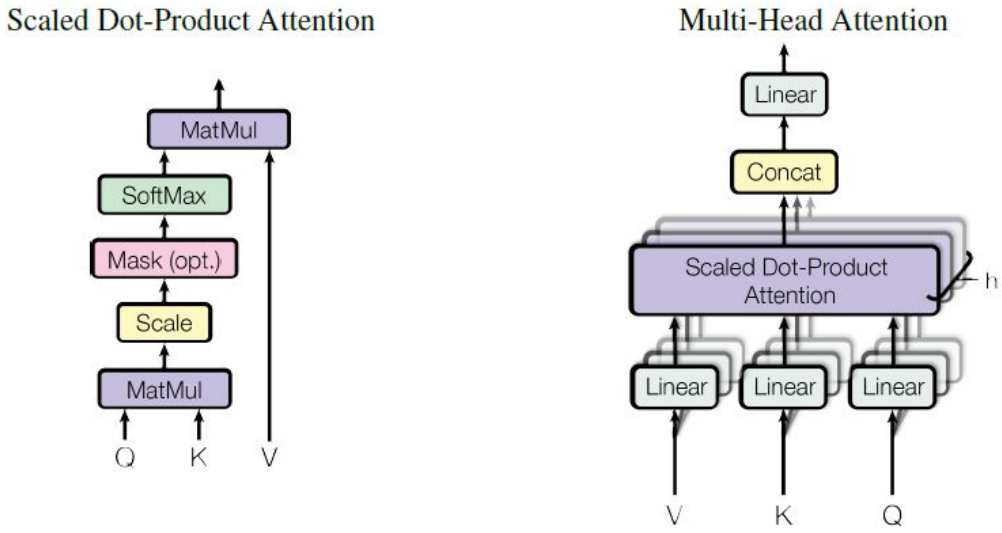
Final Project Proposal
CSE 237C

Weihong Xu
wexu@ucsd.edu

Overview: Describe your proposed final project highlighting three key elements:

- **Description:**
  Attention-based machine learning is used to model long-term dependencies in sequential data. Processing these models on long sequences can be prohibitively costly because of the large memory consumption. In this final project, I would like to use Vivado HLS to construct an efficient dataflow and accelerator to cope with the attention model.



  As shown in the above figure, I would like to optimize the scaled dot-product attention model and the multi-head attention module, two core functionalities in attention models. The work is done through making use of algorithm optimization, dataflow re-organization, and HLS programming improvements. Several techniques learnt from the CSE 237C will be used to construct the system, including matrix multiplication, CORDIC, and coalescing memory access.
  Compared to the naive implementation of attention on FPGA, the project is expected to achieve three goals: 1) memory usage reduction for attention model, 2) inference acceleration of attention model, 3) improving hardware efficiency.

- **Deliverables:**
  a. A report that describes the development flow and detailed optimization that have been accomplished during the project.
  b. Vivado HLS implementation of developed designs in Verilog and C.

c. Testbench code file to verify the functionality correctness of designs by comparing the generated computation results with the ground truth.

- **Timeline:**
    a. Day 1-2: Constructing testbench and baseline for attention model. The ground truth benchmarks will be run at Pytorch or Tensorflow.
    b. Day 3-4: Memory usage analysis and dataflow analysis of attention model. Then the efficient tiled dataflow will be developed.
    c. Day 5-8: Building up baseline implementation as well as optimized designs on Vivado HLS.
    d. Day 9-10: Collecting data and conducting comparison. Drafting report.

- **Project Requirements:**
    a. Vivado HLS (already have)
    b. Zynq Board (already have)
    c. A server with GPU card (would use Google Colab)

Your final project should demonstrate skills that you have learned in this class. Projects are varied -- they can focus more on implementation or they can be research oriented. Example projects were given in class.

This document should provide a clear picture of your project and the tasks that you must perform in order to complete your project. It will be used as a reference at the end of the quarter for grading your project.

The document should be 1-2 pages in length. The majority of the document should be devoted to articulating the deliverables and the timeline.