CROSS-LINGUAL LEARNING IN MULTILINGUAL SCENE TEXT RECOGNITION

Jeonghun Baek Yusuke Matsui Kiyoharu Aizawa
The University of Tokyo

ICASSP 2024 KOREA

Task: Multilingual Scene Text Recognition (Multilingual STR)

High-resource language: e.g. **ENG** Low-resource language: e.g. **KOR**

(a) Monolingual training on a low-resource language Korean (KOR) results **in incorrect predictions.**

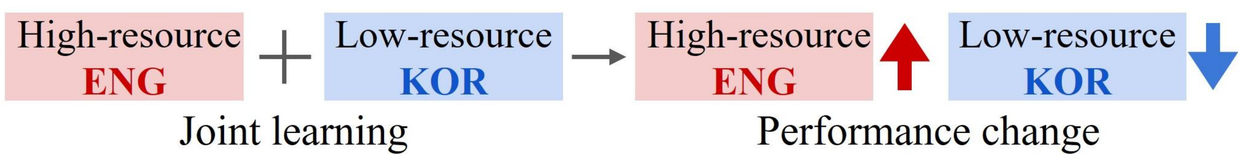
(b) Cross-Lingual Learning (CLL) from a high-resource language English (ENG) to KOR **leads to correct predictions.**

Preliminaries: (1) Multilingual STR: A task to recognize text from multiple languages in a word- or line-level scene image.
(2) Cross-Lingual Learning (CLL): A methodology for transferring knowledge from one language to another.

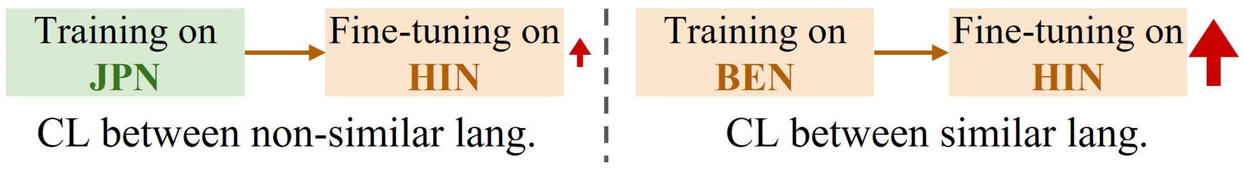
What we did:

1. Verified two general insights about CLL and showed that **the two general insights may not be applied to multilingual STR.**
2. Showed that CLL is simple and highly effective for improving performance in low-resource languages.
3. Found the condition where CLL works well: the crucial condition for CLL is **the dataset size of high-resource languages,** regardless of **the kind of high-resource languages.**

Two General Insights from Previous Works



According to MRN [1], joint learning on high- and low-resource languages may make the model biased toward high-resource languages. Therefore, it may result in poor performance in low-resource languages.



According to M-BERT [2], CLL works best between similar languages.
→ We investigate CLL performance between similar languages in linguistic typology and appearance.

Experimental Settings and Results

We investigate the effect of CLL via **joint and cascade learning.** They are simple and basic methods for CLL.

- **Joint learning** is to train a model using multiple datasets from multiple languages simultaneously.
- **Cascade learning** is to train a model in one language and then fine-tune the model in another language.

For the cascade learning part, please see our paper.

Dataset statistics: we use MLT19[3] and SynthMLT[4] datasets.

Count type	CHI	BEN	HIN	JPN	ARA	KOR	LAT
Training	1.9K	2.6K	2.7K	3.1K	3.5K	4.0K	28K
Validation	204	341	320	337	408	455	3.4K
Evaluation	248	311	349	366	421	525	3.3K
Charset	1.6K	608	696	1.3K	141	878	85
SynthMLT	33K	165K	54K	37K	202K	147K	211K

Training data	Model: CRNN (TPAMI'17) [1] CHI	BEN	HIN	JPN	ARA	KOR	LAT	Avg.	Model: SVTR (IJCAI'22) [7] CHI	BEN	HIN	JPN	ARA	KOR	LAT	Avg.
Each lang. (*Mono.*)	3.2	33.4	24.2	14.3	27.1	4.3	**85.7**	27.4	2.1	19.1	22.8	10.9	23.6	5.7	**85.0**	24.2
All lang. (*All*, 46K)	**41.3**	**69.3**	**66.9**	**44.8**	**54.7**	**61.4**	85.4	**60.5**	**35.2**	**67.8**	**68.5**	**42.9**	**53.7**	**53.0**	84.7	**58.0**
All but LAT (18K)	8.3	32.6	31.6	17.8	36.9	12.2	–	–	4.1	19.5	22.9	12.5	26.5	10.1	–	–

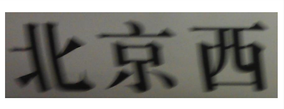
Results of joint learning: Simple joint learning drastically improves performance in low-resource languages!

Training data	Model: CRNN (TPAMI'17) [1] CHI	BEN	HIN	JPN	ARA	KOR	LAT	Avg.	Model: SVTR (IJCAI'22) [7] CHI	BEN	HIN	JPN	ARA	KOR	LAT	Avg.
Base (20K)	9.7	35.1	31.8	19.5	40.2	15.9	20.9	24.7	6.0	23.0	25.0	14.5	29.3	11.8	20.2	18.5
+ 2K S-CHI (22K)	20.1	51.8	49.7	28.0	42.8	35.5	42.2	38.6	6.5	19.3	25.6	14.3	27.1	11.2	18.9	17.6
+ 33K S-CHI (53K)	**54.8**	69.0	68.6	39.2	59.7	52.3	59.4	57.6	**56.6**	62.3	63.3	41.5	50.1	49.2	57.2	54.3
+ 2K S-BEN (22K)	10.3	40.9	37.6	19.0	40.2	15.9	23.0	26.7	6.5	23.4	27.6	14.3	27.4	13.8	20.4	19.1
+165K S-BEN (185K)	32.7	**81.3**	74.2	40.1	65.1	55.1	63.2	58.8	44.0	**84.8**	**80.8**	51.0	70.6	63.2	68.7	66.2
+ 2K S-HIN (22K)	17.7	55.1	52.3	28.7	41.7	36.0	42.5	39.2	5.8	22.3	26.3	13.9	28.1	11.1	19.2	18.1
+ 54K S-HIN (74K)	36.4	74.2	**78.7**	41.4	57.7	57.2	61.2	58.1	29.0	64.6	70.0	37.4	52.7	44.6	55.3	50.5
+ 2K S-JPN (22K)	19.4	55.9	53.9	30.5	41.8	39.2	43.5	40.6	6.0	20.9	25.2	14.6	27.3	12.7	19.0	18.0
+ 37K S-JPN (57K)	39.2	70.3	71.5	**54.8**	59.3	59.8	63.7	59.8	38.6	66.9	70.2	**56.7**	55.3	53.9	62.7	57.7
+ 2K S-ARA (22K)	13.5	46.4	43.9	26.4	41.9	31.6	39.4	34.7	6.4	19.9	25.8	14.3	28.8	13.1	19.4	18.2
+202K S-ARA (222K)	32.4	71.1	72.1	37.2	**82.2**	54.0	58.7	58.2	33.7	69.2	70.8	40.8	**82.7**	50.0	57.4	57.8
+ 2K S-KOR (22K)	10.2	37.4	35.0	19.6	38.8	19.4	20.7	25.9	6.4	23.2	25.8	14.0	28.2	14.9	20.9	19.1
+147K S-KOR (167K)	40.9	75.2	73.6	45.7	65.2	**80.4**	**64.3**	**63.6**	50.7	81.0	78.4	53.5	73.6	**82.5**	**69.6**	**69.9**

Joint learning with additional data, SynthMLT [4]:

- The crucial factor for CLL is **the number of samples in high-resource languages** rather than the similarity between languages.
- We assume this is because the essential knowledge of STR is distinguishing text in the image and can be learned from any language.

Qualitative analysis: LAT is a high-resource language (28K). The results below show the effectiveness of CLL.

Input images	Joint learning on **all but LAT**	Joint learning on **all**	Ground Truth
	الحم	الباب	الباب
	বান	রাখিবেন	রাখিবেন
	か本	北京西	北京西
	ज	मंडप	मंडप
	०り	くすり	くすり
	상온	냉면은	냉면은
	디센다	CUISINE	CUISINE

[1] Zheng et al., MRN: Multiplexed Routing Network for Incremental Multilingual Text Recognition, ICCV, 2023
[2] Pires et al., How Multilingual is Multilingual BERT?, ACL, 2019
[3] Nayef et al., ICDAR2019 Robust Reading Challenge on Multi-lingual Scene Text Detection and Recognition -- RRC-MLT-2019. ICDAR, 2019
[4] Bušta et al., E2E-MLT - an Unconstrained End-to-End Method for Multi-Language Scene Text, ACCV, 2018