Boston College
Morrissey College of Arts and Sciences

Syllabus/Contract for Individually Arranged and Courses

(Independent Study, Reading & Research, Undergraduate Research and Honors Thesis)

Student name **Riteng Zhang** **Eagle ID** **32767727**

Major(s): Computer Science BA, Mathematics BA, Philosophy minor

Instructor of record (submits grade): **Sergio A. Alvarez**
Primary mentor (if different):

Number & title of the course: **CSCI4911 Readings in Computer Science**

For Fall_____Yes_____Spring_____ Year_2023__ Number of credits: __3__

(Please fill in where applicable; add/expand sections as needed)

1) Course objectives

Study the research literature on interpreting the learned internal representations in artificial neural networks, including methods based on measuring the sensitivity of internal activations to specific aspects of the input data, and methods based on measuring the association between internal activations and data target labels. Apply different interpretation methods to well-known deep network architecture families of Inception), to develop a better understanding of the interactions between these models' structures and their performance.

More specifically,

- Compare the difference in the changes of weights in kernels with different sizes in inception-shape models during each stage (time steps) of fine-tuning or training when using different tasks using different sets of fixed hyperparameters to make a solid conclusion.
- If the difference is indeed significant, we might conclude that weights for certain kernels are more responsible for certain types of tasks theoretically, which means that their usage has a crucial influence on the results of such tasks. One can test this conclusion by comparing the results dropping on certain tasks when perturbing different kernels

(Since the kernel sizes are different, decisions on how to compare the weights changing during training fairly and how many weights or kernels exactly are going to be perturbed are important parts of the research.)

- Other methods of evaluation such as Guided Back Propagation and Deep Lift might be considered to evaluate the importance of kernels.

2) Student's intellectual preparation for the proposed work

- Took Machine Learning and all related math courses, multivariable calculus, probability, linear algebra, and statistics.
- On Coursera or other online platforms, took Deep Learning (DeepLearning.AI) Natural Language Processing (DeepLearning.AI) TensorFlow: Advanced Techniques (DeepLearning.AI) Machine Learning (Stanford University) Generative Adversarial Networks (DeepLearning.AI)

- Able to easily recreate deep learning model structures by PyTorch or TensorFlow, and extract information from these models
- Familiar with the works in interpretability of deep learning

3) What work is expected of the student? Approximately how many hours per week will the student devote to the project?

   Riteng will prepare for this independent study project during the summer by reviewing major convolutional neural network architectures and techniques, and by identifying key papers from the research literature. He will dedicate 10 hours per week to the project during the fall semester. He will report weekly on his progress. The work itself will include studying and explaining research papers in the field (which might also require developing a deeper understanding of background material), and modifying available software implementations in PyTorch or TensorFlow as needed to develop prototypes of the methods being discussed. Riteng will submit two major written reports: a midterm report by Oct. 11th, and a final report by Dec. 11th.

4) How frequently will the student meet with the instructor?

   Normally, once per week, either in person or via teleconference. In weeks in which a meeting is not possible, Riteng will provide a brief written report instead.

5) List key deadlines and describe the expected final outcome of the project criteria for evaluating student performance

   Sept. 1st, 2023 Report on relevant papers and present a detailed plan
   Sept. 8th, 2023 Course plan fixed, only minor adjustments allowed after this
   Oct. 11th, 2023 Midterm report due
   Dec. 11th, 2023 Final paper due

   Evaluation criteria:
   Project preparation (quality of preparation, creativity of the experiments) 25%
   Participation and attendance (efficiency when working on the project) 25%
   Project major reports (quality of midterm report and final paper) 50%

6) Additional comments

References:

Zeiler and Fergus, 2013
Visualizing and Understanding Convolutional Networks
https://arxiv.org/abs/1311.2901

Zintgraf et al., 2017
Visualizing Deep Neural Network Decisions: Prediction Difference Analysis
https://arxiv.org/abs/1702.04595

Learning Important Features Through Propagating Activation Differences (DeepLIFT)
http://proceedings.mlr.press/v70/shrikumar17a

Shrikumar, 2017
Striving for Simplicity: The All Convolutional Net (Guided backpropagation, deconvolution)
https://arxiv.org/abs/1412.6806

Gao, 2023
Interpretability of Machine Learning: Recent Advances and Future Prospects
https://arxiv.org/abs/2305.00537

Zeiler
Deconvolutional Networks
https://www.matthewzeiler.com/mattzeiler/deconvolutionalnetworks.pdf

O'Shea, 2015
An Introduction to Convolutional Neural Networks
https://arxiv.org/abs/1511.08458

Szegedy, 2014
Going Deeper with Convolutions
https://arxiv.org/abs/1409.4842

Both student and instructor should sign to acknowledge their agreement to and understanding of the terms above*:

Ritong Zhang Aug 22 Aug. 23, 2023

Student signature **Date** **Professor Signature** **Date**

Approval of Chairperson/Designee

Aug. 23, 2023

Chairperson/Designee **Date**

*Copies specific for each student should be filled in departmental offices. A generic version lacking individual student information can be posted on eSyllabus. (v9/10)