FID3018 - Review of Papers

Sina Sheikholeslami
sinash@kth.se

1 OVERVIEW

Many machine learning (ML) and deep learning (DL) models deal with one or several datasets as the input to their learning process. A typical example is training an image classifier through a supervised process, in which the model is given a set of labeled images, trying to "learn" features so it can correctly classify unseen images. However, even in settings where a dataset is not explicitly present, e.g., a reinforcement learning task in which an agent interacts with the environment, the model still learns from various forms of data. But how should we feed the data to our model? how would the number of data examples given to the model at each iteration impact the final performance of the model? furthermore, are all examples within a dataset of equal importance? is there such thing as "easy" or "hard" examples? if that is the case, how can we exploit it to both reduce the training time and improve the quality of our models? the three papers chosen to be reviewed below deal with these questions. In [3] the authors perform an extensive experimental evaluation on the effect of different batch size configurations on the performance of DL models that are trained with Stochastic Gradient Descent (SGD). Active Bias [1] is an approach to focus on more uncertain examples rather than easy or hard examples, in order to balance the tradeoff between robustness to outliers and training speed. Finally, in [2], a scalable ensemble AL based approach is presented, which is able to find smaller subsets of well-known image classification datasets on which the trained models can achieve even better quality compared to training on the full dataset.

2 MEASURING THE EFFECTS OF DATA PARALLELISM ON NEURAL NETWORK TRAINING

Motivation. The authors observe that faster training allows researchers to try new ideas and configurations more rapidly, thus leading to dramatic improvements in model quality. However, there has been significant disagreement regarding how does the increase in the batch size affects the out-of-sample (generalization) error, as a strong indicator of model quality. On the other hand, recent advances in hardware technology allow for much larger batch sizes to be used. Thus, there is a need to study the relationship between the batch size and the number of training steps required to reach a goal out-of-sample error, in order to find out, e.g., if it would make sense to increase the batch size beyond a certain point. This work tries to address this issue through an extensive empirical study, focused on a number of variants of Synchronous SGD.

Contributions. The authors find out that the effect of batch size on the model quality is largely based on differences in hyperparameters, compute budgets, and workloads (dataset, model, and training algorithm). They find no evidence that larger batch sizes necessarily degrade out-of-sample performance. Also, they propose that some variants of SGD (e.g., SGD with momentum) can make use of much larger batch sizes that plain SGD. In addition to that, they find that some models allow for exploiting much larger batch sizes than other models. Regarding the role of hyperparameters, they show that optimal values for hyperparameters do not consistently follow any simple relationship with batch size. But perhaps most importantly, they claim that much of the disagreement in the literature is due to the differing assumptions on the computation budgets and techniques for selecting hyperparameters.

Solution. The authors provide an extensive empirical study using seven standard datasets for image classification and language modelling tasks, and train variants of six state-of-the-art network architectures using different optimizers and learning rate schedulers, and measure the number of training steps required to reach a desired out-of-sample error.

Opinion. This manuscript is very well written, and the authors have successfully managed to leverage the length of the journal paper to provide sufficient details and discussions on various aspects of their work. However, the paper uses increase in batch size as a proxy for the degree of parallelism, and do not provide any information on the infrastructure and the schemes used for the experiment. On the other hand, though the paper restricts its focus to synchronous SGD, since the authors have released most of the material required for reproducing the experiments, researchers can try to do similar investigations for other variants of SGD (such as Asynchronous Parallel SGD), or other workloads (e.g., different network architectures).

3 ACTIVE BIAS: TRAINING MORE ACCURATE NEURAL NETWORKS BY EMPHASIZING HIGH VARIANCE SAMPLES

Motivation. The focus of this paper is is on exploiting the difference between the importance of examples in a labeled dataset. Prior to this work, there have been approaches proposed to leverage this information to speed up training and improve the model quality, e.g., Self-paced learning focuses more on "easier" examples, i.e., examples with lower loss on the model that is being trained. In contrast, hard example mining focuses more on the harder examples. However, there is a tradeoff between facilitating the training and maintaining robustness to outliers and noise: self-paced learning is more robust to noise and outliers, but since the easier examples result in smaller gradient updates, it slows down the training process. On the other hand, hard example mining accelerates SGD training when we have cleaner data, but is more prone to outliers, and we often do not know a-priori how noisy our training dataset is. Active Bias, the approach proposed by the authors, deals with this tradeoff by instead focusing on "uncertain" examples.

Contributions. The contributions of the paper include two lightweight methods to exploit uncertain sample for mini-batch SGD

classification: i) by measuring the variance of prediction probabilities, or ii) by estimating the closeness between prediction probabilities and decision thresholds. The authors argue that their approach reduces the generalization error by 1% to 18%.
Solution. The authors provide various example sampling probabilities and weighted loss functions, by focusing on either prediction variance for an example, or the closeness between the prediction for them and the decision threshold. Their approach does not require specific hyperparameter tuning.
Opinion. The authors do a very good job in explaining the required theoretical background of the work, as well as the theory behind their own solution. Their experimental study is also good. One important thing about their approach is that it does not require any hyperparameter tuning, but it would be nice to see how it would perform after tuning the hyperparameters with regards to the specific sampling or re-weighting policies.

4 TRAINING DATA SUBSET SEARCH WITH ENSEMBLE ACTIVE LEARNING

Motivation. Previous approaches for finding more efficient dataset subsets based on ensemble active learning (AL) suffer from the scaling point of view: existing studies showed no gains in performance when using ensembles of more than 10 models. Also, those studies have been limited in their experimental settings, e.g., they use small subsets of datasets (e.g., one study used 30% of ImageNet).
Contributions. The authors present an approach to scale up ensemble AL methods to hundreds of models with negligible computational overhead at training time, due to using model checkpoints as new ensemble members. They also provide a larger-scale empirical study (compared to the previous studies) on how to effectively reduce the size of existing datasets without human curation, covering image classification and object detection domains.
Solution. The approach consists of an acquisition model initialized based on a particular initialization scheme and an ensemble configuration. The acquisition model iteratively selects dataset subsets for training, and finally the target model will be trained on the desired subset of the dataset.
Opinion. The paper supports the arguments through explaining the intuitions as well as results from the experiments, however it still lacks some theoretical guarantees. Also, it is still unclear if going to such measures for finding a smaller subset of a labeled dataset is worth the extra effort.

REFERENCES

[1] Haw-Shiuan Chang, Erik Learned-Miller, and Andrew McCallum. 2017. Active bias: Training more accurate neural networks by emphasizing high variance samples. *arXiv preprint arXiv:1704.07433* (2017).

[2] Kashyap Chitta, Jose M Alvarez, Elmar Haussmann, and Clement Farabet. 2019. Training Data Distribution Search with Ensemble Active Learning. *arXiv preprint arXiv:1905.12737* (2019).

[3] Christopher J Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E Dahl. 2018. Measuring the effects of data parallelism on neural network training. *arXiv preprint arXiv:1811.03600* (2018).