CS Field Session Proposal - Summer 2023: AIML Dataset Distillation

Proposed for Colorado School of Mines CS Field Session in March 2023 by Peter Simonson, Data Scientist with CACI.

Company Background

CACI is a $6.2 billion company whose mission and enterprise technology and expertise play a vital role in our national security, safeguarding our troops, and enabling our government to deliver cost-effective and high-quality support for all Americans. CACI applies artificial intelligence (AI) and machine learning (ML) expertise and technology to accelerate mission productivity and effectiveness. We work with our customers to use AI to rapidly analyze and translate data into decisions to maximize workforce efficiency. We combine more than a decade of advanced machine learning experience with the domain expertise to help customers smartly apply AI for their business systems, intelligence, and enterprise IT missions.

Work to be Done

In recent years, deep neural networks (DNNs) have emerged as highly performant solutions to problems in computer vision, natural language processing, and other domains. These networks generally require exposure to large volumes of data during training to function well (e.g. the ImageNet dataset, often used to pretrain computer vision models, contains around 1.3M images, and is more than 140GB uncompressed).

It has been shown that much of the knowledge present in large datasets can often be distilled to smaller datasets; datasets that when given to the learning algorithm as training data produce models that approximate the performance of models trained on the original dataset. Approaches to this problem involve both subsetting and synthetic input generation. See Wang et al., 2020 for more information.

The goals of this project are as follows:

1) (Core) Investigate the possibility of dataset distillation for the Fashion MNIST classification dataset and computer vision task:
    a) Implement and train a baseline classification system (using the full 60000 member FMNIST training set) that attains better than 90% top-1 accuracy on the 10000 member FMNIST test set.
    b) Attempt to design and implement dataset distillation methods that can:
        i) Generate an FMNIST training set half the size of the original (i.e. 30000 members), that when used to train a classification system suffers less than 1 point of performance degradation on the test set relative to the baseline set in 1a.
        ii) Generate an FMNIST training set that is one one-hundredth the size of the original (i.e. 600 members), that can be used to train a classification system with better than 80% top-1 accuracy on the test set.
2) (Stretch) Investigate the possibility of dataset distillation for the IMDB Movie Reviews binary sentiment analysis natural language processing task:
    a) Implement and train a baseline classification system (using the full 25000 member IMDB training set) that attains better than 90% accuracy on the 25000 member IMDB test set.
    b) Attempt to design and implement dataset distillation methods that can:
        i) Generate an IMDB training set half the size of the original (i.e. 12500 members), that when used to train a classification system suffers less than 1 point of performance degradation on the test set relative to the baseline set in 2a.
        ii) Generate an IMDB training set that is one one-hundredth the size of the original (i.e. 250 members), that can be used to train a classification system with better than 80% accuracy on the test set.

We envision the deliverables for this project to be a collection of Python scripts/notebooks in which the method(s) are implemented and their efficacy demonstrated.

Desired Skillsets

- Python
  - Numpy
  - Matplotlib
  - PyTorch experience a plus.
- Experience with DNNs/computer vision models a plus.

Team

- Members: 3-5
- Fully remote

NDA & IP

- Non-disclosure agreements will not be required.
- Students retain all rights to any intellectual property they develop.