Are we pruning the correct channels in image-to-image translation models? (No, but we correct it.)

Yiyong Li[1], Zhun Sun[1*], Lichao[2] [*]corresponding author, zhunsun@gmail.com

[1]BIGO Ltd, [2]Tohoku University, [3]AIP, RIKEN

TL;DR

- The commonly employed lasso-based channel regularization approach prunes channels with large visual effects incorrectly.
- We build a novel perturbation model to analyze what channels should be pruned for the instance normalization (`IN`)-based models.
- Using the perturbation bound, we achieve better compression performance in both on-training and zero-shot scenarios.

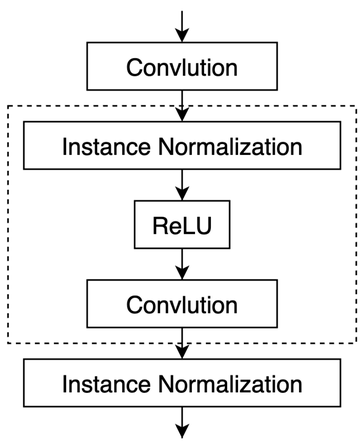
Code here.

(a) Input (b) un-pruned (c) eb=0.194 (d) eb=1.604 (e) eb=2.430

Fig. 1: Examples of `horse2zebra` generation results using a pre-trained CycleGAN with specific channels pruned. It is conspicuous that the appearances alter more severely as the perturbation error bound (eb) of the channel grows.

Background

- On **edge devices** and mobile applications, image-to-image translation GAN are commonly employed for visual effects.
- Obtaining **compact architectures** with pruning and distillation is demanded, due to the limited computational resources.
- We focus on architectures with stacked `Conv-IN-ReLU` **layers** that are edge-friendly.
- Pruning methods that are proposed for general neural networks do not fit the image-to-image translation GAN, since the visual effects are more sensitive than other tasks.



The basic building block to construct GANs.

Perturbation Analysis

Proposition (Perturbation error bound) 1 *Assume $\mathcal{Z} \in \mathbb{R}^{D\times W\times H}$ to be the output of `conv`, and $\mathcal{Z}^{\hat{i}}$ to denote the result by pruning the ith channel of the input. then the norm of perturbation $\Delta_i = \mathcal{Z} - \mathcal{Z}^{\hat{i}}$ is bounded by the following conditions: if $\gamma_i = 0$, then $\|\Delta_i\|_{\ell_1} = 0$; otherwise, we have*

$$\|\Delta_i\|_{\ell_1} \leq \begin{cases} WH\sum_{j\in[D]} F_{i,j}\left(\mathcal{W},\gamma,\mathbf{0}\right), & \beta_i \geq \tau_i \\ WH\sum_{j\in[D]} F_{i,j}\left(\mathcal{W},\gamma,\beta\right), & |\beta_i| < \tau_i \\ 0, & \text{otherwise} \end{cases}, \tag{1}$$

where $\mathbf{0} \in \mathbb{R}^C$ and $\tau_i = \sqrt{WH}|\gamma_i|$ for all $i \in [C]$, $F_{i,j}$ is the Sensitivity Measurement defined as

$$F_{i,j}\left(\mathcal{W},\gamma,\beta\right) = \sqrt{WH}|\gamma_i|\sqrt{\sum_{g,p\in[K]}\mathcal{W}(i,j,g,p)^2} + |\beta_i|\left|\sum_{g,p\in[K]}\mathcal{W}(i,j,g,p)\right|. \tag{2}$$

Step 1: Perturbation Error Bound as Loss.

Proposition 1 implies the worst-case scenario when pruning a certain channel. When applied as loss, we enforce a part of the channels have smaller bounds than the rest. **Minimizing the loss will push the model to learn more distinguish channels, which are selected to be pruned or not.**

$$L_{all} = L_{GAN} + \lambda_1 L_{dist} + \lambda_2 L_{BIG}, \tag{3}$$

$$L_{BIG}(\{\mathcal{W}_{[l]}\},\{\gamma_{[l]}\},\{\beta_{[beta]}\}) = \sum_{l\in[L]}\sum_{i\in[C_l]} P_{l,i}\left(\mathcal{W}_{[l]},\gamma_{[l]},\beta_{[l]}\right). \tag{4}$$

In the equation, $P_{l,i} = 0$ if $\gamma_{[l],i} = 0$, while if $\gamma_{[l],i} \neq 0$, then

$$P_{l,i} = \begin{cases} \sum_{j\in[D]} F_{i,j}\left(\mathcal{W}_{[l]},\gamma_{[l]},\mathbf{0}\right), & \beta_i \geq \tau_{[l],i} \\ \sum_{j\in[D]} F_{i,j}\left(\mathcal{W}_{[l]},\gamma_{[l]},\beta_{[l]}\right), & |\beta_i| < \tau_{[l],i} \\ 0, & \text{otherwise} \end{cases}, \tag{5}$$

Step 2: Update Re-scalar as Pruning.

We conduct the pruning by updating $\gamma^{(t)}_{[l],i} = \beta^{(t)}_{[l],i} = 0$, if one of the following four conditions is satisfied, we increase the threshold ρ_1, ρ_2 after the numbers of survival channels stop decreasing during training.

$$\begin{aligned} &\texttt{(i)}\quad \beta^{(t-1)}_{[l],i} \leq -\tau^{(t-1)}_{[l],i}, \quad \texttt{(ii)}\quad \gamma^{(t-1)}_{[l],i} = 0, \\ &\texttt{(iii)}\quad \frac{\sum_{j\in[D_l]} F_{i,j}\left(\mathcal{W}^{(t-1)}_{[l]},\gamma^{(t-1)}_{[l]},\mathbf{0}\right)}{\sum_{i\in[C_l]} P_{l,i}\left(\mathcal{W}^{(t-1)}_{[l]},\gamma^{(t-1)}_{[l]},\beta^{(t-1)}_{[l]}\right)} < \rho_1, \\ &\texttt{(iv)}\quad \frac{\sum_{j\in[D_l]} F_{i,j}\left(\mathcal{W}^{(t-1)}_{[l]},\gamma^{(t-1)}_{[l]},\beta^{(t-1)}_{[l]}\right)}{\sum_{i\in[C_l]} P_{l,i}\left(\mathcal{W}^{(t-1)}_{[l]},\gamma^{(t-1)}_{[l]},\beta^{(t-1)}_{[l]}\right)} < \rho_2. \end{aligned} \tag{6}$$

Results

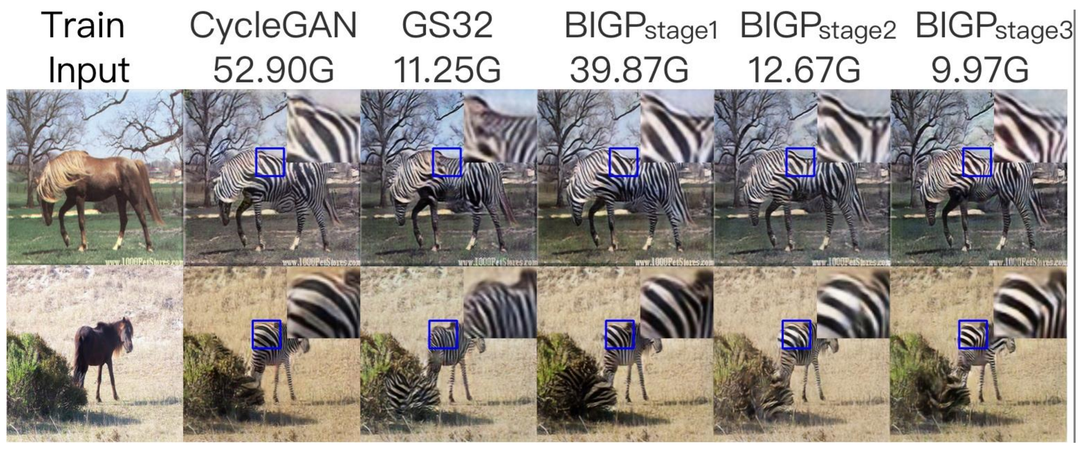


Fig. 2: Examples of CycleGAN [1] results on `horse2zebra` **train** set. The methods and their FLOPs (in G) are annotated above the images. Each row shows a different sample. Our pruned model generates samples that have similar stripes to the one generated by CycleGAN.

`horse2zebra`

Model	FLOPs	FID	# Pruned
GAN-Comp	2.67G	64.95	–
DMAD	2.41G	62.96	–
CAT	2.55G	60.18	–
GCC	2.40G	59.19	–
OMGD [2]	1.408G†	51.97	–
+ZSP	1.406G	51.70	2
+L_{BIG}	1.408G*	46.72	–
+L_{BIG}+ZSP	1.397G	47.03	10

`summer2winter`

Model	FLOPs	FID	# Pruned
Auto-GAN	4.34G	78.33	–
GAN-KD	3.20G	80.10	–
SP-KD	3.20G	76.59	–
DMAD	3.18G	78.24	–
OMGD [2]	1.408G†	73.79	–
+ZSP	1.404G	73.70	6
+L_{BIG}	1.408G*	73.12	–
+L_{BIG}+ZSP	1.398G	73.13	9

Table 1: Performance of knowledge distillation models combining the BIG loss pre-training and/or zero-shot pruning. † stands for the officially released models, * stands for our pre-trained models. Even in the currently reported best performance distillation models, there still exist channels that can be pruned without influencing the generation results.

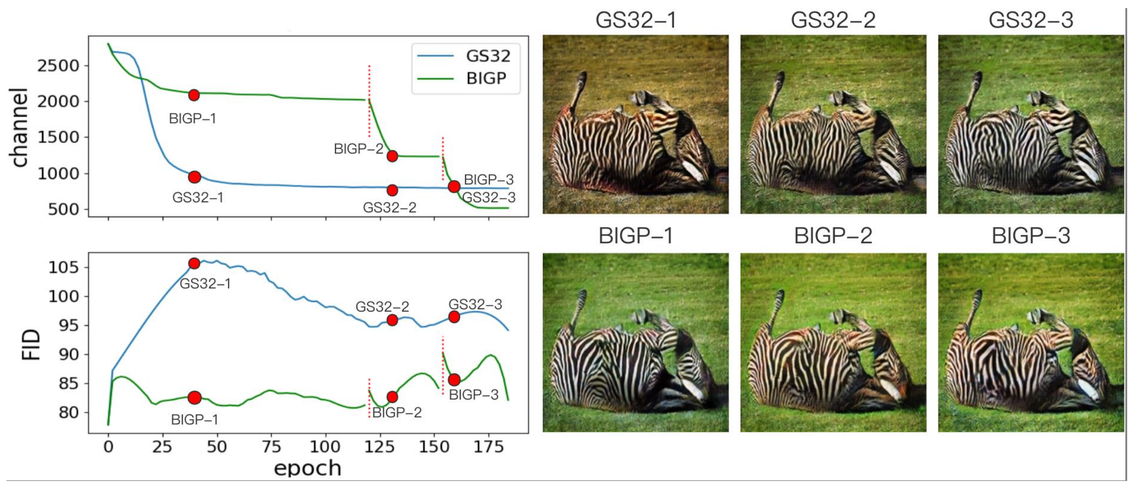


Fig. 3: (a) Learning curves of FID score and the number of channels along with the epoch of a complete pruning run. (b) The generated zebra images from the models are marked as red dots in (a), and the methods and the model ids are annotated above the images.

References

[1] Zhu, Jun-Yan and Park, Taesung and Isola, Phillip and Efros, Alexei A: *Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks, 2017*

[2] , Ren, Yuxi and Wu, Jie and Xiao, Xuefeng and Yang, Jianchao: *Online multi-granularity distillation for gan compression, 2021*