Active Learning

By
Abdelrahman Mostafa (20200827)
Abdelrahman Amin (20200311)
Mohamed Hisham (20200483)
Yousef Mohamed (20200669)

Artificial Intelligence Department
Cairo University
2023-2024

GitHub: https://github.com/MH0386/active_learning_with_deep_learning/

Table of Contents

1. INTRODUCTION 6

2. OBJECTIVES 7

3. METHODOLOGY 8

3.1. DATA 8

3.1.1. Balanced Datasets 8

3.1.1.1. FashionMNIST 8

.3.1.1.2 CIFAR10 9

3.1.1.3. IMDb 10

3.1.2. Imbalanced Datasets 10

3.1.2.1. FER2013 10

3.2. Used Active Learning Scenarios 12

3.2.1. Pool-Based Active Learning 12

3.3. Used Deep Learning 13

3.3.1. ResNet50 Model 13

4. EXPERIMENTAL RESULTS 14

4.1. BALANCED DATASET 14

4.1.1. BEFORE ACTIVE LEARNING 14

4.1.1.1. FASHIONMNIST 14

4.1.1.1.1. EVALUATIONS 14

4.1.1.1.2. CONFUSION MATRIX 15

4.1.1.1.3. PREDICTIONS 15

4.1.1.2. CIFAR10 16

4.1.1.2.1. EVALUATIONS 16

4.1.1.2.2. CONFUSION MATRIX 16

4.1.1.2.3. PREDICTIONS 17

4.1.1.3. IMDB 17

4.1.1.3.1. EVALUATIONS 17

4.1.1.3.2. CONFUSION MATRIX 18

4.1.2. AFTER ACTIVE LEARNING 19

4.1.2.1. FASHIONMNIST 19

4.1.2.1.1. ACCURACY 19

4.1.2.1.2. Loss 20
4.1.2.1.3. F1 Score 21
4.1.2.1.4. Precision 22
4.1.2.1.5. Recall 23
4.1.2.1.6. Confusion Matrix 24
4.1.2.1.6.1. Random Sampling 24
4.1.2.1.6.2. Uncertainty Sampling 25
4.1.2.1.6.3. Margin Sampling 26
4.1.2.1.6.4. Entropy Sampling 27
4.1.2.1.6.5. Query By Committee 28
4.1.2.2. CIFAR10 29
4.1.2.2.1. Accuracy 29
4.1.2.2.2. Loss 30
4.1.2.2.3. F1 Score 31
4.1.2.2.4. Precision 32
4.1.2.2.5. Recall 33
4.1.2.2.6. Confusion Matrix 34
4.1.2.2.6.1. Random Sampling 34
4.1.2.2.6.2. Uncertainty Sampling 35
4.1.2.2.6.3. Margin Sampling 36
4.1.2.2.6.4. Entropy Sampling 37
4.1.2.2.6.5. Query By Committee 38
4.1.2.3. IMDb 39
4.1.2.3.1. Accuracy 39
4.1.2.3.2. Loss 40
4.1.2.3.3. F1 Score 41
4.1.2.3.4. Precision 42
4.1.2.3.5. Recall 43
4.1.2.3.6. Confusion Matrix 44
4.1.2.3.6.1. Random Sampling 44
4.1.2.3.6.2. Uncertainty Sampling 45
4.1.2.3.6.3. Margin Sampling 46
4.1.2.3.6.4. Entropy Sampling 47
4.1.2.3.6.5. Query By Committee 48

4.2. UNBALANCED DATASET 49

4.2.1. FER2013 49

4.2.1.1. ..... BEFORE ACTIVE LEARNING 49

4.2.1.1.1. EVALUATIONS 49

4.2.1.1.2. CONFUSION MATRIX 50

4.2.1.1.3. PREDICTIONS 51

4.2.1.2. ..... AFTER ACTIVE LEARNING 52

4.2.1.2.1. ACCURACY 52

4.2.1.2.2. LOSS 53

4.2.1.2.3. F1 SCORE 54

4.2.1.2.4. PRECISION 55

4.2.1.2.5. RECALL 56

4.2.1.2.1. CONFUSION MATRIX 57

4.2.1.2.1.1. RANDOM SAMPLING 57

4.2.1.2.1.2. UNCERTAINTY SAMPLING 58

4.2.1.2.1.3. MARGIN SAMPLING 59

4.2.1.2.1.4. ENTROPY SAMPLING 60

4.2.1.2.1.5. QUERY BY COMMITTEE 61

5. CONCLUSION 61

Abstract

Active learning is a powerful technique within deep learning (DL) that allows algorithms to achieve high accuracy with a minimal amount of labeled data. This report explores the core concepts of active learning, its advantages, and disadvantages, and how it can be applied to various Deep Learning tasks.

Traditional supervised learning relies on passively consuming pre-labeled data. Active learning disrupts this paradigm, introducing a strategic approach that empowers the model to actively participate in its education. This abstract explores the core concepts of active learning in Deep Learning, highlighting its key benefits.

The core idea revolves around a pool-based learning strategy. The model starts with a limited set of labeled data and iteratively queries a vast pool of unlabeled data for the most informative points. These points are chosen based on a carefully selected query strategy, such as uncertainty sampling or committee disagreement. Human experts then label the chosen data points, significantly impacting the model's learning process. This creates an active learning loop: select, label, retrain, and repeat until the desired performance is achieved.

Active learning offers several compelling advantages. It can dramatically reduce the cost of labeling data, a significant bottleneck in many deep learning projects. Focusing on informative data can also lead to improved model performance compared to a random sampling of the entire data pool. Additionally, active learning effectively leverages the power of unlabeled data, a valuable resource often underutilized in traditional approaches.

Core Concepts

Learner: The Deep Learning algorithm responsible for selecting the most informative data points.
Unlabeled Data Pool: A vast collection of data points without corresponding labels.
Oracle (Annotator): A human expert or another information source that provides labels for the selected data points.
Active learning boils down to strategically selecting the most informative data points for labeling from a large pool of unlabeled data. Here is a quick rundown of the key ideas:

- Pool-based learning: Train on a small set of labeled data (seed), then iteratively query the unlabeled pool for informative points based on a chosen strategy.
- Query strategy: This is the "how" of choosing informative points. Examples include uncertainty sampling (focusing on unsure predictions) and query by committee (exploiting disagreements among multiple models).
- Human-in-the-loop (HITL): Humans provide labels for the chosen data points, impacting model performance significantly.
- Active learning loop: This cycle defines the process: select -> label -> retrain -> repeat until a stopping point is reached.
- Model uncertainty & data distribution: These concepts influence query strategies. Uncertainty focuses on the model's confidence, while data distribution ensures a diverse training set.

1. Introduction

The Power of Selective Learning: Introducing Active Learning in Deep Learning

Deep Learning thrives on data, but not all data is created equal. Labeling data, especially for complex tasks, can be a time-consuming and expensive bottleneck. This is where Active Learning steps in, offering a smarter approach to training Deep Learning models.

Active learning flips the script on traditional supervised learning. Instead of passively consuming a pre-labeled dataset, the model actively participates in its education. It strategically selects the most informative data points from a large pool of unlabeled data, presenting them to a human expert for labeling. These carefully chosen examples help the model learn more effectively, often achieving better performance with significantly less labeled data.

This introduction sets the stage for exploring the exciting world of Active Learning. We will delve deeper into the specific objectives of using this approach, such as reducing labeling costs, improving model performance, and leveraging the power of unlabeled data. By strategically choosing what to learn from, Active Learning empowers us to build more efficient and powerful Deep Learning models.

2. Objectives

There are several key objectives for applying active learning in Deep Learning:

1. Reduce Labeling Cost:
    - Labeling data can be a significant bottleneck in Deep Learning projects, often requiring human expertise and time. Active learning aims to strategically select the most informative data points for labeling, significantly reducing the amount of human effort required to achieve good model performance. This is especially valuable when dealing with large datasets where labeling everything is impractical.
2. Improve Model Performance:
    - By focusing on informative data points, active learning allows the model to learn from the most valuable examples. These points often lie in areas of uncertainty for the model or represent critical boundaries between classes. This targeted learning approach can lead to better model performance compared to training on a random sample of the entire data pool.
3. Leverage Unlabeled Data:
    - Real-world data often comes with a large amount of unlabeled data and a smaller set of labeled data. Active learning allows us to effectively utilize this unlabeled data by strategically incorporating it into the training process. This can be particularly advantageous when labeled data is scarce or expensive to obtain.
4. Reduce Training Time:
    - Since active learning requires labeling fewer data points, it can potentially lead to faster training times for Deep Learning models. This is especially true for complex models that require a significant amount of data to train effectively.
5. Targeted Data Collection:
    - Active learning can be used to guide data collection efforts. By identifying the most informative data points, active learning can help focus data collection on areas that will have the biggest impact on model improvement. This is useful in scenarios where actively gathering new data is possible.

In summary, active learning offers a strategic approach to training Deep Learning models by focusing resources on the most valuable data points. This leads to reduced labeling costs, improved model performance, and efficient utilization of data.

3. Methodology

3.1. Data

3.1.1. Balanced Datasets

3.1.1.1. FashionMNIST

What it is:

- A dataset of images focused on fashion items, containing 70,000 grayscale images of various clothing articles.
- Often considered a direct replacement for the classic MNIST dataset of handwritten digits.
- Used for image classification tasks in machine learning and deep learning.

Key details:

- **Size:**
  - Training set: 60,000 images
  - Test set: 10,000 images.
- **Image format:**
  - Grayscale
  - 28x28 pixel resolution (same as MNIST)
- **Labels:**
  - 10 distinct classes of clothing items:
    - T-shirt/top
    - Trouser
    - Pullover
    - Dress
    - Coat
    - Sandal
    - Shirt
    - Sneaker
    - Bag
    - Ankle boot

Why it is useful:

- **MNIST alternative:** Shares the same structure and size as MNIST, making it easy to adapt algorithms trained on MNIST to Fashion MNIST.
- **Benchmarking:** A popular dataset for evaluating image classification models in computer vision tasks.

- **Simplicity:** Small dataset with clear categories, making it a good starting point for beginners in machine learning and image classification.

3.1.1.2. CIFAR10

What it is:

- A popular dataset for computer vision tasks, containing 60,000 small color images in 10 distinct classes.
- Often used for training and evaluating image classification algorithms.
- Considered a benchmark dataset in the field of machine learning.

Key details:

- **Size:**
  - Total images: 60,000
  - Training set: 50,000 images
  - Test set: 10,000 images.
- **Image format:**
  - Color images.
  - 32x32 pixel resolution
- **Classes:**
  - 10 mutually exclusive classes of objects:
    - airplane
    - automobile
    - bird
    - cat
    - deer
    - dog
    - frog
    - horse
    - ship
    - truck

Why it is useful:

- **Small size:** Small dataset compared to others, making it computationally efficient for training and testing algorithms.
- **Balanced classes:** Equal number of images per class, providing fair representation for each category.
- **Simple categories:** Clear and distinct object classes, making it easier for algorithms to learn and recognize.

- **Widely used:** A common benchmark dataset, allowing comparison of performance across different algorithms and research efforts.

3.1.1.3. IMDb

What it is:

- A popular dataset for sentiment analysis tasks, consisting of 50,000 movie reviews from the Internet Movie Database (IMDb).
- Each review is labeled as either positive or negative.

Key details:

- **Size:**
  - Total reviews: 50,000
  - Split: 25,000 positive reviews, 25,000 negative reviews (balanced)
- **Content:**
  - Textual reviews of various movies
  - Only highly polarizing reviews are included (scores above 7 for positive and below 4 for negative on a 10-point scale)
  - Limited to a maximum of 30 reviews per movie to avoid bias.
- **Format:**
  - Plain text files

Why it is useful:

- **Sentiment analysis benchmark:** A widely used dataset for training and evaluating sentiment analysis models in Natural Language Processing (NLP).
- **Binary classification:** Straightforward task of classifying reviews as positive or negative, making it accessible for beginners in NLP.
- **Balanced classes:** Equal number of positive and negative reviews provides fair representation for both sentiment categories.
- **Large enough for training:** Sufficient data volume for training and testing various NLP models.

3.1.2. Imbalanced Datasets

3.1.2.1. FER2013

What it is:

- The FER2013 (Facial Expression Recognition 2013) dataset is a popular dataset for facial expression recognition tasks.
- It contains images of faces labeled with their corresponding emotions.
- The dataset has some limitations, including potential biases and the lack of color information.
- Despite these limitations, FER2013 remains a valuable resource for training and evaluating facial expression recognition models.

Key details:

- **Size:**
  - Training set: Approximately 28,709 images
  - Public test set: 7,066 images
- **Image format:**
  - Grayscale
  - 48x48 pixel resolution
- **Classes:**
  - 7 distinct emotions:
    - Angry
    - Disgust
    - Fear
    - Happy
    - Sad
    - Surprise
    - Neutral

Why it is useful:

- **Focus on facial expressions:** Specifically designed for recognizing emotions in facial expressions.
- **Widely used:** A common benchmark dataset in the field of facial expression recognition.
- **Balanced:** Most classes have around 5,000 samples, except Disgust with fewer images.
- **Simple format:** Grayscale images with a clear class structure, making it accessible for beginners.

3.2. Used Active Learning Scenarios

3.2.1. Pool-Based Active Learning

Pool-based active learning is a strategy within the field of active learning used to efficiently train a machine learning model by strategically choosing which data points to label. Here is a breakdown of its key aspects:

Core Idea:

- Starts with a large pool of unlabeled data.
- The model iteratively selects the most informative data points from the pool for human annotation (labeling).
- These labeled points are used to retrain the model, which then guides the selection of the next informative points.
- This cycle continues until the desired model performance is achieved, requiring less labeled data compared to training on a random sample of the entire pool.

Benefits:

- Reduced Labeling Cost: By strategically choosing informative data points, pool-based active learning can significantly reduce the amount of human effort required for labeling data, which can be expensive and time-consuming.
- Improved Model Performance: Focusing on informative data points allows the model to learn more effectively, potentially leading to better performance compared to training on randomly chosen data.

Types of Pool-Based Sampling Strategies:

There are various approaches to selecting informative data points from the unlabeled pool. Some common strategies include:

- Uncertainty Sampling: The model prioritizes data points for which it has the least confidence in its prediction. These points are most likely to help the model learn and improve its accuracy.

- Query by Committee (QBC): The model trains multiple copies of itself (committee) and selects data points where the committee disagrees the most. This disagreement suggests a challenging case that can improve the committee's overall decision-making.

Overall, pool-based active learning is a valuable technique for maximizing model performance while minimizing the need for labeled data.

3.3. Used Deep Learning

3.3.1. ResNet50 Model

What it is:

- ResNet50 is a powerful convolutional neural network (CNN) architecture designed for image classification tasks.
- It belongs to the broader family of Residual Networks (ResNet), known for their ability to train deep neural networks effectively.
- The "50" in the name refers to the number of layers in the model (although it technically has 48 convolutional layers, one MaxPool layer, and one average pool layer).

Key characteristics:

- **Deep architecture:** ResNet50's depth allows it to learn complex patterns and features within images, leading to high accuracy in image classification.
- **Residual connections:** A key innovation in ResNet architecture. These connections bypass some layers in the network, allowing the model to learn residual functions and address the vanishing gradient problem that can hinder training deeper networks.
- **Efficiency:** Despite its depth, ResNet50 uses techniques like bottleneck layers to reduce the number of parameters and computations, making it more efficient to train compared to older CNN architectures.

Applications:

- **Image classification:** ResNet50 excels at classifying images into various categories, making it suitable for tasks like object recognition, scene understanding, and content-based image retrieval.
- **Feature extraction:** Its ability to learn complex features makes it a valuable tool for other computer vision tasks like object detection, image segmentation, and video analysis.
- **Transfer learning:** Pre-trained ResNet50 models, trained on massive datasets like ImageNet, can be fine-tuned for specific tasks, leveraging the learned knowledge for better performance on smaller datasets.

4. Experimental Results

4.1. Balanced Dataset

4.1.1. Before Active Learning

4.1.1.1. FashionMNIST

4.1.1.1.1. Evaluations

4.1.1.1.2. Confusion Matrix

4.1.1.1.3. Predictions

4.1.1.2. CIFAR10

4.1.1.2.1. Evaluations

4.1.1.2.2. Confusion Matrix

4.1.1.2.3. Predictions

4.1.1.3. IMDb

4.1.1.3.1. Evaluations

- Test Accuracy is 84.5% after training on 20% of the training set (2000 samples)
- Train Accuracy is 87.6% after training on 20% of the training set (2000 samples)
- Precision is 85.7%
- Recall is 83.3%

4.1.1.3.2. Confusion Matrix

4.1.2. After Active Learning

4.1.2.1. FashionMNIST

4.1.2.1.1. Accuracy

4.1.2.1.2. Loss

Loss
Random
Uncertainty
Margin
Entropy
Committee
Value
Epochs
6
5
4
3
2
0
5
10
15
20

4.1.2.1.3. F1 Score

4.1.2.1.4. Precision

4.1.2.1.5. Recall

4.1.2.1.6. Confusion Matrix

4.1.2.1.6.1. Random Sampling

4.1.2.1.6.2. Uncertainty Sampling

4.1.2.1.6.3. Margin Sampling

4.1.2.1.6.4. Entropy Sampling

4.1.2.1.6.5. Query By Committee

4.1.2.2. CIFAR10

4.1.2.2.1. Accuracy

4.1.2.2.2. Loss

Loss
Random
Uncertainty
Margin
Entropy
Committee
Value
Epochs
18
17
16
15
14
13
12
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5

4.1.2.2.3. F1 Score

4.1.2.2.4. Precision

4.1.2.2.5. Recall

4.1.2.2.6. Confusion Matrix

4.1.2.2.6.1. Random Sampling

4.1.2.2.6.2. Uncertainty Sampling

4.1.2.2.6.3. Margin Sampling

4.1.2.2.6.4. Entropy Sampling

4.1.2.2.6.5. Query By Committee

4.1.2.3. IMDb

4.1.2.3.1. Accuracy

4.1.2.3.2. Loss

Loss
Random
Uncertainty
Margin
Entropy
Committee
Value
Epochs
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0
20
40
60
80
100

4.1.2.3.3. F1 Score

4.1.2.3.4. Precision

4.1.2.3.5. Recall

4.1.2.3.6. Confusion Matrix

4.1.2.3.6.1. Random Sampling

4.1.2.3.6.2. Uncertainty Sampling

4.1.2.3.6.3. Margin Sampling

4.1.2.3.6.4. Entropy Sampling

4.1.2.3.6.5. Query By Committee

4.2. Unbalanced Dataset

4.2.1. FER2013

4.2.1.1. Before Active Learning

4.2.1.1.1. Evaluations

4.2.1.1.2. Confusion Matrix

4.2.1.1.3. Predictions

4.2.1.2. After Active Learning

4.2.1.2.1. Accuracy

4.2.1.2.2. Loss

Loss
Random
Uncertainty
Margin
Entropy
Committee
Value
Epochs
8
7
6
5
4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

4.2.1.2.3. F1 Score

4.2.1.2.4. Precision

4.2.1.2.5. Recall

4.2.1.2.1. Confusion Matrix

4.2.1.2.1.1. Random Sampling

4.2.1.2.1.2. Uncertainty Sampling

4.2.1.2.1.3. Margin Sampling

4.2.1.2.1.4. Entropy Sampling

4.2.1.2.1.5. Query By Committee

5. Conclusion

Active learning offers a powerful approach to enhance the efficiency and performance of deep learning models. By strategically selecting the most informative data points for labeling, active learning allows for significant reductions in labeling costs, improved model performance, and reduced computational burden. As the field of active learning with deep learning continues to evolve, it has the potential to revolutionize various machine learning applications where data labeling is a critical bottleneck.

Our experiments revealed interesting observations regarding the performance of active learning query strategies on various datasets. Uncertainty sampling achieved the highest F1 score in both FashionMNIST and CIFAR10, which are balanced datasets with roughly equal distribution across classes. This suggests that uncertainty sampling performs well when the model is most uncertain about instances near the decision boundaries between classes. However, in the case of IMDb and fer2013, which are also balanced datasets, entropy sampling exhibited the highest F1 score. This might be due to the nature of the data itself. Entropy sampling prioritizes instances where the model's predicted probability distribution is more uniform, potentially indicating ambiguity in the data itself. This could be particularly relevant for datasets like IMDb where sentiment analysis involves nuanced interpretations of textual content.

Therefore, it appears that the optimal active learning query strategy may vary depending on the characteristics of the dataset. While uncertainty sampling excels in balanced datasets with clear class boundaries, entropy sampling might be more effective in scenarios where

the model encounters ambiguity or uncertainty within the data itself, such as with sentiment analysis or other tasks involving complex interpretations."

Here are some additional points you can consider incorporating:

- Briefly explain the concepts of F1 score and its significance in evaluating active learning performance.
- Discuss potential reasons why uncertainty sampling and entropy sampling might perform differently on different datasets.
- Emphasize the importance of choosing the appropriate query strategy based on the specific characteristics of the data and the task at hand.