Monitoring wildlife using AudioVisual Channels on Tiny Compute Devices

Emmanuel Azuh Mensah
UW CSE 599

Abstract

The advent of camera traps has brought high hopes regarding a more fine grained understanding of animal species and their behaviors through remote monitoring. In addition, the recent growth in the capabilities of machine learning algorithms promises to automate wildlife monitoring. This is important for working towards grand visions like the United Nations' Sustainable Development Goals on ecological protection and restoration. According to the state of the field survey conducted by WILDLABS [1], processing multiple data streams to make ecological decisions will be highly desirable in the high detail sensing movement. In this project, we experiment with audio-visual machine perception using "low" resource machine learning models as a way to improve species identification in real time use cases. The code used in this project can be found at [1]

1 Introduction

Camera traps have become a widely used way to collect data and learn from ecological species to improve our understanding of the environment. However, there's an increasing need to process multiple streams of data collected from the wild, eg audio and visual modalities [1] and with low resource devices. This is especially useful in regions where low resource devices like mobile phones are already abundant and people needing to find uses for old phones. Not only can adding audio modality assist with detection in low light conditions (such as at night), having the image along side can improve detection of certain species when it is hard to recognize them by sound alone (eg during rough weather events). Again, having a network of low resource devices such as proposed in [7] helps to provide context for detection algorithms as well as increase the chances of accurately estimating ecological measures of interest such as species density [2]. These are factors that would help disambiguate scenarios where only partial information would make estimation less precise. In addition, the setup of multimodal information sources from a network of cheap sensors improves answers to harder questions that require reasoning across space and time for sparsely sampled sensor data.

In this project, we experiment with multimodal (audio-visual) data as a way to improve animal species classification. This is a first step in a longer project to gain more insight into the high resolution ecological monitoring space. The project focuses on low resource use cases since we aim to not only situate ourselves to do near-sensor processing for real-time applications but also to make some of the otherwise computationally intensive machine learning models, more accessible to ecologists. For the rest of the paper, we provide a discussion on related works to this project, propose an initial system to guide our future work and assess initial performance of the system. Finally, we present a brief discussion.

2 Related Work

Camera traps have become a popular tool for monitoring wildlife, especially when combined with remote servers for post analysis. [13] provides a general discussion on different types of camera traps and how their percentage of positive triggers varies by distance and [14] presents factors to consider when selecting camera traps for a project. There is however, a number of projects looking at alternatives or supplements to camera traps, using low resources devices. For instance, [18] monitors birds using a camera trap with iOS mobile phone. [9] also develops a programmable camera trap that optimizes for cases where the speed of sending camera trap images to the monitoring center needs to be high while optimizing for energy. [19] uses a standard camera with a motion detect script to monitor flower-visiting animals albeit at a close up position to the flower. [6] creates a RaspberryPi-Zero based camera trap system augmented with a near IR light to continuously film animal activity and can record up to 72-hr day and night videos at >720p resolution with a 110-Wh power bank (30,000 mAh). Finally, [3] uses a mesh grid of Rpi sensor nodes to

[1] https://github.com/emma-mens/elk-recognition/tree/main/src/multimodal_species

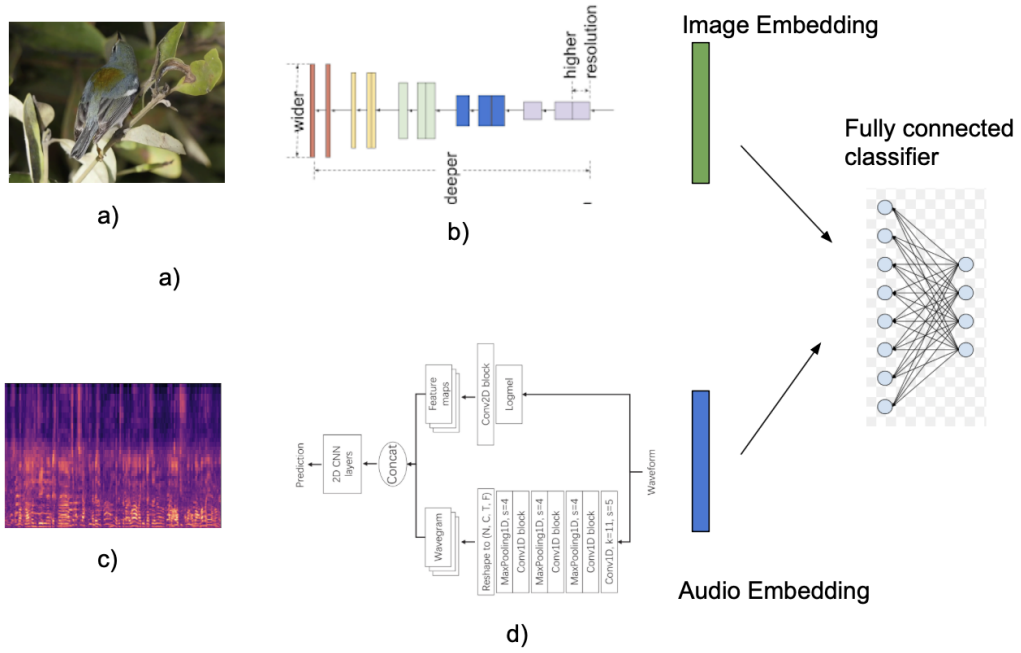

Figure 1: Our proposed system is a standard setup for a late fusion audio visual classification model. a) An input image which may be difficult to detect based on various factors such as occlusion, camouflaging, etc. b) The efficient net model as defined in [20]. c) An audio input to the d) PANNs network as defined in [10]

collect wildlife monitoring data.

The literature is slowly introducing the use of audiovisual modalities in wildlife monitoring. [4] conducts a literature review on the strengths of using either visual or audio channels for monitoring species and assesses the benefits of using the two channels jointly. [16] sets up a audio-visual monitoring system using a three tiered approach (edge device, geteway and cloud) and some low power detection algorithms like Dynamic Time Warping. [23] assesses the performance of audio-visual models that assume single visual source for sound and how they fare when tested on more challenging datasets. The SSW60 dataset [22] also motivates the use of audio-visual models by discussing cases when two species are difficult to separate visually but have distinct calls. Again, they discuss the flip case where species are difficult to differentiate by sound but are easy to tell apart visually, making joint audio-visual models a better solution for stronger discriminatory ability of models.

The field of multimodal machine learning is at this point, quite established with models like CLIP [12] demonstrating very impressive results with multimodal tasks such as co-learning from language and visual data. Typical multimodal models have separate models for each modality and combine the representations learned by the separate models either early in the network, middle or so called late fusion. Such models are also usually trained using a contrastive loss (an objective that forces representations of similar concepts to be close in a fixed vector space and disparate concepts farther apart) for unsupervised learning. In this project, we use a similar approach to train multiple mobile models but focus on a classification task.

3 Proposed System

The project went through a couple of phases during implementation but finally settled on using two recent and popular on-device machine learning models. The visual model is based on EfficientNet and the audio based on MobileNet, described further below. The main detour that didn't make it into the final paper was the use of an EfficientNet based model for audio classification named Contrastive Learning of General-Purpose Audio Representations (COLA) [15]. Even though this model setup was promising, we had difficulty replicating the paper's results in Pytorch (the machine learning framework used for this project). We eventually found a MobileNet version which performed more stably during training and is what we moved forward with.

3.1 Image Model - EfficientNet

One of the insightful papers in recent years to come out of Google presented a motivation for a more uniform scaling of model properties (number of layers, width of layers, size of convolution filters, etc) that previously were often tweaked independently. From the paper, it was shown that scaling some dimensions and leaving others tends to see under-utilization of the potential from the increased dimensions.

For more discussion on this, see [20]. For this project, we use the base model EfficientNet-B0 (which is the smallest size model proposed in the paper) and is most applicable in the mobile device setting.

3.2 Audio Model - MobileNet

As mentioned above, the inital plan to use COLA fell through due to difficulty replicating their results for an EfficientNet based model during the quarter project. We therefore pivoted to another project called PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition [10], which had an on-device version of their network based on MobileNetV2 [17] that already had a pretrained pytorch model. MobileNets were the previous most widely used mobile deep learning models out of Google. The models focused on using techniques such as depthwise separable convolutions which refactored the convolution operation into a form that greatly reduced the number of multiply-add operations needed to get the same results. For a more detailed discussion of the design, see [17]. The PANNs model used in this paper combines approaches to learn from audio waveform for frequency information while learning from spectrogram representations to learn more time domain information about audio signals.

3.3 AudioVisual Late Fusion Model

Here we use the model from the diagram in Figure 1 to test if there are any benefits to using the two modalities in this low resource setting. We take the penultimate layer (before classification layer) of each of the image and audio models. Each of these embedding layers is a vector of size 1280. These vectors are combined (using concatenation to a vector twice as long, taking a weighted average between the two vectors, or the element wise maximum to allow for experimenting with different ways to learn a joint representation). The combined embedding layers are then fed through a feed forward neural network with on hidden layer of size 1024 and finally an output layer of 60 for the number of bird species. In practice, there are far more sophisticated ways to learn a joint representation but we use this approach to have a baseline for further experimentation.

3.4 Pretraining

A common technique within the machine learning community is to train a model on a large dataset and then tune it for use cases similar to the original dataset but typically with less data available. The process of training the model with the initial large dataset is called pretraining. In this project, we use the two models described above pretrained on two famous datasets. For the visual modality, we use EfficientNet pretrained on ImageNet [5], arguably the most famous image dataset in the machine learning community, with 3.2 million images. The audio modality is pretrained on AudioSet [8] which contains 1.7 million 10 second segment audio clips with 632 classes collected from Youtube videos and spans natural sounds, animal sounds, music, sounds of things etc. Both datasets were built using human labelers to ensure correct data to class association, leading to very high performing pretrained models.

4 DataSet

For this project, we focused on the audio-visual birds dataset [22] from the Cornell Lab of Ornithology, consisting of 60 bird species. The data was carefully annotated by experts and has paired bird species images with corresponding audio clips. They present both a paired dataset (with few examples as shown in Figure 2) and an unpaired set of images and audio dataset that we didn't use in this term project but will use in future iterations.

	total	min	max	median
Audio	2597, 1264	28, 12	52, 30	45, 21
Video	3462, 1938	38, 22	68, 52	59, 31

Figure 2: SSW60 Dataset statistics with the (train, test) split presented in each column. The min, max and median columns represent descriptive statistics of data points available for each species in the dataset

We also set up the pipeline for iNaturalist dataset [21] which has 10,000 species and so is more representative of the final general purpose species classifier we are interested in. We however don't present results on the iNaturalist dataset here but will be used for next steps of the project.

5 Experiments

In this section, we explore the contributions from the two modalities to performance on fine grained species classification tasks. In future works, we will explore the benefits of finetuning the pretrained models described in Section 3 with an object detection task instead to learn richer representations.

Data preparation for audio involved stripping the audio portion from the bird video dataset and loading them at a sampling rate of 16000Hz, single channel. The PANNs model then accepts a waveform and a spectrogram representation of the audio input. Since spectrograms which are more common in the speech recognition community is made to represent audio in the audible spectrum for humans, the double representation here allows for more detailed information extraction.

For the images, we normalize using the mean and standard deviation of imagenet dataset, available in most machine learning frameworks. We then augment the training images with random rotations, horizontal flips and random blurring to reduce too much memorization of the training data which typically would lead to less generalizable models.

We train the models with a learning rate of 0.001 using Adam optimizer with default parameters. All experiments used a batch size of 32 as this was the largest we could fit on GPUs used for training. Experiments are trained for 120 epochs and with a learning rate scheduler that is reduce by a factor of 10 after 50 epochs.

6 Evaluation

6.1 Baseline Large Model Performance

To evaluate our model, we use the birds dataset and compare our result to the baseline model used [22]. In the reference paper, they used a state of the art transformer model named Multimodal Bottleneck Transformer (MBT) [11] with 86 million paramters, far more than can fit on mobile devices. Compared to MBT, our training set up uses about 7 million parameters, which is more manageable on small devices but will still need to be reduced in follow up work. MBT is pretrained on ImageNet and AudioSet and then fine-tuned in the birds paper on the SSW60 dataset. We use a similar flow of datasets in our experiments (EfficientNet pretrained on ImageNet and PANNs pretrained on AudioSet), as well as a final finetuning on the birds dataset.

6.2 Our AudioVisual Model Performance on SSW60 Test Set

We present initial results of our experiments in Table 1. The first row of the table presents the baseline performance using the state of the art audiovisual model.

6.2.1 Effects of Various Embedding Combination Methods

The second block of rows use pretrained audio and visual models and experiment the effects of using different types of embedding combination methods. We see from initial experiments that concatenating the two embedding vectors produces the best results in our model setting (likely due to freer

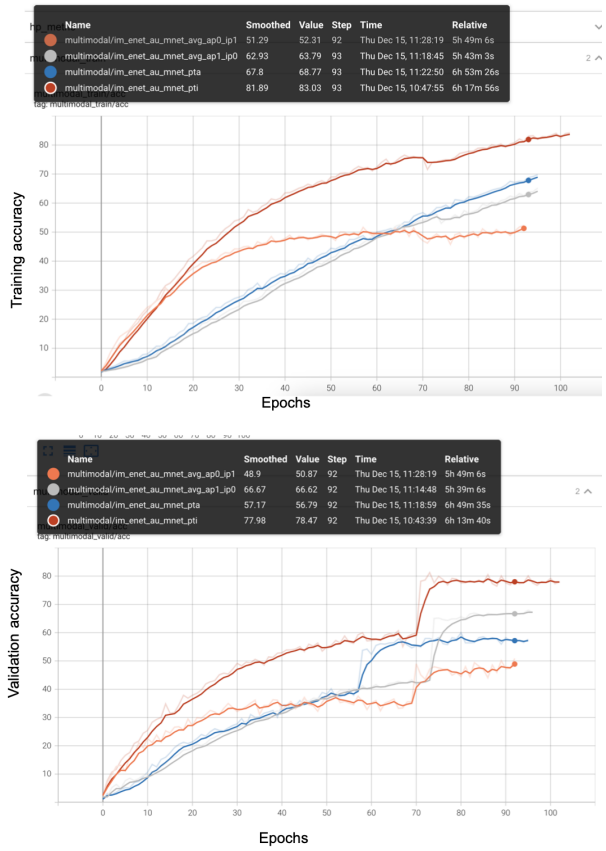

Figure 3: Training accuracy (top) and corresponding validation accuracy (bottom) of our multimodal setup. The (red) curve is for the audio visual model that averages embeddings from the two models, weighting audio with zero and image with one, while the reverse is true for the (grey) curve. The (blue) curve trains the multimodal model with only the audio model pretrained and the image model started from scratch, while the reverse is true for the (red) curve.

combination of representation vectors), although maximum produces comparable results. It is also clear that the visual modality contributes more to species detection performance in the weighted average case since higher weighting of the image embedding model produces better results.

6.2.2 Pretraining Effects

With the training setting used for our experiments, we had the lowest performance with the audio visual case with no pretraining in either model (middle row). However, the low performance presented here is lower than is possible if we trained the model much longer. Given the size of the birds dataset though, it would certainly not be sufficient to train the model even to halfway of the performance of the pretrained models.

The last row shows that starting with an image pretrained

image model	audio model	image pretrained	audio pretrained	embedding combination	accuracy
Birds baseline	Birds baseline	Yes	Yes	MBT Late Fusion	80.6
Yes	Yes	Yes	Yes	0.8 audio + 0.2 image	32.61
Yes	Yes	Yes	Yes	0.5 audio + 0.5 image	41.18
Yes	Yes	Yes	Yes	0.2 audio + 0.8 image	41.23
Yes	Yes	Yes	Yes	max	41.33
Yes	Yes	Yes	Yes	concatenate	**41.58**
Yes	Yes	No	No	concatenate	2.68
No	Yes	-	Yes	concatenate	17.18
Yes	No	Yes	-	concatenate	38.03
Yes	Yes	No	Yes	concatenate	19.71
Yes	Yes	Yes	No	concatenate	32.81

Table 1: *Effects of various model selection choices on final top 1 test accuracy for the birds species dataset. Accuracy is reported for models trained for 120 epochs (longer training might improve some of these values, especially the no pretraining context). The first row represents the baseline model used in the Sapsucker Woods Dataset baseline results. The baseline model is a state of the art Multimodal Bottleneck Transformer (MBT) model.*

model only is more important than starting with an audio only pretrained network in the multimodal setting.

6.2.3 Single Modality

The last but one row presents training results for using individual modalities. It is clear that the visual modality performs better than the audio modality alone. However, it performs slightly worse than joint audio-visual model.

6.2.4 A Brief Discussion of COLA

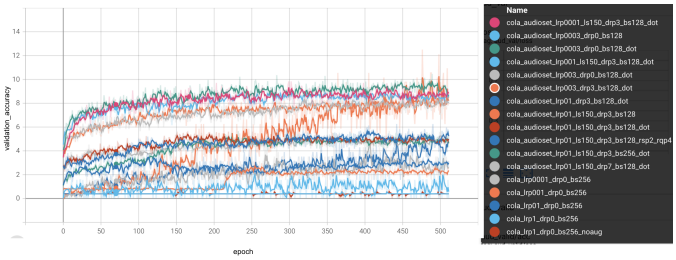

Figure 4: A sample of training experiments for COLA. Most of the experiments did not go higher than 10 percent detection rate.

Now that we have a model that trains properly towards an acceptable accuracy level for the low resource audio-visual setting, we can experiment more confidently, using the initial results as a guiding tool for sanity checking. From inspecting the initial set of experiments, one theory to validate going forward would be whether the audio data processing pipeline is working correctly. Given the experiment plot presented in Figure 4, we could tell that training COLA wasn't effective, possibly because of inaccurate hyperparameter configuration. However, we also tried to train the PANN model for audio only and had almost similar difficulty. We reported audio only result in the project as the audio-visual model with zero weighting from the visual component (which possibly works still because of more contribution from the image dataset). However, the audio-visual training failed when using COLA with the image EfficientNet model. We therefore pin the audio data as the commonality between the failed training experiments. We will investigate further if our audio extraction process from the birds video dataset is working properly.

7 Conclusion

From this project, we got the chance to setup a prototype lower resource models for audio visual learning than is currently explored in the literature especially for on-device use cases. Through the set up, we produce initial results towards benefits of using multiple information sources (modalities) for bird species classification.

In future works, we aim to experiment with opportunities to drive the size of this initial audiovisual model down further using approaches in the tinyML community (especially through finding the most important portions of the network that contribute to high performance and mainly focusing on them). We will also explore the model performance on iNaturalist dataset in order to have a more general purpose low resource species classification model to use in the wild. Finally, we will deploy the models in real world animal monitoring use cases to close the loop on making practical tiny machine learning models.

References

[1] https://wildlabs.net/state-of-conservation-technology.

[2] https://www.kaggle.com/c/iwildcam2021-fgvc8.

[3] Andrew Arnold, Paul Corapi, Michael Nasta, Kevin Wolgast, and Thomas A Babbitt. A raspberry pi sensor network for wildlife conservation. In *Proceedings of the 7th Symposium on Hot Topics in the Science of Security*, pages 1–3, 2020.

[4] Rachel T Buxton, Patrick E Lendrum, Kevin R Crooks, and George Wittemyer. Pairing camera traps and acoustic recorders to monitor the ecological impact of human disturbance. *Global Ecology and Conservation*, 16:e00493, 2018.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[6] Vincent Droissart, Laura Azandi, Eric Rostand Onguene, Marie Savignac, Thomas B Smith, and Vincent Deblauwe. Pict: A low-cost, modular, open-source camera trap system to study plant–insect interactions. *Methods in Ecology and Evolution*, 2021.

[7] Vladimir Dyo, Stephen A Ellwood, David W Macdonald, Andrew Markham, Niki Trigoni, Ricklef Wohlers, Cecilia Mascolo, Bence Pásztor, Salvatore Scellato, and Kharsim Yousef. Wildsensing: Design and deployment of a sustainable sensor network for wildlife monitoring. *ACM Transactions on Sensor Networks (TOSN)*, 8(4):1–33, 2012.

[8] Jort F. Gemmeke, Daniel P. W. 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[9] Mehmet Karaköse, CANAN Taştimur, Selim Özdemir, Merve Erol, Ahmet Tokgöz, and Erhan Akın. Development of programmable camera-trap. 2020.

[10] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020.

[11] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021.

[12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[13] Christoph Randler and Nadine Kalb. Distance and size matters: A comparison of six wildlife camera traps and their usefulness for wild birds. *Ecology and Evolution*, 8(14):7151–7163, 2018.

[14] Francesco Rovero, Fridolin Zimmermann, Duccio Berzi, and Paul Meek. " which camera trap type and how many do i need?" a review of camera features and study designs for a range of wildlife research applications. *Hystrix*, 24(2), 2013.

[15] Aaqib Saeed, David Grangier, and Neil Zeghidour. Contrastive learning of general-purpose audio representations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879. IEEE, 2021.

[16] Aya Sakhri, Oussama Hadji, Chakir Bouarrouguen, Moufida Maimour, Nasreddine Kouadria, Abderrezak Benyahia, Eric Rondeau, Noureddine Doghmane, and Saliha Harize. Audio-visual low power system for endangered waterbirds monitoring. *IFAC-PapersOnLine*, 55(5):25–30, 2022.

[17] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[18] Ronny Steen. Bird monitoring using the smartphone (ios) application videography for motion detection. *Bird Study*, 2017.

[19] Ronny Steen. Diel activity, frequency and visit duration of pollinators in focal plants: in situ automatic camera monitoring and data processing. *Methods in Ecology and Evolution*, 8(2):203–213, 2017.

[20] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[21] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.

[22] Grant Van Horn, Rui Qian, Kimberly Wilber, Hartwig Adam, Oisin Mac Aodha, and Serge Belongie. Exploring fine-grained audiovisual categorization with the ssw60 dataset. *arXiv preprint arXiv:2207.10664*, 2022.

[23] Ho-Hsiang Wu, Magdalena Fuentes, Prem Seetharaman, and Juan Pablo Bello. How to listen? rethinking visual sound localization. *arXiv preprint arXiv:2204.05156*, 2022.