GANS CAN PLAY LOTTERY TICKETS TOO

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep generative adversarial networks (GANs) have gained growing popularity in numerous scenarios, while usually suffer from high parameter complexities for resource-constrained real-world applications. However, the compression of GANs has less been explored. A few works show that heuristically applying compression techniques normally leads to unsatisfactory results, due to the notorious training instability of GANs. In parallel, the *lottery ticket hypothesis* shows prevailing success on discriminative models, in locating sparse *matching subnetworks* capable of training in isolation to full model performance. In this work, we for the first time study the existence of such trainable matching subnetworks in deep GANs. For a range of GANs, we certainly find matching subnetworks at 67%-74% sparsity. We observe that with or without pruning discriminator has a minor effect on the existence and quality of matching subnetworks, while the initialization weights used in the discriminator plays a significant role. We then show the powerful transferability of these subnetworks to unseen tasks. Furthermore, extensive experimental results demonstrate that our found subnetworks substantially outperform previous state-of-the-art GAN compression approaches in both image generation (e.g. SNGAN) and image-to-image translation GANs (e.g. CycleGAN).

1 INTRODUCTION

Generative adversarial networks (GANs) have been successfully applied to many fields like image translation (Jing et al., 2019; Isola et al., 2018; Liu & Tuzel, 2016; Shrivastava et al., 2017; Zhu et al., 2017) and image generation (Miyato et al., 2018; Radford et al., 2016; Gulrajani et al., 2017; Arjovsky et al., 2017). However, they are often heavily parameterized and often require intensive calculation at the training and inference phase. Network compressing techniques (Lecun, 1990; Wang et al., 2019; 2020b; Li et al., 2020) can be of help at inference by reducing the number of parameters or usage of memory; nonetheless, they can not save computational burden at no cost. Although they strive to maintain the performance after compressing the model, a non-negligible drop in generative capacity is usually observed. A question is raised:

Is there any way to compress a GAN model while preserving or even improving its performance?

The lottery ticket hypothesis (LTH) (Frankle & Carbin, 2019a) provides positive answers with *matching subnetworks* (Chen et al., 2020). It states that there exist matching subnetworks in dense models that can be trained to reach a comparable test accuracy to the full model within similar training iterations. The hypothesis has successfully shown its success in various fields (Yu et al., 2019; Renda et al., 2020; Chen et al., 2020), and its property has been studied widely (Malach et al., 2020; Pensia et al., 2020; Elesedy et al., 2020). However, it is never introduced to GANs and therefore the presence of matching subnetworks in generative adversarial networks still remains mysterious.

To address this gap in the literature, we investigate the lottery ticket hypothesis in GANs. One most critical challenge of extending LTH in GANs emerges: how to deal with the discriminator during compressing the generator, including *(i) whether prunes the discriminator simultaneously* and *(ii) what initialization should be adopted by discriminators during the re-training?* Previous GAN compression methods (Shu et al., 2019; Wang et al., 2019; Li et al., 2020; Wang et al., 2020b) prune the generator model only since they aim at reducing parameters in the inference stage. The effect of pruning the discriminator has never been studied by these works, which is unnecessary for them but possibly essential in finding matching subnetworks. It is because that finding matching subnetworks involves re-training the whole GAN network, in which an imbalance in generative and

discriminative power could result in degraded training results. For the same reason, the disequilibrium between initialization used in generators and discriminators incurs severe training instability and unsatisfactory results.

Another attractive property of LTH is the powerful transferability of located matching subnetworks. Although it has been well studied in discriminative models (Mehta, 2019; Morcos et al., 2019; Chen et al., 2020), an in-depth understanding of transfer learning in GAN tickets is still missing. In this work, we not only show whether the sparse matching subnetworks in GANs can transfer across multiple datasets but also study what initialization benefits more to the transferability.

To convert parameter efficiency of LTH into the advantage of computational saving, we also utilize channel pruning (He et al., 2017b) to find the structural matching subnetworks of GANs which enjoys the bonus of accelerated training and inference. Our contributions can be summarized in the following four aspects:

- Using unstructured magnitude pruning, we identify matching subnetworks at 74% sparsity in SNGAN (Miyato et al., 2018) and 67% in CycleGAN (Zhu et al., 2017). The matching subnetworks in GANs exist no matter whether pruning discriminators, while the initialization weights used in the discriminator are crucial.
- We show that the matching subnetworks found by iterative magnitude pruning outperform subnetworks extracted by randomly pruning and random initialization, in terms of the extreme sparsity and performance. To fully exploit the trained discriminator, we using the dense discriminator as a distillation source and further improve the quality of winning tickets.
- We demonstrate that the found subnetworks in GANs transfer well across diverse generative tasks.
- The matching subnetworks found by channel pruning surpass previous state-of-the-art GAN compression methods (i.e. GAN Slimming (Wang et al., 2020b)) in both efficiency and performance.

2 Related Work

GAN Compression Generative adversarial networks (GANs) have succeeded in computer vision fields, for example, image generation and translation. One significant drawback of the generative models is the high computational cost of the models' complex structure. A wide range of neural network compression techniques has been applied to generative models to address this problem. There are several categories of compression techniques, including pruning (removing some parameters), quantization (reducing the bit width), and distillation. Shu et al. (2019) proposed a channel pruning method for CycleGAN by using a co-evolution algorithm. Wang et al. (2019) proposed a quantization method for GANs based on the EM algorithm. Li et al. (2020) used a distillation method to transfer knowledge of the dense to the compressed model. Recently Wang et al. (2020b) proposed a GAN compression framework, GAN slimming, that integrated the above three mainstream compression techniques into a unified form. Previous works on GAN pruning usually aim at finding a sparse structure of the trained generator model for faster inference speed, while we are focusing on finding trainable structures of GANs following the lottery ticket hypothesis. Moreover, in existing GAN compression methods, only the generator is pruned, which could undermine the performance of re-training since the left-out discriminator may have a stronger computational ability than the pruned generator and therefore cause a degraded result due to the imparity of these two models.

The Lottery Ticket Hypothesis The lottery ticket hypothesis (LTH) (Frankle & Carbin, 2019b) claims the existence of sparse, separate trainable sub-networks in a dense network. These subnetworks are capable of reaching comparable or even better performance than full dense model, which has been evidenced in various fields, such as image classification (Frankle & Carbin, 2019b; Liu et al., 2019; Wang et al., 2020a; Evci et al., 2019; Frankle et al., 2020; Savarese et al., 2020; Yin et al., 2020; You et al., 2020), natural language processing (Gale et al., 2019), reinforcement learning (Yu et al., 2019), and adversarial robustness (Cosentino et al., 2019). Most works of LTH use unstructured weight magnitude pruning (Han et al., 2015; Frankle & Carbin, 2019b) to find the matching subnetworks, and the channel pruning also adopted in a recent work (You et al., 2019). In order to scale up LTH to larger networks and datasets, the "late rewinding" technique is proposed by Frankle et al. (2019b); Renda et al. (2020). Mehta (2019); Morcos et al. (2019); Desai et al. (2019)

are the pioneers to study the transferability of found subnetworks. However, all previous works focus on discriminative models. In this paper, we extend LTH to GANs and reveal unique findings of GAN tickets.

3 PRELIMINARIES

In this section, we describe our pruning algorithms and list related experimental settings.

Backbone Networks We use two GANs in our experiments: SNGAN (Miyato et al., 2018) and CycleGAN (Zhu et al., 2017)). SNGAN with ResNet (He et al., 2015) is one of the most popular noise-to-image GAN network and has state-of-the-art performance on several datasets like CIFAR-10. CycleGAN is a popular and well-studied image-to-image GAN network that also achieve state-of-the-art on several benchmarks. For SNGAN, let $g(\mathbf{x}; \boldsymbol{\theta_g})$ be the output of the generator network $\mathcal{G}$ with parameter $\boldsymbol{\theta_g}$ and input example $\mathbf{x} \in \mathbb{R}^{d_1}$ and $d(\mathbf{x}; \boldsymbol{\theta_d})$ be the output of the discriminator network $\mathcal{D}$ with parameter $\boldsymbol{\theta_d}$ and input example $\mathbf{x} \in \mathbb{R}^{d_2}$. For CycleGAN which is composed of two generator-discriminator pairs, we use $g(\mathbf{x}; \boldsymbol{\theta_g})$ and $\boldsymbol{\theta_g}$ again to represent the output and the weights of the two generators where $\mathbf{x} \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_1}$. The same modification can be done for the two discriminators in CycleGAN.

Datasets For image-to-image experiments, we use a widely-used benchmark horse2zebra (Zhu et al., 2017) for model training. As for noise-to-image experiments, we use CIFAR-10 (Krizhevsky et al., 2009) as the benchmark. For the transfer study, the experiments are conducted on CIFAR-10 and STL-10 (Coates et al., 2011). For better transferring, we resize the image in STL-10 to 32×32.

Subnetworks For a network $f(\cdot; \boldsymbol{\theta})$, a subnetwork is defined as $f(\cdot, \boldsymbol{m} \odot \boldsymbol{\theta})$, where $\boldsymbol{m} \in \{0, 1\}^d$ is a pruning mask for $\boldsymbol{\theta} \in \mathbb{R}^{d_1}$ and $\odot$ is the element-wise product. For GANs, two separate masks, $\boldsymbol{m_d}$ and $\boldsymbol{m_g}$, are needed for both the generator and the discriminator. Consequently, a subnetwork of GANs is consistent of: a sparse generator $g(\cdot; \boldsymbol{m_g} \odot \boldsymbol{\theta_g})$ and a sparse discriminator $d(\cdot; \boldsymbol{m_d} \odot \boldsymbol{\theta_d})$.

Following Frankle et al. (2019a), we define a *matching network* as a subnetwork that can be trained to the same accuracy as the full network within a similar iterations as training the full network; a *winning ticket* is defined as a matching subnetwork $f(\cdot; \boldsymbol{m} \odot \boldsymbol{\theta})$ where $\boldsymbol{\theta} = \boldsymbol{\theta_0}$.

Finding subnetworks Finding GAN subnetworks is to find two masks $\boldsymbol{m_g}$ and $\boldsymbol{m_d}$ for the generator and the discriminator. We use both an unstructured magnitude and structured pruning, the channel pruning (He et al., 2017a), to generate the masks.

For unstructured pruning, we follow the following steps. After we finish training the full GAN model for N iterations, we prune the weights with the lowest magnitude globally (Han et al., 2015) to obtain masks $\boldsymbol{m} = (\boldsymbol{m_g}, \boldsymbol{m_d})$, where the position of a remaining weight in $\boldsymbol{m}$ is marked as one, and the position of a pruned weight is marked as zero. The weights of the sparse generator and the sparse discriminator are then reset to the initial weights of the full network. Previous works have shown that the iterative magnitude pruning (IMP) method is better than the one-shot pruning method. So rather than pruning the network only once to reach the desired sparsity, we prune a certain amount of non-zero parameters and re-train the network several times to meet the requirement. Details of this algorithm are in Appendix A.1, Algorithm 1.

As for channel pruning, the first step is to train the full model as well. Besides using a normal loss function $\mathcal{L}_{\text{GAN}}$, we follow Liu et al. (2017) to apply a ℓ_1-norm on the trainable scale parameters γ in the normalization layers to encourage channel-level sparsity: $\mathcal{L}_{\text{cp}} = ||\gamma||_1$. To prevent the compressed network behave severely differently with the original large network, we introduce a distillation loss as Wang et al. (2020b) did: $\mathcal{L}_{\text{dist}} = \mathbb{E}_{\mathbf{x}}[\text{dist}(g(\mathbf{x}; \boldsymbol{\theta_g}), g(\mathbf{x}; \boldsymbol{m_g} \odot \boldsymbol{\theta_g}))]$. We train the GAN network with these two additional losses for N_1 epochs and get the sparse networks $g(\cdot; \boldsymbol{m_g} \odot \boldsymbol{\theta_g})$ and $d(\cdot; \boldsymbol{m_d} \odot \boldsymbol{\theta_g})$. Details of this algorithm are in Appendix A.1, Algorithm 2.

Evaluation of subnetworks After obtaining the subnetworks $g(\cdot; \boldsymbol{\theta_g} \odot \boldsymbol{m_g})$ and $d(\cdot; \boldsymbol{\theta_d} \odot \boldsymbol{m_d})$, we test whether the subnetworks are matching or not. We reset the weights to a specific step i, and train the subnetworks for N iterations and evaluate them using two specific metrics, Inception Score (Salimans et al., 2016) and Fréchet Inception Distance (Heusel et al., 2017).

Other Pruning Methods We compare the size and the performance of subnetworks found by IMP with subnetworks found by other techniques that aim at compressing the network after training to reduce computational costs at inference. We use a benchmark pruning approach named *Standard Pruning*, which iteratively prune the 20% of lowest magnitude weights, and train the network for N iterations from there (without any rewinding) until we have reached the target sparsity.

In order to verify that the statement of iterative magnitude pruning is better than one-shot pruning, we compare $\mathrm{IMP_G}$ and $\mathrm{IMP_{GD}}$ with their one-shot counterparts. Additionally, we compare IMP with some randomly pruning techniques to prove the effectiveness of IMP. They are: 1) *Randomly Pruning*: Randomly generate a sparsity mask $\boldsymbol{m'}$. 2) *Random Tickets*: Rewinding the weights to another initialization $\boldsymbol{\theta'_0}$.

4 The Existence of Winning Tickets in GAN

In this section, we will validate the existence of winning tickets in GANs with pre-trained initialization $\boldsymbol{\theta_0} := (\boldsymbol{\theta_{g_0}}, \boldsymbol{\theta_{d_0}})$. Specifically, we will empirically prove several important properties of the tickets by authenticating the following four claims:

Claim 1: Iterative Magnitude Pruning (IMP) finds winning tickets in GANs, $g(\mathbf{x}; \boldsymbol{m_g} \odot \boldsymbol{\theta_{g_0}})$ and $d(\mathbf{x}; \boldsymbol{m_d} \odot \boldsymbol{\theta_{d_0}})$. Channel pruning is also able to find wining tickets as well.

Claim 2: Whether pruning the discriminator $\mathcal{D}$ does not change the existence of winning tickets. It is the initialization used in $\mathcal{D}$ that matters. Moreover, pruning the discriminator has a slight boost of matching networks in terms of extreme sparsity and performance.

Claim 3: IMP finds winning tickets at sparsity where some other pruning methods (randomly pruning, one-shot magnitude pruning, and random tickets) are not matching.

Claim 4: The late rewinding technique (Frankle et al., 2019b) helps. Matching subnetworks that are initialized to $\boldsymbol{\theta_i}$, *i.e.*, i steps from $\boldsymbol{\theta_0}$, can outperform those initialized to $\boldsymbol{\theta_0}$. Moreover, matching subnetworks that are late rewound can be trained to match the performance of standard pruning.

Claim 1: Are there winning tickets in GANs? To answer this question, we first conduct experiments on SNGAN by pruning the generator only in the following steps: 1) Run IMP to get sequential sparsity masks $(\boldsymbol{m_{d_i}}, \boldsymbol{m_{g_i}})$ of sparsity $s_i\%$ remaining weights; 2) Apply the masks to the GAN and reset the weights of the subnetworks to the pre-trained initialization $\boldsymbol{\theta_0}$; 3) Train models to evaluate whether they are winning tickets. [1] We set $s_i\% = (1-0.8^i) \times 100\%$, which we use for all the experiments that involve iteratively pruning hereinafter. The number of training epochs for subnetworks is identical to that of training the full models.

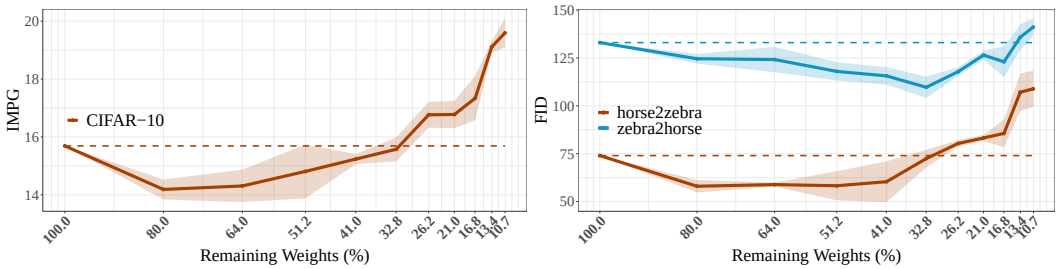

Figure 1: The Fréchet Inception Distance (FID) curve of subnetworks of SNGAN (left) and CycleGAN (right) generated by iterative magnitude pruning (IMP) on CIFAR-10 and horse2zebra. The dashed line indicates the FID score of the full model on CIFAR-10 and horse2zebra. The 95% confidence interval of 5 runs is reported.

Figure 1 verifies the existence of winning tickets in SNGAN and CycleGAN. We are able to find winning tickets by iterative pruning the generators at highest sparsity, around 74% in SNGAN, and around 67% in CycleGAN, where the FID scores of these subnetworks successfully match the FID scores of the full network respectively. The confidence interval also suggests that the winning tickets at some sparsities are statistically significantly better than the full model.

[1] The detailed description of algorithms are listed in Appendix A.1.

To show that channel pruning can find winning tickets as well, we extract several subnetworks from the trained full SNGAN and CycleGAN by varying ρ in Algorithm 2. We define the channel-pruned model's sparsity as the ratio of MFLOPs between the sparse model and the full model.

We can confirm that winning tickets can also be found by channel pruning (CP). CP is able to find winning tickets in SNGAN at sparsity around 34%. We will analyze it more carefully in Section 6.

Figure 2: Visualization by sampling and interpolation of SNGAN Winning Tickets found by IMP. Sparsity of best winning tickets : 48.80%. Extreme sparsity of matching subnetworks: 73.79%.

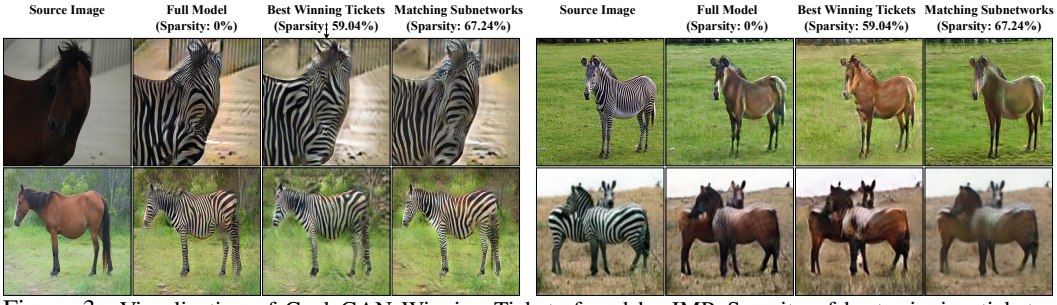

Figure 3: Visualization of CycleGAN Winning Tickets found by IMP. Sparsity of best winning tickets : 59.04%. Extreme sparsity of matching subnetworks: 67.24%. Left: visualization results on horse2zebra. Right: visualization results on zebra2horse.

Claim 2: Does the treatment of the discriminator affect the existence of winning tickets? Previous works of GAN pruning did not analyze the effect of pruning the discriminator. To study the effect, we compare two different iterative pruning settings: 1) Prune the generator only ($\mathrm{IMP_G}$) and 2) Prune both the generator and the discriminator iteratively ($\mathrm{IMP_{GD}}$). Both the generator and the discriminator are reset to the pre-trained initialization $\boldsymbol{\theta_0}$ after the masks are obtained.

The FID scores of the two experiments are shown in Figure 4. The graph suggests that the two settings share similar patterns: the minimal FID of $\mathrm{IMP_G}$ is 14.52, and the minimal FID of $\mathrm{IMP_{GD}}$ is 14.59. The difference between these two best FID is only 0.07, showing a trivial difference in generative power. The FID curve of $\mathrm{IMP_G}$ lies below that of $\mathrm{IMP_{GD}}$ at low sparsity but lies above at high sparsity, indicating that pruning the discriminator produces slightly better performance when the percent of remaining weights is small. The extreme sparsity where $\mathrm{IMP_{GD}}$ can match the performance of the full model is 73.8%. In contrast, $\mathrm{IMP_G}$ can only match no sparser than 67.2%, demonstrating that pruning the discriminator can also push the frontier of extreme sparsity where the pruned models are still able to match.

In addition, we study the effects of different initialization for the sparse discriminator. We compare different weights loading methods when applying the iterative magnitude pruning process: 1) Reset the weights of generator to $\boldsymbol{\theta_{g_0}}$ and reset the weights of discriminator to $\boldsymbol{\theta_{d_0}}$, which is identical to $\mathrm{IMP_G}$; 2) Reset the weights of generator to $\boldsymbol{\theta_{g_0}}$ and **fine-tune** the discriminator, which we will call $\mathrm{IMP_G^F}$. Figure 4 shows that resetting both the weights to $\boldsymbol{\theta_0}$ produces a much better result than only resetting the generator. The discriminator $\mathcal{D}$ without resetting its weights is too strong for the generator $\mathcal{G}$ with initial weights $\boldsymbol{\theta_{g_0}}$ that will lead to degraded performance. In summary, initialization benefits finding winning tickets.

A follow-up question arises from the previous observations: is there any way to use the weights of the dense discriminator, as the direct usage of the dense weights yields inferior results? One possible way is to use it as a "teacher" and transfer the knowledge to pruned discriminator using a consistency loss. Formally speaking, an additional regularization term is used when training the whole network:

$$\mathcal{L}_{\mathrm{KD}}(\mathbf{x}; \boldsymbol{\theta_d}, \boldsymbol{m_d}) = \mathbb{E}_{\mathbf{x}}[\mathrm{KL_{Div}}(d(\mathbf{x}; \boldsymbol{m_d} \odot \boldsymbol{\theta_d}), d(\mathbf{x}; \boldsymbol{\theta_{d_1}}))]$$

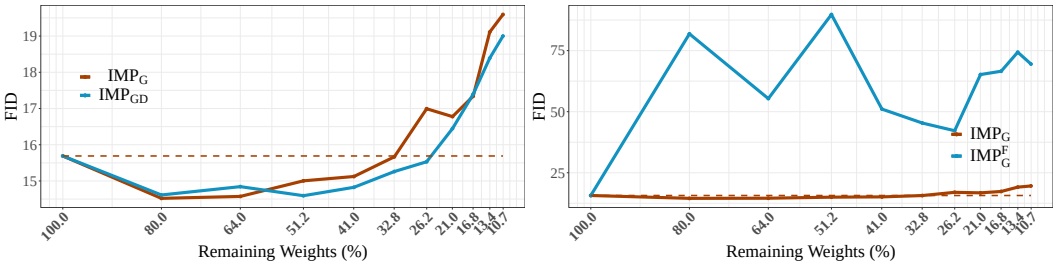

Figure 4: The FID score of Left: The FID score of best subnetworks generated by two different pruning settings: $\mathrm{IMP_G}$ and $\mathrm{IMP_{GD}}$. Right: The FID score of best subnetworks generated by two different pruning settings: $\mathrm{IMP_G}$ and $\mathrm{IMP_G^F}$.

where $\mathrm{KL_{Div}}$ denotes the KL-Divergence. We name the iterative pruning method with this additional regularization $\mathrm{IMP_{GD}^{KD}}$.

Figure 5 shows the result of pruning method $\mathrm{IMP_{GD}^{KD}}$ compared to the previous two pruning methods we proposed, $\mathrm{IMP_G}$ and $\mathrm{IMP_{GD}}$. $\mathrm{IMP_{GD}^{KD}}$ is capable of finding winning tickets at sparsity around 70%, outperforming $\mathrm{IMP_G}$, and showing comparable results with $\mathrm{IMP_{GD}}$ regarding the extreme sparsity. The FID curve of setting $\mathrm{IMP_{GD}^{KD}}$ is further mostly located below the curve of $\mathrm{IMP_{GD}}$, demonstrating a stronger generative ability than $\mathrm{IMP_{GD}}$. It suggests transferring knowledge from the full discriminator benefits to find the winning tickets.

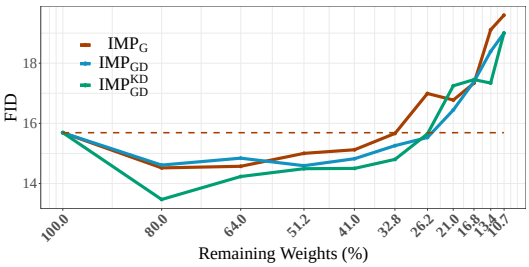

Figure 5: The FID curve of best subnetworks generated by three different pruning methods: $\mathrm{IMP_G}$, $\mathrm{IMP_{GD}}$ and $\mathrm{IMP_{GD}^{KD}}$.

Claim 3: Can IMP find matching subnetworks sparser than other pruning methods? Previous works claim that both a specific sparsity mask and a specific initialization are necessary for finding winning tickets (Frankle & Carbin, 2019a), and iterative magnitude pruning is better than one-shot pruning. To extend such a statement in the context of GANs, we compare IMP with several other benchmarks, randomly pruning (RP), one-shot magnitude pruning (OMP), and random tickets (RT), to see if IMP can find matching networks at higher sparsity.

Figure 6 and Table 1 show that iterative magnitude pruning outperforms one-shot magnitude pruning no matter pruning the discriminator or not. IMP finds winning tickets at higher sparsity (67.23% and 73.79%, respectively) than one-shot pruning (59.00% and 48.80%, respectively). The minimal FID score of subnetworks found by IMP is smaller than that of subnetworks found by OMP as well. This observation defends the statement that pruning iteratively is superior compared to one-shot pruning.

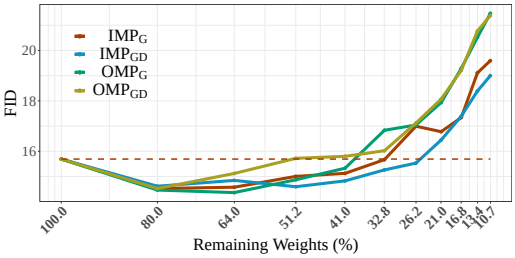

Figure 6: The FID curve of best subnetworks generated by three different pruning settings: $\mathrm{IMP_G}$, $\mathrm{IMP_{GD}}$ and $\mathrm{IMP_{GD}^{KD}}$.

Pruning Method	$\mathrm{S_{Extreme}}$	$\mathrm{FID_{Best}}$
No Pruning	-	15.69
$\mathrm{IMP_G}$	67.23%	14.20
$\mathrm{IMP_{GD}}$	73.79%	13.78
$\mathrm{OMP_G}$	59.04%	14.72
$\mathrm{OMP_{GD}}$	48.80%	15.69

Table 1: The extreme sparsity of the matching networks and the FID score of best subnetworks, found by iterative pruning and one-shot pruning.

We also list the minimal FID scores and extreme sparsity of matching networks for other pruning methods in Figure 7 and Table 2. It can be seen that IMP finds winning tickets at sparsity where some other pruning methods, randomly pruning and random initialization, cannot match. Since $\mathrm{IMP_G}$ shows the best overall result, we authenticate the previous statement that both the specific sparsity mask and the specific initialization are essential for finding winning tickets.

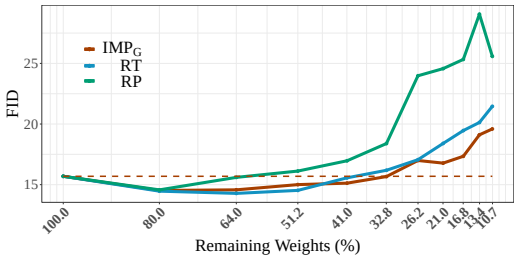

Figure 7: The FID curve of best subnetworks generated by three different pruning settings: IMP_G, IMP_GD and $\text{IMP}_\text{GD}^\text{KD}$.

Pruning Method	S_Extreme	FID_Best
No Pruning	-	15.69
IMP_G	67.23%	14.20
Random Pruning	20.00 %	14.57
Random Tickets	59.04%	14.28

Table 2: The FID score of best subnetworks and the extreme sparsity of matching networks found by Random Pruning, Random Rickets, and iterative magnitude pruning.

Claim 4: Does rewinding improve performance? In previous paragraphs, we show that we are able to find winning tickets in both SNGAN and CycleGAN. However, these subnetworks cannot match the performance of the original network at extremely high sparsity, while the subnetworks found by standard pruning can (Table 3). To find matching subnetworks at such high sparsity, we adopt the *rewinding* paradigm: after the masks are obtained, the weights of the model are rewound to $\boldsymbol{\theta_i}$, the weights after i steps of training, rather than reset to the pre-trained initialization $\boldsymbol{\theta_0}$. It was pointed out by Renda et al. (2020) that subnetworks found by IMP and rewound early in training can be trained to achieve the same accuracy at the same sparsity as subnetworks found by the standard pruning, providing a possibility that rewinding can also help GAN subnetworks.

We choose different rewinding settings: 5%, 10%, and 20% of the whole training epochs. The results are shown in Table 3. We observe that rewinding can significantly increase the extreme sparsity of matching networks. Rewinding to even only 5% of the training process can raise the extreme sparsity from 67.23% to 86.26%, and rewinding to 20% can match the performance of standard pruning. We also compare the FID score of subnetworks found at 89% sparsity. Rewind to 20% of the training process can match the performance of standard pruning at 89% sparsity, and other late rewinding settings can match the performance of the full model. This suggests that late rewinding techniques can greatly contribute to matching subnetworks with higher sparsity.

Table 3: Rewinding results. S_Extreme: Extreme sparsity where matching subnetworks exist. FID_Best: The minimal FID score of all subnetworks. $\text{FID}_{89\%}$: The FID score of subnetworks at 89% sparsity.

Setting of rewinding	S_Extreme	FID_Best	$\text{FID}_{89\%}$
Rewind 0%	67.23%	14.20	19.60
Rewind 5%	86.26%	**13.96**	15.82
Rewind 10%	86.26%	14.43	15.63
Rewind 20%	**89.26%**	14.82	15.29
Standard Pruning	**89.26%**	14.18	**15.22**

Summary Extensive experiments are conducted to examine the existence of matching subnetworks in generative adversarial models. We confirmed that there were matching subnetworks at high sparsities, and both the sparsity mask and the initialization matter for finding winning tickets. We also studied the effect of pruning the discriminator and demonstrate that pruning the discriminator can slightly boost the performance regarding the extreme sparsity and the minimal FID. We proposed a method to utilize the weights of the dense discriminator model to boost the performance further. We also compare IMP with different pruning methods, showing that IMP is superior to random tickets and random pruning. In addition, late rewinding can match the performance of standard pruning, which again shows consistency with previous works.

5 THE TRANSFER LEARNING OF GAN MATCHING NETWORKS

In the previous section, we confirm the presence of winning tickets in GANs. In this section, we will study the transferability of winning tickets. Existing works (Mehta, 2019) show that the matching networks in discriminative models can transfer across tasks. Here we evaluate this claim in GANs.

To investigate the transferability, we propose three transfer experiments from CIFAR-10 to STL-10 on SNGAN. We first identify matching subnetworks $g(\cdot, \boldsymbol{m_g} \odot \boldsymbol{\theta})$ and $d(\cdot, \boldsymbol{m_d} \odot \boldsymbol{\theta})$ on CIFAR-10, and then train and evaluate the subnetworks on STL-10. To assess whether the same pre-trained initialization is needed for transferring, we test three different weights loading method: 1) reset the weights to $\boldsymbol{\theta_0}$; 2) reset the weights to another pre-trained initialization $\boldsymbol{\theta'_0}$; 3) rewind the weights to $\boldsymbol{\theta_N}$. We train the network on STL-10 using the same hyper-parameters as on CIFAR-10. It is noteworthy that the hyper-parameters setting might not be optimal for the target task, yet it is fair to compare different transferring settings.

Table 4: Results of late rewinding experiments. θ_0: train the target model from the same initialization as the source model; θ_r: train from random initialization; θ_{Best}: train from the weights of trained source model. Baseline: full model trained on STL-10.

Model	Baseline	IMP_{G} (S = 67.23%)		IMP_{GD} (S = 73.79%)		$\text{IMP}_{\text{GD}}^{\text{KD}}$ (S = 73.79%)	
Metrics	FID_{Best}	FID_{Best}	Matching?	FID_{Best}	Matching?	FID_{Best}	Matching?
$\boldsymbol{\theta_0}$		**116.7**	✓	121.8	×	120.1	×
$\boldsymbol{\theta_r}$	115.3	119.2	×	**113.1**	✓	**115.5**	✓
$\boldsymbol{\theta_{\text{Best}}}$		204.7	×	163.0	×	179.1	×

The FID score of different settings is shown in Table 4. Subnetworks initialized by θ_0 and using masks generated by IMP_{G} can be trained to achieve comparable results with the baseline model. Surprisingly, random re-initialization $\boldsymbol{\theta_0'}$ shows better transferability than using the same initialization $\boldsymbol{\theta_0}$ in our transfer settings and outperforms the full model trained on STL-10, indicating that the combination of $\boldsymbol{\theta_0}$ and the mask generated by IMP_{GD} is more focused on the source dataset and consequently has lower transferability.

Summary In this section, we tested the transferability of IMP subnetworks. Transferring from $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_r}$ both produce matching results on the target dataset, STL-10. $\boldsymbol{\theta_0}$ works better with masks generated by IMP_{G} while the masks generated by IMP_{GD} prefer a different initialization $\boldsymbol{\theta_r}$. Given that IMP_{GD} performs better on CIFAR-10, it is reasonable that the same initialization $\boldsymbol{\theta_0}$ has lower transferability when using masks from IMP_{GD}.

6 Efficiency of GAN Winning Tickets

Unlike the unstructured magnitude pruning method, channel pruning can reduce the number of parameters in GANs. Therefore, winning tickets found by channel pruning are more efficient than the original model regarding computational cost. To fully exploit the advantage of subnetworks founded by structural pruning, we further compare our prune-and-train pipeline with a state-of-the-art GAN compression framework (Wang et al., 2020b). The pipeline is described as follows: after we extract the sparse structure generated by channel pruning from the dense model, we reset the model weights to the pre-trained initialization $\boldsymbol{\theta_0}$ and then train for the same number of epochs as the dense model used.

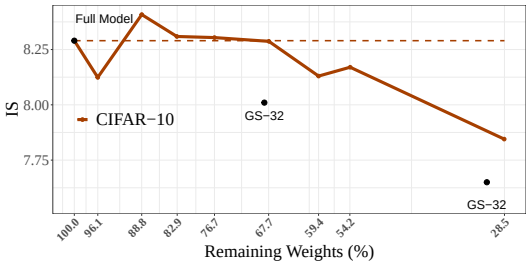

Figure 8: Relationship between the best IS score of SNGAN subnetworks generated by channel pruning and the percent of remaining weights. GS-32: GAN Slimming without quantization (Wang et al., 2020b). Full Model: Full model trained on CIFAR-10.

We can see from Figure 8 that subnetworks are founded by channel pruning method at about 67% sparsity, which provides a new path for winning tickets other than magnitude pruning. The matching networks at 67.7% outperform GS-32 regarding Inception Score by 0.25; the subnetworks at about 29% sparsity can outperform GS-32 by 0.20, setting up a new benchmark for GAN compressions.

7 Conclusion

In this paper, the lottery ticket hypothesis has been extended to GANs. We successfully identify winning tickets in GANs, which are separately trainable to match the full dense GAN performance. Pruning the discriminator, which is rarely studied before, had only slight effects on the ticket finding process, while the initialization used in the discriminator is essential. We also demonstrate that the winning tickets found can transfer across diverse tasks. Moreover, we provide a new way of finding winning tickets that alter the structure of models. Channel pruning is able to extract matching subnetworks from a dense model that can outperform the current state-of-the-art GAN compression after resetting the weights and re-training.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan, 2017.

Tianlong Chen, Jonathan Frankle, Shiyu Chang, Sijia Liu, Yang Zhang, Zhangyang Wang, and Michael Carbin. The lottery ticket hypothesis for pre-trained bert networks. *arXiv preprint arXiv:2007.12223*, 2020.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223, 2011.

Justin Cosentino, Federico Zaiter, Dan Pei, and Jun Zhu. The search for sparse, robust neural networks, 2019.

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

Bryn Elesedy, Varun Kanade, and Y. Teh. Lottery tickets in linear models: An analysis of iterative magnitude pruning. *ArXiv*, abs/2007.08243, 2020.

Utku Evci, Fabian Pedregosa, Aidan Gomez, and Erich Elsen. The difficulty of training sparse neural networks. *arXiv preprint arXiv:1906.10732*, 2019.

Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis - Finding Sparse, Trainable Neural Networks. *ICLR*, 2019a.

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019b.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M. Roy, and Michael Carbin. Linear mode connectivity and the lottery ticket hypothesis, 2019a.

Jonathan Frankle, Gintare Karolina Dziugaite, Daniel M Roy, and Michael Carbin. The lottery ticket hypothesis at scale. *arXiv preprint arXiv:1903.01611*, 2019b.

Jonathan Frankle, David J. Schwab, and Ari S. Morcos. The early phase of neural network training. In *International Conference on Learning Representations*, 2020.

Trevor Gale, Erich Elsen, and Sara Hooker. The state of sparsity in deep neural networks. *arXiv preprint arXiv:1902.09574*, 2019.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans, 2017.

Song Han, Huizi Mao, and William J. Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

Yihui He, Xiangyu Zhang, and Jian Sun. Channel Pruning for Accelerating Very Deep Neural Networks. *arXiv.org*, July 2017a.

Yihui He, Xiangyu Zhang, and Jian Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1389–1397, 2017b.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2017.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.

Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. Neural style transfer: A review. *IEEE Transactions on Visualization and Computer Graphics*, 2019.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Y Lecun. Optimal brain damage. *Neural Information Proceeding Systems*, 2(279):598–605, 1990.

Muyang Li, Ji Lin, Yaoyao Ding, Zhijian Liu, Jun-Yan Zhu, and Song Han. Gan compression: Efficient architectures for interactive conditional gans, 2020.

Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks, 2016.

Zhuang Liu, Jianguo Li, Zhiqiang Shen, Gao Huang, Shoumeng Yan, and Changshui Zhang. Learning Efficient Convolutional Networks through Network Slimming. *arXiv.org*, August 2017.

Zhuang Liu, Mingjie Sun, Tinghui Zhou, Gao Huang, and Trevor Darrell. Rethinking the value of network pruning. In *International Conference on Learning Representations*, 2019.

Eran Malach, Gilad Yehudai, Shai Shalev-Shwartz, and Ohad Shamir. Proving the lottery ticket hypothesis: Pruning is all you need, 2020.

Rahul Mehta. Sparse transfer learning via winning lottery tickets. *arXiv preprint arXiv:1905.07785*, 2019.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. *arXiv.org*, February 2018.

Ari 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*, pp. 4933–4943, 2019.

Ankit Pensia, Shashank Rajput, Alliot Nagle, Harit Vishwakarma, and Dimitris Papailiopoulos. Optimal lottery tickets via subsetsum: Logarithmic over-parameterization is sufficient. *ArXiv*, abs/2006.07990, 2020.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2016.

Alex Renda, Jonathan Frankle, and Michael Carbin. Comparing rewinding and fine-tuning in neural network pruning. *arXiv preprint arXiv:2003.02389*, 2020.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.

Pedro Savarese, Hugo Silva, and Michael Maire. Winning the lottery with continuous sparsification, 2020.

Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Josh Susskind, Wenda Wang, and Russ Webb. Learning from simulated and unsupervised images through adversarial training, 2017.

Han Shu, Yunhe Wang, Xu Jia, Kai Han, Hanting Chen, Chunjing Xu, Qi Tian, and Chang Xu. Co-Evolutionary Compression for Unpaired Image Translation. *arXiv.org*, July 2019.

Chaoqi Wang, Guodong Zhang, and Roger Grosse. Picking winning tickets before training by preserving gradient flow. *arXiv preprint arXiv:2002.07376*, 2020a.

Haotao Wang, Shupeng Gui, Haichuan Yang, Ji Liu, and Zhangyang Wang. Gan slimming: All-in-one gan compression by a unified optimization framework, 2020b.

Peiqi Wang, Dongsheng Wang, Yu Ji, Xinfeng Xie, Haoxuan Song, XuXin Liu, Yongqiang Lyu, and Yuan Xie. Qgan: Quantized generative adversarial networks, 2019.

Shihui Yin, Kyu-Hyoun Kim, Jinwook Oh, Naigang Wang, Mauricio Serrano, Jae-Sun Seo, and Jungwook Choi. The sooner the better: Investigating structure of early winning lottery tickets, 2020.

Haoran You, Chaojian Li, Pengfei Xu, Yonggan Fu, Yue Wang, Xiaohan Chen, Richard G Baraniuk, Zhangyang Wang, and Yingyan Lin. Drawing early-bird tickets: Towards more efficient training of deep networks. *arXiv.org*, September 2019.

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*, 2020.

Haonan Yu, Sergey Edunov, Yuandong Tian, and Ari S. Morcos. Playing the lottery with rewards and multiple languages: lottery tickets in rl and nlp, 2019.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv.org*, March 2017.

A MORE TECHNICAL DETAILS

A.1 ALGORITHMS

In this section, we describe the details of algorithm we used in finding lottery tickets. Two distinct pruning methods are used in Algorithm 1 and Algorithm 2.

B MORE EXPERIMENTS RESULTS AND ANALYSIS

We will provide extra experiments results and analysis in this section.

B.1 VISUALIZATION OF IMP WINNING TICKETS

We also conducted experiments to find winning tickets in CycleGAN on dataset winter2summer (Zhu et al., 2017). We observed similar patterns and found matching networks at 79.02% sparsity. We randomly sample four images from the dataset and show the translated images in Figure 9.

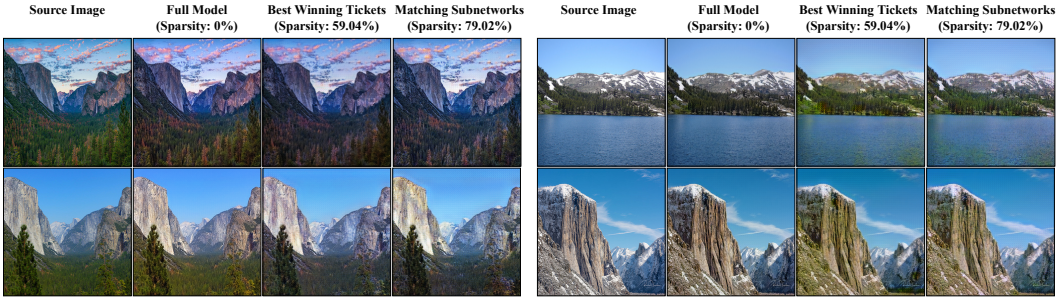

Figure 9: Visualization of CycleGAN Winning Tickets found by IMP on summer2winter. Sparsity of best winning tickets: 59.04%. Extreme sparsity of matching subnetworks: 79.02%. Left: visualization results of task summer2winter. Right: visualization results of task winter2summer.

B.2 CHANNEL PRUNING FOR SNGAN

We study the relationship between the Inception Score and the *remaining model size*, *i.e.* the ratio between the size of a channel-pruned model and its original model. The results are plotted in Figure 10. A similar conclusion can be drawn from the graph that matching networks exist, and at the same sparsity, the matching networks can be trained to outperform the current state-of-the-art GAN compression framework.

B.3 CHANNEL PRUNING FOR CYCLEGAN

We also conducted experiments on CycleGAN using channel pruning. The task we choose is horse-to-zebra, *i.e.*, we prune each of the two generators separately, which is aligned with SNGAN which

has only one generator. We prove that channel pruning is also capable of finding winning tickets in CycleGAN. Moreover, at extreme sparsity the sparse subnetwork that we obtain can be trained to reach slightly better results than the current state-of-the-art GAN compression framework without quantization.

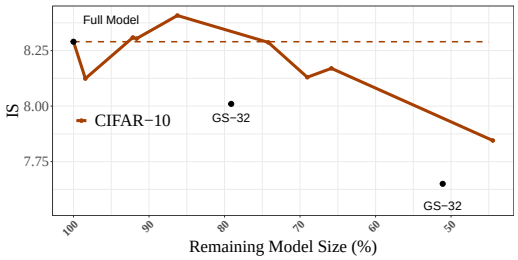

Figure 10: Relationship between the best IS score of SNGAN subnetworks generated by channel pruning and the percent of remaining model size. GS-32: GAN Slimming without quantization (Wang et al., 2020b). Full Model: Full model trained on CIFAR-10.

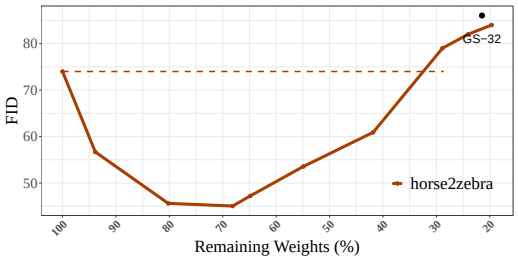

Figure 11: Relationship between the best FID score of CycleGAN subnetworks generated by channel pruning and the percent of remaining weights. GS-32: GAN Slimming without quantization (Wang et al., 2020b). Full Model: Full model trained on CIFAR-10.

Algorithm 1: Finding winning tickets by Iterative Magnitude Pruning

Input: The desired sparsity s
Output: An sparse GAN $g(\cdot; \boldsymbol{m_g} \odot \boldsymbol{\theta_g})$ and $d(\cdot; \boldsymbol{m_d} \odot \boldsymbol{\theta_d})$

Set $\boldsymbol{m_g} = \{1\}^{|\boldsymbol{\theta_{g_0}}|}$ and $\boldsymbol{m_d} = \{1\}^{|\boldsymbol{\theta_{d_0}}|}$.
Set $\boldsymbol{\theta_{g_0}} :=$ initial weights of the generator model, $\boldsymbol{\theta_{d_0}} :=$ initial weights of the discriminator model.
Iteration $i = 0$
while *the sparsity of* $\boldsymbol{m_g} < s$ **do**
  Train the generator $g(\cdot; \boldsymbol{m_g} \odot \boldsymbol{\theta_{g_0}})$ and the discriminator $d(\cdot; \boldsymbol{m_d} \odot \boldsymbol{\theta_{d_0}})$ for N epochs to get parameters $\boldsymbol{\theta^i_{g_N}}$ and $\boldsymbol{\theta^i_{d_N}}$.
  if *pruning the discriminator* **then**
    Prune 20% of the parameters in $\boldsymbol{\theta^i_{g_N}}$ and $\boldsymbol{\theta^i_{d_N}}$, creating two mask $\boldsymbol{m'_g}$ and $\boldsymbol{m'_d}$.
  else
    Prune 20% of the parameters in $\boldsymbol{\theta^i_{g_N}}$, creating a mask $\boldsymbol{m'_g}$. $\boldsymbol{m'_d}$ remains $\{1\}^{|\boldsymbol{\theta_{d_0}}|}$.
  end
  Update $\boldsymbol{m_g} = \boldsymbol{m'_g}$ and $\boldsymbol{m_d} = \boldsymbol{m'_d}$
  $i = i + 1$
end

Algorithm 2: Finding winning tickets by Channel Pruning

Input: A threshold ρ of importance score, number of steps for training N
Output: An sparse GAN $g(\cdot; \boldsymbol{m_g} \odot \boldsymbol{\theta_{g_0}})$ and $d(\cdot; \boldsymbol{m_d} \odot \boldsymbol{\theta_{d_0}})$

Randomly initialize $\boldsymbol{\gamma_g}$ for every normalization layers in generator $\mathcal{G}$ and $\boldsymbol{\gamma_d}$ for discriminator $\mathcal{D}$. $\boldsymbol{\gamma} = (\boldsymbol{\gamma_g}, \boldsymbol{\gamma_d})$
Set $\boldsymbol{\theta_{g_0}} :=$ initial weights of the generator model, $\boldsymbol{\theta_{d_0}} :=$ initial weights of the discriminator model.
$i = 0$
while $i < N$ **do**
  Compute masks $\boldsymbol{m_d}$ from $\boldsymbol{\gamma_d}$ and $\boldsymbol{m_g}$ from $\boldsymbol{\gamma_g}$.
  Compute L_{cp}, L_{GAN} and L_{dist}.
  Update $\boldsymbol{\theta_g}$ and $\boldsymbol{\theta_d}$ by training the generator $g(\cdot; \boldsymbol{m_g} \odot \boldsymbol{\theta_g})$ and the discriminator $d(\cdot; \boldsymbol{m_d} \odot \boldsymbol{\theta_d})$ for one step.
  Update $\boldsymbol{\gamma}$: $\boldsymbol{\gamma} \leftarrow \text{prox}_{\rho\eta}(\boldsymbol{\gamma} - \eta \nabla_{\boldsymbol{\gamma}} L_{\text{cp}})$, where $\text{prox}_{\lambda}(x) = \text{sgn}(x) \odot \max(|x| - \lambda \cdot \mathbf{1}, 0)$
  $i \leftarrow i + 1$
end
Calculate the final masks $\boldsymbol{m_g}$ from $\boldsymbol{\gamma_g}$ and $\boldsymbol{m_d}$ from $\boldsymbol{\gamma_d}$