Multi-attentive detection of the spider monkey whinny in the (actual) wild

Georgios Rizos, Jenna Lawson, Zhuoda Han, Duncan Butler, James Rosindell, Krystian Mikolajczyk, Cristina Banks-Leite, Björn W. Schuller

Angaben zur Veröffentlichung / Publication details:

Rizos, Georgios, Jenna Lawson, Zhuoda Han, Duncan Butler, James Rosindell, Krystian Mikolajczyk, Cristina Banks-Leite, and Björn W. Schuller. 2021. "Multi-attentive detection of the spider monkey whinny in the (actual) wild." In *Interspeech 2021, Brno, Czechia, 30 August - 3 September 2021*, edited by Hynek Heřmanský, Honza Černocký, Lukáš Burget, Lori Lamel, Odette Scharenborg, and Petr Motlicek, 471–75. Baixas: International Speech Communication Association (ISCA). https://doi.org/10.21437/interspeech.2021-1969.

INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Multi-Attentive Detection of the Spider Monkey Whinny in the (Actual) Wild

Georgios Rizos[1*], *Jenna Lawson*[2*], *Zhuoda Han*[3], *Duncan Butler*[2], *James Rosindell*[2], *Krystian Mikolajczyk*[3], *Cristina Banks-Leite*[2], *Björn W. Schuller*[1,4]

[1]GLAM – Group on Language, Audio, & Music, Imperial College London, UK
[2]Department of Life Sciences, Imperial College London, UK
[3]MatchLAB – Department of Electrical and Electronic Engineering, Imperial College London, UK
[4]EIHW – Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

georgios.rizos12@imperial.ac.uk, j.lawson17@imperial.ac.uk

Abstract

We study deep bioacoustic event detection through multi-head attention based pooling, exemplified by wildlife monitoring. In the multiple instance learning framework, a core deep neural network learns a projection of the input acoustic signal into a sequence of embeddings, each representing a segment of the input. Sequence pooling is then required to aggregate the information present in the sequence such that we have a single clip-wise representation. We propose an improvement based on Squeeze-and-Excitation mechanisms upon a recently proposed audio tagging ResNet, and show that it performs significantly better than the baseline, as well as a collection of other recent audio models. We then further enhance our model, by performing an extensive comparative study of recent sequence pooling mechanisms, and achieve our best result using multi-head self-attention followed by concatenation of the head-specific pooled embeddings – better than prediction pooling methods, as well as compared to other recent sequence pooling tricks. We perform these experiments on a novel dataset of spider monkey whinny calls we introduce here, recorded in a rainforest in the South-Pacific coast of Costa Rica, with a promising outlook pertaining to minimally invasive wildlife monitoring.

Index Terms: acoustic event detection, deep attention models, multiple instance learning, wildlife monitoring, bioacoustics

1. Introduction

Automated methods for recording and analysing bioacoustic data hold the promise for unprecedented scalability in wildlife monitoring, with the purpose of preservation through a global biodiversity crisis [1]. This has enabled biologists and engineers to perform machine learning studies on bioacoustics across a large taxonomic range, such as primates [2, 3] or other terrestrial [4, 5] or marine mammals [6, 7, 8, 9, 10], birds [11, 12, 13, 14, 15], as well as amphibians [14], in applications like call detection for verifying presence or estimating density [6, 2, 4], discerning between calls of different species [14, 15], as well as different call types of a particular animal [5, 8].

In this study, we focus on the binary call detection of the 'whinny' call of Geoffroy's spider monkey (*Ateles geoffroyi*) towards verifying their presence. This *highly sensitive* forest specialist is endangered across its range, due to forest loss for conversion into urban areas and agricultural lands, as well as also being a target for hunters for meat consumption and the pet trade, thus hampering its role as seed disperser for many large tree species which are critical to retaining healthy forests [16]. In an attempt to reduce the demanding fieldwork required to detect this rare and elusive species, we used Passive acoustic monitoring (PAM) [1] by means of *in situ* sound recorders in order to collect an acoustic dataset, while keeping perturbations to their behaviour, by our presence, to a minimal degree. We approach the spider monkey call detection task via weakly supervised, Acoustic Event Detection (AED), specifically the Multiple Instance Learning (MIL) framework [17, 18, 19, 20, 21, 22]: since our call annotations refer to the entire clip, we have to use mechanisms for pooling instance-level information to a bag-level embedding (or prediction) that characterises the entire clip.

We perform an extensive comparative study among recent acoustic core models, by controlling the pooling mechanism, and validate that a ResNet inspired from the best performing model proposed in the recent study performed by [23], is also the best performer in our dataset and task. We then show that an improvement upon this model, **SE-ResNet28**, based on the addition of Squeeze-And-Excitation (SE) mechanisms [24] leads to increase across all performance measures, with statistical significance ($p < 0.01$ in three out of four measures) using Welch's unequal variances t-test. We then perform a second comparative study, among pooling mechanisms, this time controlling the core model to be our proposed SE-ResNet28. We include in the comparison both prediction based [25, 22, 21], as well as embedding based [26, 27, 17] pooling methods. Our surprising results show that self-attention based embedding pooling is generally better than prediction pooling for our dataset, thus contradicting previous insights from similar comparisons [20, 21]. Specifically, we find multiple head attention and subsequent concatenation of the pooled embeddings from each head to be the best approach, and perform a significance analysis. Regarding machine audition research on the endangered, Geoffroy's spider monkey, we collected and introduce here the first acoustic recording dataset with call annotations, recorded using PAM methods.

2. Related Work

In a recent comparative study of convolutional neural network (CNN) layer based models applied on LogMel-Spectrograms [23], a 38-layer ResNet was found to be the best performer – more layers did not lead to improvements. We validate the great performance of ResNet28, as well as improve it using SE [24]. The latter have been used before in acoustic scene classification [28], albeit on a much smaller, 3-CNN layer VGG-style [28] architecture, comparable to the lower performers of our study.

Despite recent advances in pooling mechanisms in the MIL framework, several recent studies focusing on AED or audio tagging adopt simpler embedding pooling methods, like aver-

[*] equal contribution.

http://dx.doi.org/10.21437/Interspeech.2021-1969

Table 1: *A summary of the first A. Geoffroyi whinny call dataset.*

Partition	#Sites	#Positives	#Negatives
train	7	277	4,288
validation	3	125	2,488
test	3	189	1,886

age or max operations [29], a combination of the two [23], or concatenation of the sequence along the feature axis [2]. Two recent studies [20, 21] have performed comparisons between two groups of pooling methods: a) embedding based; which learn a weighted average of the embedding sequence produced by the core model, and b) prediction based; which apply a dense layer with an activation function on each of the embeddings to produce a corresponding sequence of prediction probabilities, and learn a weighted average of the latter. In the multi-class classification study performed in [20], the authors found that learning-free prediction pooling methods, like averaging, exponential softmax, and linear softmax (LinSoft – first proposed in [25]) outperformed attention based embedding pooling. LinSoft was later adopted by [18], and used as a competitive baseline in [21, 22]. In the binary classification study performed in [21], the max operator prediction pooling method was instead shown to outperform LinSoft; a seemingly contradictory result with respect to the previous study, which the authors attribute to the ability of the max operator to highlight rare positive event instances. More recently [22], Power Pooling was proposed, which raises the prediction probabilities to a power equal to a learnt parameter, outperforming LinSoft, and the max operator; the latter by far. With the exception of the most recent Power Pooling, no prediction pooling method has undisputedly been shown to be better, and we thus opt to include them in our comparison.

Regarding embedding pooling methods, even though single-head attention has not been shown to be better than prediction pooling [20, 21], there has been a growing number of recent studies using multi-head attention on audio classification tasks [26, 30, 17, 27]. Multiple attention heads hold the potential of learning to attend to different patterns, however, they introduce additional parameters, as well as the need to aggregate their outputs. In the speaker identification study performed in [26], the authors simply average the pooled embeddings from each head, whereas in the followup paper [30], they propose a second attention mechanism that describes a weighted average. Alternatively, the authors of [27] propose to use a different, fixed temperature parameter per head to encourage them to attend to different event durations, and then concatenate the pooled embeddings. Finally, the authors of [17] also use concatenation of the weighted averages, as well as of the weighted standard deviation of the embeddings, as proposed in [31], before applying a gating mechanism on this aggregated vector.

3. The Spider Monkey Whinny in the Wild

Our study area covered approximately 2,000 km^2 in the South-Pacific coast of Costa Rica. Data were collected at 341 sites totalling 60,000 hours of recordings at a 48 kHz sampling rate; however, in the context of this study, we annotated data from 13 sites. Of the nine recognised types of calls of *A. Geoffroyi*, the 'whinny' is the most common, representing general communication related to feeding and movement [32]. Indeed, we found that over 80 % of recorded calls were whinnies; therefore, we opted to focus on it; other species exhibit acoustic differences with respect to the type of call [5, 8]. We manually listened to 600 hours of acoustic data and isolated 591 examples of the target sound in a total of 366 sound files, which are included in this study. We included calls from both quiet and noisy backgrounds, to best represent the natural environment, annotated using the Praat software[1]. The calls are around 1 second long, and we opted to segment our recordings into 3 second clips for this study. For the positives, we calculate the earliest and latest possible timestamps that can serve as the clip starting point such that the entire call can fit. We uniformly sampled a starting point in this interval to get a clip. The segments that strictly do not contain a call, are then deterministically segmented into 3 second clips, to produce our set of negatives. We partition in a site-independent manner, to ensure that our model does not learn the characteristics of a site. The dataset characteristics are summarised in Table 1.

4. MIL Model Design

Let $x_i, y_i \sim \mathcal{D}$ be the i-th sample/label pair from the dataset. y_i is either 0 or 1, depending on whether the sample is negative or positive, and x_i can be a LogMel-Spectrogram, or the raw waveform of the audio clip, depending on the required core model $\mathcal{M}_{core}$ input. The core model processes the input, and yields a sequence of learnt, latent embeddings of length T: $H_i = \{h_{i,t}\}$. Each h_t corresponds to a segment of the input x_i, of time dependent to the degree of subsampling performed by $\mathcal{M}_{core}$, and H_i has size equal to $T \times C$, where C corresponds to the number of features/filters of each $h_{i,t}$.

4.1. Sequence pooling

Let $\mathcal{M}_{pred}$ be a prediction module that expects an input instance of size $1 \times C$, and outputs a prediction, we have two potential solutions for getting our final prediction estimate $\hat{y}_i$.

Prediction pooling: We pass all instances $h_{i,t}$ through $\mathcal{M}_{pred}$, to get T corresponding $\{\hat{y}_{i,t}\}$ probability predictions. We then need to pool this sequence, in order to get our final prediction for the entire bag of T instances; i. e., a clip-level prediction. For max pooling, we just select $\max(\hat{y}_{i,t})$. Otherwise, we use the following general form equation:

$$\hat{y}_i = \frac{\sum_t^T \hat{y}_{i,t} \cdot f(\hat{y}_{i,t})}{\sum_t^T f(\hat{y}_{i,t})}, \tag{1}$$

where $f(\cdot)$ is a learnt linear transformation, the exponential function, and the identity function for the cases of attention pooling, regular softmax, and LinSoft [25], respectively. For Power Pooling [22], the function is the power to the β-th, where β is a learnt parameter. A similar adaptive method (AutoPool) was proposed in [19], where the learnt parameter was the temperature of the regular, exponential softmax. Since regular softmax, as well as AutoPool have been outperformed [20, 21, 22] by LinSoft and Power Pooling, we will include in our comparisons here another variant: *LinSoft-Auto*, in which $f(\hat{y}_{i,t})$ is equal to $\beta \cdot \hat{y}_{i,t}$. This choice also allows for adaptation to variable call durations, but is less prone to reaching extremely high values and tending to the max operator, without the need for regularisation.

Embedding pooling: We aggregate the instance-level embeddings $h_{i,t}$ into a bag-level one h_i, which is then processed by $\mathcal{M}_{pred}$ to get the final prediction estimate $\hat{y}_i$. Since $h_{i,t}$ is an embedding in a subspace of $\mathbb{R}^C$, we can learn a linear $C \times 1$ energy transformation f that serves as an unnormalised proxy of the importance of each instance. In fact, recent work [26, 30, 17, 27] has shown the value of a plural learning of such

[1] `https://www.fon.hum.uva.nl/praat/`

Table 2: *A sketch of our proposed SE-ResNet28. The @ symbol refers to the number of filters by the respective CNN layers.*

LogMel-Spectrogram input (300×128)
(3×3-CNN @ 64, ReLU)×2 & 2×2-MaxPool
(SEBlock @ 64, ReLU)×2 & 2×2-MaxPool
(SEBlock @ 128, ReLU)×3 & 2×2-MaxPool
(SEBlock @ 256, ReLU)×5 & 2×2-MaxPool
(SEBlock @ 512, ReLU)×2 & 2×2-MaxPool
(3×3-CNN @ 2048, ReLU)×2 & 2×2-MaxPool
Reshape embedding (9, 8192)

functions – i. e., multi-head attention – in the context of audio MIL classification. Assuming K attention heads, the aggregated bag-level embedding per head is calculated as follows:

$$h_i^{(k)} = \frac{\sum_t^T h_{i,t} \cdot g^{(k)}(h_{i,t})}{\sum_t^T g^{(k)}(h_{i,t})}, \quad (2)$$

where $g^{(k)}(\cdot)$ is the linear function of the k-th attention head. Let $v^{(k)}$ be a $C \times 1$ matrix that parametrises the linear energy function. Then, $g^{(k)}(h_{i,t}) = \exp(h_{i,t}v^{(k)})$, using regular, exponential softmax to squash energies into probability based attention weights. The authors of [27] propose to include fixed temperature parameters $\beta^{(k)}$, of different values, inside the exponential function, which encourages each attention head to learn to attend to different duration patterns. We include in our comparison a variant according to which $\beta^{(k)}$ is adaptive.

Finally, there remains the issue of how to aggregate the K $\{h_i^{(k)}\}$: [17, 27] propose a concatenation $[h_i^{(1)}, h_i^{(2)}, ..., h_i^{(K)}]$, whereas the recent study in [30] proposes a higher-level single-head attention based weighted average thereof.

4.2. Core & prediction models

For our $\mathcal{M}_{core}$, we perform a comparative study among various recently published models in Sub-section 5.1. We found that the ResNet architecture that performed best in [23] was the best among published methods on our dataset as well[2]. Here, we also propose to enhance this network by using SE mechanisms [24] after each residual block [33] – hence SEBlock. Our proposed architecture, SE-ResNet28, is summarised in Table 2. As for our $\mathcal{M}_{pred}$, it is a single feedforward layer, followed by the sigmoid activation function, for a total of 29 blocks.

5. Experiments

Code to replicate the experiments is available at the project webpage[3]. We resample the audio files to 16 kHz, and calculate the LogMel-Spectrogram using a 128 ms Fast Fourier window, with 10 ms stride. In terms of data augmentation, we use: a) Random time shift, b) input jitter sampled from a zero-centred normal distribution with standard deviation equal to 1e-7, and c) SpecAugment [34] with 2 time and 2 frequency masks of size 24 and 16, respectively. We perform the former in an offline manner: during the 3 second positive clip segmentation process, we uniformly sampled 5 possible time-shifted positive clips in the training set; in validation and testing, we sampled only once, and the same clips were used throughout the experiments. The two latter augmentation methods are performed online. We use a batch size of 64, and the Adam optimiser [35], with initial learning rate of 1e-6. Finally, we use a NaN-safe cross entropy loss function.

On the validation set, we monitor Area Under the Precision-Recall curve (non-interpolated **AU-PR**) of the positive class for model selection, and use this model in testing. For the test set, we report AU-PR, as well as Macro averaged Area Under the Receiver Operator Characteristic curve (**AU-ROC**), F1 (**F1**) and recall (**Recall**) scores. In all cases, we performed 10 trials for which we report mean and standard deviation. For statistical significance testing, we use Welch's unequal variances t-test.

[2]We used one less basic residual block per group, for a total of four, such that it fit in our GeForce GTX 1080 Ti.

[3]https://github.com/glam-imperial/Spider-Monkey-Whinny-Detection

Table 3: *Core model comparison (%): We use average embedding pooling in all cases. ‡ denotes statistically significant improvement compared to the second best value in the same column at the p<0.01 level.*

Method	AU-PR	AU-ROC	F1	Recall
wavCRNN	37.54 ± 2.22	77.80 ± 1.53	66.87 ± 0.92	65.67 ± 2.12
melCRNN	59.69 ± 1.71	91.69 ± 0.30	76.78 ± 0.76	79.23 ± 2.04
CNN-3	46.15 ± 2.32	89.29 ± 0.87	70.49 ± 1.57	76.95 ± 2.36
VGG16	62.59 ± 3.95	91.42 ± 0.76	76.98 ± 1.69	77.34 ± 3.25
CNN14	63.12 ± 2.63	91.69 ± 0.51	77.18 ± 1.24	79.26 ± 1.65
ResNet28	65.76 ± 2.69	91.55 ± 0.79	77.07 ± 2.52	77.39 ± 4.60
SE-ResNet28	**71.76** ± 3.95 ‡	**93.32** ± 0.97 ‡	**80.95** ± 2.10 ‡	**81.08** ± 4.42

5.1. Results – core model comparison

In our core model comparison, we control for the sequence pooling method, by setting it equal to average embedding pooling. All the following models are applied on LogMel-Spectrograms, unless otherwise specified. We replicated them exactly based on the papers they were originally introduced, with two exceptions: i) We consistently observed a drop in performance if batch normalisation was used, even though we use double the batch size as [23], thus we do not use it anywhere, ii) the ResNet28, which is slightly adapted from the ResNet38 proposed in [23], and described in Sub-section 4.2. We consider the following core models in the comparison: a) **wavCRNN** [36]; a stack of 1-dimensional CNN layers applied on the raw audio waveform, followed by a stack of recurrent neural network layers (RNN), used in MIL based categorical emotion classification from speech, b) **melCRNN** [2]; a stack of 2-dimensional CNN layers followed by a stack of recurrent neural network layers (RNN), used in MIL based classification of Bornean gibbon calls, c) **CNN-3** [17]; a simple model using CNN layers followed by average pooling used in MIL based audio tagging – it is of similar complexity to the model used in [28], d) **VGG16** [37], e) **CNN-14** [23], f) **ResNet28** [23], and g) our proposed improvement **SE-ResNet28**. Table 3 summarises the results.

5.2. Discussion – core model comparison

The three bigger models – VGG16, CNN14, and ResNet28 – have similar performances, however, the only significant difference between them is that ResNet28 yields higher AU-PR at the p<0.1 level. Similarly, melCRNN performed relatively well in terms of AU-ROC and Recall, however, ResNet28 is better at AU-PR at the p< 0.01 level. We thus opted to apply the SE mechanism on this model, leading to the best performace of this experiment by SE-ResNet28, with significance at p<0.01 for AU-PR, AU-ROC, and F1, against the corresponding second best value. This extends the insights from [28] about the efficacy of SE in machine audition, on a higher complexity model.

Table 4: *Sequence pooling comparison (%): We use the SE-ResNet28 core model in all cases. The **upper** block summarises prediction, and the **lower** block, embedding pooling methods.*

Method	AU-PR	AU-ROC	F1	Recall
max	59.01 ± 7.25	89.97 ± 2.41	74.24 ± 3.36	76.05 ± 4.26
power	53.63 ± 11.96	89.54 ± 3.60	72.87 ± 4.07	73.91 ± 4.70
linsoft	**66.54** ± 4.29	92.01 ± 1.01	**78.55** ± 1.79	79.13 ± 1.88
linsoft-auto	66.07 ± 5.38	**92.44** ± 1.74	77.92 ± 1.90	**79.62** ± 2.87
avg-max	70.09 ± 5.23	93.02 ± 1.41	79.17 ± 3.61	75.20 ± 5.33
att-1	75.82 ± 1.77	94.17 ± 0.88	82.40 ± 1.92	80.16 ± 2.55
att-1-auto	73.32 ± 3.52	93.55 ± 2.37	80.66 ± 2.78	81.85 ± 3.50
att-4	**76.36** ± 1.79	**94.45** ± 0.87	**82.56** ± 2.12	82.00 ± 4.06
att-4-std	75.44 ± 2.54	94.14 ± 0.88	81.97 ± 1.38	80.15 ± 3.45
att-4-mul/res	74.50 ± 2.26	94.43 ± 0.63	80.77 ± 2.79	81.89 ± 2.60
att-4-auto	74.01 ± 2.96	94.25 ± 0.95	80.62 ± 2.33	**82.76** ± 3.34
double-att-4	70.63 ± 6.08	93.19 ± 1.60	79.63 ± 3.15	79.79 ± 3.11

5.3. Results – sequence pooling comparison

Table 4 summarises our sequence pooling comparison experiments, divided into two blocks: *prediction pooling (upper block)* and *embedding pooling (lower block)* methods. We include the following methods in the prediction pooling experiment: a) **max**; selecting the maximum instance prediction probability, b) **power** [22]; the recent adaptive Power Pooling method, c) **linsoft** [25]; the learning-free linear softmax method, and d) **linsoft-auto**; our own variant with adaptive temperature parameter. Finally, we include the following methods in the embedding pooling experiment: a) **avg-max** [23]; the concatenation of a max and an average pooling operation, b) **att-1**, and **att-1-auto**; single head attention mechanisms, the latter our own variant with adaptive temperature β, c) **att-4**; a four-head attention mechanism, with subsequent concatenation of the head-specific pooled embeddings, as used in [17, 27], d) **att-4-std**; as before, but also concatenating the weighted embedding standard deviation, as proposed in [31] and adopted in [17], e) **att-4-mul/res** [27]; as att-4, but with fixed temperature parameters as described in [27], f) **att-4-auto**; the temperature parameters are initialised as att-4-mul/res, but now are adaptive, g) **double-att-4** [30]; alternative aggregation of the head-specific embeddings, via weighted averaging through a higher level single attention head.

5.4. Discussion – sequence pooling comparison

The best prediction pooling methods respective to measure do not manage to outperform the average embedding pooling reported in Table 3; significantly lower for linsoft in AU-PR and F1 ($p<0.05$). Max performed better than power across all measures, contradicting [22], albeit without any kind of significance. Comparing linsoft and linsoft-auto is also inconclusive in terms of significance. However, linsoft and linsoft-auto were significantly better than max in AU-PR, AU-ROC, and Recall with $p<0.05$, and F1 with $p<0.01$; contradicting [21], but in line with [22, 20].

As for embedding pooling, we first observe that the simple avg-max method used in [23] performs worse than simple average pooling in all measures ($p<0.05$ for Recall). We thus have already offered a solid improvement upon the ResNet method proposed in [23], both in the core architecture, as well as in sequence pooling considerations. Single-head attention outperforms average pooling (AU-PR and AU-ROC with $p < 0.05$), as well as the better prediction pooling methods (AU-PR and AU-ROC with $p < 0.01$), going against previous comparisons [20, 21]. The application of adaptive temperature also performs in a similar manner to att-1, with no significance.

As for multi-head attention, the best method, which leads to the best result in this study, is the simple concatenation att-4. It is significantly better than double-att-4 [30] with $p<0.05$ at all measures except for Recall. We thus opted to compare further variants of att-4. However, att-4 proved to be better than the multiresolution based att-4-mul/res (AU-PR $p < 0.1$) and our adaptive variant (AU-PR $p < 0.05$, F1 $p < 0.1$); although att-4-auto performs better in Recall ($p \not< 0.1$). The classic att-4 outperformed att-4-std in all measures (however $p \not< 0.1$). Finally, att-4 was better than average pooling in all measures, with $p < 0.01$ for AU-PR, $p < 0.05$ for AU-ROC, and $p < 0.1$ for F1, thus validating the purpose of this series of experiments.

5.5. Discussion – on significance of results

We realise that our results may have limited effect compared to studies utilising larger datasets. Despite that, it was the larger models, and the parameter-heavy pooling methods that were the best performers overall, contrary to the conjecture made by in [23] that larger datasets are required to train them. That being said, these results are meant as they are: the comparative study of models and pooling methods on a smaller dataset (due to the rarity of the spider monkey and extensive specialist labour required for recording/annotation). Thus, we performed rigorous statistical significance testing, and are modest with our claims. Among previous pooling comparison studies [20, 22], only [21] reported number of trials and errorbars; no significance testing.

6. Conclusions & Future Work

We have achieved a two-fold computational improvement on the first machine learning study on spider monkey whinny call detection in the actual, Costa Rican wild: our SE-ResNet28 significantly outperformed the baseline [23], and, after extensive comparisons, we also gained further improvement by selecting a multiple head attention (with concatenation) embedding pooling method for sequence pooling. Surprisingly, we found that many of the recent tricks relating to attention based embedding pooling (including related to multi-head attention), were not better than this conceptually simpler approach. Adding the original definitions of the Wavegram joint spectrogram/waveform block [23], as well as the gated attention mechanism [17] on our SE-ResNet28, increased the model memory footprint beyond the capacity of the GPUs available to us, so we leave further adaptation considerations for future work. Another tangent to follow is a strong and weak joint supervision, similar to [38]. Finally, using the abundance of spider monkey call predictions as a proxy index of species abundance can unlock the potential for estimating population abundance. It is essential that we find new methods that allow for studying this species at a larger scale, enabling us to make robust conservation decisions [39].

7. Acknowledgements

This work is funded by the Engineering and Physical Sciences Research Council (EPSRC) Grant No. 2021037, and the DFG (German Research Foundation) Reinhart Koselleck-Project AUDI0NOMOUS (grant agreement No. 442218748).

8. References

[1] W. Turner, "Sensing biodiversity," *Science*, vol. 346, no. 6207, pp. 301–302, 2014.

[2] P. Tzirakis, A. Shiarella, R. Ewers, and B. W. Schuller, "Computer audition for continuous rainforest occupancy monitoring: The case of bornean gibbons' call detection," *Interspeech 2020*, pp. 1211–1215, 2020.

[3] B. W. Schuller, A. Batliner, C. Bergler, C. Mascolo, J. Han, I. Lefter, H. Kaya, S. Amiriparian, A. Baird, L. Stappen *et al.*, "The interspeech 2021 computational paralinguistics challenge: Covid-19 cough, covid-19 speech, escalation & primates," *arXiv preprint arXiv:2102.13468*, 2021.

[4] I. Himawan, M. Towsey, B. Law, and P. Roe, "Deep learning techniques for koala activity detection," *Interspeech 2018*, pp. 2107–2111, 2018.

[5] S. Hantke, N. Cummins, and B. Schuller, "What is my dog trying to tell me? the automatic recognition of the context and perceived emotion of dog barks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2018, pp. 5134–5138.

[6] C. Bergler, H. Schröter, R. X. Cheng, V. Barth, M. Weber, E. Nöth, H. Hofer, and A. Maier, "Orca-spot: An automatic killer whale sound detection toolkit using deep learning," *Scientific reports*, vol. 9, no. 1, pp. 1–17, 2019.

[7] M. Ferrari, H. Glotin, R. Marxer, and M. Asch, "Docc10: Open access dataset of marine mammal transient studies and end-to-end cnn classification," in *2020 International Joint Conference on Neural Networks*. IEEE, 2020, pp. 1–8.

[8] A. M. Usman, O. O. Ogundile, and D. J. Versfeld, "Review of automatic detection and classification techniques for cetacean vocalization," *IEEE Access*, vol. 8, pp. 105 181–105 206, 2020.

[9] P. Best, M. Ferrari, M. Poupard, S. Paris, R. Marxer, H. Symonds, P. Spong, and H. Glotin, "Deep learning and domain transfer for orca vocalization detection," in *International joint conference on neural networks*, 2020.

[10] B. W. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychosz, R. Vollmann, S.-D. Roelen, S. Schnieder, E. Bergelson *et al.*, "The interspeech 2019 computational paralinguistics challenge: Styrian dialects, continuous sleepiness, baby sounds & orca activity," *Interspeech 2019*, pp. 2378–2382, 2019.

[11] A.-M. Solomes and D. Stowell, "Efficient bird sound detection on the bela embedded system," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 746–750.

[12] T. Grill and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," in *2017 25th European Signal Processing Conference*. IEEE, 2017, pp. 1764–1768.

[13] D. Stowell, M. D. Wood, H. Pamuła, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.

[14] J. LeBien, M. Zhong, M. Campos-Cerqueira, J. P. Velev, R. Dodhia, J. L. Ferres, and T. M. Aide, "A pipeline for identification of bird and frog species in tropical soundscape recordings using a convolutional neural network," *Ecological Informatics*, vol. 59, p. 101113, 2020.

[15] Y. Shiu, K. Palmer, M. A. Roch, E. Fleishman, X. Liu, E.-M. Nosal, T. Helble, D. Cholewiak, D. Gillespie, and H. Klinck, "Deep neural networks for automated detection of marine mammal species," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[16] Ó. M. Chaves, K. E. Stoner, V. Arroyo-Rodríguez, and A. Estrada, "Effectiveness of spider monkeys (ateles geoffroyi vellerosus) as seed dispersers in continuous and fragmented rain forests in southern mexico," *International Journal of Primatology*, vol. 32, no. 1, pp. 177–192, 2011.

[17] S. Hong, Y. Zou, and W. Wang, "Gated multi-head attention pooling for weakly labelled audio tagging," *Interspeech 2020*, pp. 816–820, 2020.

[18] H. Dinkel, M. Wu, and K. Yu, "Towards duration robust weakly supervised sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 887–900, 2021.

[19] B. McFee, J. Salamon, and J. P. Bello, "Adaptive pooling operators for weakly labeled sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2180–2193, 2018.

[20] Y. Wang, J. Li, and F. Metze, "A comparison of five multiple instance learning pooling functions for sound event detection with weak labeling," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 31–35.

[21] C.-C. Kao, M. Sun, W. Wang, and C. Wang, "A comparison of pooling methods on lstm models for rare acoustic event classification," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 316–320.

[22] Y. Liu, H. Chen, P. Zhang *et al.*, "Power pooling: An adaptive pooling function for weakly labelled sound event detection," *arXiv preprint arXiv:2010.09985*, 2020.

[23] 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.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[25] A. Dang, T. H. Vu, and J.-C. Wang, "Deep learning for dcase2017 challenge," in *Technical Report, DCASE2017 Challenge*.

[26] M. India, P. Safari, and J. Hernando, "Self multi-head attention for speaker recognition," *Interspeech 2019*, pp. 4305–4309, 2019.

[27] Z. Wang, K. Yao, X. Li, and S. Fang, "Multi-resolution multi-head attention in deep speaker embedding," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 6464–6468.

[28] J. Naranjo-Alcazar, S. Perez-Castanos, P. Zuccarello, and M. Cobos, "Acoustic scene classification with squeeze-excitation residual networks," *IEEE Access*, vol. 8, pp. 112 287–112 296, 2020.

[29] C.-C. Kao, B. Shi, M. Sun, and C. Wang, "A joint framework for audio tagging and weakly supervised acoustic event detection using densenet with global average pooling," *Interspeech 2020*, pp. 846–850, 2020.

[30] M. India, P. Safari, and J. Hernando, "Double multi-head attention for speaker verification," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6144–6148.

[31] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive statistics pooling for deep speaker embedding," *Interspeech 2018*, pp. 2252–2256, 2018.

[32] C. J. Campbell, *Spider monkeys: Behavior, ecology and evolution of the genus Ateles*. Cambridge University Press, 2008, vol. 55.

[33] 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.

[34] 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," *Interspeech 2019*, pp. 2613–2617, 2019.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[36] G. Rizos, A. Baird, M. Elliott, and B. Schuller, "Stargan for emotional speech conversion: Validated by data augmentation of end-to-end emotion recognition," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 3502–3506.

[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[38] L. Lin, X. Wang, H. Liu, and Y. Qian, "Guided learning for weakly-labeled semi-supervised sound event detection," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2020, pp. 626–630.

[39] V. Arroyo-Rodríguez and L. Fahrig, "Why is a landscape perspective important in studies of primates?" *American Journal of Primatology*, vol. 76, no. 10, pp. 901–909, 2014.