Reading Chinese in Natural Scenes with a Bag-of-Radicals Prior

Yongbin Liu, Qingjie Liu, Jiaxin Chen, Yunhong Wang

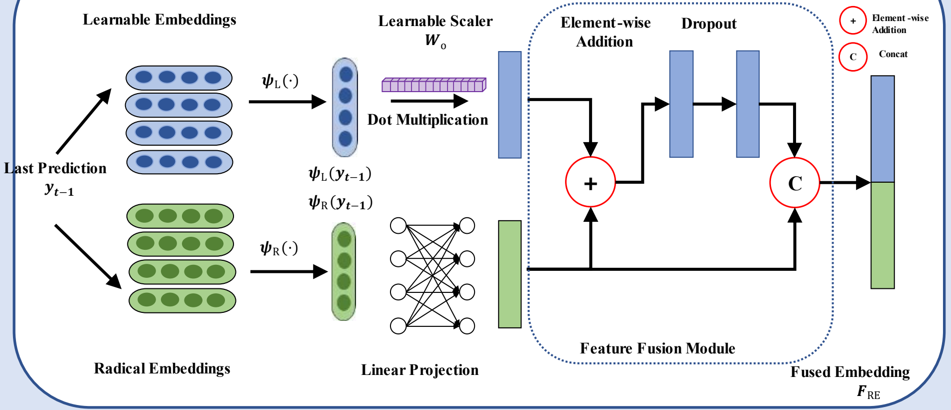
Abstract

Scene text recognition (STR) on Latin datasets has been extensively studied in recent years, and state-of-the-art (SOTA) models often reach high accuracy. However, the performance on non-Latin transcripts, such as Chinese, is not satisfactory. In this paper, we collect six open-source Chinese STR datasets and evaluate a series of classic methods performing well on Latin datasets, finding a significant performance drop. To improve the performance on Chinese datasets, we propose a novel radical-embedding (RE) representation to utilize the ideographic descriptions of Chinese characters. The ideographic descriptions of Chinese characters are firstly converted to bags of radicals and then fused with learnable character embeddings by a character-vector-fusion-module (CVFM). In addition, we utilize a bag of radicals as supervision signals for multi-task training to improve the ideographic structure perception of our model. Experiments show performance of the model with RE + CVFM + multi-task training is superior compared with the baseline on six Chinese STR datasets. In addition, we utilize a bag of radicals as supervision signals for multi-task training to improve the ideographic structure perception of our model. Experiments show performance of the model with RE + CVFM + multi-task training is superior compared with the baseline on six Chinese STR datasets.

Contributions

1. We evaluate 10 well-known STR models on six large-scale Chinese STR datasets. This result provides a fair comparison in Chinese STR field.
2. We find a performance gap between the Chinese recognition and the Latin recognition. And even worse, the models performing very well on Latin datasets may have a low performance on Chinese datasets. It indicates that simply deploying a Latin STR method in Chinese scenes may cause a performance drop.
3. A novel Chinese STR method that fuses the bags of radicals with CVFM is proposed, and we design a multi-task radical classification branch to improve the character structural perception. Experiments demonstrate that our method is superior to the baseline model.

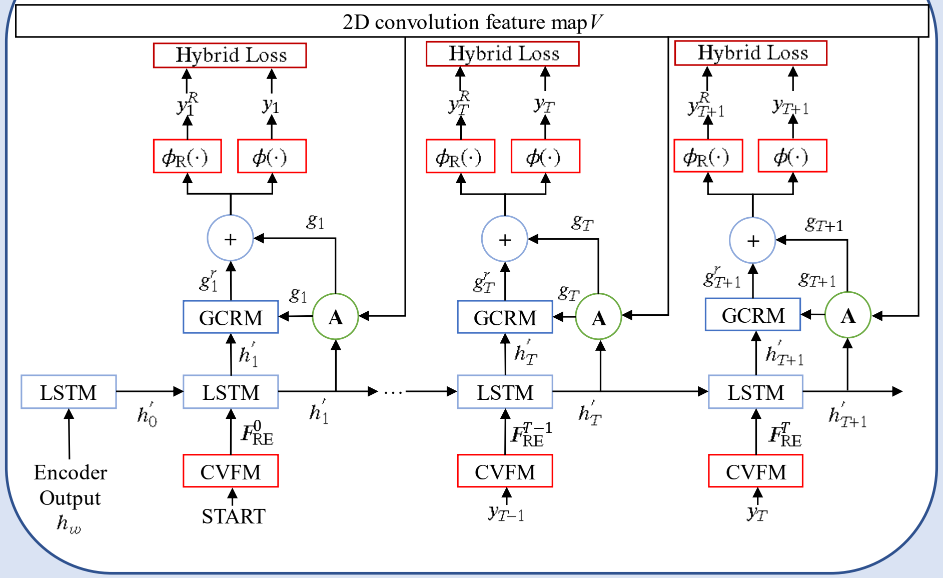
CVFM



Method

1. Radical decomposition: split each Chinese character to bag-of-radicals.
2. Convert bag-of-radicals to radical vectors.
3. Use CVFM to fuse radical vectors and contextual features.
4. Use radical classification branch to improve the structural awareness.

The Decoder Pipeline



	ArT	CASIA	ctw	LSVT	RCTW	ReCTS	Avg.
GRCNN	42.9	42.9	31.1	37.7	44.6	51.9	41.8
R2AM	63.9	51.0	53.2	57.9	55.7	66.2	57.4
CRNN	52.4	46.7	38.5	46.8	47.9	60.1	48.7
Rosetta	65.1	60.6	54.8	62.6	63.5	74.5	63.7
RARE	74.7	61.9	62.8	67.2	66.5	75.1	67.2
STARNET	51.5	53.1	49.5	53.7	57.4	69.4	56.6
TRBA	77.7	62.6	63.8	68.5	67.2	77.1	68.5
RobScan	85.2	67.0	66.8	71.8	70.4	79.7	72.0
SAR	74.2	66.7	65.2	70.9	69.7	79.8	71.3
GCAN	74.7	68.1	65.2	71.6	70.7	80.5	71.8
Ours	**79.2**	**69.3**	**67.5**	**73.6**	**72.5**	**81.5**	**73.5**

Ablation

	ArT	CASIA	ctw	LSVT	RCTW	ReCTS	Avg.
GCAN	74.7	68.1	65.2	71.6	70.7	80.5	71.8
+CVFM	78.1	69.0	67.6	73.3	72.0	81.2	73.2
+BCE	**79.2**	**69.3**	**67.5**	**73.6**	**72.5**	**81.5**	**73.5**

Loss Function

$$\hat{\mathbf{y}}_T = \phi(\mathbf{g}_T)$$
$$\hat{\mathbf{y}}_T^R = \phi_R(\mathbf{g}_T)$$
$$\mathcal{L}_O = \mathrm{CrossEntropy}(\hat{\mathbf{y}}_T, \mathbf{y}_T)$$
$$\mathcal{L}_R = \mathrm{BCE}(\hat{\mathbf{y}}_T^R, \mathbf{y}_T^R)$$
$$\mathcal{L} = \mathcal{L}_O + \mathcal{L}_R$$

Conclusions

In this work we collect six Chinese scene text datasets to evaluate the previous models performing well on Latin datasets fairly. From the experimental results we can conclude that the 2D-attention models reach high word accuracy on both Latin datasets and Chinese datasets.

In order to further improve the performance of Chinese STR, we propose RE, CVFM and hybrid radical losses for multi-task training. The experiments on the Chinese STR benchmark show that our method is superior compared with our baseline (GCAN) and other methods on six datasets.

References

[1] Jeonghun Baek, Geewook Kim, et al. What is wrong with scene text recognition model comparisons? dataset and model analysis. In ICCV, pages 4715–4723, 2019.
[2] Zhi Qiao, Xugong Qin, et al. Gaussian constrained attention network for scene text recognition. In ICPR, pages 3328–3335. IEEE, 2021.
[3] Hui Li, Peng Wang, et al. Show, attend and read: A simple and strong baseline for irregular text recognition. In AAAI, volume 33, pages 8610–8617, 2019.