The Influence of Feature Space Complexity on Optimal Data Selection in Deep Active Learning

Anonymous authors
Paper under double-blind review

Abstract

Deep active learning (DAL) studies the optimal selection of labeled data for training deep neural networks (DNNs). While data selection is mostly optimized for learning task decision boundaries, in DNN this decision boundary changes as the the feature space changes with the architecture of the DNN feature extractor. How the optimal selection of data is affected by this change is poorly understood in DAL. To shed light on this question, we present the first systematic investigation on: 1) the relative performance of representative modern DAL data selection strategies, as the architecture types and sizes change in the underlying DNN feature extractor (Focus 1), and 2) the effect of optimizing the architecture of a DNN feature extractor on DAL (Focus 2). The results suggest that the change in the DNN architecture significantly influences and outweighs the benefits of data selection in DAL. These results cautions the community in generalizing DAL findings obtained on specific architectures, while suggesting the importance to optimize the DNN architecture in order to maximize the effect of active data selection in DAL.

1 Introduction

Active learning (AL) is a long standing research area that studies how to carefully select the most informative samples to label in order to best improve learning given a limited labeling budget (Cohn et al., 1996; Settles, 2009). As the success of deep neural networks (DNNs) continues, there have seen substantial developments in bringing the success of AL to DNNs, creating a fast-growing subarea known as deep active learning (DAL) (Ren et al., 2021). In its essence, DAL starts with training a DNN on an initial pool of labelled data, followed by optimizing an *acquisition function* to select new data to be labelled to improve the DNN. This iteration continues until a labeling budget or desired accuracy is achieved (Ren et al., 2021).

The main progress in DAL has been revolving around the design and developments of acquisition functions to improve the selection of training data. Most data acquisition strategies can be categorized into efforts that exploit DNN uncertainties (Ranganathan et al., 2017; Wang et al., 2016; Li et al., 2017; He et al., 2019; Ostapuk et al., 2019; Gal et al., 2017b; Freytag et al., 2014; Käding et al., 2016; Yoo & Kweon, 2019; Huang et al., 2022), explore the diversity of unlabeled data (Wang et al., 2017; Sener & Savarese, 2017; Geifman & El-Yaniv, 2017; Shui et al., 2020; Zhang et al., 2020b; Sinha et al., 2019; Kim et al., 2021), or combine the advantages of the above two in a hybrid fashion (Liu et al., 2016; Coletta et al., 2019; Zhdanov, 2019; Ash et al., 2020; Shui et al., 2020; Kong et al., 2022; Wang et al., 2022; Settles et al., 2007; Shukla, 2022). Additional approaches also consider the generation of informative examples (Mahapatra et al., 2018; Mayer & Timofte, 2020; Zhu & Bento, 2017), utilizing the unlabeled data to pretrain the DNN feature extractor (Siméoni et al., 2021) or in a semi-supervised fashion during DAL (Siméoni et al., 2021; Gao et al., 2020), or exploiting the training dynamics of the DNN (Wang et al., 2022).

In this paper, we bring to the research community an under-explored yet important question in DAL: the effect of choice of DNN architecture on optimal selection of data during active learning. Traditional AL developments are typically associated with pre-defined features, where data selection is primarily optimized for task decision boundaries on the given feature space. In contrast, for DNNs, the feature space is unknown. As illustrated in Fig. 1A, as the architecture of the DNN feature extractor changed, the feature space also

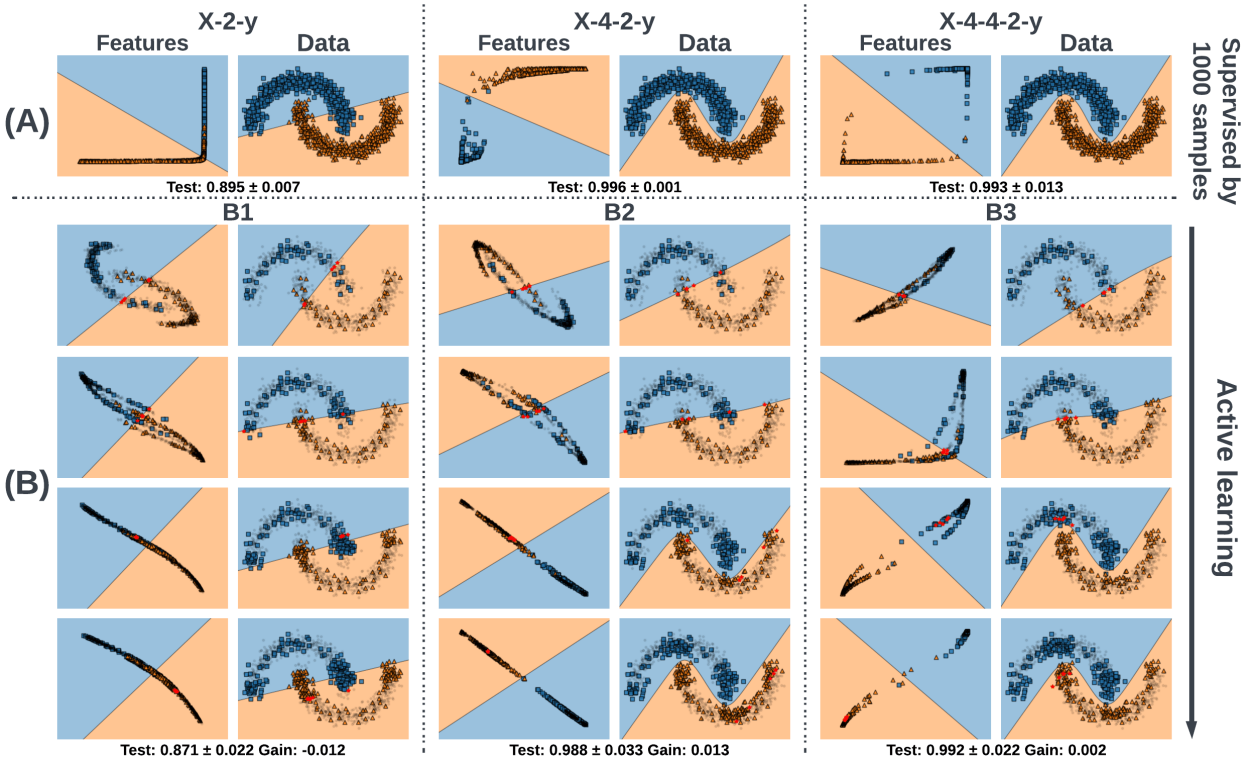

Figure 1: Illustrations of the change of optimal decision boundary due to differences in the DNN feature space (A) and its impact on DAL data selection and gain over random acquisition of data (B). Each column (separated by dotted vertical lines) represents decision boundaries achieved by different architecture of neural networks visualized in the feature (left) and data (right) space, where 2, 4-2, 4-4-2 in between input x and output y denote the number of nodes used in each hidden layer. The figures show that change in architecture changes the features and thus optimal decision boundary, which in turn affects optimal data selection in DAL.

changed which resulted in changes in the *optimal* decision boundary. When aiming to *actively* capture such decision boundary with a small amount of carefully-selected labeled data, the differences in feature space as shown in Fig. 1B had a clear impact. First and foremost, data selection for the same acquisition strategies (red) became very different, even for overly expressive DNNs (Fig. 1B2-3) that eventually arrived at similar decision boundaries. Furthermore, the gains of active learning over random acquisitions not only became substantially different among the various DNN architectures, but also appeared marginal compared to the performance gap induced by architecture differences. This strong dependence on DNN feature space, brought by change in DNN architecture, raises two critical gaps of knowledge in DAL developments and evaluation.

First, it is not clear how DAL evaluation may be dependent on (or agnostic to) the choices of DNN architectures, considering that the vast majority of DAL works are based on predefined and fixed DNN architectures (Geifman & El-Yaniv, 2018). Indeed, several recent works have investigated the reproducibility of DAL evaluations (Beck et al., 2021; Munjal et al., 2020; Mittal et al., 2019), noting that performance gaps among various DAL methods are inconsistent across experimental settings. Some also noted that the gain of DAL over random acquisition is in general marginal compared to the use of other regularization techniques (Beck et al., 2021; Munjal et al., 2020). These studies however have not focused on the effect of DNN architectures. In Focus 1 of this paper, we address this knowledge gap by systematically evaluating the relative performances of seven representative DAL acquisition functions for a given DNN architecture, across two convolution-based DNN architecture types on image datasets and three transformer-based DNN architecture types on text datasets, each with different sizes.

Second, it is not clear how DAL developments can be improved by DNN architectural optimization, or what may be the most effective approach to combine DNN training data optimization with architecture optimization. Until now, only one previous work examined the effect of incremental DNN architectural search as DAL proceeds, showing that architecture optimization brought improvements compared to the use of fixed architectures during DAL (Geifman & El-Yaniv, 2017). It however was limited to the investigation of a particular architecture optimization approach, designed for a particular type of DNN architectures (RESNET-18). Nor did it have a focus on the relative importance between optimizing the DNN architecture *versus* optimizing data selection. In parallel, there have been increasing DAL works that put an explicit focus on DNN feature space, mostly done via pre-training a feature extractor using unlabeled data (Siméoni et al., 2021) and/or simultaneously improving the learned feature during DAL using unlabeled data in addition to the acquired labeled data (Kong et al., 2022; Hacohen et al., 2022; Wang et al., 2022). All of these works, however, are performed with fixed DNN architectures. In Focus 2 of this paper, we address this knowledge gap by systematically evaluating how architecture optimization of the DNN feature extractor may affect DAL, considering three possible approaches: 1) jointly optimizing the DNN architecture with the acquired labeled data throughout the course of DAL, or 2) pre-optimizing the DNN architecture using the unlabeled data prior to DAL, or 3) pre-optimizing the DNN architecture using the initial label data prior to DAL. We conduct this investigation with three representative approaches to DNN architecture optimization that allow manageable computation when used repetitively over the course of DAL: two Bayesian architecture inference methods (KC et al., 2021; Lee et al., 2018) and one neural architecture search (NAS) method (Chen et al., 2019), all in comparison to their fixed counterparts on the same spectrum of DAL strategies considered in Focus 1.

Experimental results from Focus 1 demonstrated that the relative performance of DAL acquisition functions was substantially influenced by the underlying DNN architecture, cautioning the community in generalizing DAL findings obtained on specific architectures. Results from Focus 2 showed that optimizing DNN architecture substantially benefited DAL. More interestingly, the gain of performance induced by an optimized *vs.* pre-defined DNN architecture appeared to significantly *outweighed* the choice of DAL acquisition strategies. This leaves a potentially important implication for DAL research: while the importance of DNN architecture choices is widely accepted in standard "passive" learning, it may be especially crucial and may suggest an under-explored research avenue in DAL in order to maximize the effect of active data selection when the labeling budget is limited.

2 Related Works

Deep active learning (DAL): DAL research has flourished over the years with the design of various strategies to select data from the unlabelled pool. The strategies could be broadly divided intro three categories. Uncertainty-based strategies seek examples that a DNN is most uncertain about. A variety of measures has been proposed to represent this broadly-defined uncertainty, including entropy (Joshi et al., 2009), BALD (Houlsby et al., 2011; Shelmanov et al., 2021), least confidence based on softmax outputs (Settles, 2009), margin sampling (Scheffer et al., 2001), expected gradient length (Settles et al., 2007; Huang et al., 2016; Zhang et al., 2017; Shukla, 2022), changes in outputs in response to input perturbation(Freytag et al., 2014; Käding et al., 2016), and estimation of DNN loss (Yoo & Kweon, 2019; Huang et al., 2022). Diversity-based strategies seek samples that are representative of the unlabelled data using approaches such as density clustering (Wang et al., 2017), coreset optimization (Sener & Savarese, 2017; Geifman & El-Yaniv, 2017), and leveraging adversarial networks (Zhang et al., 2020b; Sinha et al., 2019; Kim et al., 2021). Hybrid strategies combine these two approaches to sample diverse data which the DNN is most uncertain about (Ash et al., 2020; Shui et al., 2020; Wang et al., 2022; Kong et al., 2022), such as by considering the magnitude as well as diversity of DNN gradients (Settles et al., 2007; Shukla, 2022).

In addition to these pool-based active learning where new training data is obtained by querying an unlabelled pool, there are generative approaches (Zhang et al., 2020b) that generate examples informative to the current model. These approaches leverage generative adversarial network (GAN) to generate informative data examples that has high entropy (Mayer & Timofte, 2020) or are closer to the decision boundary (Zhu & Bento, 2017) (Mahapatra et al., 2018; Mayer & Timofte, 2020; Zhu & Bento, 2017). Additional approaches include utilizing the unlabeled data to pretrain the DNN feature before DAL (Siméoni et al., 2021) or using

it in a semi-supervised fashion during DAL (Siméoni et al., 2021; Gao et al., 2020), as well as exploring DNN training dynamics (Wang et al., 2022) as measured by the derivative of training loss with respect to the number of iterations assuming that models training faster generalize better. These developments are seen in both image as well as text data domains.

Most of these existing works were conducted on specific choices of DNN architectures, such as MLPs (Ash et al., 2020), LeNet (Geifman & El-Yaniv, 2017; Hu et al., 2021), CNNs (Gal et al., 2017b), and different versions of VGG and RESNET (Ash et al., 2020; Shui et al., 2020) on image data, or BERT (Zhang et al., 2020a; Schröder et al., 2021; Wertz et al., 2022) and its two variants – DistilBERT (Schröder et al., 2021; Kirk et al., 2022) and RoBERTa (Lu & MacNamee, 2020) on text data. A lack of consistency regarding the choices of DNN architectures exist across existing studies, and it is not clear how the reported DAL evaluations may be dependent on (or agnostic to) the choices of DNN feature extractors, a critical question that will be systematically investigated in this paper.

Systematic evaluation of DAL methods: An observation emerging in recent works (Mittal et al., 2019; Munjal et al., 2020; Beck et al., 2021) is the inconsistency and reproducibility of the relative performance of DAL methods across experimental settings. The lack of unified experimental setting, such as size of the initial labeled pool, acquisition size, total labeling budget, random seeds, batch size, and optimizers have been credited for the inconsistencies of results reported (Munjal et al., 2020; Beck et al., 2021). It was further shown that the gain of DAL over random acquisition is in general marginal compared to other strategies, such as network regularization, data augmentation, and semi-supervised techniques (Munjal et al., 2020; Beck et al., 2021). This paper will add to these findings focusing on the effect of optimizing the architecture of DNN feature extractor on DAL.

DNN architecture optimization in DAL: There is a large body of literature in deterministic optimization or Bayesian inference of DNN architectures (Zoph & Le, 2016; Zoph et al., 2018; Kasim et al., 2020; Feng & Darrell, 2015; Lee et al., 2018; Dikov & Bayer, 2019; KC et al., 2021), supporting the notion that the complexity of DNN feature extractors has substantial impact when *passively* learning from given data. To date, only one work investigated the effect of optimizing DNN architecture in the context of active data selection as DAL proceeds (Geifman & El-Yaniv, 2018). Specifically, an incremental architectural search method was formulated over a modularly reduced search space customized for RESNET-18, integrated and evaluated with three existing DAL data acquisition strategies. This paper will substantially expand the scope of this previous study by: 1) investigating both supervised joint training as well as unsupervised pre-training as alternative approaches to combine DNN architecture and data optimization, 2) including a variety of DNN architectures especially convolutional architectures (VGG- and RESNET-variants) for image data and transformer architectures (BERT, DistilBERT, and RoBERTa) for text data, 3) including three architecture optimization methods representative of both Bayesian architecture inference (KC et al., 2021; Lee et al., 2018) and deterministic NAS (Chen et al., 2019) approaches, and eventually 4) deriving insights into the relationship between DNN architecture optimization and data optimization during DAL.

3 Focus 1: The Effect of DNN Architectures on DAL Performances

3.1 Methodology

In Focus 1, we systematically investigate whether and how the *relative performance* of existing DAL acquisition strategies may depend on the underlying DNN feature extractor, especially its difference in choices of architecture types and sizes. We consider classification tasks on both image and text data. On image tasks, we consider two convolutional DNN architecture types (RESNET and VGG) that are mostly commonly used in DAL literature and each with three different sizes (RESNET-18/34/50 and VGG-11/16/19). On text tasks, we consider three transformer architectures (BERT, DistilBERT, and RoBERTa) that are most commonly used in text-based active learning tasks. We consider seven state-of-the-art DAL acquisition functions representative of uncertainty-based (Settles et al., 2007; Scheffer et al., 2001; Joshi et al., 2009; Houlsby et al., 2011), diversity-based (Sener & Savarese, 2017; Ash et al., 2020), and hybrid strategies (Ash

et al., 2020), each in combination with or without unsupervised pretraining as the seventh DAL strategy presented in (Siméoni et al., 2021).

As our main focus is to understand how the performance gap among various acquisition strategies may be affected by the choice of DNN architectures, we devise metrics to measure the *relative performance* of different acquisition strategies, and examine its changes across choices of DNN architectures. We further examine how such effects change as the data selection is affected by the size of the initial labeled pool, the size of each acquisition, and the presence of data augmentation. We leave out investigations on other DNN hyperparameters or regularization techniques that have been studied in previous works (Munjal et al., 2020; Beck et al., 2021).

DAL acquisition strategies: We consider the following acquisition functions, against random sampling as the baseline:

1. *Least confidence sampling* (Settles et al., 2007) chooses instances (x^*) with the least probability scores of the predicted class $p(c)$ among class $c \in C = \{1, 2, ..., k\}$ in the output for a given unlabelled data x_i i.e. $x^* = argmin_{x_i \in U} p(c|x_i)$.

2. *Margin sampling* (Scheffer et al., 2001) selects instances (x^*) with the smallest difference between the first and second largest class label probability, assuming that a confident model is characterized by a substantial gap between the predicted label probability and the second-highest label probability:

$$x^* = argmax_{x_i \in U}[p(y = c_1|x_i) - p(y = c_2|x_i)] \tag{1}$$

3. *Entropy sampling* (Joshi et al., 2009) selects instances (x^*) a DNN is most uncertain about as measured by the entropy (H) calculated from the output softmax probabilities:

$$\begin{aligned} H &= \sum_{c \in C} p_c * log(p_c) \\ x^* &= argmax_{x_i \in U} H(x_i) \end{aligned} \tag{2}$$

4. *BALD sampling* (Houlsby et al., 2011) selects instances (x^*) that generates disagreeing predictions that the model is the most uncertain on average (I) (Gal et al., 2017b) i.e.

$$\begin{aligned} I(y; w|x, L) &= H(y|x, L) - E_{p(w|L)}[H(y|x, w, L)] \\ x^* &= argmax_{x_i \in U} I(y; w|x_i, L) \end{aligned} \tag{3}$$

5. *Coreset Sampling* (Sener & Savarese, 2017) selects unlabeled instances that are the most different from existing labelled sample based on their feature distances:

$$x^* = argmax_{i \in U} \; min_{j \in L} \triangle (x_i, x_j) \tag{4}$$

6. *BADGE sampling* (Ash et al., 2020) selects instances that generates diverse but also high gradient magnitudes in the penultimate layer of the DNN.

7. *Unsupervised Pretraining*: Based on Simeoni et al. (2021), we adopt a two-step pretraining strategy for both image and text datasets. This pretraining involves a combination of alternate unsupervised clustering task and classification task supervised by the clustering labels. We begin with random initialization of the network parameters and the features from the penultimate layers are clustered using k-means clustering. These generated pseudo- labels are then utilized as the ground truth for a subsequent supervised classification task which in turn updates the network parameters. The networks, once fully trained, serve as the initial models for all subsequent active learning experiments. This strategy is used in combination with all acquisition functions described above.

These acquisition functions are the commonly used benchmarks in DAL research. Among them, the first four are representative of uncertainty-based acquisition strategies, with Entroy and BALD calculated based on Bayesian drop-out strategies (Gal et al., 2017b). Coreset is representative of diversity-based strategies, and BADGE is a representative hybrid strategy. Each of these acquisition functions are tested without and with unsupervised pretraining as the seventh DAL strategy considered.

Network architectures: On image data, we consider two convolutional DNN architecture types that are mostly used in DAL literature, each with three different sizes: VGG-11, -16, -19; (Simonyan & Zisserman, 2014), and RESNET-18, -34, -50 (He et al., 2016). On text data, we consider three transformer architectures namely BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019) due to their prominent use in text based active learning tasks. Details of DNNs are described in the Appendix B.

Evaluation metrics: We consider two quantitative metrics: 1) labeling efficiency as described in (Beck et al., 2021), which measures the amount of data required in comparison to random acquisition (as a ratio) to achieve the same test accuracy; and 2) a new metric that measures the percentage of gain in test-accuracy over random acquisition at each acquisition round. We use these two metrics to compare the relative performance of the considered acquisition functions across DNN architecture types and sizes.

3.2 Experiments and Results

Experiments in Focus 1 were performed on four image datasets including MNIST (Deng, 2012), Fashion MNIST (Xiao et al., 2017), CIFAR10 (Krizhevsky, 2009), and SVHN (Netzer et al., 2011), and four text datasets including AGNEWS (Zhang et al., 2015), Banks77 (Casanueva et al., 2020), DBPedia (Auer et al., 2007) and QNLI (Wang et al., 2018). Let D be the complete dataset divided into initially labelled data $L = \{x_l, y_l\}_{l=1}^{|L|}$ and unlabelled pool of data $U = \{x_u\}_{u=1}^{|D|-|L|}$ of size $|L|$ and $|D| - |L|$, respectively. We initialized the networks on the labelled set L. In each acquisition round, the acquisition function of choice was used to choose $|A|$ number of data points from U to be labelled next. The newly labelled data was added back to the labelled set L to retrain the network and the process was repeated until $|B|$ number of unlabelled data has been labelled.

On image dataset, for consistent experimental settings, we used an initial balanced labelled data size of 1000 unless otherwise stated. The remaining data were divided into 90-10% training-validation split and the size of the unlabelled pool was 30,000. A total of 25 acquisition rounds were performed with an acquisition size of 1000. Each round is set to train for 500 epochs with early stopping added when the training accuracy reached 99% or validation accuracy remained unchanged for 50 epochs. On text data, we used an initial labelled data size of 100 with 25 rounds of acquisition rounds performed with an acquisition size of 100. The experiments are run for 5 epochs following (Devlin et al., 2018). All the experiments were ran for three seeds on workstations with RTX 2080Ti GPU and 32 GB of RAM. We used the DISTIL github repo (Dani et al., 2021) as a base skeleton for our image experiments and dal-toolbox github repo (Rauch et al., 2023) for text experiments.

Relative performance of acquisition functions: Figure 2 summarizes the labeling efficiency and Figure 3 summarizes the gain of DAL over random acquisition for all network architectures on the image datasets. Figure 4 – Figure 5 summarize similar results on the text datasets. As shown, across datasets and image/text tasks, both the performance gain of all DAL strategies over random acquisition, as well as the relative performance among the different DAL acquisition functions, varied substantially. For instance, DAL seemed to demonstrate less benefits over random acquisition on CIFAR10 than the other image datasets, or stronger benefits on AGNEWS than the other text datasets. Given the same DNN architecture, FASHION-MNIST (and DBPEDIA and QNLI) seemed to induce a more evident performance difference among the acquisistion function tested, compared to CIFAR10 and SVHN (AGNEWS and BANKS77) where the performance gap among the tested acquisition functions were small. More importantly, even within the same dataset, the relative performance among the acquisition functions varied with DNN types and sizes: for instance, on MNIST, data acquisition with least confidence criteria appeared to be the most advantageous using the larger RESNETs (RESNET34 and RESNET50), yet coreset based data acquisition appeared to outperform the others on RESNET18 and the variants of VGGs; similarly, on QNLI, the hybrid BADGE acquisition seemed to be favored on the BERT architecture followed by DISTILBERT, while becoming the least competitive on the ROBERTA architecture. Furthermore, the use of DISTILBERT in general seemed to induce larger gaps among the different acquisition functions compared to the other two architectures on the same datasets, although favoring different acquisitions on different datasets (least confidence based acquisistion on AGNEWS, margin based acquisistion on BANKS77, and BADGE on DBPERDIA and QNLI). Note

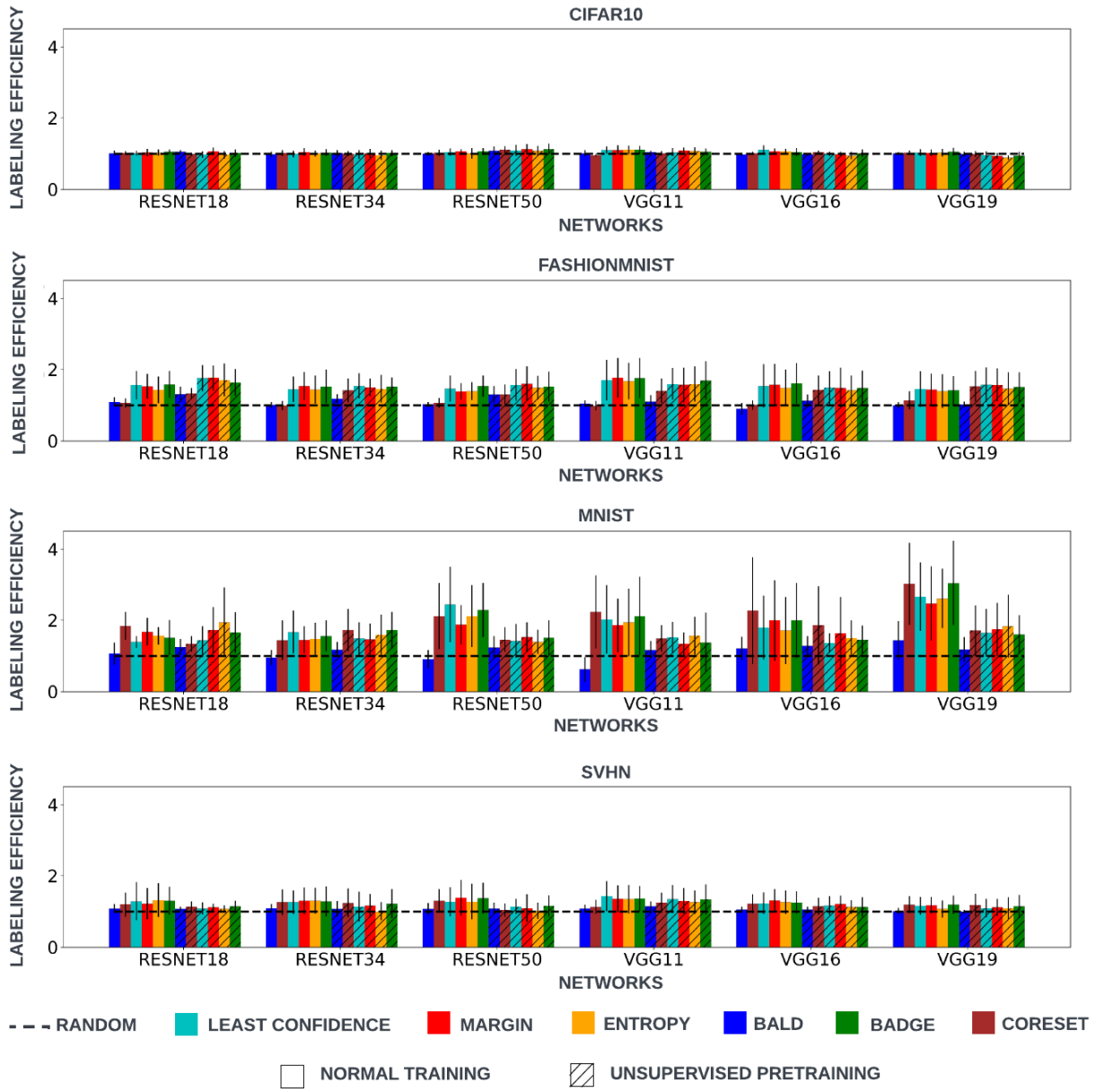

Figure 2: Comparison of labeling efficiency over random acquisition, for all six acquisition function with (hatched) and without unsupervised pretraining across four datasets and all DNN architectures on image classification tasks. The dashed horizontal line shows reference of random acquisition. Results show that the relative performance of acquisition strategies varies both across DNN choices on the same dataset, as well as across datasets.

that depending on the metric used (labeling efficiency *vs.* gain in test accuracy), the relative performance among the various acquisition functions as well as their gain over random were also different.

The use of pre-training also appeared to not only influence the relative performance of the tested acquisition functions, but also induced different gains over random acquisition in a architecture-dependent fashion, positively on some architectures and datasets (*e.g.*, RESNET50 on CIFAR10, RESNET18 on FASHION-MNIST, ROBERTA on AGNEWS and BANKS77, and DISTILBERT on QNLI), and negatively on some (*e.g.* all VGGs on MNIST, all RESNETs on SVHN, and BERT on AGNEWS and BANKS77).

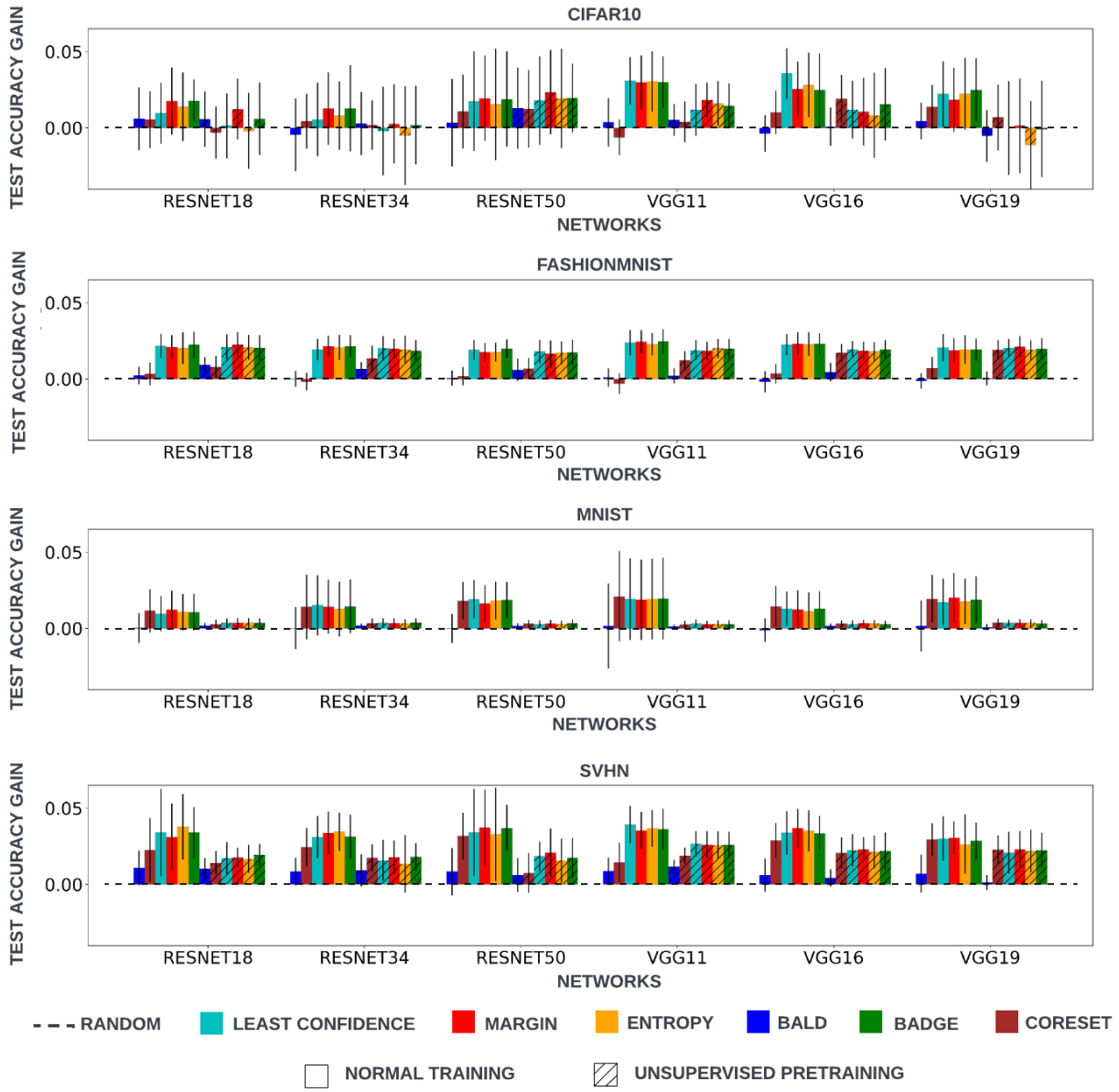

Figure 3: Comparison of gain in test-accuracy over random acquisition (as 0.00 test accuracy gain), for all six acquisition function with (hatched) and without unsupervised pretraining across four datasets and all DNN architectures on image classification tasks. Similar to Figure 2, results show that the relative performance of acquisition strategies varies both across DNN choices on the same dataset, as well as across datasets.

Ultimately, no acquisition function consistently outperformed others across datasets and architecture. This indicates that the optimality of data selection strategy depends on the choice of DNN architectures.

Effect of initialization and acquisition size: We further varied the initial pool size between 1000 and 5000, and the acquisition size among 500, 1000, 1500. As shown in Figure 6, the increase in initial pool size brought a gain in test performance for all DNNs, which then decreased for all acquisition functions as DAL proceeds. The change in acquisition size did not produce noticeable differences on the results (see Appendix C.2).

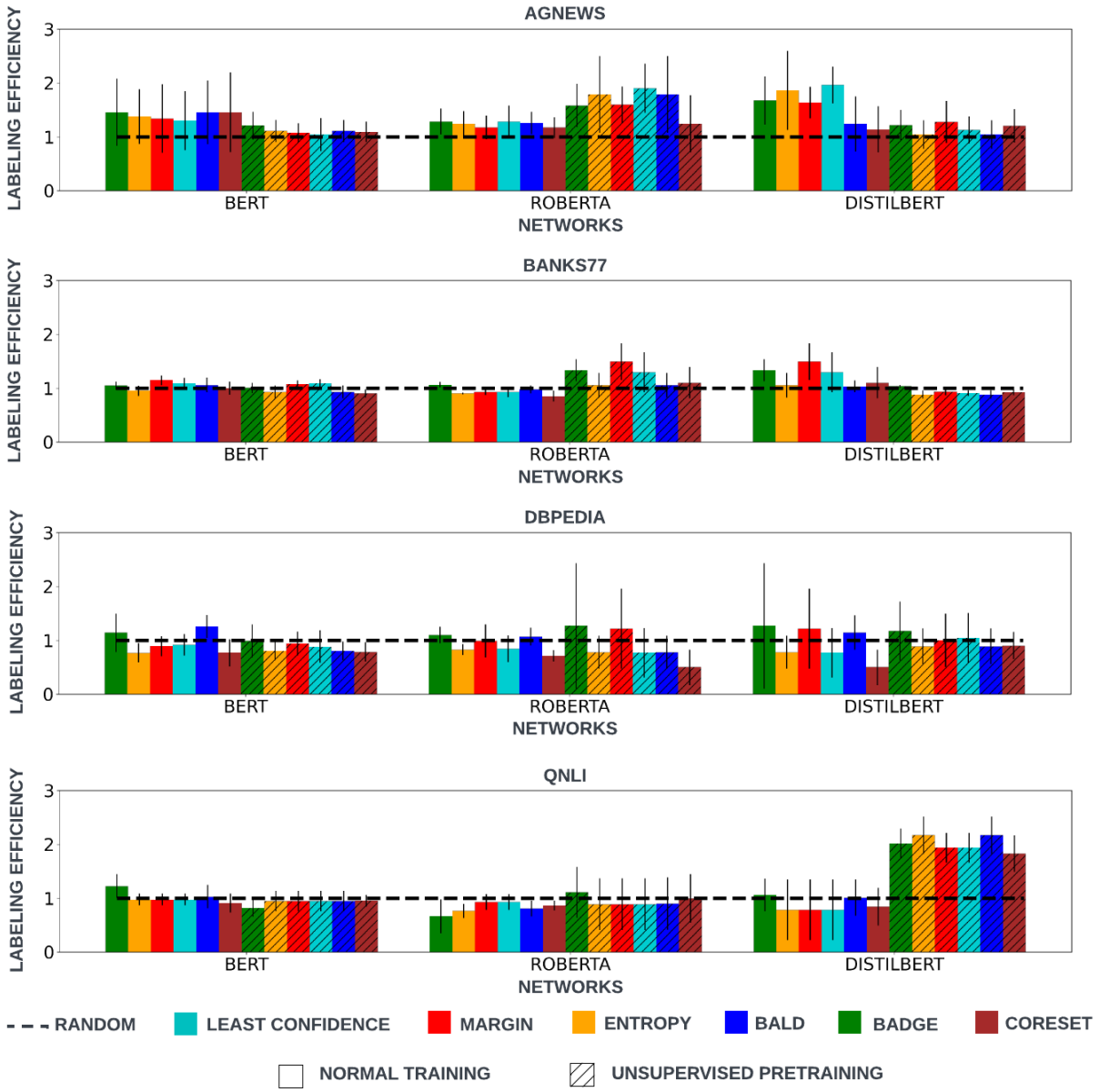

Figure 4: Comparison of labeling efficiency over random acquisition, for all six acquisition function with (hatched) and without unsupervised pretraining across four datasets and all transformer architectures on text tasks. The dashed horizontal line shows reference of random acquisition. Results show that the relative performance of acquisition strategies varies both across architecture choices on the same dataset, as well as across datasets.

Effect of data augmentation: Fig. 7 shows the test accuracy gain over random acquisition achieved by the same six acquisition functions when we applied data augmentation, including horizontal flips and random crops, to all experiments. Compared to the results without augmentation, two main differences can be observed: with data augmentation, 1) most acquisition functions tested exhibited either a substantially larger gain over the random acquisition (*e.g.*, CIFAR10 for all networks, MNIST for RESNETs and VGG11), or a reduced error in the gain (CIFAR10, SVHN for all networks) 2) the gap in the performance among different acquisition functions appeared to be minimal (*e.g.*, BADGE, entropy, least confidence, and margin) across all datasets, in comparison to the no augmentation counterpart. For FashionMNIST, the difference between

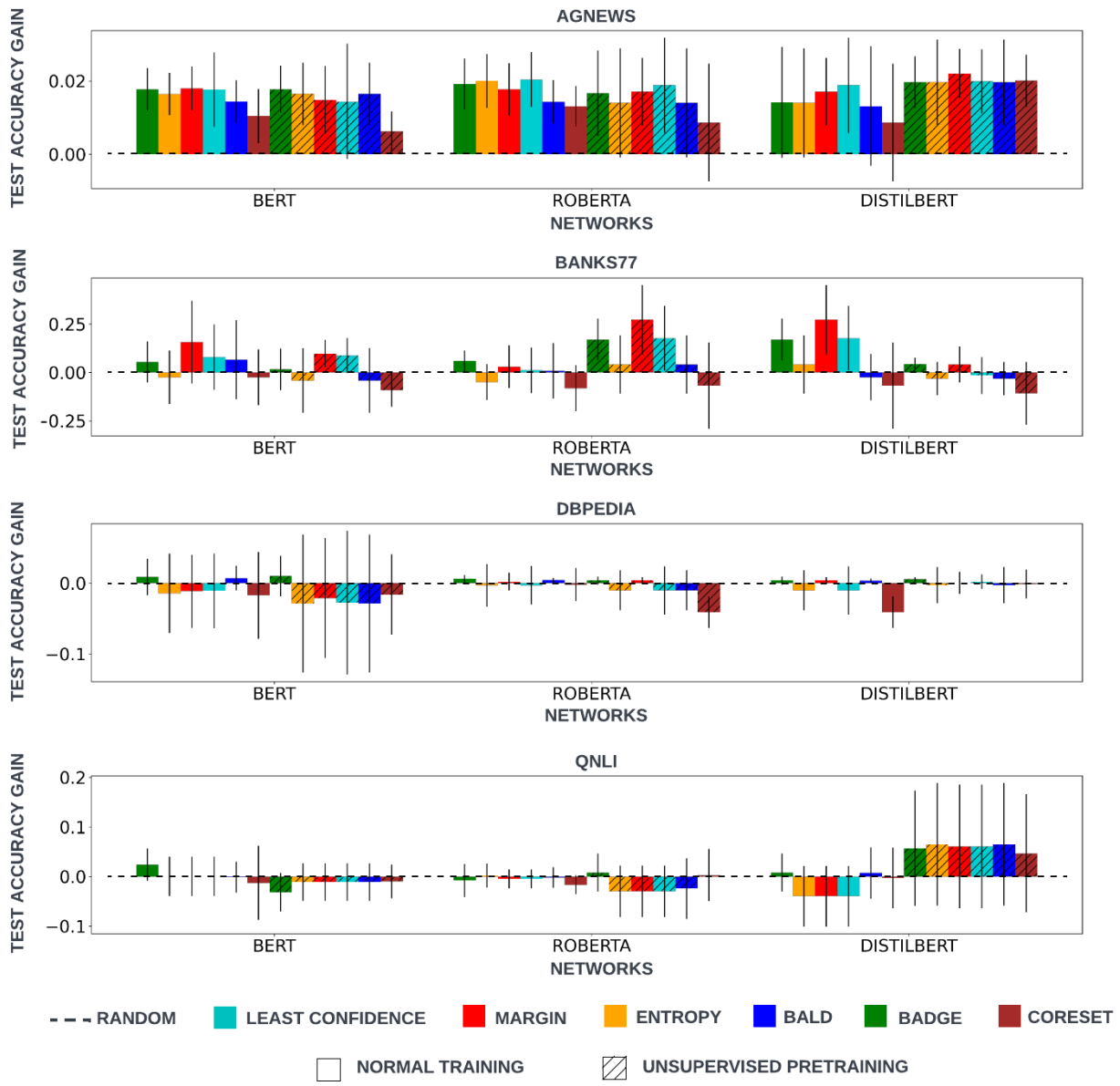

Figure 5: Comparison of gain in test-accuracy over random acquisition (as 0.00 test accuracy gain), for all six acquisition function with (hatched) and without unsupervised pretraining across four datasets and all transformer architectures on text tasks. Similar to Figure 4, results show that the relative performance of acquisition strategies varies both across architecture choices on the same dataset, as well as across datasets.

with and without augmentation appears to be minimal. Additionally (not shown on Fig. 7), consistent to existing findings (Beck et al., 2021; Mittal et al., 2019), the performance gain obtained by data augmentation was more substantial than the use of DAL.

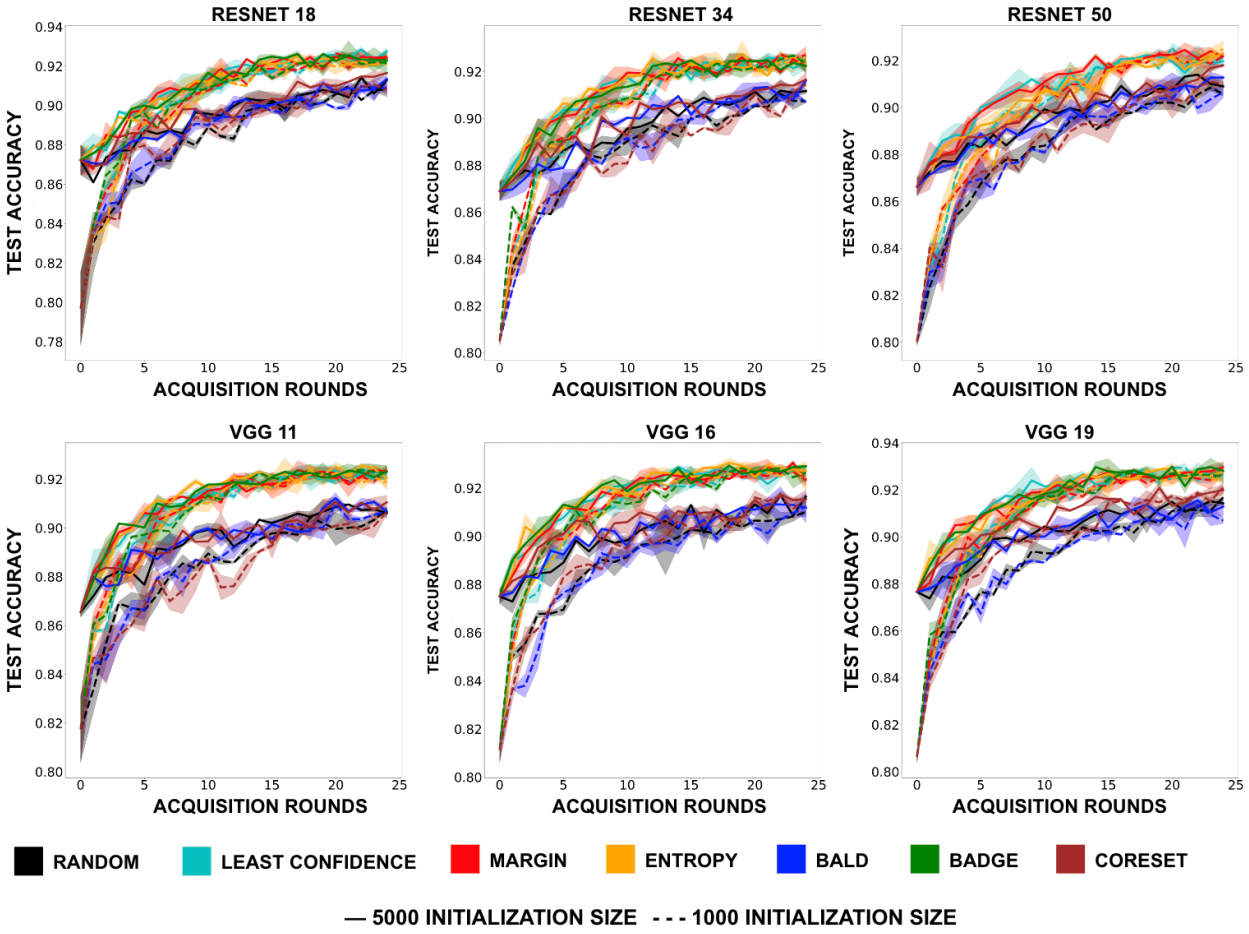

Figure 6: Comparison of test accuracy of different acquisition functions for different DNNs trained on FashionMNIST with an initialization size of 1000 (dashed) *vs.* 5000 (solid). Similar results on other datasets can be found in the Appendix C.1. The increase in initial pool size brought a gain in test performance for all DNNs, although this performance gain decreased for all acquisition functions as DAL proceeds.

4 Focus 2: DNN Architecture Optimization during DAL

4.1 Methodology

In this Focus, we investigate how DAL may be affected by the optimization of the architecture of DNN feature extractors. We consider three approaches to optimize the DNN architecture, namely supervised joint-training, supervised pre-training, and unsupervised pre-training. In *supervised joint-training*, we utilize the labeled data that are increasingly acquired during DAL. This translates to a setting that is similar to (Geifman & El-Yaniv, 2018), where DNN architecture and weight parameters are simultaneously optimized as DAL proceeds. In *supervised pre-training*, we utilize the initial labelled data to optimize the DNN architecture prior to active learning. In *unsupervised pre-training*, we utilize the unlabeled data to optimize DNN architecture prior to active learning. The pre-trained DNN is then used to initialize DAL during which the DNN's weight parameters are updated while the optimized architecture is kept fixed. This was motivated by the recent DAL work that advocated for unsupervised DNN pre-training (Siméoni et al., 2021) but on fixed architectures. In all three settings, we consider three representative architecture optimization approaches.

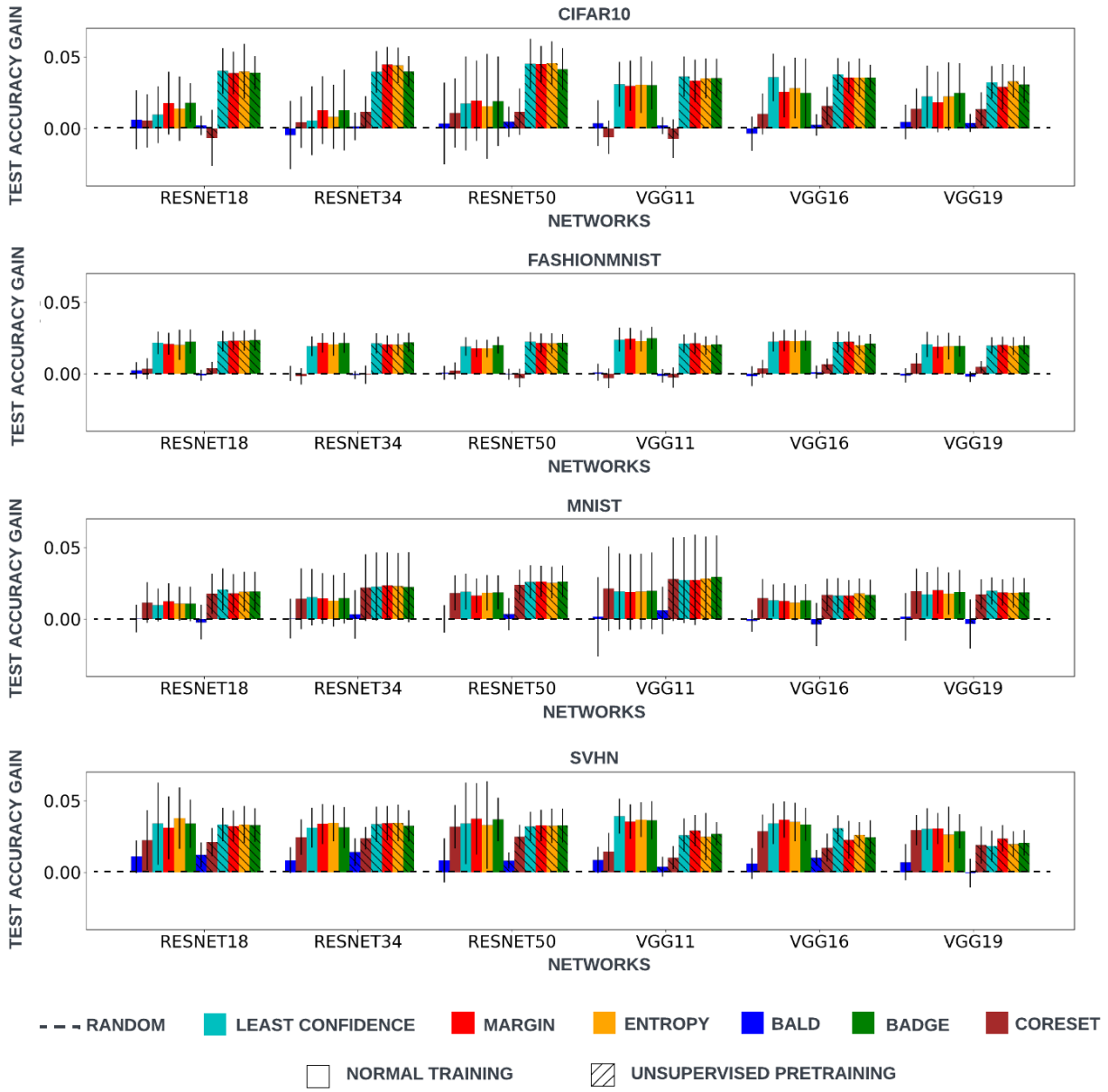

Figure 7: Gain in test-accuracy over random acquisition (as 0.00 test accuracy gain) for different acquisition functions across DNN architectures on four datasets, with (hatched) and without data augmentation. With data augmentation, most acquisition functions exhibited a similar gain over the random acquisition, and their relative performance appeared to be more clearly separated into two groups (high- *vs.* low-performing).

DNN architecture optimization approach during DAL: We consider three approaches to optimize the DNN architecture during DAL. In *supervised joint-optimization of DNN architectures and data acquisition,* at each acquisition round within DAL, we iterated between data selection given the choice of acquisition function, and the optimization of the CNN architectural and weight parameters given the new data. In *unsupervised pre-optimization (UPO) of DNN architectures,* we adopt the idea from a recent work (Siméoni et al., 2021; Caron et al., 2018) that pre-trains a DNN by iteratively clustering all unlabeled data and using the obtained clusters as pseudo-labels to train the DNN. In *supervised pre-optimization (SPO) of DNN architectures*, we use a classification task on initial labelled data to pre-train the DNN. While the original

work (Siméoni et al., 2021; Caron et al., 2018) utilized this to pre-train the weight parameters of the DNN, we use this to simultaneously optimize the architecture and weight parameters of the DNN.

The optimized DNN architecture is then kept fixed while the pre-trained weight parameters are used to initialize the DNN at each DAL acquisition.

DNN architectures: The computational cost associated with architecture optimization is high, especially for complex architectures such as RESNET and transformers. Existing works in architecture inference or optimization on transformers are also limited. Therefore we consider CNN as the choice of architecture in Focus 2.

DNN architecture optimization: We consider three methods for architecture optimizations, including two Bayesian inference methods and one NAS method, considering mainly the computational feasibility in including these methods in DAL.

For NAS, we follow the PDARTS method described in (Chen et al., 2019) that defines an over-complete network with L cells each with N nodes. Each node signifies a feature layer and two nodes are connected by operations $o \in O$. We define each subsequent node x_j as the linear combination of operations on node x_i defined by architecture parameter $\alpha^{i,j}$.:

$$x_j = \sum_{i<j} \sum_{o \in O_{i,j}} \frac{exp(\alpha_o^{i,j})}{\sum_{o' \in O} exp(\alpha_{o'}^{i,j})} o(x_i) \tag{5}$$

The optimization is split multiple steps where for each step u, the number of operations in the operations space is reduced in comparison to previous step v. In addition, the number of skip-connects are controlled by setting dropout rate which increases as we go deeper. As described in (Chen et al., 2019), the architecture parameters are optimized on validation data and the weights of the optimal network are tuned on training data. Training time varies with the amount of data available during DAL, ranging approximately from 1.5 hours using 1000 data samples to 8 hours for 30000 data samples.

We also include two Bayesian architecture inference methods described in (KC et al., 2021) (Depth-Dropout) and (Lee et al., 2018) (BBDropout) that translates to the optimization of the number of CNN layers (depth) and the number of CNN filters in each layer (width). In Depth-Dropout, the hidden layer is observed as:

$$h_l \quad = \sigma(W_l \odot h_{l-1}) \bigotimes z_l + h_{l-1} \tag{6}$$

and for BBDropout, we have:

$$h_l \quad = \sigma((W_l \odot h_{l-1}) \bigotimes z_l) + h_{l-1} \tag{7}$$

where h_l is the feature map of the lth hidden layer, W_l is the weight matrix for CNN filters in the lth layer, z_l is the activation mask for l^{th} layer, $\odot$ is a convolution operation, and $\sigma(.)$ is an activation function. We define a beta process as a prior over the number of hidden layers. A beta process sample can be denoted as (h_l, π_l), where $\pi_l \in [0, 1]$ denotes the activation probability of a hidden layer function h_l. A stick-breaking construction of the beta process can be represented as:

$$\pi_l \quad = \prod_{j=1}^{l} v_j, \quad v_l \sim Beta(\alpha, \beta) \tag{8}$$

For BBDropout, in contrast, the IBP is defined over neurons per hidden layer *i.e.*,

$$\pi_l \quad = Beta(\alpha_l, \beta_l) \tag{9}$$

We then define a conjugate Bernoulli process inducing layer-wise binary vectors z_l to drop out neurons per layer. The prior over the network structure variable Z can thus be formulated as

$$p(Z, v|\alpha, \beta) = p(Z|v)p(v|\alpha, \beta) = \prod_{m=1}^{M} Bern(z_{ml}|\pi_l) \prod_{l=1}^{\infty} Beta(v_l|\alpha, \beta) \tag{10}$$

The variational distribution with truncation number of layers K to approximate the true posterior distribution for the Depth-Dropout method is defined as:

$$q(Z, v|a_{l=1}^K, b_{l=1}^K) = q(Z|v)q(v) = \prod_{m=1}^{M} ConBern(z_{ml}|\pi_l) \prod_{l=1}^{K} Beta(v_l|a_l, b_l) \tag{11}$$

where $ConBern$ is a continuous relaxation of Bernoulli distribution. For simplicity, the number of filters in each layer is limited to M

In BBDropout, the posterior is approximated with the variational distribution of form:

$$q(Z, \pi|X) = \prod_{m=1}^{M} q(\pi_k)q(z_m|\pi_m) \tag{12}$$

where we use Kumaraswamy distribution (Kumaraswamy, 1980) for $q(\pi_m)$ and continuous relaxation of Bernoulli distribution for $q(z_m|\pi_m)$ (Lee et al., 2018; Maddison et al., 2016; Jang et al., 2016; Gal et al., 2017a) .

The distinction lies in the utilization of a beta process and its corresponding Bernoulli process independently by DepthDropout, enabling the inference of both the number of layers and nodes per layer. On the other hand, BBDropout marginalizes the beta process, inferring the number of nodes in each layer but not the depth. For both DepthDropout and BBDropout, the evidence lower bound (ELBO) of the marginal likelihood of observed data D can be derived and optimized via structured stochastic variational inference (SSVI) as described in (KC et al., 2021).

Evaluation metrics: Here we focus on the relative performance of DAL on fixed *vs.* optimized DNN architectures, as well as the relative contribution of DNN architecture optimization *vs.* data acquisition optimization to the final DNN performance. We measure this by the *spread of performance* due to the optimization of one factor when the other is controlled. This is visualized as well as quantified as the absolute difference between the maximum and minimum accuracy achieved when one of these two factors varies.

4.2 Experiments and Results

Experiments in Focus 2 were performed on four image datasets including MNIST (Deng, 2012), Fashion MNIST (Xiao et al., 2017), CIFAR10 (Krizhevsky, 2009), and SVHN (Netzer et al., 2011). For Depth-Dropout and BBDropout method, we used CNN with truncation $K = 20$ and 64 filters in each layer. A uniform distribution $U(0, 1.1)$ and $U(0, 1.0)$ was used to initialize the prior of architecture parameters a and b in equation 11 for Depth-Dropout method. The parameters a and b were initialized to 1.3133 and 1.000, respectively, in equation 9 for BBDropout. For PDARTS, we define a total of eight operations (Max pool 3x3, Average Pool 3x3, Skip-connect, Identity, Separable Convolution 3x3 and 5x5, Dilated Convolution 3x3 and 5x5) which are dropped off with dropout rate of 0.1, 0.4 and 0.7 in the subsequent training steps as used in Chen et al. (2019). All the models were trained for 500 epochs. Experiments were performed on workstations with RTX 2080Ti GPU and 32 GB of RAM as well as P8 and V100 GPU. The active learning experiments are run on four image dataset in same setting as described in 3.2.

The effect of optimized DNN architectures: Fig. 8 summarizes the DAL performance across different acquisition functions when applied to architecture optimization – Depth-Dropout (A), BBDropout (B), and PDARTs(C) – and fixed DNN architectures with 1, 5, and 9 layers on CIFAR10 and FashionMNIST. The DNN architectures were optimized either jointly or during pre-training using both supervised and unsupervised methods. Complete results on the rest of the datasets can be found in Appendix D.1. The figure depict that the optimized networks, particularly those optimized jointly or through supervised pre-optimization (SPO) approach, consistently outperform fixed pre-defined networks. SPO and joint optimization approaches, in general, showed consistent gains in all architecture optimization methods (i.e. Fig. 8A, B and C) in comparison to unsupervised pre-optimization (UPO) approach which performed least favorably. This may be attributed to the fact that the network optimization with unlabelled data was based on a clustering task different for primary DAL task.

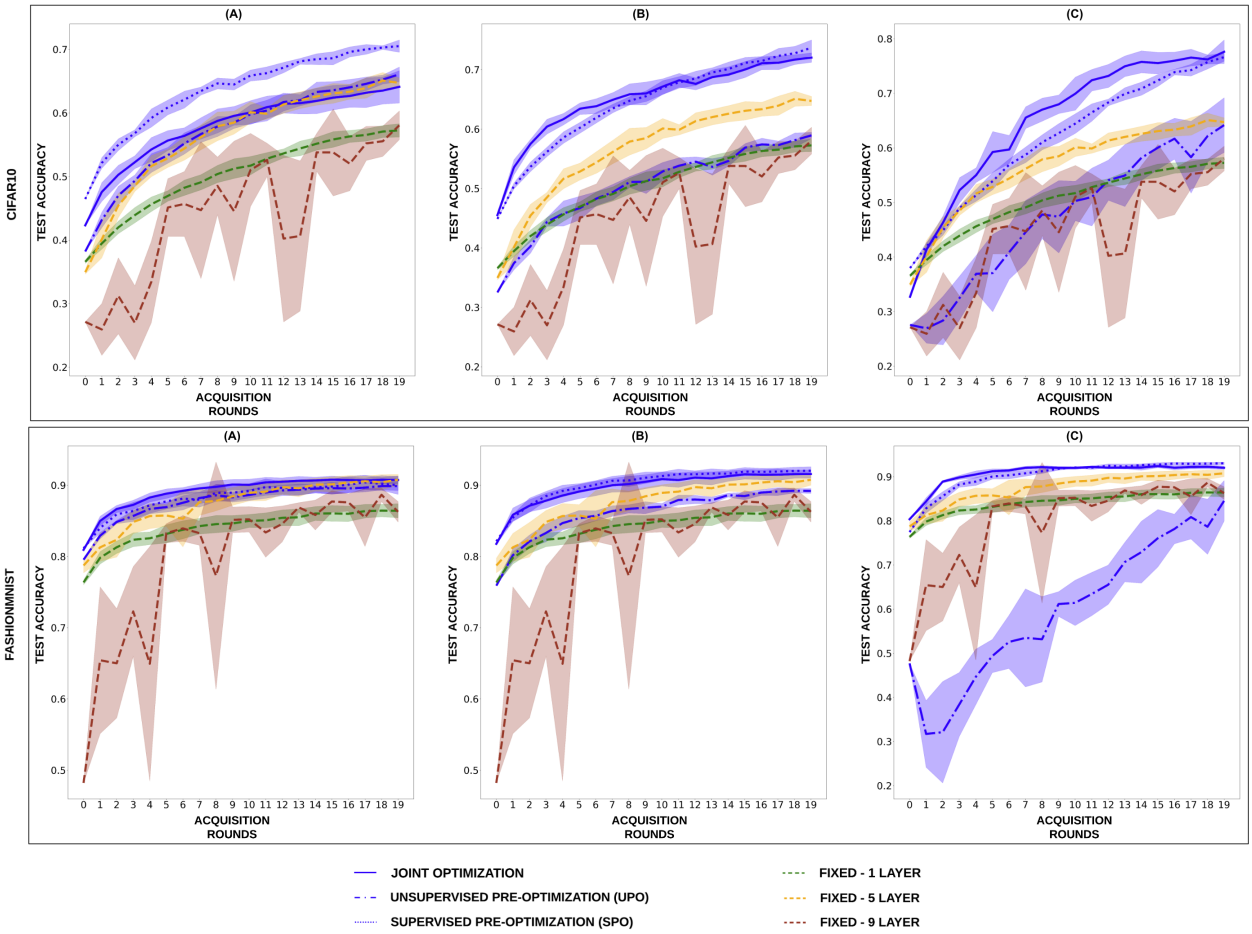

Figure 8: Spread of performance across different acquisition functions given pre-defined (fixed) CNN architecture or jointly-optimized, supervised pre-optimized (SPO), unsupervised pre-optimized (UPO) using Depth-Dropout (A), BBDropout(B) and PDARTs(C) for CIFAR10 (top row) and FashionMNIST (bottom row) . Optimization of the DNN architecture, either jointly or during pre-training, in general improved over pre-defined CNN performance.

In case of joint optimization approach, the continuous change in the size of labeled data during DAL resulted in changes in optimal DNN architecture. This, in turn, induces continuous modifications to the optimal feature space, complicating the joint optimization with data selection. To corroborate this, we experimented with more extreme changes of data size by considering an initial size of 200 and acquisition size of 100 in the same experiments. Figure 9 (A) and (B) shows the optimized architecture using Depth Dropout as DAL proceeds for CIFAR10 and FashionMNIST respectively, where the bar graph on top indicates the probability of activation of a layer and the column below indicates the activation of filters in each layer. As shown, both the depth and width of the CNN increased as DAL proceeded with adding labeled data. These results suggested that the joint optimization approach via small and growing labeled data in addition to weight optimization and active learning may complicate optimal data acquisition due to the continued change in the feature space, whereas leveraging initial labeled data to pre-optimize the feature space may offer a simpler and yet more competitive solution.

The limited performance with unsupervised pre-training using PDARTS may be attributed to the sensitivity of the unsupervised labeling task to the number of clusters used for generating pseudo-labels. Through our experiments, we observed that employing a larger cluster size with PDARTS resulted in an optimized network with minimal test accuracy ($< 30\%$), which improved as the number of clusters approached the actual number of labels in the dataset. In contrast, supervised pre-training of PDARTS with an initially

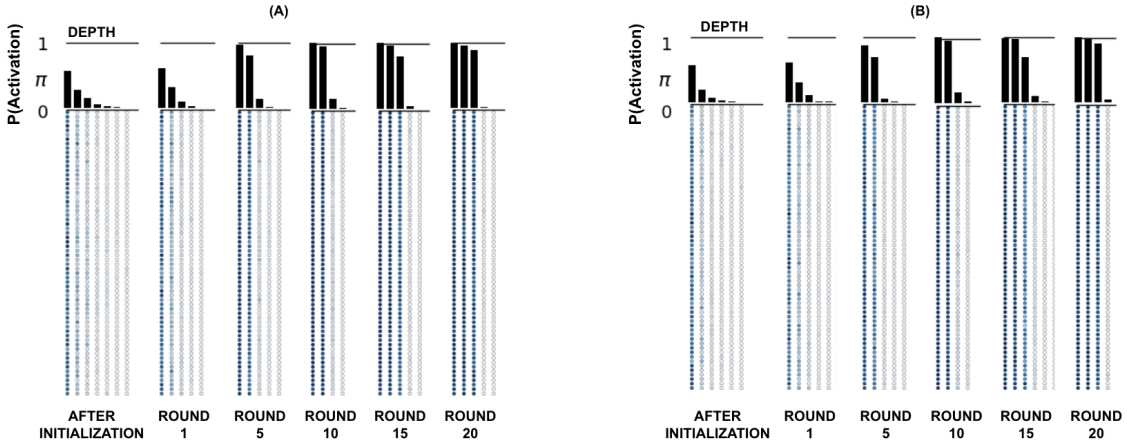

Figure 9: Visualization of optimized CNN architecture as DAL proceeds using Depth-Dropout. A: CIFAR10; B: FashionMNIST.

labeled dataset yielded a significantly superior optimized network compared to the unsupervised approach. This suggests that unsupervised pre-optimization of DNN architectures may be difficult.

Relative contribution of architecture *versus* data optimization: The shade or spread in Fig. 8 describe the spread of performance between different acquisition functions for different architecture optimizations. A closer look into Fig. 8 suggest that, with the exception of UPO on PDARTS, the performance spread across different acquisition functions is reduced using an optimized DNN architecture compared to fixed pre-defined networks. Figure 10 further summarizes the spread of performance across all optimized or fixed DNN architectures, for each given acquisition function on CIFAR10 and FashionMNIST. Complete results on the rest of the datasets can be found in Appendix D.2. As shown, different choices of architecture parameters induced a large performance gap of the DNN at any given data size for any acquisition function used. This performance spread changed as DAL proceeded, although the trend of change was not consistent among datasets: on FashionMNIST, the gap among different CNN architectures appeared to be larger at the earlier stages of DAL when the data size was smaller, whereas on CIFAR this gap appeared to increase as DAL proceeded Contrasting Fig. 8 with Figure 10, it is evident that the impact of the DNN architectures substantially outweighed and even reduced the impact of acquisition strategies. This further suggests that, when the labeling budget is small, the effort to identify optimal architecture of DNN feature extractor may be critical in order to maximize the efficacy of active data selection.

5 Discussion

Our current investigation show two important discovery. Firstly, we observe that the relative performance of DAL acquisition functions depend on the complexity and structure of the underlying DNN architectures even when applied to a given dataset. This suggests that the optimal data acquisition in DAL is particularly influenced by the feature space defined by the choice of DNN. This highlights that, while architecture optimization is an essential aspect in DNNs in general, its importance becomes even more pronounced in the context of DAL as it has a major influence on the effectiveness of data acquisition strategies. Ensuring consistent experimental setups regarding the underlying DNN architectures can be an important aspect for the DAL community to consider for reproducible and generalizable research outcome.

Secondly, optimization of DNN network architecture, either preceding DAL or jointly with DAL, plays an important part in DAL. Moreover, supervised pre-optimization of DNN architecture, utilizing the initial labeled data, may be more beneficial followed by joint architecture and data optimization and then the unsupervised pre-optimization during DAL. This underscores the substantial influence of DNN architectures in outweighing the impact of acquisition functions. These findings shed light on the importance of strate-

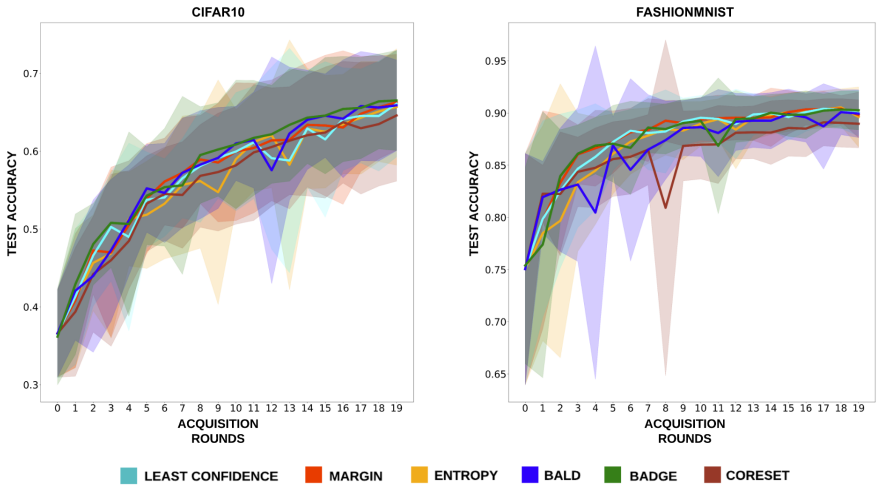

Figure 10: Spread of performance across different architectures – pre-defined architecture and architeure optimization of CNN architectures – given acquisition functions for CIFAR10 and FashionMNIST, in the form of spread of test accuracy over the course of DAL. Comparison with Fig. 8 shows that the optimization of DNN architecture outweighed the effect of data selection in the overall DAL performance. Similar results for MNIST and SVHN are shown in Appendix D.2

gic DNN architecture optimization in DAL framework, emphasizing on the necessity of considering DNN architecture optimization in DAL.

Investigations in Focus 2 of the current study is focused on CNN based architecture optimization. To generalize the findings, future studies need to extend to larger architectures such as RESNET, VGG, and transformers. Such extension will provide a more comprehensive understanding of how optimization of different network architectures influence data optimization, providing insights into the applicability of joint architecture and data optimization strategies across a variety of DNN types.

There is also room for incorporating additional data acquisition strategies in the presented study. The current study examined a range of acquisition functions, covering prevalent strategies representative of uncertainty-based, diversity-based, and hybrid strategies found in the existing literature of DAL. Future work can broaden the spectrum of acquisition strategies to more recent strategies, such as those incorporating neural tangent kernels to assess DNN training dynamics (Wang et al., 2022), and those incorporating the concept of semi-supervised learning in DAL.

In terms of approaches for optimizing DNN architectures, we considered a cell-based (NAS) method, specifically PDARTs, along with two Bayesian DNN architecture inference methods. For CNN networks considered in the Focus 2 of this study, the connections between operations within a search cell exhibit a large influence on the architecture's performance. However, cell-based NAS methods are not directly transferrable to transformers that operate on attention mechanism rather than convolution. Future work will incorporate NAS methods for transformers, such as AdaBERT and NAS-BERTs, to further test the generalizability of the findings obtained in Focus 2 regarding the effect of DNN architecture optimization on DAL. Future work can also include additional NAS methods for CNN-based architecture, such as dynamic-exploration DARTs where the dynamic architecture varies the kernel size or network path of CNN according to input data.

Finally, to investigate potential strategies for simultaneously optimizing DNN architecture and data selection, we considered unsupervised pre-optimizaiton of DNN artechitecture prior to DAL, versus supervised joint-optimization of DNN during DAL. Both approaches, however, essentially considered DNN architecture and data optimization as two separate optimization problems with their respective objective functions. An interesting yet much more challenging future research direction may be the developments of joint DNN architecture and data optimization theories, methods, and algorithms that integrate these two optimization objectives in a more coherent formulation.

6 Conclusion

In this work, we examine the influence of DNN architectures on optimal data selection in DAL. We show that the choices of DNN architecture substantially influence and outweigh data optimization in DAL, and that its optimization helps increase the benefits of active data selection. We hope that the findings help inform the research community in improving the reproducibility of DAL evaluations by taking into account the important role of DNN architecture choices in DAL, and in opening up new research avenues that better integrate DNN architecture optimization to maximize the benefits of active data selection when the labeling budget is limited.

References

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=ryghZJBKPS`.

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*, pp. 722–735. Springer, 2007.

Nathan Beck, Durga Sivasubramanian, Apurva Dani, Ganesh Ramakrishnan, and Rishabh Iyer. Effective evaluation of deep active learning on image classification tasks. *arXiv preprint arXiv:2106.15324*, 2021.

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. *arXiv preprint arXiv:2003.04807*, 2020.

Xin Chen, Lingxi Xie, Jun Wu, and Qi Tian. Progressive differentiable architecture search: Bridging the depth gap between search and evaluation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1294–1303, 2019.

David A. Cohn, Zoubin Ghahramani, Michael I. Jordan, and David Cohn. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 1996.

Luiz FS Coletta, Moacir Ponti, Eduardo R Hruschka, Ayan Acharya, and Joydeep Ghosh. Combining clustering and active learning for the detection and learning of new image classes. *Neurocomputing*, 358: 150–165, 2019.

Apurva Dani, Durga Sivasubramanian, and Rishabh Iyer. Distil. `https://github.com/decile-team/distil`, 2021.

Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Georgi Dikov and Justin Bayer. Bayesian learning of neural network architectures. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 730–738. PMLR, 2019.

Jiashi Feng and Trevor Darrell. Learning the structure of deep convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2749–2757, 2015.

Alexander Freytag, Erik Rodner, and Joachim Denzler. Selecting influential examples: Active learning with expected model output changes. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*, pp. 562–577. Springer, 2014.

Yarin Gal, Jiri Hron, and Alex Kendall. Concrete dropout. *Advances in neural information processing systems*, 30, 2017a.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pp. 1183–1192. PMLR, 2017b.

Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan Ö Arık, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pp. 510–526. Springer, 2020.

Yonatan Geifman and Ran El-Yaniv. Deep active learning over the long tail. *arXiv preprint arXiv:1711.00941*, 2017.

Yonatan Geifman and Ran El-Yaniv. Deep active learning with a neural architecture search. *arXiv preprint arXiv:1811.07579*, 2018.

Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. *arXiv preprint arXiv:2202.02794*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Tao He, Xiaoming Jin, Guiguang Ding, Lan Yi, and Chenggang Yan. Towards better uncertainty sampling: Active learning with multiple views for deep convolutional neural network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1360–1365. IEEE, 2019.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Qiang Hu, Yuejun Guo, Maxime Cordy, Xiaofei Xie, Wei Ma, Mike Papadakis, and Yves Le Traon. Towards exploring the limitations of active learning: An empirical study. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 917–929. IEEE, 2021.

Jiaji Huang, Rewon Child, Vinay Rao, Hairong Liu, Sanjeev Satheesh, and Adam Coates. Active learning for speech recognition: the power of gradients. *arXiv preprint arXiv:1612.03226*, 2016.

Siyu Huang, Tianyang Wang, Haoyi Xiong, Bihan Wen, Jun Huan, and Dejing Dou. Temporal output discrepancy for loss estimation-based active learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Ajay J Joshi, Fatih Porikli, and Nikolaos Papanikolopoulos. Multi-class active learning for image classification. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2372–2379. IEEE, 2009.

Christoph Käding, Erik Rodner, Alexander Freytag, and Joachim Denzler. Active and continuous exploration with deep neural networks and expected model output changes. *arXiv preprint arXiv:1612.06129*, 2016.

MF Kasim, D Watson-Parris, L Deaconu, S Oliver, P Hatfield, DH Froula, G Gregori, M Jarvis, S Khatiwala, J Korenaga, et al. Building high accuracy emulators for scientific simulations with deep neural architecture search. *arXiv preprint arXiv:2001.08055*, 2020.

Kishan KC, Rui Li, and MohammadMahdi Gilany. Joint inference for neural network depth and dropout regularization. *Advances in Neural Information Processing Systems*, 34, 2021.

Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8166–8175, 2021.

Hannah Rose Kirk, Bertie Vidgen, and Scott A Hale. Is more data better? re-thinking the importance of efficiency in abusive language detection with transformers-based active learning. *arXiv preprint arXiv:2209.10193*, 2022.

Seo Taek Kong, Soomin Jeon, Dongbin Na, Jaewon Lee, Hong-Seok Lee, and Kyu-Hwan Jung. A neural pre-conditioning active learning algorithm to reduce label complexity. *Advances in Neural Information Processing Systems*, 35:32842–32853, 2022.

Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

Ponnambalam Kumaraswamy. A generalized probability density function for double-bounded random processes. *Journal of hydrology*, 46(1-2):79–88, 1980.

Juho Lee, Saehoon Kim, Jaehong Yoon, Hae Beom Lee, Eunho Yang, and Sung Ju Hwang. Adaptive network sparsification with dependent variational beta-bernoulli dropout. *arXiv preprint arXiv:1805.10896*, 2018.

Ya Li, Keze Wang, Lin Nie, and Qing Wang. Face recognition via heuristic deep active learning. In *Chinese Conference on Biometric Recognition*, pp. 97–107. Springer, 2017.

Peng Liu, Hui Zhang, and Kie B Eom. Active deep learning for classification of hyperspectral images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 10(2):712–724, 2016.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Jinghui Lu and Brian MacNamee. Investigating the effectiveness of representations based on pre-trained transformer-based language models in active learning for labelling text datasets. *arXiv preprint arXiv:2004.13138*, 2020.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Dwarikanath Mahapatra, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio Reyes. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 580–588. Springer, 2018.

Christoph Mayer and Radu Timofte. Adversarial sampling for active learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3071–3079, 2020.

Sudhanshu Mittal, Maxim Tatarchenko, Özgün Çiçek, and Thomas Brox. Parting with illusions about deep active learning. *arXiv preprint arXiv:1912.05361*, 2019.

Prateek Munjal, Nasir Hayat, Munawar Hayat, Jamshid Sourati, and Shadab Khan. Towards robust and reproducible active learning using neural networks. *arXiv preprint arXiv:2002.09564*, 2020.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011.

Natalia Ostapuk, Jie Yang, and Philippe Cudré-Mauroux. Activelink: deep active learning for link prediction in knowledge graphs. In *The World Wide Web Conference*, pp. 1398–1408, 2019.

Hiranmayi Ranganathan, Hemanth Venkateswara, Shayok Chakraborty, and Sethuraman Panchanathan. Deep active learning for image classification. In *2017 IEEE International Conference on Image Processing (ICIP)*, pp. 3934–3938. IEEE, 2017.

Lukas Rauch, Matthias Aßenmacher, Denis Huseljic, Moritz Wirth, Bernd Bischl, and Bernhard Sick. dal-toolbox. `https://github.com/dhuseljic/dal-toolbox`, 2023.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Brij B Gupta, Xiaojiang Chen, and Xin Wang. A survey of deep active learning. *ACM Computing Surveys (CSUR)*, 54(9):1–40, 2021.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pp. 309–318. Springer, 2001.

Christopher Schröder, Andreas Niekler, and Martin Potthast. Revisiting uncertainty-based query strategies for active learning with transformers. *arXiv preprint arXiv:2107.05687*, 2021.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Burr Settles. Active learning literature survey. 2009.

Burr Settles, Mark Craven, and Soumya Ray. Multiple-instance active learning. *Advances in neural information processing systems*, 20:1289–1296, 2007.

Artem Shelmanov, Dmitri Puzyrev, Lyubov Kupriyanova, Denis Belyakov, Daniil Larionov, Nikita Khromov, Olga Kozlova, Ekaterina Artemova, Dmitry V Dylov, and Alexander Panchenko. Active learning for sequence tagging with deep pre-trained models and bayesian uncertainty estimates. *arXiv preprint arXiv:2101.08133*, 2021.

Changjian Shui, Fan Zhou, Christian Gagné, and Boyu Wang. Deep active learning: Unified and principled method for query and training. In *International Conference on Artificial Intelligence and Statistics*, pp. 1308–1318. PMLR, 2020.

Megh Shukla. Bayesian uncertainty and expected gradient length-regression: Two sides of the same coin? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2367–2376, 2022.

Oriane Siméoni, Mateusz Budnik, Yannis Avrithis, and Guillaume Gravier. Rethinking deep active learning: Using unlabeled data at model training. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 1220–1227. IEEE, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5972–5981, 2019.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Haonan Wang, Wei Huang, Ziwei Wu, Hanghang Tong, Andrew J Margenot, and Jingrui He. Deep active learning by leveraging training dynamics. *Advances in Neural Information Processing Systems*, 35:25171–25184, 2022.

Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.

Min Wang, Fan Min, Zhi-Heng Zhang, and Yan-Xue Wu. Active learning through density clustering. *Expert systems with applications*, 85:305–317, 2017.

Lukas Wertz, Katsiaryna Mirylenka, Jonas Kuhn, and Jasmina Bogojeska. Investigating active learning sampling strategies for extreme multi label text classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 4597–4605, 2022.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, 2017.

Donggeun Yoo and In So Kweon. Learning loss for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 93–102, 2019.

Anman Zhang, Bohan Li, Wenhuan Wang, Shuo Wan, and Weitong Chen. Mii: A novel text classification model combining deep active learning with bert. *Computers, Materials & Continua*, 63(3):1499–1514, 2020a.

Beichen Zhang, Liang Li, Shijie Yang, Shuhui Wang, Zheng-Jun Zha, and Qingming Huang. State-relabeling adversarial active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8756–8765, 2020b.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, 2015.

Ye Zhang, Matthew Lease, and Byron Wallace. Active discriminative text representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017.

Fedor Zhdanov. Diverse mini-batch active learning. *arXiv preprint arXiv:1901.05954*, 2019.

Jia-Jie Zhu and José Bento. Generative adversarial active learning. *arXiv preprint arXiv:1702.07956*, 2017.

Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8697–8710, 2018.

A Appendix

B Architectures of RESNET and VGG Networks

B.1 RESNET

Structure of VGG-11/16/19 and RESNET-18/34/50 are shown below in format Conv x, y where x is filter size and y represents number of filters.

	RESNET-18	RESNET-34	RESNET50
CONV LAYER	Conv 3, 64	Conv 3, 64	Conv 3, 64
BLOCK 1	[Conv 3, 64 Conv 3, 64] x 2	[Conv 3, 64 Conv 3, 64] x 3	[Conv 1, 64 Conv 3, 64 Conv 1, 256] x 3
BLOCK 2	[Conv 3, 128 Conv 3, 128] x 2	[Conv 3, 128 Conv 3, 128] x 4	[Conv 1, 128 Conv 3, 128 Conv 1, 512] x 4
BLOCK 3	[Conv 3, 256 Conv 3, 256] x 2	[Conv 3, 256 Conv 3, 256] x 6	[Conv 1, 256 Conv 3, 256 Conv 1, 1024] x 6
BLOCK 4	[Conv 3, 512 Conv 3, 512] x 2	[Conv 3, 512 Conv 3, 512] x 3	[Conv 1, 512 Conv 3, 512 Conv 1, 2048] x 3
FULLY CONNECTED LAYERS	FC width 512 FC width C	FC width 512 FC width C	FC width 2048 FC width C

B.2 VGG

	VGG-11	VGG-16	VGG-19
CONVOLUTION LAYERS	Conv 3, 64 Max-pool	2 x [Conv 3, 64] Max-pool	2 x [Conv 3, 64] Max-pool
	Conv 3, 128 Max-pool	2 x [Conv 3, 128] Max-pool	2 x [Conv 3, 128] Max-pool
	2 x [Conv 3, 256] Max-pool	3 x [Conv 3, 256] Max-pool	4 x [Conv 3, 256] Max-pool
	2 x [Conv 3, 512] Max-pool	3 x [Conv 3, 512] Max-pool	4 x [Conv 3, 512] Max-pool
	2 x [Conv 3, 512] Max-pool	3 x [Conv 3, 512] Max-pool	4 x [Conv 3, 512] Max-pool
FULLY CONNECTED LAYERS	2 x [FC width 4096] FC width 1000 FC width C	2 x [FC width 4096] FC width 1000 FC width C	2 x [FC width 4096] FC width 1000 FC width C

C Focus 1: The Effect of DNN Architectures on DAL Performances

C.1 Effect of Initial Training Size

C.1.1 CIFAR10

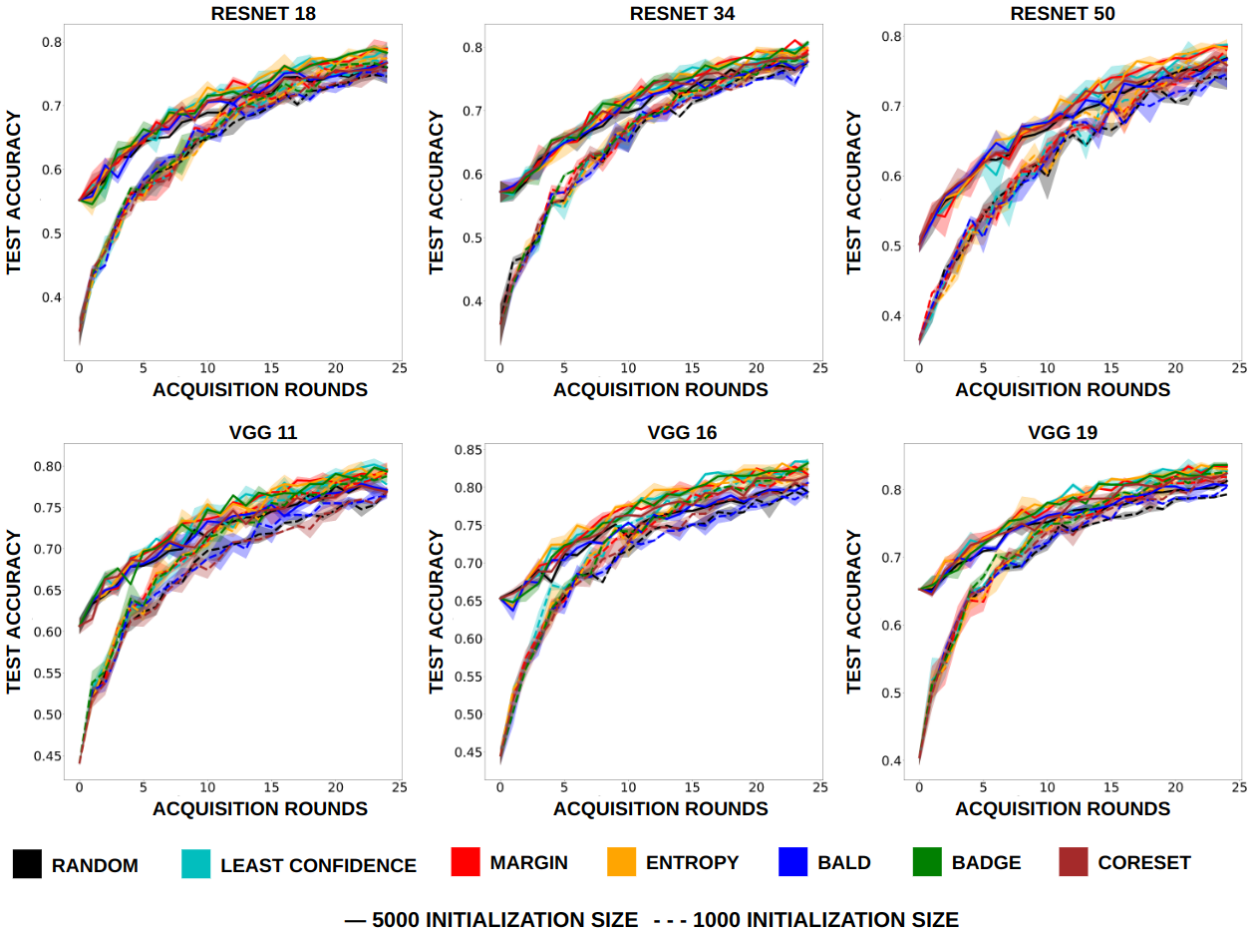

Figure 11: Comparison of test accuracy of different acquisition functions for different network trained on CIFAR10 with initialization size of 1000 (dashed) and 5000(solid).

C.1.2 FashionMNIST

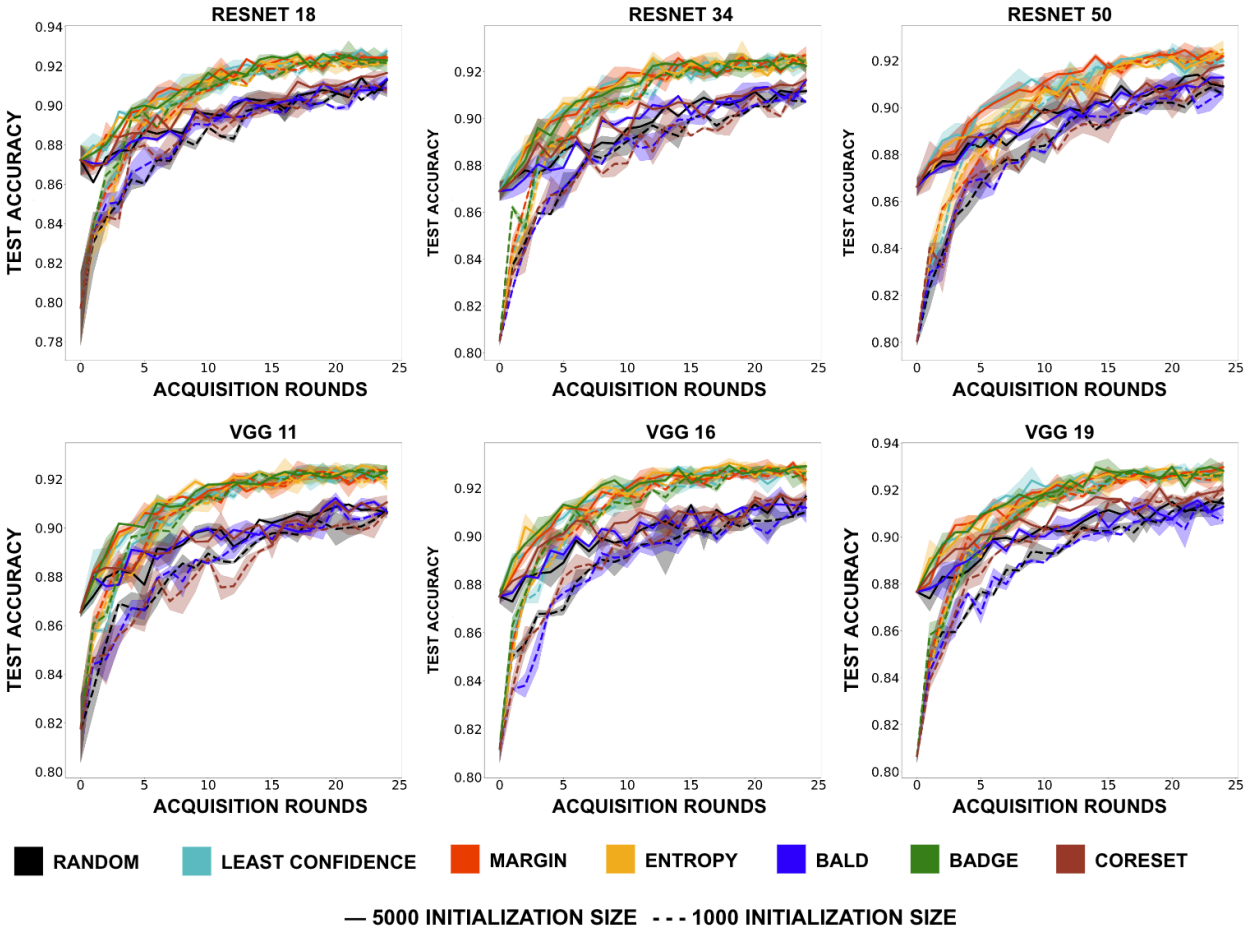

Figure 12: Comparison of test accuracy of different acquisition functions for different network trained on FashionMNIST with initialization size of 1000 (dashed) and 5000(solid).

C.1.3 MNIST

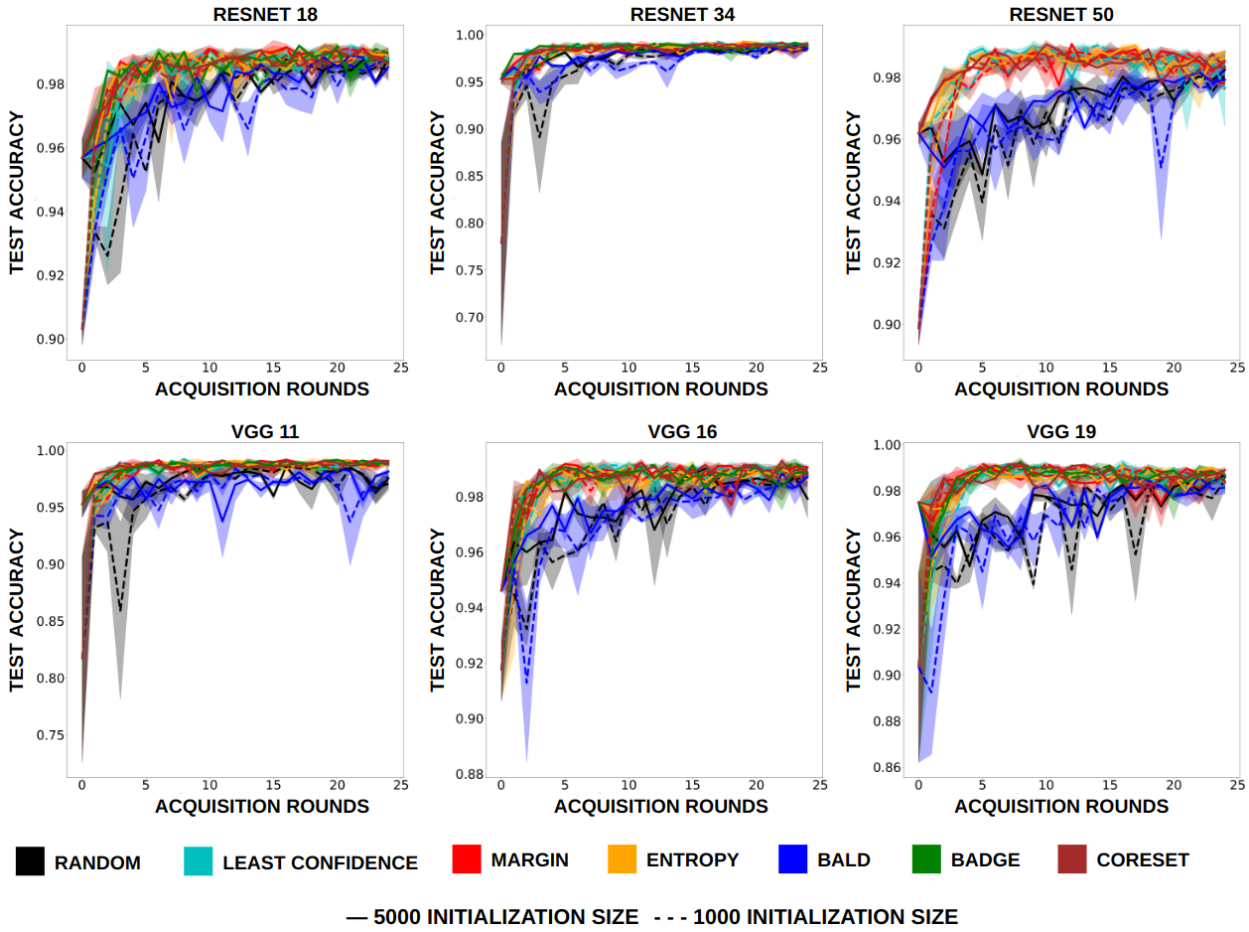

Figure 13: Comparison of test accuracy of different acquisition functions for different network trained on MNIST with initialization size of 1000 (dashed) and 5000(solid).

C.1.4 SVHN

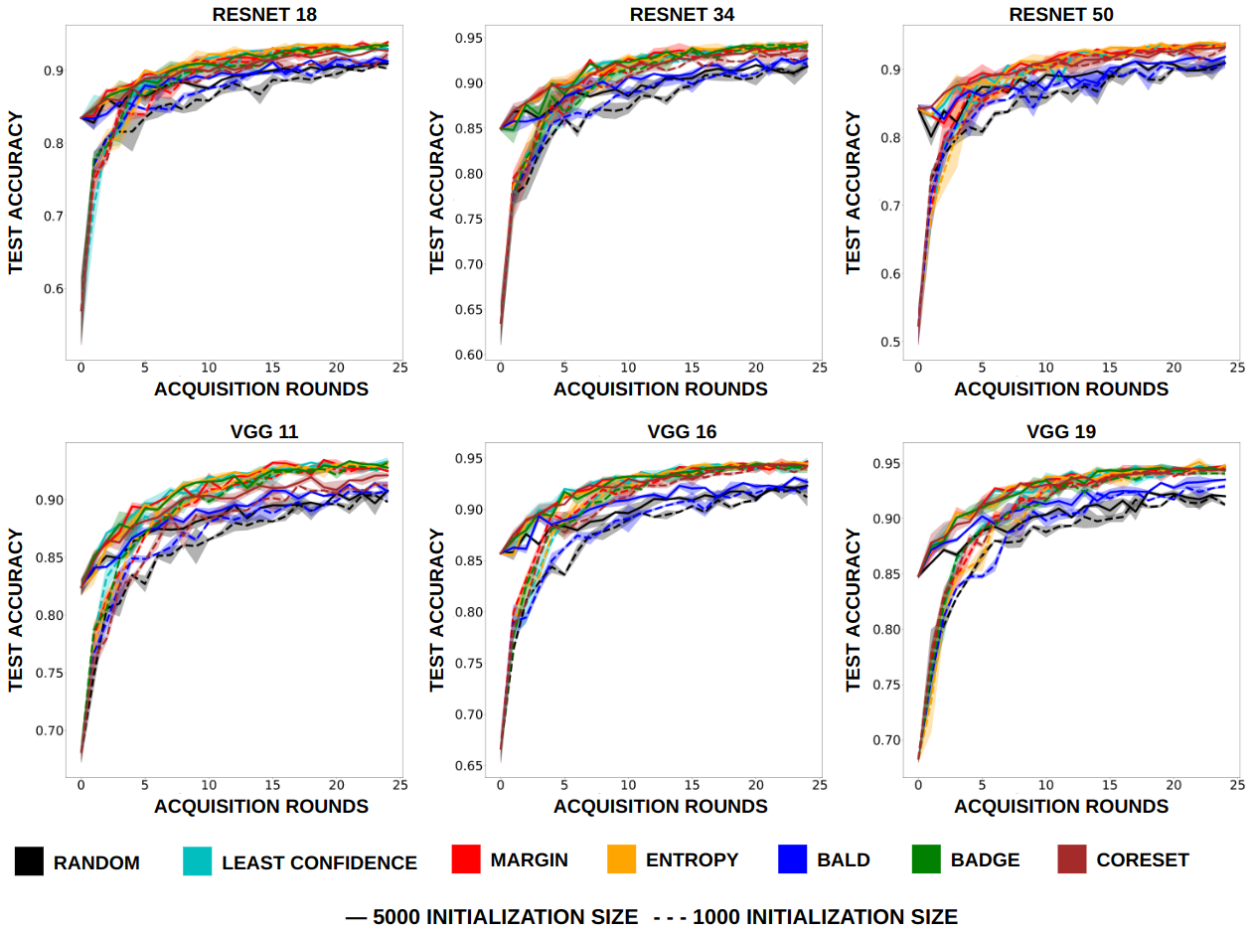

Figure 14: Comparison of test accuracy of different acquisition functions for different network trained on SVHN with initialization size of 1000 (dashed) and 5000(solid).

C.2 Effect of Acquisition size

C.2.1 CIFAR10

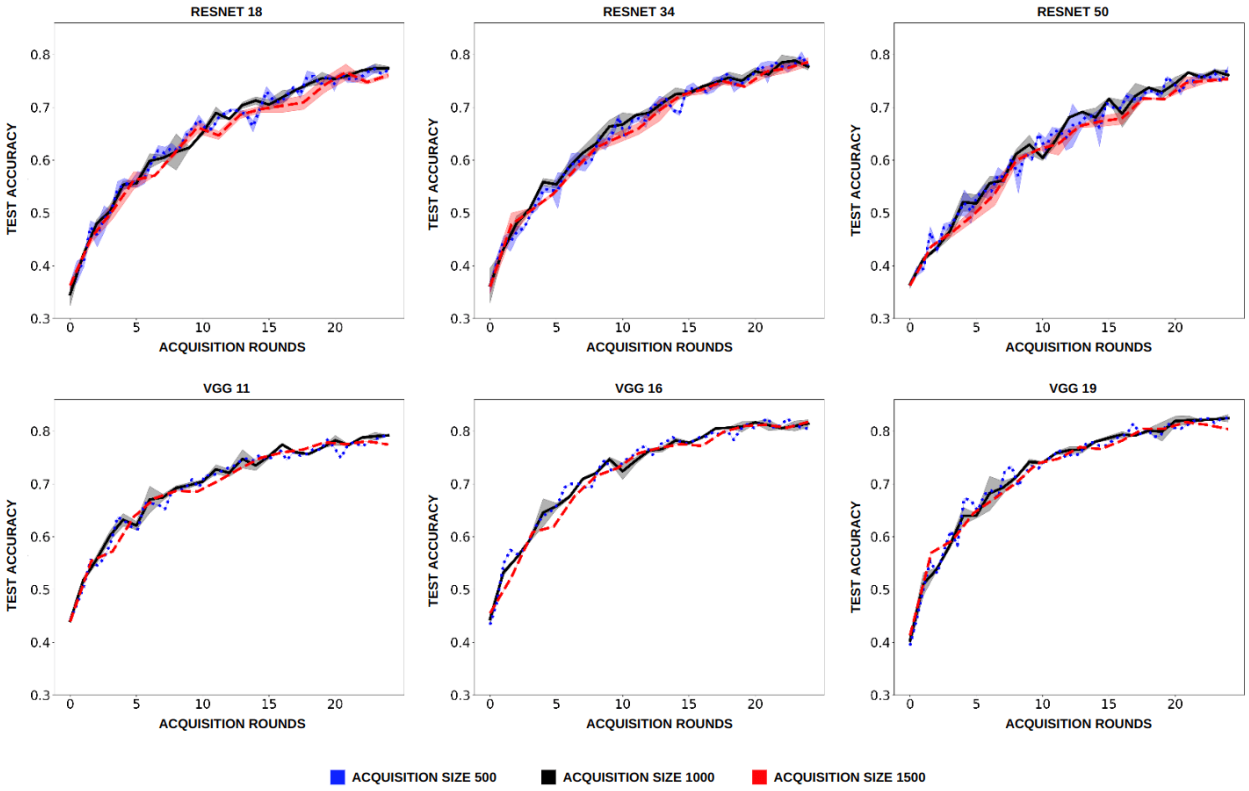

Figure 15: Visualization of performance of Entropy acquisition function on different networks trained on CIFAR10 with acquisition size of 500 (dotted blue), 1000 (solid black) and 1500 (dashed red) respectively

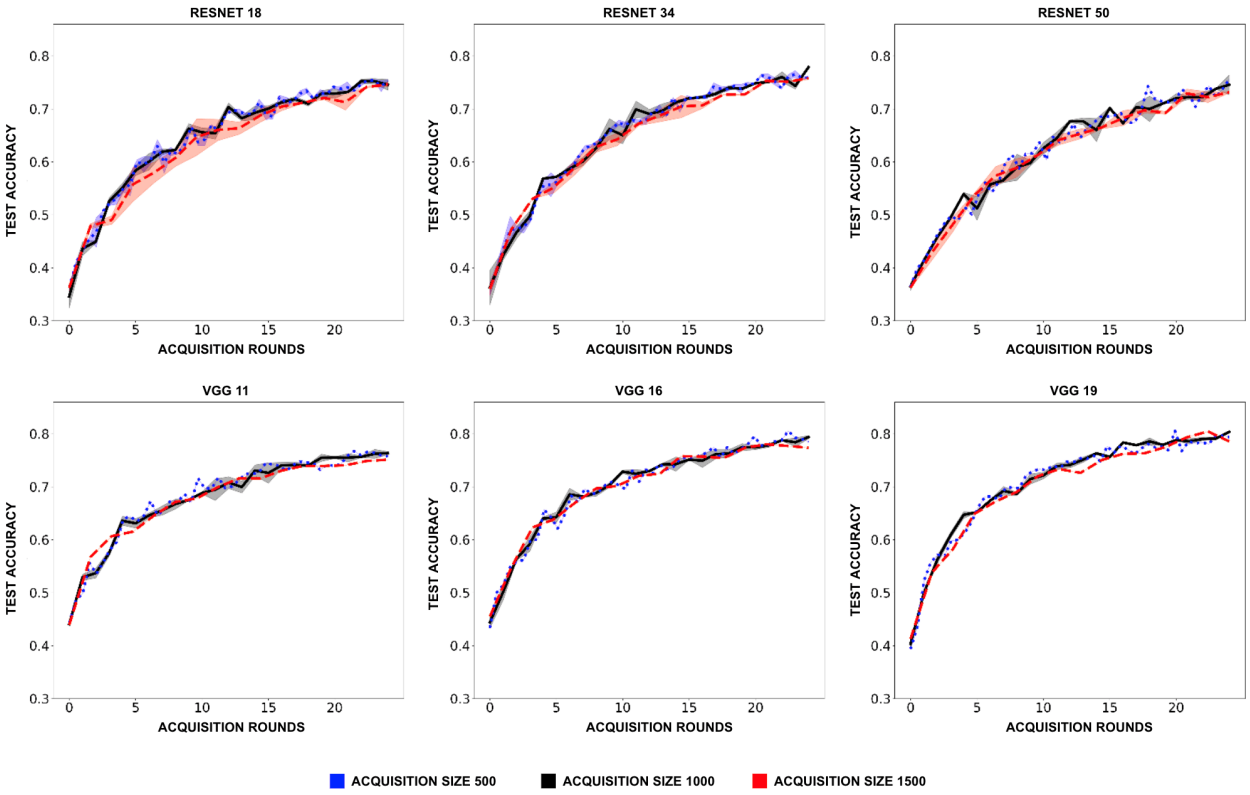

Figure 16: Visualization of performance of BALD acquisition function on different networks trained on CIFAR10 with acquisition size of 500 (dotted blue), 1000 (solid black) and 1500 (dashed red) respectively

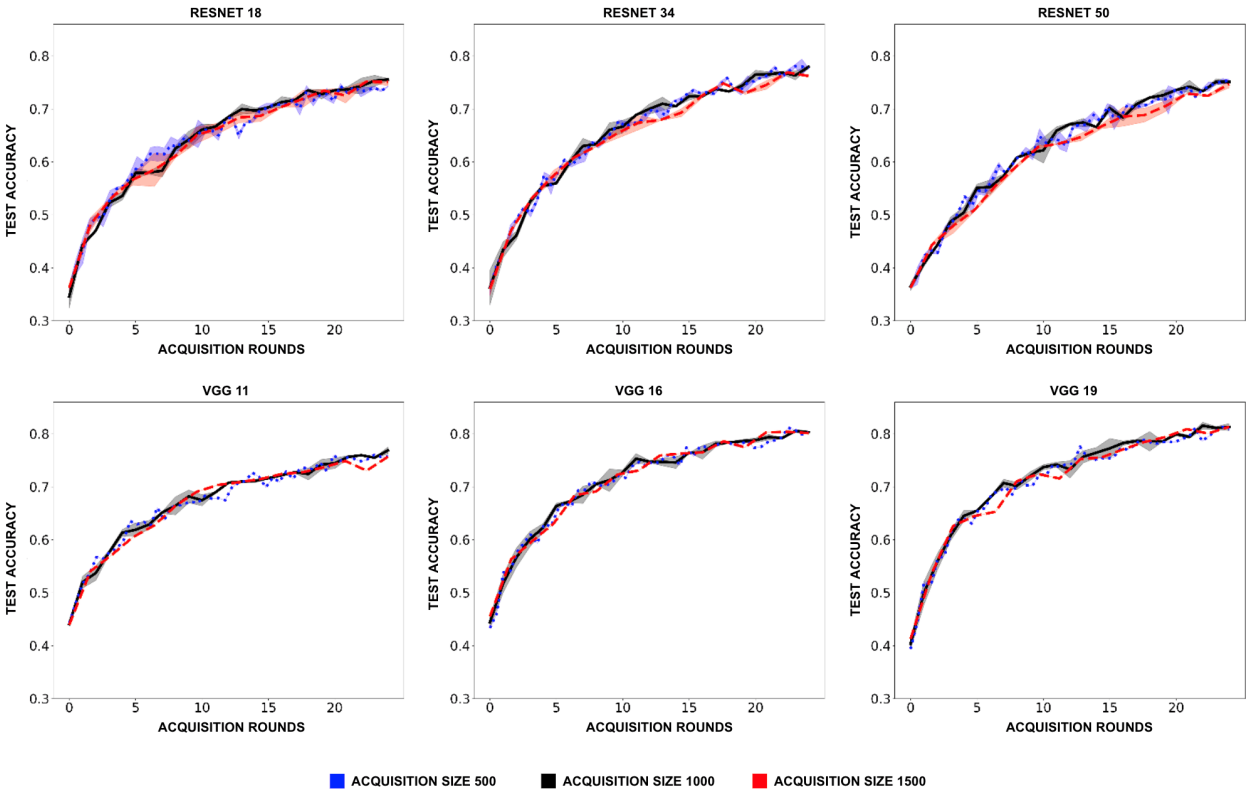

Figure 17: Visualization of performance of Coreset acquisition function on different networks trained on CIFAR10 with acquisition size of 500 (dotted blue), 1000 (solid black) and 1500 (dashed red) respectively

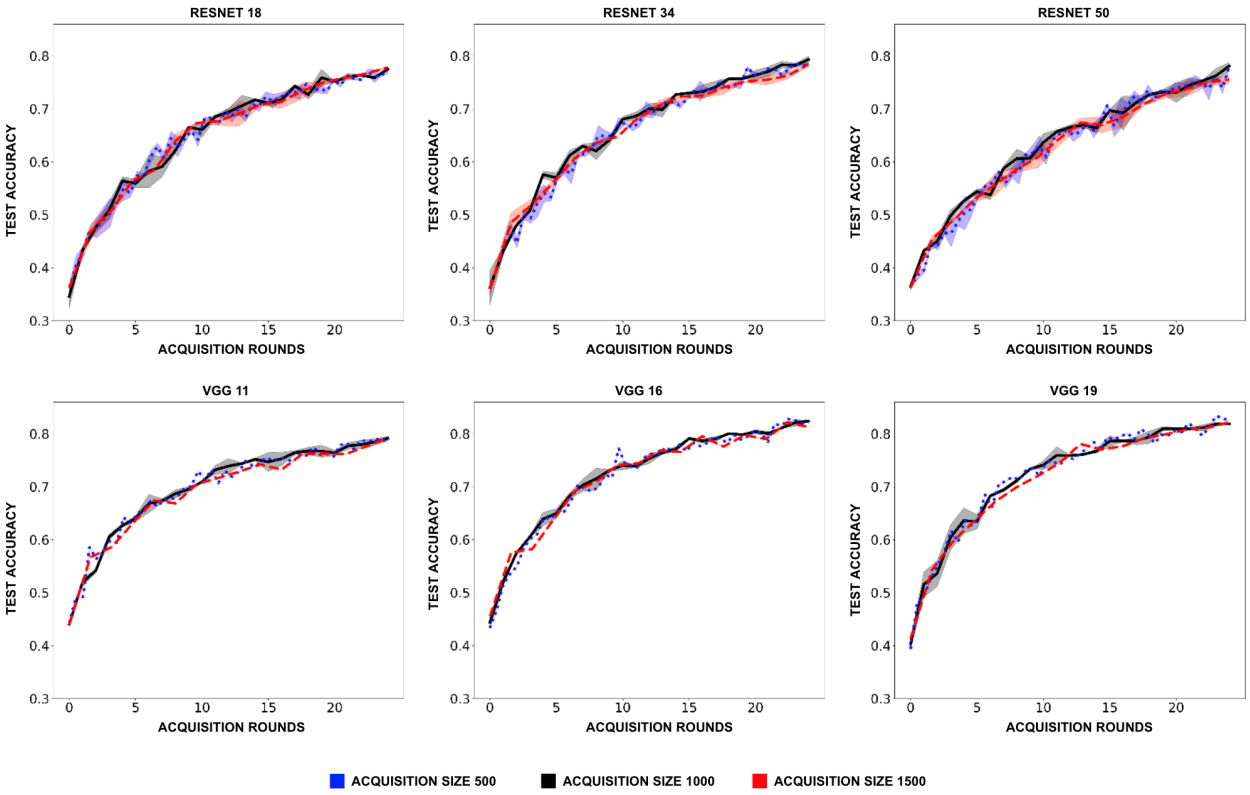

Figure 18: Visualization of performance of Margin acquisition function on different networks trained on CIFAR10 with acquisition size of 500 (dotted blue), 1000 (solid black) and 1500 (dashed red) respectively

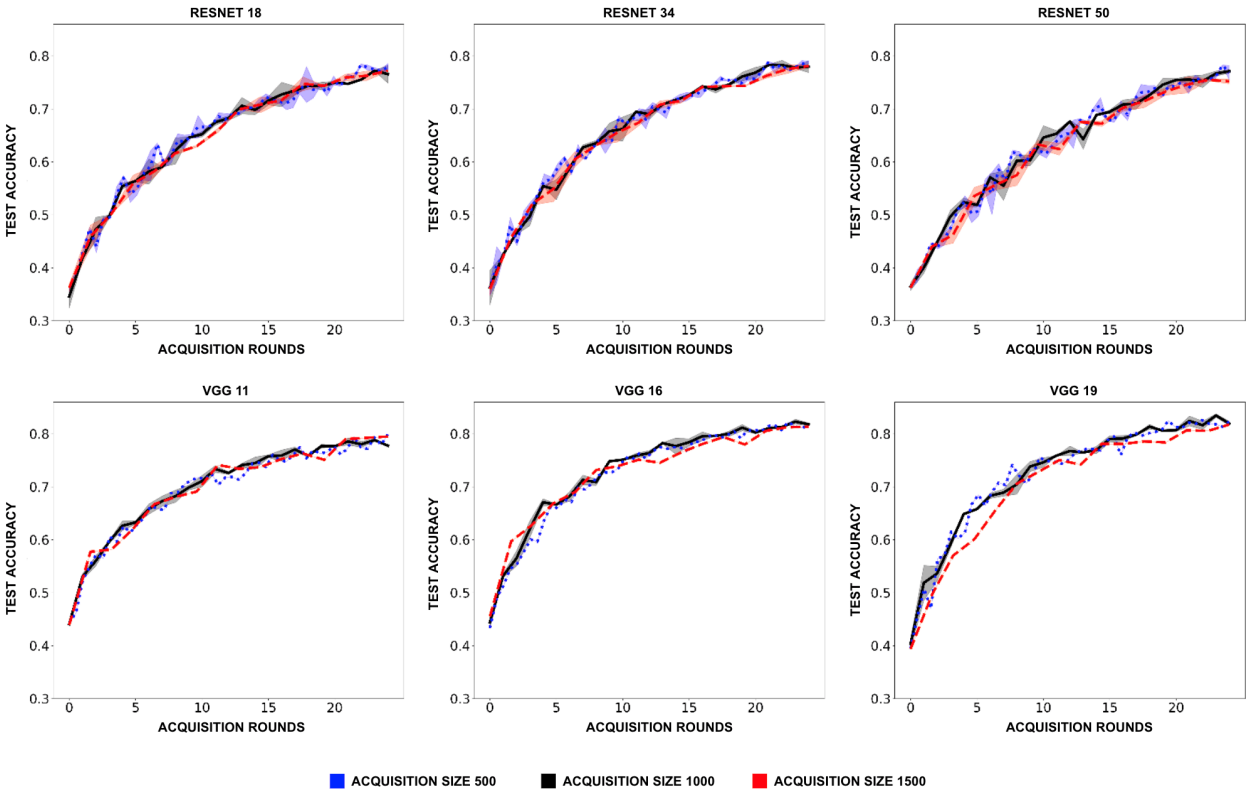

Figure 19: Visualization of performance of Least Confidence acquisition function on different networks trained on CIFAR10 with acquisition size of 500 (dotted blue), 1000 (solid black) and 1500 (dashed red) respectively

D Focus 2: DNN Architecture Optimization during DAL

D.1 Spread of performance across different acquisition function

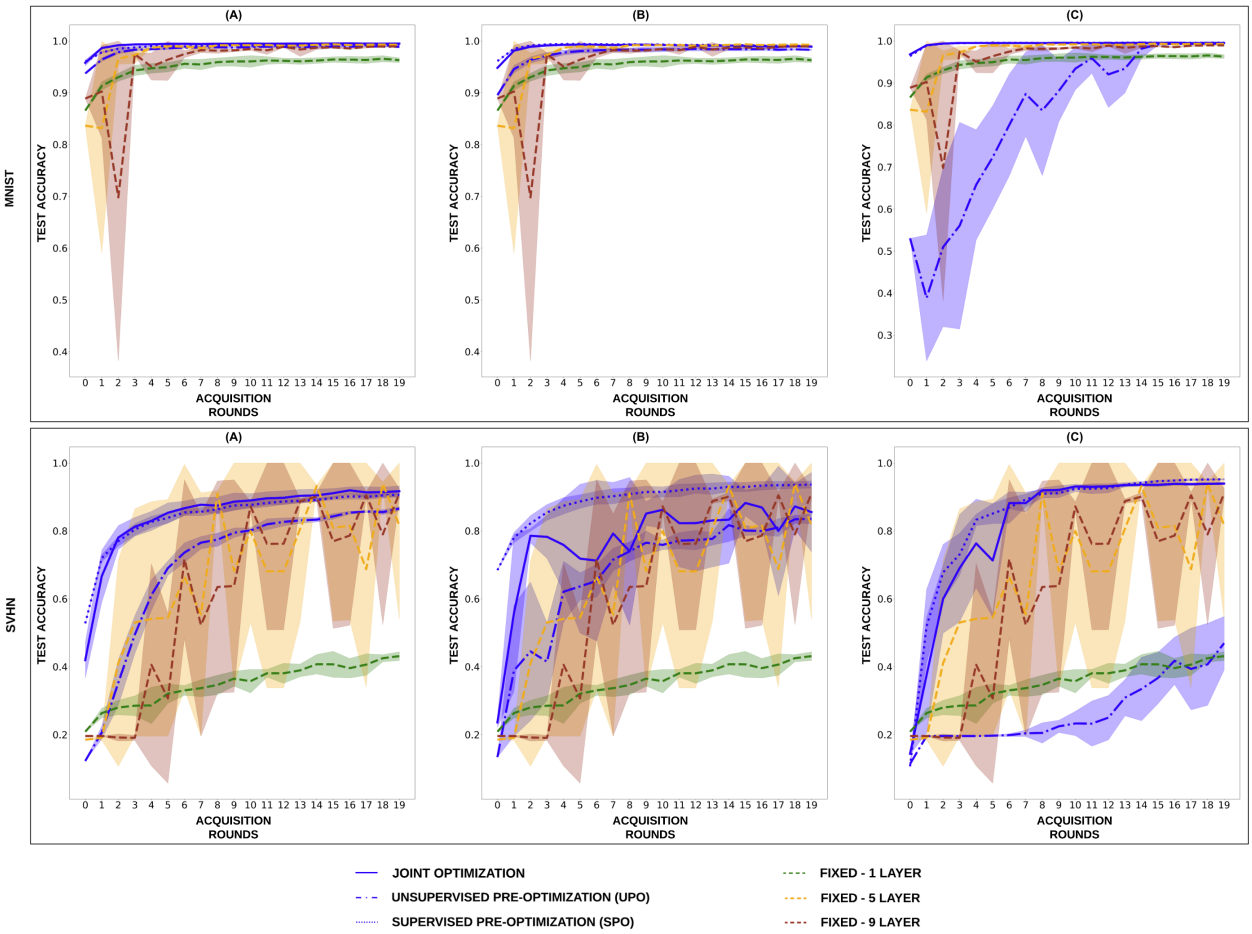

Figure 20: Spread of performance across different acquisition functions given pre-defined (fixed) CNN architecture, or jointly-optimized, supervised pre-optimized (SPO), unsupervised pre-optimized (UPO) using Depth-Dropout (A), BBDropout(B) and PDARTs(C) for MNIST (top row) and SVHN (bottom row) . Optimization of the DNN architecture, either jointly or during pre-training, in general improved over pre-defined CNN performance.

D.2 Spread of performance across different network architectures

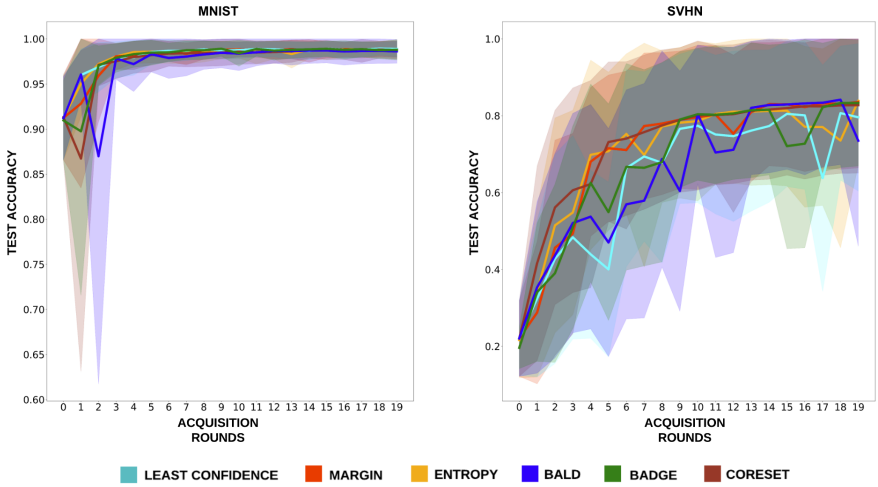

Figure 21: Spread of performance across different CNN architectures given acquisition functions for MNIST and SVHN, in the form of spread of test accuracy over the course of DAL.