Improving few-shot learning through latent space topology of deep audio generative models

Philippe Esling, Antoine Caillon

Abstract

The recent advent of *generative models* has yielded impressive results in diverse application fields, notably for audio synthesis. However, the major drawback of almost all of these methods is that they require massive datasets of examples and, consequently, extensive training times. This poses problems in cases where the amount of examples is inherently limited or very costly to obtain, but this also incurs energetic and environmental issues. Furthermore, this need for extensive datasets also questions the true generalization abilities of these models. Recently, the field of *Few-Shot Learning* (FSL) has tried to tackle these issues directly, by searching for models that could generalize from only few examples. However, most approaches are based on pre-training with large datasets or performing expensive augmentation strategies.

The goal of this PhD is to develop novel methods able to address few-shot learning in deep audio synthesis, without relying on extensive pre-training. Our main assumption is that the underlying topology of latent spaces can provide keys to understand situations where the properties of the datasets (number of examples, modes and their relative densities) are sufficient to provide accurate training. Hence, this PhD will extend on our previous works on deep audio synthesis by infusing geometry-aware methods and training. First, we will analyze how different sub-sampling of the same dataset may impact the topology of latent models and, consequently, their quality. This will also allow to study the questions of mode collapse and correlate it to the relative densities of different modes. Then, we will rely on tools from information geometry and Riemannian topology in order to regularize the learning. Our goal is to study the latent space kinematics in order to improve the generalization of deep generative models trained with only few examples. This PhD can lead to efficient audio models, but also generic approaches to improve few-shot learning and the generalization abilities of generative models.

Keywords: audio synthesis, probabilistic generative models, variational inference, few-shot learning, audio signal processing, music creation

Context

Recent advances in deep learning have provided efficient solutions in a large variety of application fields, achieving impressive accuracy in both discriminative and generative tasks. Notably, *deep generative models* have seen great success in the field of audio synthesis, enabling the generation of high-quality audio content matching the perceptual features of a given dataset [9] [5].

However, these models are becoming increasingly data greedy and require expensive training on a massive number of examples in order to generate high-quality samples [10]. For instance, the autoregressive *WaveNet* model [9] was trained on 200 hours of music to generate unstructured musical content. This necessity of large datasets represent a core issue in many applications where the amount of examples is either inherently limited or very costly to obtain. Indeed, gathering high-quality examples for real-life problems is a challenging task, which requires manual effort and expert knowledge. This is especially problematic for historical artifacts, such as musical compositions. Musical applications also often require to exploit very limited datasets that match closely a given artistic intention. Furthermore, the energy and computational costs of training on large datasets are raising crucial issues of environmental sustainability [8]. Finally, this need for very large datasets in training also questions the true generalization abilities of these models, extending the scope of this research beyond low data regime applications.

These questions have been very recently at the center of research attention, notably through the field of Few-Shot Learning (FSL) [12]. FSL is a recent and very active field of machine learning that tries to tackle this challenge of generalizing from a limited number of examples. In the vast majority of cases, a proposed solution to this question is to find a way to *transfer* the learning of a pre-trained model to the small dataset case [11]. However, this still requires to perform a first training on a very large dataset. Another direction of research is to rely on strong data augmentations or self-supervised learning. However, the success of the model highly depends on the correlation between the task and the augmentation type [12]. Finally, another line of research consists in developing models with *strong inductive bias* to ease the need for large datasets. One very successful example in audio generation is the Differentiable Digital Signal Processing (DDSP) model [6]. However, the inherent construction of DDSP strongly limits the nature of signal that can be modeled (harmonic and monophonic).

A widely accepted assumption in machine learning is that the data resides on a low-dimensional manifold embedded in the ambient high-dimensional data space [2]. Recent works suggest that latent models have low generalization capabilities when they fail to capture the manifold intrinsic geometry. This limitation might come from the assumption that their latent space is Euclidian [4]. Hence, modeling the latent space as a Riemannian manifold improves our understanding of its structure and enables better interpolation between latent points [1]. More recently, a geometry-aware Variationnal Auto-Encoder was proposed [3], taking into account the Riemannian manifold structure of the latent space to improve sample generation for medical imaging applications where only a small number of examples is available.

The main goal of this PhD is to develop new methods for enhancing the generation capabilities of generative models in low sample size and few-shot settings, with a specific emphasis on deep audio models. To do so, we will target the enrichment of latent spaces through geometrical modelling. This can lead to a better understanding of the critical characteristics required in datasets (cardinality, variety, mode densities) to ensure the generalization ability of models. Mainly, this PhD can lead to novel regularization and training methods in few-shot, small datasets situations. This can profoundly enhance the whole field of generative modeling, but also enable researchers and artists to exploit the creative potential of such models when only few examples are available.

Research design and methodology

The research methodology is constructed around three main axes. First, there is currently almost no knowledge regarding the minimal cardinality and maximal diversity of datasets required to train generative models successfully and ensure generation quality and diversity. Hence, the first step of this of this PhD will focus on understanding the dataset properties essential to the model generalisation capabilities. By operating successive sub-samplings from an initial complete dataset and evaluating the latent space kinematics, the dataset properties (cardinality, density) will be correlated with the characteristics of the learned representation. This initial work will aim at understanding how the latent space structure is transformed as examples are removed from the dataset, providing insights for the next steps of the project. This will also allow to study the questions of mode collapse related to the relative densities of different modes. To the best of our knowledge, no previous work evaluated few-shot strategies for deep generative models in audio. Hence, the second step of this PhD will consist in applying different existing low-shot transfer and data augmentation techniques from the literature to deep audio models. After constituting this baseline on state of art audio synthesis models, this research will try to overcome the need for extensive pre-training by infusing inductive biases on the generation process, striking the balance between constraining the sound type and having as few examples as possible. Also, we will adapt the recently-developed sampling and interpolation techniques relying on Riemannian geometry to enhance the generation of deep audio models [7] [3]. Based on the insights from the two initial steps of the project, the core of this PhD will be to propose strategies for enhancing the generalization capabilities of generative models in low sample size settings without relying on extensive pre-training. First, we will target the development of new augmentations techniques by relying on geometrical analysis of the latent space and associated data manifold. Based on the previously-established understanding of the latent space kinematics on low data regime, we will propose new methods to regularize the training and improve the generalisation capabilities of the models. The proposed methods will be thoroughly evaluated on a variety of datasets, characterizing their efficiency with regards to the number and diversity of training sounds. This work aims at providing efficient tools for researchers and end users of the models. Music represents an ideal testbed for this problem as some datasets are inherently limited and new examples are complex to produce. However, the approach developed in this PhD will be broadly applicable to other domains.

The proposed PhD is deeply linked with current researches conducted within the IRCAM STMS laboratory and its network of technological startup companies, and it will fully leverage the current momentum generated by the thesis director, through the SSHRC ACTOR Network and ACIDITeam Emergence(s) Ville de Paris both under the supervision of Philippe Esling.

Objectives

Here, we introduce the main tasks that will be carried out during the first two years of the PhD with regards to the proposed approaches and methods set previously.

- Characterize datasets requirements for generative models performance
- Understand the latent space kinematics under successive ablation of a complete dataset
- Implement and compare various transfer and data augmentation techniques
- Characterize how inductive biases on sound types moderate the need for massive datasets
- Exploit the Riemannian geometry of the latent spaces to improve low data regime performance
- Propose regularisation and sampling strategies to enhance models' generalisation capabilities
- Evaluate the performance of the proposed methods on real-life applications

References

[1] Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

[2] Pratik Prabhanjan Brahma, Dapeng Wu, and Yiyuan She. Why deep learning works: A manifold disentanglement perspective. *IEEE transactions on neural networks and learning systems*, 27(10):1997–2008, 2015.

[3] Clément Chadebec and Stéphanie Allassonnière. Data generation in low sample size setting using manifold sampling and a geometry-aware vae. 2021.

[4] Tim R. Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M. Tomczak. Hyperspherical variational auto-encoders. 2018.

[5] Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. *arXiv preprint arXiv:1902.08710*, 2019.

[6] Jesse Engel, Lamtharn Hantrakul, Chenjie Gu, and Adam Roberts. Ddsp: Differentiable digital signal processing. *arXiv preprint arXiv:2001.04643*, 2020.

[7] Dimitris Kalatzis, David Eklund, Georgios Arvanitidis, and Søren Hauberg. Variational autoencoders with riemannian brownian motion priors. *CoRR*, abs/2002.05227, 2020.

[8] Gaurav Menghani. Efficient deep learning: A survey on making deep learning models smaller, faster, and better. *arXiv preprint arXiv:2106.08962*, 2021.

[9] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[10] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

[11] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[12] Yaqing Wang, Quanming Yao, James T Kwok, and Lionel M Ni. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34, 2020.