Deep Active Ensemble Sampling For Image Classification

Salman Mohamadi, Gianfranco Doretto, Donald A. Adjeroh

West Virginia University, Morgantown, WV, USA

Abstract. Conventional active learning (AL) frameworks aim to reduce the cost of data annotation by actively requesting the labeling for the most informative data points. However, introducing AL to data hungry deep learning algorithms has been a challenge. Some proposed approaches include uncertainty-based techniques, geometric methods, implicit combination of uncertainty-based and geometric approaches, and more recently, frameworks based on semi/self supervised techniques. In this paper, we address two specific problems in this area. The first is the need for efficient exploitation/exploration trade-off in sample selection in AL. For this, we present an innovative integration of recent progress in both uncertainty-based and geometric frameworks to enable an efficient exploration/exploitation trade-off in sample selection strategy. To this end, we build on a computationally efficient approximate of Thompson sampling with key changes as a posterior estimator for uncertainty representation. Our framework provides two advantages: (1) accurate posterior estimation, and (2) tune-able trade-off between computational overhead and higher accuracy. The second problem is the need for improved training protocols in deep AL. For this, we use ideas from semi/self supervised learning to propose a general approach that is independent of the specific AL technique being used. Taken these together, our framework shows a significant improvement over the state-of-the-art, with results that are comparable to the performance of supervised-learning under the same setting. We show empirical results of our framework, and comparative performance with the state-of-the-art on four datasets, namely, MNIST, CIFAR10, CIFAR100 and ImageNet to establish a new baseline in two different settings.

1 Introduction

Active learning (AL) has consistently played a central role in domains where labeling cost is of great concern. The core idea of AL frameworks revolves around learning from small amounts of annotated data and sequentially choosing the most informative data sample or batch of data samples to label. To this end, after initial training using available labeled data, an acquisition function is utilized to leverage the model's uncertainty in order to explore the pool of unlabeled data for most informative data points. In parallel with advancements in AL, in the recent years, deep learning has gained tremendous attention due to its emergence as a high-performing approach, primarily conditioned on the availability

of large amounts of training data. An interesting challenge is how to efficiently incorporate data-hungry deep learning tools into supposedly data-efficient AL frameworks.

Adjusting AL algorithms for deep neural networks has been very challenging, where extending the model complexity/capacity to that of CNNs ultimately ended up with either a poor performance, or some minor improvements at the cost of querying almost all samples. On the other hand, sequential training of such expressive models as well as extending the framework to high dimensional data injects even more complexity [1–3]. This challenge was relatively under-explored, until a breakthrough work by Gal et al [4], which essentially considered the problem of incorporating deep learning into AL for high dimensional data as highly connected with that of uncertainty representation. They thus approached the problem from the perspective of uncertainty representation in deep learning for AL, and developed a Bayesian AL framework for image data. Later work (such as [5]), however, argued that the approach exhibits poor scalability to big datasets due to its limited model capacity .

Another approach that also relied on uncertainty representation, is ensemble-based AL [5]. Here, an ensemble of classifiers is used, where the classifiers independently learn from the data in parallel. The major drawback is the poor diversity (lack of exploration) even with larger ensembles. Our approach, while enjoying the power of ensembles, solves this problem by offering an inherent exploration/exploitation trade-off as classifiers maintain some dependency in the form of a shared prior. Apart from uncertainty representation, another set of emerging methods that primarily rely on geometrical data representation [6] showed improved performance in deep AL. However, similar to [7], we empirically observed that these geometric approaches typically suffer from performance degradation as the class diversity (number of classes) increases. Another recent approach is the work reported in [7] where they take advantage of adversarial training to provide improved performance over previous methods. We empirically find that their work provided a balanced performance on datasets at different scales and diversity. As we will show later, our proposed model outperforms this approach in multiple settings with significant margins, with results approaching that of supervised learning models in some cases.

In the first part of the paper, primarily motivated to efficiently integrate the advantages of uncertainty and geometrical representations, we propose an approach built upon approximate Thompson sampling. On one hand, this provides an improved representation of uncertainty over unlabeled data, and on the other hand, supports an inherent tune-able exploration/exploitation trade-off for diverse sampling [8, 9]. Unlike conventional ensemble-based methods whose performance tend to saturate quickly, under our tuneable model, adding a few more classifiers tends to improve the uncertainty and geometric representation. To mitigate the general sample diversity problem of ensemble models (see [10, 5]), we use an inclusive sample selection strategy. Our framework showed a noticeable improvement over the state-of-the-art, with performance approaching

those of supervised learning methods. Further, we explore the scope and scale of model efficiency improvements brought about by our proposed techniques.

Briefly, due to the exploration/exploitation trade-off, Thompson sampling is expected to improve both predictive uncertainty and sample diversity by computing, sampling, and updating a posterior distribution. A serious consideration, however, is that, for more expressive models such as deep convolutional neural networks (CNNs) designed for high dimensional data, Thompson sampling makes the process computationally difficult. This is primarily because computation of the posterior distribution over CNNs is complex by nature. Inferences based on Laplace approximations or Markov chain Monte Carlo approaches would be two possible alternatives. However, both approaches are still very expensive in terms of computational cost [11–13]. Lu et al [13] argue that due to the compatibility of Thompson sampling with sequential decision and updating, an approximate version of Thompson sampling could be a promising solution. Accordingly, we build an ensemble model relying on an efficient approximate of Thompson sampling, which improves the state-of-the-art. Interestingly, this model possesses both the advantage of uncertainty based deep AL approaches (exploiting most uncertain samples), and of geometric solutions (exploring for more diverse though not necessarily highly uncertain samples).

In the second part of the paper, we investigate a new line of efforts/arguments revolving around the idea of boosting AL frameworks using self/semi supervised learning techniques. We substantiate and unify these arguments and also design and perform extensive experiments on multiple baselines to assess this approach as a new general training protocol for AL frameworks. This enables our approach to be compared against recent boosted AL frameworks.

Briefly, our key contributions in this paper are as follows:

- A new framework for deep AL which enables an exploration/exploitation trade-off for sample selection and hence offers the advantages of both uncertainty-based and geometry-based methods.
- A new general training protocol for visual AL approaches, developed by substantiating and unifying recent arguments on boosting AL using self/semi supervised learning, and experimentally evaluating this approach on multiple recent baselines. We compare our framework against two sets of baselines to show its performance.

2 Background and preliminaries

Background: Early efforts on AL with image data considered mainly kernel-based approaches [14–16]. Later, AL methods with image data using CNN included uncertainty-based approaches [4, 17, 5, 18, 19], geometry-based approaches [6], or their combination [7], e.g, based on adversarial training. Generally speaking, uncertainty-based approaches focus on finding most uncertain samples to label, with the potential downside of less diversity in sample selection, while geometric approaches tend to weigh on diversity of samples, resulting in performance degradation in cases of very diverse datasets (with large number of

classes). Most recently, in a relatively different setting, Gao et al. [20] leveraged semi-supervised learning while Bengar et al. [21] applied self-supervised learning (SSL) techniques to deliver a significant performance improvement. We will compare our proposed approach against these related work, on the same problem settings. Some other recent work in this general area of modern AL with high dimensional data can be found in [22–27]. Though these are relevant, they are not as closely related to our approach.

SSL: As the second contribution of this work relates to SSL we briefly review the literature. Briefly, SSL is one of the closest modern problem domains to AL with zero labeling effort policy. Here, the goal is to leverage all unlabeled data to train a network for a pretext task so as to prepare the network for a downstream task, usually with small amounts of data [28]. Until recently, a major set of SSL baselines were contrastive baselines relying on contrasting augmented views of a sample with each other (positive contrastive pairs) and with views of other samples (negative contrastive pairs) [29, 30]. Newer baselines such as [31, 32], a.k.a non-contrastive approaches, rely on contrasting positive pairs, needless of contrasting negative pairs. Recently, Ermolov et al. [33] reported a non-contrastive method based on whitening the embedding space, which was effective, yet conceptually simple. We adopt this approach in this work.

Preliminary: We describe these two major paradigms below.

1. Uncertainty-based techniques: Two categories of well-known deep learning techniques for uncertainty representation and estimation include ensemble-based techniques (non-Bayesian) [18, 19] and Monte-Carlo (MC) dropout (Bayesian) [17, 4]. In ensemble-based methods, an ensemble of N identically structured neural networks are trained using identical training data D_{tr}, where the different random values are applied for weight initialization w_i. For a given class c out of multiple classes and input X, we then have:

$$p(y = c|x, D_{tr}) = \frac{1}{N} \sum_{i=1}^{i=N} p(y = c|x, w_i) \tag{1}$$

However, MC-dropout trains a network with dropout, and during test, implements T forward passes, each individually with a new dropout mask, resulting in T sets of weights w_t. Given input x, the average of all T softmax vectors represents the output for a desired class c.

$$p(y = c|x, D_{tr}) = \frac{1}{T} \sum_{t=1}^{t=T} p(y = c|x, w_t) \tag{2}$$

Here we briefly describe some popular effective uncertainty-based acquisition functions [4, 5] or their approximation for ensemble-based approaches, MC dropout, and our proposed framework, all based on uncertainty sampling.

A. Selecting samples with highest predictive entropy [34].

$$H[y|x, D_{tr}] := -\sum_{c} (\frac{1}{N} \sum_{n} p(y = c|x, w_n)). \log(\frac{1}{N} \sum_{n} p(y = c|x, w_n)) \tag{3}$$

B. Selecting samples with highest mutual information between their predicted labels and the weights, BALD [35, 4], which was initially applied in [4] with T forward passes in MC-dropout. It can be analogously rewritten for an ensemble with N members by replacing T with N.

$$I[y; w|x, D_{tr}] := H[y|x, D_{tr}] - \frac{1}{T}\sum_{t}\sum_{c} -p(y = c|x, w_t).\log p(y = c|x.w_t) \quad (4)$$

C. Highest Variation Ratio [36] as a measurement of non-modal predicted class labels, where f_m is the number of modal class predictions [5].

$$VR := 1 - f_m/N \quad (5)$$

We used this acquisition function in our proposed DAES framework.
2. Geometry-based techniques: Geometric or representation-based methods primarily rely on density-based acquisition functions. Typical examples include REPR [37], and Core-Set [38]. With a total of n samples, at each iteration Core-Set selects a fixed number of samples, that minimize the upper bound on the distance between point x_i in n samples, and x_j, its closest neighbour in selected subset o. The acquisition function of Core-Set is given as follows: $s = argmax_{i\in[n]\emptyset} \min_{j\in o} dist(x_i, x_j)$. See [37] for that of REPR.
3. Other techniques:
Other methods include implicit combination of uncertainty and geometry approaches, such as in [7], which designs a minimax game in the context of adversarial training. There are also methods that have used the power of pre-trained models such as [20], and to a less extent [39].

3 Deep active ensemble sampling

Our work is primarily inspired by the reports in [4, 19, 13] towards finding an uncertainty-diversity trade-off. In particular, we propose a tuneable trade-off between uncertainty-wise exploitation of samples vs exploration of less uncertain, but more diverse samples.

3.1 Thompson sampling for AL

Contextual bandit: Thompson sampling was primarily developed as a heuristic to address the Multi-armed bandit (MAB) problem, aiming for a trade-off between exploration and exploitation in sequential decision making. The core idea of Thompson sampling has a Bayesian essence (See Algorithm 1). Unlike greedy algorithms that mostly lean toward exploitation, Thompson sampling draws random samples from a posterior distribution to fine-tune between exploration and exploitation. See [9, 8] for related work on low dimensional data. New attempts towards using Thompson sampling for efficient estimation of posterior distribution for more complex models such as CNNs revealed an immediate need to find a computationally tractable approximation [13].

Deep AL: Assuming a pool-based AL setting, we initially have a set of unannotated data $U_0 = \{x_1, x_2, ...x_n\}$ and a small set of annotated data A_0, where at each iteration, an algorithm, known as *acquisition function*, looks into the whole set of unlabeled data to select a number of samples and pass them to an Oracle for labeling. In deep learning backed AL with high dimensional data such as images, the goal is to adjust the model to enable learning from a relatively small initial training set, and accordingly select a subset of most informative unlabeled data samples (in terms of uncertainty and diversity) to be labeled.

3.2 Ensemble sampling

From a geometry perspective, one ideal estimation of the desired posterior space in AL framework could be represented by a direct sum over the space. Along this line, some methods such as [24] propose splitting the input space to improve uncertainty sampling associated with the posterior distribution. Hinton et al [40] noted the fact that data points are generated by natural sources that actually inject limited complexity, rather than random sources with unlimited complexity. Therefore, unlike a random source that practically enables sampling from an infinite space, the natural source can be represented with a direct sum over the posterior space $\mathbf{S}$ with any ***finite*** number of summands Q: $\mathbf{S} = \mathbf{S_1} \oplus \mathbf{S_2} \oplus ... \oplus \mathbf{S_Q}$, where $\mathbf{S_i}$ represents the i-th subspace. Later we will see that compared with regular ensembles, ensemble sampling is closer to this direct sum as it allows a better exploration of whole representation space.

In the case of AL on a neural network with weights θ, let's say the network represents the mapping $g_\theta : \mathcal{R}^W \mapsto \mathcal{R}^K$ (W is the dimensionality of input) and the goal is to sequentially choose a fixed number of samples d_t from a pool $\mathcal{D}$ of K samples as input at each time $t = 0, 1, ...T$, where $\mathcal{D} \subseteq \mathcal{R}^W$, such that it leads to desirable output. Accordingly, with each set of samples d_t selected from $\mathcal{D}$ at time $t = 0, 1, ..., T$, an output $g_\theta(d_t)$ and random variable $w_t \sim N(0, \sigma_w^2 I)$ form the observation $y_t = g_\theta(d_t) + w_t$ which allows to update a reward $r_t = r(w_t)$ sequentially. Supposing that we have a prior on θ, $\theta \sim N(\mu_0, \Sigma_0)$, the model will become much more prone to uncertainty. Therefore, at each time t, the neural network will be fitted by d_t, y_t, and the samples are selected with the goal of converging to a trade off between immediate desirable outputs (minimizing the loss) and reducing uncertainty in θ.

With the problem presented in as above, an algorithm is required to incorporate Thompson sampling in this new context. In the case of linear bandit problem, since the conventional Thompson sampling yields an efficient solution, no approximation to Thompson sampling is needed. However, in case of neural networks, the conventional form of Thomson sampling could be computationally expensive. This calls for a more efficient implementation in terms of approximate Thompson sampling. Accordingly, Lu et. al [13] introduce an ensemble of N networks with a shared prior on their weights, as an approximate Thompson sampling. This allows efficient posterior estimation on complex models such as neural network.

3.3 Algorithms for CNNs

Here we represent ensemble sampling as an efficient approximation of Thompson sampling for neural networks. In fact, unlike in simpler cases such as linear bandit, exact Bayesian inference can not easily be performed effectively for neural networks, which necessitates an efficient approximation. First, we present the algorithm for Thompson sampling (Algorithm 1 (taken from [41])). Then, we discuss ensemble sampling as its efficient approximation, and present the algorithm for Deep Active Ensemble Sampling (Algorithm 2).

More precisely on Thompson sampling, let's assume $\mathcal{X}$ is a finite set of data points $x_1, ..x_n$, where selecting a data point x_t (or a number of data points) at time t yields a randomly generated output y_t based on a conditional probability distribution $q(.|x_t)$. Accordingly, a known function $r_t = r(y_t)$ is defined to capture the reward for the selected data point. This reward can be interpreted as a negative loss. At the beginning, the decision maker gets initialized with a prior p on θ, and as it starts to explore, updates its uncertainty representation. While greedy algorithms generally use expected value of θ with respect to p to produce model parameters $\hat{\theta}$, Thompson sampling relies on random sampling from p. Next, the algorithm will choose data points maximizing the expected reward presented as follows:

$$\mathbb{E}_{q_{\hat{\theta}}}[r(y_t)|x_t = x] = \sum_o q_{\hat{\theta}}(o|x)r(o) \tag{6}$$

Subsequently p is updated by conditioning on $\hat{y_t}$, and for θ coming from a finite set, relying on Bayes rule we will have:

$$\mathbb{P}_{p,q}(\theta = u|x_t, y_t) = \frac{p(u)q_u(y_t|x_t)}{\sum_v p(v)q_v(y_t|x_t)} \tag{7}$$

Algorithm 1 (taken from [41]) captures the above steps. As noted, this will be very time consuming, especially for neural networks.

Algorithm 1 Thompson$(\mathcal{X}, p, q, r)$

for t= 1, 2, ... , T **do**
  Sample $\hat{\theta} \sim p$
  $x_t \leftarrow \arg\max_{x \in \mathcal{X}} \mathbb{E}_{q_{\hat{\theta}}}[r(y_t)|x_t = x]$
  Input chosen x_t and observe y_t
  $p \leftarrow \mathbb{P}_{p,q}(\theta \in .|x_t, y_t)$
end for

As an efficient approximate Thompson sampling for neural networks, we use ensemble sampling, where we employ an ensemble of M networks and set priors on the weights, as presented in Algorithm 2. All networks will be trained on identical data samples while the initial shared priors on the weights makes a connection between them. Algorithm 2 is inspired by [13], with key adjustments

to make the approximate Thompson sampling adaptable to the AL framework. These changes include (1) the optimization process of ensembles; (2) selecting a set of samples rather than one sample; (3) we replace the original concept of maximizing reward in the algorithm with minimizing the loss, namely, $\bar{L}(\theta)$ in our deep active learning framework. Accordingly, it is important to mention that the optimization of the method need not to be combinatorial as in the case with combinatorial contextual bandits. Moreover, sample selection is sequential in which, each iteration of sample selection provides a batch of samples ranked by the acquisition function. Unlike classical ensemble-based approaches, the proposed deep active ensemble sampling (DAES) not only puts a joint prior on the weights of the networks (all sampled from one prior distribution rather than individual priors), but also jointly optimizes the members of an ensemble.

Algorithm 2 Deep Active Ensemble Sampling (M)

1: Ensemble $En_M\left(g, \mathcal{N}(\mu,\, \sigma^2)\right)$: $g(\theta_1), ..., g(\theta_M)$; Labeled Set: S_l^t; Unlabeled Set: S_u^t
2: **for** t= 1, 2, ... , T **do**
3: Train over S_l^t: En_M: $g(\theta_{1,t}), ..., g(\theta_{M,t})$
4: Optimize: $arg\,min_{\theta_{i,t}}(L^t) = arg\,min_{\theta_{i,t}}\,(L(\theta_{1,t}) + ... + L(\theta_{M,t}))$
5: Batch b^t selection by fixed En_M: $En_M(S_u^t)$, $VR = (1 - \frac{f_m}{M})$ via Eqn (5)
6: Update Training Set: $S_l^{t+1} = S_l^t + b^t$
7: **end for**

3.4 DAES with self-trained knowledge distillation

Consistent with primary focus of AL on less annotation effort and with the goal of establishing a new standard AL training protocol, we empirically evaluate a simple training technique which inherently empowers any active learner, regardless of the underlying approach. While this is inspired by the recent trend in [20, 21], we also argue that using pre-training, here SSL pre-training, enables any AL framework to better model uncertainty over the data, or to capture the geometry of the data, due to the prior knowledge attained by SSL. To ensure fairness of our comparisons, we apply the new training protocol to both the previous baseline AL models, and to our proposed DAES framework. The proposed training protocol could help to eventually unify this line of work with some form of knowledge distillation [42, 43].

Training protocol: The protocol is a two step process: SSL pre-training, and then active learning using the pre-training output. (See Fig. 1). Due to huge success of SSL in learning representation from unlabeled data, we adopt a most recent SSL model suitable for our setting. Thus, our proposal for training AL models is to consider a training protocol, first a pre-training is performed on the deep network (encoder in Fig. 1) as a building block for the active learning models. In this work, we tested this idea by adopting the conceptually simple,

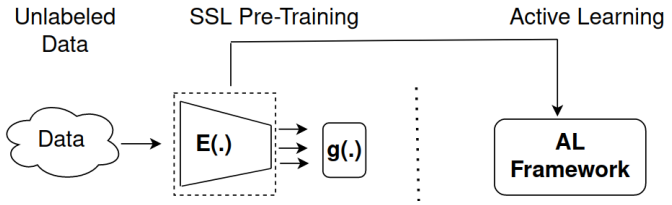

Fig. 1. SSL pre-training for deep active learning. Here, $E(.)$ is the encoder and $g(.)$ is the projection head. After pre-training, the weights of $E(.)$ will be fixed and will then be used in our AL setting, training a classifier head on top of that.

yet effective SSL model in [33] to initially train ResNet18 as the building block for the AL methods, namely, Random baseline, VAAL, Core-Set and DAES.

We explore a new setting in which a given baseline is equipped with a conceptually simple self-training as discussed above. As shown in Fig. 1, we adopt the SSL framework from [33], to leverage knowledge distilled from unlabeled data for empowering the active learner. The idea is to use whitening in SSL in order to train the encoder (ResNet18) and then freeze all layers except for head-layers which are replaced with fully connected layers to be trained.

4 Experiments and results

We conduct two sets of experiments on images classification task to evaluate our proposed DAES framework as well as compare it against state-of-the-art models. Specifically, we mainly perform the experiments on MNIST [44], CIFAR10 and CIFAR100 [45], and ImageNet [46]. To ensure the fairness of compassion scenarios, we compare the framework against **two sets** of baselines, namely, trained from scratch, and self-trained enabled by self/semi supervised learning (SSL).
Evaluation: On CIFAR10/100 and ImageNet, starting with an initial budget of 10% labeled samples, we measure the performance on sequential training using T training iterations, where in each iteration of training we add 5% labeled data from unlabeled pool to the training set (labeled data ratio of 0.1, 0.15, 0.20, ... up to 0.35 or 0.50). We assume each training iteration is from scratch unless otherwise stated. On MNIST the initial training set is 200 samples and the evaluation is performed on acquisition budget of 100 samples. The results of all our experiments on all datasets including ImageNet are averaged over three trials.
Baselines: We compare the performance of DAES against two sets of baselines. First set of approaches, specifically trained from scratch, includes Random sampling from unlabeled pool (Random), Monte-Carlo dropout (M-C Dropout) [17], deep Bayesian active learning (DBAL) [4], Core-Set [6], Ensemble with Variation Ratio (Ens-VarR) [5], and VAAL [7]. We also design and implement another set of extensive experiments on our framework as well as some of previous baselines empowered by self-training including Random, Core-Set, and VAAL to contrast against a very recent baseline taking advantage of SSL, CSSAL [20], and also later compare with a semi-supervised baseline, REVIVA [39].

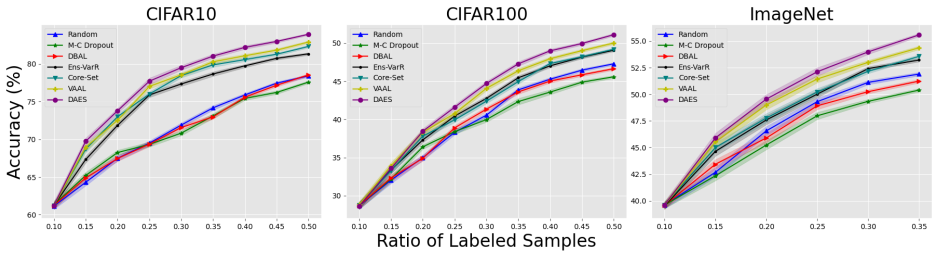

Fig. 2. Accuracy vs ratio of labeled samples from CIFAR10, CIFAR100 and ImageNet datasets.

4.1 Experimental settings

We implemented our network architectures in Pytorch. Besides our experiments, experiments of all other competitive baselines including Random baseline, on CIFAR10, CIFAR100 and ImageNet are performed with ResNet18, with similar setting of VAAL except they used VGG16 [47]. However for MNIST, we used a three-layer (two convolutional and one fully connected) network described in [4]. Specifically an ensemble includes $N = 5$ identical classifiers unless otherwise specified. We used Xavier initialization when applicable, and we utilized Adam optimizer [48] for all experiments. All experiments start with an initial balanced budget of 10% of unlabeled training pool (6000 for MNIST, 5000 for CIFAR10/100, and 128120 for ImageNet), which is then iteratively updated by adding 5% of whole training pool. Both initial training and other sequential iterations of training continue for 100 epochs. After every update, the network is trained from scratch unless otherwise specified (i.e., incremental training). Further, unlike classical ensemble-based methods, the optimization process of all classifiers in DAES is performed jointly as one loss function. Practical considerations in case of DAES with very deep networks are discussed in ablation studies.

4.2 DAES performance comparison

In this section we explain the immediate results of experiments on MNIST, CIFAR10/100 and ImageNet in two comparing scenarios, namely, AL model trained from scratch, and AL on self-trained model.

1. Trained from scratch: The conventional protocol is training from scratch.

Our results on MNIST is on par with VAAL and Core-Set where all three approaches attained $99 + \%$ accuracy with 1000 samples (1.67%) of the data. Ens-VarR, DBAL, M-C Dropout and Random baselines achieved 97.81 ± 0.12, 97.55 ± 0.18, 97.26 ± 0.14, and 95.2 ± 0.23

On CIFAR10 as shown in Fig.2, our framework tends to outperform other baselines including VAAL upon using more than 15% of the data, while the difference grows by adding more labeled samples. Our approach attains mean

accuracy of 82.98$ and 83.93% upon using 40% and 50% of the data respectively, whereas Top-1 accuracy using 100% of data is 93.27%. Second and third highly performant methods using half of the data are VAAL and Core-Set with 82.89% and 82.31% respectively. While Ens-VarR remains fairly competitive, M-C dropout as well as DBAL are evidently underperforming.

On CIFAR100 also our method starts to outperform competitive VAAL and Core-Set approaches upon using $20+\%$ of the data. The accuracy difference swiftly grows by adding more samples to the point that upon using 50% of data, our method outperforms VAAL and Core-Set by 51.33% to 50.01% and 49.03%. Note that the Top-1 accuracy using full data is 75.43%. As it is clear, due to larger number of classes, Core-Set experienced performance degradation down to performing on par with Ens-VarR .

Performance on dataset at scale: On ImageNet as a large and more challenging dataset of 1.2+ million samples of 1000 classes, our method patently outperforms former baselines upon using 15% or more of data. Compared to Top-1 mean accuracy of 71.8% using whole data, we achieve mean accuracy of 55.57% upon using only 35% of data, which is a 1.2% improvement over VAAl, (while VAAL offers only less than 1% improvement over its former baseline, Core-Set using 35% of data). Our method improves over Random baseline by mean accuracy of 3.67%. Similar to their performance on CIFAR10 and CIFAR100, Bayesian techniques, i.e., DBAL and M-C dropout, slightly underperform Random baseline.

2. Self-Training: We also evaluated the proposed use of self-supervised knowledge distillation [42] from unlabeled data as a general technique to further improve the model training process for AL methods. There are two objectives here. First, to provide a fair comparison of this SSL+AL approach when applied on our proposed DAES, and three other AL baselines (namely, VAAL, Randon and Core-Set), against two approaches [39, 20] that take advantage of knowledge distillation of unlabeled data. Second, to show that the SSL+AL protocol establishes a new standard training protocol for deep AL regardless of the underlying principle. As shown in Fig. 3, extensive experiments on Random baseline, VAAL, Core-Set and DAES on CIFAR10, CIFAR100 and ImageNet consistently confirm the performance jump due to SSL-wise leveraging of unlabeled data while still using a small percentage of labeled data. Aside from bringing some accuracy jump to VAAL, Core-Set and Random baseline, this allows our framework to outperform CSSAL [20] on CIFAR10, CIFAR100 and ImageNet by using $18+\%$, $20+\%$ and 17% of data. As can be observed, the performance of our method on all three datasets rivals Top-1 mean accuracy attained by supervised learning (having the whole data labeled, denoted by the red line in the figure). On CIFAR10 and only using 40% of data (labeled), all approaches except for Random acquisition perform above Top-1 mean accuracy of 93.27%. On CIFAR100 (50% labeled) and ImageNet (35% labeled) all methods are competitive to supervised Top-1 mean accuracy, with our method (DAES) achieving a mean accuracy of 73.55% (compared to 75.81%) and 69.92% (compared to 71.80%). Finally, compared with a recent baseline on semi-supervised learning, REVIVAL proposed

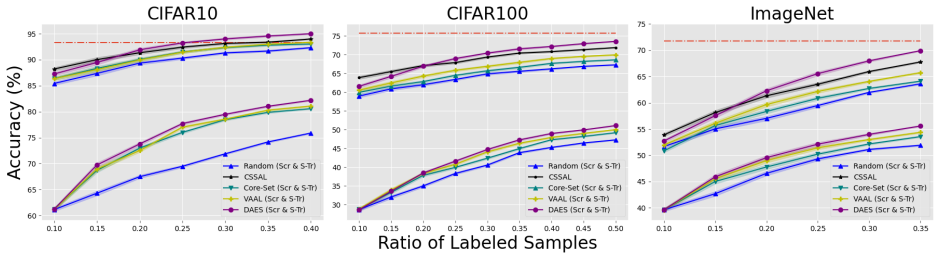

Fig. 3. Accuracy vs ratio of labeled samples from CIFAR100 and ImageNet datasets on SSL boosted networks. Top groups: results with proposed training protocol using SSL; Lower group: results with training without SSL. Red line denotes results using supervised learning with the full labeled data.

in [39], on CIFAR10 and using 40% of the data, our framework performs on par with REVIVAL. On CIFAR100 our approach (using 35% of the data) performs on par with REVIVAL (using 25% of the data).

5 Ablation study and investigative scenarios

In this section we discuss our ablation studies to assess the effect of model size on tuning the trade-off between performance and model capacity/complexity, DAES behaviour with deeper networks, and finally incremental training. For all methods, we used Variation Ratio as acquisition function, as it is empirically proved to be the most effective query strategy in the literature [4, 5].

5.1 DAES model size

One main advantage of DAES is that it can provide higher accuracy by enlarging the ensemble. Ens-VarR enjoys a performance boost only when changing the 1-member ensemble to ensemble with more than one member. Unlike Ens-VarR which lacks a malleable trade-off between computational over-head and performance, meaning that adding reasonably more classifiers to the ensemble does not lead to a proportional increase in performance, we empirically assess how larger ensembles provide desirable improvement in accuracy for DAES. As shown in Fig. 4, DAES-10 with 5 additional classifiers (total of 10), approximately doubles the former accuracy improvement on ImageNet dataset (3 times the accuracy improvement that VAAL adds to Core-Set under the same experimental setting). This is while DAES-20 with 20 classifiers brings $180+\%$ improvement over 5-member DAES. Similar experiments on CIFAR100 also confirm the proportional improvement. CIFAR10 however enjoys relatively smaller accuracy enhancement compared with the other datasets. We suspect that the underlying cause of the source of the improvement could be due to two separate reasons. First, adding more classifiers positively impacts the model's capacity on efficient sample selection. Second, training the model on full training budget,

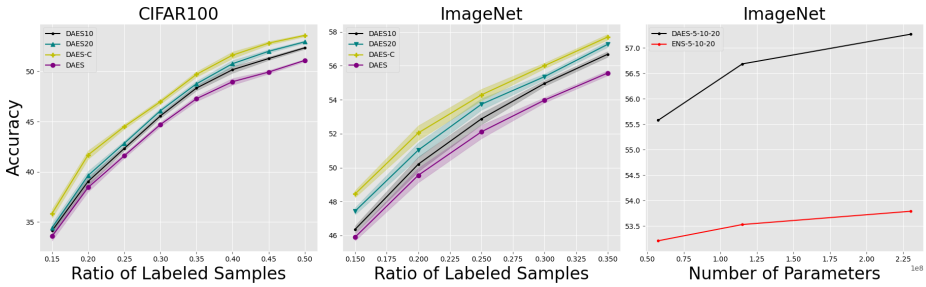

Fig. 4. Accuracy with different size of DAES and cumulative training (left and middle); and the slope of the curve representing a tunable trade-off for DAES (right).

allows classifiers to individually specialize in diverse feature representation and accordingly yields to a better generalization at test time, compared to a model with fewer classifiers. The former explanation could be intuitively conceived as the performance/behaviour spectrum of ensembles with $1, 2, ..., N$ member(s) over test time. In all experiments, cumulative training using the union of chosen samples by DAES, VAAL and Core-Set performs slightly better than DAES-20 except for CIFAR10. We see this as indicative of the superior effect of training budget size over model capacity on overall performance in this setting. Also as shown in Fig 4, compared with DAES, training VAAL with the same cumulative training budget led to lower accuracy – a clear contrast of the models' capacities.

5.2 DAES with deeper networks

We closely watch the training behavior of DAES with deeper networks such as ResNet50 and ResNet101. As an occasionally observed drawback, DAES built on very deep networks such as ResNet101, tends to take much longer convergence time than expected, or could even fail to converge. As a remedy, we found it helpful to initially pre-train the networks (or blocks) separately using initial training budget, and then train the ultimate ensemble built using pre-trained blocks. Applying this simple trick ensures the convergence of DAES.

On ImageNet and using 35% of data, DAES-5 with ResNet101 brings only approximately 1% mean accuracy improvement over DAES-5 with ResNet18 which is less than the improvement provided by DAES-10 with ResNet18. The same behavior was observed with CIFAR100. We suspect that adding more members (ensembles) to DAES leads to more improvement than replacing the blocks with deeper CNNs.

5.3 Incremental training and tunable accuracy/cost trade-off

A. Incremental training: In a standard AL experimental setting, after updating the training set, the next training iteration starts from **scratch** (here for some 100 epochs). However, we investigate incremental training of models

(VAAL, Core-Set and DAES) in which models are trained under much fewer number of epochs at each iteration while in next iteration rather than restarting the training, training continues. Specifically, we train the model for 20 epochs (formerly 100 epochs) with initial budget. Then after each data acquisition, the model first is trained on newly selected samples for as many epochs as former samples trained over, and next, the model will be trained on the updated training set for 20 epochs. This is to utilize a not fully trained model to leverage its current data representation for sample selection. Interestingly, we find that this could be a trick to speed up the active learner. Briefly, DAES, VAAL and Core-Set experience respective performance degradation of $(1.07 \pm 0.12)\%$, $(1.39 \pm 0.14)\%$ and $(1.51 \pm 0.11)\%$, respectively. In this setting Core-Set offers only 0.14% mean accuracy gain over random acquisition under previous setting. Our analysis on time complexity briefly shows that the ratio (to DAES) of average consumed time for one iteration of sample selection for DAES, VAAL, Core-Set and DAES with incremental training were 1, 0.57, 3.78, and 0.24 respectively.

B. Tunable trade-off: Consistent with the results in [5], we could not see much accuracy improvement with increasing the ensemble size in classical ensemble-based methods as shown in Fig. 4 (right figure). In other words, such classical methods do provide a tunable trade-off between accuracy and computational overhead. A satisfactory accuracy would be attained using 5 members, and increasing the number of members does not seem to proportionally improve the performance. However, active ensemble sampling showed a much robust performance in terms of exploiting more model capacity by adding more members to the ensemble. In fact, the Bayesian nature of active ensemble sampling in conjunction with its ensemble-designed structure allows achieving much higher accuracy by enlarging the ensemble at the cost of a proportional increase in computational overhead.

6 Conclusion and future work

In this paper, we introduced deep active ensemble sampling (DAES) inspired by an efficient approximation of Thompson sampling in order to combine the advantages of uncertainty-based and geometric-based approaches into one unified framework. We also examine a new training protocol formed on self-supervised knowledge distillation from unlabeled data on four baselines in order to confirm its effectiveness. Our framework is assessed on four benchmark datasets in two experimental settings to establish a new baseline. Finally we pose a few scenarios aiming for analysing DAES. We leave further theoretical and empirical analyses on DAES with asymmetric architectures for future research.

Acknowledgement This work is supported in part by grants from the US National Science Foundation (Award #1920920, #2125872).

References

1. Cohn, D., Atlas, L., Ladner, R.: Improving generalization with active learning. Machine Learning **15** (1994) 201–221

2. Balcan, M.F., Beygelzimer, A., Langford, J.: Agnostic active learning. Journal of Computer and System Sciences **75** (2009) 78–89
3. Cesa-Bianchi, N., Gentile, C., Orabona, F.: Robust bounds for classification via selective sampling. In: Proceedings of the 26th Annual International Conference on Machine Learning. (2009) 121–128
4. Gal, Y., Islam, R., Ghahramani, Z.: Deep Bayesian active learning with image data. In: International Conference on Machine Learning, PMLR (2017) 1183–1192
5. Beluch, W.H., Genewein, T., Nürnberger, A., Köhler, J.M.: The power of ensembles for active learning in image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 9368–9377
6. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations. (2018)
7. Sinha, S., Ebrahimi, S., Darrell, T.: Variational adversarial active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2019) 5972–5981
8. Bouneffouf, D., Laroche, R., Urvoy, T., Féraud, R., Allesiardo, R.: Contextual bandit for active learning: Active Thompson sampling. In: International Conference on Neural Information Processing, Springer (2014) 405–412
9. Ganti, R., Gray, A.G.: Building bridges: Viewing active learning from the multi-armed bandit lens. arXiv preprint arXiv:1309.6830 (2013)
10. Melville, P., Mooney, R.J.: Diverse ensembles for active learning. In: Proceedings of the Twenty-First International Conference on Machine Learning. (2004) 74
11. Chapelle, O., Li, L.: An empirical evaluation of Thompson sampling. Advances in Neural Information Processing Systems **24** (2011) 2249–2257
12. Brooks, S., Gelman, A., Jones, G., Meng, X.L.: Handbook of Markov Chain Monte Carlo. CRC press (2011)
13. Lu, X., Van Roy, B.: Ensemble sampling. arXiv preprint arXiv:1705.07347 (2017)
14. Zhu, X., Lafferty, J., Ghahramani, Z.: Combining active learning and semi-supervised learning using Gaussian fields and harmonic functions. In: ICML 2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining. Volume 3. (2003)
15. Li, X., Guo, Y.: Adaptive active learning for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2013) 859–866
16. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2009) 2372–2379
17. Gal, Y., Ghahramani, Z.: Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: International Conference on Machine Learning, PMLR (2016) 1050–1059
18. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. arXiv preprint arXiv:1612.01474 (2016)
19. Osband, I., Blundell, C., Pritzel, A., Van Roy, B.: Deep exploration via bootstrapped dqn. Advances in Neural Information Processing Systems **29** (2016) 4026–4034
20. Gao, M., Zhang, Z., Yu, G., Arık, S.Ö., Davis, L.S., Pfister, T.: Consistency-based semi-supervised active learning: Towards minimizing labeling cost. In: European Conference on Computer Vision, Springer (2020) 510–526

21. Bengar, J.Z., van de Weijer, J., Twardowski, B., Raducanu, B.: Reducing label effort: Self-supervised meets active learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 1631–1639
22. Yoo, D., Kweon, I.S.: Learning loss for active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2019) 93–102
23. Agarwal, S., Arora, H., Anand, S., Arora, C.: Contextual diversity for active learning. In: European Conference on Computer Vision, Springer (2020) 137–153
24. Cortes, C., DeSalvo, G., Gentile, C., Mohri, M., Zhang, N.: Adaptive region-based active learning. In: International Conference on Machine Learning, PMLR (2020) 2144–2153
25. Zhang, Y., Ling, H., Gao, J., Yin, K., Lafleche, J.F., Barriuso, A., Torralba, A., Fidler, S.: DatasetGAN: Efficient labeled data factory with minimal human effort. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 10145–10155
26. Zhang, B., Li, L., Yang, S., Wang, S., Zha, Z.J., Huang, Q.: State-relabeling adversarial active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 8756–8765
27. Ebrahimi, S., Gan, W., Chen, D., Biamby, G., Salahi, K., Laielli, M., Zhu, S., Darrell, T.: Minimax active learning. arXiv preprint arXiv:2012.10467 (2020)
28. Jing, L., Tian, Y.: Self-supervised visual feature learning with deep neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **43** (2021) 4037–4058
29. Hadsell, R., Chopra, S., LeCun, Y.: Dimensionality reduction by learning an invariant mapping. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Volume 2., IEEE (2006) 1735–1742
30. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 3733–3742
31. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al.: Bootstrap your own latent-a new approach to self-supervised learning. Advances in Neural Information Processing Systems **33** (2020) 21271–21284
32. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 15750–15758
33. Ermolov, A., Siarohin, A., Sangineto, E., Sebe, N.: Whitening for self-supervised representation learning. In: International Conference on Machine Learning, PMLR (2021) 3015–3024
34. Shannon, C.E.: A mathematical theory of communication. The Bell System Technical Journal **27** (1948) 379–423
35. Houlsby, N., Huszár, F., Ghahramani, Z., Lengyel, M.: Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745 (2011)
36. Freeman, L.C., Freeman, L.C.: Elementary Applied Statistics: For Students in Behavioral Science. New York: Wiley (1965)
37. Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z.: Suggestive annotation: A deep active learning framework for biomedical image segmentation. In: International conference on medical image computing and computer-assisted intervention, Springer (2017) 399–407
38. Sener, O., Savarese, S.: A geometric approach to active learning for convolutional neural networks. arXiv preprint arXiv:1708.00489 **7** (2017)

39. Guo, J., Shi, H., Kang, Y., Kuang, K., Tang, S., Jiang, Z., Sun, C., Wu, F., Zhuang, Y.: Semi-supervised active learning for semi-supervised models: exploit adversarial examples with graph-based virtual labels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. (2021) 2896–2905
40. Hinton, G.E., Sejnowski, T.J., et al.: Unsupervised Learning: Foundations of Neural Computation. MIT press (1999)
41. Russo, D.J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z.: A tutorial on thompson sampling. Foundations and Trends® in Machine Learning **11** (2018) 1–96
42. Xu, G., Liu, Z., Li, X., Loy, C.C.: Knowledge distillation meets self-supervision. In: European Conference on Computer Vision, Springer (2020) 588–604
43. Bhat, P., Arani, E., Zonooz, B.: Distill on the go: Online knowledge distillation in self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2021) 2678–2687
44. LeCun, Y.: The MNIST database of handwritten digits. http://yann. lecun. com/exdb/mnist/ (1998)
45. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. (2009)
46. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee (2009) 248–255
47. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
48. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)