Manual

This manual guides a reproduction of main experiment results. For data production, please refer to the Data_guide.pdf shipped in the OSOCR-data repo.

I. Setting up environments.

Step 1A. Install a clean Manjaro Linux (Archlinux should do as well)

Step 1B. Install dependencies using scripts from: https://github.com/lancercat/make_env

Step2. While waiting, grab the code and data.

1) Grab the data from kaggle (https://www.kaggle.com/vsdf2898kaggle/osocrtraining):

```
CVPR2016.zip,NIPS2014.zip,ssddata_2.zip,ctwcheval.zip,ssddata_1.tar.gz,ssddata.tar.gz
```

2) Unzip & remove the packages.

```
for i in $(ls | grep zip); do unzip $i; done
tar -xvf ssddata_1.tar.gz;
tar -xvf ssddata.tar.gz
mkdir ../packs; mv *.* ../packs;
```

Step 3C. Stop and check if you have all the following folders

Data:

```
[lasercat@lasercat-proj260 osocrdata]$ ls
artdb_seen ctwcheval CUTE80 dicts IC13_1015 lsvtdb_seen mlttrjp_hori NIPS2014 rctwtrdb_seen
ctwch ctwdb_seen CVPR2016 IC03_867 IIIT5k_3000 mlttrchlat_seen mlttrkr_hori pami_ch_fsl_hwdb SVT
```

Models:

```
[lasercat@lasercat-proj260 prfinal]$ ls ablative ctw_extra hwdb models revision_extra
```

II. Evaluate the Trained models.

1. Zero-shot Chinese character recognition

CTW dataset (ZSL)

CTW						
# characters in training set						
500	1000	1500	2000			
28.03	49.00	58.37	64.03			

basic_ctwch_CE_alter/testg2.py

test accu				
Accuracy:		CER:		
97591				
500 Done				

Accuracy:	0.583753,	AR:	0.583753,	CER:	0.416247,	WER:	0.416247
97591							
1500 Done							

CTW dataset (OSTR)

	CTW						
#NIC	50	100	200	250			
#NOC	450	400	300	250			
A(NIC)	79.3	77.1	72.6	69.6			
R(NOC)	73.3	54.7	37.7	31.5			
P(NOC)	98.9	95.9	92.4	88.6			
F(NOC)	84.2	69.7	53.5	46.5			

basic_ctwch_CE_alter/testg2-rej.py

HWDB dataset (ZSL) basic_hwdb_CE_alter/testg2.py

HWDB							
# characters in training set							
500	1000	1500	2000				
47.92	74.02	81.11	85.72				

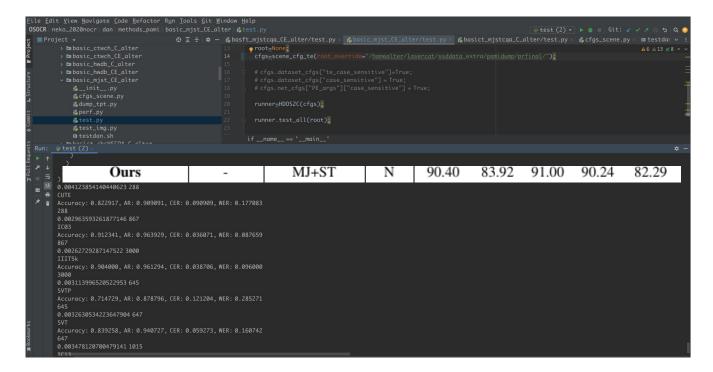
HWDB dataset(OSTR) basic_hwdb_CE_alter/testg2-rej.py

	HWDB					
#NIC	100	200	400	500		
#NOC	900	800	600	500		
A(NIC)	93.5	93.9	91.0	90.0		
R(NOC)	48.0	24.6	7.9	5.1		
P(NOC)	99.7	99.5	97.9	96.7		
F(NOC)	64.8	39.5	14.6	9.7		

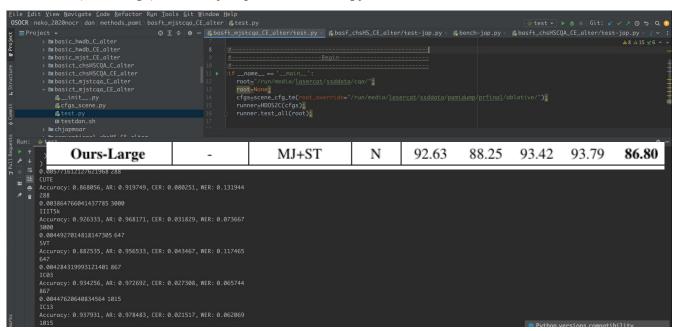
39777 dictrej400.pt Done test rej accuracy: KACR: 0.900278,URCL:0.051198, UPRE 0.967130, F 0.097248

2. Close-set text recognition

Close-set (Regular) basic_mjst_CE_alter/test.py



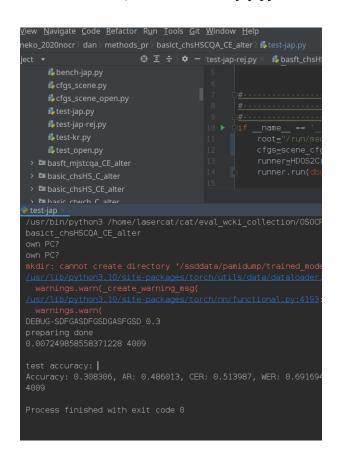
Close-set (Ours-large) basict_mjstcqa_CE_alter/test.py



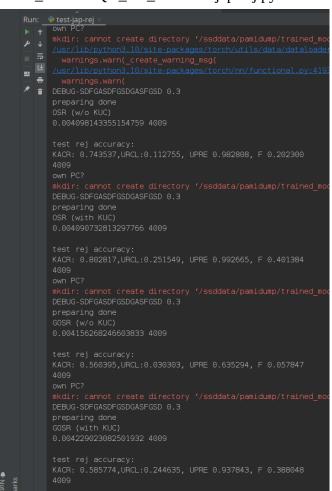
3. Open-set Text Recognition

Name	\mathbf{C}_{test}^{i}	\mathbf{C}^o_{test}	$ \mathbf{C}^i_{test} $	LA	R	P	F
GZSL	Unique Kanji, Shared	Ø	1460	30.83	-	-	-
	Kanji, Kana, Latin			,	,		
OSR	Shared Kanji, Latin	Unique Kanji,	849	74.35	11.27	98.28	20.23
w/o SOC		Kana					
OSR	Shared Kanji	Unique Kanji,	787	80.28	25.15	99.26	40.13
with SOC		Kana, Latin					
GOSR	Shared Kanji, Unique	Kana	1301	56.03	3.03	63.52	5.78
	Kanji, Latin						
OSTR	Shared Kanji, Unique	Kana, Latin	1239	58.57	24.46	93.78	38.80
	Kanji						

GZSL basict_chsHSCQA_CE_alter/test-jap.py



Others: basict_chsHSCQA_CE_alter/test-jap-rej.py



Details: chiapmoar/accr folder.py (make sure you have results from basict chsHSCQA CE alter/test-jap.py)

Name	Sample Requires	Sample Excludes	CA(%)	LA(%)
Shared Kanji	Shared Kanji	Unique Kanji, Kana	85.69	73.21
Unique Kanji	Unique Kanji	Kana	76.50	40.87
All Kanji	Unique Kanji or Shared Kanji	Kana	79.94	54.91
Kana	Hiragana or Katakana		25.10	0.72
All			54.03	30.83

```
Connected to pydev debugger (build 221.5591.52)
/run/media/lasercat/20615BC32265B955/prfinal/chs-japxl/
Accuracy: 0.308306, AR: 0.486013, CER: 0.513987, WER: 0.691694
Overall 0.5403113212380896 0.3083063108006984
/run/media/lasercat/20615BC32265B955/prfinal/chs-japxl/
Accuracy: 0.732161, AR: 0.856908, CER: 0.143092, WER: 0.267839
Seen 0.8112844997463888 0.7321613236814891
/run/media/lasercat/20615BC32265B955/prfinal/chs-japxl/
Accuracy: 0.408730, AR: 0.765065, CER: 0.234935, WER: 0.591270
Unique Kanji 0.7549325410039688 0.4087301587301587
/run/media/lasercat/20615BC32265B955/prfinal/chs-japxl/
Accuracy: 0.549169, AR: 0.799458, CER: 0.200542, WER: 0.450831
All Kanji 0.7794014876155192 0.5491692860350247
/run/media/lasercat/20615BC32265B955/prfinal/chs-japxl/
Accuracy: 0.007295, AR: 0.251016, CER: 0.748984, WER: 0.992705
Kana 0.24151569804923498 0.007295173961840628
Ours&54.03/30.83&&81.13/73.22&75.49/40.87&77.94/54.92&24.15/0.73\\
```

Process finished with exit code 0