Architecture Transformer Légère pour la Reconnaissance de Textes Anciens

Killian Barrere, Bertrand Coüasnon, Aurélie Lemaitre, Yann Soullard

Architecture Transformer Légère pour la Reconnaissance de Textes Anciens

1. INTRODUCTION

1.1 Transformers pour la reconnaissance de texte

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

⇒ **Bons résultats** [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 Problématique : manque de données

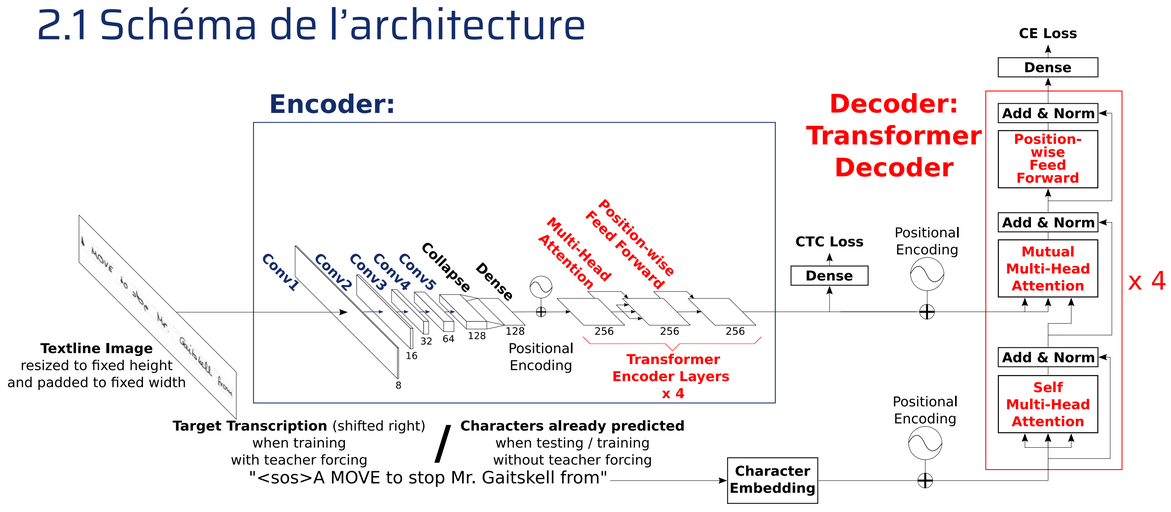
- Données annotées coûteuses et rares
- Les données ajoutées en quantité ne sont pas accessibles pour tous
- **En particulier : documents anciens**
  - **Peu de données pour la difficulté relative de la tâche**

1.3 Objectif : Transformer pour des documents anciens

- **Architecture de Transformer légère** [1] pour s'entraîner avec peu de données
- **Stratégies d’entraînement** pour pallier au manque de données annotées

2. ARCHITECTURE LÉGÈRE [1]

2.1 Schéma de l’architecture



2.2 Un modèle léger

- **5 couches de convolutions** (contre 18 ou plus (i.e. ResNet18))
- **256 neurones** dans les **Transformers** (contre 1 024 pour les plus larges)

⇒ **6.9M paramètres** (contre 100-550M)

2.3 Avantages

- Apprentissage plus rapide
- Sur du moderne (IAM) :
  - **Pas besoin de données ajoutées**
  - Résultat à l’état de l’art avec des données synthétiques

⇒ **Reconnaissance de documents anciens**

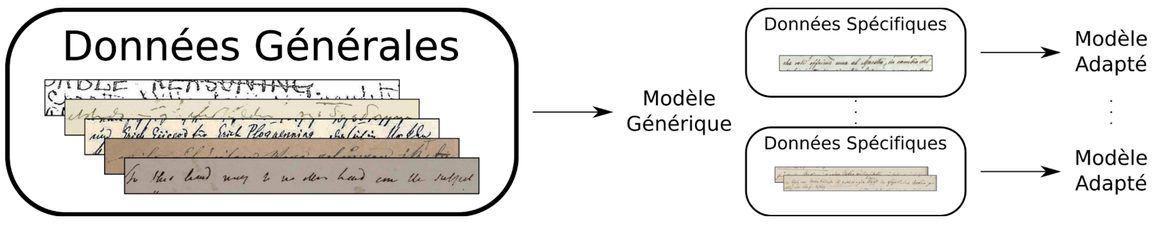
3. DATASET ICFHR READ 2018

3.1 Données générales

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

3.2 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)



4. DONNÉES SYNTHÉTIQUES

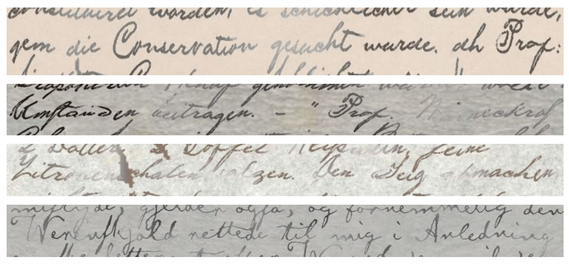
4.1 Génération

- Contenu textuel divers
- **Polices d’écritures manuscrites (21)**
- **Augmentations de données**
  - Distorsions élastiques, italique, ...
  - Coloration du texte, contraste, ...
- Générées au niveau paragraphe

4.2 Entraînement

- **Générées à la volée** à l’entraînement
- Pour une époque 50% réels 50 synth.

4.3 Exemples



5. PREMIERS RÉSULTATS

Model	CER par pages annotées				CER par document					CER Total
	0	1	4	16	Konzil. C	Schiller	Ricordi	Patzig	Schwerin	
CNN-LSTM [5]	31.39	17.73	13.27	9.02	9.39	21.10	23.27	23.17	12.98	17.86
CNN-LSTM + LM [5]	32.25	19.79	16.98	14.72	10.49	19.05	35.60	23.83	17.02	20.94
CNN-LSTM + LM [5]	35.29	22.51	16.89	11.34	9.14	25.69	30.50	25.18	18.04	21.51
FCN [6]	**25.35**	**12.63**	**8.28**	**5.82**	6.49	**13.77**	**17.33**	**14.85**	12.33	**13.02**
CNN-LSTM + LM [7]	26.57	15.47	10.00	5.82	**5.94**	14.81	21.62	18.08	**11.73**	14.46
Notre encoder	28.48	17.58	13.33	10.10	9.80	19.89	19.02	23.25	13.99	17.37
Notre decoder	29.80	18.04	13.18	9.59	8.89	18.60	21.92	24.07	13.79	17.65

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

5.1 Notre modèle

- Architecture Transformer légère
- Avec données synthétiques
  - Contenu textuel du dataset
  - Pour les données générales ou pour les données spécifiques
- Résultats au niveau de l’état de l’art

⇒ **Résultats corrects avec le peu de données disponible pour spécialiser**

5.2 Apprentissage de la langue (decoder)

- La partie decoder agit comme un modèle de langue (LM)
- Le decoder est bénéfique avec suffisamment de données
- Le decoder n’est pas bénéfique pour :
  - Nouvelle langue (Ricordi, italien)
  - Nouveau vocabulaire (Patzig)

⇒ **Le decoder est limité par le très faible nombre de données pour spécialiser**

6. PISTES ENVISAGÉES

6.1 Plus de données

- **Plus de données réelles**
  - En combinant différents datasets anciens (READ 2016, 2017 et 2018)
- **Plus de données synthétiques**
  - Basées sur le(s) dataset(s)
  - Articles Wikipédia
  - Livres anciens

⇒ Amélioration des résultats, en particulier pour le decoder

6.2 Différentes stratégies d’apprentissage

- **Comment entraîner avec plus de données ?**
  - Évolution de la proportion entre données réelles et synthétiques
- **Comment spécialiser un Transformer ?**
  - Decoder entraîné spécifiquement
  - Decoder spécifique à l’allemand
  - Par partie : spécialiser l’encoder avec un decoder fixe

7. CONCLUSION

- **Architecture Transformer sur des documents anciens**
  - Architecture légère
  - Entraînée avec des données synthétiques
- **Résultats au niveau de l’état de l’art**, malgré une faible quantité de données
  - Le modèle a de la marge de progression
  - Manque de données pour modéliser la langue correctement

  ⇒ **Besoin de plus de données pour améliorer les résultats de l’architecture**
- Plusieurs pistes prometteuses envisagées

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

References

- **[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]** Yousef M, Hussain KF, Mohammed US. Accurate, data-efficient, unconstrained text recognition with convolutional neural networks. Pattern Recognition 2020
- **[7]** 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