Mitigating the Impact of Federated Learning on Client Resources

Sebastian Caldas[1], Jakub Konečný[2], H. Brendan McMahan[2], Ameet Talwalkar[1,3]

[1]Carnegie Mellon University
[2]Google
[3]Determined AI

We bring Federated Learning (FL) to heterogeneous edge networks.

FL operates on **users' devices and networks**

FL deals with **more nodes and slower networks** than traditional distributed learning

Communication and computation bottlenecks are exacerbated.

We bring Federated Learning (FL) to heterogeneous edge networks.

FL operates on **users' devices and networks**

FL deals with **more nodes and slower networks** than traditional distribute[d] learning

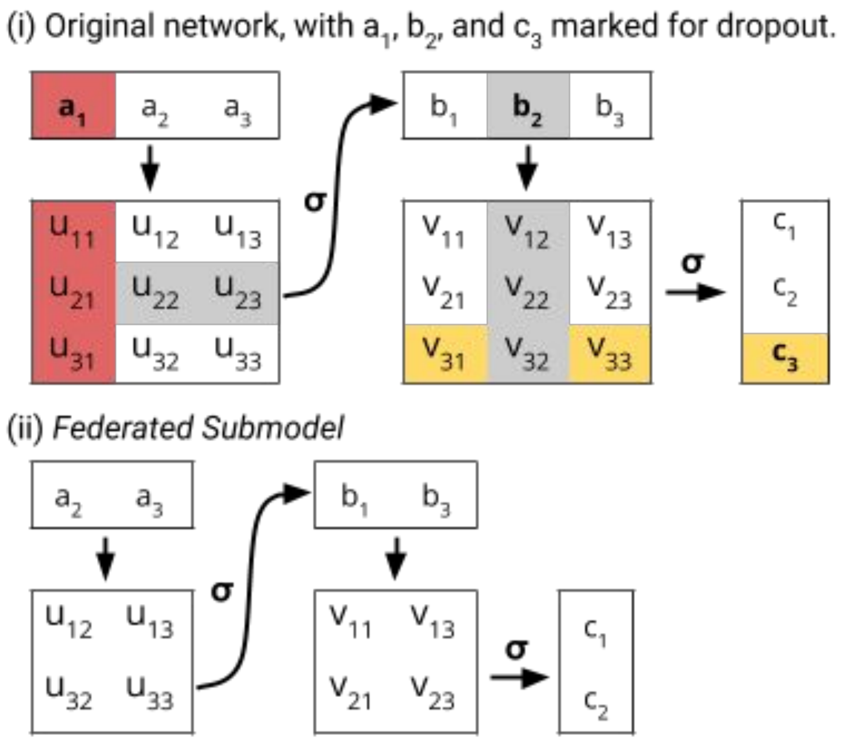
Because resources are distributed unevenly, **certain groups of clients will be systematically excluded**.

We propose strategies that reduce the communication and computation footprint of federated training (FedAvg).

Locally train Federated Submodels, smaller subsets of the full global model.

Lossy compression on the exchanges sent from server-to-client and client-to-server.

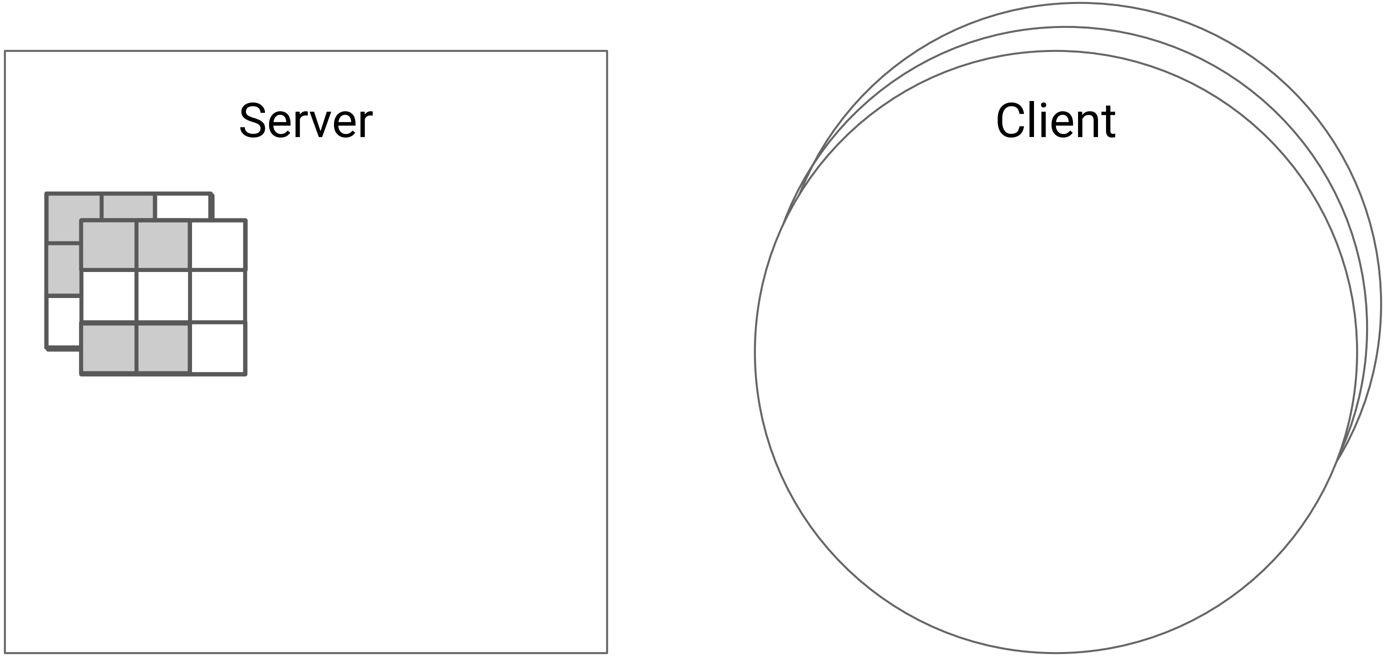
We propose strategies that reduce the communication and computation footprint of federated training.



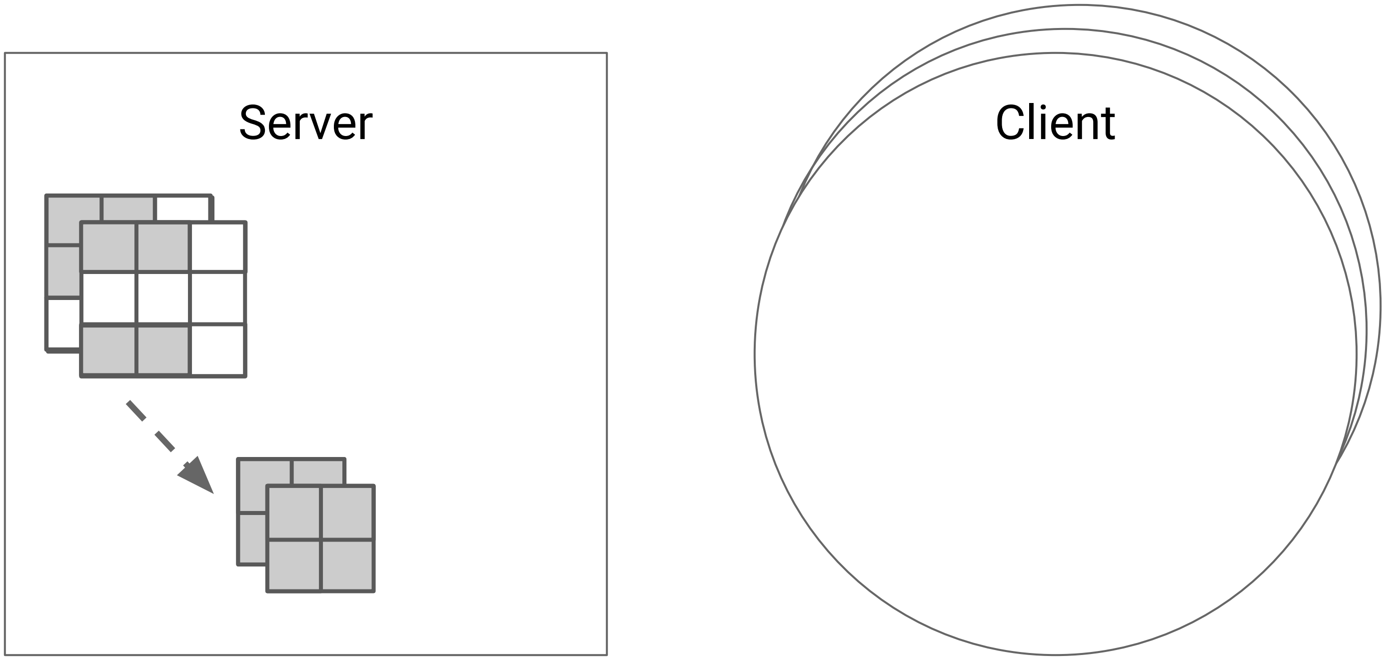
Federated Submodels

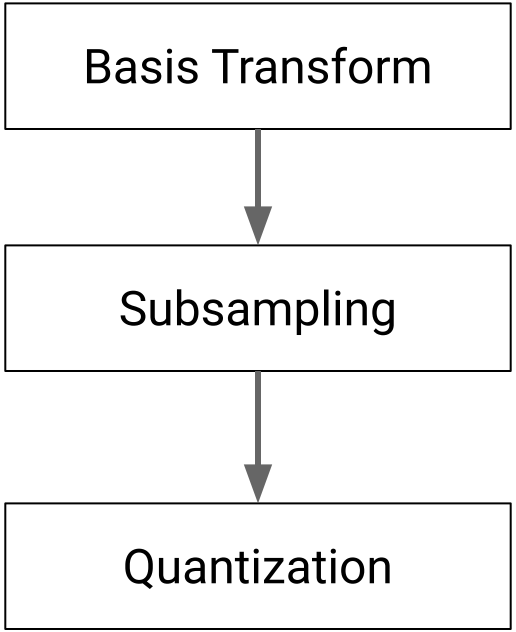
- Each client trains an update to to a subset of the global model.
- For each client, we discard a constant percentage of activations at each fully connected layer.

We propose strategies that reduce the communication and computation footprint of federated training.



We propose strategies that reduce the communication and computation footprint of federated training.

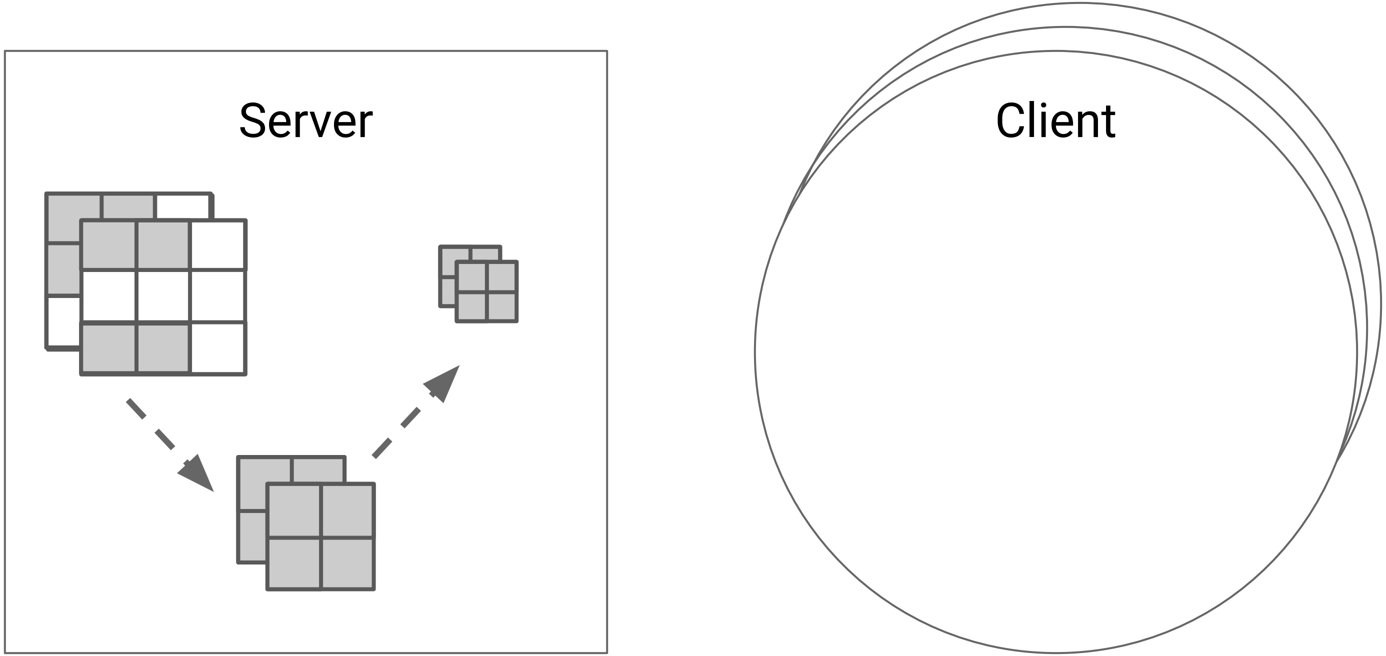




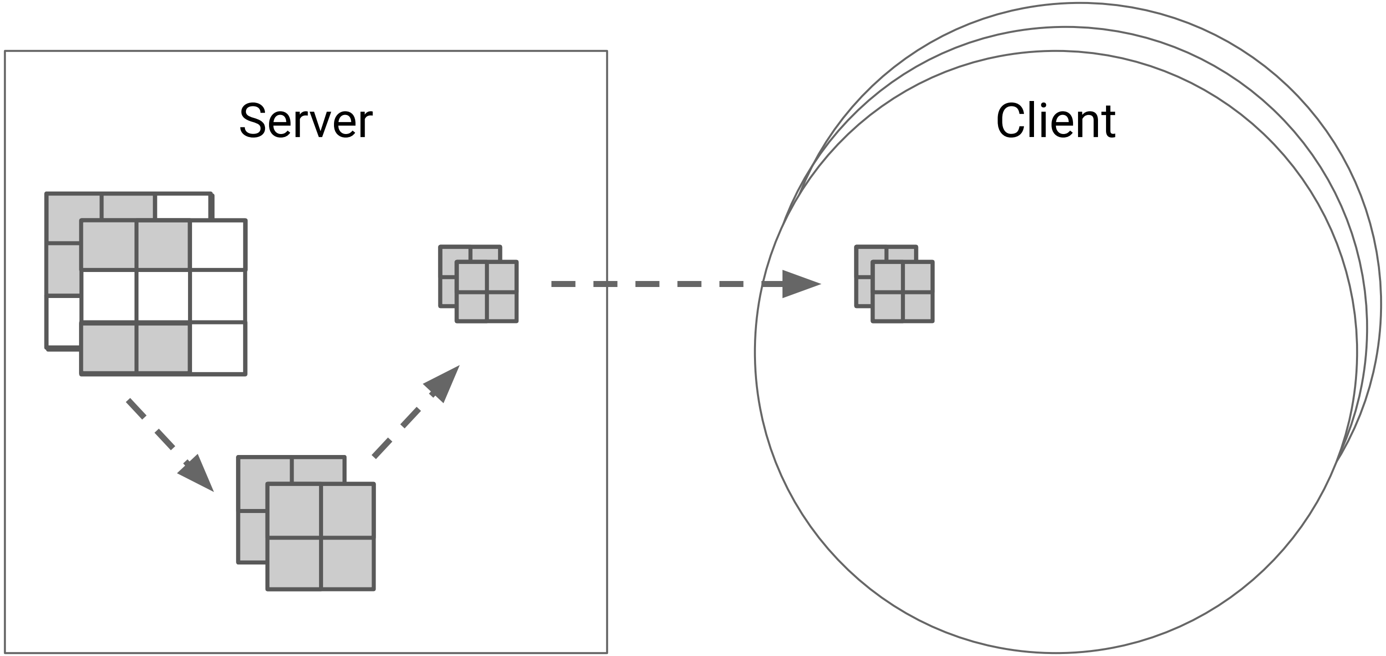
Lossy Compression

- We build upon the work of Konečný et al. (2016), which focuses on compressing gradient updates.
- We use Kashin’s representation to further mitigate the error incurred by subsequent quantization.

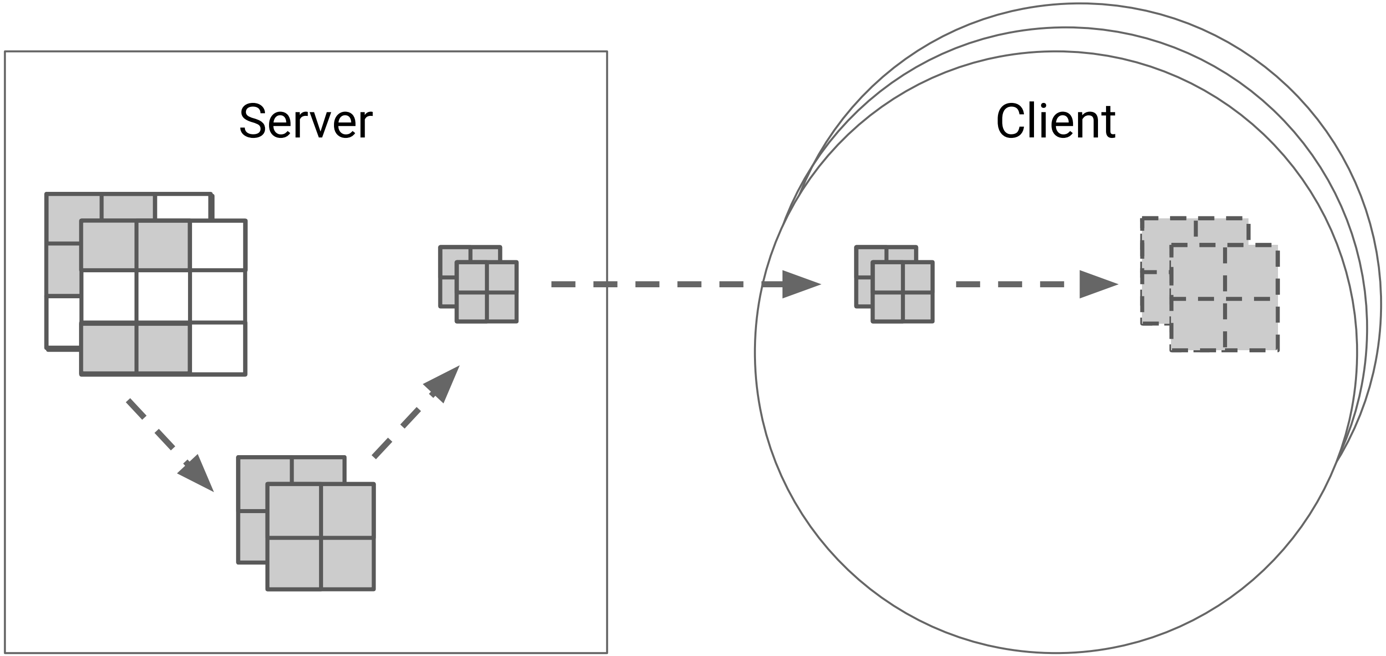
We propose strategies that reduce the communication and computation footprint of federated training.



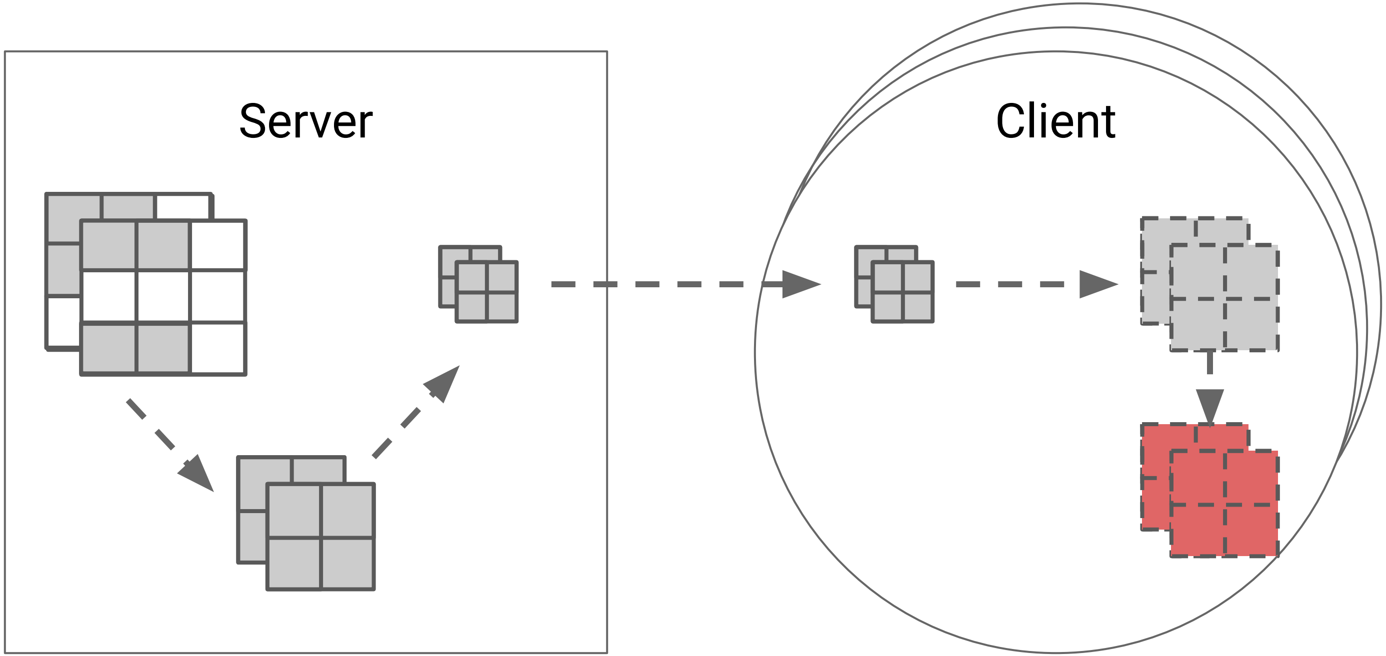
We propose strategies that reduce the communication and computation footprint of federated training.



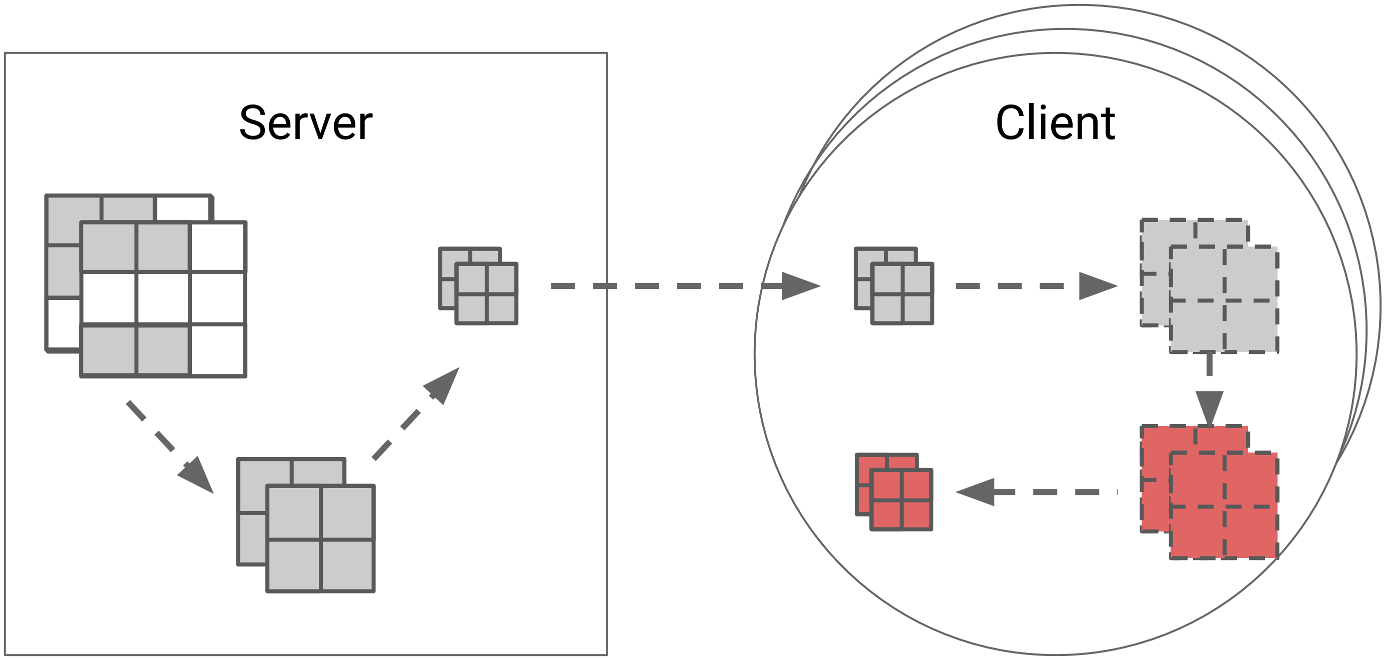
We propose strategies that reduce the communication and computation footprint of federated training.



We propose strategies that reduce the communication and computation footprint of federated training.

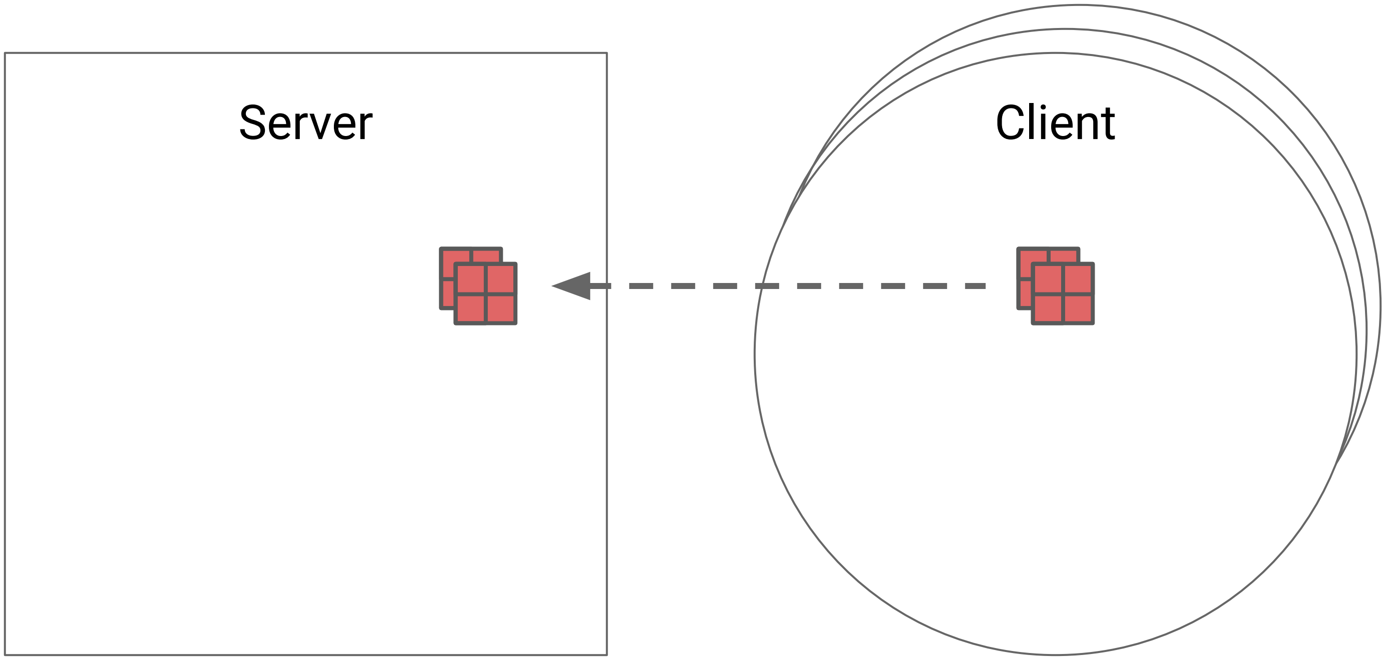


We propose strategies that reduce the communication and computation footprint of federated training.

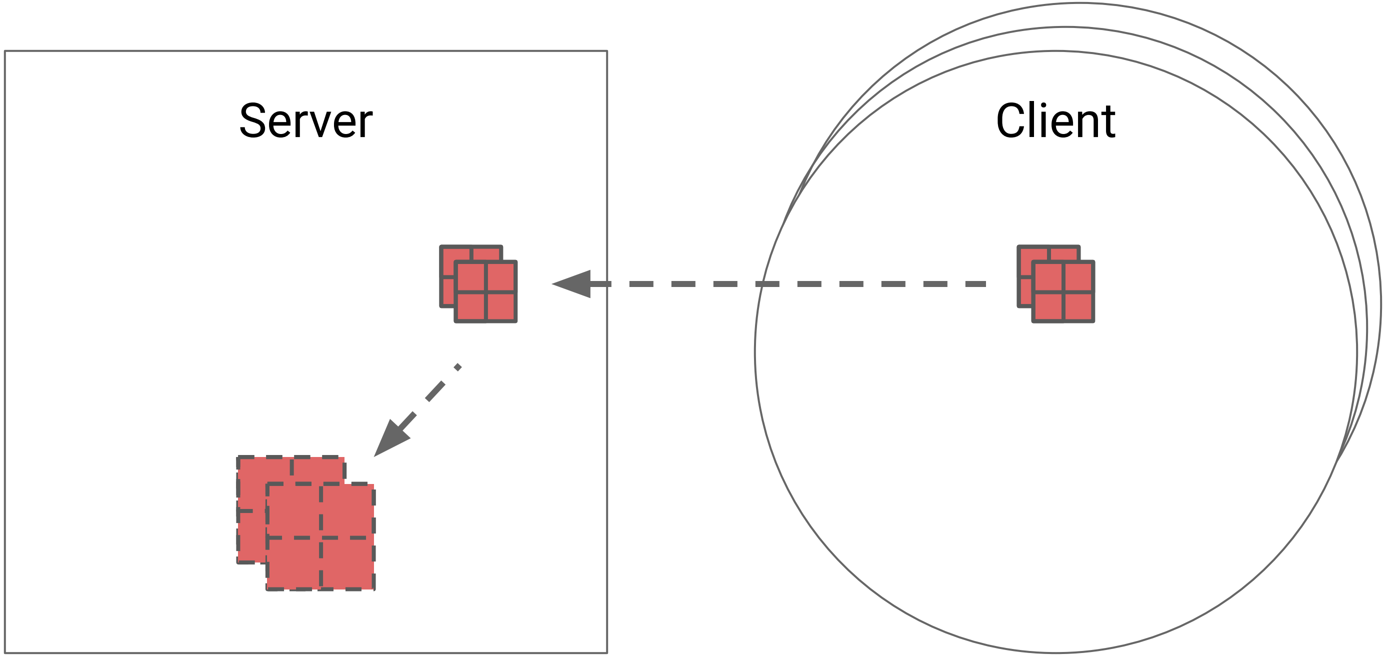


We propose strategies that reduce the communication and computation footprint of federated training.

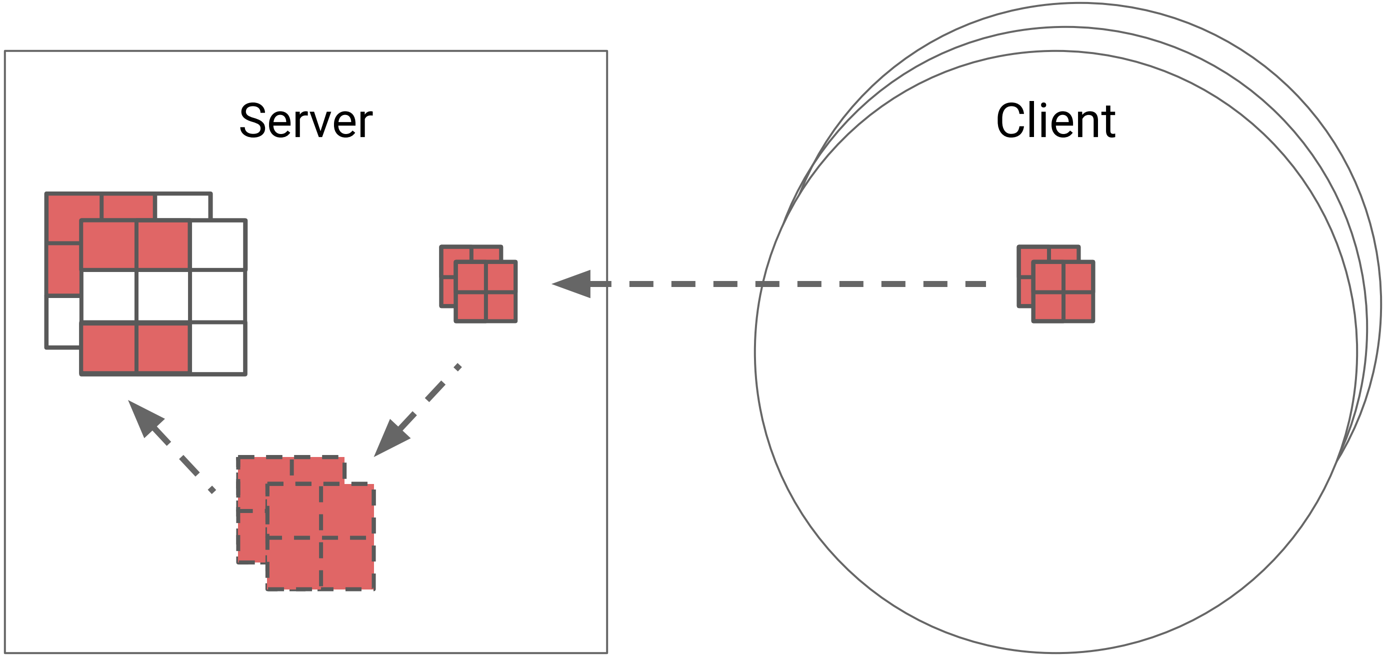
We propose strategies that reduce the communication and computation footprint of federated training.



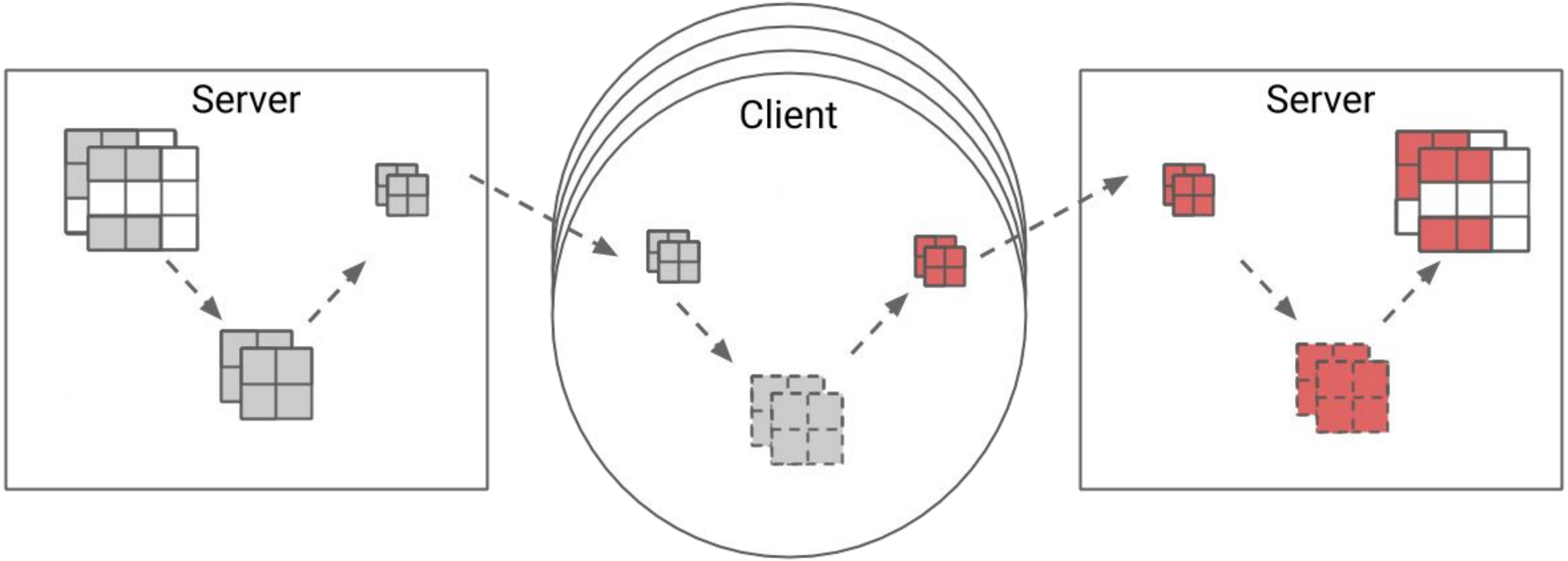
We propose strategies that reduce the communication and computation footprint of federated training.



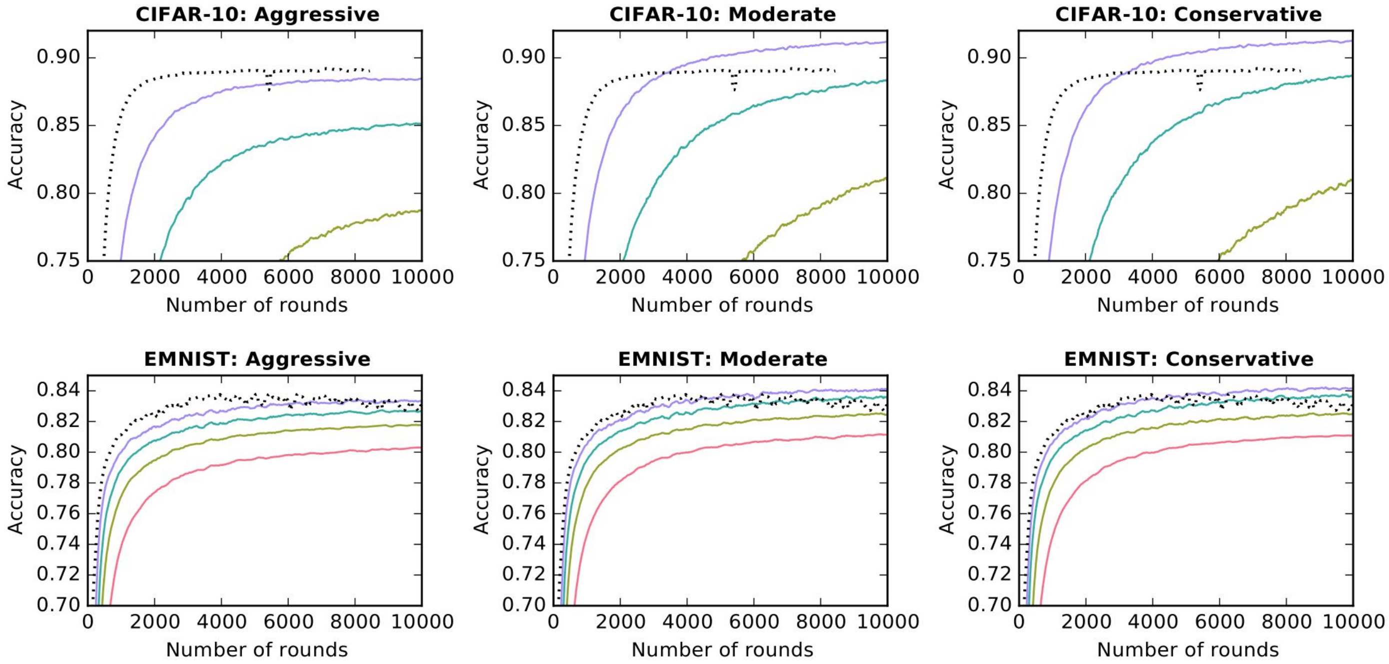
We propose strategies that reduce the communication and computation footprint of federated training.

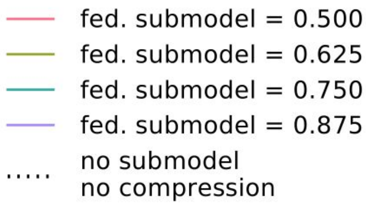


We propose strategies that reduce the communication and computation footprint of federated training.



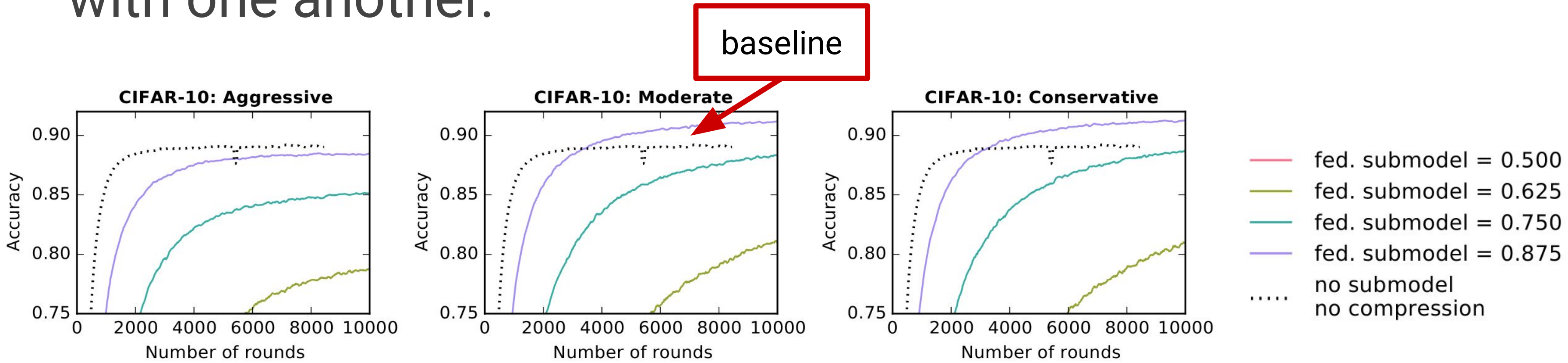
We empirically show that these approaches are compatible with one another.

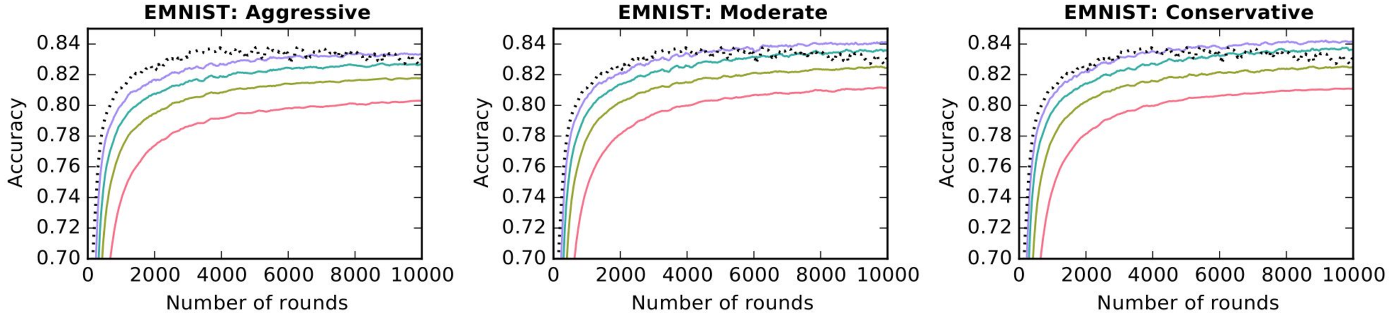




Scheme	Client-to -Server		Server-to -Client
	s	q	q
Aggressive	0.4	2	3
Moderate	0.5	4	5
Conservative	1.0	8	8

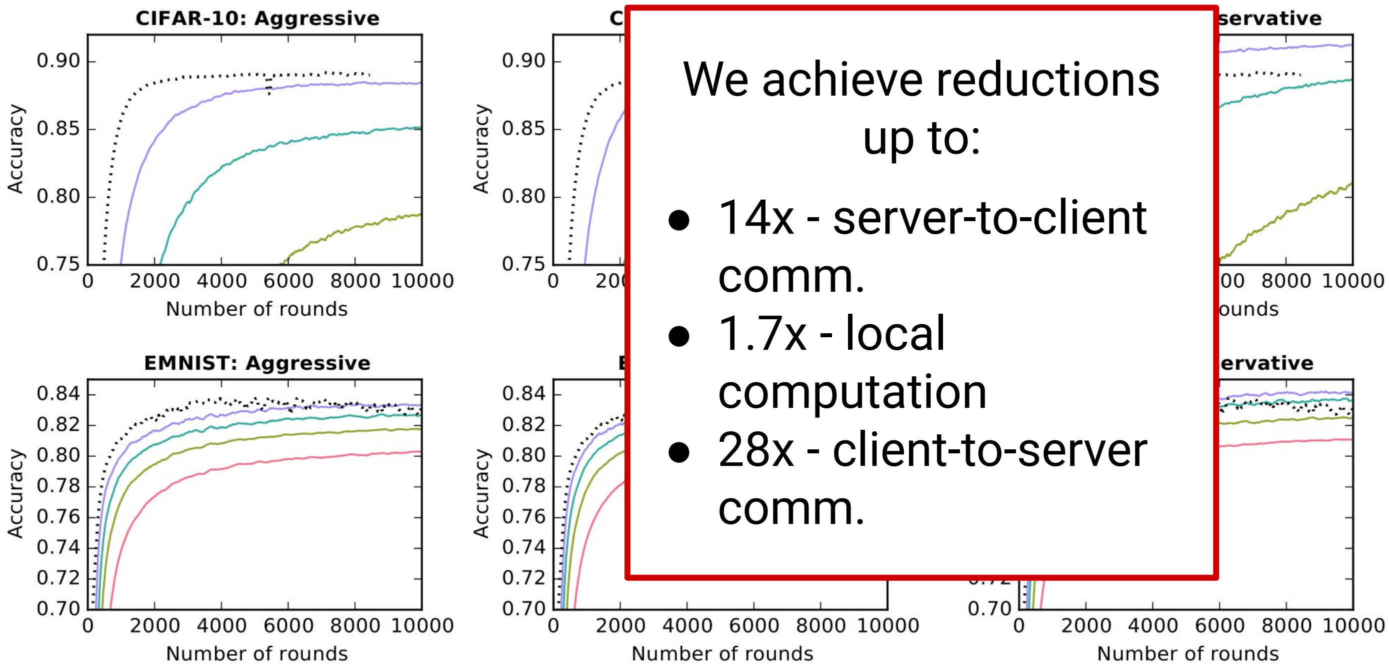
We empirically show that these approaches are compatible with one another.

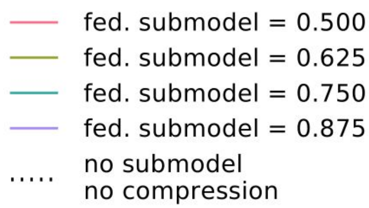




Scheme	Client-to-Server s	Client-to-Server q	Server-to-Client q
Aggressive	0.4	2	3
Moderate	0.5	4	5
Conservative	1.0	8	8

We empirically show that these approaches are compatible with one another.





We achieve reductions up to:

- 14x - server-to-client comm.
- 1.7x - local computation
- 28x - client-to-server comm.

Scheme	Client-to -Server s	Client-to -Server q	Server-to -Client q
Aggressive	0.4	2	3
Moderate	0.5	4	5
Conservative	1.0	8	8

Takeaways

In brief,

- We bring Federated Learning (FL) to realistic heterogeneous edge networks.

- We develop strategies that reduce the communication and computation footprint of any model.
    - Lossy compression
    - Federated Submodels

- We empirically show that these approaches are compatible with one another.

Thank you

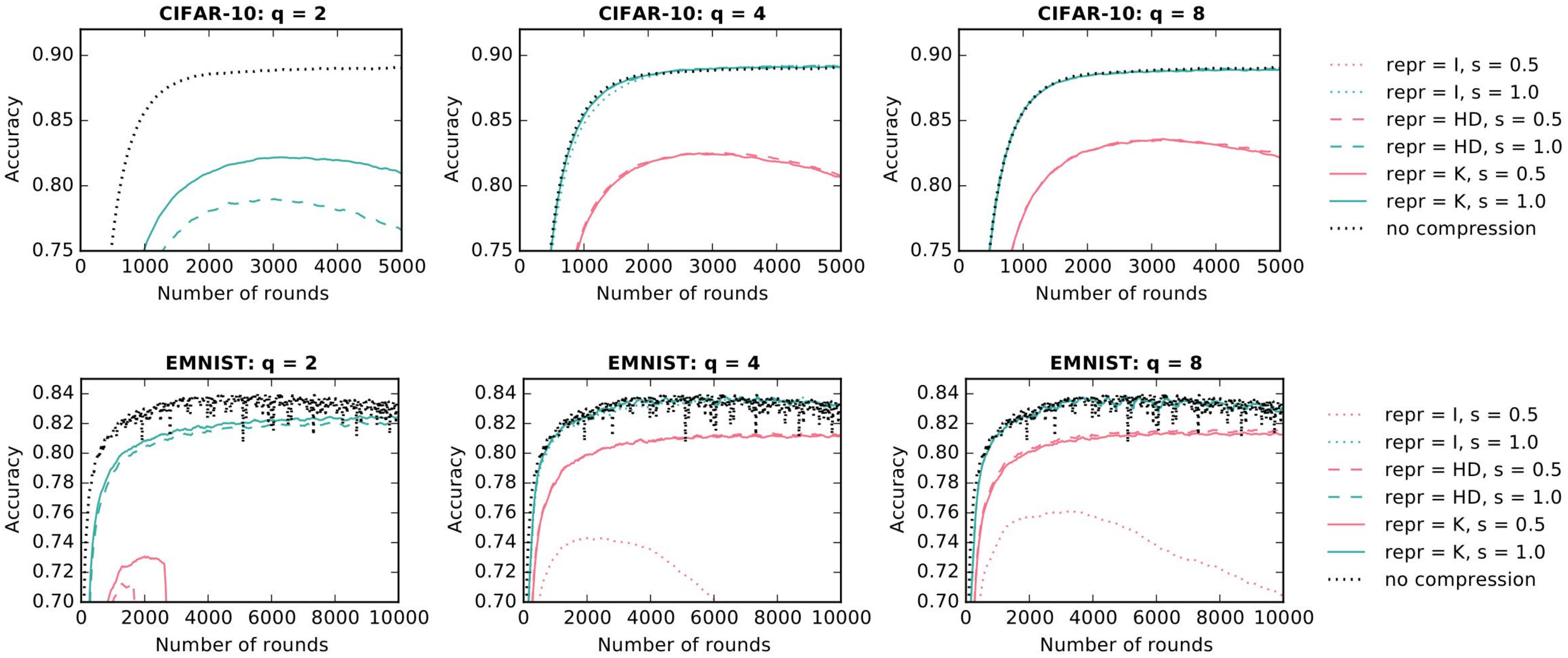
Questions

In brief,

- We bring Federated Learning (FL) to realistic heterogeneous edge networks.
- We develop strategies that reduce the communication and computation footprint of any model.
  - Lossy compression
  - Federated Submodels
- We empirically show that these approaches are compatible with one another.

Additional Slides

Experiments with only lossy compression



Experiments with only Federated Submodels