WIKITAG: WIKIPEDIA-BASED KNOWLEDGE EMBEDDINGS TOWARDS IMPROVED ACOUSTIC EVENT CLASSIFICATION

Qin Zhang[1], *Qingming Tang*[1], *Chieh-chi Kao*[1], *Ming Sun*[1], *Yang Liu*[2], *Chao Wang*[1]

[1] Alexa Speech, [2] Alexa AI, Amazon.com Inc

ABSTRACT

In this paper, we introduce Wikipedia-based text embeddings as an auxiliary information to improve the task of acoustic event classification (AEC) by aligning semantic embeddings from different views. We describe how to extract label embeddings from multiple Wikipedia texts using a POS tagging based workflow, and formulate the multi-view aligned AEC problem based on the VGGish model and AudioSet data. We show that our "wikiTAG embedding extraction" approach is a more accurate and robust alternative to label embeddings. Compared to the supervised baseline, our multi-view model achieve 7.3% and 1.3% relative improvement in mAP using 10% AudioSet data and full AudioSet for training, respectively. To the author's knowledge, this is the first work in the AEC domain on building large-scale label representations by leveraging Wikipedia data in a systematic fashion.

***Index Terms*—** Audio classification, Wikipedia, semantic embedding, multi-vew learning, AudioSet.

1. INTRODUCTION

Acoustic event classification (AEC) is the task of detecting whether certain events occur in an audio clip. It has broad applications such as acoustic sensing [1], acoustic scene understanding [2] and enhancing the robustness of ASR [3]. Current state-of-the-art AEC models are data-hungry and large amount of labeled data is needed to achieve high performance [4, 5, 6, 7], therefore limiting its potential in expanding into a wide range of events. One solution is semi-supervised or self-supervised learning that has been shown to have good performance on AEC [8, 9, 10] and non-semantic speech tasks [11, 12]. However, such methods still require a considerable amount of unlabeled audio data that is difficult to acquire due to cost and privacy concerns.

With the massive deployment of smart home devices, there is growing interest in personalizable acoustic event detectors [13, 14]. Shi, et al. [13] formulated few-shot audio tagging to enable detection of new events with very limited labeled data using a meta learning approach. Their results showed superior performance in generalization compared to its supervised counter-part, although it is difficult to interpret what the model is converging to in such a meta-learning setup. Another line of few/zero-shot AEC [15, 16, 17] leverages intermediate semantic representations for both audio samples and their label embeddings extracted by trained text embedding models such as Wrod2Vec [18] and universal sentence encoder [19]. Recently, Xie and Virtanen [15] suggest that label embeddings and sentence embeddings are very useful for zero-shot AEC. They show that hybrid concatenations of embeddings generated with different language models can further boost AEC performance. Inspired by this, we explore whether using a richer text source like Wikipedia (as opposed to labels that are only a few words) can help AEC even more.

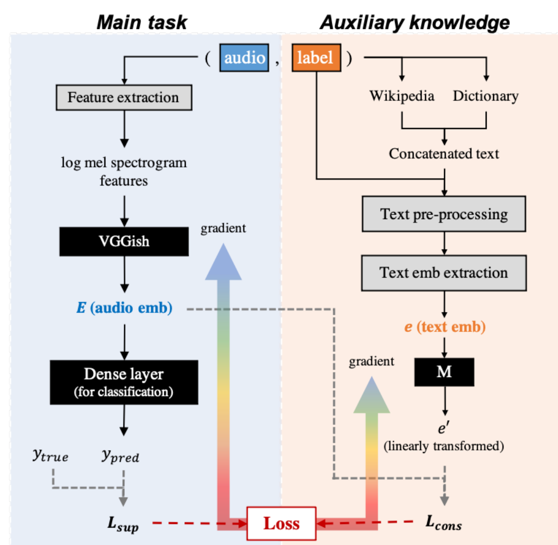

Fig. 1. Overview of Wikipedia-based multi-view aligned AEC. Black modules are trainable, and gray modules are deterministic.

In the Natural language processing (NLP) community, Wikipedia data has been widely used to improve various tasks, including but not limited to named entity recognition (NER) [20, 21], coreference [22], natural language inference [23], text classification [24] and machine translation [25]. Encouraged by these successes, we leverage the public Wikipedia data and extract a comprehensive AEC domain knowledge in the form of text embeddings as a 2^{nd} view to facilitate audio tagging. Such an approach is advantageous as it uses NLP representation learning to acquire more accurate and robust AEC label representations, which potentially capture the complex inter-class and intra-class relationships among different sound events, even when we don't have the audio recordings of a specific sound in the training data. To the authors' knowledge, there is little work done in the AEC domain to use large amounts of wiki text for multi-view learning. For the first time, we provide a way to systematically build large-scale label representations for acoustic events from Wikipedia data to support AEC knowledge expansion.

2. METHODOLOGY

In this section, we describe how to get event representations from Wikipedia and other text sources, as well as a multi-view approach to apply the text representations to benefit AEC.

2.1 Text data acquisition We use the following approach to extract relevant text data from Wikipedia. AudioSet (largest public audio dataset for AEC tasks) [26] event labels are used as an initial set of anchor titles for searching Wikipedia pages. In addition, we use WordNet [27] to find synonyms for each event label, and manually select the ones that are sensible and include their Wikipedia pages. For events that are decomposable words (e.g. "female speech"), we include the wiki pages for all words; the identification of such labels is manual. When an event has multiple wiki pages, we concatenate their content into one single text. We also filter out special words, empty spaces, math equations and citations/references as they are considered less informative. For 25 out of 527 events that can not be found on Wikipedia, we use their definitions from dictionaries such as "Oxford Languages" and "Merriam-Webster". The resulting corpus contains text data for all 527 Audioset events with each containing at least one text source. We also build a supplementary corpus based on the definitions from Oxford dictionary for all 527 events.

2.2 Text pre-processing and embedding extraction We consider 3 approaches for event-wise text embedding extraction:

- **Label embedding**: Word2Vec [18] representation (300D) of the AudioSet labels. When a label contains multiple words, we use the average (same as [15]).
- **Raw text embedding**: Universal Sentence Encoder [19] representation (512D) of the text data collected in section 2.1.
- **wikiTAG text embedding**: find all words in text data that are are nouns, verbs or adjectives using NLTK [28]'s recommended tokenizer, and keep the ones with high cosine similarities *wrt* the corresponding labels using their Word2Vec [18] representations; we then compute the average Word2Vec embedding of the selected words weighted by their occurrences in the text. This POS tagging based word selection makes the text embedding invariant to the order of concatenation of texts from various sources or multiple wiki pages.

The wikiTAG extraction is sensitive to the cosine similarity threshold. When the threshold is 0, we accept all found words; when it is 1, we only accept the ones identical to the label. Intuitively, there is a sweet spot in between which maximizes the richness of the text while not including too many irrelevant words. Given this, we select the words using the density distribution of the cosine similarity and find 0.8 to be the best percentile threshold for our corpus. It should be pointed out that we do not fine-tune any of the pre-trained text embedding models on the AEC domain.

2.3 Multiview alignment We study if aligning the learned audio representation manifold with a fixed text representation manifold extracted from Wikipedia and other text sources (auxiliary knowledge) can help AEC. We do so by adding a two-view alignment loss between text and audio embeddings as a regularizer to the supervised loss (shown in Fig 1). The log mel spectrogram features [29] and the VGGish network (without final FC layers, same as [15]) are used for extracting audio embeddings. A 527-unit dense layer with sigmoid activation is attached to VGGish for event classification as one audio may contain multiple events. In inference time, only audio modality (i.e., the left branch in Fig. 1) is need.

As the goal of this work is to see if text embeddings are useful for AEC instead of exploring new multi-view learning approaches, we simply enforce cosine similarity for the multi-view alignment. We also tried (regularized) linear CCA loss [30], however, the optimization is found to be unstable as VGGish embedding space is very high-dimensional (512D). We have not tried reducing the embeddings to a smaller space which is more friendly to CCA. The overall loss function is shown in Eq.1 where we introduce a hyper-parameter α to adjust the relative importance of the supervised loss L_{sup} and the embedding alignment loss $L_{cons}(= \frac{E_i \cdot Me_i^r}{||E_i|| \cdot ||Me_i^r||})$:

$$L = \sum_{i=1}^{t} [L_{sup}(\hat{y_i}, y_i) + \alpha \frac{1_{r \in Eve(i)} L_{cons}(E_i, Me_i^r)}{\sum_r 1_{r \in Eve(i)}}] \quad (1)$$

In Eq 1, t is batch size, $Eve(i)$ is the set of events present in audio i; y_i, $\hat{y_i}$, E_i, e_i are the label, prediction, audio embedding (512D) and text embedding (Word2Vec: 300D, Universal Sentence Encoder: 512D), respectively. M is a linear matrix to map the text embeddings into the same shape as the audio embeddings, which is shared across all 527 events.

One quick implementation for Eq.1 is to choose one event (that is present in the label) at random and use its text embeddings for calculating L_{cons}. As number of epochs becomes sufficiently large, our stochastic implementation would approximately converge to Eq.1. Please note that we do not simply calculate embedding for a multi-event sample i by averaging over $Eve(i)$. The consideration is that the text embedding models are not fine-tuned on the acoustic event domain, and it is unclear if the arithmetic mean can still preserve the semantic relationship between events. The supervised loss L_{sup} is updated regularly with all the events present in an audio.

3. ANALYSIS OF TEXT EMBEDDING QUALITY

In Fig 2, we visualize the pair-wise cosine similarities of AudioSet event embeddings extracted from various methods as well as the t-SNE projections of the embeddings color-coded by their super-categories (defined based on the AudioSet ontology [26]). Each super-category contains a list of AudioSet events belonging to this category. The list of super-categories includes "human sound", "animal sound", "music instrument","music genre","environmental sound", "vehicle and machines", "alarms" and "explosion and guns". All the super-categories combined cover 379 out of 527 AudioSet events. The audio embeddings shown in the upper left

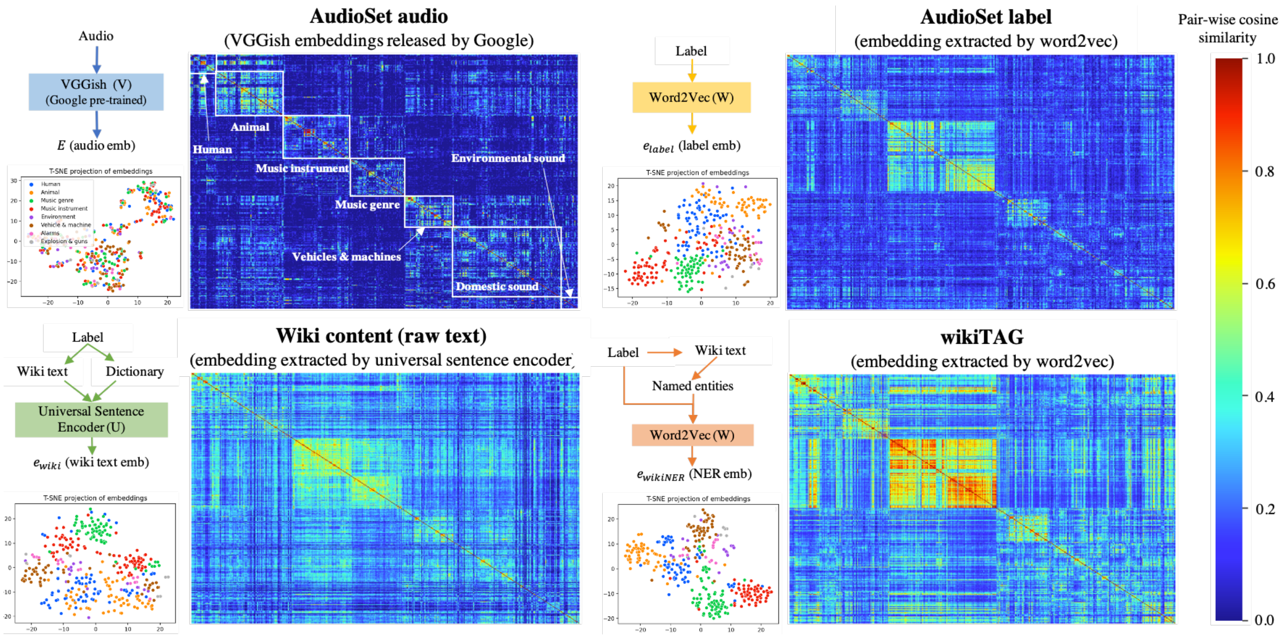

Fig. 2. Pair-wise cosine-similarities and t-SNE projections of emebddings for AudioSet events from various audio and text sources. U (universal sentence encoder [19]), V (VGGish [29]), W (Word2Vec [18]) are pretrained nets for event embedding extraction. In each subplot, x and y axes are AudioSet events in the default order where neighboring events are likely to be in the same super-category.

in Fig 2 are estimated using the mean of the audio embeddings released by Google [29] for each event in the balanced train set (containing roughly 60 audios per event) of AudioSet. It should be pointed out that the model released by Google is trained on a large YouTube dataset (later became Youtube-8M [31]) and one or multiple topic identifiers (from Knowledge Graph) from a set of 30,871 labels [29].

On a high level, all 4 methods in Fig 2 are capable of capturing the inter-class and intra-class relationship to a certain extent. Take super-categories such as "music instrument" and "music genre" as an example. As they are highly related, we not only observe strong diagonal signals within the super-categories themselves, but also cross-diagonal correlation between the two super-categories. Such a trend can also be seen in "human sound" and "animal sound". It is also observed in the T-SNE plots that different methods form clusters on different levels in the event hierarchy. For instance, the audio embeddings released by Google appear to have no distinctive structure while the wikiTAG approach seems to be able to form more compact clusters on a super-category level. One reason for this is that events belonging to the same super-categories typically share common words in their Wikipedia pages which can be picked up by our wikiTAG word selection flow. To quantify this, we measure the event embeddings by ΔS_A defined in Eq. 2, where the 1^{st} term can be viewed as inter-category compactness and the 2^{nd} term is the separateness across super-categories. Higher ΔS_A suggests better representation structuredness on super-category level.

$$\Delta S_A = \frac{\sum_{i,j \in A, i \neq j} S(e_i, e_j)}{\sum_{i,j \in A, i \neq j} 1} - \frac{\sum_{i \in A, j \notin A} S(e_i, e_j)}{\sum_{i \in A, j \notin A} 1} \quad (2)$$

In the equation, A is a super-category; i, j are indices of event labels; e_i is an embedding projection for audio i, and $S(e_i, e_j)$ is the cosine similarity between an embedding pair e_i and e_j.

	Alarm	Animal	Environment	Explosion & guns	Human	Music genre	Music instrument	Vehicles & machines	Average
Audio emb	0.125	0.252	0.103	0.197	0.184	0.271	0.334	0.151	0.203
Label emb	0.087	0.129	0.303	0.162	0.141	0.212	0.097	0.047	0.147
Raw text emb	0.227	0.057	0.103	0.204	0.085	0.195	0.216	0.123	0.151
OxfordTAG emb	0.071	0.113	0.114	0.114	0.176	0.181	0.229	0.405	0.175
wikiTAG emb	0.228	0.216	0.149	0.322	0.135	0.402	0.308	0.229	0.249

Fig. 3. ΔS_A on pre-defined super-categories.

Fig 3 shows ΔS_A *wrt* individual super-categories. As is shown, using $\Delta \bar{S}_A$ (average over super-categories) as a metric, our wikiTAG method out-performs all the other embedding extraction methods. However, the improvement in ΔS_A is not consistent across super-categories, for instance, the audio embedding baseline achieves better ΔS_A in "animal sounds", "human sounds" and "music instrument". We also apply the same word selection workflow to the supplementary corpus collected from Oxford dictionary (we call it OxfordTAG) and found its $\Delta \bar{S}_A$ worse than wikiTAG. We think that it is because Wikipedia is a better text source in terms of richness and descriptiveness for acoustic events.

4. EXPERIMENT AND RESULT

We implement the Wikipedia-based multi-view aligned AEC described in Fig.1 using VGGish network and various text embedding sources including our wikiTAG embeddings. In the following, we describe the dataset used in the experiments and show the model performance on common and rare events. The VGGish network is initialized with Google's pretrained weights [29]. The hyper-parameter α is set to be 0.01 (selected from grid search). Each experiment is repeated 5 times, and we report average mAP and standard deviation.

4.1 Data We prepare 2 datasets for training and evalaution:

1) DS1 accounts for roughly 10% of AudioSet data. This allows us to quickly conduct ablation study on a dataset decent in size and quality. In the dataset, the TRAIN partition contains 200k audios randomly selected from the unbalanced train set (2M audios). The DEV partition is the balanced train set (21k audios). The TEST set is the AudioSet eval set (20k audios). We also group the events into "common events" and "rare events", and limit their numbers in TRAIN and DEV partitions as defined in Table 1. These rare events have to be rare in AudioSet, or more generally, in the nature. They are hand-sorted from AudioSet and found to contain a considerable number of nouns used to describe sounds (e.g. "Bang", "Clang", "Biting", "Ding-dong", "Grunt", "Sigh" and etc.).

Table 1. Event partition in DS1, n_e is # of audios per event. C stands for common events, and R stands for rare events.

Event	n_e in AudioSet	n_e in TRAIN	n_e in DEV
R (34 events)	≤ 500	≤ 5	≤ 2
C (463 events)	> 500	$\gg 5$	$\gg 2$

2) DS2 is the entire AudioSet where the unbalanced train set is used as TRAIN, the balanced train set is used as DEV. For test, we evaluate only on classes with better label quality (80%+ estimated label quality) in the eval set.

4.2 Is wikiTAG useful? Table 2 compares the mAP for different text embedding sources. It is evident that our wikiTAG beats label embeddings extracted by Word2Vec [18] and raw wiki text embeddings extracted by Universal Sentence Encoder [19], outperforming the supervised baseline by 0.013 in mAP. Additionally, it is observed that using a concatenation of label embeddings and raw text embeddings (Multiview^4) performs worse than using either of them individually. It is unclear if this is due to the limited expressiveness of M (shown in Fig 1) which map the auxiliary embedding to a 512D space regardless of the text embedding dimension.

4.3 Does data size matter? Comparing Table 2 and Table 3, it appears that our multi-view AEC set-up with wikiTAG embeddings is more useful for a smaller dataset (DS1).

4.4 Does multi-view AEC help few-shot scenarios? Unfortunately, we observe no improvement in rare events compared to the supervised baseline using mAP metric. We have a few hypotheses. First, wikiTAG indeed organizes the embedding space in a more reasonable fashion for the common events and quite a few rare sound types, like "Grunt", "Throat clearing", "Light engine (high frequency)", "Bird flight, flapping wings", and etc. If we dive deep into the t-SEN plot, we can find that these events are actually grouped together with semantically similar common events. This strongly indicates that the audio event ontology information is better encoded by wikiTAG in the embedding space. However, the wikiTAG learned embeddings do not seem to provide better discriminative power, and thus does not directly contribute to mAP. This leads to our second hypothesis that many of these "audio-rare" events could also be "text-rare", thus failing to provide discriminative power in the event-level. As is mentioned previously, a considerable fraction of rare events are inherently sound events (e.g. "Clang", "Bang", "Grunt", "Breaking", "Ding-dong", they typically do not have Wikipedia pages), which are difficult to describe from the text view.

Table 2. mAP on common and rare events for models trained on DS1. MV stands for "Multiview".

Method	**mAP**, common events	**mAP**, rare events
Supervised baseline	0.177 $\pm$0.001	0.015 $\pm$0.002
MV^1 (label)	0.182 $\pm$0.004	0.013 $\pm$0.002
MV^2 (raw wiki text)	0.185 $\pm$0.006	0.013 $\pm$0.003
MV^3 (wikiTAG)	**0.190** $\pm$0.003	0.014 $\pm$0.002
MV^4 (concat)	0.181 $\pm$0.003	0.014 $\pm$0.002

Table 3. mAP on events with 80%+ estimated quality for models trained on full AudioSet (DS2).

Method	Supervised Baseline	MV^3 (wikiTAG)
mAP	0.319 $\pm$0.002	**0.323** $\pm$0.002

5. CONCLUSION AND FUTURE WORK

We extract a comprehensive AEC domain knowledge from Wikipedia in the form of text embeddings, and formulate the multi-view AEC problem based on VGGish and AudioSet. Through experiments, we find our multi-view approach using wikiTAG embeddings outperforms its supervised counter-part as well as other text embedding sources, making it a promising knowledge source for AEC. However, it is unclear if such an approach can help few-shot AEC given our simple set-up. Though we observe limitations of current wikiTAG generated embeddings, the evidences show that wikiTAG does encode rich semantic information, which provides a few directions for future research. First, we can consider a hierarchical model architecture where we apply the text embeddings which are semantically rich but not event-discriminative on a "super-category" level. This may allow us to incorporate information from the AudioSet ontology. Additionally, we may consider NER-based word selection instead of the current POS tagging based workflow as NER can potentially extract more semantically meaningful entities. Finally, we expect joint training of the text embedding network and the audio embedding network will further improve the peformance.

6. REFERENCES

[1] X. Zhuang et al, "Real-world acoustic event detection," *Pattern Recognition Letters*, pp. 1543–1551, 2010.

[2] D. Barchiesi et al., "Acoustic scene classification: Classifying environments from the sounds they produce," *IEEE Signal Processing Magazine*, pp. 16–34, 2015.

[3] A. Temko et al., *Acoustic Event Detection and Classification*, Springer, London, 2009.

[4] C. Kao et al., "R-crnn: Region-based convolutional recurrent neural network for audio event detection," *Interspeech*, 2018.

[5] A. Kumar et al., "Knowledge transfer from weakly labeled audio using convolutional neural network for sound events and scenes," *ICASSP*, 2018.

[6] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," *ICASSP*, 2019.

[7] C. Kao et al., "A comparison of pooling methods on lstm models for rare acoustic event classification," *ICASSP*, 2020.

[8] Z. Zhang and B. Schuller, "Semi-supervised learning helps in sound event classification," *ICASSP*, pp. 333–336, 2012.

[9] Tagliasacchi et al., "Self-supervised audio representation learning for mobile devices," *ArXiv*, 2019.

[10] A. Jansen et al., "Unsupervised learning of semantic audio representations," *ICASSP*, 2018.

[11] J. Shor et al., "Towards learning a universal non-semantic representation of speech," *Interspeech*, 2020.

[12] J. Peplinski et al., "Frill: A non-semantic speech embedding for mobile devices," *Interspeech*, 2021.

[13] B. Shi et al., "Few-shot acoustic event detection via meta learning," *ICASSP*, pp. 76–80, 2020.

[14] Y. Wang et al., "Few-shot sound event detection," *ICASSP*, pp. 81–85, 2020.

[15] H. Xie and T. Virtanen, "Zero-shot audio classification via semantic embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[16] H. Xie, O. Räsänen, and T. Virtanen, "Zero-shot audio classification with factored linear and nonlinear acoustic-semantic projections," *ICASSP*, 2021.

[17] M. T. Islam and S. Nirjon, "Soundsemantics: Exploiting semantic knowledge in text for embedded acoustic event classification," *IPSN*, 2019.

[18] T. Mikolov et al., "Efficient estimation of word representations in vector space," *ICLR Workshop*, 2013.

[19] D. Cer et al., "Universal sentence encoder," *ArXiv*, 2018.

[20] A. Ghaddar and P. Langlais, "WiNER: A Wikipedia annotated corpus for named entity recognition," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, 2017, pp. 413–422.

[21] H. Johannes et al., "Robust disambiguation of named entities in text," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, July 2011, pp. 782–792.

[22] V. Kocijan et al., "Wikicrem: A large unsupervised corpus for coreference resolution," in *EMNLP*, 2019.

[23] M. Chen et al., "Mining knowledge for natural language inference from wikipedia categories," *ArXiv*, vol. abs/2010.01239, 2020.

[24] Z. Chu et al., "Unsupervised label refinement improves dataless text classification," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Aug. 2021, pp. 4165–4178.

[25] H. Schwenk et al., "Wikimatrix: Mining 135m parallel sentences in 1620 language pairs from wikipedia," in *EACL*, 2021.

[26] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," *ICASSP*, 2017.

[27] G. A. Miller et al., "Introduction to wordnet: An on-line lexical database," *International Journal of Lexicography*, p. 235–244, 1990.

[28] S. Bird and E. Loper, "Nltk: The natural language toolkit," *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 214–217, 2004.

[29] S. Hershey et al., "Cnn architectures for large-scale audio classification," *ArXiv*, 2017.

[30] H. Hotelling, "Relations between two sets of variates," *Biometrika*, 1936.

[31] S. Abu-El-Haija et al., "Youtube-8m: A large-scale video classification benchmark," *ArXiv*, 2016.