

Self-Supervised GAN Compression

Chong Yu and Jeff Pool

NVIDIA

{chongy,jpool}@nvidia.com

Abstract

Deep learning’s success has led to larger and larger models to handle more and more complex tasks; trained models can contain millions of parameters. These large models are compute- and memory-intensive, which makes it a challenge to deploy them with minimized latency, throughput, and storage requirements. Some model compression methods have been successfully applied to image classification and detection or language models, but there has been very little work compressing generative adversarial networks (GANs) performing complex tasks. In this paper, we show that a standard model compression technique, weight pruning, cannot be applied to GANs using existing methods. We then develop a self-supervised compression technique which uses the trained discriminator to supervise the training of a compressed generator. We show that this framework has a compelling performance to high degrees of sparsity, can be easily applied to new tasks and models, and enables meaningful comparisons between different pruning granularities.

1. Introduction

Deep Neural Networks (DNNs) have proved successful in various tasks like computer vision, natural language processing, recommendation systems, and autonomous driving. Modern networks are comprised of millions of parameters, requiring significant storage and computational effort. Though accelerators such as GPUs make realtime performance more accessible, compressing networks for faster inference and simpler deployment is an active area of research. Compression techniques have been applied to many networks to reduce memory requirements and improve performance. Though these approaches do not always harm accuracy, aggressive compression can adversely affect the behavior of the network. Distillation (Schmidhuber, 1991; Hinton et al., 2015) can improve the accuracy of a compressed network by using information from the original, uncompressed network.

Generative Adversarial Networks (GANs) (Schmidhuber, 1990; Goodfellow et al., 2014) are a class of DNN that consist of two sub-networks: a generative model and a discriminative model. Their training process aims to achieve a Nash Equilibrium between these two sub-models. GANs have been used in semi-supervised and unsupervised learning areas, such as fake dataset synthesis (Radford et al., 2016; Brock et al., 2019), style transfer (Zhu et al., 2017b; Azadi et al., 2018), and image-to-image translation (Zhu et al., 2017a; Choi et al., 2018). As with networks used in other tasks, GANs have millions of parameters and nontrivial computational requirements.

In this work, we explore compressing the generative model of GANs for efficient deployment. We show that applying standard pruning techniques causes the generator’s behavior to no longer achieve the network’s goal and that past work targeted at compressing GANs for simple image synthesis fall short when they are applied to pruning large tasks. In some cases, this result is masked by loss curves that look identical to the original training. By modifying the loss function with a novel combination of the pre-trained discriminator and the original and compressed generators, we overcome this behavioral degradation and achieve compelling compression rates with little change in the quality of the compressed generator’s output. We apply our technique to several networks and tasks to show generality. Finally, we study the behavior of compressed generators when pruned with different amounts and types of sparsity, finding that a technique commonly used for accelerating image classification networks is not trivially applicable to GANs.

Our main contributions are:

- We illustrate that and explain why pruning the generator of a GAN with existing methods is unsatisfactory for complex tasks. (Section 3)
- We propose self-supervised compression for the generator in a GAN. (Section 4)
- We show that our technique can apply to several networks and tasks. (Section 5)
- We show and analyze qualitative differences in pruning ratio and granularities. (Section 6)

We leave performance gained by pruning GANs for future work, as it is dependent on the target hardware: there are various way to exploit fine-grained sparsity on CPUs (Elsen et al., 2019), GPUs (Chen, 2018; Zhu et al., 2018), and custom accelerators (Parashar et al., 2017; Judd et al., 2017).

2. Related Research

A common method of DNN compression is network pruning (Han et al., 2015): setting the small weights of a trained network to zero and fine-tuning the remaining weights to recover accuracy. Zhu & Gupta (2018) proposed a gradual pruning technique (AGP) to compress the model during the initial training process. Wen et al. (2016) proposed a structured sparsity learning method that uses group regularization to force weights towards zero, leading to pruning groups of weights together. Li et al. (2017) pruned entire filters and their connecting feature maps from models, allowing the smaller network to be accelerated with standard dense software libraries. Though it was initially applied to image classification networks, network pruning has been extended to natural language processing tasks (See et al., 2016; Narang et al., 2017) and to recurrent neural networks (RNNs) of all types - vanilla RNNs, GRUs (Cho et al., 2014), and LSTMs (Hochreiter & Schmidhuber, 1997). As with classification networks, structured sparsity within recurrent units has been exploited (Wen et al., 2018).

A complementary method of network compression is quantization. Sharing weight values among a collection of similar weights by hashing (Chen et al., 2015) or clustering (Han et al., 2016) can save storage and bandwidth at runtime. Changing fundamental data types affords hardware the ability to accelerate the arithmetic operations, both in training (Micikevicius et al., 2018) and inference regimes (Jain et al., 2019).

Several techniques have been devised to combat lost accuracy due to compression, since there is always the chance that the behavior of the network may change in undesirable ways when the network is compressed. Using GANs to generate unique training data (Liu et al., 2018b) and extracting knowledge from an uncompressed network, known as distillation (Hinton et al., 2015), can help keep accuracy high. Since the pruning process involves many hyperparameters, Lin et al. (2019) use a GAN to guide pruning, and Wang et al. (2019a) structure compression as a reinforcement learning problem; both remove some of the burden from the user.

3. Existing Techniques Fail for Complex Task

Though there are two networks in a single GAN, the main workload at deployment is usually from the generative model, or generator. For example, in image synthesis and style transfer tasks, the final output images are created solely

by the generator. The discriminative model (discriminator) is vital in training, but it is abandoned afterward for many tasks. So, when we try to apply state-of-the-art compression methods to GANs, we focus on the generator for efficient deployment. As we will see, the generative performance of the compressed generators is quite poor for the selected image-to-image translation task. We look at two broad categories of baseline approaches: standard pruning techniques that have been applied to other network architectures, and techniques that were devised to compress the generator of a GAN performing image synthesis. We compare the dense baseline [a] to our technique [b], as well as a small, dense network with the same number of parameters [c]. (Labels correspond to entries in Table 1, the overview of all techniques, and Figure 1, results of each technique).

Standard Pruning Techniques. To motivate GAN-specific compression methods, we try variations of two state-of-the-art pruning methods: manually pruning and fine tuning (Han et al., 2015) a trained dense model [d], and AGP (Zhu & Gupta, 2018) from scratch [e] and during fine-tuning [f]. We also include distillation (Hinton et al., 2015) to improve the performance of the pruned network with manual pruning [g] and AGP fine-tuning [h]. Distillation is typically optional for other network types, since it is possible to get decent accuracy with moderate pruning in isolation. For very aggressive compression or challenging tasks, distillation aims to extract knowledge for the compressed (student) network from original (teacher) network’s behavior. We also fix the discriminator of [g] to see if the discriminator was being weakened by the compressed generator [i].

Targeted GAN Compression. There has been some work in compressing GANs with methods other than pruning, and only one technique applied to an image-to-image translation task. For this category, we decompose each instance of prior work into two areas: the method of compression (e.g. quantization, layer removal, etc.) and the modifications required to make the compression succeed (e.g. distillation, novel training schemes, etc.). For comparisons to these techniques, we apply the modifications presented in prior research to the particular method of compression on which we focus, network pruning. We first examine two approaches similar to ours. Adversarial training (Wang et al., 2018) [j] posits that during distillation of a classification network, the student network can be thought of as a generative model attempting to produce features similar to that of the teacher model. So, a discriminator was trained alongside the student network, trying to distinguish between the student and the teacher. One could apply this technique to compress the generator of a GAN, but we find that its key shortcoming is that it trains a discriminator from scratch. Similarly, distillation has been used to compress GANs in Aguinaldo et al. (2019) [k], but again, the “teacher” discriminator was not used when teaching the “student” generator.

Table 1. GAN compression comparison (network pruning)

Technique	Generator(s)		Init Scheme	Discriminator		Loss Terms			Results	
	Compressed	Init Scheme		Fixed	L-Gc	L-Dc	L-Go	L-Do	Qualitative	FID Score
(a) No Compression	Dense	Random	Dense,Random	No	-	-	Yes	Yes	Good	6.113
(b) Self-Supervised (ours)	Dense,Sparse	From Dense	Dense,Pretrained	No	Yes	Yes	Yes	Yes	Good	6.929
(c) Small & Dense Network	Dense	Random	Dense,Random	No	-	-	Yes	Yes	Mode collapse	72.821
(d) One-shot Pruning & Fine-Tuning	Sparse	From Dense	Dense,Pretrained	No	Yes	Yes	-	-	Facial artifacts	24.404
(e) Gradual Pruning & Fine-Tuning	Sparse	From Dense	Dense,Random	No	Yes	Yes	-	-	Facial artifacts	35.677
(f) Gradual Pruning during Training	Sparse	Random	Dense,Random	No	Yes	Yes	-	-	No faces	84.941
(g) One-shot Pruning & Distillation	Dense,Sparse	From Dense	-	-	Yes	-	Yes	-	Mode collapse	45.461
(h) (d) & Distillation	Dense,Sparse	From Dense	Dense,Pretrained	No	Yes	Yes	Yes	-	Color artifacts	38.985
(i) (g) & Fix Original Loss	Dense,Sparse	From Dense	Dense,Pretrained	Yes	Yes	Yes	-	-	Facial artifacts	15.182
(j) Adversarial Learning	Dense,Sparse	Random	Dense,Random	No	Yes	Yes	Yes	Yes	Mode collapse	92.721
(k) Knowledge Distillation	Dense,Sparse	From Dense	Dense,Random	No	Yes	-	Yes	Yes	Mode collapse	103.094
(l) Distill Intermediate (LIT)	Dense,Sparse	From Dense	Dense,Pretrained	Yes	-	-	-	-	Mode collapse	61.150
(m) E-M Pruning	Dense,Sparse	From Dense	Sparse,Pretrained	No	Yes	Yes	Yes	-	Color artifacts	159.767
(n) G & D Both Pruning	Dense,Sparse	From Dense	Sparse,Pretrained	No	Yes	Yes	Yes	-	Mode collapse	46.453

Learned Intermediate Representation Training (LIT) (Koratana et al., 2019) [l] compresses StarGAN by a factor of $1.8\times$ by training a shallower network. Crucially, LIT does not use the pre-trained discriminator in any loss function. Quantized GANs (QGAN) (Wang et al., 2019b) [m] use a training process based on Expectation-Maximization to achieve impressive compression results on small generative tasks with output images of 32x32 or 64x64 pixels. Liu et al. (2018a) find that maintaining a balance between discriminator and generator is key: their approach is to selectively binarize parts of both networks in the training process on the Celeb-A generative task, up to 64x64 pixels. So, we try pruning both networks during the training process [n].

Experiments. For these experiments, we use StarGAN¹ (Choi et al., 2018) trained with the Distiller (Zmora et al., 2019) library for the pruning. StarGAN extends the image-to-image translation capability from two domains to multiple domains within a single unified model. It uses the CelebFaces Attributes (CelebA) (Liu et al., 2015) as the dataset. CelebA contains 202,599 images of celebrities’ faces, each annotated with 40 binary attributes. As in the original work, we crop the initial images from size 178×218 to 178×178 , then resize them to 128×128 and randomly select 2,000 images as the test dataset and use remaining images for training. The aim of StarGAN is facial attribute translation: given some image of a face, it generates new images with five domain attributes changed: 3 different hair colors (black, blond, brown), different gender (male/female), and different age (young/old). Our target sparsity is 50% for each approach.

We stress that we attempted to find good hyperparameters when using the existing techniques, but standard approaches like reducing the learning rate for fine-tuning (Han et al., 2015), etc., were not helpful. Further, the target sparsity, 50%, is not overly aggressive, and we do not impose any structure; other tasks readily achieve 80%-90% fine-grained

sparsity with minimal accuracy impact.

The results of these trials are shown in Figure 1. Subjectively, it is easy to see that the existing approaches (1(c) through 1(n)) produce inferior results to the original, dense generator. Translated facial images from pruning & naïve fine-tuning (1(d) and 1(e)) do give unique results for each latent variable, but the images are hardly recognizable as faces. These fine-tuning procedures, along with AGP from scratch (1(f)) and distillation from intermediate representations (1(l)), simply did not converge. One-shot pruning and traditional distillation (1(g)), adversarial learning (1(j)), knowledge distillation (1(k)), training a “smaller, dense” half-sized network from scratch (1(c)) and pruning both generator and discriminator (1(n)) keep facial features intact, but the image-to-image translation effects are lost to mode collapse (see below). There are obvious mosaic textures and color distortion on the translated images from fine-tuning & distillation (1(h)), without fine-tuning the original loss (1(i)), and from the pruned model based on the Expectation-Maximization (E-M) algorithm (1(m)). However, the translated facial images from a generator compressed with our proposed self-supervised GAN compression method (1(b)) are more natural, nearly indistinguishable from the dense baseline (1(a)), matching the quantitative Frechet Inception Distance (FID) scores (Heusel et al., 2017) in Table 1. While past approaches have worked to prune some networks on other tasks (DCGAN generating MNIST digits, see the supplementary material), we show that they do not succeed on larger image-to-image translation tasks, while our approach works on both. Similarly, though LIT (Koratana et al., 2019) [l] was able to achieve a compression rate of $1.8\times$ on this task by training a shallower network, it does not see the same success at network pruning with a higher rate, as modest as 50% sparsity is.

Discussion. It is tempting to think that the loss curves of the experiment for each technique can tell us if the result is good or not. We found that for many of these experiments, the

¹StarGAN: <https://github.com/yunjey/StarGAN>.

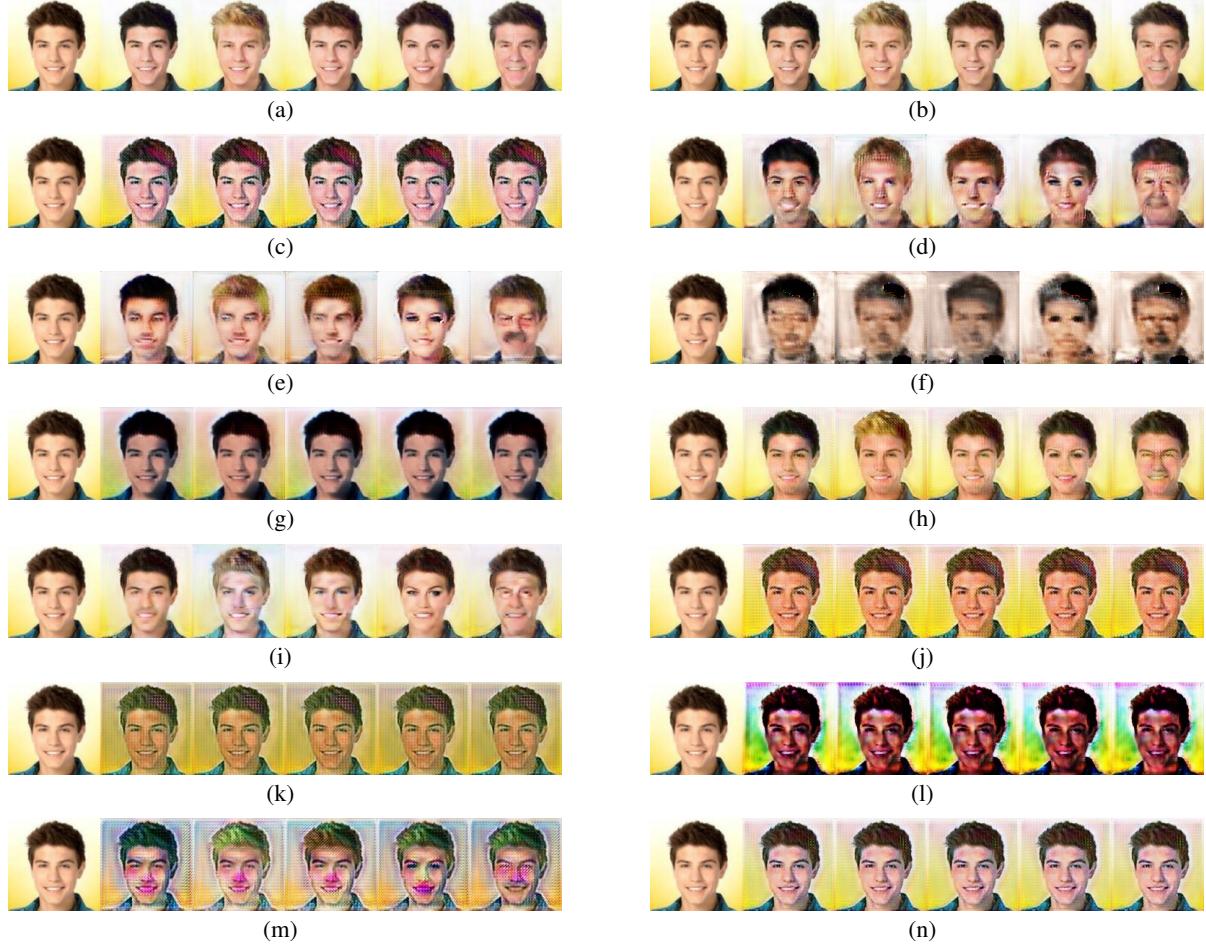


Figure 1. Various approaches to compress StarGAN with network pruning. Each group shows one input face translated with different methods of compressing the network: **a**. Uncompressed, **b**. Self-Supervised (**ours**), **c**. Small and dense, **d**. One-shot pruning and fine-tuning, **e**. AGP as fine-tuning, **f**. AGP from scratch, **g**. One-shot pruning and distilling, **h**. AGP during distillation, **i**. AGP during distillation with fixed discriminator, **j**. Adversarial learning, **k**. Knowledge distillation, **l**. Distillation on output of intermediate layers, **m**. E-M pruning, and **n**. Prune both G and D models.

loss curves correctly predicted that the final result would be poor. However, the curves for [h] and [m] look very good - the compressed generator and discriminator losses converge at 0, just as they did for baseline training. It is clear from the results of querying the generative models (Figures 1(h) and 1(m)), though, that this promising convergence is a false positive. In contrast, the curves for our technique predict good performance, and, as we prune more aggressively in Section 6, higher loss values correlate well with worsening FID scores. (Loss curves are provided in the *Appendix*.)

As pruning and distillation are very effective when compressing models for image classification tasks, why do they fail to compress this generative model? We share three potential reasons:

1. Standard pruning techniques need explicit evaluation metrics; softmax easily reflects the probability distri-

bution and classification accuracy. GANs are typically evaluated subjectively, though some imperfect quantitative metrics have been devised.

2. GAN training is relatively unstable (Arjovsky et al., 2017; Liu et al., 2018a) and sensitive to hyperparameters. The generator and discriminator must be well-matched, and pruning can disrupt this fine balance.
3. The energy of the input and output of a GAN is roughly constant, but other tasks, such as classification, produce an output (1-hot label vector) with much less entropy than the input (three-channel color image of thousands of pixels).

Elaborating on this last point, there is more tolerance in the reduced-information space for the compressed classification model to give the proper output. That is, even if the probability distribution inferred by the original and compressed classification models are not exactly the same, the classi-

fied labels *can* be the same. On the other hand, tasks like style-transfer and dataset synthesis have no obvious energy reduction. We need to keep entropy as high as possible (Kumar et al., 2019) during the compression process to avoid mode collapse – generating the same output for different inputs or tasks. Attempting to train a new discriminator to make the compressed generator behave more like the original generator (Wang et al., 2018) suffers from this issue – the new discriminator quickly falls into a low-entropy solution and cannot escape. Not only does this preclude its use on generative tasks, but it means that the compressed network for any task must also be trained from scratch during the distillation process, or the discriminator will never be able to learn.

4. Self-Supervised Generator Compression

We seek to solve each of the problems highlighted above. Let us restate the general formulation of GAN training: the purpose of the generative model is to generate new samples which are very similar to the real samples, but the purpose of the *discriminative* model is to distinguish between real samples and those synthesized by the generator. A fully-trained discriminator is good at spotting differences, but a well-trained generator will cause it to believe that the a generated sample is both real and generated with a probability of 0.5. Our main insight follows:

By using this powerful discriminator that is already well-trained on the target data set, we can allow it to stand in as a quantitative subjective judge (point 1, above) – if the discriminator can't tell the difference between real data samples and those produced by the compressed generator, then the compressed generator is of the same quality as the uncompressed generator. A human no longer needs to inspect the results to judge the quality of the compressed generator. This also addresses our second point: by starting with a trained discriminator, we know it is well-matched to the generator and will not be overpowered. Since it is so capable (there is no need to prune it to), it also helps to avoid mode collapse. As distillation progresses, it can adapt to and induce fine changes in the compressed generator, which is initialized from the uncompressed generator. Since the original discriminator is used as a proxy for a human's subjective evaluation, we refer to this as “self-supervised” compression. We illustrate the workflow in Figure 2, using a GAN charged with generating a map image from a satellite image in a domain translation task.

In the right part of Figure 2, the real satellite image (x) goes through the original generative model (\mathbf{G}_O) to produce a fake map image (\hat{y}_o). The corresponding generative loss value is $l\text{-}\mathbf{G}_O$. Accordingly, in the left part of Figure 2, the real satellite image (x) goes through the compressed generative model (\mathbf{G}_C) to produce a fake map image (\hat{y}_c).

The corresponding generative loss value is $l\text{-}\mathbf{G}_C$. This is the inference process of the original and compressed generators, expressed as follows:

$$\hat{y}_o = \mathbf{G}_O(x), \quad \hat{y}_c = \mathbf{G}_C(x) \quad (1)$$

The overall generative difference is measured between the two corresponding generative losses². We use a generative consistent loss function (L_{GC}) in the bottom of Figure 2 to represent this process.

$$L_{GC}(l\text{-}\mathbf{G}_O, l\text{-}\mathbf{G}_C) \rightarrow 0 \quad (2)$$

Since the GAN training process aims to reduce the differences between real and generated samples, we stick to this principle in the compression process. In the upper right of Figure 2, real map image (y) and fake map image (\hat{y}_o) go through the original discriminative model \mathbf{D}_O . \mathbf{D}_O tries to ensure that the distribution of \hat{y}_o is indistinguishable from y using an adversarial loss. The corresponding discriminative loss value is $l\text{-}\mathbf{D}_O$. In the upper left of Figure 2, real map image (y) and fake map image (\hat{y}_c) also go through the original discriminative model \mathbf{D}_O . In this way, we use the original discriminative model as a “self-supervisor”. The corresponding discriminative loss value is $l\text{-}\mathbf{D}_C$.

$$l\text{-}\mathbf{D}_O = \mathbf{D}_O(y, \hat{y}_o), \quad l\text{-}\mathbf{D}_C = \mathbf{D}_O(y, \hat{y}_c) \quad (3)$$

So the discriminative difference is measured between two corresponding discriminative losses. We use the discriminative consistent loss function L_{DC} in the top of Figure 2 to represent this process.

$$L_{DC}(l\text{-}\mathbf{D}_O, l\text{-}\mathbf{D}_C) \rightarrow 0 \quad (4)$$

The generative and discriminative consistent loss functions (L_{GC} and L_{DC}) use the weighted normalized Euclidean distance. Taking the StarGAN task as the example (other tasks may use different losses):

$$\begin{aligned} L_{GC}(l\text{-}\mathbf{G}_O, l\text{-}\mathbf{G}_C) &= |l\text{-}\mathbf{Gen}_O - l\text{-}\mathbf{Gen}_C| / |l\text{-}\mathbf{Gen}_O| + \\ &\quad \alpha |l\text{-}\mathbf{Cla}_O - l\text{-}\mathbf{Cla}_C| / |l\text{-}\mathbf{Cla}_O| + \\ &\quad \beta |l\text{-}\mathbf{Rec}_O - l\text{-}\mathbf{Rec}_C| / |l\text{-}\mathbf{Rec}_O| \end{aligned} \quad (5)$$

where $l\text{-}\mathbf{Gen}$ is the generation loss term, $l\text{-}\mathbf{Cla}$ is the classification loss term, and $l\text{-}\mathbf{Rec}$ is the reconstruction loss term. α and β are the weight ratios among three loss types. (We use the same values of α and β used in the original StarGAN baseline.)

$$\begin{aligned} L_{DC}(l\text{-}\mathbf{D}_O, l\text{-}\mathbf{D}_C) &= |l\text{-}\mathbf{Dis}_O - l\text{-}\mathbf{Dis}_C| / |l\text{-}\mathbf{Dis}_O| + \\ &\quad \delta |l\text{-}\mathbf{GP}_O - l\text{-}\mathbf{GP}_C| / |l\text{-}\mathbf{GP}_O| \end{aligned} \quad (6)$$

²In different GANs, the generative loss may consist of several sub-items. For example, StarGAN combines adversarial loss, domain classification loss and reconstruction loss into overall generative loss.

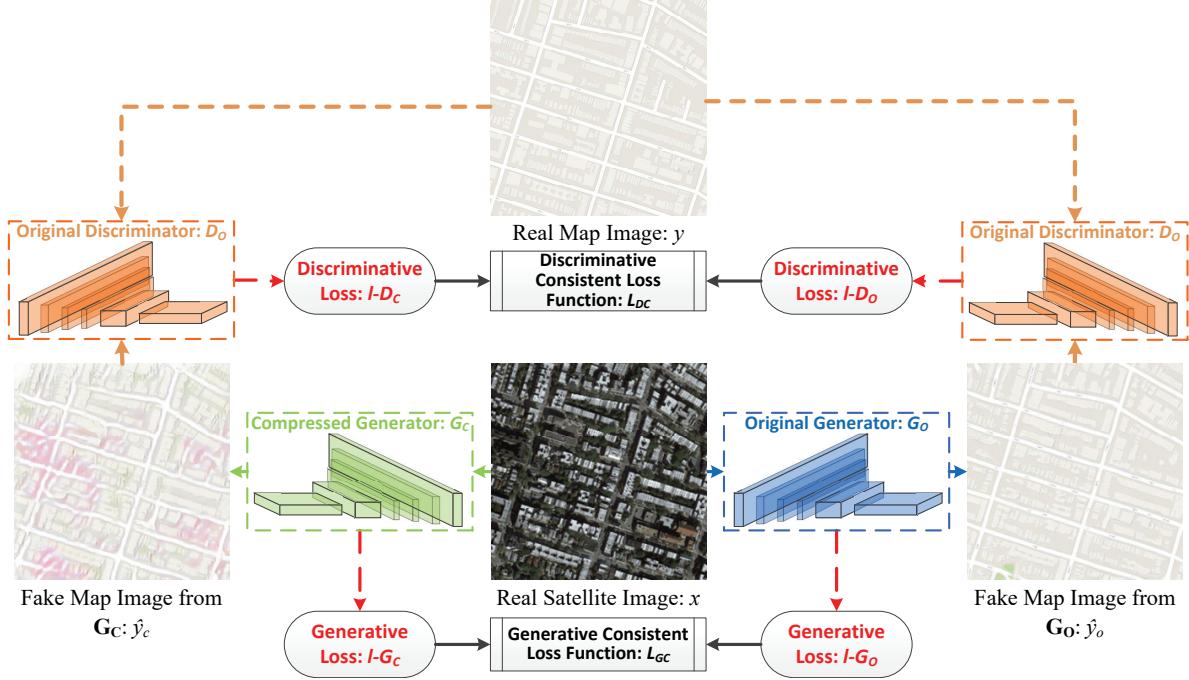


Figure 2. Workflow chart of GAN compression process.

where $l\text{-}Dis$ is the discriminative loss item, $l\text{-}GP$ is the gradient penalty loss item, and δ is a weighting factor (again, we use the same value as the baseline).

The overall loss function of GAN compression consists of generative and discriminative differences:

$$L_{Overall} = L_{GC}(l\text{-}G_O, l\text{-}G_C) + \lambda L_{DC}(l\text{-}D_O, l\text{-}D_C), \quad (7)$$

where λ is the parameter to adjust the percentages between generative and discriminative losses.

We showed promising results with this method above in the context of prior methods. In the following experiments, we investigate how well the method applies to other networks and tasks (Section 5) and how well the method works on different sparsity ratios and pruning granularities (Section 6).

5. Application to New Tasks and Networks

For the experiments in this section, we choose to prune individual weights in the generator. The final sparsity rate is 50% for all convolution and deconvolution layers in the generator (more aggressive sparsities are discussed in Section 6). Following AGP (Zhu & Gupta, 2018), we gradually increase the sparsity from 5% at the beginning to our target of 50% halfway through the self-supervised training process, and we set the loss adjustment parameter λ to 0.5 in all experiments. We use PyTorch (Paszke et al., 2017), implement the pruning and training schedules with Distiller (Zmora et al., 2019), and train and generate results with a V100

GPU (NVIDIA, 2017) using FP32 to match public baselines. In all experiments, the data sets, data preparation, and baseline training all follow from the public repositories - details are summarized in Table 2. We start by assuming an extra 10% of the original number of epochs will be required; in some cases, we reduced the overhead to only 1% while maintaining subjective quality. We include representative results for each task, but a more comprehensive collection of outputs for each experiment is included in the *Appendix*.

Image Synthesis. We apply the proposed compression method to DCGAN (Radford et al., 2016)³, a network that learns to synthesize novel images from some distribution. We task DCGAN with generating images that could belong to the MNIST data set, with results shown in Figure 3.

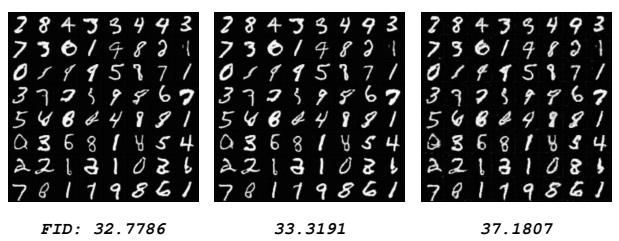


Figure 3. Image synthesis on MNIST dataset with DCGAN. Columns 1-2: Handwritten numbers generated by the original generator and pruned generator of 50%, 75% fine-grained sparsity.

³DCGAN baseline: <https://github.com/pytorch/examples/tree/master/dcgan>.

Table 2. Tasks and networks overview

Task	Network	Dataset	Resolution	FID Scores when Pruned to				
				0% (dense)	25%	50%	75%	90%
Image Synthesis	DCGAN	MNIST	64x64	50.391	50.128	50.634	50.805	51.356
Domain Translation	Pix2Pix	Sat → Map	256x256	17.636	17.897	17.990	20.235	24.892
Domain Translation	Pix2Pix	Sat ← Map	256x256	30.826	30.628	30.720	34.051	38.936
Style Transfer	CycleGAN	Monet → Photo	256x256	63.152	63.410	63.662	66.394	70.933
Style Transfer	CycleGAN	Monet ← Photo	256x256	31.987	32.102	32.346	33.913	41.409
Image-Image Translation	CycleGAN	Zebra → Horse	256x256	60.930	61.005	61.102	65.898	68.450
Image-Image Translation	CycleGAN	Zebra ← Horse	256x256	52.862	52.631	52.688	58.356	63.274
Image-Image Translation	StarGAN	CelebA	128x128	6.113	6.307	6.929	6.714	7.144
Super Resolution	SRGAN	DIV2K	≥ 512x512	14.653	15.236	16.609	17.548	18.376

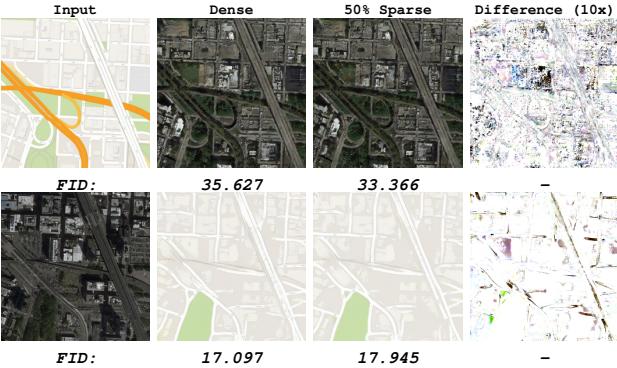


Figure 4. Representative results for domain translation: pix2pix. Row 1: map to satellite task, Row 2: satellite to map task.



Figure 5. Representative results for style transfer: CycleGAN. Row 1: Monet painting to photo, Row 2: photo to Monet painting.



Figure 6. Representative image-to-image translation results: CycleGAN. Row 1: zebra to horse, Row 2: horse to zebra.

Domain Translation. We apply the proposed compression method to pix2pix (Isola et al., 2017)⁴, an approach to learn the mapping between paired training examples by applying conditional adversarial networks. In our experiment, the task is synthesizing fake satellite images from label maps and vice-versa. Representative results of this bidirectional task are shown in Figure 4.

Style Transfer. We apply the proposed compression method to CycleGAN (Zhu et al., 2017a), used to exchange the style of images from a source domain to a target domain in the absence of paired training examples. In our experiment, the task is to transfer the style of real photos with that of the Monet’s paintings. Representative results of this bidirectional task are shown in Figure 5: photographs are given the style of Monet’s paintings and vice-versa.

Image-to-image Translation. In addition to the StarGAN results above (Section 3, Figure 1), we apply the proposed compression method to CycleGAN (Zhu et al., 2017a) performing bidirectional translation between zebra and horse images. Results are shown in Figure 6.

⁴Pix2pix, CycleGAN baseline: <https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix>.

Super Resolution. We apply self-supervised compression to SRGAN (Ledig et al., 2017)⁵, which uses a discriminator network trained to differentiate between upscaled and the original high-resolution images. We trained SRGAN on the DIV2K data set (Agustsson & Timofte, 2017), and use the

⁵SRGAN baseline: <https://github.com/xinntao/BasicSR>.

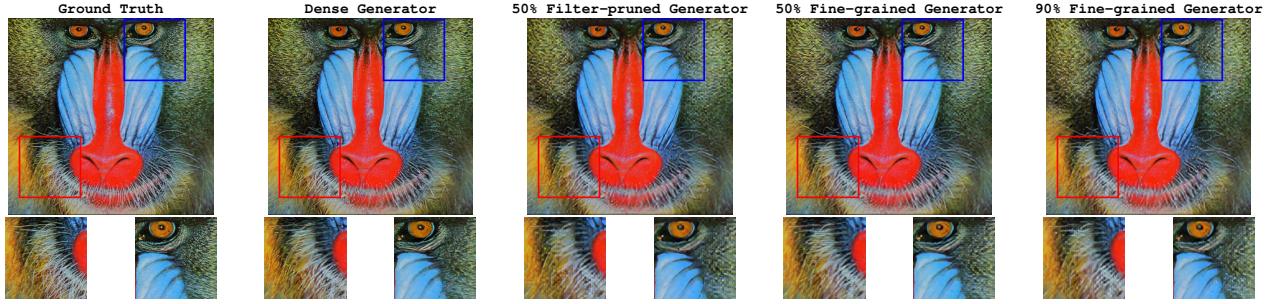


Figure 7. Representative super resolution results: SRGAN (with enlargements of boxed areas).

DIV2K validation images, as well as Set5 (Bevilacqua et al., 2012) and Set14 (Zeyde et al., 2010) to report deployment quality. In this task, quality is often evaluated by two metrics: Peak Signal-to-Noise Ratio (PSNR) (Huynh-Thu & Ghanbari, 2008) and Structural Similarity (SSIM) (Wang et al., 2004). We also show FID scores (Heusel et al., 2017) for our results in the results summarized in Table 3, and a representative output is shown in Figure 7. These results also include filter-pruned generators (see Section 6).

Table 3. PSNR, SSIM and FID indicators for Validation Datasets

Dataset	Original Generator			Filter-Compressed G			Element-Compressed G		
	PSNR	SSIM	FID	PSNR	SSIM	FID	PSNR	SSIM	FID
Set5	30.063	0.853	30.762	30.234	0.860	35.514	30.484	0.862	36.824
Set14	26.643	0.716	55.457	27.315	0.745	82.118	27.417	0.744	70.126
DIV2K	28.206	0.778	14.653	28.876	0.801	18.500	28.975	0.801	16.609

6. Effect of Pruning Ratio and Granularity

After showing that self-supervised compression applies to many tasks and networks with a moderate, fine-grained sparsity of 50%, we expand the scope of the investigation to include different pruning granularities and rates. From coarse to fine, we can compress and remove the entire filters (3D-level), kernels (2D-level), vectors (1D-level) or individual elements (0D-level). In general, finer-grained pruning results in higher accuracy for a given sparsity rate, but coarser granularities are easier to exploit for performance gains due to their regular structure. Similarly, different sparsity rates, leaving many nonzero weights or few, can result in varying levels of quality in the final network.

We pruned all tasks by removing both single elements (0D) and entire filters (3D). Further, for each granularity, we pruned to final sparsities of 25%, 50%, 75%, and 90%. Representative results for CycleGAN (Monet → Photo) and StarGAN are shown in Figure 8 and Figure 9, with results for all tasks in the *Appendix*. In general, 0D pruning is less invasive, even at higher sparsities. Up to 90% fine-grained sparsity, some fine details faded away in CycleGAN and StarGAN, but filter pruning results in drastic color shifts and loss of details at even 25% sparsity.

7. Conclusion and Future Work

Network pruning has been applied to many tasks, but never to GANs performing complex tasks. We showed that existing pruning approaches fail to retain network quality, as do training modifications aimed at compressing simple GANs by other methods applied to pruning. To solve this, we used a pre-trained discriminator to self-supervise the pruning of several GANs’ generators and showed this method performs well both qualitatively and quantitatively. Advantages of our method include:

- The results from the compressed generators are greatly improved over past work.
- The self-supervised compression is much shorter than the original GAN training process - only 1-10% of the original training time is needed.
- It is an end-to-end compression schedule that does not require objective evaluation metrics; final quality is accurately reflected in loss curves.
- We introduce a single optional hyperparameter (fixed to 0.5 for all our experiments).

We use self-supervised GAN compression to show that pruning whole filters, which can work well for image classification models (Li et al., 2017), may perform poorly for GAN applications. Even pruned at a moderate sparsity (e.g. 25% in Figure 8), the generated image has an obvious color shift and does not transfer the photorealistic style. In contrast, the fine-grained compression strategy works well for all tasks we explored. SRGAN seems to be an exception to filter-pruning’s poor results; we have to look closely to see differences, and it’s not clear which is subjectively better.

We have not tried to achieve extremely aggressive compression rates with complicated pruning strategies. Different models may be able to tolerate different amounts of pruning when applied to a task, which we leave to future work. Similarly, we have used network pruning to show the importance and utility of the proposed method, but self-supervised

Self-Supervised GAN Compression

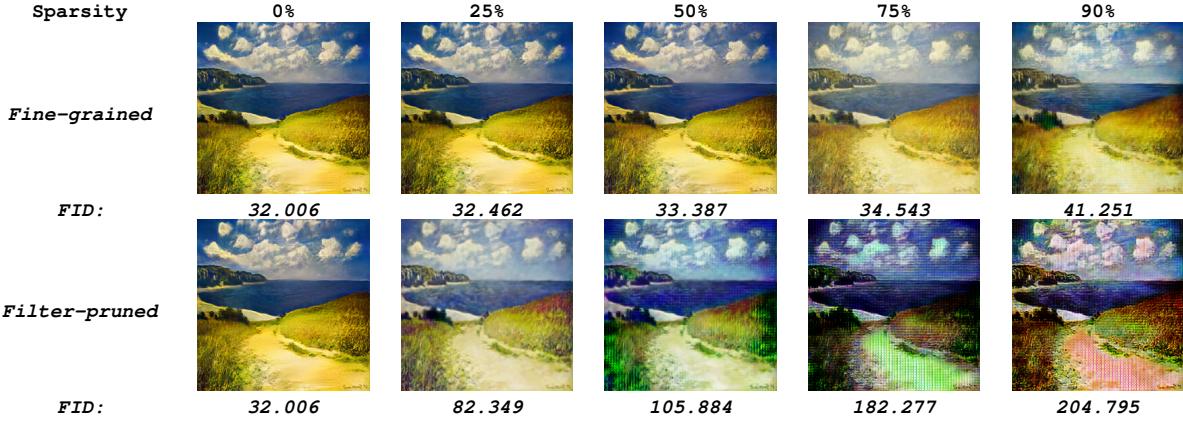


Figure 8. Representative results for pruning rate and granularity study of style transfer.

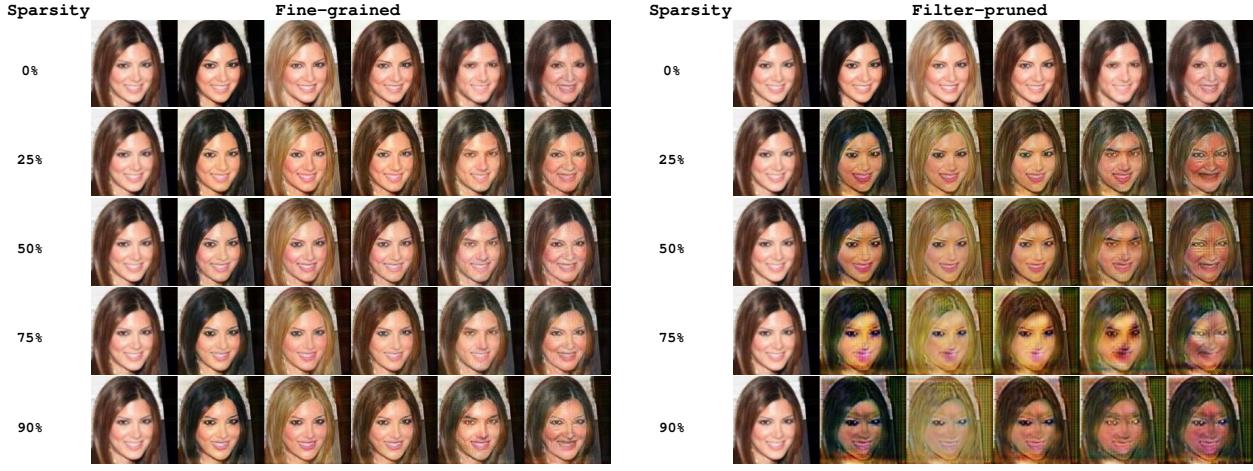


Figure 9. Representative results for pruning rate and granularity study of image-to-image translation.

compression is general to other techniques, such as quantization, weight sharing, etc. There are other tasks for which GANs can provide compelling results, and newer networks for tasks we have already explored; future work will extend our self-supervised compression method to these new areas. Finally, self-supervised compression may apply to other network types and tasks if a discriminator is trained alongside the teacher and student networks.

8. Appendix

More experiments and details can refer to the appendix file in repository: <https://gitlab.com/dxxz/Self-Supervised-GAN-Compression-Appendix>.

References

- Aguinaldo, A., Chiang, P., Gain, A., Patil, A., Pearson, K., and Feizi, S. Compressing GANs using knowledge distillation. *CoRR*, abs/1902.00159, 2019.
- Agustsson, E. and Timofte, R. NTIRE 2017 challenge on single image super-resolution: dataset and study. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2017.
- Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, 2017.
- Azadi, S., Fisher, M., Kim, V. G., Wang, Z., Shechtman, E., and Darrell, T. Multi-content GAN for few-shot font style transfer. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- Bevilacqua, M., Roumy, A., Guillemot, C., and Alberi-Morel, M. L. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. 2012.
- Brock, A., Donahue, J., and Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.
- Chen, W., Wilson, J. T., Tyree, S., Weinberger, K. Q., and Chen, Y. Compressing neural networks with the hashing trick. In *International Conference on Machine Learning*, 2015.
- Chen, X. Escoin: Efficient sparse convolutional neural network inference on GPUs. *CoRR*, abs/1802.10280, 2018.
- Cho, K., van Merriënboer, B., Gülcühre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Conference on Empirical Methods in Natural Language Processing*, 2014.
- Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
- Elsen, E., Dukhan, M., Gale, T., and Simonyan, K. Fast sparse convnets. *CoRR*, abs/1911.09723, 2019.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.
- Han, S., Pool, J., Tran, J., and Dally, W. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems*, 2015.
- Han, S., Mao, H., and Dally, W. J. Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. In *International Conference on Learning Representations*, 2016.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 2017.
- Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. In *NeurIPS Deep Learning and Representation Learning Workshop*, 2015.
- Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.
- Huynh-Thu, Q. and Ghanbari, M. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters*, 44(13):800–801, 2008.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- Jain, S. R., Gural, A., Wu, M., and Dick, C. Trained uniform quantization for accurate and efficient neural network inference on fixed-point hardware. *CoRR*, abs/1903.08066, 2019.
- Judd, P., Delmas, A., Sharify, S., and Moshovos, A. Cnvlutin2: Ineffectual-activation-and-weight-free deep neural network computing. *CoRR*, abs/1705.00125, 2017.
- Koratana, A., Kang, D., Bailis, P., and Zaharia, M. LIT: Learned intermediate representation training for model compression. In *International Conference on Machine Learning*, 2019.
- Kumar, R., Goyal, A., Courville, A., and Bengio, Y. Maximum entropy generators for energy-based models. *CoRR*, abs/1901.08508, 2019.
- Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A., Tejani, A., Totz, J., Wang, Z., et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690, 2017.
- Li, H., Kadav, A., Durdanovic, I., Samet, H., and Graf, H. P. Pruning filters for efficient convnets. In *International Conference on Learning Representations*, 2017.

- Lin, S., Ji, R., Yan, C., Zhang, B., Cao, L., Ye, Q., Huang, F., and Doermann, D. Towards optimal structured CNN pruning via generative adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.
- Liu, J., Zhang, J., Ding, Y., Xu, X., Jiang, M., and Shi, Y. PBGen: partial binarization of deconvolution based generators. *CoRR*, abs/1802.09153, 2018a.
- Liu, R., Fusi, N., and Mackey, L. Model compression with generative adversarial networks. In *International Conference on Learning Representations*, 2018b.
- Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, 2015.
- Micikevicius, P., Narang, S., Alben, J., Diamos, G. F., Elsen, E., García, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., and Wu, H. Mixed precision training. In *International Conference on Learning Representations*, 2018.
- Narang, S., Elsen, E., Diamos, G., and Sengupta, S. Exploring sparsity in recurrent neural networks. In *International Conference on Learning Representations*, 2017.
- NVIDIA. NVIDIA Tesla V100 GPU architecture. <https://images.nvidia.com/content/volta-architecture/pdf/volta-architecture-whitepaper.pdf>, 2017.
- Parashar, A., Rhu, M., Mukkara, A., Puglielli, A., Venkatesan, R., Khailany, B., Emer, J., Keckler, S. W., and Dally, W. J. SCNN: An accelerator for compressed-sparse convolutional neural networks. In *International Symposium on Computer Architecture*, 2017.
- Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in PyTorch. In *NIPS-W*, 2017.
- Radford, A., Metz, L., and Chintala, S. Unsupervised representation learning with deep convolutional generative adversarial networks. In *International Conference on Learning Representations*, 2016.
- Schmidhuber, J. Making the world differentiable: On using fully recurrent self-supervised neural networks for dynamic reinforcement learning and planning in non-stationary environments. *Institut für Informatik, Technische Universität München. Technical Report FKI-126*, 1990.
- Schmidhuber, J. Neural sequence chunkers. 1991.
- See, A., Luong, M.-T., and Manning, C. D. Compression of neural machine translation models via pruning. In *CoNLL*, 2016.
- Wang, K., Liu, Z., Lin, Y., Lin, J., and Han, S. HAQ: Hardware-aware automated quantization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019a.
- Wang, P., Wang, D., Ji, Y., Xie, X., Song, H., Liu, X., Lyu, Y., and Xie, Y. QGAN: Quantized generative adversarial networks. *CoRR*, abs/1901.08263, 2019b.
- Wang, Y., Xu, C., Xu, C., and Tao, D. Adversarial learning of portable student networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- Wen, W., Wu, C., Wang, Y., Chen, Y., and Li, H. Learning structured sparsity in deep neural networks. In *Advances in Neural Information Processing Systems*, 2016.
- Wen, W., He, Y., Rajbhandari, S., Zhang, M., Wang, W., Liu, F., Hu, B., Chen, Y., and Li, H. Learning intrinsic sparse structures within long short-term memory. In *International Conference on Learning Representations*, 2018.
- Zeyde, R., Elad, M., and Protter, M. On single image scale-up using sparse-representations. In *International Conference on Curves and Surfaces*, pp. 711–730. Springer, 2010.
- Zhu, F., Pool, J., Andersch, M., Appleyard, J., and Xie, F. Sparse persistent RNNs: Squeezing large recurrent networks on-chip. In *International Conference on Learning Representations*, 2018.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017a.
- Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., and Shechtman, E. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, 2017b.
- Zhu, M. and Gupta, S. To prune, or not to prune: exploring the efficacy of pruning for model compression. In *International Conference on Learning Representations*, 2018.
- Zmora, N., Jacob, G., Zlotnik, L., Elharar, B., and Novik, G. Neural network distiller: A python package for DNN compression research, 2019.