CHALMERS
UNIVERSITY OF TECHNOLOGY

VICMus: Variance-Invariance-Covariance Regularization for Music Representation Learning

Citation for the original published paper (version of record):
Löf, S., Hesse, C., Thomé, C. et al (2024). VICMus: Variance-Invariance-Covariance Regularization for Music Representation Learning. 2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops, ICASSPW 2024 - Proceedings: 475-479.
http://dx.doi.org/10.1109/ICASSPW62465.2024.10627508

VICMUS: VARIANCE-INVARIANCE-COVARIANCE REGULARIZATION FOR MUSIC REPRESENTATION LEARNING

Sebastian Löf [1,2] *Cody Hesse*[1] *Carl Thomé*[2] *Carlos Lordelo*[2] *Jens Ahrens*[1]

[1]Chalmers University of Technology [2]Epidemic Sound

ABSTRACT

Recent self-supervised learning methods often prevent informational collapse by implicitly regularizing information. Variance-Invariance-Covariance regularization (VICReg) was introduced as a non-contrastive loss function that explicitly maximizes information through regularization. While VICReg has garnered substantial interest in the field of computer vision, its application to the music domain remains unexplored. To address this gap, we introduce VICMus - VICReg for music representation learning. We pre-train VICMus on the Free Music Archive and achieve 36.3 mAP on MagnaTagaTune, outperforming Contrastive Learning of Musical Representations (CLMR), a recent contrastive method pre-trained on the ten times larger Million Song Dataset, which got 35.6 mAP. We evaluate VICMus on the Holistic Audio Representation Evaluation Suite (HARES)-music benchmark and achieve an average score of 51.7. Our results indicate that while VICMus may not yet achieve the performance of state-of-the-art self-supervised models, it offers a promising and computationally efficient avenue for music representation learning. Our code and models are available at https://github.com/SebastianLoef/VICMus.

Index Terms— Self-supervised learning, representation learning, regularization, music embeddings

1. INTRODUCTION

Self-supervised learning (SSL) has gained significant attention in the machine learning community, showing promise in multiple domains such as computer vision, natural language processing, audio and music recognition [1]. These methods do not rely on human annotations and excel at creating informative representations by training on large unlabeled datasets. The learned representations can later be used in various downstream tasks, as input features for simple models.

Many popular SSL frameworks are Joint Embedded Architectures (JEAs), where two or more networks project similar inputs to a joint embedding. However, while JEAs shows impressive results, a notable issue is *dimensional collapse*, in which learned representations lose their discriminative power by collapsing into a single point or a smaller subspace [2]. Many variations of contrastive methods [3, 4] have been introduced to mitigate this issue by contrasting dissimilar pairs of input. A downside with contrastive methods however, is that they require a lot of memory during training since large batch sizes are important to choose informative negatives for avoiding collapse [1, 5]. Furthermore, such methods explicitly push dissimilar inputs away from each other, which can counteract information maximization if they still have a high degree of similarity.

The Variance-Invariance-Covariance regularization (VICReg) loss function [1] is a proposed solution to information collapse without the need to contrast dissimilar embeddings. Instead, VICReg avoids collapse by directly enforcing variance and covariance regularization on each branch of the JEA separately. This self-supervised learning algorithm is simple, as information maximization is ensured through simple statistical formulations, and is thus theoretically insusceptible to collapse [6].

While VICReg has been applied primarily to computer vision tasks, its principles could be effectively translated to audio representation learning. Extending our previous work on comparing VICReg and Contrastive Learning of Musical Representations (CLMR) [7], we introduce VICMus, a specialized adaptation of VICReg for music representation learning that utilizes a ResNet50 backbone with spectrogram input, making it comparable and competitive with current music audio SSL benchmarks.

We pre-train VICMus on the medium subset of the Free Music Archive Dataset (FMA) [8], a dataset of 25k songs, and show that a simple linear classifier applied on top of the learned VICMus embeddings achieves 36.3 mean average precision (mAP) on music tagging on the MagnaTagaTune (MTAT) dataset [9]. This outperforms CLMR [10] mAP of 35.6, which is trained on the ten times larger Million Song Dataset (MSD) [11].

For evaluating VICMus, we rely on the standard downstream classification tasks for evaluating audio representations, following the Holistic Audio Representation Evaluation Suite (HARES) benchmark [12] for music data: music tagging on MTAT [9], and pitch estimation and instrument identification on NSynth Dataset [13].

This work was supported by Epidemic Sound and Google Research Cloud

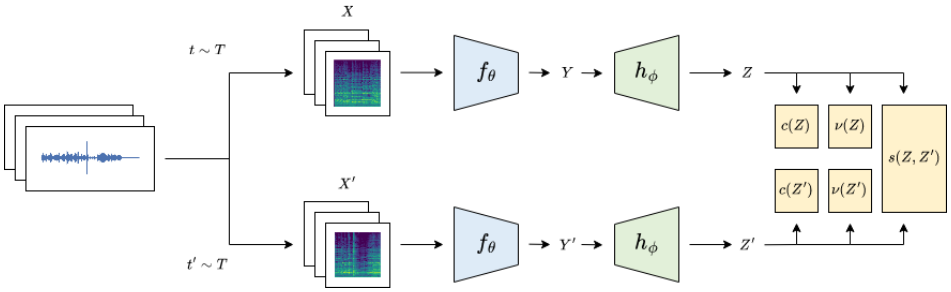

Fig. 1: VICMus architecture. From left to right: a batch of waveforms are augmented from a distribution T. The augmented waveforms are transformed into a log-mel spectrogram and finally normalized. The batch pairs are forwarded through the encoder f_θ, then the projector h_ϕ. Finally, the projector's output is used to calculate the Variance ν, Invariance s, and Covariance c regularization terms. After pre-training, the projector is discarded, and the encoder is frozen and used as input for downstream classification tasks.

2. RELATED WORK

2.1. Contrastive Learning for Audio

Contrastive methods learn information by minimizing the distance between similar audio pairs and prevent dimensional collapse by pushing dissimilar pairs away. Various contrastive methods have been adapted from the image domain to the music domain. Both COntrastive Learning for Audio (COLA) [14], Contrastive Learning of Auditory Representations (CLAR) [15], and CLMR implemented the contrastive method SimCLR [3]. While COLA uses a Siamese network architecture with spectrograms as input, CLAR uses both audio waveform and spectrogram, and CLMR only uses waveforms. Although it has been shown that those methods can achieve good performance on various tasks, they exhibit the problem of requiring substantial batch sizes to not collapse [5], or computationally expensive triplet mining to select informative negative examples [16].

2.2. Modal adaptations of VICReg

VICReg has been successfully adapted to different modalities [17, 18]. The author of [17] used a variation of VICReg loss on a combination of image and sensor input for SSL terrain representation from robot experience, achieving state-of-the-art results [17]. Similarly, Variance and Covariance regularizers from the VICReg loss were used in [18] to penalize correlated embeddings generated when finetuning BERT models [19], leading to a reduction in estimation errors in the speech-based prediction of cognitive impairment. The original VICReg publication also demonstrated that it could outperform both full-supervision and Barlow-Twins loss [4] on a general audio classification dataset ESC-50 [1]. These studies provide evidence that VICReg, while initially designed for image-based tasks, can be successfully adapted to other modalities, such as audio and sensor data, offering compelling performance in various applications.

3. VICMUS

In this work, we propose VICMus, a self-supervised method that can be used to learn useful music representations from unlabelled data. VICMus is grounded on VICReg [1], but we adapt it to process music data, which requires different data augmentation procedures. We utilize similar augmentations to CLMR [10], which have been proven to be useful for learning representations via contrastive self-supervised methods.

VICMus, like many other SSL methods, follows a siamese network architecture consisting of an encoder f_θ, which encodes audio information in a lower dimensional space, and a projector h_ϕ, which projects the representations $f_\theta(x)$ to a higher-dimensional feature space at which the VICReg loss is applied. We apply the regularization loss to a projection of the representations since this has been empirically shown to have better performance than applying the loss directly to the learned embeddings [1, 20]. Unlike VICReg, the encoder takes in normalized log-mel spectrograms as input [16]. A complete overview of VICMus can be found in Figure 1.

VICMus follows VICRegs triple loss function [1], which consists of a Variance ν, Invariance s, and a Covariance c term. Unlike most other SSL techniques, only the invariance s is applied across the branches of the Siamese architecture, and the other regularization terms are applied independently, as shown in Figure 1. In order to define these, let's first denote the batch of randomly augmented songs $X = [x_1, ..., x_n]$ and the same songs with different augmentations $X' = [x'_1, ..., x'_n]$., where n corresponds to the batch size. The augmentations of X and X' are both drawn from a distribution T. The respective resulting embeddings of the branches are denoted Z and Z', each consisting of n vectors of dimension d. z^j denotes the column-vectors at the j-th dimension in batch X. The variance term is then defined as:

$$v(Z) = \frac{1}{d}\sum_{j=1}^{d} \max(0, \gamma - S(z^j, \epsilon)). \quad (1)$$

In this equation, γ stands as the target for the standard deviation, and ϵ is a small scalar introduced to prevent numerical instability. The function S denotes a regularized standard deviation and is defined as:

$$S(x, \epsilon) = \sqrt{\mathrm{Var}(x) + \epsilon}. \tag{2}$$

This constraint ensures a constant standard deviation of γ along the batch dimension, mitigating informational collapse. The invariance term is the mean-squared Euclidian distance between pairs of vectors Z and Z':

$$s(Z, Z') = \frac{1}{n} \sum_{i} \left\| z_i - z'_i \right\|_2^2. \tag{3}$$

The covariance term is defined as

$$c(Z) = \frac{1}{d} \sum_{i \neq j} [C(Z)]_{i,j}^2, \tag{4}$$

where C is the covariance matrix defined as follows:

$$C(Z) = \frac{1}{n-1} \sum_{i=1} (z_i - \bar{z})(z_i - \bar{z})^T, \tag{5}$$

where

$$\bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i. \tag{6}$$

This decorrelates the off-diagonal elements of the covariance matrix by forcing them to be close to zero, preventing dissimilar embeddings from encoding similar information. Finally, the complete loss function is formulated as a weighted summation of the variance, invariance, and covariance terms:

$$\begin{aligned} l(Z, Z') = & \lambda s(Z, Z') \\ & + \mu [v(Z) + v(Z')] \\ & + \nu [c(Z) + c(Z')]. \end{aligned} \tag{7}$$

In the above equation, λ, μ, and ν function as hyperparameters that modulate the weight of the different components in the loss function. Following VICReg, we use 25, 25 and 1, respectively, in our experiments. Tuning those hyperparameters for music is left as future work.

4. EXPERIMENTS

4.1. Datasets

All pre-training was completed using the FMA dataset [8]. The FMA dataset is a large-scale collection of music from various genres sampled by the Free Music Archive platform, consisting of over $100,000$ tracks, spanning 161 genres. FMA is divided into three subsets based on the number of tracks: small ($8,000$ tracks), medium ($25,000$ tracks), and large ($106,574$ tracks). Its large-scale nature and extensive metadata make it particularly useful for training and evaluating machine learning models, such as self-supervised representation learning. In this study, we only use FMA medium.

The representations are evaluated using HARES-music [12]. We chose only to evaluate on the subset HARES-music since HARES also contains downstream tasks that are not related to music. HARES-music consists of three downstream classification tasks with the MagnaTagATune and Nsynth datasets.

The MagnaTagATune (MTAT) dataset [9] consists of $25,863$ music clips at 29 seconds each, with multi-label annotations provided by multiple users in a crowdsourcing tagging game called "TagATune". With this dataset, we evaluate music tagging.

Neural Synthesizer (NSynth) [13] is a synthetic dataset generated by deep neural networks. The dataset consists of $305,979$ four-second audio clips of single notes, played by 11 different instruments, covering 128 different pitches. With this dataset, we evaluate instrument recognition (NSynth Instrument) and pitch classification (NSynth Pitch).

4.2. Data Augmentations

We use the same set of data augmentations for pre-training VICMus as CLMR [10] - a contrastive based SSL method for music: RandomCrop, PolarityInversion, Noise, Gain, HighLowPass, Audio, PitchShift and Reverb. We use the same settings as CLMR [10] except for RandomCrop, with $65,024$ samples at $22,050$ hz.

Finally, we convert the augmented audio to log-mel spectrograms like in [16], but with a window size of $46\,\mathrm{ms}$, hop length of $23\,\mathrm{ms}$, and duplication over channels, resulting in an input shape of (3, 128, 128). Each input is min-max normalized to the interval $[-1, 1]$.

For training downstream tasks, we used RandomCrop for MTAT and no data augmentations for NSynth tasks.

4.3. Implementation details

Our hyperparameters during pre-training closely resemble that of VICReg [1]. We use a ResNet50 [21] as our encoder network f_θ and an MLP with dimensions *2048-8196-8196-8196* as our projector h_ϕ. We pre-train this architecture for $4,000$ epochs on FMA medium using a LARS optimizer [1] with a weight decay of 10^{-6} and a batch size of 2048. The learning rate follows a cosine decay schedule with ten warmup epochs. The learning rate is set to $lr = batch\_size/256 \times base\_lr$, where $base\_lr = 0.4$. For downstream tasks, we discard the projector and replace it with a linear classifier applied atop the pre-trained frozen encoder. All pre-training took approximately five days on a TPUv3-8 virtual machine.

Table 1: Evaluation on HARES-music downstream tasks. Scores reported on MTAT are mAP and on NSynth accuracy. HARES-music is the average score across the all tasks. All encoders are evaluated using downstream linear classification. VICMus uses the same augmentations as CLMR, other models use example mixup. Our previous work is marked by †.

Method	Model (#Params)	pre-training dataset (#Samples)	NSynth		MTAT	HARES Music
			Pitch	Instrument		
SimCLR [12]	SF NFNet-F0 (63m)	AudioSet (1.9m)	88.0	78.2	39.5	68.6
SimCLR [12]	ResNet50 (23m)	AudioSet (1.9m)	78.5	73.8	38.7	63.7
Supervised [12]	ResNet50 (23m)	-	69.3	70.7	38.7	59.6
CLMR [10]	SampleCNN (2.5m)	MSD (200k)	-	-	35.6	-
CLMR† [7]	SampleCNN (2.5m)	FMA (25k)	-	-	33.4	-
VICReg† [7]	SampleCNN (2.5m)	FMA (25k)	-	-	34.7	-
VICMus (ours)	ResNet50 (23m)	FMA (25k)	54.4	64.4	36.3	51.7

4.4. Results

We evaluated VICMus against several baseline methods on the HARES-music downstream tasks, as summarized in Table 1. For the MagnaTagATune (MTAT) task, we report mean Average Precision (mAP), and for NSynth tasks: accuracy.

To isolate the impact of the VICReg loss function, we compared VICReg and CLMR using the same SampleCNN backbone, FMA training data subset and data augmentations in our previous work [7], as seen in table 1 marked with †. VICMus extends our previous work [7] with ResNet50 as backbone utilizing spectrograms as input, which further improved performance and makes our model comparable to recent benchmarks. In table 1, VICReg outperforms CLMR when trained on FMA medium, achieving a mAP of 34.7 compared to CLMR's 33.4, both using SampleCNN (2.3m parameters) as an encoder. VICMus, employing the same loss function but with a ResNet50 architecture (23m parameters), achieves a mAP of 36.3, beating previous methods. This performance is particularly noteworthy given that it also surpasses the original CLMR [10] pre-trained on MSD, which is a dataset nearly ten times larger than FMA.

While VICMus outperforms CLMR on MTAT, it underperforms across the downstream tasks compared to its ResNet50-based supervised and SimCLR counterparts provided by HARES [12]. While there are many differences in hyperparameters and setups between these models, it is important to highlight SimCLR, which outperforms VICMus in all downstream tasks, was trained on the almost hundred times larger AudioSet dataset [22]. This highlights the potential impact of dataset size on model performance and suggests that VICMus could benefit from training on larger datasets.

Another observation is VICMus' poor performance on the NSynth Pitch downstream task. This is likely due to the PitchShift augmentation used in VICMus being detrimental to the Pitch-related downstream tasks: it trains the representations to be invariant to pitch.

In summary, our results demonstrate VICMus's significant performance gains over CLMR while also isolating hyperparameters that might otherwise confound these results. Despite its promising performance in some tasks, VICMus is outperformed by specialized or more extensively trained models in the comprehensive HARES-music benchmark.

5. CONCLUSION & FUTURE WORK

In this study we introduced VICMus, a specialized adaptation of the VICReg loss function for music representation learning, and compared its performance with baseline methods on the HARES-music downstream tasks. Our results demonstrate that VICMus not only outperforms CLMR but also does so with a smaller dataset.

However, VICReg falls short in the broader HARES-music benchmark, particularly compared to its supervised counterpart. Given the performance gains VICMus demonstrated even when trained on a smaller dataset, an exciting avenue for future research is to evaluate its capabilities when trained on larger music datasets, such as MSD. This could offer insights into scalability, and the ability to leverage more extensive data for improved performance.

Additionally, the performance of VICMus on HARES-music is decreased as a result of our augmentations, which were specifically tuned for MTAT [10]. Consequently, certain augmentations, such as PitchShift, may be detrimental to performance on NSynth Pitch; since the learned embeddings are, in this case, trained to be invariant to pitch change.

In summary, VICMus offers a promising and computationally efficient avenue for music representation learning.

6. REFERENCES

[1] A. Bardes, J. Ponce, and Y. LeCun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in *Int. Conf. on Learning Representations ICLR*, April 2022.

[2] L. Jing, P. Vincent, Y. LeCun, and Y. Tian, "Understanding dimensional collapse in contrastive self-supervised learning," in *Int. Conf. on Learning Representations (ICLR)*, 2022.

[3] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Int. Conf. on Machine Learning (ICML)*, pp. 1597–1607, 2020.

[4] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in *Int. Conf. on Machine Learning (ICML)*, pp. 12310–12320, 2021.

[5] L. Wang and A. v. d. Oord, "Multi-format contrastive learning of audio representations," in *Workshop Talk at Neural Information Processing Systems Conf. (NeurIPS)*, 2021.

[6] R. Shwartz-Ziv, R. Balestriero, K. Kawaguchi, T. G. J. Rudner, and Y. LeCun, "An information-theoretic perspective on variance-invariance-covariance regularization," in *Symposium on Advances in Approximate Bayesian Inference (AABI)*, 2023.

[7] S. Löf and C. Hesse, *Self-Supervised Learning of Musical Representations Using VICReg*. Master thesis, Chalmers University of Technology, Gothenburg, Sweden, 2023.

[8] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "FMA: A dataset for music analysis," in *18th Int. Society for Music Information Retrieval Conf. (ISMIR)*, 2017.

[9] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging," in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, (Kobe, Japan), October 2009.

[10] J. Spijkervet and J. A. Burgoyne, "Contrastive learning of musical representations," in *Int. Society for Music Information Retrieval Conf. (ISMIR)*, pp. 673–681, November 2021.

[11] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Int.Society for Music Information Retrieval Conf. (ISMIR 2011)*, 2011.

[12] L. Wang, P. Luc, Y. Wu, A. Recasens, L. Smaira, A. Brock, A. Jaegle, J.-B. Alayrac, S. Dieleman, J. Carreira, and A. van den Oord, "Towards learning universal audio representations," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4593–4597, IEEE, 2022.

[13] 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 *Int. Conf. on Machine Learning (ICML)*, pp. 1068–1077, 2017.

[14] A. Saeed, D. Grangier, and N. Zeghidour, "Contrastive learning of general-purpose audio representations," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3875–3879, IEEE, 2021.

[15] H. Al-Tahan and Y. Mohsenzadeh, "Clar: Contrastive learning of auditory representations," in *Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, pp. 2530–2538, 2021.

[16] C. Thomé, S. Piwell, and O. Utterbäck, "Musical audio similarity with self-supervised convolutional neural networks," in *Late-Breaking Demo Session on the Int. Society for Music Information Retrieval Conf.*., 2021.

[17] H. Karnan, E. Yang, D. Farkash, G. Warnell, J. Biswas, and P. Stone, "Self-supervised terrain representation learning from unconstrained robot experience," in *ICRA Workshop on Pretraining for Robotics (PT4R)*, 2023.

[18] L. Xu, K. D. Mueller, J. Liss, and V. Berisha, "Decorrelating language model embeddings for speech-based prediction of cognitive impairment," in *Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Annual Conf. of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

[20] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," in *Int. Conf. on Neural Information Processing Systems (NeurIPS)*, pp. 22243–22255, 2020.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[22] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE ICASSP 2017*, (New Orleans, LA), 2017.