An Unsupervised Way to Understand Artifact Generating Internal Units in Generative Neural Networks

Haedong Jeong[1,2], **Jiyeon Han**[2], **Jaesik Choi**[2,3]*

[1] Ulsan National Institute of Science and Technology (UNIST), South Korea
[2] Korea Advanced Institute of Science and Technology (KAIST), South Korea
[3] INEEJI, South Korea
{haedong.jeong, j.han, jaesik.choi}@kaist.ac.kr

Abstract

Despite significant improvements on the image generation performance of Generative Adversarial Networks (GANs), generations with low visual fidelity still have been observed. As widely used metrics for GANs focus more on the overall performance of the model, evaluation on the quality of individual generations or detection of defective generations is challenging. While recent studies try to detect featuremap units that cause artifacts and evaluate individual samples, these approaches require additional resources such as external networks or a number of training data to approximate the real data manifold. In this work, we propose the concept of *local activation*, and devise a metric on the local activation to detect artifact generations without additional supervision. We empirically verify that our approach can detect and correct artifact generations from GANs with various datasets. Finally, we discuss a geometrical analysis to partially reveal the relation between the proposed concept and low visual fidelity.

1 Introduction

Since the adversarial generative training scheme (Goodfellow et al. 2014) emerged, deep generative neural networks (DGNNs) have shown incredible performance on image generation tasks. From recent research with generative adversarial networks (GANs), various structures and training strategies (Brock, Donahue, and Simonyan 2019; Karras et al. 2018; Miyato et al. 2018; Karras, Laine, and Aila 2019) have been proposed to overcome the weaknesses of the adversarial training scheme (e.g., unstable training) and to accelerate the improvements of visual fidelity of the generations.

Despite significant improvements, models sometimes present undesirable outcomes such as perceptually defective generations called *artifacts*. Various metrics have been suggested to evaluate the performance of a generator (Salimans et al. 2016; Heusel et al. 2017; Sajjadi et al. 2018; Kynkäänniemi et al. 2019). However, it is non-trivial to evaluate the visual fidelity of each individual sample because existing metrics mainly focus on the distributional difference between the real dataset and the generations in the feature manifold. A research, which uses the nearest neighbor based similarity in the feature manifold (Kynkäänniemi et al. 2019), was proposed as an alternative to estimate the quality of individual generations. Although this method has shown effectiveness in evaluation, it requires a huge amount of real data and an external network for feature embedding to make the scoring process reliable.

A few studies have been conducted to understand the internal generation mechanism of GANs to detect or correct the individual generations with low visual fidelity. In GAN Dissection (Bau et al. 2019), the authors identify the defective units that mainly cause artifacts based on a set of generations on which a featuremap unit is highly activated. The authors further improve the fidelity of individual generations by zero-ablating the detected featuremap units. A similar approach trains an external classifier to extract the region with low visual fidelity in the individual generations and identifies internal units related to the extracted region (Tousi et al. 2021). On the other hand, manipulation of the latent code based on the binary linear classifier has been proposed to correct the artifact (Shen et al. 2020). While these approaches can be utilized to evaluate the fidelity of individual samples, they still require additional resources such as a human annotation process.

In this paper, we propose the concept of *local activation* to detect and correct the artifact generations in an unsupervised manner. We also discuss a geometrical analysis to partially investigate the relation between the local activation and low visual fidelity of individual generations. The main advantages of our method are twofold: (1) external networks or supervisions are unnecessary to detect and correct artifact generations. The evaluation is performed solely on the target generator by using the internal property for scoring, and (2) the proposed approach can be applicable to various structures of GANs for evaluating the visual fidelity, because the proposed approach is based on neurons that are the common basic components of neural networks. We experimentally verify that our method can detect and correct artifact effectively on PGGAN (Karras et al. 2018), and StyleGAN2 (Karras et al. 2020) with various datasets.

2 Related Work

Deep Generative Neural Networks DGNNs are the models which approximate the input distribution given a target with neural networks. Representative architectures in-

*Corresponding Author

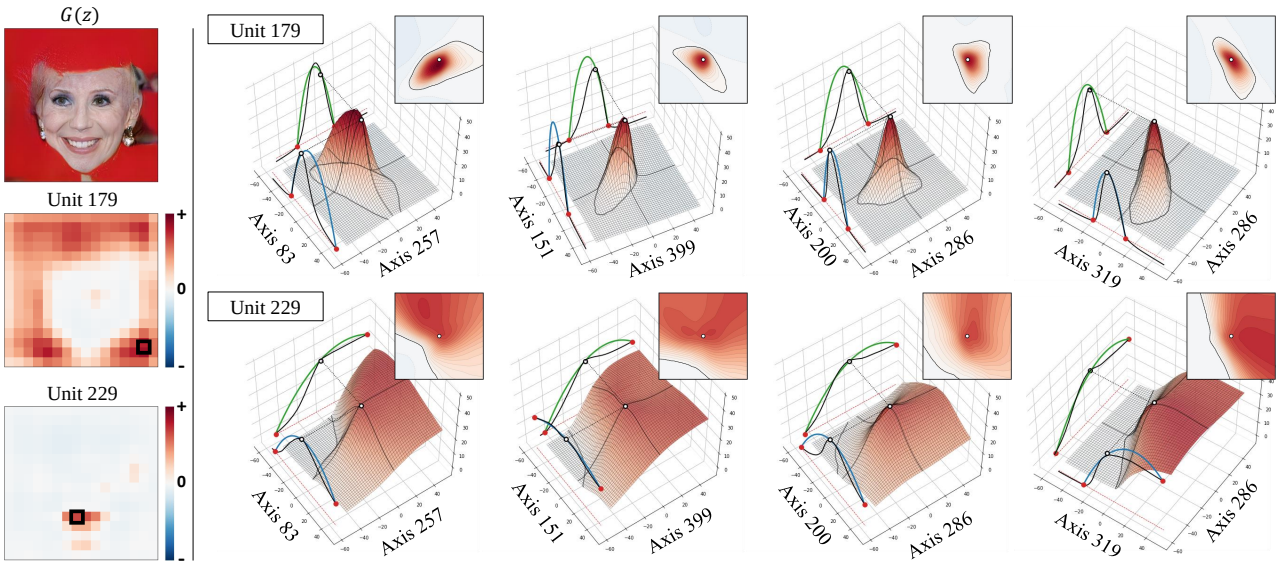


Figure 1: An illustrative example of the locally activated neuron in PGGAN model trained on CelebA-HQ. We manually select two featuremap units in layer 6 ($\in \mathbb{R}^{512\times16\times16}$) which are expected to relate to low visual fidelity (red background in $G(z)$) and mouth. (Left) The black box in each featuremap unit means the spatial information of the selected neuron. (Right) The sections of activation patterns in the latent space. The black solid line in each side means the activation pattern in the corresponding axis. The axis is randomly selected for the visualization. The red dots indicate the change points (Definition 1) and green/blue lines are the approximated curvature of local activation for each axis. The neuron in unit 179 has more locally activated pattern compared to the activation pattern of the neuron in unit 229.

clude variational autoencoder (VAE) (Kingma and Welling 2014), neural language models (Peters et al. 2018; Kenton and Toutanova 2019; Brown et al. 2020) and GANs. In particular, the adversarial training between a generator and a discriminator (Goodfellow et al. 2014) has shown impressive performance in image generation (Karras et al. 2018; Karras, Laine, and Aila 2019; Karras et al. 2020; Brock, Donahue, and Simonyan 2019).

Analysis for Interior Mechanism of GANs GAN Dissection (Bau et al. 2019) proposes a framework to investigate the generative role of each featuremap unit in GANs. It is shown that artifact generations can be improved by ablating units that are related to artifact generations. Another work (Shen et al. 2020) trains a linear classifier based on artifact-labeled data and removes artifacts by moving the latent code over the trained hyperplane. A sampling method with the trained generative boundaries was suggested to explain shared semantic information in the generator (Jeon, Jeong, and Choi 2020). Classifier-based defective internal featuremap unit identification was devised (Tousi et al. 2021). The authors increase the visual fidelity by sequentially controlling the generation flow of the identified units. Analyses for latent space of the generator were also performed to manipulate the semantic of the generation (Peebles et al. 2020; Härkönen et al. 2020). Our work focuses more on the generation process and the relation between defective generation and the internal characteristics connected from the latent space.

Metric for Generative Model Various metrics have been proposed to evaluate the performance of generative models and each properties are well-summarized in (Borji 2019). Although Fréchet Inception Distance (FID) (Heusel et al. 2017) and Inception Score (IS) (Salimans et al. 2016) have shown robustness to image distortion, they sometimes assign high scores for generations with low visual fidelity. Precision and Recall (P&R) is a surrogate metric to quantify mode dropping and inventing based on training data (Sajjadi et al. 2018; Kynkäänniemi et al. 2019). The authors also devised Realism Score (RS) to evaluate the visual fidelity of individual samples by comparing feature embeddings with training data. Perceptual path length (PPL) is another metric that quantifies the smoothness of the latent space with a hypothesis that the region in the latent space for defective generations has a small volume (Karras, Laine, and Aila 2019).

3 Locally Activated Neurons in GANs

In this section, we present our main contribution, the concept of *local activation* and its relation with low visual fidelity for individual generations. From previous research (Bau et al. 2019; Jeon, Jeong, and Choi 2020; Tousi et al. 2021), we can presume that each internal featuremap unit in the generator handles a specific object (e.g., tree, glasses) for the final generation. In particular, an artifact that has low visual fidelity can also be considered as a type of object. Thus, it is possible to identify the units causing low visual fidelity. To expand these observations, we focus on neurons as the basic component of a featuremap unit.

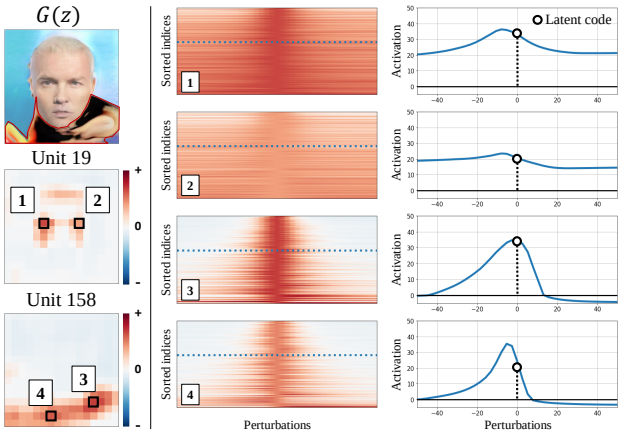


Figure 2: Activation patterns of the neurons in the manually selected featuremap units for the given latent code z. The middle column represents the heatmaps of the activation patterns around z where each row corresponds to each axis of the latent space. The rows are sorted by the local activation in the descending order from top to bottom. The right-most column represents the activation patterns for a specific axis (the dotted rows in the middle column) in the 2D representation. The neurons related to the defective region are more locally activated.

3.1 Quantification of Local Activation

We observe the neurons that correspond to the artifact region often show a bounded activation pattern. Figure 1 shows one example of artifact generation and two featuremap units in PGGAN. Unit 179 is highly correlated to the artifact region (red background) in the generation and unit 229 corresponds to the mouth part, which shows high visual fidelity. The right side of Figure 1 shows 3D representations of the activation patterns in the latent space for two neurons from unit 179 and unit 229, respectively. The neuron from unit 229 shows high activation over a large area in the latent space. In contrast, the neuron from unit 179, which corresponds to the artifact region, shows high activation only in a restricted area across various pairs of axes. Figure 2 further supports how the activation patterns are different between the artifact-related neurons and the normal neurons. In the second column, activation patterns of the artifact-related neurons (neurons 3 and 4) are more sharply concave across the latent axes compared to the normal neurons (neurons 1 and 2). The concave shape of the activation pattern suggests that the activation is bounded and concentrated around the given latent code.

From the observations in Figures 1 and 2, we suspect that the bounded activation pattern may be related to the visual fidelity of a generation. We call this bounded activation pattern *local activation* and the corresponding neuron a *locally activated neuron*. However, it is non-trivial to exactly quantify the local activation for an internal neuron in the latent space, because (1) commonly used generators have high dimensional latent space, and (2) the activation pattern forms a highly non-convex shape in the latent space. To mitigate these problems, we approximate the curvature of the local activation pattern with a line search for each latent dimension within the empirical search bound.

Let the generator G with L layers be $G(z) = g_L(g_{L-1}(\cdots(g_1(z)))) = g_{L:1}(z)$, where z is a vector in the latent space $\mathcal{Z} \subset \mathbb{R}^{D_z}$, $g_l(h_{l-1}) = \sigma(w_{g_l}^\top h_{l-1})$, $h_{l-1} = g_{l-1:1}(z)$, and $\sigma(\cdot)$ is an activation function such as LeakyReLU or ReLU[1]. One can express the bias of each layer in the homogeneous representation with this unified equation by applying an additional dimension for the bias. For the i-th neuron $g^i_{l:1}(z)$ of $g_{l:1}(z)$, we can obtain an activation pattern with the line search over the perturbation range for each latent dimension and we define the left/right change points for each activation pattern as follows.

Definition 1 (Change Point) *Let the given latent code be z_0, the dimension index be $d \in \{1, ..., D_z\}$, the search bound be $R > 0$, and the canonical basis be $\boldsymbol{e}_d = (0, ..., 0, 1, 0, ..., 0)^\top$ with the nonzero component at position d. For the set of change point $P = \{r | g^i_{l:1}(z_0 + r \cdot \boldsymbol{e}_d) = 0\} \cup \{-R, R\}$ for $r \in [-R, R]$, the right and the left change points of i-th neuron at the l-th layer are defined respectively as,*

$$ {}^r p = \min_{\forall r \in P; r \geq 0} (r) \quad \textit{and} \quad {}^l p = \max_{\forall r \in P; r \leq 0} (r). \tag{1}$$

We note that if there are no points where activation signs are changed, the search bounds are considered as the change points by Definition 1. We approximate the curvature of the local activation by computing the curvature (the coefficient of the second degree term) of the quadratic approximation of three points (the left/right change points and the given latent code z_0) for each latent axis, and averaging over the latent dimensions. The green and blue curves in Figure 1 illustrate the approximated quadratic functions for quantifying the local activation.

Definition 2 (Curvature of Local Activation (CLA)) *Let the given latent code be z_0 and the left/right change points be ${}^l p$ and ${}^r p$ respectively as in Definition 1. The right slope is defined as ${}^r s = (g^i_{l:1}(z_0 + {}^r p \cdot \boldsymbol{e}_d) - g^i_{l:1}(z_0))/{}^r p$ and the left slope is ${}^l s = (g^i_{l:1}(z_0 + {}^l p \cdot \boldsymbol{e}_d) - g^i_{l:1}(z_0))/{}^l p$ for a latent dimension d. With $C_{i,l}(d, z_0) = ({}^r s - {}^l s)/({}^r p - {}^l p)$, the curvature of local activation for the i-th neuron in the l-th layer around the given latent code z_0 is defined as,*

$$ \bar{C}_{i,l}(z_0) = \frac{1}{D_z} \sum_{d=1}^{D_z} C_{i,l}(d, z_0). \tag{2}$$

Although the definitions are constructed in the continuous space, we empirically use a grid search with search bound $R = 30$, dividing the search range by 20 for experiments throughout the paper. Details of the hyperparameter setting are provided in Appendix A.

Figure 3 shows the featuremap units that have the highest average CLA over the neurons in each unit. We can identify that the activated region for the units with a high CLA is semantically aligned with the artifact area in the generation.

[1]There are various activation functions for deep neural networks, we only consider LeakyReLU and ReLU function in this paper.

Figure 3: CLA based artifact unit identification. Bi-linear upsampled featuremap units are overlaid with the original generation.

3.2 Learning Dynamics for Local Activation

This section explores the dynamics of CLA for the epochs to validate the correlation between the visual fidelity and the magnitude of CLA. The experiments are performed with pre-trained snapshots of PGGAN model trained on CelebA-HQ[2]. First, we manually select the featuremap unit related to the defective area in layer $6 \in \mathbb{R}^{512\times16\times16}$. Next, we observe the change of local activation and the artifact emerging process during the training. Figure 4 indicates the change of the defective area (low visual fidelity) in the generation and the corresponding CLA. We can identify that the CLA increases when the activation area decreases or the activation value increases in a small area.

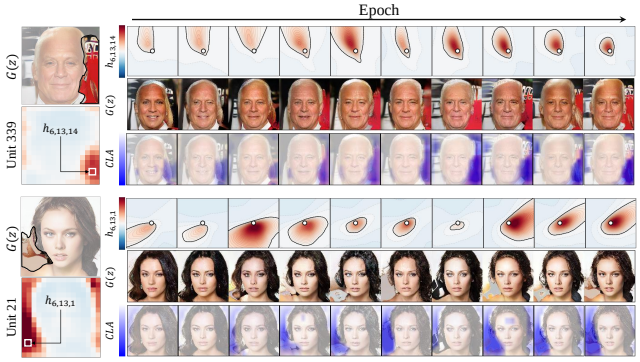


Figure 4: (First row) The visualization of activation pattern in the latent space for the neuron (white box in the featuremap unit) with selected axes. (Second row) The generation $G(z)$ for the fixed latent code. (Third row) Calculated CLA for the neurons in the unit.

4 Experimental Evaluations

This section presents analytical results of empirical relations between the low visual fidelity of individual generations and the proposed concept. We select two different GANs with various datasets. We use the pre-trained networks from the authors' official github; (1) PGGAN trained on LSUN-Bedroom, LSUN-Church outdoor (Yu et al. 2015) and CelebA-HQ[2] and (2) StyleGAN2 trained on LSUN-Car, LSUN-Cat, LSUN-Horse (Yu et al. 2015) and FFHQ[3]. To evaluate visual fidelity for individual samples, we define score as,

$$S_l(z) = \sum_i |min(\bar{C}_{i,l}(z), 0) * sign(max(h_{l,i}, 0)) + max(\bar{C}_{i,l}(z), 0) * sign(min(h_{l,i}, 0))|. \quad (3)$$

The defined score considers the degree of concavity/convexity for positive/negative activation, respectively. If the summation value of the given generation $G(z)$ is larger than other generations, we can expect that $G(z)$ has a low visual fidelity.

4.1 Qualitative Results

We randomly select 10k latent codes without truncation for each GAN and calculate the CLA on layer $4 \in \mathbb{R}^{512\times8\times8}$ for each generation. We choose the top/bottom 1k samples as High/Low CLA groups based on the score, respectively.

Artifact Detection Figure 8 depicts the results of detection in each GAN. We observe that the generations with a high CLA appear to be more defective than those with a low CLA. For example, in StyleGAN2 with LSUN-Car, we can identify that the generations which have a high CLA do not include clear information of the car compared with generations that have a low CLA. More detection results are available in the Appendix C-I.

Artifact Correction To validate that the locally activated neurons are related to the artifact, we perform an ablation study on the High CLA group as described in (Tousi et al. 2021). Instead of training an external classifier to identify the artifact causing internal units, we use the average CLA over the neurons in each unit. We set the hyperparameters as follows: stopping layer $l = 4$, the number of ablation units $n = 100$, and the maintain ratio $\lambda = 0.9$ for correction. We measure the RS after correction (last column in Table 1.). In Figure 5, we observe that when the generations contain severe artifacts, as in the cases of PGGAN with LSUN-Bedroom or LSUN-Church, we may need a more sophisticated method than simple ablation to correct the artifact. Nevertheless, we can improve the visual fidelity in most GANs in the experiments with simple ablation. From the detection and correction experiments, we believe that the locally activated neurons have a strong relationship with low visual-fidelity in the generation.

4.2 Quantitative Results

To quantify the fidelity of the detected generations, we calculate RS and PPL for each group; (1) low CLA and (2) high CLA and (3) random selection (30 trials). We use 30k real images for each model to calculate RS and set the number of neighborhood $k = 3$. For PPL, we perform interpolation in the latent space z with $\epsilon = 10^{-4}$. Table 1 indicates the scores for each group in various GANs. We can identify that the high CLA groups have low RS and high PPL compared to random groups. The results consistently show that

[2]https://github.com/tkarras/progressive_growing_of_gans

[3]StyleGAN2:https://github.com/NVlabs/stylegan2

Table 1: The detection and correction results on various GANs.

Metric	Model	Dataset	Random	Low CLA	High CLA	Correction
RS (↑ is better)	PGGAN	LSUN-Bedroom	1.028±0.003	**1.042**	1.017	1.002
		LSUN-Church	1.036±0.004	**1.059**	1.012	1.000
		CelebaA-HQ	1.076±0.004	**1.132**	1.011	1.018
	StyleGAN2	LSUN-Car	1.066±0.004	**1.084**	1.044	1.061
		LSUN-Cat	1.048±0.004	**1.071**	1.027	1.047
		LSUN-Horse	**1.056±0.004**	1.046	1.053	1.054
		FFHQ	1.075±0.004	1.077	1.069	**1.097**
PPL (↓ is better)	PGGAN	LSUN-Bedroom	423.8±7.1	**243.9**	683.3	-
		LSUN-Church	356.4±7.9	**213.7**	558.0	-
		CelebaA-HQ	243.1±12.9	**114.9**	443.7	-
	StyleGAN2	LSUN-Car	1472.6±29.3	**920.7**	1938.9	-
		LSUN-Cat	1501.3±27.7	**1053.2**	2060.4	-
		LSUN-Horse	1207.4±21.5	**885.5**	1552.5	-
		FFHQ	484.9±19.2	**377.7**	596.2	-

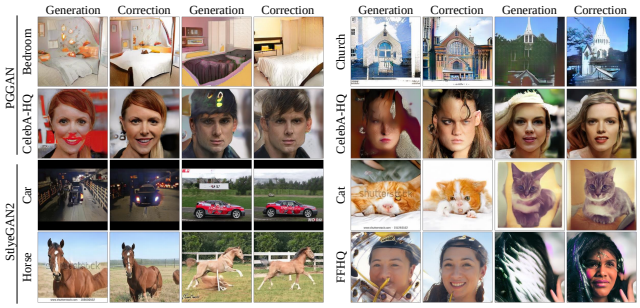
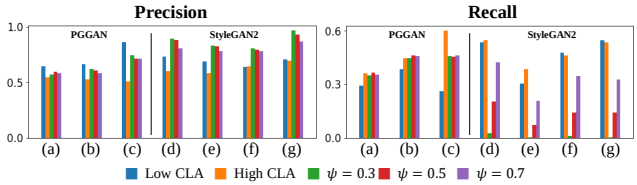


Figure 5: Examples for the correction results on various GANs for High CLA group. More examples are available in Appendix J.



(a): Bedroom (b) Church (c) CelebA-HQ (d) Car (e) Cat (f) Horse (g) FFHQ

Figure 6: Precision and recall for each group in various GANs.

the proposed method can effectively identify the generations with low visual fidelity. We note that the proposed method only uses the internal property to evaluate the visual fidelity of individual samples.

Diversity and Fidelity To compare diversity and the fidelity in each group, we calculate precision and recall (Kynkäänniemi et al. 2019) with the truncated samples as the baseline[4]. In Figure 6, we can identify that the low CLA group shows the higher precision (fidelity) with slightly lower recall (diversity) comparing to the high CLA group. However, the low CLA group shows much larger recall (diversity) comparing to the truncated samples for StyleGAN2. For PGGAN, the recall is higher on the truncated samples but the precision is higher on the low CLA group.

5 Discussion

5.1 Geometrical Interpretation of LA

In this section, we investigate the relation between the local activation and the visual fidelity under the specific condition. We begin by specifying the neurons in terms of whether they positively/negatively contribute to the discriminator's decision. The discriminator $D(x)$ can be similarly described as the generator in Section 3.1 where x is the target to generate such as a face image. The output of discriminator y for the latent code z is represented as $y = D(G(z))$. When we consider one instance z_0 and the corresponding feature vector for the $l-1$-th layer $\bar{h}_{l-1}$, one can linearize the networks G and D with a piece-wise linear activation function which is commonly used in the modern GANs. Let $\gamma \geq 0$ be the slope parameter for LeakyReLU (in ReLU case, γ is zero) and $w_{g_l,i}$ be the i-th column vector. The corresponding linearized parameter $\bar{w}_{g_l,i}$ is defined as,

$$\bar{w}_{g_l,i} = \begin{cases} w_{g_l,i} & \text{where} \quad w_{g_l,i}^\top \bar{h}_{l-1} \geq 0 \\ \gamma \cdot w_{g_l,i} & \text{otherwise} \end{cases} \tag{4}$$

We can then write the linearized generator as $\bar{G}(z_0) = W_G^\top \bar{h}_l$ and the linearized discriminator as $\bar{D}(\bar{G}(z_0)) = W_D^\top \bar{G}(z_0)$ where $W_G^\top = \bar{w}_{g_L}^\top \cdots \bar{w}_{g_{l+1}}^\top \in \mathbb{R}^{D_x \times D_l}$ and $W_D^\top = \bar{w}_{d_L}^\top \cdots \bar{w}_{d_1}^\top \in \mathbb{R}^{1 \times D_x}$. The output of the discriminator $\bar{y}$ is paraphrased as,

$$\bar{y} = \bar{D}(\bar{G}(z_0)) = W_D^\top W_G^\top \bar{h}_l = \sum_i W_D^\top W_{G,i}^\top \bar{h}_{l,i} \tag{5}$$

where $W_{G,i}^\top$ is the i-th column vector of $W_G^\top$ and $\bar{h}_{l,i}$ is the i-th element of the vector $\bar{h}_l$. We note that W_D is the nor-

[4]The latent z space for PGGAN and the w space for StyleGAN2

mal vector of the decision boundary to score the visual quality (real or fake) of the current generation $W_G^\top \bar{h}_l$. The current generation can be represented as a linear combination of $\{W_{G,i}^\top\}_i$ with the coefficients $\{\bar{h}_{l,i}\}_i$. The contribution of the i-th neuron in the l-th layer to the discriminator output $\bar{y}$ is $W_D^\top W_{G,i}^\top \bar{h}_{l,i}$. We determine that the i-th neuron has a negative/positive contribution if the contribution of the i-th neuron decreases/increases for the decision of the discriminator. For example, when $W_D^\top W_{G,i}^\top \bar{h}_{l,i} < 0$, the i-th neuron has a negative contribution.

We perform a geometrical analysis for the neurons that have a negative/positive contribution to the output of discriminator in the vanilla GAN (Goodfellow et al. 2014). We begin with describing how to update the parameters related to the direction of the contribution of each neuron and then analyze the consequences of updates. The loss function for the generator G and discriminator D is defined as,

$$\min_G \max_D V(D,G) = E_{x\sim p_{data}(x)}[\log f(D(x))] + E_{x\sim p_z(z)}[\log(1 - f(D(G(z))))] \quad (6)$$

where $f(\cdot)$ is the sigmoid function. For w_{g_l}, the updated parameter $w_{g_l}^+$ by the stochastic gradient descent of the given latent code z_0 with linearized form is described as,

$$w_{g_l}^+ = w_{g_l} + \eta c_0 f'(\bar{y}) W_D^\top W_G^\top \bar{h}_{l-1} \quad (7)$$

where $c_0 = (1 - f(D(G(z_0))))^{-1}$ and η is learning rate. The i-th column vector $w_{g_l,i}$ induces activation of the i-th neuron in the l-th layer ($\bar{h}_{l,i}$), and is updated by the direction $\bar{h}_{l-1}$ with weight $\delta_i = \eta c_0 f'(\bar{y}) W_D^\top W_{G,i}^\top \in \mathbb{R}$. Figure 7 presents geometrical illustrations of the update for four possible cases.

In Figure 7, we can observe that the perpendicular distance between the generative boundary and the feature vector $\bar{h}_{l,i}$ (colored dotted lines) decreases in the negative contribution cases when the learning rate is sufficiently small. Further, the magnitude of the activation value for $\bar{h}_{l,i}$ also decreases in the negative contribution cases ($\bar{h}_{l,i} \cdot \delta_i < 0$) as

$$\begin{aligned} \bar{h}_{l-1}^\top w_{g_l,i}^+ &= \bar{h}_{l-1}^\top (w_{g_l,i} + \delta_i \bar{h}_{l-1}) \\ &= \bar{h}_{l-1}^\top w_{g_l,i} + \delta_i \|\bar{h}_{l-1}\|^2. \end{aligned} \quad (8)$$

From the analysis on the vanilla GAN, we can presume that when a neuron in the generator negatively contributes to the discriminator output, the generator tries to deactivate the neuron by reducing the activation and distance to deceive the discriminator. As the penalization can be applied for the arbitrary latent code z, if a neuron has the negative contribution consistently during the training, the corresponding activated region in the latent space will shrink. If the locally activated region is not fully removed during the training, the corresponding neuron may generate artifacts when highly activated.

We note that although the analysis can suggest the partial explanations, the theoretical reasons for the observed relation are still remained as an open question.

5.2 Comparisons with PPL

We discuss the differences from the Perceptual Path Length (PPL), which is the most similar to the proposed method among feasibility metrics in that both measure the smoothness of the latent space of the generator. The first difference is that PPL needs an external network (e.g., pre-trained VGG16) to quantify the smoothness in the latent space. The dependency on the external network not only requires additional resources but also raises a limitation that the reliability depends on the capability of the external network. In other words, if the class of generations is not well-aligned to the external network, it becomes non-trivial to guarantee the reliability of the quantified perceptual distance in the feature space of the external network. Secondly, PPL measures the first order derivative whereas our method measures the second order derivative. Even though the path length regularization is applied in StyleGAN v2 to regularize the first order derivative, we can still observe high CLA.

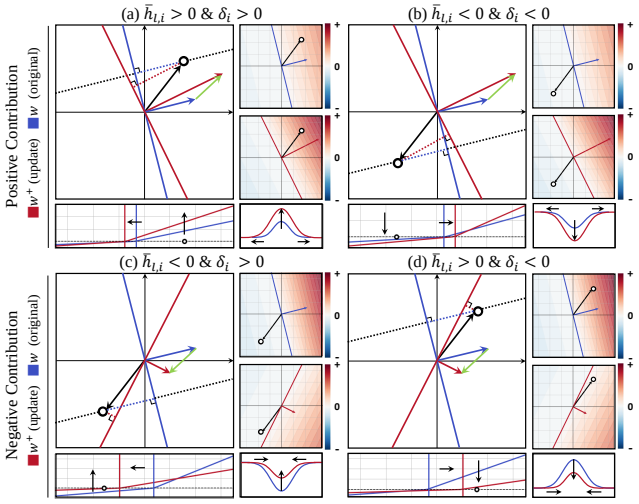


Figure 7: The geometrical illustrations of update cases ($\bar{h}_{l-1} \rightarrow \bar{h}_{l,i}$). The black arrow and the dot represent $\bar{h}_{l-1}$, the green arrow is $\delta_i \bar{h}_{l-1}$ (update term), the blue arrow is $\bar{w}_{g_l,i}$ (original parameter) and the red arrow is $\bar{w}_{g_l,i}^+$ (updated parameter). The bottom-left plot indicates the activation values on the black dashed line before and after the update. The bottom-right plot shows the conceptual representation of the activation pattern change after the update ($z_0 \rightarrow \bar{h}_{l,i}$).

6 Conclusion

In this paper, we propose the concept *local activation* on the internal neurons of GANs to evaluate the low visual fidelity of generations. We further discuss an analysis on the relationship between the proposed concept and low visual fidelity of individual generations under the restricted condition. We perform empirical studies to validate the relations on artifact detection and correction settings. The proposed method shows reasonable performance without additional supervision or resources. Because the proposed method uses the basic element (neuron) of neural networks and its internal information, we believe that the proposed approach can be extended to a wide range of deep neural networks.

Figure 8: Artifact detection results in various GANs. We select bottom 24 (good) and top 24 (bad) samples for qualitative comparison. We confirm that the generations with high CLA have lower visual fidelity compared to the the generations with low CLA. Appendix C-I provide more examples for artifact detection.

Acknowledgements

This work was conducted by Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD), Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2017-0-01779, XAI and No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)) and partly supported by KAIST-NAVER Hypercreative AI Center.

References

Bau, D.; Zhu, J.-Y.; Strobelt, H.; Zhou, B.; Tenenbaum, J. B.; Freeman, W. T.; and Torralba, A. 2019. Visualizing and Understanding Generative Adversarial Networks. In *Proceedings of the International Conference on Learning Representations*.

Borji, A. 2019. Pros and Cons of GAN Evaluation Measures. *Computer Vision and Image Understanding*, 179: 41–65.

Brock, A.; Donahue, J.; and Simonyan, K. 2019. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *Proceedings of the International Conference on Learning Representations*.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language Models are Few-shot Learners. In *Proceedings of the Advances in Neural Information Processing Systems*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative Adversarial Nets. In *Proceedings of the Advances in Neural Information Processing Systems*.

Härkönen, E.; Hertzmann, A.; Lehtinen, J.; and Paris, S. 2020. GANSpace: Discovering Interpretable GAN Controls. In *Proceedings of the Advances in Neural Information Processing Systems*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Proceedings of the Advances in Neural Information Processing Systems*.

Jeon, G.; Jeong, H.; and Choi, J. 2020. An Efficient Explorative Sampling Considering the Generative Boundaries of Deep Generative Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proceedings of the International Conference on Learning Representations*.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of Stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Kenton, J. D. M.-W. C.; and Toutanova, L. K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *Proceedings of the International Conference on Learning Representations*.

Kynkäänniemi, T.; Karras, T.; Laine, S.; Lehtinen, J.; and Aila, T. 2019. Improved Precision and Recall Metric for Assessing Generative Models. In *Proceedings of the Advances in Neural Information Processing Systems*.

Miyato, T.; Kataoka, T.; Koyama, M.; and Yoshida, Y. 2018. Spectral Normalization for Generative Adversarial Networks. In *Proceedings of the International Conference on Learning Representations*.

Peebles, W.; Peebles, J.; Zhu, J.-Y.; Efros, A.; and Torralba, A. 2020. The Hessian Penalty: A Weak Prior for Unsupervised Disentanglement. In *Proceedings of the European Conference on Computer Vision*.

Peters, M.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Sajjadi, M. S. M.; Bachem, O.; Lucic, M.; Bousquet, O.; and Gelly, S. 2018. Assessing Generative Models via Precision and Recall. In *Proceedings of the Advances in Neural Information Processing Systems*.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. In *Proceedings of the Conference on Neural Information Processing Systems*.

Shen, Y.; Gu, J.; Tang, X.; and Zhou, B. 2020. Interpreting the Latent Space of GANs for Semantic Face Editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Tousi, A.; Jeong, H.; Han, J.; Choi, H.; and Choi, J. 2021. Automatic Correction of Internal Units in Generative Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yu, F.; Zhang, Y.; Song, S.; Seff, A.; and Xiao, J. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *CoRR*.