1. OCR and Answer Space Analysis

We perform the following analysis on TextVQA’s validation set of size 5,734. We find that 44.9% (2575) of LoRRA’s predicted answers are from OCR tokens (i.e., using the copy mechanism). The remaining 55.1% of predicted answers are from the SA. This shows that our approach does in fact rely heavily on what it reads in the image, and relies on its copy mechanism to generalize and produce answers that have never been seen before or are rare in the training data. While predicting answers from OCR tokens, the model gets the entire answer string correct 27% (696) of the time, and partially correct (i.e., matches one word in answer) 11% (284) of the time. The percentage of partially correct answers shows the possibilities of getting better results by using n-grams of OCR tokens or spelling correction for fixing incorrect OCR predictions. Similarly, while predicting from the answer space, the model gets the answer correct 22.4% (709/3,159).

We find that 30.6% (1,759) of questions have their answers in OCR tokens. For these questions, LoRRA chooses to predict from OCR tokens, 68% (1,208/1,759) of the times and correctly answers 57.5% (697/1,208) of these. Similarly, 48% (2,794) of questions have their answers in SA. For these questions, LoRRA chooses to predict from LA 66.75% (1,865/2,794) of the times and gets 38% (710/1,865) of these correct.

81% questions in TextVQA’s validation set have more than or equal to 2 OCR tokens. Among these 4,645 questions, LoRRA chooses to copy from OCR tokens 49.7% (2,309) and gets 24.3% (560/2,309) of these correct. This suggests that LoRRA doesn’t randomly copy OCR token from list of available tokens.

2. TextVQA Examples and LoRRA Predictions

In Fig. 1, we show representative examples from our TextVQA dataset along with the predictions of Pythia+LoRRA. Each example shows the ground truth answer, the predictions from LoRRA, whether the answer prediction was from OCR tokens or the pre-determined answer space, and attention weights for each of the OCR tokens.

The examples indicate the following points:

- The model is able to successfully answer questions about times, dates, brands, cities and places, and is often able to correctly spell them even if the OCR tokens had them misspelled (as it is able to predict the correct answer from answer space). See Fig. 1I (short hand’s hour), Fig. 1g (birthday date), Fig. 1f (no turn on red sign), Fig. 1s (city “london” correct from the large text), Fig. 1o (samsung).

- The model is able to successfully answer questions involving colors and spatial reasoning. See Fig. 1e (player on the right), Fig. 1h (location of 25 coin), Fig. 1c (location of banner).

- Model is able to distinguish objects based on spatial positioning, colors and relative positioning well. See Fig. 1q where the model needs to identify the correct sign based on multiple colors, or Fig. 1r where the model needs to identify the correct sign in the red circle and shows that the model is not biased toward “stop” as other VQA models are, or Fig. 1a where model needs to predict the correct coin’s value based on the position mentioned.

- The model is also able to reason about basic sizes (less, greater, smallest) and shapes (circle). See Fig. 1f where the model needs to figure out which one is top and silver coin, or Fig. 1k where the model needs to figure out which one is the shorter hand, or Fig. 1q where the model needs to figure out which one is lowest measurement among four.

- The model often predicts an answer from the answer space as informed by OCR tokens. See Fig. 1k where the Pythia model which doesn’t use OCR predicts 3 in this case, but our approach predicts 4 which is the correct answer.

- The model often answers questions about cities with “new york”. See Fig. 1j where the model predicts New York instead of San Francisco. We have observed this bias in other city related questions as well.
• For yes/no questions, even though “yes” is the more common answer, the model does predict “no” frequently. See Fig. 1m, Fig. 1l.

• Sometimes when the answer is not in the answer space, but the partial answer is in OCR tokens, the model predicts the partial answer which is closest to the actual answer. See Fig. 1e where the model predicts “fly” instead of “fly emirates”, or Fig. 1g where the model predicts only the birthday date “19”, instead of “may 19”. By construction our model can only copy a single OCR token, but our TextVQA dataset contains Q/A pairs which require copying multiple OCR tokens in the right order. Explore this is an interesting direction for future work.

• The model sometimes gets seemingly simple questions wrong by predicting generic answers. See Fig. 1h where the model can’t predict “embossed” even though it is in OCR tokens, or see Fig. 1b where the model predicts most common letter “g” in the answer space instead of predicting based on “a-2” in the OCR tokens.

• The model has a strong dependency on the quality of OCR tokens produced. If the OCR module missed some text in the image, the model’s output can be wrong. See Fig. 1i or Fig. 1p where the OCR tokens do not contain the ground truth answer or see Fig. 1u where the OCR system is unable to correctly read “irig” the second time.

3. Interface Screenshots

We show three stages of data collection pipeline in Fig. 2, Fig. 3, Fig. 4, and Fig. 5. Fig. 2 and Fig. 3 shows the entrance and first stage of our task which is used to identify and remove images without text in them. Fig. 4 shows the second stage of our task which is used to collect questions on images with text. Finally, the third stage interface is shown in 5 which is used to collect the answer for a question about an image.
(a) What number is taking the shot?

- Ground Truth: 14
- Prediction: 14
- From: OCR

(b) What letter is written on the blue part of the sign?

- Ground Truth: a
- Prediction: g
- From: Answer

(c) What is the brand name on the blue banner at the top of the image?

- Ground Truth: watsons
- Prediction: watsons
- From: OCR

(d) What country is the brut from?

- Ground Truth: france
- Prediction: france
- From: Answer

(e) What is on the jersey of the player on the right?

- Ground Truth: fly
- Prediction: fly
- From: OCR

(f) How much is the top silver coin worth?

- Ground Truth: 25
- Prediction: 25
- From: OCR

(g) When is dad’s birthday?

- Ground Truth: may 19
- Prediction: 19
- From: OCR

(h) What edition is this?

- Ground Truth: embossed second
- Prediction: Answer

(i) What brand of soda on the table?

- Ground Truth: dr pepper aaro
- Prediction: Answer

(j) What city are the giants from?

- Ground Truth: san francisco
- Prediction: new york
- From: Answer

(k) What number is the small hand on?

- Ground Truth: 4
- Prediction: 4
- From: Answer

(l) Is the text about physics?

- Ground Truth: yes
- Prediction: no
- From: Answer

(m) Are all the books with same title?

- Ground Truth: no
- Prediction: no
- From: Answer

(n) What letter is written in yellow on the red sign?

- Ground Truth: m m
- Prediction: Answer

(o) What is the brand of this smartphone?

- Ground Truth: samsung
- Prediction: Answer
Figure 1: TextVQA Examples and LoRRA’s predictions on them. We show multiple examples from TextVQA, ground truth answers, along with predictions from LoRRA, attention maps on OCR tokens and where LoRRA predicted the answer from (OCR tokens or pre-determined answer space. Green, red, and blue boxes correspond to correct, incorrect, and partially correct answers, respectively. On the right side of each image, we show attention bars which depict attention weights (0-1) for each of the OCR tokens.
**Figure 3: Text detection main task.** First stage of our tasks is used to identify and remove images without text.

**Figure 4: Question task.** In the second stage, we ask workers to ask a question about an image whose answer requires reading text in the image. We provide instructions and rules to ensure that we get high quality questions.
Figure 5: **Answer task.** In the third stage, we ask workers to answer a question about the image.