Privacy Assessment on Reconstructed Images: Are Existing Evaluation Metrics Faithful to Human Perception?

Xiaoxiao Sun[†]
Australian National University

Nidham Gazagnadou[‡]
Sony AI

Vivek Sharma[‡]
Sony AI

Lingjuan Lyu[‡] ✉
Sony AI

Hongdong Li[†]
Australian National University

Liang Zheng[†]
Australian National University

A Data Annotation

In **Section 4.1**, we briefly introduced how humans annotate the reconstructed images for different datasets. In the supplementary material, we have included a graphical user interface (GUI) that was utilized by the annotators. Figure 1 displays the GUI, where **(A)** and **(B)** were specifically designed for annotating different datasets.

To minimize the influence of subjective bias, we use a relatively objective formulation: whether the reconstructed image can be correctly labeled. Specifically, for CIFAR-100, Caltech-101, and Imagenette, we provide up to 20 candidate categories and see if the annotators can correctly recognize the reconstructed image; for more difficult tasks like face recognition and fine-grained classification (Celeb-A and Stanford Dogs), we give both the original and the reconstructed images and ask the annotator if they are of the same identity or species.

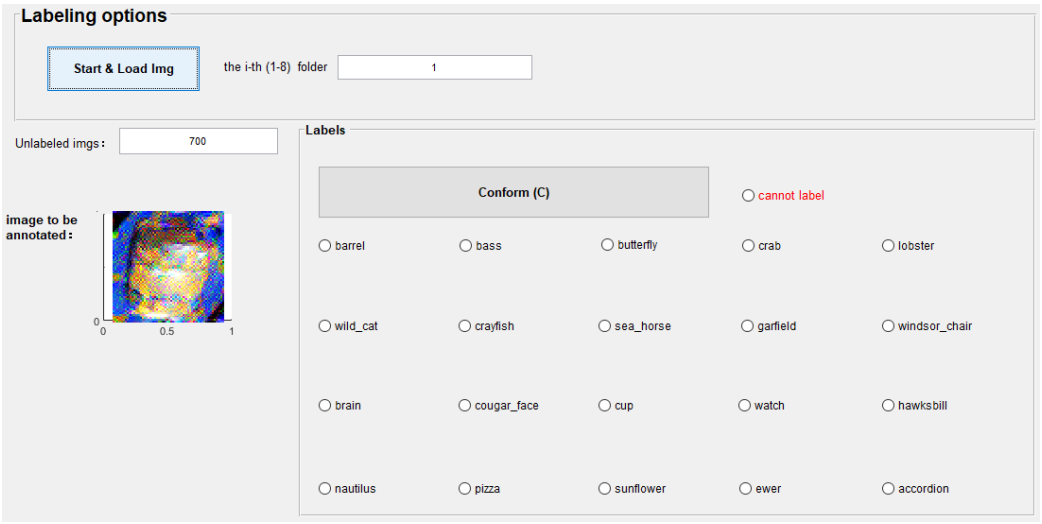
B Impact of margin value α in the triplet loss on SemSim

The effect of the margin parameter α in the triplet loss on the performance of SemSim is depicted in Figure 2. It can be observed that when α is set to a value close to 1, both Spearman's rank correlation (ρ) and Kendall's rank correlation (τ) coefficients yield better results compared to other values, on CIFAR-100 and Caltech-101 datasets. We think there are two potential reasons for this observation. Firstly, if the value of α is too small, the model may struggle to effectively learn the discriminative features that distinguish positive (recognizable reconstructed images) and negative (unrecognizable reconstructed images) samples. On the other hand, if α is set to a value that is too large, the model may become excessively confident in distinguishing between positive and negative samples. However, this can lead to convergence challenges, as the loss function may have difficulty approaching 0. In our experiments, we set α to 1. However, we acknowledge that there is potential for improved performance by carefully selecting the optimal value of α for different datasets.

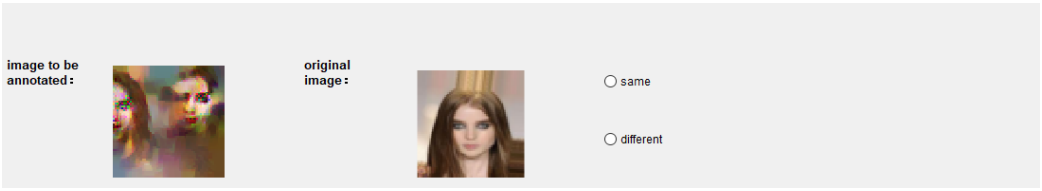
C Classification models and training detials

We conducted experiments using five datasets, CIFAR-100 [9], Caltech101 [3], CelebA [10], ImageNette [1], and Stanford dogs [8]. In our evaluation process, we considered 14 classification models for each set. Table 1 provides detailed information about these models. They were trained using stochastic gradient descent (SGD) as the optimizer, with a learning rate of 0.1.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).



(A) image classification



(B) face recognition and fine-grained image classification

Figure 1: **Graphical user interface (GUI) used in our human annodation process.** For (A) image classification, we ask annotators to give a category to the reconstructed image. In (B) face recognition and fine-grained classification, we ask annotators to tell whether the original image and its reconstruction have the same or different identity / category.

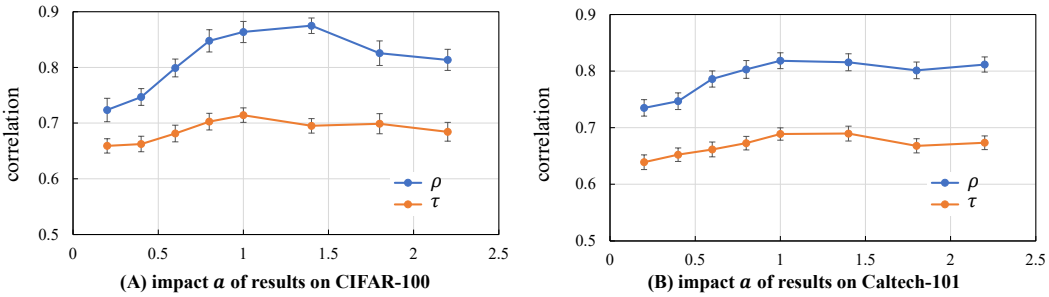


Figure 2: **Impact of α on SemSim when testing on (A) CIFAR-100 and (B) Caltech-101**. The margin value α is used in the triplet loss to ensure that negative samples are kept far apart. When evaluating the reconstructed images of CIFAR-100 or Caltech-101, we trained the ResNet50 model on the four datasets (excluding CIFAR-100 or Caltech-101) using the triplet loss. The training process involved utilizing different values of the margin parameter α for each dataset.

D Metrics for reconstruction quality

Mean squared error. Assuming $x, \bar{x} \in \mathbb{R}^{n \times m}$ are two images to compare, the mean squared error (MSE) is given by,

$$\text{MSE}(x, \bar{x}) := (1/mn) \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_{ij})^2. \tag{1}$$

The value of MSE is between 0 and $+\infty$. The lower MSE is, the closer two images are.

Table 1: **Details of classification models**. On each test set, we have two backbones trained without (vanilla, 2^{nd} column) and with different strategies, such as using data augmentation ($3^{rd} - 6^{th}$ columns) and existing defence methods ($7^{th} - 8^{th}$ columns).

Datasets	Models						
CIFAR-100	ResNet20 CoveNet8	+ Random-ResizedCrop & Random-HorizontalFlip	+ TranslateX & Invert & ranslateY	+ ranslateY & Autocontrast & Autocontrast	+ [4]	+ defense Gaussian (10^{-3})	+ defense Pruning (70%)
Caltech-101	ResNet20 DenseNet						
Imagenette	ResNet50 ResNet152						
CelebA	ResNet20 DenseNet						
Stanford Dogs	ResNet50 ResNet152						

Peak-Signal-to-Noise ratio. The Peak-Signal-to-Noise ratio (PSNR) is widely used in image quality assessment, which measures the ratio between the maximal power of a signal and its noise. Its value, expressed in dB, is given by:

$$\text{PSNR}(x, \bar{x}) = 20 \log_{10} \left(\frac{MAX_x}{\sqrt{\text{MSE}(x, \bar{x})}} \right), \tag{2}$$

where MAX_x is the maximal value in the image x (often replaced by 255 for `int8` images).

SSIM. Unlike PSNR, the structural similarity index measure (SSIM) [11] is a perception-based metric as it was designed to take into account characteristics of the human vision system through three metrics: luminance, contrast and structure of the image. It is shown that there is an analytical link between PSNR and SSIM and that it is often possible to predict one from the other for controlled perturbations (Gaussian blur, additive Gaussian noise and jpeg compressions) [6]. The above three metrics compute a pixel-wise distance between both images which is very limited when we assess semantic content of an image such as privacy leakage.

LPIPS. LPIPS [12], which stands for learned perceptual image patch similarity, is a perceptual metric based on a neural network aiming at correlating better with perceptual judgments. The authors take inspiration from neuroscience findings, where their model compares the activations between two images as neurons in a human cortex would. As explained in its `torchmetrics` documentation[1], a low LPIPS score indicates high similarity. Thus, in the context of privacy assessment, low LPIPS values for an original image and its reconstruction suggest high privacy leakage [7].

Fréchet Inception Distance. Aside from LPIPS and the aforementioned hand-crafted metrics calculated at the image level, our works also uses Fréchet inception distance (FID) [5] to measure information leakage. FID is commonly used to evaluate the domain gap between two distributions, where higher values suggest a larger domain gap. For example, FID is extensively used to evaluate the quality of images generated by generative adversarial networks (GANs) [5], by computing the distribution difference between real and generated images. In this paper, FID may reflect the difference between the original and reconstructed image distributions to reflect privacy leakage. As opposed to the pointwise metrics, FID is computed directly on image sets: $\text{InfoLeak}(\mathcal{X}, \bar{\mathcal{X}}) \propto \text{FID}(\mathcal{X}, \bar{\mathcal{X}})$.

As defined in [5, 2], given two Gaussian distributions with mean and covariance $(\boldsymbol{m}, \boldsymbol{C})$, resp. $(\bar{\boldsymbol{m}}, \bar{\boldsymbol{C}})$, FID is given by:

$$\text{FID}((\boldsymbol{m}, \boldsymbol{C}), (\bar{\boldsymbol{m}}, \bar{\boldsymbol{C}})) = \|\boldsymbol{m} - \bar{\boldsymbol{m}}\|_2^2 + \text{Tr}\left(\boldsymbol{C} + \bar{\boldsymbol{C}} - 2(\boldsymbol{C}\bar{\boldsymbol{C}})^{1/2}\right). \tag{3}$$

Its evaluation on finite sets $\mathcal{X}$ and $\bar{\mathcal{X}}$ follows verbatim by computing their empirical mean and covariance matrix. The value of FID is between 0 and $+\infty$. The lower the FID value is, the closer two distributions are.

Relationship between ℓ_2 and cosine similarity.

[1] `https://torchmetrics.readthedocs.io/en/stable/image/learned_perceptual_image_patch_similarity.html`

The ℓ_2 norm can be used as a tool to measure the distance between vectors, often embeddings of images like those produced by our SimSem model:

$$\ell_2(\mathbf{u}, \mathbf{v}) = \|\mathbf{u} - \mathbf{v}\|_2 , \tag{4}$$

The issue with distances to estimate the similarity between vectors is that they are only bounded below by zero (when $\mathbf{u} = \mathbf{v}$). This makes it hard to set a threshold above which vectors $\mathbf{u}$ and $\mathbf{v}$ can be considered dissimilar. Thus, cosine similarity is often preferred to ℓ_2 distance as its values belong to the interval $[-1, 1]$, 1 indicating proportional vectors and -1 vectors of opposite directions. Let $\mathbf{u}, \mathbf{v}$ be normalized vectors, then the relationship between cosine similarity and ℓ_2 norm is

$$\ell_2(\mathbf{u}, \mathbf{v}) = \sqrt{2(1 - \text{cossim}(\mathbf{u}, \mathbf{v}))} \ . \tag{5}$$

E More Discussions

How much extra effort needs to be paid to extend the current approach? The effort required to extend the current method to other tasks also depends on the nature of the tasks. If we evaluate privacy leakage for the counting task, it would be useful to ask human annotators to count the number of objects in the reconstructed image: if it equals the number of objects in the original image, then privacy may be considered leaked. For this particular counting task, we speculate that manageable efforts will be needed to extend our current approach. On the other hand, tasks that need specialized or professional annotations will likely require more effort, such as medical image understanding.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255, 2009.

[2] DC Dowson and BV666017 Landau. The fréchet distance between multivariate normal distributions. *Journal of multivariate analysis*, 12(3):450–455, 1982.

[3] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Pattern Recognition Workshop*, 2004.

[4] Wei Gao, Shangwei Guo, Tianwei Zhang, Han Qiu, Yonggang Wen, and Yang Liu. Privacy-preserving collaborative learning with automatic transformation search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 114–123, 2021.

[5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, 2017.

[6] Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369, 2010.

[7] Yangsibo Huang, Samyak Gupta, Zhao Song, Kai Li, and Sanjeev Arora. Evaluating gradient inversion attacks and defenses in federated learning. *Advances in Neural Information Processing Systems*, 34: 7232–7241, 2021.

[8] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011.

[9] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[10] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

[11] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[12] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.