EE599 Deep Learning – Initial Project Proposal

©B. Franzke

November 9, 2020

Project Title: Realtime Background Replacement and Super Resolution for Video Conferencing Applications

Project Team: Adityan Jothi, Aditi Hoskere Deepak, Srujana Subramanya

Project Summary: In this project we propose to generate photo-realistic background replaced images in real-time along with super resolution for video conference applications using GANs. We will collect data from Youtube-8M dataset and use some Zoom call recordings amongst the group. The project will involve experimenting on different GAN architectures and loss functions for Super Resolution and Background Replacement. A successful outcome would be to produce a 4-5 minute video call that is photorealistic with improved image quality.

Data Needs and Acquisition Plan: We would be using videos similar to teleconferencing calls, interviews, Twitch game streams, and more from the Youtube-8M dataset as a baseline and include several recorded videos of Zoom calls amongst the group. For the dataset generated by us, we would build an auto annotation tool using segmentation model to isolate subject efficiently in the video clips for labeling. The dataset is available open-source mapped to YouTube videos that can be downloaded using the youtube downloader tool.

Primary References and Codebase: We propose to build on the approach used in

- Olof Mogren, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," Computer Vision and Pattern Recognition (CVPR).
- Architecture: U-Net paper
- GitHub codebases: U-Net GAN Code, Fast-SRGAN Code

Architecture Investigation Plan: We plan to use U-Net with GANs for Background Replacement and SRGAN for Super Resolution. Furthermore we would try to construct a model that achieves both these tasks as a single end-to-end learning task and explore various methods to improve the quality of background-replaced super resolution images generated by the model.

Estimated Compute Needs: Based on the data set size in the above paper and the benchmarks in this original U-Net paper, we estimate that one training run for our initial U-Net GAN architecture will take 20 hours on a single Nvidia V100 GPU, which is the GPU resource in the AWS p3.2xlarge instance. With spot pricinsg, which is roughly 1$ per hour, we can expect 20$ per training run. For the SRGAN architecture we estimate that one training run will take 15 hours;

so we estimate 15$ per run. We expect to do several experiments to combine these two tasks into one unified architecture with different loss functions, hyperparameters. We estimate that this will cost approximate 40$. In addition, we expect to do approximately 4 full runs which brings our total estimated computing cost to roughly 250$. Pooling our resources, we expect to be about 100$ short of this value.

Team Roles: The following is the rough breakdown of roles and responsibilities we plan for our team:

- Adityan: Super Resolution GAN, Model Compression and Acceleration
- Srujana: Background Replacement U-Net, GAN and Image Matting
- Aditi: Background Matting Model using GANs, Auto-annotation model (Data collection and cleaning)

All team members will work on the final presentation, slides, and report.

Requested Mentor with Rationale: We request Professor Franzke to be our team mentor because he has expertise in Computer Vision and GANs. Oliver is our second choice because of his expertise in Computer Vision. We have a good idea of what we want to do and have a good starting point from the paper and codebase, so we are flexible regarding our mentor assignment.