Multi-modal In-Context Learning Makes an Ego-evolving Scene Text Recognizer

Supplementary Material

1. Model Architecture

Figure 1 presents the detailed model architecture of E^{2}STR. We follow the paradigm established by Flamingo [1], where we perform cross attention between the vision outputs and the language outputs in each language model layer. The language outputs serve as queries and the vision outputs serve as keys and values. The detailed configures of the vision encoder and the language decoder are summarized in Table 1. For fair comparison, we provide MAERec [2] with the same language decoder with E^{2}STR-ICL (We name this modification as MAERec†). The comparison between MAERec† and E^{2}STR is shown in Table 2.

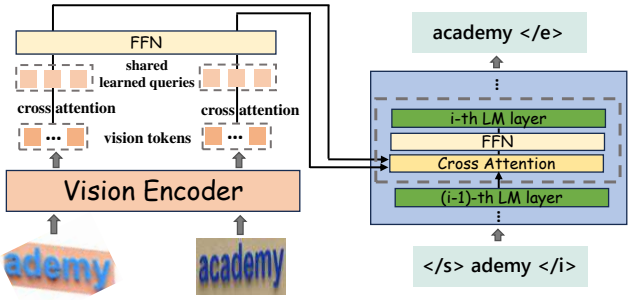

Figure 1. Detailed Model Architecture of E^{2}STR.

	Input Size	Patch Size	Embedding	Depth	Heads	Parameters
Vision Encoder	32x128	4x4	768	12	12	85M
Language Decoder	-	-	768	12	12	125M

Table 1. Model details of E^{2}STR.

	MPSC	EIST	IAM
MAERec	81.81	70.33	70.27
MAERec†	82.00	70.77	70.51
E^{2}STR-ICL	83.64	76.77	74.10

Table 2. Word Accuracy performance comparison between MAERec [2] and E^{2}STR-ICL. MAERec† refers to MAERec using the same vision encoder and the same language decoder with E^{2}STR-ICL.

2. Model Scalability

Table 3 presents the inference time change brought by the different number of in-context prompts. It is easy to find that the number of in-context prompts in E^{2}STR is scalable. For example, the inference time of E^{2}STR-ICL (where we select two prompts) is 0.094s. But When expanding the number of in-context prompts by 7 times (*i.e.*, 16 prompts), the inference time is only increased by 1.08 times (*i.e.*, 0.196s).

Prompts	0	1	2	4	8	16
Inference Time (s)	0.071	0.085	0.094	0.113	0.140	0.196

Table 3. Inference time change brought by the different number of in-context prompts.

Table 4 presents the inference time change brought by different sizes of the in-context pool. As we can see, when expanding the pool size by 4 times (*i.e.*, from 100 to 500), the inference time is only increased by 0.07 times (*i.e.*, from 0.094 to 0.101). As a result, our E^{2}STR-ICL is highly scalable in terms of both in-context pool size and the number of in-context prompts.

Pool Size	100	200	300	400	500
Inference Time (s)	0.094	0.096	0.097	0.099	0.101

Table 4. Inference time change brought by different sizes of the in-context pool.

	Prompt Domain			
	Non-context	MPSC	EIST	IAM
MPSC	81.26	**83.64**	83.00	82.96
EIST	69.66	70.30	**76.77**	70.00
IAM	69.51	72.17	71.70	**74.10**

Table 5. Performance change brought by the domain variation of the in-context pool. **Bold** values denote the best performance in a row.

3. Model Stability

Table 5 presents how the performance change when varying the domains of the in-context pool. As we can see, our E^{2}STR-ICL is stable to the change of the context prompts. On all three benchmarks, out-of-domain in-context pools still improve the performance, though the improvement is lower than in-domain in-context pools. Nevertheless, there still exists a very slim chance that E^{2}STR-ICL erroneously rectifies predictions due to misleading prompts. Shown in Figure 2, when certain areas of the prompt image is highly

	Training GPU Hours		MPSC	EIST	IAM	AVG
kNN	**415.6**	base	81.22	69.78	69.62	73.54
		ICL	82.06	70.95	71.00	**74.67**
ST	**131.2**	base	81.26	69.66	69.51	73.48
		ICL	83.64	76.77	74.10	**78.17**

Table 6. Comparisons between kNN and our ST-strategy.

similar to the test image but the ground-truth is different, E^{2}STR-ICL may erroneously rectifies the prediction.

Figure 2. Examples of erroneous rectification brought by misleading prompts.

4. Visualization

We provide more examples of the cross attention visualization in Figure 3.

Figure 3. More examples of the cross attention visualization.

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. 1

[2] Qing Jiang, Jiapeng Wang, Dezhi Peng, Chongyu Liu, and Lianwen Jin. Revisiting scene text recognition: A data perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20543–20554, 2023. 1