What is the Next Dataset Creation Pipeline?

Yuan-Hong Andrew Liao
University of Toronto and Vector Institute
andrew@cs.toronto.edu

1 Introduction

Machine learning brings profound influences on a wide range of applications from self-driving cars (Sun et al. 2020) and e-commerce (Shankar et al. 2017) to automatic medical diagnosis (Kelly et al. 2019) and wild animal monitoring (Norouzzadeh et al. 2018). As a striking example, AlphaFold2 (Senior et al. 2020) performs better than computational biologists in predicting protein structures. Much of the recent success comes from larger and better datasets, such as ImageNet (Deng et al. 2009) and CheXNet (Rajpurkar et al. 2017), and human plays an important role by providing high-quality labels.

A good and high-quality dataset includes many rounds of data labeling, data cleaning, and data analysis. The full process is labor-intensive and time-consuming. What make it even more challenging are the human factors: uncertainty and diversity. Human annotations are uncertain especially in ambiguous or complex tasks. The higher the uncertainties are, the more human feedbacks we need to supervise our machine learning models. On the other hand, human is diverse in many different perspectives which can be capsulated into *skills*. The skills are dependent on task familiarity and task-required expertise. Low-skilled workers tend to provide incorrect and uncertain feedbacks, leading to lower efficiency in supervising machine learning models.

Besides data labeling, data cleaning and data analysis also matter and there are few works focus on worker-centric dataset analysis.Worker-centric dataset analysis is more than detecting adversarial workers. It helps incentivize workers to contribute more to the dataset generation or correct undesirable behaviors. Take citizen scientists for example, e.g., iNaturist (), one of the major incentizes is that they want to push the science forward and contribute to the communities. With worker centric dataset analysis, we can provide real-time feedbacks to citizen scientists, which in return, motivates them to provide even more annotations.

This research abstract tries to approach the question: *What is the next dataset creation pipeline?* and divides the problem into three different but connected pieces as shown in Fig. 1: **R1)** How to aggregate the noisy human feedbacks? **R2)** How to efficiently allocate tasks to workers with little to no prior knowledge of workers? **R3)** To incentivize workers, how to measure the influence of each worker towards the performance of machine learning models?

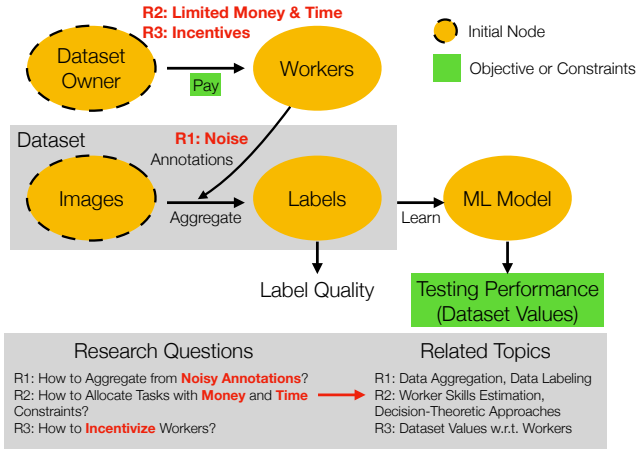

Figure 1: **Three Research Questions.** This research abstract aims to approach the question *What is the next dataset creation pipeline?* from three differnet perspectives. **R1** investiages the ways to aggregate noisy workers' annotations so as to maximize the label quality. **R2** focuses on how to perform budget(time and money)-aware task allocation such that the label quality is maximized. Lastly, **R3** analyzes the workers' annotations w.r.t. ML model preformances.

2 Background

With the surge of crowdsourcing platform (Buhrmester, Kwang, and Gosling 2011), there are plenty of works focus on label aggregation. Some focus on constructing a graphical model over the generation of annotations (Dawid and Skene 1979; Li, Rubinstein, and Cohn 2019). Other works leverage deep nerual nets to directly perform label aggregation (Rodrigues and Pereira 2017). In **R1**, we take the best of both worlds to incorporate the uncertainties measurement in graphical model and expressibility in deep neural nets.

Data labeling and data cleaning are online processes. By actively selecting which annotation (datum and worker) to sample, we can achieve better efficiency under the same budget. This is done in several works using bandit (Tran-Thanh et al. 2014; Rangi and Franceschetti 2018). The most

related topic is Active learning (Sener and Savarese 2018), which only performs sampling over datum since it assume the workers are experts. In **R2**, we want to bridge the gap between these two domains.

Data analysis is as important as data labeling and often done from the data perspective (Koh and Liang 2017; Achille et al. 2020). In **R3**, we want to evaluate the worker contributions w.r.t. model performances.

3 Efficient Dataset Generation Pipeline

We first introduce the notations and formulate the problem of efficient dataset generation. Let X be the raw data, Z be the annotations (human feedbacks), and O is the occupancy (assignment) of datum and worker. The goal is to find the best assingmnet scheme O such that the dataset values are maximized:

$$\max_{O} \mathbb{E}_{Z\sim O}\mathcal{V}(X,Z) \tag{1}$$

where $\mathcal{V}$ is the value function. A common way to define the value function is to assess the label quality, as shown in Fig. 1. The objective can be specified as minimizing the distance $\mathcal{D}(.)$ between the unobserved ground truths Y and aggregated label $\mathcal{A}(X,Z)$:

$$\max_{O} \mathbb{E}_{Z\sim O}\mathcal{D}(Y,\mathcal{A}(X,Z)) \tag{2}$$

The setting in Eq. 2 is common in the literature (Ranade and Varshney 2012). Given the thrive of machine learning, another possible way is to define the value function as the test set performance of the resulting ML models, as shown in Fig 1.

$$\begin{aligned}\max_{O} \mathbb{E}_{Z\sim O} \sum_{x,y\sim\mathcal{D}_{\text{test}}} f(x,y;\theta^*)\\ \theta^* \leftarrow \arg\min_{\theta} \mathcal{L}(X,Z)\end{aligned} \tag{3}$$

where the loss function $\mathcal{L}$ can vary depends on tasks.

The followings of this section discusses the work (Liao, Kar, and Fidler 2021) that provides us several practices on efficient label aggregation $\mathcal{A}$ (**R1**). With $\mathcal{A}$, we move on to approach the optimization problem in Eq. 3 to discuss the way to efficiently assign tasks to suitable workers (**R2**). Lastly, we are interested in analyzing individual worker's contributions toward the model performances, which can potentially incentives workers to contribute more (**R3**)

R1: Efficient Label Aggregation

In the work (Liao, Kar, and Fidler 2021), the authors provide several practices on efficient label aggregation. Given a fixed set of annotations Z, they incorporate a semi-supervised learned neural networks with a probabilistic framework. The proposed framework together with the suggested practices give us a 2.7x improvement compared to the prior art (Branson, Van Horn, and Perona 2017) on ImageNet-100. Also, it provides a realistic worker simulation that allows large-scale studies on different design factors, e.g., the number of workers, the model update frequency, the number of gold standard questions, etc. This paper stands between automatic labeling and manual labeling, therefore, giving us both the robustness and efficiency.

R2: Task Assignment

Built upon the label aggregation $\mathcal{A}(.)$ from the prior art (Liao, Kar, and Fidler 2021), we are interested in how to efficiently acquire the annotations Z. The problem is different from active learning (Sener and Savarese 2018) since it encounters the exploration and exploitation dilemma of worker skills estimation. The problem is different from bandit for crowdsourcing (Rangi and Franceschetti 2018) since it needs to consider semi-supervised learned model uncertainties. We take the first step by measuring how tightly an annotation z_{ij} and paremter posteriors θ, W (model parameters and estimated worker skills) are connected. Inspired by BALD (Houlsby et al. 2011), we measure it by computing mutual information:

$$\text{BALD: } \mathbb{I}(z_{ij},(\theta,W)) = \mathbb{H}(z_{ij}) - \mathbb{H}(z_{ij}|\theta,W) \tag{4}$$

where $\mathbb{I}$ and $\mathbb{H}$ are mutual information and entropy, respectively. We provide preliminary experiments in Sec. 4.

R3: Worker Contributions w.r.t. Models

Worker contributions are usually estimated by the number of annotation they provide. This does not consider the worker quality and might enocourage lazyness. In data-centric analysis, data contributions can be estimated by influential function (Koh and Liang 2017) or gradient in linearized networks (Achille et al. 2020). Assume the label aggregation function $\mathcal{A}(X,Z,\theta)$ is differentiable, we can derive the gradients of the loss w.r.t. annotation:

$$\begin{aligned}&\nabla_Z \sum_{x,y\sim\mathcal{D}_{\text{test}}} \mathcal{L}(x,y;\phi^*) = \nabla_{\hat{Y}} \sum_{x.y\sim\mathcal{D}_{\text{test}}} \mathcal{L}(x,y;\phi^*)\nabla_Z\mathcal{A}(X,Z)\\ &\phi^* \leftarrow \arg\min_{\phi} \mathcal{L}(X,\hat{Y}),\ \hat{Y} \leftarrow \mathcal{A}(X,Z)\end{aligned} \tag{5}$$

4 Preliminary Evaluation

We provide several preliminary evaluations of efficient label aggregation (R1) and task assignment (R2) in this section.

Label Aggregation

In the work (Liao, Kar, and Fidler 2021), the authors conduct the experiments on ImageNet-100 and the simulated workers, initialized by the feedbacks curated from AMT. The proposed method includes a semi-supervised learned model and several practices including better calibration, the feature extractor initialized by self-supervised learning, better stopping criterion, etc. The proposed method provides 2.7x efficiency compared to "Lean" (Branson, Van Horn, and Perona 2017) and 6.7x efficiency compared to manual annotation. The paper also observes that the efficiency of the label aggregation can be benefited by having a sense of dataset granularity a priori.

Task Assignment

We use a 5-class toy classification task as the testbed. The data X is a two-dimensional vector and the simulated workers are modelled by 5-by-5 confusion matrixes. The worker

skills and the model parameters are re-estimated at each time step. We compare the effectiveness of the acquisition function in Eq. 4 (BALD) with greedy and random sampling approaches. The preliminary experiments show that BALD can outperform the baselines, but fail when the workers skills are well-estimated at the beginning. Also, the cost and time constraints are not considered so far.

5 Discussion

In this research abstract, I want to emphasize the need of systematically performing data labeling, data cleaning, and data analysis. With the robust and efficient dataset creation pipeline, we can systematically improve datasets and therefore improve nearly all machine learning models at the same time.

References

[Achille et al. 2020] Achille, A.; Golatkar, A.; Ravichandran, A.; Polito, M.; and Soatto, S. 2020. Lqf: Linear quadratic fine-tuning.

[Branson, Van Horn, and Perona 2017] Branson, S.; Van Horn, G.; and Perona, P. 2017. Lean crowdsourcing: Combining humans and machines in an online system. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6109–6118.

[Buhrmester, Kwang, and Gosling 2011] Buhrmester, M.; Kwang, T.; and Gosling, S. D. 2011. Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6(1):3–5.

[Dawid and Skene 1979] Dawid, A. P., and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28(1):20–28.

[Deng et al. 2009] Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

[Houlsby et al. 2011] Houlsby, N.; Huszr, F.; Ghahramani, Z.; and Lengyel, M. 2011. Bayesian active learning for classification and preference learning.

[Kelly et al. 2019] Kelly, C.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; and King, D. 2019. Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine* 17.

[Koh and Liang 2017] Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. In Precup, D., and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*. PMLR.

[Li, Rubinstein, and Cohn 2019] Li, Y.; Rubinstein, B.; and Cohn, T. 2019. Exploiting worker correlation for label aggregation in crowdsourcing. In Chaudhuri, K., and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 3886–3895. PMLR.

[Liao, Kar, and Fidler 2021] Liao, Y.-H.; Kar, A.; and Fidler, S. 2021. Towards good practices for efficiently annotating large-scale image classification datasets.

[Norouzzadeh et al. 2018] Norouzzadeh, M. S.; Nguyen, A.; Kosmala, M.; Swanson, A.; Palmer, M. S.; Packer, C.; and Clune, J. 2018. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences* 115(25):E5716–E5725.

[Rajpurkar et al. 2017] Rajpurkar, P.; Irvin, J.; Zhu, K.; Yang, B.; Mehta, H.; Duan, T.; Ding, D.; Bagul, A.; Langlotz, C.; Shpanskaya, K.; Lungren, M. P.; and Ng, A. Y. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning.

[Ranade and Varshney 2012] Ranade, G., and Varshney, L. 2012. To crowdsource or not to crowdsource? In *HCOMP@AAAI*.

[Rangi and Franceschetti 2018] Rangi, A., and Franceschetti, M. 2018. Multi-armed bandit algorithms for crowdsourcing systems with online estimation of workers' ability. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '18. International Foundation for Autonomous Agents and Multiagent Systems.

[Rodrigues and Pereira 2017] Rodrigues, F., and Pereira, F. 2017. Deep learning from crowds.

[Sener and Savarese 2018] Sener, O., and Savarese, S. 2018. Active learning for convolutional neural networks: A core-set approach.

[Senior et al. 2020] Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; dek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; and Hassabis, D. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* 577(7792):706–710.

[Shankar et al. 2017] Shankar, D.; Narumanchi, S.; Ananya, H. A.; Kompalli, P.; and Chaudhury, K. 2017. Deep learning based large scale visual recommendation and search for e-commerce.

[Sun et al. 2020] Sun, P.; Kretzschmar, H.; Dotiwalla, X.; Chouard, A.; Patnaik, V.; Tsui, P.; Guo, J.; Zhou, Y.; Chai, Y.; Caine, B.; Vasudevan, V.; Han, W.; Ngiam, J.; Zhao, H.; Timofeev, A.; Ettinger, S.; Krivokon, M.; Gao, A.; Joshi, A.; Zhao, S.; Cheng, S.; Zhang, Y.; Shlens, J.; Chen, Z.; and Anguelov, D. 2020. Scalability in perception for autonomous driving: Waymo open dataset.

[Tran-Thanh et al. 2014] Tran-Thanh, L.; Stein, S.; Rogers, A.; and Jennings, N. R. 2014. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence* 214:89–111.