DATA SCIENCE AT PGSC

Scott Grigsby, PhD
Director, Data Science
Intelligence and Readiness Operations

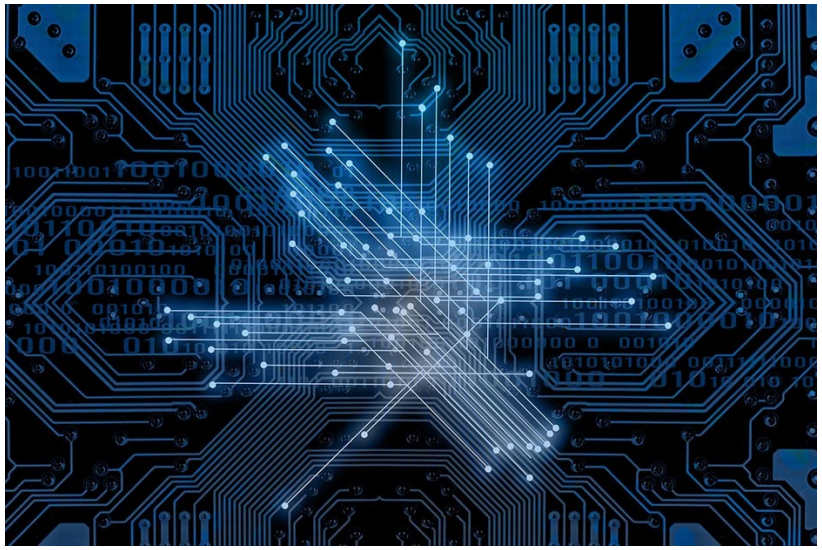
PAR GOVERNMENT™

Who is PAR?

- 430 Employees
- $70M Annual Revenue
- Headquarted in Rome, NY
- ISO 9001:2015 Certified
- Small Business under 1,500 employee size standard

PAR → Pattern Analysis and Recognition

- Company was established over 50 years ago to focus on Pattern Analysis and Recognition (PAR) problems
- **PAR Government Systems** is a subsidiary of **PAR Technology** that focuses on the Federal Government
- PAR did some of the earliest DOD research in neural networks for image processing in the 1970's and 80's; and today's customers include research (DARPA, AFRL), as well as operational and IC customers



PAR's Core Data Science Foci

AI/ML Research, Systems Integration, and Systems Engineering

PAR Government's scientists and engineers perform **basic and applied research** for our DOD and intelligence community customers in conjunction with leading **research universities**. We also perform **systems integration and engineering** on large-scale complex computing and data architectures.

Data Science & Analytics

AI / Machine Learning

Edge Computing Applications

High Performance Computing

Digital Forensics

Human-Machine Teaming

CAPABILITIES RELEVANT TO ARTS

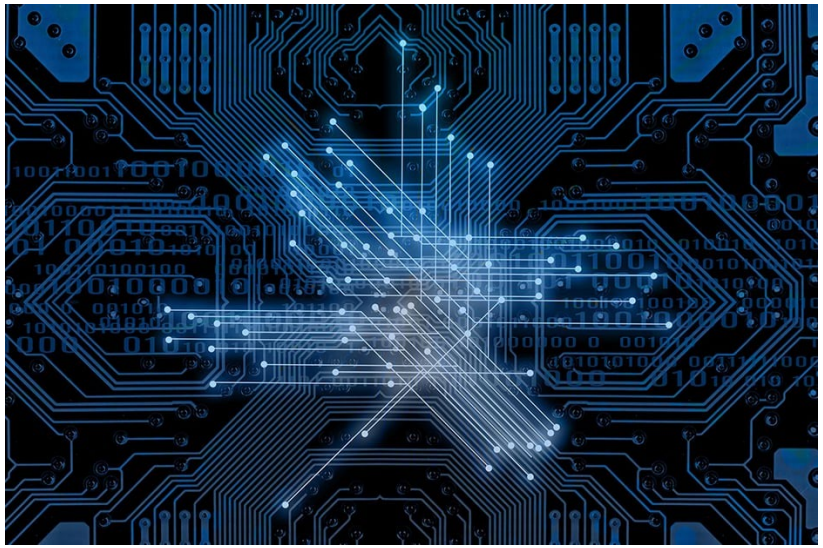
PAR GOVERNMENT™

Semantic Forensics (SemaFor) [DARPA]

Program Goal: Create rich semantic algorithms that automatically detect, attribute, and characterize falsified multi-modal media to defend against large-scale, automated disinformation attacks

PAR's Role:

- Create multi-media test datasets, test and evaluation infrastructure
  - Media includes:
    - Text
    - Audio
    - Images
    - Video
- Real (pristine and manipulated) and synthetic examples and mixtures
- Also look at semantic inconsistencies between media
- In audio domain:
  - Developed data sets of:
    - manipulated audio (cut/splice, etc.)
    - Generated/synthetic audio for detection and attribution (POIs)



Novelty in Open Worlds (SAIL-ON) [DARPA]

Program Goals

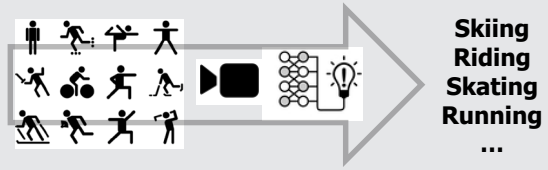
- Develop scientific principles to quantify and characterize novelty in open world domains
- Create AI systems that act appropriately and effectively in open world domains

PAR Role

- Evaluation and characterization of adaptive agents
- Curate datasets for novelty in:
  - Video Activity Recognition
  - Writer Identification (text not audio)
    - Reviewer's using false identities
    - Understand, detect, falsify a reviewer's style
    - Handwriting recognition (word choice, pen pressure, letter angle spacing, etc.)

Open worlds have novel situations that violate implicit or explicit assumptions in an agent's model of the external world, including other agents, the environment, and their interactions.

(1) Agent Task: Determine the human activity performed in video clip.

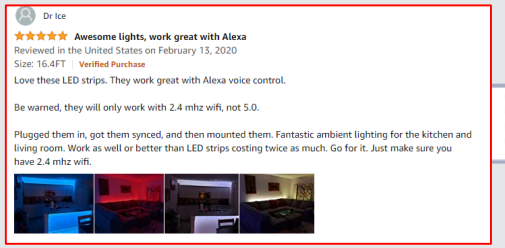


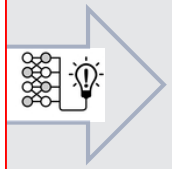
Novel Videos: Those videos of activities where generalized closed set activity recognition videos fails under novel conditions.

Approach:

1. ***Collect and annotate, or synthesize, videos of activities, with attributes defined by an ontology over visual activities and their associated environments.***
2. ***Map the Novelty Hierarchy to the associations of attributes to activity.***
3. ***Generate tests to evaluate TA2 Agents with novel presentations of these attributes.***

(2) Agent Task: Determine if the reported reviewer of a written product or service review is correct. If not, determine the real reviewer.


Dr Ice
Awesome lights, work great with Alexa
Reviewed in the United States on February 13, 2020
Size: 16.4FT | Verified Purchase
Love these LED strips. They work great with Alexa voice control.
Be warned, they will only work with 2.4 mhz wifi, not 5.0.
Plugged them in, got them synced, and then mounted them. Fantastic ambient lighting for the kitchen and living room. Work as well or better than LED strips costing twice as much. Go for it. Just make sure you have 2.4 mhz wifi.

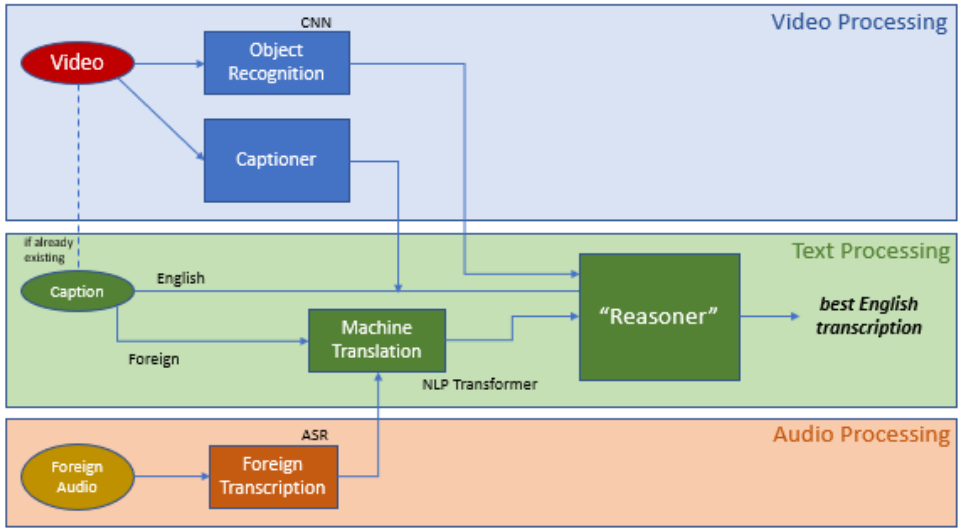


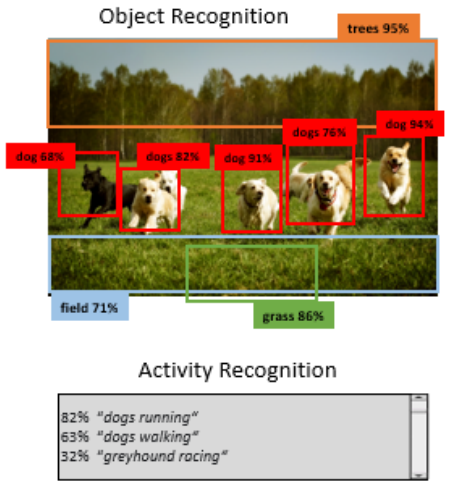
Correct/Incorrect?
Who is the real author of the review?

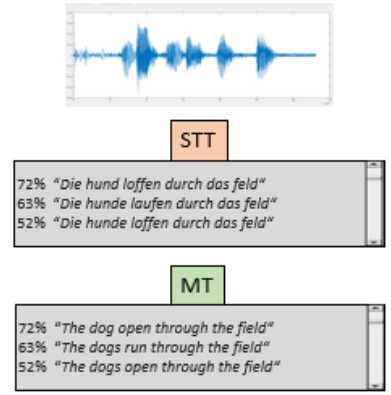
Novel Reviews: Those reviews with novel attributes and patterns out of distribution from prior known writers and reviews.

Multi-Modal Machine Translation (M3T) [AFRL/RH]

Program Goal: Can higher quality machine translation (MT) output be obtained using deep learning neural networks trained on multimodal data?







Scope

- Current MT systems take textual or audio input from a source language and provide textual translated output into a target language based on models trained on either text or audio corpora only.
- Other data types, such as corpora of bilingual audio translations or images included as part of the file are not added to the models.
- The hypothesis is that models trained on multimodal data or data derived from multiple channels will yield higher quality translations.

Approach

- Develop end-to-end multi-modal generative networks for improved machine translation
- Develop techniques for early and late merging of outputs, selecting from an N-best list of candidate translations
- Extend image processing to video processing
- Define semantic labelling of images for multimodal datasets
- Incorporate automatic speech recognition (ASR) features to augment machine translation
- Develop methods for overcoming missing/incongruent data

Summary: PAR's Experience in Data Science

Deep history in digital imagery, video, and audio – Pattern Analysis and Recognition (PAR)

- PAR's work in studying and understanding digital media goes back to the early 1970's
- Scientists who have studied human perception and emotional response to media
- Decades of published technical papers and technical reports in the subject area
- Support to multiple programs at DARPA, AFRL, and the Intel Community

Close Connection to Subject Matter of DOD/Intel applications

- Available experts in tactical operations and intelligence to support data scientists
- History of working with tactical users in training, mission rehearsal, and operations

Data Set Development

- PAR specializes in large, multi-media data sets to support ML training and testing: Large amount of multi-media data available from previous collections
- Sources include representative open source, high provenance, and synthetic data production
- Annotation tools and production-scale automation

Data Science Tools and Framework

- Development of tools to enable seamless cross HW process orchestration and data exploration from 10GB to 100TB+ scale (e.g., Arkouda) (including HPC)
- Tools for collection, synthesis, and curation of content
- Experience in data collection in LVC experiments and training environments
- Analytic test harnesses for training (with and without feedback) and testing

Data Algorithm Development

- Specializing in multi-modal algorithm development (text, audio, images, and video)
- Expertise in transformer, diffusion, and generative models for areas such as Machine Translation and Media Forensics

Questions?

Looking to team. Could Prime or Sub.

Contact Information:

Scott Grigsby, PhD

Director, Data Science

937-657-6575

scott_grigsby@partech.com

PAR Government

1430 Oak Ct, Suite 209
Beavercreek, OH 45430

160 Brooks Road
Rome, NY 13440