Absolute Variation Distance: an Inversion Attack Evaluation Metric for Federated Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

Federated Learning (FL) has emerged as a pivotal approach for training models on decentralized data sources by sharing only model gradients. However, the shared gradients in FL are susceptible to inversion attacks which can expose sensitive information. While several defense and attack strategies have been proposed, their effectiveness is often evaluated using metrics that may not necessarily reflect the success rate of an attack or information retrieval, especially in the context of multidimensional data such as images. Traditional metrics like the Structural Similarity Index (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Mean Squared Error (MSE) are typically used as lightweight metrics, assume only pixel-wise comparison, but fail to consider the semantic context of the recovered data. This paper introduces the Absolute Variation Distance (AVD), a lightweight metric derived from total variation, to assess data recovery and information leakage in FL. Unlike traditional metrics, AVD offers a continuous measure for extracting information in noisy images and aligns closely with human perception. Our results are combined with a user experience survey demonstrate that AVD provides a more accurate and consistent measure of data recovery. It also matches the accuracy of the more costly and complex Neural Network based metric, the Learned Perceptual Image Patch Similarity (LPIPS). Hence it offers an effective tool for automatic evaluation of data security in Federation and a reliable way of studying defence and inversion attacks strategies in FL.

1 Introduction

In the age of Large Models (LM), with size of billions of parameters, data play a crucial role for their continuous development. Therefore, the availability of large amounts of data for their training and fine-tuning is critical. Traditionally, data was concentrated in centralized repositories, but the increased awareness of privacy and the decentralized nature of information generation (mobile phones, multiple regional data centres) has necessitated a more nuanced approach. In this regard, Federated Learning (FL) enables models to learn from a multitude of decentralized edge devices or servers holding local data samples, obviating the need to exchange raw data. The standard FL configuration is achieved with a central aggregator node that exchanges gradients to train a centralised model (McMahan et al., 2017). Particularly, at each training step t, a client node receives neural network model weights, W_t, from an aggregator server and calculates loss l with local data (x_t, y_t) for a batch, B, which generates gradients with respect to the model weights:

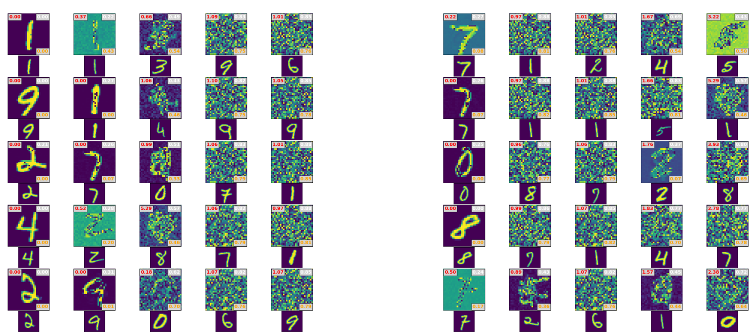
$$\Delta W_t = -\frac{\gamma}{B} \sum_{b<B} \frac{\partial l(F_{W_t}(x_{t,b}, y_{t,b}))}{\partial W_t}, \tag{1}$$

where F_{W_t} is a neural network parameterized by W_t. The gradients are typically averaged in the server with a rate, γ. Because of their flexibility and the client anonymity that they offer, FL models have been deployed in a variety of real-world applications (Yang et al., 2019; Rieke et al., 2020; Nguyen et al., 2022)

However, the gradients, ΔW_t, shared by the client are vulnerable to inversion attacks instigated by a malicious eavesdropper that can expose the original sensitive data. Existing literature on inversion attacks (Zhu et al., 2019b; Zhao et al., 2020a; Geiping et al., 2020; Yin et al., 2021; Balunović et al., 2022) have shown that these inversion attacks can be highly successful, with potentially recovering large batch of data after several rounds of learning, and at a pixel resolution Geiping et al. (2020). In general, these attacks more or less follow the same paradigm, generate a dummy dataset (usually images) and then use a loss function with priors distributions (Balunović et al., 2022) (as regulators or with additional generative models) to minimise the loss between the FL model and the dummy gradients. The success of the inversion attacks prevents FL from becoming a fully trustful framework for distributed training.

To mitigate inversion attacks in FL, several defence strategies were proposed to reduce the leakage of information (Sikandar et al., 2023; Huang et al., 2021a; Chen et al., 2022; Wainakh et al., 2022). These include, data transformation from the client side (Huang et al., 2021b), homomorphic encryption techniques (Phong et al., 2018), data sanitation methods (Zhu et al., 2019b), and defense strategies originated from Differential Privacy approaches (DP) (Dwork, 2006). In FL, DP techniques can be achieved by adding noise to the gradients shared or input data, with inevitably, a potential loss in model training performance (Zhu et al., 2019a; Zhao et al., 2020b; Eloul et al., 2022).

One important step towards the development of robust defence strategies and the prevention and understanding of such attacks is by assessing their success. Therefore, an inversion attack is studied experimentally by comparing the information revealed from the recovered input data to that of the original dataset (Huang et al., 2021a). There exist a few metrics that are commonly used to measure the reconstruction quality, with the most popular being the structural similarity index (SSIM) (Wang et al., 2004), the peak signal-to-noise ratio (PSNR) (Cahn, 1961), the learned perceptual image patch similarity (LPIPS) (Zhang et al., 2018), and the mean squared error (MSE).



(a) (b)

Figure 1: Random recovered vectors from the MNIST dataset. Sub-figure (a) shows the images sorted by the AVD metric. Sub-figure (b) shows the images sorted by the MSE metric. The smaller figures represent the original images associated with image above. The figures include complex images, such as Sub-figure (a), row 1 & column 3, that is a combination of digits 4 and 3. Sub-figure (b), row 3 & column 4, that is a combination of digits 2 and 8.

In this paper, we show how the aforementioned metrics are insufficient to properly assess the success of an inversion attack on the gradients of an FL model. Especially, for cases that the FL model generates multidimensional outputs that contain contextual information (like images). The reason is that these metrics assume spatial independence when comparing the recovered image to the original one. Therefore, in many cases they lack to produce accurate results of the semantic context, and fail to reveal minimal information out of noisy image. For example, a common defense mechanism for an FL model is to add noise to an image that depicts a number from the MNIST dataset. An attack on the gradients of this model may recover an approximate image with the same number but considerably different background colour (see Fig. 1a, row 4 & column 2). Metrics, such as SSIM, PSNR, and MSE fail to provide a consistent and accurate result and indicate this attack as unsuccessful (in the image example $MSE = 0.52$, which is considered high). That is, because of the use of Euclidean distance-per-pixel measure, they miss the fact that the attacker has recovered the most important element, the actual number; regardless of how noisy or changed the background of the image is. Therefore, these metrics discard the contextual information in the image that a human's vision would have otherwise recognized, e.g. edges and points of interests (another example is 1a, row 2 & column 3, where $MSE = 1.06$ but a human would have read the number). It becomes even a larger challenge to use these metrics as a mechanism to approve data for FL in real-time. Since the development of attacks models are mostly empirical and data dependant, it is plausible to have an automatic verification (e.g. as a smart contract/client service) to assess the security of data by applying brute-force attacks before submission of gradients. For that purpose, a reliable and lightweight metric is needed.

This problem has not gone unnoticed and efforts to address these challenges have resulted in proposal of specialized metrics that may be computationally expensive, for example, Learned Perceptual Image Patch Similarity (LPIPS) from Zhang et al. (2018). The authors use the power of deep neural networks (DNN) to create LPIPS that is aligned with human perception metric. A major downside is that LPIPS is a complex and a computationally costly metric that is difficult to interpret due to the underlying DNN themselves that may require training for new data.

This paper introduces a new distance metric to assess data recovery, the *Absolute Variation Distance* (AVD). It is derived from total variation and in contrast to standard methods (MSE, SSIM), it offers a continuous metric for extracting information in noisy images. Furthermore, we show via a user study that AVD is highly correlated with human perception, but at the same time it is computationally more efficient and interpretable compared to LPIPS. Our results show that recovery of data is more visible as AVD decreases in a continuous manner. In contrast the MSE metric for MNIST fluctuates drastically when the image is not completely clear or a blend, and can obtain various values similar or higher than the MSE for the pure noise input.

Table 1: Types of gradient inversion attacks employed in our study.

Attack Name	Main Objective Function	Description
2-norm	g^{l2} (3)	Euclidean distance and initial label determination.
Angle & var	g^{ang} + TV (4)	Geiping et al. (2020) proposed to leverage cosine similarity, total variation (TV) and initial label determination.
Angle & var & Orth_regulators	g^{ang} + TV + Orth	Cosine distance with orthogonal regulator for the input + initial label determination. (Qian et al., 2021)

(a) (b)

Figure 2: Random recovered vectors from the LFW face dataset. Sub-figure (a) shows the images sorted by the AVD metric. Sub-figure (b) shows the images sorted by the MSE metric. The smaller figures represent the original images associated with image above.

2 Absolute Variation Distance

In this paper we developed AVD, a variant of total variation metric (Rudin et al., 1992), which is a more suitable indicator to compare the spatial gradient of the recovered image and source image. Given two images v^{source} and v^{target}, we define AVD between them as following:

$$\begin{aligned}\text{AVD}(v^{source}, & v^{target}) = \\ & ||(|\nabla v^{source}| - |\nabla v^{target}|)|| + \\ & ||(|\nabla^2 v^{source}| - |\nabla^2 v^{target}|)||\end{aligned} \tag{2}$$

where $\nabla v = \frac{dv}{di} + \frac{dv}{dj}$ is the spatial gradient and $\nabla^2 v = \frac{d^2 v}{di^2} + \frac{d^2 v}{dj^2}$ is the second order gradient. Here we treat the image as a 2-D array with with values $v(i,j)$. Therefore, because AVD measures distance in gradient space it allows to consider boundaries and edges in images which are a common discriminator in visual recognition, whilst the gradient of noise remains as noise.

2.1 Inversion Attack Algorithms

In our setting (and typically) the gradient inversion attack is carried out by choosing x'_t, y'_t on a proxy model, $F'(x'_t, y'_t)$, and then minimizing an objective function that measures the distance between gradients computed the proxy model $\Delta W'_t$ and the original gradients. A typical objective can be the norm of the gradients' difference:

$$g^{l2}(x'_t, y'_t) = \min||\Delta W'_t - \Delta W_t|| \tag{3}$$

This solution searches for a model $F'(x'_t, y'_t)$ that matches the size of the gradient vector observed by the client. Although further empirical studies have found the cosine distance to provide better convergence results (Geiping et al., 2020):

$$g^{ang}(x'_t, y'_t) = \min 1 - \frac{\langle \Delta W'_t, \Delta W_t \rangle}{||\Delta W'_t|| \cdot ||\Delta W_t||} \tag{4}$$

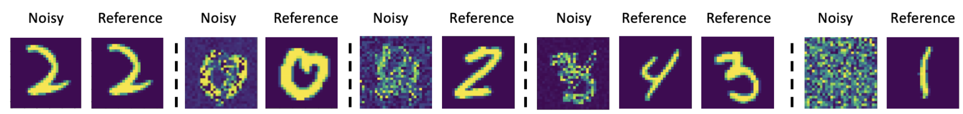
Various regularisation terms were shown to improve convergence. For example, regularisation that penalises high variations in the input images and constrains the search to high-fidelity images and de-noised solutions (Geiping et al., 2020; Yin et al., 2021). In mini-batches the orthogonality (Qian et al., 2021) between input vectors in the batch has been shown to bias the search towards different vectors in the batch. Additionally it has been found that determining the label from the gradients is important for initialisation of the numerical optimisation (Yin et al., 2021).

In our study in section 4 we apply various types of attacks and regularisation terms to provide a comprehensive analysis without any prior assumption on the performance of the attack. As summarised in Table 1, we utilise both the Euclidean distance and cosine similarity objective functions proposed by recent prior work (Zhu et al., 2019b; Geiping et al., 2020) including a selection of popular regularisation functions.

3 Experiments

We conduct gradient inversion attack experiments on two benchmnark datasets, MNIST Handwritten Digit (LeCun et al., 2010) and Labelled Faces in the Wild (LFW) (Huang et al., 2007), to illustrate how our proposed metric successfully evaluates information leakage that aligns to human perception. These two dataset are commonly used among researchers to study attacks (Zhu et al., 2019b; Zhao et al., 2020a; Melis et al., 2019; Shokri et al., 2017). We explore the privacy of the input data with the standard LeNET convolutional neural network (LeCun et al., 1990). We analyse the impact of different loss functions (MSE, LPIPS, SSIM, PSNR). For the attacks, in terms of the optimisation scheme, we utilized the standard LFBGS, with learning rate (lr) of 0.05, batch size of 4, and 300 iterations for running a proxy model to attack.

We also carried out a complementary user study by asking 10 individuals for their feedback, to rank a series of inverse attack images from $0-5$. With 0 being that they can very clearly observe underlying information (e.g. they can see the number 9 in a mnist recovered image, see example Fig. 1), to 5 that they cannot extract any useful information (e.g. the image is pure noise). We randomly generated 6 groups, with 100 images each (total 600 images); specifically, three 10×10 frames of LFW images and three 10×10 frames of MNIST images (see Appendix, Fig. 17 and 18 for an example of LFW and MNIST). The images were not ranked by noise level/clarity, but they were randomly allocated into the 10×10 frame. After the individuals ranked them then we averaged the scores they gave for each of the 100 images and we compared it to the score each image achieved from MSE, AVD, and LPIPS metrics. We used Pearson correlation (ρ) as a measure of similarity. The results can be studied in the heatmap in Fig. 3.



AVD	0.00	0.62	0.50	0.48	0.40	0.80
MSE	0.00	0.51	1.32	1.05	0.92	0.60
SSIM	0.99	0.09	0.13	0.12	0.18	0.12
PSNR	50.77	15.45	14.88	13.63	15.81	17.76
LPIPS	0.00	0.26	0.49	0.26	0.34	0.82

Figure 3: Comparison table of the most widely used metrics in FL for evaluation of inversion attacks (MSE, SSIM, PSNR, LPIPS), plus our own novel metric, the AVD. Each column compares the metrics between a noisy (attack generated) image and its reference (original) image. We included a wide array of examples, from no noise (column 1), to a mixture with two references (column 4), and complete noise (column 5).

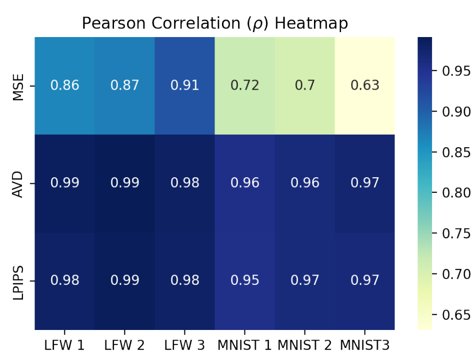


Figure 4: Pearson correlation ρ heatmap between the average ranking score (0-5) of 10 people for each image, and the MSE, LPIPS, AVD scores of these images. A ranking score of 0 means that the human can perceive very clear information in the image and a rank of 5 means the image is pure noise. The x-axis shows 6 groups (LFW1, LFW2, LFW3, MNIST1, MNIST2, MNIST3) of images. Each group has a 10×10 frame, 100 images in each frame. The y-axis shows the the three inversion attack metrics. The correlation is between the metrics (y-axis) and the average ranking score by the users. For more details refer to Section 3.

4 Results

For Fig. 1 we ran the experiment for the MNIST dataset. The first sub-figure (a) sorts the images by AVD value. For comparison, the second sub-figure (b) sorts the images by the MSE. The smaller images represent the original MNIST data. Low score ranking represents that the attack was successful and the image matches the original. High score ranking shows that the attack was unsuccessful and the generated attack images are very noisy with no observable pattern. From Fig. 1b, row one and columns four and five, the $MSE_{1,4} = 1.67$ and $MSE_{1,5} = 3.22$. These images clearly have a pattern, a user might be able to infer that the number of the last image relates to the number 5. So the attacker can extract private information from an FL model. But, according to the MSE these images are more private (noisy) when compared against images two and three from the same row ($MSE_{1,2} = 0.97$ and $MSE_{1,2} = 1.01$). On the other hand, our metric AVD captures these irregularities, with $AVD_{1,4} = 0.69$ and $AVD_{1,5} = 0.43$ being lower than $AVD_{1,2} = 0.88$ and $AVD_{1,3} = 0.85$. The AVD scores also agree with the LPIPS benchmark in these images, which indicates the AVD follows the human perception to evaluate the success of an inversion attack. In the LFW dataset, Fig. 2, we can observe the same phenomenon when using the MSE as a score to evaluate FL inversion attacks. In Fig. 3 we compare different inversion attack metrics, including PSNR, SSIM, MSE, LPIPS, and AVD. The LPIPS and AVD results are consistent and agree very well with human consensus; they attribute the lowest value ($LPIPS = 0\&AVD = 0$) to column one that the two images are identical, and the largest value at column five ($LPIPS = 0.82\&AVD = 0.80$), where the generated image is just noise.

In our final experiment, we contacted a qualitative survey amongst 10 people, Fiq 3. When we evaluate inversion attacks in multidimensional data that exhibit strong intercorrelation amongst the data-points, such as images, then a contextual interpretation of the image is imperative for an accurate evaluation. Therefore, the similarity metric should be able to showcase human-like perception. Our survey results further support the quantitative analysis that we conducted in Fig. 3 and show that the AVD is highly correlated ($\rho \geq 0.96$) with how a human would have recognised information from an attack generated image. For the LFW group of images, the MSE had a correlation between $0.86 \leq \rho \leq 0.91$ with the average human score. On the other hand, for the same images, the LPIPS and the AVD achieved very high levels of correlation, $0.98 \leq \rho \leq 0.99$. For the MNIST group, the MSE showed correlation between $0.63 \leq \rho \leq 0.72$. The AVD though retained consistently high levels of correlation with the human score, $0.96 \leq \rho \leq 0.97$. It seems that the mixing of numbers and the change of the background that we observed in the MNIST examples (Fig. 1 and Fig. 3) "confuse" the MSE score and drive the results further away from human perception, reducing its accuracy.

5 Conclusion

In this paper, we have addressed a significant challenge in the field of FL - the evaluation of the success of inversion attacks and the effectiveness of defense strategies. Traditional metrics such as the SSIM, PSNR, and MSE have been shown to be insufficient for accurately assessing the success of these attacks, particularly in the context of multidimensional outputs like images. These metrics, which assume spatial independence, fail to consider the semantic context of the recovered data, leading to potentially misleading evaluations.

To overcome these limitations, we introduced the AVD, a metric for assessing data recovery and information leakage in FL. Derived from total variation, AVD offers a continuous measure for extracting information in noisy images, aligning closely with human perception. It is computationally more efficient and mathematically more interpretable than the LPIPS, a deep learning-based metric. The quantitative experiments demonstrated that AVD provides a more accurate and consistent measure of data recovery, thereby offering a more reliable tool for evaluating defense strategies against inversion attacks in FL. Also, the survey that we contacted amongst 10 people, asking them to rank random generated images by scoring the success of the recovered image, showed that human perception had high correlation with the AVD scores.

By providing a more accurate measure of data recovery, AVD allows researchers to better understand the effectiveness of their defense strategies and to develop robust FL evaluation of data security. We hope that our work will inspire further advancements in the field of FL and contribute to the development of more secure and reliable distributed learning systems.

References

Balunović, M., Dimitrov, D. I., Staab, R., and Vechev, M. Bayesian framework for gradient leakage, 2022.

Cahn, C. A note on signal-to-noise ratio in band-pass limiters. *IRE Transactions on Information Theory*, 7(1):39–43, 1961. doi: 10.1109/TIT.1961.1057616.

Chen, Y., Gui, Y., Lin, H., Gan, W., and Wu, Y. Federated learning attacks and defenses: A survey, 2022.

Dwork, C. Differential privacy. In *33rd International Colloquium on Automata, Languages and Programming, part II (ICALP 2006)*, volume 4052 of *Lecture Notes in Computer Science*, pp. 1–12. Springer Verlag, July 2006. ISBN 3-540-35907-9. URL `https://www.microsoft.com/en-us/research/publication/differential-privacy/`.

Eloul, S., Silavong, F., Kamthe, S., Georgiadis, A., and Moran, S. J. Enhancing privacy against inversion attacks in federated learning by using mixing gradients strategies, 2022.

Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. Inverting Gradients - How easy is it to break privacy in federated learning? In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/c4ede56bbd98819ae6112b20ac6bf145-Abstract.html`.

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

Huang, Y., Gupta, S., Song, Z., Li, K., and Arora, S. Evaluating gradient inversion attacks and defenses in federated learning. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021a. URL `https://openreview.net/forum?id=0CDKgyYaxC8`.

Huang, Y., Song, Z., Li, K., and Arora, S. Instahide: Instance-hiding schemes for private distributed learning, 2021b.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., and Jackel, L. Handwritten Digit Recognition with a Back-Propagation Network. In *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1990. URL `https://proceedings.neurips.cc/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf`.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

McMahan, B., Moore, E., Ramage, D., Hampson, S., and Arcas, B. A. y. Communication-Efficient Learning of Deep Networks from Decentralized Data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017. URL `https://proceedings.mlr.press/v54/mcmahan17a.html`.

Melis, L., Song, C., Cristofaro, E. D., and Shmatikov, V. Exploiting unintended feature leakage in collaborative learning. In *2019 IEEE Symposium on Security and Privacy, SP 2019, San Francisco, CA, USA, May 19-23, 2019*, pp. 691–706. IEEE, 2019. doi: 10.1109/SP.2019.00029. URL `https://doi.org/10.1109/SP.2019.00029`.

Nguyen, D. C., Pham, Q.-V., Pathirana, P. N., Ding, M., Seneviratne, A., Lin, Z., Dobre, O., and Hwang, W.-J. Federated learning for smart healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(3):1–37, 2022.

Phong, L. T., Aono, Y., Hayashi, T., Wang, L., and Moriai, S. Privacy-preserving deep learning via additively homomorphic encryption. *IEEE Transactions on Information Forensics and Security*, 13(5):1333–1345, 2018. doi: 10.1109/TIFS.2017.2787987.

Qian, J., Nassar, H., and Hansen, L. K. Minimal model structure analysis for input reconstruction in federated learning, 2021.

Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H., Albarqouni, S., Bakas, S., Galtier, M., Landman, B., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R., Trask, A., Xu, D., Baust, M., and Cardoso, M. The future of Digital Health with Federated Learning. *npj Digital Medicine*, 3(1), December 2020. ISSN 2398-6352. doi: 10.1038/s41746-020-00323-1.

Rudin, L. I., Osher, S., and Fatemi, E. Nonlinear total variation based noise removal algorithms. In *Proceedings of the Eleventh Annual International Conference of the Center for Nonlinear Studies on Experimental Mathematics: Computational Issues in Nonlinear Science: Computational Issues in Nonlinear Science*, pp. 259–268, USA, 1992. Elsevier North-Holland, Inc.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 3–18. IEEE Computer Society, 2017. doi: 10.1109/SP.2017.41. URL `https://doi.org/10.1109/SP.2017.41`.

Sikandar, H. S., Waheed, H., Tahir, S., Malik, S. U. R., and Rafique, W. A detailed survey on federated learning attacks and defenses. *Electronics*, 12(2), 2023. ISSN 2079-9292. doi: 10.3390/electronics12020260. URL `https://www.mdpi.com/2079-9292/12/2/260`.

Wainakh, A., Zimmer, E., Subedi, S., Keim, J., Grube, T., Karuppayah, S., Guinea, A. S., and Mühlhäuser, M. Federated learning attacks revisited: A critical discussion of gaps, assumptions, and evaluation setups, 2022.

Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated Machine Learning: Concept and Applications. *ACM Trans. Intell. Syst. Technol.*, 10(2), jan 2019. ISSN 2157-6904. doi: 10.1145/3298981. URL `https://doi.org/10.1145/3298981`.

Yin, H., Mallya, A., Vahdat, A., Alvarez, J., Kautz, J., and Molchanov, P. See through Gradients: Image Batch Recovery via GradInversion. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16332–16341, 2021. doi: 10.1109/CVPR46437.2021.01607.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

Zhao, B., Mopuri, K. R., and Bilen, H. idlg: Improved deep leakage from gradients, 2020a.

Zhao, Y., Zhao, J., Yang, M., Wang, T., Wang, N., Lyu, L., Niyato, D., and Lam, K.-Y. Local differential privacy based federated learning for internet of things, 2020b.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a. URL `https://proceedings.neurips.cc/paper_files/paper/2019/file/60a6c4002cc7b29142def8871531281a-Paper.pdf`.

Zhu, L., Liu, Z., and Han, S. Deep leakage from gradients. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL `https://proceedings.neurips.cc/paper/2019/file/60a6c4002cc7b29142def8871531281a-Paper.pdf`.

Appendix

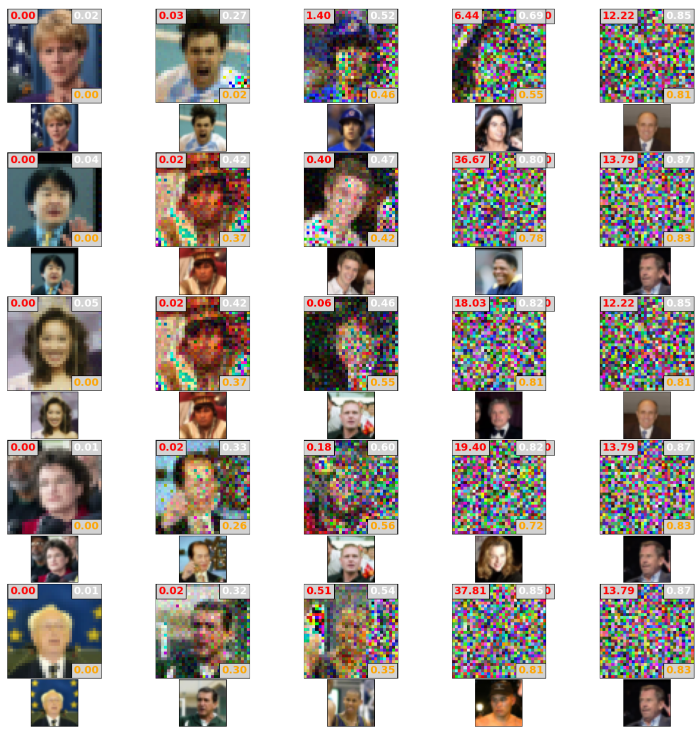


Figure 5: Random recovered vectors from LFW datasets, column-wise sorted via the AVD.

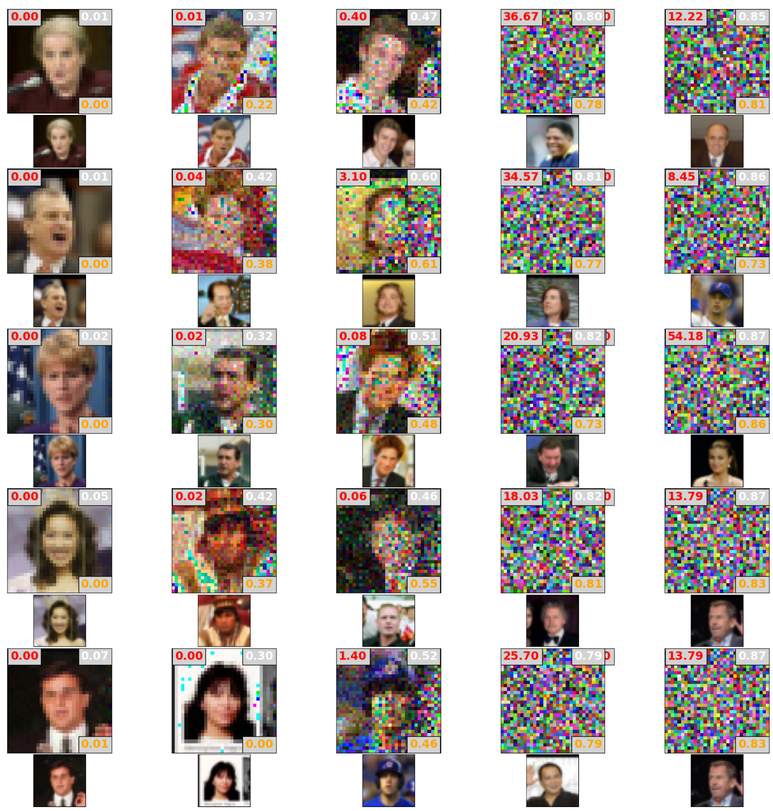


Figure 6: Random recovered vectors from LFW datasets, column-wise sorted via the AVD.

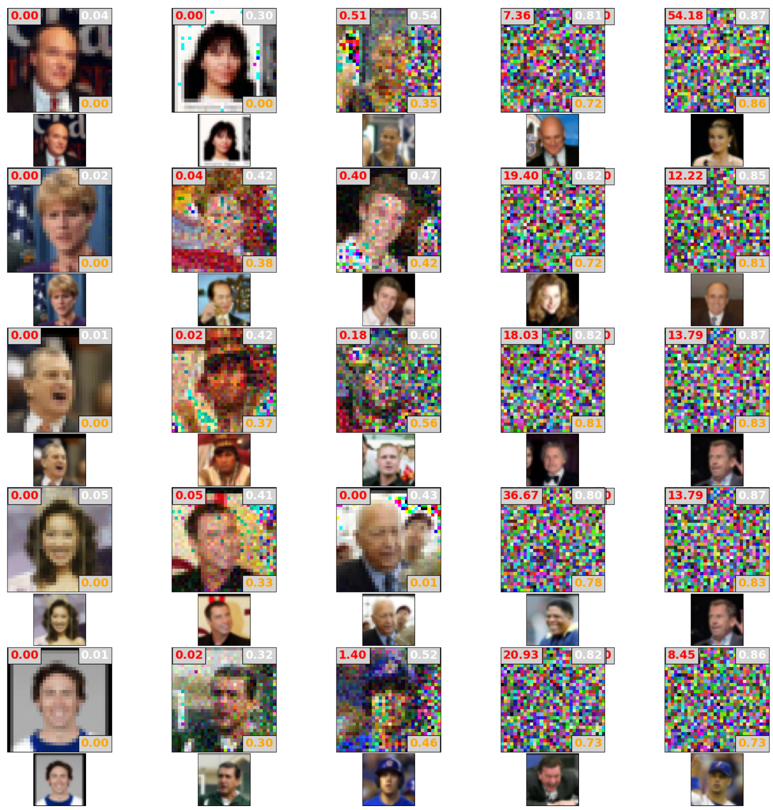


Figure 7: Random recovered vectors from LFW datasets, column-wise sorted via the AVD.

Figure 8: Random recovered vectors from LFW datasets, column-wise sorted via the MSE.

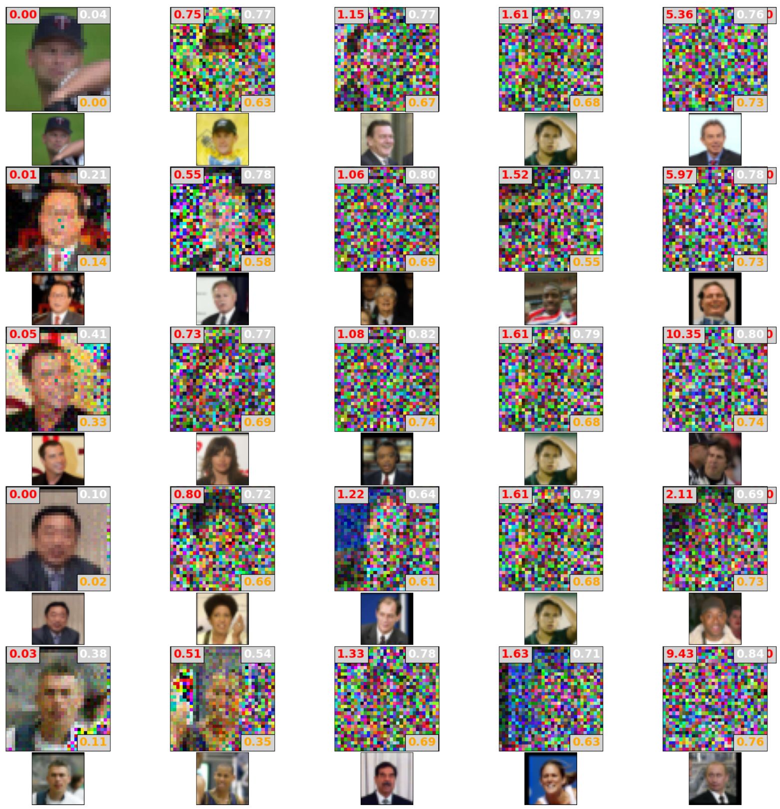


Figure 9: Random recovered vectors from LFW datasets, column-wise sorted via the MSE.

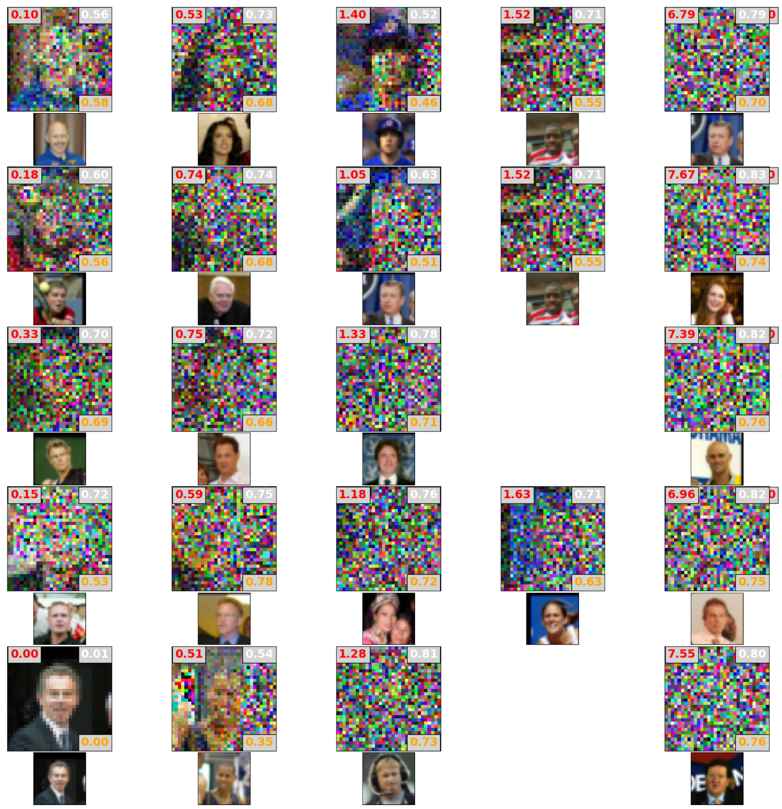


Figure 10: Random recovered vectors from LFW datasets, column-wise sorted via the MSE.

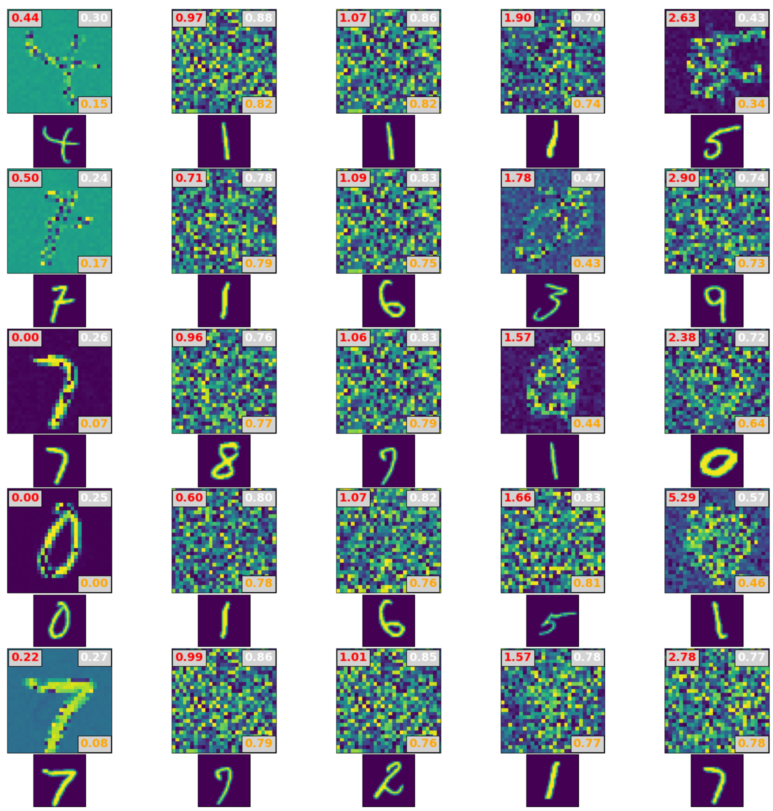


Figure 11: Random recovered vectors from MNIST datasets, column-wise sorted via the MSE.

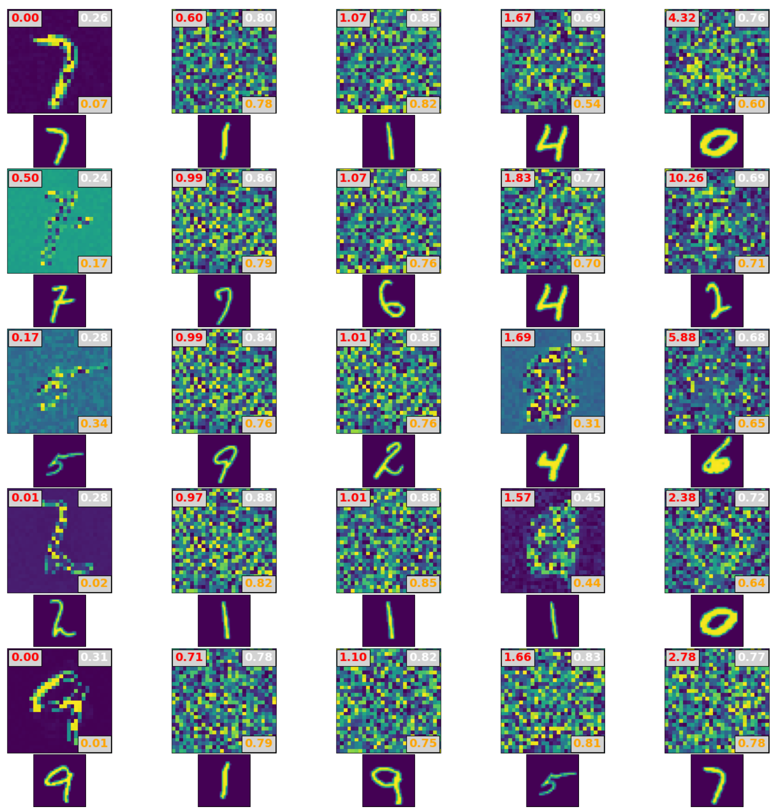


Figure 12: Random recovered vectors from MNIST datasets, column-wise sorted via the MSE.

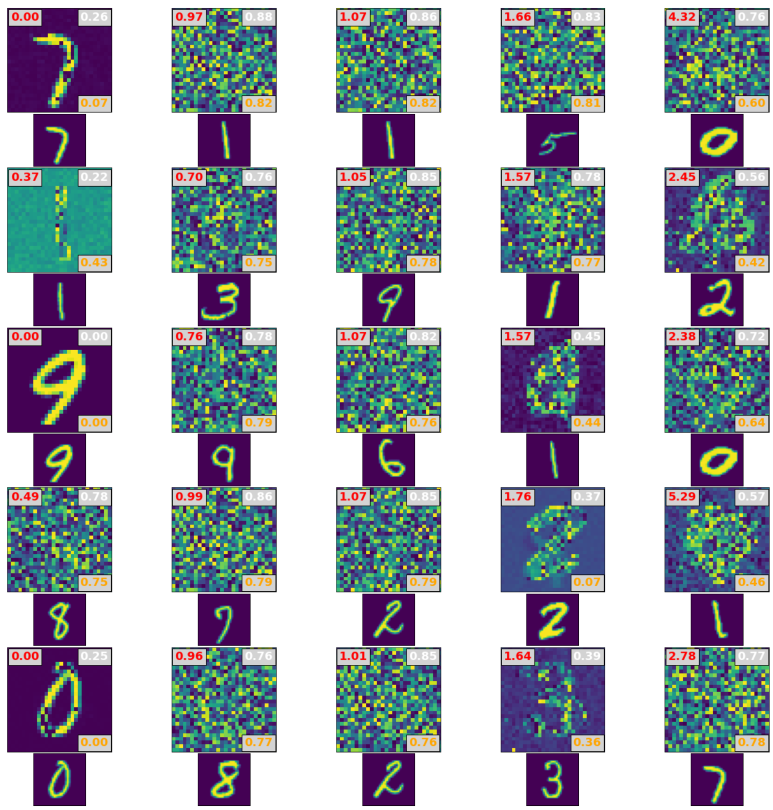


Figure 13: Random recovered vectors from MNIST datasets, column-wise sorted via the MSE.

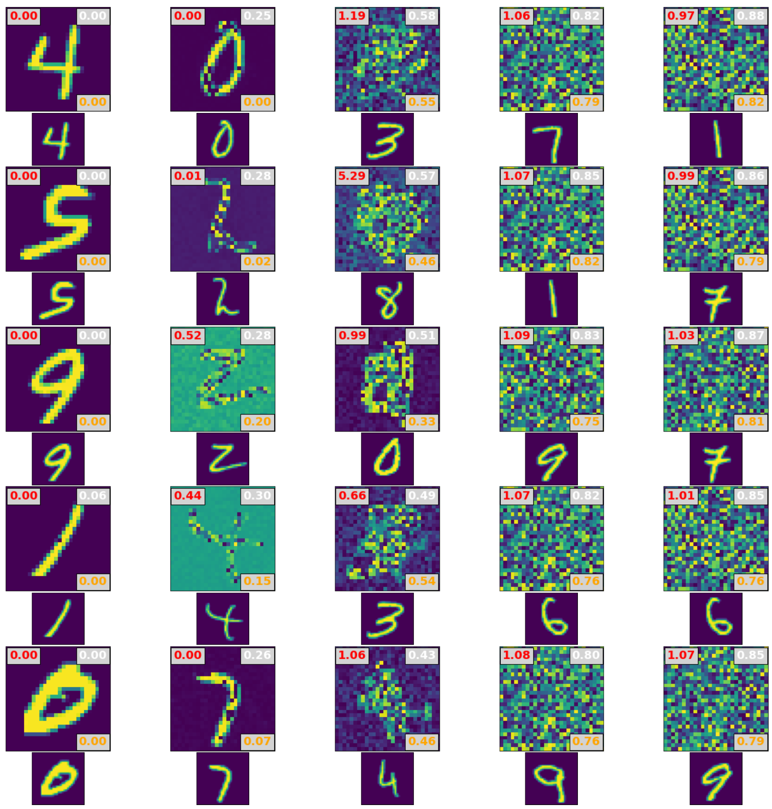


Figure 14: Random recovered vectors from MNIST datasets, column-wise sorted via the AVD.

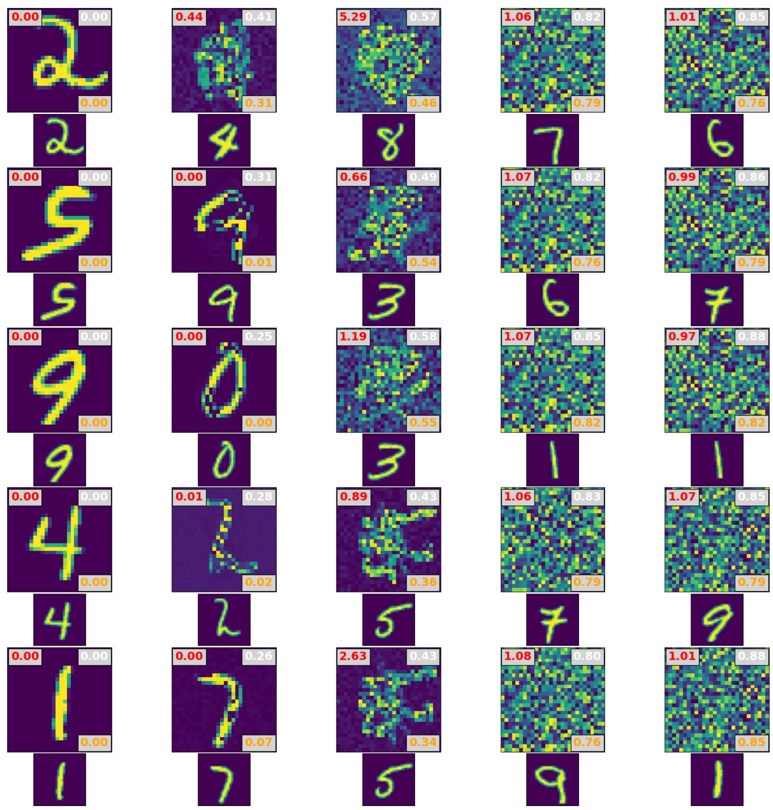


Figure 15: Random recovered vectors from MNIST datasets, column-wise sorted via the AVD.

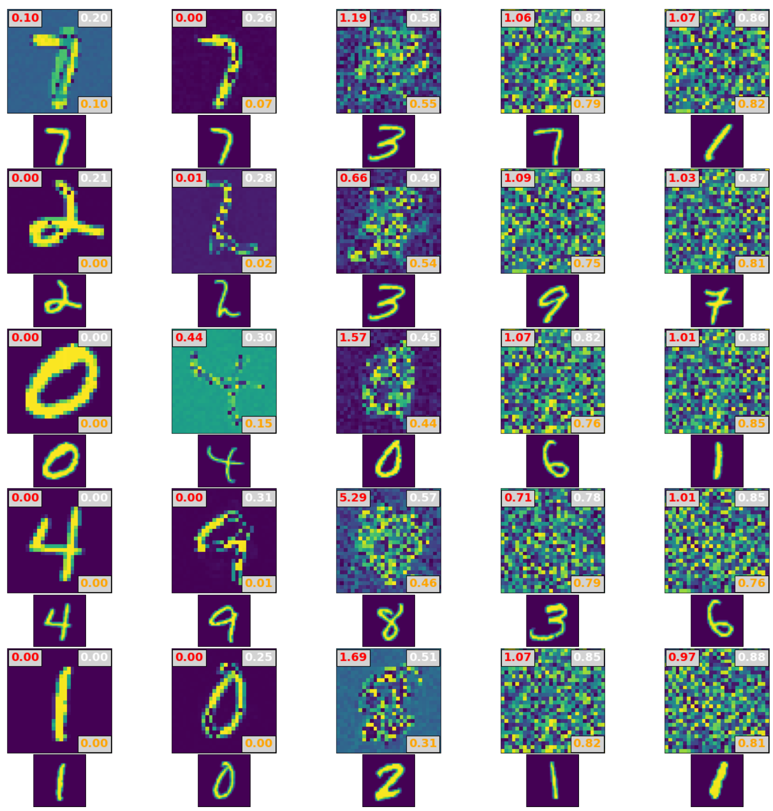


Figure 16: Random recovered vectors from MNIST datasets, column-wise sorted via the AVD.

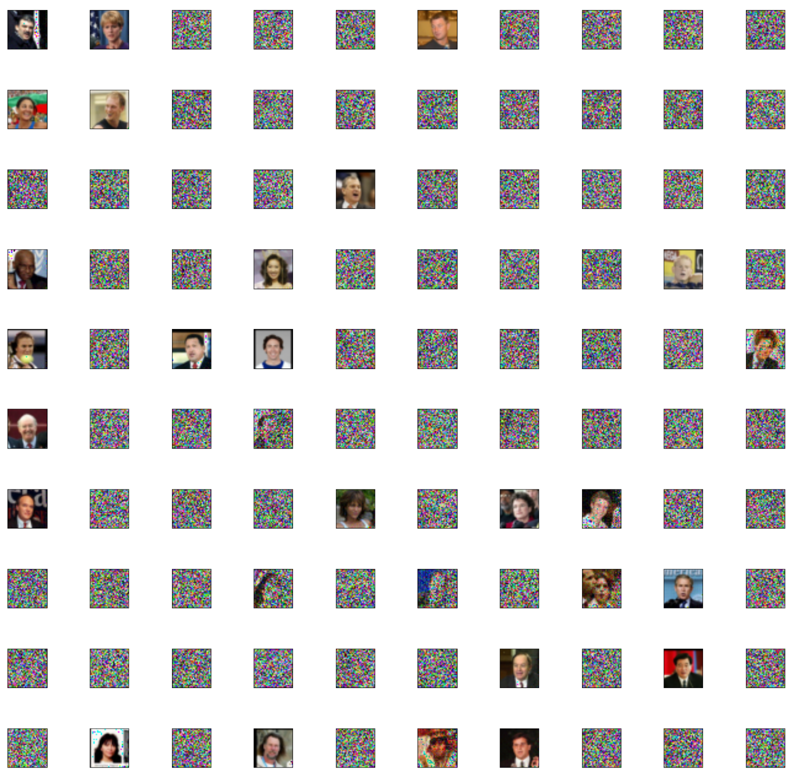


Figure 17: Random recovered vectors from MNIST datasets, column-wise sorted via the AVD.

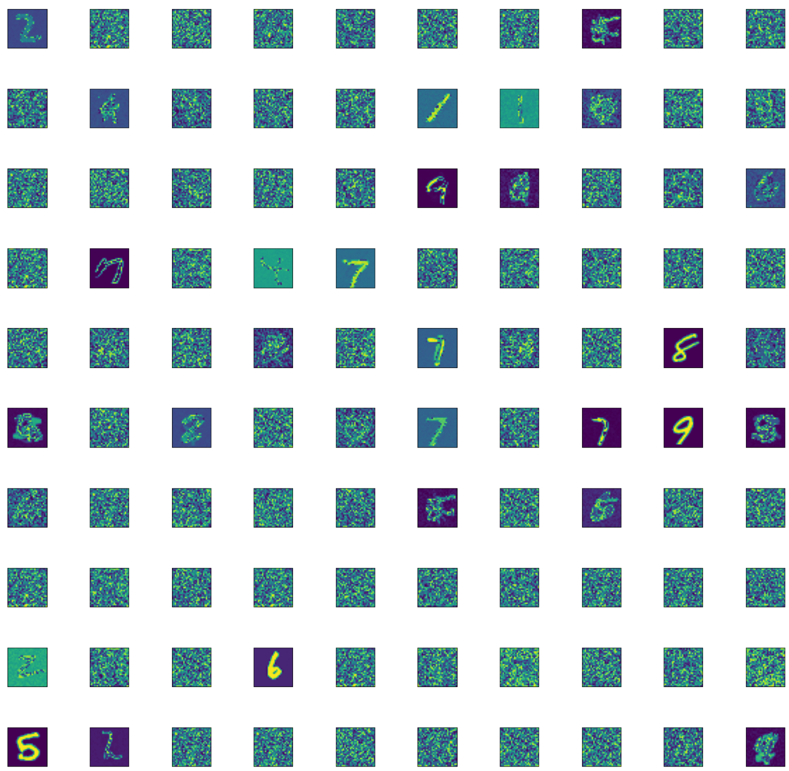


Figure 18: Random recovered vectors from MNIST datasets, column-wise sorted via the AVD.