LaTr: Layout-Aware Transformer for Scene-Text VQA

Ali Furkan Biten[1]*† Ron Litman[2]* Yusheng Xie[2] Srikar Appalaraju[2] R. Manmatha[2]
[1]Computer Vision Center, UAB, Spain, [2]AWS AI Labs
abiten@cvc.uab.es {litmanr, yushx, srikara, manmatha}@amazon.com

Abstract

We propose a novel multimodal architecture for Scene Text Visual Question Answering (STVQA), named Layout-Aware Transformer (LaTr). The task of STVQA requires models to reason over different modalities. Thus, we first investigate the impact of each modality, and reveal the importance of the language module, especially when enriched with layout information. Accounting for this, we propose a single objective pre-training scheme that requires only text and spatial cues. We show that applying this pre-training scheme on scanned documents has certain advantages over using natural images, despite the domain gap. Scanned documents are easy to procure, text-dense and have a variety of layouts, helping the model learn various spatial cues (e.g. left-of, below etc.) by tying together language and layout information. Compared to existing approaches, our method performs vocabulary-free decoding and, as shown, generalizes well beyond the training vocabulary. We further demonstrate that LaTr improves robustness towards OCR errors, a common reason for failure cases in STVQA. In addition, by leveraging a vision transformer, we eliminate the need for an external object detector. LaTr outperforms state-of-the-art STVQA methods on multiple datasets. In particular, +7.6% on TextVQA, +10.8% on ST-VQA and +4.0% on OCR-VQA (all absolute accuracy numbers).

1. Introduction

Scene-Text VQA (STVQA) aims to answer questions by utilizing the scene text in the image. It requires reasoning over rich semantic information conveyed by various modalities – vision, language and scene text. Fig. 1 (a) depicts representative samples in STVQA, showcasing a model's desired abilities, including; (1) a-priori information and world knowledge such as knowing what a website looks like (left image); and (2) the capability to use language, layout, and visual information (middle and right images).

In this work, we introduce Layout-Aware Transformer

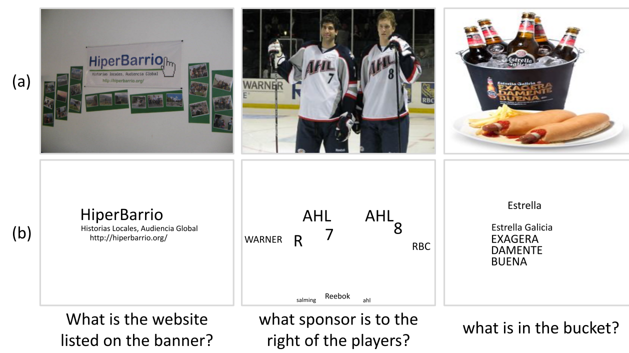

Figure 1. **The Role of Language and Layout in STVQA.** (a) Representative samples from TextVQA. (b) We visualize the information extracted by the OCR system, showing that some questions only require text features, some require both text and layout information and only some need beyond that. Accounting for this, we propose a *layout-aware* pre-training and architecture.

(LaTr), a multimodal encoder-decoder transformer based model for STVQA. We begin by exploring how far language and layout information can take us in STVQA . In Fig. 1 (b) we visualize the information extracted by the optical character recognition (OCR) system [1, 6, 14, 38], exhibiting three question categories: the first type can be answered with just the text tokens; the second type can be answered with text and layout information (*right* vs *left*); the third can only be answered by utilizing text, spatial and visual features all together. We quantitatively show that in the current datasets, most questions fall under the first two categories. To methodologically show this, we first evaluate a zero-shot language model on STVQA benchmarks, and then show that LaTr can already correctly answer over 50% of the questions with only text tokens. Next, we show the performance gain achieved by enriching the language modality with layout information via our propose *layout-aware* pre-training and architecture.

Recently, Yang et al. [74] demonstrated the advantages in pre-training STVQA models on natural images, proposing *text-aware* pre-training (TAP) scheme, which is designed to foster multi-modal collaboration. Acquiring large quantities of natural images with text is challenging and hard to scale, as most natural images do not contain scene text.

*Authors contribute equally.
†Work done during an internship at Amazon.

Even when they do, the amount of text is often sparse (previous statistics suggest a median of only 6 words per image [67, 74]). In addition, and more importantly, TAP did not account for the importance of aligning the layout information with the semantic representations when designing the pre-training objectives.

To counter these drawbacks, we propose *layout-aware* pre-training based on a single objective using only text and spatial cues as input. Our pre-training forces the model to learn a joint representation which accounts for the interactions between text and layout information, benefiting the down-stream task of STVQA. Despite the domain gap, we find that pre-training on documents has certain advantages over natural images. Scanned documents contain more text compared to natural images, therefore it is easier to scale the experiment and expose the model to more data. Words in documents are usually complete sentences, helping the model better learn semantics beyond a simple bag of words. Moreover, scanned documents provide varied layouts, leading to effective alignment between language and spatial features. Lastly, performing pre-training without visual features reduces computational complexity substantially.

Our model utilizes a vision transformer [13] for extracting visual features, thus replacing the extensive need for an external object detector [21, 25, 74]. Moreover, in practice, current STVQA models exploit a dataset-specific vocabulary with a pointer mechanism for decoding [17, 21, 24, 25, 71, 74–76], creating an over-reliance on the fixed vocabulary and leaving no room for fixing OCR errors. Our model performs vocabulary-free decoding, does well even on answers out-of-vocabulary, and even overcomes OCR errors in some cases. LaTr outperforms the state-of-the-art STVQA methods by large margins on multiple public benchmarks. To summarize, the key contributions of our work are:

1. We recognize the key role language and layout play in STVQA and propose a *layout-aware* pre-training and architecture to account for that.
2. We pinpoint a new symbiosis between documents and STVQA via pre-training. We show empirically that documents are beneficial for tying together language and layout information despite the huge domain gap.
3. We show that existing methods perform poorly on out-of-vocabulary answers. LaTr does not require a vocabulary, does well even on answers that are not in the training vocabulary, and can even overcome OCR errors.
4. We provide extensive experimentation and show the effectiveness of our method by advancing the state-of-the-art by +7.6% on TextVQA and +10.8% on ST-VQA and +4.0% in OCR-VQA dataset.

2. Related Work

Pre-training and Language Models. The low cost of obtaining language text combined with the success of pre-training, language models [12,40,52,53] has shown remarkable success in machine translation, natural language understanding, question answering and more. Recently, numerous studies [2, 10, 22, 28, 34–37, 42, 43, 61, 62, 77] showed the benefits of pre-training multi-modal architectures for vision and language tasks. Yang et al. [74] demonstrated, for the first time, the effectiveness of pre-training in scene text VQA by using masked language modeling and image-text matching as pretext tasks. In this paper, we show that tying together language and layout information via a simple *layout-aware* pre-training scheme is beneficial for scene text VQA. Moreover, we perform pre-training over scanned documents and discover that, despite the domain gap, documents can be leveraged for task of STVQA.

Vision-language tasks incorporating scene text. Recently, integrating reading into the vision and language tasks has become imperative, especially in VQA and captioning where the models were known to be illiterate [8,58]. Since the usage of text can be quite distinct in terms of the environment, several papers introduce new datasets for various contexts in which text appears; ST-VQA [9], TextVQA [58] in natural images; OCR-VQA [49] in book and movie covers; DocVQA [47] in scanned documents; InfoVQA [46] in info-graphics. Moreover, STE-VQA [70] is proposed for multi-lingual VQA and TextCaps [57] for captioning on natural images. There are several papers published on scene text VQA. LoRRa [58] extended Pythia [23] with a pointer network [68] to select either from a fixed vocabulary or from OCR tokens. M4C [21] also used pointer networks but instead used multi-modal transformers [66] to encode all modalities together. SA-M4C [25] build on top of M4C by providing supervision on self-attention weights. MM-GNN [16] builds separate graphs for different modalities by utilizing graph neural networks [29]. Instead of having separate graphs for each modality, SMA [15] introduces a single graph that encodes all modalities. [78] proposes to use an attention mechanism to fuse pairwise modalities.

LaTr enriches the language modality with layout information via pre-training to achieve state-of-the-art performance across multiple benchmarks. Our model is generative in nature and as such alleviates the problem of vocabulary reliance current methods suffer from. In addition, we will show that LaTr is more robust to OCR errors, one of the most common reasons for failure cases in STVQA [21, 74].

3. Method

In this section, we describe in detail our model architecture and our pre-training strategy, as seen in Fig. 2. LaTr consists of three main building blocks. First, a language model pre-trained on only text. Second, use of spatial embedding for OCR tokens bounding box in conjunction with further *layout-aware* pre-training on documents, as depicted in Fig. 2 (a). Finally, a ViT architecture [13] for obtaining

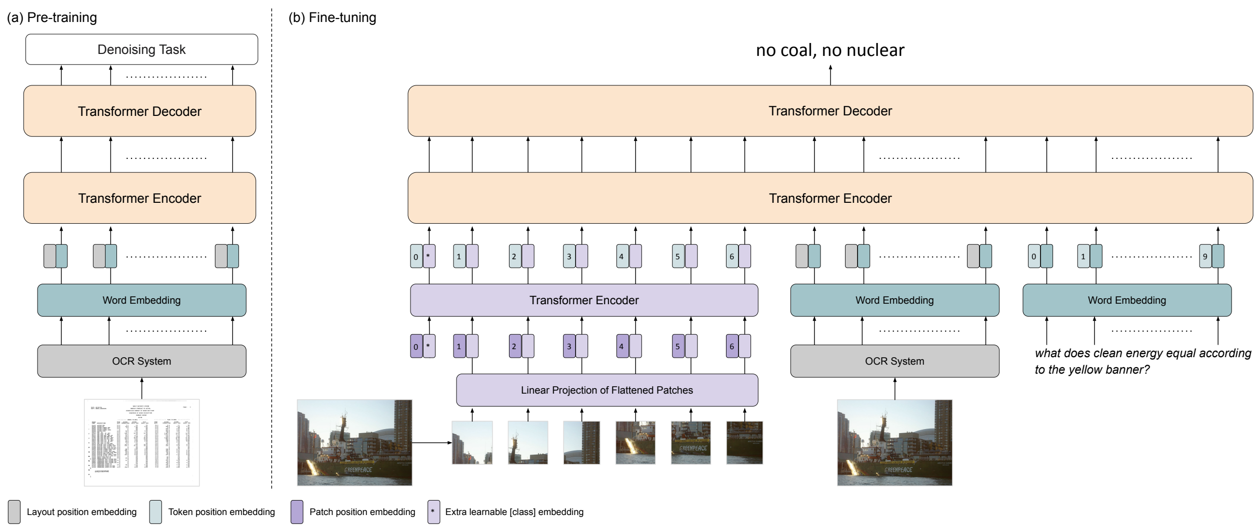

Figure 2. **An overview of LaTr.** (a) In pre-training, we only train the language modality with text and spatial cues to jointly model interactions between text and layout information. Pre-training is done on large amounts of documents. Documents are a text rich environment with a variety of layouts. (b) In fine-tuning, we add visual features from a ViT, thus eliminating the need for an external object detector.

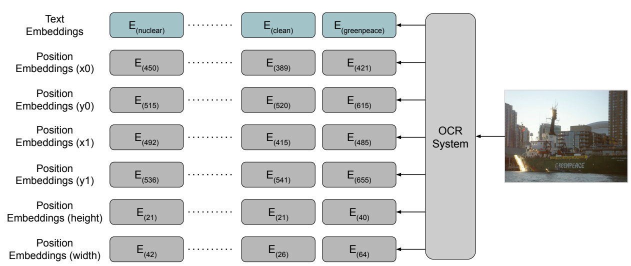

Figure 3. **Layout Position Embedding.** 2-D position embeddings representing the text layout in the image are leveraged to enrich the semantic representations.

visual features. We first explain each of the modules and then describe how all the modules come together as a whole.

The Language Model We base our LaTr architecture on the encoder-decoder transformer architecture of *Text-to-Text Transfer Transformer* (T5 [53]). Apart from minor modifications, T5's architecture is roughly equivalent to the original transformer proposed by [66], which makes it easy to extend in various ways. In addition, the vast amount of pre-training data used in the T5 pretraining makes it attractive for STVQA as model initialization. In particular, [53] used Common Crawl publicly-available web archive to obtain a subset of 750 GB cleaned English text data, which they term Colossal Clean Crawled Corpus (C4). Pre-training on C4 is done with a de-noising task, which is a variant of masked-language modeling (MLM [12]). We follow the implementation and use the weights from HuggingFace [63] [1].

[1] https://huggingface.co/transformers/model_doc/t5.html

2-D Spatial Embedding Recent document understanding literature [5, 72, 73] prove the value of layout information when working with Transformers. The key idea is to associate and couple the 2-D positional information of the text with the language representation, *i.e.* creating better alignment between the layout information and the semantic representation. Unlike words in a document, scene text in natural images may appear in arbitrary shapes and angles (e.g., as on a watch face). Therefore, we include the height and width of the text to indicate the reading order.

Formally, as seen in Fig. 3, given an OCR token O_i, the associated word bounding box may be defined by $(x_0^i, y_0^i, x_1^i, y_1^i, h^i, w^i)$, where (x_0^i, y_0^i) corresponds to the position of the upper left corner of the bounding box, (x_1^i, y_1^i) represents the position of the lower right corner, and (h^i, w^i) represents the height and width with respect to the reading order. To embed bounding box information, we use a lookup table commonly used for continuous encoding one-hot representations (e.g. nn.Embedding in PyTorch). Before we feed the word representation into the transformer encoder, we sum up all the representations together:

$$\begin{aligned}\mathcal{E}_i = E_O(O_i) + E_x(x_0^i) + E_y(y_0^i) + \\ E_x(x_1^i) + E_y(y_1^i) + E_w(w^i) + E_h(h^i)\end{aligned} \tag{1}$$

where $\mathcal{E}_i$ is the encoded representation for an OCR token O_i and E_O, E_x, E_y, E_w, E_h are the learnable look-up tables.

Layout-Aware Pre-Training As T5 was trained on just text data, we perform further pre-training to effectively align the layout information (in form of the 2-D spatial embedding) and the semantic representations. To the best of our knowledge, we are the first to propose pre-training on documents instead of natural images for the task of scene text VQA. The motivation for selecting documents is that they are a source of rich text environment in a variety of complex layouts. Inspired by [53], we perform a *layout-aware* de-noising pre-training task, which includes the 2-D spatial embedding, as seen in Fig. 2 (a). This enables the use of weak data with no answer annotations in the pre-training stage. Like the normal de-noising task, our *layout-aware* de-noising task masks a span of tokens and forces the model to predict the masked spans. Unlike the normal de-noising task, we also give the model access to the rough location of the masked tokens, which encourages the model to fully utilize the layout information when completing this task.

More formally, let $\mathcal{O} = \{O_1, O_2, ..., O_n\}$ be the set of all OCR tokens (strings) and $\mathcal{B} = \{B_1, B_2 ..., B_n\}$ be the corresponding bounding box information, where $B_j = (x_0^j, y_0^j, x_1^j, y_1^j, w^j, h^j)$. Now, let $\mathcal{M}_l = \{j, j+1, ..., j+k\}$ be the l^{th} mask span where j is the starting index to mask such that $\max(M_l) < \min(M_{l+1})$. Then, $\{O_j, ..., O_{j+k}\}$ and $\{B_j, ..., B_{j+k}\}$ are replaced by $\tilde{O}_i$ (a special indexed mask token) and $\tilde{B}_i$ (the span's minimal containing bounding box) in the following manner:

$$\begin{aligned} \tilde{O}_i &= \text{<extra\_id\_}l\text{>}, \text{where } l \in \{0, ..., k-1\} \\ \tilde{B}_i &= (\min(\{x_0^i\}), \min(\{y_0^i\}), \\ &\quad \max(\{x_1^i\}), \max(\{y_1^i\})) \\ &\quad \text{where } j \leq i \leq j+k \end{aligned} \tag{2}$$

where the height and width of the masked tokens' bounding box are calculate with the coordinates of $\tilde{B}_i$.

Essentially, we have replaced a span of words tokens $\{O_j, ..., O_{j+k}\}$ and their corresponding bounding boxes $\{B_j, ..., B_{j+k}\}$ with a special token $\tilde{O}_i$ and a corresponding "loose" bounding box. In other words, when we mask the span of words, we select the minimum of the top-left coordinates and the maximum of the bottom-right ones. The reasons are twofold. First, we do not want our model to know precise token boxes because that would reveal how many tokens are masked. Second, we choose not to mask the bounding boxes completely because then the model does not know where the text should appear in the document and cannot use the correct spatial context effectively. So, we prevent the model from taking shortcuts, but at the same time give it enough information to learn. The masked token $\tilde{O}_i$ and its bounding box $\tilde{B}_i$ are then embedded using Eq. (1) like any other regular token. We use cross-entropy loss to predict all the masked tokens' original text.

Visual Features Most previous methods utilized an external pre-trained object detector [21, 74] for extracting objects labels, visual object features and visual OCR features. In this work, we diverge from the literature and leverage a Vision Transformer (ViT) [13]. The ViT is an image classification network which is pre-trained and fine-tuned on ImageNet [11]. We utilize ViT in our architecture only in the fine-tuning stage, and we freeze all the layers except the last fully connected projection layer we add. Formally, an image I having the dimension of $H \times W \times C$ is reshaped into 2D patches of size $N \times (p^2 \cdot C)$, where (H, W) is the height and width, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the final number of patches. As depicted in Fig. 2 (b), we utilize a linear projection layer to map the flattened patches to D dimensional space and feed them to the ViT. We pass the full ViT output (containing $[class]$ token) sequence to a trainable linear projection layer and then feed it to the transformer encoder. Position embeddings are added to the patch embeddings to retain positional information. We denote the final visual output as $\mathcal{V} = \{V_0, ..., V_N\}$.

LaTr So far, we explained the building blocks of our method, now we describe how we put it all together, as depicted in Fig. 2 (b). After pre-training the language modality of the model with layout information, we input all three modalities, namely; image, OCR information and question to the transformer encoder. Let $\mathcal{V} = \{V_0, ..., V_N\}$ be a set of visual patch features such that V_0 is the $[class]$ embedding, $\mathcal{Q} = \{W_1, ..., W_m\}$ be the question tokenized into W_i and $\mathcal{O} = \{O_1, O_2, ..., O_n\}$ be the OCR tokens. We embed the OCR tokens and questions using Eq. (1) to obtain encoded OCR tokens $\mathcal{E}$ and encoded question features $\mathcal{E}^q$. For the 2-D spatial embedding of each W_i, we use fixed values ($x_0 = y_0 = 0; x_1 = y_1 = 1000$). Finally, we concatenate all the inputs $[\mathcal{V}; \mathcal{E}; \mathcal{E}^q]$ to feed to the multimodal transformer encoder-decoder architecture. Cross entropy loss is used to fine-tune our model.

4. Experiments

In this section, we experimentally examine our method, comparing its performance with state-of-the-art methods. We consider the standard benchmarks of TextVQA [58], ST-VQA[2] [9] and OCR-VQA [49]. For pre-training we consider the same datasets used in [7, 74] with the addition of the Industrial Document Library (IDL)[3]. The IDL is a collection of industry documents hosted by UCSF. It hosts millions of documents publicly disclosed from various industries like tobacco, drug, food etc. The data from the website amounts to about 13M documents, translating to about 64M pages of various document images. We further extracted OCR for each document using Textract OCR[4]. Implementation details and further information on all datasets

[2] We use ST-VQA for denoting the dataset proposed in [9], and STVQA for denoting the general task of scene text VQA.

[3] https://www.industrydocuments.ucsf.edu/

[4] https://aws.amazon.com/textract/

Method	OCR System	Pre-Training Data	Extra Finetune	No. of Param.	Val Acc.	Test Acc.
M4C [21]	Rosetta-en	✗	✗	200M	39.40	39.01
SMA [15]	Rosetta-en	✗	✗	-	40.05	40.66
CRN [39]	Rosetta-en	✗	✗	-	40.39	40.96
LaAP-Net [20]	Rosetta-en	✗	✗	-	40.68	40.54
TAP [74]	Rosetta-en	TextVQA	✗	200M	44.06	-
LaTr -Small	Rosetta-en	✗	✗	149M	41.84	-
LaTr -Base	Rosetta-en	✗	✗	311M	44.06	-
LaTr -Base	Rosetta-en	IDL	✗	311M	48.38	-
SA-M4C [25]	Google-OCR	✗	ST-VQA	200M	45.4	44.6
SMA [15]	SBD-Trans OCR	✗	ST-VQA	-	-	45.51
M4C [21, 74]	Microsoft-OCR	✗	ST-VQA	200M	45.22	-
TAP [74]	Microsoft-OCR	TextVQA	✗	200M	49.91	49.71
TAP [74]	Microsoft-OCR	TextVQA, ST-VQA	ST-VQA	200M	50.57	50.71
LOGOS [44]	Microsoft-OCR	✗	ST-VQA	-	51.53	51.08
TAP [74]	Microsoft-OCR	TextVQA, ST-VQA, TextCaps, OCR-CC	ST-VQA	200M	54.71	53.97
M4C [21]	Amazon-OCR	✗	✗	200M	47.84	-
LaTr-Base	Amazon-OCR	✗	✗	311M	52.29	-
LaTr-Base	Amazon-OCR	IDL	✗	311M	58.03	58.86
LaTr‡-Base	Amazon-OCR	IDL	ST-VQA	311M	59.53	59.55
LaTr-Large	Amazon-OCR	IDL	✗	856M	59.76	59.24
LaTr‡-Large	Amazon-OCR	IDL	ST-VQA	856M	**61.05**	**61.60**

Table 1. **Results on the TextVQA dataset [58]**. As commonly done, the top part of the table presents results in the constrained setting that only uses TextVQA for training and Rosetta for OCR detection, while the bottom part is the unconstrained settings. LaTr advances the state-of-the-art performance, specifically by +6.43% and +7.63% on validation and test, respectively.

can be found in Appendix A and B, respectively. We note that throughout the rest of the paper, ‡ refers to the models fine-tuned with both TextVQA and ST-VQA, at the same time. "-Small", "-Base" and "-Large" model sizes refer to architectures that have 6+6, 12+12 and 24+24 layers in encoder and decoder, respectively. For convenience, we refer to LaTr-Base as LaTr.

TextVQA Results Similar to previous work [74], we define two evaluation settings. The former is the constrained setting that only uses TextVQA for training and Rosetta for OCR detection. The latter is the unconstrained setting, in which we present our best performance with the state-of-the-art. The first part of Tab. 1 reports the accuracy under the constrained setting. As can be appreciated, LaTr-Small outperforms M4C (**+2.44%**), with fewer parameters. Increasing the model capacity to LaTr results in a performance gain of **+2.22%** (additional discussion on the model capacity can be found in Appendix D). In addition, LaTr achieves the same performance as TAP [74] without any pre-training, demonstrating the effectiveness of our model. Furthermore, when LaTr is pre-trained on IDL, performance increase from 44.06% to 48.38% (**+4.32%**) using the Rosetta OCR. This clearly shows the effectiveness of *layout-aware* pre-training on scanned documents to the task of scene text VQA, even in the constrained setting.

In the bottom part of Tab. 1 we modify the OCR system to a more recent one than Rosetta and gradually add additional training datasets (unconstrained settings). In this work, we experiment with Amazon Text-in-Image (Amazon-OCR)[5] [65]. As seen, when using Amazon-OCR our method outperforms the M4C baseline, improving performance from 47.84% to 52.29% (**+4.45%**). Furthermore, when enabling pre-training, LaTr outperforms the previous art [74] by large margins from 54.71% to 58.03% (**+3.32%**) on validation and from 53.97% to 58.86% (**+4.89%**) on the test. We note that for [74] there is a -0.74% decrease between validation and test while for LaTr we observe an increase of +0.83%, demonstrating better generalization. Another critical point is that LaTr can benefit more when ST-VQA dataset is added as an extra fine-tune data. We believe this point to be critical since we do not have to train separate models for TextVQA and ST-VQA but rather one model that can get the best performance on both dataset. Finally, increasing our model capacity to LaTr-Large further boosts performance to 61.6% (**+7.6%** from [74]).

ST-VQA Results Tab. 2 presents the accuracy on ST-VQA [9] in the unconstrained setting. LaTr uses the Amazon-OCR and is pre-trained on IDL and fine-tuned on the training set of ST-VQA. LaTr‡ is also fine-tuned with TextVQA. The behaviour observed in TextVQA is consistent with ST-VQA dataset, LaTr‡-Base and LaTr‡-Large outperforming the previous art [74] by **+8.26%** and **+10.81%**, respectively. Moreover, we show a similar trend on OCR-VQA [49] dataset where the discussion and the numbers can be found in Appendix E.

Qualitative Analysis In Fig. 4 we depict five different question categories which are representative of the capa-

[5] https://docs.aws.amazon.com/rekognition/index.html

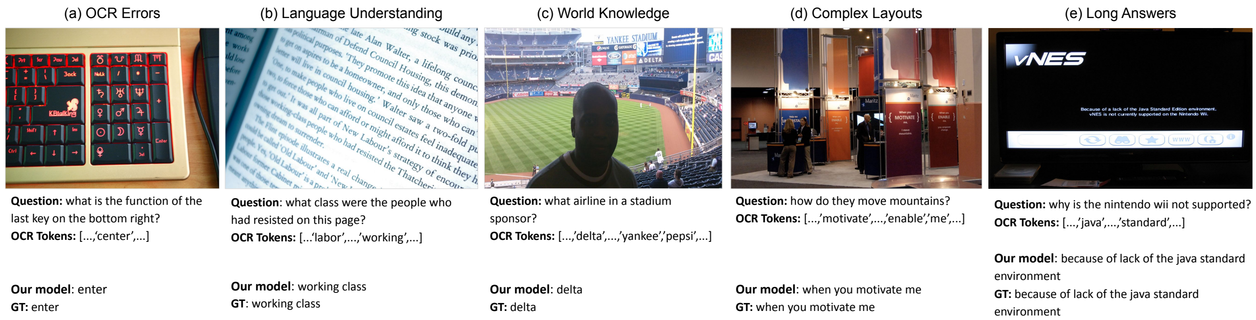

Figure 4. **Why is STVQA hard?** Current state-of-the-art methods struggle to acquire various abilities which are needed for scene text VQA. We depict five representative abilities; fixing OCR errors, language understating, world knowledge, understating complex layouts and the ability to produces long answers. Our model is able to correctly answer each one of the these examples. We refer the reader to more qualitative results and comparisons to previous art in Appendix F.

Method	Val Acc.	Val ANLS	Test ANLS
M4C [21]	38.05	0.472	0.462
SA-M4C [25]	42.23	0.512	0.504
SMA [15]	-	-	0.466
CRN [39]	-	-	0.483
LaAP-Net [20]	39.74	0.497	0.485
LOGOS [44]	48.63	0.581	0.579
TAP [74]	50.83	0.598	0.597
LaTr-Base	58.41	0.675	0.668
LaTr‡-Base	59.09	0.683	0.684
LaTr‡-Large	**61.64**	**0.702**	**0.696**

Table 2. **Results on the ST-VQA Dataset [9].** Our model advances the state-of-the-art performance by +10.81%.

bilities STVQA models need. We start with the ability to correct OCR errors (Fig. 4 (a)). Most state-of-the-art OCR systems for scene text [6,14,38,50] operate on a word-level, and thus are unable to utilize image-level context. Current STVQA methods depend on a pointer network for decoding, which means they are bounded by the performance of the OCR system at hand. Contrary to that, LaTr leverages image-level context and jointly with its generative nature, is able to correct OCR errors. Next, scene text VQA models are required to have the ability to understand language together with world knowledge (Fig. 4 (b)(c)). Both requirements are met in LaTr thanks to its extensive pre-training.

As seen in Fig. 4 (d), answering questions often requires reasoning over the relative spatial positions of the text in the image. Over the years several methods aimed at developing spatially aware models were proposed [25, 44]. However, most of those methods are complex, not easy to implement and eventually led to minimal performance improvements. LaTr is pre-trained on documents with layout information, which leads to a spatially aware model without any complex architectural changes. The last category we analyze is long answers (Fig. 4 (e)). In practice, the existing pointer network decoding mechanism is also limited in ability to produce long answers. Furthermore, when pre-training is done

Model	OCR	Acc.
T5-Base	Rosetta-en	16.05
T5-Base	Amazon-OCR	21.93
T5-Base	GT text	25.45

Table 3. **Zero Shot Performance of T5 Language Model on TextVQA.** In this setting, T5-Base is pre-trained on C4 and fine-tuned on SQuAD [54], a reading comprehension dataset. Showing that a "blind" pre-trained language model can get up to 25.45%.

on natural images as in [74], the model hardly encounters long sentences. LaTr does not rely on a pointer network and is pre-trained on documents, in which text appears in a variety of lengths.

We provide further qualitative analysis and comparisons to previous work [21] in Appendix F. In addition, we display failure cases of our method on the TextVQA dataset. The failure cases are mostly composed of OCR errors, compositionality of spatial reasoning and visual attributes.

5. Ablation Studies

In this section, we provide insightful experiments which we deem crucial for the STVQA task and its future development. We start off by showing the significance of language understanding in STVQA . Then, we show the effectiveness of language and layout information and discuss the biases existing in STVQA benchmarks. Next, we study the effect of pre-training as a function of dataset size and type. Finally, we showcase our model's robustness towards vocabulary and OCR errors. All the numbers are obtained by using the TextVQA validation set.

Zero-shot Language Models on TextVQA To quantify the importance of language understanding in STVQA , we devise a novel zero-shot setting where we use the T5 language model pre-trained on C4 and only fine-tuned on SQuAD [54], a reading comprehension dataset. Tab. 3 presents the performance of this setting while varying the OCR system. Interestingly, even without any visual fea-

Model	2-D	Pre-training	OCR	Visual	Acc.
LaTr	✗	✗	✗	✗	11.18
	✗	✗	✗	*V*	11.74
	✗	✗	*random*	✗	41.77
	✗	✗	✓	✗	50.37
	✓	✗	✓	✗	51.22
	✓	✗	✓	*V*	52.29
	✓	✓	✓	✗	57.38
	✓	✓	✓	*F*	**58.11**
	✓	✓	✓	*V*	58.03
LaTr‡	✓	✓	✓	✗	58.92
	✓	✓	✓	*F*	58.45
	✓	✓	✓	*V*	**59.53**

Table 4. **LaTr Ablation Studies on TextVQA.** We ablate LaTr - Base by varying the building blocks of our method, including pre-training, input types and fine-tuning data. V refers to ViT and F refers to FRCNN as visual backbone, *random* means OCR tokens are provided but presented in a random reading order.

Model	Pre-training Data	Acc.
LaTr-Base	✗	50.37
	TextVQA	51.81
	TextVQA,ST-VQA,TextCaps,OCR-CC	54.22
	IDL - 1M	55.12
	IDL - 11M	56.28
	IDL - 64M	58.03
	IDL-64M,TextVQA,ST-VQA,TextCaps,OCR-CC	**58.51**
LaTr‡-Base	IDL - 64M	**59.53**
	IDL-64M,TextVQA,ST-VQA,TextCaps,OCR-CC	59.06

Table 5. **The Effect of Pre-training.** Ablation studies on pre-training as a function of different datasets type and size.

tures or fine-tuning, T5 reaches a performance of 16.05% and 21.93% with Rosetta and Amazon-OCR, respectively. More importantly, a zero-shot "blind" model with the perfect OCR (ground truth OCR annotation [59]) can get to as high as 25%, experimentally demonstrating the need for language understanding in STVQA . However, one needs to be careful attributing the entirety of the performance to language understanding since deep models are known to exploit dataset biases [64]. Thus, we investigate if there are any biases in the data and if it is possible to categorize them.

Dataset Bias or Task Definition? To get a better sense of the biases in TextVQA, we start by training a model where only questions are given as input. As can be seen in Tab. 4, our model is able to achieve 11.18% in a task that requires reading and reasoning about the text without *the text*. Next, we study the effect of the OCR system by dividing the information provided by it into text token transcription, reading order and 2-D positional information. Reading order is the order where OCR tokens are extracted from left to right and top to bottom with respect to line boxes or text blocks. Reading order is so intertwined with OCR systems that it is not thought of as a detached feature.

As shown in Tab. 4, adding OCR tokens without any reading order gives us 41.77% and a fixed reading order already gets us to 50.37%, showing the importance of reading order for given OCR tokens. The gain becomes marginal when adding the 2-D positional and visual information without pre-training, +0.85% and +1.09%, respectively. However, when performing *layout-aware* pre-training on documents, obtaining alignment between the layout information and the semantic representations, LaTr's performance increases significantly by +7.01% to 57.38%. In other words, we can already achieve SOTA on a *Visual* Question Answering task without any visual features (other than using the images for OCR extraction). Finally, adding visual features still *marginally* increases performance by around +0.7%. Recently, [69] showed a similar phenomenon using the M4C [21] architecture, where visual information only slightly contributed to the performance, validating that this is not specific to our technique.

Regarding the comparison of the different visual backbones, we train our model with visual features extracted either from FRCNN [4] or ViT [13]. We note that the performance difference is very marginal when only TextVQA is used in fine-tuning. However, when TextVQA and ST-VQA are used together, the model with FRCNN features perform worse than the model without any visual features while ViT increases performance by +0.61%, demonstrating that ViT features can scale better with more data.

At this point, we would like to take a step back and discuss STVQA as a task. As we see it, our analysis can be interpreted from two viewpoints. The first viewpoint is how STVQA is defined as a task. In particular, is the STVQA task defined such that all (or a majority of) questions should require reasoning over all modalities (including visual features)? Regardless of the answer, we present a second viewpoint, a dataset bias. To better explore the bias perspective, in Appendix G we visualize question-image pairs sorted by the information required to answer them. Clearly, generating questions from the final category (*i.e.* questions which require reasoning over all modalities) is not an easy task. Furthermore, we quantitatively showed that at-least 60% of the questions do not fall under the final category, allowing the model to extensively exploit language priors and make educative guesses. Both viewpoints lead us to wonder are visual features even needed for STVQA? Or better yet, is vision an artifact in STVQA task? We believe that visual features are of importance for the task of STVQA, however current benchmarks do not reflect it, making it harder to evaluate how much V matters in STVQA.

Model	All 5000	InVoc. 3731	OutVoc. 1269	Gap
M4C [21]	47.84	51.07	38.37	**12.7**
LaTr-Base	59.53	59.93	58.35	1.58

Table 6. **Vocabulary Reliance.** Accuracy gap between answers with words in and out of vocabulary used by [21, 25, 74]. InVoc. and OutVoc. stand for in and outside the vocabulary, respectively.

The Effect of Large-Scale Pre-Training Tab. 5 demonstrates the benefits of pre-training while varying the datasets type and scale. First, we explore the effect of pre-training on natural images with visual features (as done in [74]) using our architecture. In particular, we add the image-text matching objective and leverage the same datasets (which we term TAP-datasets) as in [74]. Pre-training only on TextVQA (Tab. 5), provides only +1.5% improvement for us compared to [74] reporting +5%. The same behaviour of diminished gain is also observed with TAP-datasets.

Next, we compare IDL and TAP-datasets in pre-training. Even pre-training on 1M documents, LaTr's performance increases by almost +5%, which is more than the combination of all TAP-datasets. This is inspiring for two reasons, one of which is 1M documents are less than two thirds the size of TAP-datasets [74]. Secondly, our model is pre-trained with a simple de-noising objective and no visual features, making the pre-training significantly faster (around 23 times) compared to TAP [74] which is pre-trained with visual features, scene text features and multiple losses. We also argue that IDL is a better bed for *layout-aware* pre-training since it provides varied layouts to better align with language. Finally, we discuss the effect of increasing the size of IDL. Adding an order of magnitude more data only result in +1% or +2% increase. We emphasize that 64M documents hardly seems the saturation point for LaTr, *i.e.* more pre-training data can still improve the performance, especially when also increasing the model capacity.

Vocabulary Reliance and Robustness Towards OCR Errors Current state-of-the-art methods predict the answer through an amalgamation of a pointer mechanism and a dataset-specific 5K most frequent vocabulary. The usage of a vocabulary is limiting in a real-world scenario and may result in high performance on in-vocabulary answers but lead to poor performance on out-of-vocabulary ones, in other words, lack of generalization. This is clearly observed in Tab. 6 where M4C [21] exhibits a heavy reliance on the fixed vocabulary as the gap between categories is **-12.7%**. Contrary to that, LaTr is not limited to any hand-crafted dataset-specific vocabulary. Its gap between in and out of the training vocabulary is only **-1.58%**.

Finally, we experimentally display that our model is more robust to OCR errors compared to M4C architecture.

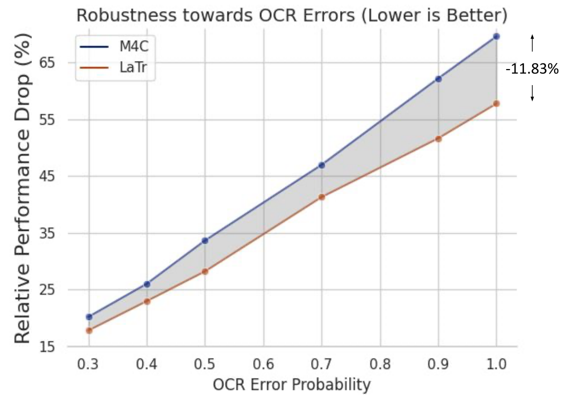

Figure 5. **Robustness towards OCR Errors.** OCR Error Probability refers to the percentage of OCR tokens that we replace a single character by a random one, simulating OCR engine errors. LaTr's relative robustness is higher compared to [21] and increases with the probability of OCR errors.

To validate our claim we introduce a new setting where we replace a single character for certain amount of OCR tokens. Whether to replace a character in each word is decided according to the threshold from a Bernoulli distribution, called OCR Error Probability in Fig. 5. To simulate real-world OCR errors, we utilized the publicly available nlp-augmenter from [45]. LaTr is more robust than [21] and in fact the lead increases as more OCR errors are added.

6. Conclusion

We convey a couple of important take-home messages for the STVQA community. Firstly, *language and layout are essential*. Language indirectly is utilized for questions that need world/prior knowledge or simply for language understanding. Layout information allows the model to reason over spatial relations. In our work, we methodologically demonstrated their importance to STVQA. Secondly, we propose a *layout-aware* pre-training and show a new symbiosis between scanned documents and scene text, where the layout information of scanned documents promotes a better understanding of scene text information. This is exciting news since scanned documents are more abundantly available than natural images that contains scene text. Text in documents appears in a variety of complex layouts, making our model spatially aware without any complex architectural changes. Last but not least, we replace the extensive need of FRCNN for feature extraction. We exhibit that using a ViT as a feature extractor can scale better than FRCNN, *i.e.* leading to better performance. However, perhaps more crucially, we diagnose a condition in which STVQA models (ours included) make use of the visual features *marginally*. This begs the question whether this is because of the dataset bias, and we as a community need to make V matter again in VQA.

References

[1] Aviad Aberdam, Ron Litman, Shahar Tsiper, Oron Anschel, Ron Slossberg, Shai Mazor, R Manmatha, and Pietro Perona. Sequence-to-sequence contrastive learning for text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15302–15312, 2021. 1, 12

[2] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. In *EMNLP*, pages 2131–2140, 2019. 2

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016. 12

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. 7

[5] Srikar Appalaraju, Bhavan Jasani, Bhargava Urala Kota, Yusheng Xie, and R Manmatha. Docformer: End-to-end transformer for document understanding. *IEEE/CVF International Conference on Computer Vision, ICCV*, 2021. 3

[6] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4715–4723, 2019. 1, 6

[7] Ali Furkan Biten, Rubèn Tito, Lluis Gomez, Ernest Valveny, and Dimosthenis Karatzas. Ocr-idl: Ocr annotations for industry document library dataset. *arXiv preprint arXiv:2202.12985*, 2022. 4

[8] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1563–1570. IEEE, 2019. 2

[9] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In *ICCV*, pages 4291–4301, 2019. 2, 4, 5, 6, 12

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 4, 12

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 3

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 4, 7, 12

[14] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7098–7107, 2021. 1, 6

[15] Chenyu Gao, Qi Zhu, Peng Wang, Hui Li, Yuliang Liu, Anton van den Hengel, and Qi Wu. Structured multimodal attentions for textvqa. *arXiv preprint arXiv:2006.00753*, 2020. 2, 5, 6

[16] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. Multi-modal graph neural network for joint reasoning on vision and scene text. In *CVPR*, pages 12746–12756, 2020. 2

[17] Lluís Gómez, Ali Furkan Biten, Rubèn Tito, Andrés Mafla, Marçal Rusiñol, Ernest Valveny, and Dimosthenis Karatzas. Multimodal grid features and cell pointers for scene text visual question answering. *Pattern Recognition Letters*, 150:242–249, 2021. 2

[18] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 12

[19] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018. 12

[20] Wei Han, Hantao Huang, and Tao Han. Finding the evidence: Localization-aware answer prediction for text visual question answering. *arXiv preprint arXiv:2010.02582*, 2020. 5, 6

[21] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 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, 2020. 2, 4, 5, 6, 7, 8, 12, 13, 14, 15

[22] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849*, 2020. 2

[23] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018. 2

[24] Zan-Xia Jin, Heran Wu, Chun Yang, Fang Zhou, Jingyan Qin, Lei Xiao, and Xu-Cheng Yin. Ruart: A novel text-centered solution for text-based visual question answering. *IEEE Transactions on Multimedia*, 2021. 2

[25] Yash Kant, Dhruv Batra, Peter Anderson, Alex Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. Spatially aware multimodal transformers for textvqa. In *ECCV*, 2020. 2, 5, 6, 8

[26] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 12

[27] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *2013 12th International Conference on Document Analysis and Recognition*, pages 1484–1493. IEEE, 2013. 12

[28] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021. 2

[29] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2

[30] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 12

[31] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*, 2018. 12

[32] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*, 2018. 12

[33] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4. *International Journal of Computer Vision*, 128(7):1956–1981, 2020. 12

[34] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020. 2

[35] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *arXiv preprint arXiv:2107.07651*, 2021. 2

[36] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[37] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020. 2

[38] Ron Litman, Oron Anschel, Shahar Tsiper, Roee Litman, Shai Mazor, and R Manmatha. Scatter: selective context attentional scene text recognizer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11962–11972, 2020. 1, 6, 12

[39] Fen Liu, Guanghui Xu, Qi Wu, Qing Du, Wei Jia, and Mingkui Tan. Cascade reasoning network for text-based visual question answering. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4060–4069, 2020. 5, 6

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 2

[41] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 12

[42] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019. 2

[43] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *CVPR*, pages 10437–10446, 2020. 2

[44] Xiaopeng Lu, Zhen Fan, Yansen Wang, Jean Oh, and Carolyn P Rose. Localize, group, and select: Boosting text-vqa by scene text modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2631–2639, 2021. 5, 6

[45] Edward Ma. Nlp augmentation. https://github.com/makcedward/nlpaug, 2019. 8

[46] Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. *arXiv preprint arXiv:2104.12756*, 2021. 2

[47] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209, 2021. 2

[48] A. Mishra, K. Alahari, and C. V. Jawahar. Image retrieval using textual cues. In *ICCV*, 2013. 12

[49] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 947–952. IEEE, 2019. 2, 4, 5, 13

[50] Oren Nuriel, Sharon Fogel, and Ron Litman. Textadain: Fine-grained adain for robust text recognition. *arXiv preprint arXiv:2105.03906*, 2021. 6, 12

[51] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 12

[52] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. 2

[53] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. 2, 3, 4

[54] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016. 6

[55] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 12

[56] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 13

[57] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. *arXiv preprint arXiv:2003.12462*, 2020. 2, 12

[58] 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*, pages 8317–8326, 2019. 2, 4, 5, 12, 14, 15

[59] Amanpreet Singh, Guan Pang, Mandy Toh, Jing Huang, Wojciech Galuba, and Tal Hassner. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8802–8812, 2021. 7

[60] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. From strings to things: Knowledge-enabled vqa model that can read and reason. In *CVPR*, pages 4602–4612, 2019. 12

[61] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019. 2

[62] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, pages 5103–5114, 2019. 2

[63] Wolf Thomas, Debut Lysandre, Sanh Victor, Chaumond Julien, Delangue Clement, Moi Anthony, Cistac Pierric, Rault Tim, Louf Rémi, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019. 3, 12

[64] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011. 7

[65] William Ughetta. The old bailey, us reports, and ocr: Benchmarking aws, azure, and gcp on 360,000 page images. 2021. 5, 13

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2, 3

[67] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 2, 12

[68] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. *arXiv preprint arXiv:1506.03134*, 2015. 2

[69] Qingqing Wang, Liqiang Xiao, Yue Lu, Yaohui Jin, and Hao He. Towards reasoning ability in scene text visual question answering. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2281–2289, 2021. 7

[70] Xinyu Wang, Yuliang Liu, Chunhua Shen, Chun Chet Ng, Canjie Luo, Lianwen Jin, Chee Seng Chan, Anton van den Hengel, and Liangwei Wang. On the general value of evidence, and bilingual scene-text visual question answering. In *CVPR*, pages 10126–10135, 2020. 2

[71] Jiajia Wu, Jun Du, Fengren Wang, Chen Yang, Xinzhe Jiang, Jinshui Hu, Bing Yin, Jianshu Zhang, and Lirong Dai. A multimodal attention fusion network with a dynamic vocabulary for textvqa. *Pattern Recognition*, 122:108214, 2022. 2

[72] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm: Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1192–1200, 2020. 3

[73] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*, 2020. 3

[74] Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8751–8761, 2021. 1, 2, 4, 5, 6, 8, 12, 13

[75] Gangyan Zeng, Yuan Zhang, Yu Zhou, and Xiaomeng Yang. Beyond ocr+ vqa: Involving ocr into the flow for robust and accurate textvqa. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 376–385, 2021. 2

[76] Xuanyu Zhang and Qing Yang. Position-augmented transformers with entity-aligned mesh for textvqa. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2519–2528, 2021. 2

[77] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *AAAI*, pages 13041–13049, 2020. 2

[78] Qi Zhu, Chenyu Gao, Peng Wang, and Qi Wu. Simple is not easy: A simple strong baseline for textvqa and textcaps. *arXiv preprint arXiv:2012.05153*, 2020. 2