SPARSITY IMPROVES UNSUPERVISED ATTRIBUTE DISCOVERY IN STYLEGAN

Shusen Liu, Rushil Anirudh, Jayaraman J. Thiagarajan, Peer-Timo Bremer

Center for Applied Scientific Computing (CASC),
Lawrence Livermore National Laboratory

ABSTRACT

Rich semantics exist in latent spaces inferred using deep generative models. The ability to extract and interpret them is not only essential for understanding the underlying factors of variation in the data distribution, but also crucial for controlled image generation. Several methods have been proposed to identify semantically meaningful linear directions, either through existing annotations, or relying on identifying directions of large variation that arise from the data representation of the network. In this paper, we identify a new criterion, representation sparsity, that allows us to produce extremely efficient yet diverse semantic directions in GAN (generative adversarial network) latent spaces. The observation also reveals a potential deeper connection between representation sparsity and semantics in deep neural networks that worth further exploration.

***Index Terms*—** StyleGAN, semantic discovery, XAI, sparsity, deep learning

1. INTRODUCTION

Recent developments [1, 2, 3] in generative adversarial networks [4] (GANs), e.g., StyleGAN, greatly improve the quality of synthesized images, opening up a wide range of potential applications. Once trained, these generative models sample from a prior distribution in the latent space, to generate high-quality images. To have better control over the quality and diversity of generated images, discovering and interpreting semantics in the latent space of GANs has gained a lot of attention recently [5, 6]. Quantifying the kinds of changes induced by moving in the latent space serves both as a tool for understanding the mechanism of the generative model, as well as an avenue for the inverse problem of generating desirable images with given specifications. Often the semantics are described as a linear direction in the latent representation (along a subspace), and walking along this direction is expected to increase (or decrease) a particular attribute (e.g., hair color, age, eye, or expression in a faces dataset) in the generated output.

The notion of encoding a concept via a vector direction in a latent space can be traced back to the widely discussed analogy pair example from the word2vec [7]. Subsequently, Kim *et al.* [8] formalized the notion of concept vectors, in which supervised information is leveraged to learn a vector representation of the semantic of interest. A linear direction, though limited in its expressiveness, facilitates intuitive human interpretation and manipulation. Therefore, the semantic discovery in GAN latent spaces often revolve around how to obtain a set of meaningful vector directions in the latent space, such that they correspond to concepts that are readily recognizable and preferably disentangled. The most straightforward approach for this is through external labeling or annotation, where a linear direction can be easily learned [9]. However, the limitation is apparent, as such a scheme not only requires a large amount of labeling information but can also only emulate already known concepts. Alternatively, Voynov et al. propose an approach [10] for identifying directions in the latent space based on how predictive the changes they induced are. Other works on semantics and GANs have explored ideas such as steerability [11], memorability and aesthetics [12], structure of semantics [13], conditional generation [14] and explanation [15].

The problem of interest in this paper is *unsupervised semantic discovery* – i.e., automatically identifying meaningful, disentangled factors of variation in the StyleGAN latent space. Among existing works, one crucial intuition has proved useful for finding these directions – as the generator transforms the input noise distribution to the final image, an isotropic prior distribution (e.g., Gaussian or uniform) is stretched and squeezed to form a more complicated representation, during which large variations in intermediate latent representations can signal important semantic directions. Methods such as GANSpace [5] and SeFa [6] exploit this intuition, where the former achieves it by computing PCA directions from latent representations of data samples, while the latter relies on direct matrix factorization of network weights.

However, a fundamental challenge with these approaches is that directions of large variances in the latent space do not necessarily correspond well to *distinctive* concepts. If we are solely aiming for capturing the largest variance, we may inadvertently combine several concept directions that could, and should be separated. Furthermore, these directions of large variance are also known to be sensitive to outliers or nuisance attributes in the dataset, leading to sub-optimal results. In Fig.1, we illustrate an extreme case of this idea, where the di-

rection of maximum variance is misaligned with the regions of maximum density resulting in distinctive concepts (manifesting as unique directions) being represented by an average vector, as long as they contribute to the overall variance.

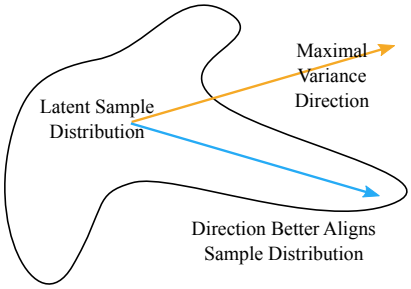

Fig. 1. Large variance directions do not necessarily align with the geometry of the underlying sampling distribution, which is where the semantics arise. In our work, instead of focusing solely on finding large variances, we propose to utilize sparsity as a key criterion to uncover directions that align with the sample distribution, thereby better capturing the geometry in the latent space.

Contributions In this paper, we propose a novel approach for addressing this challenge via an alternative intuition that the concept vector directions should follow the modes of the sample distribution closely, which can be achieved by imposing *sparsity* in the basis directions, for example, using sparse PCA [16]. By design, a non-linear generator network leverages sparsity of the representation (in a basis spanned by the neural network), since it transforms a dense prior distribution (i.e., a full dimensional Gaussian or uniform) into a very high-dimensional yet sparse output manifold (i.e., natural images) [17]. Such an observation indicates that semantics emerges as the network transforms the representation while making them increasingly sparser. We hypothesize that a sparse concept direction will more likely stay close to samples in the distribution, and provide a disentangled and concise direction compared to methods only focusing on variance. A direct consequence of this is that moving along these directions is expected to produce larger changes in the output space, for the same amount of distance moved – in other words, these directions end up being more *efficient* to obtain desired attribute changes.

2. PROPOSED APPROACH

Our key intuition in this paper is to exploit sparsity for meaningful semantic discovery. Instead of solely relying on factors of variation in the intermediate representations, we hypothesize that a combination of both variance and sparsity objectives will yield better directions. We expect these attribute directions to be better aligned with the sample distribution and therefore the semantics. While PCA finds a basis to maximize explained variance, we additionally enforce sparsity of the basis as a constraint. This has most commonly been studied in the context of sparse PCA and there have been several techniques to achieve this over the years [16, 18, 19], with varying degrees of success, which typically trade-off sparsity with the total explained variance. However, as mentioned earlier, an ideal basis for attribute discovery is expected to be sparse while also explaining most of the variance (i.e., closer to PCA). Hence, we adopt the recently proposed SCA [20], which was found to improve on the total explained variance, while still maintaining desired sparsity with the use of a new (rotated) basis. We outline our method to use SCA to find meaningful directions in StyleGAN's latent space next and demonstrate its utility over existing methods in unsupervised attribute discovery.

Sparse component analysis (SCA) Let the samples in an intermediate layer of StyleGAN be represented as $X \in \mathbb{R}^{n \times p}$, where n is the number of samples, and p is the dimensionality of features in the layer. A PCA basis optimizes the following cost: $\max_Y \|XY\|_F \quad \text{subject to} \quad Y \in \nu(p,k)$, where ν is the Stiefel manifold, i.e., $\nu(p,k) = \{Y \in \mathbb{R}^{p \times k} | Y^T Y = I_k\}$ and k is the number of principle components. However, both the coefficients of Y (often referred to as *loading*), and $S = XY, S \in \mathbb{R}^{n \times k}$ (i.e., the *score*), are likely not sparse. This also means most samples are not close to the principle components (as the coordinates will be dense).

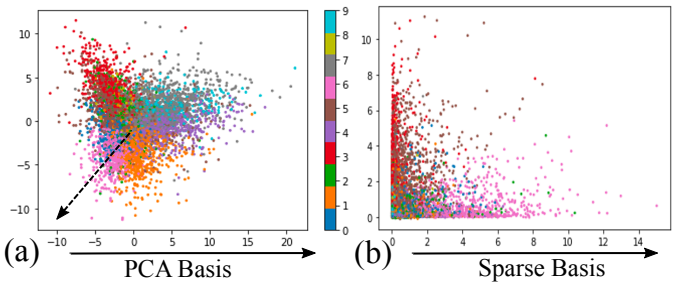

Fig. 2. Visualizing (intermediate) latent representations from a GAN trained on MNIST. The colors indicate different digits. The PCA basis (a) reveals interesting overall structure but the axis/basis do not directly align with semantic concepts of interest. By utilizing a sparser basis (b), we are able to better align with semantic concepts and the sample distribution.

To better illustrate this point, we compute PCA on the latent representations of all data samples in the layer after the input Z space of a standard GAN trained using MNIST digits. Subsequently, a classifier is applied to the corresponding final output image of the sample in order to assign a ground truth label to them. As shown in Fig.2 (a), where the samples are colored by the digit label, the first two PC directions capture the meaningful structure of the latent space. Yet, the actual PC does not go along any of the digits of interest. A more desirable basis that better describes the semantics should for example go along the points that belong to a single digit, e.g., the dotted arrow in the figure.

To address this potential mis-alignment between PC and

interpretable semantic concepts, we propose to produce a basis that requires sparse *loading* and *score* matrices. Here, we leverage some recent work [20] in the sparse PCA research that was able to effectively enforce sparsity, while better explaining variance compared to state-of-the-art sparse PCA methods [16], by utilizing a basis rotation step before imposing any sparsity constraint. To perform such an optimization, we can rewrite the PCA formulation as follows:

$$\begin{aligned} \min_{Z,B,Y} \quad & \left\|X - ZBY^T\right\|_F \\ \text{Subject to} \quad & Z \in \nu(n,k), Y \in \nu(p,k) \end{aligned} \tag{1}$$

This formulation will be identical to PCA, provided B is diagonal. However, by allowing flexibility in B, we can facilitate orthogonal rotation to be expressed in the optimization, i.e., $X \approx UDV^T = (UO)(O^TDR)(VR)^T = ZBY^T$, where O and R are the rotations. By adding a sparsity constraint on Y, we can then try to find the right rotation to make the column space of V sparse.

By solving this optimization, we can obtain a more sparse basis that is better aligned with the existing sample distribution. As shown in Fig. 2 (b), the sparse basis can perfectly express distinct digits, e.g., the x-axis (one of the sparse bases) directly captures the variations of digit-6, whereas the PCA bases express variance that arises from multiple different digits. Once we construct these bases in intermediate layers, similar to GANSpace[5], we can then assign scores to each sample and fit a set of vectors that express similar semantics in the W space of the styleGAN for image generation.

Max-pooling for scalable semantic search Interesting semantics and structure emerge in the later layers of the deep network. However, one challenge associated with going deeper into a GAN's latent representation is the explosion of dimension that will make any subsequent analysis intractable. For image-based generative models, one crucial observation enables us to simplify this. For example, in styleGAN the high dimensionality in the latent space comes predominantly from an increased spatial resolution rather than the number of feature detectors used. If we are only interested in semantic information, it does not matter where the signal is presented in the image, so max-pooling can be used to significantly reduce the latent space dimensionality.

3. EXPERIMENTS

We carried out experiments on the state-of-the-art styleGAN [2] trained with the FFHQ dataset [2]. To compute the semantic directions, we first generated 10000 random samples in the Z space then mapped them into later layers; we then computed PCA and SCA in these intermediate latent spaces and then identified the best-matched linear direction in W space via least square fitting [5]. To quantify semantic changes for evaluation, we independently trained an auxiliary multi-classes face attribute classifier using the CelebA dataset. In Fig.3, we show a few examples to validate the semantics captured by our approach. Like existing approaches, we are able to find many common facial attributes, e.g., hair color, glass, smiling, facial hair, etc, moreover, it also captures other more subtle semantics such as changes in lighting conditions (or shadows).

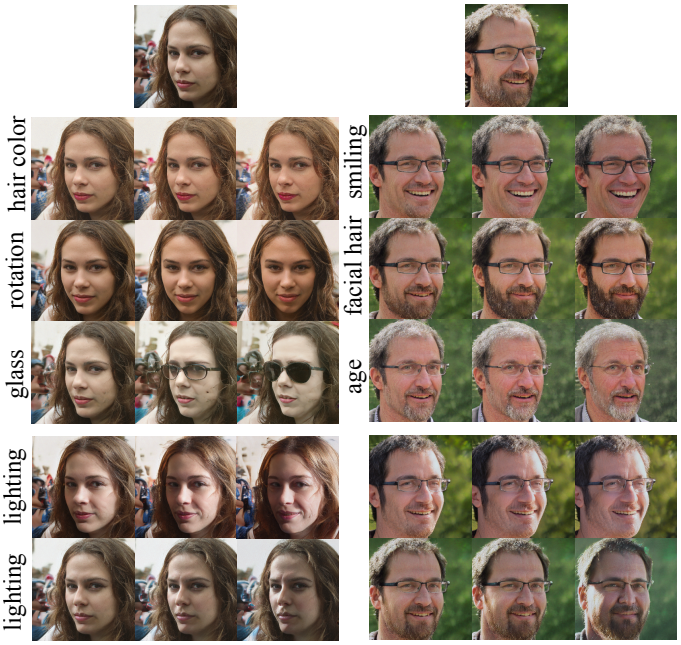

Fig. 3. Examples of walks in latent space along discovered semantic directions. Our approach can identify a diverse set of semantic direction, including subtle ones such as light direction changes (bottom two rows).

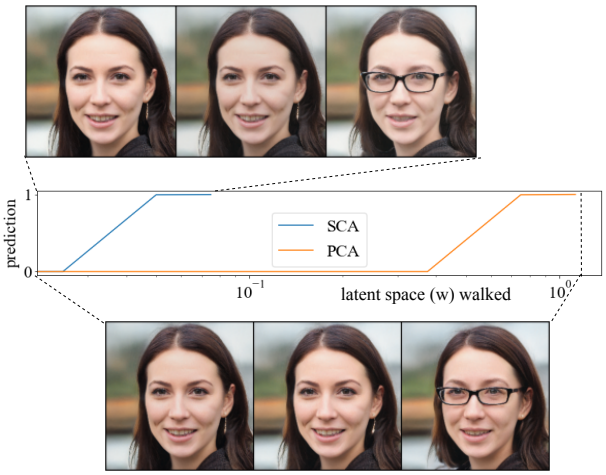

Fig. 4. Our method enables us to find a more effective direction, i.e., walk short distance in the latent space, to achieve the desirable semantic changes.

Walking efficiently in StyleGAN's latent space: One of the key strengths of our method is its ability to produce much more *efficient* direction – i.e., the same amount of distance traveled in a direction identified by SCA (proposed here) and PCA yields vastly different amount of changes in the

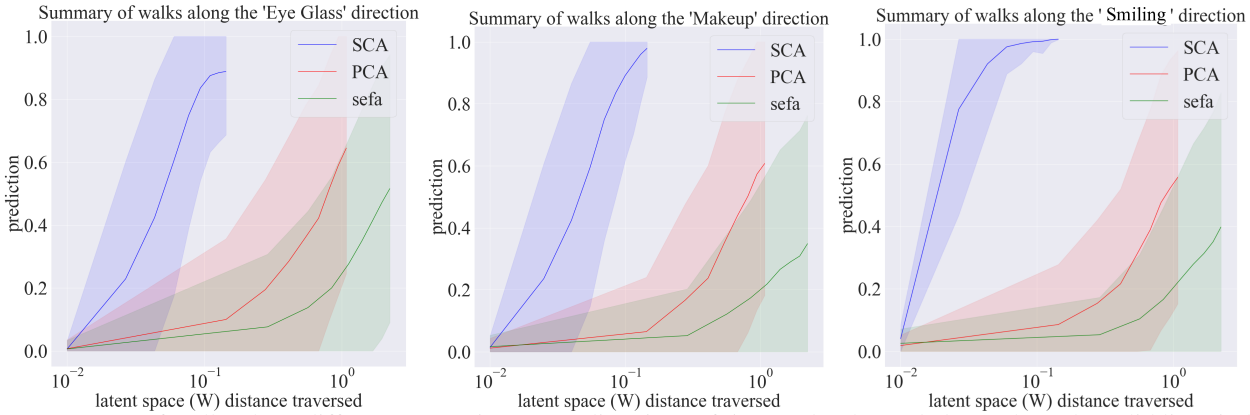

Fig. 5. Summary of walks along different semantics vector directions of 500 randomly seeded samples. The solid lines indicate the mean, the colored region illustrate the standard deviation.

generated image, even though both correspond to the exact same pre-trained model. This is illustrated in Fig.4, where we show how far each method needs to traverse to explore these semantic concepts (prediction) changes. We make a surprising finding that using sparsity as a key criterion, we are able to achieve the same amount of desired semantic shifts by traveling only 1/10$^{\text{th}}$ of the distance that is required by the directions discovered using existing approaches such as GANspace [5].

Moreover, as shown in Fig.5, such an improvement can be readily observed across samples and for different semantics. Here, we show a summary gathered from 500 randomly seeded samples (filtered by negative predictions of the output images, so that we can see the semantic shift). The solid lines indicate the mean prediction as we walk along the semantic directions, the color regions illustrate the corresponding standard deviation. The proposed method requires significantly shorter traversal distances, in the same latent space to reach desirable semantic changes.

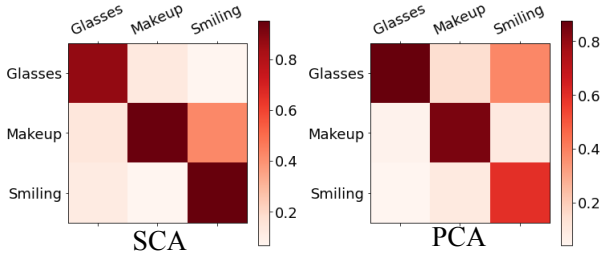

Fig. 6. This illustration shows how disentangled the semantic directions identified using SCA and PCA are.

In addition to revealing interpretable directions, the other desirable quality is disentanglement, i.e., walking along a semantic direction should not affect other unrelated attributes. As illustrated in Fig.6, we show how much unrelated semantic changes happen when we walk along a specific direction. In the visualization, the rows correspond to predictions for different labels along the same direction. Ideally, one would expect to see all changes concentrated on the diagonal. The proposed attribute discovery method (SCA) produces a relatively similar disentanglement result compared to PCA. In other words, our approach can easily produce much more efficient directions without sacrificing the disentanglement property. In fact, from the quantitative evaluation presented next, our method even outperforms the PCA in terms of disentanglement metric.

To formally quantify and evaluate the above observed properties, we proposed two metrics that directly measure 1) how efficient the direction is, i.e., the ratio of semantic changes and the distance walked, $S_e = \Delta p / \Delta d$, where Δp is the change in prediction, and Δd is the latent space distance traversed; 2) how disentangled the directions are, i.e., the ratio between the intended semantic changes and sum of all semantic changes, $S_d = \text{Tr}(A) / \sum_{i,j} A_{ij}$, where i^{th} row of $\mathbf{A}$ corresponds to prediction of all concepts along the i^{th} semantics. As shown in Table 1, the proposed method clearly out-performs pure-variance based methods [5, 20].

Table 1. Comparing efficiency and disentanglement metrics.

	PCA [5]	sefa [20]	Ours
Efficiency score	0.492	0.162	**6.234**
Disentanglement score	0.740	0.684	**0.745**

4. CONCLUSION

We proposed an new unsupervised technique to discover meaningful semantic directions in styleGAN latent space. In particular, we showed that representation sparsity is a useful prior to identify directions that are more sensitive/efficient compared to existing methods on the same pre-trained styleGAN. These findings are potentially useful to a wide range of applications relying on image manipulation with StyleGAN, considering that our approach achieves manipulation *faster*, thereby needing fewer steps in the latent space to achieve the same desired semantic change. This also opens several avenues of potential future work – for e.g., to study the role sparsity plays in concept formulation and investigate how to go beyond sparse PCA for solving the concept discovery problem; and examining the potential connection between sparsity and semantics in supervised models.

Acknowledgement

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Released under LLNL-PROC-827511.

5. REFERENCES

[1] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[2] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[3] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of styleGAN," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[4] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[5] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris, "GANspace: Discovering interpretable GAN controls," *arXiv preprint arXiv:2004.02546*, 2020.

[6] Yujun Shen and Bolei Zhou, "Closed-form factorization of latent semantics in GANs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1532–1540.

[7] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[8] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al., "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *International conference on machine learning*. PMLR, 2018, pp. 2668–2677.

[9] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou, "InterfaceGAN: Interpreting the disentangled face representation learned by GANs," *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[10] Andrey Voynov and Artem Babenko, "Unsupervised discovery of interpretable directions in the GAN latent space," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9786–9796.

[11] Ali Jahanian, Lucy Chai, and Phillip Isola, "On the" steerability" of generative adversarial networks," *arXiv preprint arXiv:1907.07171*, 2019.

[12] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola, "GANalyze: Toward visual definitions of cognitive image properties," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5744–5753.

[13] Ceyuan Yang, Yujun Shen, and Bolei Zhou, "Semantic hierarchy emerges in deep generative representations for scene synthesis," *International Journal of Computer Vision*, vol. 129, no. 5, pp. 1451–1466, 2021.

[14] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE transactions on image processing*, vol. 28, no. 11, pp. 5464–5478, 2019.

[15] Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han, "Generative counterfactual introspection for explainable deep learning," in *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2019, pp. 1–5.

[16] Hui Zou, Trevor Hastie, and Robert Tibshirani, "Sparse principal component analysis," *Journal of computational and graphical statistics*, vol. 15, no. 2, pp. 265–286, 2006.

[17] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros, "Generative visual manipulation on the natural image manifold," in *European conference on computer vision*. Springer, 2016, pp. 597–613.

[18] Alexandre d'Aspremont, Laurent El Ghaoui, Michael I Jordan, and Gert RG Lanckriet, "A direct formulation for sparse pca using semidefinite programming," *SIAM review*, vol. 49, no. 3, pp. 434–448, 2007.

[19] Daniela M Witten, Robert Tibshirani, and Trevor Hastie, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515–534, 2009.

[20] Fan Chen and Karl Rohe, "A new basis for sparse PCA," *arXiv preprint arXiv:2007.00596*, 2020.