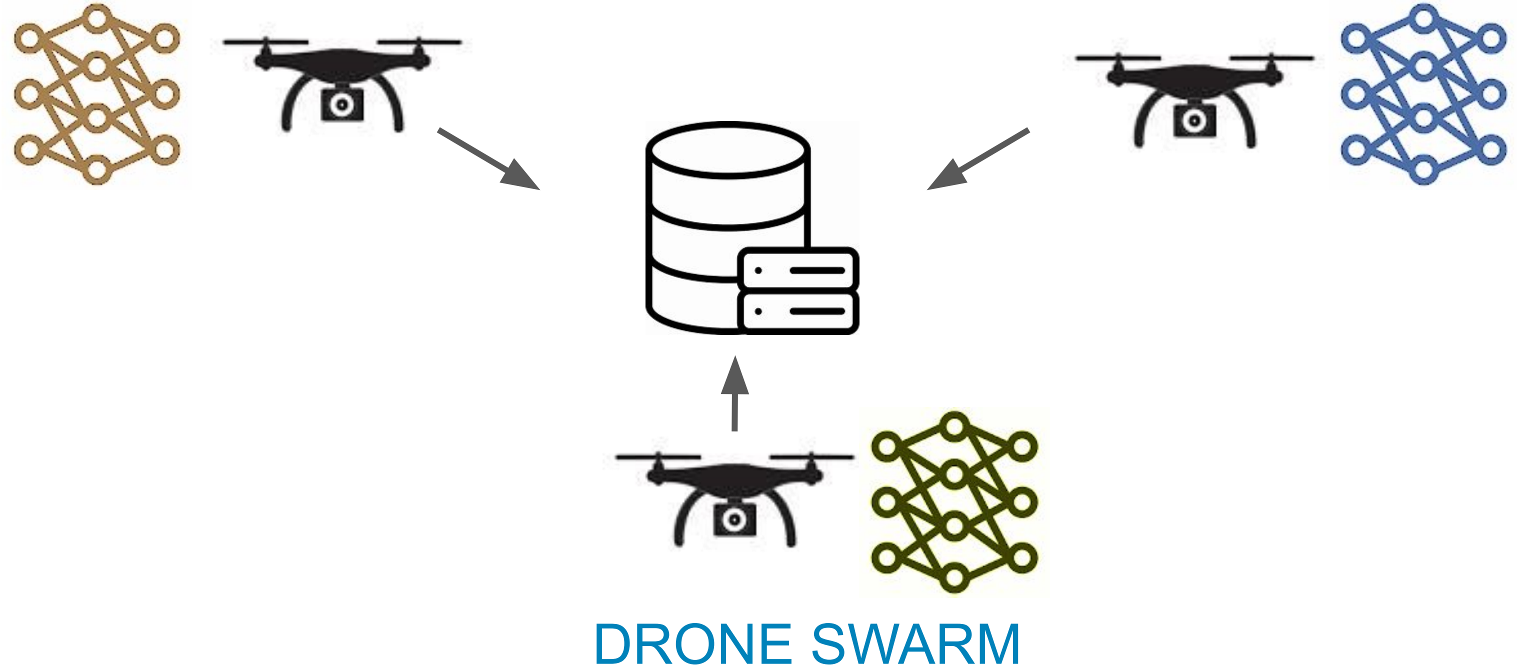
Semantic Label Reconstruction

How to Breach Privacy in Federated Learning

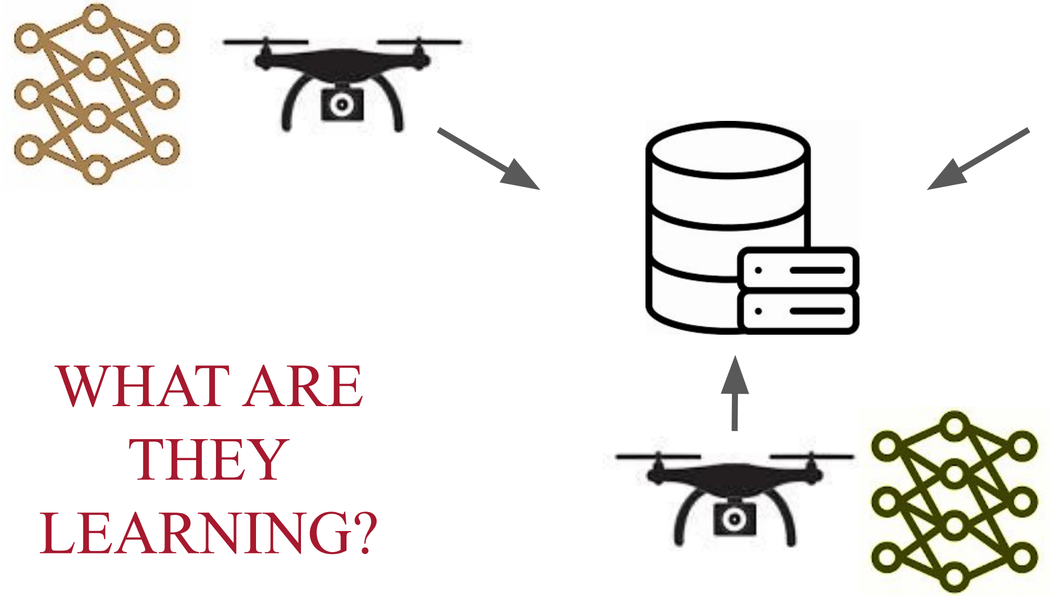
Rafał Malcewicz[1,2], Ignacy Stępka[1], Abby Turner[1], Artur Dubrawski[1]

[1]Auton Lab, Carnegie Mellon University
[2]Aalto University

Why This Matters



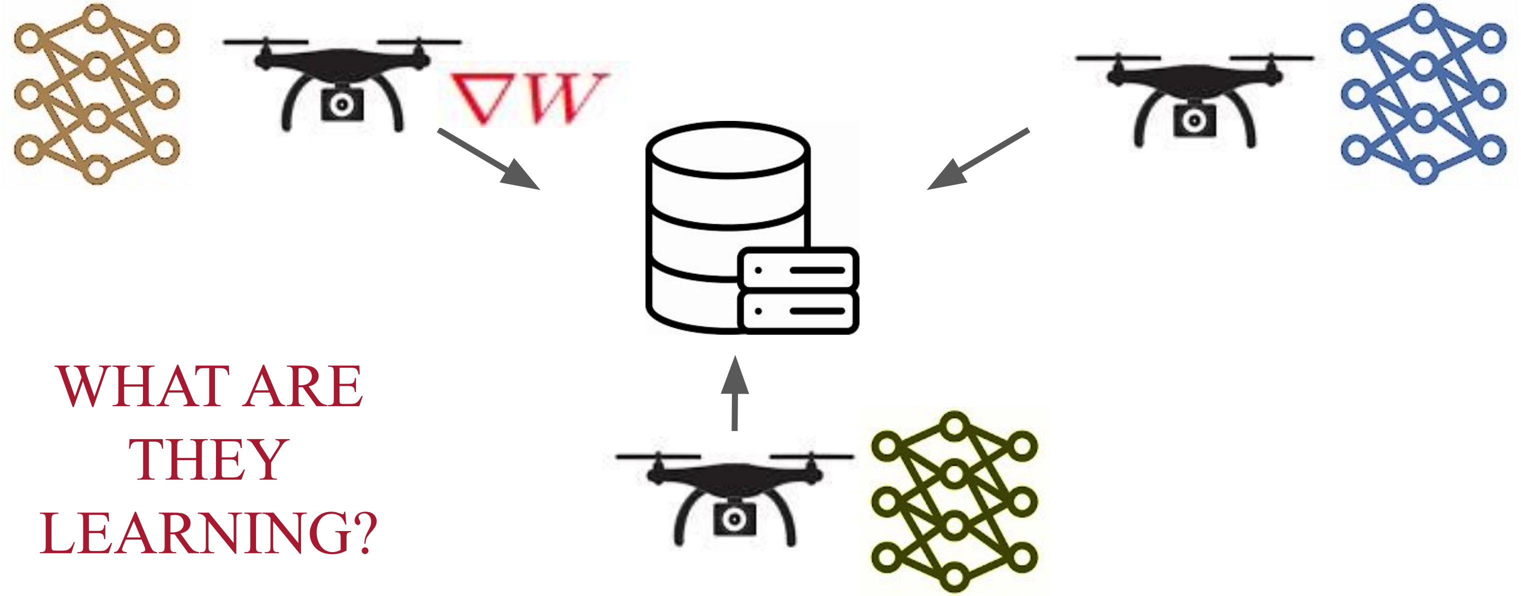
Why This Matters



WHAT ARE THEY LEARNING?

DRONE SWARM

Why This Matters

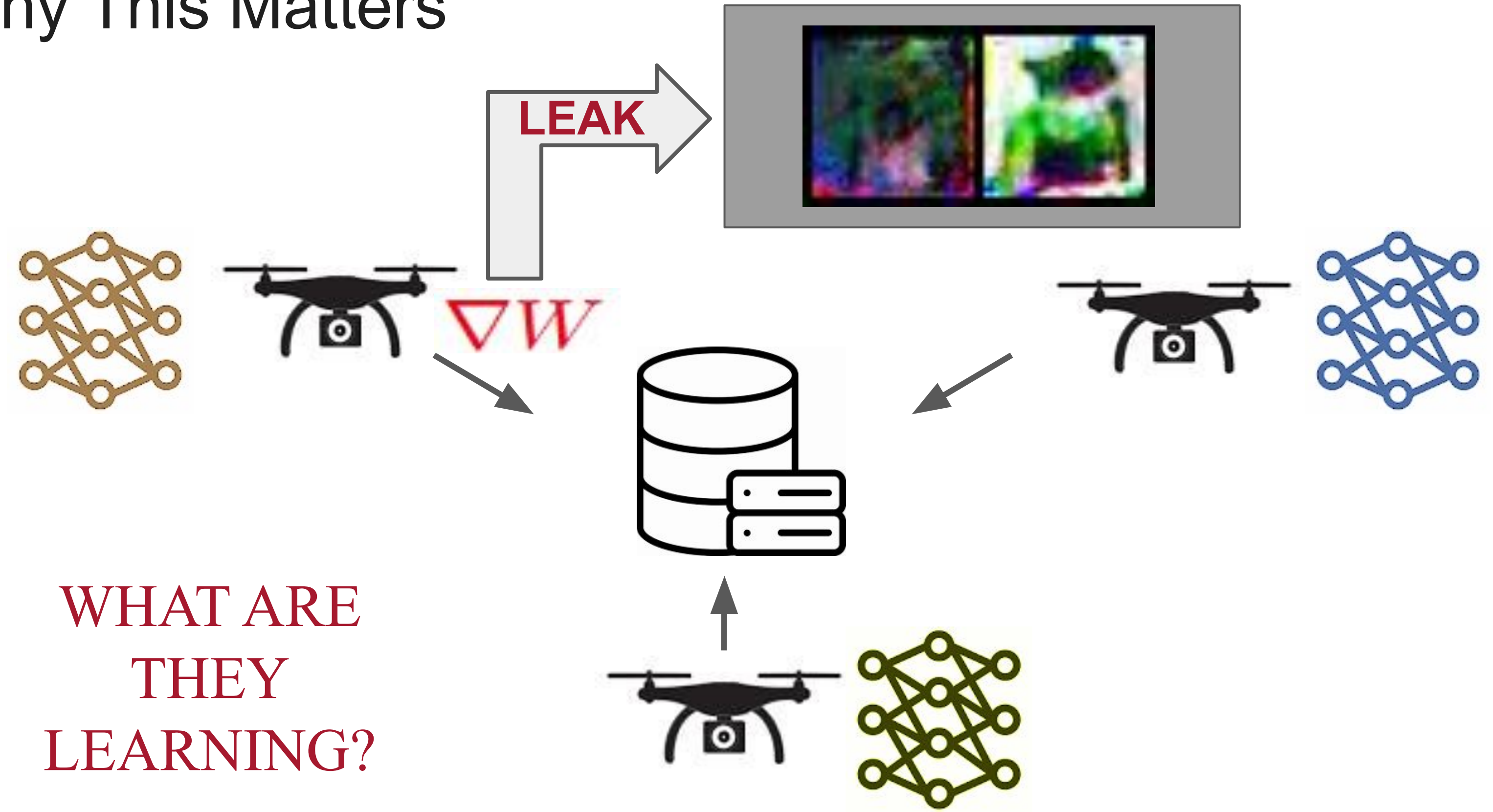


WHAT ARE THEY LEARNING?

DRONE SWARM

Why This Matters

Zhu et al., 2019 [1]



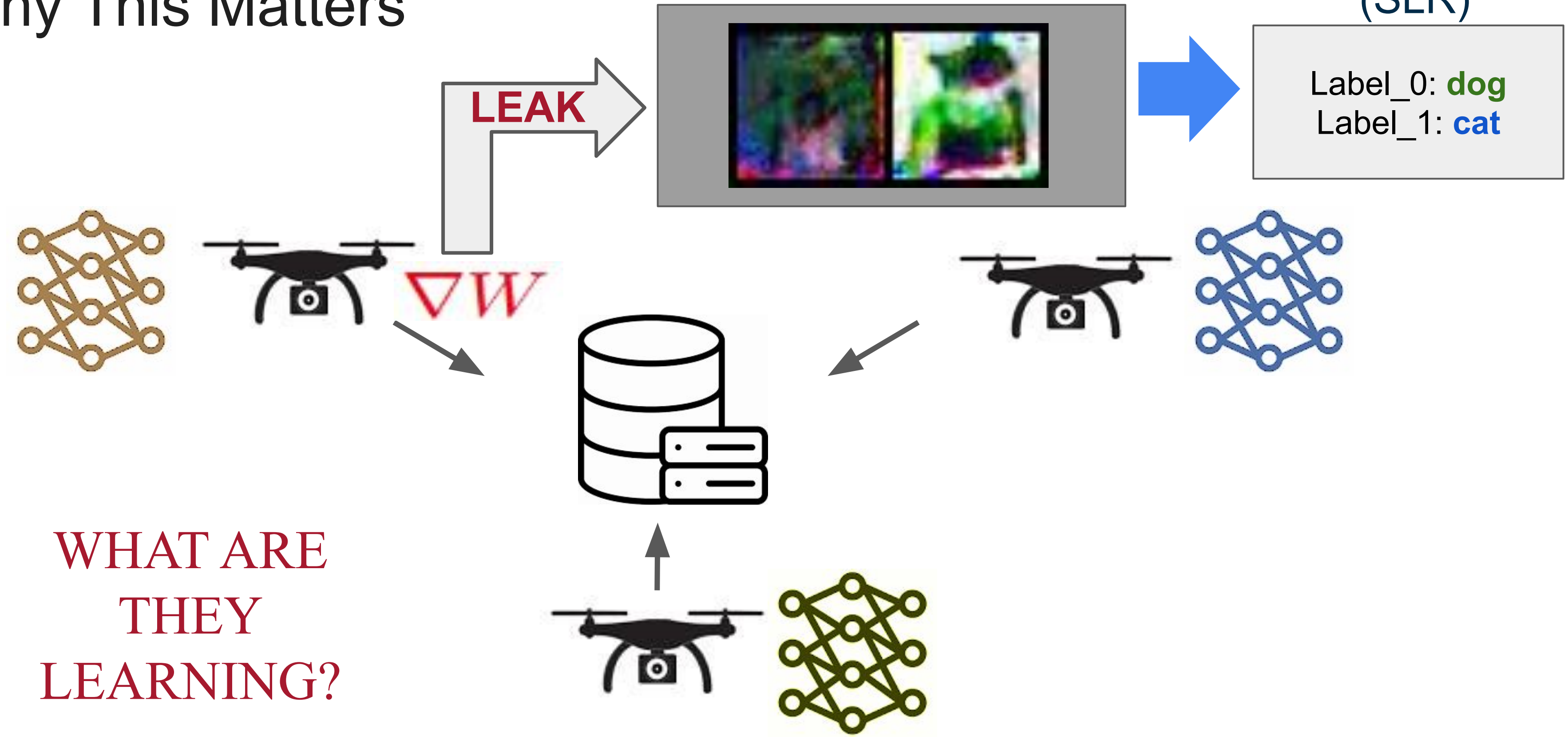
DRONE SWARM

[1] Zhu, Ligeng, Zhijian Liu, and Song Han. “Deep Leakage from Gradients.” arXiv, December 19, 2019.

Why This Matters

Zhu et al., 2019 [1]

Semantic Label Reconstruction (SLR)



DRONE SWARM

[1] Zhu, Ligeng, Zhijian Liu, and Song Han. “Deep Leakage from Gradients.” arXiv, December 19, 2019.

Plan for Today

Background

- What are Gradient Inversion Attacks (GIA's)?
- How do we evaluate the success of the reconstruction?

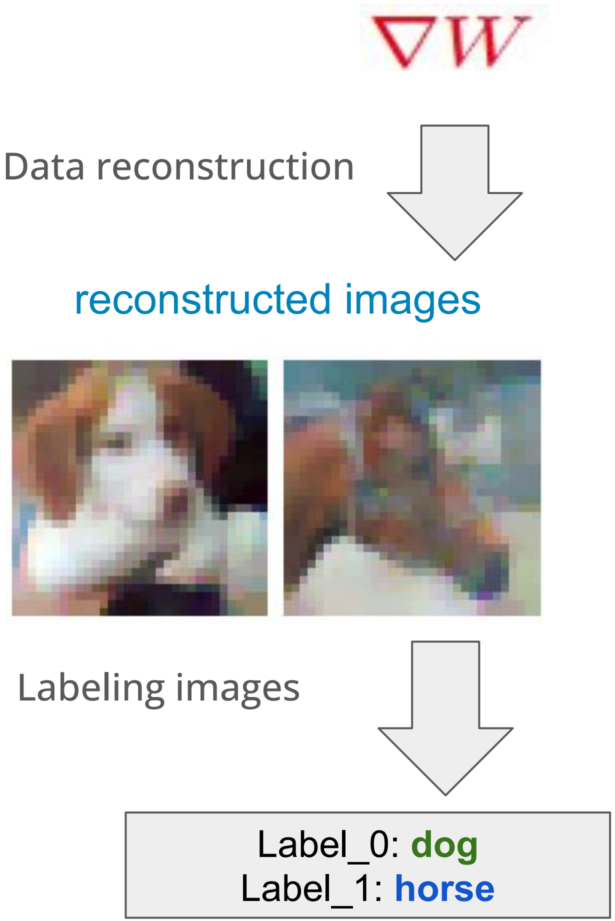
Semantic Label Reconstruction

- Label recovery with CLIP (Contrastive Language-Image Pre-training), Radford et al., 2021 [2]
- CLIP guided reconstruction

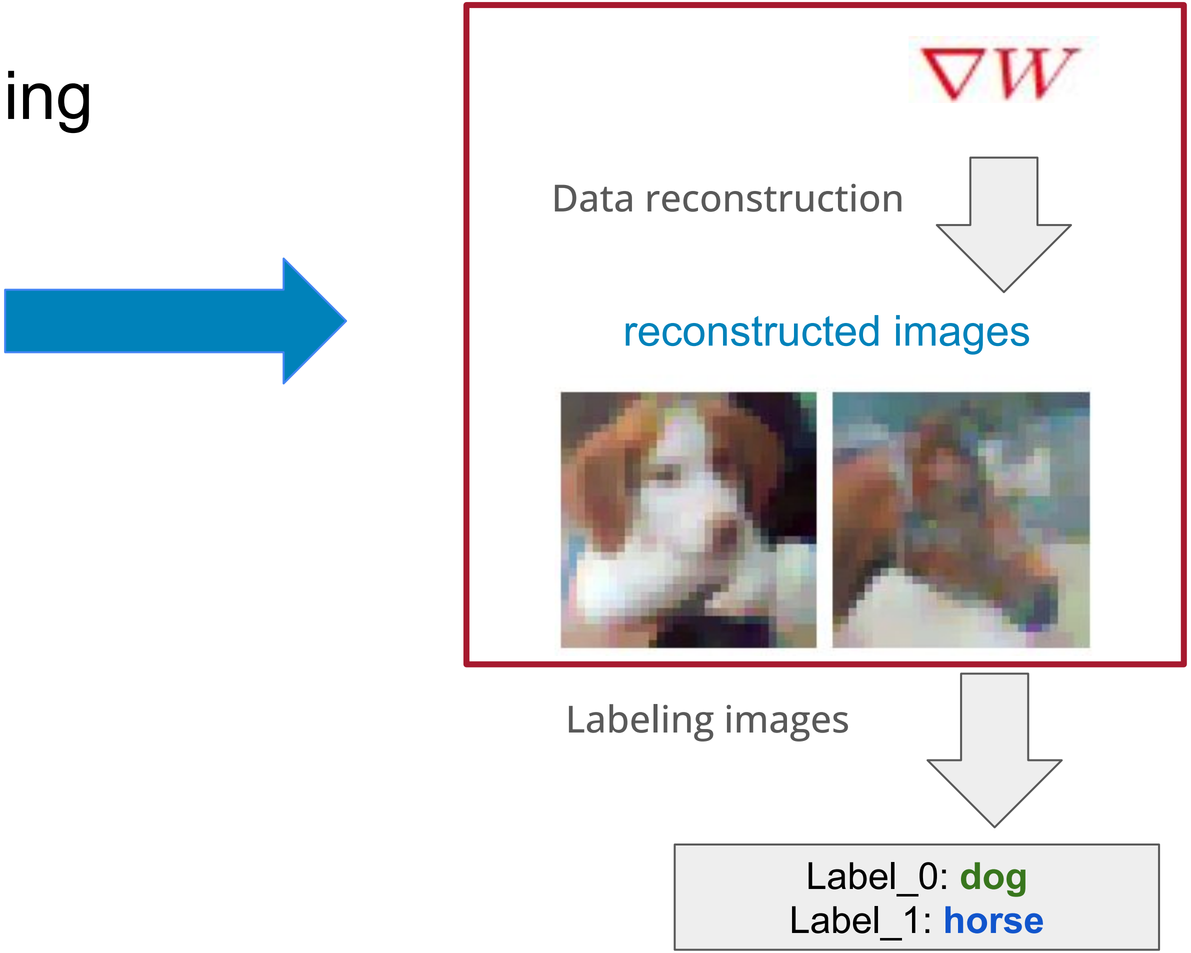
[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020

From Pixels to Meaning

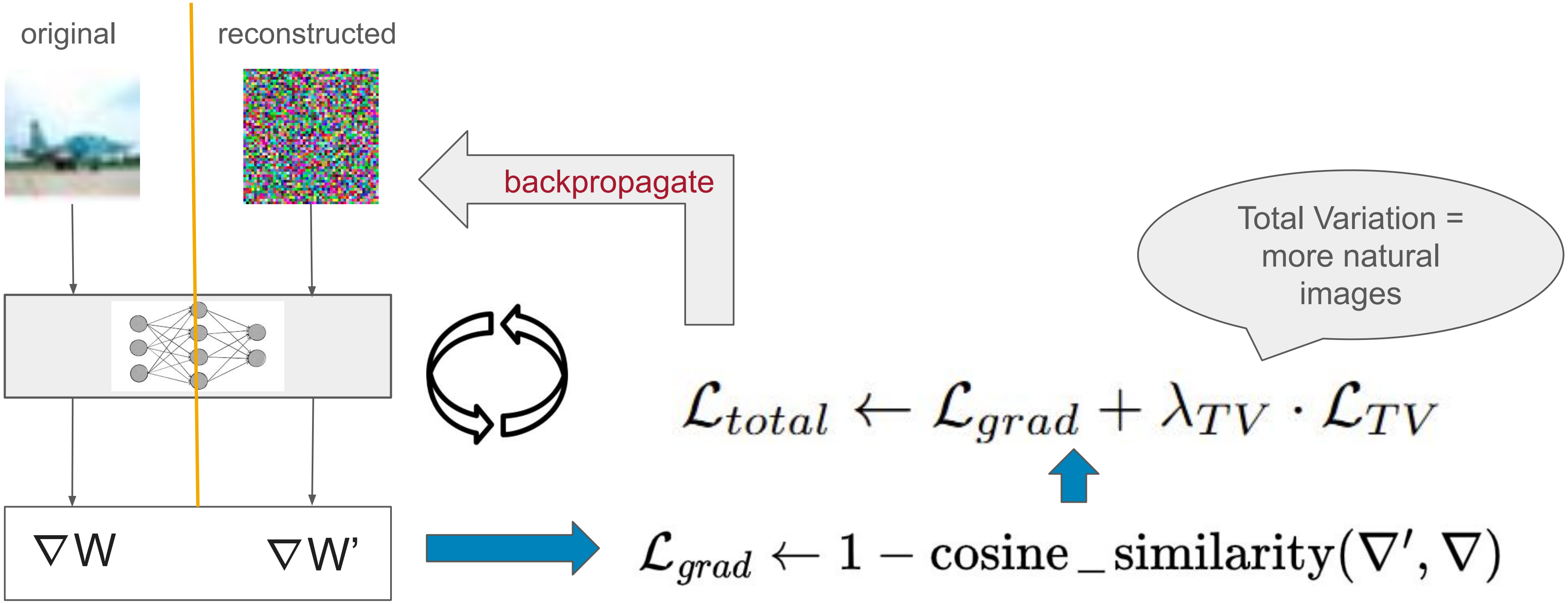
- Start with eavesdropped gradient
- Reconstruct images
- Label the the reconstructed image to retrieve semantic labels of the training data



From Pixels to Meaning



How Gradient Inversion Attack Work



$$\mathcal{L}_{total} \leftarrow \mathcal{L}_{grad} + \lambda_{TV} \cdot \mathcal{L}_{TV}$$

$$\mathcal{L}_{grad} \leftarrow 1 - \text{cosine\_similarity}(\nabla', \nabla)$$

Different Approaches to GIA’s

Approach	Paper	Method / Key Idea
Gradient Matching Loss	Zhu et al., 2019 [1]	Used **L2** distance
	Geiping et al., 2020 [2]	Used **cosine similarity**
Additional Loss Terms	Zhu et al., 2019 [1]	None
	Geiping et al., 2020 [2]	Added **Total Variation (TV)** loss
	Jeon et al., 2021 [3]	Added **Batch Normalization statistics**
Label Distribution Recovery	Zhao et al., 2020 [4]	Worked only for **batch size = 1**
	Ma et al., 2023 [5]	Solved **system of linear equations**
Latent Space Optimization	Fang et al., 2023 [6]	Used **GAN** to optimize latent space

Different Approaches to GIA’s

Approach	Paper	Method / Key Idea
Gradient Matching Loss	Zhu et al., 2019 [1]	Used **L2** distance
	Geiping et al., 2020 [2]	Used **cosine similarity**
Additional Loss Terms	Zhu et al., 2019 [1]	None
	Geiping et al., 2020 [2]	Added **Total Variation (TV)** loss
	Jeon et al., 2021 [3]	Added **Batch Normalization statistics**
Label Distribution Recovery	Zhao et al., 2020 [4]	Worked only for **batch size = 1**
	Ma et al., 2023 [5]	Solved **system of linear equations**
Latent Space Optimization	Fang et al., 2023 [6]	Used **GAN** to optimize latent space

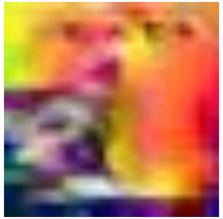
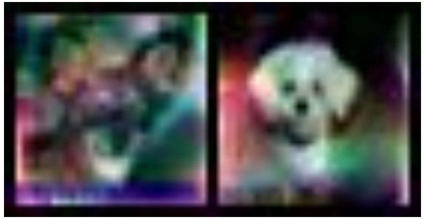
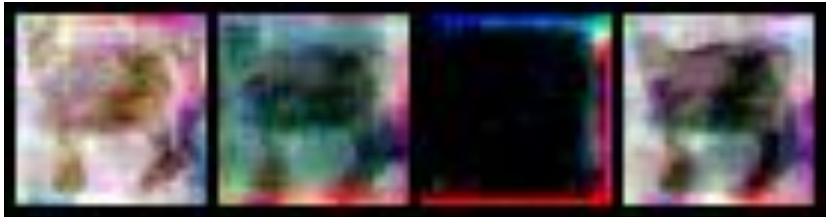
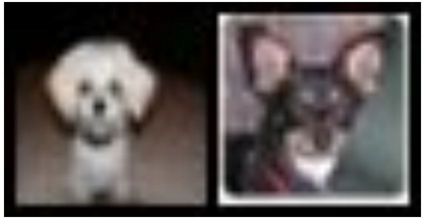
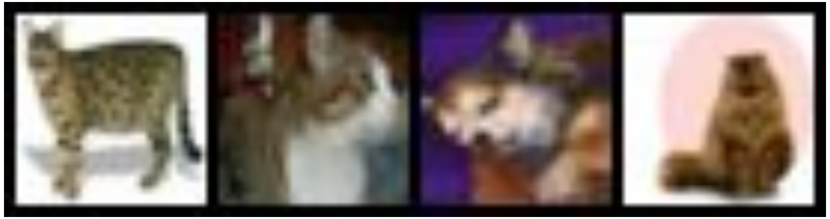
Different Approaches to GIA's

Approach	Paper	Method / Key Idea
Gradient Matching Loss	Zhu et al., 2019 [1]	Used **L2** distance
	Geiping et al., 2020 [2]	Used **cosine similarity**
Additional Loss Terms	Zhu et al., 2019 [1]	None
	Geiping et al., 2020 [2]	Added **Total Variation (TV)** loss
	Jeon et al., 2021 [3]	Added **Batch Normalization statistics**
Label Distribution Recovery	Zhao et al., 2020 [4]	Worked only for **batch size = 1**
	Ma et al., 2023 [5]	Solved **system of linear equations**
Latent Space Optimization	Fang et al., 2023 [6]	Used **GAN** to optimize latent space

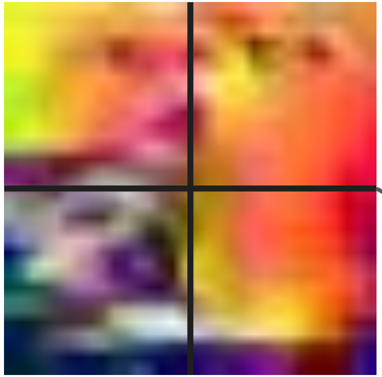
Different Approaches to GIA’s

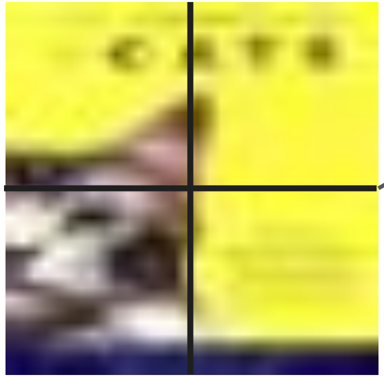
Approach	Paper	Method / Key Idea
Gradient Matching Loss	Zhu et al., 2019 [1]	Used **L2** distance
	Geiping et al., 2020 [2]	Used **cosine similarity**
Additional Loss Terms	Zhu et al., 2019 [1]	None
	Geiping et al., 2020 [2]	Added **Total Variation (TV)** loss
	Jeon et al., 2021 [3]	Added **Batch Normalization statistics**
Label Distribution Recovery	Zhao et al., 2020 [4]	Worked only for **batch size = 1**
	Ma et al., 2023 [5]	Solved **system of linear equations**
Latent Space Optimization	Fang et al., 2023 [6]	Used **GAN** to optimize latent space

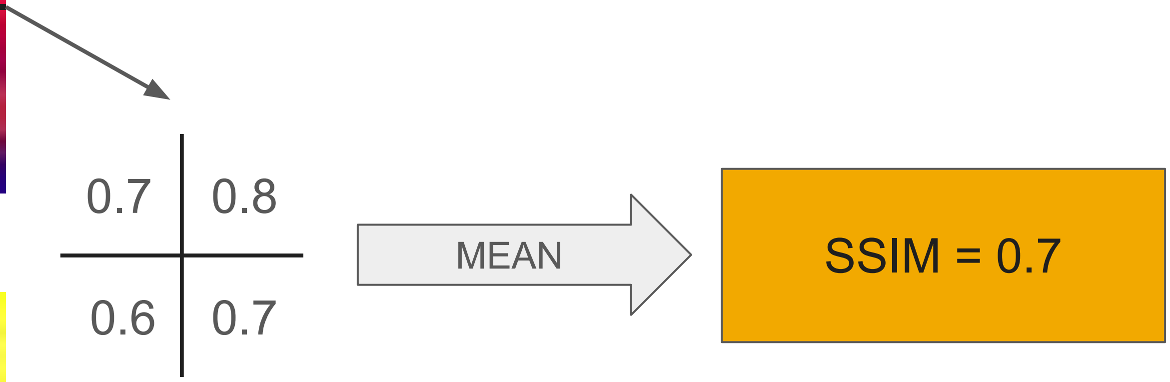
Results of the Reconstruction

	Batch Size		
	1	**2**	**4**
reconstructed			
original			

SSIM (Structural Similarity Index Measure)

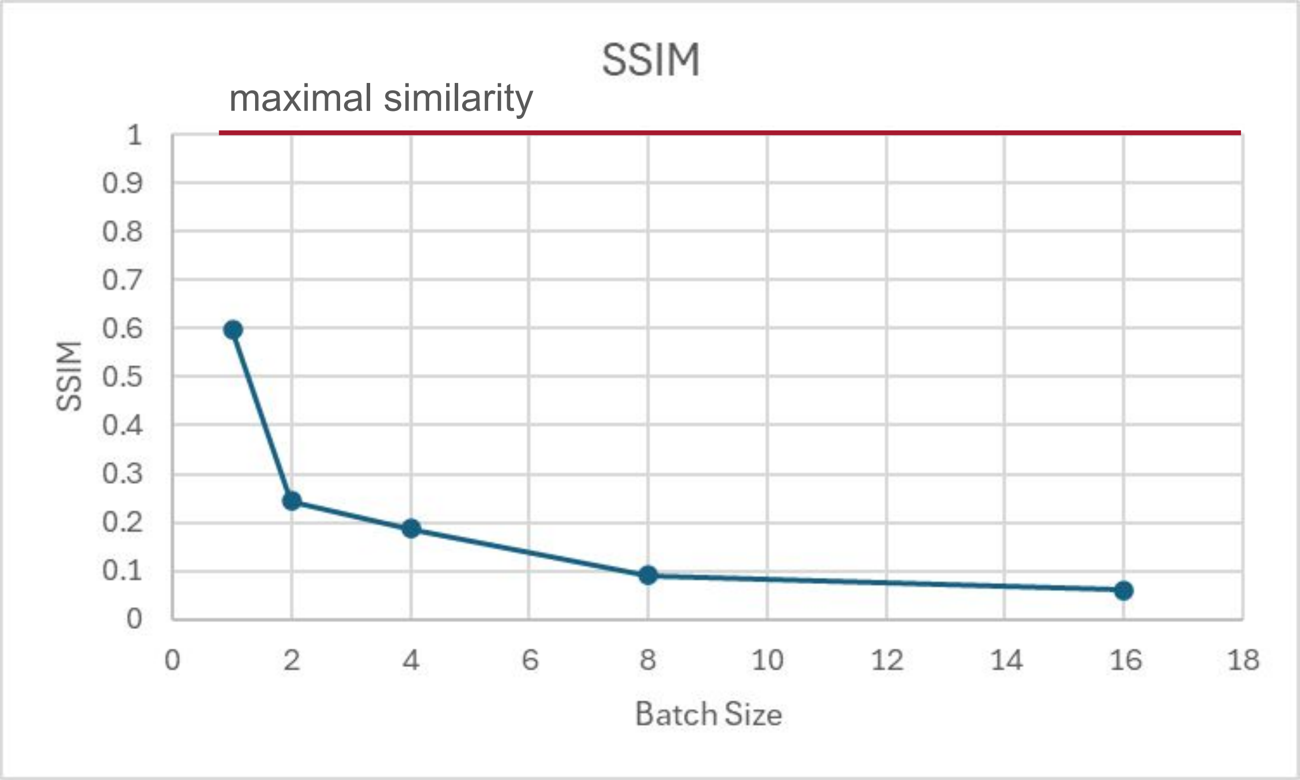






$$\mathrm{SSIM}(x, y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

SSIM (Structural Similarity Index Measure)



Permutation Agnostic SSIM

1st assignment

$$SSIM_1 = \frac{SSIM(\;\cdot\;,\;\cdot\;) + SSIM(\;\cdot\;,\;\cdot\;)}{2}$$

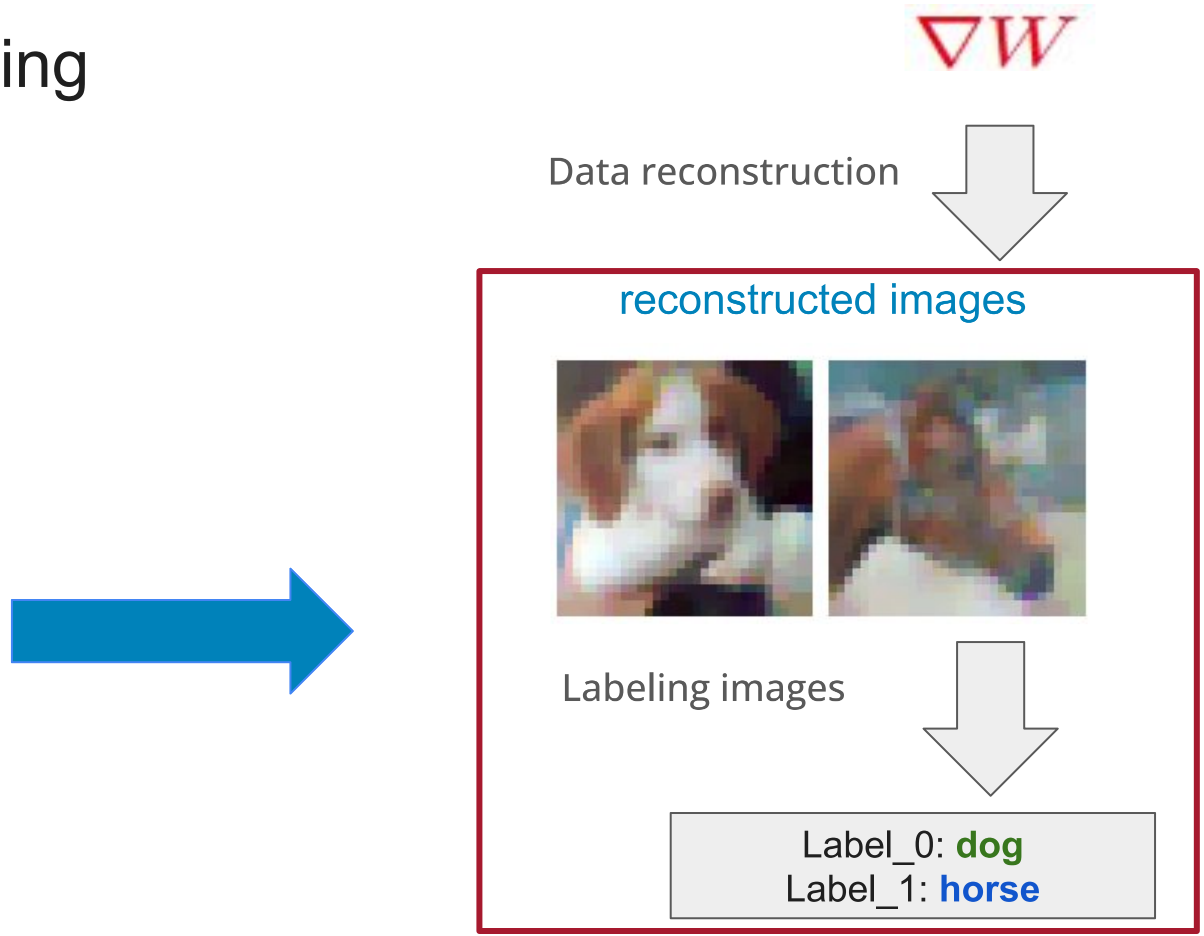
2nd assignment

$$SSIM_2 = \frac{SSIM(\;\cdot\;,\;\cdot\;) + SSIM(\;\cdot\;,\;\cdot\;)}{2}$$

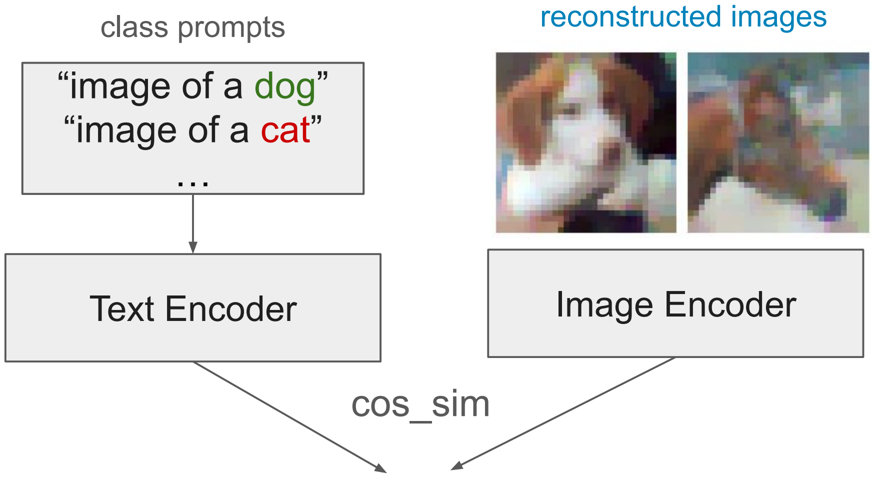
$$\mathbf{SSIM_{SMART}} = \max(SSIM_1, SSIM_2)$$

Permutation Agnostic SSIM

From Pixels to Meaning



CLIP for Label Recovery



Similarity Matrix

Image	Dog	Frog	Cat	Truck	...
A	**0.7**	0.1	0.05	0.05	...
B	**0.65**	0.12	0.05	0.05	...
C	0.08	**0.75**	0.1	0.02	...
D	0.1	**0.7**	0.07	0.02	...

(A, B: label_0; C, D: label_1)

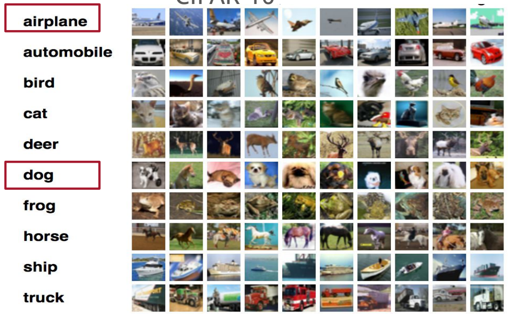
Label	Dog	Frog	Cat	Truck	...
0	**0.675**	0.11	0.05	0.05	...
1	0.09	**0.725**	0.085	0.02	…

Setup & Evaluation

Train model on some subset of classes

Setup & Evaluation

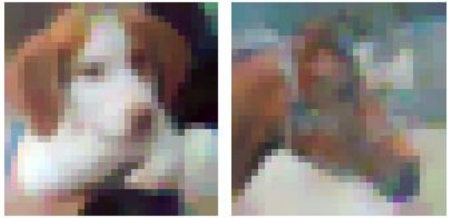
CIFAR-10



Train model on some subset of classes

$$\nabla W$$

Data reconstruction

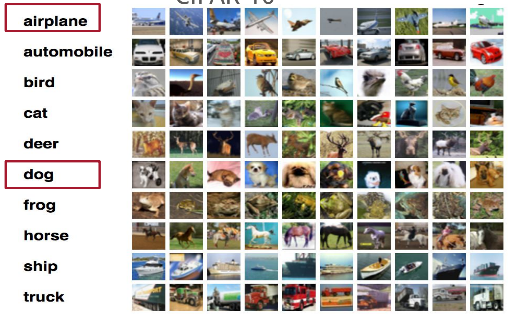


Labeling images

Label_0: **dog**
Label_1: **horse**

Setup & Evaluation

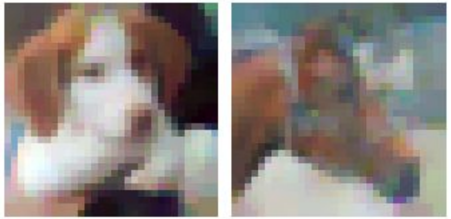
CIFAR-10



Train model on some subset of classes

∇W

Data reconstruction



Labeling images

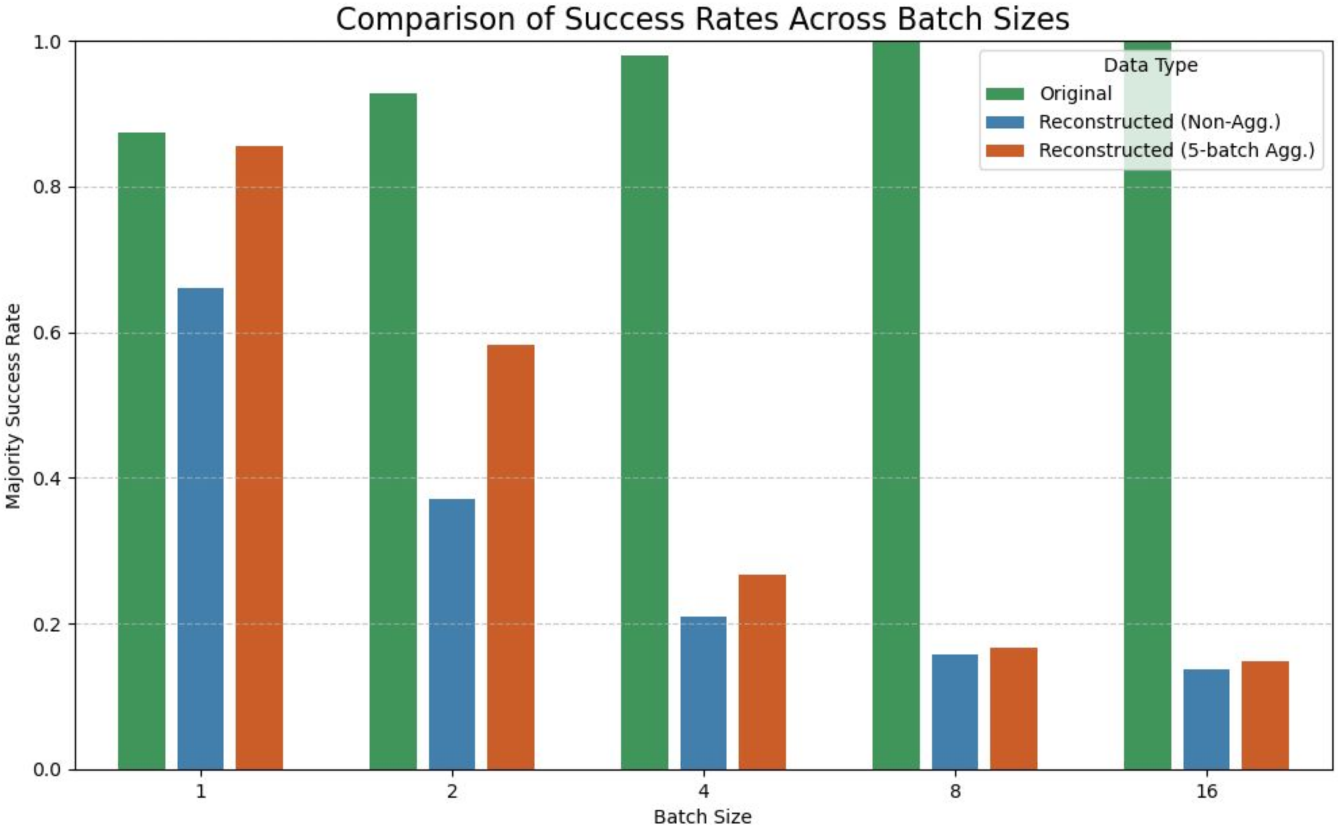
Metric:

- Success of recovery of majority class (0/1)

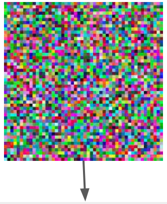
Label_0: **dog**
Label_1: **horse**

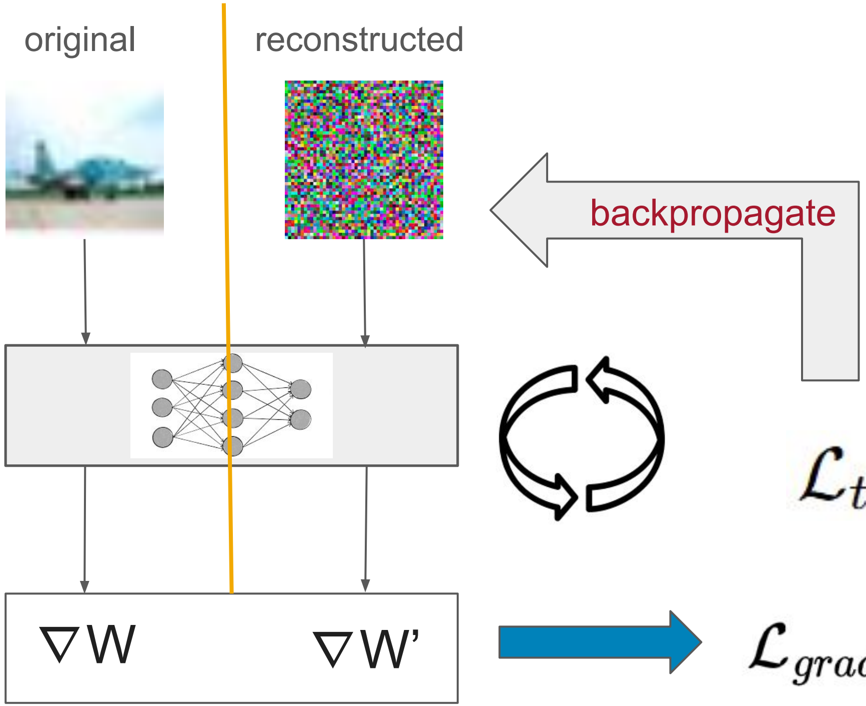
Leakage Drops as Batch Size Increases

- Original - almost perfect
- Aggregation improves quality
- Batch Size 4 still 2x better then random



CLIP Guided Reconstruction





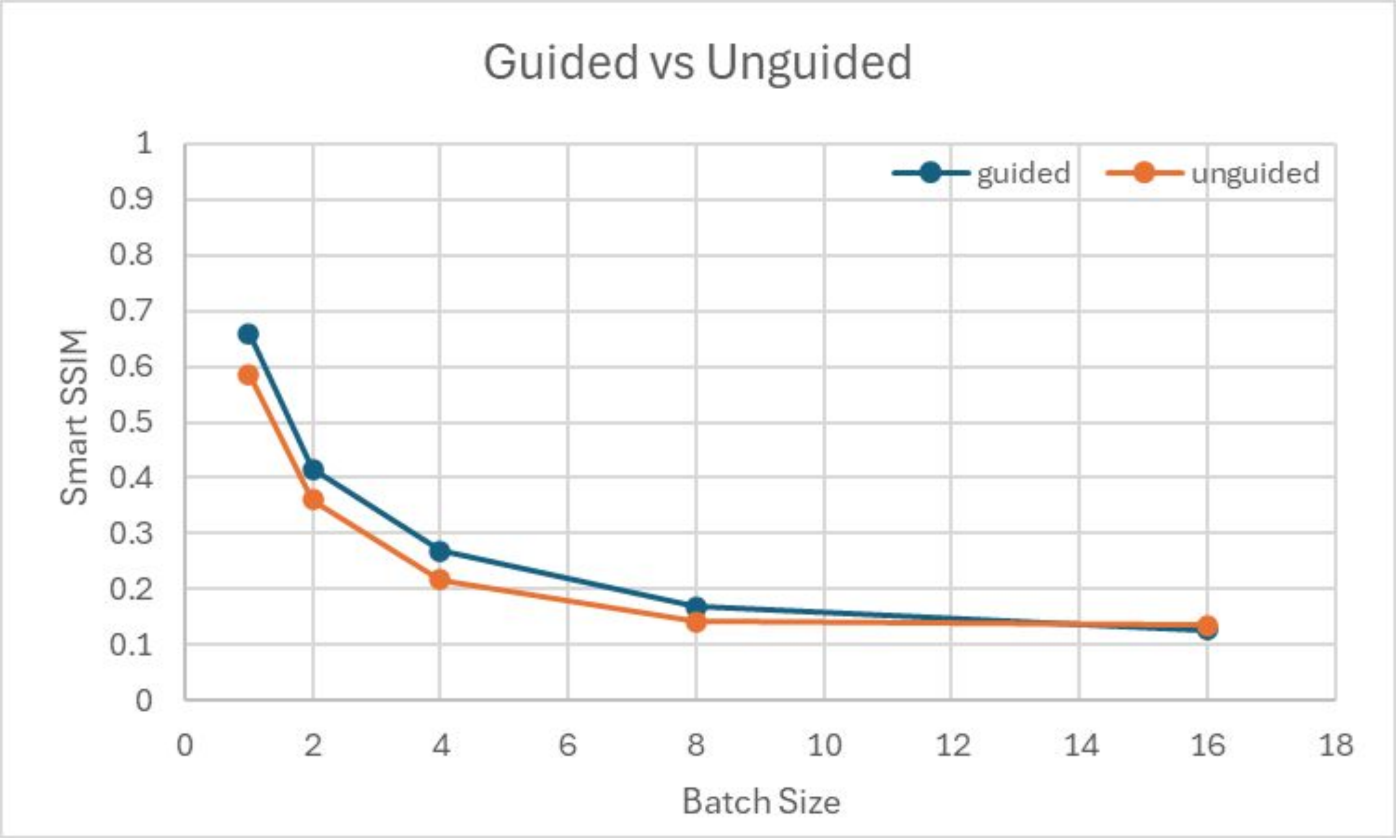
Guidance: “image of an airplane”

$$\mathcal{L}_{CLIP} = \frac{1}{N}\sum_{i=1}^{N}\Big(1 - cosine\_simmilarity(img^{(i)}, guidance\_prompt^{(i)})\Big)$$

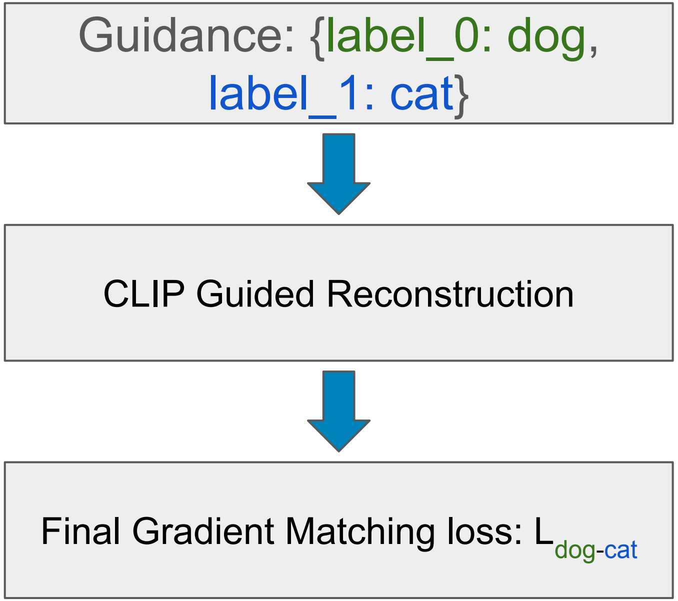
$$\mathcal{L}_{total} \leftarrow \mathcal{L}_{grad} + \lambda_{TV} \cdot \mathcal{L}_{TV}$$

$$\mathcal{L}_{grad} \leftarrow 1 - \text{cosine\_similarity}(\nabla', \nabla)$$

Oracle Guidance Improves Quality

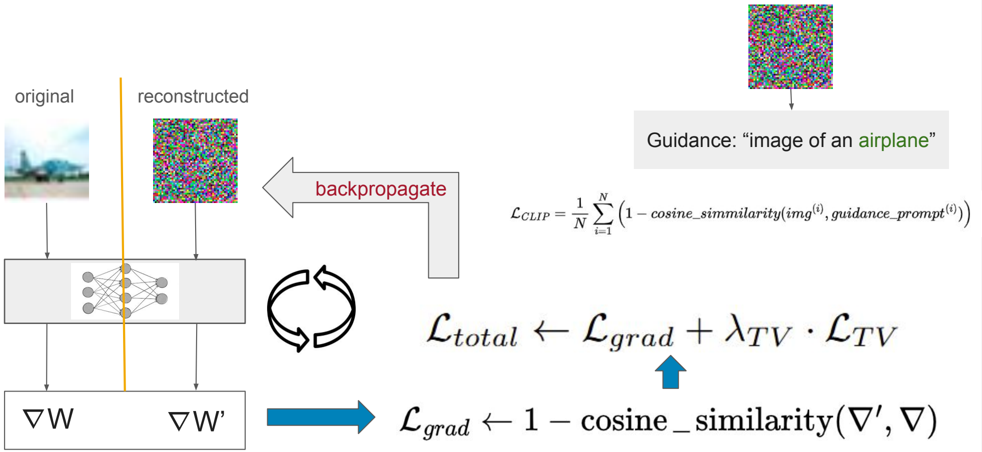


Recovering Semantic Labels from CLIP Guidance



Repeat for each possible guidance and report one that leads to lowest loss

Thank you!



Semantic Label Reconstruction

How to Breach Privacy in Federated Learning

Summary

- Gradients leak semantic information
- Increasing batch size is an effective defensive tool
- Aggregation across multiple batches improves label recovery

References

[1] Zhu, Ligeng, Zhijian Liu, and Song Han. “Deep Leakage from Gradients.” arXiv, December 19, 2019.

[2] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020

[3] Geiping, Jonas, Hartmut Bauermeister, Hannah Dröge, and Michael Moeller. “Inverting Gradients -- How Easy Is It to Break Privacy in Federated Learning?” arXiv, September 11, 2020.

[4] Jeon, Jinwoo, jaechang Kim, Kangwook Lee, Sewoong Oh, and Jungseul Ok. “Gradient Inversion with Generative Image Prior.” In Advances in Neural Information Processing Systems, 34:29898–908. Curran Associates, Inc., 2021.

[5] Zhao, Bo, Konda Reddy Mopuri, and Hakan Bilen. “iDLG: Improved Deep Leakage from Gradients.” arXiv, January 8, 2020.

[6] Ma, K., Sun, Y., Cui, J., Li, D., Guan, Z., & Liu, J. (2023). Instance-wise batch label restoration via gradients in federated learning. In Proceedings of the International Conference on Learning Representations (ICLR 2023).

[7] Fang, Hao, Bin Chen, Xuan Wang, Zhi Wang, and Shu-Tao Xia. “GIFD: A Generative Gradient Inversion Method with Feature Domain Optimization.” In Proceedings of the IEEE/CVF International Conference on Computer Vision, 4967–76, 2023.