# Exploring pretraining tasks for multimodal methods in DocVQA

Recently, VQA has been applied to documents as a generic natural language interface for different information extraction tasks.

This project will consist of two milestones. First, it has been discussed that explicit visual features are not specially relevant to understand documents, which has yield some works to ignore visual pretraining tasks which results in exploiting even less this modality. To start, the student will use an already existing model (examples are visual T5 [7] and Hi-VT5 [8]) with textual pretraining, and extend it with visual pretraining tasks to investigate the actual performance boost that explicit visual features can provide for DocVQA [5].

Then, with several different pretraining strategies proposed since the start of DocVQA and Document Intelligence in general [1-4, 6, 9-10]. Not all the works have provided clear ablation studies showing the effectiveness of such tasks compared to already existing ones. Therefore, the student will implement multiple pretraining strategies and perform a comparative analysis on the final results.

To conduct this work, the student will need to perform an extensive literature review on this field to identify the most promising pretraining tasks both in visual, textual and multimodal alignment pretraining tasks.

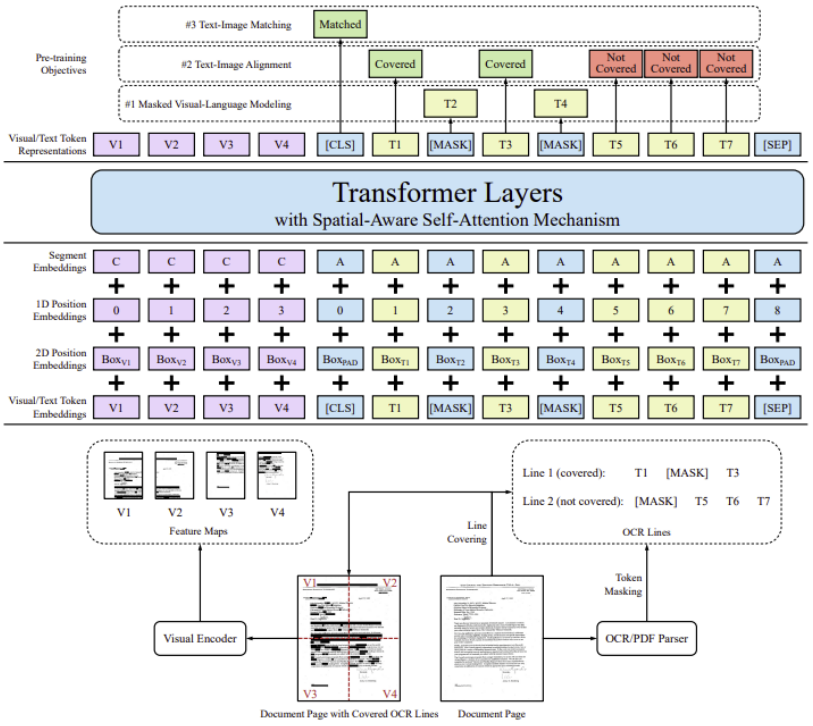


Figure 1: LayoutLMv3 [3] is a multi-modal transformer model for DocVQA that performs pretraining in both textual and multimodal alignment pretraining tasks

# References


1. Appalaraju, S., Jasani, B., Kota, B.U., Xie, Y., Manmatha, R.: Docformer: End-to-end transformer for document understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 993–1003 (2021)
2. Davis, B., Morse, B., Price, B., Tensmeyer, C., Wigington, C., Morariu, V.: End-to-end document recognition and understanding with dessurt. arXiv e-prints pp. arXiv–2203 (2022)
3. Huang, Y., Lv, T., Cui, L., Lu, Y., Wei, F.: Layoutlmv3: Pre-training for document ai with unified text and image masking. arXiv preprint arXiv:2204.08387 (2022)
4. Kim, G., Hong, T., Yim, M., Nam, J., Park, J., Yim, J., Hwang, W., Yun, S., Han, D., Park, S.: Ocr-free document understanding transformer. In: European Conference on Computer Vision. pp. 498–517. Springer (2022)
5. Mathew, M., Karatzas, D., Jawahar, C.: Docvqa: A dataset for vqa on document images. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2200–2209 (2021)
6. Powalski, R., Borchmann, L., Jurkiewicz, D., Dwojak, T., Pietruszka, M., Palka, G.: Going full-tilt boogie on document understanding with text-image-layout transformer. In: International Conference on Document Analysis and Recognition. pp. 732–747. Springer (2021)
7. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. 21(140), 1–67 (2020)
8. Tito, R., Karatzas, D., Valveny, E.: Hierarchical multimodal transformers for multi-page docvqa. arXiv preprint arXiv:2212.05935 (2022)
9. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. arXiv preprint arXiv:2012.14740 (2020)
10. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: 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. pp. 1192–1200 (2020)