ACTIVE LEARNING FOR EFFICIENT AUDIO ANNOTATION AND CLASSIFICATION WITH A LARGE AMOUNT OF UNLABELED DATA

Yu Wang, Ana Elisa Mendez Mendez, Mark Cartwright, and Juan Pablo Bello

New York University

ABSTRACT

There are many sound classification problems that have target classes which are rare or unique to the context of the problem. For these problems, existing data sets are not sufficient and we must create new problem-specific datasets to train classification models. However, annotating a new dataset for every new problem is costly. Active learning could potentially reduce this annotation cost, but it has been understudied in the context of audio annotation. In this work, we investigate active learning to reduce the annotation cost of a sound classification dataset unique to a particular problem. We evaluate three certainty-based active learning query strategies and propose a new strategy: alternating confidence sampling. Using this strategy, we demonstrate reduced annotation costs when actively training models with both experts and non-experts, and we perform a qualitative analysis on 20k unlabeled recordings to show our approach results in a model that generalizes well to unseen data.

Index Terms— active learning, sound classification, audio annotations, machine listening

1. INTRODUCTION

Sound classification is an important topic in machine listening, having a wide range of applications such as noise monitoring [1, 2], animal call classification [3, 4], and music information retrieval [5, 6]. Modern sound classification models are typically trained using supervised learning. However, supervised learning requires a large amount of labeled data to train a robust model. While labeled audio data can be acquired through human annotation, it can cost a significant amount of effort. One can justify this cost if the data can be reused for several different problems (e.g., AudioSet [7]), but there are many problems that have sound classes of interest that are idiosyncratic to the problem, e.g. unusual machine or sensor failures. For such problems, existing data is of little value, and we must collect new data that may have minimal utility for other tasks—thus increasing the annotation cost per task.

One example of an idiosyncratic sound classification task appears in the Sounds of New York City (SONYC) project [8]. SONYC aims to monitor, analyze, and mitigate urban noise pollution by deploying a city-wide acoustic sensor network and leveraging machine listening, big data analytics, and citizen science. The project uses machine listening to continuously detect sound events of interest within urban acoustic environments. Thus far, SONYC has deployed 54 sensors and has recorded approximately 28 years worth of audio for analysis and model training.

However, within these unlabeled audio recordings, a distortion artifact of unknown origin has been discovered that is present in an estimated 7% of the data. This noisy, unknown sound may confuse machine listening models, causing them to produce inaccurate predictions. To remedy this situation, we need labeled examples of the artifact noise to train models to classify and disambiguate this new sound source. Since this sound class appears unique to our data, we cannot use an existing model or labeled dataset—we must annotate and develop a new problem-specific dataset. In addition, since this rare class only occurs in roughly 7% of the data, we must use an efficient annotation method that does not oversample negative examples. We propose to tackle this problem using active learning.

Active learning (AL) is a machine learning method that actively chooses the data to learn from. An AL algorithm queries an oracle—in our case, human annotators—for labels of the most informative instances to improve model performance. By doing so, AL maximizes the effectiveness of each annotation and can significantly reduce the number of labels, and thus the human effort, to develop a robust classifier. The effectiveness of AL has been shown in research areas such as natural language processing [9, 10], computer vision [11, 12], and speech recognition [13, 14]. Recently, AL and other interactive learning frameworks have also been applied to reducing audio annotation time [15], sound event classification for bird sounds [16, 17] and for environmental sound [18, 19]. Most AL research efforts simulate the AL querying process by using pre-labeled ground truth as the oracle. In contrast, we apply AL to real, unlabeled urban sound data from SONYC with actual human annotators in the loop.

In this work, we demonstrate the use of AL in a real-world scenario to train a binary sound classification model to identify an unusual, problem-specific noise source. In doing so, we systematically evaluate certainty-based AL query strategies, propose a new sampling strategy: *alternating confidence sampling*, and evaluate our approach on a population of 15 annotators. The proposed approach results in a model with comparable performance to a reference model but trained more efficiently on far fewer annotated recordings.

2. ACTIVE LEARNING FRAMEWORK

To train our AL models, we use a pool-based AL framework which consists of the following steps in a loop:

1. The labeled training pool is initialized to contain one positive and one negative example.
2. A random forest classifier is trained using the labeled training pool.
3. Using the trained model, a sampler selects a single query from the unlabeled data pool using a certainty-based query strategy (see Section 3).
4. A human annotator listens to and labels the queried audio clip.
5. The newly labeled example is then added to the training set, completing one iteration.

This work was partially supported by National Science Foundation award 1544753.

6. Steps 2 to 5 are repeated until the desired model performance is achieved.

In our experiments, we perform 100 training iterations to train each model. Note that we use a random forest because it is robust to training with a small dataset and can be quickly updated with new data. In addition, only one query is labeled in each iteration instead of a batch of queries since our goal is to reduce annotation effort.

3. QUERY STRATEGIES

The most critical part of AL is defining the informativeness of each unlabeled example and querying instances accordingly. While there are several different families of query strategies, including *certainty-based sampling* [20], *query-by-committee* [21], *expected error reduction* [22], and *medoid-based AL* [18], we focus on the certainty-based sampling strategy since there is a direct relation between "certainty" and the prediction probability given by a random forest. We systematically evaluate three variations of certainty-based sampling: *least-confidence sampling*, *semi-supervised active learning*, and *alternating confidence sampling*.

3.1. Least-confidence sampling (LC)

Least-confidence sampling (LC) is a popular certainty-based sampling strategy [20, 23] in which the model queries the instance that it is least confident about how to label. In the binary classification case, it is equivalent to two other common certainty-based sampling: *maximum entropy* [24] and *smallest margin* [25]. It is also equivalent to querying the instance whose prediction probability is closest to the decision threshold T_d in a random forest model. In our experiments, we evaluate two methods to determine T_d for sampling: 1) fixed threshold, setting $T_d = 0.5$, 2) varied threshold, setting T_d at each iteration to the value that gives the best validation performance. This relatively unrealistic condition is setup to see if decision threshold significantly affects AL model performance.

3.2. Semi-supervised active learning (SSAL)

Previous works have demonstrated the benefit of using AL combined with *semi-supervised learning* (SSL) [18, 19], a technique that uses a pre-trained model on a smaller set of labeled data to automatically annotate a larger set of unlabeled data. Here we use the semi-supervised active learning strategy similar to [19]. The model is updated twice in each training iteration, once by AL and once by SSL. In the AL stage, the model goes through one regular AL training iteration using the LC query strategy with a fixed threshold T_d. Then in the SSL stage, the model predicts labels for the unlabeled data pool. Predicted labels with confidence higher than a set threshold T_{SSL} are then added to the labeled data pool for retraining the model. In our experiments, the model was always more confident about negative predictions. To counter this prediction bias, we select a balanced subset of the confident predicted labels. We evaluate two different T_{SSL} in our experiments: 0.95 and 0.98.

3.3. Alternating confidence sampling (AC)

While LC is effective at selecting the most informative instance, it is also sensitive to annotation error. For example, if an annotator mislabels an example during training, the model may subsequently misclassify some examples with high confidence, making it difficult to correct the error since the instances with high confidence will not be queried by LC. To make the model more robust to such errors, we propose a new query strategy: *alternating confidence sampling* (AC). With this strategy, the model not only queries the instances about which it is least confident, but it also queries instances about which it is very confident. In our setting, the model occasionally queries instances with high confidence by randomly sampling from the set with prediction probabilities higher than a threshold $T_{HC} = 0.85$. This method accounts for the possibility of the model making confident errors in its prediction during training and allows the human annotator to check and fix those errors. Two frequencies f_{HC} for sampling high-confidence instances are tested: drawing one high-confidence example 1) every 5 iterations and 2) every 10 iterations. All other iterations use LC sampling a with fixed threshold T_d.

4. EXPERIMENTS

To develop a sound classification model to detect our artifact noise, we conducted a set of experiments using the AL framework described in Section 2. In Section 4.2, we present a systematic evaluation of the query strategies described in Section 3 using one expert annotator, who has the domain knowledge and experience of identifying target sound. In Section 4.3, we compare the best strategy from the evaluation to baseline methods. In Section 4.4, we investigate whether models trained with non-expert annotators can achieve similar performance as our expert-trained models. And lastly, in Section 4.5, we qualitatively evaluate our best AL-trained model using an unlabeled dataset to investigate the generalizability of the model.

4.1. Data preparation and experimental setup

To train models with AL we need an unlabeled dataset as a sampling pool, and to train models without AL (i.e., our reference baseline) we need a labeled dataset. We created our labeled dataset by manually labeling 300 positive (i.e., with the artifact noise present) and 300 negative one-second audio clips which were evenly sampled from 15 different SONYC sensors. The total 600 labeled clips were conditionally partitioned by sensor ID into balanced training, validation, and test sets with a 3:1:1 ratio for 5-fold cross-validation. Within each training set, two clips, one positive and one negative, were randomly drawn as the initial training data for the AL experiments. We kept the initial training set extremely small to more clearly see the performance trends when comparing models. The unlabeled dataset was built by randomly sampling 100,000 one-second audio clips from the remaining sensors.

We extracted input features from both datasets using a pre-trained VGGish audio model [26]. The VGGish audio classification model was trained over the YouTube-8M dataset and can be used as a pre-trained feature embedding model to generate a compact, discriminative 128-dimensional feature vector.

For each experiment, we ran 5-fold cross validation to estimate model performance, using F-measure as the evaluation metric. We tuned the decision threshold of the random forest model after each training iteration using mean model performance on the validation set, and we reported the mean model performance on the test set.

4.2. Comparing query strategies

To compare query strategies and their variations, we trained separate models for each query strategy variation using the training framework described in Section 2. For each model, the mean and maximum F-measure within 100 training iterations are reported in Ta-

Query Strategy	Mean	Max	Iteration
LC, fixed T_d	0.904	0.947	77
LC, varied T_d	0.891	0.924	86
SSAL, $T_{SSL} = 0.95$	0.857	0.910	24
SSAL, $T_{SSL} = 0.98$	0.892	0.928	56
AC, $f_{HC} = 1/5$	0.897	0.939	48
AC, $f_{HC} = 1/10$	**0.913**	**0.962**	90

Table 1. Mean F-measure, maximum F-measure, and the iteration where maximum is reached from the classifiers trained with different AL query strategies.

ble 1, as well as the iteration at which the maximum performance is reached.

First, we compared LC with fixed T_d and varied T_d. The results in Table 1 (first two rows) show that training with fixed T_d gives better model performance. We noticed during training that queries sampled with varied T_d are often biased towards negative examples compared to those sampled with fixed T_d. Classifiers trained with varied T_d quickly learn about negative examples and become much more confident in their negative predictions. This lowers the effective decision threshold and reinforces the situation since new queries are subsequently sampled using the updated decision threshold. Fixing the decision threshold at 0.5 for sampling forces the model to learn about positive and negative examples equally and return more balanced queries.

Second, we compared SSAL with two different T_{SSL}. The result shown in the Table 1 (middle two rows) indicates that adding an SSL stage does not improve model performance in our case. Higher T_{SSL} values result in better performance, but when testing on T_{SSL} higher than 0.98, the positive predictions do not have high enough confidence to update the training set in the SSL stage. Therefore, it is equivalent to AL without SSL. It has been shown in previous work that T_{SSL} needs to be carefully tuned for specific tasks in order to improve performance [19]. SSL also generally requires an accurate initial model, otherwise adding too many false predictions can significantly affect model performance. This is especially crucial in our case where the initial model is trained on only two labeled examples, and only one ground truth example is given in each iteration.

Lastly, we compare AC with two different sampling frequencies for high-confidence instances. Table 1 (bottom two rows) shows that sampling one high-confidence instance every 10 iterations achieves better model performance than sampling it every 5 iterations. While every 10 iterations seems frequent enough to correct errors, every 5 iterations may be too frequent, reducing the average information per label and hindering learning.

Overall, we found the best model performance when training using the AC strategy with $f_{HC} = 1/10$. Our newly proposed certainty-based strategy enables annotators to correct possible mistakes from the model, which helps the model generalize better on unseen data. We use this top-performing AC strategy to train the AL models in all of the remaining experiments.

4.3. Comparing AL and baseline methods

Next, we compare AL to two baseline methods. We use random sampling as the first baseline method, using the same framework as AL: two initial training examples and 100 training iterations. At each iteration, the query is randomly sampled from the unlabeled data pool instead of using certainty-based sampling. We also trained a reference model on the entire labeled set under the same partitioning and

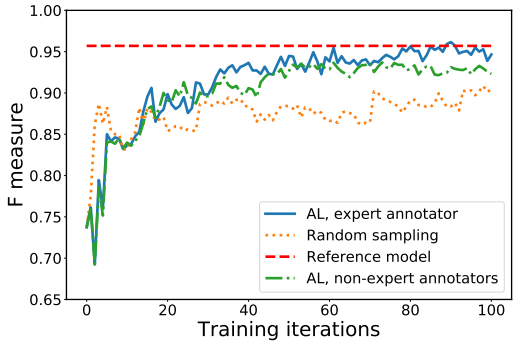


Fig. 1. Model performance at each training iteration for different training methods.

Training Method	Mean	Max	Iteration	Positive Queries
AL, expert annotator	0.913	**0.962**	**90**	**42%**
Random Sampling	0.875	0.908	98	3%
Reference Model	0.957	0.957	-	-
AL, non-expert annotators	0.903	0.937	73	43%

Table 2. Mean F-measure, maximum F-measure, the iteration where maximum is reached, and the percentage of positive queries from the classifiers trained with AL and baseline methods.

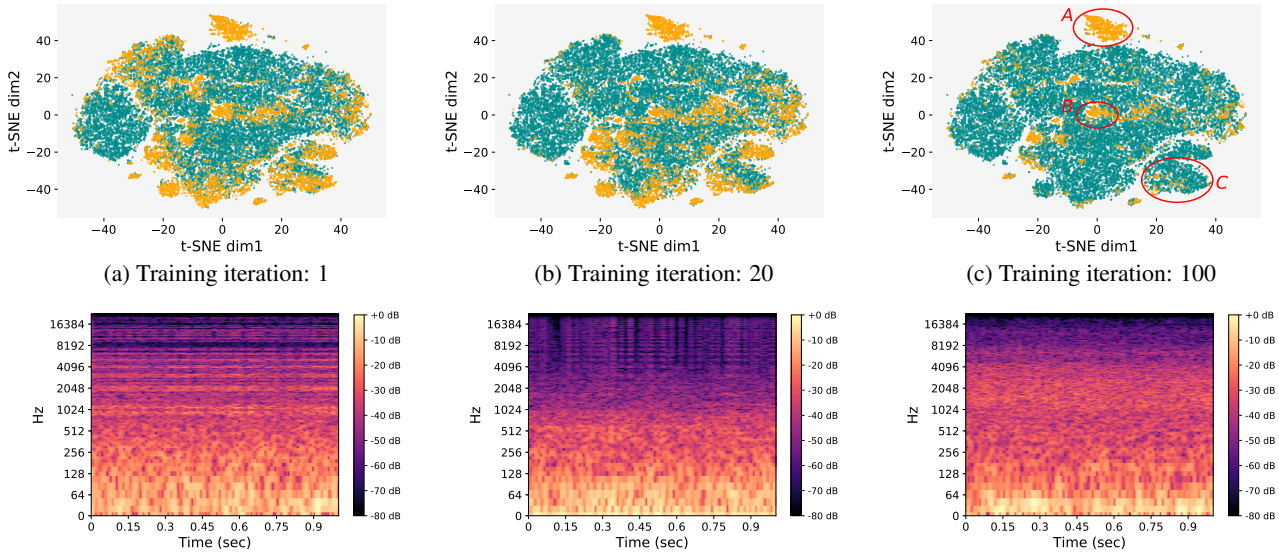
cross-validation setup to demonstrate the best possible performance using all the 600 labeled examples.

Figure 1 shows model performance at each training iteration for different training methods. The model trained with AL and expert annotator outperforms the model trained with random sampling, and soon reaches a comparable performance to the reference model after 50 training iterations. Table 2 shows the mean and maximum F-measures within 100 iterations, and the iteration when the best performance occurred for each model. The classifier trained with AL and expert annotator reaches its best performance at iteration 90, which means the model is only trained on 92 labeled examples, including the two initial training examples. Although the classifier was trained on far fewer labeled examples than the reference model (600 examples), AL effectively allows the model to utilize information from a much larger unlabeled data pool, which produces more efficient training.

Table 2 also shows the percentage of positive examples in the 100 queries during training. AL returns a much more balanced set of queries than random sampling. More balanced queries lead to a more efficient annotation process and a more robust model.

4.4. Annotator generalizability

We recruited 15 non-expert annotators to test whether models trained by non-experts could achieve similar performance as those trained by expert annotator. Each participant annotated one fold, resulting in a total of three annotations per fold. We calculated the optimal decision threshold per iteration based on best mean F-measure across all 15 models (3 models per fold, 5 folds in total). Results shown in Figure 1 demonstrate that the models trained with expert and non-expert annotators result in similar performance trends. Although the non-expert trained models never reach reference performance as the expert-trained models do, they still perform better than the expert-trained random model. Table 2 shows the average percentage of



(a) Training iteration: 1 (b) Training iteration: 20 (c) Training iteration: 100

(d) Power spectrogram, A (e) Power spectrogram, B (f) Power spectrogram, C

Fig. 2. (a)–(c) t-SNE plots colored by predicted labels from models at training iteration 1, 20, and 100. Orange: predicted positive; Green: predicted negative. (d)–(f) Power spectrogram from one of the excerpts sampled from region A, B, and C.

positive queries returned by AL during the 100 training iterations, also showing a similar trend to that of expert annotator.

4.5. Qualitative evaluation on unlabeled data

To investigate the generalizability of the model, we used our best-performing noise classifier trained with AL to predict labels for the unseen, unlabeled data. Since there is no ground truth, we used t-SNE [27] to visualize the results in Figures 2(a)–(c). For each plot, we used t-SNE to project 20,000 data points (sampled from the unlabeled data pool) to a two-dimensional space, using VGGish features as the t-SNE input representation and a t-SNE perplexity of 30. The color of each data point is assigned based on its label predicted by the classifier: orange for positive, green for negative. Figures 2(a)–(c) visualize model predictions at three different training iterations: 1, 20, and 100. The distribution of predicted labels changes while training iterations increase, and the positive predictions converge to a small number of clusters.

We listened to 10 randomly selected audio excerpts from the 3 regions indicated in Figure 2(c). A spectrogram from one of the excerpts in each region is also shown respectively in Figures 2(d)–(f). All of the excerpts sampled from regions A and B contain the target artifact noise. The noise in region A is very prominent, resulting in the clear harmonic bands in upper midrange on the spectrogram as shown in Figure 2(d). The noise in region B is softer than the noise in region A and is accompanied by other environmental sounds in the lower frequency range. Its spectrogram in Figure 2(e) shows fainter harmonic bands with energy focused on lower frequencies. Excerpts sampled from region C do not contain the artifact noise—they are environmental sounds in a similar frequency range as the noise. The spectrogram in Figure 2(f) shows that although there is no harmonic pattern, there is also energy distributed in the upper midrange. This could explain why there are many positive predictions in region C in early training iterations.

This qualitative evaluation shows how the classifier learns and refines its prediction during the AL process. It also shows that the AL-trained classifier is generalizable to unseen data and can identify target artifact noise within different contexts, enabling us to explore the target noise in the SONYC data.

5. CONCLUSION

In this work, we used active learning to reduce annotation costs for an idiosyncratic sound classification task for which existing datasets were not usable. We developed a target artifact noise classifier with unlabeled audio data by using a pool-based active-learning framework with actual human annotators in the loop. In doing so, we proposed a new certainty-based query-sampling strategy, *alternating confidence sampling*, and found that it improved model performance over two other certainty-based strategies. This new query sampling strategy allows annotators to check and fix classification errors by occasionally sampling high-confidence instances. We found that models trained with the proposed strategy outperform a baseline model trained with random sampling, and with far fewer labeled training examples, they reach performance comparable to the reference model. Using active learning, the artifact noise classifier reached an F-measure of 0.962 after training on only 92 labeled examples. We also evaluated our sampling strategy by training a model with 15 non-expert annotators and showed that it performed similarly to the model trained with one expert annotator. Lastly, we qualitatively showed that our noise classifier trained with active learning is generalizable to unseen data and can identify the target noise within different contexts. This work showed that active learning can improve training efficiency and significantly reduce annotation effort in a real-world scenario. Future works include analyzing the effectiveness of alternating confidence sampling quantitatively, and experimenting more query strategies with non-expert annotators. We hope this work will encourage others to utilize active learning when developing costly problem-specific datasets.

6. REFERENCES

[1] Paul Gaunard, Corine Ginette Mubikangiey, Christophe Couvreur, and Vincent Fontaine, "Automatic classification of environmental noise events by hidden markov models," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 1998, pp. 3609–3612.

[2] Antonio J. Torija and Diego P. Ruiz, "Automated classification of urban locations for environmental noise impact assessment on the basis of road-traffic content," *Expert Syst. Appl.*, vol. 53, no. C, pp. 1–13, July 2016.

[3] Herve Glotin, Julien Ricard, and Randall Balestriero, "Fast Chirplet Transform to Enhance CNN Machine Listening - Validation on Animal calls and Speech," 2016.

[4] Lior Shamir, Carol Yerby, Robert Simpson, Alexander M. von Benda-Beckmann, Peter Tyack, Filipa Samarra, Patrick Miller, and John Wallin, "Classification of large acoustic datasets using machine learning and crowdsourcing: Application to whale calls," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 953–962, Feb 2014.

[5] Ichiro Fujinaga and Karl MacMillan, "Realtime recognition of orchestral instruments.," in *ICMC*, 2000.

[6] Janet Marques and Pedro J Moreno, "A study of musical instrument classification using gaussian mixture models and support vector machines," *Cambridge Research Laboratory Technical Report Series CRL*, vol. 4, 1999.

[7] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[8] Juan Pablo Bello, Cláudio T. Silva, Oded Nov, R. Luke DuBois, Anish Arora, Justin Salamon, Charles Mydlarz, and Harish Doraiswamy, "SONYC: A system for the monitoring, analysis and mitigation of urban noise pollution," *CoRR*, vol. abs/1805.00889, 2018.

[9] Cynthia A. Thompson, Mary Elaine Califf, and Raymond J. Mooney, "Active learning for natural language parsing and information extraction," in *Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, Bled, Slovenia, June 1999, pp. 406–414.

[10] Andrew Kachites Mccallum, "Employing em in pool-based active learning for text classification," in *In Proceedings of the 15th International Conference on Machine Learning*. 1998, pp. 350–358, Morgan Kaufmann.

[11] D. Tuia, F. Ratle, F. Pacifici, M. F. Kanevski, and W. J. Emery, "Active learning methods for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 7, pp. 2218–2232, July 2009.

[12] Xin Li and Yuhong Guo, "Adaptive active learning for image classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[13] D. Hakkani-Tr, G. Riccardi, and A. Gorin, "Active learning for automatic speech recognition," in *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, May 2002, vol. 4, pp. IV–3904–IV–3907.

[14] Dong Yu, Balakrishnan Varadarajan, Li Deng, and Alex Acero, "Active learning and semi-supervised learning for speech recognition: A unified framework using the global entropy reduction maximization criterion," *Computer Speech & Language*, vol. 24, no. 3, pp. 433 – 444, 2010, Emergent Artificial Intelligence Approaches for Pattern Recognition in Speech and Language Processing.

[15] Bongjun Kim and Bryan Pardo, "A human-in-the-loop system for sound event detection and annotation," *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 2, pp. 13:1–13:23, June 2018.

[16] Kun Qian, Zixing Zhang, Alice Baird, and Bjrn Schuller, "Active learning for bird sound classification via a kernel-based extreme learning machine," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1796–1804, 2017.

[17] Kun Qian, Zixing Zhang, Alice Baird, and Björn Schuller, "Active learning for bird sounds classification," *Acta Acustica united with Acustica*, vol. 103, no. 3, pp. 361–364, 2017.

[18] Z Shuyang, T Heittola, and T Virtanen, "Active learning for sound event classification by clustering unlabeled data," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 751–755.

[19] Wenjing Han, Eduardo Coutinho, Huabin Ruan, Haifeng Li, Björn Schuller, Xiaojie Yu, and Xuan Zhu, "Semi-Supervised Active Learning for Sound Classification in Hybrid Learning Environments.," *PLoS ONE*, vol. 11, no. 9, pp. 1–23, Sep 2016.

[20] David D. Lewis and William A. Gale, "A sequential algorithm for training text classifiers," *CoRR*, vol. abs/cmp-lg/9407020, 1994.

[21] H. S. Seung, M. Opper, and H. Sompolinsky, "Query by committee," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, New York, NY, USA, 1992, COLT '92, pp. 287–294, ACM.

[22] Nicholas Roy and Andrew McCallum, "Toward optimal active learning through sampling estimation of error reduction," in *Proceedings of the Eighteenth International Conference on Machine Learning*, San Francisco, CA, USA, 2001, ICML '01, pp. 441–448, Morgan Kaufmann Publishers Inc.

[23] David D. Lewis and Jason Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *In Proceedings of the Eleventh International Conference on Machine Learning*. 1994, pp. 148–156, Morgan Kaufmann.

[24] Claude Elwood Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[25] Tobias Scheffer, Christian Decomain, and Stefan Wrobel, "Active hidden markov models for information extraction," in *Advances in Intelligent Data Analysis*, Frank Hoffmann, David J. Hand, Niall Adams, Douglas Fisher, and Gabriela Guimaraes, Eds., Berlin, Heidelberg, 2001, pp. 309–318, Springer Berlin Heidelberg.

[26] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[27] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.