SCALABLE DATA MINING 2021 TERM PROJECT LIST

1. Scalable k means ++

Abstract: Over half a century old and showing no signs of aging, k-means remains one of the most popular data processing algorithms. As is well-known, a proper initialization of k-means is crucial for obtaining a good final solution. The recently proposed k-means++ initialization algorithm achieves this, obtaining an initial set of centers that is provably close to the optimum solution. A major downside of the k-means++ is its inherent sequential nature, which limits its applicability to massive data: one must make k passes over the data to find a good initial set of centers. In this work we show how to drastically reduce the number of passes needed to obtain, in parallel, a good initialization. This is unlike prevailing efforts on parallelizing k-means that have mostly focused on the post-initialization phases of k-means. We prove that our proposed initialization algorithm k-means|| obtains a nearly optimal solution after a logarithmic number of passes, and then show that in practice a constant number of passes suffices. Experimental evaluation on real world large-scale data demonstrates that k-means|| outperforms k-means++ in both sequential and parallel settings.

Links: https://arxiv.org/pdf/1203.6402.pdf

Mentor: Soumi Das , **Email-id:** soumid.04@gmail.com

2. Beyond Backprop: Online Alternating Minimization with Auxiliary Variables

Abstract: Despite significant recent advances in deep neural networks, training them remains a challenge due to the highly non-convex nature of the objective function. State-of-the-art methods rely on error backpropagation, which suffers from several well known issues, such as vanishing and exploding gradients, inability to handle non-differentiable nonlinearities and to parallelize weight-updates across layers, and biological implausibility. These limitations continue to motivate exploration of alternative training algorithms, including several recently proposed auxiliary-variable methods which break the complex nested objective function into local subproblems. However, those techniques are mainly offline (batch), which limits their applicability to extremely large datasets, as well as to online, continual or reinforcement learning. The main contribution of our work is a novel online (stochastic/mini-batch) alternating minimization (AM) approach for training deep neural networks, together with the first theoretical convergence guarantees for AM in stochastic settings and promising empirical results on a variety of architectures and datasets.

Links: http://proceedings.mlr.press/v97/choromanska19a/choromanska19a.pdf

Mentor: Soumi Das , **Email-id:** soumid.04@gmail.com

3. Top-K Entity Resolution with Adaptive Locality-Sensitive Hashing

Abstract: Given a set of records, entity resolution algorithms find all the records referring to each entity. In top-k entity resolution, the goal is to find all the records referring to the k largest (in terms of number of records) entities. Top-k entity resolution is driven by many modern applications that operate over just the few most popular entities in a dataset. In this paper we introduce the problem of top-k entity resolution and we summarize a novel approach for this problem; full details are presented in a technical report. Our approach is based on locality-sensitive hashing, and can very rapidly and accurately process massive datasets. Our key insight is to adaptively decide how much processing each record requires to ascertain if it refers to a top-k entity or not: the less likely a record is to refer to a top-k entity, the less it is processed. The heavily reduced amount of processing for the vast majority of records that do not refer to top-k entities, leads to significant speedups. Our experiments with images, web articles, and scientific publications show a 2× to 25× speedup compared to traditional approaches for high-dimensional data.

Links:
https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8731463&casa_token=OqaiOFhnFNUAAAAA:EQf5dG3IqmlNVrDA0HcTmy3Kb00VPNzsjnwmOV8HThcGl6damHmqUr18N24XhlW9iv3jyc3esfM&tag=1

Mentor: Soumi Das , **Email-id:** soumid.04@gmail.com

4. Adaptivity without Compromise: A Momentumized, Adaptive, Dual Averaged Gradient Method for Stochastic Optimization

Abstract: We introduce MADGRAD, a novel optimization method in the family of AdaGrad adaptive gradient methods. MADGRAD shows excellent performance on deep learning optimization problems from multiple fields, including classification and image-to-image tasks in vision, and recurrent and bidirectionally-masked models in natural language processing. For each of these tasks, MADGRAD matches or outperforms both SGD and ADAM in test set performance, even on problems for which adaptive methods normally perform poorly.

Links: https://arxiv.org/pdf/2101.11075.pdf

Mentor: Soumi Das , **Email-id:** soumid.04@gmail.com

5. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization

Abstract: We propose a technique for producing 'visual explanations' for decisions from a large class of Convolutional Neural Network (CNN)-based models, making them more transparent. Our approach - Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept (say logits for 'dog' or even a caption), flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept. Unlike previous approaches, Grad-CAM is applicable to a wide variety of CNN model-families: (1) CNNs with fully-connected layers (e.g. VGG), (2) CNNs used for structured outputs (e.g. captioning), (3) CNNs used in tasks with multi-modal inputs (e.g. VQA) or reinforcement learning, and needs no architectural changes or re-training. We combine Grad-CAM with existing fine-grained visualizations to create a high-resolution class-discriminative visualization and apply it to image classification, image captioning, and visual question answering (VQA) models, including ResNet-based architectures. In the context of image classification models, our visualizations (a) lend insights into failure modes of these models (showing that seemingly unreasonable predictions have reasonable explanations), (b) outperform previous methods on the ILSVRC-15 weakly-supervised localization task, (c) are more faithful to the underlying model, and (d) help achieve model generalization by identifying dataset bias. For image captioning and VQA, our visualizations show that even non-attention based models can localize inputs. Finally, we design and conduct human studies to measure if Grad-CAM explanations help users establish appropriate trust in predictions from deep networks and show that Grad-CAM helps untrained users successfully discern a 'stronger' deep network from a 'weaker' one even when both make identical predictions.

Links:https://openaccess.thecvf.com/content_iccv_2017/html/Selvaraju_Grad-CAM_Visual_Explanations_ICCV_2017_paper.html

Mentor: Soumi Das , **Email-id:** soumid.04@gmail.com

6. Adversarially Robust Few-Shot Learning: A Meta-Learning Approach

Abstract: Previous work on adversarially robust neural networks for image classification requires large training sets and computationally expensive training procedures. On the other hand, few-shot learning methods are highly vulnerable to adversarial examples. The goal of our work is to produce networks which both perform well at few-shot classification tasks and are simultaneously robust to adversarial examples. We develop an algorithm, called Adversarial Querying (AQ), for producing adversarially robust meta-learners, and we thoroughly investigate the causes for adversarial vulnerability. Moreover, our method achieves far superior robust performance on few-shot image classification tasks, such as Mini-ImageNet and CIFAR-FS, than robust transfer learning.

Links: https://arxiv.org/abs/1910.00982

Mentor: Kiran Purohit , **Email-id:** kiran.purohit789@gmail.com

7. SCAFFOLD: Stochastic Controlled Averaging for Federated Learning

Abstract: Federated learning is a key scenario in modern large-scale machine learning where the data remains distributed over a large number of clients and the task is to learn a centralized model without transmitting the client data. The standard optimization algorithm used in this setting is Federated Averaging (FedAvg) due to its low communication cost. We obtain a tight characterization of the convergence of FedAvg and prove that heterogeneity (non-iid-ness) in the client's data results in a 'drift' in the local updates resulting in poor performance. As a solution, we propose a new algorithm (SCAFFOLD) which uses control variates (variance reduction) to correct for the 'client drift'. We prove that SCAFFOLD requires significantly fewer communication rounds and is not affected by data heterogeneity or client sampling. Further, we show that (for quadratics) SCAFFOLD can take advantage of similarity in the client's data yielding even faster convergence. The latter is the first result to quantify the usefulness of local-steps in distributed optimization.

Links: http://proceedings.mlr.press/v119/karimireddy20a.html

Mentor: Kiran Purohit , **Email-id:** kiran.purohit789@gmail.com

8. Universal Adversarial Training

Abstract: Standard adversarial attacks change the predicted class label of a selected image by adding specially tailored small perturbations to its pixels. In contrast, a universal perturbation is an update that can be added to any image in a broad class of images, while still changing the predicted class label. We study the efficient generation of universal adversarial perturbations, and also efficient methods for hardening networks to these attacks. We propose a simple optimization-based universal attack that reduces the top-1 accuracy of various network architectures on ImageNet to less than 20%, while learning the universal perturbation 13× faster than the standard method. To defend against these perturbations, we propose universal adversarial training, which models the problem of robust classifier generation as a two-player min-max game, and produces robust models with only 2× the cost of natural training. We also propose a simultaneous stochastic gradient method that is almost free of extra computation, which allows us to do universal adversarial training on ImageNet.

Links: https://ojs.aaai.org//index.php/AAAI/article/view/6017

Mentor: Kiran Purohit , **Email-id:** kiran.purohit789@gmail.com

9. Coresets for Data-efficient Training of Machine Learning Models

Abstract: Incremental gradient (IG) methods, such as stochastic gradient descent and its variants are commonly used for large scale optimization in machine learning. Despite the sustained effort to make IG methods more data-efficient, it remains an open question how to select a training data subset that can theoretically and practically perform on par with the full dataset. Here we develop CRAIG, a method to select a weighted subset (or coreset) of training data that closely estimates the full gradient by maximizing a submodular function. We prove that applying IG to this subset is guaranteed to converge to the (near) optimal solution with the same convergence rate as that of IG for convex optimization. As a result, CRAIG achieves a speedup that is inversely proportional to the size of the subset. To our knowledge, this is the first rigorous method for data-efficient training of general machine learning models. Our extensive set of experiments show that CRAIG, while achieving practically the same solution, speeds up various IG methods by up to 6x for logistic regression and 3x for training deep neural networks.

Links: https://arxiv.org/pdf/1906.01827v3.pdf

Mentor: Kiran Purohit , **Email-id:** kiran.purohit789@gmail.com

10. Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics

Abstract: Large datasets have become commonplace in NLP research. However, the increased emphasis on data quantity has made it challenging to assess the quality of data. We introduce Data Maps—a model-based tool to characterize and diagnose datasets. We leverage a largely ignored source of information: the behavior of the model on individual instances during training (training dynamics) for building data maps. This yields two intuitive measures for each example—the model's confidence in the true class, and the variability of this confidence across epochs—obtained in a single run of training. Experiments across four datasets show that these model-dependent measures reveal three distinct regions in the data map, each with pronounced characteristics. First, our data maps show the presence of ambiguous regions with respect to the model, which contribute the most towards out-of-distribution generalization. Second, the most populous regions in the data are easy to learn for the model, and play an important role in model optimization. Finally, data maps uncover a region with instances that the model finds hard to learn; these often correspond to labeling errors. Our results indicate that a shift in focus from quantity to quality of data could lead to robust models and improved out-of-distribution generalization.

Links: https://arxiv.org/pdf/2009.10795.pdf

Mentor: Kiran Purohit , **Email-id:** kiran.purohit789@gmail.com

11. Adversarial Attacks and Defenses in Images, Graphs and Text

Abstract: In this work, we will be implementing a few attacks and defenses provided in the paper. Our aim will be to reproduce the results provided by the author using their github repository (deep robust)

Links: Paper ppt DeepRobust-github

Mentor: Kiran Purohit , **Email-id:** kiran.purohit789@gmail.com

12. Interpretable Convolutional Neural Network

Abstract: Understanding interpretable convolutional neural networks.Implementation of the paper "Convolutional Dynamic Alignment Networks for Interpretable Classifications ".

Links:https://openaccess.thecvf.com/content/CVPR2021/papers/Bohle_Convolutional_Dynamic_Alignment_Networks_for_Interpretable_Classifications_CVPR_2021_paper.pdf

Mentor: Soumi Das , **Email-id:** soumid.04@gmail.com

13. Clustering by Sum of Norms: Stochastic Incremental Algorithm

Abstract: Standard clustering methods such as K–means, Gaussian mixture models, and hierarchical clustering, are beset by local minima, which are sometimes drastically suboptimal. Moreover the number of clusters K must be known in advance. The recently introduced sum–of–norms (SON) or Clusterpath convex relaxation of k-means and hierarchical clustering shrinks cluster centroids toward one another and ensure a unique global minimizer. We give a scalable stochastic incremental algorithm based on proximal iterations to solve the SON problem with convergence guarantees. We also show that the algorithm recovers clusters under quite general conditions which have a similar form to the unifying proximity condition introduced in the approximation algorithms community (that covers paradigm cases such as Gaussian mixtures and planted partition models). We give experimental results to confirm that our algorithm scales much better than previous methods while producing clusters of comparable quality.

Links: http://proceedings.mlr.press/v70/panahi17a/panahi17a.pdf

Mentor: Dr. Sourangshu Bhattacharya , **Email-id:** sourangshu@gmail.com

14. Multi target fear mongering detection

Abstract: Hate Speech is a widespread problem in society and relevant datasets are available to tackle the problem. Similar datasets and novel methods are not available for fear mongering detection. However, recent studies show fear leads to hate. We want to verify that claim in a multicountry setting and create a new dataset and implement baseline methods.

Links: https://dl.acm.org/doi/10.1145/3442381.3450137

Mentor: Souvic Chakraborty , **Email-id:** chakra.souvic@gmail.com

15. Automated Curation of a Large Scale Multimodal Dataset by augmenting sections of an E-Manual with related YouTube Videos

Abstract: Reading E-Manuals is a cumbersome task due to its vastness and complex language. Seghayer et al (2001) suggests that - "video better builds a mental image, better creates curiosity leading to increased concentration, and embodies an advantageous combination of modalities (vivid or dynamic image, sound, and printed text)". This would basically be the reverse of video tagging/annotation, i.e. each section of an E-Manual would be assigned a youtube video that matches its content. The challenges would include - search query optimization and efficient text-based video retrieval.

Links:https://paperswithcode.com/paper/fine-grained-video-text-retrieval-with,
https://paperswithcode.com/paper/clip4clip-an-empirical-study-of-clip-for-end,
https://paperswithcode.com/paper/noscope-optimizing-neural-network-queries

Mentor: Abhilash Nandy , **Email-id:** nandyabhilash@gmail.com