TRANSFERABLE AUDIO LOTTERY TICKETS: GRADIENT ACCUMULATION FOR EXTREME SPARSITY

Hyunjae Kim, Juhan Nam, Kyung Myun Lee

Korea Advanced Institute of Science and Technology
Daejeon, Republic of Korea

{present, juhan.nam, kmlee2}@kaist.ac.kr

ABSTRACT

Large-scale neural networks have achieved impressive performance across diverse audio domains, but their growing size raises the need for lightweight alternatives. The Lottery Ticket Hypothesis (LTH) offers a compelling direction by revealing sparse subnetworks that can match full network performance, yet its potential across various audio subdomains remains underexplored. In this work, we examine this potential by evaluating sparse subnetworks on diverse audio classification tasks, spanning speech, music, and environmental sound. In particular, we propose a simple modification to the training process that incorporates momentum-like gradient accumulation during subnetwork search. We show that this strategy enables finding extremely sparse subnetworks with less than 1.0% of the initial parameters remaining, while still retaining up to 90% of dense model performance without layer collapse even under severe unstructured pruning. Furthermore, these subnetworks were effectively transferred across different audio subdomains while sustaining their sparsity-robust characteristics.

***Index Terms*—** Lottery Ticket Hypothesis, Pruning, Gradient Accumulation, Transferability, Audio Classification

1. INTRODUCTION

Recent advances in deep learning have led to remarkable performance across a wide range of tasks, often driven by highly over-parameterized models trained on massive datasets. However, their heavy computational and memory demands limit practical deployment in many real-world settings, prompting efforts to reduce model complexity without sacrificing performance. Among these efforts, the Lottery Ticket Hypothesis (LTH) has emerged as a compelling approach. It posits that within a sufficiently over-parameterized dense network, there already exist sparse subnetworks at initialization, referred to as *winning tickets (WT)*, that can reach the performance of the full network [1]. LTH has since been explored in various domains, demonstrating not only the feasibility of compact subnetworks but also their ability to transfer effectively across different datasets [2, 3, 4].

In the audio domain, LTH has been employed across various tasks including speech recognition, acoustic scene classification, and music information retrieval [5, 6, 7]. These studies confirm that WTs can indeed emerge in specific audio tasks, achieving competitive performance with only a fraction of the original model size. In some cases, they also suggest potential benefits of WTs in terms of noise robustness and transferability within the speech recognition task [6, 8]. However, these prior studies paid little attention to methods for discovering more effective subnetworks. Moreover, they have remained confined to individual subdomains such as speech or music only, and the possibility of audio-general LTH remains unexplored.

In this context, we explore the potential of audio-general WTs in classification tasks across three representative audio subdomains: speech, music, and environmental sound. We first show that WTs consistently emerge in each case. Furthermore, our proposed strategy of gradient accumulation in the WT search process enables the discovery of extremely sparse subnetworks, retaining as few as 1.0% of the original weights while still achieving comparable accuracy to the full model. Finally, we demonstrate that these subnetworks and their sparsity-robust characteristics are not restricted to their original datasets but can transfer between distinct audio subdomains.

2. RELATED WORK

For speech recognition, [6] showed that WTs are transferable across different ASR datasets even with improved noise robustness. [8] further reported that language-specific WTs share a larger fraction of weights across different languages and outperform dense models. In addition, [9] showed similar benefits in spoken language understanding. Although these studies provided partial evidence for the transferability of audio WTs, their scope remained limited to the speech domain. Beyond speech, LTH has also been applied to acoustic scene classification [7], music information retrieval [5], and audio-visual wake-word spotting [10], mainly for practical

lightweight deployment. While these studies demonstrate the effectiveness of LTH in audio tasks, their investigations are still limited to specific audio categories. In this work, we take a first step toward a cross-subdomain setting, examining transferable WTs across speech, music, and environmental sound.

In a separate line of work, recent studies have highlighted the importance of gradient handling in discovering effective subnetworks. [11] showed that preserving gradient flow at initialization yields more trainable subnetworks. [12] demonstrated that pruning masks stabilize early in training, indicating that gradient signals at this stage provide reliable cues for winning ticket search. Meanwhile, [13] argued that LTH does not exploit the benefit of enhanced gradient flow. Despite these insights, gradient-aware strategies have not been considered in audio-domain LTH; motivated by this gap, we introduce a simple gradient accumulation strategy that helps to find effective WTs under extreme sparsity.

3. PROPOSED APPROACH

We first overview the LTH procedure, then introduce a gradient accumulation strategy to discover sparsity-robust subnetworks, and finally outline transfer experiments across audio subdomains.

3.1. Lottery Ticket Hypothesis (LTH) for Audio

The Lottery Ticket Hypothesis (LTH), originally proposed by Frankle and Carbin [1], states that within a sufficiently over-parameterized network, there exist sparse subnetworks—called *winning tickets (WT)*—that, when trained in isolation from their original initialization, can match or even surpass the performance of the full dense model. WTs are identified by iteratively pruning a trained dense model and *rewinding* the remaining weights to their initial values.

Formally, consider a dense model $f(x;\theta)$ with parameters θ initialized as θ_0. After training for j iterations, the parameters are updated to θ_j. We then prune the network, retaining $(p \times 100)\%$ of the parameters by magnitude, and obtain a binary mask $m \in \{0,1\}^{|\theta|}$. The surviving parameters are rewound to their initial values from θ_0, resulting in a pruned model $f(x; m \odot \theta_0)$.

Repeating this procedure r times (i.e. r rounds) produces progressively sparser subnetworks, $f(x; m_r \odot \theta_0)$, with mask m_r and sparsity $s_r = 1 - p^r$. Among these subnetworks, those that achieve performance comparable to or surpassing the dense model are regarded as WTs. In this work, we apply iterative magnitude pruning up to $r = 15$ rounds with $p = 0.6$.

As baselines, we include unstructured magnitude pruning (UMP) and layer-wise magnitude pruning (LMP) [14, 15], evaluated at the same sparsity levels as the corresponding WTs.

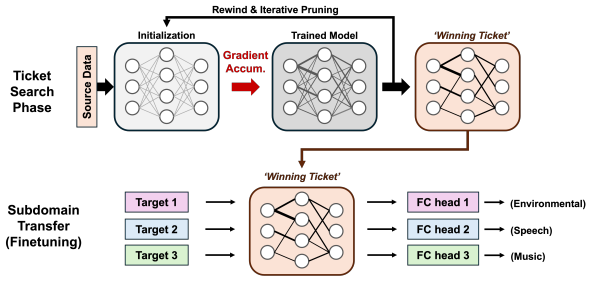

Fig. 1. Schematic illustration of winning ticket discovery with gradient accumulation and subdomain transfer.

3.2. Gradient Accumulation for LTH (GA-LTH)

Sparse subnetworks often suffer from poor gradient flow, leading to unstable optimization [16]. Although the rewind step in LTH facilitates convergence [13], it does not address gradient flow itself, which remains a separate limitation. To facilitate more effective WT search, we propose a momentum-like gradient accumulation strategy (GA-LTH). During the ticket search phase, the accumulated gradient at each training iteration is computed as

$$\tilde{g}_t = g_t + \alpha\, \tilde{g}_{t-1}, \tag{1}$$

and the parameters are updated by

$$\theta_{t+1} = \theta_t - \eta_t\, H_t(\tilde{g}_t), \tag{2}$$

where g_t denotes the gradient at training step t, α is a decay factor controlling the contribution of past gradients, and $H_t(\cdot)$ represents the optimizer-specific preconditioning (e.g., exponential moving average of gradients and squared gradients in AdamW). When applied with plain SGD, this update becomes equivalent to SGD with momentum. In our study, however, we employ the AdamW optimizer to ensure stable training, while leveraging the proposed gradient accumulation to amplify the effective learning dynamics of sparse models. Although mechanisms such as gradient clipping are often required to prevent gradient explosion, in our training setup we did not observe such instability and found that controlling the decay factor α was sufficient to maintain stable optimization.

3.3. Transferable Winning Tickets

Unlike prior audio LTH studies that mainly focused on intra-domain conditions, we examine the transferability of WTs to out-of-domain datasets by evaluating them across distinct audio subdomains—speech, music, and environmental sound (Fig. 1). In the finetuning phase, subnetworks derived from three source datasets are transferred to three additional target datasets for each sparsity level. The backbone parameters and pruning masks obtained from the source models are employed, while the output head is newly initialized to match the label space of the target dataset. Training is then performed on the entire network, excluding the pruned weights.

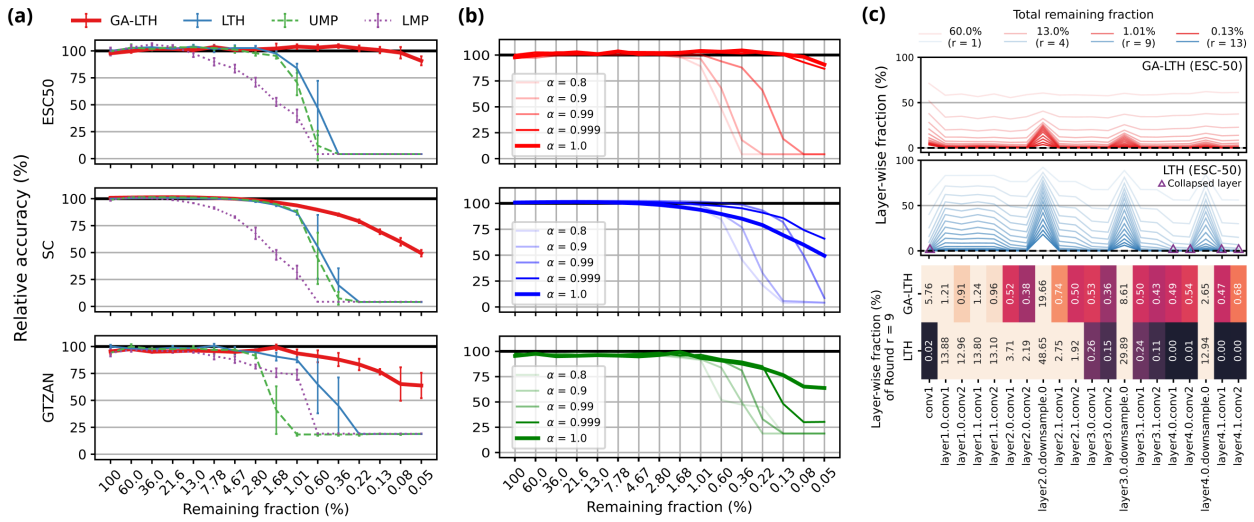

Fig. 2. Effects of gradient accumulation on winning ticket discovery. (a) Relative accuracy across sparsity levels for GA-LTH, LTH, and pruning baselines (UMP, LMP) on ESC-50, SC, and GTZAN. (b) Effect of accumulation decay factor α. (c) Layer-wise sparsity of WTs from ESC-50; the heatmap shows fractions of subnetworks at 1.01% total remaining fraction ($r = 9$).

4. EXPERIMENTAL SETUP

4.1. Dataset

We selected three representative audio classification datasets as source datasets: ESC-50 for environmental sound, Speech Commands for speech, and GTZAN for music. To further evaluate the transferability of WTs from source datasets, we adopted three additional target datasets—UrbanSound8k, LibriCount, and Nsynth. A brief description is as follows: **ESC-50** [17]: 2,000 five-second clips across 50 environmental sound classes; official 5-fold split used. **Speech Commands v0.02 (SC)** [18]: 105,829 one-second utterances of 35 words; official train/validation/test split used. **GTZAN** [19]: 1,000 thirty-second clips spanning 10 music genres; following fault-filtered split of [20] adopted. **LibriCount (LC)** [21]: 5,720 five-second clips derived from LibriSpeech, 11-class speaker-count task (0–10); HEAR configuration adopted [22]. **UrbanSound8K (US8k)** [23]: 8,732 clips (≤4 s) over 10 everyday sound classes; official 10-fold split used. **NSynth-pitch (NSynth)** [24]: 5,000 four-second clips of instrument notes across diverse instruments, labeled with 88 pitch classes (MIDI 21–108); HEAR configuration adopted [22].

4.2. Implementation Details

To control input size variability, all audio clips were downsampled to 16 kHz and converted into mel-spectrograms with 64 mel bins using 32 ms window and 8 ms hop. For training, 1-s segments (128 frames) were randomly cropped from each clip, while testing used fixed center crops, except for NSynth where cropping always started at the beginning.

We employed a ResNet18 model [25] with single-channel mel-spectrogram input. Models were trained for 5,000 iterations with a batch size of 64 using AdamW (β_1=0.9, β_2=0.999, learning rate = 1e-4, weight decay = 3e-4). Early stopping with a patience of 2,000 iterations was applied. All experiments were conducted with four different random seeds and averaged.

For the ticket search phase, LTH pruning iteration was performed for 15 rounds, retaining 60% of the weights per round, resulting in subnetworks with remaining fraction down to about $0.6^{15} \times 100 \approx 0.05\%$ at round $r = 15$. Pruning was applied only to the convolutional backbone, excluding the final fully connected (FC) layer to ensure compatibility across datasets. For all datasets, the dense model baselines were obtained separately by standard training (without pruning or gradient accumulation), and relative accuracy was computed as $\frac{\text{subnetwork accuracy}}{\text{dense accuracy}} \times 100\%$.

As a default, gradient accumulation was applied with decay factor $\alpha = 1.0$ except in the α manipulation experiment (Fig. 2b). Thus, GA-LTH indicates WTs discovered with $\alpha = 1.0$, while standard LTH corresponds to $\alpha = 0.0$. As pruning baselines, we use unstructured magnitude pruning (UMP) and layer-wise magnitude pruning (LMP), which prune the trained dense model once and then fine-tune the pruned network.

5. RESULT AND DISCUSSION

5.1. Effects of Gradient Accumulation for Audio LTH

First, we investigate the existence of WTs across three source datasets and assess the impact of gradient accumulation in audio classification. As shown in Fig. 2(a), we observe that subnetworks maintain comparable performance in both GA-

		UrbanSound8k			LibriCount			Nsynth-pitch		
		remaining fraction (%)			remaining fraction (%)			remaining fraction (%)		
	Source Dataset	13.0	1.01	0.13	13.0	1.01	0.13	13.0	1.03	0.13
GA-LTH	ESC50	68.1	**68.3**	**66.2**	58.9	**58.1**	**54.2**	74.8	**73.8**	**70.3**
	Speech Commands	67.9	66.0	58.4	59.0	57.1	50.5	74.4	73.7	64.4
	GTZAN	67.8	65.5	49.8	57.6	56.2	45.7	74.9	73.6	34.3
LTH	ESC50	68.0	60.5	12.0	58.5	53.8	8.4	74.4	64.7	1.3
	Speech Commands	66.8	62.5	12.0	58.6	54.8	8.4	75.0	66.1	1.3
	GTZAN	68.3	63.1	12.0	57.6	54.5	8.4	74.8	67.8	1.3
UMP		**68.4**	16.4	12.0	**60.3**	51.7	9.1	**75.8**	66.1	1.4
LMP		65.9	47.8	13.2	57.6	47.0	9.9	71.5	9.9	1.1

Table 1. Transfer experiment results across target datasets under three sparsity levels (13.0%, 1.01%, and 0.13%) for subnetworks obtained from source datasets by GA-LTH and LTH, and pruning baselines (UMP, LMP).

LTH and standard LTH up to pruning round $r = 8$ (1.68% remaining fraction) across all three source datasets, indicating that WTs exist across audio subdomains even if the gains over UMP are modest.

However, at sparsity levels beyond 99% (round $r \geq 9$), all other pruning methods show a sharp accuracy drop, while GA-LTH clearly sustains performance even under extreme compression. Notably, on ESC-50, GA-LTH discovers WTs that retain near-dense accuracy with only 0.08% of parameters (Fig. 2(a)).

Fig. 2(c) additionally shows that standard LTH suffers from layer collapse (black boxes in the heatmap), whereas GA-LTH avoids such collapse by preserving a balanced layer-wise distribution even in an unstructured pruning manner. Importantly, while LMP inherently avoids collapse, it still exhibits poor sparsity-robust performance, suggesting that GA-LTH does not simply prevent collapse but also facilitates the discovery of more meaningful subnetwork structures.

To further investigate the effects of gradient accumulation, we varied its decay factor (Fig. 2b). The results show a general tendency that larger accumulation (higher decay factors) improves sparsity-robustness. Nevertheless, the effect of accumulation varies across datasets: ESC-50 achieves its best WT performance at $\alpha = 1.0$, whereas SC performs best at $\alpha = 0.99$ and declines at $\alpha = 1.0$. These observations suggest that proper gradient handling supports effective ticket search, but the optimal setting can depend on the specific subdomain, dataset, or experimental setup.

For reference, the dense model baselines reach 47.5% on ESC-50, 92.9% on SC, and 56.9% on GTZAN. It should be noted that all experiments were conducted on downsampled, 1-s cropped audio; thus, our experimental goal is not to achieve state-of-the-art performance but to ensure comparability under consistent experimental settings.

5.2. Transferability of Winning Tickets

The results of transfer experiments are presented in Table 1, where WTs identified from ESC-50, SC, and GTZAN are fine-tuned on US8k, LC, and NSynth. The dense model baselines are 68.8% on US8k, 59.0% on LC, and 77.0% on NSynth, obtained under the same controlled conditions as above. We report results under three sparsity levels: 13.0%, 1.01%, and 0.13% remaining fractions.

At 13.0% remaining, no notable differences are observed among pruning methods. When only 1.01% of the parameters are retained, both GA-LTH and LTH sustain relatively strong performance, with GA-LTH exhibiting greater stability, whereas UMP and LMP undergo substantial accuracy drops, particularly on US8k. Finally, at the extreme sparsity of 0.13% remaining, GA-LTH demonstrates remarkable robustness: ESC-50 tickets exhibit only a 2–5% accuracy drop, while all other methods degrade to chance-level performance. These results suggest that LTH-based approaches can uncover subnetworks with more audio-general representations than conventional pruning methods. In particular, our proposed gradient accumulation strategy facilitates the discovery of extremely sparse and transferable WTs across diverse audio subdomains.

6. CONCLUSION

In this work, we investigated the potential of LTH-based approaches for sparsity-robustness and transferability in audio classification tasks. By incorporating gradient accumulation into the LTH procedure, we identified extremely sparse subnetworks that sustain comparable accuracy even with less than 1.0% of the parameters without suffering from layer collapse. Moreover, our transfer experiments demonstrate that these subnetworks can be effectively applied across distinct audio subdomains while preserving their sparsity-robust characteristics, underscoring the potential of LTH approaches as domain-general audio representations. As future work, we plan to further extend these approaches in more practical scenarios with advanced architectures and pretrained foundation models.

7. ACKNOWLEDGMENTS

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (RS-2023-00222383).

8. REFERENCES

[1] Jonathan Frankle and Michael Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *International Conference on Learning Representations (ICLR)*, 2019.

[2] Ari S. Morcos, Haonan Yu, Michela Paganini, and Yuandong Tian, "One ticket to win them all: Generalizing lottery ticket initializations across datasets and optimizers," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[3] Sharath Girish, Shishira R. Maiya, Kamal Gupta, Hao Chen, Larry Davis, and Abhinav Shrivastava, "The lottery ticket hypothesis for object recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 762–771.

[4] Arthur da Cunha, Emanuele Natale, and Laurent Viennot, "Proving the lottery ticket hypothesis for convolutional neural networks," in *International Conference on Learning Representations (ICLR)*, 2022.

[5] Philippe Esling, Theis Bazin, Adrien Bitton, Tristan Carsault, and Ninon Devis, "Ultra-light deep mir by trimming lottery tickets," in *International Society for Music Information Retrieval Conference (ISMIR)*, 2020.

[6] Shaojin Ding, Tianlong Chen, and Zhangyang Wang, "Audio lottery: Speech recognition made ultra-lightweight, transferable, and noise-robust," in *International Conference on Learning Representations (ICLR)*, 2022.

[7] Hao Yen, Chao-Han Huck Yang, Hu Hu, Sabato Marco Siniscalchi, Qing Wang, Yuyang Wang, Xianjun Xia, Yuanjun Zhao, Yuzhong Wu, Yannan Wang, Jun Du, and Chin-Hui Lee, "A lottery ticket hypothesis framework for low-complexity device-robust neural acoustic scene classification," in *DCASE Workshop*, 2021.

[8] Mu Yang, Andros Tjandra, Chunxi Liu, David Zhang, Duc Le, and Ozlem Kalinli, "Learning asr pathways: A sparse multilingual asr model," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[9] Hayato Futami, Siddhant Arora, Yosuke Kashiwagi, Emiru Tsunoo, and Shinji Watanabe, "Finding task-specific subnetworks in multi-task spoken language understanding model," in *Interspeech*, 2024.

[10] Hengshun Zhou, Jun Du, Chao-Han Huck Yang, Shifu Xiong, and Chin-Hui Lee, "A study of designing compact audio-visual wake word spotting system based on iterative fine-tuning in neural network pruning," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7572–7576.

[11] Chaoqi Wang, Guodong Zhang, and Roger Grosse, "Picking winning tickets before training by preserving gradient flow," in *International Conference on Learning Representations (ICLR)*, 2020.

[12] Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Richard G. Baraniuk, Zhangyang Wang, and Yingyan Celine Lin, "Drawing early-bird tickets: Towards more efficient training of deep networks," in *International Conference on Learning Representations (ICLR)*, 2020.

[13] Utku Evci, Yani A. Ioannou, Cem Keskin, and Yann Dauphin, "Gradient flow in sparse neural networks and how lottery tickets win," in *AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[14] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag, "What is the state of neural network pruning?," *Proceedings of machine learning and systems*, vol. 2, pp. 129–146, 2020.

[15] Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi, "A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[16] Kale-ab Tessera, Sara Hooker, and Benjamin Rosman, "Keep the gradients flowing: Using gradient flow to study sparse network optimization," *arXiv preprint arXiv:2102.01670*, 2021.

[17] Karol J. Piczak, "Esc: Dataset for environmental sound classification," in *ACM International Conference on Multimedia*, 2015, pp. 1015–1018.

[18] Pete Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[19] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[20] Corey Kereliuk, Bob L. Sturm, and Jan Larsen, "Deep learning and music adversaries," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2059–2071, 2015.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[22] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, Max Henry, Nicolas Pinto, Camille Noufi, Christian Clough, Dorien Herremans, Eduardo Fonseca, Jesse Engel, Justin Salamon, Philippe Esling, Pranay Manocha, Shinji Watanabe, Zeyu Jin, and Yonatan Bisk, "Hear: Holistic evaluation of audio representations," in *Proceedings of the NeurIPS 2021 Competitions and Demonstrations Track*. 2022, pp. 125–145, PMLR.

[23] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello, "A dataset and taxonomy for urban sound research," in *ACM International Conference on Multimedia*, 2014, pp. 1041–1044.

[24] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan, "Neural audio synthesis of musical notes with wavenet autoencoders," in *International Conference on Machine Learning (ICML)*, 2017, pp. 1068–1077.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.