Transferring General Multimodal Pretrained Models to Text Recognition

Junyang Lin, Xuancheng Ren, Yichang Zhang, Gao Liu,
Peng Wang, An Yang, Chang Zhou
DAMO Academy, Alibaba Group
junyang.ljy@alibaba-inc.com

Abstract

This paper proposes a new method, OFA-OCR, to transfer multimodal pretrained models to text recognition. Specifically, we recast text recognition as image captioning and directly transfer a unified vision-language pretrained model to the end task. Without pretraining on large-scale annotated or synthetic text recognition data, OFA-OCR outperforms the baselines and achieves state-of-the-art performance in the Chinese text recognition benchmark. Additionally, we construct an OCR pipeline with OFA-OCR, and we demonstrate that it can achieve competitive performance with the product-level API. The code[1] and demo[2] are publicly available.

1 Introduction

Optical character recognition (OCR) plays an important role in the real-world applications. It helps users or developers extract text contents from different types of images, including photos, scanned documents, etc. In practice, building a tool for OCR needs a pipeline consisting of a text localization module and a text recognition module.

In this work, we focus on improving the accuracy of text recognition. Text recognition has often been regarded as a key challenge owing to the room for improvements in recognition accuracy. In the deep learning era, the classical methods are mostly based on CNN and RNN, which are responsible for visual feature extraction and sequence modeling, respectively (Shi et al., 2017a, 2019; Luo et al., 2019). Recently, with the rise of Transformer (Vaswani et al., 2017), researchers applied the Transformer encoder-decoder framework to text recognition and achieved outperforming results over the baselines (Li et al., 2021; Lyu et al., 2022). However, most methods are based on large-scale pretraining on human-annotated or synthetic OCR data. It is hard for other researchers to collect or create such data for reproduction. Furthermore, the methods often include complex model or objective designs, like DETR-like decoder (Carion et al., 2020), CTC loss (Graves et al., 2006), etc. These components also might hinder reproduction as they increase the difficulty in training. Therefore, we naturally raise a question: *Is there any way to achieve high recognition accuracy without complex designs on data and model?*

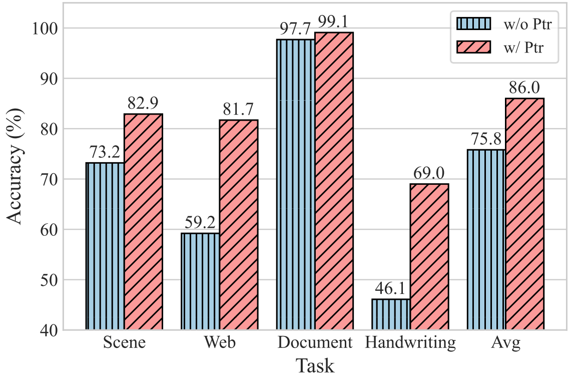

Figure 1: **An comparison between the performance with or without general vision-language pretraining.** On 4 subtasks of text recognition, OFA-OCR with general-domain vision-language pretraining significantly outperforms the baseline without one.

Inspired by the recent progress in multimodal pretraining, we argue that the transfer of a unified multimodal pretrained model is a possible solution. Multimodal pretraining has proved significant to the performance of downstream tasks, and thanks to the rise of unified multimodal pretrained models, they can perform both cross-modal understanding and generation and achieve state-of-the-art performance (Wang et al., 2022a,b; Lu et al., 2022). We therefore propose to transfer the unified multimodal pretrained model by finetuning the pretrained model on the text recognition datasets with the task of image captioning, which is essentially

[1] https://github.com/OFA-Sys/OFA

[2] https://modelscope.cn/studios/damo/ofa_ocr_pipeline/summary

Findings of the Association for Computational Linguistics: ACL 2023, pages 588–597
July 9-14, 2023 ©2023 Association for Computational Linguistics

a simple sequence-to-sequence learning task with maximum likelihood estimation for optimization.

To support the effectiveness of the proposed method, we have conducted extensive experiments on the Chinese text recognition benchmark (Chen et al., 2021b) covering multiple scenarios, including scene, web, document, and handwriting. Specifically, we finetune the open-source Chinese multimodal pretrained model OFA (Wang et al., 2022a) on text recognition, and we name the model OFA-OCR. Figure 1 demonstrates the results of methods with or without general-domain pretraining. It shows that multimodal pretraining on general-domain vision-language data can effectively boost downstream performance in text recognition. To achieve the best performance, we apply the multitask + single-task finetuning to OFA-OCR, and it outperforms the previous state-of-the-art methods on the benchmark. Furthermore, through the ablation studies, we demonstrate the effectiveness of our method designs, including multitask + single-task finetuning, data augmentation, etc. Furthermore, to enable deployment for real-world applications, we construct a pipeline with both OFA-OCR and a simple text localization module. We find that this simple pipeline can provide high-quality OCR performance, competitive with a product-level API.

2 Method

2.1 Pretraining

To leverage the capability of the multimodal pretrained model for image captioning, we employ the unified multimodal pretrained model architecture. Specifically, we implement our models on OFA (Wang et al., 2022a), an open-source state-of-the-art unified multimodal pretrained model with the release of Chinese models.

The model is mainly based on the Transformer encoder-decoder framework (Vaswani et al., 2017). To make information from different modalities adaptable to the Transformer, there are adaptors for images and texts, which are visual backbones, e.g., ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2021), etc., and word embeddings, respectively. The information from modalities is encoded as discrete tokens so that the decoder can perform their generation.

For Chinese multimodal pretraining, OFA-Chinese was pretrained on a large-scale dataset, which consists of LAION-5B (Schuhmann et al., 2022), Wukong dataset, as well as translated datasets from MSCOCO (Chen et al., 2015), Visual Genome (Krishna et al., 2017), VQA (Goyal et al., 2017), RefCOCO (Yu et al., 2016), etc.

Note that this work is different from previous pretraining-related methods, which pretrain the model on large-scale human-annotated or synthetic data. We show that through pretraining on general-domain data, the model can obtain the potential of text recognition by finetuning on small datasets.

2.2 Finetuning with Image Captioning

It is natural to recast text recognition as image captioning, as text recognition also requires the model to generate a piece of text based on the input image. It is equivalent to finetuning on different image captioning datasets, where the target refers to the text on the image. We finetune the model with maximum likelihood estimation for optimization.

Furthermore, to better alleviate the discrepancy between upstream and downstream data, we apply a transformation to the input images to make them square, e.g., a resolution of 480×480. Specifically, we first resize the image to a longer edge of the specified resolution while keeping the original height-width ratio of the image, and we make the image square by padding on all sides with the edge value. The lengths for the directions are random, and thus this method can play as data augmentation in this context. We demonstrate the pseudo code in Sec. A.3.

For better performance in the downstream tasks, we often use a larger resolution in the finetuning stage, and thus we encounter issues with the positional embedding. In our practice, we still use the same one from pretraining but apply interpolation to adapt to images of a larger resolution.

2.3 Multitask Finetuning

There are multiple subtasks in text recognition, concerning different scenarios, e.g., scene, document, etc. Our experiments are implemented on the Chinese text recognition benchmark consisting of 4 subtasks. In our practice, we implement multitask finetuning and single-task finetuning for comparison. Specifically, as the data of all subtasks are organized with the same format, we directly build a mixture of datasets for multitask finetuning. We find that directly applying multitask finetuning can help OFA-OCR achieve outstanding performance on all datasets. To further boost its performance, we additionally apply single-task finetuning after

Metrics	Scene	Web	Document	Handwriting	Average
CRNN (Shi et al., 2017a)	53.4	54.5	97.5	46.4	67.0
ASTER (Shi et al., 2019)	54.5	52.3	93.1	38.9	64.7
MORAN (Luo et al., 2019)	51.8	49.9	95.8	39.7	64.3
SAR (Li et al., 2019)	62.5	54.3	93.8	31.4	67.3
TransOCR (Chen et al., 2021a)	63.3	62.3	96.9	53.4	72.8
MaskOCR$_{\text{ViT-B}}$	73.9	74.8	99.3	63.7	80.8
MaskOCR$_{\text{ViT-L}}$	76.2	76.8	99.4	67.9	82.6
OFA-OCR$_{\text{Base}}$	82.9	81.7	99.1	69.0	86.0
OFA-OCR$_{\text{Large}}$	83.7	82.6	99.2	67.7	86.3

Table 1: **Experimental results on the Chinese text recognition benchmark.** Results show that the base-size OFA-OCR model can outperform the previous state-of-the-art, and the large-size model achieves the best performance on average.

multitask finetuning, and we find that this pushes its performance to the new state-of-the-art.

3 Experiments

3.1 Datasets and Metrics

We implement OFA-OCR on the Chinese text recognition benchmark (Chen et al., 2021b). This benchmark consists of multiple subtasks of text recognition, which are text recognition in different scenarios, including scene, web, document, and handwriting. The details of the datasets are provided in Sec. A.1. The evaluation metric includes accuracy, which refers to the ratio of exact match.

3.2 Experimental Results

The experimental results are demonstrated in Table 1. We compare our method with baseline models of OCR, including the previous state-of-the-art MaskOCR (Lyu et al., 2022). It can be found that with no regard to the scale of models, the base-size OFA-OCR, which is finetuned from the pretrained Chinese OFA$_{\text{Base}}$, can outperform both the base-size and large-size MaskOCR models. Specifically, it shows the advantages of 9.0, 6.9, and 5.3 absolute improvements in the scenarios of scene, web, and handwriting. On average, the base-size OFA-OCR outperforms the base-size MaksOCR by 5.2 and the large-size MaskOCR by 3.4. Scaling up the model size can consistently bring steady improvement in the downstream performance. On average, OFA$_{\text{Large}}$ reaches the best results of 86.3.

Specifically, we find that the advantage in the scene dataset is the largest among the tasks. This may be attributed to the pretraining on general-domain data, where there are images of street views, and some of them might contain texts. Similarly, the pretraining dataset consists of web images that resemble those in the web dataset, and thus the gaps between OFA-OCR and the previous methods are large. However, text recognition for documents should be a simpler task as the texts are more regular in fonts and there is often much less noise in the background. Thus, even the conventional method like CRNN can achieve a high accuracy.

3.3 Ablation Study of Training Strategies

To check how the multitask learning influences the final performance, we conduct an ablation study to evaluate its effects. Specifically, the experiments are conducted with the base-size OFA-OCR. We provide experiments in 4 setups, which are training from scratch (scratch), single-task finetuning (ft), multitask-finetuning (mt), and multitask + single-task finetuning (mt+ft), respectively. Experimental results are shown in Figure 2. It can be found that on average, the addition of the initialization of the pretrained OFA model significantly boosts the performance on the datasets. Surprisingly, multitask finetuning alone can outperform single-task finetuning on all 4 tasks, and the advantage in the web dataset is the most obvious. We assume that this is attributed to the small amount of supervised training data for downstream transfer. A mixture of datasets of related subtasks can encourage performance on all subtasks. Furthermore, the combination of multitask finetuning and single-task finetuning is the best solution owing to its outstanding performance, while multitask finetuning on the mixture of datasets is the most cost-efficient.

3.4 Ablation Study of Data Augmentation

The preprocessing of images for this task can play as data augmentation. To validate its effects, we use a simple resizing to the specified resolution as

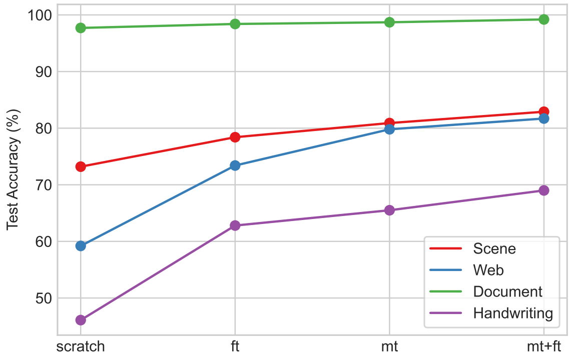

Figure 2: **Performance of OFA-OCR in different setups.** We validate the model performance on the 4 datasets in the setups of training from scratch (scratch), single-task finetuning (ft), multitask-finetuning (mt), and multitask + single-task finetuning (mt+ft). We observe consistent performance growth with the addition of the pretrained weight initialization and multitask finetuning.

Method	w/o Aug.	w/ Aug.
Scene	77.0	78.4
Web	72.3	73.4
Document	98.2	98.4
Web	60.4	62.8
Avg	81.0	82.1

Table 2: **Performance comparison with or without data augmentation for images.** The experiments are conducted in the setup of single-task finetuning on the base-size model.

a baseline. We also implement experiments on the 4 datasets, and for simplicity we implement the experiments in the setup of single-task finetuning on the base-size models. Results are demonstrated in Table 2. We use "Aug." to indicate the preprocessing method mentioned in Sec. 2. The results indicate that the introduced technique for data preprocessing can effectively boost the performance.

3.5 Deployment

To construct an OCR system applicable in real-world scenarios, a strong text recognition model is not sufficient, and we need to build a pipeline with both the text detection and text recognition module. While the former one is not the focus of this research, we directly use a light-weight model from EasyOCR[3] for detection. After detecting all the bounding boxes which possibly contain texts, we crop them with boxes to create a batch of new images. The final step is to process the images with OFA-OCR for the generation of text recognition results. Through our case study, we find that the simple OCR pipeline based on OFA-OCR can achieve competitive performance with the product-level API. Examples are demonstrated in Sec. A.4.

4 Related Work

We focus on the review of text recognition methods and multimodal pretraining. Conventional methods based on CNN and RNN have demonstrated great effectiveness (Shi et al., 2017a; Luo et al., 2019; Shi et al., 2019; Yu et al., 2020; Li et al., 2019; Fang et al., 2021). Recent methods have turned to the use of Transformer and achieved improved performance (Atienza, 2021; Li et al., 2021; Zhang et al., 2022; Lyu et al., 2022). However, before this work, we have not witnessed the direct transfer of general-domain vision-language pretrained models to text recognition. Vision-language pretraining has proved a success as it has leveled up the model performance on a series of downstream tasks (Chen et al., 2019; Lu et al., 2019; Radford et al., 2021; Wang et al., 2021), and the unified models capable of both understanding and generation have become popular and achieved the best performance (Wang et al., 2022a,b). Yet, there are only a few unified multimodal pretrained models in Chinese (Lin et al., 2021; Wang et al., 2022a).

5 Conclusion

In this work, we propose a simple method called **OFA-OCR**, which leverages the unified multimodal pretrained model and transfers it to text recognition by image captioning. To be more specific, we utilize the Chinese multimodal pretrained model OFA without pretraining on OCR data and transfer it to text recognition with multitask + single-task finetuning. Through extensive experiments, we demonstrate that OFA-OCR can achieve state-of-the-art performance on the Chinese text recognition benchmark. Additionally, we build a pipeline of OCR by integrating an existent simple text detection module and OFA-OCR. The deployed pipeline achieves competitive performance in comparison with a product-level API. We hope that this research sheds light on the application of general-domain multimodal pretraining, and also helps OCR practitioners.

[3] https://github.com/JaidedAI/EasyOCR

Limitations

This section discusses the limitations of this work for more insights on the research in this track. Though OFA-OCR achieves high accuracy on multiple text recognition datasets, its costs are larger than the non-Transformer baselines. In practice, it is difficult to deploy such large models. Thus in our future work, we will discover how to distill or compress OFA-OCR to a light-weight model with high efficiency.

Ethics Statement

Our method is essentially based on a generation model, and thus the OCR results should be taken as AI-generated contents. As the generated results should be aligned with the input, we have not noticed deliberate harmful contents, e.g., hate speech, bias, etc. However, the model maintains such ability, which might be triggered. Although after finetuning on the public datasets the risk of such phenomena is extremely low, we still take it into account. In the future research, besides focusing on improving downstream performance, we will study how to increase the controllability on the generation.

References

Rowel Atienza. 2021. Vision Transformer for fast and efficient scene text recognition. In ICDAR (1), volume 12821 of Lecture Notes in Computer Science, pages 319–334. Springer.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In European conference on computer vision, pages 213–229. Springer.

Jingye Chen, Bin Li, and Xiangyang Xue. 2021a. Scene text telescope: Text-focused scene image super-resolution. In CVPR, pages 12026–12035. Computer Vision Foundation / IEEE.

Jingye Chen, Haiyang Yu, Jianqi Ma, Mengnan Guan, Xixi Xu, Xiaocong Wang, Shaobo Qu, Bin Li, and Xiangyang Xue. 2021b. Benchmarking Chinese text recognition: Datasets, baselines, and an empirical study. CoRR, abs/2112.15093.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. 2015. Microsoft COCO captions: Data collection and evaluation server. CoRR, abs/1504.00325.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Uniter: Universal image-text representation learning. In European Conference on Computer Vision.

Chee-Kheng Chng, Yuliang Liu, Yipeng Sun, Chun Chet Ng, Canjie Luo, Zihan Ni, Chuanming Fang, Shuaitao Zhang, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, Chee Seng Chan, and Lianwen Jin. 2019. Icdar2019 robust reading challenge on arbitrary-shaped text - rrc-art. 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 1571–1576.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In ICLR. OpenReview.net.

Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. 2021. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7094–7103.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In CVPR, pages 6325–6334. IEEE Computer Society.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In Proceedings of the 23rd international conference on Machine learning, pages 369–376.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In CVPR, pages 770–778. IEEE Computer Society.

Mengchao He, Yuliang Liu, Zhibo Yang, Sheng Zhang, Canjie Luo, Feiyu Gao, Qi Zheng, Yongpan Wang, Xin Zhang, and Lianwen Jin. 2018. Icpr2018 contest on robust reading for multi-type web images. 2018 24th International Conference on Pattern Recognition (ICPR), pages 7–12.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int. J. Comput. Vis., 123(1):32–73.

Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. 2019. Show, attend and read: A simple and strong baseline for irregular text recognition. In AAAI, pages 8610–8617. AAAI Press.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. TrOCR: Transformer-based optical character recognition with pre-trained models. CoRR, abs/2109.10282.

Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, Jie Zhang, Jianwei Zhang, Xu Zou, Zhikang Li, Xiaodong Deng, Jie Liu, Jinbao Xue, Huiling Zhou, Jianxin Ma, Jin Yu, Yong Li, Wei Lin, Jingren Zhou, Jie Tang, and Hongxia Yang. 2021. M6: A Chinese multimodal pretrainer. CoRR, abs/2103.00823.

Xi Liu, Rui Zhang, Yongsheng Zhou, Qianyi Jiang, Qi Song, Nan Li, Kai Zhou, Lei Wang, Dong Wang, Minghui Liao, Mingkun Yang, Xiang Bai, Baoguang Shi, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. Icdar 2019 robust reading challenge on reading chinese text on signboard. 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 1577–1581.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In ICLR 2019.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Neural Information Processing Systems.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022. Unified-io: A unified model for vision, language, and multi-modal tasks. arXiv preprint arXiv:2206.08916.

Canjie Luo, Lianwen Jin, and Zenghui Sun. 2019. MORAN: A multi-object rectified attention network for scene text recognition. Pattern Recognit., 90:109–118.

Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2022. MaskOCR: Text recognition with masked encoder-decoder pretraining. CoRR, abs/2206.00311.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In ICML, volume 139 of Proceedings of Machine Learning Research, pages 8748–8763. PMLR.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. ArXiv, abs/2210.08402.

Baoguang Shi, Xiang Bai, and Cong Yao. 2017a. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE Trans. Pattern Anal. Mach. Intell., 39(11):2298–2304.

Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. 2019. ASTER: An attentional scene text recognizer with flexible rectification. IEEE Trans. Pattern Anal. Mach. Intell., 41(9):2035–2048.

Baoguang Shi, Cong Yao, Minghui Liao, Mingkun Yang, Pei Xu, Linyan Cui, Serge Belongie, Shijian Lu, and Xiang Bai. 2017b. Icdar2017 competition on reading chinese text in the wild (rctw-17). In 2017 14th iapr international conference on document analysis and recognition (ICDAR), volume 1, pages 1429–1434. IEEE.

Yipeng Sun, Zihan Ni, Chee-Kheng Chng, Yuliang Liu, Canjie Luo, Chun Chet Ng, Junyu Han, Errui Ding, Jingtuo Liu, Dimosthenis Karatzas, Chee Seng Chan, and Lianwen Jin. 2019. Icdar 2019 competition on large-scale street view text with partial labeling - rrc-lsvt. 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 1557–1562.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In NIPS, pages 5998–6008.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In ICML, volume 162 of Proceedings of Machine Learning Research, pages 23318–23340. PMLR.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2022b. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks. CoRR, abs/2208.10442.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. ArXiv, abs/2108.10904.

Deli Yu, Xuan Li, Chengquan Zhang, Tao Liu, Junyu Han, Jingtuo Liu, and Errui Ding. 2020. Towards accurate scene text recognition with semantic reasoning networks. In CVPR, pages 12110–12119. Computer Vision Foundation / IEEE.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C. Berg, and Tamara L. Berg. 2016. Modeling context in referring expressions. In ECCV (2), volume 9906 of Lecture Notes in Computer Science, pages 69–85. Springer.

Tailing Yuan, Zhe Zhu, Kun Xu, Cheng-Jun Li, Tai-Jiang Mu, and Shimin Hu. 2019. A large chinese text dataset in the wild. Journal of Computer Science and Technology, 34:509–521.

Hesuo Zhang, Lingyu Liang, and Lianwen Jin. 2020. Scut-hccdoc: A new benchmark dataset of handwritten chinese text in unconstrained camera-captured documents. Pattern Recognit., 108:107559.

Xinyun Zhang, Binwu Zhu, Xufeng Yao, Qi Sun, Ruiyu Li, and Bei Yu. 2022. Context-based contrastive learning for scene text recognition. In AAAI, pages 3353–3361. AAAI Press.

A Appendix

A.1 Datasets

The Chinese text recognition benchmark consists of 4 subtasks, which are scene, web, document, and handwriting. The scene dataset consists of multiple datasets, including RCTW (Shi et al., 2017b), ReCTS (Liu et al., 2019), LSVT (Sun et al., 2019), ArT (Chng et al., 2019), and CTW (Yuan et al., 2019). It consists of 509,164 samples for training, 63,645 for validation, and 63,646 for testing. The web dataset is derived from MTWI (He et al., 2018), and it has 112,471 samples for training, 14,059 for validation, and 14,059 for testing. The document dataset is constructed with synthetic data created with Text Renderer[4], and it has 400,000 samples for training, 50,000 for validation, and 50,000 for testing. The handwriting dataset is collected from SCUT-HCCDoc (Zhang et al., 2020), and it has 74,603 samples for training, 18,651 for validation, and 23,389 for testing.

A.2 Evaluation

We calculate the ratio of exact match as the accuracy for the evaluation. For the average score on the 4 subtasks, we calculate the average score weighted by the number of testing samples (Lyu et al., 2022).

A.3 Implementation Details

For single-task, multitask, and multitask + single-task finetuning, we finetune the pretrained base-size and large-size OFA for 100 epochs. We use the AdamW (Loshchilov and Hutter, 2019) optimizer for training. For the base-size model, the batch size is 256 and the peak learning rate is 5×10^{-5}, and for the large-size model, the batch size is 512 and the peak learning rate is 2×10^{-5}.

Here we provide more details about the preprocessing for images. The specified resolution is 480×480, and as the pretrained models were pretrained on images of the resolution of 224×224, we apply interpolation to the positional embedding.

As to the data augmentation, we demonstrate the process with the pseudo code below.

```
import torch
from torchvision.transforms import
    InterpolationMode
from torchvision.transforms import functional
    as F
```

[4] https://github.com/Sanster/text_renderer

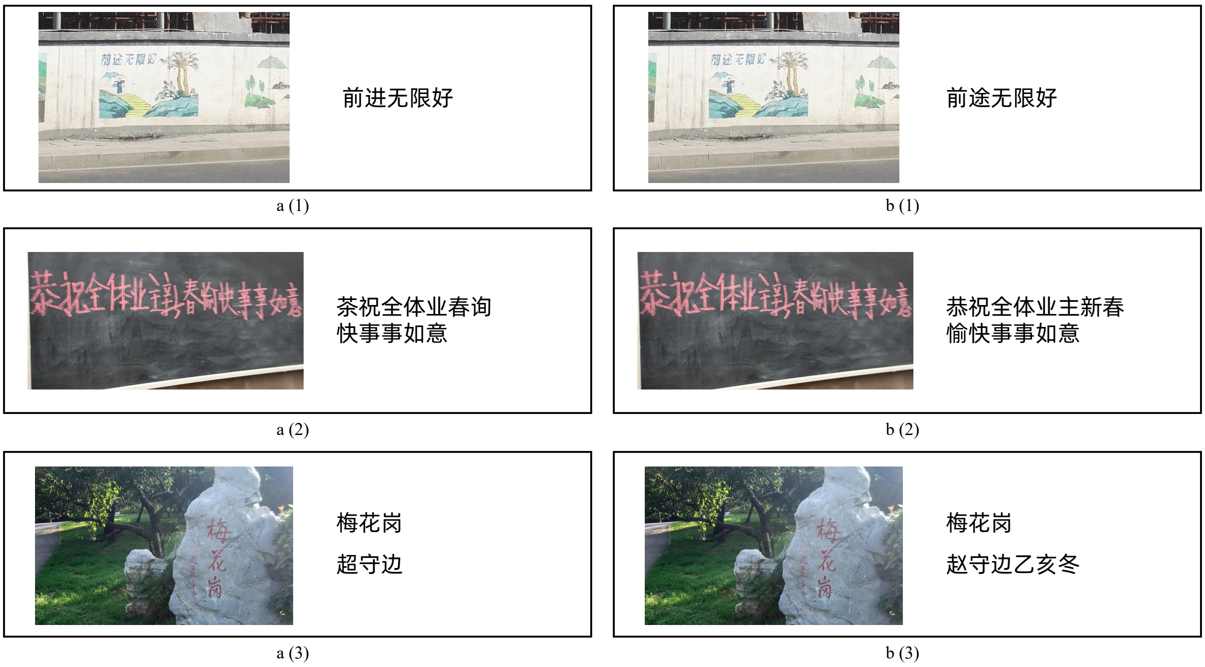

Figure 3: **A Case study of different OCR demos.** We compare a product-level API (a) with OFA-OCR (b). Through the case study, we find that OFA-OCR can reach a competitive performance.

```
def ocr_resize(img, resolution=480,
     is_document=False):
    img = img.convert("RGB")
    width, height = img.size

    if width >= height:
        new_width = max(64, resolution)
        new_height = max(64, int(resolution *
             (height / width)))
        top = random.randint(0, resolution -
             new_height)
        bottom = resolution - new_height - top
        left, right = 0, 0
    else:
        new_height = max(64, resolution)
        new_width = max(64, int(resolution *
             (width / height)))
        left = random.randint(0, resolution -
             new_width)
        right = resolution - new_width - left
        top, bottom = 0, 0

    img_new = F.resize(
        img,
        [new_height, new_width],
        interpolation=InterpolationMode.BICUBIC,
    )

    img_new = F.pad(img_new, padding=[left, top,
         right, bottom], padding_mode="edge")

    return img_new
```

A.4 Case Study

Here we evaluate the performance of the constructed simple OCR pipeline. For comparison, we use a product-level API[5] as the baseline. Figure 3 demonstrates the cases comparison. It can be found that on the 3 cases while the baseline makes mistakes by different extents, OFA-OCR makes the correct prediction of all characters, even if there are missing strokes or the text is in hard-to-recognize handwriting style.

[5] https://www.paddlepaddle.org.cn/modelsDetail?modelId=17

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
See Section Limitations after the conclusion

☑ A2. Did you discuss any potential risks of your work?
See Section Ethics Statement after the conclusion.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
See Abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

See Section 2 and Appendix.

☑ B1. Did you cite the creators of artifacts you used?
See Section 2.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
See Section 2 and Appendix.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
See Section 1, 2 and 5.

☒ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
We directly use the public benchmark datasets for the evaluation.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
See Section 2 and Appendix

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
See Section A.1

C ☑ Did you run computational experiments?

See Section 3.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
See Section 3 and Section A.3

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
See Section A.3

☒ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
See Section 3.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
See Section 3.5

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.