Active Learning for Image Classification

Sergey Ilizirov 211624788
Michal Barsht 212673784
Amir Steiner 302342498

November 2024

1 Abstract

Active learning has emerged as a critical approach to mitigate the dependency of deep learning models on large labeled datasets by intelligently selecting the most informative samples for labeling. In this paper, we propose a new active learning strategy tailored for image classification, addressing the challenges of limited labeling budgets in advanced fields like medicine that require highly trained professionals for labeling. Our method integrates an uncertainty-based sampling criterion with the feature representations of images, rather than selecting the top K samples with the highest uncertainty, we extract and analyze their CNN-based feature representations to identify common feature patterns and use them to choose the images to label. Experimental results on standard image classification benchmark data sets reveal that our approach is competitive with existing active learning methods.

2 Introduction and Previous Works

Deep learning methods have demonstrated remarkable success in image classification tasks. Yet, their reliance on large labeled data sets remains a significant bottleneck, especially in fields where annotation requires specialized expertise, such as medical imaging and industrial diagnostics. Active learning has emerged as a practical solution to this challenge, aiming to minimize labeling costs by intelligently selecting the most informative samples for annotation. By focusing labeling efforts on these samples, active learning methods can achieve competitive model performance with significantly fewer labeled instances. Traditional active learning strategies, such as uncertainty sampling, random sampling, and query-by-committee (QBC), have been widely studied. Uncertainty sampling, a foundational technique introduced by Lewis and Gale (1994), prioritizes samples for which the model's predictions exhibit the highest uncertainty, often measured through metrics like entropy or prediction probability margins. Settles (2009) provided a comprehensive overview of active learning methods, emphasizing the role of uncertainty-based strategies in reducing annotation costs. Furthermore, pretrained CNN features have proven highly effective in unsupervised learning, with studies like Guérin et al. (2018) demonstrating that features from models trained on datasets like ImageNet outperform traditional clustering methods for unsupervised classification. This aligns with the active learning paradigm, where efficient use of feature representations can guide the selection of informative samples, as explored in this paper.

3 Active Learning Methods

3.1 System Overview

Let $M \in \mathbb{R}^{(H,W)}$ be the set of available images, $M_L \subseteq M$ the labeled images, $M_U \subseteq M$ the unlabeled images, K the number of images that the data selection algorithm A, is selecting for the oracle to label in each iteration $t \in T$ where T is the total number of iterations. The proposed active learning system is an iterative framework designed to optimize the performance of an image classification model while minimizing the labeling effort. The workflow comprises the following steps:

1. Initialization: A subset of the dataset M, denoted as $M_L \subseteq M$, is randomly sampled and labeled. This serves as the initial labeled data pool, while the remaining images form the unlabeled data pool, $M_U = M \setminus M_L$.

2. Model Training: The image classification model f_θ, parameterized by $\boldsymbol{\theta}$, in this project, the focus is on the CNN model described later, and is trained using the labeled data M_L. The objective is to minimize the cross-entropy loss.

3. Uncertainty Estimation: For each unlabeled sample $x \in M_U$, an uncertainty score is computed based on the model's predictive distribution.

4. Data Selection Algorithm: Using the uncertainty scores and feature representations $\boldsymbol{f}(x)$ extracted from the model, a data selection algorithm A that will be presented further, identifies a subset $L \subseteq M_U$ of size K. These samples are the most informative for improving the model.

5. Expert Labeling: The selected samples L are labeled by an expert/oracle and added to the labeled data pool:

$$M_L \leftarrow M_L \cup L, \quad M_U \leftarrow M_U \setminus L.$$

6. Model Update: The model f_θ is fine-tuned using the updated labeled data pool M_L.

7. Iteration: Steps 3–6 are repeated for $t \in \{1, \dots, T\}$, where T is the total number of iterations. The process terminates either when a predefined budget limit is reached or when the model's performance converges.

This iterative mechanism ensures that the labeling effort is focused on the most informative samples, leading to a more efficient utilization of the labeling budget while progressively enhancing model performance.

3.2 Uncertainty Estimation

Uncertainty estimation is crucial in active learning, guiding the selection of samples the model finds most challenging. We consider three common methods: Least Confidence, Least Margin, and Entropy, each leveraging the output probabilities from the softmax layer of a CNN to quantify uncertainty.

1. Least Confidence: Measures uncertainty as the negative of the highest predicted probability:

$$\mathrm{lc}_i = -\max_j p_j(X_i, y_i)$$

where $p_j(X_i, y_i)$ is the predicted probability of the j-th class for the i-th instance. The smaller the maximum predicted probability, the higher the uncertainty of the prediction.

2. Least Margin: Captures uncertainty as the negative margin between the top two predicted probabilities:

$$\mathrm{lm}_i = -\left(p_{j_1}(X_i, y_i) - p_{j_2}(X_i, y_i)\right)$$

where j_1 and j_2 are the classes with the highest and second-highest predicted probabilities, respectively.

3. Entropy: Quantifies uncertainty based on the distribution spread across all classes:

$$\mathrm{en}_i = -\sum_{j=1}^{C} p_j(X_i, y_i) \log p_j(X_i, y_i)$$

where C is the total number of classes. For our study, we have chosen to use the **Entropy** method. This decision was made after considering the trade-offs in complexity and computation.

3.3 Data Selection Algorithm

The fundamental strategy in active learning that incorporates uncertainty estimation involves selecting K samples from the unlabeled dataset M_U, which are deemed most uncertain based on a predefined uncertainty criterion. These samples, denoted by $\mathcal{S}_u \subset M_U$, are then presented to an expert for labeling. In our proposed method, we enhance this strategy by utilizing visual information embedded in the feature representations of the images, extracted using a Convolutional Neural Network (CNN). Let $\alpha \in (0, 1]$ denote the proportion of samples selected purely based on uncertainty. We define the set of highly uncertain samples as:

$$\mathcal{S}_u = \{x_i \in M_U \mid u(x_i) \text{ is among the top-}\alpha K \text{ highest uncertainties}\},$$

where $u(x_i)$ quantifies the uncertainty of sample x_i (e.g., entropy, margin, or least confidence). Next, we want to find the features that characterize the most uncertain samples and choose the remaining images for labeling from images with those features, for this purpose, we incorporate a clustering step. Let $\mathbf{F} = \{\mathbf{f}_i \mid x_i \in \mathcal{S}_u\}$ represent the feature vectors of the uncertain samples, where $\mathbf{f}_i \in \mathbb{R}^d$ is the feature vector of sample x_i, extracted from the final convolutional layer of the CNN before the fully connected layer. Using the K-means clustering algorithm, we partition $\mathbf{F}$ into n clusters, denoted as $\{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_n\}$. The additional $(1-\alpha)K$ samples, denoted $\mathcal{S}_c$, are selected as follows: for each image $x_i \in M_U \setminus \mathcal{S}_u$ with feature vector $\mathbf{f}_i$, we determine the cluster $\mathcal{C}_j$ with the closest centroid $\mathbf{c}_j$, where:

$$j^* = \arg\min_j \|\mathbf{f}_i - \mathbf{c}_j\|_2^2.$$

The distance of x_i to its closest cluster centroid is denoted as:

$$d_i = \min_j \|\mathbf{f}_i - \mathbf{c}_j\|_2^2.$$

Then we select the images that have the least distance to their assigned cluster, reflecting the similarity of their feature representations to those of the images for which the model exhibited the highest uncertainty. This approach aims to assist the model in addressing the features that cause the most confusion, thereby enhancing its overall clarity and performance. Finally, the labeled set is updated as:

$$M_L \leftarrow M_L \cup (\mathcal{S}_u \cup \mathcal{S}_c),$$

and the model is fine-tuned using the updated labeled set M_L. To further enhance performance, the CNN used for feature extraction can be a pre-trained model (e.g., ResNet or VGG), which has been trained on a large and diverse dataset. This allows the features to capture generalizable patterns. Additionally, this pre-trained model can be fine-tuned iteratively during the active learning process to align the feature representations more closely with the target task, thereby improving clustering and uncertainty estimation quality over successive iterations.

3.4 Image Classification Model

The Active learning project is based on the CNN architecture and consists of three convolutional blocks, each comprising two convolutional layers followed by a max-pooling operation. Given an input tensor $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, where C, H, and W denote the number of channels, height, and width, the convolutional layers iteratively transform the input into feature maps $\mathbf{F}_i$ using kernels $\mathbf{K}_i$, followed by ReLU activations:

$$\mathbf{F}_{i+1} = \text{ReLU}(\text{Conv2D}(\mathbf{F}_i, \mathbf{K}_i)).$$

After the final convolutional block, the feature maps are flattened into a vector $\mathbf{z} \in \mathbb{R}^d$, which is passed through fully connected layers for classification. To enhance feature extraction, we considered using a pre-trained model such as ResNet-18, VGG-16, or MobileNet-V2. This idea was dropped due to time constraints, and this is left for improvements in future works

4 Experiments

4.1 Data review

The first experiment was conducted on the "CIFAR-10 dataset" which consists of 60000 32x32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. The dataset is divided into five training batches and one test batch, each with 10000 images. The training batches contain the remaining images in random order, but some training batches may contain more images from one class than another. Between them, the training batches contain exactly 5000 images from each class.

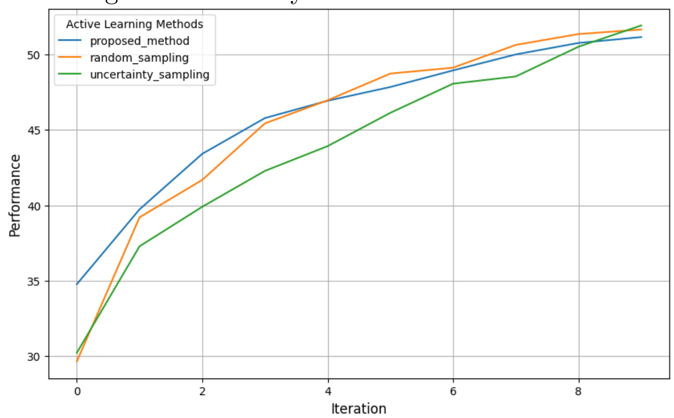
4.2 Baseline

To evaluate the performance of the proposed method, we will compare it against the baseline active learning benchmark, which is the random sampling method, as well as the uncertainty sampling method referenced throughout the paper. The evaluation will focus on the accuracy of the model in the training set.

4.3 Results

In this study, we performed a grid search across a range of parameters, as detailed in the accompanying notebook.

Figure 1: Accuracy as a Function of Iterations



The optimal results were achieved using the parameter set depicted in the plot. As shown in Figure 1, the proposed method demonstrates competitiveness with the baseline approaches, exhibiting a significant advantage over the standard uncertainty sampling method in the majority of iterations. Notably, this superiority is particularly evident in the regions with the fewest labeled samples, which aligns with the core objective of active learning. These results suggest that the underlying algorithm is promising. However, the performance of the proposed method initially shows limited improvement over random sampling, indicating potential areas for refinement in future research. Our code is implemented and can be viewed here `https://github.com/michalbar031/Active_Learning`

Table 1: Summary of Experiments

Dataset	**Model**	**Method**	**Final Accuracy (%)**
CIFAR10	CNN	proposed_method	51.15
CIFAR10	CNN	random_sampling	51.65
CIFAR10	CNN	uncertainty_sampling	51.92

References

1. **CIFAR-10 Dataset.**
   `https://www.cs.toronto.edu/~kriz/cifar.html`

2. Ash, J. T., et al. (2020). *Deep Batch Active Learning by Diverse, Uncertain Gradient Lower Bounds.*

3. Gal, Y., Islam, R., & Ghahramani, Z. (2017). *Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.*

4. Guérin, J., Gibaru, O., Thiery, S., & Nyiri, E. (2017). *CNN Features Are Also Great at Unsupervised Classification.* arXiv preprint arXiv:1707.01700.

5. Holzinger, A., et al. (2016). *Interactive Machine Learning: Experimental Evidence for the Human in the Algorithmic Loop.*

6. Lewis, D. D., & Gale, W. A. (1994). *A Sequential Algorithm for Training Text Classifiers.*

7. Sener, O., & Savarese, S. (2018). *Active Learning for Convolutional Neural Networks: A Core-Set Approach.*

8. Settles, B. (2009). *Active Learning Literature Survey.*

9. Sinha, S., et al. (2019). *Variational Adversarial Active Learning.*

10. Yoo, D., & Kweon, I. S. (2019). *Learning Loss for Active Learning.*