Active Metric Learning and Classification using Similarity Queries

Namrata Nadagouda **Austin Xu** **Mark A. Davenport**

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, Georgia, USA

Abstract

Active learning is commonly used to train label-efficient models by adaptively selecting the most informative queries. However, most active learning strategies are designed to either learn a representation of the data (e.g., embedding or metric learning) or perform well on a task (e.g., classification) on the data. However, many machine learning tasks involve a combination of both representation learning and a task-specific goal. Motivated by this, we propose a novel unified query framework that can be applied to any problem in which a key component is learning a representation of the data that reflects similarity. Our approach builds on similarity or nearest neighbor (NN) queries which seek to select samples that result in improved embeddings. The queries consist of a reference and a set of objects, with an oracle selecting the object most similar (i.e., nearest) to the reference. In order to reduce the number of solicited queries, they are chosen adaptively according to an information theoretic criterion. We demonstrate the effectiveness of the proposed strategy on two tasks – active metric learning and active classification – using a variety of synthetic and real world datasets. In particular, we demonstrate that actively selected NN queries outperform recently developed active triplet selection methods in a deep metric learning setting. Further, we show that in classification, actively selecting class labels can be reformulated as a process of selecting the most informative NN query, allowing direct application of our method.

1 INTRODUCTION

A defining feature of modern machine learning is a reliance on large volumes of human-labeled data. Perhaps the most prominent example is the existence of massive hand-labeled image datasets, but the task of acquiring large amounts of human-provided data is nearly ubiquitous in machine learning. However, such data is not free; it is often tedious and expensive to gather a sufficient number of query responses to satisfy data hungry machine learning models.

Active learning (AL) [Settles, 2009] seeks to mitigate this issue by carefully selecting only the most informative samples to be labeled. More generally, AL attempts to identify the most informative queries to pose to an oracle. These queries can include asking for a class label or rating, or more general relational queries such as the similarity (or dissimilarity) of different items. In this paper, we focus on metric learning from perceptual similarity queries and classification, two prominent application areas for AL, and show that despite the different queries being posed to the oracle (labels in classification vs. similarity judgements for metric learning), there is a fundamental connection between the two problems.

Learning an embedding or representation of the data that accurately reflects similarity between items is the goal of metric learning. Many approaches in metric learning aim to make inter-class item distances small and intra-class item distances large by using triplets of items consisting of an anchor point, a positive sample of the same class as the anchor, and a negative point of a different class [Hoffer and Ailon, 2015]. Class labels are used as a proxy for item similarity/dissimilarity, which is only feasible if class labels are widely available. However, when given a new (unlabeled) dataset, we cannot apply this approach without manually labeling large amounts of data, and it is far from clear that class labels are the most effective mechanism for learning about similarity. We focus on one way to avoid this issue, which is to directly query an oracle for perceptual similarity information, as is done in Kumari et al. [2020], where triplets of the form "Is item B or item C more similar to item A?" are actively selected for learning an embedding of items. Active deep metric learning (DML) builds on this idea by finding the most informative queries to ask the oracle.

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

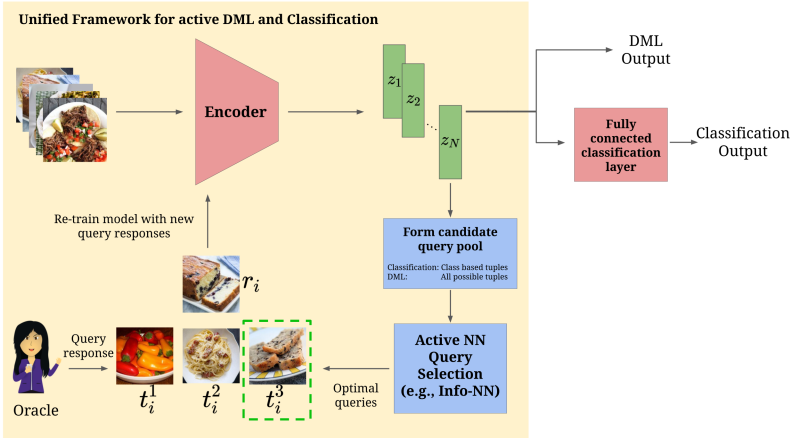

Figure 1: Visualization of the unified query solicitation framework with an example query. Candidate NN queries to be evaluated by the active NN query selection method are formed based on the setting (metric learning or classification). The oracle then responds to the most informative of these queries. In the case of metric learning, the response is utilized to place the reference closer to the similar item while for classification, the response is equivalent to a label corresponding to the reference (e.g., cake in this case).

While seemingly dissimilar from metric learning, contemporary classification relies on models (e.g., neural networks) with the ability to learn good representations of the data from training data. Active classification focuses on how to best solicit labels for unlabeled data points, with many modern approaches either implicitly or explicitly relying on representations learned by the model to determine the most informative label. Methods that use metrics based on the predicted class probabilities, such as uncertainty [Gal et al., 2017] or consistency [Gao et al., 2019], implicitly rely on such representations, whereas core-set based approaches [Sener and Savarese, 2017, Pinsler et al., 2019] directly use learned representations to select diverse samples. Thus, if we seek the most informative labels with respect to improving the learned representation of our classification model, the goals of active classification and active metric learning are aligned. Despite these commonalities and virtually identical learning frameworks for the two problems, to the best of our knowledge, there is no approach for query selection that is problem agnostic. In this paper, we present a unified framework, which is made feasible by a novel type of similarity query that applies to both DML and classification.

Specifically, we consider the *nearest neighbor* (NN) query, which, given a reference data point r, asks an *oracle* (e.g., a human expert) to select the most similar point from among a set of M alternatives $t^1, t^2, \ldots, t^M$. We denote this a length M NN query. With the goal of minimizing the required number of queries, we adapt an *active* query selection strategy to this query type. We take an information theoretic approach and estimate the gain in *mutual information* (conditioned on previous query responses) as the criteria for selecting the most informative query, an approach that we dub *Info-NN*.

To the best of our knowledge, we are the first to study this query type. Similar ideas have been explored before, such as using UI configurations to collect multiple triplets at once [Wilber et al., 2014], enforcing a class-similarity based quadruplet loss (one anchor, one positive point, two negative points) [Chen et al., 2017], and soliciting ranking queries [Canal et al., 2020]. Of these approaches, Chen et al. [2017] is the most similar, but NN queries are 1) not confined to a particular fixed length, and 2) not restricted to using class information. The first difference allows us to generalize to any classification problem and the second allows us to collect similarity information of items of the same class, or in cases where class labels are not available.

Contributions. Our main contributions are as follows.

1. We propose a novel type of similarity query, called the NN query (Sec. 3).
2. We re-cast active classification as finding the most informative NN query, which allows us to unify active classification and active DML under one framework. This framework is flexible enough to accommodate *any* active NN query selection method (Secs. 3.1 and 3.2).
3. We empirically validate [1] DML and classification performance using our unified framework and novel NN query selection method (Secs. 4.1 and 4.2).

2 BACKGROUND AND RELATED WORK

Metric learning. Learning embeddings from similarity-based comparisons has been previously studied in a variety of scenarios [Agarwal et al., 2007, Van Der Maaten and Weinberger, 2012, Terada and Luxburg, 2014, Amid

[1] Code available at `https://github.com/nnadagouda95/InfoNN`.

and Ukkonen, 2015, Kleindessner and von Luxburg, 2017, Karaletsos et al., 2015, Veit et al., 2017, Ma et al., 2019, Ghosh et al., 2019], spanning everything from utilizing non-metric multidimensional scaling (MDS) to accommodating noisy/corrupted triplets to examining deeper connections to kernels. The importance of learning meaningful embeddings is shown in various applications such as face verification [Sankaranarayanan et al., 2016], fine-grained classification [Wah et al., 2014], extracting usable information from crowd-sourcing [Kajino et al., 2012], and even fashion recommendations [Vasileva et al., 2018]. To complement these techniques, active query selection methods have been developed which examine uncertainty [Tamuz et al., 2011], exploit a low-dimensionality [Jamieson and Nowak, 2011], incorporate auxiliary features [Heim et al., 2015], and utilize Bayesian techniques [Lohaus et al., 2019]. However, all of these methods are designed for non-parametric embedding techniques (e.g., MDS) which cannot easily generate a corresponding embedding given new items.

More recently, deep metric learning (DML) has aimed to overcome these limitations [Kaya and Bilge, 2019]. DML trains a neural network to learn an embedded representation that respects similarity information. In particular, many triplet-based DML methods assume knowledge of class labels for items, and attempt to minimize inter-class distances while maximizing intra-class distances [Hoffer and Ailon, 2015, Ge, 2018, Chen et al., 2017]. Although class labels may not always be available, very few works consider the case of DML with perceptual similarity queries, especially in an active manner. Recently, active similarity query selection methods for DML that focus on finding batches of non-redundant *triplets* have been proposed [Kumari et al., 2020] by encouraging both informativeness (measured by entropy) and diversity (through a variety of heuristic approaches) within the selected batch. Our method adopts a similar framework as Kumari et al. [2020], but we utilize mutual information to find informative NN queries.

Classification. Traditionally, active learning has been used with support vector machines and Gaussian processes for image classification [Joshi et al., 2009, Tuia et al., 2009, Kapoor et al., 2007, Houlsby et al., 2011]. More recently, a variety of active methods based on uncertainty [Gal et al., 2017, Wang et al., 2016, Kirsch et al., 2019, Song et al., 2019], diversity [Sener and Savarese, 2017, Pinsler et al., 2019, Kirsch et al., 2019], and consistency [Gao et al., 2019] have been used for training deep neural networks in the supervised and semi-supervised classification settings. In these settings, the goal is to learn a model for predicting the class probabilities on a dataset consisting of points belonging to C classes. We assume access to an initial labeled and unlabeled set of samples. The samples from the unlabeled pool are iteratively evaluated for informativeness and labeled accordingly. Based on feedback from the oracle, we can learn a model in either supervised (using only the labeled data) or in semi-supervised (using all data) settings.

Some active classification approaches [Houlsby et al., 2011, Kirsch et al., 2019] consider mutual information between the model parameters and the predicted class probabilities to select the most informative samples, while some others [Sener and Savarese, 2017, Pinsler et al., 2019] follow a coreset based approach to select a subset of diverse samples such that the model learned with these samples best approximates the one learned on the entire data. In Sener and Savarese [2017], the authors use the features learned by the model to select the samples such that the maximum distance between an unlabeled sample and its nearest labeled sample is minimized. The method in Pinsler et al. [2019] chooses samples such that the model posterior with the selected samples best approximates the posterior with the complete data.

Our method derives inspiration from Houlsby et al. [2011], Gal et al. [2017], Kirsch et al. [2019] in using mutual information to evaluate informativeness, but we consider mutual information between the features and the predicted class probabilities computed based on the inter-sample distances in the feature space. Our approach is similar to the work of Sener and Savarese [2017] in that both use the Euclidean distances of the features learned by the neural network. However, their focus is only on coverage of the entire feature space, whereas we select samples with the goal of improving the learned embedding. Apart from these, there are a few works that focus on active discriminative representation learning. In Zhang et al. [2017], the authors propose an AL approach for text classification that selects instances containing words which are likely to most affect the embeddings by computing the *expected gradient length* with respect to the embeddings. A multi-armed bandit based method that uses networking data and learned representations for adaptively labeling informative nodes is suggested in Gao et al. [2018] to learn network representations. However, to the best of our knowledge, no framework of active representation learning has been applied to image classification before and none of the above methods propose a generalized querying strategy.

3 UNIFIED FRAMEWORK AND ACTIVE QUERY SELECTION

In this section, we provide an overview of our proposed generalized query framework. Specifically, we show that in any classification setting where a latent representation of the data is learned, querying an oracle for a class label can be re-formulated as soliciting the oracle's feedback for an NN query, allowing us to draw the connection to metric learning. We also present Info-NN, an active method of selecting NN queries using information theoretic criterion.

Formally, an NN query $Q_i = r_i \cup T_i$ of length M consists of a reference data point r_i and a set of data points $T_i =$

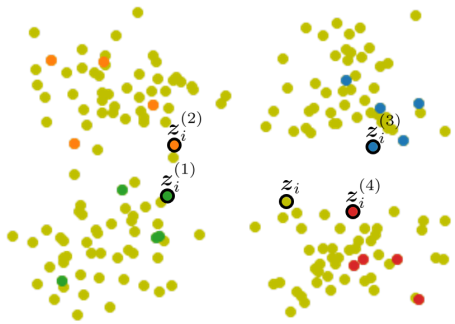

Figure 2: Example of an unlabeled $\boldsymbol{z}_i$ and the nearest labeled neighbors to $\boldsymbol{z}_i$ from each class: $\boldsymbol{z}_i^{(1)}, \boldsymbol{z}_i^{(2)}, \boldsymbol{z}_i^{(3)}, \boldsymbol{z}_i^{(4)}$. In this example, we might expect that the most likely label would be $y_i = 4$, which could be interpreted as a nearest neighbor query response (that $\boldsymbol{z}_i^{(4)}$ is the nearest neighbor).

$\{t_i^{(1)}, t_i^{(2)}, \ldots, t_i^{(M)}\}$, from which the oracle picks the point most similar to the reference r_i. Let $Y_i \in \{1, 2, \ldots, M\}$ be the random variable indicating the oracle's response to the ith query. When $Y_i = m$, this indicates that the oracle selected $t_i^{(m)} \in T_i$ as the most similar to the reference r_i. A visual example of the NN query can be found in Fig. 1.

3.1 CLASSIFICATION AS AN NN QUERY SELECTION PROBLEM

We approach AL for classification as one chiefly of selecting labels that will improve the feature representation, as most modern classification techniques (e.g., neural networks) can be interpreted as learning an embedding that enables simple linear classifiers to be effective. We do this via an analogy in which obtaining the class label for an unlabeled sample is equivalent to a particular NN query response.

Consider a dataset $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$ consisting of points belonging to C classes, $\{y_i\}_{i=1}^N \in \{1, 2, \ldots, C\}$. We assume access to an initial labeled, $\mathcal{L} = \{\boldsymbol{x}_i, y_i\}_{i=1}^j$ and unlabeled, $\mathcal{U} = \{\boldsymbol{x}_i\}_{i=j+1}^N$ set of samples. Let $\boldsymbol{Z} = \{\boldsymbol{z}_i\}_{i=1}^N$ represent initial estimates of the embeddings for the dataset according to a model learned on an initial set of labeled samples. Now suppose we want to choose a new point $\boldsymbol{x}_{j+1}$ from $\mathcal{U}$ whose label y_{j+1} we will obtain. For any $\boldsymbol{x}_i$ in $\mathcal{U}$, consider its embedding $\boldsymbol{z}_i$ in the feature space and the nearest neighbor to $\boldsymbol{z}_i$ from $\mathcal{L}$ for each class, i.e.,

$$\boldsymbol{z}_i^{(c)} = \underset{\boldsymbol{z}_\ell \in \mathcal{L}_c}{\arg\min} \|\boldsymbol{z}_\ell - \boldsymbol{z}_i\|_2,$$

where $\mathcal{L}_c = \{\boldsymbol{z}_\ell : (\boldsymbol{x}_\ell, y_\ell) \in \mathcal{L}, y_\ell = c\}$. An example of an unlabeled $\boldsymbol{z}_i$ and the nearest labeled neighbors to $\boldsymbol{z}_i$ from each class is illustrated in Fig. 2.

Note that if the embedding that we have learned does a reasonable job of representing similarity (as it pertains to the task of classification), then we would expect that the most likely label for $\boldsymbol{z}_i$ would correspond to the class c for which $\boldsymbol{z}_i^{(c)}$ is closest to $\boldsymbol{z}_i$. Thus, we can interpret the label y_i as a response to a length C nearest neighbor query in which $\boldsymbol{z}_i$ is the reference to which $\boldsymbol{z}_i^{(1)}, \boldsymbol{z}_i^{(2)}, \ldots, \boldsymbol{z}_i^{(C)}$ are compared. (For computational reasons, one may choose to not use all C nearest neighbors in practice.) Because this NN query response reveals information about the relative locations of items in the learned representation, retraining the classification model with the new oracle response should improve the representation. **This is the key idea behind our approach: select NN queries (or equivalently, points to label) that result in the best improvement of the embedding.**

3.2 UNIFIED FRAMEWORK FOR ACTIVE CLASSIFICATION AND METRIC LEARNING

This view of active classification gives rise to a unified framework which can be used in either active classification or active DML: from a pool of candidate NN queries, choose the most informative query to ask the oracle, then re-train the model to incorporate the newly acquired query response. **Despite each problem seemingly requiring fundamentally different oracle responses (similarity information vs. labels), both problems can be tackled utilizing NN queries, and thus, the same active query selection strategy.** The main difference is the pool of candidate queries. In active DML, we can query the oracle for similarity information about any set of items, whereas in active classification, the pool of candidate NN queries is restricted to queries that contain one item from every class. This pool of candidate queries is formed by setting every $\boldsymbol{z}_i$ corresponding to an $\boldsymbol{x}_i \in \mathcal{U}$ as the reference point, and finding (up to) C nearest neighbors of differing classes. **A critical feature of this unified framework is that it does not depend on which measure of "informativeness" is used. This allows for a practitioner to plug-in their desired active query selection criteria without making any modifications to the framework**, as depicted in Fig. 1. In our experiments, we select the queries that maximize *mutual information* for both active DML and classification experiments. In particular, we utilize two methods for computing mutual information, including a novel approach dubbed Info-NN.

3.3 ACTIVE QUERY SELECTION VIA INFO-NN

The main idea behind our selection strategy is to select queries that are maximally informative about the embedding while avoiding ones that do not provide new information. This goal is achieved by using mutual information between the embedding and a query as a measure of the informativeness of the query. Let $y^{i-1} = \{y_1, y_2, \ldots, y_{i-1}\}$ denote the set of all responses (true labels for the selected samples) obtained after $i-1$ queries. We denote Y_i to be the random variable corresponding to the oracle's response to query Q_i.

Algorithm 1 Info-NN-embedding

Input: Embedding $\boldsymbol{Z}$, candidate queries Q, num. samples n_s, variance σ^2

$I \leftarrow$ empty list of size $|Q|$ (Mutual information values for candidate queries)
$p_k, H_k \leftarrow$ empty lists of size $|Q|$
for $n = 1$ **to** n_s **do**
  $\tilde{\boldsymbol{Z}}_n \leftarrow \boldsymbol{Z} + \boldsymbol{G}$, elements of $\boldsymbol{G}$ drawn i.i.d from $\mathcal{N}(0, \sigma^2)$.
  for $Q_i \in Q$ **do**
    $r_i \leftarrow$ first element of Q_i
    $T_i \leftarrow Q_i \backslash \{r_i\}$
    $D_{Q_i} \leftarrow$ distance of every item in T_i to r_i in $\tilde{\boldsymbol{Z}}_n$
    $Y_i \leftarrow$ query response using D_{Q_i}
    $p_k[Q_i] \leftarrow p_k[Q_i] + p(Y_i|D_{Q_i})$ (cumulative probability)
    $H_k[Q_i] \leftarrow H_k[Q_i] + H[p(Y_i|D_{Q_i})]$ (cumulative entropy)
  end for
end for
for $Q_i \in Q$ **do**
  $I[Q_i] \leftarrow H\left[\frac{p_k[Q_i]}{n_s}\right] - \frac{H_k[Q_i]}{n_s}$
end for

Output: I

The Plackett-Luce (PL) model [Turner et al., 2018], which is an extension of the triplet model commonly used with similarity comparisons [Tamuz et al., 2011], is used to model the response. For an NN query of length M, the probability of $t_i^{(m)}$ being the nearest or the most similar point to r_i is modeled as

$$p(Y_i = m \mid D_{Q_i}) = \frac{(D_{im}^2 + \mu)^{-1}}{\sum_{j=1}^{M} (D_{ij}^2 + \mu)^{-1}} \tag{1}$$

where $D_{Q_i} \coloneqq \{D_{im} \ : \ t_i^{(m)} \in T_i\}$, D_{im} denotes the distance between the embeddings of r_i and $t_i^{(m)}$, and μ is a parameter set by the user. This model captures uncertainty in the oracle responses as well as uncertainty in our current estimate of the embedding (and hence distances). The parameter μ is indicative of our confidence in the distances. Note that even though we use this model in our query selection strategy, we do *not* require that query responses are generated according to the PL model.

Now consider the mutual information between the embedding $\boldsymbol{Z}$ and the response Y_i:

$$I(\boldsymbol{Z}; Y_i \mid y^{i-1}) = H[\boldsymbol{Z} \mid y^{i-1}] - \mathop{\mathbb{E}}_{Y_i}(H[\boldsymbol{Z} \mid Y_i, y^{i-1}]). \tag{2}$$

This quantity measures how much information the response to query Q_i would provide about the embedding, conditioned on the fact that we have already acquired the responses y^{i-1} to the previous queries. This is exactly what

Algorithm 2 Info-NN-distances

Input: Embedding $\boldsymbol{Z}$, candidate queries Q, num. samples n_s, variance σ^2

$I \leftarrow$ empty list of size $|Q|$ (Mutual information values for candidate queries)
for $Q_i \in Q$ **do**
  $r_i \leftarrow$ first element of Q_i
  $T_i \leftarrow Q_i \backslash \{r_i\}$
  $D_{Q_i} \leftarrow$ distance of every item in T_i to r_i in $\boldsymbol{Z}$
  $Y_i \leftarrow$ query response using D_{Q_i}
  $D_s \leftarrow n_s$ copies of $\mathcal{N}(D_{Q_i}, \sigma^2)$
  $I_{Q_i} \leftarrow H\left[\sum_{D \in D_s} \frac{(p(Y_i \mid D))}{n_s}\right] - \sum_{D \in D_s} \frac{H[p(Y_i \mid D)]}{n_s}$
  $I[Q_i] \leftarrow I_{Q_i}$
end for

Output: I

we would like to use to select informative queries, but computing this quantity in the above form is computationally expensive. To compute this in a naïve manner we would need to find the estimate of the embedding for every possible response to the query and compute the entropies of these estimates in the high dimensional embedding space. Fortunately, using an approach similar to Houlsby et al. [2011], we can use the symmetry of mutual information to re-write (2) as

$$I(Y_i; \boldsymbol{Z} \mid y^{i-1}) = H[Y_i \mid y^{i-1}] - \mathop{\mathbb{E}}_{\boldsymbol{Z}}(H[Y_i \mid \boldsymbol{Z}, y^{i-1}]). \tag{3}$$

We can now compute entropies in the response space, which is usually much smaller than the embedding space. This second form of mutual information also provides an interesting insight about the selection strategy. The first term, which denotes the entropy of the predicted response, encourages the selection of queries which are highly uncertain for the current estimate of the embedding. The second term denotes the expected entropy of the responses predicted by the individual samples from the distribution over the embedding estimate and encourages queries for which the individual samples are fairly confident. This simultaneously avoids the acquisition of redundant queries and queries for which the oracle response is likely to be uncertain.

Estimation of mutual information. Computing the mutual information as in (3) requires a probabilistic estimate of the embedding. However, in many learning scenarios, only point estimates are computed. We place the assumption of normal distributions and utilize two Monte Carlo sampling based methods for tractable computation of mutual information. The first method, which we refer to as *Info-NN-embedding*, assumes that the embedding values are normally distributed, with mean equal to the previous estimate of the embedding. With this assumption, we have a tractable means of computing the mutual information. We

can further increase computational efficiency by making the stronger assumption that inter-item distances in the embedding are normally distributed, with mean equal to the previous estimates of the distances. We refer to this approach as *Info-NN-distances*. In general, we use *Info-NN-distances* for experiments dealing with real-world data, and *Info-NN-embedding* for synthetic experiments. The two methods are presented in Alg. 1 and Alg. 2, respectively.

To enable efficient computation of mutual information in practice, we make a few more simplifying assumptions. We follow a similar approach as the one presented in Canal et al. [2020] and adapt it to NN queries. For Info-NN-embedding, the computation of $H[Y_i \mid y^{i-1}]$ and the corresponding assumptions are described below.

$$\begin{aligned} H[Y_i \mid y^{i-1}] &= H[\mathop{\mathbb{E}}_{\boldsymbol{Z}}(p(Y_i|\boldsymbol{Z}, y^{i-1})|y^{i-1})] \\ &= H[\mathop{\mathbb{E}}_{\boldsymbol{Z}}(p(Y_i|\boldsymbol{Z})|y^{i-1})] && \text{(I)} \\ &= H[\mathop{\mathbb{E}}_{\boldsymbol{Z}_{Q_i}}(p(Y_i|\boldsymbol{Z}_{Q_i})|y^{i-1})] && \text{(II)} \\ &= H[\mathop{\mathbb{E}}_{\boldsymbol{Z}_{Q_i}}(p(Y_i|\boldsymbol{Z}_{Q_i})|\boldsymbol{Z}^{i-1})] && \text{(III)} \\ &= H[\mathop{\mathbb{E}}_{\boldsymbol{Z}_{Q_i}\sim\mathcal{N}(\boldsymbol{Z}^{i-1}_{Q_i},\sigma^2_{i-1})}(p(Y_i|\boldsymbol{Z}_{Q_i}))] && \text{(IV)} \end{aligned}$$

(I) The response Y_i is independent of past responses y^{i-1}, when conditioned on $\boldsymbol{Z}$.

(II) The oracle's response conditioned on $\boldsymbol{Z}$, depends only on $\boldsymbol{Z}_{Q_i}$ - embeddings of the items involved in the query and is independent of the embeddings $\boldsymbol{Z}_{s\notin Q_i}$.

(III) $\boldsymbol{Z}$ is independent of y^{i-1}. given the previous estimate of the embedding $\boldsymbol{Z}^{i-1}$.

(IV) Conditioned on $\boldsymbol{Z}^{i-1}$, the $(a,b)^{th}$ entry of $\boldsymbol{Z}$, $\boldsymbol{Z}_{a,b}$, is distributed normally with mean $\boldsymbol{Z}^{i-1}_{a,b}$ and variance σ^2_{i-1}. We will slightly abuse notation, and write $\boldsymbol{Z} \sim \mathcal{N}(\boldsymbol{Z}^{i-1},\ \sigma^2_{i-1})$.

Following a similar process, we have

$$\mathop{\mathbb{E}}_{\boldsymbol{Z}}(H[Y_i \mid \boldsymbol{Z}, y^{i-1}]) = \mathop{\mathbb{E}}_{\boldsymbol{Z}_{Q_i}\sim\mathcal{N}(\boldsymbol{Z}^{i-1}_{Q_i},\sigma^2_{i-1})}(H[p(Y_i|\boldsymbol{Z}_{Q_i})]).$$

We can now utilize Monte Carlo sampling methods for estimating the entropies, as presented in Alg. 1

For Info-NN-distances, we make the same assumptions as above, except for (IV). Instead, we assume that the distances between data points are distributed normally with the mean for each pair set equal to the distance computed from the estimated embedding matrix. This assumption enables an efficient method of estimating the posterior distribution over the distances.

These, along with assumptions on conditional independence of the oracle responses and the distance estimates with respect to the previous responses y^{i-1}, enable efficient estimation of the mutual information. Specifically, the entropies in Eq. 3 can be computed as follows:

$$\begin{aligned} H[Y_i \mid y^{i-1}] &= H\left[\mathop{\mathbb{E}}_{\boldsymbol{Z}}\left(p(Y_i \mid \boldsymbol{Z}, y^{i-1}) \mid y^{i-1}\right)\right] \\ &= H\left[\mathop{\mathbb{E}}_{D_{Q_i}\sim\mathcal{N}^{i-1}_{Q_i}}\left(p(Y_i \mid D_{Q_i})\right)\right] \end{aligned} \tag{4}$$

and

$$\begin{aligned} \mathop{\mathbb{E}}_{\boldsymbol{Z}}(H[Y_i \mid \boldsymbol{Z}, y^{i-1}]) &= \mathop{\mathbb{E}}_{\boldsymbol{Z}}\left(H\left[p(Y_i \mid \boldsymbol{Z}, y^{i-1}) \mid y^{i-1}\right]\right) \\ &= \mathop{\mathbb{E}}_{D_{Q_i}\sim\mathcal{N}^{i-1}_{Q_i}}\left(H\left[p(Y_i \mid D_{Q_i})\right]\right), \end{aligned} \tag{5}$$

where $\mathcal{N}^{i-1}_{Q_i} := \mathcal{N}(D_{Q_i}, \sigma^2_{i-1})$. Due to this normal distribution assumption, the entropy computations in (4) and (5) are straightforward calculations. The full procedure is shown in Alg. 2

4 EXPERIMENTS

4.1 DEEP METRIC LEARNING

In this section, we directly query an oracle with NN queries and learn a similarity embedding from query responses using a Deep Metric Learning (DML) framework.

Active embedding framework. We utilize a neural network to learn an embedding that matches the oracle's responses to similarity queries. Because a length M NN query can be decomposed into $M-1$ triplets, we utilize a triplet loss [Weinberger et al., 2006]. We initialize our network with a random batch of triplets, then select batches of B queries, receive oracle responses to the selected queries, add the new queries to the pool of already answered queries, and re-train our network for 100 epochs using all prior query responses. For each experiment, we select a pool of $20,000$ training length-3 NN queries and $20,000$ testing length-3 NN queries from the set of all possible queries (decomposing NN queries into triplets for triplet based methods).

In scenarios where re-training the network many times is computationally expensive, batch methods that select multiple queries at once are preferable. We compare the performance of Info-NN to recently developed triplet batch methods [Kumari et al., 2020]. While Info-NN can identify informative queries, batches of the most informative queries at a fixed instance may result in poor diversity of queries, as the most informative queries often cover the same areas of the space. Therefore, we utilize a very simple batch extension for DML experiments. For a batch of B queries, we select the top $B' \leq B$ most informative queries, then select $B - B'$ queries uniformly at random from the query pool. We show that simply augmenting randomly selected queries with a set of the most informative queries can outperform methods designed specifically for batch query selection.

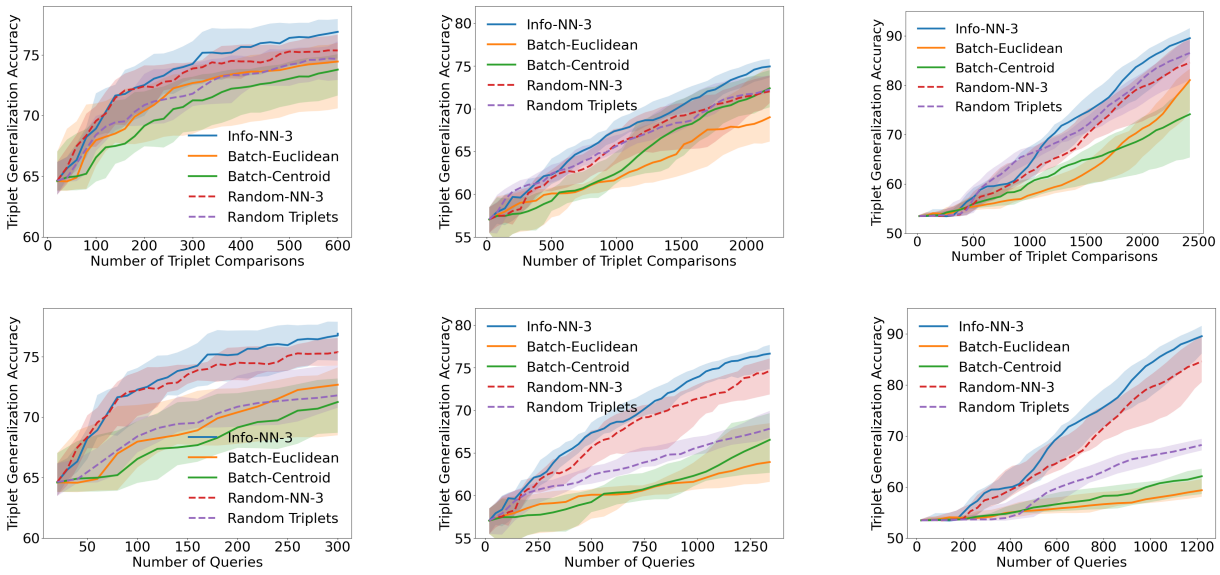

Figure 3: Per-triplet (top) and per-query (bottom) TGA comparison of Info-NN against active batch triplet methods and random queries on synthetic (left) and food (center), and Graduate Admission (right) datasets. Info-NN outperforms random and batch methods, and NN queries exhibit far superior per-query performance, requiring less interactions with the oracle.

Figure 4: Visualization of food embedding learned using queries selected with Info-NN, generated using t-SNE Maaten and Hinton [2008]. Similarly tasting objects are generally grouped together, such as vegetables (center) and fruits (top left)

In our experiments, *Info-NN-M* means the batch variant of Info-NN described above was used to select NN queries of length M, while *Batch-Euclidean/Centroid* indicate methods proposed in Kumari et al. [2020]. Finally, *Random* means the query type (NN or triplet) was constructed by selecting queries uniformly at random from the training set. Precise experimental details can be found in the appendix.

Datasets and evaluation metrics. We test our active embedding technique on a variety of datasets:

- **Synthetic Mahalanobis Dataset:** We generate $N = 100$ items of dimension $D = 10$ from a standard normal distribution. The oracle makes perception judgements based on some randomly generated Mahalanobis metric. We introduce artificial noise by corrupting 25% of all training queries to assess the robustness of our embedding method. We collect batches of size $B = 10$. Info-NN-embedding is used in these experiments.
- **Food73 Dataset:** This dataset contains $72,148$ crowdsourced triplets gathered for 73 different food items [Wilber et al., 2014]. We utilize 6 L1 normalized features (bitterness, saltiness, savoriness, spiciness, sourness, and sweetness) for each food item and form

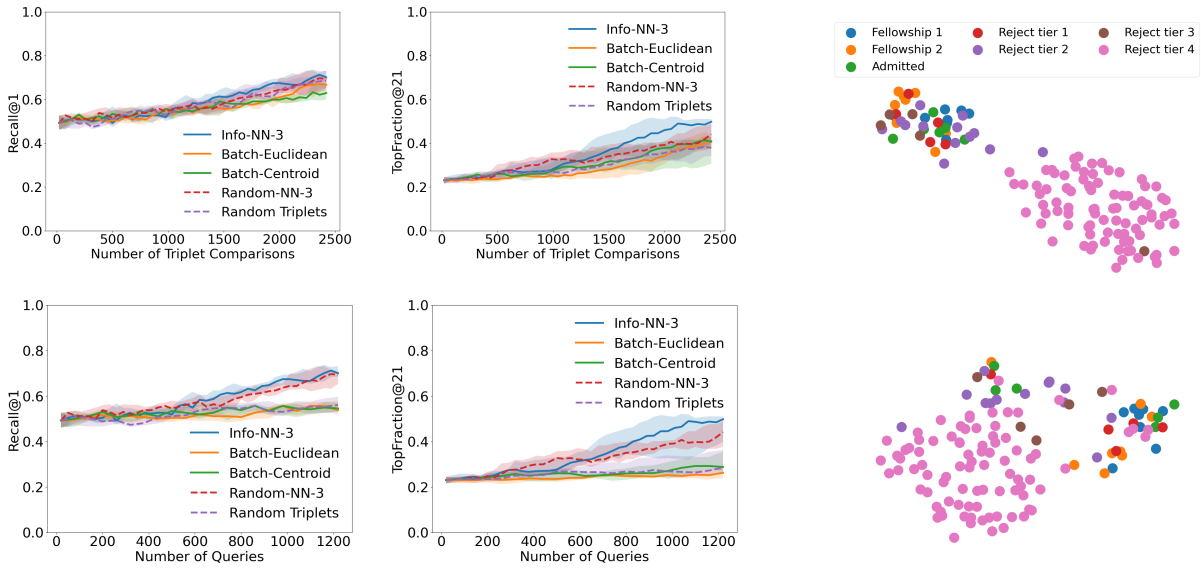

Figure 5: Per-triplet (top) and per-query (bottom) comparison for Info-NN against other methods Recall@1 (left) and TopFraction@21 (center). NN queries result in objects of the same class to be more nearby and group admitted students together, with Info-NN exhibiting the best performance of all methods tested. Visualization of embedding learned using Info-NN (right-top) and Batch-Centroid (right-bottom), generated using t-SNE [Maaten and Hinton, 2008]. Info-NN groups more highly rated candidates closer together.

$1,047,251$ length 3 NN queries from the collected triplets. The collected triplets, and, as a result, the formed NN queries contain inconsistencies. We collect batches of size $B = 30$. Info-NN-distances is used in these experiments.

- **Ranked Graduate Admissions Dataset:** We obtained partially ranked lists of 133 Ph.D. applicants to Georgia Tech School of Electrical and Computer Engineering. The top 22 candidates were accepted for admission, with the top 18 candidates individually ranked and the the rest of the candidates sorted into 5 tiers of varying sizes. Candidates fall into one of 7 classes: Admitted with fellowship 1, admitted with fellowship 2, admitted without fellowship, reject (sorted into 4 tiers). For each candidate, we have access to 25 features including GPA, letters of recommendation scores, and external fellowship application status. We form triplets across among the ranked candidates and between candidates of different tiers, resulting in $434,470$ triplets and $21,634,487$ length 3 NN queries, and randomly corrupt 25% of all queries to assess robustness. We collect batches of size $B = 30$. Info-NN-distances is used in these experiments.

To measure the performance of our embedding learning algorithm, we use *triplet generalization accuracy*, which records the fraction of test triplets whose ordering is consistent with the learned embedding. Furthermore, for the Graduate Admissions Dataset, because we have access to class labels, we record Recall@K. Furthermore, to get a sense of how the algorithms group the admitted students, we compute TopFraction@K, which denotes the fraction of the K nearest neighbors of the top 22 (admitted) students that are admitted students. Because NN queries can be decomposed into triplets, we compare performance against triplet-based methods on both a *per-triplet* basis and a *per-query* basis (number of queries posed to the oracle). We report the median and 25% and 75% quantile over 20 (synthetic), 10 (food), and 10 (admissions) trials.

Experiment results. As seen in Fig. 3, both versions of Info-NN are able to outperform recent methods developed specifically for batch query selection on both synthetic and challenging real-world datasets on both a per-triplet and a per-query basis. **This also demonstrates the flexibility of the unified framework; multiple active query selection methods can be plugged into the framework with consistently strong performance.** Furthermore, there seems to be minimal performance difference in selecting random NN queries vs. random triplets on a per-triplet basis, but using NN queries requires far fewer interactions with the oracle. From these experiments, it appears that the methods in Kumari et al. [2020] require more of a "warm up" to catch up to random query performance, whereas Info-NN can consistently outperform random. Inspecting the visualization in Fig. 4 of the learned food embedding also reveals that the embedding learned using Info-NN nicely separates savory foods from sweet foods, and can even group together similar foods, such as vegetables and fruits. Beyond triplet generalization accuracy, we can see in Fig. 5 that Info-NN is able

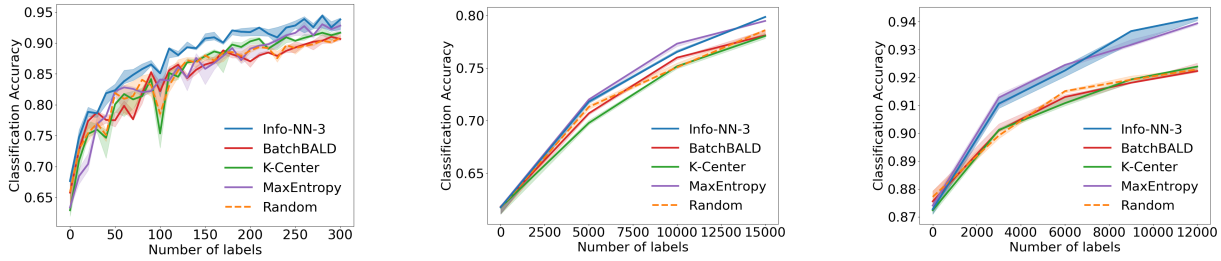

Figure 6: Active classification performance comparison on MNIST (left), CIFAR-10 (center) and SVHN (right) datasets.

to outperform the same methods on both a per-triplet and per-query basis in Recall@1 and TopFraction@21, which suggests that Info-NN is more capable of grouping admitted students together. This can be visualized in Fig. 5, where top-rated students are more clearly grouped together in the embedding learned using Info-NN compared to the embedding learned with Batch-Centroid. Results for varying values of K for Recall@K and TopFraction@K can be found in the appendix. We also note that Info-NN can be utilized in non-DML settings, such as using MDS Tamuz et al. [2011], and performs on par with more complex ranking queries (see appendix).

4.2 CLASSIFICATION

We perform experiments on active image classification in a supervised setting, using NN queries to acquire labels iteratively. Info-NN-distances is used in these experiments.

Label selection and experimental framework. To select samples using Info-NN, for every unlabeled sample, we form the corresponding nearest neighbor query and compute an estimate of the information gain provided by that query. We then request a label for the unlabeled sample corresponding to the most informative query. A simple batch extension of our query acquisition strategy, which performs a K-Means clustering of the unlabeled samples in the embedding space and selects the most informative samples from every cluster, is used in the experiments. *Info-NN-M* means the batch variant of Info-NN was used to select NN queries of length M. We use Euclidean distances between the features learned by the last hidden layer to compute distances for the probability model. We experimented with the length of the queries and illustrate plots for the best performing values. We plot the median of the accuracy values along with the 25% and 75% quantiles over 3 trials. More details can be found in the appendix.

We conduct experiments on **MNIST** [LeCun et al., 1998], **CIFAR-**10 [Krizhevsky et al., 2009] and **SVHN** [Netzer et al., 2011] datasets using CNNs to demonstrate the performance of our active learning method with supervised classification. The experiments on MNIST have an initial balanced labeled set of 30 samples, 3 from every class, chosen at random and an acquisition batch of size 10 is used. For CIFAR-10 and SVHN, we start with initial balanced labeled sets of $5,000$ and $3,000$ respectively, and acquire batches of size $5,000$ and $3,000$. We compare the performance of Info-NN with *BatchBALD* [Kirsch et al., 2019], *K-Center* [Sener and Savarese, 2017], *MaxEntropy*, and *Random* methods.

Experiment results. While our method outperforms all the baselines on MNIST, on CIFAR-10 and SVHN, it performs almost on par with MaxEntropy. A possible explanation here is the embeddings learned might be of lower quality. The embeddings directly affect the informativeness and diversity measures, thereby impacting the quality of labels chosen. Using pre-trained networks instead could result in better quality embeddings. We do not explore this direction here but is an interesting avenue for future work.

5 CONCLUSION

In this paper, we introduce a generalized similarity based active learning framework for selecting informative queries for both metric learning and classification. In a deep metric learning setting, we demonstrated that our framework is capable of outperforming recently developed methods for selecting batches of triplets on a both per-triplet and per-query basis. For classification, our framework for active label selection resulted in a better performance compared to the baselines. As shown by strong empirical performance, this framework marks the first step in developing generalized active learning methods capable of performing well in multiple problem areas. An avenue for future work is studying the proposed active label selection method for regression and improving the generalizability of the framework.

Acknowledgements

This work was supported, in part, by DARPA grant FA8750-19-C020 and NSF grant CCF-2107455.

References

Sameer Agarwal, Josh Wills, Lawrence Cayton, Gert Lanckriet, David Kriegman, and Serge Belongie. Generalized non-metric multidimensional scaling. In *Artificial Intelligence and Statistics*, pages 11–18, 2007.

Ehsan Amid and Antti Ukkonen. Multiview triplet embedding: Learning attributes in multiple maps. In *International Conference on Machine Learning*, pages 1472–1480. PMLR, 2015.

Gregory Canal, Stefano Fenu, and Christopher Rozell. Active ordinal querying for tuplewise similarity learning. In *AAAI*, pages 3332–3340, 2020.

Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 403–412, 2017.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International Conference on Machine Learning*, pages 1183–1192. PMLR, 2017.

Li Gao, Hong Yang, Chuan Zhou, Jia Wu, Shirui Pan, and Yue Hu. Active discriminative network representation learning. In *IJCAI International Joint Conference on Artificial Intelligence*, 2018.

Mingfei Gao, Zizhao Zhang, Guo Yu, Sercan O Arik, Larry S Davis, and Tomas Pfister. Consistency-based semi-supervised active learning: Towards minimizing labeling cost. *arXiv preprint arXiv:1910.07153*, 2019.

Weifeng Ge. Deep metric learning with hierarchical triplet loss. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 269–285, 2018.

Nikhil Ghosh, Yuxin Chen, and Yisong Yue. Landmark ordinal embedding. In *Advances in Neural Information Processing Systems*, pages 11506–11515, 2019.

Eric Heim, Matthew Berger, Lee Seversky, and Milos Hauskrecht. Active perceptual similarity modeling with auxiliary information. *arXiv preprint arXiv:1511.02254*, 2015.

Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

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

Kevin G Jamieson and Robert D Nowak. Low-dimensional embedding using adaptively selected ordinal data. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1077–1084. IEEE, 2011.

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

Hiroshi Kajino, Yuta Tsuboi, and Hisashi Kashima. A convex formulation for learning from crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2012.

Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Active learning with gaussian processes for object categorization. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

Theofanis Karaletsos, Serge Belongie, and Gunnar Rätsch. Bayesian representation learning with oracle constraints. *arXiv preprint arXiv:1506.05011*, 2015.

Mahmut Kaya and Hasan Şakir Bilge. Deep metric learning: A survey. *Symmetry*, 11(9):1066, 2019.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. In *Advances in Neural Information Processing Systems*, pages 7026–7037, 2019.

Matthäus Kleindessner and Ulrike von Luxburg. Kernel functions based on triplet comparisons. In *Advances in neural information processing systems*, pages 6807–6817, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Priyadarshini Kumari, Ritesh Goru, Siddhartha Chaudhuri, and Subhasis Chaudhuri. Batch decorrelation for active metric learning. In *IJCAI*, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Michael Lohaus, Philipp Hennig, and Ulrike von Luxburg. Uncertainty estimates for ordinal embeddings. *arXiv preprint arXiv:1906.11655*, 2019.

Ke Ma, Qianqian Xu, and Xiaochun Cao. Robust ordinal embedding from contaminated relative comparisons. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7908–7915, 2019.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9 (Nov):2579–2605, 2008.

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

Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. Bayesian batch active learning as sparse subset approximation. In *Advances in Neural Information Processing Systems*, pages 6359–6370, 2019.

Swami Sankaranarayanan, Azadeh Alavi, Carlos D Castillo, and Rama Chellappa. Triplet probabilistic embedding for face verification and clustering. In *2016 IEEE 8th international conference on biometrics theory, applications and systems (BTAS)*, pages 1–8. IEEE, 2016.

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

Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Shuang Song, David Berthelot, and Afshin Rostamizadeh. Combining mixmatch and active learning for better accuracy with fewer labels. *arXiv preprint arXiv:1912.00594*, 2019.

Omer Tamuz, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai. Adaptively learning the crowd kernel. *arXiv preprint arXiv:1105.1033*, 2011.

Yoshikazu Terada and Ulrike Luxburg. Local ordinal embedding. In *International Conference on Machine Learning*, pages 847–855, 2014.

Devis Tuia, Frédéric Ratle, Fabio Pacifici, Mikhail F Kanevski, and William J Emery. Active learning methods for remote sensing image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232, 2009.

Heather Turner, Jacob van Etten, David Firth, and Ioannis Kosmidis. Introduction to plackettluce, 2018.

Laurens Van Der Maaten and Kilian Weinberger. Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2012.

Mariya I Vasileva, Bryan A Plummer, Krishna Dusad, Shreya Rajpal, Ranjitha Kumar, and David Forsyth. Learning type-aware embeddings for fashion compatibility. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 390–405, 2018.

Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Catherine Wah, Grant Van Horn, Steve Branson, Subhransu Maji, Pietro Perona, and Serge Belongie. Similarity comparisons for interactive fine-grained categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

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

Kilian Q Weinberger, John Blitzer, and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2006.

Michael Wilber, Iljung Kwak, and Serge Belongie. Cost-effective hits for relative similarity comparisons. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 2014.

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