On the Relationship of Train and Test Dataset Similarity on SAR-ATR Model Performance

University of CINCINNATI

Sponsored By: Etegent Technologies

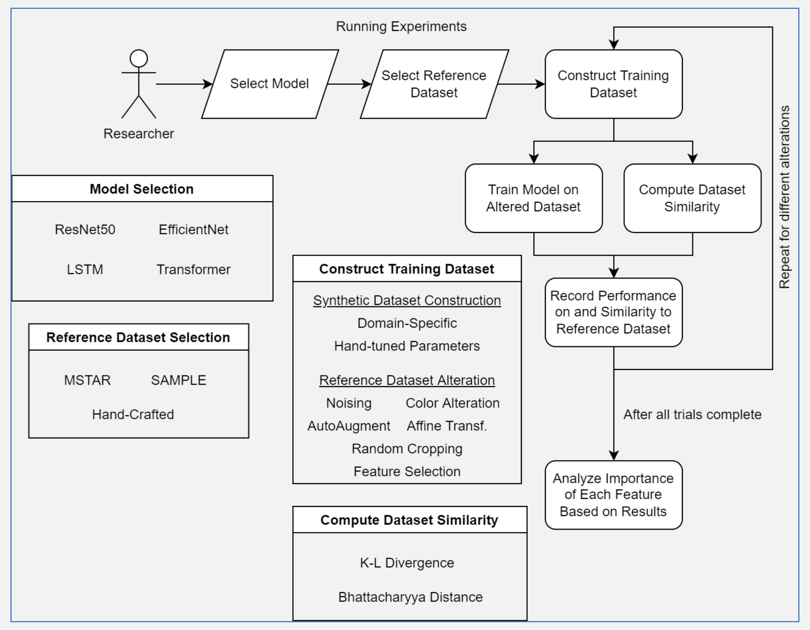
MOTIVATION

Machine Learning requires a lot of data, but for many problems, collecting real-world data (“measured” data) is not feasible or practical. One method of alleviating this is investment into “low-shot” supervised learning (that is, learning with few training examples). However, another technique which has been examined less thoroughly in the scientific literature is usage of wholly synthetic datasets to augment training. This research intends to study this relatively under-appreciated area to increase the scope of machine learning applications.

OBJECTIVE

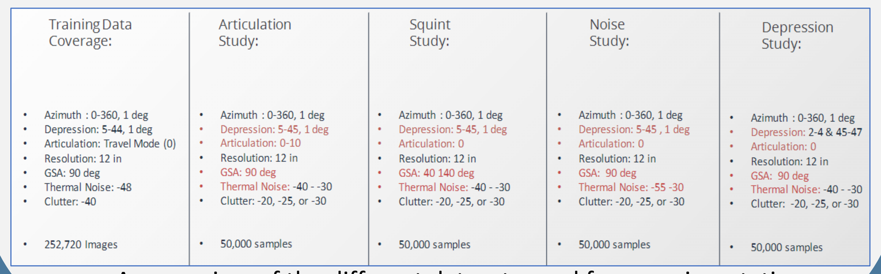
The objective of this research is to determine the relationship between the similarity of training and test datasets and eventual performance achieved by a deep computer vision model on recognition tasks. This would allow future researchers and practitioners to estimate how much effort needs to be expended in dataset curation to achieve a particular level of performance.

EXPERIMENTAL OVERVIEW



Flowchart of the Experimental Process

The overall experimental framework involves training a computer vision model on a synthetic dataset, computing the similarity of the synthetic dataset to a reference dataset, and noting the performance that model achieves on the baseline dataset. Then, by varying the dataset used for training, a relationship between these factors can be discovered.



An overview of the different datasets used for experimentation

COMPUTING DATASET SIMILARITY

The second major challenge of this research was determining a rigorous method for computing the similarity between synthetic SAR-ATR datasets. To accomplish this, we devised a method to adapt existing statistical techniques to our use case.

KL-Divergence is a common statistical technique for reporting the amount of divergence (analogous to distance) between two statistical distributions of the same domain.

$$D_{KL}(p||q) = \int_x p(x) log \frac{p(x)}{q(x)} dx$$

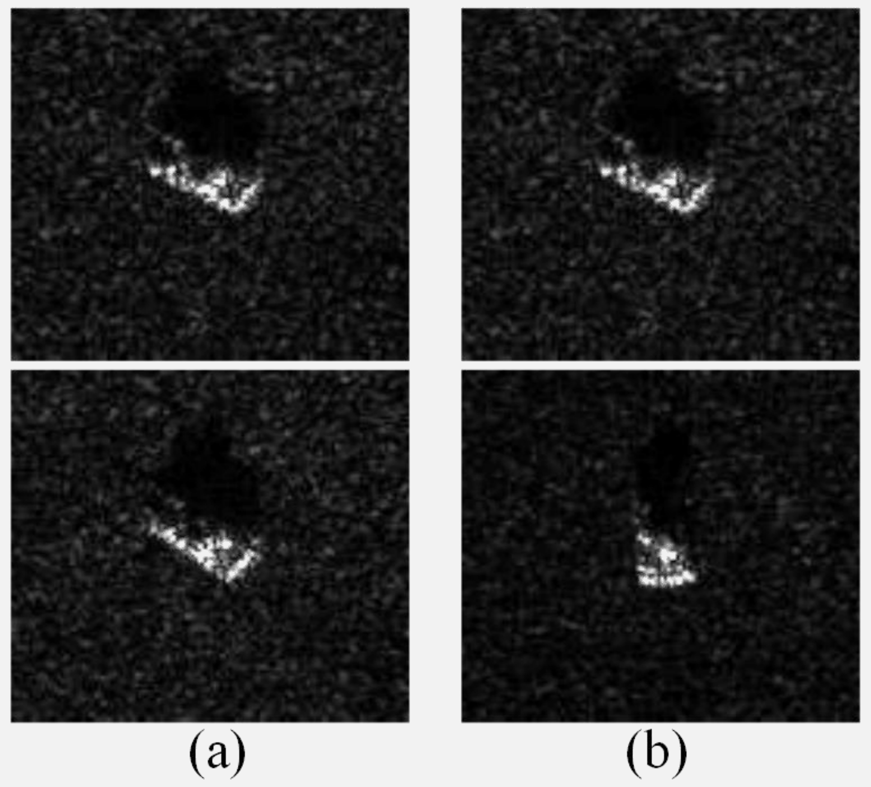
Critically, KL-Divergence is summable between independent distributions, easing the computational burden. This method is then applied to distributions of operating-characteristics in each dataset, rather than distributions of any pixel-level data.

SYNTHETIC APERTURE RADAR

The research conducted in this project focused on the domain of Synthetic Aperture Radar imagery for purposes of Automatic Target Recognition (SAR-ATR).

SAR is an active remote sensing technique where electromagnetic waves are broadcast at a surface and the reflected waves are used to construct a representation of that surface.

In Automatic Target Recognition systems, the goal of the model is to identify the type of vehicle in the image. It should be able to do regardless of the parameters of the imagery collection (e.g., azimuth, elevation, and so on).



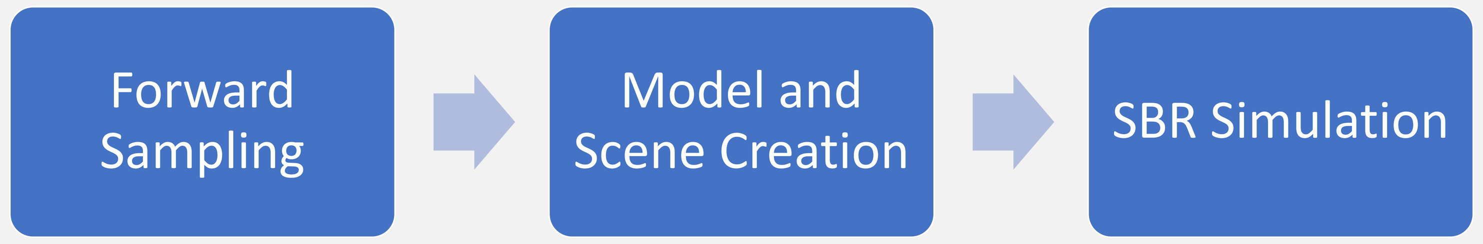
SAR image of a T-72 tank from the MSTAR dataset. (a) Two images from similar azimuthal directions. (b) Two images from different azimuthal directions.

SYNTHETIC DATASET GENERATION

Generating synthetic SAR data with varied, but reasonable and internally-consistent parameters was a significant challenge. Our solution for the SAR modality was a multi-step-process:

1) Generate the operating characteristics of each synthetic image by forward sampling from a hand-tuned Bayesian Network which governs the distributions and relationships between parameters such as azimuth, elevation, type of vehicle, degree of perturbation to the vehicle, amount of noise, and others.
2) Use the list of operating characteristics for each image to parametrically modify a 3D model of the vehicle and its immediate environment.
3) Feed remaining operating characteristics and the 3D model to a shooting-and-bouncing ray simulation to estimate the vehicle’s radar cross-section.

We believe this approach could be generalized to other domains within remote sensing and is a principal contribution of this research.



RESULTS

Computational trials suggest a positive, but non-linear relationship between dataset similarity and model performance. Using the five datasets outlined in the **Experimental Overview** panel, testing suggests that in the SAR-ATR space, degree of articulation has the greatest impact on model performance among the OCs studied, such that mixing that feature into the training set had the most dramatic impact on model accuracy.

More precise results are awaiting public release and publication in the *Journal of Applied Remote Sensing.*

Trial	Accuracy
Baseline	70%
Baseline + Studies	96%
+ Depression Study	87%
+ Squint Study	87%
+ Noise Study	88%
+ Articulation Study	91%

TEAM INFORMATION

Charles Hill
Computer Science

Adam Nolan, Ph.D.
Advisor - CTO