On the Impact of Train/Test Dataset Similarity on SAR-ATR Performance

Charles Hill | hill3cj@mail.uc.edu

Advisor – Adam Nolan, Ph.D. | adam.nolan@etegent.com

Goals

- Quantify the relationship between similarity of datasets used for training and evaluation for Synthetic Aperture Radar modality Automatic Target Recognition (SAR-ATR) systems
- Determine which aspects of synthetic data have the greatest impact on model performance
- Enable evaluating cost efficacy of allocating additional compute time to data generation
- Create platform to perform analysis on other modalities and domains

Intellectual Merits

Research in computer vision tends to largely be model-centric, with emphasis on discovering new model architectures or tweaks to existing architectures. Data-centric research provides a lesser pursued, but nonetheless worthwhile avenue to enhancing machine learning model performance.

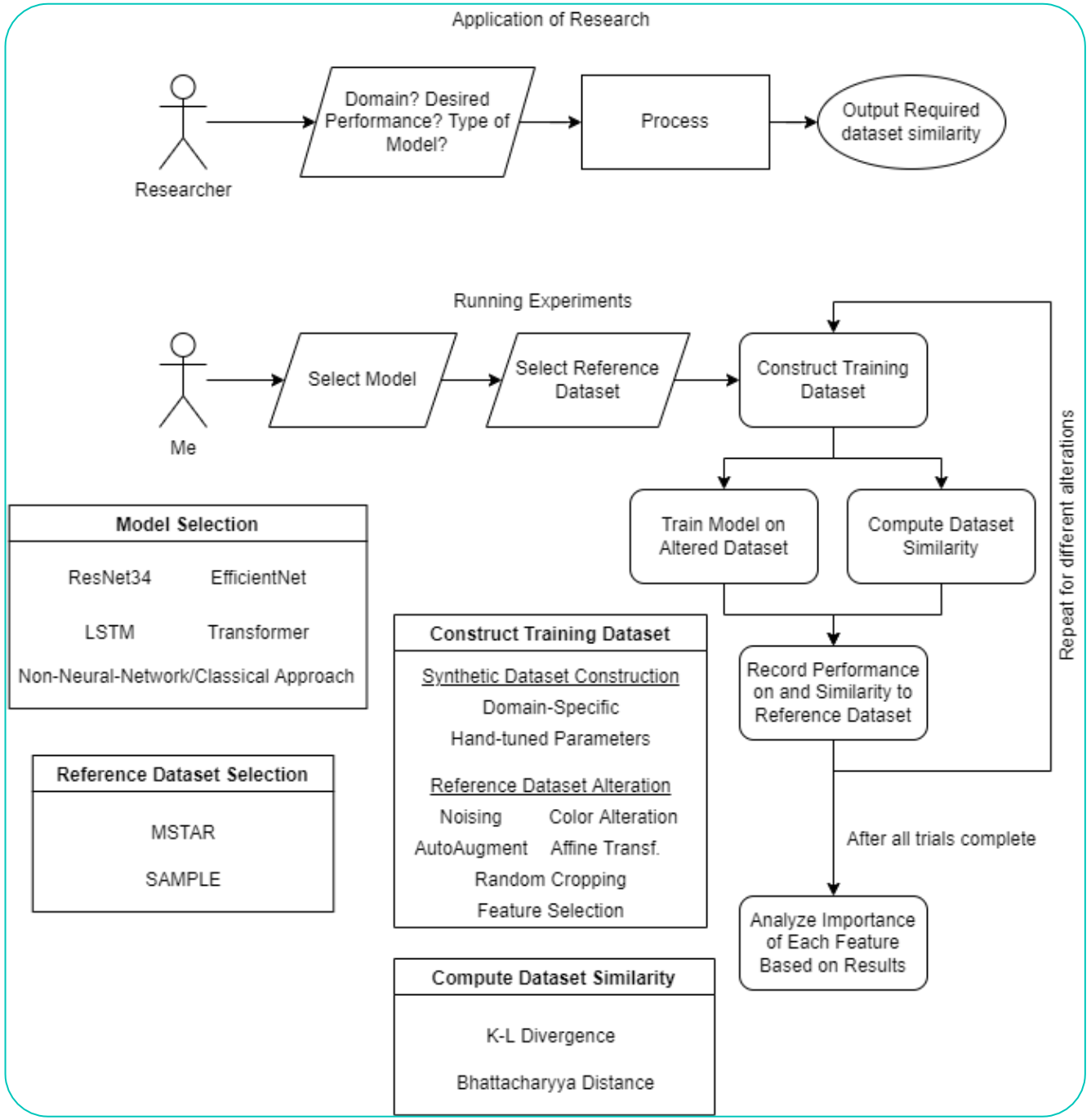
The topic of this project, which studies the degree of similarity necessary between synthetic and measured data to achieve a particular level of model performance, has been henceforth largely unstudied in the scientific literature and is thus a novel approach to improving model performance in industry.

Broader Impacts

For many modalities, it is simply infeasible to curate a training dataset of sufficient size and complexity in order to train modern computer vision models. Greater utilization of synthetic datasets can empower researchers and practitioners to use deep learning techniques to solve problems in these domains and modalities at greater computational cost, but lesser overall costs.

This project aims to provide data on the SAR modality specifically, and to create a pipeline other researchers can use for computing target metrics for their modalities of interest

Design Diagrams



System Overview

The structure of experimentation is to train a computer-vision model on SAR-ATR (i.e., image recognition) tasks using synthetic data. Once training has completed, it's performance is evaluated on either more computationally sophisticated synthetic data or measured (i.e., real) data. The performance of the learner is then correlated with the computed similarity between the two datasets using domain-specific implementations of KL-Divergence and Bhattacharyya Distance measures. Bootstrapping is further utilized on the training dataset to determine the degree of error in these measurements.

The fundamental design of the experiment structure is framework agnostic, however, discussion on specific technologies utilized in the research is discussed on the following slide.

Technologies

The research conducted in this project used Python3/PyTorch for the entirety of the model code. Experimental trials were arranged, orchestrated, and documented using an internal, proprietary deep learning experiment management tool, ATLAS. During testbed creation, numerous modules were contributed to this framework, such as modules which handle loading and pre-processing SAR data and handle the unique needs of computing dataset similarity measures.

Data processing techniques to allow usage of KL-Divergence and Bhattacharyya Distance on SAR imagery were also developed, and entail operating on the distribution of higher-level features in the imagery rather than on the pixel-level data.

Milestones

- 31 Dec 2021 – Complete Background Research [Complete]
- 14 Jan 2022 – Complete Testbed Development [Complete]
- 15 Feb 2022 – Complete Computation Trials

  [Delayed but completed]
- 15 Mar 2022 – Complete Data Analysis and other Project Deliverables
- 1 Apr 2022 – Get Public Release Clearance from DoD

  [Forthcoming]

Results

- Completed integration into Etegent Technologies' proprietary deep learning experiment framework, ATLAS.
- Implemented dataset similarity measures appropriate for SAR imagery
- Determined which features of SAR data are most important for recognition performance – Getting imagery of similar articulation
- Publication in the Journal of Applied Remote Sensing is forthcoming.

Challenges

- One of the most significant challenges overcame on this problem was determining how to compute dataset similarity for SAR-ATR tasks. I overcame this challenge by adapting standard statistical techniques for distribution similarity/distance (e.g., KL-Divergence) to operate on distributions of higher-level features of each image which are governed by parameters of the synthetic data generator, This dramatically reduces the dimensionality of the problem and makes this exercise feasible.
- Creating a system to generate synthetic data with controllable but internally-consistent and random distributions of higher-level features was a difficult problem but was overcame using a Bayesian network with reasonable defaults to generate image parameters, and proprietary software to simulate remote sensing.