Self-supervised Learning for Music Genre Classification using Image Augmentation

TAEMI KIM, Korea Advanced Institute of Science and Technology, South Korea
SUNGRAE HONG, Korea Advanced Institute of Science and Technology, South Korea
SOL LEE, Korea Advanced Institute of Science and Technology, South Korea

Self-supervised learning is in the spotlight as a way to solve the problem of data-hunger of deep learning. Especially, in the music domain, self-supervised learning is attracting solution as a way to solve the high-cost and high-labor problem of labeling numerous songs. However, self-supervised learning has a collapsing problem that outputs a constant solution. To solve this problem, a method has been devised to give the network a variety of augmentation that is difficult to match. Until now, a method of applying augmentation to the time axis or frequency axis of audio has been used, but we propose a method of using image augmentation by converting audio into a spectrogram. This method is easier, and has advantages in performance. The source code of our study is in https://github.com/HongSungRae/KSE527.

Additional Key Words and Phrases: Self-Supervised Learning, Augmentation, Music

ACM Reference Format:
TaeMi Kim, SungRae Hong, and Sol Lee. 2022. Self-supervised Learning for Music Genre Classification using Image Augmentation. In . ACM, New York, NY, USA, 7 pages. https://doi.org/123.456

1 INTRODUCTION

Deep learning has achieved great success in many fields, such as computer vision and natural language processing. However, it has the disadvantage of heavy dependence on manual labels, which requires a lot of costs, so only a small number of companies led to the large and private dataset. For example, Google made JFT-300M which contains 300 million labeled images and META made 1G-1B-Targeted which consists of 940 million public images with 1.5K hashtags matching with 1000 ImageNet1K synsets. To make cost-effective learning and prevent data unfairness problems, some researchers have begun to pay attention to learning using data without labels. Self-supervised learning is in some sense a type of unsupervised learning, it learns useful representation from large amounts of label-free data. SimCLR[2] is a methodology using augmentation and contrastive loss. It learns representation by strong augmentation that one input is augmented differently, and all other samples are considered negative samples. In the case of BYOL[6], collapsing does not occurred without negative samples. The online network is learned to follow up the output of the target network. In the case of Dino[1], the concept of Teacher and Student is introduced to update the network. It solved the collapsing problem through the centering process without the negative samples. In the case of SimSiam[3], the Siam network architecture was used. The collapsing problem was solved by using a stop-gradient to one side of the network. However, since those studies have focused on the self-supervised learning methodology itself, it has been conducted only in image domains where data is easy to augmented and processed. On the music domain, thousands

Manuscript submitted to ACM

of new songs and video audio are uploaded everyday. To make these musics into the dataset, requires tagging the labels. Music labels include such things as genres, instruments, mood, and so on, including the tagger's judgment, and the possibility of mislabeling is quite high[4]. Plus, it also needs a lot of time and money too. For this reason, there have been Attempts to apply self-supervised learning to several downstream tasks in the music domain.[7][8][10][13] Although data augmentation is an important process to prevent the collapsing problem, which is a major problem in self-supervised learning. But previous studies are used only a limited augmentation method. So, more augmentation needs to be applied. Therefore, we study the way that applying self-supervised learning well to downstream tasks in the music domain. After performing self-supervised learning for the music genre classification tasks by preprocessing audio data into images and applying diverse image augmentation methods. Finally, we identify the possibility of applying more diverse image augmentation methodologies to audio data. We adopted Simsiam [4] as the baseline for self-supervised learning. Resnet50, Resnet101 and Resnet152 are used as backborns of the Simsiam architecture. Here we propose to utilize two methods: (i) Audio Effect Augmentation (ii) Spectrogram Image Augmentation. There are two contributions to this study. (i) A self-supervised learning structure without a negative sample is applied to the music domain, (ii) To prevent collapsing, It is the first attempt to use the image augmentation technique for audio data in the self-supervised application of the music domain.

2 RELATED WORK

2.1 Self-Supervised Learning

Self-supervised learning is a learning method in which a model is trained to learn the data representations based on a large amount of unlabeled data and transferred to other main tasks such as classification. This approach typically sets the pretext task for learning data features in a heuristic way and use the model trained with the pretext task for a downstream task. Examples of the pretext task include relative image patch prediction, image rotation angle prediction or image color jittering. However, the pretext task method has the problem of relying on empirical definement. As a result, most of the recent self-supervised learning approaches are based on the Siamese network.

The Siamese network is a general architecture for comparing the two data inputs. This network typically consists of an encoder that combines the CNN backbone and a projector and is trained by increasing the similarity between the output vectors from the two augmented views of input data. However, Siamese networks generally have a collapsing problem. Collapsing refers to a phenomenon in which all networks' outputs converge to a constant, so loss converges without properly learning representations. To prevent the collapsing, SimCLR[2] uses contrastive loss with negative samples in a batch of large size. BYOL[6] is also based on a similar architecture but use momentum encoder without negative samples, while Simsiam[3] showed that Siamese networks can perform well enough with only the Siamese structure itself and stop-gradient without contrastive learning or momentum encoder.

2.2 Self-Supervised Learning for Music Classification

Self-supervised learning for music has been mainly based on pretext task and contrastive method. Jiyoung et al.[7] trained Siamese DCNN using artist labels and applied transfer-learning to music classification and retrieval tasks. Not only artists but also meta information originally annotated in music such as album and track information has been used to pre-train networks for representation learning. In addition, Wu, Ho Hsiang et al.[10] combined multiple pretext tasks using re-weighted loss for pretraining the network to classify music. On the other hand, Janne et al.[8] classified music genres with musical representations learned from SimCLR and Zhao, Han et al.[13] combined the Swin Transformer with a contrastive self supervised learning method, MoCo using negative samples.

Most of previous self-supervied music classification studies have pretrained the model using very large size of dataset and audio augmentation is applied only in the form of audio data. However, audio augmentation is more limited than image augmentation and has not been studied in various ways. Thus, we focus on self-supervised music representation learning based on spectrogram images from audio using Simsiam architecture and image augmentation operations. We prove that feature learning in image form performs better than learning in audio form in a self-supervised way.

3 METHOD

3.1 Architecture

Our architecture named Siamusic is based on the SimSiam. x_1 and x_2, two augmented views of data x, are entered into the model. x can be audio or image data. The overall structure of the model is shown in Figure 1. The two views x_1 and x_2 pass through an encoder consisting of a backbone and a projector. CNN-based models such as ResNet, or Transformer's encoder structure can be used for the backbone. The projector is an MLP consisting of 3 layers. Encoder shares weight between the x_1 and x_2. Predictor is also an MLP structure, which predicts output using the vector transformed from x_1 through the encoder to match the x_2. For a projected vector of x_2 that has not passed through the predictor, a stop-gradient operation is applied. The stop-gradient which does not propagate gradient is significant to prevent the collapsing problem of the Siamese network. The goal of Siamusic is to minimize the negative cosine similarity between the two outputs from x_1 and x_2. We define the loss to train the model as follows. h is the predictor and sg is the stop-gradient operation.

$$S(x_1, x_2) = -\frac{x_1}{\|x_1\|_2} \cdot \frac{x_2}{\|x_2\|_2} \tag{1}$$

$$Loss = \frac{S(sg(x_1), h(x_2)) + S(h(x_1), sg(x_2))}{2} \tag{2}$$

3.2 Augmentation

Type	Augmentation	Lambda
Audio	Mixup	0.6
	Pitch Shift	0.6
	Time Stretching	0.6
Image	CutMix[11]	0.6
	Mixup[12]	0.6
	Mask	0.4
	Horizontal Flip	0.8
	Vertical Flip	0.8
	CutOut[5]	0.8

Table 1. Augmentations

The details of the audio augmentation and image augmentation are proposed in Table 1. Here, lambda refers to the probability that each augmentation is applied.

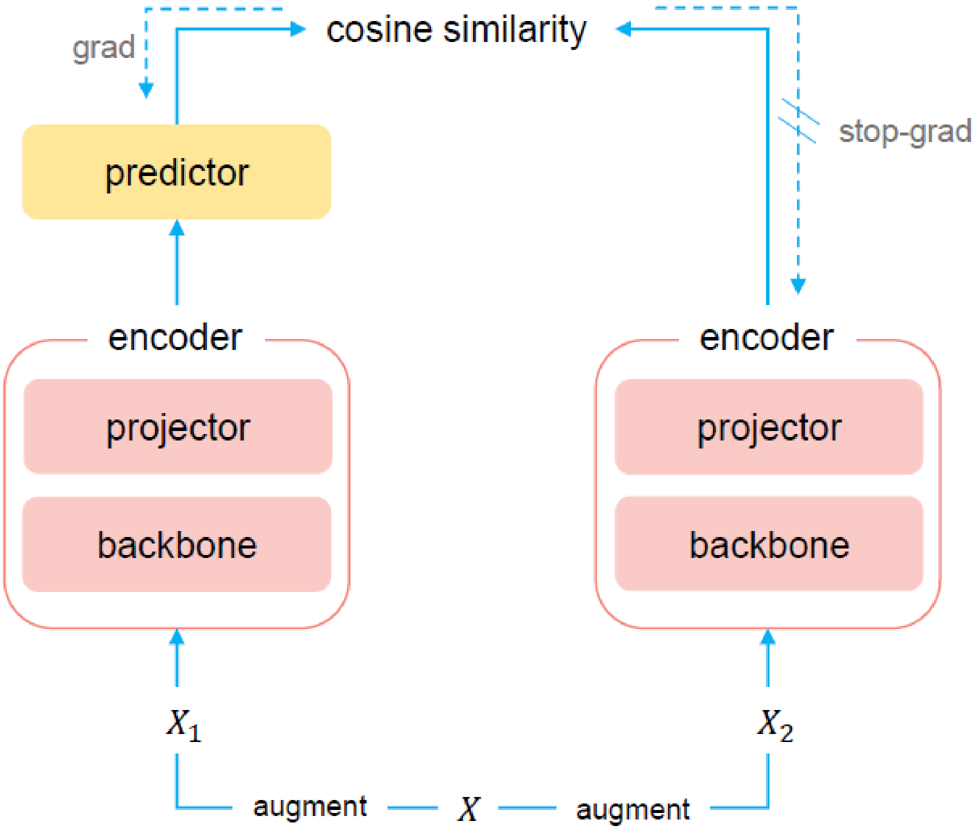

Fig. 1. SimSiam Architecture

The results of augmentation are shown in Figure 2. The top image is the original audio. The middle image is the result of applying image augmentation, and the last image is the result of applying audio augmentation.

3.3 Implementation details

In the experiment, four backbone networks were used: *ResNet50, ResNet101* and *ResNet152.* And two augmentations were used: *Audio Effect* and *Image Augmentation for Spectrogram.*

Adam optimizer was used with weight decay 0.00001. Except for some experiments, pre-training was trained 100epochs and fine tuning was trained 50epochs. All operations were computed on two GPUs.

4 EXPERIMENT

4.1 Experiment Settings

We planned an experiment to verify image aggregation for audio that we propose. First, we compare the performance by differentiating only the augmentation under the same conditions. Second, various backbone networks also compared

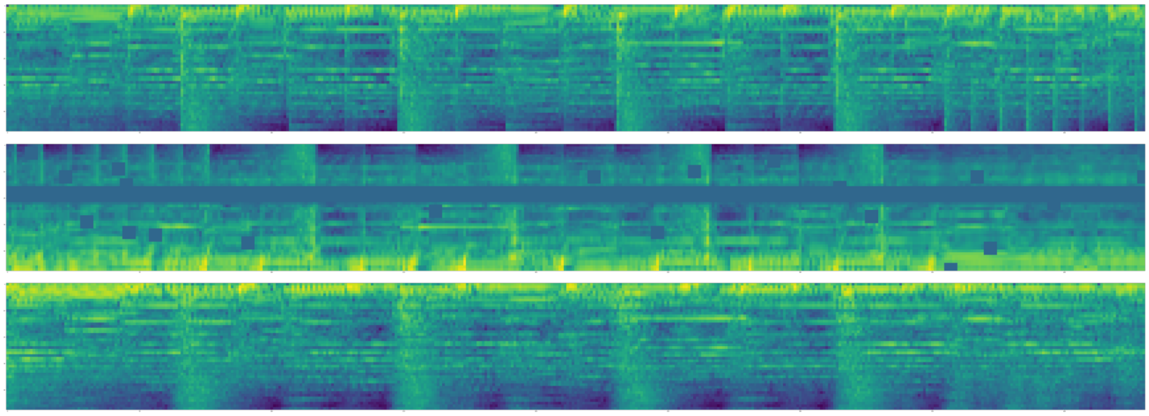

Fig. 2. An original audio and augmented audio

to compare proposed method with the traditional audio augmentation. The pre-training dataset used in all experiments are FMA_Medium or FMA_SMALL, and the dataset used for fine-tuning is FMA_SMALL only.

4.2 Audio Augmentation versus Image Augmentation

We compared the effects of audio augmentation and image augmentation with ResNet50 as a backbone network. The results of the experiment are shown in the Table 2. Image augmentation was performed better. Remind that [9] showed 42.67% of accuracy with supervised learning. We performed 0.354 of accuracy with unsupervised learning.

Backbone	Augmentation	Pre-training Dataset	Transfer Learning Dataset	Accuracy	Recall	Precision	F1
ResNet50	Audio(Baseline)	FMA_medium	FMA_small	0.324	0.325	0.314	0.319
	Image(Ours)	FMA_medium	FMA_small	**0.354**	**0.359**	**0.361**	**0.359**
	Audio(Baseline)	FMA_small	FMA_small	0.253	0.255	0.229	0.241
	Image(Ours)	FMA_small	FMA_small	**0.262**	**0.267**	**0.279**	**0.273**

Table 2. Audio and Image Augmentation Performance

4.3 Image Augmentation For Various Backbone

We verified whether image augmentation can perform better than audio aggregation even if the backbone network changes. All, but one case of ResNet101, had better performance when the backbones used the image augmentation that we proposed. In the Table 3, each bold word represents the best performance in each metric, and the underbar represents the better performance compared to the same backbone and fine-tuning dataset.

5 DISCUSSION

In this study, we derived better performance to self-supervised learning of the music genre classification task with Spectrogram image augmentation for ResNet50, ResNet101. It clearly shows that self-supervised learning can be performed sufficiently well if an image augmentation method is applied to audio data. Due to the lack of computing

Backbone	Augmentation	Pre-training Dataset	Transfer Learning Dataset	Accuracy	Recall	Precision	F1
ResNet50	Audio	FMA_medium	FMA_small	0.324	0.325	0.314	0.319
	Image(Ours)			0.354	0.359	**0.361**	**0.359**
	Audio	FMA_small		0.253	0.255	0.229	0.241
	Image(Ours)			0.262	0.267	0.279	0.273
ResNet101	**Audio**	FMA_medium		0.332	0.335	0.336	0.335
	Image(Ours)			0.327	0.331	0.331	0.330
	Audio	FMA_small		0.242	0.246	0.219	0.231
	Image(Ours)			0.270	0.278	0.272	0.274
ResNet152	Audio	FMA_medium		0.352	0.353	0.352	0.352
	Image(Ours)			**0.357**	**0.362**	0.356	0.358
	Audio	FMA_small		0.234	0.238	0.208	0.222
	Image(Ours)			0.251	0.258	0.222	0.239

Table 3. Various Backbones And Performance

resources, pre-train could not be performed using larger data. Other related studies have derived good classification performance to utilize very large data, such as the audio set of 2.1M size and 1 million songs in the pre-train stage. On the other hand, we could not derive better performance because we pre-trained with only as little data as 25K. Plus, we did not identify the effect of each image augmentation or their combination on improving classification performance. If these were done, it would have been a more practical study to apply self-supervised learning using audio data.

6 CONCLUSION

Through this study, we confirmed that self-supervised learning, which is being actively studied in the image classification task, can also be applied to audio domains. In addition, it was the first attempt to apply image augmentation methods to music data to prevent the collapsing problem of Siamese network architecture. We drive slightly better performance compared to using audio Augmentation. This fully shows that it is worth attempting an image augmentation for SSL learning using audio data. And since Image Augmentation is the most studied area, it suggests the need for related research in that it can be applied in various ways to audio in the future.

REFERENCES

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294* (2021).

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[3] Xinlei Chen and Kaiming He. 2021. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15750–15758.

[4] Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler. 2018. The effects of noisy labels on deep convolutional neural networks for music tagging. *IEEE Transactions on Emerging Topics in Computational Intelligence* 2, 2 (2018), 139–149.

[5] Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* (2017).

[6] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. 2020. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733* (2020).

[7] Jiyoung Park, Jongpil Lee, Jangyeon Park, Jung-Woo Ha, and Juhan Nam. 2017. Representation learning of music using artist labels. *arXiv preprint arXiv:1710.06648* (2017).

[8] Janne Spijkervet and John Ashley Burgoyne. 2021. Contrastive learning of musical representations. *arXiv preprint arXiv:2103.09410* (2021).

[9] Dat Thanh Tran, Nikolaos Passalis, Anastasios Tefas, Moncef Gabbouj, and Alexandros Iosifidis. 2022. Attention-based neural bag-of-features learning for sequence data. *IEEE Access* 10 (2022), 45542–45552.

[10] Ho-Hsiang Wu, Chieh-Chi Kao, Qingming Tang, Ming Sun, Brian McFee, Juan Pablo Bello, and Chao Wang. 2021. Multi-task self-supervised pre-training for music classification. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 556–560.

[11] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*. 6023–6032.

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

[13] Hang Zhao, Chen Zhang, Bilei Zhu, Zejun Ma, and Kejun Zhang. 2022. S3T: Self-Supervised Pre-training with Swin Transformer for Music Classification. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 606–610.