FID3018 Advanced Course in Data Mining and Analytics
Overview of Selected Papers

Sina Sheikholeslami

May 2021

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 deal with these questions.

In (Shallue et al. 2018) 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 (Chang, Learned-Miller, and McCallum 2017) is an approach to focus on more uncertain examples rather than easy or hard examples, in order to balance the trade-off between robustness to outliers and training speed. Finally, in (Chitta et al. 2019), a scalable ensemble active learning-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.

References

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

Chitta, Kashyap et al. (2019). "Training Data Distribution Search with Ensemble Active Learning". In: *arXiv preprint arXiv:1905.12737*.

Shallue, Christopher J et al. (2018). "Measuring the effects of data parallelism on neural network training". In: *arXiv preprint arXiv:1811.03600*.