**Data Collection for SpringBoard Capstone Project**

The project is an OCR task which involves text extraction from PDF’s or images with strong noise, blur or low quality images by using Convolution Neural Networks.

For a model to perform well for text recognition on low quality images it should be trained well on low quality images, but unfortunately the current open source datasets for OCR task are accommodated with decent to high quality pictures.

**Brno Mobile OCR Dataset:**

Brno Mobile OCT dataset is one of its kind , it’s a collection of low quality images taken from 23 different mobile cameras of different resolutions, lighting conditions, camera positions.

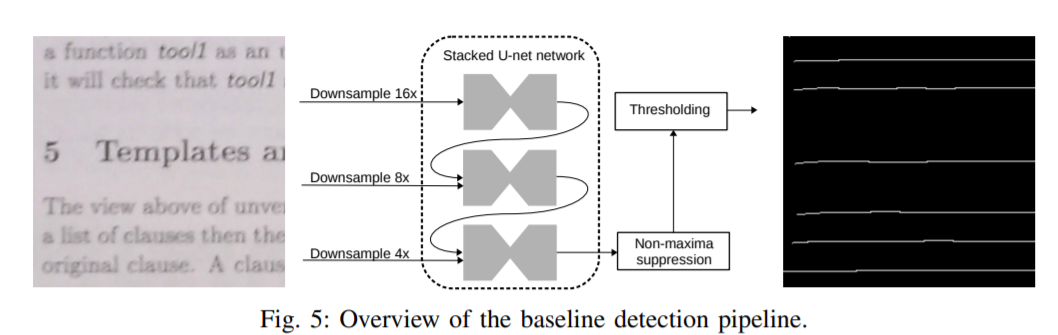
**Characteristics of dataset:**

* Combination of low quality images with non-uniform lighting, image blur, strong noise, built-in de-noising, sharpening, and compression.
* The dataset contains 2115 unique pages from random scientific papers, which were photographed by multiple people using 23 different mobile devices. Out of these image collection 19,725 images of various visual quality are selected.
* This dataset also provides text annotations for around 500K text lines along with their positional information.
* The state of art text recognition model build on this dataset using Convolution neural network (text localization), Recurrent Neural Networks (text recognition) on CTC loss function gives error rates of 2%, 22% , 73% word error rates on easy , medium and hard datasets segregated based on image quality.

**Drawbacks of Existing Datasets:**

* Datasets such as SmartDoc, SmartDoc-QA, SmartATID are also collection of images captured by mobile cameras but main drawbacks are they are relatively small, uniform and lack line level text annotations.

**Dataset Preparation:**

1. **Line Level Ground Truth:**
   * Ground truth for the text inside the images are obtained by using ABBYY FineReader and Tesseract.
   * ABBYY Fine Reader is a Deep learning based OCR product which is used to extract text from any formats and Tesseract is a similar OCR product.
   * Results from both of the OCR products are combined to obtain the ground truth text values.
2. **Text line position Refinement.**
   * Around 25 manually annotated rectified photographs are used to train a baseline neural network model to predict the baseline points.
   * The predicted baselines are then used by RANSAC algorithm to output a text line.
   * The neural network architecture consists of stacked u-net network , followed by Non maximum suppression and thresholding.  
     

**Separation of Images into Training, Validation and Development sets.**

* All the images are bucked into either of Easy, medium and High categories by using a network trained on all the images.
* The predictions for each text line is obtained and the images containing the text lines with no text recognition errors are categorized into easy bucket, images with text recognition error up to 20% error are categorized into medium bucker and the rest are categorized into hard bucket.
* Network used for text recognition is combination of convolution and recurrent neural networks with CTC loss. Recurrent network is combination of Encoder and decoder with attention mechanism.
* Images are further categorized into train, validation and test sets in such a way that texts are a different in training when compared with validation and testing sets.
* The below image shows the examples of easy, medium and hard images.

