TAP: Text-Aware Pre-training for Text-VQA and Text-Caption (Supplementary Material)

A. The OCR-CC Dataset

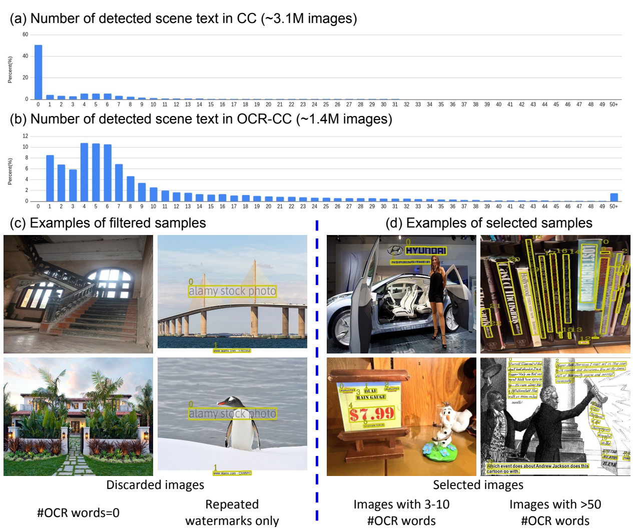

Figure A. **(a,b)** The distribution of the detected scene text number by Microsoft-OCR on the Conceptual Captioning (CC) dataset [5] and our OCR-CC dataset. **(c,d)** Representative examples of discarded and selected images. We draw the OCR box over multiple related words for visualization purposes. We note that each scene text region contains a single word, *e.g.*, four words "HYUNDAI," "INSPIRING," "THE," "FL" in the top left sub-figure of (d).

In this section, we introduce the details of building the OCR-CC dataset based on the Conceptual Captioning (CC) dataset [5]. First, we run the Microsoft Azure OCR system on all CC images (around 3.1 million). Then, we discard the images that don't have scene text (around half of the CC images) or have watermark "text" only (around 5% of the CC images). These watermark "text" records the source image website/provider and are thus not related to the image content. Figure A (c) shows examples of the discarded images, which either have no detected scene text or have watermark "text" only. In the end, we select $1,367,170$ images from CC as the images in our OCR-CC dataset. We pair each selected image with a caption $\mathbf{w}$ for pre-training. The caption text $\mathbf{w}$ is the concatenation of the original image caption $\mathbf{w^q}$ in CC, the detected object labels $\mathbf{w^{obj}}$, and the detected scene text words $\mathbf{w^{ocr}}$. Figures A (a,b) visualize the distribution of the scene text number in CC and our OCR-CC, respectively. Similar to the distribution on TextVQA [7] and ST-VQA [2], the majority of images contains 3-10 detected scene text regions, while a small portion of images has a large number of scene text regions. Figure A (d) shows some representative selected images.

Table A. Results on the TextCaps [6] validation set. B-4, M, R, S, C short for BLEU, METEOR, ROUGE_L, SPICE, CIDEr, respectively. The oracle analyses are shown in the gray text color.

Method	B-4	M	R	S	C
BUTD [1]	20.1	17.8	42.9	11.7	41.9
AoANet [4]	20.4	18.9	42.9	13.2	42.7
M4C [6]	23.3	22.0	46.2	15.6	89.6
MMA-SR [8]	24.6	23.0	47.3	16.2	98.0
M4C† [6]	24.3	22.9	47.3	16.5	99.9
TAP (Ours)	25.2	23.4	47.7	16.9	105.0
TAP†† (Ours)	**25.8**	**23.8**	**47.9**	**17.1**	**109.2**
M4C (GT OCR) [6]	26.0	23.2	47.8	16.2	104.3

Table B. Results on the TextCaps [6] test set.

Method	B-4	M	R	S	C
BUTD [1]	14.9	15.2	39.9	8.8	33.8
AoANet [4]	15.9	16.6	40.4	10.5	34.6
M4C [6]	18.9	19.8	43.2	12.8	81.0
CNMT[9]	20.0	20.9	44.4	13.5	93.0
M4C† [6]	20.4	20.7	44.6	13.6	93.4
TAP (Ours)	21.5	21.7	45.4	14.5	99.5
TAP†† (Ours)	**21.9**	**21.8**	**45.6**	**14.6**	**103.2**
M4C (GT OCR) [6]	21.3	21.1	45.0	13.5	97.2
Human [6]	24.4	26.1	47.0	18.8	125.5

B. TextCaps Results

Tables A, B present the full results on TextCaps [6] to supplement the abstracted results in the main paper's Table 3. We draw similar conclusions from Tables A, B as the ones in the main paper. Specifically, "TAP" significantly improves the non-TAP baseline "M4C†" in all metrics with the identical network architecture and training data. Our TAP approach also outperforms the previous state of the art [6, 8, 9] by large margins.

Furthermore, we compare TAP with the oracle numbers, as shown in the gray text color at the bottom part of Tables A, B. "TAP" outperforms the "M4C (GT OCR)" that uses ground-truth scene text detection in training and inference. Meanwhile, there still exists a gap between "TAP" and human performance. We expect future studies focusing on captioning to further reduce the gap, *e.g.*, with better decoding step pre-training designed especially for captioning.

C. Hyper-parameters

We summarize the hyper-parameters used in the "TAP" and "TAP††" experiments. We conduct experiments based on the M4C [3, 6] and follow most of its hyper-parameter selections, as shown in Table C. We highlight the changed parameters in bold in the table.

- First, the max length of the extended text input $\mathbf{w} = [\mathbf{w^q}, \mathbf{w^{obj}}, \mathbf{w^{ocr}}]$ is set to $20 + 100 + 100 = 220$.
- Second, we increase the max length of scene text $\mathbf{v^{ocr}}$ from 50 to 100 when experimented with Microsoft-OCR. Compared with Rosetta, Microsoft-OCR generates more detected scene text regions in each image. For example, in the TextVQA dataset, the mean and median of scene text numbers are 12.8 and 8 with Rosetta, and are 23.1 and 12 with Microsoft-OCR. With Rosetta, 3.5% of images contain more than 50 scene text regions detected, while the percentage is 14.3% with Microsoft-OCR. To cover more detected scene text, we increase the max length of scene text $\mathbf{v^{ocr}}$ from 50 to 100 when experimented with Microsoft-OCR.
- In the experiment of "pre-training without extra data" ("TAP"), we follow the same learning rate step and maximum iteration settings as used in the fine-tuning. In pre-training with OCR-CC ("TAP††"), we pre-train the model for a maximum iteration of $480K$ and scale the learning rate steps linearly.

D. Pre-train + Fine-tune *vs*. Joint-train

Results in the main paper's Section 4.3 show that TAP works well even without extra data. We hypothesize that we can view TAP as a multi-task learning framework, and obtain similar improvement by using the pre-training tasks (MLM, ITM, RPP) as the auxiliary training loss. Therefore, we explore an alternative training pipeline named "joint train," where the pre-training tasks are used as the auxiliary losses together with the main answer/caption loss. Because MLM and ITM tasks require "polluting" the input sequence, we randomly select 50% of the samples in a batch to compute the pre-training loss and keep the remaining 50% unchanged for the answer/caption loss.

Studies show that these two training pipelines can achieve similar performances, *i.e.*, 49.91% for "pre-train + fine-tune" and 49.46% for "joint train" on TextVQA. Both methods significantly outperform the non-TAP baseline (44.50%). For "joint train," we train the framework for 120K iterations. Compared with "joint train," one advantage of the "pre-train + fine-tune" pipeline in the main paper is that the extra weak data with no answer/caption annotations can be more easily used.

The effectiveness of different TAP pipelines implies the potential of improving other multi-modal tasks by incorporating pre-training tasks. Specifically, the pre-training tasks can be used either in the "joint-train" approach to best preserve the main task's training pipeline, or in the "pre-train + fine-tune" approach to benefit from the large-scale weak pre-training data.

Table C. Hyper-parameters of the TAP experiments with and without OCR-CC pre-training, *i.e.*, "TAP††" and "TAP." We conduct the experiments based on M4C [3, 6] and highlight the changed parameters in bold. We detail these changes in Section C.

Hyper-parameter	Value
(a) General parameters	
max length of text word $\mathbf{w}$	**220**
max length of visual object $\mathbf{v^{obj}}$	100
max length of scene text $\mathbf{v^{ocr}}$	**100**
optimizer	Adam
batch size	128
base learning rate	1e-4
warm-up learning rate factor	0.2
warm-up iterations	2000
max gradient L2-norm for clipping	0.25
learning rate decay	0.1
(b) Pre-training parameters	
learning rate steps ("TAP," VQA)	14K, 19K
max iterations ("TAP," VQA)	24K
learning rate steps ("TAP," Caption)	10K, 11K
max iterations ("TAP," Caption)	12K
learning rate steps ("TAP††")	**280K, 380K**
max iterations ("TAP††")	**480K**
(c) Text-VQA Fine-tuning (TextVQA, ST-VQA)	
max length of decoding step	12
learning rate steps	14K, 19K
max iterations	24K
(d) Text-Caption Fine-tuning (TextCaps)	
max length of decoding step	30
learning rate steps	10K, 11K
max iterations	12K

E. Qualitative Results

In this section, we present additional qualitative examples. Figure B shows the failure cases that can be corrected by OCR detection. Figure C presents the failure cases of our method. "TAP" occasionally fails on samples that require complex reasoning (Figures C (a,b)) or have incorrect scene text detection (Figures C (c,d)). For example, in Fig-

(a) what is the name of the bar?

M4C†: 15
Ours: moon bar
GT: moon bar

(b) what type of beer is in the blue can?

M4C†: unanswerable
Ours: bud light
GT: bud light

(c) what name is on the patch?

M4C†: unanswerable
Ours: clemson
GT: clemson

(d) what company made most of these books?

M4C†: timehop
Ours: marvel
GT: marvel

Figure B. Failure cases that can be corrected by scene text detection. The top and bottom rows visualize the detected scene text by Rosetta-OCR and Microsoft-OCR, respectively. We draw adjacent words into the same box for visualization purposes and highlight the key scene text regions for the question, *e.g.*, "moon bar," "bud light," "clemson," and "marvel."

(a) what brand is on the white bag?

M4C†: uni
Ours: cutfittep
GT: aldo

(b) what is the jersey number of the man in green on the far right?

M4C†: 2
Ours: 6
GT: 5

(c) what candy bar is down there on the bottom?

M4C†: jack daniels
Ours: honey
GT: hershey's

(d) what is the largest measurement we can see on this ruler?

M4C†: 40
Ours: 40
GT: 50

Figure C. Representative failure cases of "TAP." We highlight the key scene text regions for each question.

ure C (a), TAP selects the scene text "cutfittep" on the black bag as the answer, instead of the correct scene text "aldo" on the referred white bag.

References

[1] 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

[2] 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

[3] 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] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, pages 4634–4643, 2019. 1

[5] 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. 1

[6] 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

[7] 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

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

[9] Zhaokai Wang, Renda Bao, Qi Wu, and Si Liu. Confidence-aware non-repetitive multimodal transformers for textcaps. In *AAAI*, 2021. 1