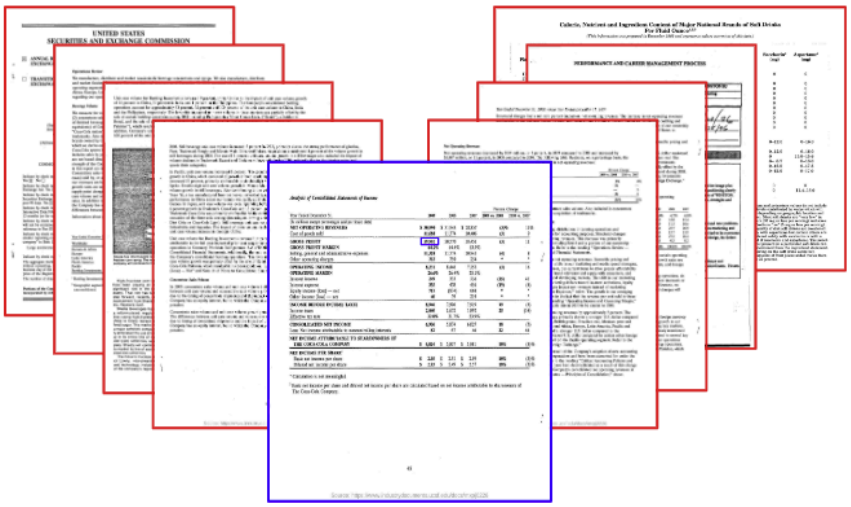
Can We Read a Book without Opening it? A New Perspective towards Multi-Page Document Visual Question Answering

Supr: Dr Lei Kang, (lkang@cvc.uab.cat)
Co-Supr: Dr Dimosthenis Karatzas, (dimos@cvc.uab.cat)
Computer Vision Center (CVC), Universitat Autònoma de Barcelona (UAB)

1. **Dataset**
   Document Visual Question Answering (DocumentVQA) refers to the task of answering questions from document images. Existing work on DocVQA only considers single-page documents. However, in real scenarios documents are mostly composed of multiple pages that should be processed altogether. Thus, our team has recently proposed a Multi-Page DocumentVQA dataset, namely MP-DocVQA (https://rrc.cvc.uab.es/?ch=17&com=tasks). Please refer to our paper for more details (https://arxiv.org/pdf/2212.05935.pdf).

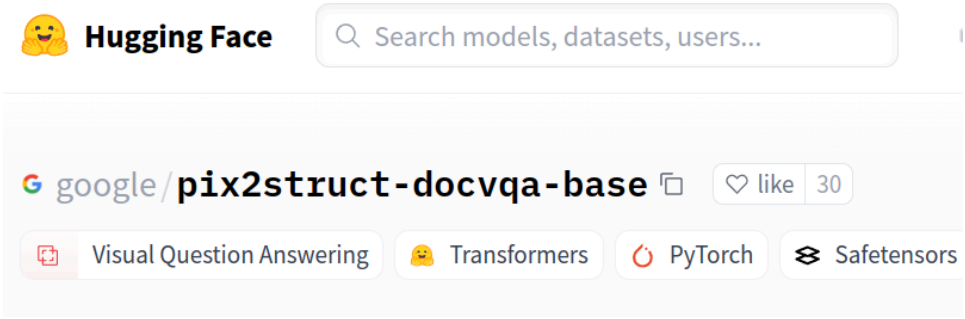


Q: What was the gross profit in the year 2009?
A: $19,902

2. **Challenge**
   Improving DocumentVQA methods to accommodate multi-page scenarios requires increased computational resources as the number of pages expands.
   Instead of squeezing a sequence of document images to fit the maximum input length of a model, can we concatenate the document images together vertically (where each page is treated as a channel) to accommodate the maximum input length without sacrificing high resolution?

3. **Possible Solution**
   Instead of reading a document from page to page, we choose to close the “book” and let the machine read the whole document all at once.

4. **Baseline Model**
   Pix2Struct (https://huggingface.co/docs/transformers/en/model_doc/pix2struct) is a visual-only DocumentVQA model, with a pre-trained model available on the single-page scenario.



5. **Master Thesis Schedule ("W" denotes "Week")**
   - Play with the Pix2Struct model on a single-page scenario, and understand MP-DocVQA dataset. [W1-W2]
   - Concatenate all the pages of each document into one “closed” document (maximum 20 pages per document), and modify the Pix2Struct model to fit the new input. [W3-W8]
   - Try different hyper-parameters to tune the model for a better performance. [W9-W11]
   - Finalize all the experiments and write the master thesis. [W12]

Excellent students might be given the possibility to continue for a PhD at the Computer Vision Center following the successful completion of this project.