NEURAL AUDIO CAPTIONING BASED ON CONDITIONAL SEQUENCE-TO-SEQUENCE MODEL

Shota Ikawa[1], *Kunio Kashino*[1,2]

[1] Graduate School of Information Science and Technology, The University of Tokyo, Japan
[2] NTT Communication Science Laboratories, NTT Corporation, Japan

ABSTRACT

We propose an audio captioning system that describes non-speech audio signals in the form of natural language. Unlike existing systems, this system can generate a sentence describing sounds, rather than an object label or onomatopoeia. This allows the description to include more information, such as how the sound is heard and how the tone or volume changes over time, and can accommodate unknown sounds. A major problem in realizing this capability is that the validity of the description depends not only on the sound itself but also on the situation or context. To address this problem, a conditional sequence-to-sequence model is proposed. In this model, a parameter called "specificity" is introduced as a condition to control the amount of information contained in the output text and generate an appropriate description. Experiments show that the model works effectively.

Index Terms— audio captioning, unknown sounds, sequence-to-sequence model, cross-modal embedding

1. INTRODUCTION

Sound plays an important role in our daily life. It helps us to understand the events around us. In the realm of computational auditory scene analysis, the major topics have been sound source separation, acoustic event detection, and its classification [1]. For example, studies on environmental sounds include the detection or classification of acoustic events [2, 3], acoustic scene classification [4], and abnormal sound detection [5, 6]. However, little work has been done regarding detailed description of sounds.

Against this background, here we address audio captioning for non-speech audio signals. Audio captioning here means generating texts describing sounds given an audio signal as an input. Such captions can include more information than just an acoustic event label can, such as how the sound is heard and how the tone or volume changes over time.

An audio caption is a way to visualize acoustic information so that we can understand what is happening at a glance, even without actually hearing the sound. Therefore, it will be useful for multimedia content search, sound effect search, abnormality search, and closed captioning systems that can describe non-speech sounds. To the best of our knowledge, no work has been reported regarding automatic audio captioning systems that can generate a sound description in the form of a full sentence.

This paper is organized as follows. Section 2 details the audio captioning problem. Section 3 describes the proposed audio captioning models: the basic model and the conditional model. Section 4 explains the experimental results, which show the effectiveness of the proposed model. Section 5 concludes the paper.

2. PROBLEM OF AUDIO CAPTIONING

2.1. Related Works

Recently, an onomatopoeia generation system has been proposed [7, 8]. Here, onomatopoeia means a word or a sequence of phonemes that directly imitates a sound. Based on an encoder-decoder model, the system produces valid onomatopoeias for various input sounds. Onomatopoeia generation can be viewed as a kind of natural language generation for sounds. However, an audio caption is a sequence of words rather than phonemes, and the correspondence to the input sound is highly indirect. Whether such an indirect sequence conversion is possible or not has been an open problem.

Another related task is image captioning. Compared to object recognition, image captioning produces not only a list of the object labels contained in an image but also sentences that may include their attributes or the relationships among them. Recently, systems based on the encoder-decoder model [9, 10] have achieved reasonably good accuracy [11, 12]. In those studies, conditional neural networks (CNN) pre-trained for an image classification task were employed as the encoder, and the recurrent language model (RLM) [13] was used for caption generation based on a fixed size vector. Video captioning has also been addressed, and the long short-term memory (LSTM) was shown to effectively deal with an input with variable length [14].

However, information contained in audio signals can be much more ambiguous than that in images. In fact, it is often difficult even for humans to accurately recognize the objects in an audio signal. Moreover, how to decide the best description is not obvious for given audio because the validity of the description generally depends on the situation or context as well as the sound itself. For example, a short warning may be more appropriate than a long description and vice versa. It is important to note that such problems particularly come to light in the audio captioning task.

2.2. Specificity Conditioning

To deal with the avobe-mentioned nature of the audio captioning problem, we introduce a specificity measure of the output text based on the amount of information that the text carries.

Let p_w be the probability of appearance of a word w in a language. The amount of information carried by a word w is defined as a negative logarithm of its probability:

$$I_w \equiv -\log p_w \tag{1}$$

Given an arbitrary natural language corpus, or a dataset of audio captions, we can estimate p_w by $p_w = N_w/N$, where N_w is the

https://doi.org/10.33682/7bay-bj41

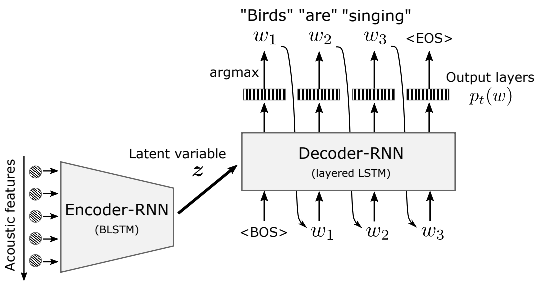

Figure 1: Block diagram of SCG.

number of appearances of w, and N is the total number of words in a language corpus or training dataset.

We consider I_w as specificity of w, and define I_s, the specificity of an audio caption s consisting of words $w_1, w_2, \ldots, w_n$, by the sum of the information values with respect to the words in $\boldsymbol{s}$:

$$I_{\boldsymbol{s}} \equiv \sum_{t=1}^{n} I_{w_t} \tag{2}$$

Obviously, $I_{\boldsymbol{s}}$ becomes high when infrequent words are contained in the caption or the caption is long in terms of the number of words.

In the audio captioning, more (or less) specificity is not always better, and therefore, the ability to control the text specificity is essential in generating a valid output text. In the following section, we first propose a caption generator based on a plain encoder-decoder model, and then we extend it to a conditional encoder-decoder model where the specificity is treated as a condition for caption generation.

3. PROPOSED MODEL

3.1. Sequence-to-Sequence Caption Generator

Figure 1 shows the audio caption generator with the plain sequence-to-sequence caption generator (SCG).

A series of acoustic features $\boldsymbol{x}$ is input to the encoder consisting of recurrent neural network (RNN) and embedded within a fixed length vector $\boldsymbol{z}$, which is a latent variable that serves as latent features of the input acoustic signal. Then, the decoder is initialized based on the derived latent variable. The decoder serves as an RLM, which calculates estimated probabilities of every word in each step and chooses the word with the highest probability w_t as input for the next step. The generated audio caption $\hat{\boldsymbol{s}} = (w_1, w_2, \ldots)$ is the series of the chosen words. The input in the first step is "BOS (beginning of the sentence)", and the generation of the audio caption finishes in a step when "EOS (end of the sentence)" is chosen.

The SCG can be viewed as an approximation of the generative model for the following optimal audio caption $\bar{s}$:

$$\bar{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} p(\boldsymbol{s} \mid \boldsymbol{z}), \quad z = f(\boldsymbol{x}), \tag{3}$$

where f is a mapping to derive latent variables from acoustic signals and corresponds to the encoder in the model. $p(\boldsymbol{s} \mid \boldsymbol{z})$ is a probability distribution in which each audio caption is generated when the latent variable is given. The decoder is expected to generate audio captions with the highest probability.

Pairs comprising an acoustic signal and audio caption for the signal are used for learning this model. Given an acoustic signal as input, the model calculates cross entropy between the output layer of the decoder and the corresponding word of the target audio caption in each step of the decoder. Summation of all the cross entropy values is used as a loss function $\mathcal{L}_{\text{gen.}}$, which is viewed as the loss of the audio caption generation. Let $\boldsymbol{o_t}$ be the vector provided by the output layer in step t, $\boldsymbol{y}_t$ be the one-hot vector representing w_t, the tth word in the current training sentence, and n be the number of words in the sentence. The error function is then expressed as follows:

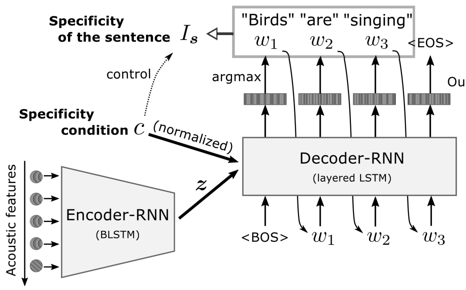

Figure 2: Block diagram of CSCG. Specificity condition c is given to the decoder, so that the resulting output sentence has the specificity close to the value of c.

$$\mathcal{L}_{\text{gen.}} \equiv \sum_{t=1}^{n} \text{cross entropy}(\boldsymbol{o}_t, \boldsymbol{y}_t) \tag{4}$$

$$= \sum_{t=1}^{n} -\log(\hat{p}_t(w_t)) \tag{5}$$

Then, the model is optimized by backpropagation based on $\mathcal{L}_{\text{gen.}}$.

3.2. Conditional Sequence-to-Sequence Caption Generator

Inspired by the conditional generative models that have been successfully applied in various works [15, 16, 17, 18, 19], here we propose a conditional sequence-to-sequence caption generator (CSCG) to control the specificity of the generated audio captions.

As illustrated in Figure 2, the encoder of the CSCG works in the same manner as that of the SCG. In addition, specificity condition c is given to the decoder in addition to the latent variable derived from the audio signal. The CSCG is trained to generate the following optimal audio caption $\bar{s}$:

$$\bar{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} p(\boldsymbol{s} \mid \boldsymbol{z}, c) \tag{6}$$

We expect $\bar{s}$ to have a specificity close to c and, at the same time, correctly correspond to the input acoustic signal.

We can train the CSCG by alternatively performing the following two steps. In the first step, audio caption generation and the specificity are learned simultaneously. The pairs of acoustic signals and audio captions are used for the learning. The specificity of an audio caption of these pairs, I_s, is input to the decoder as the specificity condition, and the model is trained by backpropagation. To control the specificity of the generated captions, we introduce the specificity loss $\mathcal{L}_{\text{sp.}}$. The total loss function in this step, $\mathcal{L}_{\text{SC-1}}$, is defined as the weighted sum of $\mathcal{L}_{\text{gen.}}$ and $\mathcal{L}_{\text{sp.}}$:

$$\mathcal{L}_{\text{SC-1}} \equiv \mathcal{L}_{\text{gen.}} + \lambda \mathcal{L}_{\text{sp.}}, \tag{7}$$

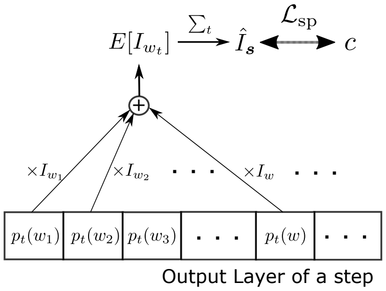

Figure 3: Estimation of specificity of a generated caption and specificity loss. Each value at the output unit, $p_t(w)$, denotes the estimated probability for the corresponding word w at step t.

where λ is a hyperparameter to balance the two loss values. Audio captioning is an ill-posed problem with potentially multiple solutions, and the second term has the role of regularization to determine the unique solution by adding constraints to the specificity of the generated caption.

The text generation with the decoder includes a discrete process to choose one word in each step, and that makes it impossible to backpropagate the losses. To solve this problem, we approximate the specificity of generated captions without discrete processes. Figure 3 illustrates the estimation process. The expectation of the amount of information of $\hat{w}_t$ corresponding to the tth step of the decoder can be calculated using its output layer, each unit of which encodes the probability of corresponding words.

$$E[I_{\hat{w}_t}] = \sum_w \hat{p}_t(w) I_w \ . \tag{8}$$

The summation of $E[I_{\hat{w}_t}]$ for all steps is $\hat{I}_{\hat{\boldsymbol{s}}}$, which is the estimated value of the specificity of the generated caption $\hat{\boldsymbol{s}}$.

$$\hat{I}_{\hat{\boldsymbol{s}}} = \sum_{t=1}^{n} E[I_{\hat{w}_t}] \tag{9}$$

Here, we define specificity loss $\mathcal{L}_{\text{sp.}}$ as follows:

$$\mathcal{L}_{\text{sp.}} \equiv (\hat{I}_{\hat{\boldsymbol{s}}} - c)^2 \ . \tag{10}$$

We can optimize the model by backpropagation using $\mathcal{L}_{\text{sp}}$, because it is calculated from the vectors obtained from the output layer of the decoder by using multiplication and addition only.

The second step is introduced to alleviate overfitting with respect to specificity. In this step, the decoder is trained only with texts rather than audio-text pairs. First, a latent variable $\boldsymbol{z}$ is extracted in advance from an audio signal by using the encoder with the current parameter. This means that we sampled $\boldsymbol{z}$ from real audio signals, rather than using random vectors, but signals not associated with any audio captions can be used here. Then, the specificity condition c is generated randomly. As training sentences, any captions with the closest specificity value to c can be used. We train only the decoder using backpropagation based on the following loss function $\mathcal{L}_{\text{SC-2}}$:

$$\mathcal{L}_{\text{SC-2}} \equiv \lambda' \mathcal{L}_{\text{gen.}} + \lambda \mathcal{L}_{\text{sp.}} \tag{11}$$

Hyperparameter λ' smaller than 1 is chosen. Even when the audio caption for calculating $\mathcal{L}_{\text{gen.}}$ does not correspond to the input signal, the first term has the role of regularization to suppress the generation of unnatural sentences.

Table 1: Experimental conditions.

Decoder LSTM layers	3
LSTM cells	512
Latent variable dimensions	256
Output word labels	1177
Normalization of c	division ($\max(I_s) \rightarrow 2.0$)
Batchsize	200
Total epoch	400
Hyper-parameter λ	2.0×10^{-2}
Hyper-parameter λ'	1.0×10^{-2}
Optimizer	ADAM [20]
MFCC dimensions	80
FFT window (MFCC)	2048 samples
FFT shift (MFCC)	512 samples

4. EVALUATION

To evaluate the effectiveness of the proposed model, we performed objective and subjective experiments.

4.1. Dataset

We used a part of the audio signals contained in Free Sound Dataset Kaggle 2018 [21], which is a subset of FSD [22] and includes various sound samples digitized at 44.1 kHz with linear PCM of 16 bit accuracy. We chose 392 signals for the training set and 29 for the test set. These signals are not longer than 6 s in length and include various everyday sounds.

To build the training set, audio captions were collected from human listeners. All the collected captions were in Japanese. Since one audio signal can correspond to various captions with various specificity values, multiple audio captions were attached to each audio signal. To accomplish this, 72 Japanese speakers were asked to describe the sound in Japanese text. We associated one to four audio captions for each training signal, and five audio captions for each test signal. The total numbers of captions were 1,113 in the training dataset and 145 in the test dataset. Then, the captions for the training signals were augmented to be expanded to 21,726, by manually deleting or replacing the words.

4.2. Conditions

Table 1 lists the experimental conditions. We used a series of mel-frequency cepstral coefficients (MFCC) and f0 as the input. The vocabulary size for the system was 1,440, as there were 1,437 kinds of words in the audio captions for training, and three special symbols “BOS”, “EOS”, and “UNK” (unknown word).

4.3. Examples

Table 2 shows some examples of the captions generated from the test signals. They were manually translated from the Japanese output.

4.4. Controllability of Specificity

Table 3 lists the averages and the standard deviations of the specificities for generated captions. Since the SCG does not deal with the specificity, the standard deviation is relatively large. On the other hand, the specificity values with the CSCG on average are quite close to the conditioned input c. This shows that the proposed conditioning mechanism works effectively.

Table 2: Examples of generated audio captions (English translation).

Sound source	Methods	c	Generated captions
Bell	SCG		A high-pitched metallic sound echoes.
	CSCG	20	A loud sound.
		50	A high sound like friction of metals.
		80	A metal bell is hit only once and makes loud, high and sustained sound.
		110	A high-pitched sound sounds as if a metal is hit, first loudly and then gradually fades out.
Bass drum	SCG		A low sound rings for a moment.
	CSCG	20	A low sound sounds for a moment.
		50	A light and low-pitched sound as if something is dashed on the mat for a moment.
		80	A drum is uninterestedly played, making a faint, very low-pitched sound only once.
		110	A faint, low-pitched sound sounds as if something is hit dully, and it soon disappears.
Glass	SCG		High-pitched sounds continue as if a metal is rolling
	CSCG	20	Glass is broken.
		50	A dry sound of breaking glass sounds once a little loudly.
		80	A high-pitched sound as if glass is breaking diminishes in a moment.
		110	A high, cold sound as if glass is breaking is heard for one or two seconds.

Table 3: Specificity of the generated captions.

Methods	c	Average	SD
SCG		38.0	21.2
CSCG	20	21.7	2.4
	50	57.7	5.0
	80	90.5	9.5
	110	107.2	20.6

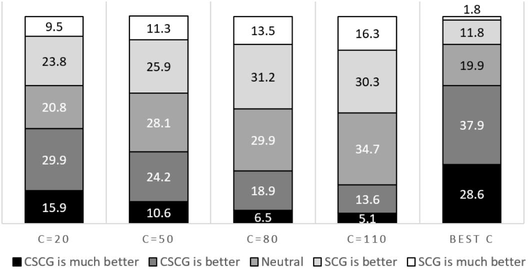

Figure 4: Comparisons of SCG and CSCG. The "BEST c" bar shows the results obtained by using the best c value (among the four) for each signal. The numbers in the bars show the percentage.

4.5. Objective Scores

Table 4 shows the BLEU scores. The CSCG with $c = 50$ marks the best BLEU, 17.01%, but it is still lower than that of human captions. Note that BLEU has a penalty for short sentences, which adversely affected the BLEU of the CSCG with low c values.

Table 4: BLEU Scores.

Methods	c	BLEU [%]
SCG		13.02
CSCG	20	5.83
	50	17.01
	80	12.52
	110	11.21
Human		22.35

4.6. Subjective Evaluation

We evaluated the proposed methods with two kinds of subjective evaluations.

Evaluation 1 investigated acceptability for the generated captions. The test audio signals and corresponding generated captions were presented to 41 subjects who understand Japanese. The subjects evaluated the captions in four levels: "very suitable", "suitable", "partly suitable" and "unsuitable". These answers were converted to points of 3, 2, 1 and 0, and the values of the average were the metric of acceptability. The captions given by humans were also evaluated for comparison. All the subjects responded to the 29 sound sources, for a total of 1,189 responses. Table 5 shows the results. The average scores of all methods are over 1.0, which is higher than the point of "partly suitable". The CSCG with $c = 20$ has the best acceptability within the proposed method.

Table 5: Acceptability scores.

Methods	c	Average	SD
SCG		1.45	1.13
CSCG	20	1.69	1.17
	50	1.29	1.11
	80	1.14	1.16
	110	1.07	1.07
Human		2.11	0.87

Evaluation 2 compared the SCG and the CSCG models. The subjects were presented with one audio signal and two audio captions, "A" and "B." They were then asked to choose one of the five options: "A is much better", "A is better", "Neutral", "B is better", or "B is much better". Either "A" or "B" (randomly selected) was the audio caption generated with the SCG and the other was the one generated with the CSCG. Figure 4 shows the results. With an appropriate choice of c, CSCG outperformed SCG for about 2/3 of the test samples. That is, if the optimal c value is known somehow in advance, CSCG can produce better captions compared with SCG.

5. CONCLUSION

This paper proposed a neural audio captioning system for audio signals. The experiments showed that two versions of the proposed method, SCG and CSCG, work effectively and that the conditional version (CSCG) can successfully control the amount of information contained in the output sentence. It was also shown that CSCG generated subjectively better captions than SCG when we could choose the best specificity value for each signal. Unlike the existing audio classification systems, the proposed system does not solve the classification problem but performs sentence generation using the learned vocabulary, as in machine translation. For this reason, it tends to perform reasonably well even for unknown or ambiguous sounds. In our future work, we will investigate a specificity adaptation method for individual sounds, situations, and applications.

6. REFERENCES

[1] Mark D. Plumbley, Christian Kroos, Juan P. Bello, Gael Richard, Daniel P.W. Ellis, and Annamaria Mesaros: *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*. Tampere University of Technology. Laboratory of Signal Processing (2018).

[2] Haomin Zhang, Ian McLoughlin, and Yan Song: Robust sound event recognition using convolutional neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 559–563 (2015).

[3] Emre Cakir, Toni Heittola, Heikki Huttunen, and Tuomas Virtanen Polyphonic sound event detection using multi label deep neural networks. In *2015 International Joint Conference on Neural Networks (IJCNN)*, 1–7 (2015).

[4] Daniele Barchiesi, Dimitrios Giannoulis, Dan Stowell, and Mark D. Plumbley Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Processing Magazine*, 32, 3, 16–34 (2015).

[5] Cristhian Potes, Saman Parvaneh, Asif Rahman, and Bryan Conroy: Ensemble of feature-based and deep learning-based classifiers for detection of abnormal heart sounds. In *2016 Computing in Cardiology Conference (CinC)*, 621–624 (2016).

[6] Yuma Koizumi, Shoichiro Saito, Hisashi Uematsu, and Noboru Harada: Optimizing acoustic feature extractor for anomalous sound detection based on Neyman-Pearson lemma. In *2017 25th European Signal Processing Conference (EUSIPCO)*, 698–702 (2017).

[7] Shota Ikawa and Kunio Kashino: Generating Sound Words from Audio Signals of Acoustic Events with Sequence-to-Sequence Model. In *Acoustics, Speech and Signal Processing (ICASSP)*, 346 – 350 (2018).

[8] Shota Ikawa and Kunio Kashino: Acoustic event search with an onomatopoeic query: measuring distance between onomatopoeic words and sounds. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE)*, 59–63 (2018).

[9] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le: Sequence to Sequence Learning with Neural Networks. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 3104–3112 (2014).

[10] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio: Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. *CoRR* abs/1406.1078 (2014).

[11] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan: Show and Tell: A Neural Image Caption Generator. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2015).

[12] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio: Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2048–2057 (2015).

[13] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Cernockỳ, and Sanjeev Khudanpur: Recurrent neural network based language model. In *INTERSPEECH*, 1045–1048 (2010).

[14] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. : Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2625–2634 (2015).

[15] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve J. Young: Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems. *CoRR* abs/1508.01745 (2015).

[16] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan: A Persona-Based Neural Conversation Model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 1, 994–1003 (2016).

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros.: Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1125–1134 (2017).

[18] Mehdi Mirza and Simon Osindero: Conditional generative adversarial nets. *arXiv:1411.1784* (2014).

[19] Emily L. Denton, Soumith Chintala, Rob Fergus, et al.: Deep generative image models using a Laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*, 1486–1494 (2015).

[20] Diederik P. Kingma and Jimmy Ba: Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference on Learning Representation (ICLR)* (2015).

[21] FSD: Freesound General-Purpose Audio Tagging Challenge — Kaggle. `https://www.kaggle.com/c/freesound-audio-tagging/data`. accessed 2018/12/31.

[22] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andrés Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra: Freesound Datasets: a platform for the creation of open audio datasets. In *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 486–493 (2017).