AND: Audio Network Dissection for Interpreting Deep Acoustic Models

Tung-Yu Wu [*1] **Yu-Xiang Lin** [*1] **Tsui-Wei Weng** [2]

Abstract

Neuron-level interpretations aim to explain network behaviors and properties by investigating neurons responsive to specific perceptual or structural input patterns. Although there is emerging work in the vision and language domains, none is explored for acoustic models. To bridge the gap, we introduce AND, the first **A**udio **N**etwork **D**issection framework that automatically establishes natural language explanations of acoustic neurons based on highly-responsive audio. AND features the use of LLMs to summarize mutual acoustic features and identities among audio. Extensive experiments are conducted to verify AND's precise and informative descriptions. In addition, we demonstrate a potential use of AND for audio machine unlearning by conducting concept-specific pruning based on the generated descriptions. Finally, we highlight two acoustic model behaviors with analysis by AND: (i) models discriminate audio with a combination of basic acoustic features rather than high-level abstract concepts; (ii) training strategies affect model behaviors and neuron interpretability – supervised training guides neurons to gradually narrow their attention, while self-supervised learning encourages neurons to be polysemantic for exploring high-level features.

1. Introduction

Deep Neural Networks (DNNs) have achieved remarkable success across various tasks. However, their inherent nature of high non-linearity poses great challenges in understanding model behaviors and neuron functionalities. To tackle this longstanding issue, several lines of work have attempted to acquire deeper knowledge about DNNs, ranging from decision explanation for the input (Simonyan et al., 2013; Selvaraju et al., 2017; Chattopadhay et al., 2018), property observation of layer-wise features (Pasad et al., 2021; 2023), to neuron-level interpretations (Bau et al., 2017; Hernandez et al., 2021; Oikarinen & Weng, 2023; Bills et al., 2023; Lee et al., 2023; Anonymous, 2024; Bai et al., 2024).

Among them, neuron-level interpretation tools automatically analyze the functionality of the neurons inside a model. One classical strategy is to examine each neuron's activation to inputs with different perceptual, structural, and semantic patterns (Hernandez et al., 2021; Oikarinen & Weng, 2023; Kalibhat et al., 2023). Knowing neuron's responsive features brings benefits to understanding fine-grained model behaviors. For instance, FALCON (Kalibhat et al., 2023) observes that a group of neurons with varied responsive features may together build a more concise and interpretable feature; MILAN (Hernandez et al., 2021), which explains neurons with natural language descriptions, finds that visual neurons in shallow layers requires more adjectives to describe, and these neurons substantially affect model performance. In addition, neuron-level interpretability can also be applied to tasks that require neuron-level knowledge, such as LLM factual editing (Meng et al., 2022), anonymized models auditing and spurious features editing (Hernandez et al., 2021).

Nevertheless, current neuron-level interpretability frameworks are all designed for visual (Bau et al., 2017; Hernandez et al., 2021; Oikarinen & Weng, 2023; Bai et al., 2024) and language modalities (Bills et al., 2023; Lee et al., 2023), rendering them incompatible with acoustic models. For example, some rely on external vision models such as image segmentation (Bau et al., 2020) and CLIP (Oikarinen & Weng, 2023). While sound event detection models (Nam et al., 2022; Shao et al., 2024) could offer timestamp annotations analogous to pixel-wise image segmentation, current models suffer from narrow detectable classes due to smaller dataset scale (Turpault et al., 2019; Serizel et al., 2020). This limitation also hinders the development of text-audio constrastive models (Guzhov et al., 2022; Wu* et al., 2023), causing difficulties for the direct transfer of previous interpretability tools from vision modality to audio modality.

Hence, to fill in the gap, in this paper we present AND,

[*]Equal contribution [1]National Taiwan University, Taipei, Taiwan [2]HDSI, UC San Diego, CA, USA. Correspondence to: Tsui-Wei Weng <lweng@ucsd.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

Our source code is available at https://github.com/Trustworthy-ML-Lab/Audio_Network_Dissection

an elegant and effective framework for **A**udio **N**etwork **D**issection using natural language descriptions. AND features an LLM-based pipeline, along with three specialized proposed modules, to capture responsive acoustic features of neurons in an audio network. The outputs of AND, including the closed-set concept, open-set concept, and LLM-generated summary, serve as mediums to inquire activities and properties of an audio network. Extensive experiments are presented in Section 4 to showcase the quality and usefulness of AND's outputs. In particular, in Section 4.1, we validate the efficacy of closed-concept identification by considering last-layer network dissection. In Section 4.2, we provide human evaluation to verify closed-concept identification and summary calibration. In Section 4.3, we discuss the middle-layer concept-specific pruning using AND by demonstrating a potential use of AND for audio machine unlearning, achieved by leveraging the open-concept identification module. Finally, we investigate critical properties of audio networks with AND in Section 4.4 and Section 4.5. In particular we analyze feature importance of different acoustic features by looking into parts of speech (POS) (Kumar et al., 2017) in Section 4.4, and argue the influence of different training strategies on neuron and model behaviors in Section 4.5.

In conclusion, this work's contributions are threefold:

- We present AND, the first automatic **A**udio **N**etwork **D**issection framework to provide natural language descriptions of acoustic neurons activities. Extensive experiments are conducted to verify AND's efficacy.
- We showcase the effect of middle-layer concept-specific pruning conducted by AND and discuss its relations to audio machine unlearning as a potential use case of AND.
- We leverage AND to analyze the feature importance of different acoustic features and the influence of training strategy on model behaviors as well as neuron interpretability.

2. Related Work

2.1. Neuron-level Interpretations

Neuron-level Interpretations aim to investigate neurons' activities to acquire fine-grained knowledge of deep networks as well as high-level network properties by inquiring groups or layers of neurons. Network Dissection (Bau et al., 2017) makes the first attempt to analyze visual models' behaviors by automatically exploring individual neurons' functionalities, with a line of follow-up work (Hernandez et al., 2021; Oikarinen & Weng, 2023; Kalibhat et al., 2023; Anonymous, 2024; Bai et al., 2024). In particular, MILAN (Hernandez et al., 2021) utilizes LSTMs to directly generate natural language descriptions conditioned on a set of patched highly-activated images. To remove the need of concept labels in (Bau et al., 2017; Hernandez et al., 2021) and accelerate the dissection process, CLIP-Dissect (Oikarinen & Weng, 2023) identify neuron concepts by computing similarity between neuron activations and representative features of open-vocabulary concept sets through probing dataset and CLIP. FALCON (Kalibhat et al., 2023) identifies critical concepts by scoring nouns, verbs and adjectives in predefined image descriptions of highly-activated images. Additionally, lowly-activated images are also utilized to address spurious and vague concepts, with overlapping concepts between the two sets being removed. Recently, LLM-based open-domain automatic description generation (Anonymous, 2024; Bai et al., 2024) is also proposed for interpreting visual models. Besides dissecting vision networks, there has being growing interest to interpret LLMs (Meng et al., 2022; Bills et al., 2023; Lee et al., 2023). ROME (Meng et al., 2022) employs causal mediation analysis (CMA) to locate neurons influential to specific factual knowledge. (Bills et al., 2023) proposes to use LLMs to generate the explanations for each neuron in language models and assess dissecting qualities. (Lee et al., 2023) further improves (Bills et al., 2023) by proposing efficient prompting techniques to obtain high quality neuron descriptions at lower computation cost.

2.2. Audio and Speech Network Interpretability

Previous efforts for interpreting audio and speech models primarily focus on input-specific explanations (Mishra et al., 2017; Becker et al., 2018) and layer-wise analysis (Pasad et al., 2021; 2023; Li et al., 2023). The former utilizes local interpretability tools, such as local interpretable model-agnostic explanations (LIME) (Ribeiro et al., 2016) and layer-wise relevance propagation (LRP) (Bach et al., 2015) to visualize models' attention to the MFCC or mel-spectrogram of a specific audio. The latter, mostly conducted on self-supervised learning (SSL) speech models, examines the acoustic, phonetic, and word-level properties encoded in the representations of each transformer layer by correlation-based analysis. Specifically, it is found that models pretrained with hidden discrete units prediction, such as HuBERT (Hsu et al., 2021) and WavLM (Chen et al., 2022), learn to encode richer phonetic and word information in deeper transformer layers (Pasad et al., 2023). While layer-wise findings provide reliable guidance for utilizing the embeddings of pretrained SSL models to downstream speech tasks, they provide limited utility for neuron-level tasks, such as model editing and unlearning. On the other hand, previous work (Kumar et al., 2017) have pointed out relations between acoustic concepts and natural languages, with a large set of audio concepts such as "glass breaking" and "sound of honking cars" exploited through a designed

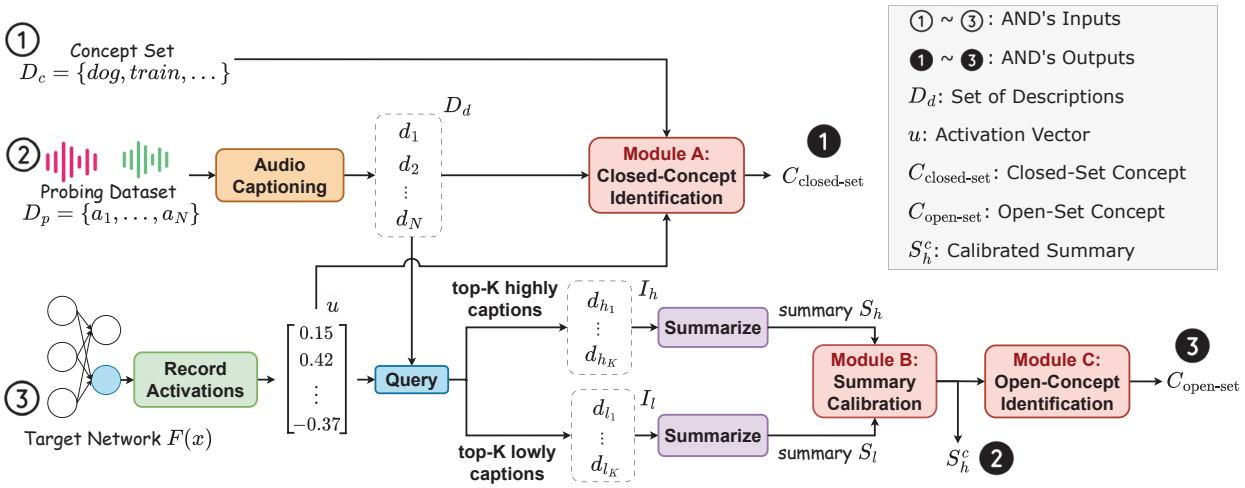


Figure 1. The proposed framework of AND. Taking concept set D_c, probing dataset D_p, and target network $F(\cdot)$ as inputs, AND employs a coarse-to-fine LLM-based pipeline to analyze each neuron's highly-responsive acoustic concepts by three specialized modules: (A) closed-concept identification, (B) summary calibration, and (C) open-concept identification. Closed-ended concept $C_{\text{closed-set}}$, calibrated summary S_h^c, and open-ended concept $C_{\text{open-set}}$ are generated as outputs of AND.

pipeline including techniques such as part-of-speech (POS) tagging. Inspired by this, in contrast to existing research, we propose the first neuron-level description-based interpretability tool, AND, to understand audio network behaviors from a more fine-grained aspect but with informative natural language descriptions. Our framework is elegant and instructive, as described in Section 3 and verified in Section 4 with extensive experiments.

3. Audio Network Dissection (AND)

3.1. Framework Overview

Inputs and Outputs As shown in Figure 1, AND takes a target network $F(\cdot)$, a predefined concept set D_c, $|D_c| = M$, with concepts $c_1, \dots, c_M$, and a probing dataset D_p, $|D_p| = N$, with audio clips $a_1, \dots, a_N$, as inputs. AND then dissects neurons by observing and summarizing the shared acoustic properties and identities among top-K highly-activated audio. We design an LLM-based pipeline with three specialized modules in AND: (A) closed-concept identification, (B) summary calibration, and (C) open-concept identification. AND provides 3 types of output corresponding to each module: a set of closed-set concepts $C_{\text{closed-set}}$ selected out of D_c, a natural language summary S_h^c describing commonalities among highly-activated audio, and a set of open-set concepts $C_{\text{open-set}}$ acquired from S_h^c. A dissection example of AND is provided in Appendix B.

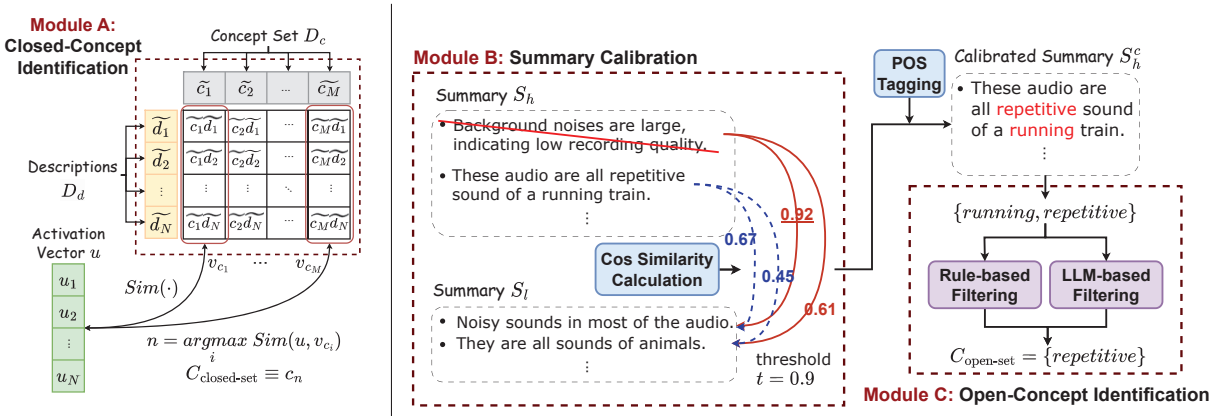
Pipeline AND leverages closed-concept identification, summary calibration, and open-concept identification (marked as **Module A**, **Module B**, and **Module C** in Figure 1), to acquire closed-concept $C_{\text{closed-set}}$, summary S_h^c, and open-concept $C_{\text{open-set}}$. Inputs of the three modules are acquired from concept set D_c, probing dataset D_p, and target network $F(\cdot)$, which are three original inputs of AND.

Specifically, closed-concept identification (module A) takes D_c, activation vector $u = [u_1, \dots, u_N]^T$, and descriptions $D_d = \{d_1, \dots, d_N\}$ as inputs, where d_i is the open-domain description of audio a_i generated by an audio captioning model; u_i is the activation value of interested neuron on audio a_i. Closed-concept identification generates $C_{\text{closed-set}}$ as the output. Summary calibration (module B) takes S_h and S_l as inputs, which are summaries of top-K highly-activated and lowly-activated audio, with calibrated summary S_h^c being the output. The retrieval of highly-activated and lowly-activated audio is achieved by querying D_d based on the values in u. Finally, S_h^c serves as inputs for open-concept identification (module C) to extract critical concepts as $C_{\text{open-set}}$. We elaborate each module in Section 3.3-3.5 respectively, with detailed illustration in Figure 2.

3.2. Input Preprocessing for Each Module

This section describes the preprocessing of AND's inputs to generate inputs for three specialized modules. In particular we discuss the generation of D_d, S_h, and S_l.

To obtain caption D_d of audio clips in D_p, we adopt SALMONN (Tang et al., 2024) as the open-domain audio captioning model. We feed each audio a_i into SALMONN



(a) Illustration of closed-concept identification. Audio are represented as its descriptions rather than waveform feature embeddings in CLAP (Wu* et al., 2023).

(b) Illustration of summary calibration and open-concept identification. Pairwise sentence similarity between sentences of S_h and S_l is computed to remove homogeneous concepts and generate calibrated summary S_h^c. Critical concepts in S_h^c are captured by filtering out concepts unrelated to acoustics (e.g. "running"). Adjectives are used as an example here.

Figure 2. Detailed illustration of AND's three specialized modules: (A) closed-concept identification, (B) summary calibration, and (C) open-concept identification.

to acquire d_i, forming $D_d = \{d_1, \ldots, d_N\}$. For S_h and S_l, given an acoustic neuron $f(\cdot)$ in $F(\cdot)$, $u_i = f(a_i)$ is the activation value of the neuron for the audio clip a_i. It is the descriptions of audio with top-K highest and lowest activation values that constitute the high-activation descriptions I_h:

$$I_h = \{d_i \mid u_i \ \textit{is of top-K highest value in u}\},$$

and the low-activation descriptions I_l:

$$I_l = \{d_i \mid u_i \ \textit{is of top-K lowest value in u}\}.$$

I_h and I_l are subsequently used to generate the summary S_h and S_l by instructing Llama-2-chat-13B (Touvron et al., 2023) to summarize and list down the commonalities among these audio descriptions in the set. The generated D_d, S_h, and S_l serve as inputs for implementing closed-concept identification (module A), summary calibration (module B), and open-concept identification (module C), which are detailed in Section 3.3, Section 3.4, and Section 3.5.

3.3. Module A: Closed-Concept Identification

Closed-concept identification takes concept set $D_c = \{c_1, \ldots, c_M\}$, set of descriptions $D_d = \{d_1, \ldots, d_N\}$, and activation vector u as inputs. The output is a concept $C_{\text{closed-set}}$ from D_c to label the interested neuron, similar to CLIP-Dissect's spirits. Note that the term "closed-set" is due to pre-defined concept set, but it can be an open vocabulary set as needed.

For each concept c_i, AND generates the representative feature vector v_{c_i}, where $[v_{c_i}]_j = \widetilde{c_i} \cdot \widetilde{d_j}, j = 1, \ldots, N$, as shown in Figure 2a. CLIP's text encoder $E_{text}(\cdot)$ is used to encode concept c_i into $\widetilde{c_i}$ and descriptions d_j into $\widetilde{d_j}$. That is, $\widetilde{c_i} = E_{text}(c_i)$ and $\widetilde{d_j} = E_{text}(d_j)$. Then, a similarity function $Sim(\cdot)$, such as cosine similarity or weighted pointwise mutual information (WPMI), is applied to measure the similarity between u and v_{c_i}. The concept with the highest similarity score is considered the most-matched closed-set concept, denoted as $C_{\text{closed-set}}$. We refer to this proposed approach as the description-based (DB) method.

In addition to DB, we present two straightforward alternative approaches than Figure 2a to achieve closed-concept identification. The first idea is using in-context learning (ICL) to query the LLM to directly select a concept from D_c that best matches information in the calibrated summary S_h^c. We provide the instruction and examples for ICL in Appendix A.3 The second solution is to adopt a text-audio contrastive learning model, such as CLAP, and perform cross-modal retrieval as CLIP-Dissect (Oikarinen & Weng, 2023) does but in a text-audio-based (TAB) manner. Notably, the TAB method can be considered as the variant of CLIP-Dissect for the acoustic model. Performance comparisons of ICL, DB, and TAB methods are presented in Section 4.1 and Table 2.

3.4. Module B: Summary Calibration

As shown in Figure 2b, the summary calibration module takes high-activation summary S_h and low-activation summary S_l as inputs. S_l is used to filter out spurious concepts that are ambiguous or hallucinating in S_h. For instance, in scenarios where all audio in D_p contain no noise, S_h and S_l would hold sentences emphasizing the clarity of sound. We then remove this redundant information in S_h. In the calibration process, for each mentioned point p in

S_h, we compute the cosine similarity between its sentence feature (Reimers & Gurevych, 2019) and those in S_l. If the similarity value between p and any point in S_l exceeds a predefined threshold t, we discard p. The calibrated S_h is the output of this module, denoted as S_h^c.

3.5. Module C: Open-Concept Identification

As illustrated in Figure 2b, the open-concept identification takes calibrated summary S_h^c as the input and extracts open-set concept $C_{\text{open-set}}$. Then naming of "open-set" is because $C_{\text{open-set}}$ is derived from the open-domain summary S_h^c.

Open-concept identification aims to identify critical concepts that are related to acoustics. This is implemented by applying designed filtering to S_h^c. Using adjectives as an example, we first apply POS tagging to draw out all adjectives from S_h^c, while some may be irrelevant to acoustics such as the word "running". Second, we leverage rule-based and LLM-based filtering to remove trivial adjectives. The former eliminates common stop words, while the latter queries Llama-2-chat-13B to determine whether the adjective describes acoustic properties. For LLM-based filtering, a designed hard prompt is used[1]. The resulting filtered set $C_{\text{open-set}}$ comprises sound-related concepts, reflecting shared acoustic properties among top-K highly-activated audio. Figure 3 displays the distribution of top-10 most common adjectives for all linear layer neurons within an Audio Spectrogram Transformer (AST) (Gong et al., 2021) trained on the ESC50 (Piczak, 2015) dataset. More results are in Appendix D.

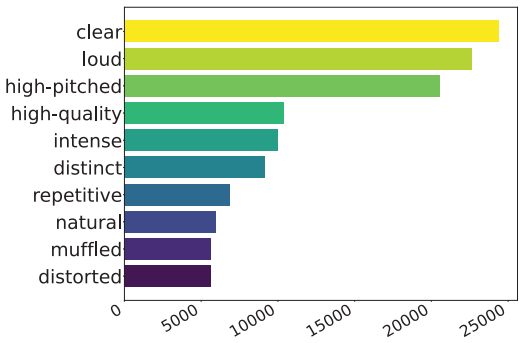


Figure 3. Counting of adjectives for AST's all linear layer neurons generated by open-concept identification module. We show the top-10 most-used adjectives here.

[1]In practice, after applying POS tagging to all neurons' calibrated summaries, AND forms a universal set of adjectives. Llama-2-chat-13B is then used to examine adjectives in the set to prevent repeated queries to the same words.

4. Experiments

Overview This section aims to verify the dissection quality of AND and investigate acoustic model behaviors using AND. First, we validate the dissection quality of AND in Section 4.1 before diving into model analysis. Following CLIP-Dissect, Section 4.1 measures last layer dissection accuracy as an indicator of the interpretation quality of AND's closed-concept identification module (module A) because neurons in the output layer have 1-1 corresponding with classes in the training dataset and thus can serve as ground-truth. In Section 4.2, we conduct a human evaluation to assess AND's performance on middle-layer neurons with model-unseen concepts. Also ,we provide qualitative results in Appendix B.

Next, we verify $C_{\text{open-set}}$ from module C as well as introduce a use case of AND regarding machine unlearning (Golatkar et al., 2020; Nguyen et al., 2022; Zhang et al., 2022) in Section 4.3. Notably, classical strategies for machine unlearning involve ablating individual neurons that are influential to the target concept (Pochinkov & Schoots, 2024). While there exist works focusing on unlearning vision and language models (Meng et al., 2022; Gandikota et al., 2023; Zhang et al., 2023), audio machine unlearning remains relatively underexplored. To achieve this challenging task, we consider linguistically parsing each neuron's S_h^c generated by AND's module B and leverage acquired $C_{\text{open-set}}$ to conduct neuron pruning based on a given target concept. More concretely, we prune all of the neurons dissected with the target concept and examine audio models' perception abilities towards both related and unrelated concepts after ablation. Results are reported in Section 4.3.

Finally, Section 4.4 and Section 4.5 are analyses and findings of acoustic model behaviors using AND. In Section 4.4, we follow MILAN's experimental settings but on acoustic networks with AND to investigate how it is different from or the same as vision network's observations discovered in MILAN. In Section 4.5, we analyze neuron polysemanticity (uninterpretability) (Mu & Andreas, 2020; Oikarinen & Weng, 2023; Bills et al., 2023), i.e., neurons' diverse attention to seemingly unrelated features, affected by network training strategies.

For experiments in Section 4.1- Section 4.5, we utilize the ESC50 (Piczak, 2015) as D_p and consider all its 50 audio classes as D_c. The only exception is that we use open-ended acoustic concept set proposed by Kumar et al. (2017) as D_c in Section 4.2. For the target networks $F(\cdot)$, as shown in Table 1, we select the Audio Spectrum Transformer (AST) (Gong et al., 2021) and BEATs (Chen et al., 2023). BEATs is an SSL audio model pretrained with Masked Audio Modeling (MLM) (Chen et al., 2023), which allows us to use either by finetuning the whole model (BEATs-finetuned) or training only the last linear layer (BEATs-

Table 1. Training settings of audio models used in the experiments.

	AST	BEATs-finetuned	BEATs-frozen
Pretraining on Audioset	Supervised (all layers)	Self-supervised (all layers)	Self-supervised (all layers)
Fine-tuning on ESC50	Supervised (all layers)	Supervised (all layers)	Supervised (final linear layer)

Table 2. Last layer network dissection accuracy of AST, BEATs-frozen, and BEATs-finetuned on the ESC50 dataset, with the highest performance marked in bold. As discussed in Section 3.3, ICL refers to querying LLM to choose a best-matched concept for the summary, and TAB/DB refers to the text-audio-based/description-based method. DB achieves the best results among all metrics and models.

Model	AST			BEATs-finetuned			BEATs-frozen		
Method	Top-1 Acc	Top-5 Acc	Cos	Top-1 Acc	Top-5 Acc	Cos	Top-1 Acc	Top-5 Acc	Cos
AND (module A: ICL)	72.0	60.0	0.83	52.0	56.0	0.68	16.0	16.0	0.37
AND (module A: TAB)	96.0	**100.0**	0.98	74.0	92.0	0.82	46.0	72.0	0.60
AND (module A: DB)	**100.0**	**100.0**	**1.00**	**76.0**	**100.0**	**0.83**	**58.0**	**82.0**	**0.69**

frozen). We adopt both versions. These target networks are trained on the ESC50, achieving testing accuracies of 95.0% for AST, 89.75% for BEATs-finetuned, and 84.75% for BEATs-frozen, respectively.

4.1. Evaluation by Last Layer Dissection

Experimental Settings We evaluate ICL, DB, and TAB methods for closed-concept identification (module A), as discussed in Section 3.3. For the ICL-based approach, we utilize ICL to query Llama-2-chat-13B to select one class from the given concept set that best matches the calibrated summary, with some hand-crafted examples provided. For TAB, we use CLAP to extract audio features in D_p and concept features in D_c. For DB, we use CLIP to extract concept/description features.

We consider five similarity functions: cosine similarity, cubed cosine similarity, rank reorder, WPMI, and softWPMI, as discussed in CLIP-Dissect (Oikarinen & Weng, 2023) and label-free concept bottleneck models (Oikarinen et al., 2023). Each method's best performance among the five functions is presented, with detailed results provided in Appendix E. We report the top-1 and top-5 classification accuracy as well as the CLIP-based cosine similarity between the predicted concept and corresponding ground-truth class for each final-layer neuron.

Results Table 2 demonstrates DB's superiority compared with ICL and TAB across all three target models. Particularly, perfect dissection results are achieved on the AST. On BEATs-frozen, DB outperforms TAB by 12% in top-1 accuracy and 10% in top-5 accuracy. This underscores the advantages of similarity calculation within the text feature space, as proposed in DB to generate audio descriptions, mitigating potential imperfections in cross-modal feature projection between audio and text features that might arise from CLAP as in TAB.

Notably, the dissection accuracy typically decreases as the classification ability of the target models on the probing dataset declines. For instance, DB's top-1 accuracy drops from 100.0% to 76.0% and 58.0%, corresponding to testing accuracies of 95.0%, 89.75%, and 84.75% for AST, BEATs-finetuned, and BEATs-frozen, respectively. This decline is attributed to incorrect activation values (logits) of the output layer neurons for misclassified samples, which in turn leads to erroneous calculation of the similarity function. However, despite this decrease, DB still consistently outperforms ICL and TAB in all metrics.

4.2. Human Evaluation

Experimental Settings This section evaluates AND on middle-layer neurons with model-unseen concepts. Specifically, we replace D_c with a large-scale acoustic concept set proposed by Kumar et al. (2017) to evaluate the quality of the calibrated summary S_h^c from module B and the concept identified by DB and TAB in module A. Evaluating the middle-layer interpretation is challenging due to the lack of intrinsic labels for neurons in middle layers, driving us to conduct human evaluation. Following MILAN (Hernandez et al., 2021) and CLIP-Dissect (Oikarinen & Weng, 2023), we ask human annotators to write summaries for top-K high-activated audio as well as score the description generated by AND. Notably, previous data collected through small-scale experiments on Amazon Mechanical Turk are not satisfying despite providing instructions and examples. This is probably due to considerable challenges for people to precisely describe potential acoustic characteristics among a group of audio in natural language. As the middle-

layer assessment requires professional writers, we turn to author-based experiments. For transparency and fairness, the human study results and the associated audio clips are provided in our code repository for the reader's reference.

We assess the dissection quality of 10 randomly selected neurons per layer in AST, with a total of 120 neurons selected. We as the human evaluators to (1) score whether the produced description matches the highly activated audio samples and (2) write the shared properties of highly activated audio samples. For each sampled neuron, five highly activated audio clips are given. For (1), We rate the quality of AND's description on a scale of 1 to 5, with 1 being strongly disagree and 5 being strongly agree, in response to the question, "Does the given description accurately describe most of these audio clips?" For (2), the collected summaries are evaluated through the semantic similarity (e.g. cos similarity or BERTScore (Zhang et al., 2020)) with descriptions from AND.

Results Firstly, scores across neurons are averaged to serve as a performance indicator. Second, we evaluate AND by computing cos similarity and BERTScore on the generated description, aiming to compare the acoustic similarity between the human-written summaries and descriptions. Results are shown in Table 3. Calibrated summaries S_h^c from module B achieve the highest quality among all methods, reflecting the effectiveness of open-domain interpretation.

Table 3. Results of human evaluation to measure AND's capability of dissecting middle-layer neurons. Rating is the mean of scores (1-5) across neurons. Cos similarity and BERTScore are computed between AND's descriptions and human's written descriptions.

	Rating	Cos	BERTScore
AND (Moduel A: TAB)	2.73	0.24	0.42
AND (Module A: DB)	3.24	0.57	0.45
AND (Module B: SUM)	**3.49**	**0.72**	**0.46**

4.3. Use Case: Audio Machine Unlearning

Experimental Settings We observe models' change of confidence for samples in the ESC50 testing set after neuron ablation. For instance, upon choosing the class "water drops" as the target concept, we prune out all neurons dissected to have "water drops" as their responsive features. Then, samples of all 50 classes in the ESC50 testing set, such as "pouring water" and "cow", are fed to the pruned network to observe the change of logits. This process is repeated 50 times by using each of ESC50's classes as the target concept, and the values of confidence change on the target concept and non-target concepts are averaged.

There are several intuitive strategies to determine whether a neuron is relevant to the target concept. First method is to extract non-trivial open-domain entities in S_h^c as $C_{\text{open-set}}$ (module C). If $C_{\text{open-set}}$ includes the target concept, we mask out this neuron. We refer to this method as open-concept pruning (OCP). The second method is using closed-set concept $C_{\text{closed-set}}$. Similarly, if the target concept is included in $C_{\text{closed-set}}$, we mask out this neuron. To align the pruning numbers between open-concept pruning and closed-concept pruning, we select the top-3 most matched concepts when constructing $C_{\text{closed-set}}$, i.e., $C_{\text{closed-set}}$ contains three most matched concepts drawn from D_c. Notably, ICL, DB, and TAB in module A are all potential options to construct $C_{\text{closed-set}}$, as discussed in Section 4.1. Since DB and TAB are considerably better than ICL for last-layer dissection, we adopt DB and TAB in this experiment.

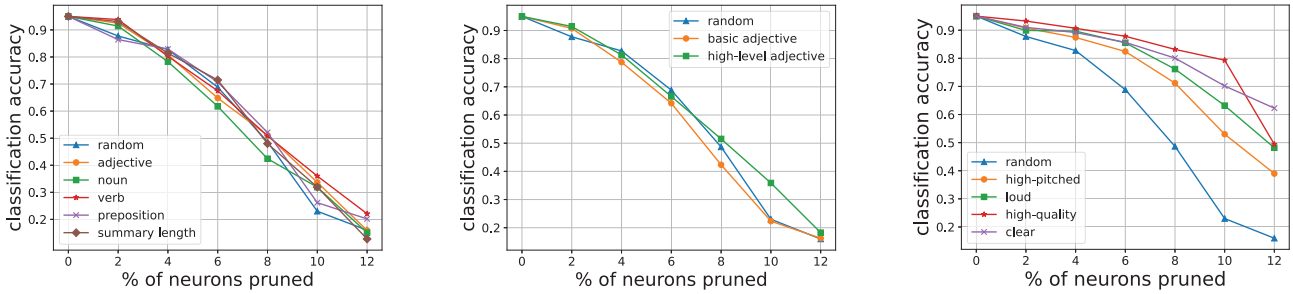
Results The effects of ablating neurons are displayed in Table 4. Results show that OCP based on module C is most efficient in middle-layer concept pruning. Although one-pass pruning may incur backup neurons in the network to show up and complement the pruned ability (McGrath et al., 2023), OCP still imposes considerable effects on models' perceptions of the target concept. Specifically, for AST, the classification confidence drop of the target concept is higher than that of non-target concepts, with a gap of 3.9 achieved. This trend is also observed in BEATs-finetuned and BEATs-frozen. Notably, BEATs-frozen, whose model weights are from SSL pretraining, is much more robust to pruning. This may indicate the diverse attention of SSL pretrained neurons with stronger "backup" abilities (McGrath et al., 2023) compared with models trained in a supervised manner. Figure 10 in Appendix F illustrates the change of confidence with "water drops" as the target concept, BEATs-finetuned as the target network, and OCP as the ablating strategy. After pruning, BEATs-finetuned's classification abilities on water-related concepts, such as "toilet flush" and "pouring water", are heavily affected, with a smaller influence on other unrelated concepts.

4.4. Analyzing Acoustic Feature Importance

Experimental Settings Similar to MILAN (Hernandez et al., 2021), we linguistically parse descriptions to fine-grained information and investigate factors influencing model performance by neuron ablation with certain criteria. Among all linear layer neurons of AST, we prune $r\%$ neurons with the highest number of nouns, adjectives, verbs, prepositions, and the longest calibrated summary, where r is a parameter. Second, we categorize adjectives into two groups: basic adjectives and high-level adjectives. The former comprises "high-pitched", "high-quality", "clear", and "loud", which are common in generated descriptions and depict fundamental acoustic features such as frequency and amplitude. We group all the remaining adjectives as high-level ones with more abstract concepts such as "dramatic" and "repetitive". We again ablate neurons based on either the number of basic adjectives or high-level adjectives in

Table 4. Averaged change of confidence after neuron ablation, with each class being the target concept. Avg, ΔA, and ΔR refer to the averaged pruned numbers of neurons, confidence change on ablating class samples, and confidence change on remaining class samples, respectively. OCP refers to open-concept pruning by leveraging $C_{\text{open-set}}$.

Model	AST				BEATs-finetuned				BEATs-frozen			
Method	Avg	ΔA	ΔR	ΔA - ΔR ↑	Avg	ΔA	ΔR	ΔA - ΔR ↑	Avg	ΔA	ΔR	ΔA - ΔR ↑
Random baseline	3000	-13.52	-13.48	0.04	3000	-0.55	-0.55	0.00	3000	-0.72	-0.73	-0.01
AND (module A: TAB)	3317	-4.19	-4.03	0.16	2765	-0.50	-0.49	0.01	2765	-0.42	-0.51	-0.09
AND (module A: DB)	3317	-7.18	-7.41	-0.23	2765	-0.53	-0.57	-0.04	2765	-0.48	-0.58	-0.10
AND (module C: OCP)	2808	-10.73	-6.83	**3.90**	2677	-2.00	-0.51	**1.49**	2906	-0.65	-0.55	**0.10**



(a) Ablating neurons with the highest number of POS and longest summary.

(b) Ablating neurons with the highest number of basic and high-level adjectives.

(c) Ablating neurons if a certain basic adjective is contained in $C_{\text{open-set}}$.

Figure 4. Feature importance analysis of AST on ESC50, measured by module C in AND. The x-axis is the percentage of ablated linear layer neurons, and the y-axis is the testing performance on ESC50 after pruning.

their summaries. Finally, we study the individual role of four basic adjectives. For each adjective, we prune a neuron if it is in $C_{\text{open-set}}$. If the number of neurons meeting the criterion exceeds $r\%$ of the total neurons, we conduct random pruning among them. Experiments are conducted three times under three different random seeds and the numbers are averaged.

Results Our observations in acoustic neurons diverge from MILAN's findings in visual neurons, which observe that visual neurons with more adjectives in the descriptions are more influential to model performance. As shown in Figure 4a, pruning neurons with most adjectives does not necessarily lead to a more significant performance drop. While pruning nouns exhibits a higher impact than random pruning for 4%, 6%, and 8% neurons, it falls short for 2% and 10%. When further categorizing adjectives into basic and high-level ones, as illustrated in Figure 4b, we find that basic adjectives are consistently more influential than high-level ones, with up to over 10% performance difference for the case of 10% pruning. Moreover, pruning basic adjectives incurs a slightly larger performance drop than random pruning, except for the 2% case. This may indicate that acoustics are identified more with its fundamental features rather than abstract concepts such as repetitiveness and emotions, which is intuitive especially when the models are supervisedly trained to classify audio entities, such as AST on ESC50. Further examinations on the four individuals discover much smaller pruning effects for ablating neurons with a specific basic adjective, as shown in Figure 4c. Comparing the results in Figure 4b, we conclude that models distinguish audio more based on a combination of basic acoustic features rather than individual ones or high-level concepts, particularly for the task of entity classification.

As shown in Figure 5, we also compute the averaged number of adjectives per linear layer neuron for each transformer block as in MILAN. In particular, we analyze AST, BEATs-finetuned, and BEATs-frozen, corresponding to three different training strategies: supervised training, SSL pretraining with whole-model supervised finetuning, and SSL pretraining with final layer finetuning. We observed that AST's attention to different acoustic features, represented as adjectives, narrows down in deeper transformer blocks. This trend is analogous to MILAN's observation for supervised ResNet18 on ImageNet. However, when considering SSL pretraining, the diverse-to-accordant trend is mitigated, as evidenced by the case of BEATs-finetuned. Moreover, for BEATs-frozen, all transformer layers exhibit the same level of attention to adjectives, and the overall numbers are lower.

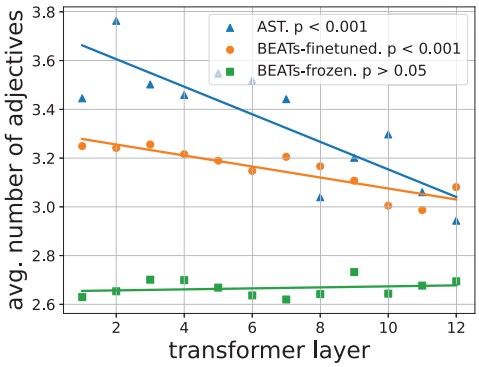


Figure 5. Number of averaged adjectives per neuron in different transformer blocks of AST, BEATs-finetuned, and BEATs-frozen.

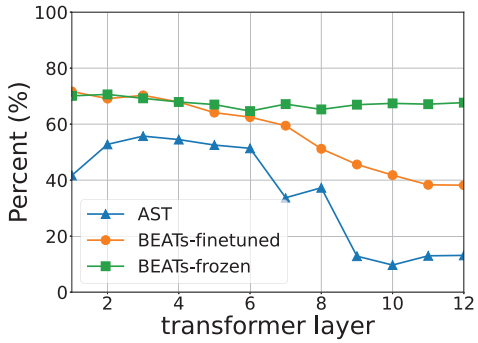


Figure 6. Percentage of uninterpretable neurons in different transformer blocks of AST, BEATs-finetuned, and BEATs-frozen. More results are provided in Appendix G.

This suggests a potential effect of training strategies on model behaviors. Neurons in BEATs-frozen are more diverse and do not represent mutually similar concepts. While the supervised training leads to a more capable yet convergent neuron layout. Further discussions on the effects of training strategies are in Section 4.5.

4.5. Training Strategy Affects Neuron Interpretability

Experimental Settings Pseudo code for the proposed pipeline is displayed in Appendix G.1. We first pretrain a K-means model $K(\cdot)$ with sentences of audio descriptions d_i for each audio a_i in D_p. Number of clusters is set to 11, decided by the elbow method (Thorndike, 1953). Then, for an interested neuron $f(\cdot)$, we acquire its S_h and project sentences of each description to the 11 groups using $K(\cdot)$. We say a description is related to a cluster if at least one sentence within the description belongs to that cluster. If at least τ descriptions are related to the same cluster, i.e., descriptions of highly-activated audio share common information, we label the neuron as "*interpretable*". Otherwise, it is regarded as "*uninterpretable*". τ is a parameter. Higher τ denotes a stricter criterion. The semantics of found clusters are discussed in Appendix G.2.

Results The percentages of uninterpretable neurons for each transformer block of three models are illustrated in Figure 6, with $\tau = 4$. AST demonstrates a decrease in the percentage of uninterpretable neurons from shallow to deep transformer blocks. Neurons in shallow layers are more diverse and gradually concentrate on certain concepts in deeper layers, with more overlapped information among highly-activated audio, to conduct the classification task. On the other hand, BEATs-frozen has consistent percentages of unexplainable neurons among all layers. From the scale of individual neurons, this manifests the effect of SSL pretraining on diversifying models' abilities. When finetuning the entire model, as in BEATs-finetuned, the supervision attempts to guide the model to converge its attention in deeper layers by narrowing neurons' responsive features. This finding links to the previously reported attentive overfitting of supervised models (Zagoruyko & Komodakis, 2017; Ericsson et al., 2021), which SSL networks are less vulnerable to (Ericsson et al., 2021). While previous discussions on attentive overfitting typically rely on observing performance metrics or using input-specific interpretation tools like Grad-cam (Selvaraju et al., 2017), we provide neuron-level explanations for this phenomenon using AND. Finally, this is also related to the averaged number of adjectives per neuron in Figure 5, showing that a decrease in acoustic features extracted from natural language descriptions may represent a narrowed neuron polysemanticity. Figures under different top-K and τ are presented in Appendix G.3, with similar trends observed. Similar experiments but adopting GTZAN Music Genre (Tzanetakis & Cook, 2002) as the probing and training dataset are presented in Appendix G.4.

5. Conclusion

In this work, we have introduced AND, the first **A**udio **N**etwork **D**issection framework, which applies an LLM-based pipeline incorporated with three specialized modules to identify the responsive features of acoustic neurons. Extensive experiments are conducted to verify AND's dissection quality and discover acoustic model behaviors using AND. Specifically, last layer dissection and human evaluation demonstrate AND's high-quality dissection. Second, we examine the effect of concept-specific pruning conducted using AND and discuss its potential use case of model unlearning. Third, we investigate the roles of acoustic features in the model's perception ability and make a comparison with previous observations in the vision domain. Finally, we discuss the impact of different training strategies on neuron-level model behaviors, where the results align with the analyses of acoustic features.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. Particularly, this work targets acoustic model interpretability, with a network dissection tool developed, aiming to help the community gain deeper knowledge of the properties of acoustic neural networks.

Acknowledgement

The authors thank the anonymous reviewers for insightful feedback and suggestions. This work was supported in part by National Science Foundation (NSF) awards CNS-1730158, ACI-1540112, ACI-1541349, OAC-1826967, OAC-2112167, CNS-2100237, CNS-2120019, the University of California Office of the President, and the University of California San Diego's California Institute for Telecommunications and Information Technology/Qualcomm Institute. Thanks to CENIC for the 100Gbps networks. T.-W. Weng is supported by National Science Foundation under Grant No. 2107189 and 2313105. T.-W. Weng also thanks the Hellman Fellowship for providing research support.

References

Anonymous. TeLLMe what you see: Using LLMs to explain neurons in vision models, 2024. URL https://openreview.net/forum?id=01ep65umEr.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Bai, N., Iyer, R. A., Oikarinen, T., and Weng, T.-W. Describe-and-dissect: Interpreting neurons in vision networks with language models. *arXiv preprint arXiv:2403.13771*, 2024.

Bau, D., Zhou, B., Khosla, A., Oliva, A., and Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6541–6549, 2017.

Bau, D., Zhu, J.-Y., Strobelt, H., Lapedriza, A., Zhou, B., and Torralba, A. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020.

Becker, S., Ackermann, M., Lapuschkin, S., Müller, K.-R., and Samek, W. Interpreting and explaining deep neural networks for classification of audio signals. *arXiv preprint arXiv:1807.03418*, 2018.

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., and Saunders, W. Language models can explain neurons in language models. 2023.

Chattopadhay, A., Sarkar, A., Howlader, P., and Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 839–847, 2018.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing (JSTSP)*, 16(6):1505–1518, 2022.

Chen, S., Wu, Y., Wang, C., Liu, S., Tompkins, D., Chen, Z., Che, W., Yu, X., and Wei, F. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 5178–5193. PMLR, 2023.

Ericsson, L., Gouk, H., and Hospedales, T. M. How well do self-supervised models transfer? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5414–5423, 2021.

Gandikota, R., Materzyńska, J., Fiotto-Kaufman, J., and Bau, D. Erasing concepts from diffusion models. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2426–2436, 2023.

Golatkar, A., Achille, A., and Soatto, S. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9304–9312, 2020.

Gong, Y., Chung, Y.-A., and Glass, J. AST: Audio Spectrogram Transformer. In *Proceedings of INTERSPEECH*, pp. 571–575, 2021.

Guzhov, A., Raue, F., Hees, J., and Dengel, A. Audioclip: Extending clip to image, text and audio. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 976–980, 2022.

Hernandez, E., Schwettmann, S., Bau, D., Bagashvili, T., Torralba, A., and Andreas, J. Natural language descriptions of deep visual features. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 29: 3451–3460, 2021.

Kalibhat, N., Bhardwaj, S., Bruss, C. B., Firooz, H., Sanjabi, M., and Feizi, S. Identifying interpretable subspaces in image representations. In *Proceedings of International Conference on Machine Learning (ICML)*, pp. 15623–15638. PMLR, 2023.

Kumar, A., Raj, B., and Nakashole, N. Discovering sound concepts and acoustic relations in text. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 631–635, 2017.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*, pp. 611–626, 2023.

Lee, J., Oikarinen, T., Chatha, A., Chang, K.-C., Chen, Y., and Weng, T.-W. The importance of prompt tuning for automated neuron explanations. *NeurIPS ATTRIB workshop*, 2023.

Li, Y., Anumanchipalli, G. K., Mohamed, A., Chen, P., Carney, L. H., Lu, J., Wu, J., and Chang, E. F. Dissecting neural computations in the human auditory pathway using deep neural networks for speech. *Nature Neuroscience*, pp. 1–13, 2023.

McGrath, T., Rahtz, M., Kramar, J., Mikulik, V., and Legg, S. The hydra effect: Emergent self-repair in language model computations. *arXiv preprint arXiv:2307.15771*, 2023.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in gpt. volume 35, pp. 17359–17372, 2022.

Mishra, S., Sturm, B. L., and Dixon, S. Local interpretable model-agnostic explanations for music content analysis. In *ISMIR*, volume 53, pp. 537–543, 2017.

Mu, J. and Andreas, J. Compositional explanations of neurons. volume 33, pp. 17153–17163, 2020.

Nam, H., Kim, S.-H., Ko, B.-Y., and Park, Y.-H. Frequency Dynamic Convolution: Frequency-Adaptive Pattern Recognition for Sound Event Detection. In *Proceedings of INTERSPEECH*, pp. 2763–2767, 2022.

Nguyen, T. T., Huynh, T. T., Nguyen, P. L., Liew, A. W.-C., Yin, H., and Nguyen, Q. V. H. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Oesper, L., Merico, D., Isserlin, R., and Bader, G. D. Wordcloud: a cytoscape plugin to create a visual semantic summary of networks. *Source Code for Biology and Medicine*, 6(1):7, 2011.

Oikarinen, T. and Weng, T.-W. Clip-dissect: Automatic description of neuron representations in deep vision networks. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2023.

Oikarinen, T., Das, S., Nguyen, L., and Weng, T.-W. Label-free concept bottleneck models. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2023.

Pasad, A., Chou, J.-C., and Livescu, K. Layer-wise analysis of a self-supervised speech representation model. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921, 2021.

Pasad, A., Shi, B., and Livescu, K. Comparative layer-wise analysis of self-supervised speech models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Piczak, K. J. Esc: Dataset for environmental sound classification. In *Proceedings of the ACM international Conference on Multimedia (ACM MM)*, pp. 1015–1018, 2015.

Pochinkov, N. and Schoots, N. Dissecting language models: Machine unlearning via selective pruning. *arXiv preprint arXiv:2403.01267*, 2024.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3982–3992, 11 2019.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 1135–1144, 2016.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Serizel, R., Turpault, N., Shah, A., and Salamon, J. Sound event detection in synthetic domestic environments. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 86–90, 2020.

Shao, N., Li, X., and Li, X. Fine-tune the pretrained atst model for sound event detection. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 911–915, 2024.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Tang, C., Yu, W., Sun, G., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., and Zhang, C. Salmonn: Towards generic hearing abilities for large language models. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024.

Thorndike, R. L. Who belongs in the family? *Psychometrika*, 18(4):267–276, 1953.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Turpault, N., Serizel, R., Shah, A. P., and Salamon, J. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2019.

Tzanetakis, G. and Cook, P. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing (TASLP)*, 10(5):293–302, 2002.

Wu*, Y., Chen*, K., Zhang*, T., Hui*, Y., Berg-Kirkpatrick, T., and Dubnov, S. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

Zagoruyko, S. and Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

Zhang, E., Wang, K., Xu, X., Wang, Z., and Shi, H. Forget-me-not: Learning to forget in text-to-image diffusion models. *arXiv preprint arXiv:2303.17591*, 2023.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. Bertscore: Evaluating text generation with BERT. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2020.

Zhang, Z., Zhou, Y., Zhao, X., Che, T., and Lyu, L. Prompt certified machine unlearning with randomized gradient smoothing and quantization. volume 35, pp. 13433–13455, 2022.

A. Implementation Details

A.1. Audio Captioning Model

To obtain natural language descriptions of audio, we utilize SALMONN (Tang et al., 2024), which is a pre-trained LLM-based multi-task audio model. One of SALMONN's pretraining tasks is audio captioning, endowing it with remarkable open-domain audio captioning abilities. Following its official guidance, we use the prompt *"Please describe the audio in detail."* and employ SALMONN-13B. As SALMONN is newly proposed, its implementation requires further refinement, notably in the absence of batch inference capabilities. Captioning the entire ESC50 dataset takes approximately 40 hours on an NVIDIA-A6000 GPU, with batch size set to be 1. There might be a substantial efficiency gain once batch inference is available.

A.2. Large Language Model

We adopt Llama-2-chat-13B (Touvron et al., 2023) for all LLM-related experiments. The vllm package (Kwon et al., 2023) is employed to boost the inference efficiency, which integrates efficient attention mechanism and other speed-up techniques to create a memory-efficient LLM inference engine. The summarization process for all linear layers in AST and BEATs takes around 12 and 10 hours respectively on an NVIDIA RTX A6000 GPU. This includes generating summaries for both highly and lowly activated samples.

For the description summarization, we instruct the LLM with prompt *"Here are descriptions of some audio clips. Please summarize these descriptions by identifying their commonalities."* to ask LLM for summarization. Additionally, the LLM-based non-acoustic-word filtering is conducted with prompt *"Can the adjective be used to describe the tone, emotion, or acoustic features of audio, music, or any other form of sound? Answer yes or no and give the reason."*.

A.3. Instruction and Examples of In-context Learning

In Section 3.3, we introduce a method called ICL in AND's module A, which uses LLM to interpret the calibrated summary S_c^h and then explain neurons in target model $F(\cdot)$. This method is built upon in-context learning, enabling LLM to select a concept $C_{\text{close-set}}$ from D_c as output. We have experimented with 1-shot and 2-shot learning. The instructions and examples used are provided in Table 5.

Table 5. Instruction and examples used in ICL (module A in AND).

Instruction	You have a set of object classnames: [concept set D_c] The following is a description about some audio clips. Based on the description, select a classname out of the above classnames that matches the description most.
Example 1	**Description**: The audio features a car meowing: All of the clips contain the sound of a cat meowing. Loud sound: These clips are all of loud sound but with varying degrees of intensity. Repetitive barking: Clips 1 and 4 are repetitive, with the cat meowing multiple times in each clip. Poor audio quality: All clips have poor audio quality, with either distortion, muffling, or apparent background noises. **Response**: We know these clips are about the class "cat" in the concept set. We can get this answer since the description mentions All of the clips contain the sound of a cat meowing. **Answer**: cat
Example 2	**Description**: They all feature a person snoring loudly. The snoring starts off slow and gets louder over time. The audio is recorded in mono. There are no other sounds in the background. The snoring is described as loud and intense. The audio clips differ in the following ways. The first clip features a man snoring, while the second and fourth clips feature a person snoring (gender not specified). The third clip features a zombie growling and snarling, while the other clips only feature snoring. The third clip is described as scary and creepy, while the other clips are not. The third clip is intended for use in a horror movie or zombie video game, while the other clips do not have specific intended uses stated. The third clip is of poor quality, while the other clips are not specified as such. **Response**: Based on the description, the most suitable classname for the audio clips would be "snoring" or "zombie growling and snarling". Both of these classnames match the description of loud sounds with a strong emotional impact, specifically fear and terror. But "zombie growling and snarling" is not in the given classname set. So the answer is "snoring" **Answer**: snoring

A.4. Others

For the CLIP model, we employ ViT-B/32. For the CLAP model, we utilize the 630k-audioset-best version. To capture the representation of textual artifacts in AND, we use all-MiniLM-L12-v2 pre-trained model provided by Reimers & Gurevych (2019). For all experiments, we use $K = 5$ to select top-K highly/lowly activated samples, $t = 0.7$ to remove similar sentences in summary calibration module.

B. Qualitative Results

B.1. Dissection Pipeline

We provide an example to illustrate the pipeline in Figure 7. We randomly select a neuron from the first transformer encoder's *output* layer in AST with its corresponding top-5 highly and lowly activated audio samples. We highlight the predicted audio source with bold font and similar concepts with the same text color. SALMONN is shown to yield high-quality descriptions that capture correct audio sound source and its detailed acoustic properties for both top-5 highly and lowly activated audio samples. Llama-2-chat subsequently identifies mutual information among these textual representations. Notably, Llama-2-chat tends to produce some greeting messages due to its training objective, we remove these meaningless contents by rule-based filtering. Then, summary calibration utilizes summary of lowly-activated samples to adjust summary of highly-activated samples. For instance, the property *"no background noise"* is removed due to its presence in both summaries. Finally, the calibrated summary of highly-activated samples serves as the output.

B.2. Outputs from Different Modules in AND

To make things clear, we randomly select some neurons in AST, using module A and B of AND to generate dissection descriptions. The results are shown in Table 6. We adopt ESC50 (Piczak, 2015) as probing dataset D_p, and the open-domain acoustic concept set proposed by Kumar et al. (2017) as D_c here.

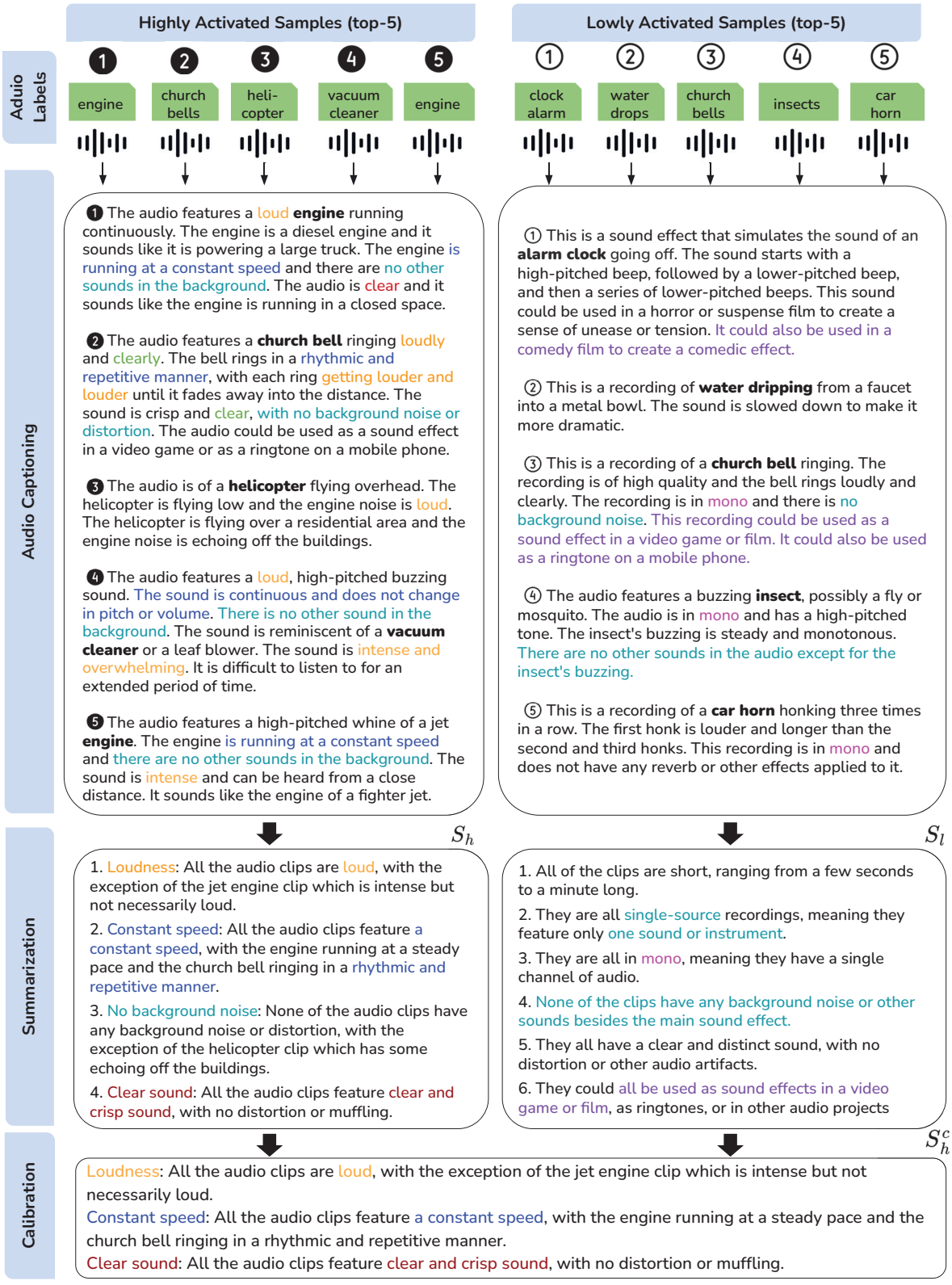


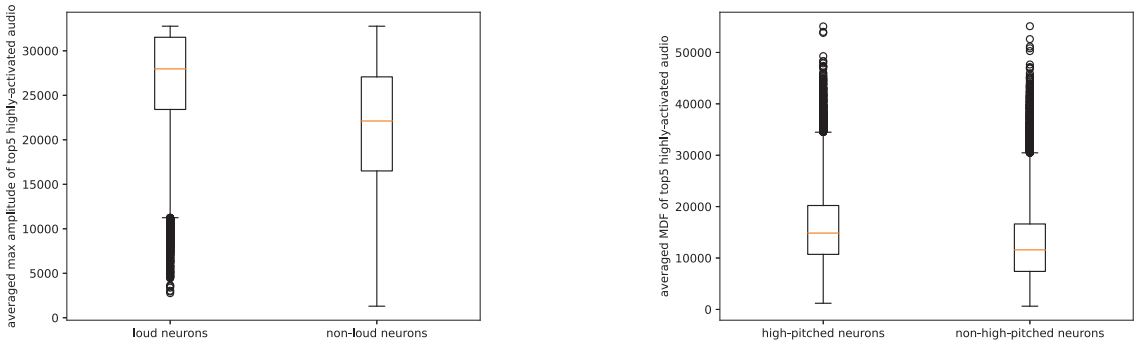
Figure 7. The pipeline of AND's module B. Texts with the same color means they refer to the same concept or property. In this example, a property *"no background noise"* is removed during calibration because it appears in both summaries.

Table 6. Outputs from different modules in AND

Top-5 highly activated audio label	AND (Module A: TAB)	AND (Module A: DB)	AND (Module B: SUM)
• church bells • clock alarm • clapping • door wood creaks • clapping	steam rushing	bell-ringers crucifying	1. All of the clips are recordings of sounds: The first clip is a recording of a bell ringing, the second is an alarm clock ringing, the third is a frog croaking, and the fourth is a recording of a group of people clapping. 2. High quality recordings: All of the clips are of high quality, meaning they are clear and well-recorded. 3. Loud and clear sounds: Each of the clips features loud and clear sounds that are easy to hear. 4. Gradual increase in volume: Three of the clips (the alarm clock, the frog, and the group of people clapping) feature a gradual increase in volume over time. 5. Potential use in video games or movies: All of the clips could be used as sound effects in video games or movies, based on their descriptions.
• glass breaking • water drops • glass breaking • mouse click • glass breaking	ribbon being	glass chattering	1. Loud and intense sounds: All of the audio clips feature loud and intense sounds, which suggests that they may be used to create a sense of drama or urgency in a video or film. 2. High-pitched sounds: Four of the five audio clips feature high-pitched sounds, which could be used to create a sense of tension or excitement. 3. Breaking or tapping sounds: Three of the audio clips feature sounds of breaking or tapping, which could be used to create a sense of impact or action in a video or film. 4. Mono audio: Three of the audio clips are mono, which means they have a single audio channel and may be used to create a more intimate or focused sound. 5. Short duration: Four of the audio clips are short, lasting only a few seconds, which could be used to create a quick impact or effect in a video or film. Overall, these commonalities suggest that the audio clips could be used to create a sense of drama, tension, or action in a video or film, and could be particularly effective when used in quick succession or in combination with other audio or visual elements.

C. Middle Layer Analysis from Basic Acoustic Properties

We measure the basic acoustic features of the top-K highly-activated audio for each neuron in the AST, with "loud" and "high-pitched" adopted. Firstly, we group the neurons based on whether their $C_{\text{open-set}}$ include the word "loud". There are 22612 and 32734 neurons dissected with and without this word respectively. Then, the averaged waveform amplitude of top-K highly-activated audio for each neuron is calculated. As shown in Figure 8a, highly-activated audio of "loud" neurons typically have larger sounds. On the other hand, we use median frequency (MDF) as a measure of audio's representative frequency. As shown in Figure 8b, neurons dissected with the word "high-pitched" tend to be more responsive to high-frequency audio. While the trend is not as significant as the case of "loud" due to this indirect measure of overall frequency.

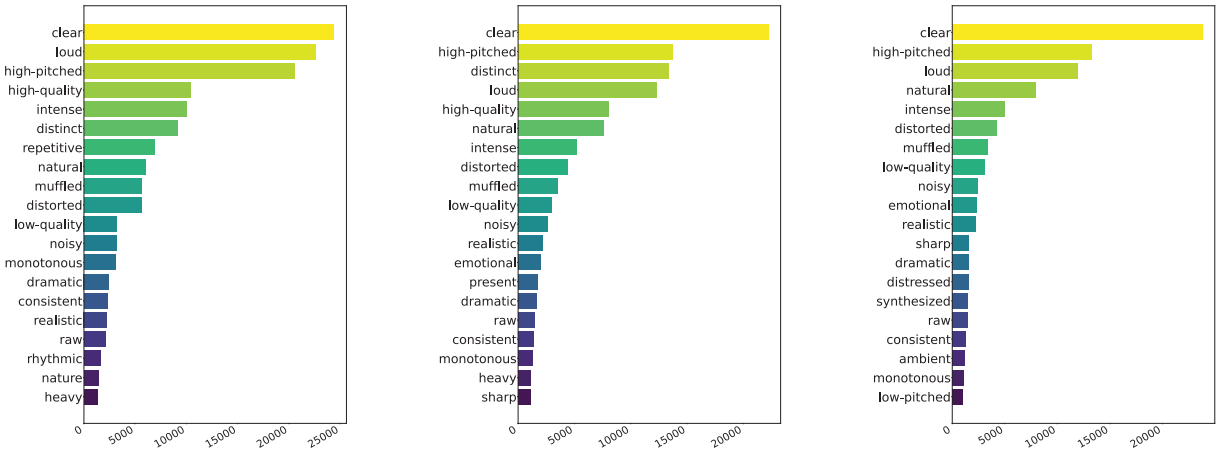


(a) Amplitude. (b) MDF.

Figure 8. The averaged waveform amplitude and the median frequency (MDF) of top-K highly-activated audio for all neurons in the AST.

D. Adjective Distribution of Different Target Networks

Figure 9 displays the distribution of top-20 most common adjectives of all neurons in AST, BEATs-finetuned, and BEATs-frozen on the ESC50, extracted from $C_{\text{open-set}}$. Several common words, such as "loud" and "high-pitched", can be observed across models. We provide an analysis on neuron's properties with respect to these two words as examples in Appendix C.



(a) AST. (b) BEATs-finetuned. (c) BEATs-frozen.

Figure 9. The distribution of top-20 most common adjectives across all linear layer neurons in AST, BEATs-finetuned, and BEATs-frozen on the ESC50. These adjectives are extracted from $C_{\text{open-set}}$

E. Detailed Results of Last Layer Dissection

Table 7 presents detailed results of experiments in Section 4.1. In particular, we evaluate ICL, TAB, and DB of module A on five similarity functions on AST, BEATs-finetuned, and BEATs-frozen. Each method's performance is considered the best result achieved among all similarity functions and is reported in Table 2.

Table 7. Last layer dissection accuracy of ICL, TAB, and DB of module A on AST, BEATs-frozen, and BEATs-finetuned when adopting five different similarity functions.

	cos similarity	cos similarity cubed	rank reorder	wpmi	soft wpmi
Model	**AST**				
AND (Module A: ICL)	72 / 60 / 0.83				
AND (Module A: TAB)	2 / 16 / 0.23	92 / 98 / 0.97	84 / 92 / 0.91	**96** / 100 / 0.98	**96** / 98 / 0.98
AND (Module A: DB)	80 / 82 / 0.85	**100** / 100 / 1.00	86 / 100 / 0.93	88 / 100 / 0.95	2 / 10 / 0.24
Model	**BEATs-finetuned**				
AND (Module A: ICL)	52 / 56 / 0.68				
AND (Module A: TAB)	50 / 80 / 0.64	66 / 94 / 0.77	52 / 80 / 0.67	68 / 94 / 0.77	**74** / 92 / 0.82
AND (Module A: DB)	**76** / 100 / 0.83	74 / 96 / 0.82	56 / 90 / 0.71	62 / 94 / 0.75	2 / 10 / 0.23
Model	**BEATs-frozen**				
AND (Module A: ICL)	16 / 16 / 0.37				
AND (Module A: TAB)	14 / 34 / 0.35	38 / 76 / 0.56	18 / 48 / 0.42	**46** / 72 / 0.60	42 / 80 / 0.60
AND (Module A: DB)	**58** / 82 / 0.69	46 / 82 / 0.62	36 / 72 / 0.56	34 / 68 / 0.53	2 / 10 / 0.23

F. Audio Machine Unlearning Example

In Section 4.3, we discuss the potential use case for audio machine unlearning by AND. We provide an example in Figure 10. When we use OCP (module C) to prune out neurons in BEATs-finetuned associated with "water drops", the classification abilities of BEATs-finetuned on water-related concepts, such as "toilet flush" and "pouring water", are heavily affected, with a smaller influence on other unrelated concepts.

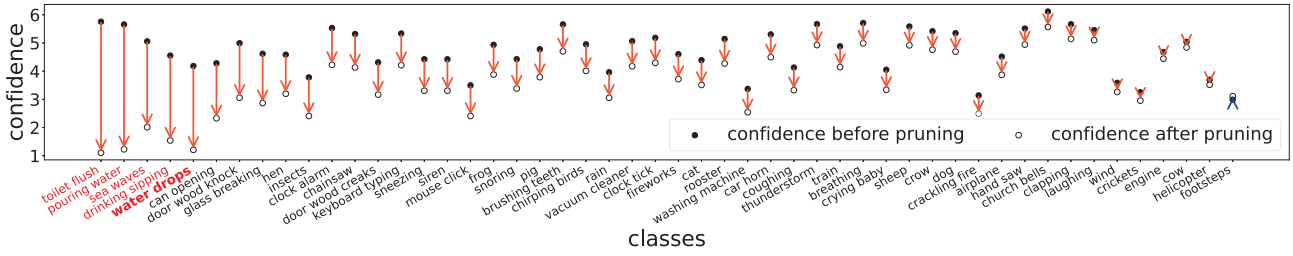


Figure 10. Change of model confidence when neurons associated with **"water drops"** are ablated. Confidence in recognizing water-related audio (with class names labeled in red) decreases, while other sounds are not significantly affected.

G. Neuron Interpretability

G.1. Algorithm

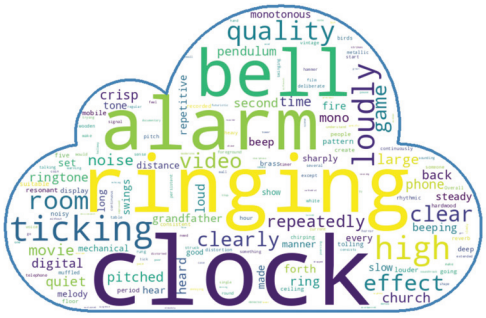
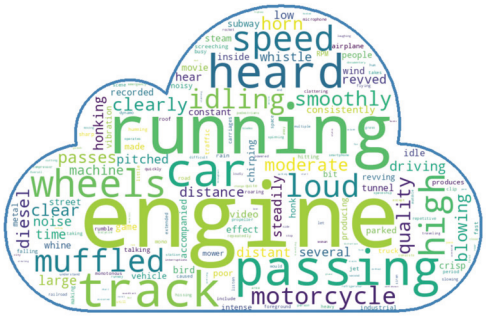
Algorithm 1 GET-UNINTERPRETABLE-NEURONS

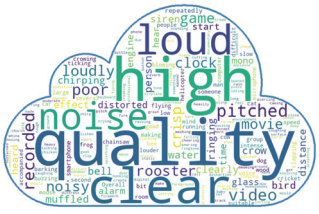
Input: Probing dataset D_p, target network $F(\cdot)$, threshold τ
Initialize sentence pool S, interpreatable neuron pool N_i, and uninterpreatable neuron pool N_u to be empty
for description $d_i \in D_p$ **do**
    for sentence $s_j \in d_i$ **do**
        Add s_j to S
    end for
end for
Train a K-means model $K(\cdot)$ with S
for neuron $f(\cdot) \in F(\cdot)$ **do**
    Get $S_h = d_1, \ldots, d_k$ of $f(\cdot)$
    for $d_i \in S_h$ **do**
        Initialize an empty multiset C_i
        for $s_j \in d_i$ **do**
            Add $K(s_j)$ to C_i
        end for
    end for
    if $|\bigcap_{i=1}^{k} C_i| \geq \tau$ **then**
        Add $f(\cdot)$ into N_i
    else
        Add $f(\cdot)$ into N_u
    end if
end for
return N_i, N_u

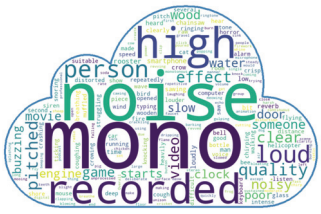
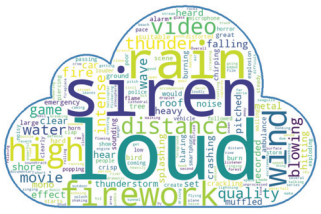
G.2. Clusters in ESC50

After applying K-means clustering, audio captions D_d generated by SALMONN (Tang et al., 2024) are categorized into 11 distinct groups. To delve into the underlying mechanism, we leverage wordcloud package (Oesper et al., 2011) to visualize word distribution of these groups, as presented in Table 8. The wordclouds are generated based on the word frequency within each cluster's corpus: the more frequently a word appears, the larger its font size. The numbers following the cluster id refer to the count of audio captions / sentences in the corresponding group.

For instance, cluster 1 exhibits a strong association with repetitive and high-pitched sounds, such as *clock alarm*, *ringing*, and *bell*. Cluster 2 encompasses various words related to traffic noise, such as *engine*, *motorcycle*, *track*. These wordcloud illustrations vividly depict the effectiveness and rationality of the clustering process.

Cluster 1 (119 / 565)	Cluster 2 (144 / 722)
	

Cluster 3 (424 / 636)	Cluster 4 (819 / 1207)	Cluster 5 (493 / 1449)
		

Cluster 6 (532 / 651)	Cluster 7 (91 / 582)	Cluster 8 (357 / 977)
		

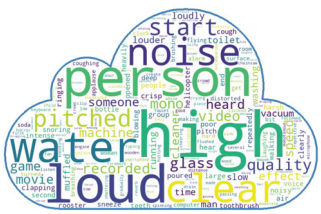
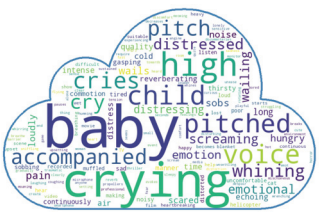
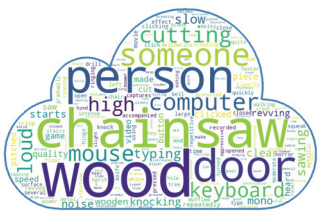
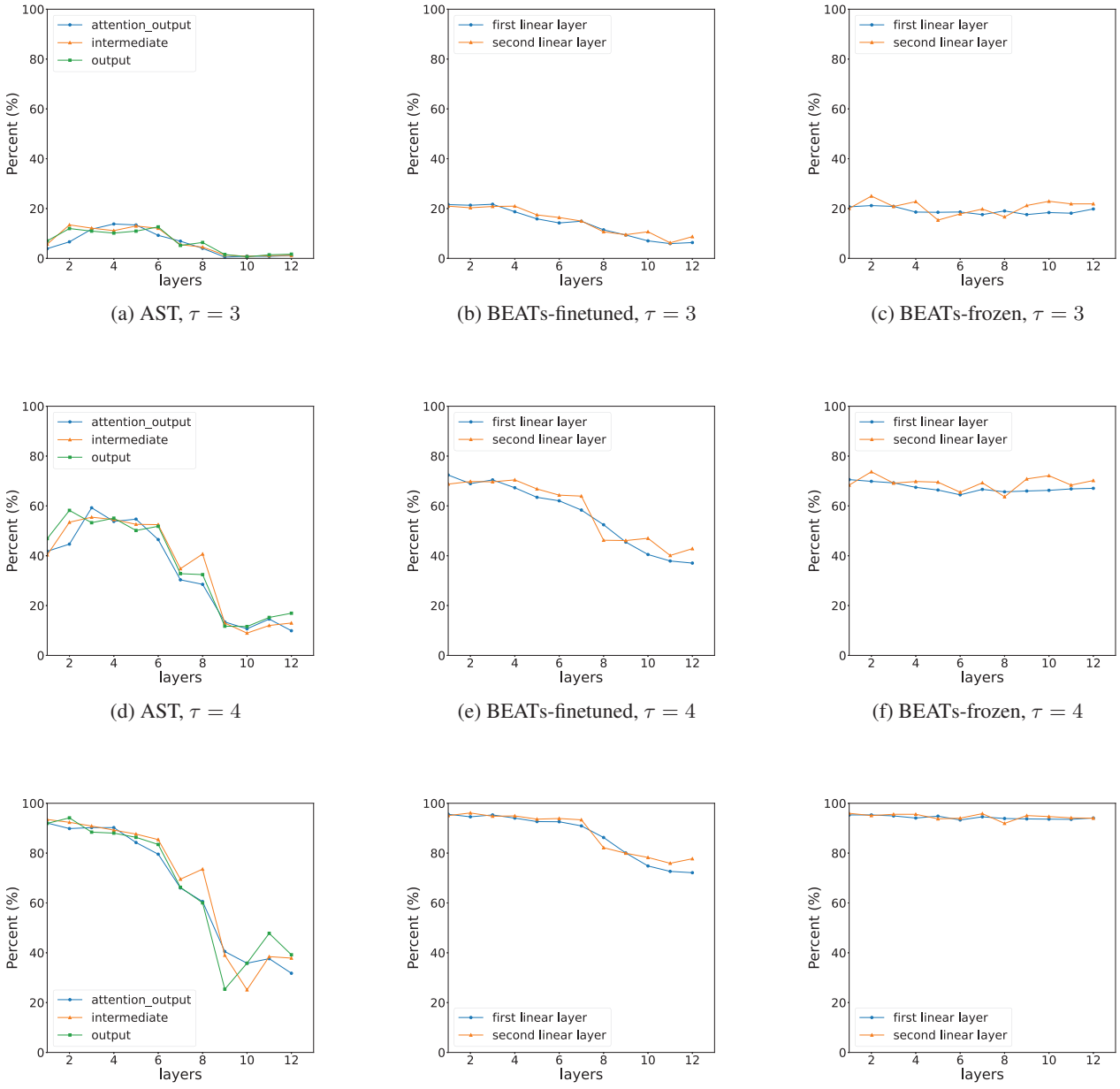
Cluster 9 (891 / 2084)	Cluster 10 (43 / 235)	Cluster 11 (281 / 747)
		

Table 8. Wordclouds of different clusters. The numbers in parentheses refers to the count of audio descriptions / sentences in this cluster.

G.3. Measuring Neuron Interpretability under different τ and top-K

In AND, a threshold τ introduced in Section 4.5 is used to help determine the polysemanticity of neurons. Also, there are some uncertainties about whether selecting top-K highly activated samples causes different trends. Hence, this section conducts further experiments to evaluate the impact of varying τ and top-K. The results are shown in Figure 11, Figure 12, and Figure 13, with top-5, top-10, and top-20 samples selected, respectively. Although the percentage of polysemantic neurons varies with different τ and top-K values, consistent trends as in Section 4.5 are observed.

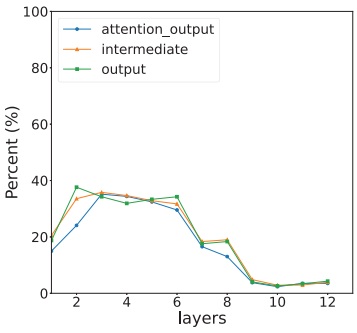


(a) AST, $\tau = 3$ (b) BEATs-finetuned, $\tau = 3$ (c) BEATs-frozen, $\tau = 3$

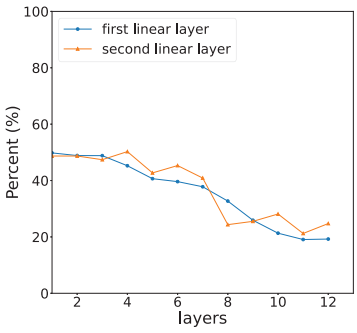
(d) AST, $\tau = 4$ (e) BEATs-finetuned, $\tau = 4$ (f) BEATs-frozen, $\tau = 4$

(g) AST, $\tau = 5$ (h) BEATs-finetuned, $\tau = 5$ (i) BEATs-frozen, $\tau = 5$

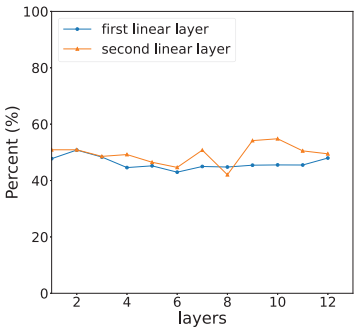
Figure 11. Percentage of polysemantic neurons when adopting $\tau = 3, 4, 5$ and top-$K = 5$



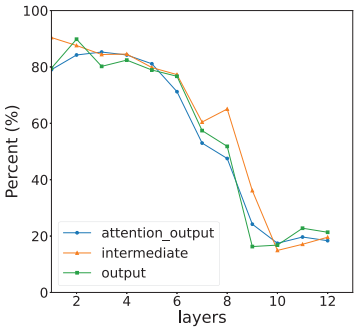
(a) AST, $\tau = 6$



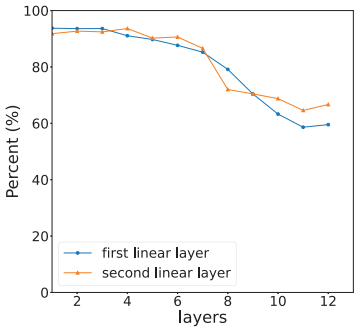
(b) BEATs-finetuned, $\tau = 6$



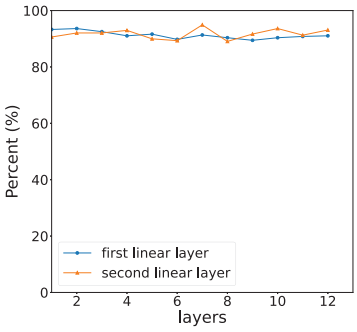
(c) BEATs-frozen, $\tau = 6$



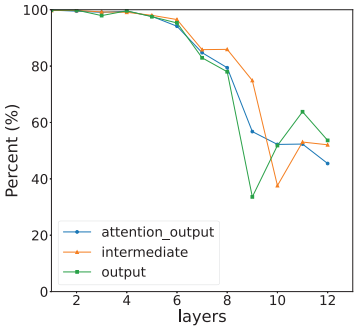
(d) AST, $\tau = 8$



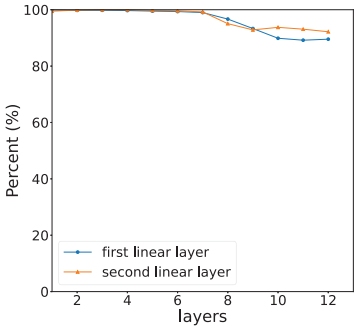
(e) BEATs-finetuned, $\tau = 8$



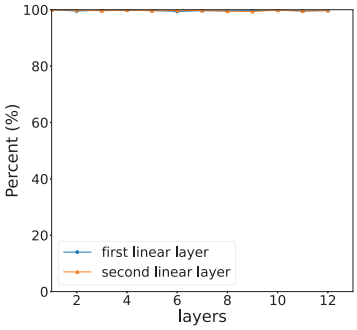
(f) BEATs-frozen, $\tau = 8$



(g) AST, $\tau = 10$

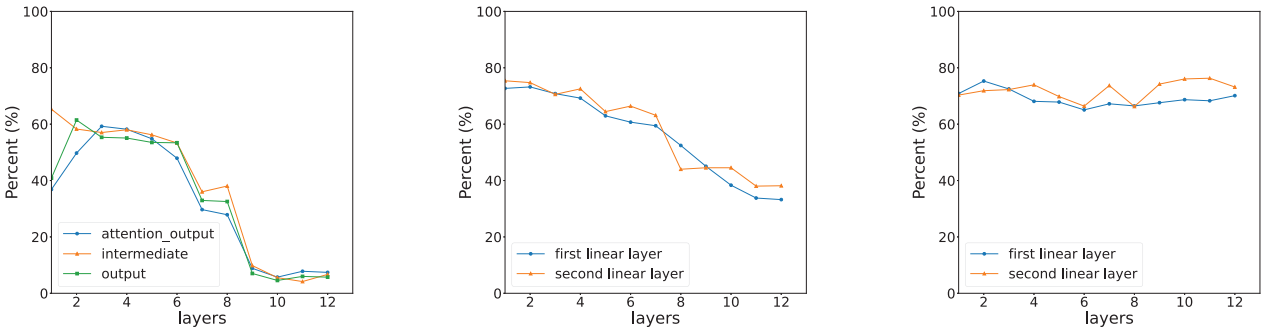


(h) BEATs-finetuned, $\tau = 10$

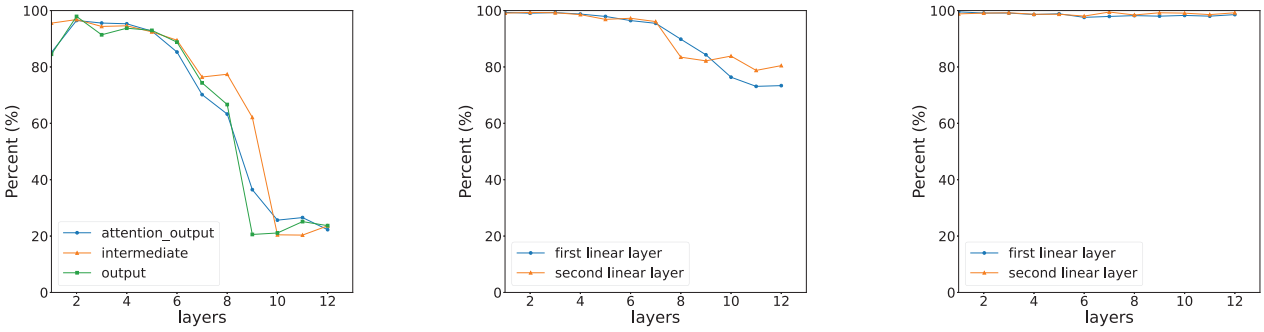


(i) BEATs-frozen, $\tau = 10$

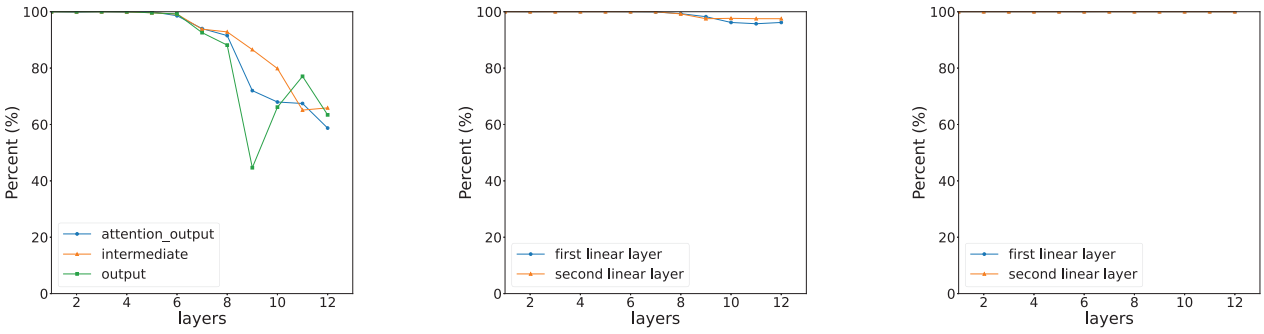
Figure 12. Percentage of polysemantic neurons when adopting $\tau = 6, 8, 10$ and top-$K = 10$.



(a) AST, $\tau = 12$ (b) BEATs-finetuned, $\tau = 12$ (c) BEATs-frozen, $\tau = 12$



(d) AST, $\tau = 16$ (e) BEATs-finetuned, $\tau = 16$ (f) BEATs-frozen, $\tau = 16$

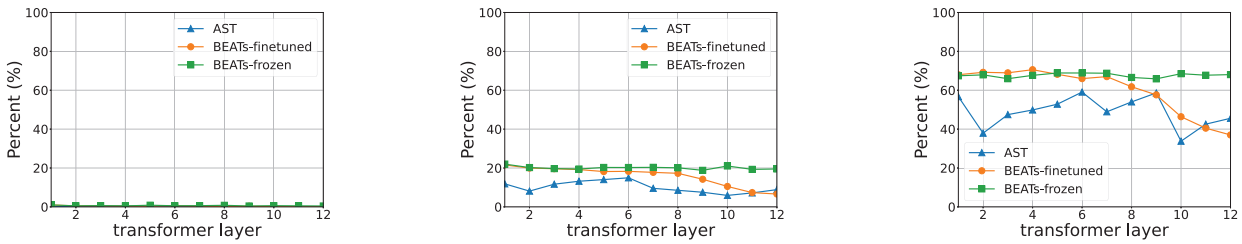


(g) AST, $\tau = 20$ (h) BEATs-finetuned, $\tau = 20$ (i) BEATs-frozen, $\tau = 20$

Figure 13. Percentage of polysemantic neurons when adopting $\tau = 12, 16, 20$ and top-$K = 20$.

G.4. Neuron Interpretability under Different Training Strategies - GTZAN Music Genre

This section conducts the neuron interpretability experiments as in Section 4.5 but adopts GTZAN Music Genre (Tzanetakis & Cook, 2002) dataset as the probing dataset and training dataset of AST, BEATs-finetuned, and BEATs-frozen. The results are shown in Figure 14. Since the GTZAN Music Genre is a simpler dataset with only 10 classes, neuron behaviors are more explainable when responding to samples in this dataset, with no more than 20% neurons classified as "uninterpretable" for $\tau = 4$. Strengthening the criterion to $\tau = 5$ produces a clearer trend, with BERTs-frozen being always diverse, and AST being gradually concentrating, coinciding with the findings on the ESC50 dataset.



(a) $\tau = 3$ (b) $\tau = 4$ (c) $\tau = 5$

Figure 14. Percentage of polysemantic neurons when adopting $\tau = 3, 4, 5$ and top-$K = 5$, with GTZAN Music Genre being the training dataset and probing dataset of AST, BEATs-finetuned, and BEATs-frozen.