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

An evaluation of data augmentation methods for sound scene geotagging

Helen L. Bear[1], *Veronica Morfi*[1], *Emmanouil Benetos*[1,2]

[1]Centre for Digital Music, Queen Mary University of London, London, UK
[2]The Alan Turing Institute, London, UK
{h.bear, g.v.morfi, emmanouil.benetos}@qmul.ac.uk

Abstract

Sound scene geotagging is a new topic of research which has evolved from acoustic scene classification. It is motivated by the idea of audio surveillance. Not content with only describing a scene in a recording, a machine which can locate where the recording was captured would be of use to many. In this paper we explore a series of common audio data augmentation methods to evaluate which best improves the accuracy of audio geotagging classifiers.

Our work improves on the state-of-the-art city geotagging method by 23% in terms of classification accuracy.

Index Terms: computational sound scene analysis, data augmentation, sound scene geotagging, city classification

1. Introduction

Sound scene geotagging is a relatively novel topic of research. It means to correctly identify the geographical position where a recording was captured [1] and as such it can also be described as either audio geotagging or as a more specific task such as audio-based city classification. Subject to the precision of the dataset annotations, it could be formulated as a coarse target problem (e.g. at the level of a city) or an investigation of an even more precise target (e.g. at the level of coordinates or postcodes), if such labels are provided. To the best of our knowledge no such dataset exists with labels on a more granular level than cities. Hence, for this paper, city classification from sound scenes is the primary geotagging task.

Sound scene geotagging is a task with many future real-world applications, such as automatically knowing a precise geo-location from the background sounds of an emergency call that could potentially decrease the response time [2] or it could position a person in a location with no cameras at a specific time for criminal investigations. However, the current state-of-the-art in audio geotagging that uses multi-task learning, achieves only 56% accuracy [1] on city classification.

In this paper, we seek to improve the state-of-the-art performance in sound scene geotagging. One approach is to train on more data. However, data collection is expensive, and more so when one would have to travel to collect recordings from different cities. Thus, we seek to achieve a dataset expansion by using a series of data augmentation methods.

Data augmentation is a common way of improving accuracy scores in deep learning methods. In audio some examples are: genre classification [3], speech recognition [4], and sound event detection [5]. In this work, we explore a series of common audio data augmentation methods to evaluate which ones best improve city classification accuracy. This work is conducted using data which is varied in the types of scenes recorded in each possible city. Through our experiments we found that augmentations that can preserve all the information of the original signal, such as time shift (referred to as *cyclic* augmentation in this manuscript), provide the highest improvement on the performance of our sound scene geotagging method.

The rest of this paper is as follows: a short background in Section 2 discussing prior work is followed by Section 3 presenting the data available for sound scene geotagging and the augmentation methods used in our analysis. Section 4 follows describing our methodology and the results. Finally, Section 6 concludes this paper with our observations on these experiments and possible future work.

2. Background

Prior to the work in [1], there was little investigation of audio geotagging, and to the best of our knowledge, not as a classification task. Examples by Elizalde et al. [6] and Kumar et al. [7] used sound scenes generated with sound events not specific to the locations. This seems non-intuitive for city classification, as the sound events which are unique to a city would be easily learnt by a classifier. In addition to this, in [6] and [7] the problem of geotagging is not formulated as a classification task but rather as retrieval task seeking the most similar city in a clustering model.

Whilst sound scene geotagging has received limited attention, there is work which uses audio to help geotag videos for visual scene understanding or in multi-media systems [8]. In videos it can be rather straight-forward to identify visual features in a scene that correspond to audio tracks and use those for geotagging a location. Estimated geo-tags can be used as metadata which enables better prediction of events in videos [9]. Furthermore, they can be used for modelling relationships between spatio-temporal segments and entities in multimedia data [10].

In [1], the authors use a dataset labelled for both audio geotagging (in a city level) and acoustic scene classification. As such, the annotations function as a many-to-many connection between the city and scene classes. Each recording maps to both one scene and one city. This results in multiple types of scenes appearing in a single city and multiple cities containing each scene. Both labelling schemes are based on how humans describe a scene or identify a location [11]. The fact that the labels are not data driven means there can be features in the recordings which are confounding variables that a classifier has to learn to discriminate [12]. This dataset is also used in our experiments and it is described in the following section along with the augmentation methods applied on the recordings.

3. Data

The DCASE 2018 Acoustic Scene Classification (ASC) subtask 1A [13] contains recordings from six cities: Barcelona, Helsinki, London, Paris, Stockholm, and Vienna. The training data partition consists of approximately 70% of the recordings from each city. Recordings include ten predefined scenes [13]

http://dx.doi.org/10.21437/Interspeech.2021-1837

Table 1: *The CNN structure and parameters for a bigger model able to cope with the greater volume of training data. Adam optimiser with* $LR = 0.0001$, $beta_1 = 0.9$, $beta_2 = 0.999$, $epsilon = None$, $decay = 0.0$, $amsgrad = False$.

	Layer	params
1	Convolutional2D	filters=64, kernel=(7,7)
2	BatchNormalization	
3	MaxPooling2D	pool_size=(5,5), strides=2, padding='same'
4	Dropout	prob_drop_conv=0.3
5	Convolutional2D	filters=128, kernel=(7,7)
6	BatchNormalization	
7	MaxPooling2D	pool_size=(4,4), strides=2, padding='same'
8	Dropout	prob_drop_conv=0.3
9	Convolutional2D	filters=256, kernel=(5,5)
10	BatchNormalization	
11	MaxPooling2D	pool_size=(3,3), strides=2, padding='same'
12	Dropout	prob_drop_conv=0.3
13	Convolutional2D	filters=512, kernel=(3,3)
14	BatchNormalization	
15	MaxPooling2D	pool_size=(2,2), strides=2, padding='same'
16	Dropout	prob_drop_conv=0.3
	layer 16 output feeds into two separate output blocks, 17 and 19	
17	GlobalAveragePooling2D	
18	Dense	filters=10, activation='sigmoid', kernal_regularizer = L2(0.001)
19	GlobalAveragePooling2D	
20	Dense	filters=6, activation='sigmoid', kernal_regularizer = L2(0.001)

at random times between 9am and 9pm on different weekdays. The development set consists of 8640 audio segments in total, split into a training and an evaluation subset containing 6122 and 2518 segments, respectively. In our experiments, we use the evaluation subset for validation. The test set consists of 2518 audio segments, previously held back in order to review the performance of the DCASE challenge submissions. The development dataset contains in total 24 hours of audio, while the testing set contains in total 7 hours of audio. The equipment used for recording consists of a binaural Soundman OKM II Klassik/studio A3 electret in-ear microphone and a Zoom F8 audio recorder using 48kHz sampling rate and 24-bit resolution.

3.1. Data Preparation

Before any data augmentation is performed, we first downsample the audio files to 22050Hz. Then, in order to enhance our training set, we apply three different data augmentation methods, two on the waveforms and one on the time-frequency representations (of FFT window length of 2048), followed by trimming the lowest 100Hz and highest 100Hz frequencies. Finally, we compute the log mel-spectrogram (128 mel bands), which will be used as input to our network. Augmentations are applied only on the training set, while the validation and testing sets are kept the same.

1. Waveform augmentations:
   Cyclic. We shift each audio sample in time by 50% of its length. In a waveform this means that we cut it into two parts, at 50% of its length, and place the second part in front of the first. This augmentation is able to preserve all the information of the original waveform. Cyclic augmentation produces one new audio sample per input audio sample.

   Drop. We perform time interval dropout by skipping a random number of samples (between 1 and length of recording) in a random index in the waveform. We do this twice per audio segment, using different random numbers of skipped samples and different random indices in the waveform. Drop augmentation produces two new audio samples per input audio sample.

2. Spectrogram augmentations:
   Stretch. We apply time and frequency stretching by resizing a random number of columns and rows at a random position followed by image resizing using bilinear interpolation. This is performed four times on each audio segments with different random number of columns and rows to stretch and stretch factor. Stretch augmentation produces four new samples per input audio sample.

Table 2: *Training data variation by data augmentation method.*

Augmentation	**#Training Samples**
none	6,122
cyclic	6,122
drop	12,244
stretch	24,488
Total	48,976

The number of samples produced by each augmentation method can be found in Table 2. These will be used in addition to the original 6122 training samples for building CNNs.

These waveform augmentations methods are widely used in audio applications when retaining as much information about the original signal as possible is important (such as for animal sounds [14]). Cyclic augmentation can enhance an audio dataset by preserving the original signal information, while drop

Table 3: *City classification accuracy by data augmentation method with different CNN models.*

		cyclic	drop	stretch	cyclic+stretch	all	benchmark
Accuracy (%) obtained with the multi-task CNN from [1]							
Scene	train	85	74	86	86	83	
	validation	46	17	49	54	55	
	test	47	16	49	53	55	57
City	train	80	64	82	83	82	
	validation	56	30	53	55	60	
	test	56	29	52	55	60	56
Accuracy (%) obtained with a larger multi-task CNN							
		cyclic	drop	stretch	cyclic+stretch	all	benchmark
Scene	train	99	95	99	99	99	
	validation	70	19	62	67	65	
	test	70	19	63	67	66	57
City	train	99	94	99	99	99	
	validation	79	46	75	72	73	
	test	**79**	47	75	71	74	56
Accuracy (%) obtained with a larger single-task CNN							
		cyclic	drop	stretch	cyclic+ctretch	All	benchmark
City	train	99	99	99	99	91	
	validation	76	69	72	77	59	
	test	75	69	71	77	59	56

creates more variety of new audio signals by removing time steps. On the other hand, stretch in the time-frequency domain is adjusted from the image stretching for augmentation. We assume that with small stretching factors the overall scene and city information can be preserved in the spectrogram.

3.2. Mono vs Stereo

During the process of applying the data augmentations we experienced some challenges worthy of note; our recordings are stereo two channel audio. We tested experimentally and learned that augmentations are more effective when they are applied to each channel separately (taking care to ensure that they are consistent for aspects such as temporal alignment) before the channels are averaged into a single channel feature matrix, rather than averaging before applying the augmentation.

4. Experiments

We use three different CNN architectures. For the first one, we use the same multi-task model architecture as presented in [1]. We train a different model for each augmentation method. Each model is trained on the original data and the selected augmented data; namely *cyclic*, *drop* and *stretch*. The benchmark we seek to outperform is 56% accuracy for city classification [1]. Additionally, we train a model with all the augmentations that outperform the baseline and the original data (referred to as *cyclic+stretch*). Finally, we train a model that includes all augmentations and the original data (referred to as *all*).

For the second architecture, the same series of experiments are repeated using a larger architecture than the baseline model. The parameters and architecture of this model are detailed in Table 1. As the focus of our work is audio geotagging rather than acoustic scene classification, we also use a third architecture which does not perform multi-task predictions but solely focuses on the city classification task. The single-task architecture is similar to the multi-task model from Table 1 with the scene classification branch being removed (layers 17 and 18).

5. Results

The results for all our experiments are reported in Table 3. Best performance on city classification is achieved with the larger CNN model with the use of only the cyclic augmentation. The results of the first architecture based on [1] are shown in the top of the table. The results of the larger CNN architecture from Table 1) are in the middle of the table and the final single-task CNN results are presented in the bottom of the table.

With the architecture of [1], we struggle to significantly outperform the original benchmarks; for acoustic scene classification (ASC) none of the augmentations outperform the 57% benchmark but for sound scene geotagging (city classification), we match the benchmark with cyclic augmentations and improve on it by 4% by using all three. Using drop augmentations significantly reduces the performance, possibly due to the method removing the entirety of some shorter sound events that might be unique to particular locations. Further experiments using different drop factors could be implemented to test this hypothesis. We sought to outperform the all augments result by removing the drop augment in the cyclic+stretch trained model, however this was not successful by only scoring 55%, one percent less than the benchmark.

With our larger CNN architecture, different behaviour is observed. Whilst drop augmentation is still detrimental compared to the benchmarks for both ASC and audio geotagging (city classification), both cyclic and stretch achieved large increases in accuracy, $+13\%$ and $+6\%$ respectively for ASC and $+23\%$ $+19\%$ for audio geotagging. The performance of cyclic augmentations for this model provides us with the best overall accuracy of 79%. Whilst the drop augmentation is still the lowest performing one, it is improved by the larger model architecture with 47% test accuracy compared to the previous model with 29% test accuracy.

Finally, for the single-task CNN performing only city classification, all augmentation methods, including drop, achieve an accuracy greater than the benchmark. However, the model using all augmentation methods together has a lower performance than any of the models using single augmentations. The better

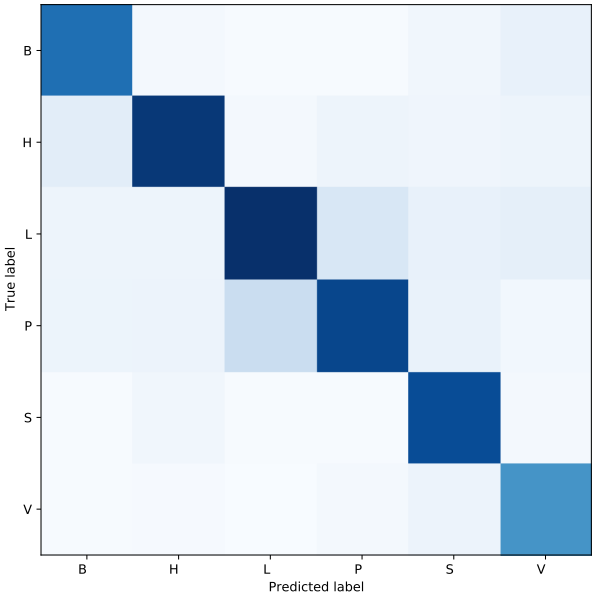

Figure 1: *Confusion matrix for our best model (cyclic multi-task CNN). Classes are denoted by their first initial: B(arcelona), H(elsinki), L(ondon), P(aris), S(tockholm), V(ienna)*

performance of drop augmentations for single-task city classification implies that this augmentation is not fitting for scene classification, which in the multi-task model acts as a regulariser over the city classification task, reducing overall performance. Despite this robust performance by all augmentations, the cyclic and stretch single-task CNNs are not performing as well as the multi-task models, but the cyclic+stretch accuracy increases by $+6\%$ to 77% accuracy.

Overall, cyclic augmentations have the best performance compared to all other augmentation methods. Drop augmentations appear to reduce the accuracy for ASC models leading us to believe that this type of augmentation might not be appropriate to preserve some important information for audio scenes. It is possible that a more conservative number of frames to drop would perform better, as in current experiments the number of frames can be up to the length of the recording.

Our larger multi-task CNN architecture has overall a better performance for both scene and city classification compared to the CNN model proposed in [1]. The best model for city classification is the larger multi-task model presented in Table 1 with an accuracy of 79%. Furthermore, except for the single-task model using all augmentations, the models using augmented data converged in less epochs than the baseline without any augmentation. Whilst this was not the original goal of this paper it is a useful observation for optimising training geotagging classifiers.

Figure 1 shows a city-wise confusion matrix for the multi-task model trained using cyclic augmentations which scored 79% accuracy in round two. The bulk of the accuracy comes from the London and Paris recordings which are very robustly classified. This is consistent with the observations in [1] which uses the same dataset. Also consistent with these prior observations is that Vienna is the least well identified of the six cities.

Finally, it was observed in [12] that the difficulty of predicting cities can vary by the scene itself. Therefore, to understand the quality of our predictive models, we list the city classification accuracy of our best performing model for each scene in Table 4. The number of test files for each scene is in the last column. It is apparent that the *airport* and *park* scenes are having significant difficulty in separating out which cities they are recorded in with 59% and 47% accuracy, respectively. This might be due to the international/formulaic organisation and behaviour of airports all over the world, and the quiet outdoors nature of park recordings which are unlikely to contain distinctive features. This would be an area worthy of further research in the future for tackling audio geotagging in difficult scenes.

Table 4: *City classification test Accuracy (%) by Scene with the best model*

Scene	City Accuracy	#testfiles
Airport	59	265
Bus	91	242
Metro	90	261
Metro Station	89	259
Park	47	242
Pedestrian Street	74	247
Public Square	81	216
Shopping Mall	77	279
Traffic	86	246
Tram	91	261

6. Conclusions and Future work

In this work we have completed a comprehensive review of common data augmentation methods with a goal of improving audio geotagging (city classification on sound scenes). We have learned that applying the augmentations to stereo input instead of mono input and then mixing the two channels (with care taken for time alignment where necessary) is the optimal method of preparing the data for training.

Our results suggest that while different augmentations work for each task in the multi-task model, for example *drop* is not good for ASC but *cyclic* is best for geotagging, using data augmentations overall can improve the performance for city classification. Also, supplementing training data with data augmentation enables faster converging of CNN geotagging classifiers.

Other future work includes better model development and parameter optimisation, and testing on the latest twelve city dataset from DCASE 2019. As discussed in Section 5 we will further evaluate augmentation parameters such as the *drop* length. We will further improve scene specific geotagging for difficult scenes, such as *park* and *airport* with non-discriminative features.

Ultimately, we have reviewed the most common audio data augmentation methods and shown that with *cyclic* augmentation and a multi-task CNN, we have improved the state-of-the-art sound scene geotagging by $+23\%$ to 79% accuracy.

7. Acknowledgements

VM is supported by the BBSRC grant BB/R008736/1. EB is supported by a Turing Fellowship under the EPSRC grant EP/N510129/1.

8. References

[1] H. L. Bear, T. Heittola, A. Mesaros, E. Benetos, and T. Virtanen, "City classification from multiple real-world sound scenes,"

in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2019, pp. 11–15.

[2] R. G. Malkin and A. Waibel, "Classifying user environment for mobile applications using linear autoencoding of ambient audio," in *Proc' IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, March 2005, pp. v/509–v/512 Vol. 5.

[3] R. L. Aguiar, Y. M. Costa, and C. N. Silla, "Exploring data augmentation to improve music genre classification with convnets," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[4] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[5] Y. Chen and H. Jin, "Rare sound event detection using deep learning and data augmentation." in *INTERSPEECH*, 2019, pp. 619–623.

[6] B. Elizalde, G. Chao, M. Zeng, and I. Lane, "City-identification of flickr videos using semantic acoustic features," in *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, April 2016, pp. 303–306.

[7] A. Kumar, B. Elizalde, and B. Raj, "Audio content based geotagging in multimedia," *Proc. Interspeech 2017*, pp. 1874–1878, 2017.

[8] J. Luo, D. Joshi, J. Yu, and A. Gallagher, "Geotagging in multimedia and computer vision—a survey," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 187–211, 2011.

[9] D. Joshi and J. Luo, "Inferring generic activities and events from image content and bags of geo-tags," in *Proceedings of the 2008 international conference on Content-based image and video retrieval*, 2008, pp. 37–46.

[10] J. Luo, M. Boutell, and C. Brown, "Pictures are not taken in a vacuum-an overview of exploiting context for semantic scene content understanding," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 101–114, 2006.

[11] H. Bear and E. Benetos, "An extensible cluster-graph taxonomy for open set sound scene analysis," in *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2018.

[12] D. Heise and H. L. Bear, "Visually exploring multi-purpose audio data," 2021.

[13] A. Mesaros, T. Heittola, and T. Virtanen, "A multi-device dataset for urban acoustic scene classification," in *arXiv preprint arXiv:1807.09840*, November 2018, pp. 9–13.

[14] L. Nanni, G. Maguolo, and M. Paci, "Data augmentation approaches for improving animal audio classification," *Ecological Informatics*, vol. 57, p. 101084, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1574954120300340