A Transformer-based Radical Analysis Attention Network for Chinese Character Recognition

Chen Yang, Qing Wang, Jun Du, Jianshu Zhang, Changjie Wu, Jiaming Wang
National Engineering Laboratory for Speech and Language Information Processing
University of Science and Technology of China
Hefei, Anhui, China
yc24@mail.ustc.edu.cn, xiaosong@mail.ustc.edu.cn, jundu@ustc.edu.cn, xysszjs@mail.ustc.edu.cn, wucj@mail.ustc.edu.cn, jmwang66@mail.ustc.edu.cn

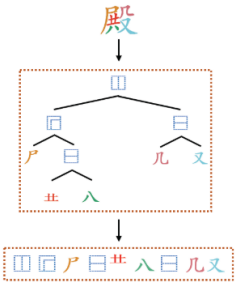
1. Background & Motivation

The recent studies of Chinese character recognition (CCR) can be roughly divided into two categories: **character-based CCR** and **radical-based CCR**.

- Character-based methods can perform well on common Chinese characters with a lot of training data. But they have difficulty in dealing with zero-shot learning problem and perform poorly when handling complex characters.
- Radical-based methods has the ability to recognize unseen Chinese characters, However, to recognize more complicated radical structures or learn the composition rules of low-frequency samples.

Therefore, it is important to build a more powerful recognition system for CCR.

- Self-attention mechanism can capture long-range dependencies and the detailed internal pattern
- The Transformer is composed of stacked blocks and aggregates the input context for each block, which naturally provides us with more hierarchical representations.



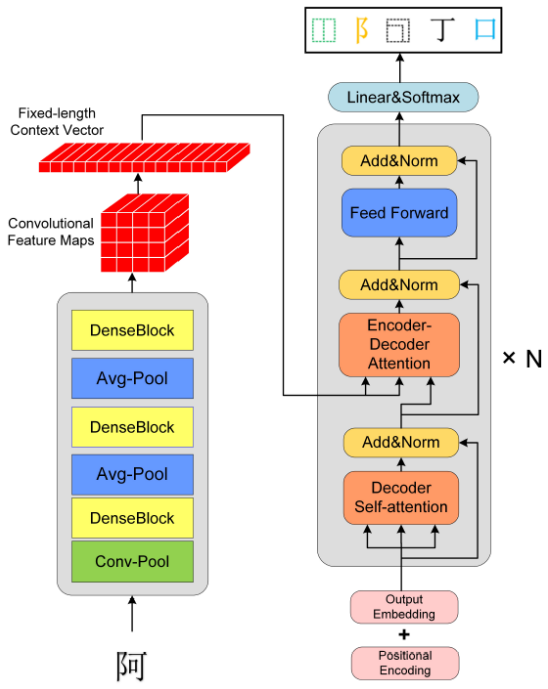
★ Hierarchical radical structure of an example Chinese character

2. The Proposed Method

Architectures of the **Transformer-based radical analysis network (RTN)** contains two components:

(1) a dense encoder which takes the image as input to produce a fixed-length context vector;

(2) a transformer decoder which takes the context vector as input to generate a variable-length symbol sequence.



➢ A dense encoder

$$\mathbf{A} = \{\mathbf{a}_1, \ldots, \mathbf{a}_L\}, \mathbf{a}_l \in \mathbb{R}^D$$

➢ A transformer decoder

$$\mathbf{s}_i^n = \mathrm{MA}(\mathbf{c}_i^{n-1}, \mathbf{C}_{<i}^{n-1}, \mathbf{C}_{<i}^{n-1})$$

$$\mathbf{z}_i^n = \mathrm{MA}(\mathbf{s}_i^n, \mathbf{A}, \mathbf{A})$$

$$\mathbf{c}_i^n = \max(0, \mathbf{W}_1^n \mathbf{z}_i^n + \mathbf{b}_1^n)\,\mathbf{W}_2^n + \mathbf{b}_1^n$$

➢ Training Objective

$$\mathrm{P}(\mathbf{Y}|\mathbf{X}) = \prod_{i=1}^{|\mathbf{Y}|} \mathrm{P}(\mathbf{y}_i|\mathbf{Y}_{<i}, \mathbf{X}) = \prod_{i=1}^{|\mathbf{Y}|} \mathrm{P}(\mathbf{y}_i|\mathbf{c}_i^N)$$

What is self-attention?

$$\mathrm{Attn}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \mathrm{softmax}\left(\frac{\mathbf{q}\mathbf{K}^T}{\sqrt{\mathrm{d_k}}}\right)\mathbf{V}$$

$$\mathrm{head_j} = \mathrm{Attn}\left(\mathbf{q}\mathbf{W}_j^{\mathrm{Q}}, \mathbf{K}\mathbf{W}_j^{\mathrm{K}}, \mathbf{V}\mathbf{W}_j^{\mathrm{V}}\right)$$

$$\mathrm{MA}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = [\mathrm{head}_1, \mathrm{head}_2, \ldots, \mathrm{head}_j]\mathbf{W}^{\mathrm{O}}$$

3. Experiments and Results

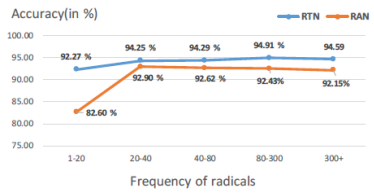
Our experiments are conducted on both printed Chinese character dataset and natural scene Chinese character dataset.

➢ Experiments on single-font printed Chinese characters

- **Comparison of accuracy rate between RTN and RAN with different caption lengths on different font-style unseen Chinese characters respectively.**

Font Style	RAN(%)			RTN(%)		
	ALL	≤ 6	> 6	ALL	≤ 6	> 6
Song	92.21	93.65	90.78	**94.54**	**94.93**	**94.16**
FangSong	91.04	91.98	90.11	**94.21**	**94.84**	**93.57**
Hei	90.41	91.34	89.50	**92.79**	**92.41**	**93.11**
Kaiti	88.57	90.59	86.58	**91.31**	**92.96**	**89.67**

- **Comparison of character-level accuracy between RTN and RAN with respect to the frequency of radicals.**



➢ Experiments on natural scene Chinese characters

- **Comparison of the performance of RAN and RTN with the different appearance frequency of character-level categories.**

Frequency	≤ 20	≤ 50	≤ 100	HF	ALL
Categories	398	511	335	1044	2015
Samples	1128	1229	1663	48745	52765
RAN	25.88%	47.92%	65.12%	89.01%	85.95%
RTN	**41.84%**	**61.51%**	**71.67%**	**89.76%**	**87.51%**

- **Comparison of the recognition performance of RTN and RAN with different caption lengths on the CTW valid database.**

Model	Caption length		
	ALL	≤ 4	> 4
RAN	85.95%	89.33%	82.31%
RTN	**87.51%**	**90.02%**	**84.80%**
Accuracy↑	**1.56%**	0.69%	**2.49%**

- **Comparison of the recognition performance of RAN and RTN with respect to 6 attributes on the CTW test dataset.**

Attributes	Training Samples	RAN(%)	RTN(%)
All	760107	85.56	87.31
occluded	101393	71.55	73.94
background	218560	82.84	84.57
distorted	192481	71.55	83.60
3D raised	199066	76.17	78.06
wordart	65983	87.11	84.25
handwritten	6661	63.58	66.70

4. Conclusion

- ✓ **We explore the option to improve the capability of RAN by employing the Transformer architecture.**
- ✓ **The proposed model achieves significant performance improvements on both printed Chinese character database and natural scene Chinese character database.**
- ✓ **Further analysis proves that RTN is more effective and robust than RAN for recognizing complicated and low-frequency samples**