ACTIVE LEARNING OF NON-SEMANTIC SPEECH TASKS WITH PRETRAINED MODELS

Harlin Lee[1], *Aaqib Saeed*[2], *Andrea L. Bertozzi*[1]

[1]University of California Los Angeles, Los Angeles, CA, USA
[2]Philips Research, Eindhoven, The Netherlands

ABSTRACT

Pretraining neural networks with massive unlabeled datasets has become popular as it equips the deep models with a better prior to solve downstream tasks. However, this approach generally assumes that for downstream tasks we have access to annotated data of sufficient size. In this work, we propose ALOE, a novel system for improving data- and label-efficiency of non-semantic speech tasks with active learning (AL). ALOE uses pretrained models in conjunction with active learning to incrementally label data and learn classifiers for downstream tasks, thereby mitigating the need to acquire labeled data beforehand. We demonstrate the effectiveness of ALOE on a wide range of tasks, uncertainty-based acquisition functions, and model architectures. Training a linear classifier on top of a frozen encoder with ALOE is shown to achieve performance similar to several baselines that utilize the entire labeled data.

***Index Terms*—** active learning, audio, non-semantic speech, self-supervised learning, transfer learning

1. INTRODUCTION

Deep neural networks require a large amount of well-annotated training data to generalize well. In the real world, access to labeled datasets is limited, and collecting abundant examples requires significant investment in terms of both finance and time. Further, the expertise required to collect high-quality labels can be limited in domains like health monitoring. Pretraining neural networks with massive unlabeled datasets have become a popular choice to tackle this issue, as it equips the deep models with a better prior for downstream problems. In particular, the self-supervision strategy has shown to achieve tremendous success across data modalities including audio and speech domains [1, 2, 3]. Self-supervised learning tasks a neural network to solve an auxiliary learning problem for which supervision can be acquired from the unlabeled input itself, and this pushes the model to learn useful representations from unlabeled data. One can then use the pretrained model (also known as an encoder) either as a fixed feature extractor or as initialization in transfer learning.

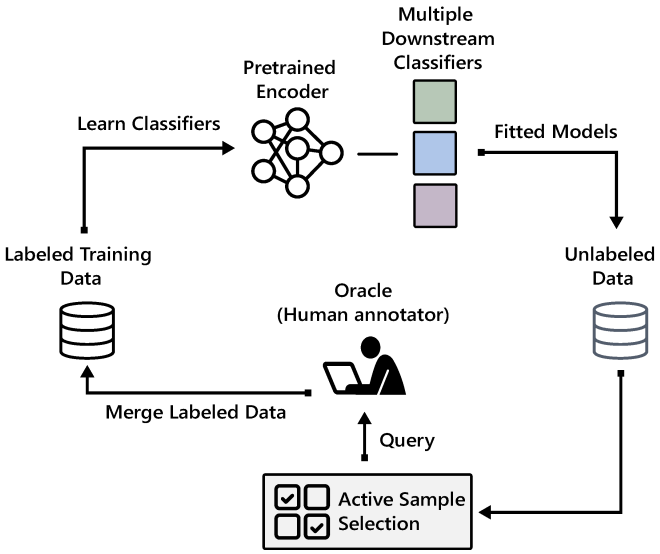


Fig. 1: Illustration of ALOE: active learning with a pretrained model. The encoder parameters are frozen, allowing the use of the same encoder across multiple downstream tasks.

Still, the downstream learning tasks require sufficient annotated samples for one to train a classifier on top of the encoder from the self-supervised or unsupervised pretraining phase. We note that for some tasks, pretrained models may require only a small percentage (e.g. 10%-20%) of labeled data to match the performance of a supervised model, but the number of labeled instances can still be huge. As mentioned previously, data labeling is cumbersome and expensive, and it is unclear a priori for which instances we should get annotations, as not all examples carry useful information for learning. In comparison, getting unlabeled data for the desired task is much easier.

In light of these observations and challenges, we design ALOE (**A**ctive **L**earning of n**o**n-s**e**mantic speech tasks with pretrained models) to improve the label- and data-efficiency of non-semantic speech tasks [1]. Figure 1 describes how ALOE exploits pretrained models with *active learning* (AL) for gradual labeling of data and learning classifiers for downstream tasks. To the best of our knowledge, our method is a first

This work was partially supported by NSF grants DMS-1952339 and DMS-2152717, and NGA award HM0476-21-1-0003. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NGA. Emails: harlin@math.ucla.edu, aaqib.saeed@philips.com, bertozzi@math.ucla.edu.

attempt at improving label efficiency for non-semantic speech tasks with AL. In principle, AL provides a systematic way to procure labels for the most informative instances, e.g. a class for which a model is highly uncertain [4, 5]. AL framework comprises of three stages: a) acquisition—selecting relevant examples, b) annotation—getting ground-truth labels from a human annotator, and c) retraining—learning model on acquired and existing labeled instances. In practice, these steps are time-consuming and computationally expensive because a model has to be trained many times during the AL cycle. We improve the efficiency of AL with the power of semantic representations from pretrained models while also enhancing its acquisition capability, which is otherwise limited when training a model from scratch. Furthermore, we are privy to a better view of the annotation process with human-in-the-loop.

We demonstrate the effectiveness of our approach on a broad range of tasks, uncertainty-based acquisition functions, and model architectures. Training a linear classifier on top of a frozen encoder with AL uses only a fraction of the examples yet results in a similar performance as several baselines that utilize the entire labeled data. ALOE provides an end-to-end system for exploiting pretrained models more effectively and mitigates the need to acquire large labeled data beforehand.

2. METHODOLOGY

We propose to use a pretrained (unsupervised or self-supervised) model with active learning (AL) to improve the data- and label-efficiency of deep models in non-semantic speech tasks. Importantly, ALOE has a shared fixed pretrained encoder with separate shallow classifiers for each end-task. Our approach leverages generic speech representations learned from massive amounts of unlabeled data and identifies key samples from an intended end-task that are to be labeled by a human annotator (i.e. an oracle) depending on their uncertainty scores.

We consider an AL setup with pool-based acquisition as commonly studied in the literature [5, 6]. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be d-dimensional speech data samples, $\mathcal{Y}$ be labels of a non-semantic speech classification task, and $\mathcal{F}_\theta : \mathcal{X} \rightarrow \mathbb{R}^m$ be a pretrained encoder, whose embedding dimension $m \ll d$. Given some labeled instances $\mathcal{D}_l \subset \mathcal{X} \times \mathcal{Y}$ and a large amount of unlabeled data $\mathcal{D}_u \subset \mathcal{X}$ with $|\mathcal{D}_l| \ll |\mathcal{D}_u|$, the aim is to incrementally label samples in $\mathcal{D}_u$ to minimize the cost of annotation, better understand the annotation process, and ultimately improve model generalization with few labels.

For each task, we initiate the AL process by acquiring $\mathcal{D}_l$ to train a shallow classifier $\mathcal{G}_\omega : \mathbb{R}^m \rightarrow \mathbb{R}^{|\mathcal{Y}|}$ on top of the m-dimensional representations from $\mathcal{F}_\theta$. $\mathcal{G}_\omega$ is a fully-connected linear model with softmax activation such that it outputs the probability an input $x \in \mathcal{X}$ belongs to each class $y \in \mathcal{Y}$:

$$\mathcal{G}_\omega(\mathcal{F}_\theta(x)) = [P(y_1|x), P(y_2|x), \ldots, P(y_{|\mathcal{Y}|}|x)]. \quad (1)$$

Once $\mathcal{G}_\omega$ is trained on $\mathcal{D}_l$, model weights ω is fixed and predictions for each instance in $\mathcal{D}_u$ are generated. ALOE then uses an acquisition function $\mathcal{A} : \mathbb{R}^{|\mathcal{Y}|} \rightarrow \mathcal{X}$ to select the most valuable examples of $\mathcal{D}_u$ for label generation within a certain budget. Once these labels are acquired, this batch of examples is merged into the existing $\mathcal{D}_l$. It is followed by retraining $\mathcal{G}_\omega$ using the updated $\mathcal{D}_l$, and we repeat this process for a fixed number of AL acquisition steps. We note that instead of keeping $\mathcal{F}_\theta$ fixed during AL, one can fine-tune the entire model (i.e. $\mathcal{F}_\theta$ and $\mathcal{G}_\omega$) end-to-end, but this process can be very time-consuming as retraining has to be performed after each label acquisition step, and the learnable parameters also increase significantly. Similarly, the feature extraction model $\mathcal{F}_\theta$ will become task-specific; it may lose its generalizable nature and not perform well for other downstream tasks.

For the acquisition function $\mathcal{A}$, we use an uncertainty sampling-based method called *smallest margin*, which is relatively simple yet shown to be effective in identifying informative examples for annotation [5, 7]. Specifically, $\mathcal{A}$ selects

$$x^* = \arg\min_{x \in D_u} P(y_{(1)}|x) - P(y_{(2)}|x), \quad (2)$$

where $P(y_{(i)}|x)$ is the ith largest probability in (1), e.g. the predicted label for the sample is $y_{(1)}$. That is, (2) chooses samples that have a "very close second prediction." Thus the model is less certain about them, and acquiring their labels from the oracle will provide more information about where the decision boundary of the updated model should be.

3. EXPERIMENTS

For $\mathcal{F}_\theta$, we leverage publicly available pretrained models from the TRILLsson family, in particular, the Audio Spectrogram Transformer (AST) [9]. TRILLsson models are trained via knowledge distillation using large-scale unlabeled data with a massive self-supervised conformer-based teacher model. They provide state-of-the-art performance on a broad spectrum of downstream non-semantic speech tasks while being smaller than the teacher model. All pretrained models mentioned in this paper are from TensorFlow Hub[1]. $\mathcal{F}_\theta$ embeds the entire audio clip into a single vector with dimension $m = 1024$.

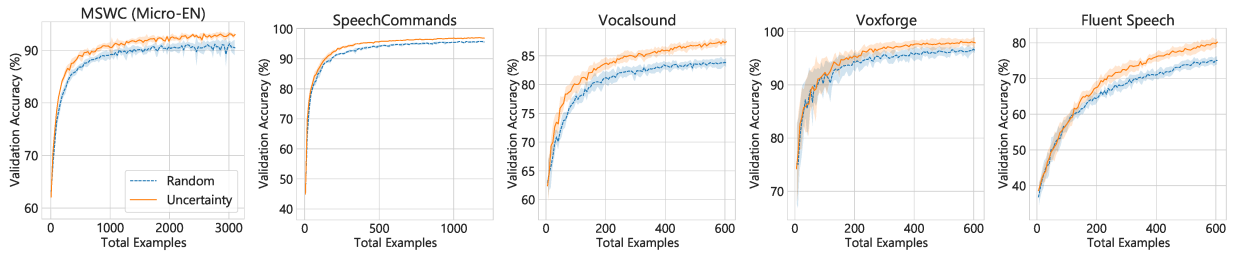
For each task, we report results aggregated from 10 independent runs of the experiment. For each run of the experiment, there are 100 AL acquisition steps, and at each AL acquisition step, $\mathcal{G}_\omega$ is trained for 100 epochs with Adam and a learning rate of 0.001. As the size of ω is quite small, the added benefit of our approach is that AL executes at a rapid pace. $\mathcal{D}_l$ is initially seeded with 5 labeled examples per class. At each round of AL, *class-aware* sampling selects one example per class based on the predicted labels, while *class-agnostic* sampling picks a single example from all of $\mathcal{D}_u$. Note that this means they acquire $|\mathcal{Y}|$ labels and 1 label at each step, respectively.

Several publicly available datasets are used to evaluate audio recognition models with active learning. Specifically, we focus on non-semantic speech tasks [1, 9] as the feature

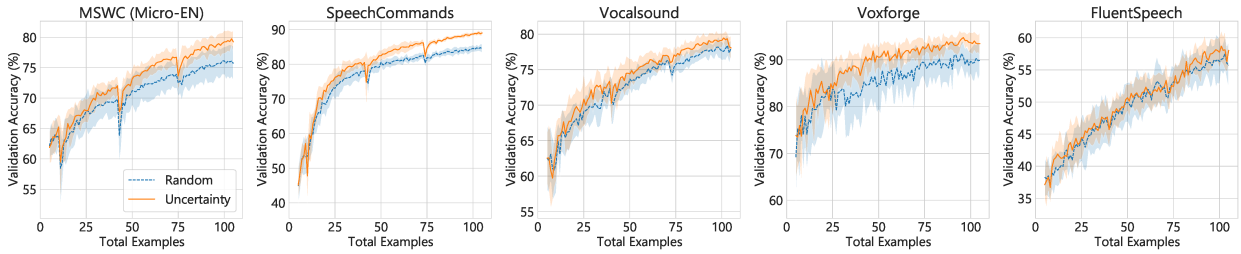
[1] https://tfhub.dev/

Table 1: Comparison of *test set* accuracy (%) of our approach with other baselines on different end-tasks. ALOE achieves similar recognition rate with class-aware sampling, while using several folds less labeled examples. The baseline results are from [8], where available; else we train a linear classifier on time-averaged representations using published models.

Dataset	**TRILL [1]**	**TRILL-Dist [1]**	**FRILL [8]**	**TRILLsson [9]**	**ALOE (Ours)**			
					Class-Aware		*Class-Agnostic*	
					Random	*Uncertainty*	*Random*	*Uncertainty*
MSWC (Micro-EN)	81.3	74.4	79.1	93.7	90.6	93.0	76.0	79.7
SpeechCommands	81.9	80.2	79.7	96.4	92.6	94.9	70.3	78.6
Vocalsound	88.2	85.8	86.7	91.1	84.9	88.2	79.0	79.1
Voxforge	84.5	80.0	76.9	99.6	98.4	99.2	94.4	96.2
FluentSpeech	69.3	62.3	64.9	97.5	83.5	87.5	60.8	62.6



(a) Class-Aware Sampling



(b) Class-Agnostic Sampling

Fig. 2: Uncertainty sampling outperforms random sampling at every AL round, as measured by *validation set* accuracy (%).

extraction models we use are pretrained on speech datasets. We perform experiments on MSWC (Micro-English) [10] and SpeechCommands [11] for keyword spotting, comprising of (96,099 samples, 31 classes) and (100,503 samples, 12 classes), respectively. For spoken language identification, we use Voxforge [12] containing 176,436 audio clips from 6 different languages. Similarly, for human vocal sounds monitoring, e.g. coughing and sneezing, we use Vocalsound [13] that has 21,024 audio examples from 6 classes. Finally, we evaluate on the task of action detection (30,043 samples, 6 classes) using FluentSpeech [14]. In all cases, we use the standard train, validation, and test splits provided with the original datasets, with audio sampled at 16kHz.

We compare ALOE against many baseline models. First, we investigate several other self-supervised or distillation-based methods. In TRILL [1], TRILL-Dist [1], FRILL [8], and TRILLsson (AST), the linear classifiers are trained using the entire labeled data from the downstream task. These baselines, particularly the AST model, establish an upper-bound performance that can be achieved with a simple classifier trained with all labeled examples from a specific task. Additionally, we explore other uncertainty sampling methods [5, 6] such as largest margin, least confidence, entropy, and norm, as well as random acquisition, i.e. sampling from a uniform distribution on the pool $\mathcal{D}_u$ at each acquisition step. Lastly, we compare the performance of AST as $\mathcal{F}_\theta$ to other neural network architectures such as ResNet-50 and EfficientNetv2 (B3) from the TRILLsson family [9]. Unless otherwise specified, ALOE in this work uses the TRILLsson (AST) pretrained model with class-aware uncertainty-based sampling, specifically the smallest margin sampling.

4. RESULTS AND ANALYSIS

Table 1 compares the *test set* performance of ALOE after 100 rounds of AL to other self-supervised or distillation-based models. As expected, TRILLsson has the highest recognition rate (often above 95%) across a wide range of non-semantic

Table 2: Uncertainty sampling provides better performance on *test set* than random sampling across different architectures.

Dataset	ResNet-50		EfficientNet-v2 (B3)		Spectrogram Transformer (AST)	
	Random	*Uncertainty*	*Random*	*Uncertainty*	*Random*	*Uncertainty*
MSWC (Micro-EN)	91.4± 0.70	**92.8± 0.58**	88.2± 1.11	**91.9± 0.56**	90.6± 1.48	**93.0± 0.50**
SpeechCommands	92.4± 1.37	**93.1± 0.89**	90.4± 0.58	**92.5± 0.96**	92.6± 1.70	**94.9± 1.12**
Vocalsound	82.4± 0.92	**85.6± 0.80**	84.2± 0.73	**87.6± 0.38**	84.4± 0.94	**88.2± 0.35**
Voxforge	95.7± 0.41	**97.4± 0.23**	97.0± 0.29	**98.4± 0.07**	98.4± 0.17	**99.2± 0.06**
FluentSpeech	79.0± 1.90	**82.7± 1.07**	80.9± 1.95	**85.1± 0.96**	83.5± 1.14	**87.5± 0.96**

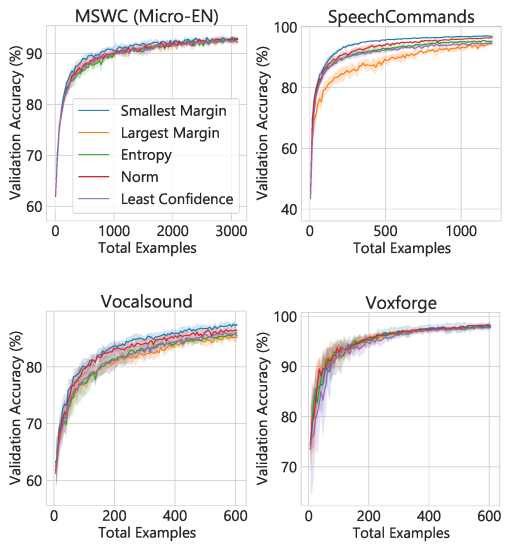


Fig. 3: Different uncertainty sampling methods paired with pretrained model perform similarly well on *validation set*.

speech classification tasks. But ALOE with class-aware uncertainty-based sampling achieves accuracy comparable to TRILLsson while using several times fewer labeled examples. For example, ALOE arrives at 99.2% accuracy on Voxforge after only acquiring labels for 600 samples, surprisingly close to the 99.6% achieved by the upper-bound model that used the entire labeled dataset. This suggests that with the pretrained TRILLsson (AST) model as $\mathcal{F}_\theta$, the linear classifier $\mathcal{G}_\omega$ needs only a couple hundred labeled examples to be near perfect in this task. The implications are impressive: when deployed in the real world, this human-in-the-loop approach may eliminate the cost of acquiring *excess* labels in the first place, taking the guesswork out of data collection. We note that ALOE gets to within 0.7% to 2.9% of TRILLsson's accuracy in MSWC, SpeechCommands, and Vocalsound as well, supporting the notion that AL with pretrained models can improve data- and label-efficiency in many non-semantic speech tasks.

Next, we discuss the effect of uncertainty-based sampling in ALOE. Comparing the neighboring random and uncertainty columns in Table 1 shows that although the extent may differ (0.1% to 8.3%), the uncertainty-based sampling method often leads to a higher test set accuracy for different tasks in both class-aware and class-agnostic settings. Figure 2 confirms that this is also observed during the AL phase, as measured by validation set accuracy. These results indicate that ALOE successfully selects examples that are more informative than random picks. Within uncertainty-based sampling, the different acquisition functions perform similarly well; see Figure 3. Furthermore, we explore the effect of AST as the default pretrained model by switching it out with different neural network architectures. Results on test set are summarized in Table 2. The different pretrained TRILLsson models performed similarly, but AST showed a slight advantage. We point out from Table 2 that uncertainty sampling performs better than random sampling across different architectures as well.

Recall that the class-aware setting selects more examples than the class-agnostic setting at every acquisition step, and therefore these two cannot be directly compared to each other. Instead, users should choose one of the approaches based on their AL labeling budgets and computational costs. They also need to choose the total number of AL acquisition steps. While we fixed that number at 100 for all experiments in this work, users can easily monitor validation accuracy as in Figure 2 to determine different stopping points. For instance, they may decide to do early stopping when the performance plateaus (e.g. SpeechCommands with class-aware sampling) or let the model train longer (e.g. Fluent Speech).

To reiterate, ALOE does not fine-tune the encoder during AL as our motivation is to use a generic feature extractor for more than one task, and updating the parameters with few samples can have catastrophic consequences for the learned representations from large-scale data. However, we note that when a sizeable portion of the data is labeled at the end of the AL phase, it may be used for end-to-end model fine-tuning.

5. CONCLUSIONS

We proposed ALOE, a novel approach to improve label- and sample-efficiency of non-semantic speech tasks with active learning. It provides an end-to-end system that exploits pretrained models more effectively, consequently mitigating the need to prepare large labeled data for downstream tasks. We plan to extend this framework to graph-based active learning.

6. REFERENCES

[1] Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Félix de Chaumont Quitry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv, "Towards Learning a Universal Non-Semantic Representation of Speech," in *Proc. Interspeech 2020*, 2020, pp. 140–144.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12449–12460, 2020.

[3] Aaqib Saeed, David Grangier, and Neil Zeghidour, "Contrastive learning of general-purpose audio representations," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3875–3879.

[4] David D. Lewis and William A. Gale, "A sequential algorithm for training text classifiers," 1994, pp. 3–12, Springer-Verlag.

[5] Burr Settles, "Active learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 6, no. 1, pp. 1–114, 2012.

[6] Kevin Miller, Jack Mauro, Jason Setiadi, Xoaquin Baca, Zhan Shi, Jeff Calder, and Andrea L Bertozzi, "Graph-based active learning for semi-supervised classification of sar data," in *Algorithms for Synthetic Aperture Radar Imagery XXIX*. SPIE, 2022, vol. 12095, pp. 126–139.

[7] David D Lewis and Jason Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Machine learning proceedings 1994*, pp. 148–156. Elsevier, 1994.

[8] Jacob Peplinski, Joel Shor, Sachin Joglekar, Jake Garrison, and Shwetak Patel, "Frill: A non-semantic speech embedding for mobile devices," *arXiv preprint arXiv:2011.04609*, 2020.

[9] Joel Shor and Subhashini Venugopalan, "Trillsson: Distilled universal paralinguistic speech representations," *arXiv preprint arXiv:2203.00236*, 2022.

[10] Mark Mazumder, Sharad Chitlangia, Colby Banbury, Yiping Kang, Juan Manuel Ciro, Keith Achorn, Daniel Galvez, Mark Sabini, Peter Mattson, David Kanter, et al., "Multilingual spoken words corpus," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[11] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[12] Ken MacLean, "Voxforge," *Ken MacLean.[Online]. Available: http://www. voxforge. org/home.[Acedido em 2012]*, 2018.

[13] Yuan Gong, Jin Yu, and James Glass, "Vocalsound: A dataset for improving human vocal sounds recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 151–155.

[14] Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.