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 and Related Work

Supervised learning often requires large, annotated datasets, but labeling data can be costly and time-intensive, especially when expert knowledge is needed. Active learning addresses this challenge of reducing the overall labeling costs through the selection of the most informative samples for annotation based on a query strategy. This approach enables machine learning models to perform well in tasks with limited data and high labeling costs.
The effectiveness of active learning depends heavily on the chosen query strategy, which determines which samples are selected for annotation. Two main frameworks are commonly used:

1. Pool-based active learning: A model is trained on a small labeled dataset and used to evaluate a large pool of unlabeled data. The most informative samples are selected based on a query strategy, labeled by an oracle, and added to the training set. The model is retrained with the updated training set and the process is repeated until a stopping criterion is met.
2. In stream-based active learning, unlabeled data arrives sequentially. Thus, the model decides whether to query the label of an incoming instance using an evaluation metric like the uncertainty of the prediction and checking whether a threshold is reached. If not, the instance is discarded.

This report focuses on comparing several pool-based as well as stream-based query strategies. The modular Active Learning framework (modAL) (Danka & Horvath, 2018) is used in the following experiments as it simplified the implementation of active learning pipelines, contains multiple already implemented querying strategies and is compatible with models from the Scikit-learn library (Pedregosa et al., 2011).
Several previous works explore different query methods for different tasks.
(Schröder et al., 2022) compares various uncertainty based query strategies in the context of fine-tuning transformer models for text classification, while (Zhan et al., 2022) compares several query methods on MNIST and CIFAR.
According to (Ueno et al., 2023) most papers concentrate on two image-based datasets: MNIST (Lecun et al., 1998) and CIFAR (Krizhevsky et al., 2009), while their paper evaluates strategies on six different datasets, including medical and visual inspection images. They also verify the effectiveness of query strategies independently of the training process, underfitting and randomness due to early-stage training in an experiment where fully trained models are used to select the most informative samples according to various query strategies and to construct a labeled dataset.
(Werner et al., 2023) agrees with their statement that experimental settings are not standardized across papers which makes comparisons of different existing and new query strategies difficult. 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 (Ueno et al., 2023) evaluates strategies across several image-based datasets from a wide range of image-domains like the medical and visual inspection images, (Werner et al., 2023) focuses on a wide range of data-domains including images, text and vector-based datasets.

*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.

This report focuses on comparing strategies in the image domain. It compares query strategies for a simple logistic regression model and a more complex CNN to study the effect of model complexity on the optimal query strategy.
It is uncertain whether query strategies which work well on simple models generalize well to more complex models like Neural Networks (Schröder & Niekler, 2020). To guarantee real-world applicability, knowledge about this needs to be gained as model complexities vary across different tasks and datasets. Furthermore, the impact of different sizes of the initial annotated dataset is additionally investigated which is something that few studies analyze in combination. The effect of the combination between model complexity and initial annotated dataset size is a gap not researched in previously mentioned papers.
By examining different starting sizes, this work simulates various real-world scenarios: (1) starting with no labeled data, (2) having a small pool of labeled data, and (3) working with a well-trained model to identify data points for fine-tuning.
The contributions of this paper are as follows:

- **Systematic evaluation of active learning query strategies** on MNIST with a logistic regression model and CIFAR-10 using a CNN to analyze their behavior across datasets and model complexities.
- **Investigation of model complexity's influence** on the effectiveness of query strategies to provide knowledge into how shallow and deep models respond differently to active learning methods.
- **Analysis of the initial labeled dataset size's impact** on the performance of query strategies to simulate real-world scenarios such as starting completely from scratch, having a small labeled dataset or fine-tuning a model.

In section 2 an overview of the used datasets, processing steps and models is given, while section 3 introduces the three conducted experiments followed by the results and discussion in section 4 and 5, conclusion in section 6

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, ...)

3. Experiments

Setup: details for reproducibility (random number generators)? Random seed = 42 (**?**) mention reproducibility problems in models even when setting random number generators before each run. By training models multiple times in succession and checking whether the same configurations lead to different outcomes, it is shown that in this scenario, reproducibility is guaranteed. TODO: Schnell ein skript aufsetzen und dann finale accuracy und so weiter vergleichen.
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 perforamnce 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
- Accuracy vs. fraction of data size in relation to the datset (percentage of used data points)

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.

9. Comparison to the other papers

Previous papers do not: Is everything here now mentioned previously? if yes, then it can be deleted.

- (Schröder et al., 2022): Does not compare across different sizes of the pool, and for simple models.
- (**?**): Does not compare impact of different initial train set size and strategies on shallow classifiers.
- According to (Ueno et al., 2023): check if true: Does not compare different sizes of the initial starting sets as well as how simple and more complex perform with the query methods
- (Werner et al., 2023): They focus on a wide range of domains, this paper only compares methods in the image domain. I compare strategies for a simple logistic regression model and a complex CNN across different starting sizes of the annotated dataset.

References

Danka, T. and Horvath, P. modal: A modular active learning framework for python. 2018. URL https://arxiv.org/abs/1805.00979.

Krizhevsky, A. et al. Learning multiple layers of features from tiny images. 2009.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.

Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Schröder, C. and Niekler, A. A survey of active learning for text classification using deep neural networks, 2020. URL https://arxiv.org/abs/2008.07267.

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.