LARGE-SCALE AUDIO EVENT DISCOVERY IN ONE MILLION YOUTUBE VIDEOS

Aren Jansen, Jort F. Gemmeke, Daniel P. W. Ellis, Xiaofeng Liu, Wade Lawrence, Dylan Freedman

Google, Inc., Mountain View, CA, and New York, NY, USA

{arenjansen,jgemmeke,dpwe,xiaofengliu,wadelawrence,freedmand}@google.com

ABSTRACT

Internet videos provide a virtually boundless source of audio with a conspicuous lack of localized annotations, presenting an ideal setting for unsupervised methods. With this motivation, we perform an unprecedented exploration into the large-scale discovery of recurring audio events in a diverse corpus of one million YouTube videos (45K hours of audio). Our approach is to apply a streaming, nonparametric clustering algorithm to both spectral features and out-of-domain neural audio embeddings. We use a small portion of manually annotated audio events to quantitatively estimate the intrinsic clustering performance. In addition to providing a useful mechanism for unsupervised active learning, we demonstrate the effectiveness of the discovered audio event clusters in two downstream applications. The first is weakly-supervised learning, where we exploit the association of video-level metadata and cluster occurrences to temporally localize audio events. The second is informative activity detection, an unsupervised method for semantic saliency based on the corpus statistics of the discovered event clusters.

Index Terms— Audio event discovery, unsupervised learning, weakly-supervised learning, streaming clustering algorithms

1. INTRODUCTION

The detection of audio events using supervised classifiers is a well-studied problem where a wide array of learning methods [1, 2, 3, 4, 5] have been evaluated on several small academic datasets [6, 7]. Meanwhile, the Internet is amassing virtually unlimited stores of unannotated multimedia content and researchers have just scratched surface of the opportunities this presents. Collecting this audio is easy, but manually annotating audio events for model training is notoriously difficult. With this motivation, we chronicle a first attempt to leverage the power of unsupervised methods at-scale in the service of audio event modeling. Our task is large-scale audio event *discovery*, the unsupervised identification of repeated acoustic patterns that are realizations of some semantic category. Our approach is to apply a state-of-the-art nonparametric clustering algorithm to a corpus of one million YouTube videos (45K hours of audio), each weakly labeled using a diverse set of 200 audio classes. The result is millions of audio clusters with a wide range of sizes, distributions across the documents, and co-occurrence patterns with the weak labels.

Such large-scale clustering has many potential uses. If the clusters are semantically pure, then they enable a form of unsupervised active learning where annotating only a single example per cluster is required to cover the dataset. In conjunction with a partition of the stream into documents and any associated document-level metadata (e.g. tags, titles), we can derive criteria from cluster occurrence patterns that predict their semantic salience. Specifically, if a cluster consistently co-occurs with a document-level tag, it is evidence that the cluster contains audio events associated with the semantic concept implied by the tag. On the other hand, if a cluster co-occurs equally with all tags, then it is likely that the underlying audio events are not salient for semantic analysis and can be safely ignored.

By including a small collection of manually annotated segments for the 200 categories, we estimate the intrinsic clustering performance. This demonstrates the utility of the discovery method in streamlining large-scale human annotation efforts. We also use cluster co-occurrence with video-level tags as a mechanism for weak-supervision, obtaining competitive performance competitive with a state-of-the-art deep neural network (DNN) trained in-domain. Finally, we propose informative activity detection, a method that estimates the semantic saliency of a cluster based on the class-independent document frequencies across the million video set. We demonstrate that filtering clusters with both abnormally high and abnormally low document frequencies reduces processing requirements while not negatively impacting classification performance.

2. RELATED WORK

In the audio classification domain, unsupervised learning has been explored as a means of data-driven feature extraction in several contexts. The most common is the audio bag-of-words [8, 9] representation, which uses a vector quantization (VQ) of the acoustic space, typically learned with k-means, and represents each audio recording as a histogram over the codebook units. This concept was extended by replacing the simple VQ methods with an elegant hierarchical model, producing large performance improvements [10]. However, these studies were not directed at or evaluated on the direct discovery localized instances of individual audio events. They were also applied to small datasets and were limited to thousands of units, an insufficient number to cover a diverse domain like YouTube.

Inspired by standard information retrieval and natural language processing methods, [11] introduced the concept of *acoustic* stopwords. In analogy to the text-processing concept of removing stopwords (high frequency non-content words, like “a” and “the”), this work demonstrated the value of removing high frequency codebook units from the acoustic bag-of-words representation. Our proposed informative activity detection approach generalizes this filtering to include extremely low frequency discovered units, since their rarity also precludes the conveyance of semantic information. The unprecedented scale of our dataset and the nonparametric nature of our clustering algorithms enable this ability to assert a discovered unit is indeed rare. Finally, recent work has also explored training of models from Internet videos in a weakly-supervised fashion by framing it as a multiple instance learning problem [12].

3. LARGE-SCALE AUDIO EVENT DISCOVERY

3.1. Clustering Algorithm

Given the scale and richness that 45K hours of YouTube videos implies, we have two strong requirements of our clustering algorithm.

First, it must be computationally efficient enough to process $n \sim$ one hundred million input frames. Second, it must support an extremely large number of clusters that grows with the size of the input to flexibly cover the full range of acoustic events that may be present in a diverse domain like YouTube. Classical clustering algorithms that have been considered for audio processing applications in the past [8, 11] do not satisfy these needs. The most commonly used is k-means, which has time complexity $O(nki)$, where i is the number of iterations, and requires k to be explicitly set in advance [13]. With these motivations, we consider the popular streaming clustering algorithm DenStream [14]. Critically, it is nonparametric, allowing the number of clusters to grow over time as the data stream evolves.

DenStream operates over a sequence of input vectors $x_1, \ldots, x_T$, where each $x_t \in \mathbb{R}^d$. The clustering state of the algorithm at each point in time includes a collection of microcluster centroids $M = \{m_1, \ldots, m_j\}$. To process each incoming x_t, a list of the nearest microcluster centroids are retrieved. If there exists any microclusters whose distance (we use cosine distance for all experiments) is less than a threshold ρ, the point gets assigned to the nearest microcluster. If no such microcluster exists, a new microcluster is spawned with initial centroid set to the input x_t. To produce a set of output clusters, $C = \{c_1, \ldots, c_k\}$, online microcluster assignment is followed by an offline microcluster merging step that learns the partition of M into C. Like k-means, DenStream is natively linear in both the number of inputs and the number of clusters (though it does not require several passes over the data). Since we are interested in permitting an arbitrarily large number of clusters k, we introduce two optimizations described below.

First, to reduce the microcluster lookup time, we apply a variant of locality sensitive hashing (LSH) [15] to both the microcluster centroids and each incoming vector to enable (approximate) nearest neighbor search in sublinear time. Each B-bit hash is determined by thresholding B random projections defined by samples from a normal distribution. The threshold used for each projection is set to the median value measured for a small subset of training data. We use 4 hashes of $B = 16$-bits each, and require a maximum Hamming distance of 1 to the input vector using all four hashes in order to consider a microcluster for assignment. Second, we use a witness-based online microcluster merging strategy to avoid expensive offline post-processing of the microclusters. Specifically, we maintain a witness table $W : M \times M \rightarrow \mathbb{R}$, that (sparsely) keeps track of the number of times a pair of microclusters is simultaneously observed in proximity to an incoming vector. Each time an incoming vector falls within distance threshold δ of two microclusters, it is taken as evidence those two microclusters should be merged into the same output cluster. When an entry in this witness table exceeds a specified threshold, their containing output clusters are merged.

The computational savings from these optimizations is dramatic. For example, using 128-dimensional features at frame rate 1.04 Hz, our 1 million video dataset amounts to a stream of $T = 165$M frames. A multithreaded implementation running on 40 cores processes the stream into 30M clusters in under 10 hours. Without the optimizations, it would take weeks.

3.2. Weakly Supervised Learning

When provided video-level metadata, we can derive weak labels that can indicate the categorical audio events likely contained in the video. For example, if a video has a title *Two Dogs Barking*, there is a high chance that there are one or more dog bark audio events contained. If given 5,000 similarly titled videos, one can identify acoustic patterns that are present in a higher proportion than would be found in the general population. This will include patterns of interest (e.g. dog barks), but depending on the specificity of the label, they can include other sounds that are commonly coincident (e.g. growling or whimpering). Even though this sort of weak supervision is provided at the document level, we can use the cluster assignments to temporally localize the regions that are most likely associated with each semantic category. This property can be quantified by the posterior probability $P(l|c)$ of the label l for a given cluster $c \in C$, which can be defined as either the fraction of frames or fraction of documents in each cluster $c \in C$ that have attached the weak label l. Once these posteriors are computed, each frame in a given cluster inherits the set of posterior scores for that cluster. Since every frame is assigned to exactly one cluster, we can reconstruct a posterior vector time series (posteriorgram) for each video, where regions of high score for a given label are likely audio event instances for that semantic category.

3.3. Informative Activity Detection

Our goal of informative activity detection is the isolation of audio segments that are useful for predicting semantic content to save processing time by ignoring everything else. Given a clustering of frames over a weakly labeled set of videos, we can quantify the informativeness of each region of the acoustic space by the maximum pointwise mutual information (PMI) measured between a cluster $c \in C$ and all weak labels in set L, defined as

$$\mathrm{maxPMI}(c) = \max_{l \in L} \mathrm{PMI}(c, l) = \max_{l \in L} \log \frac{P(c, l)}{P(c)P(l)}. \quad (1)$$

Here, $P(c, l)$ is the proportion of documents that have weak label l and are assigned to cluster c; $P(c)$ and $P(l)$ are document-level priors. The PMI is a normalized measure of associativity between a cluster id and a label. Thus, high maxPMI values for a cluster indicates a high degree of positive semantic relevance for at least one category, qualifying it as semantically informative.

Now, the definition of "informative" varies across applications, so ideally we can define a fully task- and label-independent method to make this determination. A low maxPMI cluster must appear in documents with a large variety of semantic categories, while high maxPMI clusters can only occur mainly in documents drawn from a small number of semantic categories. Taken together, these constraints imply a negative correlation between maxPMI and class-independent document frequency of each cluster (the proportion of documents containing the cluster). Thus, we consider the principle behind acoustic stopword removal [11], using lower document frequency as a proxy for higher maxPMI in our determination of informativeness. However, this alone does not address the impact of rare acoustic events. Consider a singleton cluster $c \in C$ whose sole frame occurs in a video with weak label l. In this case, $\mathrm{maxPMI}(c)$ defined in Eq. 1 reduces to $1/P(l)$, which can attain very high values for low-prior classes. While the math is correct that these rare patterns are highly associated with the label, it is not observed enough times to judge whether that association is spurious. This can be mitigated by imposing a minimum document frequency criterion for informativeness, which we evaluate below.

4. EXPERIMENTS

4.1. Dataset and Features

Our evaluation set was drawn from the newly introduced *Audio Set*, a collection of manually annotated audio events [16]. This dataset

consists of 10 second audio clips from YouTube videos, each labeled with one or more audio classes from an comprehensive ontology of 635 categories. We use a subset of 200 classes that had at least 100 verified segments. This included the speech and music top-level classes; 4 gender and age-based speech subclasses; 44 music genres such as rock, heavy metal, hip-hop, and electronic; 65 musical instruments including piano, cymbal, and acoustic guitar; 9 nonspeech human vocalizations like crying, laughing, and sneezing; 15 animal sounds including purring and barking; 17 vehicle sounds including jet engine and car engine; and 44 miscellaneous specific audio events like sawing, vacuum cleaner, and bell. In addition to the 100 verified positives, we also collected 100 verified negative segments for each class (not included in *Audio Set*).

In the experiments below, each class was evaluated independently using just its own explicit positive and negative segments. These labels were used for evaluation only and were not included in the document frequency or posterior calculations described above. To construct the larger weakly-labeled train set, we collected 5000 YouTube videos for each of the 200 classes selected above. The class labels were derived from video-level tags that were automatically extracted from the associated metadata and visual stream [17]. This set is multilabeled since multiple tags can apply to a single video.

We considered two feature representations. The first is non-overlapping 0.25 second context windows of log mel spectrogram (25 ms analysis frames, at a frame rate of 100 Hz). We used 40 mel bins, bringing the dimension of each patch to 1000. To explore the impact of using semantically informed representation, we also considered a data-driven bottleneck feature representation extracted from a VGG-architecture [18] deep neural network audio model [5]. This model was also trained on weakly labeled YouTube data (disjoint from our cluster training set), but using a visual entity detection dataset similar to [19] for which audio bottlenecks served as helpful side information. Only 86 of our 200 classes set were also included in the set of 4923 visual entities of that out-of-domain task, with the overlap mostly limited to musical instruments and animal categories. The model takes as input 0.96 second non-overlapping patches of log mel spectrograms. Here, a longer context window was made possible by the convolutional architecture. The 128-dimensional embedding is defined by the linear (i.e., pre-nonlinearity) output of a 128-unit bottleneck layer.

4.2. Intrinsic Evaluation

The unprecedented scale and multilabeled nature of our YouTube dataset precludes the application of standard cluster evaluation methodology [20], which relies on each frame having exactly one mutually-exclusive groundtruth assignment. Since only a small portion of the data we cluster is annotated, we can only measure the quality of a cluster restricted to elements in it that have a label. Given our labeled segments were randomly selected for verification, we can assume measuring performance with them alone will be representative. However, given the multilabel nature of the data, a present judgment for one class does not imply the absence of the remaining 199.

With these considerations in mind, our performance estimate proceeds as follows. First, we simultaneously cluster the train and test sets. We retain cluster assignments for frames in the evaluation set, which excludes the majority of clusters that only contain train set frames. Next, to accommodate the multilabel nature of the data, we evaluate each label independently. However, this cannot be treated as a binary clustering problem since we cannot make the assumption that negative segments for a label would cluster. This rules out popular single-value summary metrics like adjusted rand index and V-measure [20]. Instead, we measure two quantities: positive label purity (fraction of frames in a cluster from positive segments) and fragmentation (number of clusters frames from positive segments are distributed across) for each cluster. Since our positive-labeled segments are themselves weakly labeled (not all 10 seconds are expressions of the class), we restrict our purity and fragmentation calculation to as many of the largest clusters necessary to include frames of 50 of the 100 positives for each class (denoted Purity@50 and Frag@50, respectively). This prevents metrics from being dominated by clusters of non-target frames from positive segments.

Table 1. Median clustering performance as a function of input feature and microcluster threshold ρ, with microcluster merging turned off ($\delta = 0$). Here, k is the total number of clusters generated by the algorithm.

Features	ρ	k	Frag@50	Purity@50
LogMelSpec	0.1	33.2M	12.0	79.2
	0.15	22.7M	10.0	78.3
	0.2	15.5M	9.0	77.2
Embeddings	0.05	95.1M	35.5	97.8
	0.1	30.1M	15.0	94.1
	0.15	9.7M	10.0	92.6
	0.2	3.6M	8.0	90.9
	0.25	1.7M	6.0	90.0

Table 1 shows median performance across the 200 classes as a function of both input feature types and microcluster radius threshold. Several trends are apparent. First, the use of features based on semantically-informed, out-of-domain audio embeddings greatly improves the purity of the clusters at a equivalent amount of fragmentation. This makes a good deal of sense, since the raw spectral features do not encode even simple normalization transforms (e.g. volume, duration) necessary to reduce within-class variation. This result points to the promise of neural audio embeddings for cross-domain transfer to arbitrary audio processing tasks. Second, we find that increasing microcluster radius threshold controls the usual trade-off between fragmentation and purity that results from varying the number and size of clusters. In this way, the selection of number of clusters used by other clustering algorithms is replaced by a selection of microcluster radius threshold. Finally, we also evaluated the witness table approach for online merging of microclusters, measuring small improvements for some classes when using conservative witness threshold δ values. However, we also found performance to be highly sensitive to this witness threshold; if set too large, the highest occupancy microclusters precipitously merge into a mega-cluster, having the expected deleterious impact on performance. Therefore, we turn off microcluster merging ($\delta = 0$) in the downstream evaluations.

When using audio embeddings, for more than half (as implied by median) of the classes we are able to maintain a purity of at least 90% when fragmenting half of the target segments into only 6 clusters. As a method to nominate segments for manual annotation, this would reduce the manual annotation cost by 88% while introducing less than 10% label noise. Given the promise of even weaker supervision demonstrated in the next section, this may be a quite an acceptable level of label noise for training highly accurate audio event models. Finally, our inspection of per-class performance indicated a bi-modal distribution about this median value. Acoustically specific classes (e.g. accordion, engine knocking) have purities approaching 100% and fragmentations as low as 2. Semantically broad classes (e.g.

Table 2. Audio event classification performance for the baseline DNN model and proposed cluster-based method, using optimal hyperparameter values in all cases.

Method	Mean EER	MAP	Mean P@20
Inception DNN	25.9	79.4	86.1
LogMelSpec Clusters	37.4	65.2	70.5
Embedding Clusters	26.6	78.0	84.6

Table 3. Impact of document frequency (DF) based cluster filtering on classification performance, using the embedding cluster method with optimal hyperparameters.

Min DF	Max DF	% Filtered	MAP
0.0	1.0	0.0	78.0
10^{-5}	1.0	4.4	78.2
0.0001	1.0	6.6	78.3
0.0005	1.0	19.5	77.3
0.0	0.05	2.8	78.2
0.0	0.02	7.2	78.2
0.0	0.01	21.4	77.9
0.0001	0.02	13.8	78.5

video game music, farm animals) generally fall well below median purity values and are not well suited for this methodology.

4.3. Downstream Evaluations

To evaluate utility of our clustering procedure for weak supervision, we again jointly clustered the train and test, leaving the test labels out of the per-cluster posterior calculations. Each test frame then inherits the per-class posteriors of the cluster to which it is assigned. We found that calculating cluster posteriors using document counts gave slightly better performance than frame counts, so we report the former in all cases. We compute segment-level scores for each test segment by averaging the frame-level posteriors for the class(es) the segment was verified for. Thus for each class, we have 100 positive segment scores and 100 negative segment scores with which we can compute the standard classifier metrics equal error rate (EER), average precision (AP), and precision at 20 (P@20). To put our cluster-based performance in context, we also evaluated the performance of a state-of-the-art DNN model based on the Inception convolutional neural network architecture (this was found to be the best architecture in [5]). This weakly-supervised DNN was trained on our train set with the assumption that each video-level label applies to each constituent frame. Like our cluster-based approach, the model produces posterior scores for each test segment frame, which are averaged to produce a segment-level scores.

Table 2 shows the performance average across the classes of the cluster methods and the in-domain Inception DNN model. The performance using optimal hyperparameters are reported in all three cases. We again find that the VGG embeddings provide a significant improvement in the cluster-based performance, indicating the value of semantically informed feature representations (even if they are learned out-of-domain). We also find that the embedding-based performance nearly matches the state-of-the-art Inception DNN model trained on the in-domain data. This provides further evidence that our clusters are successfully discovering a semantically meaningful categorical structure, which enables nearly state-of-the-art classification when coupled with weak-label information.

Finally, we evaluated our proposed informative activity detection method by measuring the impact of filtering clusters before

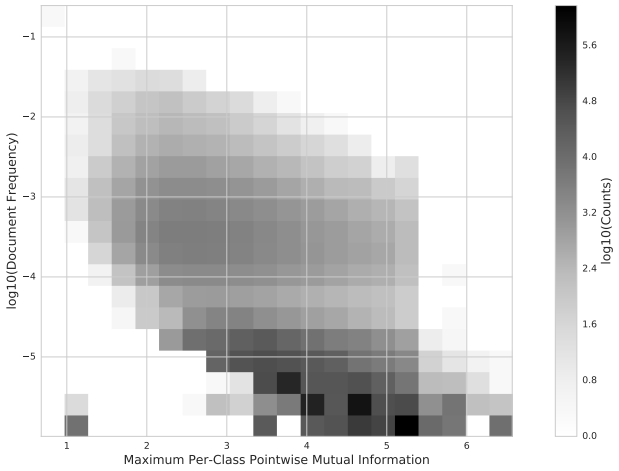

Fig. 1. 2D Histogram of cluster document frequencies and maximum per-class pointwise mutual information values.

segment-level score averaging performed in our weakly supervised classification evaluation. In Section 3.3, we provided an argument for the negative correlation between maxPMI (Eq. 1) and document frequency. Figure 1 shows a 2D histogram of the maxPMI and document frequency pairs for the 3.6 million embedding clusters resulting from the $(\rho = 0.2, \delta = 0.0)$ hyperparameter settings. The negative correlation is apparent, but the exception is in the rare cluster band below 10^{-4}, where outlier frames produce both some of the lowest and highest measured maxPMI values. This unpredictability indicates a potential utility in filtering low document frequency clusters. Using the embedding cluster-based system reported in Table 2 as a baseline, Table 3 lists the percentage of the audio filtered and the corresponding classification MAP using a variety of minimum and maximum document frequency thresholds. By removing clusters that either occur in more than 2% of documents (including the largest cluster, which has document frequency of 25% and is made up entirely digital silence) or in less than 100 documents, we filter 13.8 percent of the dataset duration (over 6K hours) with no degradation of our ability to recognize the audio events. This result indicates substantial utility for domains with higher proportions of uninformative audio than is typical in YouTube videos.

5. CONCLUSIONS

We have performed what is to-date the largest-scale investigation of the unsupervised discovery of recurring audio events. Using a streaming, nonparametric clustering algorithm to discover millions of audio pattern clusters, we have demonstrated the technology's promise for unsupervised active learning, weakly supervised audio modeling, and unsupervised informative activity detection. We expect each of these directions to be useful components in developing high-quality audio event detectors at scale.

6. ACKNOWLEDGMENTS

The authors would like to thank Shawn Hershey and Justin Paul, both with Google, for their help in training the DNN audio models used in this work. We also thank Manoj Plakal, also at Google, for his invaluable technical discussions and infrastructure assistance.

7. REFERENCES

[1] Anurag Kumar, Pranay Dighe, Rita Singh, Sourish Chaudhuri, and Bhiksha Raj, "Audio event detection from acoustic unit occurrence patterns.," in *Proceedings of ICASSP*, 2012, pp. 489–492.

[2] Jort F Gemmeke, Lode Vuegen, Peter Karsmakers, Bart Vanrumste, et al., "An exemplar-based nmf approach to audio event detection," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2013, pp. 1–4.

[3] Tomoki Hayashi, Shinji Watanabe, Tomoki Toda, Takaaki Hori, Jonathan Le Roux, and Kazuya Takeda, "Bidirectional LSTM-HMM hybrid system for polyphonic sound event detection," in *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, September 2016, pp. 35–39.

[4] Huy Phan, Lars Hertel, Marco Maass, and Alfred Mertins, "Robust audio event recognition with 1-max pooling convolutional neural networks," *arXiv preprint arXiv:1604.06338*, 2016.

[5] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson, "CNN architectures for large-scale audio classification," in *Proceedings of ICASSP*, 2017.

[6] Andrey Temko, Robert Malkin, Christian Zieger, Dušan Macho, Climent Nadeu, and Maurizio Omologo, "Clear evaluation of acoustic event detection and classification systems," in *International Evaluation Workshop on Classification of Events, Activities and Relationships*. Springer, 2006, pp. 311–322.

[7] Dan Stowell, Dimitrios Giannoulis, Emmanouil Benetos, Mathieu Lagrange, and Mark D Plumbley, "Detection and classification of acoustic scenes and events," *IEEE Transactions on Multimedia*, vol. 17, no. 10, pp. 1733–1746, 2015.

[8] Stephanie Pancoast and Murat Akbacak, "Bag-of-audio-words approach for multimedia event classification.," in *Proceedings of Interspeech*, 2012, pp. 2105–2108.

[9] Stephanie Pancoast and Murat Akbacak, "Softening quantization in bag-of-audio-words," in *Proceedings of ICASSP*, 2014, pp. 1370–1374.

[10] Sourish Chaudhuri and Bhiksha Raj, "Unsupervised structure discovery for semantic analysis of audio," in *Advances in Neural Information Processing Systems*, 2012, pp. 1178–1186.

[11] Samuel Kim, Shiva Sundaram, Panayiotis Georgiou, and Shrikanth Narayanan, "Acoustic stopwords for unstructured audio information retrieval," in *Proceedings of European Signal Processing Conference*, 2010, pp. 1277–1280.

[12] Anurag Kumar and Bhiksha Raj, "Weakly supervised scalable audio content analysis," *arXiv preprint arXiv:1606.03664*, 2016.

[13] David Arthur, Bodo Manthey, and Heiko Röglin, "Smoothed analysis of the k-means method," *Journal of the ACM*, vol. 58, no. 5, pp. 19, 2011.

[14] Feng Cao, Martin Ester, Weining Qian, and Aoying Zhou, "Density-based clustering over an evolving data stream with noise.," in *Proceedings of SIAM International Conference on Data Mining*, 2006, vol. 6, pp. 328–339.

[15] Piotr Indyk and Rajeev Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *Proceedings of the ACM Symposium on Theory of Computing*. ACM, 1998, pp. 604–613.

[16] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: A strongly labeled dataset of audio events," in *Proceedings of ICASSP*, 2017.

[17] "Google I/O 2013 - semantic video annotations in the Youtube Topics API: Theory and applications," `https://www.youtube.com/watch?v=wf_77z1H-vQ`.

[18] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of CVPR*, 2015, pp. 4694–4702.

[20] Andrew Rosenberg and Julia Hirschberg, "V-measure: A conditional entropy-based external cluster evaluation measure.," in *Proceedings of EMNLP-CoNLL*, 2007, vol. 7, pp. 410–420.