Utilisation & Fine-Tuning de Modèles de Fondation pour l'OCR et l'Extraction d'Informations dans des Corpus Manuscrits Complexes

Les avancées récentes des modèles de fondations, mêlant différentes modalités de données, ouvrent des opportunités pour la reconnaissance optique de caractères (OCR) et en extraction d'informations structurées. Les modèles comme GPT, Llama [DJP+24] ou LayoutLM [XLC+20, HLC+22] intègrent des données textuelles et visuelles qui peuvent être exploités pour analyser des documents complexes. Cependant, leur efficacité est limitée lorsqu'il s'agit de corpus manuscrits difficilement lisibles et de formulaires spécifiques nécessitant une compréhension approfondie des structures.

Dans ce contexte, ce stage vise à explorer et à développer des solutions pour traiter des documents présentant des écritures manuscrites peu lisibles, souvent accompagnées de formulaires structurés, tout en tirant parti de connaissances a priori pour guider l'extraction.

Objectifs

1. Étude des modèles existants :
   — Explorer des modèles de fondation pour l'analyse de documents, tels que LayoutLM et leurs variantes.
2. Reconnaissance de manuscrits difficiles :
   — Étudier les limitations des modèles actuels face à des écritures manuscrites peu lisibles.
3. Exploitation de connaissances a priori :
   — Intégrer des connaissances spécifiques sur la structure des formulaires pour guider les modèles d'extraction.
4. Développement de modèles few-shot :
   — Concevoir et entraîner des modèles capables de traiter efficacement des formulaires spécifiques à partir d'un nombre limité d'exemples annotés.
   — Évaluer l'utilisation de techniques telles que le transfert d'apprentissage et les adaptations légères de modèles préentraînés.

Objectifs pédagogiques : prendre en main les modèles de fondations, explorer le prompting puis réaliser des opérations de fine-tuning. Réaliser des campagnes d'expériences.
Encadrement : Vincent Guigue (Professeur d'informatique à AgroParisTech), Tanguy Herserant (Doctorant)

Références

[DJP+24] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv :2407.21783*, 2024.

[HLC+22] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3 : Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4083–4091, 2022.

[XLC+20] Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, and Ming Zhou. Layoutlm : Pre-training of text and layout for document image understanding. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1192–1200, 2020.