An Empirical Study of Scaling Law for Scene Text Recognition

Supplementary Material

A. More Experiment Analysis

A.1. The impact of model training details

Regarding how to train the optimal model, we also conduct relevant research on batch sizes and model depth used in the training process.

BatchSize we focus on examining the impact of various batch sizes on the accuracy of the PARSeq-B model. This investigation is integral to determining the model's optimal training conditions. The findings, as presented in Table 1, reveal that the model reaches its optimal performance, with an accuracy of 96.35%, when the batch size is configured to 1024. This result corroborates the conclusions from the CLIP4STR[33]. it is underscoring the significant role that larger batch sizes play in enhancing model accuracy. Notably, it is also observed that an excessively large batch size leads to a reduction in accuracy, indicating a critical balance in batch size selection for optimal model training.

Model	Backbone	Batch	Word Acc
PARSeq	ViT-B	1344	96.28
PARSeq	ViT-B	1024	**96.35**
PARSeq	ViT-B	896	96.33
PARSeq	ViT-B	448	96.3

Table 1. Average accuracy using different batch sizes on common benchmarks, training data is the real dataset (3.3M).

Depth PARSeq is divided into encoder and decoder. The encoder leverages the widely-recognized Vision Transformer (ViT) series, specifically employing the ViT-S variant. Conversely, the decoder is subject to more intricate fine-tuning, particularly concerning its depth. This aspect of the model architecture is a focal point of our research.

Our empirical investigations, as detailed in Table 2, centered on the interplay between the encoder's ViT-S configuration and varying depths of the decoder. The experimental findings are revealing. With the encoder consistently utilizing ViT-S, we observe that setting the decoder's depth to 1 layers resulted in optimal model accuracy. This suggests a significant relationship between decoder depth and model performance, underlining the importance of carefully calibrated model architecture in achieving high STR accuracy. Our results contribute to a deeper understanding of the architectural nuances in Transformer-based STR models and their impact on performance.

A.2. Benefits of pretraining in different languages

In this supplementary section, we conduct a thorough examination of the impact of language-specific pretraining on STR models, with a particular focus on fine-tuning for English datasets. Our approach involved utilizing models pretrained in Arabic, Latin, and a hybrid of Chinese-English, each trained on a dataset comprising 300,000 entries drawn from private sources. The core architecture for these models is based on the CMT-S framework, as detailed in Guo et al. (2022) [9]. Subsequent secondary training is conducted on the REB dataset, a subset of REBU-Syn, wherein different language-specific pretrained models are employed. Notably, the final classification layer's parameters are not loaded from these pretrained models to ensure a fair comparison.

Model	Encoder	Decoder-Depth	Word Acc
PARSeq	ViT-S	1	**95.56**
PARSeq	ViT-S	2	95.31
PARSeq	ViT-S	3	94.77
PARSeq	ViT-S	4	94.50
PARSeq	ViT-S	5	93.77

Table 2. Average accuracy using different depth for decoder on benchmark test set, training model in real dataset.

As illustrated in Table 4, our results reveal pronounced improvements in models pretrained in Latin, Chinese, and English, with Latin demonstrating the most substantial enhancement. This improvement is likely due to the visual congruence between Latin and English scripts, emphasizing the STR models' dependency on visual features for effective recognition. Meanwhile, the performance of models pretrained in Chinese and English, though slightly lower by a margin of 0.01% compared to the Latin model, indicates a potential bias introduced by the inclusion of Chinese data in the pretraining phase.

Intriguingly, models pretrained in Arabic does not exhibit significant benefits over their non-pretrained counterparts. This can be attributed to the stark visual differences between Arabic and English scripts, reinforcing the notion that visual similarity plays a crucial role in the efficacy of pretraining for STR tasks. Collectively, these findings suggest that pretraining STR models with languages visually akin to the target language offers enhanced benefits. Conversely, a pronounced visual dissimilarity between the scripts negates the advantages of pretraining, a critical consideration for the training models.

B. Comparisons on Union14M benchmark

To evaluate the generalization capabilities of our model, we conducted an extensive assessment using the Union14M benchmark dataset [12]. This benchmark is particularly

Method	Training data	Artistic	Contextless	Curve	General	Multi-Oriented	Multi-Words	Salient	Avg
CRNN [24]	MJ+ST	20.7	25.6	7.5	32.0	0.9	25.6	13.9	18.0
SVTR [7]	MJ+ST	37.9	44.2	63.0	52.8	32.1	49.1	67.5	49.5
MORAN [19]	MJ+ST	29.4	20.7	8.9	35.2	0.7	23.8	17.9	19.5
ASTER [25]	MJ+ST	27.7	33.0	34.0	39.8	10.2	27.6	48.2	31.5
NRTR [23]	MJ+ST	36.6	37.3	31.7	48.0	4.4	54.9	30.6	34.8
SAR [16]	MJ+ST	42.6	44.2	44.3	50.5	7.7	51.2	44.0	40.6
DAN [28]	MJ+ST	35.0	40.3	26.7	42.1	1.5	42.2	36.5	32.0
SATRN [14]	MJ+ST	48.0	45.3	51.1	58.5	15.8	52.5	62.7	47.7
RobustScanner [32]	MJ+ST	41.2	42.6	43.6	39.5	7.9	46.9	44.9	38.1
SRN [31]	MJ+ST	34.1	28.7	63.4	46.3	25.3	26.7	56.5	40.1
ABINet [8]	MJ+ST	43.3	38.3	59.5	55.6	12.7	50.8	62.0	46.0
VisionLAN [29]	MJ+ST	47.8	48.0	57.7	52.1	14.2	47.9	64.0	47.4
MATRN [22]	MJ+ST	43.8	41.9	63.1	57.0	13.4	53.2	66.4	48.4
CRNN [24]	Union14M	31.9	39.3	18.9	58.1	4.3	21.5	15.1	27.0
SVTR [7]	Union14M	50.2	63.0	70.5	74.7	66.6	42.6	71.4	62.7
MORAN [19]	Union14M	44.3	51.1	42.4	42.9	12.4	36.8	41.0	38.7
ASTER [25]	Union14M	39.2	47.9	37.4	64.4	12.5	34.5	30.2	38.0
NRTR [23]	Union14M	51.8	65.1	47.9	72.9	39.1	51.4	40.1	52.6
SAR [16]	Union14M	58.0	69.0	66.9	73.7	54.7	51.2	57.0	61.5
DAN [28]	Union14M	47.0	56.6	44.6	66.7	22.1	39.8	41.5	45.5
SATRN [14]	Union14M	64.3	71.1	73.0	78.8	64.7	47.4	69.2	66.9
RobustScanner [32]	Union14M	58.7	72.7	64.2	73.5	52.8	47.8	56.9	60.9
SRN [31]	Union14M	47.6	57.9	48.7	60.7	20.0	27.9	41.6	43.5
ABINet [8]	Union14M	62.2	66.3	73.0	75.6	59.6	43.1	69.5	64.2
VisionLAN [29]	Union14M	54.4	60.1	68.8	72.1	55.2	37.9	64.7	59.0
MATRN [22]	Union14M	67.3	71.0	79.3	78.4	66.0	53.8	74.9	70.0
MAERec-S [12]	Union14M-L	68.9	77.8	79.3	80.4	69.5	51.9	75.1	71.8
MAERec-B [12]	Union14M-L	75.9	80.7	86.6	83.8	82.1	56.2	82.2	78.2
PARSeq-S [3]	R	81.7	86.5	91.1	86.5	89.3	85.3	84.6	86.5
CLIP4STR-B [33]	R	86.5	**92.2**	96.3	89.9	96.1	88.9	91.2	91.6
CLIP4STR-L [33]	R	87.2	91.0	**97.0**	**90.3**	96.6	89.9	91.5	91.9
PARSeq-S*	REBU-Syn	85.2	89.4	94.0	88.0	93.1	89.9	89.8	89.9
CLIP4STR-B*	REBU-Syn	88.6	90.1	96.4	89.1	96.3	**92.2**	91.9	92.1
CLIP4STR-L*	REBU-Syn	**88.6**	90.4	96.4	89.3	**97.2**	90.7	**92.7**	**92.2**

Table 3. Word accuracy on Union14M benchmark, * indicates training with REBU-Syn.

Pretrain	Model	Datasets	Word Acc
From Scratch	PARSeq	REB	95.60
Arabic	PARSeq	REB	95.62
Cn-En	PARSeq	REB	95.81
Latin	PARSeq	REB	**95.82**

Table 4. Average accuracy using language-specific pretraining on benchmark test set, training model in real dataset of REB.

comprehensive, encompassing a vast array of real-world textual data, systematically categorized into seven distinct subsets: Artistic, Contextless, Curve, General, Multi-Oriented, Multi-Words and Salient. The results of this evaluation, presented in Table 3, demonstrate the model's robust and consistent performance across a range of scenarios. Notably, in comparative evaluations against standard benchmarks and the multifaceted Union14M dataset, the CLIP4STR-L* model emerges as a standout performer. This model demonstrates exceptional accuracy across the majority of datasets. Its ability to consistently deliver high-quality results, particularly in the context of the challenging Union14M benchmark, underscores its robustness and versatility. Such performance highlights the efficacy of the CLIP4STR-L* architecture in handling a diverse range of textual data scenarios, making it a benchmark in the field.

C. Visulization Analysis

In Fig 1, we present a visualization of our model's performance across the seven major categories of the Union14M benchmark. The results demonstrate that our model outperforms in the majority of datasets. However, a slight dip in effectiveness is noted in the Contextless dataset. This can be attributed to the limitations of the text encoder in processing texts lacking semantic information.

Despite this, our model distinguishes itself from other contemporary STR models through its enhanced ability to accurately interpret and navigate a diverse range of complex real-world scenarios. This advancement significantly bolsters the robustness of STR models, enabling them to operate with greater reliability in varied and challenging environments. The enhanced robustness of our model not only

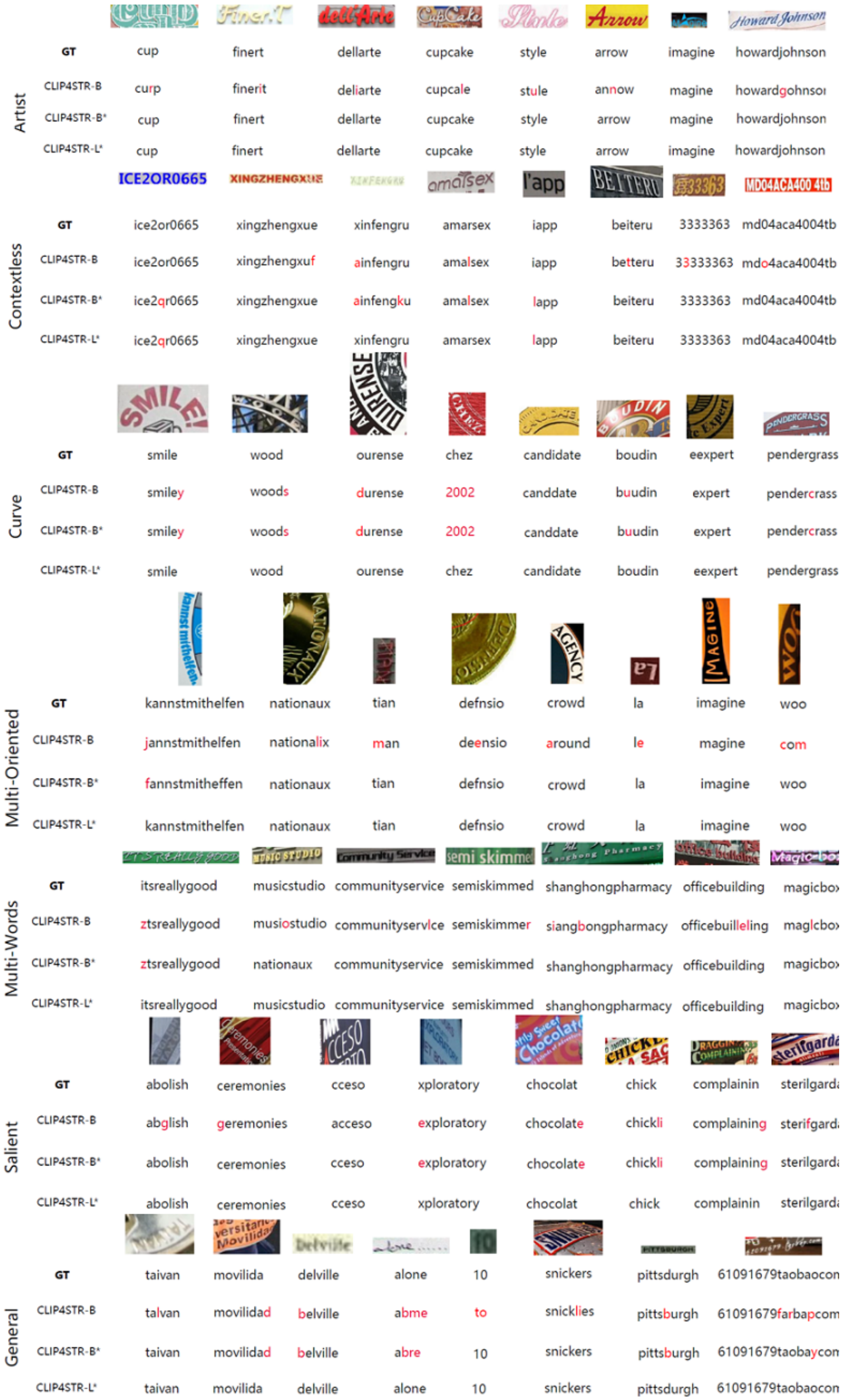

Figure 1. Error analysis of the Union14M benchmark. We select three representative models and show their prediction results (Text in black represents correct prediction and red text vice versa).

showcases its technical excellence but also emphasizes its practical applicability in real-world settings characterized by high variability and complexity.

D. STR Enhanced LMM

In the realm of large-scale models, we observe a distinct bifurcation into two primary categories: Large Language Models (LLMs) and Large Multimodal Models (LMMs). It is crucial to acknowledge that while LLMs are devoid of a visual component, LMMs' visual branches demonstrate room for enhancement in terms of text recognition capabilities[26]. This observation underscores the relative underdevelopment of text recogition proficiency within large-scale models. Scene Text Recognition (STR) tasks, however, offer a promising avenue to address this shortfall, thereby motivating our investigation into the benefits of integrating STR models with these models.

Dataset and Metric Our analysis utilized a diverse range of tasks from the Visual Question Answering (VQA) series, specifically STVQA [4], TextVQA [27], DocVQA [20] and InfoVQA [21]. While STVQA and TextVQA are geared towards natural scenes, DocVQA and InfoVQA focus on general document contexts. Here are some details of evaluation dataset:

- **STVQA** contains 31K questions that require understanding the scene text, based on 23K images from : ICDAR2013 and ICDAR2015, ImageNet [6], VizWiz [10], IIIT Scene Text Retrieval, Visual Genome [13] and COCO-Text.
- **TextVQA** contains 45K questions that need to read and reason the text in images, based on 28K images from natural images.
- **DocVQA** contains 50K questions and 12K images from industry documents.
- **InfoVQA** contains 30K questions that require understanding the document text, based on 5.4K images combining textual, graphical and visual elements from Infographics.

We employed the Average Normalized Levenshtein Similarity (ANLS) as our evaluation metric, a standard in the VQA domain.

Experiment Setting For the large-scale model, we selected the recently unveiled Qwen-VL-chat [2], a state-of-the-art multimodal model. In terms of STR, we utilized Rosetta [5] for detection, and CLIP4STR-L* for recognition. We began by concatenating the text recognized through coordinate information to generate STR tokens. These tokens, combined with the question, formed our prompts. The prompt format was meticulously refined to: 'STR token: $\{ocrtokens\}$, please answer the following questions based on STR tokens and pictures, $\{question\}$'. This approach involved inputting both the prompt and images into the large-scale model.

Model	STVQA	TextVQA	DocVQA	InfoVQA
BLIP-2 [1]	21.7	32.2	4.9	-
LLaVAR [18]	39.2	48.5	11.6	-
InstructBLIP [17]	-	50.7	-	38.3
LLaMA-7B [30]	-	52.6	62.2	38.2
Pix2Struct-base [15]	-	-	72.1	38.2
Qwen-VL-Chat	50.25	61.5	63.41	31.7
Qwen-VL-Chat with STR	**70.32**	**69.64**	**73.44**	**38.48**

Table 5. Result on benchmarks for VQA tasks using LMM models with or without STR, all result are ANLS on the val split.

Result and Analyze We performed a detailed comparative analysis to assess the accuracy of the QWen-VL-chat model, examining its performance with and without STR integration, as delineated in Table 5. Our results reveal a significant improvement in the accuracy of the model for scene-based VQA tasks upon the integration of STR. Additionally, there is a noticeable enhancement in document-based VQA tasks. These findings suggest that the incorporation of STR not only enhances the model's accuracy but also extends its generalization capabilities across diverse VQA scenarios. This evidence distinctly highlights the vital role that STR inputs play in augmenting the performance of LVLM for downstream tasks. Furthermore, the improved accuracy with STR integration underscores the model's enhanced ability to interpret and analyze combined visual and textual data, thereby validating the efficacy of multimodal approaches in tackling complex analytical challenges.

VQA Visulization Analysis Our visual analysis of QWen-VL-Chat, with and without the STR module, across varied datasets offers critical insights. In natural scene Visual Question Answering (VQA) tasks, QWen-VL-Chat encounters difficulties in detecting small text in images. The upper left corner of Fig 2, the model overlooks pertinent content, erroneously indicating its absence. Moreover, its tendency to inaccurately complete blurred text stems from its sophisticated linguistic abilities. This is evident in the upper right corner of Fig 2, where 'dr' in '45th parallel dr' is incorrectly expanded to 'drive'. Notably, the model's text misidentification issues, such as converting 'honghe' to 'Hongte' on a cigarette pack as depicted in the lower left corner of Fig 2 (mistaking the second 'h' for a 't'), are significantly mitigated by integrating our STR module.

In general document scenarios involving dense textual information, the performance of QWen-VL-Chat remains suboptimal.In the left of Fig 3, when tasked with identifying brands in advertisements amidst extensive text, the model erroneously generates non-existent words from the image. Incorporating STR crucially directs the model towards accurate text recognition. This pattern is consistent in table-based VQA Tasks in the right of Fig 3, where the model frequently errs in its responses. The inclusion of STR proves instrumental in steering the model towards correct answers. This comprehensive analysis highlights the pivotal role of STR in augmenting LMM models' comprehension

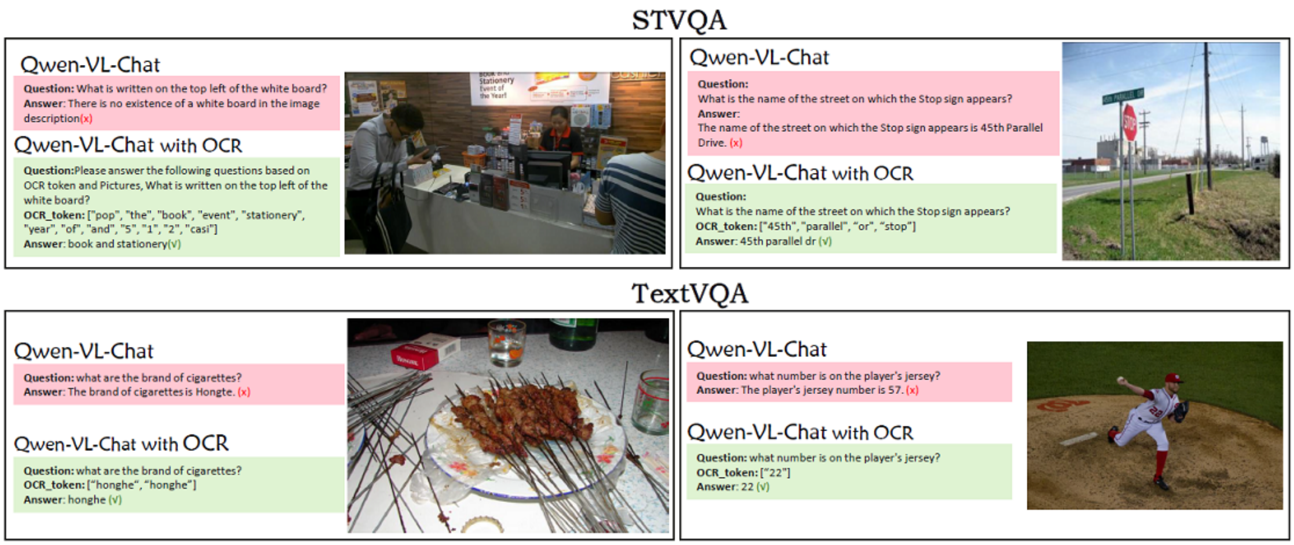

Figure 2. Visual anwer comparison for QWen-VL-Chat with or without STR tokens in natural scenes VQA.

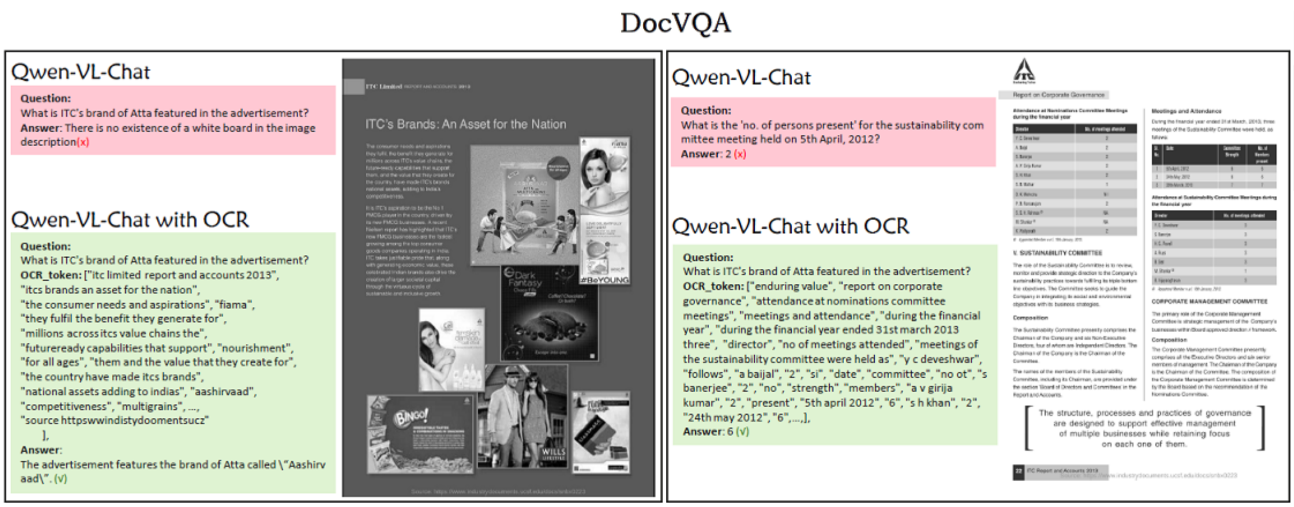

Figure 3. Visual anwer comparison for QWen-VL-Chat with or without STR tokens in Document VQA.

and recognition capabilities within intricate visual-textual contexts.

E. Scaling law algorithm description

We formalize the power law of performance in terms of scaling factors, and the implemented details are shown in Algorithm 1.

F. The scaling law on Union14M benchmark

We supplement the experiments with scaling laws on the Union14M benchmark. The parameters and accuracy of PARSeq-(S/B/L) and CLIP4STR-(S/B/L) on the Union14M benchmark are shown in Table 6 and Table 7 respectively. The curves of scaling law on CLIP4STR and PARSeq models are shown in Fig 4. It demonstrates that the scaling law is still applicable on the Union14M benchmark.

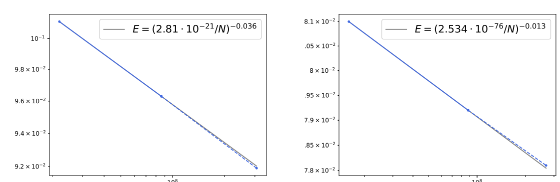

Figure 4. **Left**: PARSeq-(S/B/L) results on Union14M. **Right**: CLIP4STR-(S/B/L) results on Union14M.

G. Applicability in document contexts

We also validate the power law using scaling model sizes on Moreover, apart from applying the STR benchmark, we

Algorithm 1: the power-law function

input : x-axis data for data volume, model size or compute time X, word error rate E.

output: a_0, a_1 are the coefficients of the power law function $E(\cdot) = (a_0 * X)^{a_1}$, v is used to determine whether the power law holds.

$X' \leftarrow logX$, $E' \leftarrow logE$;
define $LineFunc(X', E') = k * X' + b$;
for $i \leftarrow 1$ to $t - 1$ **do**
  Use the first t-1 points to fit the straight line equation $LineFunc(X', E')$ and obtain the coefficients, k and b.
end
// Replace (X', E') in the straight line formula $LineFunc$ with (X, E) to obtain the coefficients (a_0, a_1) of the power law function $E(\cdot) = (a_0 * X)^{a_1}$.
$(a_0, a_1) \leftarrow logE = k * logX + b$
// Verify that (X_t, E_t) is on the equation of the power law function $E(\cdot) = (a_0 * X)^{a_1}$.
$E_t^{pred} \leftarrow (a_0 * X_t)^{a_1}$;
$dev \leftarrow E_t^{pred} - E_t$;
if $dev < 0.1$ **then** $v \leftarrow 1$;
else $v \leftarrow 0$;

Method	Param (M)	Avg
PARSeq-S	22.5	89.89
PARSeq-B	104.0	90.37
PARSeq-L	335.9	**90.81**

Table 6. Word accuracy with different model sizes of PARSeq. Test data: Union14M.

Method	Param (M)	Avg
CLIP4STR-S	43.6	91.90
CLIP4STR-B	86.7	92.08
CLIP4STR-L	268.2	**92.19**

Table 7. Word accuracy with different model sizes of CLIP4STR. Test data: Union14M.

further extend the application of the scaling law to a document dataset in order to authenticate its validity and reliability. The FUNSD [11] dataset contains a large number of scanned documents, and each sample is annotated with detailed text, word bounding boxes, and structured tags. It is intended to support the development and assessment of model performance by researchers for the purpose of processing and comprehending information from scanned documents in noisy, real-world. Notably, CLIP4STR-L* achieved a SOTA accuracy of 96.5%, surpassing the previous best, CLIP4STR-L. The experimental results are shown in Table 8. These results highlight the robustness of CLIP4STR-L* in both scene and document text recognition tasks.

Model	Word Acc
CLIP4STR-L	96.02
CLIP4STR-L*	**96.50**

Table 8. Accuracy for CLIP4STR-L on FUNSD. * indicates training with REBU-Syn.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35:23716–23736, 2022. 4

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966, 2023. 4

[3] Darwin Bautista and Rowel Atienza. Scene text recognition with permuted autoregressive sequence models. 2022. 2

[4] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. In Proceedings of the IEEE/CVF international conference on computer vision, pages 4291–4301, 2019. 4

[5] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, pages 71–79, 2018. 4

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. 4

[7] Yongkun Du, Zhineng Chen, Caiyan Jia, Xiaoting Yin, Tianlun Zheng, Chenxia Li, Yuning Du, and Yu-Gang Jiang. Svtr: Scene text recognition with a single visual model. arXiv preprint arXiv:2205.00159, 2022. 2

[8] Shancheng Fang, Hongtao Xie, Yuxin Wang, Zhendong Mao, and Yongdong Zhang. Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition. 2021. 2

[9] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. Cmt: Convolutional neural networks meet vision transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12175–12185, 2022. 1

[10] 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, 2018. 4

[11] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. Funsd: A dataset for form understanding in noisy scanned documents, 2019. 6

[12] Qing Jiang, Jiapeng Wang, Dezhi Peng, Chongyu Liu, and Lianwen Jin. Revisiting scene text recognition: A data perspective. 2023. 1, 2

[13] 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 Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016. 4

[14] Junyeop Lee, Sungrae Park, Jeonghun Baek, Seong Joon Oh, Seonghyeon Kim, and Hwalsuk Lee. On recognizing texts of arbitrary shapes with 2d self-attention. 2019. 2

[15] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In International Conference on Machine Learning, pages 18893–18912. PMLR, 2023. 4

[16] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. 2019. 2

[17] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. arXiv preprint arXiv:2310.03744, 2023. 4

[18] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. arXiv preprint arXiv:2304.08485, 2023. 4

[19] Canjie Luo, Lianwen Jin, and Zenghui Sun. A multi-object rectified attention network for scene text recognition. 2019. 2

[20] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In Proceedings of the IEEE/CVF winter conference on applications of computer vision, pages 2200–2209, 2021. 4

[21] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 1697–1706, 2022. 4

[22] Byeonghu Na, Yoonsik Kim, and Sungrae Park. Multimodal text recognition networks: Interactive enhancements between visual and semantic features. 2022. 2

[23] Fenfen Sheng, Zhineng Chen, and Bo Xu. Nrtr: A no-recurrence sequence-to-sequence model for scene text recognition. In 2019 International conference on document analysis and recognition (ICDAR), pages 781–786. IEEE, 2019. 2

[24] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. 2015. 2

[25] Baoguang Shi, Mingkun Yang, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Aster: An attentional scene text recognizer with flexible rectification. IEEE transactions on pattern analysis and machine intelligence, 41(9):2035–2048, 2018. 2

[26] Yongxin Shi, Dezhi Peng, Wenhui Liao, Zening Lin, Xinhong Chen, Chongyu Liu, Yuyi Zhang, and Lianwen Jin. Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation. arXiv preprint arXiv:2310.16809, 2023. 4

[27] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 8317–8326, 2019. 4

[28] Tianwei Wang, Yuanzhi Zhu, Lianwen Jin, Canjie Luo, Xiaoxue Chen, Yaqiang Wu, Qianying Wang, and Mingxiang Cai. Decoupled attention network for text recognition. 2019. 2

[29] Yuxin Wang, Hongtao Xie, Shancheng Fang, Jing Wang, Shenggao Zhu, and Yongdong Zhang. From two to one: A new scene text recognizer with visual language modeling network. 2021. 2

[30] Jiabo Ye, Anwen Hu, Haiyang Xu, Qinghao Ye, Ming Yan, Yuhao Dan, Chenlin Zhao, Guohai Xu, Chenliang Li, Junfeng Tian, et al. mplug-docowl: Modularized multimodal large language model for document understanding. arXiv preprint arXiv:2307.02499, 2023. 4

[31] Deli Yu, Xuan Li, Chengquan Zhang, Junyu Han, Jingtuo Liu, and Errui Ding. Towards accurate scene text recognition with semantic reasoning networks. CoRR, abs/2003.12294, 2020. 2

[32] Xiaoyu Yue, Zhanghui Kuang, Chenhao Lin, Hongbin Sun, and Wayne Zhang. Robustscanner: Dynamically enhancing positional clues for robust text recognition. 2020. 2

[33] Shuai Zhao, Xiaohan Wang, Linchao Zhu, Ruijie Quan, and Yi Yang. Clip4str: A simple baseline for scene text recognition with pre-trained vision-language model. 2023. 1, 2