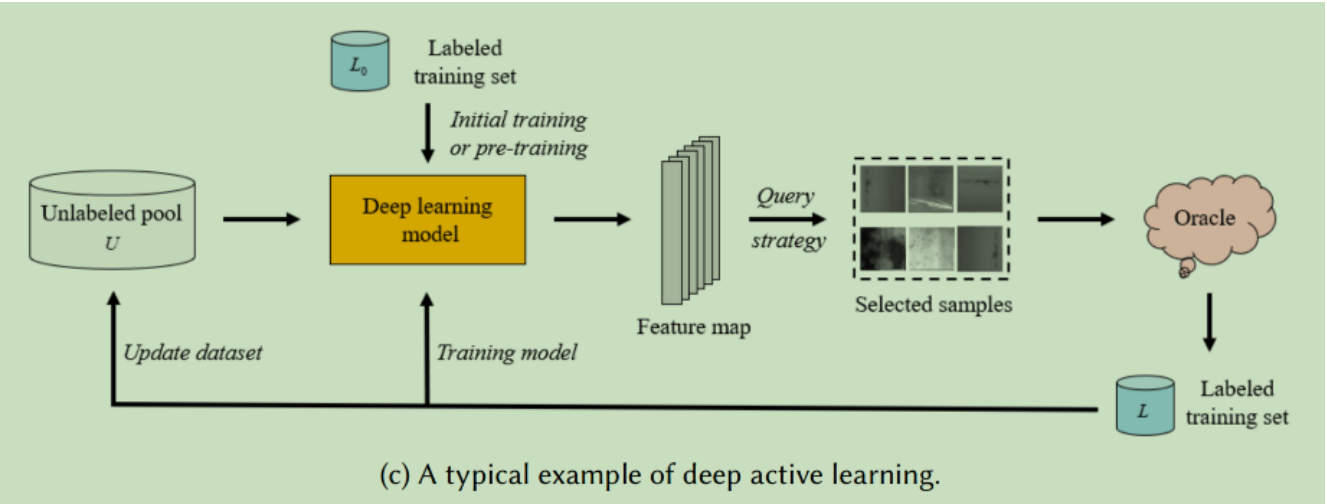
A Survey of Deep Active Learning

Introduction

- DL has strong learning capabilities due to its complex structure, but this also means that DL requires a large number of labeled samples to complete the corresponding training
- AL focuses on the study of data sets, and it is also known as query learning
- **AL assumes that different samples in the same data set have different values for the update of the current model, and tries to select the samples with the highest value to construct the training set**
- However, the classic AL algorithm also finds it difficult to handle high-dimensional data
- Beginning in 2017, DL gradually shifted from the initial feature extraction automation to the automation of model architecture design
- AL approaches can be divided into membership query synthesis, stream-based selective sampling and pool-based AL from application scenarios
- Moreover, the key difference between stream-based selective sampling and pool-based sampling is that the former makes an independent judgment on whether each sample in the data stream needs to query the labels of unlabeled samples, while the latter chooses the best query sample based on the evaluation and ranking of the entire dataset
-



(c) A typical example of deep active learning.

- Stream-based selective sampling is mainly aimed at the application scenarios of small mobile devices that require timeliness, because these small devices often have limited storage and computing capabilities
- **The more common pool-based sampling strategy in the paper related to AL research is more suitable for large devices with sufficient computing and storage resources**
- The main query strategies include the
    - Uncertainty-based approach
    - Diversity-based approach
    - Expected model change

The Necessity And Challenge Of Combining DL And Al

- Model uncertainty in Deep Learning
- Insufficient data for labeled samples
- Processing pipeline inconsistency

Deep Active Learning

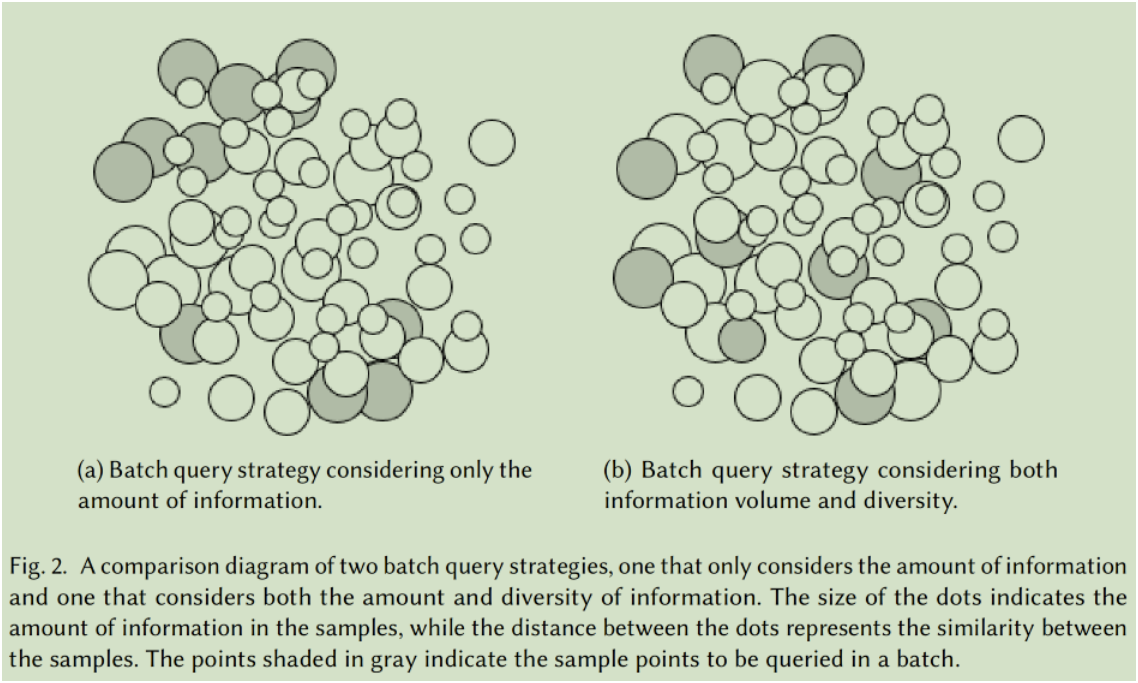
Query Strategy Optimization in DeepAL

- In the pool-based method, we define $U^n = \{\mathcal{X}, \mathcal{Y}\}$ as an unlabeled dataset with n samples; here, $\mathcal{X}$ is the sample space, $\mathcal{Y}$ is the label space, and $P(x, y)$ is a potential distribution, where $x \in \mathcal{X}, y \in \mathcal{Y}$. $L^m = \{X, Y\}$ is the current labeled training set with m samples, where $\mathrm{x} \in X, \mathrm{y} \in Y$. Under the standard supervision environment of DeepAL, our main goal is to design a query strategy Q, $U^n \xrightarrow{Q} L^m$, using the deep model $f \in \mathcal{F}, f : \mathcal{X} \rightarrow \mathcal{Y}$. The optimization problem of DeepAL in a supervised environment can be expressed as follows:

$$\arg\min_{L^m \subseteq U^n, (\mathrm{x},\mathrm{y}) \in L^m, (x,y) \in U^n} \mathbb{E}_{(x,y)}[\ell(f(\mathrm{x}), \mathrm{y})], \tag{1}$$

where $\ell(\cdot) \in \mathcal{R}^+$ is the given loss equation, and we expect that $m \ll n$. Our goal is to make m as small as possible while ensuring a predetermined level of accuracy. Therefore, the query strategy Q in DeepAL is crucial to reduce the labeling cost. Next, we will conduct a comprehensive and systematic review of DeepAL's query strategy from the following five aspects.

- Batch Mode DeepAL (BMDAL)
  - The batch-based query strategy is the foundation of DeepAL
  - The one-by-one sample query strategy in traditional AL is inefficient and not applicable to DeepAL, so it is replaced by batch-based query strategy
    - In traditional AL, most algorithms use a one-by-one query method, which leads to frequent training of the learning model but little change in the training data
  - **需要确认待标数据之间的互信息，这样可以考虑到分布情况**
- Uncertainty-based and Hybrid Query Strategies
  - Uncertainty-based query strategy refers to the model based on sample uncertainty ranking to select the sample to be queried. The greater the uncertainty of the sample, the easier it is to be selected
  - However, this is likely to ignore the relationship between samples. Therefore, the method that considers multiple sample attributes is called the hybrid query strategy
  - For example, in the margin sampling, margin M is defined as the difference between the predicted highest probability and the predicted second highest probability of an sample as follows: $M = P(y1 \mid x) - P(y2 \mid x)$, where $y1$ and $y2$ are the first and second most probable labels predicted for the sample x under the current model

- A feasible strategy would thus be to use a hybrid query strategy in a batch query, taking into account both the information volume and diversity of samples in either an explicit or implicit manner
- The performance of early Batch Mode Active Learning (BMAL) algorithms are often excessively reliant on the measurement of similarity between samples. In addition, these algorithms are often only good at exploitation (learners tend to focus only on samples near the current decision boundary, corresponding to high-information query strategies), meaning that the samples in the query batch sample set cannot represent the true data distribution of the feature space (due to the insufficient diversity of batch sample sets)
  - Exploration-P
    - Using a deep neural network to learn the feature representation of the samples, then explicitly calculates the similarity between the samples
  - DMBAL (Diverse Mini-Batch Active Learning)
    - Adding informativeness to the optimization goal of K-means by weight, and further presents an in-depth study of a hybrid query strategy that considers the sample information volume and diversity under the mini-batch sample query setting
  - 

    (a) Batch query strategy considering only the amount of information.

    (b) Batch query strategy considering both information volume and diversity.

    Fig. 2. A comparison diagram of two batch query strategies, one that only considers the amount of information and one that considers both the amount and diversity of information. The size of the dots indicates the amount of information in the samples, while the distance between the dots represents the similarity between the samples. The points shaded in gray indicate the sample points to be queried in a batch.
- Deep Bayesian Active Learning (DBAL)
  - Traditional DL methods rarely represent such model uncertainty
  - $f(\mathrm{x};\theta)$, $p(\theta)$ is a prior on the parameter space θ (usually Gaussian), and the likelihood $p(\mathrm{y}=c|\mathrm{x},\theta)$ is usually given by $softmax(f(\mathrm{x};\theta))$. Our goal is to obtain the posterior distribution over θ, as follows:

    $$p(\theta|X,Y)=\frac{p(Y|X,\theta)p(\theta)}{p(Y|X)}. \tag{8}$$

    For a given new data point x^*, $\hat{\mathrm{y}}$ is predicted by:

    $$p\left(\hat{\mathrm{y}}|\mathrm{x}^*,X,Y\right)=\int p\left(\hat{\mathrm{y}}|\mathrm{x},\theta\right)p(\theta|X,Y)d\theta=\mathbb{E}_{\theta\sim p(\theta|X,Y)}[f(\mathrm{x};\theta)]. \tag{9}$$
- Density-based Methods
  - The density-based method is a query strategy that attempts to find a core subset representing the distribution of the entire dataset from the perspective of the dataset to reduce the cost of

annotation
    - **This idea is mainly inspired by the compression idea of the core set dataset and attempts to use the core set to represent the distribution of the feature space of the entire original dataset, thereby reducing the labeling cost of AL**
    - The Core-set approach attempts to solve this problem by constructing a core subset
    - Previous core-set-based methods often simply try to query data points as far as possible to cover all points of the data manifold without considering the density, which results in the queried data points overly representing sample points from manifold sparse areas
- Automated Design of DeepAL

Data Expansion of Labeled Samples in DeepAL

- CEAL (Cost-Effective Active Learning) enriches the training set by assigning pseudo-labels to samples with high confidence in model prediction in addition to the labeled dataset sampled by the query strategy
- GNN 的方式也比较流行

DeepAL Generic Framework

- CEAL (Cost-Effective Active Learning) enriches the training set by assigning pseudo-labels to samples with high confidence in model prediction in addition to the labeled dataset sampled by the query strategy
- GNN 的方式也比较流行
- CEAL is one of the first works to combine AL and DL in order to solve the problem of depth image classification
- CEAL merges deep convolutional neural networks into AL, and consequently proposes a novel DeepAL framework. It sends samples from the unlabeled dataset to the CNN step by step, after which the CNN classifier outputs two types of samples: a small number of uncertain samples and a large number of samples with high prediction confidence. A small number of uncertain samples are labeled by the oracle, and the CNN classifier is used to automatically assign pseudo-labels to a large number of high-prediction-confidence samples

DeepAL Stopping Strategy

- Most DeepALs often use the predefined stopping criterion, and when the criterion is satisfied, they stop querying labels from the oracle
    - Maximum number of iterations
    - Minimum threshold for changing classification accuracy
    - Minimum number of labeled samples
    - Expected accuracy value
- Although these stopping criteria are simple, these predefined stopping criteria are likely to cause DeepAL to fail to achieve optimal performance

Application Of DeepAL in Fields Such As Vision And NLP

Visual Data Processing

- Image classification and recognition
    - That DeepAL faces in the field of image vision tasks is that of how to efficiently query samples of high-dimensional data (an area in which traditional AL performs poorly) and obtain satisfactory performance at the smallest possible labeling cost
- Object detection and semantic segmentation
- Video processing

NLP

- Machine translation
    - [Active Learning for Neural Machine Translation] proposes to use the AL framework to select information source sentences to construct a parallel corpus. It proposes two effective sentence selection methods for AL: selection based on semantic similarity and decoder probability
- Text Classification
    - [Active Discriminative Text Representation Learning] focuses on selecting those samples that have the greatest impact on the embedding space
        - It proposes a method for sentence classification that selects instances containing words whose embeddings are likely to be updated with the greatest magnitude, thereby rapidly learning discriminative, task-specific embeddings
        - 上面这个描述可以用关键词的 embedding 替代
    - [Sampling Bias in Deep Active Classification: An Empirical Study] focuses on the problem ofsampling bias in deep active classification and apply active text classification on the large-scale text corpora of
- Semantic Analysis
- Information Extraction
- Question Answering

Discussion And Future Directions

- DeepAL combines the common advantages of DL and AL: it inherits not only DL's ability to process high-dimensional image data and conduct automatic feature extraction but also AL's potential to effectively reduce annotation costs