Effective Version Space Reduction for Convolutional Neural Networks

Jiayu Liu[1]✉, Ioannis Chiotellis[1], Rudolph Triebel[1,2], and Daniel Cremers[1]

[1] Department of Informatics, Technical University of Munich, Munich, Germany
{liuji,chiotell,triebel,cremers}@in.tum.de
[2] German Aerospace Center (DLR), Wessling, Germany

Abstract. In active learning, sampling bias could pose a serious inconsistency problem and hinder the algorithm from finding the optimal hypothesis. However, many methods for neural networks are hypothesis space agnostic and do not address this problem. We examine active learning with convolutional neural networks through the principled lens of version space reduction. We identify the connection between two approaches – prior mass reduction and diameter reduction – and propose a new diameter-based querying method – the minimum Gibbs-vote disagreement. By estimating version space diameter and bias, we illustrate how version space of neural networks evolves and examine the realizability assumption. With experiments on MNIST, Fashion-MNIST, SVHN and STL-10 datasets, we demonstrate that diameter reduction methods reduce the version space more effectively and perform better than prior mass reduction and other baselines, and that the Gibbs vote disagreement is on par with the best query method.

Keywords: active learning · deep learning · version space · diameter reduction

1 Introduction

Active learning is a supervised learning framework in which the learner is given access to a pool or stream of unlabeled samples and is allowed to selectively query labels from an oracle (e.g., a human annotator). In each query round, the learner queries the labels of some unlabeled samples and trains on the augmented labeled set to obtain new classifiers. The goal is to learn a good classifier or *hypothesis* using as few labels as possible. This setting is relevant in many real-world problems, where labeled data are scarce or expensive to obtain, but unlabeled data are cheap and abundant.

Many active learning methods for neural networks rely on measures of the "informativeness" of a query, in the form of classifier uncertainty, margin [17, 10] or information gain [16, 12, 19]. Other methods capture the informativeness by representativeness of the query set using geometry-based [23] or discriminative [13] methods. However, most of these methods ignore the notion of the hypothesis space and do not address the problem of sampling bias [8], which plague many
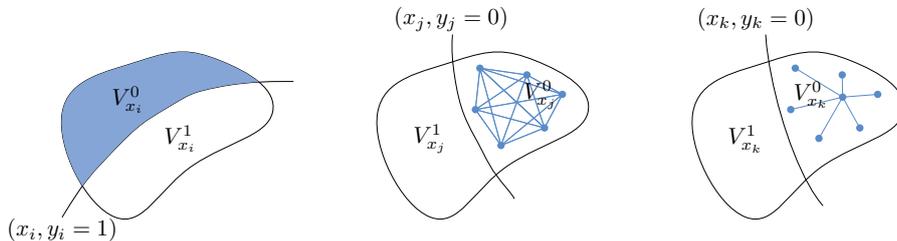
Fig. 1: Version space reduction for binary classification. Upon observing the label of x, the current version space V is split into subspaces V_x^0 and V_x^1, one of which will be removed and the other remains. **Left**: Prior mass reduction methods remove approximately half of the mass. **Middle**: Diameter reduction methods, like pairwise disagreement, query a sample that lead to sub-spaces of small diameter. **Right**: Proposed method, the *Gibbs-vote disagreement*, measures diameter by the expected distance between random hypotheses and their majority vote.

active learning methods. Without carefully handling this problem, an active learning algorithm is not guaranteed to be *consistent*, i.e., capable of finding the optimal classifier in the hypothesis space.

We consider the hypothesis space of convolutional neural networks (ConvNets) and study version space reduction methods. Version space reduction works by removing hypotheses that are inconsistent with the observed labels from a predefined hypothesis space and maintaining the consistent sub-space, the *version space*. A key condition called the *realizability assumption* is that the hypothesis space contains the classifier that provides the ground truth – if not, there are no guarantees that the best hypothesis will not be removed, because a hypothesis might make mistakes on the queried samples but perform well on the data distribution.

For neural networks, the realizability assumption may not hold for all cases. For instance, no neural networks can achieve arbitrarily small test error on some classification datasets. A workaround is to consider the effective labelings on a set of i.i.d. pool samples. To avoid the problem of an unreasonably large effective hypothesis space, as implied by the result of [25], we only consider the labelings achievable by training on unaltered samples and correct labels. We examine experimentally whether the realizability holds with this restriction and analyze its implications on version space reduction methods.

Prior mass reduction [6, 14, 4] and diameter reduction [7, 24] are two widely used version space reduction approaches. See Fig. 1 for illustration. However, prior mass reduction is not an appropriate objective for active learning [24] since any intermediate version spaces containing more than one hypothesis may still have a large diameter, i.e., large error rate in the worst-case scenario, despite

having substantially reduced mass. We derive connections between prior mass and diameter reduction and introduce a new interpretation of diameter reduction as prior mass "reducibility reduction".

We propose a new diameter measure called the *Gibbs-vote disagreement*, which equals the expected distance between the random hypotheses and their majority vote classifier. We show its relation to a common diameter measure, the *pairwise disagreement*, and discuss under which situations the former may be advantageous. We show experimentally on four image classification datasets that diameter reduction methods perform better than all baselines and that prior mass reduction [6, 14, 4] and other baselines like [16, 12, 23, 10] do not perform consistently better than random query and sometimes fail completely.

2   Related Work

A lot of research has been conducted to study the label complexity for active learning and optimality guarantees for greedy version space reduction. Hanneke [15] and Balcan et al. [1] prove upper-bounds on the label complexity in the realizable and non-realizable cases, using a parameter called the disagreement coefficient. Tosh and Dasgupta [24] propose a diameter-based active learning algorithm and characterize its sample complexity using a parameter called the splitting index. Dasgupta [6] shows that a greedy strategy maximizing the worst-case prior mass reduction is approximately as good as the optimal strategy. Golovin and Krause [14] show that the prior mass reduction utility function is adaptive submodular and a greedy algorithm is guaranteed to obtain near-optimal solutions in the average-case scenario. Cuong et al. [5] prove a worst-case optimality guarantee for pointwise submodular functions.

A variety of methods relying on the informativeness of a query have been proposed for neural networks. Gal et al. [12] use the Monte Carlo dropout to approximate the mutual information between predictions and model posterior [16] in a Bayesian setting. Kirsch et al. [19] extend [16, 12] to a batch query method. Ducoffe and Precioso [10] use adversarial attacks to generate samples close to the decision boundaries. Sener and Savarese [23] adopt a core-set approach to select representative samples for query. Gissin and Shalev-Shwartz [13] use a discriminative method to select samples such that the labeled and the unlabeled set are indistinguishable. Pinsler et al. [22] formulate batch query as a sparse approximation to the expected complete data posterior of model parameters in a Bayesian setting. Beluch et al. [3] show that ensemble methods consistently outperform geometry-based [23] and the Monte Carlo dropout method [11, 12].

3   Preliminaries

Let $\mathcal{X}$ be the input feature space and $\mathcal{Y}$ the label space. Let $\mathcal{H}$ be a hypothesis space of functions $h : \mathcal{X} \to \mathcal{Y}$ and assume a prior π over $\mathcal{H}$. A hypothesis randomly drawn from the prior is called a *Gibbs classifier*. Denote $S = \{(x_i, y_i)\}_{i=1}^n$

a pool of i.i.d. samples from the data distribution P_{XY} and $Q \subseteq S$ the set of queried labeled samples. Define the version space V corresponding to Q as

$$V := \{h \in \mathcal{H} : h(x) = y, \ \forall (x, y) \in Q\} . \tag{1}$$

Denote the subset of V that is consistent with x being labeled as y as

$$V_x^y := \{h \in \mathcal{H} : h(x) = y, \ h \in V\} , \tag{2}$$

and the pseudo-metric induced by the marginal distribution P_X as

$$d(h, h') := \Pr_x(h(x) \neq h'(x)) . \tag{3}$$

The disagreement and agreement region are defined as

$$\mathrm{DIS}(V) := \{x \in \mathcal{X} : \exists h, h' \in V, h(x) \neq h'(x)\} , \tag{4}$$

$$\mathrm{AGR}(V) := \mathcal{X} \setminus \mathrm{DIS}(V). \tag{5}$$

4   Prior Mass Reduction

4.1   Gibbs Error

The Gibbs error [4] of an unlabeled sample x is the average-case relative prior mass reduction:

$$\mathrm{GE}(x|V) := \mathbb{E}_y\big[1 - \Pr_{h \sim \pi|_V}(h(x) = y)\big] = \mathbb{E}_y[1 - \pi|_V(V_x^y)] , \tag{6}$$

where $\pi|_V(h) = \pi(h)/\pi(V)$ is the conditional distribution of $\mathcal{H}$ restricted to V. Gibbs error measures the proportion of inconsistent hypotheses taking expectation over all possible labelings of x, achievable by hypotheses in the version space. A greedy strategy that considers maximizing the average-case absolute prior mass reduction in each query can equivalently select the unlabeled sample that maximizes the Gibbs error

$$\arg\max_x \mathrm{GE}(x|V). \tag{7}$$

Define the prior mass reduction utility function as

$$f(Q) := 1 - \Pr\left(\{h \in \mathcal{H} : h(x) = y, \ \forall(x, y) \in Q\}\right) = 1 - \pi(V). \tag{8}$$

The optimization problem in (7) can be written, up to a scaling factor, as

$$\arg\max_x \pi(V)\mathrm{GE}(x|V) = \arg\max_x \mathbb{E}_y[f(Q \cup \{(x, y)\}) - f(Q)] \tag{9}$$

$$= \arg\max_x \Delta_{\mathrm{avg}}(x|Q), \tag{10}$$

where the notation $\Delta_{\mathrm{avg}}(x|Q)$ denotes the expected marginal gain of x in terms of prior mass reduction given the labeled samples in Q.

A closely related objective for active learning is the label entropy given x. It can be shown that the Gibbs error lower bounds the entropy. However, a greedy strategy that maximizes the entropy is not guaranteed to be near-optimal in the adaptive case [5]. Furthermore, empirically this criterion performs similarly or worse than the maximum Gibbs error. For the sake of simplicity, we do not consider this method in this paper.

4.2 Variation Ratio

The variation ratio of an unlabeled sample x is the worst-case relative prior mass reduction upon the reveal of its label:

$$\mathrm{VR}(x|V) := \min_y \left[1 - \mathrm{Pr}_{h \sim \pi|_V}(h(x) = y)\right] = \min_y \left[1 - \pi|_V(V_x^y)\right]. \qquad (11)$$

It measures the proportion of inconsistent hypotheses considering the worst-case labeling of x and is a lower bound on the Gibbs error. A greedy strategy that considers maximizing the worst-case absolute prior mass reduction in each query selects the unlabeled sample that maximizes the variation ratio

$$\arg\max_x \mathrm{VR}(x|V), \qquad (12)$$

which can be expressed in terms of the prior mass reduction utility function, up to a scaling factor, as

$$\arg\max_x \pi(V)\mathrm{VR}(x|V) = \arg\max_x \min_y \left[f(Q \cup \{(x,y)\}) - f(Q)\right] \qquad (13)$$

$$= \arg\max_x \Delta_{\mathrm{wc}}(x|Q), \qquad (14)$$

where the notation $\Delta_{\mathrm{wc}}(x|Q)$ denotes the worst-case marginal gain of x in terms of prior mass reduction given the labeled samples in Q.

5 Diameter Reduction

5.1 Worst-Case Pairwise Disagreement

The size of the version space can be measured by the expected pairwise disagreement between hypotheses drawn from the conditional distribution:

$$\mathrm{PWD}(V) := \mathbb{E}_{h,h' \sim \pi|_V}\left[d(h,h')\right]. \qquad (15)$$

It is the *average diameter* of the version space. A greedy strategy selects the unlabeled sample that minimizes the worst-case pairwise disagreement

$$\arg\min_x \max_y \mathrm{PWD}(V_x^y) = \arg\min_x \max_y \mathbb{E}_{h,h' \sim \pi|_{V_x^y}}\left[d(h,h')\right]. \qquad (16)$$

Other measures of diameter based on the supremum distance [18,7] are not amenable to implementation because evaluation of such diameters involves optimization. The pairwise disagreement can be estimated from a finite set of sample hypotheses from the version space.

5.2   Worst-Case Gibbs-Vote Disagreement

We propose a new diameter measure called the *Gibbs-vote disagreement*. It is the expected disagreement between random hypotheses and their majority vote:

$$\text{GVD}(V) \coloneqq \mathbb{E}_{h \sim \pi|_V}[d(h, h_{\text{vote}}|_V)], \tag{17}$$

where $h_{\text{vote}}|_V$ is the majority vote classifier of hypotheses from V. For each x, it induces a prediction

$$h_{\text{vote}}|_V(x) = \arg\max_y \mathbb{E}_{h \sim \pi|_V}[p(y|x; h)], \tag{18}$$

where $p(y|x; h)$ is the predicted probability of x belonging to class y given by a hypothesis h. The majority vote classifier is the deterministic classifier that has the smallest expected distance to the Gibbs classifier [18, 9]:

$$\mathbb{E}_{h'}[d(h', h_{\text{vote}})] = \min_h \mathbb{E}_{h'}[d(h', h)]. \tag{19}$$

Hence the Gibbs-vote disagreement measures the size of the version space by the expected distance of the random hypotheses to their "center". Further, the following relation holds

$$\frac{1}{2} \text{PWD}(V) \leq \text{GVD}(V) \leq \text{PWD}(V) \tag{20}$$

We defer the proof to the appendix[3]. Essentially, Equation (20) reveals that the Gibbs-vote disagreement is sandwiched between the average radius and diameter.

A greedy strategy selects the unlabeled sample that minimizes the worst-case Gibbs-vote disagreement

$$\arg\min_x \max_y \text{GVD}(V_x^y) = \arg\min_x \max_y \mathbb{E}_{h \sim \pi|_{V_x^y}}[d(h, h_{\text{vote}}|_{V_x^y})], \tag{21}$$

where $h_{\text{vote}}|_{V_x^y}$ is the majority vote of hypotheses from the subspace V_x^y.

5.3   Diameter Reduction as Reducibility Reduction

Pairwise disagreement shares a simple relation with Gibbs error – it is the expected Gibbs error:

$$\text{PWD}(V) = \mathbb{E}_{h,h' \sim \pi|_V}[\mathbb{E}_x[\mathbb{1}(h(x) \neq h'(x))]] \tag{22}$$
$$= \mathbb{E}_x[\mathbb{E}_{h \sim \pi|_V}[\Pr_{h' \sim \pi|_V}(h(x) \neq h'(x))]] \tag{23}$$
$$= \mathbb{E}_x[\text{GE}(x|V)]. \tag{24}$$

A similar relation holds between Gibbs-vote disagreement and the variation ratio:

$$\text{GVD}(V) = \mathbb{E}_{h \sim \pi|_V}[\mathbb{E}_x[\mathbb{1}(h(x) \neq h_{\text{vote}}|_V(x))]] \tag{25}$$
$$= \mathbb{E}_x[\mathbb{E}_{h \sim \pi|_V}[\mathbb{1}(h(x) \neq h_{\text{vote}}|_V(x))]] \tag{26}$$
$$= \mathbb{E}_x[\text{VR}(x|V)], \tag{27}$$

---

[3] The paper with appendix is available at https://arxiv.org/abs/2006.12456.

where the last equality holds because the predictions of the majority vote classifier are always the worst-case labels for prior mass reduction. Diameter reduction selects samples such that, upon revealing their labels, the induced subspaces have minimum possibility to be further reduced by a potential random query. Thus, it can be thought of as reducing the expected prior mass "reducibility".

Prior mass reduction finds splits in directions that evenly partition the version space, but could result in version spaces that have irregular shapes, in the sense that the space can be whittled down finely in some directions while being under-split in others. The worst-case error rate of the resulted version space could still be large. Diameter reduction correctly resolve this issue. Fig. 1 illustrates the differences between prior mass and diameter reduction.

5.4 Weighted Diameter Reduction

Tosh and Dasgupta [24] show that in general average diameter cannot be decreased at steady rate and propose to query the unlabeled samples that minimize the diameter weighted by the *squared* prior mass in the worst-case scenario

$$\arg\min_x \max_y \mathbb{E}_{h,h'\sim\pi}\big[\mathbb{1}(h,h' \in V_x^y)\, d(h,h')\big] \tag{28}$$

$$= \arg\min_x \max_y \pi(V_x^y)^2\, \mathbb{E}_{h,h'\sim\pi|_{V_x^y}}\big[d(h,h')\big]. \tag{29}$$

The potential to be minimized is a surrogate for the "amount" of edges between hypotheses and is closely related to the splittablity of version space [24, 7].

6 Realizability Assumption

Even though neural networks are capable of fitting an arbitrary pool set, we show experimentally that the version space obtained by training on a subset of the pool set with stochastic gradient descent – the "samplable" version space – is biased and not likely to contain the correct labeling of the pool set. Indeed, the distance from the *Bayes classifier*, which provides the ground truth labeling, to the "boundary" of the version space is non-negligible.

Let $h_\perp^*$ be the projection of the Bayes classifier h^* to the set of hypotheses $\tilde{V}$ that agree with V on $\mathrm{AGR}(V)$ (see the left plot of Fig. 2), i.e.,

$$h_\perp^* := \arg\min_{h\in\tilde{V}} d(h,h^*), \tag{30}$$

$$\tilde{V} := \{h : h(\mathrm{AGR(V)}) = h'(\mathrm{AGR(V)}), h' \in V\}. \tag{31}$$

It is easy to see that $h_\perp^*$ provides the ground truth on $\mathrm{DIS}(V)$ and predicts the same labels on $\mathrm{AGR}(V)$ as hypotheses in V do, hence

$$d(h_\perp^*, h^*) = d(h, h^*; \mathrm{AGR(V)}) = \mathbb{E}_x[\mathbb{1}(x \in \mathrm{AGV}(V))\mathbb{1}(h(x) \neq h^*(x))],\ \forall h \in V. \tag{32}$$
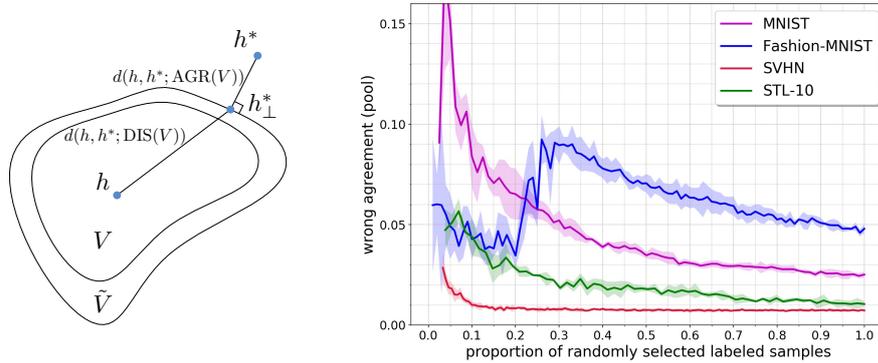
Fig. 2: **Left**: Projection of h^* to the samplable version space. **Right**: Wrong agreement of version spaces trained on random samples. Total numbers of samples are 400, 1000, 3000 and 2580 for MNIST, Fashion-MNIST, SVHN and STL-10 respectively.

where $d(h, h^*; \mathrm{AGR}(V))$ is the disagreement probability restricted to AGR(V), or equivalently the *wrong agreement* of hypotheses in V.

We show the evolution of wrong agreement in the right plot of Fig. 2. As more random samples are queried, the wrong agreement decreases for all datasets, but for some much slower than the others. In Fig. 3, we show for MNIST a 2-D embedding of version spaces using Multi-Dimensional Scaling (MDS) [20], which finds a low-dimensional representation of potentially high-dimensional data by preserving pairwise distances between the data points. The Bayes classifier is not contained in any of the samplable version spaces although the distances between them decrease steadily.

In general neural networks trained with a random subset do not automatically predict all labels in the pool set correctly, unless a relatively large proportion of samples are used for training. However, this fact does not render version space reduction inconsistent, because the samplable version space is not fixed, but it shifts towards the correct labeling and finally covers it when the whole pool set has been used.

We conjecture that the dynamics of active learning with neural networks have two major components: (1) shrinkage of the samplable version space, which is explicitly optimized by the learning algorithm and (2) reduction of bias, which is not directly controllable. Empirical evidence is provided in the next section.

7   Evaluation

Datasets and Architectures We conduct active learning experiments[4] on four image classification datasets: MNIST, Fashion-MNIST, SVHN and STL-10.

---

[4] Source code is available at `https://github.com/jiayu-liu/effective-version-space-reduction-for-convnets`.
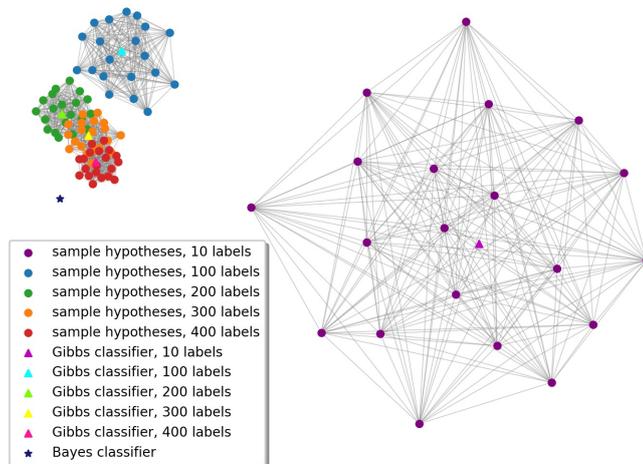
Fig. 3: Embedding of version spaces on MNIST using MDS. As more random samples are used for training, the samplable version spaces move closer to the Bayes classifier but hardly cover it.

Neural network architectures are chosen to be competent for each dataset but as simple as possible in the hope of controlling the model complexity and mitigating the effect of overfitting. See Table 1 for the complete experiment settings.

Active Learning Methods We compare nine querying methods: Random, variation ratio (VR), Gibbs error (GE), Bayesian Active Learning by Disagreement with Monte Carlo dropout (BALD-MCD) [16, 12], Core-Set [23], Deep-Fool Active Learning (DFAL) [10], pairwise disagreement (PWD), Gibbs-vote disagreement (GVD), and double-weighted pairwise disagreement (M^2-PWD) [24]. For each method on each dataset, at least three runs of active learning with different random balanced initial training set are performed.

Ensemble Size We train networks multiple times from scratch to obtain sample hypotheses and use them for prior mass and diameter estimation. Since diame-

Table 1: Settings for each dataset used in the active learning experiments.

Dataset	Pool/Val/Test	Model	Ensemble Size	Init/Query/Total	Runs
MNIST	45000/5000/10000	2-conv-layer ConvNet	20	10/5/400	4
Fashion-MNIST	55000/5000/10000	3-conv-layer ConvNet	20	10/10/1000	4
SVHN	40000/5000/15000	6-conv-layer ConvNet	20	100/20/3000	4
STL-10	4000/1000/8000	ResNet18	20	100/40/2580	3
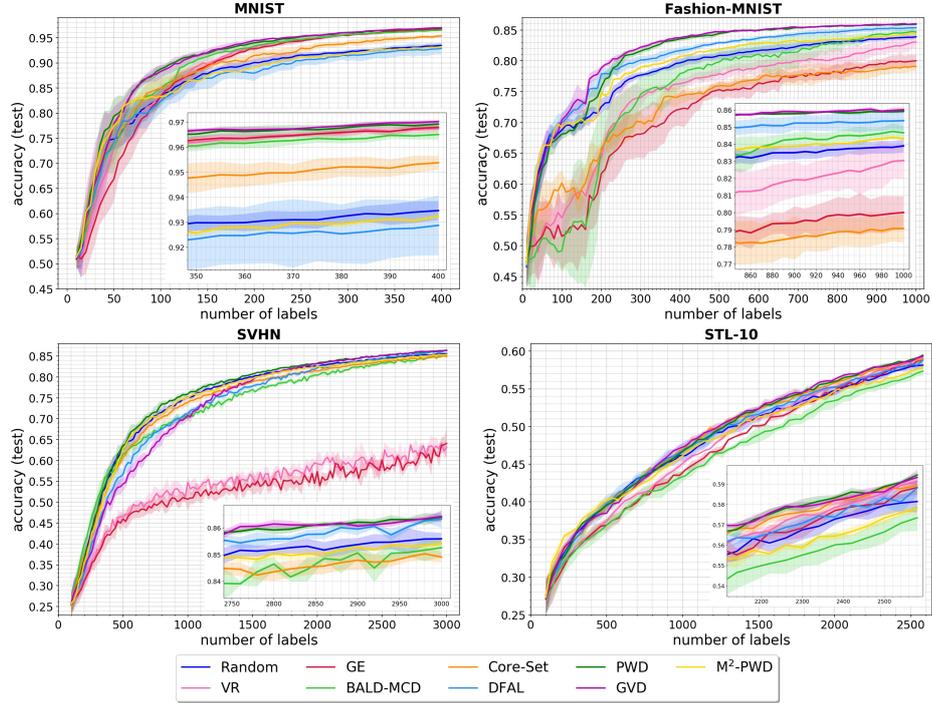
Fig. 4: Accuracy over number of queried labels on the test set. Direct diameter reduction methods PWD and GVD are consistently better than Random and are among the best methods. Weighted diameter reduction M^2-PWD is on par with Random. Other baselines are effective on some datasets but inferior to Random on the others. Note that PWD, GVD and M^2-PWD exhibit smaller variances than the others.

ters are estimated by considering partitioned version spaces, the ensemble size should be at least in the order of number of classes. We set the size to 20. Larger ensemble improves estimation but at the cost of longer training time. In preliminary experiments, we tried larger ensembles (40) and did not observe significant differences. Hence we do not include experiments on changing this hyper-parameter in the paper.

Query Size We set a small query budget for each round to reduce the correlation between queries. Larger budget may alleviate the pressure of frequent retraining, but the effect of each query can not be estimated and examined reliably. We observed in preliminary experiments that using larger budget (one or two orders larger) hides the differences between methods.

7.1   Diameter Reduction is More Effective Than Prior Mass Reduction

Fig. 4 and Table 2 show that direct diameter reduction methods PWD and GVD are consistently better than Random and achieve higher accuracy than other baselines while weighted diameter reduction M2-PWD is on par with Random. Diameter reduction methods usually exhibit less variances because training on samples queried by PWD, GVD and M2-PWD yields version spaces with smaller diameters and less diverse sample hypotheses. Prior mass reduction is not always effective and even fails on SVHN. This failure is an example of prior mass reduction being incapable of reducing the diameter, and provides empirical evidence that it may not be an appropriate objective for active learning.

Table 2: Accuracy on the test set in percentage.

	MNIST	Fashion-MNIST	SVHN	STL-10
#labels	400	1000	3000	2580
Random	93.47 ± 0.38	83.90 ± 0.38	85.60 ± 0.23	58.15 ± 0.54
VR	96.74 ± 0.15	83.05 ± 1.09	63.23 ± 1.99	59.13 ± 0.21
GE	96.79 ± 0.10	80.01 ± 0.94	64.08 ± 3.77	58.84 ± 0.34
BALD-MCD	96.51 ± 0.22	84.67 ± 0.41	85.26 ± 0.34	57.35 ± 0.64
Core-Set	95.38 ± 0.28	79.08 ± 0.82	84.91 ± 0.20	58.93 ± 0.33
DFAL	92.88 ± 1.19	85.38 ± 0.60	86.34 ± 0.33	58.81 ± 0.37
PWD	96.92 ± 0.12	85.92 ± 0.10	86.41 ± 0.12	**59.45 ± 0.11**
GVD	**97.02 ± 0.06**	**86.01 ± 0.15**	**86.44 ± 0.20**	59.33 ± 0.37
M^2-PWD	93.24 ± 0.09	84.33 ± 0.03	85.42 ± 0.16	57.81 ± 0.20

Table 3: Diameter (pairwise disagreement) on the test set in percentage.

	MNIST	Fashion-MNIST	SVHN	STL-10
#labels	400	1000	3000	2580
Random	2.86 ± 0.18	7.55 ± 0.26	13.13 ± 0.29	32.88 ± 0.43
VR	2.27 ± 0.18	10.64 ± 0.72	46.88 ± 2.76	34.21 ± 0.08
GE	2.30 ± 0.04	11.38 ± 1.52	44.87 ± 4.25	34.25 ± 0.09
BALD-MCD	2.39 ± 0.15	8.11 ± 0.51	16.58 ± 0.42	33.55 ± 0.44
Core-Set	2.91 ± 0.18	10.79 ± 1.34	14.66 ± 0.47	33.13 ± 0.64
DFAL	3.79 ± 0.60	7.06 ± 0.60	13.98 ± 0.31	32.41 ± 0.27
PWD	**1.93 ± 0.04**	**6.91 ± 0.16**	**12.80 ± 0.08**	**32.25 ± 0.26**
GVD	1.98 ± 0.05	6.98 ± 0.26	12.88 ± 0.25	32.96 ± 0.48
M^2-PWD	3.37 ± 0.13	7.22 ± 0.08	13.31 ± 0.13	33.23 ± 0.18

7.2   Comparison to Other Baselines

BALD-MCD, Core-Set and DFAL are not consistently better than Random although each of them achieves comparative test accuracy on certain dataset. Their inferiority to Random in terms of test accuracy usually correlates with higher diameter (See description in Fig. 5 and Table 3). BALD-MCD and DFAL are highly related to prior mass reduction methods in that BALD [16] seeks samples for which the model parameters under the posterior disagree the most about the prediction [16], and that DFAL, inspired by margin-based active learning [2], tries to locate the decision boundary with fewer labels which is essentially removing inconsistent hypotheses in the realizable case. However, none of them explicitly minimize the diameter, neither does Core-Set.

Note that for a fair comparison, we do not augment the training set by also adding the adversarial samples as the original DFAL paper [10] does. Samples with minimum adversarial perturbation are then verified reliably to be less effective than those lead to minimum diameter. The original Core-Set paper [23] uses a large query batch size (in the order of 1000). However, many baselines rely on greedy selection and do not perform any batch optimization. To reduce query correlation, we adopt as small batch size as possible. This allows reliable evaluation of the effectiveness of queried samples as in the online setting. We are therefore able to identify one major cause of inferiority to Random as failing to effectively reduce the version space diameter.

7.3   Evolution of Samplable Version Space and its Implications

As shown in Fig. 5 and 3, the samplable version space shifts closer to the correct labeling while reducing its diameter as more labels are queried. These two processes together result in smaller test error.

No Direct Control Over Reduction of Version Space Bias Interestingly, the Core-Set method, which queries representative samples from the pool set by solving a k-center problem in the feature space learned by neural networks, is incapable of achieving negligible wrong agreement on the learned version spaces. Indeed, it suffers larger version space bias than the direct diameter reduction methods. After all, random queries which are i.i.d. by assumption fail to achieve this goal as concluded in Section 6 and other attempts without augmenting the training data seem doomed.

Prior Mass Induced by Stochastic Gradient Descent May Not Be a Reliable Surrogate Measure The continued decline in wrong agreement indicates that the distribution over labelings changes over time. This fact of shifting density over samplable labelings renders the notion of prior mass problematic, hence all notions relying on prior mass may not be well-defined. A direct consequence is that an estimate of the worst-case version space reduction would be more reliable than the average-case one. For example, VR provides a more

reliable estimate of version space reduction than GE does.

Inferiority of Weighted Diameter Reduction Method The estimation of weighted diameter involves estimating the prior mass. Hence, the inferiority of M^2-PWD to PWD and GVD can be attributed to the intrinsic difficulty of obtaining unbiased samplable version spaces and the resulted density shift. A supportive evidence can be seen by noting that on MNIST and Fashion-MNIST, where the wrong agreement is large (hence large density shift), the weighted variant performs worse, while on SVHN and STL-10, where the wrong agreement is small (hence small shift), the gap is less significant.

7.4   Gibbs-Vote Disagreement

The Gibbs-vote disagreement is among the best methods on all datasets, except for the early learning stage on SVHN. Its effectiveness can be ascribed to an interesting phenomenon – majority voting reduces mistakes. Although it need not necessarily be the case, this phenomenon occurs in many situations and the boost to accuracy depends on the variance of errors of Gibbs classifiers [21]. We show empirically that the majority vote classifier indeed has smaller error rate than random hypotheses in the version space in Fig. 6. Hence, optimizing the Gibbs-vote disagreement not only reduces the diameter but also implicitly moves the consistent hypotheses closer to the correct labeling, which is useful when the samplable version spaces are biased and do not contain the Bayes classifier.
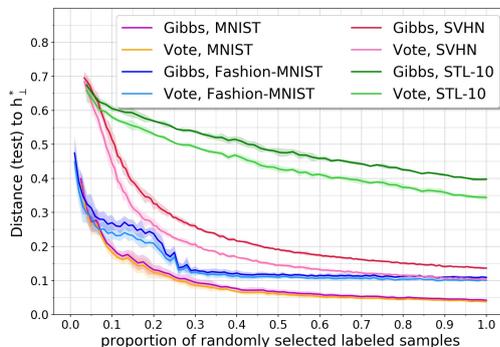
Fig. 6: Distance from the Gibbs and the majority vote classifier to the projection of h^*. On four datasets, the majority vote classifier has a smaller distance, hence smaller error rate. See description of Fig. 2 for total numbers of random samples.

8   Conclusion

In this work, we studied version space reduction for convolutional neural networks. We revealed the differences and connections between prior mass and diameter reduction methods and proposed the Gibbs-vote disagreement as a new effective diameter-reduction method. With experiments on four datasets, we shed light into how version space reduction works in the deep active learning setting and demonstrated the superiority of diameter reduction over prior mass reduction methods and other baselines.
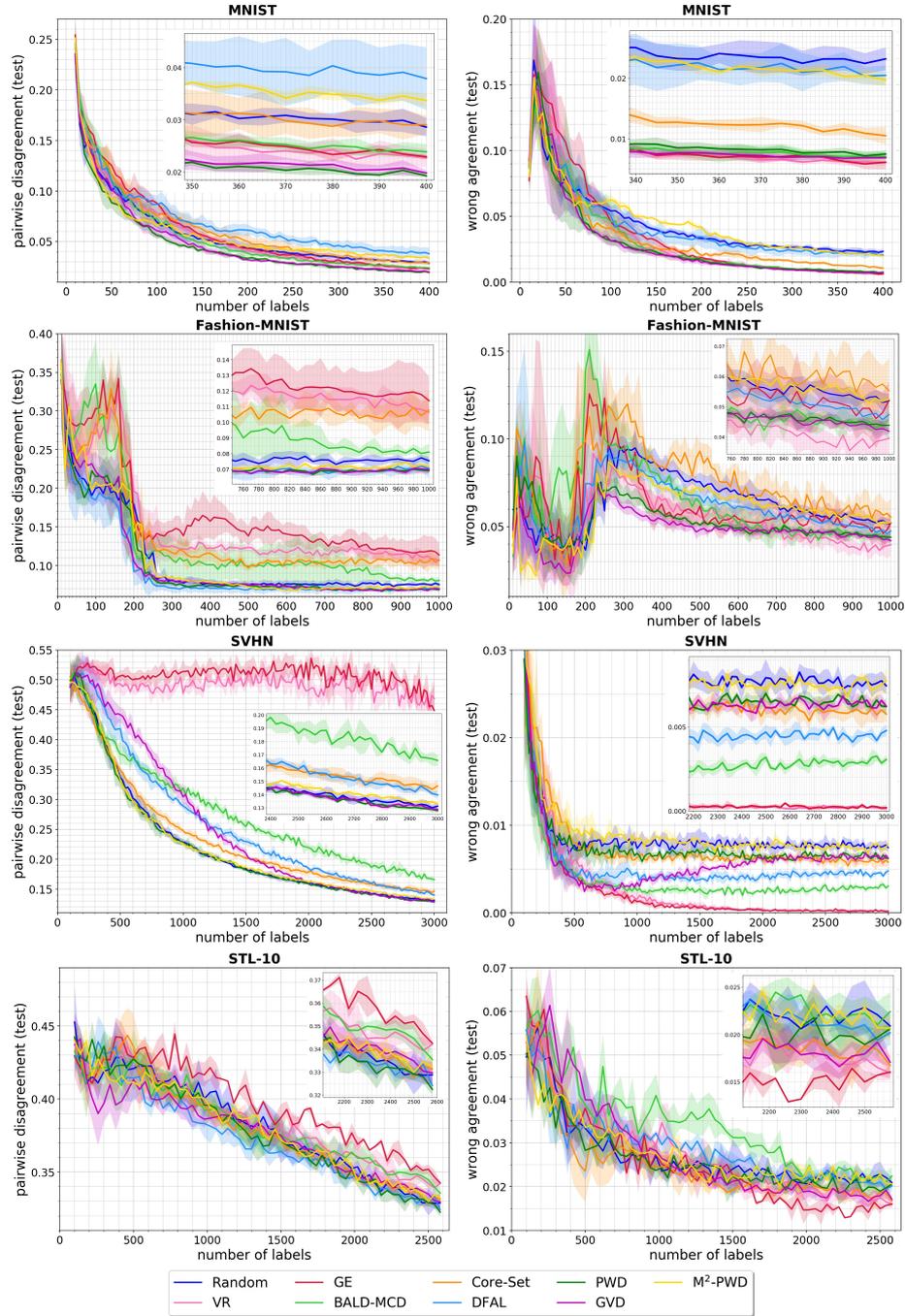
Fig. 5: Pairwise disagreement and wrong agreement over number of queried labels on the test set. Except direct diameter reduction methods PWD and GVD, other baselines are not consistently better than or on par with Random at reducing version space diameter. Performing worse than Random: GE, VR and BALD-MCD on datasets except MNIST, Core-Set on Fashion-MNIST and SVHN, and DFAL on MNIST and SVHN, and M^2-PWD on MNIST.

Acknowledgments

References

1. Balcan, M.F., Beygelzimer, A., Langford, J.: Agnostic active learning. In: Proceedings of the 23rd International Conference on Machine Learning. pp. 65—72 (2006)
2. Balcan, M.F., Broder, A., Zhang, T.: Margin based active learning. In: Proceedings of the 20th Annual Conference on Learning Theory. pp. 35–50 (2007)
3. 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. pp. 9368–9377 (2018)
4. Cuong, N.V., Lee, W.S., Ye, N.: Near-optimal adaptive pool-based active learning with general loss. In: Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence. pp. 122–131 (2014)
5. Cuong, N.V., Lee, W.S., Ye, N., Chai, K.M.A., Chieu, H.L.: Active learning for probabilistic hypotheses using the maximum gibbs error criterion. In: Advances in Neural Information Processing Systems. pp. 1457–1465 (2013)
6. Dasgupta, S.: Analysis of a greedy active learning strategy. In: Advances in Neural Information Processing Systems. pp. 337–344 (2005)
7. Dasgupta, S.: Coarse sample complexity bounds for active learning. In: Advances in Neural Information Processing Systems. pp. 235–242 (2006)
8. Dasgupta, S.: Two faces of active learning. In: Proceedings of the 20th International Conference on Algorithmic Learning Theory pp. 1–1 (2009)
9. Devroye, L., Györfi, L., Lugosi, G.: A probabilistic theory of pattern recognition, vol. 31. Springer Science & Business Media (2013)
10. Ducoffe, M., Precioso, F.: Adversarial active learning for deep networks: a margin based approach. arXiv preprint arXiv:1802.09841 (2018)
11. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the 33rd International Conference on Machine Learning. pp. 1050–1059 (2016)
12. Gal, Y., Islam, R., Ghahramani, Z.: Deep bayesian active learning with image data. In: Proceedings of the 34th International Conference on Machine Learning. pp. 1183–1192 (2017)
13. Gissin, D., Shalev-Shwartz, S.: Discriminative active learning. arXiv preprint arXiv:1907.06347 (2019)
14. Golovin, D., Krause, A.: Adaptive submodularity: A new approach to active learning and stochastic optimization. Journal of Artificial Intelligence Research pp. 427–486 (2011)
15. Hanneke, S.: A bound on the label complexity of agnostic active learning. In: Proceedings of the 24th International Conference on Machine Learning. pp. 353–360 (2007)
16. Houlsby, N., Huszár, F., Ghahramani, Z., Lengyel, M.: Bayesian active learning for classification and preference learning. Computing Research Repository abs/1112.5745 (2011)

17. Joshi, A.J., Porikli, F., Papanikolopoulos, N.: Multi-class active learning for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2372–2379 (2009)
18. Kääriäinen, M.: Generalization error bounds using unlabeled data. In: Proceedings of the 18th Annual Conference on Learning Theory. pp. 127–142 (2005)
19. Kirsch, A., van Amersfoort, J., Gal, Y.: Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In: Advances in Neural Information Processing Systems. pp. 7024–7035 (2019)
20. Kruskal, J.B.: Multidimensional scaling. No. 11, Sage (1978)
21. Lacasse, A., Laviolette, F., Marchand, M., Germain, P., Usunier, N.: Pac-bayes bounds for the risk of the majority vote and the variance of the gibbs classifier. In: Advances in Neural Information Processing Systems. pp. 769–776 (2007)
22. Pinsler, R., Gordon, J., Nalisnick, E., Hernández-Lobato, J.M.: Bayesian batch active learning as sparse subset approximation. In: Advances in Neural Information Processing Systems. pp. 6356–6367 (2019)
23. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: Proceedings of 6th the International Conference on Learning Representations (2018)
24. Tosh, C., Dasgupta, S.: Diameter-based active learning. In: Proceedings of the 34th International Conference on Machine Learning. pp. 3444–3452 (2017)
25. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: Proceedings of the 5th International Conference on Learning Representations (2017)