AUDIO AUGMENTATIONS FOR SEMI-SUPERVISED LEARNING WITH FIXMATCH

Sascha Grollmisch
Fraunhofer IDMT
goh@idmt.fraunhofer.de

Estefanía Cano
Songquito UG

Jakob Abeßer
Fraunhofer IDMT

ABSTRACT

FixMatch, a semi-supervised learning method proposed for image classification, includes unlabeled data instances into the training procedure by predicting labels for differently augmented versions of the unlabeled data. In our previous work, we adapted FixMatch to audio classification by applying image augmentations to spectral representations of the audio signal. While this approach matched the performance of the supervised baseline with only a fraction of the training data, the performance of audio-specific augmentation techniques, and their effect on the FixMatch approach was not evaluated. In this work, we replace all image-based augmentation techniques with audio-specific ones and keep the feature extraction unchanged. The audio-specific approach improved upon the supervised baseline which confirms the effectiveness of the FixMatch approach for semi-supervised learning even with a completely different set of augmentations. However, the image-based approach outperforms the audio-based approach on the three audio classification tasks evaluated.

1. INTRODUCTION

Semi-supervised learning (SSL) is a widely used learning paradigm that aims to reduce the amount of labeled data required to train a model by including unlabeled data into the training procedure. While labeling data often entails an expensive and time-consuming process, unlabeled data is often abundant and easier to collect. In recent years, SSL approaches such as MixMatch and FixMatch have produced competitive results in various image classification tasks [1–3].

FixMatch (FM) relies on the augmentation of unlabeled data for which pseudo-labels are generated iteratively during training time. To achieve this, the unlabeled data in each batch is first weakly augmented and passed through the model. Only those training examples above a confidence threshold of 0.95 are considered and binarized to pseudo-labels. These unlabeled examples are also strongly augmented and passed through the model using the pseudo-labels as ground-truth. The categorical cross-entropy (CCE) is calculated for the unlabeled data, and added to the CCE of the labeled data.

In our previous work [4], we adapted FM for audio classification and evaluated the approach on three different tasks: (1) instrument family recognition (10 classes) using the *NSynth* dataset [5], (2) acoustic scene classification (15 classes) using the *TUT2017* dataset [6, 7], and (3) industrial sound classification (3 classes) with the IDMT_ISA_METAL_BALLS (*MB*) dataset [8]. The Resnet-based `CNN420` architecture [4] was used, with log Mel spectrograms as input. 13 common image augmentation techniques such as translation in x / y direction, additive Gaussian noise, and sharpening were considered. Our results demonstrated that the selection of the weak augmentation is crucial for performance. With this in mind, we proposed a simple yet effective method to chose the weak augmentation method: the system is first trained on the available non-augmented labeled data and evaluated on augmented versions of the same data. The method with the smallest loss in accuracy compared to the non-augmented data is used as weak augmentation. As strong augmentations, two augmentation methods are randomly selected and each applied with a probability of $p = 0.5$ and random magnitude. This is followed by SpecAugment [9] which replaces the masked time-frequency regions with zeros. This adapted FM version also reached the accuracy of the supervised systems that had all labels available with less than 5% of labeled data for two of the three datasets. For more details we refer the reader to the initial publication [4].

Our work demonstrated the efficacy of FM for audio classification tasks; however, one important question remained unanswered: how do augmentations of the raw audio signal such as pitch shifting and time stretching affect the performance of FM? In this work we replace all image-based augmentations with audio-based ones and evaluate the performance of FM by keeping the rest of the process unchanged. [1]

2. EXPERIMENTAL PROCEDURE

In our previous work, the log Mel spectrograms were extracted using the Librosa python library [10] and used as input to the `CNN420`. To enable runtime audio aug-

[1] This work was supported by the German Research Foundation (AB 675/2-2).

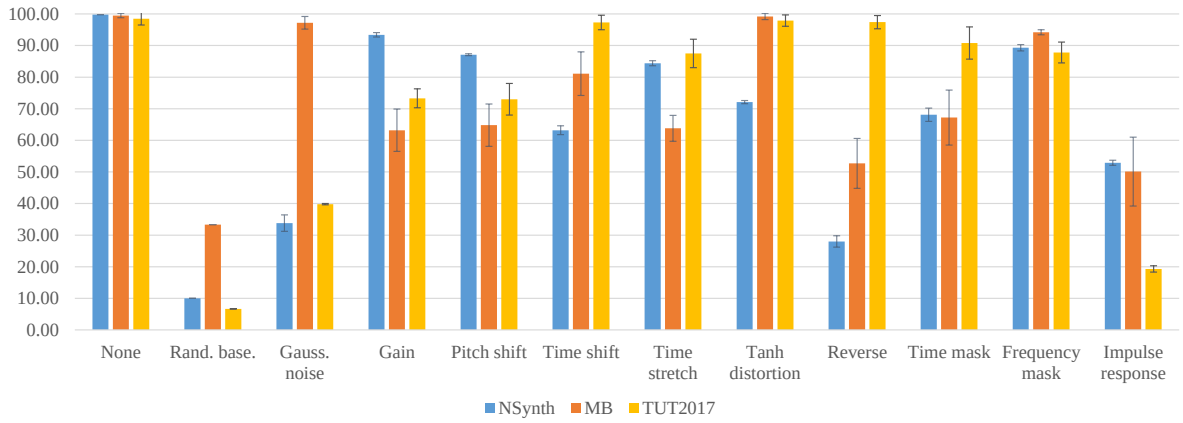

Figure 1. Mean file-wise accuracy and standard deviation (error bars) for all audio augmentation methods on 5% of the labeled training data of each dataset.

mentation and faster processing in this version of FM, we included the feature extraction to the `CNN420` using the Kapre python library [11]. Kapre aims at replicating Librosa extraction methods using Tensorflow operations. The extraction hyperparameters are the same, except for *TUT2017* where Per-channel energy normalization (PCEN) [12] was replaced with logarithmic magnitude compression since PCEN is not available in Kapre. The image-based results for *TUT2017* were repeated with the adjusted feature extraction and file-wise accuracy decreased by 3 percentage points. The augmentations are performed for each batch with Audiomentations python library [13]. Similar to the image-based FM, the single weakest augmentation method is selected on the training set and the two strong augmentations are randomly picked with random magnitude. As shown in Figure 1, the 10 applied audio augmentations contain pitch shifting, time stretching, reverb (with impulse responses from the Echothief library [14]), among others. SpecAugment is replaced by a combination of time and frequency masking.

To measure the impact of audio augmentations, 5 % of the training data are used as labeled examples and all training data as unlabeled data, similar to our previous work. Every experiment is repeated three times to account for randomness in data selection and network initialization.

3. RESULTS

Similar to our previous results using image augmentations, the accuracy on the already seen non-augmented training data is close to 100 %. Each audio augmentation method has a different impact on the performance for each dataset, as can be seen in Figure 1. The impact of the audio augmentations on the performance ranges from small (Tanh distortion) to big (impulse response), which indicates that the selected methods create sufficient data variability to be used for audio-based FM. As weak augmentation, gain was selected for *NSynth* and Tanh distortion for *MB* and *TUT2017*. The validity of the audio augmentation pipeline was confirmed with *MB*, where the supervised baseline on all training data with random audio augmentations achieved 100 % on the test data.

Table 1. Mean accuracy and standard deviation in % for image- and audio-based augmentations with 5 % of the training data for the supervised baseline (Sup.) and FM.

Dataset	Sup. image	Sup. audio	FM image	FM audio
NSynth	70.4±0.6	64.6±2.3	75.8±0.5	71.4±0.6
MB	87.1±2.4	87.4±4.4	99.8±0.3	97.7±2.3
TUT2017	48.0±2.5	36.2±2.0	66.9±1.9	50.1±1.1

The results with 5 % of the training data for the supervised baseline and FM with image and audio augmentations are shown in Table 1. While the audio-based supervised approach showed comparable results to FM on *MB*, the performance is considerably lower for *TUT2017* and *NSynth*. These results demonstrate that a simple replacement of the image augmentation methods does not necessarily improve the accuracy for supervised audio classification. The same can be observed for FM with audio augmentations: The proposed system performed slightly worse than its image-based counterpart on *NSynth* and *MB*, and underperforms for *TUT2017* by a large margin.

It must be noted, that this was investigated for 5 % of the training data and results might differ with more labeled examples. Furthermore, fewer audio than image augmentations were included. Adding more augmentations methods might add more variability and change the results. As a side note, training with raw audio data also has a negative impact on training time and memory usage. Mel spectrograms are a compressed representation of audio data: On *NSynth* for example, 61 x 64 image patches are processed while raw audio data takes 64k input values (16 kHz x 4 sec) for each audio clip. On the upside, training with audio-based FM always improved the corresponding supervised baseline. FM adds useful information to the model from the unlabeled data and can also be used with a different set of augmentations.

These results show that more optimization is needed in terms of the selected audio augmentations, their parameters, their (random) combination for supervised and semi-supervised learning, and finally the neural network architecture itself. CNNs are originally designed for images which might be a reason why image-based augmentations were more effective in the conducted experiments.

4. REFERENCES

[1] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "MixMatch: A holistic approach to Semi-supervised Learning," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2019, pp. 5049–5059.

[2] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "FixMatch: Simplifying Semi-Supervised Learning with consistency and confidence," in *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, Online, 2020, pp. 596–608.

[3] M. Assran, M. Caron, I. Misra, P. Bojanowski, A. Joulin, N. Ballas, and M. Rabbat, "Semi-Supervised Learning of visual features by non-parametrically predicting view assignments with support samples," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, Online, 2021, pp. 8423–8432.

[4] S. Grollmisch and E. Cano, "Improving Semi-Supervised Learning for audio classification with FixMatch," *Electronics*, vol. 10, no. 15, 2021.

[5] J. Engel, C. Resnick, A. Roberts, S. Dieleman, M. Norouzi, D. Eck, and K. Simonyan, "Neural audio synthesis of musical notes with WaveNet autoencoders," in *Proceedings of the International Conference on Machine Learning (ICML)*, Sydney, Australia, 2017, pp. 1068–1077.

[6] A. Mesaros, T. Heittola, and T. Virtanen, "TUT Acoustic scenes 2017, Development dataset," Website https://zenodo.org/record/400515, last accessed 23/08/2022, 2017.

[7] ——, "TUT Acoustic scenes 2017, Evaluation dataset," Website https://zenodo.org/record/1040168, last accessed 23/08/2022, 2017.

[8] S. Grollmisch, J. Abeßer, J. Liebetrau, and H. Lukashevich, "Sounding industry: Challenges and datasets for Industrial Sound Analysis," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, A Coruña, Spain, 2019, pp. 1–5.

[9] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple augmentation method for automatic speech recognition," in *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria, 2019, pp. 2613–2617.

[10] B. McFee *et al.*, "librosa/librosa: 0.8.0," Website https://zenodo.org/record/3955228, last accessed 23/08/2022, 2020.

[11] K. Choi, D. Joo, and J. Kim, "Kapre: On-gpu audio preprocessing layers for a quick implementation of deep neural network models with Keras," in *Machine Learning for Music Discovery Workshop at International Conference on Machine Learning (ICML)*, Sydney, Australia, 2017.

[12] Y. Wang, P. Getreuer, T. Hughes, R. F. Lyon, and R. A. Saurous, "Trainable frontend for robust and far-field keyword spotting," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 5670–5674.

[13] I. Jordal, *et al.*, "iver56/audiomentations: v0.25.1," Website https://zenodo.org/record/6645998, last accessed 23/08/2022, 2022.

[14] C. Warren, "Echothief impulse response library," Website http://www.echothief.com/, last accessed 30/08/2022, 2020.