# Active Learning for Sound Event Classification using Monte-Carlo Dropout and PANN Embeddings

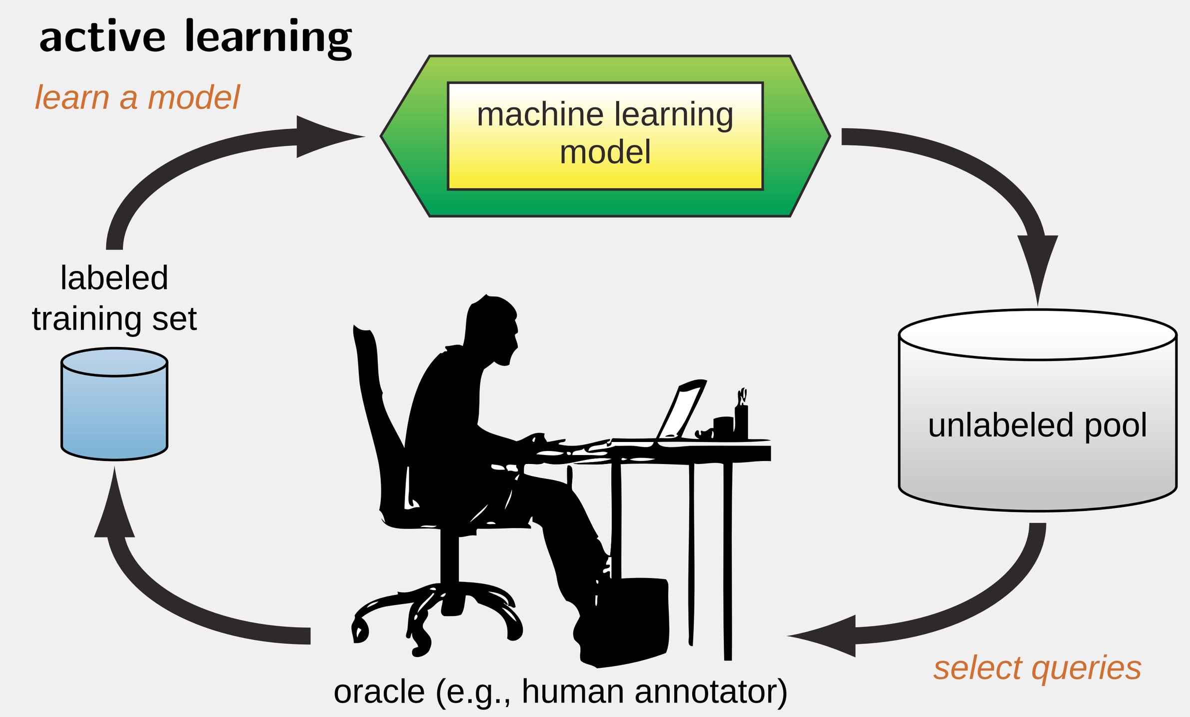
Stepan Shishkin[1], Danilo Hollosi[1], Simon Doclo[1,2], Stefan Goetze[3]

[1] Fraunhofer Institute for Digital Media Technology IDMT, Division Hearing, Speech and Audio Technology, Oldenburg, Germany
[2] University of Oldenburg, Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all, Oldenburg, Germany
[3] The University of Sheffield, Dept. of Computer Science, Speech and Hearing (SPandH), Sheffield, United Kingdom

## 1 Motivation

- problem: labeling audio material by hand is tedious
- solution: employ **active learning** to train a machine learning system to classify sound segments from few provided examples

**active learning**




- iterative process involving a trainable *classifier* and a human *annotator*
- stops when the number of manual labels has reached the *labeling budget*

  → annotator does not have to label the entire dataset

B. Settles, 2009. Active learning literature survey. *Technical report. University of Wisconsin-Madison*

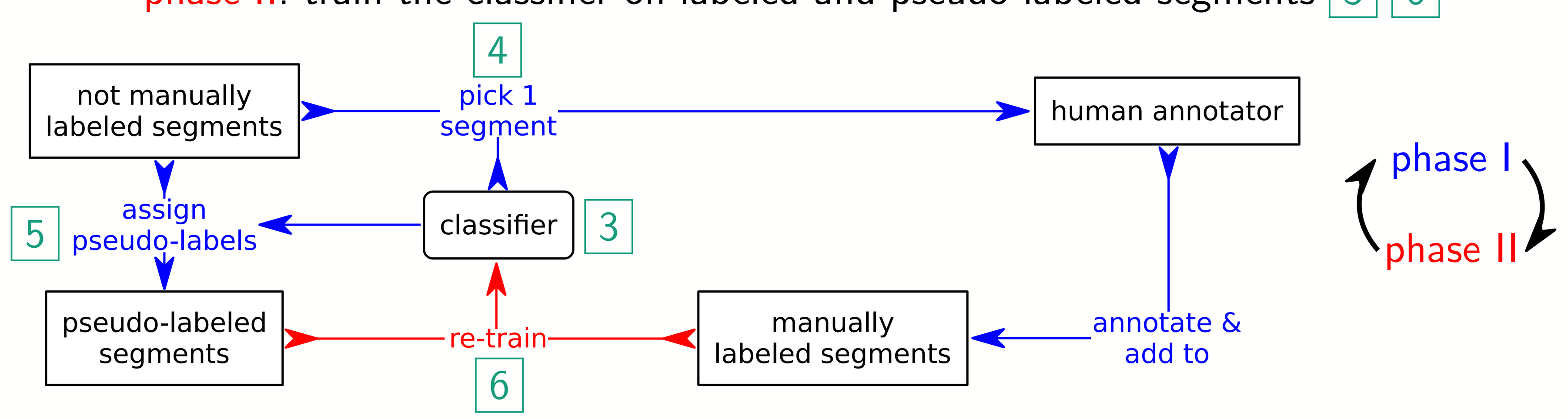
## 2 Overview

We present a dropout-based active learning system for classification of sound segments (**DAL**), which utilizes

- *transfer learning* via PANN[1] embeddings 3
- *semi-supervised learning* via pseudo-labeling 4 6
- Bayesian modeling via Monte-Carlo dropout[2] 3

## 2 Dropout-based active learning (DAL) workflow

- start by training a classifier on some initially provided set of labeled segments
- iterate between
  - phase I: assign pseudo-labels to some unlabeled segments 5 and pick one unlabeled segment to be presented to the annotator 4
  - phase II: train the classifier on labeled and pseudo-labeled segments 3 6




## 3 Bayesian neural network classifier

- probabilistic classifier via a Bayesian neural network

PANN embedding[1] $\mathbf{x}$ of a sound segment

↓

| dropout ($p = 0.5$; random dropout mask $\mathbf{d}$) |
|---|
| dense |
| softmax |

↓

class probability distribution $P(c|\mathbf{x}, \mathbf{d})$

- dropout layer is kept in stochastic mode at all times
- processing an input $\mathbf{x}$ with a random dropout mask $\mathbf{d}$ ≡ evaluating a hypothesis from a variational Bayesian posterior[2]
- posterior class distribution $P(c|\mathbf{x}) = \mathbb{E}_{\mathbf{d}}[P(c|\mathbf{x}, \mathbf{d})]$
- predicted class $\hat{l}(\mathbf{x}) = \text{argmax}_c P(c|\mathbf{x})$

## 4 Picking segment for manual annotation

- idea: pick the segment where the classifier is most uncertain
- each hypothesis (≡ each sampled dropout mask) casts a *vote* in favor of one class $c$: $v(\mathbf{x}, \mathbf{d}) = \text{argmax}_c P(c|\mathbf{x}, \mathbf{d})$
- collecting individual votes results in the *vote distribution*: $\tilde{P}(c|\mathbf{x}) = \mathbb{E}_{\mathbf{d}}[\delta_{v(\mathbf{x},\mathbf{d}),c}]$ with $\delta$ the Kronecker-delta
- *disagreement* is measured as the entropy of the vote distribution: $H_{\tilde{P}}(\mathbf{x}) = -\sum_c \tilde{P}(c|\mathbf{x}) \cdot \log \tilde{P}(c|\mathbf{x})$
- the segment with the highest disagreement is picked and presented to the annotator

## 5 Pseudo-labeling

- idea: assign pseudo-labels to those unlabeled segments where the classifier 3 is confident
- classifier is confident iff the probability of the predicted class is above some threshold $P(\hat{l}|\mathbf{x}) > \Theta \Leftrightarrow$ assign pseudo-label $\hat{l}$ to $\mathbf{x}$ ($\Theta$ is a parameter of DAL)
- special cases:
  - $\Theta = 0$: always assign pseudo-labels to all unlabeled segments
  - $\Theta = 1$: never assign pseudo-labels

## 6 Training

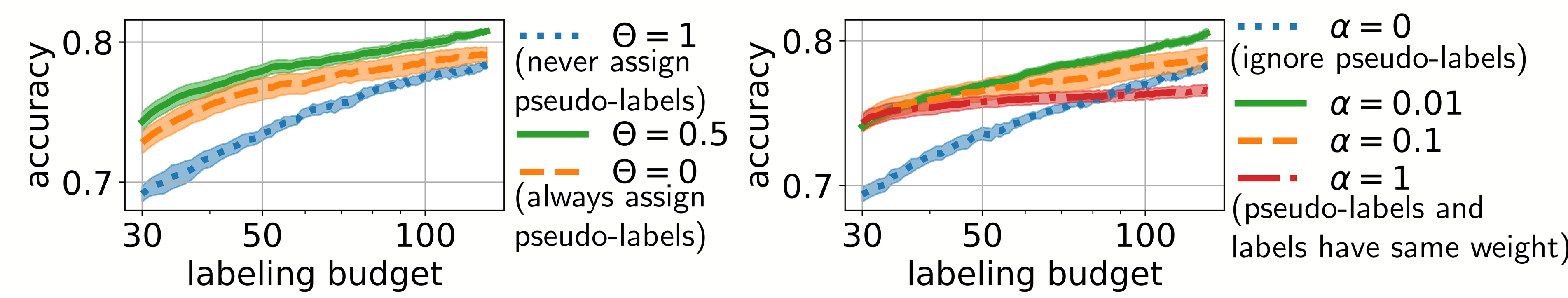
- idea: limit the impact of pseudo-labels to avoid self-amplifying misclassifications
- labeled and pseudo-labeled segments are sampled into minibatches and the cross-entropy loss is minimized via stochastic gradient descent
- chance of a pseudo-labeled segment to be drawn into a minibatch is $\alpha^{-1}$ times smaller than the chance of a manually labeled segment ($\alpha$ is a parameter of DAL)
- special cases:
  - $\alpha = 0$: pseudo-labeled segments are not used for training
  - $\alpha = 1$: pseudo-labeled and labeled segments are weighted the same

## 7 Experiments

### setup

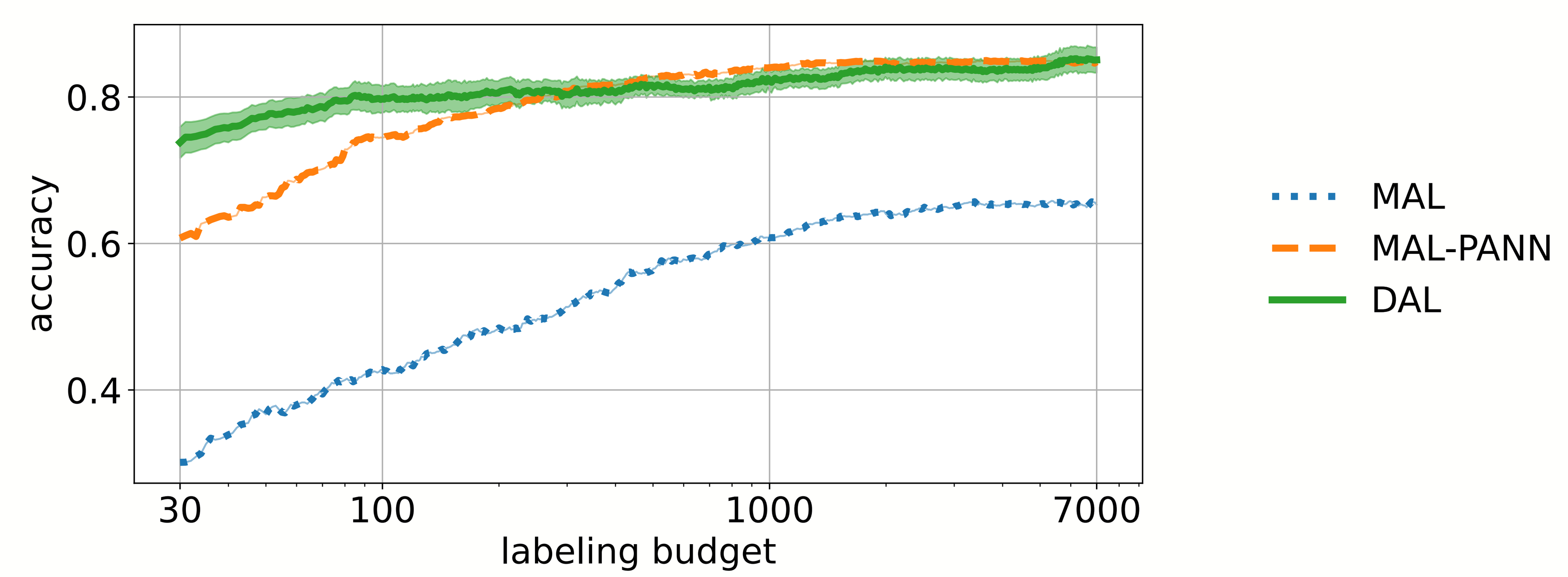
- dataset: UrbanSound8K
  - 8732 sound segments, up to 4 seconds each
  - 10 classes: air conditioner, car horn, children playing, dog bark, drilling, engine idling, gun shot, jackhammer, siren, and street music
- DAL starts with 3 labeled examples for each class (chosen randomly)
- DAL parameters: $\Theta = 0.5$ 5; $\alpha = 0.01$ 6
- human annotator is simulated by looking up ground-truth labels
- performance metric: accuracy (macro-recall) of the classifier for different labeling budgets

### DAL performance sensitivity to $\Theta$ 5 and $\alpha$ 6




### comparison to benchmarks

- baseline: medoid-based active learning (**MAL**)[3]
  1. group sound segments into small clusters using MFCC-based features
  2. manually annotate medoids of $N$ largest clusters, where $N$ is the labeling budget
  3. propagate labels to other cluster members
  4. train SVM on manual & propagated labels
- **MAL-PANN**, a modification of MAL which uses PANN embeddings[1] instead of MFCC-based features




## 8 Conclusions

- Performance of dropout-based active learning depends on the choice of pseudo-labeling confidence threshold $\Theta$ 5 and the rel. weighting of pseudo-labeled segments $\alpha$ 5.
- In our experiments, dropout-based active learning outperforms benchmark methods especially for low labeling budgets.

[1] Kong et al. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *TASLP* **28**

[2] Gal et al. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *PMLR* **48**

[3] Shuyang et al. 2017. Active learning for sound event classification by clustering unlabeled data. *ICASSP*