Role of Hyperparameters in Deep Active Learning

Denis Huseljic*, Marek Herde, Paul Hahn and Bernhard Sick

University of Kassel, Intelligent Embedded Systems, Wilhelmshöher Allee 73, Kassel, 34121, Germany

Abstract

Despite considerable research on pool-based *deep active learning* (DAL), deploying DAL into real applications still has several challenges. A frequently neglected aspect is the choice of training *hyperparameters* (HPs), such as the learning rate. Since these HPs determine how the deep neural network learns in each cycle, they must be chosen carefully. In this article, we analyze the role of HPs in DAL. We find that optimizing HPs reduces the performance gap between DAL strategies. Conversely, we highlight challenges in finding optimal HPs when using datasets selected via DAL strategies.

1. Introduction

Deep neural networks (DNNs) typically require large amounts of annotated data to achieve excellent performance in supervised learning tasks [1]. Pool-based *deep active learning* (DAL) aims to reduce the associated annotation cost by intelligently querying (human) annotators for instances' annotations (e.g., labels) [2]. In an iterative process, a selection strategy decides which instances should be annotated to yield high-performance improvements for the DNN.

While a considerable amount of research focuses on developing new selection strategies [3, 4, 5], deploying DAL into real applications remains difficult [6, 7]. Recent studies [8, 6, 9] demonstrate that many DAL strategies perform similarly or worse when compared to a random selection of instances. Especially [8] show this observation when training hyperparameters (HPs), such as weight decay, are appropriately tuned in each DAL cycle. Typically, each DAL cycle consists of a new learning problem, i.e., a change in the training data distribution, in which the optimal training HPs may also vary. Neglecting these HPs by fixing them at the start of DAL could lead to false conclusions regarding selection strategies' performance. Moreover, it might also negatively affect the usefulness of acquired instances regarding DNN's performance.

Figure 1 depicts DAL learning curves of a ResNet18 [1] on CIFAR-10 [10] under various settings of learning rate and weight decay (two influential HPs) fixed at the beginning of DAL. We employ a random instance selection and the state-of-the-art DAL strategy Badge [3]. The random selection results in Fig. 1 (a) reveal considerable differences between learning curves, indicating that the choice of training HPs plays a critical role in how the DNN learns in each cycle. When we focus on Badge in Fig. 1 (b), the training HPs' influence on a DAL selection

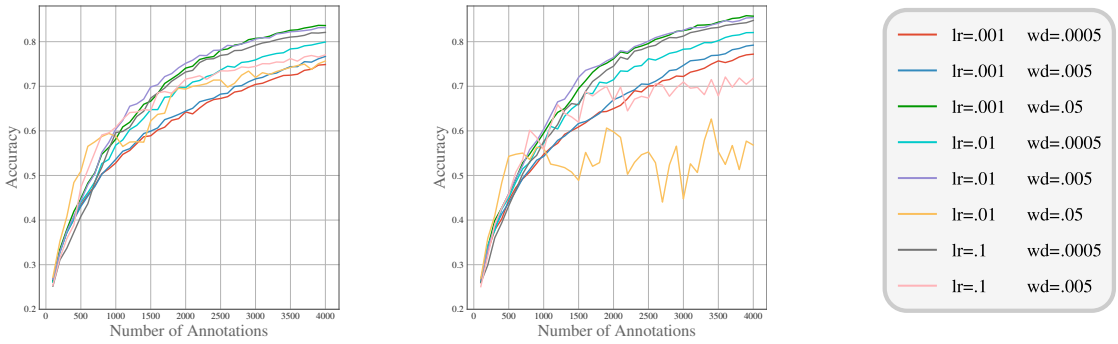
IAL@ECML-PKDD'23: 7th Intl. Worksh. & Tutorial on Interactive Adaptive Learning, Sep. 22nd, 2023, Torino, Italy

*Corresponding author.

dhuseljic@uni-kassel.de (D. Huseljic); marek.herde@uni-kassel.de (M. Herde); paul.hahn@uni-kassel.de (P. Hahn); bsick@uni-kassel.de (B. Sick)

0000-0001-6207-1494 (D. Huseljic); 0000-0003-4908-122X (M. Herde); 0009-0005-1012-1281 (P. Hahn); 0000-0001-9467-656X (B. Sick)

CEUR Workshop Proceedings (CEUR-WS.org)



(a) Random (b) Badge (c) HP Settings

Figure 1: Learning curves (number of annotations vs. accuracy) of a ResNet18 on CIFAR-10 with eight HP settings of learning rate (lr) and weight decay (wd). Results are averaged over three repetitions.

strategy becomes apparent. In DAL, the DNN's predictions are typically used to select instances for annotations. Suboptimal HPs could result in less accurate model predictions, negatively affecting DAL strategies. Learning curves show a greater variance than random when using a DAL strategy. Notably, two specific HP settings resulted in a significant decrease in performance. This experiment suggests the need for carefully selecting HPs when deploying DAL strategies.

We present new studies regarding HPs' influence on DAL with the following contributions:

- We briefly review influential articles in the field of DAL, especially focusing on how these handle training HP in their experimental design.
- We present experiments analyzing training HPs' influence on DAL by evaluating the reusability of labeled pools acquired by DAL strategies for DNNs with optimized HPs.
- We provide a roadmap to guide future research on choosing training HPs in DAL.

2. Related Work

Foundational DAL Research: Most highly-cited DAL research underemphasizes HP choice, despite the presence of a randomly selected validation dataset in their experiments. For example, although dropout [11] is an essential component in [12], they only optimize weight decay in each DAL cycle. In contrast, [13], [14], [15], and [16] fix training HPs and only evaluate validation accuracy to select the best-performing DNN during training. While [17] fix training HPs, they employ a grid search to optimize HPs specific to their strategy. The authors in [3] argue that DAL should simply work given fixed HPs. However, their results show significant accuracy discrepancies compared to the literature. For example, in their CIFAR-10 experiment with 10k annotations and a ResNet-18, the average test accuracy is over 50% worse than in [18].

Contemporary DAL Research: Recent research [4, 5] still primarily fixes training HP at the beginning of DAL. The authors of [8] highlight that this hinders comparability across various studies and DAL strategies. Furthermore, they empirically demonstrate that HP optimization during DAL reduces accuracy improvements over a random instance selection. Accordingly, they recommend HP optimization during DAL to foster comparability. Other studies, such as [19], argue that HP optimization per cycle requires much computational effort and does not

Table 1
Test accuracies on CIFAR-10 after HP optimization on datasets that were acquired via DAL strategies. The DAL column reports test accuracies of the DNN with pre-defined HPs obtained with DAL. The BO column reports test accuracies of a DNN with the same architecture that resulted from applying BO to the labeled pool obtained from DAL. Results are averaged over 3 seeds. Colors indicate the best, second best, third best, and worst accuracies across DAL strategies.

$\lvert\mathscr{L}\rvert$	Strategies	HP 1 (lr = .001 wd = .05)		HP 2 (lr = .01 wd = .0005)		HP 3 (lr = .01 wd = .05)		HP 4 (lr = .1 wd = .005)	
		DAL	BO	DAL	BO	DAL	BO	DAL	BO
2k	Random	74.07 ± 0.66	76.96 ± 0.43	69.69 ± 0.24	77.72 ± 0.17	69.35 ± 0.78	77.14 ± 0.59	71.61 ± 0.61	77.33 ± 0.22
	Entropy	74.54 ± 0.31	77.63 ± 0.34	69.12 ± 1.77	77.78 ± 0.76	26.08 ± 0.21	77.12 ± 0.24	16.61 ± 9.34	76.61 ± 1.17
	Core-Sets	75.86 ± 0.51	77.42 ± 0.09	68.75 ± 1.88	76.47 ± 0.83	43.57 ± 7.43	76.76 ± 1.83	68.89 ± 1.77	76.98 ± 0.88
	Badge	76.14 ± 0.61	78.56 ± 0.44	71.47 ± 1.58	78.91 ± 0.48	59.79 ± 4.74	77.61 ± 0.61	66.69 ± 3.88	78.27 ± 1.37
4k	Random	83.61 ± 0.44	83.41 ± 0.52	79.91 ± 0.11	83.47 ± 0.28	75.65 ± 0.85	83.68 ± 0.25	77.04 ± 0.58	83.64 ± 0.28
	Entropy	85.46 ± 0.29	84.07 ± 0.88	81.69 ± 0.27	84.57 ± 0.86	47.23 ± 11.23	84.02 ± 0.61	58.57 ± 17.07	84.05 ± 0.92
	Core-Sets	85.42 ± 0.62	84.98 ± 0.37	81.47 ± 0.78	84.46 ± 0.90	47.98 ± 7.70	83.42 ± 0.77	71.58 ± 1.58	84.30 ± 0.18
	Badge	85.72 ± 0.17	85.12 ± 0.72	82.04 ± 0.25	84.58 ± 0.29	56.81 ± 7.04	84.02 ± 0.23	71.79 ± 1.57	85.50 ± 0.29

affect the DAL strategies' ranking. Thus, they suggest fixing HPs at the start of DAL, providing recommendations for specific HP settings.

3. Experiments

We examine image classification problems using the ResNet-18 [1] as DNN and conduct experiments on CIFAR-10 [10], which contains a predefined training and test split. Throughout this article, we employ the training split for DAL. We consider a DAL scenario with an unlabeled pool $\mathscr{U}$ and a labeled pool $\mathscr{L}$. We start with 100 randomly sampled labeled instances. Afterward, each cycle consists of DNN training followed by selecting 100 instances from $\mathscr{U}$ based on a DAL strategy. Selected instances are removed from $\mathscr{U}$ and added to $\mathscr{L}$ with their respective labels. The cycle is repeated until a budget of 4000 instances is reached. In addition to random instance selection, we investigate three selection strategies: i) entropy, which is uncertainty-based, ii) core-sets [14], which is representation-based, and iii) BADGE [3], which is a hybrid of both. All results are averaged over three repetitions.

To assess the effect of training HPs (here: learning rate and weight decay) on DAL's selection, we examine the labeled pools collected by DAL strategies that use different HP settings. Specifically, for each strategy, we fix four unique HP settings at the beginning of DAL, each leading to a distinct labeled pool. We focus on labeled pools in two stages of the DAL process: in the middle (2k instances at 50%) and at the end (4k instances at 100%). These are then employed to evaluate if varying HP settings substantially influence the performance of strategies, e.g., if a poor-performing strategy due to bad HPs would still underperform with good HPs. To this end, once a labeled pool is collected via DAL, we subsequently optimize HPs with Bayesian optimization (BO) [20]. To mimic a realistic scenario, we randomly sample a validation split of 10% of the labeled pool. As the sizes of 2k and 4k are relatively small, we use three validation splits for each HP. For a more detailed experimental setup and additional results on CIFAR-100, we refer to our implementation at https://github.com/dhuseljic/dal-toolbox.

Table 1 details the experimental results. Focusing on DAL accuracies first, we see *inconsistent*

improvements of DAL strategies over a random selection across different HP settings. Especially HP 3 & 4 demonstrate considerable performance drops of strategies to random selection, suggesting that DAL strategies may even fail with a poor HP choice. Considering the ranking of the DAL accuracies across HPs, we observe that *different HPs can lead to varying rankings*. For example, DAL accuracies of Badge are best when considering HP 1 & 2, while random is best for the other HPs. Once we consider BO accuracies, we notice that a *poor HP choice during DAL does not necessarily lead to a worse selection*, indicated by decreasing performance gaps to random. Notably, Badge performs robustly, achieving the best BO accuracy in 7 out of 8 cases. Comparing the ranking of DAL and BO accuracy pairs for fixed HP settings, we see that *optimizing HPs after DAL impacts strategies' ranking*. For example, when considering HP 4, Badge becomes the best performing DAL strategy after BO, despite not being the best during DAL. This also suggests that *a suboptimal HP choice at the beginning of DAL does not necessarily lead to a bad instance selection of DAL*. Interestingly, comparing DAL and BO accuracies of HP 1 for 4k instances, we see *no accuracy improvement of BO over DAL, which we believe is because a well-working HP leads to a strong sampling bias in the labeled pool*. For example, the labeled pool of entropy sampling consists of uncertain instances which do not match the true data distribution leading to a validation dataset unsuitable for HP optimization.

4. Takeaways and Future Research

We provided an empirical study examining the influence of HPs on the instance selection of different DAL strategies. Our main takeaways are: i) Varying HP settings at the beginning of DAL can lead to inconsistent performance improvements, strategy rankings, and even failing DAL strategies. However, suboptimal HPs do not necessarily lead to a worse instance selection. ii) Choosing proper HPs at the beginning of DAL does not guarantee suitable labeled pools for real applications. The resulting datasets might not represent the data distribution due to sampling bias, making a representative validation split difficult.

As datasets persist and research continuously develops new DNN architectures, the reusability of labeled pools collected by DAL strategies needs a stronger focus. Specifically, reusing these for new DNN architectures with possibly different HPs is crucial for effective integration into practical applications [21]. Therefore, it is essential that future research addresses:

- **Research perspective**: HP optimization in the context of DAL requires deeper investigations. Our study suggests that HP optimization at the end of DAL reduces the performance improvement of strategies over random instance selection, corroborating the findings of [8]. Research should be extended toward more alternatives (e.g., optimize HP after regular intervals within DAL) with various architectures and strategies.
- **Application perspective**: There should be a stronger focus on how to proceed after a dataset was acquired via DAL, especially due to requiring a validation dataset that is suitable for HP optimization. This aspect is critical for enabling DAL's successful deployment into real applications.

To this end, it might be necessary to design DAL strategies that select not only informative instances but also representative ones, which can be used for effective validation. Consequently, this would also minimize the validation dataset's annotation costs.

References

[1] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: Computer Vision and Pattern Recognition, IEEE, Las Vegas, NV, USA, 2016, pp. 770–778.

[2] B. Settles, Active Learning Literature Survey, Technical Report, University of Wisconsin, Department of Computer Science, 2009.

[3] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, A. Agarwal, Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds, in: International Conference on Learning Representations, 2020.

[4] G. Hacohen, A. Dekel, D. Weinshall, Active Learning on a Budget: Opposite Strategies Suit High and Low Budgets, in: International Conference on Machine Learning, PMLR, 2022, pp. 8175–8195.

[5] O. Yehuda, A. Dekel, G. Hacohen, D. Weinshall, Active Learning Through a Covering Lens, in: Advances in Neural Information Processing Systems, 2022.

[6] S. Mittal, M. Tatarchenko, O. Cicek, T. Brox, Parting with illusions about deep active learning, arXiv preprint arXiv:1912.05361 (2019).

[7] M. Herde, D. Huseljic, B. Sick, A. Calma, A survey on cost types, interaction schemes, and annotator performance models in selection algorithms for active learning in classification, IEEE Access 9 (2021) 166970–166989.

[8] P. Munjal, N. Hayat, M. Hayat, J. Sourati, S. Khan, Towards robust and reproducible active learning using neural networks, in: Computer Vision and Pattern Recognition, 2022, pp. 223–232.

[9] Y. Li, M. Chen, Y. Liu, D. He, Q. Xu, An Empirical Study on the Efficacy of Deep Active Learning for Image Classification, arXiv preprint arXiv:2212.03088 (2022).

[10] A. Krizhevsky, Learning Multiple Layers of Features from Tiny Images, Master's thesis, University of Toronto, 2009.

[11] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A Simple Way to Prevent Neural Networks from Overfitting, Journal of Machine Learning Research 15 (2014) 1929–1958.

[12] Y. Gal, R. Islam, Z. Ghahramani, Deep bayesian active learning with image data, in: International Conference on Machine Learning, PMLR, 2017, pp. 1183–1192.

[13] W. H. Beluch, T. Genewein, A. Nurnberger, J. M. Kohler, The Power of Ensembles for Active Learning in Image Classification, in: Computer Vision and Pattern Recognition, IEEE, 2018, pp. 9368–9377.

[14] O. Sener, S. Savarese, Active Learning for Convolutional Neural Networks: A Core-Set Approach, in: International Conference on Learning Representations, 2018.

[15] D. Gissin, S. Shalev-Shwartz, Discriminative active learning, arXiv preprint arXiv:1907.06347 (2019).

[16] A. Kirsch, J. Van Amersfoort, Y. Gal, Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning, in: Advances in Neural Information Processing Systems, 2019.

[17] S. Sinha, S. Ebrahimi, T. Darrell, Variational Adversarial Active Learning, in: International Conference on Computer Vision, 2019, pp. 5972–5981.

[18] A. Lang, C. Mayer, R. Timofte, Best Practices in Pool-based Active Learning for Image

Classification, OpenReview, 2022. URL: https://openreview.net/forum?id=7Rnf1F7rQhR.
[19] Y. Ji, D. Kaestner, O. Wirth, C. Wressnegger, Randomness is the Root of All Evil: More Reliable Evaluation of Deep Active Learning, in: Winter Conference on Applications of Computer Vision, 2023, pp. 3943–3952.
[20] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2623–2631.
[21] D. Lowell, Z. C. Lipton, B. C. Wallace, Practical Obstacles to Deploying Active Learning, in: Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing, 2019.