Mutual Information-based Analysis of Latent Neural Representations for Interpretable Computer Vision

This proposal is being submitted to BAA Topic **EL-24 (paragraph 2)** with reference number **BAA-EL-2024-0007**.

Duration of the project: 1 year

Principal Investigator: Ryan T. White, PhD – (321) 848-8301 – rwhite@fit.edu
Assistant Professor, Mathematics & Director, NEural TransmissionS (NETS) Lab
Department of Mathematics and Systems Engineering, Florida Institute of Technology

Executive Summary

In recent years, artificial intelligence (AI) has permeated nearly every scientific and engineering discipline. The commercial technology sector is largely built on its recent developments and capabilities. AI has become essential, from the way we manage national defense to the everyday use of our personal devices.

Most recent developments in AI are based on the extraordinarily powerful machine learning paradigm of deep learning (i.e., neural networks). For imaging and visual tasks, from image classification to object detection and beyond, neural networks have far surpassed competing approaches nearly uniformly.

Despite their performance, neural networks for computer vision have some shortcomings:

- Neural network decision-making process is not generally human-interpretable.
- Building neural works that perform well and generalize accurately is not straightforward.
- Neural networks trained on still images do not always work well on video feeds.

The proposed research will dissect convolutional neural network (CNN)-based computer vision models to address all three challenges. We will develop information theoretic techniques to visualize the complex web of interconnections within the networks to shed light on the decision-making process to increase interpretability. This will lead to an understanding of optimal information flow patterns in well-performing models, which will be used to guide design choices for effective neural models. This will further be exploited to upgrade static image-based computer vision models to effective temporal models that work well on video feeds. This work is important for learning from visual data across critical applications from environmental and vegetation assessment by UAVs to geospatial intelligence to visual inspection of infrastructure.

The proposed work is broken into two phases, one that develops new tools to analyze neural networks to identify and encourage optimal information flow patterns, which enables practical applications of the tools in the second phase.

- **Phase 1:** Expand beyond entropy-based metrics and implement GPU-accelerated paired-pixel mutual information (PPMI) to analyze the hidden representations within deep learning models to build an understanding of the impact on mutual information patterns on performance. Develop *mutual information-informed neural networks* (MI2N2s) that exploit patterns found in mutual information in their architectures and loss functions.
- **Phase 2:** Explore applications of both PPMI and (M)I2N2s by incorporating loss functions and visualization techniques in training schemes. Use localized PPMI to perform semi-supervised

object tracking. Develop entropy- and MI-based model diversification techniques to build powerful ensemble models for diverse vision tasks.

Prior Work

Early work done by M. Meni through the NSF Mathematical Sciences Graduate Internship with the U.S. Army Corps of Engineers showed how neural networks took highly disordered input data and transformed them into orderly binary representations. Under previous ERDC collaboration, M. Meni and R. White then derived new mathematical formulas that map the changes in entropy as inputs flow through the layers of dense and convolutional neural network layers. They used the optimal patterns of entropy flow experimentally found to exploit these mathematical formulas through a novel loss function. This loss function controls the entropy flow in designated layers to either encourage or discourage entropy gain. It was shown that networks trained with this loss function converged faster, achieved increased accuracy, and reduced the model's complexity.

While this work showed entropy flow patterns through the neural networks, it was unknown what this meant for how the input images were being transformed. To explore visualization of the data as it moved through these models, M. Meni and R. White developed a technique called Probabilistic Explanations of Entropic Knowledge extraction (PEEK). This method is a low-compute calculation that analyzes the latent representations of inputs as they move through trained networks. This method provides the ability for real-time monitoring. Also, it was shown that it can automatically locate biases in experimental data collection and offer potential solutions to common failure modes of object detection models. In addition, M. Meni and R. White utilized PEEK for diverse visual tasks, such as weakly-supervised object detection and improving well-known segmentation model U-Net by allowing it to learn with PEEK. This prior work has offered some insights into cause of bias, new training processes, and a way to visualize patterns inside of neural networks.

Statement of Work

Phase 1: Mutual Information-based Analysis and Guidance of Neural Networks ***(Months 0-6)***

The first stage seeks to uncover patterns in the internal data representations and processing in effective deep learning models. We developed the paired-pixel mutual information (PPMI) measure as an analog of entropy-based methods that uses mutual information to compare “pixels” of different features maps in CNNs. We have demonstrated PPMI encodes more useful information, but it is computationally expensive. We propose to develop a GPU-accelerated version to make it practical for large-scale models. PPMI will then be used to identify properties that promote MLP/CNN performance and develop a PPMI-based loss function to construct mutual information-informed neural networks (MI2N2s).

Tasks: At least 1 seminar talk at Florida Tech. **At least 1 collaborative scientific paper** will be submitted to a national conference or journal. **2 conference talks**.

Milestone 1 (3 months): Develop GPU-accelerated PPMI. Validate on and visualize internal data representations of baseline models. Identify optimal PPMI patterns in large neural networks.

Deliverable 1: Presentation and report on the findings to ERDC. Slides summarizing findings and computer code (in Python) will be submitted to ERDC.

Milestone 2 (6 months): Develop a novel PPMI-based loss function to promote optimal relationships between input data, substructures of neural networks, and outputs.

Deliverable 2: Presentation and report on the findings to ERDC. Slides summarizing findings and computer code (in Python) will be submitted to ERDC.

Phase 2: Applications of Mutual Information-Informed Neural Networks ***(Months 7-12)***

Based on the information-based understanding of effective CNNs and losses developed in Phase 1, we will explore applications of these techniques with real data. Object detection models (e.g. YOLO) simultaneously localize and classify objects of interest in images. PEEK was shown to focus on objects of interest, even in video feeds where YOLO sometimes fails. We will use the temporal consistency of PEEK to upgrade the static object detectors to effective object *trackers*. We will then explore PPMI to associate data representations through time and space for further improvements. Recent work shows different optimizers can result in models with similar performance with substantially different internal data representations as revealed by PEEK. Diversity of representations in different high-accuracy models promotes their performance as an ensemble for a given task. We will develop automated techniques for developing ensembles by adding a loss function that encourages PEEK and PPMI maps inside a new model to differ from previous models added to the ensemble.

Tasks: At least 1 seminar talk at Florida Tech. **At least 1 collaborative scientific paper** will be submitted to a national conference or journal. **1 senior capstone project** will be supervised by R. White and M. Meni and presented at Florida Tech's Senior Design Showcase. **2 conference talks**.

Milestone 3 (9 months): Develop an object tracker that uses PEEK on internal representations inside YOLO when processing video frames. Then, develop a PPMI version. Test both on large-scale benchmark tracking datasets and NETS data for tracking satellite components.

Deliverable 3: Presentation and report on the findings to ERDC. In addition, slides summarizing findings and computer code (in Python) will be submitted to ERDC.

Milestone 4 (12 months): Develop a loss function and training scheme to encourage diverse PEEK/PPMI maps within individual neural networks. Develop a cross-model training approach to construct diverse models to form an ensemble. Test results on benchmark datasets on diverse tasks.

Deliverable 4: Presentation and final report on the entire project findings to ERDC. Slides on findings, all computer code and academic manuscripts produced during the project will be submitted to ERDC.

Additional Benefits from the Collaboration

The NETS Lab is equipped to seek and acquire further funding to support longer-term collaboration with ERDC, through a wide range of grant programs. Over the past year, NETS submitted 8 proposals and has been funded by US Army ERDC, AFRL/USSF, Energy Management Aerospace, NSF, NVIDIA, ESA, and Google. NETS has contacts with AFRL, LANL, AFTAC, potential industrial STTR partners, and can secure letters of support from NVIDIA through membership in the Applied Research Accelerator Program.

In addition, NETS has multiple projects with direct use-cases for I2N2s with academic interest and demand from potential funders. These provide opportunities to incorporate innovations into ongoing projects for multidisciplinary work, joint publications, joint proposals, and acknowledgments to ERDC.

Personnel

PI: Dr. Ryan T. White, Assistant Professor, NETS Lab at Florida Tech