TAP: Text-Aware Pre-training for Text-VQA and Text-Caption

Zhengyuan Yang[1*] Yijuan Lu[2] Jianfeng Wang [2] Xi Yin[2]
Dinei Florencio[2] Lijuan Wang[2] Cha Zhang[2] Lei Zhang[2] Jiebo Luo[1]
[1]University of Rochester [2]Microsoft Corporation
{yijlu,jianfw,dinei,lijuanw,chazhang,leizhang}@microsoft.com
{zyang39, jluo}@cs.rochester.edu yinxi.whu@gmail.com

Abstract

In this paper, we propose Text-Aware Pre-training (TAP) for Text-VQA and Text-Caption tasks. These two tasks aim at reading and understanding scene text in images for question answering and image caption generation, respectively. In contrast to conventional vision-language pre-training that fails to capture scene text and its relationship with the visual and text modalities, TAP explicitly incorporates scene text (generated from OCR engines) during pre-training. With three pre-training tasks, including masked language modeling (MLM), image-text (contrastive) matching (ITM), and relative (spatial) position prediction (RPP), pre-training with scene text effectively helps the model learn a better aligned representation among the three modalities: text word, visual object, and scene text. Due to this aligned representation learning, even pre-trained on the same downstream task dataset, TAP already boosts the absolute accuracy on the TextVQA dataset by $+5.4\%$, compared with a non-TAP baseline. To further improve the performance, we build a large-scale scene text-related image-text dataset based on the Conceptual Caption dataset, named OCR-CC, which contains 1.4 million images with scene text. Pre-trained on this OCR-CC dataset, our approach outperforms the state of the art by large margins on multiple tasks, i.e., *$+8.3\%$ accuracy on TextVQA, $+8.6\%$ accuracy on ST-VQA, and $+10.2$ CIDEr score on TextCaps.*

1. Introduction

The *Vision-language tasks incorporating scene text* [7, 18, 49, 46], *e.g.*, Text-VQA [49, 8, 40, 56] and Text-Caption [46], pose new challenges to vision-language models of reading and understanding scene text in image context. Extended from Visual Question Answering (VQA) [6], Text-VQA aims to answer questions by understanding the scene text in the image-question context. Text-Caption seeks to generate an image caption [54, 4] that describes both the visual and scene text information in the image, as shown in Figure 1 (a). These tasks have many potential applications, including robotics [5], document understanding [40], assisting visually-impaired people [7, 18], *etc*.

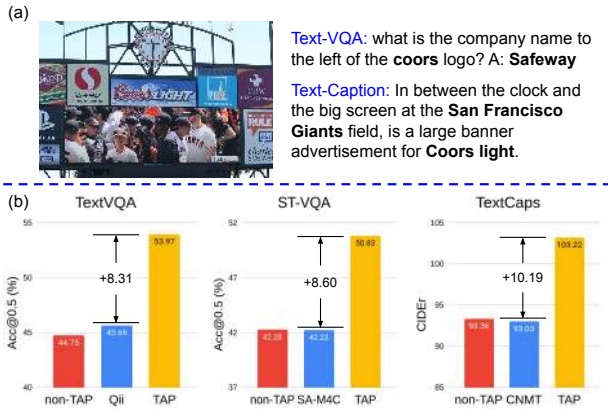

Figure 1. **(a)** Text-VQA and Text-Caption tasks aim at reading and understanding scene text in images for question answering and image caption generation, respectively. We highlight the scene text-related words in bold. **(b)** By explicitly incorporating scene text in pre-training, Text-Aware Pre-training (TAP) significantly outperforms both the non-TAP baseline and previous state of the art on multiple tasks (bars shown in red and blue colors, respectively).

A typical Text-VQA/Text-Caption framework consists of 1) a feature encoder for each single modality (text word, visual object, and scene text), 2) a multi-modal fusion module, and 3) a decoding module for prediction generation. Previous studies [49, 17, 16, 20, 25, 46, 55] improve the model's performance by designing stronger network architectures. Among them, LoRRA [49] added an OCR attention branch for scene text encoding to a VQA model [24]. M4C [20, 46] proposed a transformer-based multi-modal fusion module [52] and a multi-step multi-choice decoding module. Despite the effective network design, most previous models are optimized with a sole ob-

*This work was done while Z.Yang was an intern at Microsoft.

jective directly towards the correct answer/caption. Such a single answer/caption loss tries to predict each word in the ground-truth but is less effective in learning a joint representation among text word, visual object, and scene text. Without a good joint representation, directly optimizing for question-answering/image-captioning could be challenging. Inspired by the success of Vision-Language Pre-training (VLP) [37, 32, 12, 51, 34, 23, 11] in image-text joint representation learning, we leverage the effective Text-VQA/Text-Caption network designs and explore to further improve Text-VQA/Text-Caption by pre-training.

Vision-Language Pre-training (VLP) shows its effectiveness in learning task-agnostic joint representations of image and text. The main idea is to first pre-train the model with pre-training tasks on image-caption datasets [45, 29, 54, 41, 43], and then fine-tune the model for a specific vision-language task [6, 60, 28, 54]. However, conventional VLP methods are designed intuitively for vision-language tasks and do not include scene text in pre-training. Therefore, previous methods fail to capture the scene text modality and its relationship with the visual and text modalities, and are thus less effective in Text-VQA/Text-Caption.

In this study, we propose *Text-Aware Pre-training* (TAP), which incorporates the scene text modality in pre-training to learn a joint representation of text word, visual object, and scene text. In TAP, we design text-aware pre-training tasks to better fuse scene text (including both scene text words and their visual regions detected by OCR) with the *text words* and *visual objects*. For the former, we refine the pre-training tasks in VLP [37, 34] to support the extra scene text input. We find it particularly important to include the detected scene text words as extra language inputs. The extra inputs anchor the scene text and language modalities and make the aligned representation learning easier. For the latter, previous studies [25, 55] show that the spatial relationships between scene text and object regions are important, *e.g.*, the relationship "left" in Figure 1 (a). Therefore, we propose a "relative (spatial) position prediction" task that learns regions' spatial relationships by predicting their relative spatial positions in pre-training.

The extra scene text modality, together with the specially designed pre-training tasks, effectively helps the model learn a better aligned representation among the three modalities: text word, visual object, and scene text. This aligned representation learning, even pre-trained and fine-tuned on the same downstream task dataset, leads to significant improvement over the non-TAP baseline and helps the TAP model achieve the new state of the art.

To further unleash the power of TAP, we clean and generate a large-scale scene text-related image-caption dataset for pre-training. In general image-caption datasets [45, 29, 54, 41, 43], many image-text pairs contain either no scene text-related visual regions or no scene text-related language referring, and are thus less helpful to Text-VQA/Text-Caption. On the visual side, we run an OCR detector to filter out images with no scene text. On the language side, we include the detected OCR text tokens as the additional caption input to obtain scene text-related language descriptions. In the end, we build a large-scale dataset named OCR-CC with around 1.4 million scene text-related image-text pairs based on the Conceptual Captioning dataset [45]. By using this large-scale dataset for pre-training, we observe further improvement on the Text-VQA and Text-Caption tasks.

We experiment with the TAP approach on the M4C network architecture [20] and benchmark it on the TextVQA [49], ST-VQA [8], and TextCaps [46] datasets. With the identical network architecture and training data, TAP improves the accuracy on the TextVQA dataset [49] from 44.50% to 49.91%, compared with a non-TAP baseline. Our final model ranks **No.1**[1] on multiple Text-VQA/Text-Caption challenges, and outperforms previous methods by large margins: TextVQA [49] ($+8.3\%$ in absolute accuracy), ST-VQA [8] ($+8.6\%$ in absolute accuracy), and TextCaps [46] ($+10.2$ in CIDEr score).

Our main contributions are:

- To the best of our knowledge, we are the first to explore pre-training for Text-VQA and Text-Caption.
- By explicitly incorporating scene text with three specially designed pre-training tasks, Text-Aware Pre-training (TAP) effectively learns a better aligned representation that leads to significant performance improvement on Text-VQA/Text-Caption.
- We build a large-scale dataset named OCR-CC with around 1.4 million scene text-related image-text pairs. TAP with OCR-CC leads to the new state of the art on multiple tasks: TextVQA [49] ($+8.3\%$ in absolute accuracy), ST-VQA [8] ($+8.6\%$ in absolute accuracy), and TextCaps [46] ($+10.2$ in CIDEr score). We will release the dataset and the models.

2. Related Work

Vision-language tasks incorporating scene text. Text-VQA [49, 8, 40, 56] and Text-Caption [46] aim at reading and understanding scene text in images for question answering and image caption generation. Various datasets [49, 8, 40] are built for the Text-VQA task, *e.g.*, the TextVQA dataset [49], the ST-VQA dataset [8], *etc*. TextCaps [46] is a dataset recently proposed for the Text-Caption task.

Recent studies [49, 17, 16, 20, 25, 55, 36, 19] proposed various network architectures to improve the Text-VQA/Text-Caption performance. Among them, LoRRA [49] approached Text-VQA by extending a VQA model Pythia [24] with an OCR attention branch. The answer vocabulary is a combination of a static vocabulary

[1] According to the official leader-boards (Nov. 2020)

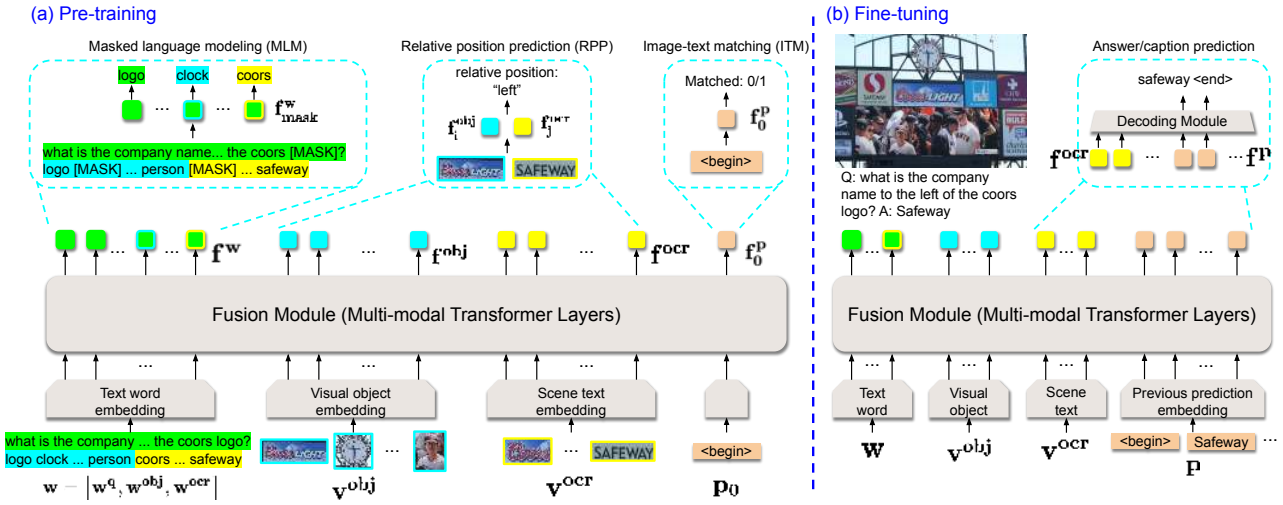

Figure 2. An overview of Text-Aware Pre-training (TAP). **(a)** In pre-training, the framework takes text words $\mathbf{w}$, visual objects $\mathbf{v}^{\mathbf{obj}}$, scene text $\mathbf{v}^{\mathbf{ocr}}$, and a special `begin` token $\mathbf{p_0}$ as inputs, and improves the aligned representation learning by performing pre-training tasks (MLM, ITM, RPP) on fused feature $\mathbf{f}$. **(b)** In fine-tuning, we train the same model to step-by-step generate the answer/caption prediction, conditioned on $\mathbf{w}$, $\mathbf{v}^{\mathbf{obj}}$, $\mathbf{v}^{\mathbf{ocr}}$, and the previous word predictions $\mathbf{p} = \{\mathbf{p_t}\}_{t=0}^{T-1}$ at decoding step T. Text word, visual object, and scene text-related tokens are highlighted by the green, cyan, and yellow colors, respectively.

and detected OCR tokens. Multi-modal Multi-Copy Mesh (M4C) [20] boosted the Text-VQA performance by proposing a transformer-based multi-modal fusion module [52] and a multi-step multi-choice decoding module that supports multi-step answer decoding. M4C's variants M4C-Captioner [46] set a strong baseline on TextCaps [46] with the question text inputs removed. SA-M4C [25] further improved M4C by encoding the spatial relationships among visual regions as the attention masks in the multi-modal transformer. Similar explorations [55] on the spatial relationships are studied in the Text-Caption task.

Despite the effective network design, all previous studies directly optimize towards the sole objective for the Text-VQA/Text-Caption task. We contend that such a single answer/caption loss could be ineffective in aligned representation learning and thus limits the Text-VQA/Text-Caption performance. In this study, we leverage the effective network designs and explore to further improve Text-VQA/Text-Caption by pre-training.

Vision-Language Pre-training (VLP). VLP [37, 32, 1, 31, 51, 50, 61, 12, 38, 34, 23] shows its effectiveness in learning task-agnostic vision-language joint representations. Most studies [37, 51, 12] focused on vision-language understanding tasks, *e.g.*, image-text retrieval [60], visual question answering [6], visual grounding [28, 58], *etc*. Recent studies [61, 34, 21] unified the pre-training framework to cover generation tasks, *e.g.*, image captioning [54, 4].

However, conventional VLP methods do not capture scene text during pre-training and are therefore less effective for Text-VQA/Text-Caption. The proposed Text-aware Pre-training (TAP) explicitly incorporates scene text to learn a better aligned representation among the three modalities: text word, visual object, and scene text.

3. Text-Aware Pre-training (TAP)

TAP explicitly incorporates scene text in pre-training to improve Text-VQA/Text-Caption. We first pre-train the model with the scene text-aware pre-training tasks and then fine-tune it for a specific downstream task.

In this section, we first introduce the design of scene text-aware pre-training tasks. We then present the data corpus used for TAP and our proposed OCR-CC dataset. We postpone the model details to Section 4.2.

3.1. Text-aware pre-training tasks

Figure 2 overviews TAP in pre-training and fine-tuning. In pre-training, the input to the fusion module are embeddings of K text words $\mathbf{w}$, M object regions $\mathbf{v}^{\mathbf{obj}}$, N scene text regions $\mathbf{v}^{\mathbf{ocr}}$, and a special `begin` token $\mathbf{p_0}$. In the text word embedding, each word in the extended text input $\mathbf{w} = \left[\mathbf{w}^{\mathbf{q}}, \mathbf{w}^{\mathbf{obj}}, \mathbf{w}^{\mathbf{ocr}}\right]$ is encoded as a feature vector, where $\mathbf{w}^{\mathbf{q}}, \mathbf{w}^{\mathbf{obj}}, \mathbf{w}^{\mathbf{ocr}}$ are the question text, detected object labels, and detected scene text words. In the object and scene text embedding, object and scene text regions are detected and encoded by object detectors and OCR engines.

Taking the fused feature $\mathbf{f} = \left[\mathbf{f}^{\mathbf{w}}, \mathbf{f}^{\mathbf{obj}}, \mathbf{f}^{\mathbf{ocr}}, \mathbf{f}^{\mathbf{p}}\right]$ as inputs, TAP improves multi-modal fusion by performing text-aware pre-training tasks. The proposed pre-training tasks consist of two parts, focusing on fusing scene text $\mathbf{v}^{\mathbf{ocr}}$ with text words $\mathbf{w}$ and visual objects $\mathbf{v}^{\mathbf{obj}}$, respectively.

Scene-text language pre-training tasks. To better fuse the scene text $\mathbf{v}^{\mathbf{ocr}}$ with the text words $\mathbf{w}$, we design two scene-text language pre-training tasks based on the masked language modeling (MLM) and image-text (contrastive) matching (ITM) tasks in VLP [15, 37, 12]. For MLM on the extended text input $\mathbf{w} = \left[\mathbf{w}^{\mathbf{q}}, \mathbf{w}^{\mathbf{obj}}, \mathbf{w}^{\mathbf{ocr}}\right]$, we randomly

mask each text token in $\mathbf{w}$ with a probability of 15%. The masked words $\mathbf{w_{mask}}$ are replaced with a special MASK token 80% of the time, a random word 10%, and remains unchanged 10%. The MLM task takes the fused feature at the masked position $\mathbf{f^w_{mask}}$ as the input, and aims to recover the masked word $\mathbf{w_{mask}}$ with two fully-connected layers. For ITM, $\mathbf{w}$ is polluted 50% of the time by replacing text subsequence $\mathbf{w^q}$, $\mathbf{w^{obj}}$, or $\mathbf{w^{ocr}}$ with a randomly-selected one from another image. The polluted text words $\mathbf{w}$ are thus not paired with the visual regions $\mathbf{v^{obj}}$ and $\mathbf{v^{ocr}}$. The ITM task takes the sequence feature $\mathbf{f^p_0}$ as the input and aims to predict if the sequence has been polluted or not.

We find that the extra scene text word input $\mathbf{w^{ocr}}$ is critical for learning the scene-text language aligned representation. As a comparison to the extended text input $\mathbf{w}$, pre-training with the original MLM and ITM [15, 37] on question text $\mathbf{w^q}$ leads to limited improvement over the non-pre-training baseline. The failure is due to the limited number of scene text-related words in the language input $\mathbf{w^q}$. In this case, since many randomly masked words $\mathbf{w^q_{mask}}$ and polluted sequences are not relevant to scene text, scene text regions $\mathbf{v^{ocr}}$ are less important for solving the pre-training tasks (MLM, ITM) and are thus often overlooked. $\mathbf{w^{ocr}}$ in the extended text input $\mathbf{w}$ generates extra scene text referring in the language modality and thus makes TAP effective.

Scene-text visual pre-training tasks. Understanding the spatial relationships between the visual object $\mathbf{v^{obj}}$ and scene text $\mathbf{v^{ocr}}$ benefits Text-VQA/Text-Caption [25, 55]. The extra feature input of bounding box coordinates helps the spatial relationship learning [20, 17, 16], but hasn't fully solved the problem. Recent studies [25, 55] hard code the coordinate features as the regions' relationships in feature fusion and obtain further improvement. In this study, we explore spatial relationship learning by pre-training.

Specifically, we design a scene-text visual pre-training task in TAP. The main idea is to predict the relative spatial position between two randomly sampled visual regions. Therefore, we refer to the task as "relative (spatial) position prediction" (RPP). The input to the pre-training task is a randomly sampled visual object feature $\mathbf{f_i^{obj}}$ and scene text feature $\mathbf{f_j^{ocr}}$, where $i \in \{1, \cdots, M\}$ and $j \in \{1, \cdots, N\}$. The objective is to predict the relative spatial position between the two sampled regions $\mathbf{v_i^{obj}}$ and $\mathbf{v_j^{ocr}}$. We start with a single relationship of whether "scene text region $\mathbf{v_j^{ocr}}$ is *on* object $\mathbf{v_i^{obj}}$," and thus model RPP as a binary classification problem. We then extend the task to a 12-class relative position prediction problem with the classes defined by Yao *et al*. [59], including on, cover, overlap, eight-way relative orientation, and unrelated.

3.2. Pre-training corpus

TAP works well even without extra pre-training data. We first experiment with "TAP without extra data," where we only use the downstream Text-VQA/Text-Caption dataset for pre-training, *i.e*., the training set of the TextVQA [49], ST-VQA [8], or TextCaps [46] datasets. These datasets [49, 8, 46] all contain less than 30K images and 150K image-text pairs. We detail the pre-training and fine-tuning pipeline for each downstream task in Section 4.2.

We then experiment with "TAP with large-scale data." We build a large-scale scene text-related image-caption dataset named *OCR-CC* based on the Conceptual Caption (CC) dataset [45], and use the dataset for pre-training. Among the image-caption datasets [45, 29, 54, 41, 43], only the CC dataset contains a reasonable portion of images with meaningful scene text regions. Therefore, we run the Microsoft Azure OCR system[2] on all images in the CC dataset and filter out the images with no scene text, watermarks only, and tiny scene text regions only. In the end, we obtain 1.367 million image-caption pairs with a mean and median of 11.4 and 6 scene text detected per image. As a reference, the mean and median are 23.1 and 12 in the TextVQA dataset [20], and 8.03 and 6 in the ST-VQA dataset [8]. We adopt the same region feature extraction method used in the TextVQA dataset [49] to provide object and scene text region embedding. By including scene text words $\mathbf{w^{ocr}}$ as additional text inputs, OCR-CC provides scene text-related image-caption pairs for TAP. We keep the caption text from CC in OCR-CC and use it as the question text $\mathbf{w^q}$ in pre-training. We show the details of dataset collection, scene text number distribution, and additional qualitative examples of OCR-CC in the supplementary material.

4. Experiments

We benchmark TAP for both the Text-VQA task on the TextVQA [49] and ST-VQA [8] datasets, and the Text-Caption task on the TextCaps dataset [46]. We use our proposed OCR-CC dataset for large-scale pre-training.

4.1. Datasets

TextVQA. The TextVQA dataset [49] contains 28,408 images from the Open Images dataset [30]. We follow the same training/validation/test split used in the previous work [49] in our experiments. The methods are evaluated by the soft-voting accuracy of 10 answers.

ST-VQA. The ST-VQA dataset [8] contains 21,892 images from multiple sources including ICDAR 2013 [27], ICDAR 2015 [26], ImageNet [13], VizWiz [18], IIIT STR [39], Visual Genome [29], and COCO-Text [54]. The methods are evaluated by both accuracy and Average Normalized Levenshtein Similarity (ANLS) [8].

TextCaps. The TextCaps dataset [46] augments the 28,408 images in TextVQA [49] with 145,329 captions. The cap-

[2]Public Microsoft OCR API: https://docs.microsoft.com/en-us/azure/cognitive-services/computer-vision/concept-recognizing-text

tions are evaluated by the caption metrics (BLEU [42], METEOR [14], ROUGE_L [35], SPICE [3], and CIDEr [53]).

OCR-CC. Our OCR-CC dataset contains 1.367 million scene text-related image-caption pairs from the Conceptual Captioning (CC) dataset [45]. More details of OCR-CC are in the supplementary material.

4.2. Experiment settings

Network architecture. We conduct experiments based on the M4C network architecture [20, 48, 47]. We extend text input $\mathbf{w_q}$ with object label $\mathbf{w^{obj}}$ and scene text word $\mathbf{w^{ocr}}$. We keep all remaining settings the same as in the original M4C [20], including the feature embedding, network architecture, training parameters, and layer initialization.

M4C's text encoder is a three-layer trainable transformer [52] initialized from the first three layers of $\text{BERT}_{\text{BASE}}$ [15]. A pre-trained Faster R-CNN [44] detects objects and represents the detected region with its visual and coordinate features. The final layer (fc7) of the detector is fine-tuned. An offline OCR detector [10] detects scene text regions and represents the region with its visual, coordinates, FastText [9], and Pyramidal Histogram of Characters (PHOC) [2] features. The fusion module in M4C is a four-layer multi-modal transformer that has the same hyper-parameters as $\text{BERT}_{\text{BASE}}$. The fusion module is initialized from scratch. A multi-step decoding module then takes fused features $\mathbf{f^{ocr}}, \mathbf{f^{p}}$ as inputs, and word-by-word predicts the final answer. The predicted answer word at each decoding step T is selected either from a fixed frequent word vocabulary or from the dynamic OCR tokens. The word classification loss is applied to each decoding step.

Adapting to Text-VQA. By taking the fused feature $\mathbf{f}$ as input, we pre-train the feature encoder and fusion module with the pre-training tasks (MLM, ITM, RPP). MLM is only computed on the sequences that have not been polluted by ITM. The pre-trained model with the highest pre-training task accuracy is used to initialize the feature encoder and fusion module. In fine-tuning, the model step-by-step predicts the answer with an extra decoding module, and is trained with the answer classification loss in each step.

Adapting to Text-Caption. We keep the framework architecture the same for Text-Caption as for Text-VQA, except increasing the maximum answer decoding length from 12 words [20] to 30 words [46]. $\mathbf{w^q}$ is left blank in both pre-training and fine-tuning. The input text sequence $\mathbf{w}$ consists of $\mathbf{w^{ocr}}$, $\mathbf{w^{obj}}$, and the blank $\mathbf{w^q}$. During fine-tuning, the framework is trained with the same multi-step word classification loss as used in Text-VQA.

Compared methods. We compare TAP with other state of the art [49, 17, 20, 25, 16, 36, 19, 55] and systematically study the following baselines and variants of our method.

- **TAP (Ours).** We first experiment with "TAP without extra pre-training data." We use the same downstream task dataset for both pre-training and fine-tuning, and follow the same training parameters as used in M4C. For the Text-VQA task, we pre-train the model for 24K iterations with the pre-training tasks (MLM, ITM, RPP) and then fine-tune it with the answer loss for another 24K iterations. The numbers of pre-training and fine-tuning iterations are both 12K for the Text-Caption task following M4C-Captioner [46].
- **M4C†.** "M4C†" is the non-TAP baseline. Based on M4C, we include the detected object labels $\mathbf{w^{obj}}$ and scene text tokens $\mathbf{w^{ocr}}$ as the additional text input following "TAP." We train the model for 48K iterations with the answer loss to match TAP's total iteration number. Compared with "TAP," the only difference is that "M4C†" trains the first 24K iterations with the answer loss, instead of the pre-training tasks.
- **TAP†† (Ours).** "TAP††" reports our best performance achieved with extra pre-training data (TextVQA, ST-VQA, TextCaps, OCR-CC) and other minor modifications. We pre-train "TAP††" for 480K iterations. Section 4.4 details the benefits of each extra data source.

4.3. Text-VQA/Text-Caption results

TextVQA. Table 1 reports the accuracy on the TextVQA dataset [49]. The **top part** of the table shows the results in the constrained setting that only uses TextVQA for training and Rosetta [10] for OCR detection. The **bottom** compares our best performance with the state of the art [49, 17, 20, 25, 16, 36, 19, 55] in the unconstrained setting.

We list the adopted OCR detector in the "OCR system" column. LoRRA [49] and M4C [20] adopted the Rosetta OCR system [10]. SA-M4C [25] and SMA [16] experiment with both Rosetta and other OCR systems (Google-OCR, SBD-Trans OCR). In this study, we experiment with Rosetta and the Microsoft Azure OCR system (Microsoft-OCR). We use Microsoft-OCR to detect the single OCR words appeared in the image, *i.e.*, each detected scene text region contains only a single word. The "Extra data" column shows the used training data other than the TextVQA dataset. Previous methods [20, 25, 16] adopt the ST-VQA dataset for joint training. Other than ST-VQA, TAP enables the use of weak data with no ground-truth answer in pre-training, *e.g.*, TextCaps and OCR-CC. "TAP††" reports the final performance with all extra datasets.

Three major observations can be made from Table 1: **1)** "TAP" significantly outperforms the non-TAP baseline "M4C†" with the identical training data and network architecture, in both the constrained setting (top part of Table 1) and the unconstrained setting (bottom part). In the constrained setting, TAP improves the non-TAP baseline accuracy from 39.55% to 44.06%. In the unconstrained setting, "TAP" with Microsoft-OCR obtain 5.4% and 5.3% absolute accuracy improvement over the corresponding non-TAP

Table 1. Text-VQA results on the TextVQA dataset [49]. The top part reports results in the constrained setting that only uses TextVQA for training and Rosetta for OCR detection. The bottom part compares our best performance with other state-of-the-art methods in the unconstrained setting. The methods "M4C†," "TAP," "TAP††" are detailed in Section 4.2.

Method	OCR System	Extra Data	Val Acc.	Test Acc.
LoRRA [49]	Rosetta-ml	✗	26.56	27.63
MM-GNN [17]	Rosetta-ml	✗	31.44	31.10
M4C [20]	Rosetta-en	✗	39.40	39.01
SMA [16]	Rosetta-en	✗	40.05	40.66
CRN [36]	Rosetta-en	✗	40.39	40.96
LaAP-Net [19]	Rosetta-en	✗	40.68	40.54
M4C† [20]	Rosetta-en	✗	39.55	-
TAP (Ours)	Rosetta-en	✗	44.06	-
M4C [20]	Rosetta-en	ST-VQA	40.55	40.46
LaAP-Net [19]	Rosetta-en	ST-VQA	41.02	40.54
SA-M4C [25]	Google-OCR	ST-VQA	45.4	44.6
SMA [16]	SBD-Trans OCR	ST-VQA	-	45.51
M4C† [20]	Microsoft-OCR	✗	44.50	44.75
M4C† [20]	Microsoft-OCR	ST-VQA	45.22	-
TAP (Ours)	Microsoft-OCR	✗	49.91	49.71
TAP (Ours)	Microsoft-OCR	ST-VQA	50.57	50.71
TAP†† (Ours)	Microsoft-OCR	ST-VQA, TextCaps, OCR-CC	**54.71**	**53.97**

Table 2. Text-VQA results on the ST-VQA dataset [8].

Method	Val Acc.	Val ANLS	Test ANLS
SAN+STR [8]	-	-	0.135
M4C [20]	38.05	0.472	0.462
SA-M4C [25]	42.23	0.512	0.504
SMA [16]	-	-	0.466
CRN [36]	-	-	0.483
LaAP-Net [19]	39.74	0.497	0.485
M4C† [20]	42.28	0.517	0.517
TAP (Ours)	45.29	0.551	0.543
TAP†† (Ours)	**50.83**	**0.598**	**0.597**

Table 3. Text-Caption CIDEr scores on the TextCaps dataset [46]. The full result table can be found in the supplementary material.

Method	Val CIDEr	Test CIDEr
BUTD [4]	41.9	33.8
AoANet [22]	42.7	34.6
M4C [46]	89.6	81.0
MMA-SR [55]	98.0	88.0
CNMT [57]	101.7	93.0
M4C† [46]	99.89	93.36
TAP (Ours)	105.05	99.49
TAP†† (Ours)	**109.16**	**103.22**

baselines "M4C†" and "M4C† +STVQA," respectively. The improvement achieved with the same network and training data validates the effectiveness of our pre-training approach for Text-VQA/Text-Caption. **2)** "TAP" outperforms the previous state of the art [49, 17, 20, 16, 36, 19] by large margins, even without large-scale pre-training. **3)** Large-scale pre-training with the OCR-CC dataset further improves the accuracy. "TAP††" adopts OCR-CC in pre-training and improves the accuracy from 49.91% to 54.71%. The improvement shows that TAP benefits from extra training data, and indicates the effectiveness of our proposed OCR-CC.

ST-VQA. Table 2 shows the Text-VQA accuracy on the ST-VQA dataset [8] in the unconstrained setting. "TAP" uses the Microsoft-OCR and is pre-trained and fine-tuned on the training set of ST-VQA. "TAP††" uses TextVQA, ST-VQA, TextCaps, and OCR-CC in pre-training. Similar conclusions as in Table 1 can be drawn from Table 2. First, "TAP" outperforms the state of the art [20, 25, 16, 36, 19] by large margins, and significantly improves the non-TAP baseline "M4C†." Second, large-scale pre-training further improves the accuracy by $+5.5\%$ as shown in bottom two rows.

TextCaps. Table 3 shows the CIDEr score on the TextCaps dataset [46]. We report only the CIDEr score in the table and present the full table with other metrics in the supplementary material. We draw similar observations that with the same training data, "TAP" improves the CIDEr score of "M4C†" from 99.89 to 105.05. Large-scale pre-training "TAP††" further improves the CIDEr score to 109.16.

4.4. Ablation studies

Pre-training tasks. We experiment with different pre-training tasks (MLM, ITM, RPP) as well as their variants. We conduct ablation studies on TextVQA with Microsoft-OCR and no extra data. We examine the effectiveness of scene-text language pre-training (MLM, ITM) and scene-text visual pre-training (RPP). We verify the importance of the extra scene-text token input $\mathbf{w}^{\mathbf{ocr}}$ in MLM and ITM.

As shown in Table 4, the scene-text language pre-training in row (d) and scene-text visual pre-training in row (e) improve the non-TAP baseline (row (b)) from 44.50% to 49.01% and 46.42%, respectively. "TAP" performs all pre-training tasks and further improves the accuracy to 49.91%.

The extra scene text token input $\mathbf{w}^{\mathbf{ocr}}$ is essential for TAP. Rows $(a$-$d)$ in Table 4 show that neither extra $\mathbf{w}^{\mathbf{ocr}}$ inputs ($c.f.$ rows (a, b)) nor pre-training ($c.f.$ rows (b, c)) alone lead to an improvement from the Non-TAP base-

Table 4. Ablation studies on different pre-training tasks (MLM, ITM, RPP), and the variant of excluding the extra scene-text token input $\mathbf{w}^{\mathbf{ocr}}$ in MLM and ITM. We highlight "TAP" by underline.

	+MLM,ITM	+RPP	Val Acc.
(a) Non-TAP w/o $\mathbf{w}^{\mathbf{ocr}}$	-	-	44.48
(b) Non-TAP	-	-	44.50
(c) + MLM,ITM w/o $\mathbf{w}^{\mathbf{ocr}}$	✓	-	44.63
(d) + MLM,ITM	✓	-	49.01
(e) + RPP	-	✓	46.42
(f) TAP	✓	✓	<u>49.91</u>

Table 5. Ablation studies on pre-training with extra data. We use the listed data only in pre-training and then fine-tune the model with the TextVQA dataset only. $(3, 4)$ and $(0, 12)$ indicate the layer numbers of the text and multi-modal transformers, respectively. We highlight "TAP" and "TAP††" by underline and bold.

	TextVQA	ST-VQA	TextCaps	OCR-CC	Val Acc. $(3,4)$	Val Acc. $(0,12)$
(a)	✓	-	-	-	<u>49.91</u>	48.78
(b)	✓	✓	-	-	50.57	49.64
(c)	✓	✓	✓	-	51.86	50.13
(d)	-	-	-	✓	52.10	54.03
(e)	✓	✓	✓	✓	52.90	**54.71**

line (row (b)). In contrast, TAP with the extra $\mathbf{w}^{\mathbf{ocr}}$ input (row (d)) boosts the accuracy to 49.01%. The bottom rows (e, f) show the effectiveness of RPP. RPP with a single spatial relationship "on" improves the accuracy from 44.50% to 46.42% (*c.f.* rows (b, e)). Combining RPP with MLM and ITM improves the accuracy from 49.01% to 49.91% (*c.f.* rows (d, f)). Extending spatial relationship classes to 12 [59] leads to an improvement from 49.91% to 50.17%.

Pre-training with extra data Table 5 breaks down the benefits of adopting different sources of extra data. We conduct experiments on the TextVQA dataset with Microsoft-OCR. TAP enables the use of weak data with no answer annotations in the pre-training stage such like TextCaps and OCR-CC, in addition to the Text-VQA datasets. Compared with "TAP" with no extra data, pre-training with ST-VQA and TextCaps improves the accuracy from 49.91% to 50.57% and 51.86% (*c.f.*, rows (a, b), rows (b, c)). The large-scale pre-training with OCR-CC (row (d)) achieves the accuracy of 52.10%. Including all data during pre-training (row (e)) further improves the accuracy to 52.90%.

Furthermore, we find that the extra data benefits the use of large models. The original architecture consists of a 3-layer text-only transformer and a 4-layer multi-modal transformer. We experiment with a 12-layer multi-modal transformer with the same structure as $\text{BERT}_{\text{BASE}}$ [15]. We initialize the model from $\text{BERT}_{\text{BASE}}$ and remove the separate text transformer. We represent the two architectures as $(3, 4)$ and $(0, 12)$ in Table 5, where the numbers indicate the text and multi-modal transformer layer numbers. With extra transformer layers, the accuracy without extra data drops from 49.91% to 48.78% (row (a)), while the accuracy with extra data increases from 52.90% to 54.71% (row (e)).

Table 6. The coreference scores with and without TAP. Numbers represent the attention score between two semantically corresponded tokens, averaged across all such token pairs in TextVQA. Higher coreference scores imply a better aligned representation.

Coref Type	W/O TAP	With TAP
Text Word $\rightarrow$ Scene Text	0.0477	**0.3514**
Scene Text $\rightarrow$ Text Word	0.0473	**0.5206**
Visual Object $\rightarrow$ Scene Text	0.0045	**0.0130**
Scene Text $\rightarrow$ Visual Object	0.0337	**0.0680**

4.5. How does TAP help?

In this section, we analyze how TAP helps Text-VQA/Text-Caption. We empirically show that with TAP, certain attention heads in the multi-modal transformer ground the scene text $\mathbf{v}^{\mathbf{ocr}}$ to the semantically corresponded text word $\mathbf{w}$ or visual object $\mathbf{v}^{\mathbf{obj}}$. By learning such latent alignments, TAP improves the aligned representation learning and thus helps Text-VQA/Text-Caption.

Recent VLP analyses [11, 33] show that VLP [51, 12, 32] learns the latent alignments between the semantically corresponded region-word or region-region pairs. Specifically, certain attention heads in the transformer generate higher attention scores between such corresponded pairs. The attention scores between corresponded pairs are also referred to as coreference scores [11]. Similarly, we analyze the change in the coreference score of scene text-related pairs to better understand TAP.

There exist $(4 \text{ layers} \times 12 \text{ heads}) = 48$ attention scores between any two positions in our multi-modal transformer. Following VALUE [11], we define the coreference score as the maximum attention score among all 48 heads between two semantically corresponded positions. A text word and a scene text region are corresponded if they refer to the same scene text token, *e.g.*, the text word and scene text region "coors" in Figure 3. We collect all corresponded pairs between the extended text input $\mathbf{w}$ and scene text regions $\mathbf{v}^{\mathbf{ocr}}$ in the TextVQA dataset, and report the averaged score over all pairs. A scene text $\mathbf{v}^{\mathbf{ocr}}$ and a visual object $\mathbf{v}^{\mathbf{obj}}$ are corresponded if they share the spatial relationship "on."

As shown in Table 6, we analyze TAP by comparing the change in the coreference score before and after TAP, *i.e.*, "M4C†" and "TAP." The first two rows show that TAP improves the scene-text language coreference scores by seven times. The bottom two rows show that TAP increases the scene-text visual coreference scores by two times. These increases validate that TAP successfully learns the latent alignment and thus improves joint representation learning.

Furthermore, Figure 3 visualizes the attention score between a text word and all visual regions. Qualitatively, we observe a higher coreference score with TAP (bottom row) than the non-TAP baseline (top row). For example, in Figure 3 (a), TAP grounds the text word "must" and "survive" to the corresponded scene text regions.

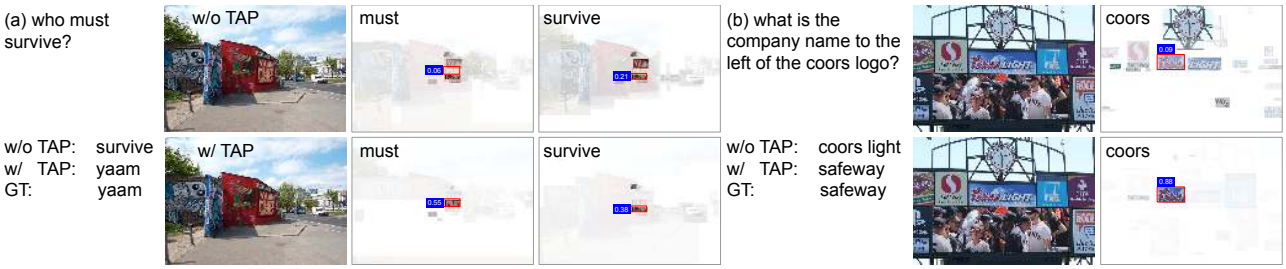

Figure 3. Visualization of region attention scores with respect to each word in the question text $\mathbf{w}$, extracted from the multi-modal fusion transformers with (bottom row) and without (top row) TAP. The score by a region indicates its attention strength. TAP generates interpretable attentions on scene text-related question words like "must" and "survive."

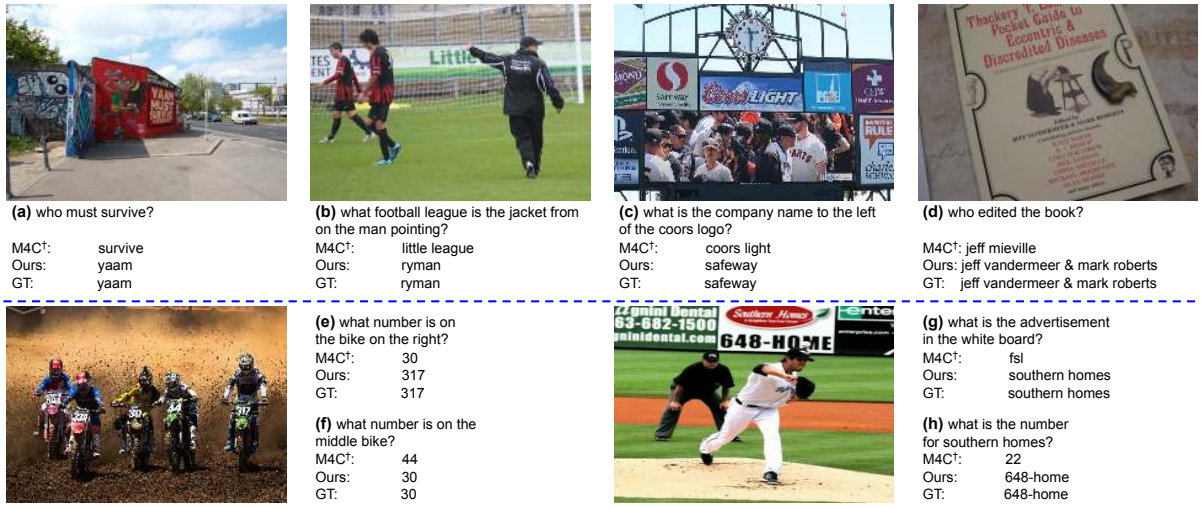

Figure 4. Failure cases of the non-TAP baseline "M4C†" that can be corrected by "TAP."

4.6. Qualitative results

Figure 4 shows representative failure cases of the non-TAP baseline "M4C†" that can be corrected by "TAP." These cases show that TAP improves Text-VQA/Text-Caption by learning better aligned representations.

- TAP shows a good performance on challenging questions that require paraphrasing the scene text sentences. For example, in Figure 4 (a), the model answers "who *must survive*" by the scene text "yaam must survive" in the image. The attention in Figure 3 further visualizes the latent region-word alignments.
- TAP also performs better on questions that refer to a scene text via an intermediate object. For example, in Figure 4 (b), the model grounds the object region "the jacket on the man pointing" and generates the correct answer "ryman" with the scene text "ryman football league" on the man's jacket.
- Figure 4 (c) shows an example that TAP correctly understands the relative spatial relationship in question.
- Furthermore, TAP helps the model read a large piece of text. For example, in Figure 4 (d), the model correctly answers the question "who edited the book" by finding the editors' names "jeff vandermeer & mark roberts." We note that each word is detected as a separate scene text region, *e.g.*, "jeff," "&," *etc.*, which makes the answer sequence prediction non-trivia.

The bottom row of Figure 4 shows examples of multiple questions on the same image. For example, (e,f) (g,h) show that the model selects correct scene text regions as the answer based on the input questions. More qualitative results are included in the supplementary material.

5. Conclusion

We have presented Text-Aware Pre-training (TAP) that explicitly incorporates scene text in pre-training and effectively learns a better aligned multi-modality representation for Text-VQA/Text-Caption. With the identical framework and training data, TAP boosts the non-TAP baselines by $+5.4\%$ in absolute accuracy on the TextVQA challenge. Furthermore, we build a large-scale dataset named OCR-CC and further improve the TAP performance. TAP outperforms the state-of-the-art methods by large margins. Analyses show that TAP helps the aligned representation learning among text word, visual object, and scene text.

Acknowledgment

This work is supported in part by NSF awards IIS-1704337 and IIS-1813709, as well as our corporate sponsors.

References

[1] 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. 3

[2] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2552–2566, 2014. 5

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016. 5

[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. 1, 3, 6

[5] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018. 1

[6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *ICCV*, 2015. 1, 2, 3

[7] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samual White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pages 333–342, 2010. 1

[8] 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. 1, 2, 4, 6

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

[10] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *SIGKDD*, pages 71–79, 2018. 5

[11] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *ECCV*, 2020. 2, 7

[12] 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, 3, 7

[13] 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

[14] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 5

[15] 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. 3, 4, 5, 7

[16] 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. 1, 2, 4, 5, 6

[17] 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. 1, 2, 4, 5, 6

[18] 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. 1, 4

[19] 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. 2, 5, 6

[20] 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. 1, 2, 3, 4, 5, 6

[21] Xiaowei Hu, Xi Yin, Kevin Lin, Lijuan Wang, Lei Zhang, Jianfeng Gao, and Zicheng Liu. Vivo: Surpassing human performance in novel object captioning with visual vocabulary pre-training. *arXiv preprint arXiv:2009.13682*, 2020. 3

[22] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, pages 4634–4643, 2019. 6

[23] 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, 3

[24] 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. 1, 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. 1, 2, 3, 4, 5, 6

[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. 4

[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. 4

[28] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 2, 3

[29] 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. 2, 4

[30] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 4

[31] 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. 3

[32] 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, 3, 7

[33] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. What does bert with vision look at? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5265–5275, 2020. 7

[34] 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, 3

[35] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 5

[36] 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. 2, 5, 6

[37] 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, 3, 4

[38] 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. 3

[39] Anand Mishra, Karteek Alahari, and CV Jawahar. Image retrieval using textual cues. In *ICCV*, pages 3040–3047, 2013. 4

[40] 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. 1, 2

[41] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Advances in neural information processing systems*, pages 1143–1151, 2011. 2, 4

[42] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002. 5

[43] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, 2015. 2, 4

[44] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 5

[45] 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. 2, 4, 5

[46] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. *arXiv preprint arXiv:2003.12462*, 2020. 1, 2, 3, 4, 5, 6

[47] Amanpreet Singh, Vedanuj Goswami, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Mmf: A multimodal framework for vision and language research. https://github.com/facebookresearch/mmf, 2020. 5

[48] Amanpreet Singh, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia-a platform for vision & language research. In *SysML Workshop, NeurIPS*, volume 2018, 2018. 5

[49] 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. 1, 2, 4, 5, 6

[50] 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. 3

[51] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, pages 5103–5114, 2019. 2, 3, 7

[52] 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. 1, 3, 5

[53] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015. 5

[54] 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. 1, 2, 3, 4

[55] Jing Wang, Tang Jinhui, and Luo Jiebo. Multimodal attention with image text spatial relationship for ocr-based image captioning. In *ACMMM*, 2020. 1, 2, 3, 4, 5, 6

[56] 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. 1, 2

[57] Zhaokai Wang, Renda Bao, Qi Wu, and Si Liu. Confidence-aware non-repetitive multimodal transformers for textcaps. In *AAAI*, 2021. 6

[58] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019. 3

[59] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, pages 684–699, 2018. 4, 7

[60] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2, 3

[61] 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. 3