My senior design project is to study how similarity between training and testing datasets impacts computer vision model performance in recognition tasks. It is a given in the study of creating supervised neural network models that in order to have good performance in real-world scenarios, the model needs to be trained using an ample amount of high-quality data. High-quality in this context can refer to many things, but principally, the training data should be a representative sample of the real-world data. What I would like to study is how *unrepresentative* the training data can be while still allowing a model to generalize to real-world data. A further goal is to quantify this relationship in such a way that we can make predictions of model performance for a given level of dataset quality (i.e., similarity). By accomplishing, it would allow machine learning practitioners to make informed decisions on how much resources (both time and money) need to be invested in dataset creation to achieve a given level of performance in machine learning tasks.

In my college experience, I have a taken a few different courses that have helped me develop skills necessary for actualizing this project’s goals. One example course is CS5173 – Deep Learning, which acted as my gateway into the discipline of deep neural networks. My project is primarily concerned with deep neural networks for computer vision, so the initial experience and theoretical foundations I developed in this course will be very helpful. Another course I have taken which will help me on this project is CS5135 – Learning Probabilistic Models. This course involved the study of statistical models whose parameters are learned from training data. During the course, I developed some insights into the importance of dataset quality, learning algorithms, and of machine learning and dataset evaluation from a more rigorous academic lens. These skills will likely prove essential to studying my research problem of relating dataset quality to model performance.

I did of all of my co-ops at the same firm, Etegent Technologies, where I was Junior Engineer working in Algorithm Research and Development, Full Stack Web Development, and Computer Vision. It was my experience with a mentor at the company which inspired me to pursue a research topic in computer vision, and I’m glad to have that mentor also act as my advisor on this project. At my co-op, I have learned about technical communication and developed practical skills in machine learning, such as how to handle datasets and perform exploratory data analysis, as well as use industry-standard machine-learning libraries such as PyTorch. I currently intend to use this library for this project, so my experience at work will prove very valuable. I also was able to gain insight into synthetic data generation will which be a key technique used to produce training sets of different levels of “quality”.

I’m interested to participate in this project since it presents a great opportunity to perform research in a field I have grown passionate about during my last year or two of coop and classes. Additionally, it presents a chance to examine the core machine learning problem from a different perspective. Much scholarship on computer vision and neural networks in general is model-focused, with much research focused on creating more sophisticated models, training algorithms, objective functions, and so on. Less scholarship is focused on data augmentation, synthetic data generation, and the “inputs” of machine learning in general. In addition to being a topic in a less-studied facet of machine learning, it also provides an opportunity to provide machine learning practitioners with data needed to determine how much company resources, both in terms of dollars and compute-time, need to be allocated into dataset acquisition. For many modalities (e.g., Synthetic Aperture Radar), “real-world” data acquisition is prohibitively expensive, and it is impractical to collect large, high-quality datasets. Crafting a “good enough” synthetic dataset and determining what this “good enough” threshold is can increase the utilization of this powerful artificial technique for numerous fields where data is hard to come by.

My preliminary approach to this project is focus on computer vision in modalities where real-world datasets are impractical to collect (like Synthetic Aperture Radar). Developing a methodology for quantifying the similarity of datasets for computer vision, as well as a pipeline for creating synthetic datasets of varying levels of similarity will be critical steps. With these pillars in place, research can be performed where models trained on different synthetic datasets can be contrasted, and their performance related to the amount of compute resources which went in to constructing their training datasets. A final step or stretch goal would be to generalize these findings to other computer vision and potentially any deep neural network-appropriate tasks. I expect to see that more similar datasets will perform better, but I also suspect that synthetic datasets and somewhat dissimilar datasets will perform well enough to satisfy baseline performance goals. I will know when I am done with the project when I would have to means to answer a question along the lines of “How much investment is needed to achieve a model which has performance up to X% of what we would expect if it was trained with testing data?”