# TODO: Funky active learning pun

**Lukas Weber** [1]

## Abstract

A lot of unlabelled data generally available but annotation is costly. A goal in fields where labelled data is not available is to maximize performance with minimum amount of labeled data instances. Active learning selects instances for labelling based on a selection strategy.

- Objective: Compare active learning query methods using modAL.
- Methods: Experiments with Logistic Regression (MNIST) and a CNN (CIFAR).
- Experiments: Vary initial dataset sizes (small, moderate, large).
- Key findings: Highlight main conclusions about performance in different scenarios.

## 1. Introduction

Start with a broader context like: Active learning aims to minimize the labeling cost by selecting the most informative samples for annotation by an oracle. However, the choice of an appropriate query strategy significantly impacts the effectiveness of active learning.

Motivation (Prolem i want to solve and why is the problem important)

- Labelling expensive task
- important to reduce labelled data requirement in machine learning to broaded to be able to use it in data sparse environments
- role of active learning in reducing annotation costs
- significance of evaluating query strategies

[*]Equal contribution [1]MSc Computer Science, Tübingen, GER. Correspondence to: Lukas Weber <lukas2.weber@student.uni-tuebingen.de>.

*Project report for the "Data Literacy" course at the University of Tübingen, Winter 2023/24 (Module ML4201).* Style template based on the ICML style files 2023. 

Define Active Learning and why query strategies are important. FOCUS IS ON COMPARING STRATEGIES UNDER DIFFERENT DATASET SIZES AND MODEL COMPLEXITIES
Objective

- Compare different active learning strategies using the modAL framework

**Related Work:** What did the research do? prior comparisons of query strategies and highlight the gap that i am adressing

- Previous works have explored strategies such as ..., .., .... While they have been widely tested, most evaluations focus on single dataset scenarios or specific models
- Then the gaps: However, few studies analyze the impact of initial dataset size on query strategy performance, and even fewer investigate their behavior on simple versus complex models.

Previous papers:

- (Schröder et al., 2022) Compares various uncertainty based query strategies in context of fine-tuning transformer models for text classification.
  Does not compare across different sizes of the pool, and for simple models.
- (Ueno et al., 2023) TOWARDS FUTURE DEEP ACTIVE LEARNING": Most papers concentrate on cifar or mnist, this paper evaluates six different datasets, including medical and visual inspection images. Also tackles the problem that experimental settings are not standardized, making evaluation of existing methods difficult. Also did a verification experiment where fully trained models are used to select most informative samples according to various query strategies and construct a labeled dataset. This is done to isolate and evaluate the effectiveness of query strategies indepently of training process/ underfitting or randomness due to early-stage training.

- (Zhan et al., 2022): compares different query methods on mnist and cifar with a cnn (resnet18). Does not compare impact of different initial train set size and strategies on shallow classifiers.

- (Werner et al., 2023) highlight critical challenges in active learning research, including the lack of reproducible experiments and fair comparisons across domains. They propose a benchmark framework evaluating active learning strategies on tabular, image, and text datasets using robust evaluation protocols to mitigate variance and ensure comparability. While they focus on a wide range of domais, this paper only compares methods in the image domain. Furthermore, this paper focuses on comparing the strategies for a simple logistic regression model and a CNN across different starting sizes of the annotated dataset.
  Comparing the different starting sizes simulates different possible usages of models. (1) Starting completely without labeled data, (2) Having a small pool of labeled data (3) Having a well-trained model and searching data points for fine-tuning.

**Modal**: description and relevance for research maybe? The modAL framework(ref) has simplified the implementation of active learning pipelines, yet its comparative use across different strategies remains underexplored.

Prior studies

- existing studies that compare active learning methods

- do they have gaps in the literature? (maybe: comparison under different dataset sizes, use with cnns, comparison of simple or complicated models)

**Contributions** Comparative study of popular active learning query strategies using modAL framework. Contributions are as follows:

- systemativ evaluation on two datasets: MNIST (Logistic Regression) and CIFAR (CNN)

- analysis across different scenarios: small, moderate and large initial datasets

- insight into performance impact of model complexity and initial dataset size on active learning performance

**Structure of the paper**

- outline paper

# 2. Methods

## 2.1. Datasets

- MNIST and CIFAR describe

- processing steps
  MNIST: Vectorized, flattened (describe input size), normalized

## 2.2. Models

- Logistic Regression: Details on implementation and hyperparameters, how were they chosen?

- same for CNN

## 2.3. Active Learning

- How are the splits (initial pool, initial training data) generated

- numbmer of iterations for active learning

- Used active learning query strategies

Metrics for evaluation (Accuracy, train error, confusion matrix, ...)

Tool: modAL, scikit

# 3. Experiments

Setup: details for reproducibility (random number generators)? Random seed = 42
Did I perform multiple runs?

Scenarios

- describe experiments and their settings and purposes

# 4. Results

Quantitative results: Tables or plots comparing the query strategies across scenarios. Also maybe key peroramnce differences (accuracy vs number of labeled examples)

# 5. Discussion

Qualitative analysis: Why certain strategies perform better or worse in specific settings. Trends between simple models and complex models

Impact of initial dataset size

Most methods require knowledge about the balance of the dataset and might be disadvantageous to use in case

of a disbalanced dataset. In real world scenarios, this is not known.
Only image-based datasets were tested. Not vector- or text-based.

## 6. Conclusion

Summary of findings and insights (maybe that a strategy works best with small datasets while another one is robust on bigger ones)
Implications: practical recommendations?
Future work? (More diverse datasets, exploring additional strategies, applying in real-world tasks)

## 7. Cool ideas

- Show the typical framework of active learning in a figure in the introduction

## 8. Good formulations

From the active learning survey: Entropy is an information-theoretic measure that represents the amount of information needed to "encode" a distri- bution. As such, it is often thought of as a measure of uncertainty or impurity in machine learning.

## References


Schröder, C., Niekler, A., and Potthast, M. Revisiting uncertainty-based query strategies for active learning with transformers, 2022. URL https://arxiv.org/abs/2107.05687.

Ueno, S., Yamada, Y., Nakatsuka, S., and Kato, K. Benchmarking of query strategies: Towards future deep active learning, 2023. URL https://arxiv.org/abs/2312.05751.

Werner, T., Burchert, J., and Schmidt-Thieme, L. Towards comparable active learning, 2023. URL https://arxiv.org/abs/2311.18356.

Zhan, X., Wang, Q., hao Huang, K., Xiong, H., Dou, D., and Chan, A. B. A comparative survey of deep active learning, 2022. URL https://arxiv.org/abs/2203.13450.