FACEBOOK ACOUSTIC EVENTS DATASET

Haoqi Fan, Jiatong Zhou, Christian Fuegen

Facebook
1 Hacker Way, Menlo Park
{haoqifan,jiatong,fuegen}@fb.com

ABSTRACT

The introduction of large scale datasets such as ImageNet, AudioSet, YouTube-8M and Kinetics has greatly advanced the state-of-the-art in machine perception. These datasets primarily focus on single modalities of audio or visual cues. We seek to broaden the scope in machine perception to multi-modal understanding with this work, which introduces the Facebook Acoustic Events dataset. This is a human labeled dataset which contains acoustic event labels of 500K segments from a random sample of public Facebook videos. Combined with its visual counterpart, labeled with scenes, objects and actions, we hope to make research in multi-modal learning and video understanding more accessible and convenient. We provide a well balanced dataset for acoustic event classification together with comprehensive benchmarks on both single and multimodal experiments on acoustic event detection using novel CNN based architectures.

Index Terms— acoustic event detection, multi-modal video understanding, machine perception, video database

1. INTRODUCTION

In the past, we've seen that the collation of larger and larger datasets directly enabling the success of large scale perceptual understanding. However, most datasets focus on a single modality of either audio or visual cues. We provide the Facebook Acoustic Events dataset, a large scale human labeled audio dataset for acoustic event classification. Combined with the Facebook Visual Events dataset, which provides labels for scenes, objects, and actions of the same video clips, we hope to accelerate research in the area of multimodal representation learning.

The Facebook Acoustic Events dataset consists of 500K publicly available video clips each about 10s long and labeled with 529 different audio classes. In order to provide the most value to the research community, we have chosen to use the AudioSet [1] ontology for labeling which provides a comprehensive catalog of acoustic events. This also enables research across different datasets.

While AudioSet can be considered the closest work to ours, both from the perspective of the data and scale, we believe that our dataset can provide additional value by

- pairing it with the Facebook Visual Events dataset that expands the label set to visual annotations of objects, scenes, and actions
- a more balanced distribution of acoustic events with a smoother long tail driven by a different distribution of Facebook public available videos.

A review of the audio literature reveals a rich history in the development of datasets and taxonomies. Nakatani et al. [2] developed an ontology for computational auditory scene analysis, Burger et al. [3] introduced a set of 42 distinct noisemes for manual annotation of noise segments, Salamon et al. [4] provided a dataset and taxonomy for urban sound research, Sager et al. [5] developed a large scale semantic ontology for audio content analysis, and Piczak [6] introduced a dataset for environmental sound classification. However, these datasets are typically orders of magnitude smaller than the 500k publicly available videos and approximate 2 million annotations of the Facebook Acoustic Events dataset.

In the context of video understanding, there have been many previous work such as [7], [8], [9], [10], [11], and [12] which explore using multiple modalities. However, current datasets such as YouTube8M [13], AVA [14], and Kinetics [15] tend to focus on a single modality only, i.e. audio, video, and human actions respectively. With the combination of the Facebook Acoustic Events Dataset and the Facebook Visual Events dataset, we hope to facilitate advancements in multi-modal video understanding, fusion and representation learning.

In addition to describing the creation of the dataset and analyzing its distribution, we also provide experimental results for acoustic event detection using audio and audio-visual features with comparisons across various CNN based architectures.

2. DATASET

The Facebook Acoustic Events dataset provides audio labels to 500k randomly selected video segments of roughly 10 second duration from publicly available Facebook videos. While this paper focuses on acoustic events, the visual counterpart of this dataset with labeled scenes, objects, and actions.

On average, each of the 10 second video segments have 3 to 4 acoustic event labels, capturing the information of both

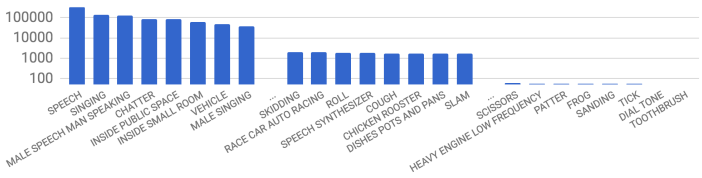

Fig. 1: Distribution of the number of annotations per category in the Facebook Acoustic Events dataset.

foreground and background sounds. The distribution of the frequency of labels in the dataset is shown in Fig. 1, where classes like SPEECH and VEHICLE can be seen to be of the most frequently occurring classes. Music categories were not the focus of this initial offering, but we are continuing to work on expanding labels for future releases. More details on the dataset distribution is shared in section 2.2.

The Facebook Acoustic Events dataset inherits the same ontology as AudioSet, with one additional label: SLOW MOTION. We decided to use the AudioSet ontology because of (1) its high coverage on different acoustic events occurring in video data, and (2) to facilitate research across different datasets and the ability for knowledge transfer between the two different distributions of data. The reason that we introduced the SLOW MOTION label is because of the prevalence of videos which were taken in slow motion mode, making the corresponding audio either very hard to distinguish or possess different acoustic characteristics than the original video.

2.1. Dataset Construction

In this section we describe the construction of the Facebook Acoustic Events dataset. This can be broken down into:

1. Generating the candidate set of videos to be included in the dataset
2. Training a deep model on AudioSet data to provide potential acoustic event labels for each video
3. A manual human review to remove incorrect labels
4. A second pass of human review to add back or revise concepts to improve precision and recall

The video candidates are randomly selected from public available Facebook video posts and shot boundary detection [16] is used to identify 10s segments of visual activity in each video. The same segments are used for audio and visual labeling.

To accelerate the effort of human review, we trained a deep convolutional neural network on the AudioSet data and used this model to bootstrap a set of candidate annotations per video clip. In order to remove the bias caused by the unbalanced AudioSet data, we re-calibrate the model confidence based on the recall of the ground truth labels as follows: For each acoustic event class, we generate a calibration table that maps model confidence to ground truth label recall. We pick a recall threshold r and find the corresponding model confidence c_i for each label at r. Now for each clip we use the model to produce a score s_i for each label and a ratio s_i/c_i. We select the label candidates for each video as the labels with the top k ratios. To define r and k we ran A/B tests with the human annotators, asking them to rate which configuration produced the best human rated labels. By doing this, we found $r = 0.3$ and $k = 10$ to work the best.

2.1.1. Quality Assurance

After pre-populating each video clip with acoustic event candidates, we do a manual review pass over the data by asking the annotators to remove any incorrect concepts. Since there are certain sounds which are hard to distinguish, we let the annotators watch the 10 second video clip as well. However, annotators are specifically told to only use the visual signal to confirm acoustic concepts and to be careful to not accidentally label an acoustic event because of the existence of matching visual cues, e.g. annotating a "dog barking" just because a dog is seen in the video, but without barking being present in the audio.

In order to calibrate and train annotators, we used a high quality 2-pass annotated subset of video-clips. In addition, different annotators overlapped on small subsets of the data. The overlapping annotation results are used to compute a performance metric for each annotator. To compute this metric, we calculate the average percentage of overlapping labels between the annotator and all other annotators labeling that video. During auditing by a specially trained review team, this performance metric is used to detect under performing reviewers. In addition, a random subset of clips for each annotator is reviewed as well. In case an underperforming annotator has been detected, we roll back all annotation results for re-labeling in order to ensure a high quality dataset.

After the first review pass to remove all incorrectly auto-generated labels, we perform a second pass that adds in missing labels to improve recall.

2.2. Dataset Properties

With the above mentioned data collection methodology, we have been able to annotate a total of 521, 547 video clips. A detailed histogram of the number of annotations per video can be seen in Fig. 2. On average, each video in the Facebook Acoustic Events dataset has been annotated with 3.1 labels. Given that video clips have a duration of approximate 10 seconds, we can see that the data is quite densely labeled.

Fig. 3 dives deeper into the data and compares the distribution of the top 30 most frequent classes of the Facebook Acoustic Events dataset with AudioSet. The histogram in blue belongs to the Facebook Acoustic Events dataset, while the histogram in red belongs to AudioSet. It can be observed that the class distribution in the Facebook Acoustic Events dataset is more balanced than that of AudioSet. This helps to

significantly mitigate the bias of a prior data distribution.

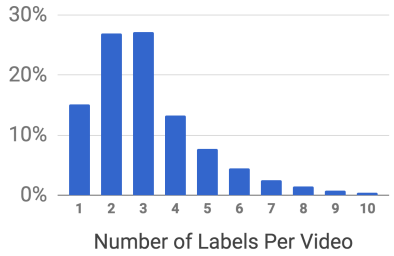

Fig. 2: Histogram of the number of annotations per video. A smooth distribution can be observed. Most of the videos have around 2-4 labels, with an average of 3.1.

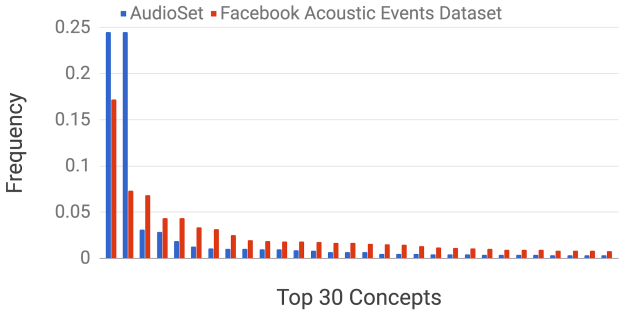

Fig. 3: Visualization of the label frequency distribution on the 30 most occurring classes. Red is AudioSet, blue is Facebook Acoustic Event dataset. The top 30 concepts are different in the different datasets.

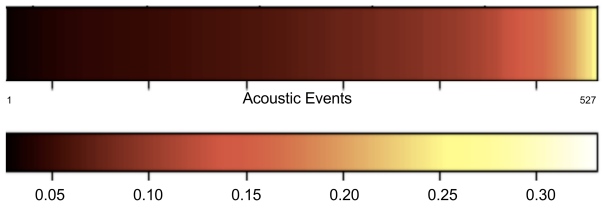

Fig. 4: Correlation between each acoustic concept and its most correlated visual concept in sorted order. The small bright region to the right represents acoustic signals which are highly correlated with one or more visual labels, whilst the larger, darker region on the left shows a significant amount of acoustic events are uncorrelated with any visual label.

To analyze the relationship between visual and acoustic labels in our dataset, we computed the class-wise correlations between the acoustic and visual concepts.

Given K videos labeled with M acoustic classes and N visual classes, we use matrix $A \in R^{M \times K}$ (where $A_{m,k}$ represents the m_{th} label on the k_{th} video) to represent the audio labels and $V \in R^{N \times K}$ to represent the visual labels. We calculate the correlation of the m_{th} acoustic concept and the n_{th} visual concept as:

$$corr = \frac{K(\sum A_m V_n) - (\sum A_m)(\sum V_n)}{[K \sum A_m^2 - (\sum A_m)^2][K \sum V_n^2 - (\sum V_n)]}$$

where A_m is the m_{th} column of A and V_n is the n_{th} column of V.

The correlations of each acoustic event with its most correlated visual signal are shown in Fig. 4. Each unit on the x axis represents a different acoustic event and the brighter the color, the higher the correlation to one or more visual labels. We see a limited number of acoustic signals for which there is high synchronization with visual signals. These are the classes that one would intuitively expect, such as *ocean* (visual concept) and *waves_surf* (acoustic concept), *computer_keyboard* and *typing*, *thunder* and *thunderstorm* etc. However, over 75.3% of the concepts actually have correlation less than 0.1 as seen in the large region of dark red on the left of the figure. This suggests that the audio signal is bringing in new information, that is not available from the visual modality. Because of this, we believe that this introduces some evidence that using multimodal signals will be beneficial over single-modality signals for video understanding.

2.3. Dataset Release

The Facebook Acoustic Events dataset will be released as a structured file containing links to the publicly available videos together with start and end times of the 10s segment labeled.

3. BENCHMARKS

Metrics		
Model	MAP	AUC
Single Visual Stream	10.83	54.62
Single Acoustic Stream [17]	30.91	**80.29**
Single Acoustic Stream VGG like	**31.22**	80.07
Two Stream with Simple Concat	31.73	78.08
Two Stream with Simple Add	32.05	**81.56**
Two Stream with Content Gating	**33.46**	81.51
Two Stream with Multi-Task Training	**34.27**	**81.58**

With the release of the Facebook Acoustic Events dataset, we provide benchmarks that help evaluating other models trained on the same dataset and demonstrate the value of multimodal training for acoustic event detection. In section 3.1, we describe a baseline architecture for the acoustic event detection task and compare the numbers to the modified ResNet in [17]. This provides baseline numbers for understanding the complexiy of the dataset. In section 3.2, we present and evaluate several simple late fusion strategies for multimodal joint training. We show that improvements in the accuracy of acoustic event detection can be obtained by using the raw visual features for training as well as by using visual labels as supervision and training towards labels of both modalities. The audio signal is preprocessed to 128 dimensional log-mel spectrogram while the visual signal follows the configuration of [18], using a single RGB image as input.

The 521547 video clips have been divided into a training and test set. 90% of the video clips are used for training, the remaining 10% for testing.

3.1. Acoustic Event Understanding

We have evaluated both a VGG like model and the modified ResNet model from [17] The VGG like model is composed of 12 convolutional layers and 5 pooling layers. Each of the 3×3 (stride 1×1) convolutional layers is followed by batch-normaliztion and a ReLU activation. After each two such convolutional layers, we perform pooling with a kernel of 2×2 and stride of 2×2. The ResNet model that was evaluated is the same as described in [17], with the stride of 2 removed from first 7×7 convolution. Both models have a global pooling layer at the end. The last convolutional layer is followed by a sigmoid activation and the model is trained with cross-entropy loss. From table 3, we can see the VGG like model performs 0.31% better on the Facebook Acoustic Events dataset.

3.2. Multi-Modal Joint Training

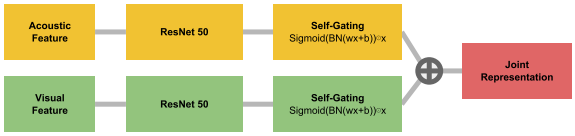

Fig. 5: Multi-Modal Joint Training Architecture.

To analyze the potential of the dataset for research on joint modeling, we evaluated the performance of raw acoustic and visual features alone for detecting acoustic events. To fairly compare the performance between using only acoustic or only visual features, we opted to use a ResNet-50 like architecture for both modalities. When using only the visual signal to predict acoustic labels with a single stream ResNet-50, we obtain a MAP of 10.83%. Compared to the baseline of 30.91% MAP of using acoustic signal as input to predict acoustic labels, we can see that visual information alone has its limitations in differentiating the acoustic information. This is in line with our observation in section 2.2 that there are a significant number of visual and acoustic event which have low correlation with any of the concepts in the other modality.

As a next step, we evaluated the use of both visual and acoustic information. We decided to use a two stream architecture [18] [19], where one stream is used for the acoustic information while the other stream is used for the visual information. We used the similar ResNet-50 architecture for both streams of the network, but for the acoustic stream, we enlarged the size of the final pooling kernel to make sure the entire audio can be fed into the network. For the visual stream, we randomly selected one frame from the clip as per Simonyan et al. in [18]. We then applied different strategies to fuse the two streams (late fusion). These strategies include concatenation, outer product, element-wise multiplication, and element-wise addition. We found that the MAP when using a simple concatenation is already slightly better than the single stream model. However, using element-wise addition outperforms the other approaches with a MAP of 32.05% clearly showing that visual information adds additional value for acoustic event detection. We omit the results of the other fusion strategies for brevity since they were not as performant.

We also explored more complex architectures and are proposing an architecture that uses a self-gated activation function [20] with batch normalization [21] as a learnable content gating [22] layer on each stream. This is then fused by element-wise addition. The self-gated activation function can be formulated as: given input x and learnable weights w and bias b, it computes the output $y = sigmoid(BN(wx + b)) \odot x$ [20] as a learnable content gate [22] on each stream, where $\odot$ is the Hadamard product. We observed that end-to-end training without batch-normalization in the self-gated activation function yielded poor performance. With this model, we achieved an additional 1.31% gain in MAP compared to the best performing element-wise fusion. We believe that the context gating is learning to suppress features from both modalities that are irrelevant to the task and emphasize the important features.

Finally, we adopt our best performing model and conduct multi-task training. Both acoustic and visual signals are used as supervision to predict acoustic and visual concepts. In order to have consistent evaluation metrics, we only report the performance on the acoustic events. We can see that this achieves the best performance with an MAP of 34.27%, proving that the additional signal from a different modality can help to learn a more discriminative representation [23] and gain a better understanding of the task.

4. CONCLUSION

With this paper, we introduce the Facebook Acoustic Events dataset, a collection of 500K labeled 10s publicly available video clips with a label ontology inherited from AudioSet. Combined with the Facebook Visual Events dataset, which provides labels for scenes, objects, and actions of the same video clips, we hope to accelerate research in the areas of machine perception such as acoustic event detection as well as multimodal video understanding. Moreover, we have provided benchmarks using CNNs for acoustic event detection using audio as well as audio-visual modalities. We showed that (1) models trained on audio-visual signals outperform audio or visual only models and (2) the best performance of 34.27% MAP and an AUC of 81.58% can be achieved using a two stream architecture with content gating trained in a multi-task fashion.

5. ACKNOWLEDGEMENT

We would like to thank Maksim Khadkevich for providing the initial acoustic event detection models that have been used to bootstrap the labels for this effort.

References

[1] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Acoustics, Speech and Signal Processing, IEEE International Conference on*. IEEE, 2017, pp. 776–780.

[2] Tomohiro Nakatani and Hiroshi G Okuno, "Sound ontology for computational auditory scence analysis," pp. 1004–1010, 1998.

[3] Susanne Burger, Qin Jin, Peter F Schulam, and Florian Metze, "Noisemes: Manual annotation of environmental noise in audio streams," 2012.

[4] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 1041–1044.

[5] Sebastian Sager, Damian Borth, Benjamin Elizalde, Christian Schulze, Bhiksha Raj, Ian Lane, and Andreas Dengel, "Audiosentibank: Large-scale semantic ontology of acoustic concepts for audio content analysis," *IEEE/ACM Transactions on audio, speech, and language processing*, 2016.

[6] Karol J Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015.

[7] Liqiang Nie, Xiang Wang, Jianglong Zhang, Xiangnan He, Hanwang Zhang, Richang Hong, and Qi Tian, "Enhancing micro-video understanding by harnessing external sounds," in *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 2017, pp. 1192–1200.

[8] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "See, hear, and read: Deep aligned representations," *arxiv*, p. 1706.00932, 2017.

[9] Yusuf Aytar, Carl Vondrick, and Antonio Torralba, "Soundnet: Learning sound representations from unlabeled video," in *Advances in Neural Information Processing Systems*, 2016, pp. 892–900.

[10] Jun Ye, Kai Li, and Kien A Hua, "Wta hash-based multimodal feature fusion for 3d human action recognition," in *Multimedia (ISM), 2015 IEEE International Symposium on*. IEEE, 2015, pp. 184–190.

[11] Kai Li, Jun Ye, and Kien A Hua, "What's making that sound?," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 147–156.

[12] Yun Wang, Leonardo Neves, and Florian Metze, "Audio-based multimedia event detection using deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2742–2746.

[13] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan, "Youtube-8m: A large-scale video classification benchmark," *arXiv preprint arXiv:1609.08675*, 2016.

[14] Chunhui Gu, Chen Sun, Sudheendra Vijayanarasimhan, Caroline Pantofaru, David A Ross, George Toderici, Yeqing Li, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, et al., "Ava: A video dataset of spatio-temporally localized atomic visual actions," *arXiv preprint arXiv:1705.08421*, 2017.

[15] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al., "The kinetics human action video dataset," *International Conference on Computer Vision and Pattern Recognition*, 2017.

[16] Alan Hanjalic, "Shot-boundary detection: Unraveled and resolved?," *IEEE transactions on circuits and systems for video technology*, 2002.

[17] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., "Cnn architectures for large-scale audio classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 131–135.

[18] Karen Simonyan and Andrew Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[19] Minghuang Ma, Haoqi Fan, and Kris M Kitani, "Going deeper into first-person activity recognition," *arXiv preprint arXiv:1605.03688*, 2016.

[20] Prajit Ramachandran, Barret Zoph, and Quoc V Le, "Swish: a self-gated activation function," *arXiv preprint arXiv:1710.05941*, 2017.

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

[22] Antoine Miech, Ivan Laptev, and Josef Sivic, "Learnable pooling with context gating for video classification," *arXiv preprint arXiv:1706.06905*, 2017.

[23] Donghyun Yoo, Haoqi Fan, Vishnu Naresh Boddeti, and Kris M Kitani, "Efficient k-shot learning with regularized deep networks," *arXiv preprint arXiv:1710.02277*, 2017.