LaTr: Layout-Aware Transformer for Scene-Text VQA
Supplementary Material

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

A. Implementation Details

In this section we detail the implementation specifics of our paper divided into three parts; (1) pre-training; (2) fine-tuning; (3) ablation studies. In our work all models are pre-trained on 8 A100 GPUs and are implemented using PyTorch [20]. T5 uses SentencePiece [13] to encode the text as WordPiece tokens [12,21], we use a vocabulary of 32,000 wordpieces for all experiments.

Pre-training. For the base-size model, we utilize a batch size of 25 for each GPU with the maximum OCR token length set to 512 and pre-training is done for 2.2M steps. For the large-size model, we use a batch size of 28 for each GPU with the maximum OCR token length set to 384 and pre-training is done for 0.9M steps. In both models, the learning rate is increased linearly over the warm-up period of 100K steps to 1e-4 learning rate and then linearly decayed to 0 at the end of the training, and we enable gradient accumulation. For our *layout-aware* de-noising task, we corrupt 15% of the original text sequence, with a span length which vary as a function of the amount of text in each sample.

Fine-tuning. We train all of our models for 100K steps and use AdamW [16] optimizer with 1e-4 max learning rate. Warm-up period is set to 1,000 steps and again is linearly decayed to zero. The same batch sizes that were used for pre-training are also used in this stage. We use a ViT [5] to extract visual features. The ViT is pre-trained and fine-tuned on ImageNet [4] for classification. We follow the implementation and use the weights from HuggingFace [26] [1].

Ablation studies. For ablating the visual backbone, we follow the common practice [8, 24, 25, 29] of detecting objects with a Faster R-CNN detector [2] which is pre-trained on the Visual Genome dataset [11]. We keep the 100 top-scoring objects per image and, similarly to previous work, only fine-tune the last layer. We now detail the specifics of our pre-training ablation studies. When exploring the effect of pre-training with visual features, we combine the de-noising pre-training task with an image-text (contrastive) matching (ITM) task. For the ITM taks, we follow the same implementation as in [29], the text input is polluted 50% of the time by replacing the whole text sequence with a randomly-selected one from another batch. The polluted text words are thus not paired with the visual patch features from the ViT. The ITM task takes the sequence feature as the input and aims to predict if the sequence has been polluted or not. One important point to mention is that for the de-noising task, we compute the gradients for both encoder and decoder. Yet, for the ITM task, we merely compute the gradients for our encoder.

For the vocabulary reliance experiment, we collect the top 5000 frequent words from the answers in the training set as our answer vocabulary as done by [8, 29].

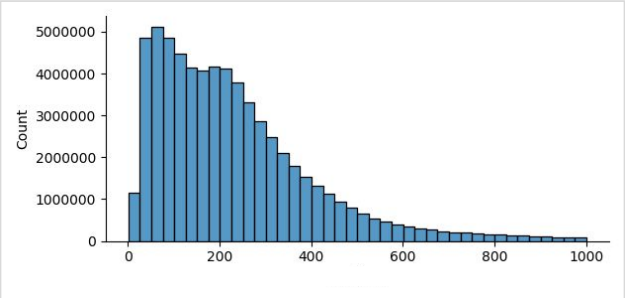
B. Datasets

TextVQA [24] contains 28k images from the Open Images [14] dataset. The questions and answers are collected through Amazon Mechanical Turk (AMT) where the workers are instructed to come up with questions that require reasoning about the scene text in the image. Following VQAv2 [6], 10 answers were collected for each question. In total, there are 45k questions divided into 34,602, 5,000, and 5,734 for train, validation and test set, respectively.

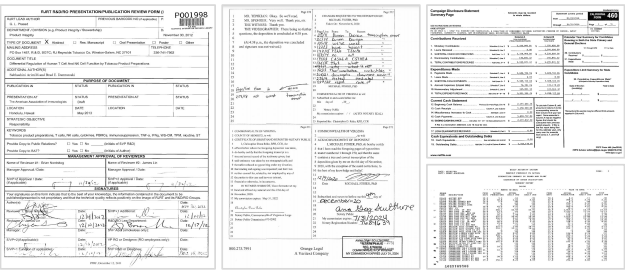
ST-VQA [3] is an amalgamation of well-known computer vision datasets, namely: ICDAR 2013 [10], ICDAR2015 [9], ImageNet [4], VizWiz [7], IIIT Scene Text Retrieval [17], Visual Genome [11] and COCO-Text [28]. ST-VQA is also collected through AMT, asking workers to come up with questions so that the answer is always the scene text in the image. In total, there are 31k questions, separated into 26k questions for training and 5k questions for testing.

TextCaps [23] is composed of 28,408 images, when there

*Authors contribute equally.
†Work done during an internship at Amazon.
[1] https://huggingface.co/transformers/model_doc/vit.html



(a) Number of detected scene text in IDL



(b) Examples of images in IDL

Figure 1. **IDL dataset.** (a) We show the distribution of the detected OCR number by Textract OCR [1,15,19] on the IDL dataset. (b) We visualize representative examples from the dataset.

are 5 captions per image, amounting to a total of 142,040 captions. The images are taken from TextVQA [24] dataset. The dataset is annotated with AMT. The AMT annotators are asked to provide captions that are based on the text in the image. In other words, the captions can not be generated without having OCR tokens, however, the provided captions do not necessarily contain the OCR tokens.

OCR-VQA [18] is composed of 207,572 images of book covers and contains more than 1 million question-answer pairs about these images. The questions are template-based, asking about information on the book such as title, author, year. The questions are all can be answered by inferring the book cover images.

OCR-CC [29] is a subset of Conceptual Captions (CC) [22] dataset proposed by [29]. This subset is compromised of 1.367 million scene text-related image-caption pairs. To obtain OCR-CC, [29] used the Microsoft Azure OCR system to extract the text in the image, then any image that does not contain any text or any image that only has watermarks is discard. As this subset is not publicly released, we follow the same process to create it. However, we use Amazon-OCR[2] as our main OCR system. As was presented in [29], the distribution of the detected scene text in the original CC datasets is that only 45.16% of the images contain text. Out of the images that do contain text, the data has a mean and median of 11.4 and 6 scene text detected per image.

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

C. The Industrial Document Library dataset

In this subsection, we present more details on the Industrial Document Library (IDL)[3] dataset. As mentioned in the main paper, the IDL is a digital archive of documents created by industries which influence public health. The IDL is hosted by the University of California, San Francisco Library. It hosts millions of documents publicly disclosed from various industries like drug, chemical, food and fossil fuel. The data from the website is crawled, leading to about 13M documents, which translate to about 70M pages (64M usable) of various document images. IDL has various documents (like forms, tables, letters) with varied layouts as seen in Fig. 1 (b). We extracted OCR for each document using Textract OCR[4] [27].

The crawled and OCR'ed IDL data was pre-processed before consuming for pre-training. We removed all documents which had less than 10 words or the image was unreadable. In addition, to weed out documents having a majority of erroneous OCR text and documents with non-English content, we considered a fixed English dictionary with a 350K-sized vocabulary and check if each OCR word is part of that dictionary with either exact-match or edit-distance of 1. We do not apply this filter if the word is either a number, float, currency or date (as those are unlikely to be present in the fixed English dictionary and would inflate the error count if considered). If the number of erroneous words are $\geq 50\%$ for that document, we ignore it. After all this filtering we are left with about 64M documents (roughly 6M are discarded) which are used for pre-training. The subsets used in Table 5 in the main paper are uniform random samples of this larger 64M data.

We show in Fig. 1 (a) the detected OCR word distribution across all the 64M documents. The plot roughly looks like a right-skewed normal distribution, with the majority of documents lying in the hump (having 20 to 400 words per doc). Unlike OCR-CC, documents by definition contain words, and thus we are able to use over 91% of the original IDL dataset (compared to 45.16% for OCR-CC). In addition, as clearly seen, there are much more words an average in IDL than OCR-CC which is extremely beneficial for pre-training in scene text VQA tasks. In Fig. 1 (b) we depict representative examples from the IDL dataset.

D. Model Capacity

The number of model parameters in M4C ([8]) is 200M (90M for BERT and 110M for FRCNN), while LaTr-Small has 149M (60M for T5, 86M for ViT and 3M for spatial embedding). As seen in Table 1 in the main paper, LaTr-Small without pre-training achieves 41.84% accuracy when trained and evaluated with Rosetta-en and still outperforms

[3] https://www.industrydocuments.ucsf.edu/
[4] ttps://aws.amazon.com/textract/

Method	Val Acc.	Test Acc.
CNN [18]	-	14.3
BLOCK [18]	-	42.0
BLOCK+CNN [18]	-	41.5
BLOCK+CNN+W2V [18]	-	48.3
M4C [8]	63.5	63.9
LaTr-Base	**67.5**	**67.9**

Table 1. **Results on the OCR-VQA Dataset [18]**. We use our base model pretrained on IDL and utilize Rosetta OCR system so that it is comparable across all the models. LaTr improves the state-of-the-art by +4.0%.

M4C (**+2.44%**), showing the gain achieved by our architecture. LaTr-Base has 311M (220M for T5, 86M for ViT and 3M for spatial embedding). We note, only a **+2.22%** is obtained by increasing the model capacity to LaTr-Base compared with LaTr-Small. The significant gain comes from our proposed pre-training strategy, resulting in **+8%** gain, as seen in Table 4 in the main paper.

E. OCR-VQA Results

As commonly done by previous work [8], we only evaluate our model using the constrained setting. In this setting, we do not change the OCR system, *i.e*. we use Rosetta OCR system. Similarly to TextVQA and ST-VQA datasets, LaTr-Base outperforms the previous state of the art [8] by a large margin, specifically, from 63.5% to 67.5% (**+4.0%**).

F. Qualitative Examples

In this section, we present additional qualitative examples of our method compared with M4C [8]. In the first four columns of Fig. 2, we display examples in which our model is successful while M4C fails. Compared to M4C, our model clearly has better natural language understanding (top left image). In addition, our model has the ability to reason over layout information significantly better than M4C (third image in row 3). This is both attributed to the extensive pre-training and the fact that we leveraged documents for performing *layout-aware* pre-training with 2-D spatial position embedding.

Out of the cases displayed, we wish to further discuss two types of observed biases in the data. The first is for the question asking "*what is the handwritten message?*". Our model successfully answers this question, both with and without visual features. This indicates that, at-least for the model without the visual features, the model is just guessing based on some heuristic. In this case, it could be that the largest OCR bounding box is the most probable answer. As all the datasets were created by AMT it is possible that the annotators created most of the questions base on the largest or the clearest text in the image. The second type of observed bias is the fact that most images contain only a few pieces of text. Thus, the model can make a lot of educated guesses. For example, the question asking "*What is the number on the rear of the white car?*". There are only two numbers in the image, thus giving the model at-least 50% chance of guessing correctly. Similarly, more than 85% "Yes/No" questions are with answers "Yes" in TextVQA dataset, given the model a strong (and incorrect) prior knowledge, allowing easy guesses.

An additional interesting observation is with regard to questions about reading the time from an analog watch. We observed that both our model and the M4C model, in most cases, predict the time of 10:10 regardless of the actual time in the image. This is a bias the models developed from a common marketing trick. Watch sellers displays watches aimed to 10:10 as business marketing research showed it increases sales, and therefore, our model can't actually read the time but just guesses the most likely time based on the pre-training prior knowledge.

In the final column of Fig. 2, we display our model's failure cases. The failure cases are mostly composed of OCR errors, compositionality of spatial reasoning and visual attributes. We wish to further discuss the last example (bottom right) as we believe this is an example of a question which requires a higher level of "intelligence" than the other examples. To answer this question, the model has to not only reason over both the image and the text, but also to understand that the soda wish to be like the regular coca-cola as it is "imagining" its reflection in the mirror.

G. Dataset Bias or Task Definition?

In the main paper, we showed that STVQA models (ours included) make use of the visual features marginally. This begs the question whether this is because of a dataset bias, or is it simply the task nature. To explore this, we attempt to categorize the type of questions current benchmarks consist of. We divide the questions in TextVQA [24] into four different categories. The questions categories are defined by the information type required to answer them. The first category consist of all questions that can be answered with just an order-less bag of words. In Fig. 3 (1) we depict examples from this category, *i.e*. question that do not require anything beyond the order-less bag-of-words and some world knowledge. Base on the analysis presented in the main paper, this category amounts to over 40% of the test data and include the questions that can be answered with just the questions ($\approx$11%). The second category consists of questions which require an ordered bag-of-words. Currently, most papers treat the OCR system as a black-box and reading order is so intertwined with OCR systems that it is not thought of as a detached feature. We make the distinction between the information types extracted from the OCR system and demonstrate that an additional 10% of the questions can be

answered by just adding the reading order. Examples from this category are depicted in Fig. 3 (2).

The next category requires to reason over both word tokens and their 2-D spatial layout. In the main paper, we showed that via *layout-aware* pre-training, we are able to leverage the additional layout information to boost performance by over 7%. Base on a qualitative analysis, we believe that 7% is the lower bound of this category size and more questions can be answered by just reasoning over the text and its layout. Examples from this category can be found in Fig. 3 (3). The last category consists of question which require reasoning over all modalities, specifically the text, the layout information and the image itself. Generating such questions is not an easy task, and therefore in current benchmarks most question do not fall under this category. We believe that in order to advance the field of STVQA this issue needs to be addressed. We propose a simple mechanism for determining whether an image falls under the last category. In this mechanism the question is given to the annotator with just the words and layout visualization (third column of Fig. 3), if the question can still be answered it should be dropped. Examples from this category are depicted in Fig. 3 (4).

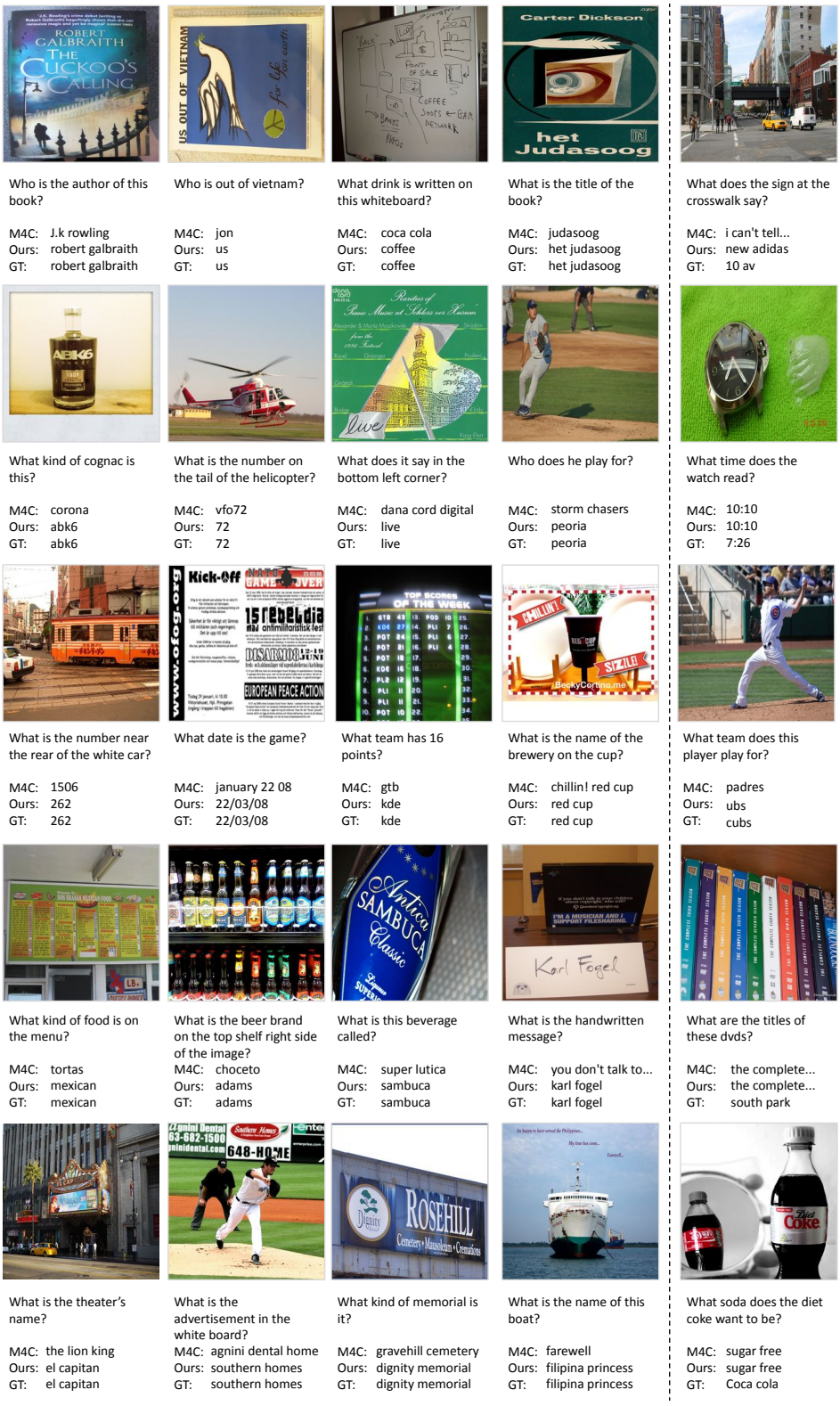


Figure 2. **Qualitative Examples.** The first four columns displays failure cases of M4C [8] in which our model is successful. As can be seen, LaTr is able to outperform M4C on a variety of different question types, including, layout, world knowledge, natural language understand and more. In the last column, we present fail cases of our model, demonstrating representative failure cases of LaTr. We note that we present the questions as they are originally appear in the TextVQA dataset [24]

Figure 3. **Dataset Bias or Task Definition?.** We depict four different questions types based on the information needed to answer them. Questions which require; (a) order-less bag-of-words; (b) ordered bag-of-words; (c) words and their 2-D spatial layout; (d) words, their 2-D spatial layout and the image.

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. 2

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, pages 382–398. Springer, 2016. 1

[3] 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

[4] 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. 1

[5] 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. 1

[6] 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. 1

[7] 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

[8] 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, 5

[9] 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. 1

[10] 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. 1

[11] 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. 1

[12] Taku Kudo. Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*, 2018. 1

[13] 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. 1

[14] 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. 1

[15] 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. 2

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[17] A. Mishra, K. Alahari, and C. V. Jawahar. Image retrieval using textual cues. In *ICCV*, 2013. 1

[18] 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, 3

[19] Oren Nuriel, Sharon Fogel, and Ron Litman. Textadain: Fine-grained adain for robust text recognition. *arXiv preprint arXiv:2105.03906*, 2021. 2

[20] 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. 1

[21] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015. 1

[22] 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

[23] 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

[24] 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, 3, 5

[25] 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. 1

[26] 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. 1

[27] William Ughetta. The old bailey, us reports, and ocr: Benchmarking aws, azure, and gcp on 360,000 page images. 2021. 2

[28] 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

[29] 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