Group: Brendan Neal (nealbre1), Ashkan Kiyomarsi (kiyomars)

CSC420 Project Proposal: LATEX Formula Extraction

General Idea: given a picture of a page with math equations, extract and then translate the equations into their raw LATEX source code.

Pipeline

1. Page Orientation Correction / Perspective Elimination
   - Motivation: prepare the page for later pipeline stages where the images/equations are assumed to have no perspective distortions
   - Implementation: we will create our own implementation of the techniques described by DropBox for correcting page orientation
     - https://blogs.dropbox.com/tech/2016/08/fast-and-accurate-document-detection-for-scanning/
       1. Use edge detection and a Hough transform to mark potential quadrilaterals
       2. Score each quadrilateral according to how their edges align with detected edges
       3. Pick quadrilateral with highest score as the page
       4. Determine the homography that would be able to correct for perspective
       5. Apply the homography to the page to yield a non-warped view
2. Formula Detection and Extraction
   - Motivation: want to focus on the relevant sections of the page (the equations)
   - Implementation: we will create and train a bounding-box detector for equations using automatically generated training data:
     - https://tex.stackexchange.com/questions/20575/attractive-boxed-equations
       1. This link describes how to create colored bounding boxes around equations in LATEX
       2. We will automatically interleave the formulas from the translation training set into fake paragraphs of text. These equations will have bounding boxes drawn around them in one color channel.
       3. By extracting the color channel of the bounding boxes, we will have an equation/not-equation Boolean mask that can be used for training the box detector
       4. The text color channel can be taken in isolation to hide these bounding boxes from the input data (to make training non-trivial)
3. Formula Translation
   - Motivation: want to translate the equations into LATEX source code
   - Implementation:
     - Architecture: https://arxiv.org/pdf/1609.04938v1.pdf
       1. We will be implementing the architecture described in this paper, possibly experimenting with different architecture decisions and hyper parameters
     - Training Data: https://zenodo.org/record/56198#.W9TQapNKiUl
       1. We will train the network using the im2latex-100k training dataset.
       2. The equations in this dataset will also be used to in the generation of the train data for the previous pipeline step