On the detection of acoustic events for public security: the challenges of the counter-terrorism domain

Anna Pompili[1]*, Tiago Luís*[1]*, Nuno Monteiro*[1]*, João Miranda*[1]*, Carlos Mendes*[1]*, Sérgio Paulo*[1]

[1]VoiceInteraction, Portugal

anna.pompili@voiceinteraction.pt, tiago.luis@voiceinteraction.pt

Abstract

Massive amounts of audio-visual contents are shared in public platforms everyday. These contents are created with many purposes, from entertaining or teaching, to extremist propaganda. Civil security actors need to monitor these platforms to detect and neutralize security threats. Generating actionable knowledge from multimedia contents requires the extraction of multiple information, from linguistic data to sounds and background noises. Information extraction demands audio-visual annotations, a costly, time-consuming task when performed manually, which hinders the analysis of such an overwhelming amount of data. This work, performed in the context of the EU Horizon 2020 Project AIDA, addresses the challenge of building a robust sound detector focused on events relevant to the counter-terrorism domain. Our classification framework combines PLP features with a convolutional architecture to train a scalable model on a large number of events that is later fine-tuned on the subset of interest. The fusion of different corpora was also investigated, showing the difficulties posed by this task. With our framework, results attained an average F1-score of 0.53% on the target set of events. Of relevance, during the fine-tune phase a general-purpose class was introduced, which allowed the model to generalize on 'unseen' events, highlighting the importance of a robust fine-tune.

Index Terms: sound events detection, DNN architecture, CNN architecture, robust fine-tune, counter-terrorism, AIDA project

1. Introduction

Over the last years, the detection of sound events gained an increasing interest from both the research community and the software industry, as confirmed also by the inception of a yearly-based event, the Detection and Classification of Acoustic Scenes and Events (DCASE) Challenge. This growing relevance is motivated by the large number of scenarios in which this type of technology may be useful, from inclusive technology [1, 2], to health condition identification [3], surveillance [4, 5], and public security [6–8]. Among these areas, the last is experiencing an urgent need of robust sound events detectors in order to actively participate in the fight against crime and terrorism. Nowadays, multimedia data are created and shared at a very fast pace, generating a large volume of data whose manual analysis hampers a timely intervention. Nevertheless, the immediate block and removal of terrorist content from online platforms is of vital importance to avoid the spread and diffusion of this type of information. To this end, the use of audio processing technology become fundamental to allow the automatic analysis of large amount of data.

This work focuses on the detection of sound events, a twofold task requiring to determine first the presence and temporal location of sounds in audio signals and then to establish their nature. Sound events have many different acoustic characteristics, some may be very short, transient-like, while others may have longer duration. Additionally, events usually are polyphonic, meaning that multiple sounds occur concurrently, leading to partial or full temporal overlap. Due to the combination of multiple sources, the detection of overlapping events is much more challenging than the detection of isolated ones. Early works in this area combined Mel-frequency cepstral coefficients (MFCC) features with Hidden Markov Models (HMM) [9] or Gaussian Mixture Models (GMM) [10]. More recently, the release of large sound events datasets provided within the DCASE challenges, allowed the adoption of deep architectures, such as Convolutional Neural networks (CNN) [11, 12] or Convolutional Recurrent Neural Networks (CRNN) [13, 14].

Overall, the majority of the research works reviewed use datasets that were made available in the context of the yearly-based DCASE series, for which benchmark results are also provided. These datasets are usually focused on specific domains of interest (e.g., urban or domestic sounds). In this study, on the other hand, we are interested in a specific set of events that are relevant to Law Enforcement Agencies to rapidly identify extremist propaganda content. Since there are no similar resources available, we had to create our own dataset from large sound events corpora. The lack of an adequate dataset raised several challenges, not only for the inability to compare the outcome of this research with existing benchmarks, but also in terms of heterogeneity and amount of training data available for each class. Thus, our approach investigated the fusion of different corpora, in order to collect enough data to support deep learning, data-hungry techniques. Using an incremental approach, we explored a scalable strategy based on the creation of a general model trained on a large number of classes, later fine-tuned on the subset of interest. To allow the model to generalize on 'unseen' sound events, in the fine-tune phase a general-purpose class has been introduced, allowing, this way, to reach an average F1-score of 0.53%, our best classification result.

2. Related work

Mesaros *et al.* [9] targeted the task of detecting and recognizing 61 isolated acoustic events from real life environments recordings. The authors modeled the problem with a three-state left-to-right HMMs and MFCCs as feature representation. On the classification task alone, the maximum accuracy achieved was of 54%, while on the joint task of detecting and recognizing events, the accuracy dropped to 24%, with an overall error of 84% provided by the detection phase. More recently, Gauvain *et al.* [10], performed a research on the detection of acoustic events to support the fight against terrorism, similarly to the purposes of this work. Due to the lack of dedicated corpora and to the large disparities found in the available data, the authors focused on four acoustic events: explosions, gunshots, machine

guns, and Nasheed (singing). The problem was modeled using GMMs, results were presented by accounting for the number of correct detection and false alarm, which hampers a direct comparison. The Nasheed class achieved the best performance, while for the other three categories there were more false alarms than correct detections. According to the authors, this behavior was possibly due to the presence of other polyphony instances in the data, with continuous events (i.e., speech or wind) overlapping with the impulsive ones (i.e., machine gun).

Traditional approaches, like the ones described above, are not suited to detect multiple events at the same time, representing a major limitation since sound events usually co-occur in time. In contrast, neural network-based models, especially deep architectures, lend themselves naturally to multi-label classification since multiple output neurons are active at the same time. Thus, with increasing dataset sizes, deep neural networks become the dominant strategy in this area. Gorin et al. [12] combined log Mel filter bank features and CNN for detecting and classifying 18 polyphonic sound events. The system, developed in the context of the DCASE 2016 challenge, was focused on sound events detection in real life audio. Since limited data was provided by the organizers, the authors explored data augmentation techniques such as audio mixing and speed perturbation, which showed a marginal improvement in the performance. Overall, the system developed was able to achieve an F1-score of 44.1%. Phan et al. [14] investigated a multi-label, multi-task framework based on a CRNN to combine both isolated and overlapped audio events. The evaluation was carried out on two datasets, the first contained isolated events, the second was synthetically created in the context of the DCASE challenge to investigate overlapped events. On these datasets, the proposed approach achieved an average F1-score of 97.6% and 64%, respectively.

3. Methodology

In this work, we target the detection of polyphonic sound events using deep neural network models. The lack of publicly available corpora on the domain of interest, however, required the creation of a dedicated dataset. This process raised several questions regarding the amount of data, the type of classes to select, and the training approach to follow in order to build a robust sound events detector. To answers these issues, our methodology investigated an incremental approach based on the use of three different corpora. Additionally, the lack of benchmarks relevant to our purposes motivated the use of two different architectures, a DNN and a CNN. The first was previously developed in the scope of a national civil security project, AGATHA, and thus used as a baseline. The corpora and the network architectures are described in detail in the remainder of this Section.

3.1. Corpora

To build the dataset required for the project AIDA[1] we investigated the sound events classes contained in three large corpora publicly available: AudioSet [15], Freesound Dataset 50k (FSD50K) [16], and VGG-Sound [17].

AudioSet comprises a collection of 2,084,320 multi-label, human-labeled, 10-second sound clips drawn from YouTube videos and an ontology[2] specified as a hierarchical graph of event categories, for a total of 527 sound events classes. The corpus is divided in three sets, a balanced train, an unbalanced train, and an evaluation set containing, respectively, 22.176, 2.042.985, and 20.383 samples. Labels are provided at the clip-level (i.e., weak labels), the labeling process could be considered correct, but potentially incomplete. In this study, we used the balanced train and the evaluation partitions.

FSD50K is a multi-label corpus of human-labeled sound events containing 51,197 variable length (i.e., from 0.3 to 30s) clips, for a total of 108.3 hours. The corpus encompasses 200 sound classes drawn from the AudioSet Ontology. Data is collected from Freesound, an online collaborative audio clip sharing site [18]. Ground truth labels are provided at the clip-level. Audio clips are grouped into a development and an evaluation set, containing, respectively 40,966 and 10,231 samples. In the development set labels are correct, but could be occasionally incomplete. The evaluation set is instead labeled exhaustively, meaning that labels are correct and complete for the considered vocabulary.

VGG-Sound is an audio-visual correspondent (i.e., the sound source is visually evident in the clip) corpus consisting of short clips of audio sounds, extracted from videos uploaded to YouTube. The dataset contains over 200,000 clips for 309 different sound classes. Each clip is 10 seconds long. Differently from AudioSet and FSD50K, the set of sound labels is flat (i.e., there is no hierarchy). Audio clips are grouped into a training and a test set. In this research, only a subset of 17 classes was selected from this corpus, leading to 20,468 audio clips.

From these corpora we are interested in 17 classes that include sounds related with explosion, (i.e., artillery fire, explosion, gunshot, and machine gun), vehicle (i.e., airplane, train, engine, and siren), and some specific human sounds (i.e., crowd shouting, screaming), among others. We will refer to the dataset created with these 17 events as *AIDA_17C*. To achieve good classification results on this set of events however, we also created different datasets of variable size, according to the training strategy with which they were used.

3.2. Data pre-processing

The first stage of this work addressed the analysis and standardization of the three corpora. In fact, as mentioned above, the set of classes existing in the FSD50K corpus is drawn from the AudioSet ontology, while the same do not apply for the VGG-Sound corpus. For this reason, it was needed to create a mapping between the classes of this corpus that were considered relevant and the ones contained in the AudioSet ontology. Then, the data from the three corpora were re-distributed since they originally provide two partitions, one for development or training purposes, and one for evaluation. Nevertheless, to comply with our baseline approach, the two partitions of each corpus were merged and then divided into a training, validation, and evaluation partitions. The amount of data retained in each partition was of 80%, 10%, and 10%, respectively. Finally, audio files were down-sampled to 16kHz to compute 26 Perceptual Linear Prediction (PLP) features.

3.3. Baseline and extended DNN architecture

As briefly mentioned, the baseline DNN architecture was previously developed in the scope of a national civil security project, AGATHA. It consists of a sequence of five densely connected layers (with 1024, 1024, 512, 512, and 128 hidden units), followed by a global max pooling layer. Batch Normalization [19] and Dropout [20] was also applied before the non-linearity (ReLU) of each fully connected layer. The final layer uses a softmax activation function, making it unsuitable for multi-

[1] https://www.project-aida.eu/

[2] https://github.com/audioset/ontology

Table 1: *Evaluation on AIDA_17C of the models trained with the extended DNN.*

	Individual Models			Joint Model		
	Precision	Recall	F1	Precision	Recall	F1
VGG-Sound	0.59	0.82	0.68	0.54	0.85	0.66
FSD50K	0.52	0.71	0.60	0.49	0.70	0.58
AudioSet	0.38	0.53	0.44	0.45	0.72	0.55

Table 2: *Evaluation on AIDA_17C of the model Agatha trained with the baseline DNN.*

	Agatha		
	Precision	Recall	F1
VGG-Sound	0.91	0.53	0.67
FSD50K	0.89	0.50	0.64
AudioSet	0.95	0.70	0.80

label classification. The model trained with this architecture was also used as a baseline. It was focused on the identification of generic sound events, containing 8 classes, out of 50, that were pertinent to the purposes of the AIDA project. We will refer to this model as *Agatha*.

In this research, the baseline DNN architecture was extended to support multi-label classification. To this end, the loss function and the last layer of the original model were updated to use binary-crossentropy [21] and the sigmoid activation function, respectively. The whole file is provided as input to the network, which is analyzed with a window size of 0.5 second.

3.4. CNN architecture

In the experimental phase a VGG-like architecture was also exploited, a model widely used in computer vision and recently successfully employed also in sound events detection tasks. The network comprises seven convolutional layers, three of 48 filters, two of 96 filters, and the last two of 128 filters. Output feature maps are summarized by concatenating global max pooling and global average pooling per channel. The first five convolutional layers have a receptive field of (3,3), while for the last two layers this value is decreased to (2, 2). All convolutional layers are followed by Batch Normalization and ReLU activation. Between each group of convolutional layers with the same number of filters, max-pooling of size (2,2) is applied. Output feature maps are summarized by concatenating global max pooling and global average pooling per channel. Input to the network is provided as chunks of 1 second length. Each chunk inherits the label of the clip, independently if it contains the event or not.

4. Experiments and results

In the following sections, first the evaluation performed with the extended DNN architecture is described. These experiments were mainly focused on the union of different corpora and on the assessment of the impact that the use of variable-size datasets may have on classification performances. Then, further experiments were conducted with the DNN architecture, focusing on a specific corpus. Results were assessed using precision, recall, and F1-score, computed according to their standard definition. More specifically, metrics were computed globally, by counting the total true positives, false negatives and false positives.

4.1. Deep neural network

4.1.1. Models focused on a reduced set of events

To have a general understanding of the performance achievable with the data of the three corpora individually, three preliminary models were trained using the extended DNN architecture on each *AIDA_17C* dataset. These results are presented in Table 1. The model *Agatha* was then assessed on the same sets of data (see Table 2). A comparison of these outcomes showed that the three individual models achieved, on average, a slightly lower F1-score with respect to the baseline model *Agatha*. In particular, while *Agatha* generally achieved higher precision and lower recall, the three new models presented an opposite behavior (i.e., low precision, high recall).

Then, to take advantage of the greater amount of data available, the three individual datasets *AIDA_17C*, generated from the three corpora, were combined to train a new model. The evaluation of this joint model was performed on each corpus individually. From the results, also reported in Table 1, it was generally observed a reduction of the performance with respect to previous experiments, except for the dataset generated with the AudioSet corpus, which is more challenging and has benefited from the union of the other datasets.

With a small set of sound events classes, like *AIDA_17C*, the model may be unable to generalize on sound events not seen before, and thus may introduce a considerable number of false positives, as was indeed shown by the results. To mitigate this problem, a general-purpose class, containing the complement of sound events not selected in *AIDA_17C*, was introduced. This dataset is referred as *AIDA_18C*. The drawback of this approach, however, is that the model may be prone to classify every event as the general-purpose class, which was confirmed from preliminary experiments. To reduce the importance of the general-purpose class, both automatically estimated class weights and a sampling strategy were implemented. In both cases the classification results were more balanced in terms of precision and recall, but the overall performance of the models was lower than the ones reported in Table 1.

From the results of this preliminary evaluation is it possible to observe that, both for each corpus in isolation and on a joint collection of the three corpora, the dataset *AIDA_17C* appeared to be inadequate to build a robust sound events detector. This conclusion is corroborated by the consideration that the model *Agatha* is indeed more generic, containing 50 heterogeneous sound events.

4.1.2. General model, robust fine-tune approach

The conclusions of the previous experiments suggested the creation of a more general model, targeting a wide set of different sound events. To this end, we built the dataset *AIDA_414C*, which contained all the events of the AudioSet ontology with the exclusions of the nodes 'Music', 'Human voice', and their children. With this dataset, several models were trained in order to experiment different parameters, including the number of hidden layers and neurons, and the batch size. The use of two regularization techniques was also investigated, namely Mixup [22] and Dropout. These models were then fine-tuned on the smaller dataset *AIDA_17C*. The fine-tune phase experimented the incremental freezing of the weights of the last four layers. Overall, the evaluation of these models has shown a worsening of the performance in comparison to the models trained on the three datasets *AIDA_17C*. We hypothesized that this outcome could be due either to a problem of data imbalance, or to inter-corpus different acoustic and recording set-

Table 3: *Evaluation on AIDA_17C of the models Agatha and the CNN-based.*

	Agatha			**AIDA_138C fine-tuned on AIDA_17C**		
	Precision	Recall	F1	Precision	Recall	F1
FSD50K	0.89	0.50	0.64	0.80	0.68	0.74

Table 4: *Evaluation on AIDA_18C of the models Agatha and the CNN-based.*

	Agatha			**AIDA_138C fine-tuned on AIDA_18C**		
	Precision	Recall	F1	Precision	Recall	F1
FSD50K	0.40	0.50	0.44	0.81	0.39	0.53

tings. Challenges may derive not only from the number of samples contained in each class, but also from the duration of different events, which may vary from very short events like a closing door, to more sustained one, like wind. Such irregularities in the training data may lead to a disproportional representation of samples [23]. To partially mitigate a possible data imbalance problem, the algorithm that inspects the ontology to select children's nodes was modified to discard those ones that were present only in a minority of the three corpora. Unfortunately, the evaluation of the models built on this dataset did not confirm our expectations. Classification results did not improve with respect to our initial experiments, achieving an F1-score of 0.63% (VGG-Sound), 0.50% (FSD50K), and 0.54% (AudioSet).

While it is not completely clear the impact that the number of classes and their variety may have, the outcomes of these experiments appeared to confirm that the union of the three corpora does not provide the expected improvement in terms of performance.

4.2. Convolutional neural network

The results achieved so far, although do not discourage to pursue the research on the training of a general model, point out the need to focus on a single corpus. We choose FSD50K, since particular care was provided in the design of this resource. In fact, we recall that this corpus was created using human-labeled data and an exhaustive labelling procedure for the evaluation set, which guarantees that annotations are correct and complete, while labeling error in AudioSet is estimated at above 50% for $\approx$18% of the classes [16]. At this stage, preliminary experiments with the extended DNN and the CNN architectures were performed on the whole FSD50K corpus. The results showed an F1-score of 0.36% and 0.40% respectively, justifying the adoption of the latter. Of relevance, the re-distribution of data described in Section 3.2, led to an absolute improvement of 11%, achieving an F1-score of 0.51%.

In order to determine the best subset of sound events classes for the general model, and thus prune events frequently misclassified, classes-specific performances were analyzed. To this end, we built a 200D matrix using the classification results. However, the creation of such a matrix in a multi-label setup rises the issue of properly identifying and assigning the errors produced by the model, especially in the case of false positive. More in details, consider an audio with the following ground truth labels: 'Gunshot', 'Explosion', 'Rail Transport, and 'Vehicle', while the labels predicted by the model are: 'Gunshot', 'Rail Transport', and 'Chatter'. In this case, it is not possible to determine to which class assign the false positive 'Chatter'. Thus, the solution implemented equally redistributed the set of false positives (FP) among the set of false negatives using the weight $1/\#FP$. Additionally, the number of false positive and false negative produced in each class were also considered. From the results of this analysis, a new dataset referred as *AIDA_138C* was created, which filtered out 62 classes, most of which were children of the category 'Musical instrument' (see Section 3.1 for a link to the AudioSet ontology). The evaluation of a model trained on this dataset showed an absolute improvement of 3%, achieving an F1-score of 0.54%. Although further investigations should be performed, this result appears to indicate that an appropriate selection of sound classes may actually be of benefit to improve model's robustness.

This model was then fine-tuned on the dataset *AIDA_17C*. For a fairer comparison with the model *Agatha*, the evaluation was performed on the subset of events in common between the two models. Although the results were promising (see Table 3), a closer inspection showed that, due to the reduced number of events existing in *AIDA_17C*, the model was not able to generalize on unseen events. This is a common issue in transfer learning setups, when the pre-trained model is much larger than the size of the target dataset, the fine-tune is prone to "memorize" training corpus. To address this problem, a fine-tune with the dataset *AIDA_18C* was performed. Results, presented in Table 4, on one side confirmed the improvement of the model built on the dataset *AIDA_138C* with respect to the model *Agatha*, on the other side showed that the use of the general-purpose class now helped attaining better generalization performance. In particular, comparing Tables 3 and 4, one can observe that the model *Agatha* achieved a lower precision, while the model fine-tuned a lower recall. These outcomes can be explained considering that the general-purpose class included in *AIDA_18C* contains a wide set of classes that do not exist in the model *Agatha*, confounding it, but this class helped the fine-tuned model to discard false positive instances, lowering the recall.

5. Conclusions and future work

In this work, we addressed the development of a robust detector of sound events for the counter-terrorism domain to be used by Law Enforcement Agencies in their daily activity. Due to the lack of existing resources, we targeted the creation of a dedicated dataset using existing corpora, showing the inherent challenges of this task. Furthermore, we investigated the use of DNN and CNN architectures, with the latter exhibiting the best results on the same set of data. Notably, our experiments illustrated the importance of a robust fine-tune approach, which may conceal serious issues in the modeling phase. In this way, our classification framework was able to achieve an F1-score of 0.53% on the set of identified sound events. Future directions demand to continue the research on the creation of a balanced dataset tailored to the terrorism domain.

6. Acknowledgements

This work has received funding from the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 883596 related to the project AIDA. The content of the publication herein is the sole responsibility of the authors and it does not necessarily represent the views expressed by the European Commission or its services.

7. References

[1] J. Snyder, *The visual made verbal: A comprehensive training manual and guide to the history and applications of audio description*. Æ Academic Publishing, 2020.

[2] J. P. Udo and D. I. Fels, "The rogue poster-children of universal design: closed captioning and audio description," *Journal of Engineering Design*, vol. 21, no. 2-3, pp. 207–221, 2010.

[3] E. Messner, M. Fediuk, P. Swatek, S. Scheidl, F.-M. Smolle-Jüttner, H. Olschewski, and F. Pernkopf, "Multi-channel lung sound classification with convolutional recurrent neural networks," *Computers in Biology and Medicine*, vol. 122, p. 103831, 2020.

[4] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," *ACM Comput. Surv.*, vol. 48, no. 4, feb 2016. [Online]. Available: https://doi.org/10.1145/2871183

[5] M. Jung and S. Chi, "Human activity classification based on sound recognition and residual convolutional neural network," *Automation in Construction*, vol. 114, p. 103177, 2020.

[6] M. Lojka, M. Pleva, E. Kiktová, J. Juhár, and A. Čižmár, "Efficient acoustic detector of gunshots and glass breaking," *Multimedia Tools and Applications*, vol. 75, no. 17, pp. 10 441–10 469, 2016.

[7] J. Gauvain, L. Lamel, V. B. Le, J. Despres, J. Gauvain, A. Messaoudi, B. Vieru, and W. B. Kheder, "Challenges in audio processing of terrorist-related data," in *MultiMedia Modeling - 25th International Conference, MMM 2019, Thessaloniki, Greece, January 8-11, 2019, Proceedings, Part II*, ser. Lecture Notes in Computer Science, I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W. Cheng, and S. Vrochidis, Eds., vol. 11296. Springer, 2019, pp. 80–92. [Online]. Available: https://doi.org/10.1007/978-3-030-05716-9_7

[8] H. Oh, "Countermeasure of Uumanned Aerial Vehicle (UAV) against terrorist's attacks in South Korea for the public crowded places," *Journal of the Society of Disaster Information*, vol. 15, no. 1, pp. 49–66, 2019.

[9] A. Mesaros, T. Heittola, A. Eronen, and T. Virtanen, "Acoustic event detection in real life recordings," in *2010 18th European signal processing conference*. IEEE, 2010, pp. 1267–1271.

[10] J. Gauvain, L. Lamel, V. B. Le, J. Despres, J.-L. Gauvain, A. Messaoudi, B. Vieru, and W. B. Kheder, "Challenges in audio processing of terrorist-related data," in *International Conference on Multimedia Modeling*. Springer, 2019, pp. 80–92.

[11] I. Ozer, Z. Ozer, and O. Findik, "Noise robust sound event classification with convolutional neural network," *Neurocomputing*, vol. 272, pp. 505–512, 2018.

[12] A. Gorin, N. Makhazhanov, and N. Shmyrev, "DCASE 2016 sound event detection system based on convolutional neural network," *IEEE AASP Challenge: Detection and Classification of Acoustic Scenes and Events*, 2016.

[13] S. Adavanne, P. Pertilä, and T. Virtanen, "Sound event detection using spatial features and convolutional recurrent neural network," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 771–775.

[14] H. Phan, O. Y. Chén, P. Koch, L. Pham, I. McLoughlin, A. Mertins, and M. De Vos, "Unifying isolated and overlapping audio event detection with multi-label multi-task convolutional recurrent neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 51–55.

[15] J. F. Gemmeke, D. P. 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)*. IEEE, 2017, pp. 776–780.

[16] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "FSD50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2022.

[17] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A Large-Scale Audio-Visual Dataset," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 721–725.

[18] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proceedings of the 21st ACM international conference on Multimedia*, 2013, pp. 411–412.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. PMLR, 2015, pp. 448–456.

[20] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus, "Regularization of Neural Networks using DropConnect," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 3. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 1058–1066. [Online]. Available: https://proceedings.mlr.press/v28/wan13.html

[21] M. Yi-de, L. Qing, and Q. Zhi-Bai, "Automated image segmentation using improved PCNN model based on cross-entropy," in *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.* IEEE, 2004, pp. 743–746.

[22] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *CoRR*, vol. abs/1710.09412, 2017. [Online]. Available: http://arxiv.org/abs/1710.09412

[23] S. Park and M. Elhilali, "Time-Balanced Focal Loss for Audio Event Detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 311–315.