# CSC420 Project: Captioning images of equations into LaTeX

**Sooham Rafiz and Keith Shiran Yang**

Disclaimer: The following proposal makes forward-looking statements reflecting the group's expectation of the final project. The final project may not implement all features laid out here due to time constraints.

We propose using CV to caption images of equations into valid LaTeX, which is based on an OpenAI research recommendation.

**General pipeline:**

We first normalize scenes containing equations. This will be done by computing the homography transformations and removing noise and binarizing the image. The second step is to compute bounding boxes around equations in the image, for which we will experiment with different object-detection models such as YOLO, RCNN, and Single Shot Multibox detectors, if time permits. The final step is to explore between two to three LaTeX captioning models on detected equations: The first model is Sequence to Sequence (Seq2Seq) with attention, with the fixed size image turned into a sequence of vectors over regions of the image. The second model is a Multimodal Neural Language Model (taught in CSC321). The third model uses HOG (or another traditional) feature descriptor over the image in conjunction with a recursive neural network (RNN, GRU, LSTM). The latter model is different as it uses precomputed features instead of learned features. Finally, we shall compare and contrast variations of the models and hyperparameters quantitatively using popular deep learning metrics: precision and recall, accuracy, BLEU score and more. and qualitatively on corner cases, training and inference time etc.

We plan to use the following datasets: im2latex-100k dataset, the sum of all CROHME datasets, MNIST dataset and Mathematical Symbols Dataset for the purpose of training our sequence models or use as symbol images.

We anticipate the following difficulties in our project: We believe adapting to images of various sizes is a cumbersome task. Finding differentiable loss function to compare our predicted translation, which being more meaningful than the Levenshtein Distance will be a challenge, as there are many loss functions to explore here. For operators which affect a scope of other symbols (i.e fraction, square root etc.), we believe determining the scope will be difficult. Finally running our experiments on limited compute resources will be a challenge.

**Key related works:**

Im2Latex-100k dataset

Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks

SSD: Single Shot MultiBox Detector

You Only Look Once: Unified, Real-Time Object Detection

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

Multimodal Neural Language Models

Attention Is All You Need

An End-to-End Trainable Neural Network for Image-based Sequence Recognition and Its Application to Scene Text Recognition