Efficient deep generative models through topology-merging subnetworks for embedded audio synthesis

Philippe Esling[1], Philippe Codognet[1, 2], Tom Hurlin[3]

[1] IRCAM – UMR 9912, Sorbonne, [2] JFLI – CNRS IRL 3527, University of Tokyo, [3] Squarp Instruments

Abstract

Deep learning models have provided extremely successful methods in most application fields, obtaining unprecedented accuracy in various tasks. However, the consistently overlooked downside of deep models is their massive complexity and tremendous computation cost. Besides hampering the possibility for model interpretability, the energy and computational costs of such architectures are raising crucial issues of environmental sustainability. This aspect is especially critical in audio applications, which heavily relies on specialized embedded hardware with real-time constraints. Although deep generative models are now able to synthesize waveform data with unprecedented quality, they still require costly specialized hardware and large inference times. Hence, the lack of work on efficient lightweight deep models is a significant limitation for the real-life use of deep models on resource-constrained hardware.

The goal of this PhD is to explore the recently-defined *lottery ticket hypothesis*, in the case of generating highly variable and complex audio data with multiple modes. This hypothesis states that randomly-initialized neural networks already contain extremely sparse subnetworks that could have higher accuracy than their larger counterparts if they were trained in isolation. Hence, finding these subnetworks implies that the same task could be solved in a lightweight, memory and energy-efficient way. However, most of the researches in this direction remain applied on large and homogeneous datasets. The goal of this PhD is to extend and analyze these methods in case of high dataset variability with low amount of available data (low-shot) per mode. To do so, we will explore the possibility to obtain different subnetworks for specific data modes, while avoiding the need for repeated training cycles. We will then study the *mode connectivity* between both the obtained representation spaces and the weights of resulting subnetworks with tools from the *information bottleneck theory*. Based on this, we will explore the notion of topology-merging subnetworks, where efficient light networks are trained jointly for different modes, while trying to merge their respective computation. This PhD will lead to the development of innovative musical instruments, providing users with diverse creative control over deep generative audio models on constrained hardware and embedded audio synthesizers.

Keywords : Deep generative models, audio synthesis, digital signal processing, network compression, information bottleneck theory, embedded machine learning

1 Context

The recent progress in the field of deep neural networks have yielded impressive results in a wide variety of tasks. In particular, *deep generative models* have provided increasingly accurate and groundbreaking results in various generative tasks, notably in *musical audio synthesis*. Models such as the Differentiable Digital Signal Processing (DDSP) [6] or the Real-time Audio Variational Autoencoder (RAVE) [2] are now capable of real-time audio generation while matching the perceptual features of a given dataset. This opens a whole set of new prospects for a variety of musical applications. However, the consistently overlooked downside of these models is their massive complexity, which seems concomitantly crucial to their success. In particular, in order to achieve accurate generation, deep generative models have become increasingly over-parameterized. For instance, the MegatronLM network reaches up to 8.3 billion parameters [16]. This exploding size leads to profound issues in both the use and understanding of these models. Indeed, increasingly complex models require ever longer training time [1] raising serious environmental issues and also lead to slower inference, precluding their implementation in end-user embedded systems which prevails in audio applications. Finally, such complexity indubitably decreases the potential interpretability of these models.

Recently, the *lottery ticket hypothesis* [9] has provided empirical evidence that randomly-initialized neural networks already contain powerful subnetworks that could reach the same or even higher accuracy than the original networks if they were trained in isolation. Furthermore, these subnetworks being easier to analyze, they could simplify future works towards explainability [10]. Several studies have analyzed different properties of this hypothesis, for instance in terms of the potential for transferring these subnetworks [14] to other domains. Notably, recent researches by the supervising team has shown that the lottery ticket could be successfully applied to musical data [7] and also to deep audio generative models [8], even when using

structured pruning, which ensures a true compression of the models. However, this hypothesis has two major flaws. First, it has a large training cost, as finding subnetworks seems to be only stable when the training is repeated multiple times over iteratively smaller networks [20]. Second, another major issue (shared with generative models) is that, while networks usually perform well when trained on homogeneous datasets, their performances seriously drop when trained on datasets with high data variability [17] and low amount of data per mode. Many researches suggest that the subnetworks do not depend on the training data nor on the generative model, and that they rely solely on the network topology [13, 4]. Hence, recent studies have aimed to develop generic compression techniques [19, 18], regardless of the training dataset.

The goal of this PhD is to leverage these recent works, while improving the theoretical framework surrounding their use. To do so, we will study the connections between the latent spaces and weights of different subnetworks, trained on specific modes of highly variable data. By relying on tools from the *information bottleneck theory*, we aim to improve the discovery of efficient subnetworks. Then, we will evaluate the relationships between different subnetworks, both in terms of latent space and mode connectivity of their weights. Based on this analysis framework, we will evaluate the possibility to perform the joint training of efficient subnetworks and various data modes. Then, we will evaluate the possibility for topology-merging of these networks while maintaining their efficiency. This PhD will lead to the development of lightweight deep learning models, creative tools and enhanced approaches that would allow to democratize deep models on lightweight and resource-constrained embedded hardware.

2 Research design and methodology

This PhD aims to develop novel methods for improving the discovery of efficient subnetworks in the context of generative audio models. Regarding the model itself, the PhD will fully leverage the previous research of the supervising team, which developed the RAVE model [2] already providing high-quality audio generation at 48kHz sampling rate. The improvements to efficient subnetworks are specifically targeted on the objectives of *handling heterogeneous audio datasets*, *avoiding the need of multiple cycles of training* and *embedding the resulting models on resource-constrained architectures*. Based on the aforementioned objectives, the research methodology of this PhD is structured around three major and complementary axes. First, the PhD will improve efficient subnetwork discovery (**axis 1**) but on a single data mode (*homogeneous* training data), trying to avoid multiple training cycles and studying the relations between found subnetworks. Then, we will work on developing methods for merging these efficient subnetworks (**axis 2**), effectively producing models that are able to generate a variety of controllable output sounds (*heterogeneous* training data), while remaining lightweight. Finally, the PhD will work on implementing these efficient models on resource-constrained architectures (**axis 3**), leading to novel and creative embedded deep audio synthesizers.

Axis 1. Improving efficient subnetworks discovery for generative audio models.

Even though the lottery ticket hypothesis has yielded significant compression rates on a variety of networks and tasks, the reasons behind its effectiveness still remain unclear and an active field of research. Recent studies [11, 10] suggest that the efficiency of pruning algorithms is highly correlated to the stability of the optimization path in the objective landscape. Hence, efficient sparse subnetworks can thus be found very early in training, as confirmed by follow-up researches [20, 19]. However, most experiments so far have been conducted on classification tasks and with highly homogeneous data. On the other hand, our previous work [8] suggest that deep generative models can also be largely pruned without altering the generation accuracy. Hence, the first step of this PhD will be to study how these different early pruning techniques can generalize to deep generative models, in the particular case of audio synthesis. This would bypass the need for expensive multiple training cycles of the subnetworks. One direction of theoretical improvement is that the training of deep models seems to follow a two-stage process [15]. Therefore, the efficiency of pruning methods should lie in the transition between *fitting* and *compression* phases. Hence, we will study how the *information bottleneck theory* can be leveraged to exploit global properties on this transition, by relying on the *mutual information* between representations [15]. This can lead to more efficient methods for the discovery of efficient subnetworks when training on a single data mode.

Axis 2. Topology merging of efficient subnetworks with the information bottleneck theory.

Based on the insights from the previous axis, we will study the relations between subnetworks trained on different modes of a dataset. Recent researches [14, 4] suggest that subnetworks generalize successfully across different data distributions. Yet, the links between the resulting networks and latent representations still remain unclear. Hence, we aim at studying the shared properties of subnetworks when trained on different modes of a dataset. In particular, we will study the *mode connectivity* between corresponding layers [12], and *mutual information* [15] between intermediate representations. On the basis of these properties, the core of

this PhD is to find strategies to merge tickets from different distributions of heterogeneous datasets in order to achieve a high generation variety within a single model. Specifically, we will study how the previously established connectivity properties can be leveraged to train a single network on multiple data modes. We will take inspiration from recent advances in *lifelong learning* with sparse networks [3]. This would allow to keep separate modes as increasingly complex information comes in order to prevent from *catastrophic collapse*. Hence, the PhD student will develop regularization strategies to combine information bottlenecks while keeping mode-specific information within the layers. This can lead to develop specific strategies to perform *topology-merging* between subnetworks trained on different modes. A fallback solution will be to first study the use of mixture models, considering each subnetwork as a base distribution inside the mixture (merged network). This work can deeply contribute to the knowledge on the information structure within deep neural networks, while providing new tools for improving accuracy of models on diverse tasks.

Axis 3. Embedded efficient deep audio synthesis for innovative musical instruments.

Finally, based on the methods for merging efficient subnetworks, this PhD aims to effectively embed the resulting models on constrained architectures. Specifically, we will target very light and inexpensive architectures that can be used for the development of audio synthesizers (e.g. *Raspberry Pi*). In this axis, the PhD will fully leverage the collaboration with *Squarp Instruments* and also previous prototypes from the supervising team [5, 8] for the development of specifically-tailored hardware prototypes. The PhD is expected to evaluate the computational costs of efficient deep audio generators to make them fit to the target interface. Thus, the implementation on this hardware will also require to ensure that models resulting from the previous axes truly allow real-time generation on embedded hardware. These prototypes will then be extended in order to ensure intuitive control over the generation, and the creation of versatile and diverse sounds. To do so, we will collaborate with musicians and composers to have a direct evaluation feedback loop for the model design. Hence, the potential for both creative and industrial applications to this PhD is very high, and can lead to a whole new category of musical instruments.

3 Objectives

The following main objectives will frame the work and the research of the PhD student :

- Improving strategies for discovering efficient subnetworks on single data modes.
- Extending the theoretical framework by incorporating ideas from the information bottleneck.
- Establishing global properties for identifying sparse sub-networks without relying on multiple training.
- Analyzing shared properties (latent spaces and weights) between subnetworks on single data modes.
- Generalizing this analysis to subnetworks trained on different data modes.
- Developing new approaches for jointly training subnetworks on highly diverse datasets.
- Develop methods to perform topology-merging subnetworks training.
- Deploying lightweight models on hardware with limited resources.
- Analyzing the creative potential of these novel instruments with composers and musicians.

Although a specific emphasis will be put on audio generative models, the approaches and methods developed in this PhD are expected to be widely expandable to many other fields, which can give rise to potential applications to a large variety of tasks.

4 Supervision context

The PhD will mainly take place at IRCAM (Institut de Recherche et de Coordination Acoustique/Musique - CNRS UMR 9912), in the Artificial Creative Intelligence and Data Science (ACIDS) team directed by Philippe Esling (associate professor at Sorbonne Université), which is one of the foremost research group in deep audio generative models. Hence, the proposed PhD is deeply linked with current research objectives conducted within the IRCAM STMS laboratory, as it seeks to develop new interactive and user-based tools for musical creation. Thus, as this PhD is focused around developing deep learning models that provide musicians with creative control over the generation, it is directly in continuity and will fully leverage previous researches conducted by this team, but also its extensive experience and network of users. Furthermore, the PhD will also be co-supervised by Tom Hurlin, co-founder of the company Squarp Instruments, which specializes in developing Eurorack modules, which will provide precious knowledge and skills around the

hardware and embedded programming questions of this PhD. Finally, this PhD will also be co-supervised by Philippe Codognet, which is head of the JFLI laboratory (University of Tokyo), providing an expertise in theoretical aspects of machine learning and opportunity for international mobility during the PhD.

Therefore, the PhD ideally sits at the crossroads of diverse experiences of the supervisors and will fully leverage both the current momentum generated by the thesis director through various funded projects (SSHRC ACTOR Network and ACIDITeam Emergence), of the industrial prospects provided by Squarp Instruments (Tom Hurlin), and the international opening offered by the Japanese JFLI (Philippe Codognet).

References

[1] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *arXiv preprint arXiv:2103.04922*, 2021.

[2] Antoine Caillon and Philippe Esling. Rave: A variational autoencoder for fast and high-quality neural audio synthesis. *arXiv preprint arXiv:2111.05011*, 2021.

[3] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Long live the lottery: The existence of winning tickets in lifelong learning. In *International Conference on Learning Representations*, 2020.

[4] Shrey Desai, Hongyuan Zhan, and Ahmed Aly. Evaluating lottery tickets under distributional shifts. *arXiv preprint arXiv:1910.12708*, 2019.

[5] Ninon Devis and Philippe Esling. Neurorack: deep audio learning in hardware synthesizers. In *EPFL Workshop on Human factors in Digital Humanities*, number CONF, 2021.

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

[7] Philippe Esling, Theis Bazin, Adrien Bitton, Tristan Carsault, and Ninon Devis. Ultra-light deep mir by trimming lottery tickets. In *International Symposium on Music Information Retrieval (ISMIR)*, 2020.

[8] Philippe Esling, Ninon Devis, Adrien Bitton, Antoine Caillon, Constance Douwes, et al. Diet deep generative audio models with structured lottery. *arXiv preprint arXiv:2007.16170*, 2020.

[9] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[10] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis. In *International Conference on Machine Learning*, pages 3259–3269. PMLR, 2020.

[11] Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. Stabilizing the lottery ticket hypothesis. *arXiv preprint arXiv:1903.01611*, 2019.

[12] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

[13] Neha Mukund Kalibhat, Yogesh Balaji, and Soheil Feizi. Winning lottery tickets in deep generative models. *arXiv preprint arXiv:2010.02350*, 2020.

[14] Ari Morcos, Haonan Yu, Michela Paganini, and Yuandong Tian. One ticket to win them all: generalizing lottery ticket initializations across datasets and optimizers. *Advances in neural information processing systems*, 32, 2019.

[15] Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

[16] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[17] Akash Srivastava, Lazar Valkov, Chris Russell, Michael U Gutmann, and Charles Sutton. Veegan: Reducing mode collapse in gans using implicit variational learning. *Advances in neural information processing systems*, 30, 2017.

[18] Jingtong Su, Yihang Chen, Tianle Cai, Tianhao Wu, Ruiqi Gao, Liwei Wang, and Jason D Lee. Sanity-checking pruning methods: Random tickets can win the jackpot. *Advances in Neural Information Processing Systems*, 33:20390–20401, 2020.

[19] Hidenori Tanaka, Daniel Kunin, Daniel L Yamins, and Surya Ganguli. Pruning neural networks without any data by iteratively conserving synaptic flow. *Advances in Neural Information Processing Systems*, 33:6377–6389, 2020.

[20] Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Richard G Baraniuk, Zhangyang Wang, and Yingyan Lin. Drawing early-bird tickets: Toward more efficient training of deep networks. In *International Conference on Learning Representations*, 2019.