Charles Hill

Senior Design

On the Impact of Train/Test Dataset Similarity on Computer Vision Model Performance

What we are studying

The Team

- Charles Hill

  | hill3cj@mail.uc.edu

- Adam Nolan, Ph.D.

  | adam.nolan@etegent.com

Abstract

- Collecting measured datasets for **machine learning** problems is logistically infeasible for many modalities and domains of research. Using **synthetic data** presents an opportunity to bring machine learning to these domains at an economic scale. The **degree of similarity** necessary between synthetic or altered data and measured data to achieve a particular level of performance has been henceforth unstudied, however. This research intends to fill this gap by examining this important, but under-studied problem.

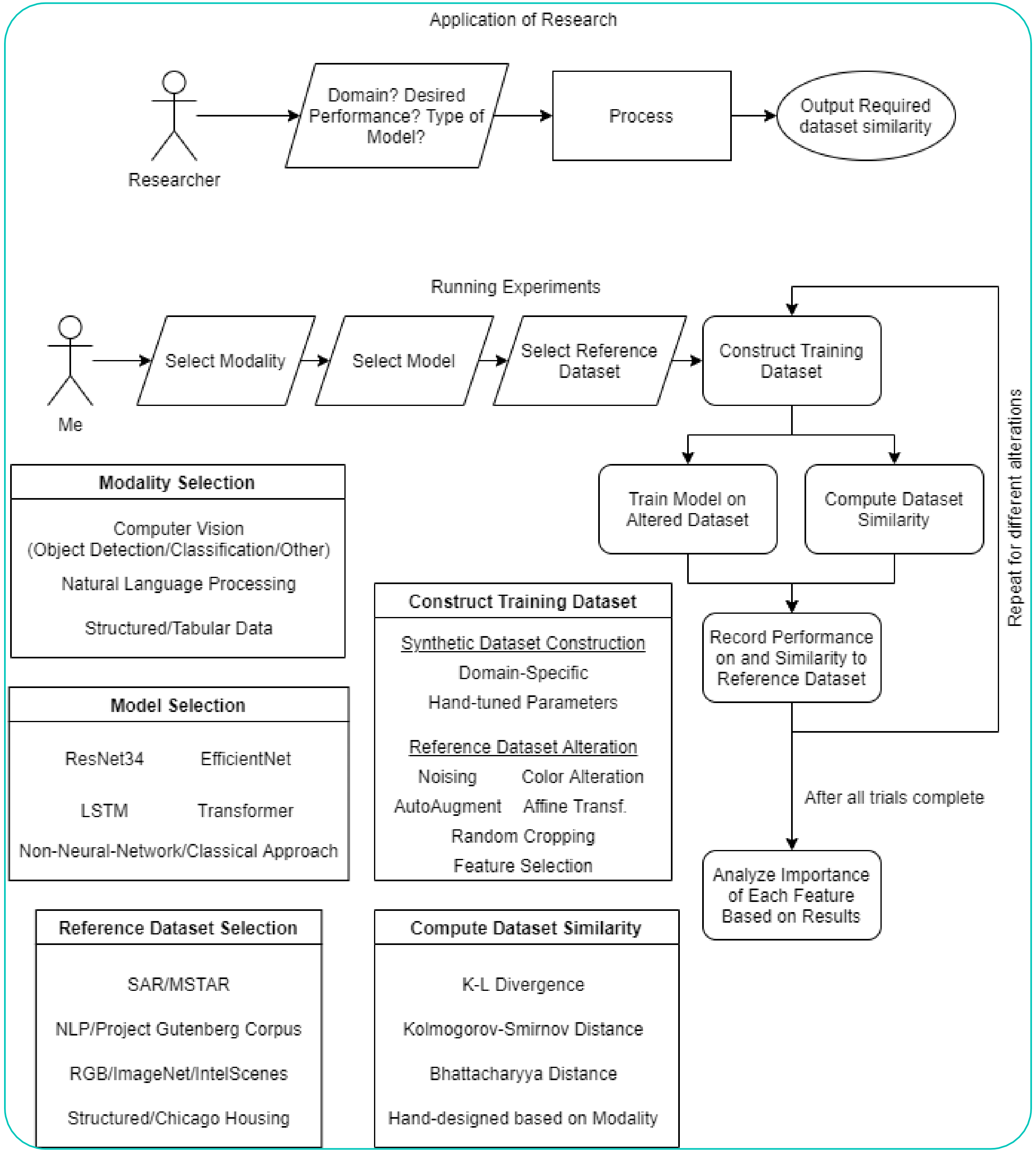
User Stories

- As a researcher, I want to know how to generate synthetic data for my experiment, so I do not have to acquire costly measured data.

- As a company, I want to know how to train a model on synthetic data, so I can use machine learning techniques in domains where measured data is hard to come by.

- As a company, I want to how to know what mixture of real and synthetic data I need to reach a given level of performance on a machine learning task, so I know how to best allocate my limited resources.

Design Diagrams



Application of Research

Researcher

Domain? Desired Performance? Type of Model?

Process

Output Required dataset similarity

Running Experiments

Me

Select Modality

Select Model

Select Reference Dataset

Construct Training Dataset

Train Model on Altered Dataset

Compute Dataset Similarity

Record Performance on and Similarity to Reference Dataset

Repeat for different alterations

After all trials complete

Analyze Importance of Each Feature Based on Results

Modality Selection

Computer Vision
(Object Detection/Classification/Other)

Natural Language Processing

Structured/Tabular Data

Model Selection

ResNet34 EfficientNet

LSTM Transformer

Non-Neural-Network/Classical Approach

Reference Dataset Selection

SAR/MSTAR

NLP/Project Gutenberg Corpus

RGB/ImageNet/IntelScenes

Structured/Chicago Housing

Construct Training Dataset

Synthetic Dataset Construction

Domain-Specific

Hand-tuned Parameters

Reference Dataset Alteration

Noising Color Alteration

AutoAugment Affine Transf.

Random Cropping

Feature Selection

Compute Dataset Similarity

K-L Divergence

Kolmogorov-Smirnov Distance

Bhattacharyya Distance

Hand-designed based on Modality

Major Project Constraints

- Economic Costs
  - Compute facilities
  - Possible cost of proprietary datasets and libraries
- Scope
  - Ambiguity in time requirements
- Professional and Technical Expertise
  - Machine Learning is a constant evolving field

Review of Project Progress

- Prerequisite Planning
  - Initial project infrastructure

- Researching requisite background knowledge
  - Statistical methods
  - Publicly-available datasets
  - Dataset alteration and synthetic dataset generation techniques

Expectations for the End of the Term

- Complete All Research Tasks:
  - Compile lists of publicly-available datasets
  - Compile list of suitable data augmentation strategies for each modality
  - Determine the range of data augmentations for experimentation
  - Research methods for constructing synthetic datasets for each modality
  - Research methods for computing dataset similarity for each modality
  - Research reference ML models for each modality
- Get Results for the SAR Modality Published in SPIE – Journal of Applied Remote Sensing

Expected Demonstration for the Expo

- Presentation of Research Results and Graphics Deliverables
  - Discussion of dataset similarity methodologies
  - Discussion on which modalities are most compatible with synthetic data
  - Knee-in-the-curve analysis of dataset similarity vs. model performance
- Demonstration of Published Findings