Data Pre-Processing and Generation for Real-World Neural Architecture Search Applications

Project Supervision Team:

Stephen McGough (Stephen.McGough@newcastle.ac.uk)
David Towers (David.Towers@newcastle.ac.uk)

Executive Summary:

A big disparity between real-world and academic data science projects is the availability and reliability of the data. CIFAR-10, a common benchmark dataset, contains images all the same size and modality, it is perfectly class balanced, and contains no missing values. Working with our industry partners we are working to create high-quality training datasets and accurate models. One of the aims is to explore and enhance their data, creating clear labels, and generating data samples to make up for the missing use cases and biased training sets. We will investigate how style transfer can be used to expand these datasets.

Details:

The goal of this project is to advance an already existing dataset making it ready for public release, currently the data is unlabelled and there are not enough data samples to train models for challenging real-world tasks. The project will require brief EDA to identify the bias in the sample data, and the use of advanced DL techniques such as style transfer to create data samples to help address this bias. This will require the identification of suitable publicly available datasets that contain the appropriate information to be transferred onto the sample data. A dataset with additional samples will also be generated using Generative AI techniques to help reduce bias, an approach previously considered by our industry partners, to compare against the dataset using style transfer, to investigate which is more practical for real-world problems and if there is any performance trade-off.

This project will help develop Exploratory Data Analysis, Deep Learning, and Prompt Engineering skills. It will also give the student experience working with real work data on a mission critical project.

Timeline:

Week1: Exploratory Data Analysis of the data.
Week 2: Labelling of the current data.
Week 3: Identification of the bias in the dataset and what needs to be done.
Week 4: Identification of suitable datasets for style transference for each task.
Week 5: Development of Cycle-GANs (or other similar techniques) to perform style transfer.
Week 6: Generation of similar data using image Generative AI techniques.
Week 7: Evaluation and comparison of performance between style transfer and Generative AI images on standard NAS approaches.
Week 8: Report and documentation, hopefully producing a paper.

Anticipated outcome:

The outcome is a labelled dataset that will be used for Neural Architecture Search purposes and an analysis of the usability of generative AI compared to more traditional image data imputation techniques.

Background Reading:

1) Neural Architecture Search with Reinforcement Learning (https://arxiv.org/abs/1611.01578)
2) Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks (https://arxiv.org/abs/1703.10593)
3) An Empirical Study of GPT-4o Image Generation Capabilities (https://arxiv.org/abs/2504.05979)