Name: Benedict Aaron Tjandra (bat34)

Title: **Multimodal Relational Reasoning for Visual Question Answering**

Supervisor: Catalina Cangea

Project Description:

The efforts of deep learning has seen an explosive improvement in the efficiency and accuracy of mono-modal tasks. For instance, Convolutional Neural Networks (CNNs) have had great successes in object detection and segmentation when combined with Region Proposal Networks (RPNs) and Multilayer Perceptrons (MLPs), at times surpassing human performance \cite{performance}. Recurrent Neural Networks (RNNs) have seen similar improvements when dealing with data that is sequential; for example, in audio and textual processing. These advances are necessarily *mono*-modal in that they only accept one type of input --- in this case, visual or textual.

However, high-level cognitive tasks that humans perform in their day-to-day lives are *multimodal*; therefore, models that aim to solve multimodal tasks require effective representations for input data sources of different types.

One example of such a task is VQA (Visual Question Answering): answering an arbitrary question about an image, which requires the comprehension of textual and visual input.

This project aims to implement a network that aims to tackle VQA. The first step is to process inputs accordingly. For the visual input, the network will use an object detection model to detect objects and their bounding boxes in the image. As for the textual input, the network will use a sentence encoder to encode the question into a vector.

Crops of the objects detected in the image, their bounding boxes, as well as the encoded question are then fed to the MuRel cell. This will perform bilinear fusion to combine the visual and textual inputs and construct a pairwise relational graph on which it will perform relational reasoning. The output of the network is a vector of probabilities over all answers that the question entails, and the answer chosen by the network is the one that is assigned the highest probability.