ROCHESTER
DATA SCIENCE CONSORTIUM

2019 ENVI Analytics Symposium

W. A. L. Johnson, PhD
Executive Director
August 15, 2019

EAS
ENVI ANALYTICS SYMPOSIUM

Our Charter

- Non-profit organization founded by University of Rochester and Harris Corporation in 2017
- Funded by New York State beginning in September 2017
- Goals: To use Data Science (”Big Data”, AI, Machine Learning, etc.) to
  - Create jobs in the FLX region
  - Create economic impact by enabling local companies to understand and implement DS solutions for pressing business needs (anywhere)
- Not just another Academic-Industry Partnership

Drivers

- Progress driven by advances in Data Science – AI, ML, Virtual Reality, Autonomous Transportation, etc., are now changing the world at a rapid pace
- As a new discipline, “Big Data” is difficult to understand, implement, and leverage – even though the results can be game-changing
- Worse, there is a scarcity of senior talent with commercial experience
  - Mostly resides in Academia or High-Tech Hubs
  - Unaffordable for most SMBs
  - Grass roots
- “Greed and Fear”

University of Rochester Data Science

GIDS

Goergen Institute of Data Science

Faculty,
Bachelors
&
Masters
Programs

NY State
Center of Excellence
in Data Science

RDSC

Companies
&
Institutions

Academic → Commercial

Current Membership

Staffing

We have technical depth in Statistics, Bayesian Statistics, Deep Learning, Artificial Intelligence, Machine Learning, Image Understanding and Computer Vision, Natural Language Processing, and Commercial and Academic project scoping & execution

- 10 PhD research scientists
  - Deep Learning (AI), Computer Vision, Signal Processing, Bayesian Statistics, Natural Language Processing, Geospatial Analytics
- 2 MS research scientists
  - Machine Learning and Natural Language Processing
- 3 Full Time Administrative Staff
- 13 Current Interns (BS, MS, PhD), many others graduated/exited

Data Assurance – L3Harris

As we get more and more information from different sources the possibility of “fake data” being consumed grows

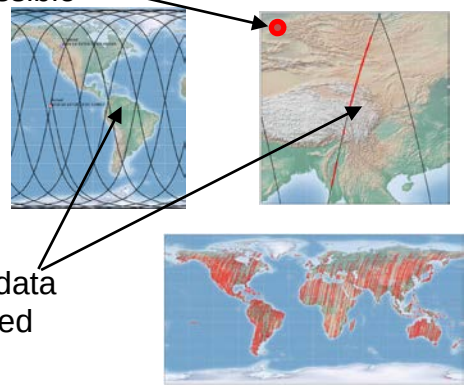
- Harris is characterizing different types and levels of attacks on the data
- Assessing deep fake detection used in social media
- Looking for an end-to-end solution
- AI techniques for detecting forgery
- Watermarking/hashing/block chain for chain of custody/protection
- The better the Fake the more complicated the detection algorithm will be

Original

Fake Image

Fake/Edit Metadata:
Don’t allow users to find the right data.

Not possible

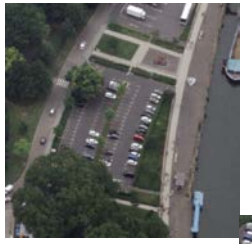


Metadata Derived Orbit

Change Date or Location in Metadata

- Simple changes can be detected against system’s orbit
- We will need more information if the hacker has knowledge of orbits and capability

Edit Imagery:
Add or Remove objects to hide information.



How do you Detect

- Simple Copy Move
- Some Detected with internet tools

While Humans would not be fooled with this fake image - AI detections are easily fooled

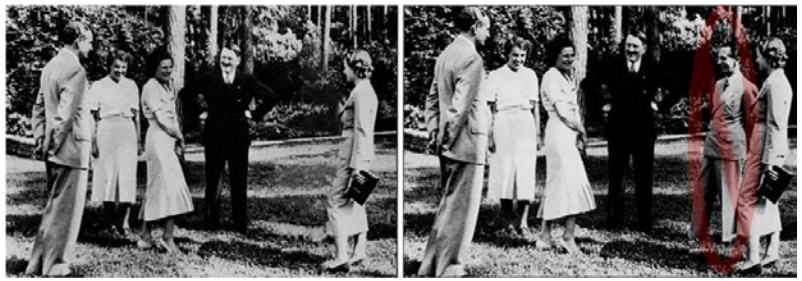
The Deep Fake tools being developed for internet fakes will not work on good fake overhead imagery

Classic Disinformation of Governments with Image Fakery

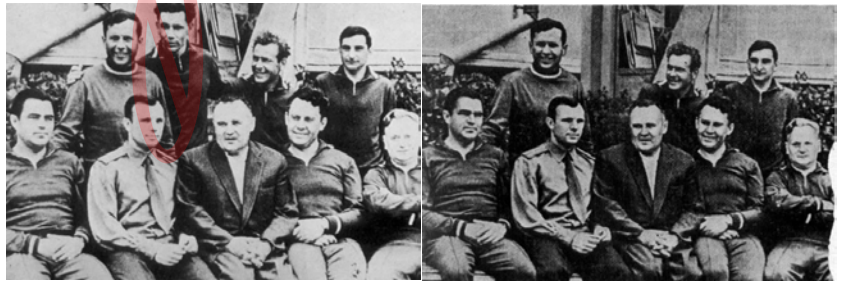
Many of the classic film techniques were advanced by the Soviet Union and Nazi Germany

- Main purpose of these images were to deceive citizens of their own country
- Digital technology makes it easier

Joseph Goebbles – Where did he go?



The missing cosmonaut (1963)



Grigoriy Nelyubov was removed from the space program because of bad behavior and from this historical picture of the original cosmonauts.

Photo from the unveiling ceremony in Tehran

+

Mount Damavand from PickyWallpapers web page

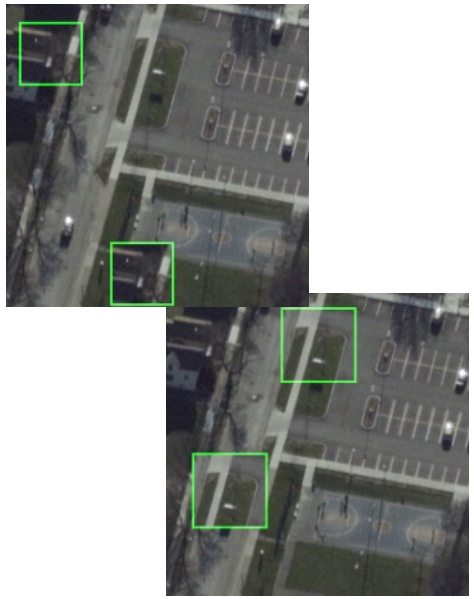
=

Iranian government released image of their Qaher-313 stealth fighter in flight (February 13, 2013)

Examples of Forged Data

Developing training and test data to help develop AI detection algorithms

- Automatic copy move – different size, number of copies
- High quality copy move
- Synthetic targets inserted (DIRSIGSimulations)



University of Rochester Medical Center

Parkinson's Disease Progression Modeling

Exploring Parkinson's Disease Progression using Deep Generative Models Built on DaTscan Brain Imagery

Trevor W. Richardson, Beilei Xu, Wencheng Wu, Lei Lin, Charles Venuto, Monica Javidnia, Edgar A. Bernal

Rochester Data Science Consortium, Center for Health & Technology, Rochester, NY

ABSTRACT

For Parkinson's Disease (PD), which affects one percent of the population over sixty, no objective biomarkers for diagnosis or progression have been validated to date. The current approach to measuring PD disease progression is the Unified Parkinson Disease Rating Scale (UPDRS). The UPDRS scoring mechanism is a subjective quantitative scale, and previous research has shown that the variance of an individual's score assigned by different doctors is unreasonably high. This measurement variance adversely affects the quality of the data collected from expensive and lengthy clinical trials on PD and impedes the community of researchers focused on solving problems centered on PD. We propose constructing a new method for scoring disease progression and diagnosis using state of the art data-driven deep generate models in the hopes of alleviating the aforementioned problems. The approach exploits previously underutilized image data, such as DaTscan imagery, provided by the Parkinson's Progression Markers Initiative (PPMI) dataset. Not only would this methodological approach provide a way to minimize or eliminate variance injected by subjective human measurements, but would also potentially provide a novel avenue for constructing new medical tools and expand our understanding of PD progression.

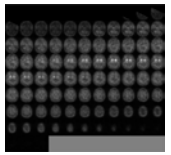
Background

Parkinson's Disease is a degenerative disorder that affects the central nervous system. Major symptoms include and are not limited to tremor, loss of automatic movement rigid muscles, depression and cognitive impairment. Recently, a large amount of resources have been focused on developing proper clinical and observational trials in which many forms of longitudinal data are collected for further research. Parkinson's Progression Markers Initiative (PPMI) is one of these landmark studies that is still following cohorts of interest. One focus of PPMI is on studying disease progression and some of the data types collected for subsets of their patient and healthy populations include medical images, biological samples and clinical assessments.

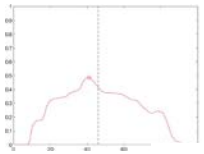
Data

This research limited data to the subpopulation of PPMI where a patient had both DaTscans and UPDRS scores.

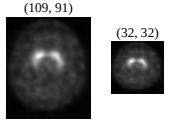
Individual DaTscan Image



Peak selection method for individual input image to model

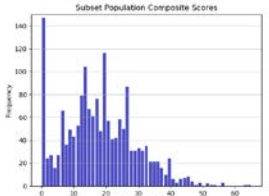


Each brain is resized for computational purposes



Tremor (UPDRS) score

0 = Absent
1 = Slight and infrequently present
2 = Moderate; bothersome to patient
3 = Severe; interferes with many activities
4 = Marked; interferes with most activities



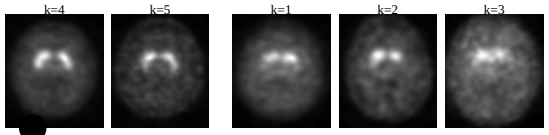
METHODS

Flow-based deep generative models are the primary methodological approach for learning the pdf of the DaTscan image distribution. This method was chosen over other generative modeling approaches due to the following mathematical properties of normalizing flow models: 1. exact log-likelihood computation; 2. exact inference of latent variables; 3. exact sampling; and 4. interpretable latent space. The target prior distribution chosen for Z is a spherical multivariate gaussian. The transformation f is an invertible composition of functions learned using the change of variables rule. X is mapped to Z using f. On top of the modeling, other techniques have been used to analyze and validate the approach such as PCA, clustering analysis and correlation analysis.

$$X \sim p^*(X) = ? \quad Z \sim p_\theta(Z) = \mathcal{N}(\mu, \sigma^2) \quad X_i = f_\theta(Z_i; \theta)$$

RESULTS

Clustering the Embedding Space



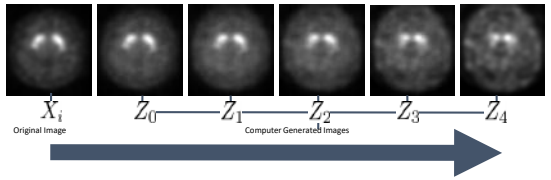
K-medoids clustering used on embedding space in order to find the most representative Tscan images

Healthy control (HC) image appears at *k* = 4; implying higher intra-class variance for PD DaTscan images than between PD and HC

Simulating Disease Progression -- Manipulating Latent Features

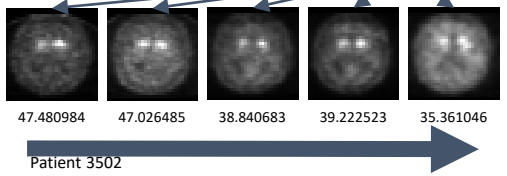
Acquire embedding Z from image X and then scale by directional tensor d

$$X_i = f_\theta(Z_i; \theta) \quad Z_{i+1} = Z_i + \alpha \odot d$$

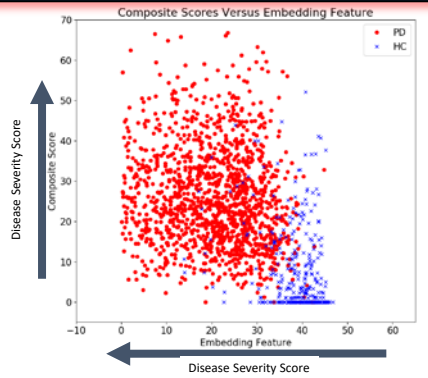


Euclidian Distance to the Medoid Embedding

$$\|Z_{medoid} - Z_i\|_2^2$$



Composite Scores Versus Embedding Feature



Cross-correlation - 0.382

$$e = \|d - p\| \qquad p = d\frac{x^i \cdot d}{(\|d\|)^2}$$

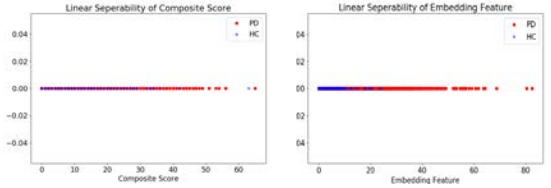
$$\bar{x}_{pd} = \frac{\sum(x^i_{pd})}{N_{pd}} \quad \bar{x}_{hc} = \frac{\sum(x^i_{hc})}{N_{hc}} \qquad d = \bar{x}_{hc} - \bar{x}_{pd}$$

Logistic Regressor Predicting Diagnosis Using Single Feature

Composite Score : 90%

Embedding Feature : 93%

Both 94%



Future Work

The direction of this research is now on designing new loss functions that incorporate information from labels into the learned transformation of the images into the embedding space. In addition, we are interested in developing models that identify individuals who have a high probability of fast progressing Parkinson's disease.

REFERENCES

Burke et al. BMC Cancer (2015) 15:929

Cause and Effect Strategic Marketing

- Sr. PGA
  - Goal: Develop model-based marketing target predictor
- Dealer Location prediction
  - A large Pro-Sumer manufacturing company
    - Goal: Model driven predictor for new dealer locations
    - Result: 300% uplift in responses compared to their predictions

Thank you!