Active Domain Adaptation via Clustering Uncertainty-weighted Embeddings

Viraj Prabhu\textsuperscript{1}  Arjun Chandrasekaran\textsuperscript{2,\ast}  Kate Saenko\textsuperscript{3}  Judy Hoffman\textsuperscript{1}
\textsuperscript{1}Georgia Tech  \textsuperscript{2}Max Planck Institute for Intelligent Systems, Tübingen  \textsuperscript{3}Boston University
\{virajp,judy\}@gatech.edu  achandrasekaran@tue.mpg.de  saenko@bu.edu

Abstract

Generalizing deep neural networks to new target domains is critical to their real-world utility. In practice, it may be feasible to get some target data labeled, but to be cost-effective it is desirable to select a maximally-informative subset via active learning (AL). We study the problem of AL under a domain shift, called Active Domain Adaptation (Active DA). We demonstrate how existing AL approaches based solely on model uncertainty or diversity sampling are less effective for Active DA. We propose Clustering Uncertainty-weighted Embeddings (CLUE), a novel label acquisition strategy for Active DA that performs uncertainty-weighted clustering to identify target instances for labeling that are both uncertain under the model and diverse in feature space. CLUE consistently outperforms competing label acquisition strategies for Active DA and AL across learning settings on 6 diverse domain shifts for image classification. Our code is available at https://github.com/virajprabhu/CLUE.

1. Introduction

Deep neural networks excel at learning from large labeled datasets but struggle to generalize this knowledge to new target domains [32, 42]. This limits their real-world utility, as it is impractical to collect a large new dataset for every new deployment domain. Further, all target instances are usually not equally informative, and it is far more cost-effective to identify maximally informative target instances for labeling. While Active Learning [2, 6, 8, 9, 35, 36] has extensively studied the problem of identifying informative instances for labeling, it typically focuses on learning a model from scratch and does not operate under a domain shift. In many practical scenarios, models are trained in a source domain and deployed in a different target domain, often with additional domain adaptation [10, 17, 32, 44]. In this work, we study the problem of active learning under such a domain shift, called Active Domain Adaptation [29] (Active DA).

Concretely, given i) labeled data in a source domain, ii) unlabeled data in a target domain, and iii) the ability to obtain labels for a fixed budget of target instances, the goal of Active DA is to select target instances for labeling and learn a model with high accuracy on the target test set. Active DA has widespread utility as a means of cost-effective adaptation from cheaper to more expensive sources of labels (e.g. synthetic to real data), as well as when the quantity (e.g. autonomous driving) or cost (e.g. medical diagnosis) of labeling in the target domain is prohibitive. Despite its practical utility, it is a challenging task that has seen limited follow-up work since its introduction over ten years ago [5, 29, 40].

The traditional AL setting typically focuses on techniques to select samples to efficiently learn a model from scratch, rather than adapting under a domain shift [36]. As a result, existing state-of-the-art AL methods based on either uncertainty or diversity sampling are less effective for Active DA. Uncertainty sampling selects instances that are highly uncertain under the model’s beliefs [8, 9, 20, 41]. Under a domain shift, uncertainty estimates on the target domain may...
be miscalibrated [39] and lead to sampling uninformative, outlier, or redundant instances (Fig. 1, top). A parallel line of work in AL based on diversity sampling instead selects instances dissimilar to one another in a learned embedding space [13, 35, 38]. In Active DA, this can lead to sampling uninformative instances from regions of the feature space that are already well-aligned across domains (Fig. 1, middle). As a result, solely using uncertainty or diversity sampling is suboptimal for Active DA, as we demonstrate in Sec 4.4.

Recent work in AL and Active DA has sought to combine uncertainty and diversity sampling. AADA [40], the state-of-the-art Active DA method, combines uncertainty with diversity measured by ‘targetness’ under a learned domain discriminator. However, targetness does not ensure that the selected instances are representative of the entire target data distribution (i.e. not outliers), or dissimilar to one another. Ash et al. [2] instead propose performing clustering in a hallucinated “gradient embedding” space. However, they rely on distance-based clustering in high-dimensional spaces, which often leads to suboptimal results.

In this work, we propose a novel label acquisition strategy for Active DA that combines uncertainty and diversity sampling in a principled manner without the need for complex gradient or domain discriminator-based diversity measures. Our approach, Clustering Uncertainty-weighted Embeddings (CLUE), identifies informative and representative target instances from dense regions of the feature space. To do so, CLUE clusters deep embeddings of target instances weighted by the corresponding uncertainty of the target model. Our weighting scheme effectively increases the density of instances proportional to their uncertainty. To construct non-redundant batches, CLUE then selects nearest neighbors to the inferred cluster centroids for labeling. Our algorithm then leverages the acquired target labels and, optionally, the labeled source and unlabeled target data, to update the model, consistently leading to more cost-effective domain alignment than competing (and frequently more complex) alternatives.

Contributions:

1. We benchmark the performance of state-of-the-art methods for active learning on challenging domain shifts, and find that methods based purely on uncertainty or diversity sampling are not effective for Active DA.
2. We present CLUE, a novel and easy-to-implement label acquisition strategy for Active DA that uses uncertainty-weighted clustering to identify instances that are both uncertain under the model and diverse in feature space.
3. We present results on 6 diverse domain shifts from the DomainNet [27], Office [32], and DIGITS [22, 26] benchmarks for image classification. Our method CLUE improves upon both the previous state-of-the-art in Active DA across shifts (by as much as 9% in some cases), as well as state-of-the-art methods for active learning, across multiple learning strategies.

2. Related Work

Active Learning (AL) for CNN’s. AL for CNN’s has focused on the batch-mode setting due to the instability associated with single-instance updates. The two most successful paradigms in AL have been uncertainty sampling and diversity sampling [2]. Uncertainty-based methods select instances with the highest uncertainty under the current model [8, 9, 34, 41], using measures such as entropy [45], classification margins [30], or confidence. Diversity-based methods select instances that are representative of the entire dataset, and optimize for diversity in a learned embedding space, via clustering, or core-set selection [12, 13, 35, 38].


Domain Adaptation. The task of transferring models trained on a labeled source domain to an unlabeled [10, 17, 32, 44] or partially-labeled [7, 33, 47] target domain has been studied extensively. Initial approaches aligned feature spaces by optimizing discrepancy statistics between the source and target [23, 44], while in recent years adversarial learning of a feature space encoder alongside a domain discriminator has become a popular alignment strategy [10, 11, 43]. In this work, we propose a label acquisition strategy for active learning under a domain shift that generalizes across multiple domain adaptation strategies.

Active Domain Adaptation (Active DA). Unlike semi-supervised domain adaptation which assumes labels for a random subset of target instances, Active DA focuses on selecting target instances to label for domain adaptation. Rai et al. [29] first studied the task of Active DA applied to sentiment classification from text data. They propose ALDA, which samples instances based on model uncertainty and a learned domain separator. Chattopadhyay et al. [5] select target instances and learn importance weights for source points by solving a convex optimization problem of minimizing MMD between features. More recently, Su et al. [40] study Active DA in the context of deep CNN’s and propose AADA, wherein target instances are selected based on predictive entropy and targetness measured by an adversarially trained domain discriminator, followed by adversarial domain adaptation via DANN [11]. We propose CLUE, a novel label acquisition strategy for Active DA that identifies uncertain and diverse instances for labeling that outperforms prior work on diverse shifts across multiple learning strategies.
3. Approach

We address active domain adaptation (Active DA), where the goal is to generalize a model trained on a source domain to an unlabeled target domain, with the option to query an oracle for labels for a subset of target instances. While individual facets of this task – adapting to a new domain and selective acquisition of labels, have been well-studied as the problems of Domain Adaptation (DA) and Active Learning (AL) respectively, Active DA presents the new challenge of identifying target instances that will, once labeled, result in the most sample-efficient domain alignment. Further, the answer to this question may vary based on the properties of the specific domain shift. In this section, we present CLUE, a novel label acquisition strategy for Active DA which performs consistently well across diverse domain shifts.

3.1. Notation & Preliminaries

In Active DA, the learning algorithm has access to labeled instances from the source domain \((X_S, Y_S)\) (solid pink in Fig. 2), unlabeled instances from the target domain \(X_{UT}\) (blue outline in Fig. 2), and a budget \(B (= 3 \text{ in Fig. 2})\) which is much smaller than the amount of unlabeled target data. The learning algorithm may query an oracle to obtain labels for at most \(B\) instances from \(X_{UT}\), and add them to the set of labeled target instances \(X_{LT}\). The entire target domain data is \(X_T = X_{LT} \cup X_{UT}\). The task is to learn a function \(h: X \to Y\) (a convolutional neural network (CNN) parameterized by \(\Theta\)) that achieves good predictive performance on the target. In this work, we consider Active DA in the context of \(C\)-way image classification – the samples \(x_S \in X_S, x_T \in X_T\) are images, and labels \(y_S \in Y_S, y_T \in Y_T\) are categorical variables \(y \in \{1, 2, \ldots, C\}\).

**Active Learning.** The goal of active learning (AL) is to identify target instances that, once labeled and used for training the model, minimize its expected future loss. In practice, prior works in AL identify such instances based primarily on two proxy measures, uncertainty and diversity (see Sec. 2). We first revisit these terms in the context of Active DA.

**Uncertainty.** Prior work in AL has proposed using several measures of model uncertainty as a proxy for informativeness (see Sec. 2). However, in the context of Active DA, using model uncertainty to select informative samples presents a conundrum. On the one hand, models benefit from initialization on a related source domain rather than learning from scratch. On the other hand, under a strong distribution shift, model uncertainty may often be miscalibrated [39]. Unfortunately however, without access to target labels, it is impossible to evaluate the reliability of model uncertainty!

**Diversity.** Acquiring labels solely based on uncertainty often leads to sampling batches of similar instances with high redundancy, or to sampling outliers. A parallel line of work in active learning instead proposes sampling diverse instances that are representative of the unlabeled pool of data. Several definitions of “diverse” exist in the literature: some works define diversity as coverage in feature [35] or “gradient embedding” space [2], while prior work in Active DA measures diversity by how “target-like” an instance is [40]. In Active DA, training on a related source domain (optionally followed by unsupervised domain alignment), results in some classes being better aligned across domains than others. Thus, in order to be cost-efficient it is important to avoid sampling from already well-learned regions of the feature space. However, purely diversity-based AL methods are unable to account for this, and lead to sampling redundant instances.
While sampling instances that are either uncertain or diverse may be useful to learning, an optimal label acquisition strategy for Active DA would ideally capture both jointly. We now introduce CLUE, a label acquisition strategy for Active DA that captures both uncertainty and diversity.

3.2. Clustering Uncertainty-weighted Embeddings

To measure informativeness we use predictive entropy \( H(Y|X; \Theta) \) [45] (\( H(Y|X) \) for brevity), which for \( C \)-way classification, is defined as:

\[
H(Y|X) = - \sum_{c=1}^{C} p_\Theta(Y = c|X) \log p_\Theta(Y = c|X)
\]  

(1)

Under a domain shift, entropy can be viewed as capturing both uncertainty and domainness. Rather than training an explicit domain discriminator [10, 40], we consider an implicit domain classifier \( d(x) \) [33] based on entropy thresholding:

\[
d(x) = \begin{cases} 
1, & \text{if } H(Y|X) \geq \gamma \\
0, & \text{otherwise}
\end{cases}
\]  

(2)

where 1 and 0 denote target and source domain labels, and \( \gamma \) is a threshold value. The probability of an instance belonging to the target domain is thus given by:

\[
p(d(x) = 1) = \frac{H(Y|X)}{\log(C)} \times H(Y|X) \quad \text{[C is constant]}
\]  

(3)

where \( \log(C) \) is the maximum possible entropy of a \( C \)-way distribution. Next, we measure diversity based on feature-space coverage. Let \( \phi(x) \) denote feature embeddings extracted from model \( h \). We identify diverse instances by partitioning \( X_T \) into \( K \) diverse sets via a partition function \( S : X_T \rightarrow \{X_1, X_2, \ldots, X_K\} \). Let \( \{\mu_1, \mu_2, \ldots, \mu_K\} \) denote the corresponding centroid of each set. Each set \( X_k \) should have a small variance \( \sigma^2(X_k) \). Expressed in terms of pairs of samples, \( \sigma^2(X_k) = \frac{1}{|X_k|^2} \sum_{x_i, x_j \in X_k} ||\phi(x_i) - \phi(x_j)||^2 \) [48]. The goal is to group target instances that are similar in the CNN’s feature space, into a set \( X_k \). However, while \( \sigma^2(X_k) \) is a function of the target data distribution and feature space \( \phi(\cdot) \), it does not account for uncertainty.

To jointly capture both diversity and uncertainty, we propose weighting samples based on their uncertainty (given by Eq. 1), and compute the weighted population variance [28]. The overall set-partitioning objective is:

\[
\arg\min_{S, \mu} \sum_{k=1}^{K} \frac{1}{Z_k} \sum_{x \in X_k} H(Y|X)||\phi(x) - \mu_k||^2
\]  

(4)

where the normalization \( Z_k = \sum_{x \in X_k} H(Y|X) \). Our weighted set partitioning can also be viewed as standard set partitioning in an alternate feature space, where the density of instances is artificially increased proportional to their predictive entropy. Intuitively, this emphasizes representative sampling from uncertain regions of the feature space.

Since the objective in Eq. 4 is NP-hard, we approximate it using a Weighted K-Means algorithm [19] (see Algorithm 1 - uncertainty-weighting is used in the update step). We set \( K = B \) (budget), and use activations from the penultimate CNN layer as \( \phi(x) \). After clustering, to select representative instances (i.e., non-outliers), we acquire labels for the nearest neighbor to the weighted-mean of each set \( \mu_k \) in Eq. 4. Note that Eq. 4 equivalently maximizes the sum of squared deviations between instances in different sets [21], ensuring that the constructed batch of instances has minimum redundancy.

### Trading-off Uncertainty and Diversity

CLUE captures an implicit tradeoff between model uncertainty (via entropy-weighting) and feature-space coverage (via clustering). Consider the predictive probability distribution for instance \( x \):

\[
p_\Theta(Y|x) = \sigma \left( \frac{h(x)}{T} \right)
\]  

(5)

where \( \sigma \) denotes the softmax function and \( T \) denotes its temperature. We observe that by modulating \( T \), we can control the uncertainty-diversity tradeoff. For example, by

**Algorithm 1 CLUE:** Our proposed Active DA method, which uses Clustering Uncertainty-weighted Embeddings (CLUE) to select instances for labeling followed by a model update via semi-supervised domain adaptation.

1. **Require:** Neural network \( h = f(\phi(\cdot)) \), parameterized by \( \Theta \), labeled source data \( \{X_S, Y_S\} \), unlabeled target data \( X_T \). Per-round budget \( B \), Total rounds \( R \).
2. **Define:** Labeled target set \( X_{LT} = \emptyset \)
3. **Train source model** \( \Theta^1 \) on \( \{X_S, Y_S\} \).
4. **Adapt model** to unlabeled target domain (optional).
5. **for** \( \rho = 1 \) to \( R \) **do**
6. **CLUE:** For all instances \( x \in X_T \setminus X_{LT} \):
   1. **Compute deep embedding** \( \phi(x) \)
   2. **Run Weighted K-Means** until convergence (Eq. 4):
      a. Init. \( K (=B) \) centroids \( \{\mu_i\}_{i=1}^B \) (KMeans++)
      b. **Assign:**
         \[ X_k \leftarrow \{x|k = \arg\min_{i=1,\ldots,K} \sum_{x \in X_k} H(Y|\phi(x))|\phi(x) - \mu_i|^2 \} \forall x \]
      c. **Update:** \( \mu_k \leftarrow \frac{\sum_{x \in X_k} \phi(x)}{|X_k|} \forall k \)
   3. **Acquire labels** for nearest-neighbor to centroids \( X_{LT} \leftarrow \{\text{NN}(\mu_i)\}_{i=1}^B \)
   4. **X_{LT} = X_{LT} \cup X_{LT}^\rho**
7. **Semi-supervised DA:** Update model \( \Theta^\rho+1 \).
8. **Return:** Final model parameters \( \Theta^{R+1} \).
increasing $T$, we obtain more diffuse softmax distributions for all points leading to similar uncertainty estimates across points; correspondingly, we expect diversity to play a bigger role. Similarly, at lower values of $T$ we expect uncertainty to have greater influence.

Our full label acquisition approach, Clustering Uncertainty-weighted Embeddings (CLUE), thus identifies instances that are both uncertain and diverse (see Fig. 2).

**Domain adaptation.** After acquiring labels via CLUE, we proceed to the next step of active adaptation: we update the model using the acquired target labels and optionally, the labeled source and unlabeled target data (see Fig. 2, right). In our main experiments (Sec. 4.4), we experiment with 3 learning strategies: i) finetuning on target labels, ii) domain-adversarial learning via DANN [10] with an additional target cross-entropy loss, and iii) semi-supervised adaptation via minimax entropy (MME [33]). In Sec 4.6 we also combine CLUE with additional DA methods from the literature.

Algorithm 1 describes our full approach when using CLUE in combination with semi-supervised domain adaptation. Given a model trained on labeled source instances, we align its representations with unlabeled target instances via unsupervised domain adaptation. For $R$ rounds with per-round budget $B$, we iteratively i) acquire labels for $B$ target instances that are identified via our proposed sampling approach (CLUE), and ii) Update the model using a semi-supervised domain alignment strategy.

4. Experiments

We begin by describing our datasets and metrics, implementation details, and baselines (Sec 4.1- 4.3). Next, we benchmark the performance of CLUE across 6 domain shifts of varying difficulty against state-of-the art methods for Active DA and AL, across different learning settings (Sec 4.4). We then ablate our method, analyze its sensitivity to various hyperparameters, and visualize its behavior (Sec 4.5). Finally, we combine our method with various DA strategies, and study its effectiveness in learning from scratch (Sec 4.6). We follow the standard batch active learning setting [4], in which we perform multiple rounds of batch active sampling, label acquisition, and model updates.

4.1. Datasets and Metrics

**DomainNet.** DomainNet [27] is a large domain adaptation benchmark for image classification, containing 0.6 million images from 6 distinct domains spanning 345 categories. We study four shifts of increasing difficulty as measured by source→target transfer accuracy (TA): Real→Clipart (easy, TA=40.6%), Clipart→Sketch (moderate, TA=34.7%), Sketch→Painting (hard, TA=30.3%), and Clipart→Quickdraw (very hard, TA=11.9%).

**DIGITS and Office.** We also report performance on the SVHN [26]→MNIST [22] and DSLR→Amazon [32] shifts.

**Metric.** We compute model accuracy on the target test split versus the number of labels used from the target train split at each round. We run each experiment 3 times and report mean accuracies. For clarity, we report performance at 3 randomly chosen intermediate budgets in the main paper and include full plots (mean accuracies and 1 standard deviation over all rounds) in the supplementary.

4.2. Implementation details

**DomainNet.** We use a ResNet34 [16] CNN, and perform 10 rounds of Active DA with a randomly selected per-round budget $= 500$ instances (total of 5000 labels). On DomainNet, we use the Clipart→Sketch shift as a validation shift and use a small target validation set to select a softmax temperature of $T = 0.1$ which we use for all other DomainNet shifts (details in supplementary). We include a sensitivity analysis over $T$ and $B$ in Sec. 4.5.

**DIGITS.** We match the experimental setting to Su et al. [40]; we use a modified LeNet architecture [17], and perform 30 rounds of Active DA with $B=10$.

**Office.** We use a ResNet34 CNN and perform 10 rounds of Active DA with $B=30$. On DIGITS and Office, we use the default value of $T = 1.0$. Across datasets, we use penultimate layer embeddings for CLUE and implement weighted K-means with $K=B$. All models are first trained on the labeled source domain. When adapting via semi-supervised domain adaptation, we additionally employ unsupervised feature alignment to the target domain at round 0. For additional details see supplementary.

4.3. Baselines

We compare CLUE against several state-of-the-art methods for Active DA and Active Learning.

1) **AADA:** Active Adversarial Domain Adaptation [40] (AADA) is a state-of-the-art Active DA method which performs alternate rounds of active sampling and adversarial domain adaptation via DANN [11]. It samples points with high predictive entropy and high probability of belonging to the target domain as predicted by the domain discriminator. Further, we also benchmark the performance of 4 diverse AL strategies from prior work in the Active DA setting.

2) **entropy** [45]: Selects instances for which the model has highest predictive entropy.

3) **margin** [30]: Selects instances for which the score difference between the model’s top-2 predictions is the smallest.

4) **coreset** [35]: Core-set formulates active sampling as a set-cover problem, and solves the K-Center [46] problem. We use the greedy version proposed in Sener et al. [35].

5) **BADGE** [2]: BADGE is a recently proposed state-of-the-art active learning strategy that constructs diverse batches by running KMeans++ [1] on “gradient embeddings” that incorporate model uncertainty and diversity.

Methods (2) and (3) are uncertainty based, (4) is diversity-based, and (1) and (5) are hybrid approaches.
We evaluate all methods across three ways of learning in the (S), Painting (P) and Quickdraw (Q). We perform 10 rounds of Active DA with

Table 1: Accuracies on target test set for 4 DomainNet shifts of increasing difficulty spanning 5 domains: Real (R), Clipart (C), Sketch (D), DIGITS, and Office. We make the following observations:

- Finetuning a model trained on the source
- Margin-based (margin [30], coreset [35], BADGE [2]) and Active DA (AADA), spanning different AL paradigms: uncertainty sampling (U), diversity sampling (D), and hybrid (H) combinations of the two. We use multiple learning strategies: finetuning (ft), Minimax entropy [33] (MME [33] (state-of-the-art semi-supervised DA method), and semi-supervised DA via DANN [10]. Best performance is in bold, gray rows are our method.

Table 2: Active DA accuracies on target test set at 3 intermediate budgets (30, 60, 150) for: Middle: 30 rounds with B = 10 from SVHN→MNIST (DIGITS). Right: 10 rounds with B = 30 from DSLR→Amazon (Office). Best performance is in bold, gray rows are our method. For full plots see supplementary.

4.4. Results

We evaluate all methods across three ways of learning in the presence of a domain shift with the acquired labels:

1) **FT from source**: Finetuning a model trained on the source domain with acquired target labels.

2) **MME [33] from source**: Minimax entropy [33] (MME) is a state-of-the-art semi-supervised DA method that starts from a source model and minimizes an adversarial entropy loss for unsupervised domain alignment in addition to finetuning on labeled source and target data.

Tables 1 and 2 demonstrate our results on DomainNet, DIGITS, and Office. We make the following observations:

- Uncertainty and diversity sampling are less effective for Active DA, frequently underperforming even random sampling. Approaches solely based on uncertainty (e.g. margin [30]) work well on relatively easier shifts (R→C with MME, SVHN→MNIST), but overall we find uncertainty-based (margin, entropy), and diversity-based (coreset) approaches generalize poorly to challenging shifts (e.g. S→P, C→Q), frequently underperforming even random sampling! On the other hand, hybrid approaches (CLUE and BADGE) that combine uncertainty and diversity are versatile across shift difficulties.

- **CLUE outperforms prior AL methods in the Active DA setting.** Across learning strategies, shifts, benchmarks, and most rounds, CLUE consistently performs best. Averaged over 4 DomainNet shifts, CLUE outperforms margin-based uncertainty sampling and coreset-based diversity sampling at $B = 2k$ by 1.4% and 3% when finetuning, and 1.3% and 2.3% when adapting via MME (Tab. 1). Similarly, CLUE outperforms the next best-performing method (BADGE) by 0.7% at $B = 2k$ when finetuning and 0.8% with MME (4 shift average). While BADGE [2] is also a hybrid AL method that combines uncertainty and diversity sampling, it does so by clustering in a high-dimensional "gradient-embedding" space (~176k dimensions on C→S with a ResNet34, versus 512-dimensional embeddings used by CLUE, details in Tab. 3), in which distance-based diversity measures may note here that DomainNet is a complex benchmark with 345

- **CLUE is also a hybrid AL method** by...
across diverse shifts without shift-specific tuning.

On DIGITS and Office, CLUE’s gains are even more significant (Tab. 2). For instance at $B = 30$ on SVHN→MNIST, CLUE improves upon margin, coreset, and BADGE when finetuning by 1.9%, 12.3%, and 5.2%, and 4%, 2.5% and 0.6% on DSLR→Amazon.

▷ Additional unsupervised adaptation helps in the Active DA setting. Across AL methods, we observe adaptation with MME to consistently outperform finetuning (e.g. by 2.4-2.7% accuracy on DomainNet).

▷ CLUE significantly outperforms the state-of-the-art Active DA method AADA. AADA [40] acquires labels by using a domain classifier learned via DANN [10]. Thus, it is undefined in the FT and MME settings. For an apples-to-apples comparison, we report performance of CLUE +DANN in the last 2 rows of Tabs. 1 and 2. As seen, CLUE +DANN consistently outperforms AADA, the state-of-the-art Active DA method, e.g. by 0.4%-2% on DomainNet. Further, we find the performance gap between our method and AADA [40] increases with increasing shift difficulty, as predictive uncertainty becomes increasingly unreliable (3.4% gain at $B = 2k$ on the very hard C→Q shift). We observe similar improvements over AADA on the DIGITS and Office (Tab. 2) benchmarks, e.g. 2.4% and 7.9% at $B = 60$. Further, our best performing CLUE + MME strategy improves the gains still further to as much as 3.5% at $B = 2k$ on DomainNet and 9% at $B = 60$ on Office!

As discussed in Sec. 3, the optimal label acquisition criterion may vary across shifts and stages of training as the model’s uncertainty estimates and feature space evolve, and it is challenging for a single approach to work well consistently. Despite this, CLUE effectively trades-off uncertainty and diversity to generalize reliably across shifts.

4.5. Analyzing and Ablating CLUE

Visualizing CLUE via t-SNE. We provide an illustrative comparison of sampling strategies using t-SNE [25]. Fig. 3 shows an initial feature landscape together with points selected by entropy-based uncertainty sampling, diversity-based coreset sampling, and CLUE at Round 0 on the SVHN→MNIST shift. We find that entropy [45] (left) samples uncertain but redundant points, coreset [35] samples diverse but not necessarily uncertain points, while our method, CLUE, samples both diverse and uncertain points. In the supplementary, we include visualizations over several rounds and find that CLUE consistently selects diverse target instances from dense, uncertain regions of the feature space.

Varying uncertainty measure in CLUE. In Fig. 4a, we consider alternative uncertainty measures for CLUE on the C→S shift. We show that our proposed use of sample entropy significantly outperforms a uniform sample weight and narrowly outperforms an alternative uncertainty measure - sample margin score (difference between scores for top-2 most likely classes). This illustrates the importance of using uncertainty-weighting to bias CLUE towards informative samples. We also experimented (not shown) with last-layer embeddings (instead of penultimate) for CLUE, and observed near-identical performance across multiple shifts, suggesting that CLUE is not sensitive to this choice.

Sensitivity to parameters. In Fig. 4 we measure CLUE’s sensitivity to two parameters: the softmax temperature hyperparameter $T$ and experimental parameter budget $B$.

i) Sensitivity to softmax temperature $T$. Recall from Sec. 3 that by tuning the softmax temperature in CLUE, we can vary the trade-off between uncertainty and diversity. In Fig. 4b we run a sweep over temperature values used for CLUE on C→S. As seen, lower values of temperature (which emphasizes the role of uncertainty) improve performance, particularly at later rounds when uncertainty estimates are more reliable. We note that $T$ is an optional hyperparameter that may be tuned if a small target validation set is available, but CLUE obtains strong state-of-the-art results across across DIGITS, Office, and DomainNet even with the default value of $T=1.0$. On DomainNet, we further improve performance...
Figure 4: (a), (b), (c): Ablating and analyzing CLUE on C→S. (d): Combining CLUE with different DA strategies on C→S. Best viewed in color. We perform 10 rounds of Active DA with B=500, and report accuracy mean and 1 standard deviation (via shading) over 3 runs.

by selecting $T = 0.1$ via grid search on a single C→S shift and find that it generalizes to other DomainNet shifts.

ii) Sensitivity to budget $B$. We now vary the per-round budget (and consequently the total number of active adaptation rounds) and report performance on the Clipart→Sketch shift. As seen in Fig. 4c, CLUE performs well across budget values of 100, 500, 1k and 2.5k. We also observe consistent performance with a different budget ($B = 30$) on the SVHN→MNIST shift (details in supplementary).

Time complexity. Table 3 shows the average case complexity and AL querying time-per-round on SVHN→MNIST and C→S. CLUE and BADGE, which achieve the best accuracy, are slower to run due to a (CPU) clustering step. CLUE can be optimized further via GPU acceleration, using last-(instead of penultimate) layer embeddings, or pre-filtering data before clustering.

<table>
<thead>
<tr>
<th>AL Strategy</th>
<th>Query Complexity</th>
<th>Query Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLUE (Ours)</td>
<td>$O(tNB^D)$</td>
<td>(60s, 16.2m)</td>
</tr>
<tr>
<td>BADGE [2]</td>
<td>$O(NBD^C)$</td>
<td>(103s, 16.3m)</td>
</tr>
<tr>
<td>coreset [35]</td>
<td>$O(CNB)$</td>
<td>(52s, 2.8m)</td>
</tr>
<tr>
<td>AADA [40]</td>
<td>$O(NC)$</td>
<td>(3.7s, 139s)</td>
</tr>
<tr>
<td>entropy [45]</td>
<td>$O(NC)$</td>
<td>(3.5s, 45s)</td>
</tr>
<tr>
<td>margin [30]</td>
<td>$O(NC)$</td>
<td>(3.2s, 45s)</td>
</tr>
</tbody>
</table>

Table 3: Query complexity and time-per-round for CLUE and prior AL strategies. $C$ and $N$ denotes number of classes and instances, $D$ denotes embedding dimensionality, $B$ denotes budget, $t =$ number of clustering iterations, and fwd stands for forward pass.

4.6. CLUE across learning strategies

CLUE with different DA strategies. We now study CLUE’s compatibility with a few additional domain adaptation strategies from the literature. In addition to the finetuning, DANN [10] and MME [33] already studied in Sec. 4.4, we fix our sampling strategy to CLUE and vary the learning strategy to: i) MMD [24], a discrepancy-statistic based DA method, and ii) VADA [37], a domain-classifier based method that uses virtual adversarial training. iii) ENT [14]: A variant of the MME method using standard rather than adversarial entropy minimization. Initial performance varies across methods since we employ unsupervised DA at Round 0.

In Fig. 4d, we observe that domain alignment with MME significantly outperforms all alternative methods. With all DA methods except VADA, we observe improvements over finetuning; however, MME clearly performs best. The improvements over DANN and VADA are consistent with Saito et al. [33], who find that domain-classifier based methods are less effective when some target labels are available.

How well does CLUE learn from scratch? While CLUE is designed for active learning under a domain shift, for completeness we also evaluate its performance against prior work when learning from “scratch” as is conventional in AL. We find that it outperforms prior work when finetuning using ImageNet [31] initialization on C→S, and performs on par with competing methods when finetuning from scratch on SVHN [26] (details in supplementary).

5. Conclusion

We address active domain adaptation, where the goal is to select target instances for labeling so as to generalize a trained source model to a new target domain. We show how existing active learning strategies based solely on uncertainty or diversity sampling are not effective for Active DA. We present CLUE, a novel label acquisition strategy for active sampling under a domain shift, that performs uncertainty-weighted clustering to select diverse, informative target instances for labeling from dense regions of the feature space. We demonstrate CLUE’s effectiveness over competing active learning and Active DA methods across learning settings and domain shifts, and comprehensively analyze its behavior.

Acknowledgements. This work was supported by the DARPA LwLL program. We would like to thank Devi Parikh for guidance, and Prithvijit Chattopadhyay, Cornelia Köhler, and Shruti Venkatram for feedback on the draft.
References


