A Mutli-Objective Approach for Sustainable Deep Learning

Constance Douwes[1] and Philippe Esling[1]

[1] *IRCAM, 1 Place Igor Stravinsky, Paris, France*

Depuis plusieurs années, les recherches en Intelligence Artificielle (IA), et plus particulièrement sur les réseaux de neurones, ont permis des avancées remarquables dans un grand nombre de domaine. Ce n'est pas sans conséquence sur le coût computationnel, qui n'a cessé de croître et atteint aujourd'hui des niveaux non négligeables. Au coeur de ce problème, vient notre manière d'évaluer et de mesurer leur performance : actuellement, les chercheurs se concentrent sur l'amélioration de la qualité des résultats générés, occultant ainsi le coût de calcul et l'impact environnemental de ces nouvelles technologies.

Nous introduisons ici l'emploi d'une mesure *multi-objective* permettant d'évaluer simultanément la qualité des modèles et leur efficacité énergétique. En appliquant cette mesure à plusieurs modèles de l'état de l'art en generation audio, nous montrons qu'elle change radicalement notre manière de juger les modèles, encourageant des techniques d'entrainement plus "verts" ainsi qu'une allocation plus optimales des ressources. Nous espérons que ce type de mesure sera largement adopté au sein de cette comunnauté, afin de mettre les coûts de calcul et les émissions de carbone sous les projecteurs de la recherche en apprentissage profond.

Mots-clefs : Informatique verte, empreinte carbone, optimisation multi-objective, generation audio

1 Context

Between 2012 and 2018, the amount of computation used in deep learning grew by a factor of *300,000* [1]. This exponential growth might have permitted to achieve impressive results across a wide variety of tasks, but it also strongly increased the demand for energy production, responsible for approximately 35% of total greenhouse gas emissions in 2010. If this trend continues, it is fairly logical to predict that deep learning will be a significant contributor to climate change. Moreover, research institutes and individuals can lack sufficient resources, due to the demand of countless types of specialized hardware (GPUs, TPUs), often running continuously for several days and even up to weeks. Hence, obtaining a quality similar to state-of-the-art models is becoming an unattainable goal, both financially and ecologically [9].

Generally speaking, the absence of energy-based criteria for generative models falls within the broader lack of suitable evaluation methods, notably for assessing the quality of the generated content [10]. In this study, we propose a new method to evaluate both accuracy and energy efficiency of deep generative models, and focus on raw audio synthesis. This task is a major challenge given that audio signals have strong temporal dependencies, composed of complex structures at both local and global scales. Most of the recent advances produced by deep approaches rely on a significant increase in terms of both size and complexity [5], as well as an ever-growing number of training examples.

In this study, we first present estimations of training costs in terms of CO_2 emissions for all state-of-the-art models we had enough training details among the twenty-five studied ones: *SampleRNN* [8], *SING* [2], *WaveGAN* [3], *GANSynth* [4] and *FloWaveNet* [6]. We then propose the use of a multi-objective Pareto criterion to provide fair comparisons regarding both accuracy and energy efficiency when publishing new models. We compute a subjective score for accuracy, and present two Pareto fronts, one for the training based on our CO_2 estimation, and one for the inference based on the number of parameters.

2 Methodology

First, we estimate carbon emissions of each of the five original papers' training procedures. Since we do not have all of the specific hardware they used, we make the assumption of the worst-case scenario and take the maximum power consumption p_{max} for each of the GPUs according to their technical specifications, as does the "Machine Learning Impact Calculator" [7]. We then multiply it by g, the number of GPUs used for training and by t in hours, to get the kilo-Watt hours consumption. As carbon emissions are location-dependent, we took a carbon intensity factor of 0.437 $kgCO_2eq/kWh$ as it is the global yearly average of 2018[†] to convert kilowatt-hours to carbon emissions. We ended up with the following formula to estimate the carbon emission (CO_2e) of a whole training as $CO_2e = 0,437 \times gp_{max} \times t$.

Increasing the size of a model and the number of training examples generally increases its accuracy, but also the energetic cost of its training. As these objectives are clearly conflicting, our idea is to rely on Pareto optimality, in order to evaluate a model according to both its accuracy and its environmental impact. Given two different models A and B with the same accuracy, if A is more energy-efficient than B, A is said to *dominate* B. If there is no better solution than A, it is *Pareto optimal*. Hence, we aim to find the set of all Pareto optimal models to form a Pareto front and remove non-optimal models.

As discussed earlier, measuring the accuracy of generative models is a daunting task. Here, we rely on the Mean Opinion Score (MOS) to illustrate our multi-objective proposal, as it is the most popular measure among our surveyed papers. As this score is highly dependent on each experimental setup, we compute $\%\text{MOS} = \frac{\text{MOS}_M}{\text{MOS}_{GT}}$ to allow more accurate comparisons, where MOS_M and MOS_{GT} stands respectively for the MOS obtained by the model and the one obtained by the respective "ground truth" from each original paper. The higher the perceived quality of the sound produced by the model, the closer this ratio will be to 1, and conversely the lower the perceived quality, the closer it will be to 0. The goal is to maximize this ratio, and thus to minimize $1 - \%\text{MOS}$. We consider this last measure as our subjective accuracy score.

For the energy-efficiency score, we separate training from inference. Regarding training, we take the previously introduced measure of energy consumption. Regarding inference, we rely on the number of parameters of the models. This count is highly correlated to the computational complexity and is independent of the device used to perform inference.

We display in Figure 1 the multi-objective space, where we plot the Pareto front for training (left) and for inference (right). The three models FloWaveNet, SING and WaveGAN are Pareto optimal in training, whereas WaveGAN is dominated by SING in inference and, therefore, is sub-optimal.

Since our goal is to propose a new tool for sustainable evaluation of models, we did not re-train the models to make our work more consistent and greener. Therefore, we would like to clarify that we rely on approximations and hand-crafted measures; these figures support our overall approach, but it warrants more extensive and reliable analyses, with a larger array of models. However, it should be noted that our approach is generic, and could be applied to any type of model or input data.

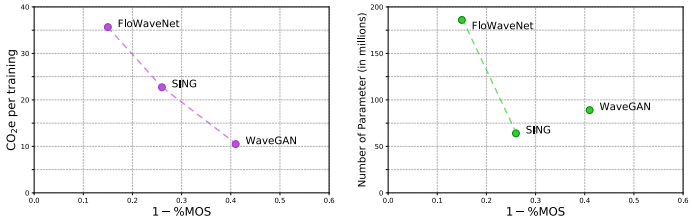

Figure 1. Representation of two Pareto Fronts (dotted lines). The objective is to minimize the subjective score ($1 - \%\text{MOS}$) along with the energy efficiency of either the training (left) with the measure of the carbon emission (CO_2e) per training, or the inference (right) with the number of parameters.

[†] https://www.carbonfootprint.com

References

[1] DARIO, A., AND DANNY, H. Ai and compute, 2018.

[2] DÉFOSSEZ, A., ZEGHIDOUR, N., USUNIER, N., BOTTOU, L., AND BACH, F. Sing: Symbol-to-instrument neural generator. *Advances in Neural Information Processing Systems 2018-Decem*, Nips (2018), 9041–9051.

[3] DONAHUE, C., MCAULEY, J., AND PUCKETTE, M. Adversarial audio synthesis. *7th International Conference on Learning Representations, ICLR 2019* (2019), 1–16.

[4] ENGEL, J., AGRAWAL, K. K., CHEN, S., GULRAJANI, I., DONAHUE, C., AND ROBERTS, A. GANSynth: Adversarial neural audio synthesis. *7th International Conference on Learning Representations, ICLR 2019* (2019), 1–17.

[5] HERNANDEZ, D., AND BROWN, T. B. Measuring the Algorithmic Efficiency of Neural Networks.

[6] KIM, T., LEE, J., AND NAM, J. Comparison and Analysis of SampleCNN Architectures for Audio Classification. *IEEE Journal on Selected Topics in Signal Processing 13*, 2 (2019), 285–297.

[7] LACOSTE, A., LUCCIONI, A., SCHMIDT, V., AND DANDRES, T. Quantifying the Carbon Emissions of Machine Learning.

[8] MEHRI, S., KUMAR, K., GULRAJANI, I., KUMAR, R., JAIN, S., SOTELO, J., COURVILLE, A., AND BENGIO, Y. Samplernn: An unconditional end-to-end neural audio generation model. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings* (2019), 1–11.

[9] SCHWARTZ, R., DODGE, J., SMITH, N. A., AND ETZIONI, O. Green AI. 1–12.

[10] THEIS, L., VAN DEN OORD, A., AND BETHGE, M. A note on the evaluation of generative models. *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings* (2016), 1–10.