MAXIMIZING AUDIO EVENT DETECTION MODEL PERFORMANCE ON SMALL DATASETS THROUGH KNOWLEDGE TRANSFER, DATA AUGMENTATION, AND PRETRAINING: AN ABLATION STUDY

Daniel Tompkins, Kshitiz Kumar, Jian Wu*

Microsoft

ABSTRACT

An Xception model reaches state-of-the-art (SOTA) accuracy on the ESC-50 dataset for audio event detection through knowledge transfer from ImageNet weights, pretraining on AudioSet, and an on-the-fly data augmentation pipeline. This paper presents an ablation study that analyzes which components contribute to the boost in performance and training time. A smaller Xception model is also presented which nears SOTA performance with almost a third of the parameters.

Index Terms— Audio Event Detection, Data Augmentation, Knowledge Transfer, Ablation Study

1. INTRODUCTION

Audio Event Detection (AED) has greatly benefited from deep-learning methods with CNN-based models and, more recently, Transformer-based models providing significant increases in classification accuracy and raising the state-of-the-art (SOTA). However, many AED datasets have a small number of labeled examples, especially when compared to other domains such as vision and language. This limits the ability of large models to train directly on datasets without over- or under-fitting.

Solutions to the small data problem include data augmentation, knowledge transfer, and pretraining on a larger labeled dataset such as AudioSet [1] or a self-supervised learning approach with substantial amounts of unlabeled data [2]. However, further study is necessary to determine which of these techniques or technique combinations provide the optimal solution. This paper analyzes the effects of knowledge transfer, data augmentation, and pretraining on a popular small dataset, ESC-50 [3]. We train two models, Xception and Xception-small, and conduct an ablation study of different combinations of the mentioned techniques and record the ESC-50 accuracy.

Our Xception model with knowledge transfer from ImageNet weights, pretraining on AudioSet, with on-the-fly data augmentation reaches SOTA while our Xception-small model reaches near SOTA despite being much smaller, providing an option for low-compute scenarios.

2. PREVIOUS WORK

The development of Pretrained Audio Neural Networks (PANNS) has presented impact of different architectures, augmentation, and other training and dataset options and how it impacts AudioSet tagging [4]. The PANNS study also evaluated on ESC-50 and reached high quality. Other studies of data augmentation for AED and other audio-tagging tasks include [5, 6, 7]. Some tools have become standard for augmenting data and creating synthetic data for AED, most notably [8].

The Audio Spectrogram Transformer (AST) paper applies ImageNet weights to a ViT-type architecture for audio [9]. Their evaluation on the ESC-50 dataset is currently SOTA, and they found applying ImageNet weights significantly improved their results. However, they implied this improvement might only be available for very large models. Initializing models with ImageNet-trained weights has also been explored in [10, 11, 12] for CNN-based models.

There have been several other models trained on large datasets, AudioSet or other large dataset, such as [13, 14]. Recently, a wav2vec approach with a large billion-parameter model was trained with self-supervised learning (SSL) on unlabeled data that has been used as a general embedding model for AED and several other speech tasks [2]. A near SOTA of ESC-50 was reached in [15] by using a sequential self-teaching model which provided a significant SOTA lead in the ESC-50 leaderboard at the time of its publication.

Xception models have also been used for AED, such as in [7, 16], as the model provides high performance with relatively few parameters compared to other top-performing models.

Our paper builds on the work described above but focuses on an ablation study of how each component—data augmentation, knowledge transfer with ImageNet, pretraining with AudioSet—impacts performance, with a specific focus on small datasets (ESC-50) and relatively small models (Xception and Xception-small).

Corresponding author: daniel.tompkins@microsoft.com

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

3. DATASETS

We focus our evaluation metrics on the ESC-50 dataset [3] because it has a comprehensive and current leaderboard and because it has relatively few examples per class. The ESC-50 dataset is structured in 5 folds, so we trained and evaluated our models 5 times, rotating the withheld fold for evaluation. We average the evaluation folds to obtain our accuracy scores. ESC-50 contains 50 balanced classes, 8 files of each class per fold for a total of 2000 files. Each example is exactly five seconds long and contains only one class.

For pretraining, we use AudioSet's unbalanced dataset [1], which contains over 2 million labeled examples covering 527 classes, which are structured as a hierarchical ontology. Most examples are 10 seconds in length and contain multiple labels. We use AudioSet's evaluation dataset to evaluate the model after each epoch of training on the unbalanced dataset.

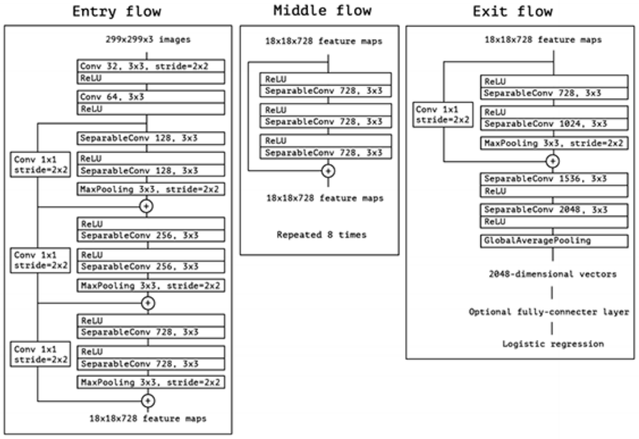


Fig. 1. The Xception architecture, as described in [17]. Our Xception-small model does not repeat the middle flow.

4. KNOWLEDGE TRANSFER AND AUGMENTATION

In the absence of a large labeled AED dataset, knowledge transfer from another domain and data augmentation are common ways of approaching the problem small datasets such as ESC-50 pose to deep learning models.

4.1. ImageNet Weight Initialization

The Xception model has been trained with ImageNet [18], which is a large dataset of millions of images. Several previous studies [10, 11, 12, 9] have applied ImageNet weights to models as a replacement for weight initialization. Despite the difference between image data and audio data, using ImageNet weights is often shown to give a performance increase. To apply ImageNet weights to the Xception model, we used a method similar to [9], which includes averaging the first three channels of the model input into one channel (RGB image channels to one spectrogram channel).

We attempted to apply ImageNet weights to the Xception-small model by starting with the full Xception model with ImageNet weights and deleting the middle flow repetitions, but we did not see any boost in performance. Further study is needed on methods of altering a pretrained model while keeping the benefits.

4.2. On-The-Fly Audio Data Augmentation

We developed an on-the-fly data augmentation pipeline that gives each incoming audio example a probability of being altered. Alterations include varying pitch, volume, and speed, adding noise, randomly zeroing frames, bandpass filters, resampling, mixup [19, 7], and negative data augmentation (NDA) [20].

Unlike previous applications of mixup where the proportion of mix of two examples is reflected in the labels, we consider two mixed items to have positive labels for all classes mixed together. For example, mixing an example of speech with an example of an engine at a ratio of .6/.4 will result in both classes having a label of 1 rather than reflecting mix ratio.

We made NDA audio-specific rather than directly using image-based NDA. Flipped spectrograms, which are often label-preserving in images, are converted to negative examples. However, we create negative examples from: shuffled spectrograms (where the frequency bins or the time bins are shuffled randomly), jigsaw (2D areas are shuffled randomly), and cutout (randomly zeroing a portion of the spectrogram). The motivation of NDA is to encourage the model to learn global features rather than very local ones.

5. MODELS, FEATURES, AND TRAINING PIPELINE

For our experiments, we use the Xception model architecture as described in [17], which is a depthwise-separable CNN with residual connections. The full model can be seen reproduced from [17] in Figure 1. We also introduce an Xception-small model (Xception-s) that is identical to Xception but does not repeat the middle flow convolution layers. This change reduces the model size from 21 million parameters to 8 million. A comparison of the Xception models' parameter sizes and other top-performing models on ESC-50 is shown in Table 1.

The input features for all Xception and Xception-small models are a single-channel 80-bin log-mel filterbank. All audio was resampled to 16kHz. The two-dimensional shape of filterbanks provide some analogous properties to image recognition. The final output layer size is determined by the number of classes in each dataset. The final layer is 50 for ESC-50 and 527 for AudioSet.

For training on the ESC-50 dataset, we use adam optimizer with an initial learning rate of 0.001 with a decay of 0.8 every epoch for 25 epochs. We use cross-entropy loss. The training process is repeated 5 times while rotating the training folds and evaluation fold as required by the ESC-50 leaderboard criterion. Data augmentation as described in Section 4.2 is applied but without mixup or NDA due to the single-label structure of the dataset.

For pretraining on AudioSet, the same configuration of optimizer and learning rate is used. However, we use binary cross-entropy loss because AudioSet is a multi-label dataset where many examples include more than one positive labels. Data augmentation includes mixup and NDA. The AudioSet evaluation loss and mAP score is tracked, and training is stopped when the evaluation loss and mAP stagnates and the learning rate decay has passed below 1e-6. The epoch with the lowest evaluation loss is selected as the model to test. We also apply class weights to the criterion due to the strong imbalance in the AudioSet classes.

To convert an AudioSet model for fine-tuning on ESC-50, we remove the prediction layer from the AudioSet-trained model and apply a randomly-initialized linear layer of size 50. During fine-tuning, all parameters are updated rather than only updating the final layer(s). We have found this provides a higher accuracy on ESC-50 than freezing internal layers.

Model	N. Params (mil)
BigSSL-XXL [2]	1000
AST [9]	87
PANNs [4]	81
Xception (ours)	21
Xception-s (ours)	8

Table 1. Comparison of model sizes and types.

6. RESULTS

The full results of the Xception models with various configurations of knowledge transfer, pretraining with AudioSet, and data augmentation can be found in Table 2 along with other recent top-performing models. Visualizations of the average validation accuracy (averaged 5 folds per ESC-50 policy) for the first 25 epochs are shown in Figure 2. The average validation losses for the first 25 epochs are shown in Figure 3.

6.1. Xception

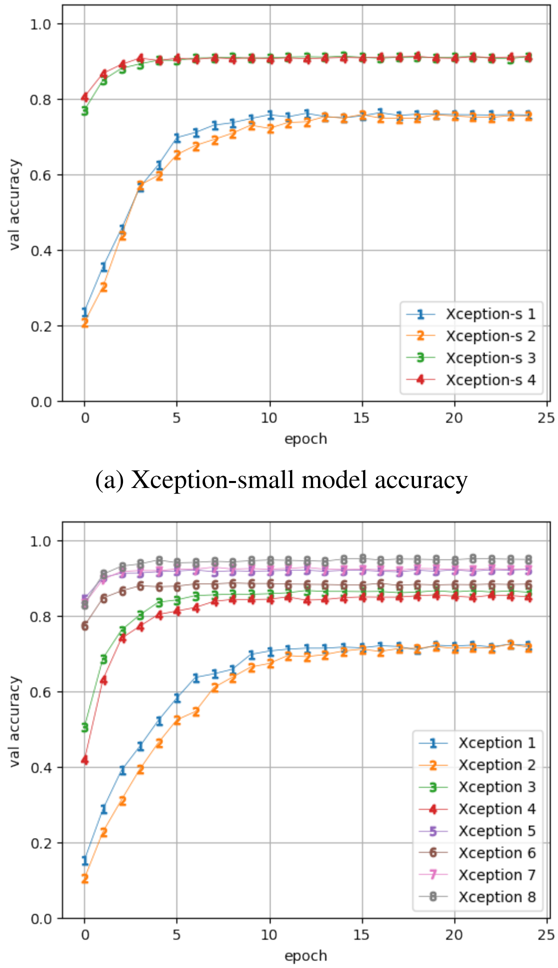
Xception 8, which was initialized with ImageNet weights, pretrained with AudioSet, which was had on-the-fly augmentation, reaches SOTA. However, most of the models pretrained with AudioSet were close to SOTA such that small alterations to the ESC-50 fine-tuning pipeline could change the order of ranking. When the Xception model was pretrained with AudioSet, there were no performance gains from augmenting the ESC-50 data during fine-tuning. There is clear performance differences between models trained from scratch (Xception 1, 2), which scored the lowest, models initialized with ImageNet weights only (Xception 3, 4), which performed over 10 points higher and the remaining models pretrained with AudioSet, which achieve the highest scores. Initializing with ImageNet weights followed by training with AudioSet (Xception 7, 8) yields a slightly higher accuracy than only pretraining on AudioSet. Figure 2 clearly shows this separation not only with accuracy scores but how quickly each model reaches optimal accuracy. These results are also reflected in the running validation losses in Figure 3.

Model	Pretr.	Aug.	ESC-50 acc.
AST (SOTA)	IN + AS	AS + ESC	95.6
PANNs	AS	AS	94.7
BigSSL-XXL	UL	-	90.9
Xception-s 1	-	-	76.3
Xception-s 2	-	ESC	77.4
Xception-s 3	AS	-	92.0
Xception-s 4	AS	AS	94.2
Xception 1	-	-	72.5
Xception 2	-	ESC	72.5
Xception 3	IN	-	86.8
Xception 4	IN	ESC	86.1
Xception 5	AS	-	92.7
Xception 6	AS	AS	89.9
Xception 7	IN + AS	-	95.6
Xception 8	**IN + AS**	**AS**	**95.8**

Table 2. Accuracy scores from the ESC-50 dataset compared with other top-scoring models. Comparing pretraining options: ImageNet weights (IN), AudioSet (AS); and augmentation options: AS, ESC-50 (ESC), and Unlabeled (UL). The Xception model pretrained with IN and AS with data augmentation during AS pretraining achieves SOTA results.

6.2. Xception-small

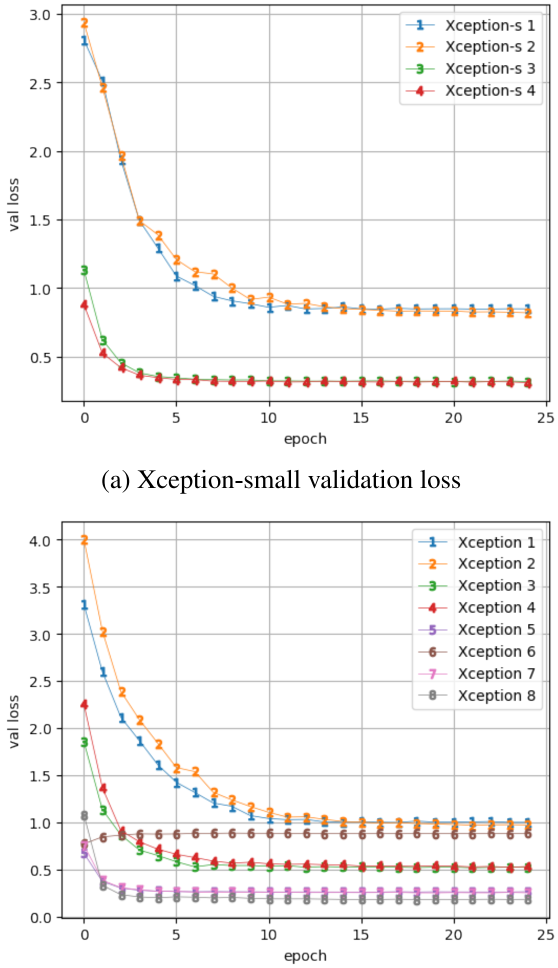
Similar to the Xception models, the Xception-small models that were pretrained with AudioSet scored much higher than those trained from scratch. The best Xception-small model scores near the other top-scoring models despite having a tenth or fewer parameters. While the full Xception model did not show any impact with data augmentation, the Xception-small models showed modest improvement in performance by including data augmentation. When comparing training on ESC-50 from scratch, Xception-small performs better than the full Xception model, likely because of it has fewer parameters and is less likely to over-fit.



(a) Xception-small model accuracy

(b) Xception model accuracy

Fig. 2. Average validation accuracy over 5 folds of ESC-50 over 25 epochs for Xception and Xception-small models.



(a) Xception-small validation loss

(b) Xception validation loss

Fig. 3. Average validation loss over 5 folds of ESC-50 over 25 epochs for Xception and Xception-small models.

7. CONCLUSION AND FURTHER WORK

We have shown that Xception models can reach SOTA on ESC-50 when they are pretrained with AudioSet and are much smaller than other top-performing models. In the absence of a larger in-domain dataset, applying knowledge transfer from an outside domain such as image recognition gives a better result than training on a small dataset directly. Data augmentation only benefits the Xception-small model, and only by a small amount.

Future work will include evaluating on other datasets, especially multi-label datasets. We would also like to analyze the individual components of the data augmentation pipeline to find which, if any, are most beneficial. Furthermore, we would like to attempt knowledge transfer from other domains than image recognition.

Further work should be done in reducing the size of models, such as converting Xception to Xception-small, while retaining the benefit of pretrained ImageNet weights. Deleting the layers directly removed any benefit of the ImageNet models.

Additional study of model performance relative to parameter size and runtime on low-compute devices would be beneficial, especially with the increase of running AED models on low-compute devices.

8. REFERENCES

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

[2] Yu Zhang, Daniel S Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, et al., "Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2109.13226*, 2021.

[3] Karol J. Piczak, "ESC: Dataset for Environmental Sound Classification," in *Proceedings of the 23rd Annual ACM Conference on Multimedia*. 2015, pp. 1015–1018, ACM Press.

[4] 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*, vol. 28, pp. 2880–2894, 2020.

[5] Shengyun Wei, Kele Xu, Dezhi Wang, Feifan Liao, Huaimin Wang, and Qiuqiang Kong, "Sample mixed-based data augmentation for domestic audio tagging," *arXiv preprint arXiv:1808.03883*, 2018.

[6] Naoya Takahashi, Michael Gygli, Beat Pfister, and Luc Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," *arXiv preprint arXiv:1604.07160*, 2016.

[7] Kele Xu, Dawei Feng, Haibo Mi, Boqing Zhu, Dezhi Wang, Lilun Zhang, Hengxing Cai, and Shuwen Liu, "Mixup-based acoustic scene classification using multi-channel convolutional neural network," in *Pacific Rim conference on multimedia*. Springer, 2018, pp. 14–23.

[8] Justin Salamon, Duncan MacConnell, Mark Cartwright, Peter Li, and Juan Pablo Bello, "Scaper: A library for soundscape synthesis and augmentation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 344–348.

[9] Yuan Gong, Yu-An Chung, and James Glass, "Ast: Audio spectrogram transformer," *arXiv preprint arXiv:2104.01778*, 2021.

[10] Kamalesh Palanisamy, Dipika Singhania, and Angela Yao, "Rethinking cnn models for audio classification," *arXiv preprint arXiv:2007.11154*, 2020.

[11] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel, "Esresnet: Environmental sound classification based on visual domain models," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 4933–4940.

[12] Grzegorz Gwardys and Daniel Michał Grzywczak, "Deep image features in music information retrieval," *International Journal of Electronics and Telecommunications*, vol. 60, no. 4, pp. 321–326, 2014.

[13] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson, "Cnn architectures for large-scale audio classification," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[15] Anurag Kumar and Vamsi Ithapu, "A sequential self teaching approach for improving generalization in sound event recognition," in *International Conference on Machine Learning*. PMLR, 2020, pp. 5447–5457.

[16] Tomas Gajarsky and Hendrik Purwins, "An xception residual recurrent neural network for audio event detection and tagging," in *15th International Sound & Music Computing Conference*. Sound and Music Computing Network, 2018, pp. 210–216.

[17] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[19] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, "mixup: Beyond empirical risk minimization," *arXiv preprint arXiv:1710.09412*, 2017.

[20] Abhishek Sinha, Kumar Ayush, Jiaming Song, Burak Uzkent, Hongxia Jin, and Stefano Ermon, "Negative data augmentation," *arXiv preprint arXiv:2102.05113*, 2021.