Efficient Precision and Recall for Assessing Generative Models using Hubness-aware Sampling

Yuanbang Liang [1] Jing Wu [1] Yu-Kun Lai [1] Yipeng Qin [1]

[1]School of Computer Science and Informatics, Cardiff University

Motivation & Contribution

- We propose *efficient precision and recall* (eP&R) metrics for assessing generative models, which give almost identical results as the original P&R [1] but consume much less time and space. Theoretically, our eP&R run in $O(mn \log n)$ time and consume $O(mn)$ space (m is the number of of hubs samples and $m < n$), which are much more efficient than the original P&R metrics that run in $O(n^2 \log n)$ time and consumes $O(n^2)$ space.
- We identify two important types of redundancies in the original P&R metrics and uncover that both of them can be effectively removed by hubness-aware sampling [2, 3]. In addition, the insensitivity of hubness-aware sampling to exact k-nearest neighbor (k-NN) results allows for further efficiency improvement by using approximate k-NN methods.
- Extensive experimental results demonstrate the effectiveness of our eP&R metrics.

Preliminaries

The precision and recall (P&R) metrics for assessing generative models [1] are defined as:

$$\text{precision}(\mathbf{\Phi}_r, \mathbf{\Phi}_g) = \frac{1}{|\mathbf{\Phi}_g|} \sum_{\phi_g \in \mathbf{\Phi}_g} f(\phi_g, \mathbf{\Phi}_r), \tag{1}$$

$$\text{recall}(\mathbf{\Phi}_r, \mathbf{\Phi}_g) = \frac{1}{|\mathbf{\Phi}_r|} \sum_{\phi_r \in \mathbf{\Phi}_r} f(\phi_r, \mathbf{\Phi}_g) \tag{2}$$

where $\mathbf{\Phi_g}$ and $\mathbf{\Phi_r}$ are the sets of feature vectors corresponding to the generated and real image samples, respectively; $|\mathbf{\Phi}|$ denotes the number of samples in set $\mathbf{\Phi}$ and $|\mathbf{\Phi}_g| = |\mathbf{\Phi}_r|$; $f(\phi, \mathbf{\Phi})$ is a binary function determining whether a sample ϕ lies on a manifold represented by $\mathbf{\Phi}$:

$$f(\phi, \mathbf{\Phi}) = \begin{cases} 1, & \text{if } \|\phi - \phi'\|_2 \le \|\phi' - \text{NN}_k(\phi', \mathbf{\Phi})\|_2 \text{ for } \textbf{at least one } \phi' \in \mathbf{\Phi} \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

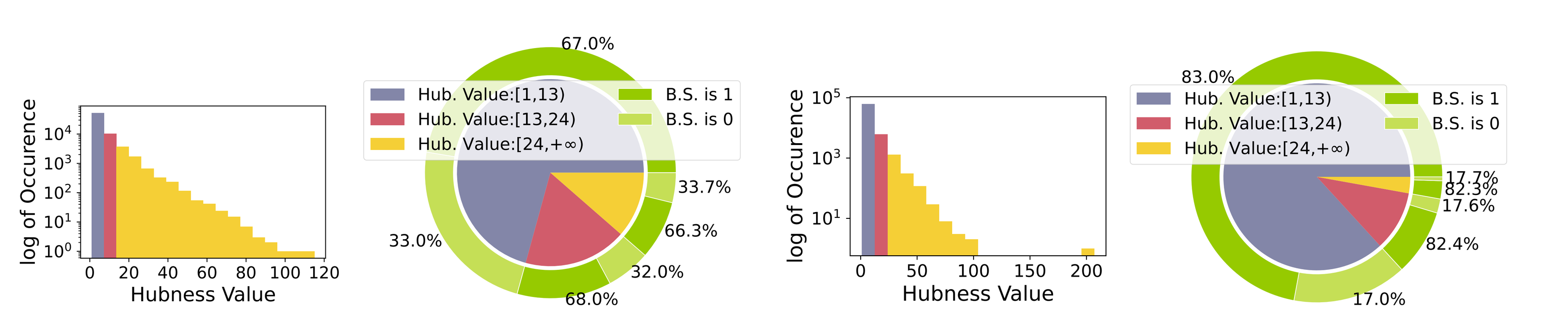
where $\text{NN}_k(\phi', \mathbf{\Phi})$ denotes the kth nearest neighbour of ϕ' in $\mathbf{\Phi}$.

The Redundancies in Precision and Recall

Observation 1 [Redundancy in Ratio Estimation] As Eq. 2 shows, the P&R metrics are essentially ratios of the number of samples in a set $\mathbf{\Phi}$ that lie on a given manifold to the number of all samples in $\mathbf{\Phi}$. Thus, we can obtain similar P&R ratios by using *representative samples* of $\mathbf{\Phi}$ with the rest as redundant.

Observation 2 [Redundancy in Inside/Outside Manifold Identification] As shown in Eq. 3, $f(\phi, \mathbf{\Phi})$ is 1 as long as ϕ is within the k-NN hypersphere of *at least one* sample $\phi' \in \mathbf{\Phi}$. This means that we only need to find one valid ϕ' for each ϕ and all the other ϕ's are redundant.

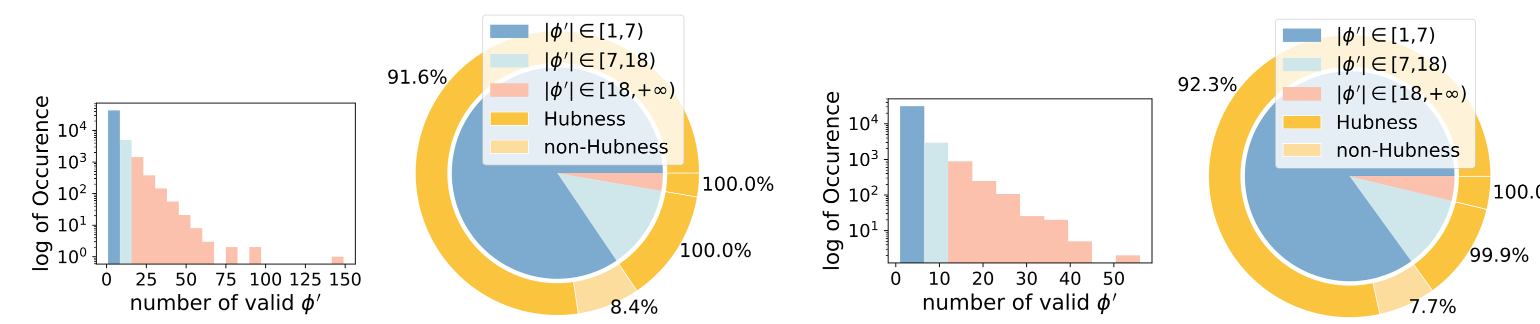
Redundancy Reduction using Hubness-aware Sampling



(a) All 70k images in the FFHQ dataset (b) 70k images generated by StyleGAN2

Figure 1. Samples with similar hubness values are effective representative samples in terms of P&R ratio calculation. (a) Left: Histogram of sample occurrences *vs.* hubness value. Right: Pie chart showing that all three groups share similar ratios of samples identified as 1 vs. 0 using Eq. 3 for recall calculation. (b) The same experiment as (a) but on StyleGAN-generated samples for precision calculation.

Redundancy Reduction using Hubness-aware Sampling (Cont'd)



(a) All 70k images in the FFHQ dataset (b) 70k images generated by StyleGAN2

Figure 2. Most samples ϕ with $f(\phi, \mathbf{\Phi}) = 1$ (Eq. 3) are included in the k-NN hypersphere of at least one hubs sample ($t = 3$) of the other distribution. (a) Left: Histogram of sample occurrences (log scale) *vs.* the times a sample is included in the k-NN hypersphere of a sample of the other distribution, *i.e.*, valid ϕ'; the illustration can be checked in Fig. 3. Right: Pie chart showing the ratio of samples within the k-NN hypersphere of *hubness* vs. *non-hubness* samples from the other distribution, to the total number of samples ϕ with $f(\phi, \mathbf{\Phi}) = 1$ in each group.

Rationale

- For Observation 1 and Fig. 1, we find that samples with similar hubness values are effective representative samples of set $\mathbf{\Phi}$ in terms of P&R ratios as they share similar ratios of samples identified as 1 *vs.* 0 by Eq. 3, indicating that we can use a small number of hubs samples to approximate P&R;
- for Observation 2 and Fig. 2, we find that most ϕ with $f(\phi, \mathbf{\Phi}) = 1$ (Eq. 3) are included in the k-NN hypersphere of at least one ϕ' with high hubness values, *i.e.*, hubs samples, indicating that we can obtain similar outputs of Eq. 3 using a small number of hubs samples.

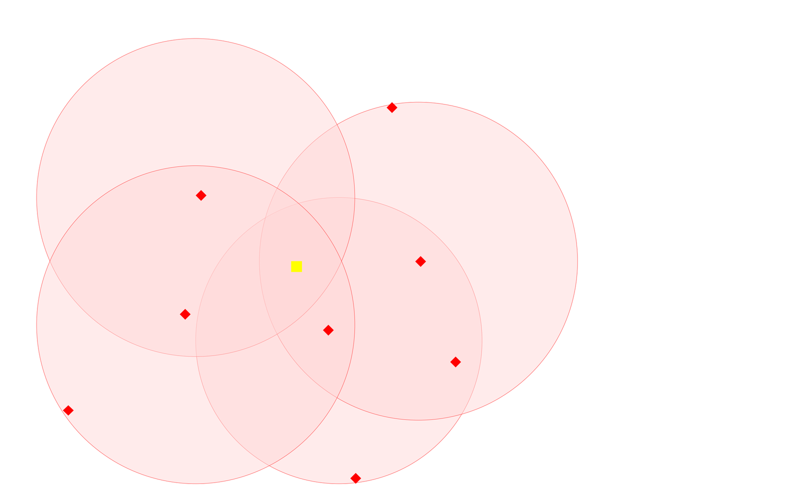
Thus, our **efficient P&R metrics (eP&R)** can be defined as:

$$\text{precision}^{hub}(\mathbf{\Phi}_r, \mathbf{\Phi}_g) = \frac{1}{|\mathbf{\Phi}_g^{hub}|} \sum_{\phi_g^{hub} \in \mathbf{\Phi}_g^{hub}} f(\phi_g^{hub}, \mathbf{\Phi}_r^{hub}) \tag{4}$$

$$\text{recall}^{hub}(\mathbf{\Phi}_r, \mathbf{\Phi}_g) = \frac{1}{|\mathbf{\Phi}_r^{hub}|} \sum_{\phi_r^{hub} \in \mathbf{\Phi}_r^{hub}} f(\phi_r^{hub}, \mathbf{\Phi}_g^{hub}) \tag{5}$$

where $\mathbf{\Phi}_g^{hub}$ and $\mathbf{\Phi}_r^{hub}$ are the sets of feature vectors with hubness values $m > t$ corresponding to the generated and real image samples, respectively; t is a threshold hyper-parameter.

Illustration for valid ϕ'



As Fig. 3 shows, by "the times a sample is included in the k-NN hypersphere of a sample of the other distribution, *i.e.*, valid ϕ'", we count the number of times ϕ (yellow cube) is within the k-NN hypersphere of $\phi' \in \Phi$ (red rhombuses).

Figure 3. Illustration of valid ϕ'. ϕ is represented by a yellow cube and $\phi' \in \Phi$ set are represented by red rhombuses.

Error Analysis (Partial Results)

Table 1. Approximation errors compared to the original Precision and Recall (P&R) metrics.

	FFHQ		LSUN-Car		LSUN-Church	
	Precision	Recall	Precision	Recall	Precision	Recall
eP&R	0.719±0.002	0.501±0.002	0.732±0.001	0.422±0.002	0.608±0.002	0.392±0.003
B.L.	0.716±0.001	0.493±0.001	0.725±0.001	0.426±0.001	0.592±0.001	0.389±0.002
Error(%)	0.4%	1.6%	0.9%	0.9%	1.9%	0.7%

Computational Complexity Analysis (Partial Results)

B.L.: the original P&R metrics as the baseline; eP&R: our efficient P&R metrics; DM: Distance Matrix; A. hubs: the approximate hubness value; $m_x = \max\{m_r, m_g\}$ and $|\mathbf{\Phi}_g| = m_g, |\mathbf{\Phi}_r| = m_r$.

Profiling	B.L. Time	B.L. Memory	Profiling	eP&R Time	eP&R Memory
DMs ($\mathbf{\Phi}_r, \mathbf{\Phi}_g$)	$O(n^2)$	$O(n^2)$	Subspace ($\mathbf{\Phi}_r, \mathbf{\Phi}_g$)	$O(\log n)$	$O(n)$
			A. hubs ($\mathbf{\Phi}_r^{hub}, \mathbf{\Phi}_g^{hub}$)	$O(m_x)$	–
			eDMs	$O(m_x n)$	$O(m_x n)$
Sorting	$O(n^2 \log n)$	–	eSorting	$O(m_x n \log n)$	–
Radii	$O(n)$	$O(n)$	Radii	$O(m_x)$	$O(m_x)$
DM ($\mathbf{\Phi}_r \leftrightarrow \mathbf{\Phi}_g$)	$O(n^2)$	$O(n^2)$	eDM ($\mathbf{\Phi}_r^{hub} \leftrightarrow \mathbf{\Phi}_g^{hub}$)	$O(m_r m_g)$	$O(m_r m_g)$
P&R	$O(n^2)$	–	eP&R	$O(m_x^2)$	$O(m_x^2)$
Total/Peak	$O(n^2 \log n)$	$O(n^2)$	Total/Peak	$O(m_x n \log n)$	$O(m_x n)$

Theoretically, the proposed eP&R metrics run in $\max(O(m_r n \log n), O(m_g n \log n))$ time and consumes $\max(O(m_r n), O(m_g n))$ space while the original P&R metrics run in $O(n^2 \log n)$ time and consumes $O(n^2)$ space. Since $m_r < n$, $m_g < n$, the proposed eP&R metrics are far more efficient than the original P&R metrics.

Table 2. Time and space consumption of our eP&R metrics V.S the original P&R metrics [1] on the FFHQ. Time (S): serial implementation. Time (P): parallel implementation using CUDA.

Profiling	B.L. Time (S)	B.L. Time (P)	B.L. Memory	Profiling	eP&R Time (S)	eP&R Time (P)	eP&R Memory
DMs ($\mathbf{\Phi}_r, \mathbf{\Phi}_g$)	160s	66s	15.84 GB	Subspace ($\mathbf{\Phi}_r, \mathbf{\Phi}_g$)	4s	3s	3.01 GB
				A. hubs ($\mathbf{\Phi}_r^{hub}, \mathbf{\Phi}_g^{hub}$)	2s	1.2s	–
				eDMs	72s	32s	11.23 GB
Sorting	104s	22s	–	eSorting	50s	12s	–
Radii	2.2s	2.2s	0.58 GB	Radii	1.7s	1.7s	0.30 GB
DM ($\mathbf{\Phi}_r \leftrightarrow \mathbf{\Phi}_g$)	85s	34s	19.24 GB	eDM ($\mathbf{\Phi}_r^{hub} \leftrightarrow \mathbf{\Phi}_g^{hub}$)	18s	9s	8.74 GB
P&R	48s	28s	–	eP&R	11s	6s	–
Total/Peak	399s	144s	19.90 GB	Total/Peak	165s	**75s**	**14.24 GB**

References

[1] T. Kynkäänniemi, T. Karras, S. Laine, J. Lehtinen, and T. Aila, "Improved precision and recall metric for assessing generative models," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[2] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Hubs in space: Popular nearest neighbors in high-dimensional data," *Journal of Machine Learning Research*, vol. 11, no. sept, pp. 2487–2531, 2010.

[3] Y. Liang, J. Wu, Y.-K. Lai, and Y. Qin, "Exploring and exploiting hubness priors for high-quality GAN latent sampling," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 13271–13284, PMLR, 17–23 Jul 2022.

QinY16@cardiff.ac.uk