LODENet:
A Holistic Approach to Offline Handwritten Chinese and Japanese Text Line Recognition

ICPR 2020, January 10th-15th, 2021

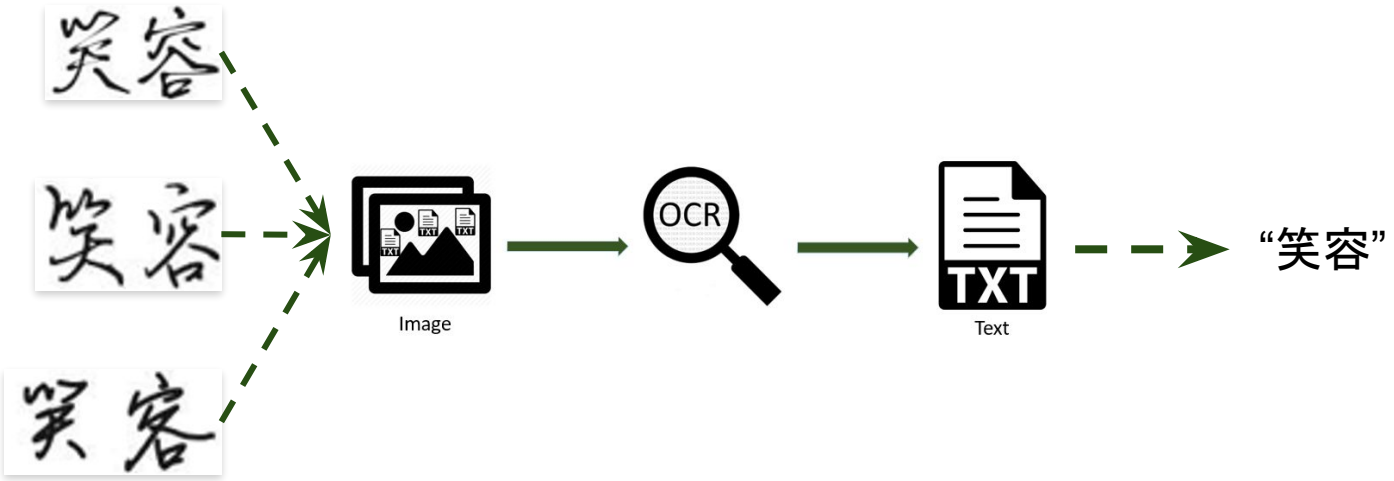
Huu-Tin Hoang*, Chun-Jen Peng*, Hung Vinh Tran, Hung Le
Cinnamon AI
Minato, Tokyo 105-0001, Japan
Email: {tin, larry, xing, toni}@cinnamon.is

Huy Hoang Nguyen
University of Oulu
Oulu, Finland
Email: huy.nguyen@oulu.fi

*Corresponding author

Handwritten OCR challenges

Vast style variety of handwritten texts, especially for Chinese and Japanese



Motivation: Kanji input method

cinnamon AI

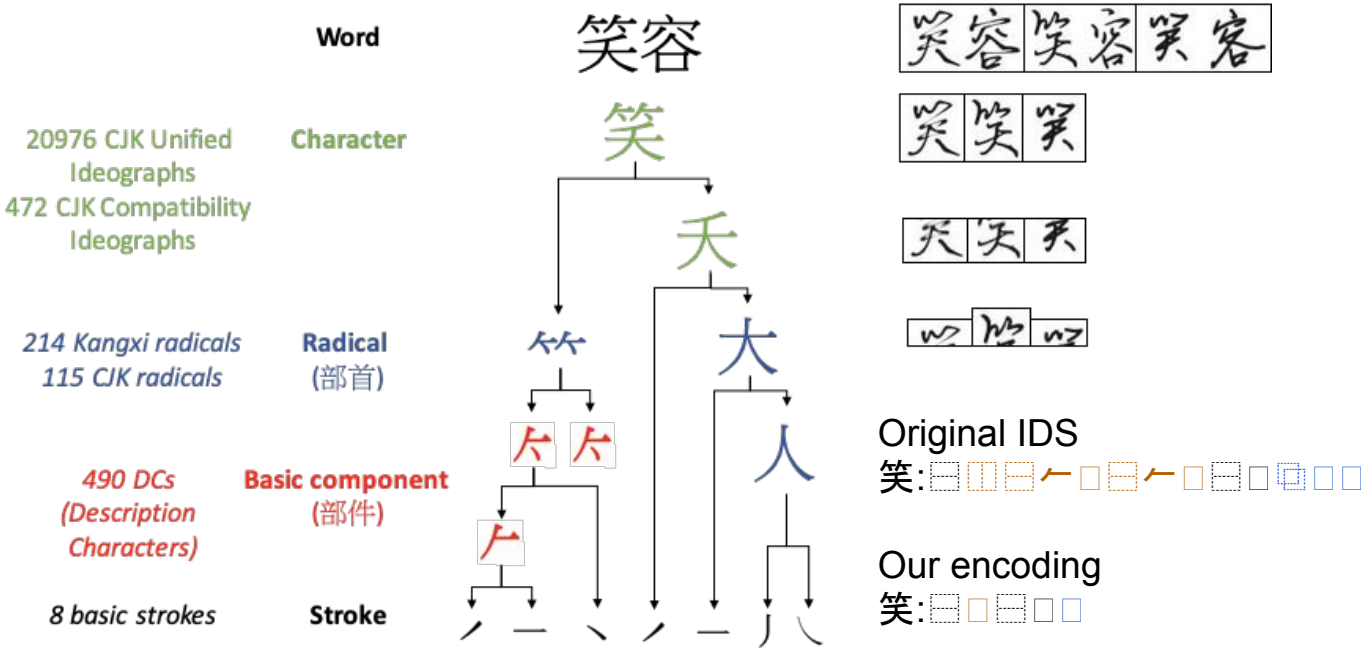
There are over 47,000 kanji characters!
… and here is the one-hot encoding keyboard.

https://knowyourmeme.com/photos/479661-wtf

Level of decomposition

Difference between our encoding (radicals + basic components) and IDS

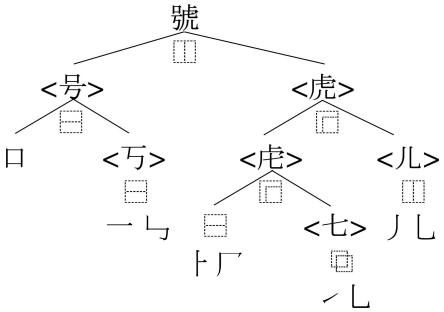
- **Radicals**: stopped at *214 Kangxi radicals* and *115 CJK radicals*.
- **Basic components**: If no radical occurred, stopped at leaf nodes



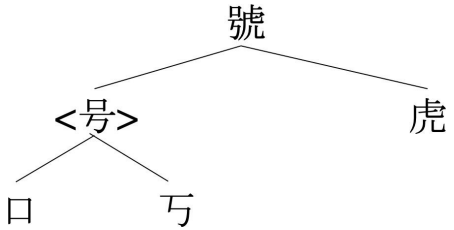
LODEC vs Other encodings

(a) Handwritten Sample

(b) Cangjie



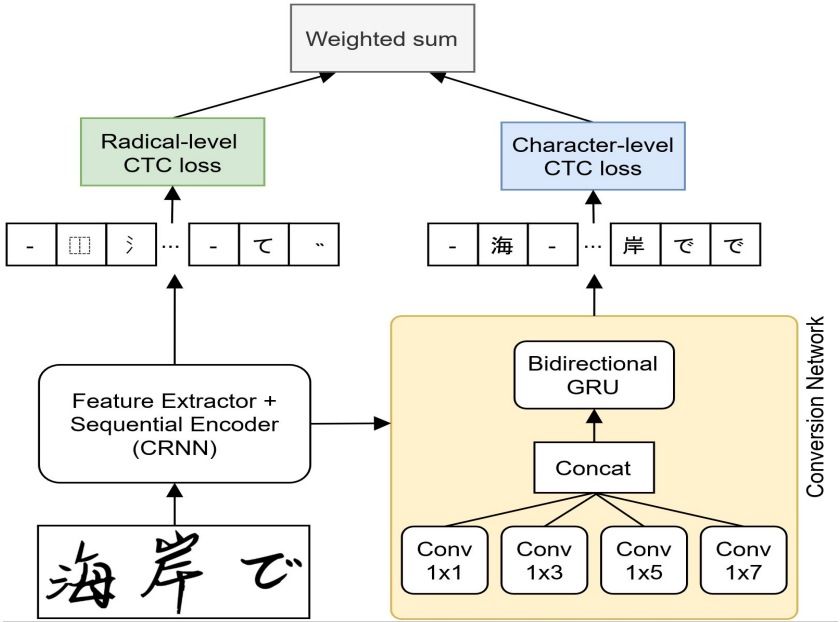
(c) IDS



(d) LODEC

(a) A handwritten sample of the character 號 and the corresponding encoding methods:
(b) Cangjie, (c) Ideographic Description Sequences (IDS) and (d) LODEC.

LODENet: Overview



LODENet overview picture

Feature Extractor and Sequential Decoder: CRNN architecture inspired by *Baoguang Shi et.al.**, for predicting radicals

Conversion Network: composed of inception convolutional and sequential layers, for predicting logograms

** Baoguang Shi et.al., An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition*

Experimental Dataset

	CASIA	SCUT-EPT	Private*
Dataset information			
Language	Chinese	Chinese	Japanese
Writers	1,080	2,986	-
Input height *h*	128	64	64
Classes	7,355	4,251	2,227
Amount of data (lines/characters)			
Train	52,229 / 1.4M	40,000 / 1.0M	14,595 / 248K
Synthesis-Random texts	120,000 / 1.9M	120,000 / 1.9M	-
Synthesis-Wiki texts	120,000 / 1.9M	120,000 / 1.9M	-
Test	3,432 / 92K	10,000 / 249K	2,941 / 50K

Description of 3 datasets used in this study.

**D. Nguyen et al. '19, "Improving long handwritten text line recognition with convolutional multi-way associative memory"*

Result: SCUT-EPT Dataset

Model	Character (%)		Radical (%)	
	AR ↑	CR ↑	AR ↑	CR↑
CNN + MDLSTM + CTC, *Messina et al. '15*	73.26	78.30		
CNN + MDirLSTM + CTC, *Wu et al. '17*	73.65	78.53		
CNN + LSTM + CTC, *Y. Zhu et al. '18*	75.97	80.26		
CMAM, *D. Nguyen et al. '19*	74.45	82.14		
LODENet (Ours)	76.61	82.91	77.81	85.40
LODENet + Random texts (Ours)	77.36	**84.64**	**79.27**	**87.47**
LODENet + Wikipedia texts (Ours)	**77.61**	83.74	79.25	86.33

Testing results on SCUT-EPT.

Result: CASIA Dataset

Model	Character (%)		Radical (%)	
	AR ↑	CR ↑	AR ↑	CR↑
HIT-1, *C. Szegedy et al. '15*	83.58	86.15		
HIT-2, *C. Szegedy et al. '15*	86.73	88.76		
Wang et al. '16	88.79	90.67		
CNN + SMDLSTM + CTC, *Wu et al. '17*	86.64	87.43		
LODENet (Ours)	86.82	87.83	88.51	91.09
LODENet + Random texts (Ours)	92.00	92.89	**92.79**	**95.25**
LODENet + Wikipedia texts (Ours)	**92.16**	**93.04**	92.77	94.65

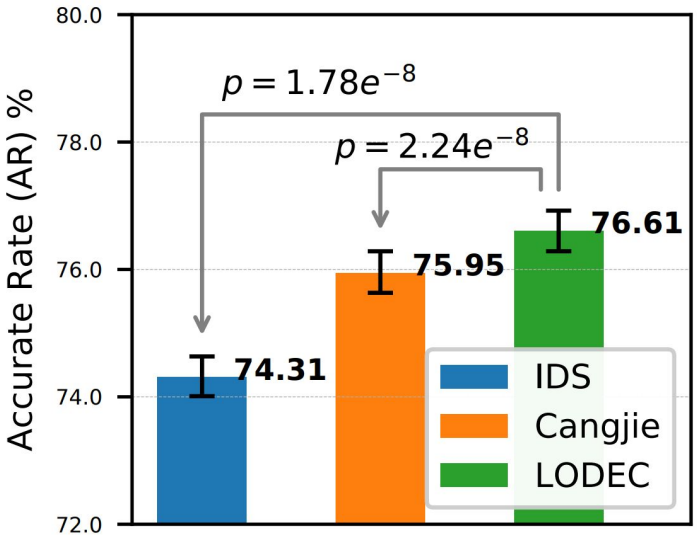
Testing results on CASIA.

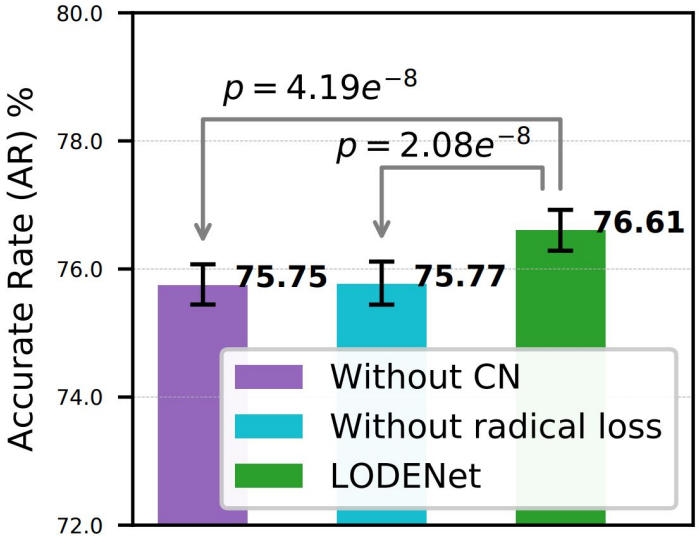
Result: Japanese dataset

Model	Character (%)	
	Validation ↓	Test ↓
CMAM, *D. Nguyen et al. '19*	17.55	12.99
LODENet w/o conversion network (Ours)	12.36	6.25
LODENet (Ours)	**11.70**	**5.47**

CERs on the Japanese dataset.

Ablation study: Statistical Tests





Ablation studies on different components of our method.

Qualitative result

Ground truth	与其城市文化有很大相关性
CRNN	与其城市菊 (8)
LODENet (Ours)	与其城市文化有有大相关性 (1)
Radical output	与一甘一八土戊冂亠小亻匕亻木月丷一大人忄青

Ground truth	责任在肩头	善假于物也
CRNN	责近"寿高头 (4)	苦物也 (3)
LODENet (Ours)	责任"在肩头 (1)	善假于物也 (0)
Radical output	丰贝亻一士"土一尸月头	一丬口亻十巳コ日又于牛勹丿丿也

LODENet and CRNN on SCUT-EPT test set

Conclusions

cinnamon AI

1. **LODEC encoding** that can fully represent all logograms and syllabic characters of Chinese and Japanese.
2. An **end-to-end training scheme** that can be plugged in any sequential architecture and radical-based encoding method.
3. **LODENet architecture** equipped with the conversion network that learns to transcribe Japanese and Chinese contents from radical-based features.
4. **SOTA results** on CASIA and SCUT-EPT, and one private Japanese dataset.

Thank You!

Poster Session: PS T4.1 - 1319

Appendix

Evaluation metrics: AR, CR and CER

Accurate rate (AR) and correct rate (CR), which were used in the ICDAR 2013 competition.

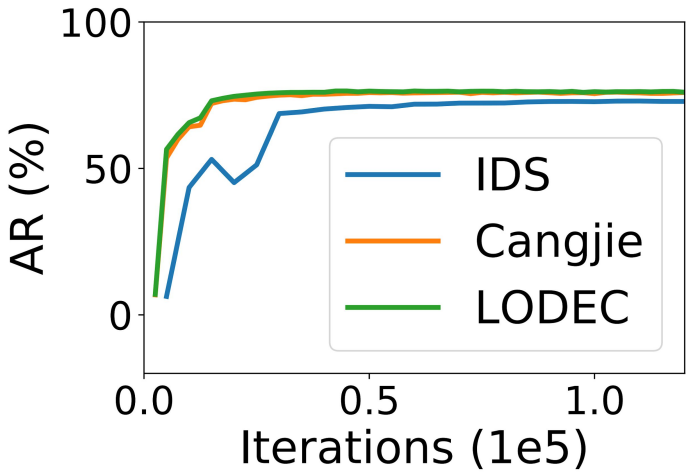
$$CR = \frac{N - Sub - Del}{N}$$

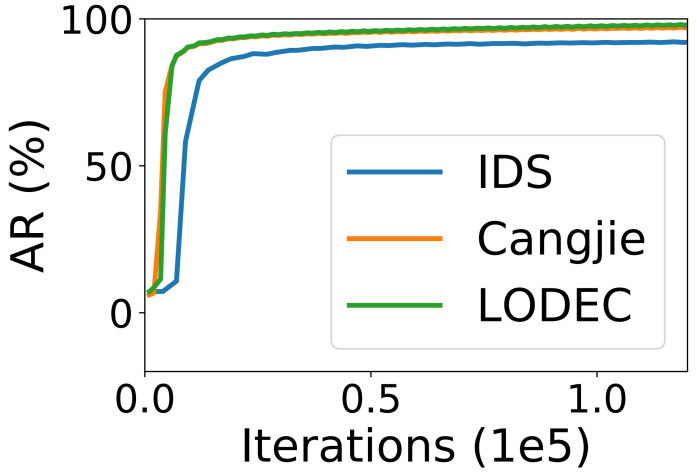
$$AR = \frac{N - Sub - Ins - Del}{N}$$

Character error rate (CER) on the Japanese dataset.

$$CER = \frac{Sub + Ins + Del}{N}$$

Learning Curve





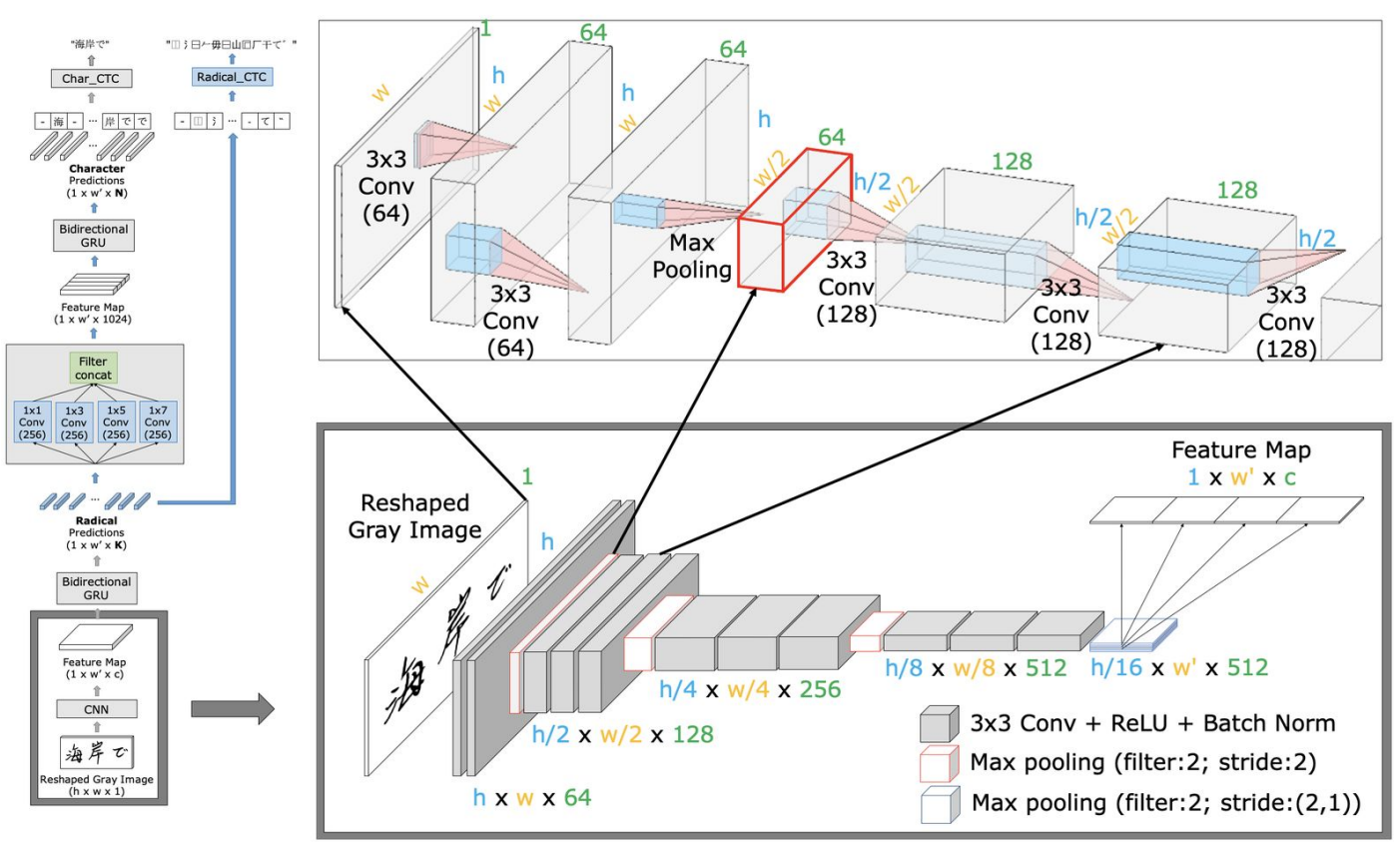
Learning curves of LODENet with different encoding methods on the SCUT-EPT dataset.

Encodings statistic

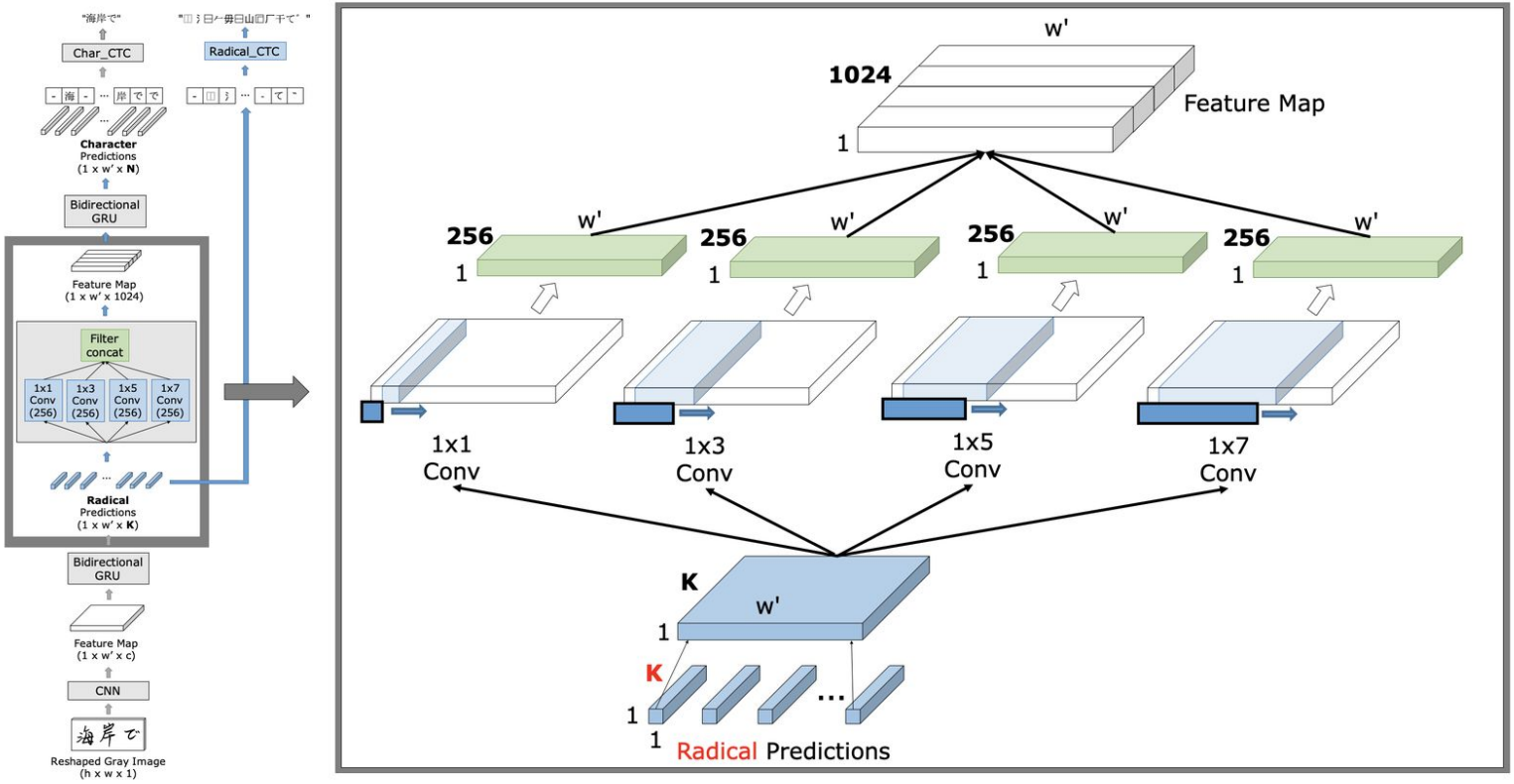
Method	No. of codes	Bijective function	Encoding length Mean & Std.
One-hot	21448	Yes	21448±0
Cangjie	26	No	109±19
IDS	387	No	3618±1722
LODEC (Ours)	520	Yes	1856±691

Table I: The amounts of codes and encoding lengths to represent 21,448 Kanji characters having Unicode correspondences in different encoding methods.

LODENet: Feature Extractor



LODENet: Conversion Network



Inspiration: Kanji input method

Problem: There are over 47,000 kanji characters!!

Similar repetitive components appear in different characters

Radical	Code	Auxiliary Shape	Examples
日	A	日 日 四	明 晶 晌 昭 昌 星 昃 昆 巴 色 象 像

http://202.175.185.186/it-school/mbc/association/sec/it/cj_note.htm

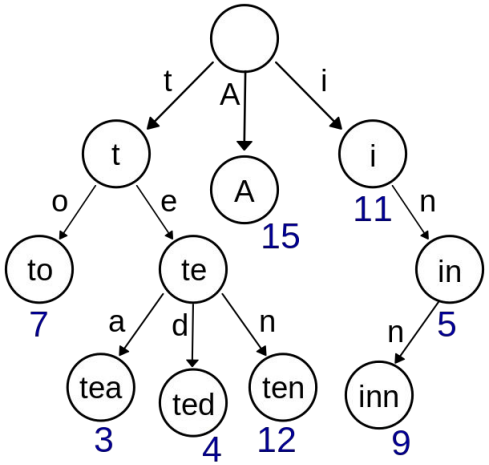
Cangjie input method

- First Chinese input method on QWERTY, invented by Chu Bong-Foo in 1976
- Composed of **24 radicals** associated with **77 auxiliary shapes** (total 101 basic components)

http://www.binarywebidea.com/blog/technology/20161027/

Cangjie input method

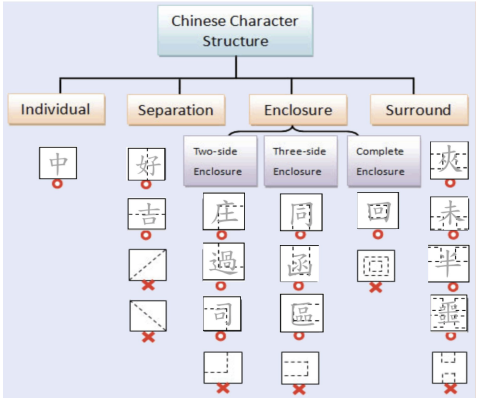
https://www.wikiwand.com/en/Cangjie_input_method



https://www.wikiwand.com/en/Trie

Ideographic Description Sequences (IDS)

- Unicode defines **12 IDCs** (Ideographic Description Characters)
- Different encoding for same component
- Not fully decomposable



Unicode	IDC	Example	IDS
2FF0	⿰	相	⿰木目
2FF1	⿱	志	⿱士心
2FF2	⿲	湘	⿲氵木目
2FF3	⿳	糞	⿳米田共
2FF4	⿴	回	⿴口口
2FF5	⿵	问	⿵门口
2FF6	⿶	凶	⿶凵乂
2FF7	⿷	匝	⿷匚口
2FF8	⿸	病	⿸疒丙
2FF9	⿹	戴	⿹[異
2FFA	⿺	起	⿺走巳
2FFB	⿻	巫	⿻工从

LODEC for Japanese

- Codec package for JApanese character DEComposition

1. One-hot : 輝

2. Cangjie : FUBJJ

3. IDC_Cangjie : ⿰FUBJJ

4. DCs : 丶丿’□□□

5. IDC_DCs : ⿰丶丿’□□□

6. IDCs_DCs : ⿰⿱丶丿’□⿱□□

- Hiragana & Katakana decomposition:
  Isolate **dakuten** and **handakuten**
  - ホ -> ["ホ"]
  - ボ -> ["ホ", "゛"]
  - ポ -> ["ホ", "゜"]

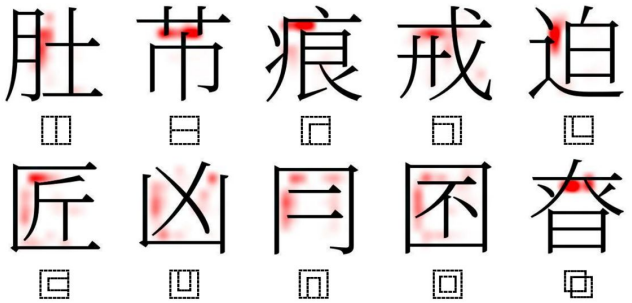
7. original_DCs : 丶丿’□□□□□

8. original_IDC_DCs :⿰丶丿’□□□□□

9. original_IDCs_DCs :⿰⿱丶丿’⿱□⿰□□⿱□□

Discussion

1. Why do we use 1st Level IDC ?
   - Most Kanjis are simple 2-level IDS tree
   - The major space between radicals (described by 1st level IDC) is an important feature to learn



1. Difference between our encoding (radicals + basic components) and IDS
   - **Radicals**: We stopped at *214 Kangxi radicals* and *115 CJK radicals*.
   - **Basic components**: If no radical occured, we stopped at leaf nodes