Entraînement d'architectures type Transformer sur peu de données : application à la reconnaissance de texte manuscrit ancien

Killian Barrere, Yann Soullard, Aurélie Lemaitre, Bertrand Coüasnon

Entraînement d'architectures type Transformer sur peu de données : application à la reconnaissance de texte manuscrit ancien

UMR IRISA

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES RENNES

1. INTRODUCTION

1.1 Transformers pour la reconnaissance de texte

Principe :
- Architectures type “encoder-decoder”
- Apprentissage conjoint :
  - Reconnaissance optique
  - Modélisation de la langue

⇒ **Résultats** à l’**état de l’art** [1,2,3,4]

Tendance actuelle :
- **Architectures de plus en plus larges** (jusqu’à 550M de paramètres [2,3])
- Apprentissages avec **beaucoup de données**
  - Données synthétiques
  - Ajout de données réelles

1.2 Documents anciens : manque de données

- Écriture et langues anciennes
- Dégradation des documents
- Documents très variés

⇒ **Données annotées coûteuses et rares**

1.3 Objectif : Transformer pour des documents anciens

- **Transformer léger** ⇒ entraînement avec peu de données, limiter le surapprentissage
- **Stratégies d’entraînement** pour pallier au manque de données annotées

2. DATASET ICFHR READ 2018

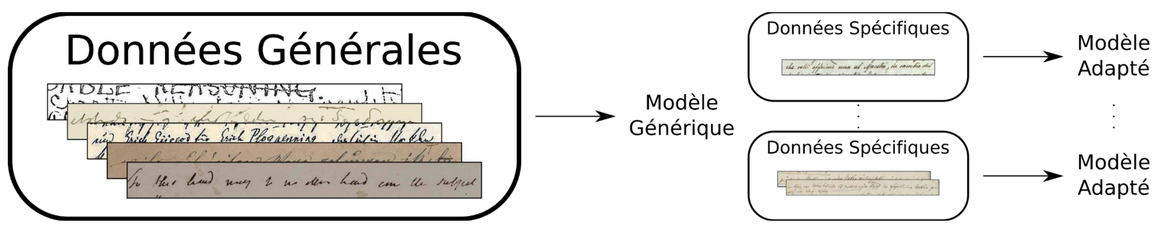
2.1 Données générales

But : s’entraîner sur **un corpus large**
- 11 903 images de texte ancien
- 4 langues (allemand, anglais, danois, suédois)
- ⇒ **Bonnes capacités d’apprentissage**

2.2 Données spécifiques

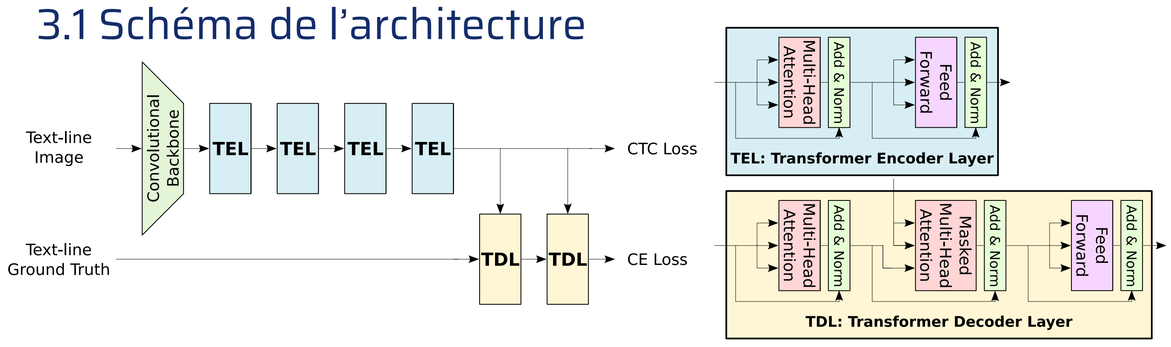
But : se spécialiser sur **peu de données**
- 5 documents (4 allemands, 1 italien)
- 4 scénarios par document
  - (0, 1, 4 ou 16 pages annotées)

⇒ Bonnes capacités d’**adaptation avec peu de données**



3. VERY LIGHT TRANSFORMER

3.1 Schéma de l’architecture



3.2 Un modèle léger

- **5 couches de convolutions** (contre 18 ou plus [2] (i.e. ResNet18))
- **256 neurones** dans les **Transformers** (contre 1 024 [2] pour les plus larges)
- **2 couches de décodeur** pour limiter le surapprentissage

⇒**5.6M paramètres** (contre 100-550M)

3.3 Avantages

- Apprentissage plus rapide
- Moins sujet au surapprentissage

⇒ **adapté aux documents anciens**

4. STRATÉGIES

4.1 Stratégie d'entraînement et de spécialisation

Entraînement général
- Encodeur entraîné
  ⇒ reconnaissance de caractères
- Décodeur entraîné
  ⇒ modèle de langue générique

Spécification sur un document
- Encodeur ré-entraîné
  ⇒ **spécialisation sur l’écriture de l’auteur** et le type de document
- Décodeur non ré-entraîné
  ⇒ garder un **modèle de langue générique**, limiter le surapprentissage

4.2 Stratégie de prédiction à base de votes

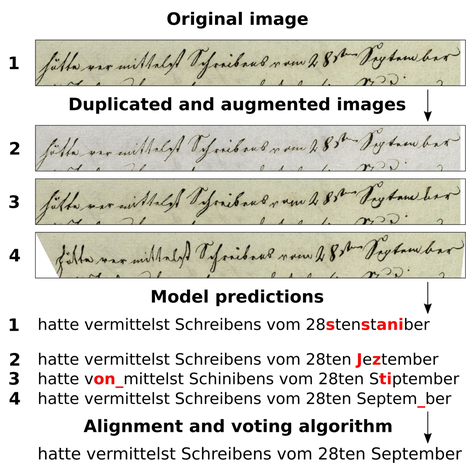
- Inspiré de [6]
- But : **correction d’éventuelles erreurs** avec de multiples prédictions

Augmentations en test
- L’image originale est dupliquée
- Augmentations aléatoires

⇒ Plusieurs prédictions

Fusion des prédictions
- Alignement des prédictions, basé sur la distance d’édition
- Votes des caractères

⇒ **Prédiction finale consolidée**

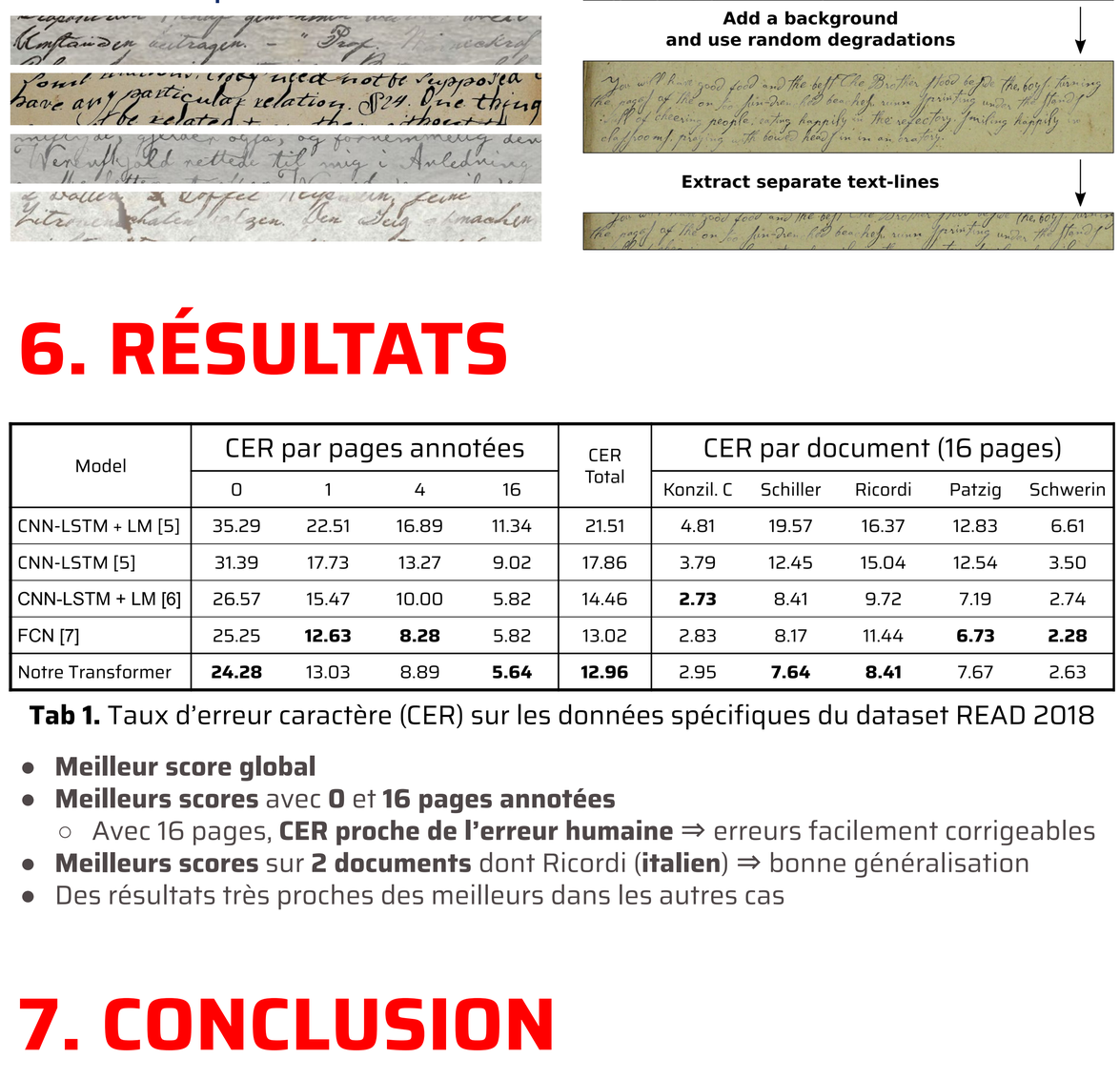


5. DONNÉES SYNTHÉTIQUES

5.1 Génération

- Contenu textuel divers
- **Polices d’écritures manuscrites (21)**
- **Augmentations de données**
  - Distorsions élastiques, italique, ...
  - Coloration du texte, contraste, ...
- Générées au niveau paragraphe
- **Générées à la volée** à l’entraînement
- Pour une époque 50% réel 50% synth.

5.2 Exemples



6. RÉSULTATS

Model	CER par pages annotées				CER Total	CER par document (16 pages)				
	0	1	4	16		Konzil. C	Schiller	Ricordi	Patzig	Schwerin
CNN-LSTM + LM [5]	35.29	22.51	16.89	11.34	21.51	4.81	19.57	16.37	12.83	6.61
CNN-LSTM [5]	31.39	17.73	13.27	9.02	17.86	3.79	12.45	15.04	12.54	3.50
CNN-LSTM + LM [6]	26.57	15.47	10.00	5.82	14.46	**2.73**	8.41	9.72	7.19	2.74
FCN [7]	25.25	**12.63**	**8.28**	5.82	13.02	2.83	8.17	11.44	**6.73**	**2.28**
Notre Transformer	**24.28**	13.03	8.89	**5.64**	**12.96**	2.95	**7.64**	**8.41**	7.67	2.63

Tab 1. Taux d’erreur caractère (CER) sur les données spécifiques du dataset READ 2018

- **Meilleur score global**
- **Meilleurs scores** avec **0** et **16 pages annotées**
  - Avec 16 pages, **CER proche de l’erreur humaine** ⇒ erreurs facilement corrigeables
- **Meilleurs scores** sur **2 documents** dont Ricordi (**italien**) ⇒ bonne généralisation
- Des résultats très proches des meilleurs dans les autres cas

7. CONCLUSION

Architecture Transformer pour des documents anciens
- Architecture de Transformer **légère**
- Entraînée avec **diverses stratégies**
  - Données synthétiques pour les documents anciens
  - Stratégies d’entraînement et de spécification ⇒ adaptation à l’écriture
  - Stratégie de prédiction ⇒ correction d’erreurs

⇒ **Bonne généralisation** et bonne **capacité d’adaptation**

Killian Barrere
Univ Rennes, CNRS, IRISA, France
Yann Soullard
Univ Rennes, CNRS, IRISA, France
Aurélie Lemaitre
Univ Rennes, CNRS, IRISA, France
Bertrand Coüasnon
Univ Rennes, CNRS, IRISA, France

Contact
killian.barrere@irisa.fr

Références

- **[1]** Barrere K, Soullard Y, Lemaitre A, Coüasnon B. A Light Transformer-Based Architecture for Handwritten Text Recognition. In International Workshop on Document Analysis Systems (DAS) 2022
- **[2]** Kang L, Riba P, Rusiñol M, Fornés A, Villegas M. Pay attention to what you read: non-recurrent handwritten text-line recognition. Pattern Recognition. 2022
- **[3]** Li M, Lv T, Cui L, Lu Y, Florencio D, Zhang C, Li Z, Wei F. Trocr: Transformer-based optical character recognition with pre-trained models. arXiv preprint arXiv:2109.10282. 2021
- **[4]** Coquenet D, Chatelain C, Paquet T. DAN: a Segmentation-free Document Attention Network for Handwritten Document Recognition. arXiv preprint arXiv:2203.12273. 2022
- **[5]** Strauß T, Leifert G, Labahn R, Hodel T, Mühlberger G. ICFHR2018 competition on automated text recognition on a READ dataset. In 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR) 2018
- **[6]** Soullard Y, Swaileh W, Tranouez P, Paquet T, Chatelain C. Improving text recognition using optical and language model writer adaptation. In 2019 International Conference on Document Analysis and Recognition (ICDAR) 2019
- **[7]** Yousef M, Hussain KF, Mohammed US. Accurate, data-efficient, unconstrained text recognition with convolutional neural networks. Pattern Recognition 2020