Fundamentals of AI and Large Models

Chengzhong Xu, University of Macau

September 2024

Our research on AI could be tracked back to year 1999, when we developed a "Parallel backpropagation learning algorithm for urban traffic congestion measurement" (ANNIE'99). Ten years later (since 2009), we developed a series of resource management strategies, based on Reinforcement Learning and Bayesian Learning, for servers, datacenters, as well as clouds, and graduated three Ph.D. in the area of ***ML for systems***. Representative work includes VCONF (ICAC'2009), URL (JPDC'2012), ILA (HPDC'2013), and hPREFECT (SC'2007). Recently, we also investigate the performance (utilization, latency, throughput, etc) and scalability issues of ***Systems for Large Models and Deep Learning***; *stay tuned for for our representative work in this domain.*

Since 2019, we started to address challenge issues of AI models directly, including learning effectiveness, data efficiency, architecture pruning and time complexity, adversarial attack robustness, and other robustness issues. Examples of the AI architectures include reinforcement learning, transfer learning, and federated learning. We summarize the representative results in the following:

Reinforcement Learning

We developed techniques in application of RL for resource management in servers and cloud centers. We also provided conditions for convergence of parallel implementation of RL in asynchronous manner. See early project of "ACM: Autonomic Cloud Management" for details.

Recently, we provided theoretical treatment of RL further, from perspective of iteration policy and state-action value. RL is an iterative approximation process in either value iteration or policy iteration. In PI, the initial policy may not be available when the system dynamics is completely unknown. In contrast, VI has no requirements for initial stabilizing policy, but converges much slower than PI. \lambda-Policy strikes a balance and offers advantages of both PI and VI policy. In [TNNLS'2023-Yang], we presented model-free extension of the \lambda-PI using the off-policy reinforcement learning technique, and shown that the off-policy variants of the algorithm are robust against the probing noise. (Model-free λ-policy iteration for discrete-time linear quadratic regulation)

In Q-learning networks, the iteration process relies on Q values of state-action pairs to search for the best action with maximal Q value. Traditional RL defines Q value to be a scalar. Recent development is distributional RL, which extends Q-value to be random variable so that a value distribution is used to replace the scalar Q value. In [TNNLS'2023-Wang], we proposed a distributional framework for multiagents to explicitly consider the correlations between cooperative agents. Under the framework, a model-free, online and off-policy method was developed to characterize the distributed dynamics of multiagent systems. (A Distributional Perspective on Multiagent Cooperation With Deep Reinforcement Learning)

Transfer learning

Transfer learning is to transfer knowledge obtained in auxiliary tasks to a desired task or fine tune a large pre-trained model for specific application domains for model accuracy and/or training data efficiency in target domains. However, the training process is not necessarily stable or generalizable, due to the limited target training data. We studied the model accuracy and generalization issues from the perspectives of model prior design and data augmentation.

Existing priors used pre-trained weight as a center, which could pose a severe risk of insufficient adaptation to the target data. In [ICML2020-Li], we proposed a RIFLE method that would actively forget what has been learned by re-initializing fully connected layer during fine-tuning, so as to enhance target adaptation. Stability of the learning process is mostly due to model collapse or abnormal mutual information of input and output in information theory. Model collapse can be detected by changes in cross-layer mutual information. In [ICASSP 2023-Li], we supplemented existing priors on BERT with an inductive bias, based on the information, to favor networks with stable information propagation so as to reduce model collapse rate and improve the model stability. We further developed new priors with tractable generalization bound to encourage the fine-tuned model to be insensitive to input noise [NAACL2021-Li]. (Noise stability regularization for iproving BERT fine-tuning)

Mixup is an effective way for data augmentation in general. But it would harm fine-tuning when training data are limited because few sample Mixup would fail to achieve generalizable interpolation effects and cause severe knowledge loss in transfer learning. In [Transactions on Machine Learning Research, 2023], we developed a sample-to-feature Mixup approach, in which the mixed features and source labels achieve both the goals of generalizing linear behaviors and preserving transferrable knowledge.

We developed a semi-supervised transfer learning framework by integrating a low-density prior realized by representation consistency and behavior-based distance constraint around the pre-trained weight with adaptive unlabeled data selection and demonstrated its superior performance in comparison with SOTA approaches [CVPR'21-Li] (Adaptive consistency regularization for semi-supervised transfer learning.)

From the learning architecture perspective, transfer learning is to preserve the source knowledge or features retained by important neurons. Unimportant neurons could be pruned. Conventional learning architectures transfer knowledge either equally for each neuron with the same regularized weights, or determine the strength of regularization using only the source dataset. But the source domain could be vastly larger than the target, giving importance to weights that are irrelevant to the target task. In [ICLR'2020-Wang], we proposed a method, called attentive feature distillation and selection (AFDS) to transfer source knowledge and meanwhile pruning unimportant neurons. (Pay attention to features, transfer learn from faster CNNS.)

Federated Learning

Federated learning leverages distributed data from multiple clients to jointly train a global shared model under the coordination of a central server, without sharing clients' data. Clients tend to be heterogeneous in processing power and their data are not necessarily independent and identical distributed. When directly applied on imbalanced data under heterogeneous systems, FL encounters the problems like disperse gradient updates of each client, which poses the risks of cancelling the training progress of each other, slowing down the convergence speed, and resulting in compromised accuracy. From the data level, we need to address the data heterogeneity and annotation scarcity for faster convergence. From the model level, to achieve effective global aggregation, we need to solve the biased knowledge transfer between global and local models. From the system level, it is challenging to tackle the system heterogeneity problem to enable the federated training on massive devices.

We dealt with the heterogeneity by customizing models for different clients because neuron training be heavily dependent on clients' data. Clients may waste computing resources in neurons that lead to conflicting update trajectories. Accordingly, in [CVPR'2023-Liao] we proposed Flado approach that optimizes channel activation probabilities to sparsify client neurons with trajectory alignment towards the global trajectory. It adapts sparsities for each neuron in each client model, depending on the training trajectories.

The challenge of non-i.i.d. could be dealt with at a high level of distributed optimization assuming same network architecture for clients. The model parameters are trained adaptively in response to the characteristics of client data. In [CVPR2022-Gao], we proposed a federated learning algorithm with local drift decoupling and correction (FeDC) approach that modifies the local training phase of each client model so as to utilize an auxiliary local drift variable to track the gap between the local and global model parameters.

We also empirically and theoretically investigated the output probability distributions of local clients. We revealed that the majority classes generally dominate the output probability. Motivated by this, we proposed a model agnostic instance reweighing (MAIR) method under the bi-level optimization framework to reweigh the learning of each class. This promotes the balanced optimization of each client model and reduces conflicting updates among clients.

To address the label deficiency issue in FL, we offered empirical and theoretical insights into the challenges caused by federated semi-supervised algorithms. Motivated by the empirical observations, we proposed federated learning with progressive distribution matching (FedPDM) that regularizes the distribution of pseudo-labels. It progressively reshapes the distribution to align with the ground truth distribution. We also showed that a widely-used data-free knowledge distillation (DFKD) technique has fundamental problems when applied onto models pretrained from heterogeneous data in federated learning. We revealed that adversarial DFKD algorithms favour minority classes, while causing a disastrous impact on majority classes. We theoretically proved that a biased teacher could cause severe disparity on different groups of synthetic data in adversarial distillation. To tackle this problem, we proposed a class-adaptive regularization method that encourages impartial representation learning of a generator among different classes under a constrained learning formulation. We devised a primal-dual algorithm to solve the target optimization problem [Liao-Thesis2024].

In [SmartPC, RTSS2019-Li], we investigated implementation issues of federal learning in real-time resource-constrained devices where the training time and model accuracy must be balanced in an energy-efficient manner. We proposed a global/local hierarchical online pace control framework that selects devices based on their running status and assigns them a virtual deadline for each training round.

Adversarial Attack Evaluation Strategies for Robustness

Deep learning are susceptible to adversarial attacks. They can be easily deceived to give an incorrect output by adding a tiny perturbation to the input. In [CVPR2021-Yu], we introduced a unified l_∞-norm white box attack algorithm which harnesses latent features in its gradient descent steps. This adversary is beyond any current defense mechanisms, and also computationally efficient for attacks.

Ensemble defense is trained to minimize attack transferability among sub-models. It is widely regarded promising direction to improve robustness against adversarial attacks while maintaining a high accuracy on natural inputs. However, existing attack strategies cannot reliably evaluate ensemble defenses and overestimate the robustness of the ensemble defense. In [NeurIPS2021-Yu], we introduce MORA, a model-reweighting attack to steer adversarial example synthesis by reweighing the importance of sub-model gradients. It not only demonstrates the weak robustness of SOTA defenses, but also provides a leaderboard of ensemble defenses under various attack strategies.

Model's robustness is often evaluated by gradient-based attacks. In [CVPR'2023-Yu], we found that relative errors in calculated gradients are fundamental reason why gradient-based attacks fail to accurately assess the model's robustness. Although hard to eliminate relative errors, we could control their effects on the attacks. We proposed an efficient loss function to minimize the impacts of floating-point errors so as to improve the efficiency and accuracy of gradient-based attacks [CVPR'2023-Yu].

Model Pruning and Compression

This part of research aims to reduce model complexity without over-sacrificing accuracy. It's observed that importance of features produced by deep models is highly input-dependent. Different images would excite neurons of a network with different channel weights. We developed a Feature Boosting and Suppression (FBS) method to predictively amplify salient channels and skip unimportant ones at run-time. It was demonstrated significant improvements over the traditional pruning methods[ICLR'2019-Gao] (Dynamic channel pruning).

Channel pruning introduces various degrees of sparsity to different layers. Traditional shirt quantization becomes a poor choice for certain layers in sparse models, as most near-zero quantization levels are under-utilized. In [NeurIPS'2019-Gao], we proposed a method, Focused Quantization, to exploit the statistical properties of weights in pruned models to quantize them efficiently and effectively. (Focused quantiztion for sparse CNNS)

Shift operation facilitates HW implementation. HW design tends to use flattened streaming arch for inference acceleration. Flatten accelerators isolate layer-wise computing, offering chance to use different arithmetic and precisions for each layer's computation. In [FPT'2019-Gao], we proposed Tomato HW/SW co-design method that deploys hybrid quantization to automate the selection of arithmetic and precisions in FPGA implementation for different layers of a model, so as to map all the layers onto a single or multiple FPGAs. (FPGA acceleration for multi-precision multi-arithmetic CNN.)

More fundamental results can be seen in the following (citation info is as of August 2024):

- Adaptive Fuzzy Leader-Follower Synchronization of Constrained Heterogeneous Multiagent Systems, TFS'2020 (citation 65)
- Robust Actor-Critic Learning for Continuous-Time Nonlinear Systems With Unmodeled Dynamics, TFS 2021 (citation 117)
- Hamiltonian-Driven Adaptive Dynamic Programming With Approximation Errors, T-Cybernetics, 2022 (citation 81)
- Model-Free λ-Policy Iteration for Discrete-Time Linear Quadratic Regulation T NNLS, 2021 (citation 124)