CNN-based Audio Event Recognition for Automated Violence Classification and Rating for Prime Video Content

Tarun Gupta[1] *Mayank Sharma* [2] *Kenny Qiu* [1] *Xiang Hao*[1] *Raffay Hamid*[1]

[1] Prime Video Compliance and Classification, Amazon.com
[2] Prime Video International Expansion, Amazon.com

targupt@amazon.com, mysharm@amazon.com, kennyqiu@amazon.com, xianghao@amazon.com, raffay@amazon.com

Abstract

Automated violence detection in Digital Entertainment Content (DEC) uses computer vision and natural language processing methods on visual and textual modalities. These methods face difficulty in detecting violence due to diversity, ambiguity and multilingual nature of data. Hence, we introduce a method based on audio to augment existing methods for violence and rating classification. We develop a generic Audio Event Detector model (AED) using open-source and Prime Video proprietary corpora which is used as a feature extractor. Our feature set includes global semantic embedding and sparse local audio event probabilities extracted from AED. We demonstrate that a global-local feature view of audio results in best detection performance. Next, we present a multi-modal detector by fusing several learners across modalities. Our training and evaluation set is also at least an order of magnitude larger than previous literature. Furthermore, we show that, (a) audio based approach results in superior performance compared to other baselines, (b) benefit due to audio model is more pronounced on global multi-lingual data compared to English data and (c) the multi-modal model results in 63% rating accuracy and provides the ability to back-fill top 90% Stream Weighted Coverage titles in PV catalog with 88% coverage at 91% accuracy.

Index Terms: Audio Classification, Violence Detection, CNN, Digital Entertainment Content

1. Introduction

Amazon Prime Video (PV) is responsible for generating Content Based Rating (CBR) that provides appropriate age designation for PV titles. CBR consists of maturity rating (Example: rated 'R') and content descriptors (CDs) (Example: 'mild violence'). Maturity rating data is a key component to enabling customer and regulator trust, thereby allowing informed viewing decisions to be made on a per title basis. The CDs accompany maturity rating so that the customer understands why a title has received a specific rating. The industry standard is to provide visibility to the amount of sexuality, violence, offensive language, and substance use in each title. To manually generate CBR value for a title, an operator looks for elements related to each CD and assign one of the five levels, namely *None, Mild, Moderate, Strong* and *Severe* to the CD elements they found. Highest CD severity is mapped to corresponding CBR value (all ages, 7+, 13+, 16+ and 18+) and the final CBR rating for a title is determined. Typically, human review process is used to assess the CBR, which is expensive and time consuming. Moreover, expanding catalog and marketplaces necessitates automatic CD classification. In this paper, we present a method for automatic violence and rating classification in PV content based on audio information. A common approach for violence detection in video content is to use computer vision (CV) and natural language processing (NLP) methods on visual and textual modalities of the video [1–4]. However, these methods face difficulty in detecting violence due to diversity and ambiguity in visual and textual signal, global coverage of subtitles and multilingual support for textual models. Given the large size of PV catalog and variability in PV content, PV definition of violence covers a wide diversity of sub-categories and the distribution has a long tail.

While most previous work focuses on a narrow set of hand-picked audio-related events, we use a generic AED model as a feature extractor and probability generator. We experiment with a combination of a sparse set of local audio event categories and probabilities along with dense semantic neural embeddings to obtain a unified local-global fixed length representation for a title. This representation is used to train an audio classifier on a proprietary high quality manually annotated violence CD dataset. Finally, we combine other modalities classifiers with audio to generate a fusion classifier. The models presented in the paper are evaluated on datasets, where the number of videos is orders of magnitude higher than in existing literature.

The main contributions of this paper are: (a) we present unified local-global features, where global features are useful for learning similarities and discovering related patterns (i.e. generalization), and local features are better for matching against specific events (i.e. memorization), (b) we describe several interesting model learnings, revealing the obvious and less-obvious audio events correlated with violence, (c) we show competitive violence detection performance of audio model compared to several baseline methods based on non-audio modality and, (d) we present a multi-modal violence detection and rating classification model using audio, text and video modalities.

2. Related Works

Violence definitions, datasets and methods based on audio from previous work are limited in scale ($\approx$ 100 movies) and do not provide coverage of diversity in regions of interest. In [1, 5], authors define a very specific fixed set of audio-based violence categories such as 'Shots, Screams, Speech, Gunshot, Environmental sounds and Fights' on very small datasets ($\approx$ 10 films). The violent scenes dataset (VSD) [6] is a recently proposed benchmark dataset for violence detection limited to only 3 audio-related categories 'explosions, screams, gunshots' and 6 video-related categories 'blood, fire, firearms, cold arms, car chases and gory scenes' based on 32 movies. The violent flows-crowd violence dataset consists of 246 short YouTube video clips [7] containing 'crowd violence'. Authors in [3] focus on detection of aggressive behaviors and analyzed on CareMedia aggression dataset containing 42 aggressive clips. Datta et. al, [4] exploited the accelerate motion vector to detect fist fighting, kicking using

a manually recorded dataset with 8 people. Cheng et. al, [8] proposed a hierarchical approach to recognizing gunshots, explosions, and car-braking and evaluated on dataset of 5-min movie segments from 5 movies; Lin et. al, [9] describe a weakly-supervised audio violence classifier combined using co-training with a motion, explosion and blood video classifier to detect violent scenes in movies on a tiny dataset of 3 movies. There has been work on violence detection using verbal signals [2].

3. Methodology

In this section, we describe the proposed technique for audio-based violence detection. First, we outline the design and architecture of the Audio Event Detection (AED) model. Second, we present the violence classification approach which leverages the pre-trained AED model as a feature extractor.

3.1. Audio Event Detection

To detect presence of audio events, we pre-train a CNN based AED model on a combination of proprietary Digital Entertainment Content (DEC) audio event corpora consisting of 2900 movies (DEC-1100 and DEC-1800) and their corresponding subtitles, publicly available FSDKaggle2019 [10] and Google Audioset [11]. This model classifies 120 categories of sounds which are most frequently displayed as a part of subtitle files for deaf and hard of hearing (SDH), also known as captions. The audio categories include sound effects, music and other background noises which are important to the on-screen action. We use the timing of captions or plot pertinent sounds (such as dog barking, traffic sounds, gunshots, explosion, etc.) from subtitle file and extract the audio corresponding to those timings. For other datasets, we use the clips containing these 120 categories. We further divide the datasets into 2s (s: seconds) overlapping clips (sampled at 48 kHz) with 50% overlap. The clip timing of 2s is chosen because 90% of the caption durations in our proprietary datasets are < 2.3 s. This results in the following distribution of 2s clips: a) DEC-1100: 51K, b) DEC-1800: 90K, c) FSDKaggle2019: 151K, d) Google Audioset: 9K. We discard existing labels and re-label the clips by 2 annotators to minimize human error by retaining the clips with agreement between the two annotators and discarding the rest, resulting in 200K clips across 120 classes. We use log scaled mel-spectrogram (log mel STFT) as the feature input [12]. The log mel STFT consists of 128 components (bands) covering the frequency range (0-48 kHz). We use a window size of 25 ms (1200 samples at 48 kHz) and hop length of 15 ms (720 samples at 48 kHz), resulting in a signal of size $R^{134\times128}$. However, we observe a huge data imbalance across classes ($\approx$ 7500 in worst case). Hence, we use spec-Augment [13] to re-balance the dataset. This results in 1.5M samples across 120 classes. Finally, we divide the pre-training dataset into train, test and validation set with ratios 75%, 15% and 10%.

We compare 4 architectures for AED model. First, we use a VGGish CNN with Time Distributed layer (CNN-TD) [14] with 4 CNN blocks consisting of 2 layers each. Each block contains two 2D (3×3) conv layers with 1st, 2nd, 3rd and 4th block containing 64, 128, 256 and 512 channels respectively along with Batch Normalization and PReLU activation function followed by (3×3) MaxPool with stride of 2. Following a time dimension average pooling layer (TD), we add a 2048 dimensional dense layer and a 120 dimensional classification layer resulting in 13M parameters. Second, we use a 2 layered bidirectional GRU network [15] with two fully connected layers (256 and 120 dimensions) consisting of 0.6M parameters. Third, we use ResNeXt [16] containing 27 conv layers, followed by 1 dense layer (120 dimensions) and having 8.4M parameters. Fourth, we use PANN consisting of CNN14 architecture described in [17]. Table 1 presents the performance comparison of various AED methods across metrics including Accuracy, weighted AUC and weighted Recall. The weighted metrics take into account the class imbalance. We observe that CNN-TD results in best AUC and weighted Recall. Hence, we use CNN-TD as our generic AED model. Next, we describe the violence classification model.

Table 1: *Comparison of various AED methods on the pre-training test set.*

Model	Accuracy	AUC	weighted Recall
GRU	54.7%	0.972	63.7%
ResNeXt	63.8%	0.984	73.1%
CNN-TD	71.8%	**0.9876**	**77.1%**
PANNs	**73.22%**	0.9546	58.31%

3.2. Violence Classification from audio

The violence classification and rating generation from audio consists of feature extraction and rating generation from AED. **Feature Extraction and training:** The violence classifier uses two levels of features extracted from AED as shown in figure 1. First, it uses audio event category probabilities ($x^i \in \Re^{120}$) max-pooled over every 2s clips ($i \in [T]$; T is total number of 2s clips in a given video) in the title ($x = \max_i x^i$) and second, it uses mean pooled embeddings $h = \frac{1}{T}\sum_{i=1}^{T} h^i;\ h, h^i \in \Re^{2048}$ from the AED across every 2s clips. We experimented with several pooling and combination techniques and found the aforementioned combination to perform the best. The event category probability correlates with the predefined AED categories and the embedding representations have the potential to identify ambiguous mix of violence events throughout the title or the categories which aren't present as a part of AED train set. Finally, we freeze the AED model weights, append the two embeddings ($z = [x; h]$) and train the title level binary violence detection model using binary cross-entropy loss on our proprietary dataset. **Rating Generation** Since, our maturity ratings are ordered (*none, mild, moderate, strong, severe* violence), we perform a simple threshold-based ordinal regression to transform the title-level violence probability into a 5-way rating.

4. Baseline Violence Classification Models

We also define several binary classifier baselines across video, image and text modalities to compare performance of our proposed model.

Image: Following [18], we use a single frame architecture by treating a video as a bag of frames extracted at 1 s intervals. We use a pre-trained ResNet50 [19] as a feature extractor and train a frame level violence model.

Early fusion: This model fuses title level video summary features and subtitles to train a title level violence classifier. It employs temporal mean and max pooling of ResNet50 embedding across all video frames to generate video level embedding. Additionally, we extract TF-IDF features from the movie subtitles and synopsis. Finally, we train a classifier using the concatenated set of visual and textual embeddings. However, pooling over a title results in a loss of contextual information.

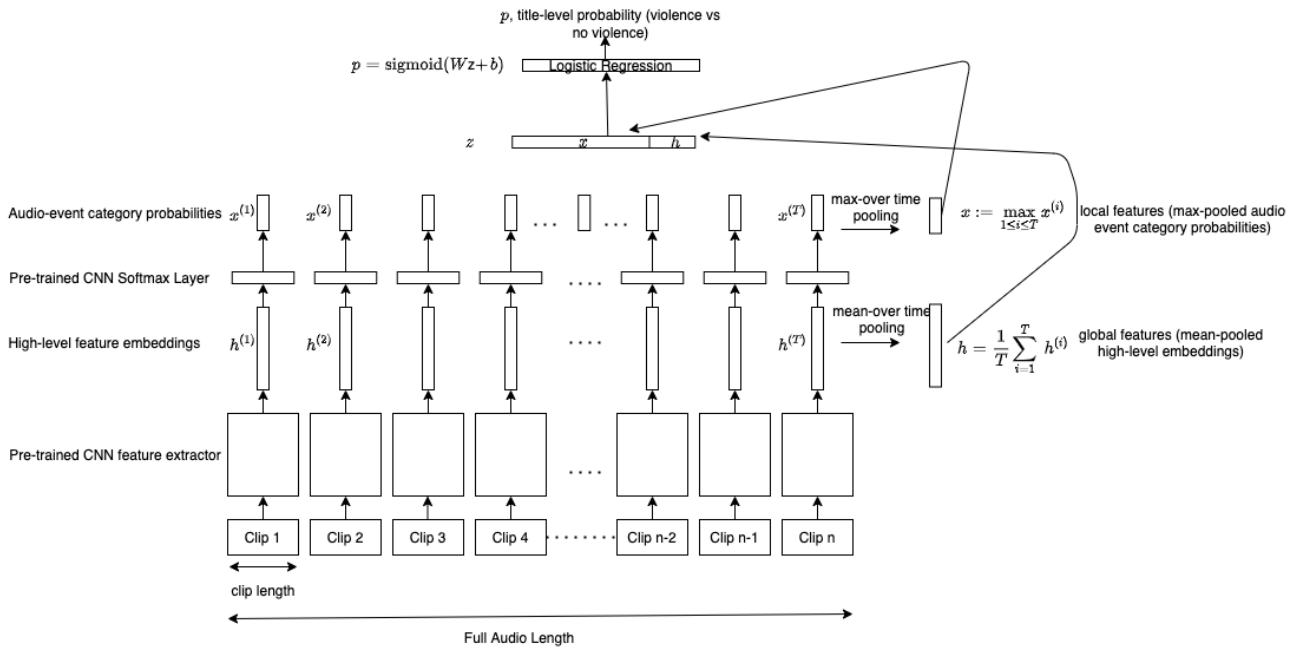

Figure 1: *Generation of local-global audio representations for titles using pre-trained CNN-TD AED followed by violence modeling.*

Clip: To understand short context better compared to early fusion or image models, we use video summary features from short-clips of up to 2 min duration. During inference, the model is executed on all 2 min segments of the video to generate clip-level scores, which are averaged to obtain title-level score.

BM25: A BM25 scorer [20] is employed to estimate violence presence within subtitles using keywords as a search query. The keywords are selected from training data based on Chi-squared feature selection method. This model generates an unbounded score in range $[-\infty, +\infty]$. We use sigmoid function to convert score in the range [0, 1].

Genre: Title genre can provide information about violence. For example, an action movie is likely to contain violent parts while a children's video may not. We exploit this information by adding a logistic regression model on the title genre information encoded as a multi-hot vector.

5. Experiment Setup

Datasets: In this work, we use manually annotated proprietary PV DEC corpora for training and evaluation. For violence model training with binary labels 'Contains' and 'None', we annotate a dataset of 3K titles from US marketplace. For US marketplace evaluation, we use another 3K titles annotated with binary and multi-level labels from Top-90% SWC titles. For Global marketplace evaluation, we use a dataset of 7K titles with binary labels from Top-90% SWC titles from United States (US), United Kingdom (UK), Germany (DE), Japan (JP) and rest-of-world (ROW) marketplaces. The annotators also identify the time duration associated with violent scenes which are used to train and validate the baseline classifiers. Each dataset is fairly balanced across two categories. For multi-level CD rating classification, we use 25% of 3K titles from US marketplace evaluation set for training and validating the ordinal regressor and remaining 75% for evaluation. We tune the weight decay hyperparameter of all the models on the validation set of US marketplace dataset from the set {1e-04, 1e-03, 1e-02, 1e-01, 1, 10, 100} and select the best model based on evaluation measures.

Evaluation measures For binary classification, it is important to control the chance of predicting a 'Contains' title as 'None', because it is riskier to present an adult content to kids. Thus, we define the metrics as the recall at precision of 97% (R @ P97) for 'None' classification, and use recall at precision of 90% (R @ P90) for 'Contains' classification by choosing two thresholds. During testing, any data in between the two thresholds is deemed 'unknown' and sent for human evaluation. For multi-level rating, we use prediction accuracy.

Fusion model: While the majority of the existing methods focus on one of the modalities, we experiment with a fusion (multi-modal) of all the aforementioned independent learners along with audio using probability averaging. The fusion unit helps tackle the diversity in input data and violence types.

6. Results

6.1. Violence Classification Results

Model Insights: Audio Event Category Feature Importance: In order to assess the feature importance of audio event category in violence detection, we experiment with two methods. First, we perform a L1 Logistic Regression (LR) based feature selection on standardized max-pooled audio-event probability features space. We progressively increase the L1 regularization until there are only ≈10 features selected by the model. Second, we plot random forest (RF) feature importance using the gini importance method [21]. These results are plotted in the figures 2 and 3 respectively. As expected, gunfire and explosion are strongly predictive of violence. However, we observe several interesting less-obvious events. Tone refers to strength of signal and appears to be positively correlated with violence. Additional, less-obvious categories include siren and thud which could be symptomatic of violence. Negative predictive features like instrument-play and cheer do not indicate violence.

Best model comparison against baseline: Table 2 reports

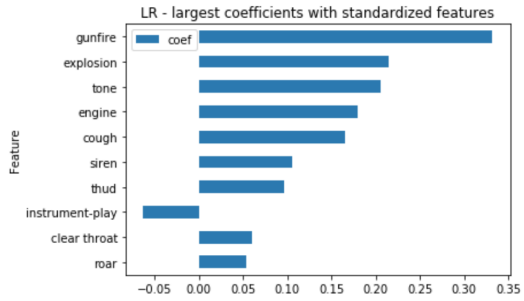

Figure 2: *Feature importance plots for L1 logistic regression (L1-LR).*

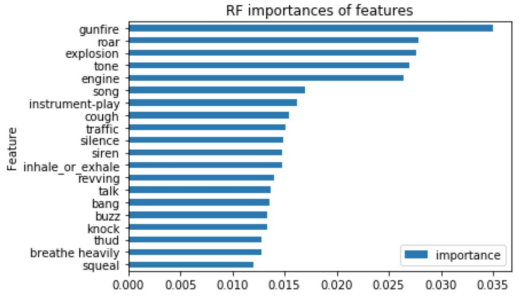

Figure 3: *Feature importance plots for Random Forest (RF).*

performance on evaluation data from US and Global marketplaces using our audio method (section 3) and baseline (section 4) methods. Audio based approach is independently able to predict violence and outperforms most existing visual and textual models across one or more metrics. Especially for R @ P97 for 'None' category, our approach outperforms the second best method by 29% and 45% (relative) respectively for the two evaluation sets. For 'Contains', early fusion is the best independent method closely followed by audio.

Table 2: *Performance of individual models on US and Global evaluation data*

Marketplace	US		Global	
base learner	R @ P97 (None)	R @ P90 (Contains)	R @ P97 (None)	R @ P90 (Contains)
early fusion	22.44%	**87.14%**	6%	**87.53%**
clip	7.87%	72.81%	8.98%	82.69%
bm25	0%	79.85%	0%	54.29%
genre	5.71%	68.78%	2.92%	69.63%
image	4.23%	51.28%	0.67%	61.67%
audio	**31.86%**	80.05%	**11.08%**	78.22%

Fusion performance: Table 3 presents a comparison of the fusion model's performance with and without audio method. Results indicate superior performance by adding audio modality. We also perform a 2x over-weighting of audio and early fusion methods as they are top performing models and observe best performance. Each of the 6 classifiers in the fusion model, generates a score in the range of [0,1]. The base fusion model averages these scores to generate a title level score between [0,1] which is thresholded to generate violence/no-violence category for the given title. With 2x overweighting, we compute a weighted average of the individual classifier scores with weight of scores of the audio and early fusion models is set to $\frac{1}{4}$ which is twice that of other classifiers (set to $\frac{1}{8}$) as they are among the top performing models. Hence, final 2x weighting score is given by the eq. 1:

$$\text{score}_{\text{fusion}} = \frac{1}{4}\text{score}_{\text{audio}} + \frac{1}{4}\text{score}_{\text{early-fusion}} + \frac{1}{8}\text{score}_{\text{clip}} + \frac{1}{8}\text{score}_{\text{bm25}} + \frac{1}{8}\text{score}_{\text{genre}} + \frac{1}{8}\text{score}_{\text{image}}. \quad (1)$$

Table 3: *Results of fusion models on US and Global data.*

Marketplace	US		Global	
model	R @ P97 (None)	R @ P90 (Contains)	R @ P97 (None)	R @ P90 (Contains)
Fusion without audio	33.66%	87.19%	18.93%	82.06%
Fusion with audio	37.01%	88.78%	23.86%	85.80%
Fusion (2x overweight)	**38.29%**	**89.75%**	**25.76%**	**88.74%**

Table 4: *Title-level multi-level rating accuracy on US data*

Method	Accuracy
early fusion	58.60%
clip	52.73%
bm25	50%
genre	54.93%
image	48.92%
audio	59.10%
Fusion (2x overweight)	**62.32%**

As audio model is language agnostic, the benefit due to audio model is more pronounced on Global data (26% relative improvement in R @ P97 metric for Global compared to 12% for US for 'None'). The final fused model allows us to backfill top 90% SWC titles at 88% coverage and 91% accuracy in PV.

5-way ratings: We also present 5-way rating prediction performance in the table 4. Similar to observations for binary classification, audio and early fusion are top-2 methods. The 2x over-weighted fusion method delivers a 5% (relative) lift in performance compared to best individual method.

7. Conclusions and Future Work

In this paper, we presented a violence detector based on a pretrained AED model which generates a high quality audio representation. Our violence detector uses a combination of explicit sparse audio event category features and an implicit set of dense feature embeddings extracted from AED. Our evaluations on datasets from US and Global PV catalog show that audio model leads to significant improvement over individual modalities. The proposed audio-based model outperforms all baselines across one or more metrics. Experiments also indicate that the benefit due to audio based approach is more pronounced on Global data compared to US data. Finally, we developed a stronger multi-modal detector by fusing the individual learners. On a 5-way CD classification, the fusion model outperforms other independent learners by at least 5% (relative). To address some of the problematic cases like slapstick and peril violence, we plan to collect more granular labeled data both in terms of inputs (i.e. image level and clip level labels) and labels (e.g. Slapstick - Slippery Skid). Following which, we plan to build multi-class multi-modal models to address these subcategories where data is sparse.

8. Acknowledgements

Authors would like to thank Amazon Web Services (AWS) for providing relevant infrastructures for experiment.

9. References

[1] O. Deniz, I. Serrano, G. Bueno, and T.-K. Kim, "Fast violence detection in video," in *2014 international conference on computer vision theory and applications (VISAPP)*, vol. 2. IEEE, 2014, pp. 478–485.

[2] V. R. Martinez, K. Somandepalli, K. Singla, A. Ramakrishna, Y. T. Uhls, and S. Narayanan, "Violence rating prediction from movie scripts," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 671–678.

[3] D. Chen, H. Wactlar, M.-y. Chen, C. Gao, A. Bharucha, and A. Hauptmann, "Recognition of aggressive human behavior using binary local motion descriptors," in *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE, 2008, pp. 5238–5241.

[4] A. Datta, M. Shah, and N. D. V. Lobo, "Person-on-person violence detection in video data," in *Object recognition supported by user interaction for service robots*, vol. 1. IEEE, 2002, pp. 433–438.

[5] Y. Potharaju, M. Kamsali, and C. R. Kesavari, "Classification of ontological violence content detection through audio features and supervised learning," *International Journal of Intelligent Engineering and Systems*, vol. 12, no. 3, pp. 20–230, 2019.

[6] M. Schedi, M. Sjöberg, I. Mironică, B. Ionescu, V. L. Quang, Y.-G. Jiang, and C.-H. Demarty, "Vsd2014: a dataset for violent scenes detection in hollywood movies and web videos," in *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*. IEEE, 2015, pp. 1–6.

[7] Y. I. T. Hassner and O. Kliper-Gross, "Violent flows: Real-time detection of violent crowd behavior," in *3rd IEEE International Workshop on Socially Intelligent Surveillance and Monitoring (SISM) at the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2012. [Online]. Available: www.openu.ac.il/home/hassner/data/violentflows/

[8] W.-H. Cheng, W.-T. Chu, and J.-L. Wu, "Semantic context detection based on hierarchical audio models," in *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, 2003, pp. 109–115.

[9] J. Lin and W. Wang, "Weakly-supervised violence detection in movies with audio and video based co-training," in *Pacific-Rim Conference on Multimedia*. Springer, 2009, pp. 930–935.

[10] E. Fonseca, M. Plakal, F. Font, D. P. W. Ellis, and X. Serra, "Audio tagging with noisy labels and minimal supervision," in *Proceedings of the Workshop on Detection and Classification of Acoustic Scenes and Events 2019 (DCASE 2019), New York University, NY, USA, October 2019*, M. I. Mandel, J. Salamon, and D. P. W. Ellis, Eds., 2019, pp. 69–73. [Online]. Available: http://dcase.community/documents/workshop2019/proceedings/DCASE2019Workshop_Fonseca_33.pdf

[11] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. IEEE, 2017, pp. 776–780. [Online]. Available: https://doi.org/10.1109/ICASSP.2017.7952261

[12] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," *IEEE Signal processing letters*, vol. 24, no. 3, pp. 279–283, 2017.

[13] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[14] R. Hebbar, K. Somandepalli, and S. Narayanan, "Robust speech activity detection in movie audio: Data resources and experimental evaluation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4105–4109.

[15] H. Phan, P. Koch, F. Katzberg, M. Maass, R. Mazur, and A. Mertins, "Audio scene classification with deep recurrent neural networks," *arXiv preprint arXiv:1703.04770*, 2017.

[16] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[17] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[18] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[20] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford *et al.*, "Okapi at trec-3," *Nist Special Publication Sp*, vol. 109, p. 109, 1995.

[21] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.