UMD FAIRNESS RESEARCH GROUP MEETING NOTES

DAVID YANG

Note: † denotes equal contribution.

1. Bilevel Optimization Problems in ML and Efficient Solvers (June 15th)

Led by Mucong Ding.

Definition 1 (Bilevel Optimization)
The **Bilevel Optimization** problem is

$$\min_{v,\theta} f(v,\theta) \text{ s.t. } \theta \in \operatorname*{argmin}_{\theta'} g(v,\theta')$$

where $\min_{v,\theta} f(v,\theta)$ is the **outer function** and $\theta \in \operatorname*{argmin}_{\theta'} g(v,\theta')$ is the **inner function**.

Note that the bilevel optimization problem can even be framed as a hyperparameter tuning problem in Machine Learning. The bilevel optimization problem can also be summarized as solving two interdependent problems where the outer problem depends on the inner problem.

Bilevel Optimization Problems in Machine Learning include the following:

(1) **Dataset Condensation/Distillation**: we want to learn a synthetic dataset such that the model trained on it has comparable performance to the model trained on the original dataset.
(2) **Coreset Selection**: similar to Dataset Condensation, but the learnable set is a subset of the original training set.
(3) **Targeted Dataset Poisoning**: modify the training data to cause reclassification of the unmodified test image.
(4) **Learnable Dataset Augmentation**: the learning of dataset augmentation can be formulated to minimize the loss of the trained model on the validation split.

Definition 2 (Efficient Solvers)
BO problems that enjoy convergence guarantee under mild conditions are **exact solvers**. The main approaches include Hypergradient descent methods, stationary, seeking methods, and value-function methods.

To summarize, Bilevel Optimization arises in many ML problems, each of which have unique setups and characteristics.

Date: Summer 2023.

2. Upcoming Projects: (Extreme Multi-label Compression/Online Data Pruning/GNNs for Tensor Completion)

Led by Tahseen Rabbani.

Definition 3 (Extreme Multi-label Learning)

The eXtreme Multi-label Learning (XML) addresses the problem of learning a classifier which can automatically tag a data sample with the most relevant subset of labels from a large label set.

Note: the above definition was taken from the following Deep Extreme Multi-label Learning paper.

Examples of Extreme Multi-label Datasets include "Delicious" (200k), Amazon (670k) – a product to product recommender, and Wiki (500k) – excerpts of Wikipedia articles. One of the important steps in this direction is to determine how to perform Extreme Multi-label Compression for better ML on these datasets.

A conventional compression strategy is described in Zhou et. al (2012): Compressed labeling on distilled labelsets for multi-label learning. A few key weaknesses/observations include:

- The Recovery algorithm $(y*)^{-1}$ restores several classes at once (distillates collates frequently occurring label patterns) all with equal likelihood.
- **We should try to predict a single most likely class/bit in the true label, i.e. P@1.**
- SVMS are not appropriate when p (label size) is large. For Delicious-200k, even at a 99% compression, we would need to train over 200k SVMs over samples with 700k features.
- Distillation scheme presented in Zhou et. al will not work for our regime since label size is greater than sample count and there is minimal overlap between labels.

They propose training a DNN over compressed labels. Further observations include: "if we believe that round(y^*), y^* is in R^c, is close to the SGP (Signed Gaussian Projection) of y, then we can find the most likely label i in y by looking at the SGP of the e_i."

Thus, they propose finding the most likely label by comparing only a subset of bits within the SGP of y that were predicted with high accuracy with the equivalent subset in e_i.

Other topics they are interested in include **Tensor Completion via GNNs** and **Better Data Pruning for Large Models**. For the latter topic, some motivation includes the fact that "For large vision transformers, an additional 2 billion pre-training data points (starting from 1 billion) leads to an accuracy gain on ImageNet of a few percentage points." Consequently, there is an observed "power law" between test-accuracy and the number of examples for large datasets, which can be improved on with better data pruning for large models.

3. Scene Detection in Vision Language Models

Led by Yuancheng Xu.

Note: since Yuancheng preferred to not have his presentation recorded, I refrained from sharing notes for confidentiality.

4. Hyper-Dimensional Function Encoding and its Extensions

Definition 4 (Continuous Objects)

Continuous objects are objects that can be sampled [].

The sample distribution and resolution may vary between the training and testing phase.

Suppose one wants to perform Machine Learning tasks on coninuous objects (regression, classification, continuous object prediction, function-to-function mapping), it would be great to encode a continuous objects into a fixed-length vector (so that it can be passed into something like a Neural Network). In practice, the continuous objects will be captured by the samples of them, and so traditional Neural Network methods may not work since input dimensions may change.

Let R be the continuous object to encode with $\{x_1, x_2, \ldots, x_n\}$ being n i.i.d samples from R (with the samples are drawn from a distribution p i.e. $x_i \sim p(R)$ with $p(x) > 0$ for all $x \in \mathbb{R}$). We want an encoder $E(\{x_1, \ldots, x_n\}) \to \mathbb{C}^N$.

- **Fixed-Length Representation**: E maps arbitrarily many samples into fixed length vector.
- **Sample Invariance**: As $n \to \infty$, $E(\{x_1, x_2, \ldots, x_n\}) \to V_\infty$ where V_∞ is independent of p.
- **Decodable**: Suppose $z = E(R)$, one can determine $Prob(x \in R)$ purely from z.

Previous work includes Support Vector Regression and Vector Function Architecture (mathematical details left out). Dehao then presented some methodology on his approach and the differences between his approach and previous ones.

A high-dimensional experiement found that the performance of an encoding does not depend on the explicit dimension d, but depends on the complexity of the function.

5. Explaining Embeddings from Model Outputs

Led by Bang An.

6. Source-Free Domain Adaptation

Led by Hesun Chen.

Datapoints from different datasets confer different distributions, which is the idea behind Distribution Shifts.

In classic DA (Domain Adaptation), the source and target data are referenced at the same time. This cross reference may cause privacy leaks, and so a better way is to keep data local. Furthermore, frequently referencing data results in heavy communication costs. In source-free DA, source and target data are referenced separately (first, models are trained on source domains and then it is passed to the target domain). See below for the classic DA Setup:

Classic DA Problem

- Basic Setup:
  - Domain (Dataset) : $\mathcal{D}_i = \{(X, y)|(X, y) \sim P^i_{XY}\}$
  - Source Domain: $\mathcal{S}_i = \{(X, y)\}$; Target Domain: $\mathcal{T}_j = \{(X, -)\}$
  - Goal: Given $\mathcal{S} = \{\mathcal{S}_i\}_{i=1}^m$ and $\mathcal{T} = \{\mathcal{T}_j\}_{j=1}^n$, find a mapping f_θ :

$$\underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{\mathcal{D}_j \sim \mathcal{S} \cup \mathcal{T}} \mathbb{E}_{(X_i, y_i) \sim \mathcal{D}_j} Acc(y_i, f_\theta(X_i))$$

i) Labels from target domains are inaccessible only during training stage.
ii) Usually we only discuss $\mathcal{D}_j \sim \mathcal{T}$, here we have $\mathcal{S}$ included for further discussion.

Pseudo labeling is important since labeled data becomes inaccessible in later stages. Three types of pseudo label methods include hidden structure based, data augmentation based, and knowledge distillation based.

7. Diverse Reinforcement Learning

Led by Pankayaraj Pathmanathan.

Definition 5 (Diverse RL)
Diverse RL aims to induce diversity in policies as well as online adapation of the diverse policies.

Problem Formulation: Inducing Diversity: We want to generate multiple policies so that each policy will be different from each other. Related/previous work includes information based and successor feature based methods.

Pan then presents his own ideal method for Diverse RL (details left out, to be presented in future papers). Flaws of previous objective functions include that

(1) It makes the support for a certain policy weaker.

(2) It adds a level of stochasticity to the overall reward.

He introduces an alternative stable objective function that can be intrepreted as a cross entropy between two distributions with a constant difference. The new objective will be without an indicator function term. With this objective, the stability factor is a bigger factor when it comes to training (even compared to "ideal" objectives).