ARIA
APPLIED RESEARCH IN ACTION

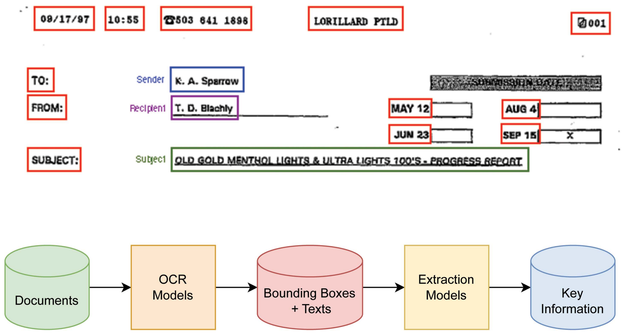
Two-Stage Deep-Learning Framework for Understanding Scanned Documents

Using a deep OCR model to identify texts and a transformer language model to extract relevant texts

Tsz Fung (Aiden) Yau

Gerald Penn
ACADEMIC SUPERVISOR

Simrandeep Singh
INDUSTRY SUPERVISOR



PROJECT SUMMARY

Scotiabank gets a lot of documents for various kinds of applications like mortgage approval and credit card approval. It takes a lot of manual work and time to validate them by customer agents. To speed it up, we built an automated document-understanding pipeline to extract relevant information from documents across various domains, like electronic and scanned forms, with minimal manual effort. Current rule-based approaches, however, cannot properly handle noises or rotations in scanned documents. To address this issue, we propose a machine-learning framework that can capture complex patterns of noisy and rotated text.

Our two-stage framework employs a custom-made deep OCR model using open-source EasyOCR and PaddleOCR libraries to identify texts from documents, plus another entity extraction model LayoutLMv2 to tag the target information from extracted texts. We observe that it achieved a 90% recall in OCR tasks on noisy banking documents as compared to 15% in the baseline Tesseract model. The result demonstrates that the machine-learning approach can handle scanned documents well by reducing the OCR errors which propagate in extraction models. We continue to explore other approaches such as OCR-less models for further improvement on scanned items.

REFERENCES

[1] G. Kim, T. Hong, M. Yim, et al., OCR-free document understanding trans- former, Oct. 6, 2022. doi: 10.48550/arXiv.2111.15664. arXiv: 2111. 15664[cs]. [Online]. Available: http://arxiv.org/abs/2111.15664 (visited on 03/21/2023).

[2] F. Lebourgeois, Z. Bublinski, and H. Emptoz, "A fast and efficient method for extracting text paragraphs and graphics from unconstrained documents," in Proceedings., 11th IAPR International Conference on Pattern Recognition. Vol.II. Conference B: Pattern Recognition Methodology and Systems, Aug. 1992, pp. 272–276. doi: 10.1109/ICPR.1992.201771.

[3] J. Ha, R. Haralick, and I. Phillips, "Document page decomposition by the bounding-box project," in Proceedings of 3rd International Conference on Document Analysis and Recognition, vol. 2, Aug. 1995, 1119–1122 vol.2. doi: 10.1109/ICDAR.1995.602115.

[4] J. Ha, R. Haralick, and I. Phillips, "Recursive x-y cut using bounding boxes of connected components," in Proceedings of 3rd International Con- ference on Document Analysis and Recognition, vol. 2, Aug. 1995, 952–955 vol.2. doi: 10.1109/ICDAR.1995.602059.

Computer Science
UNIVERSITY OF TORONTO
Master of Science in Applied Computing