

LIDLAW SCHOLARS LEADERSHIP AND RESEARCH PROGRAM

Research Report

Can AI Still Learn When Data Cannot Be Shared?

Learning without sharing private data

Dimitri Urbain

Scalable Computing Systems Laboratory

École Polytechnique Fédérale de Lausanne

October 2026



Abstract

Training a powerful AI model requires diverse training examples and substantial computing resources. Public datasets provide broad coverage, while private records capture cases specific to an institution and help address gaps in public data. When local computing capacity is insufficient, renting external servers provides additional resources, but conventional training requires transferring the training data to those servers. This creates a barrier when private records must remain within the institution, preventing both sources from being brought together externally. We study a hybrid approach in which two models of the same architecture train in parallel: one learns from public data on external servers, while the other learns from private data locally. The private environment periodically imports the public model and averages its learned parameters with those of the private model, combining their learning without sending private data or training updates outside. In an image-classification experiment, the hybrid approach raises private test accuracy from 22.08% with public-only training to 87.54%, while accuracy across the combined public and private test images increases from 80.03% to 81.88%. Further experiments show that the gains are strongest when private examples fill gaps in public coverage, and that the averaging weights control the balance between private and overall performance. These results demonstrate that combining external public training with local private learning can substantially improve performance on locally important cases while preserving overall accuracy.

1 Introduction

An AI model needs a large number of training examples that cover the situations it will encounter in use. Public datasets provide broad coverage, but can leave gaps for a particular institution, whose private records contain examples of the cases it actually handles. Conventionally, these complementary sources would be brought together to train a single model. However, training at this scale requires powerful computing infrastructure that the institution may not have in-house, leading it to rent external servers instead. Renting provides access to the necessary computing power without purchasing and maintaining the hardware permanently, making it particularly attractive for occasional training workloads. The difficulty is that conventional training on these servers also requires transferring the data to them, which is incompatible with private records that must remain within the institution.

This restriction reflects concerns about who can access and control sensitive information once it leaves the institution. NIST’s guidelines on public cloud computing identify security and privacy challenges involving trust in the provider, data protection and legal compliance [1]. The U.S. CLOUD Act adds a concrete example of the legal considerations: covered providers subject to U.S. jurisdiction can be required, through valid legal process, to disclose data in their possession, custody or control, even when those data are stored outside the United States [2]. In the setting we study, the institution therefore uses external computing resources for public data while keeping private records and the training that uses them on trusted local infrastructure.

A hospital specialising in paediatrics illustrates why both sources matter. Its model needs to work well on children while remaining useful for other patients, yet public medical datasets do not necessarily provide enough examples of the hospital’s main patient population. A review of 203 public medical imaging datasets found that children represented less than 2% of patients in datasets with available age information [3]. The hospital’s own records can fill this gap, but their sensitivity prevents simply transferring them to the external servers. Keeping all training within the hospital would avoid that transfer yet would bring back the original difficulty: obtaining the computing power needed to train on both sources locally.

The simplest option is therefore to train externally using public data alone, respecting the privacy boundary but leaving out the local examples that could address these gaps. As a result, the model can remain poorly adapted to cases that are particularly important to the institution. Fine-tuning offers a way to include them by taking a publicly trained model and continuing its training on private data locally. Although this can improve performance on the institution’s own cases, adapting a model to one distribution can also reduce its performance on others [4]. For the hospital, becoming better at handling children is therefore only part of the objective, since the model must also remain useful for other patients.

Federated learning offers another way to combine learning while keeping the underlying data on participating devices. In standard FedAvg, the server aggregates locally learned updates and redistributes the resulting model to the participants for the next round of training [5]. Keeping raw records local does not, however, prevent information about them from being exposed: gradient inversion attacks have demonstrated that training examples can be reconstructed from shared gradients under certain conditions [6, 7]. Hosting the aggregation server within the institution would keep private updates inside the trusted environment, but public workers would still receive a combined model whose parameters are influenced by private training. This outward transfer conflicts with our requirement that information learned from private records remain within the institution.

We address this constraint by training public and private models in parallel, periodically importing the public model into the private environment and averaging its learned parameters with those of the local model. The combined model continues learning privately, while external workers train their public model independently without receiving private contributions. We investigate whether this approach improves performance on locally important cases while preserving broader performance. To study this question, we use image-classification benchmarks rather than medical records, varying both the differences between the datasets and the influence given to each model when they are combined.

2 Method and evaluation

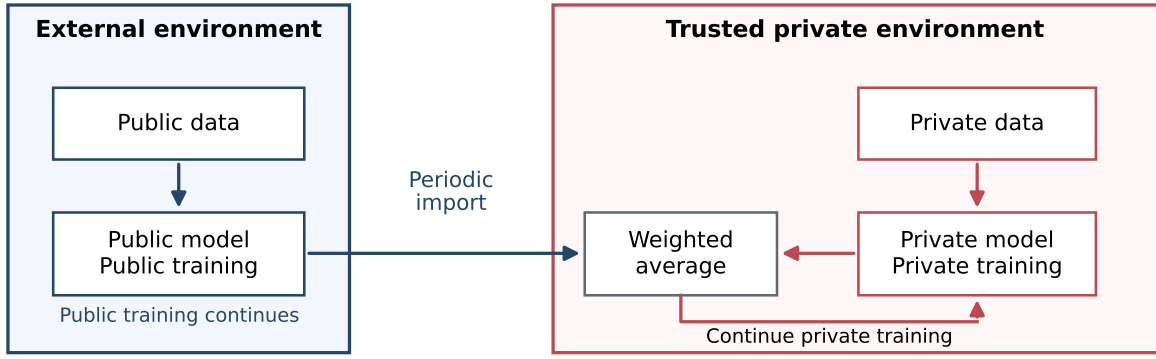
2.1 Combining public and private learning

The method maintains two copies of the same neural network, one in each environment. During training, each copy adjusts its parameters, the numerical weights that determine how it processes an input. Their shared architecture makes it possible to combine corresponding parameters through a weighted average. At regular intervals, the private environment imports the available public model and updates its local model according to

$$\theta_{\text{new}} = \lambda \theta_{\text{private}} + (1 - \lambda) \theta_{\text{public}}.$$

Here, θ denotes the model parameters and λ , pronounced lambda, sets the weight given to the private model. With $\lambda = 0.6$, each new parameter receives 60% of its value from the private model and 40% from the public model. The resulting hybrid resumes private training until the next import. Meanwhile, the public model continues learning exclusively from public data, without receiving any contribution from private training. Repeated imports allow the local model to benefit from further public learning throughout its own adaptation.

This arrangement assumes that the institution has enough local resources to train a full copy of the model on its smaller private dataset, even though training on the combined dataset would require more computing capacity. Processing the larger public dataset externally reduces the work assigned locally. The institution still needs enough memory to hold the private model and its training state, since using fewer examples does not make the model itself smaller.



Public and private models share an architecture. No private training contributions leave the trusted environment.

Figure 1. Public training uses only public data and receives no private contributions. The private environment periodically imports the public model, averages its parameters with the local model and continues private training.

Figure 1 shows how this communication boundary is maintained. Averaging takes place inside the trusted private environment, where private examples, gradients and model updates remain throughout training. The final hybrid model also stays local. The external system therefore receives no private training contributions through this protocol, provided that the local environment remains trusted and the hybrid model is not subsequently released externally.

2.2 Constructing complementary datasets

We evaluate the method on CIFAR-10, which contains small colour images from ten categories, including animals and vehicles [8]. The task is to predict the category of each image. Public and private datasets are constructed from separate subsets of these benchmark images, with some categories made scarce publicly but common privately. This recreates the coverage gap motivating the hospital example without using confidential medical records.

Two quantities control the split. The private fraction p sets the proportion of retained training images assigned privately, with values of 5%, 15% and 25%. The parameter α controls the difference between the category proportions in the two datasets: $\alpha = 0.1$ produces strongly different distributions, $\alpha = 0.5$ an intermediate difference, and $\alpha = 5$ more similar distributions. A Dirichlet-based allocation procedure generates these proportions, then retains as many images as possible while respecting the target private share and the available examples in each category.

Some images remain unused to preserve the intended imbalance, since assigning every remaining example publicly would refill the coverage gaps being studied. Figure 2 shows the actual training split for $\alpha = 0.1$ and $p = 0.15$. Dogs and trucks together account for approximately 96% of the private examples but less than 1% of the public examples, making the two sources strongly complementary.

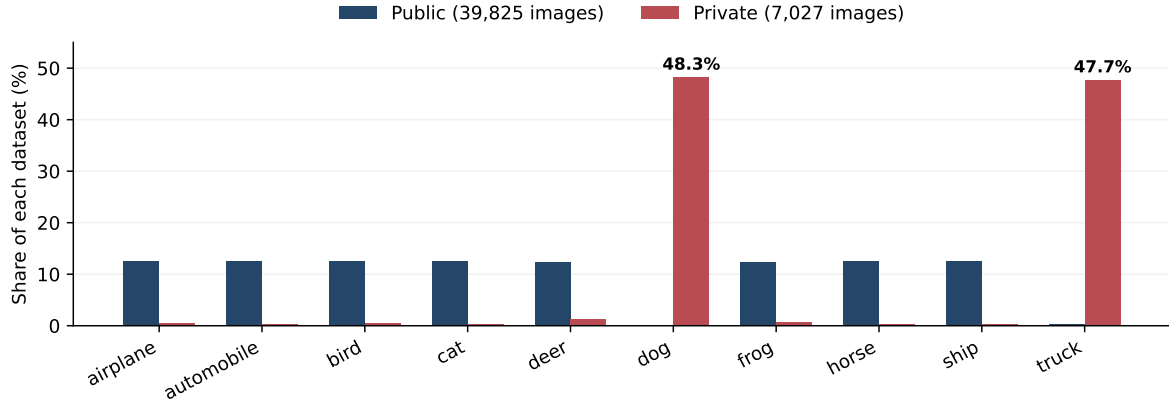


Figure 2. Actual CIFAR-10 training split for $\alpha = 0.1$ and $p = 0.15$ (seed 1236): 39,825 public images, 7,027 private images and 3,148 unused images. Percentages are calculated within each dataset. Dogs and trucks dominate the private dataset but are rare publicly.

2.3 Training configuration

We use LeNet and ResNet-20 [9], two image-classification networks. Training is distributed across 13 public and 3 private processes, each working on part of its environment’s data. After every pass through their assigned images, the public processes produce a common public model and the private processes produce a common private model. The private environment additionally imports the public model after every four private passes. Each pass is called an epoch.

ResNet-20 trains for up to 160 public and 120 private epochs; LeNet uses 96 and 72. Both use stochastic gradient descent with batches of 16 images and public/private learning rates of 0.05/0.03 for ResNet-20 and 0.03/0.018 for LeNet. The ResNet-20 public learning rate decreases tenfold halfway through training and again at three quarters of its budget. Evaluation occurs every eight private epochs. We compare constant mixing weights of 0.4, 0.5 and 0.6, with $\lambda = 0.6$ retained for the main comparisons. The training and data-split seeds are 1234 and 1236; configurations have not been repeated across multiple seeds. All results discussed here concern CIFAR-10.

2.4 Metrics and model selection

Accuracy measures the percentage of test images classified correctly. Private accuracy measures performance on locally important cases, while overall accuracy measures performance across the retained public and private test images. These test subsets are separate from the training images and follow their corresponding distributions. The combined score weights each group by its number of test images, so a private share near 15% gives private performance approximately 15% of the overall weight.

Overall accuracy is labelled Retained in the experiment records. It represents performance when the data encountered in use follow the constructed training distribution, excluding unused images. Full accuracy instead uses the original balanced CIFAR-10 test set, giving equal weight to all ten classes for a broader evaluation. Because the class proportions differ, Full and Retained measure performance on different populations and must be reported separately. Public accuracy additionally reveals losses outside the private group.

We compare the hybrid with public-only training, private fine-tuning and an oracle reference. Fine-tuning starts from the selected public-only model and continues learning on private data

locally. The oracle trains on the union of the retained public and private datasets, providing a joint-training comparison without the separation constraint.

For all four methods, we select the recorded checkpoint with the highest Full accuracy, using the same criterion to assess general performance across the ten classes. Private, public and overall scores are then reported at that checkpoint, which may occur at a different stage for each method. Results explicitly described as final are measured at the end of training. These exploratory comparisons use test results for checkpoint selection; future evaluations will use a separate validation set for selection.

3 Results

3.1 Improving private accuracy while preserving overall performance

The main ResNet-20 experiment uses strongly different datasets, a 15% private share and a private mixing weight of 60% ($\alpha = 0.1$, $p = 0.15$, $\lambda = 0.6$). The public-only model reaches 90.25% accuracy on public test images but only 22.08% on private images. Its overall accuracy nevertheless remains 80.03%, illustrating how the larger public group can hide a serious gap in local performance.

The hybrid raises private accuracy to 87.54%, a gain of 65.46 percentage points, while public accuracy falls to 80.88%. The private improvement more than offsets this public loss in the combined test set, bringing overall accuracy to 81.88%, an increase of 1.85 points. The result handles locally important cases much better while slightly improving the combined score.

Figure 3 compares this result with fine-tuning and joint training, rounding scores to whole percentages for readability. Fine-tuning reaches 98% private accuracy but only 58% overall, compared with 88% and 82% for the hybrid. The oracle reaches 84% privately and 89% overall, giving it the strongest combined performance among these configurations. The hybrid offers a substantially better balance than the tested fine-tuning configuration, while retaining an approximately seven-point overall gap to joint training.

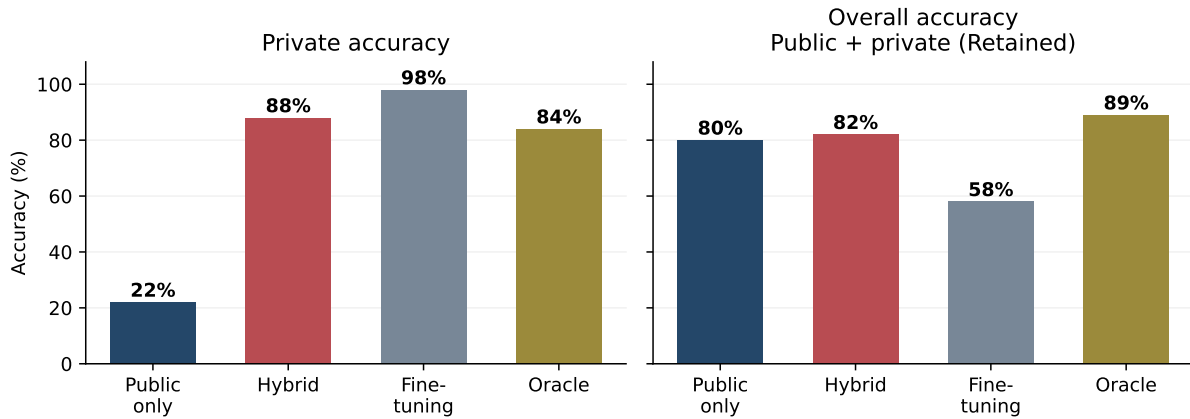


Figure 3. Comparison with public-only training, private fine-tuning and joint training (oracle). Overall accuracy measures the retained public and private test images. Values are rounded to whole percentages.

A complementary analysis at the end of this run shows how adaptation affects individual categories (measured on the full test set). Accuracy on dogs rises from 0% with public-only training to 83.3% with the hybrid, and accuracy on trucks rises from 38.6% to 94.7%. These are the categories concentrated in the private dataset, while accuracy on cats falls from 82.0% to

38.7%. The local gains therefore come with losses elsewhere, which the combined score alone cannot fully describe.

3.2 When private examples fill a gap in public coverage

To examine when private learning contributes most, we vary α while fixing $p = 0.15$ and $\lambda = 0.6$. Table 1 shows that the private gain reaches 65.46 points when the distributions differ strongly, but falls to 6.97 points at $\alpha = 0.5$, where overall accuracy also declines. At $\alpha = 5$, the public-only model already performs well on private cases and the hybrid improves neither private nor overall accuracy.

Table 1. ResNet-20 on CIFAR-10 with $p = 0.15$ and the retained mixing weight $\lambda = 0.6$. Smaller α means stronger distribution differences. Overall accuracy refers to Retained. All scores use checkpoints selected on Full, as described in Section 2.4.

α	Private accuracy (%)		Overall accuracy (%)	
	Public only	Hybrid	Public only	Hybrid
0.1	22.08	87.54	80.03	81.88
0.5	81.23	88.20	86.27	84.80
5	87.94	86.78	88.04	87.83

LeNet shows the same dependence on complementary data. With $\alpha = 0.1$ and $\lambda = 0.6$, private accuracy rises from 9.40% to 79.99%, while overall accuracy remains almost unchanged, moving from 68.95% to 68.78%. At $\alpha = 0.5$, private accuracy rises from 59.34% to 74.18%, but overall accuracy falls from 72.40% to 71.07%. With similar distributions, both changes are below one percentage point. The largest private gains thus occur where public training leaves the largest gaps, although they do not always increase the combined score.

3.3 The effect of the private data share

The balance also changes with the private share. Keeping $\alpha = 0.1$ and $\lambda = 0.6$, ResNet-20 gains 70.57 private accuracy points at $p = 0.05$ but loses 9.57 points overall. At $p = 0.15$, the gains are 65.46 private points and 1.85 overall points; at $p = 0.25$, they reach 73.60 and 9.23 points, respectively. Differences are calculated from the unrounded scores.

Private performance improves substantially in all three settings, but the improvement does not grow consistently with p . Changing the private share changes both the training datasets and the weight of private examples in the combined test score. These comparisons therefore describe different public-private settings, rather than the effect of simply adding private examples to an otherwise unchanged experiment. Overall performance improves at the two larger shares, where local cases contribute more to the combined score.

3.4 Choosing a constant mixing weight

We hypothesise that giving the private model more weight will strengthen local adaptation, potentially at the expense of broader performance. To examine this, λ remains constant throughout each run and is set empirically to 0.4, 0.5 or 0.6 in separate experiments. These values compare different balances between public and private learning; they do not establish a universally optimal weight.

For ResNet-20 with $\alpha = 0.1$ and $p = 0.15$, increasing λ from 0.4 to 0.6 raises private accuracy from 66.31% to 87.54%, while overall accuracy falls from 84.41% to 81.88%. Equal mixing gives

78.85% privately and 83.72% overall. All three settings improve both scores over public-only training in this example, but the strongest private and overall results occur at different weights.

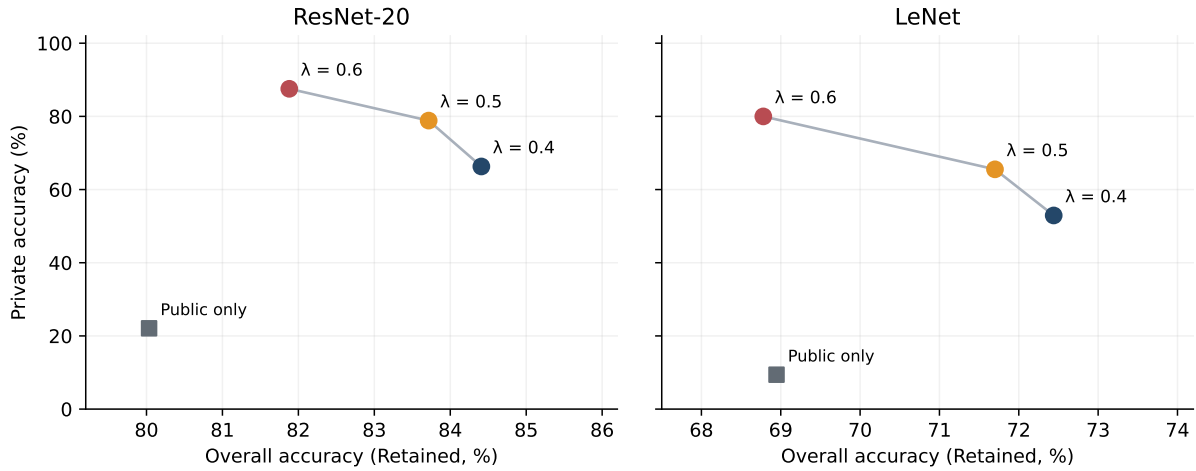


Figure 4. Constant mixing weights on CIFAR-10 with $\alpha = 0.1$ and $p = 0.15$. Increasing λ improves private accuracy but reduces overall accuracy for both architectures. Each point reports scores from one checkpoint selected on Full; lines connect the three tested weights.

LeNet confirms this trade-off for the strongly different datasets shown in Figure 4. Increasing λ from 0.4 to 0.6 raises private accuracy from 52.92% to 79.99%, while overall accuracy falls from 72.44% to 68.78%. Table 2 shows the nine LeNet configurations at $p = 0.15$. The same direction appears at $\alpha = 0.5$, whereas at $\alpha = 5$ the changes are small and greater private influence no longer improves private accuracy. Selecting λ therefore requires considering both the data distribution and the institution’s priorities.

Table 2. Nine LeNet configurations on CIFAR-10 with $p = 0.15$. All scores refer to the hybrid model at its selected Full checkpoint. Overall accuracy refers to Retained; Full uses the original balanced test set.

α	λ	Public (%)	Private (%)	Overall (%)	Full (%)
0.1	0.4	75.88	52.92	72.44	71.02
0.1	0.5	72.79	65.53	71.70	71.32
0.1	0.6	66.80	79.99	68.78	69.39
0.5	0.4	73.17	70.25	72.73	72.56
0.5	0.5	72.29	72.95	72.39	72.83
0.5	0.6	70.52	74.18	71.07	72.49
5	0.4	75.28	75.00	75.24	75.07
5	0.5	75.25	74.80	75.18	75.03
5	0.6	75.23	74.39	75.11	74.92

4 Discussion and next steps

The experiments show that combining public and private learning is most valuable when the two sources are complementary. Private examples then address cases poorly learned publicly, while public imports offer a way to retain knowledge beyond those local cases. When the distributions already resemble each other, the extra private training produces much smaller benefits. The choice of mixing weight matters because stronger local adaptation can reduce public performance even when the combined score improves.

A natural next step is to adapt λ during training, since the relative value of public and private contributions may change as the two models learn. The sensitivity observed with constant weights motivates this hypothesis, but initial dynamic strategies did not produce consistent improvements. Developing an adaptive rule and comparing it with well-chosen constant weights will establish whether it can improve the balance further.

Repeating experiments across multiple seeds and using a separate validation set for model selection will strengthen the evaluation. Comparisons should also include a wider range of fine-tuning configurations and an adaptation of WiSE-FT, which averages a pretrained model with its fine-tuned version [4]. Applied locally, that alternative can respect the same privacy boundary and help isolate the benefit of repeatedly importing a public model that continues learning.

Evaluating training time, local memory requirements and communication costs will establish which workloads benefit from this division of computation. Tests on institutional datasets can then assess whether the controlled coverage gaps studied here reflect practical needs in healthcare, finance and industry, where local populations or operating conditions may differ from publicly available examples.

5 Conclusion

Private data are most useful when they contribute examples that public training fails to cover. In the main experiment, the hybrid improves private accuracy from 22.08% with public-only training to 87.54%, while overall accuracy rises from 80.03% to 81.88%. Across the experiments, however, the benefit depends on the data split and on how much weight private learning receives. Choosing that weight is therefore part of defining the desired balance between local adaptation and broader performance. An adaptive choice during training is a promising direction to investigate, with the aim of improving this balance beyond the constant weights tested here.

Acknowledgements

I thank Prof. Anne-Marie Kermarrec for hosting this research in the Scalable Computing Systems Laboratory (SaCS) at EPFL. I am grateful to Sayan Biswas, Martijn de Vos, Jade Garcia Bourrée, Milos Vujasinovic, and Rafael Pires for their contributions, discussions, and feedback throughout the project. I also thank the Laidlaw Scholars Leadership and Research Program for supporting this research.

References

- [1] Jansen, W., and Grance, T. (2011). Guidelines on Security and Privacy in Public Cloud Computing. NIST Special Publication 800-144. <https://doi.org/10.6028/NIST.SP.800-144>
- [2] U.S. Department of Justice (2019). Promoting Public Safety, Privacy, and the Rule of Law Around the World: The Purpose and Impact of the CLOUD Act. White paper, April 2019. <https://www.justice.gov/criminal/media/999601/dl?inline=>
- [3] Hua, S. B. Z., Heller, N., He, P., Towbin, A. J., Chen, I. Y., Lu, A. X., and Erdman, L. (2026). Underrepresentation of children in public medical imaging datasets. *Nature Health* 1, 888–897. <https://doi.org/10.1038/s44360-026-00111-3>
- [4] Wortsman, M., et al. (2022). Robust Fine-Tuning of Zero-Shot Models. CVPR, 7959–7971. https://openaccess.thecvf.com/content/CVPR2022/html/Wortsman_Robust_Fine-Tuning_of_Zero-Shot_Models_CVPR_2022_paper.html
- [5] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and Agüera y Arcas, B. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. AISTATS, PMLR 54, 1273–1282. <https://proceedings.mlr.press/v54/mcmahan17a.html>
- [6] Zhu, L., Liu, Z., and Han, S. (2019). Deep Leakage from Gradients. *Advances in Neural Information Processing Systems* 32. https://papers.neurips.cc/paper_files/paper/2019/hash/60a6c4002cc7b29142def8871531281a-Abstract.html
- [7] Geiping, J., Bauermeister, H., Dröge, H., and Moeller, M. (2020). Inverting Gradients—How Easy Is It to Break Privacy in Federated Learning? *Advances in Neural Information Processing Systems* 33. https://papers.neurips.cc/paper_files/paper/2020/hash/c4ede56bbd98819ae6112b20ac6bf145-Abstract.html
- [8] Krizhevsky, A. (2009). Learning Multiple Layers of Features from Tiny Images. Technical report, University of Toronto. Dataset documentation: <https://www.cs.toronto.edu/~kriz/cifar.html>
- [9] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. CVPR, 770–778. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html