Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Demonstrations Track

Visualizing and Interacting with Model Representation Space for Human-Centric Active Learning

Rida Saghir[1,2] , **Thiago S. Gouvêa**[1,2]
Daniel Sonntag[1,2]
[1]German Research Center for Artificial Intelligence (DFKI), Germany
[2]University of Oldenburg, Germany
{rida.saghir, thiago.gouvea, daniel.sonntag}@dfki.de,

Abstract

Active learning reduces annotation effort by selecting informative samples, yet most approaches remain model-driven, offering users little control over training or support for understanding model behaviour. Human-centric active learning brings users further into the loop by introducing additional points of interaction, particularly in the sample selection process. However, such systems are typically demonstrated using fixed feature projections or visualizations of shallow classifier outputs. We present a representation-centric active learning tool in which interaction takes place directly within the model's representation space. By operating in the same space the model uses for decision making, the interface supports the co-evolution of representations and user understanding. We additionally report initial qualitative (think-aloud) and quantitative findings from a pilot study, illustrating that such representation-centric frameworks can achieve comparable performance to standard baselines while fostering improved human–model collaboration. (Video and the code available at https://cst.dfki.de/demo-interacting-model-space)

1 Introduction

Active Learning (AL) aims to reduce annotation costs by selecting samples that are expected to be most informative for improving a model performance [Settles, 2009]. Traditional AL pipelines are largely model-centric where the system determines which instances to query based on predefined selection criteria, while users primarily serve as labelling oracles [Settles, 2009; Kath *et al.*, 2023]. While effective in optimizing learning efficiency, this paradigm offers limited support for user-driven exploration, inspection of model behaviour, or intervention beyond annotation.

To address these limitations, recent work has explored interactive and human-centric active learning systems that combine iterative model training with visual access to data structure and model outputs, enabling users to have greater control over sample selection and steer the model's learning process [Legg *et al.*, 2019]. Studies have shown that visual interactive labelling can be competitive with algorithm-driven querying using fixed low-dimensional projections of static feature representations [Bernard *et al.*, 2017]. Further work analyses user-driven sampling strategies in visual-interactive labelling, demonstrating how different user behaviours can inform active learning algorithms, mitigate early bootstrap issues, and in some cases outperform traditional selection heuristics [Bernard *et al.*, 2018]. Related approaches allow users to guide selection through visualizations of classifier confidence or posterior probabilities [Seifert and Granitzer, 2010], or support inspection of model behaviour by displaying decision boundaries, margins, or uncertainty over dimensionally reduced feature spaces [Heimerl *et al.*, 2012; Huang *et al.*, 2017].

While these systems vary in interaction design and visual encodings, they commonly rely on fixed feature projections or visualization spaces derived from shallow classifier outputs. For models operating on fixed feature representations, decision boundaries and margin-based views provide a faithful abstraction of model behavior and are well suited for interaction. However, modern active learning increasingly relies on representation-learning models [Wang *et al.*, 2016], where performance depends on a learned representation space that evolves as new annotations are incorporated. This creates opportunities for representation-centric interfaces that allow humans to interact with the model through the same representation space that model uses as a decision space. Related work beyond active learning has explored direct interaction with learned latent spaces, such as SpaceEditing [Wei *et al.*, 2024], which allows users and models to have a common decision space. This further motivates representation-centric interfaces as a medium for human–model collaboration.

To investigate this form of interaction in active learning, we present an interactive, web-based system that places a trainable representation space at the centre of human–model interaction. The system visualizes a representation space learned end-to-end with a classifier, which evolves as user annotations are incorporated. To expose informative samples, the system employs established active learning sampling criteria; such as entropy based uncertainty [Settles, 2009; Lewis, 1995], diversity [Monarch, 2021], and density [Zhu *et al.*, 2008] which have been shown to support effective learning. These sampling scores highlight informative samples while leaving the final annotation decision to the user.

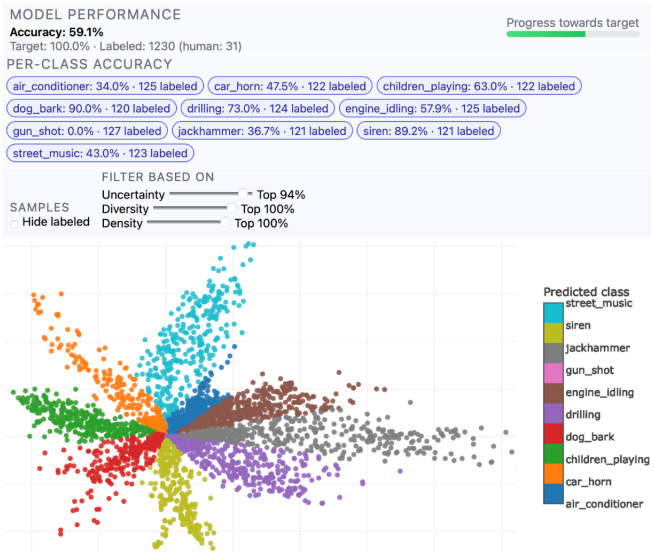

Figure 1: Tool view of the human-centric active learning interface.

2 System Overview

We present a web-based interactive system for human-centric active learning. The frontend is implemented using HTML and JavaScript, while the backend is built in Python using PyTorch for model training and inference, and handles embedding updates as well as the computation of active learning sampling scores. The system is designed to support audio datasets and allows users to explore, listen to, and annotate samples through a two-dimensional representation space.

2.1 Machine Learning Backend

The machine learning component employs the pretrained audio model YAMNet[1] as a feature extractor to obtain high-dimensional representations of the audio samples. YAMNet operates on fixed-length audio windows of approximately 0.96 s and aggregates predictions over sliding windows to produce clip-level representations. These representations are passed to a multi-layer perceptron classifier with a 256-dimensional hidden layer and a two-dimensional output layer, resulting in a trainable 2D representation space. The model is trained using a standard cross-entropy classification loss, and the representation space is optimized jointly with this objective.

2.2 Interface Design

The main interaction view of the tool is a two-dimensional representation space (Fig 1). Each point corresponds to an audio sample in the dataset and is colored according to its predicted class. To support focused exploration, the interface provides sliding filters that allow users to downsample the displayed samples based on their informativeness, following the sampling strategies described in section 2.3. At the top of the interface, overall model accuracy and per-class accuracy are displayed to provide performance feedback. Interaction is initiated by clicking on a data point, which triggers playback of the corresponding audio sample and prompts the user to provide a label. As new annotations are submitted, they are incorporated into the training process, and the model enters a retraining phase during which further annotations are temporarily disabled to ensure consistency. During retraining, both the classifier parameters and the representation space are updated, allowing the visualization to reflect the evolving structure of the model's learned representation over time.

2.3 Sampling Filters

To guide user attention toward potentially informative samples without automating query selection, the system employs multiple active learning sampling strategies. Specifically, for each unlabelled sample i, we compute:

- **Uncertainty**, quantified using the entropy of the predicted class distribution:

$$u(i) = -\sum_{c} p_{i,c} \log p_{i,c}, \tag{1}$$

  where $p_{i,c}$ denotes the model-predicted probability of class c for sample i.

- **Diversity**, defined as the minimum Euclidean distance to any labeled sample in the representation space:

$$d(i) = \min_{j \in \mathcal{L}} \|\mathbf{z}_i - \mathbf{z}_j\|, \tag{2}$$

  where $\mathcal{L}$ is the set of labelled samples and $\mathbf{z}_i$ denotes the low-dimensional embedding of sample i.

- **Density**, estimated as the inverse of the average distance to the k nearest unlabelled neighbours:

$$\rho(i) = \frac{1}{\frac{1}{k}\sum_{j \in \mathcal{N}_k(i)} \|\mathbf{z}_i - \mathbf{z}_j\|}, \tag{3}$$

  where $\mathcal{N}_k(i)$ denotes the set of the k nearest unlabelled neighbors of sample i in the representation space.

Each score is min-max normalized across the current unlabelled pool and exposed to users through a percentile-based *slider* control. Given a user-selected value $q \in [0, 1]$, the system computes the $(1 - q)$-quantile of the corresponding score distribution and retains only samples whose normalized scores exceed this cutoff, corresponding to the top q fraction of samples. When multiple sliders are active, samples must satisfy all corresponding percentile cutoffs (logical AND), allowing users to interactively narrow the candidate set.

3 Demonstration and Preliminary Results

We demonstrate the system using the UrbanSound8K [2] dataset, which contains 8,732 audio clips from ten everyday sound classes (e.g., car horn, dog bark, drilling) that are generally familiar to non-expert users. The audio clips vary in duration between approximately 1 and 4 seconds. Although UrbanSound8K is typically evaluated using pre-defined ten-fold cross-validation, our goal is to study an active learning setting rather than fully supervised generalization. We therefore reserve one fold exclusively for evaluation, and treat the

[1] YAMNet

[2] UrbanSound8k

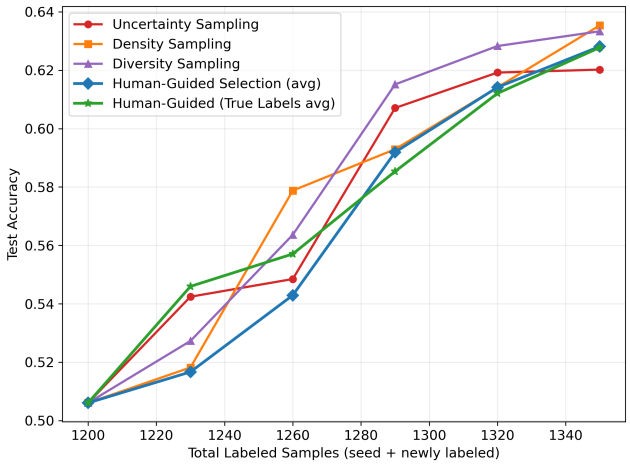

Figure 2: Model performance on the test set as a function of the number of labeled samples. The blue curve shows average performance across the two users, compared against active learning baselines (uncertainty, density, diversity sampling). The green curve reports performance using human-selected samples with ground-truth labels.

remaining folds as a shared pool from which both the initial labelled set and subsequent queried samples are drawn. Performance is reported on the held-out test fold using overall and per-class accuracy. To initialize the learning process, we bootstrapped the model by training on a small seed set containing 120 labelled samples per class equally from all train folds. We conducted a small-scale pilot study with two participants from the data science field. After a brief task introduction, they listened to five example audio samples per class and explored the tool to gain familiarity before starting annotation. Each participant labelled up to 150 samples, with the model incrementally retrained after every 30 newly provided annotations. The study was evaluated both quantitatively, using model performance metrics, and qualitatively, through think-aloud sessions.

3.1 Quantitative Results

We report the average model performance across the two users (blue curve) as a function of the number of labelled samples and compare it to active learning baselines using uncertainty, density, and diversity sampling (Fig 2). Because human annotations were not always accurate due to confusing classes and noisy audio, we additionally report model performance using the human-selected samples with ground-truth labels (green curve). Overall, we observe that human-guided selection achieves performance comparable to algorithm-driven querying strategies across labelling rounds.

3.2 Observations from Think-Aloud Sessions

We conducted think-aloud sessions with the participants while they interacted with the tool and summarized recurring themes below.

Annotation Guided by Spatial Structure. Participants treated the representation space as a form of the model's mental map, assuming that their labeling actions could influence its structure and reorganize regions of uncertainty. For example, User 1 said: *"I will label tightly packed samples having different prediction colours."*. The same participant further noted: *"I will take samples from all coloured regions so that model learns for all of them."* The same participant also noted: *"Some classes have more outliers, I will label more from them."* User 2 had similar observations: *"Samples are tightly packed in top 30% uncertain samples."* and *"The more they are apart from centre, the more certain they are."*

Aligning Spatial Proximity with Acoustic Similarity. Participants often interpreted spatial proximity in the representation space in terms of acoustic similarity. User 1 observed: *"Jackhammer and drilling are very close and lie in uncertain regions. Makes sense, their audio is hard to distinguish even for me."*. User 2 further observed: *"Some classes have fewer samples in uncertain region, which means model learn about these classes much easier, like siren."*

Making Sense of Sampling with Filters. Filtering actions promoted understanding of active learning common sampling strategies. User 1 remarked: *"I understand why this middle region is uncertain specially in top 1-2% uncertain pool."*. *"Diversity looks like trying to cover whole embedding space."* These reflections suggest that participants used the filtering mechanisms to understand how different sampling criteria operate within the representation space.

Perceived Transformation of the Space. Across annotation rounds, participants reported observing structural change. User 1 noted: *"In this round some classes lie in low uncertainty... before they were taking more space in uncertain pool."*. *"In beginning tightly packed uncertain samples were in middle... now gone and spread outwards."* Participants also formed expectations about how the space would evolve over subsequent training rounds: *"I expect the tightly packed samples in middle to go further away in further epochs."*

Overall, the think-aloud sessions suggest that participants treated the representation space as a platform for understanding the model, teaching it through sampling and labelling, and strategically influencing its structure.

4 Limitations and Future Work

While a two-dimensional representation space enables intuitive and accessible interaction, it cannot capture complex information. Although our aim was to have a common human interaction and model's decision space for better understanding, reducing the representation to two dimensions inevitably limits the model performance due to information loss. Models could operate in higher-dimensional spaces and project them to two dimensions at each training round using dimensionality reduction techniques. However, this approach would have introduced additional computational overhead and was challenging to integrate within the scope of our pilot study. Therefore, we plan to explore similar interaction paradigms for higher-dimensional representations and validate the framework through larger and more systematic user studies.

Acknowledgments

This research is part of the Computational Sustainability & Technology project area[3], and has been supported by the Lower Saxony Ministry of Science and Culture (MWK) in zukunft.niedersachsen program, the Federal Ministry of Research, Technology and Space (BMFTR) under grant number 16IW26002 (AMENABLE), and the Endowed Chair of AAI at University of Oldenburg.

References

[Bernard *et al.*, 2017] Jürgen Bernard, Marco Hutter, Matthias Zeppelzauer, Dieter Fellner, and Michael Sedlmair. Comparing visual-interactive labeling with active learning: An experimental study. *IEEE transactions on visualization and computer graphics*, 24(1):298–308, 2017.

[Bernard *et al.*, 2018] Jürgen Bernard, Matthias Zeppelzauer, Markus Lehmann, Martin Müller, and Michael Sedlmair. Towards user-centered active learning algorithms. In *Computer Graphics Forum*, volume 37, pages 121–132. Wiley Online Library, 2018.

[Heimerl *et al.*, 2012] Florian Heimerl, Steffen Koch, Harald Bosch, and Thomas Ertl. Visual classifier training for text document retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848, 2012.

[Huang *et al.*, 2017] Lulu Huang, Stan Matwin, Eder J de Carvalho, and Rosane Minghim. Active learning with visualization for text data. In *Proceedings of the 2017 ACM Workshop on Exploratory Search and Interactive Data Analytics*, pages 69–74, 2017.

[Kath *et al.*, 2023] Hannes Kath, Thiago S Gouvêa, and Daniel Sonntag. A human-in-the-loop tool for annotating passive acoustic monitoring datasets. In *IJCAI*, pages 7140–7144, 2023.

[Legg *et al.*, 2019] Phil Legg, Jim Smith, and Alexander Downing. Visual analytics for collaborative human-machine confidence in human-centric active learning tasks. *Human-centric computing and information sciences*, 9(1):5, 2019.

[Lewis, 1995] David D. Lewis. A sequential algorithm for training text classifiers: corrigendum and additional data. *SIGIR Forum*, 29(2):13–19, September 1995.

[Monarch, 2021] Robert Munro Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

[Seifert and Granitzer, 2010] Christin Seifert and Michael Granitzer. User-based active learning. In *2010 IEEE International Conference on Data Mining Workshops*, pages 418–425. IEEE, 2010.

[Settles, 2009] Burr Settles. Active learning literature survey. 2009.

[3]https://cst.dfki.de/

[Wang *et al.*, 2016] Keze Wang, Dongyu Zhang, Ya Li, Ruimao Zhang, and Liang Lin. Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600, 2016.

[Wei *et al.*, 2024] Jiafu Wei, Ding Xia, Haoran Xie, Chia-Ming Chang, Chuntao Li, and Xi Yang. Spaceediting: A latent space editing interface for integrating human knowledge into deep neural networks. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 489–503, 2024.

[Zhu *et al.*, 2008] Jingbo Zhu, Huizhen Wang, Tianshun Yao, and Benjamin K Tsou. Active learning with sampling by uncertainty and density for word sense disambiguation and text classification. In *22nd international conference on computational linguistics, coling 2008*, pages 1137–1144, 2008.