La Région Auvergne-Rhône-Alpes · INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON · LIRIS · cnrs · IAPR

DAS 2022
May 22-25, La Rochelle

Challenging children handwriting recognition study exploiting synthetic, mixed and real data

Sofiane Medjram, Véronique Eglin et Stéphane Bres
Université de Lyon INSA Lyon, LIRIS, UMR5205, F-69621, France

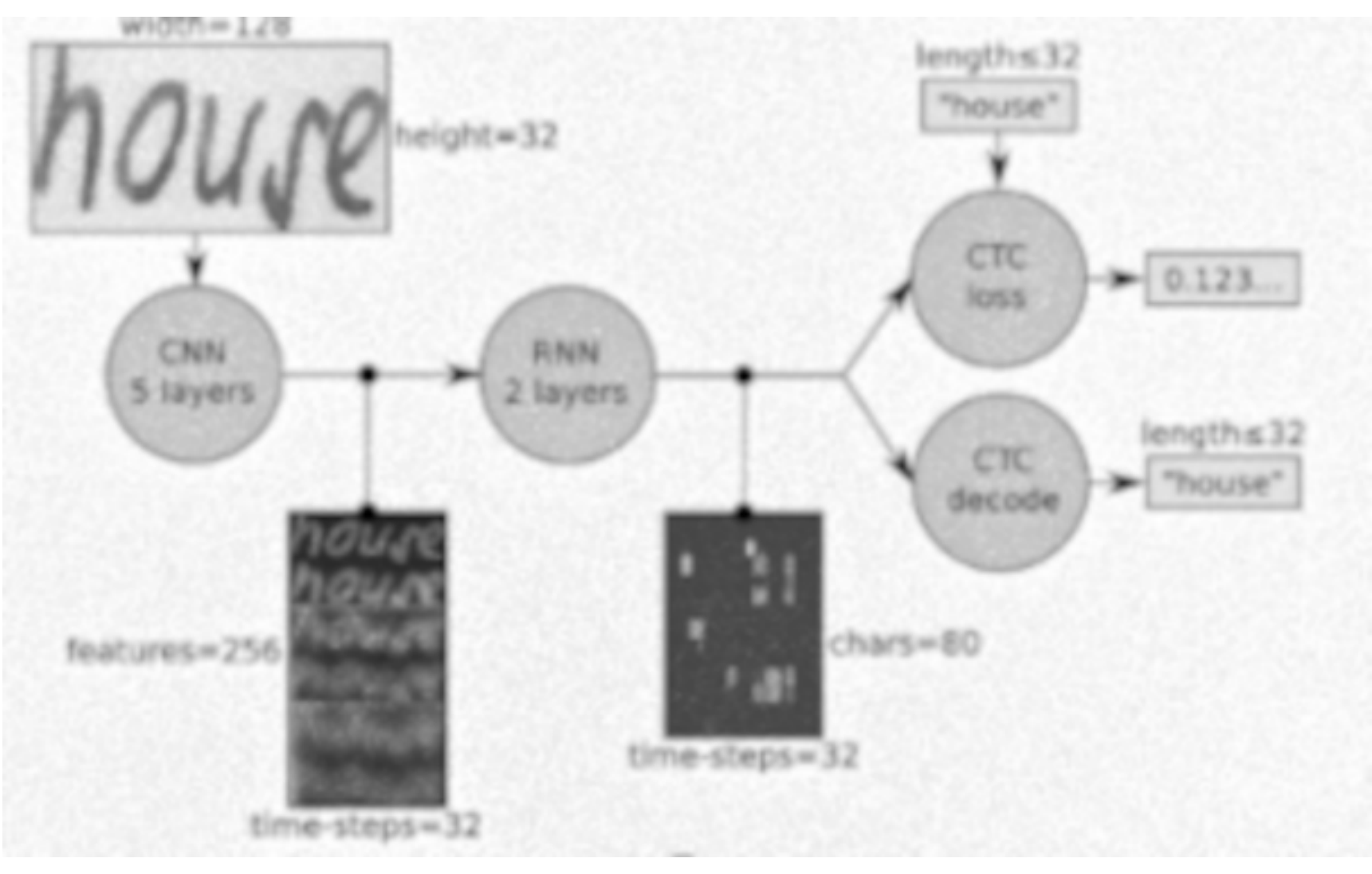
In this paper, we investigate the behavior of a MDLSTM-RNN architecture to recognize challenging children handwriting in French language. The system is trained across compositions of synthetic adult handwriting and small collections of real children dictations gathered from first classes elementary school. The paper presents the results of investigations concerning handwriting recognition in a context of weak annotated dataset and synthetic images generation for data augmentation

Problematic

Handwritten Text Recognition (HTR) domain is still a challenging research field in a context where training data are at least weakly annotated if not missing. Children's handwriting recognition represents an example that falls into this ontext of weakly annotated data, making deep learning based HTR systems ineffective.

MDLSTM-RNN Model

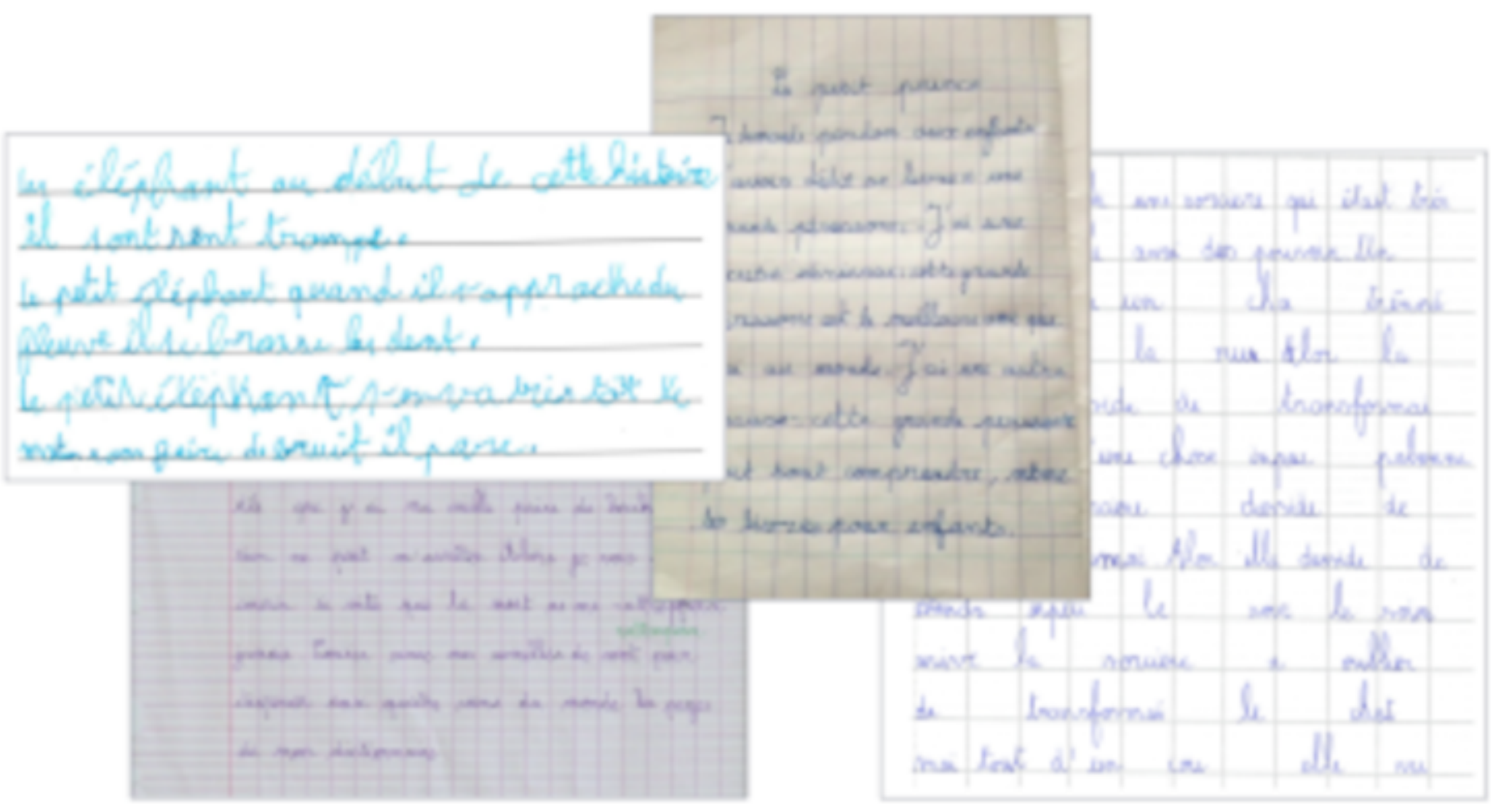
- Model fits well on IAM adults handwriting dataset
- CRNN decoder-decoder architecture based
- B-LSTM encoder based
- CTC decoder based
- 32 timesteps
- 256 encoding features



Available Data

Collected from ScolEdit dataset children copies of CP classes

- 250 clean copies without guidlines
- 71 children handwriting s
- 3832 detected words
- Annotated to IAM format



Empirical Study Strategies

Strategy I:

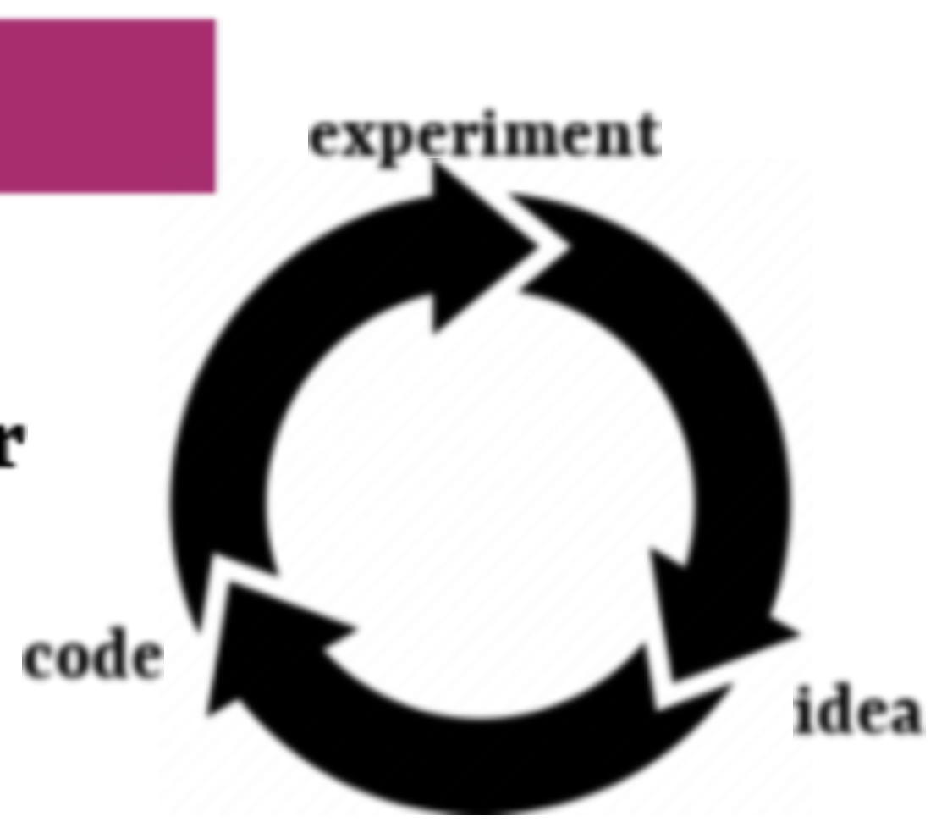
- Supervised validation and domain transfer

Strategy II:

- Automaton for dictation word

Strategy III:

- Large lexicon training with transfer



Data analysis

71 children
3832 words

95% train
5% val & test

CV
cross validation
random split

Manu
validation
manual split

Test
children
real words

GSV · GSVg
GSV: GAN small vocabulary
GSVg: GSV augmented
3 Adult styles
6 words, 320 repetitions

GLV · GLVch
GLV: GAN large vocabulary
GLVch: GLV with real children
3 Adult styles
1736 unique words each

GAN · IAM
Datasets for
transfert learning

Supervised validation and domain transfer

Dataset	TL	Val CER%	Val WAR%	Test CER%	Test WAR%
CV	No	17.52%	62.74%	35.07%	49.45%
Manu	No	9.50%	77.08%	31.09%	54.68%
Manu	GAN	10.46%	74.47%	44.01%	41.66%
Manu	IAM	6.38%	82.81%	28.87%	56.25%
IAM	No	9.71%	78.50%	-	-

Automaton for dictation word

Dataset	TL	Val CER%	Val WAR%	Test CER%	Test WAR%
GSV	No	0%	100%	34.93%	56.25%
GSVg	No	0%	100%	20.36%	74.57%

Large lexicon training with transfer

Dataset	TL	Val CER%	Val WAR%	Test CER%	Test WAR%
GLV	No	1.16%	95.70%	82.31%	3.12%
GLVch	No	7.18%	80.46%	8.22%	74.47%
GLVch	IAM	5.47%	85.67%	5.52%	84.11%

Conclusions

- MDLSTM-RNN model adapts well to Children handwriting recognition
- Synthetic data and augmentation are essential to improve performances
- HTR Performances are significant where validation and test sets are well selected
- When the amount of data is small and the complexity of writing important, it is convenient to choose carefully training and validation sets for a better generalization and model convergence.