300-dim FastText vector (Bojanowski et al., 2017), 2048-dim output from fc6 features/ fc7 weights from Faster R-CNN detector for the bounding box region, 604-dim Pyramidal Histogram of Characters (PHOC) vector (Almazán et al., 2014), and 4-dim relative location feature $[x_{min}/W_{im}, y_{min}/H_{im}, x_{max}/W_{im}, y_{max}/H_{im}]$. Features are projected to a common d-dimensional semantic space used for decoding and prediction. The prediction takes place through a dynamic pointer network (Vinyals et al., 2015) that allows to either predict from a fixed vocabulary or from OCR tokens extracted from the image.

4.3 Multireferences for Textual Explanations

Neural text generation tasks such as machine translation, image captioning and summarization typically only consider a single reference for each example during training (Zheng et al., 2018). In our case however, considering just a single reference for training is insufficient because of the inherently subjective nature of textual explanations. Thus we leverage the multi-references we have collected in the TextVQA-X dataset during both training and evaluation. We use the *sample one* technique for incorporating multi-references during training. We randomly pick one of the available references in each training epoch.

4.4 Visual Explanations through Semantic Segmentation

Visual explanations are obtained through a semantic segmentation module (Feature Pyramid Network - FPN (Kirillov et al., 2017)). They are made an explicit and natural component of end-to-end training by leveraging ground truth label supervision. Incorporating explicit visual explanations is known to achieve state-of-the-art results on semantic segmentation benchmarks (Li et al., 2018).

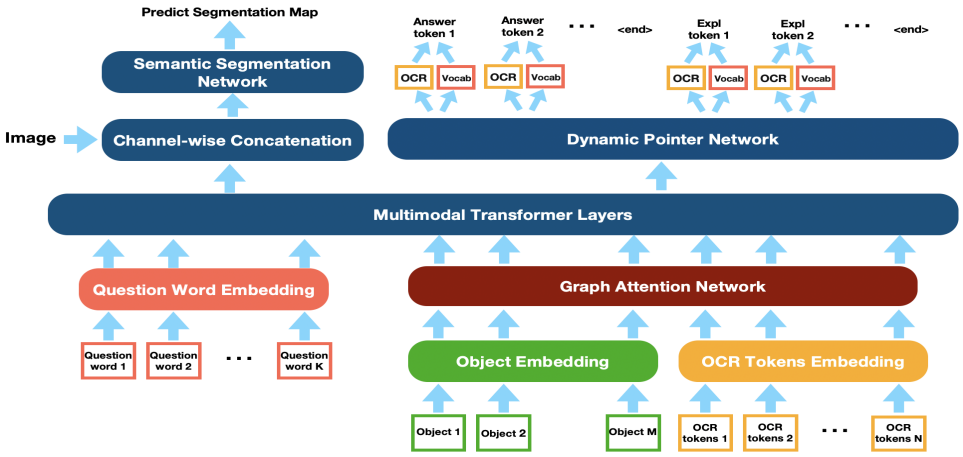

Figure 3: Our Multimodal Text-in-Image Explanation Model (MTXNet) architecture generates multimodal explanations. Explanations and Answers are utilized as a part of the iterative autoregressive decoding procedure.

Moreover, this allows the model to explain the image region in focus, while also providing a means for feedback. On another note, in the complimentary domain of NLP, the use of attention as a means of model explanation has been a topic of considerable debate (Wiegreffe and Pinter, 2019; Jain and Wallace, 2019). We thus leverage ground truth label supervision and explicitly ensure the visual explanation to be part of the training objective. To incorporate the multimodal embedding from the MMT into the segmentation module, we reshape, pad and concatenate the output with the raw input image along the channel. Thus, the overall input channels for the segmentation module increases to five, with 3 color channels and 2 multimodal channels. The output of the segmentation model is a continuous mask with a higher value implying greater relevancy to the inputs. The mask may be binarized through thresholding.

4.5 Training

The MTXNet architecture is end-to-end trainable with three distinct tasks (1) answer prediction (2) textual explanation generation and (3) visual explanation through semantic segmentation. We ensure cross-modal feedback between the textual explanations and predicted answers by leveraging a phased training process where we randomly choose between one of three choices (1) predict answer then textual explanation (2) predict textual explanation then answer and (3) predict both answer and textual explanation independently. Each task corresponds to an individual part of the training objective. For the losses of answer prediction ($\mathscr{L}_{\text{ans}}$) and textual explanation generation ($\mathscr{L}_{\text{text}}$) we use the *binary cross entropy with logits* [2]. For semantic segmentation ($\mathscr{L}_{\text{vis}}$) we use the *dice loss* (Sudre et al., 2017). The naive approach to combine multiple losses is to use a predetermined weighted linear sum of the individual losses. However, the model performance is sensitive to the weights which are hyperparameters and expensive to tune. We thus use a multitask learning loss with homoscedastic uncertainty as proposed by Kendall et al. (2018). The overall objective is present in Equation 1. The weights $\{w_{ans}, w_{text}, w_{vis}\}$ corresponding to the loss terms of the three individual tasks are learned.

$$\mathscr{L} = \sum_i \mathscr{L}_i exp(-w_i) + w_i, i \in \{ans, text, vis\} \quad (1)$$

5 Experiments

In this section, we detail the experimental setup, present quantitative results with ablations and finally analyze qualitative results.

5.1 Experimental Setup

This subsection discusses the dataset splits, model training, hyperparameter settings and evaluation metrics.

Dataset Splits. We use the TextVQA-X dataset described in Section 3. We choose a random 80/20 split for train and test. The dataset split statistics are present in Table 2. Each question is associated with a single image, one or more textual explanations and a single visual explanation. The OCR tokens and object regions are already present in the original TextVQA dataset.

[2] https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html

Split	#Img.	#Ques.	#Text Expl.	#Vis. Expl.
train	10379	14475	53536	14475
test	3354	3619	13507	3619

Table 2: Train / Test Splits of TextVQA-X Dataset

Preprocessing. The dynamic pointer network is allowed to choose between a fixed 5000 word vocabulary and a maximum of 100 OCR tokens per image. For each image, we use the top 36 possible objects extracted by Faster R-CNN sorted in descending order of confidence score attribute. The average number of edges per image was 104. Each image included an average of 13 OCR tokens. The text explanations and answers are capped to a maximum length of 16 and 12 tokens respectively. For the visual explanations, we use a FPN decoder with ResNeXt50 encoder and $320 \times 320 \times 5$ input feature size. The MMT consists of 4 layers and 12 attention heads. The dimension of the joint embedding space is 184×768 which is padded and resized to $320 \times 320 \times 2$ and concatenated with the 3-channel image input.

Model training and hyperparameters. We train the MTXNet model end-to-end in a supervised setting using the Pythia [3] framework. We use a batch size of 128 and train for a maximum of 8500 epochs using Adam optimizer. The learning rate is set to $1e-4$ with no weight decay. The best model is chosen corresponding to the lowest train loss at an evaluation granularity of every 100 epochs. The entire training task varies from 14-20 hours on 8 Nvidia K80 GPUs.

Evaluation Metrics. Each question in the TextVQA dataset has 10 human-annotated answers, and the predicted answer accuracy is measured via a soft voting in accordance with the VQA task evaluation script [4]. We evaluate the textual explanations using the standard BLEU-4 (Papineni et al., 2002), ROUGE (Lin, 2004), METEOR (Banerjee and Lavie, 2005) and CIDEr (Vedantam et al., 2015) metrics computed with the `coco-caption` [5] code . All the text generation metrics account for multi-references by averaging the individual scores. Finally, we evaluate the visual explanations using IoU (Intersection over Union) score with a threshold of 0.5.

[3] https://github.com/facebookresearch/mmf

[4] https://visualqa.org/evaluation

[5] https://github.com/tylin/coco-caption

5.2 Ablation Study

We ablate MTXNet and compare quantitatively with a related model on our TextVQA-X dataset through automatic evaluations for answers and explanations. The results are present in Table 3.

Comparison with existing baselines. We compute the performance of the baseline model M4C (Hu et al., 2020) on the TextVQA-X test set (without explanations) and obtain an answer accuracy of 35.23%. Using the MTXNet architecture and evaluating on the TextVQA-X test set, we obtain an answer accuracy of 36.27%. The addition of explanations thus complements the MTXNet performance.

Unimodal vs. Multimodal explanations We notice that each modality mutually influences the other as the model learns to jointly optimize for both modalities of explanations and the answer prediction. Excluding visual explanations results in the largest drop of up to 7% in CIDEr scores of the textual explanations. Similarly, the absence of text explanations results in a 2% drop in IoU of visual explanations. More importantly, we notice that the multimodal explanations provide visual and textual rationale into a models decision. This further accentuates the value of designing multimodal explanation systems.

GAT better captures structural dependencies. The removal of GAT from the MTXNet architecture adversely impacts the quality of explanations and answers. The greatest drop of 7% is observed for the CIDEr metric. We believe the GAT helps better encode the relationship between objects and OCR tokens enhancing the relationship reasoning ability. The image region corresponding to the text is also highlighted better as seen in the 2% increase in IoU when GAT is included in MTXNet.

Multi-reference training improves text generation. Training with multi-references significantly outperforms training with a single randomly chosen sample fixed for all epochs. The largest increase of up to 25% was noticed in CIDEr score, with the increase being consistent across all text generation metrics. This underscores the benefits of having multi-references for both training and evaluation and designing systems that utilize this effectively.

5.3 Qualitative Samples

As can be seen in Figure 5, the MTXNet is able to accurately answer the given question while also justifying its decision through textual and visual

Ablation	Approach	Visual Explanation	Textual Explanation			
		IoU	B	R	M	C
No visual explanation (VE)	MTXNet (GAT + MR + TE)	-	25.16	47.63	21.76	88.43
No textual explanation (TE)	MTXNet (GAT + MR + VE)	16.10	-	-	-	-
No graph attention (GAT)	MTXNet (MR + TE + VE)	16.55	27.87	49.28	21.61	88.57
No multireferences (MR)	MTXNet (GAT + TE + VE)	17.52	5.92	28.05	11.65	70.60
Consolidated architecture	MTXNet (GAT + MR + TE + VE)	**18.86**	**31.07**	**53.87**	**22.06**	**95.07**

Table 3: Quantitative Evaluation of Answer and Explanations. All metrics are in %. VE: visual explanation, TE: textual explanation, GAT: graph attention network, MR: multi-references. Evaluated automatic metrics: Intersection over Union (IoU), BLEU-4 (B), METEOR (M), ROUGE (R), CIDEr (C).

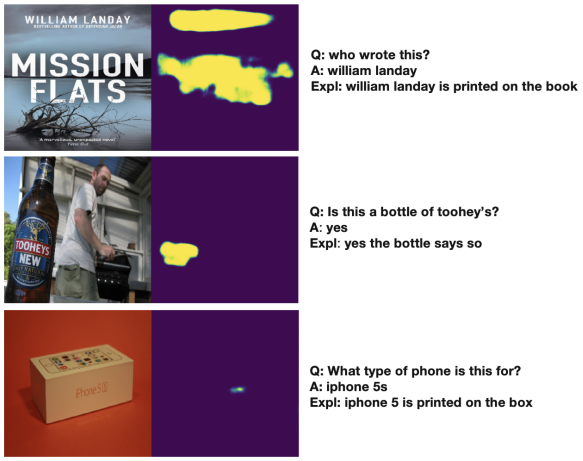

Figure 5: Examples where the MTXNet model produces high quality explanations.

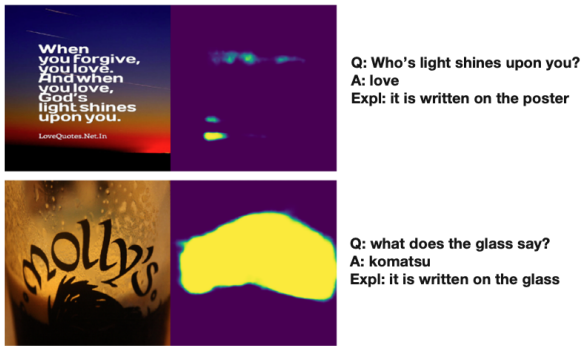

Figure 6: Examples where the MTXNet model fails.

explanations. In certain cases, the OCR engine could be inaccurate and lead to wrong tokens being predicted, but the overall answer and explanations are correct. Figure 6 depicts two failure cases. The upper subimage indicates this could be due to incorrect visual localization while the lower subimage indicates a potential OCR prediction error, although the visual explanation is correct. Despite being generic and dull the textual explanations are correct. In other cases, the model fails due to incorrect visual localization as seen in Figure 7.

Explanations help explain incorrect decisions of model. In Figure 7, we see that the right answer to the question is "target". However, the model

Figure 7: Example where the explanation is consistent with an incorrect prediction.

predicts "dollar tree". From the visual and textual explanations we see that the image region localized is incorrect and the model fails to grasp the meaning of "fading". This potentially results in it focusing on the more prominent "dollar tree" text. Such an analysis provides insights into the component of the system that is failing and deserves further attention.

6 Applications to E-Commerce Businesses

E-commerce businesses need to comply with industry-wide, and country-specific regulations, to provide accurate and useful information of products to improve customer experience that leads to more business. Our long-term goal with explainable multimodal architectures is to automate and reduce manual effort required for compliance and product detail checks. This will enable businesses to scale compliance and customer experience improvement efficiently without linear increases in cost. Further, these architectures help validate if models are performing as intended and used for the right purposes.

A potential customer experience issue arises when the physical product in a warehouse is different from that uploaded by a seller on the product details page. A possible reason could be that

the seller or manufacturer labeled the product erroneously when they packaged it. Many sellers taking advantage of lower cost of manufacturing in a global supply chain, may not be able to audit every batch of product leaving the factory. Such discrepancies will almost certainly lead to product returns, because the customer didn't get what they wanted and increases costs. Such discrepancies may also be due to more nefarious reasons, such as opportunistic bad actors taking advantage of sellers that have successful products by introducing poorer quality or mismatched offers at a lower price to unsuspecting customers. Examples of compliance issues include detecting products that contain batteries and chemicals to comply with transportation and logistics regulations, as well as identifying products that require additional safety documentation and checks, such as products that may have unintended use by children (e.g. toys and products that may end up as toys should not have heavy metals or other poisons that cause illness or death when accidentally ingested). While not all answers can be obtained with product images alone, manual investigation processes utilize these images to identify potential risks that warrant additional steps in the process (e.g. lab testing).

Rather than manually auditing products in a warehouse, product images can be automatically captured at scale, and passed through models that detect such discrepancies. With the help of subject matter experts, attributes such as quantity, color and brand names, and other common misleading attributes are identified apriori. Relevant questions that target these attributes are formulated. The image and question are then inputs to a multimodal explainable system (such as MTXNet) that can provide an answer and justify its prediction through multimodal explanations. Answers can then be compared against the information extracted from the product detail pages on the website. Any discrepancies found can be noted and a selling partner can be provided evidence through the multimodal explanations to take corrective steps.

An example use-case is as follows. Given a large container of cereal, with smaller boxes within, a potential question is: "How many cereal boxes are within the container?" . This information is usually written on the larger container present in the warehouse and can be answered based on reading the text in the image. If there is any discrepancy encountered in the number of boxes of cereal in the warehouse and that listed on the website, appropriate action can be taken. Other similar questions include: "How heavy is the product?", "Is the chair red?", "Does the item contain allergens?", and "Did the product pass the lead test?".

The challenges with the use of such explainable systems are two-fold. First, since there can be multiple stakeholders with diverse expertise and expectations, we need to clearly define the level of abstraction at which they interact with the system. For instance, while a scientist can use the explanations to improve the model, a business operations associate may use the explanations to identify and audit product discrepancies. Second, we need fine grained evaluation methodologies and metrics that take into account the stakeholders as well.

7 Conclusion

A central tenet of explainable AI is to create a suite of tools and frameworks that result in explainable models without sacrificing learning performance and allow humans to understand and trust AI models. As Miller et al. (2017) argues, for explainable AI to succeed, we should draw upon existing principles and create strategies that are more people-centric. Unfortunately most prior explanation approaches have been post-hoc, unimodal, ignore text present in the image and not always in accordance with human interpretation. Further, there is a paucity of labeled multimodal explanation datasets. The research presented in this paper shows that existing TextVQA systems can be rather easily adapted to produce multimodal explanations that focus on the text in the image when given access to ground truth annotations. We curate the TextVQA-X dataset consisting of visual and textual explanations. We then present a novel end-to-end trainable architecture, MTXNet, that generates multimodal explanations focusing on the text in the image, in line with human interpretation and surpasses unimodal baselines (7% in CIDEr scores and 2% in IoU) while complimenting model performance. We also show how the system may be applicable in the e-commerce space to reduce effort for manual audit of compliance checks and improve customer experience. Results of this research open the door to design of explanainable models part of the original system design that effectively takes advantage of available ground truth multimodal explanation annotations. Future work involves incorporating visual features as part of the transformer architecture.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. *arXiv preprint arXiv:1606.07356*.

Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. 2014. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2552–2566.

Amazon-AWS. 2018. SageMaker Ground Truth. https://aws.amazon.com/sagemaker/groundtruth/.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4291–4301.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Mark G Core, H Chad Lane, Michael Van Lent, Dave Gomboc, Steve Solomon, and Milton Rosenberg. 2006. Building explainable artificial intelligence systems. In *AAAI*, pages 1766–1773.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

David K Duvenaud, Dougal Maclaurin, Jorge Iparraguirre, Rafael Bombarell, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. 2015. Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems*, pages 2224–2232.

Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. 2019. Graph neural networks for social recommendation. In *The World Wide Web Conference*, pages 417–426.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*.

Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton van den Hengel, and Qi Wu. 2020. Structured multimodal attentions for textvqa. *arXiv preprint arXiv:2006.00753*.

Bryce Goodman and Seth Flaxman. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI magazine*, 38(3):50–57.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Yash Goyal, Akrit Mohapatra, Devi Parikh, and Dhruv Batra. 2016. Towards transparent ai systems: Interpreting visual question answering models. *arXiv preprint arXiv:1608.08974*.

Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. 2016. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer.

Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. 2018. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 264–279.

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9992–10002.

Drew A Hudson and Christopher D Manning. 2019. Gqa: a new dataset for compositional question answering over real-world images. *arXiv preprint arXiv:1902.09506*, 3(8).

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8779–8788.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Yash Kant, Dhruv Batra, Peter Anderson, Alex Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. 2020. Spatially aware multimodal transformers for textvqa. *arXiv preprint arXiv:2007.12146*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7482–7491.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Alexander Kirillov, Kaiming He, Ross Girshick, and Piotr Dollár. 2017. A unified architecture for instance and semantic segmentation.

H Chad Lane, Mark G Core, Michael Van Lent, Steve Solomon, and Dave Gomboc. 2005. Explainable artificial intelligence for training and tutoring. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE ....

Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo. 2018. Tell-and-answer: Towards explainable visual question answering using attributes and captions. *arXiv preprint arXiv:1801.09041*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *arXiv preprint arXiv:1712.00547*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952. IEEE.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626.

Edward H Shortliffe and Bruce G Buchanan. 1975. A model of inexact reasoning in medicine. *Mathematical biosciences*, 23(3-4):351–379.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 240–248. Springer.

Michael Van Lent, William Fisher, and Michael Mancuso. 2004. An explainable artificial intelligence

system for small-unit tactical behavior. In *Proceedings of the national conference on artificial intelligence*, pages 900–907. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in neural information processing systems*, pages 2692–2700.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Jialin Wu and Raymond Mooney. 2019. Faithful multimodal explanation for visual question answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 103–112.

Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

Renjie Zheng, Mingbo Ma, and Liang Huang. 2018. Multi-reference training with pseudo-references for neural translation and text generation. *arXiv preprint arXiv:1808.09564*.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1097–1100.