THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

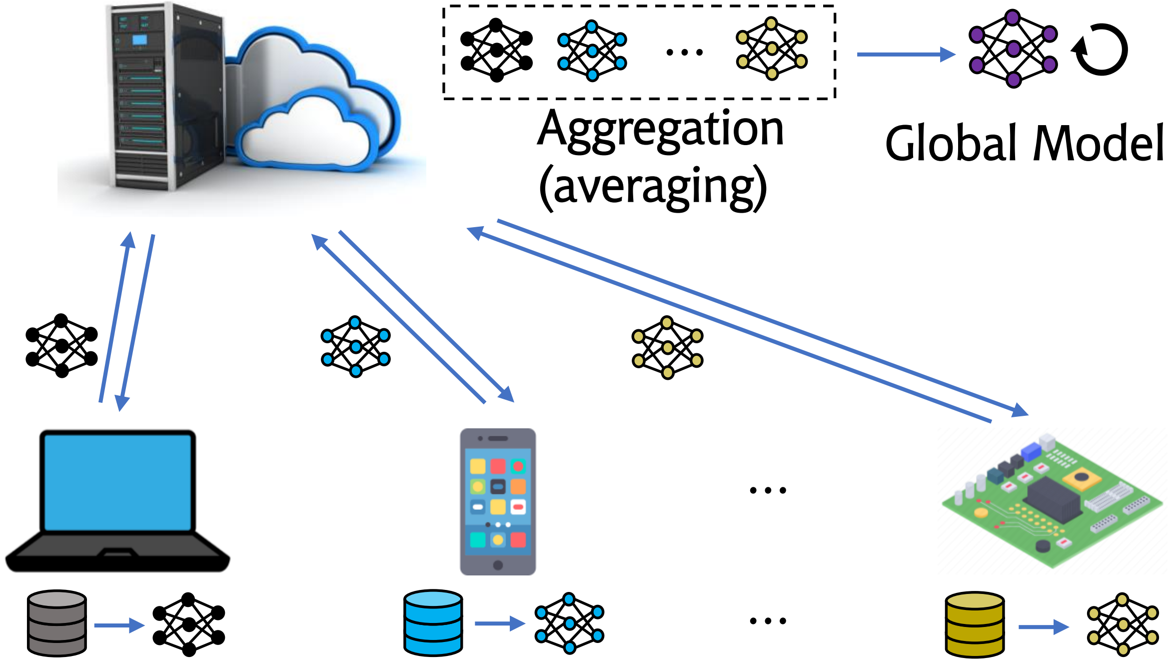
FedConv: A Learning-on-Model Paradigm for Heterogeneous Federated Clients

Leming Shen[1], Qiang Yang[1,2], Kaiyan Cui[1,3], Yuanqing Zheng[1], Xiao-Yong Wei[4,1], Jianwei Liu[5], Jinsong Han[5]

[1]The Hong Kong Polytechnic University, [2]University of Cambridge, [3]Nanjing University of Posts and Telecommunications, [4]Sichuan University, [5]Zhejiang University

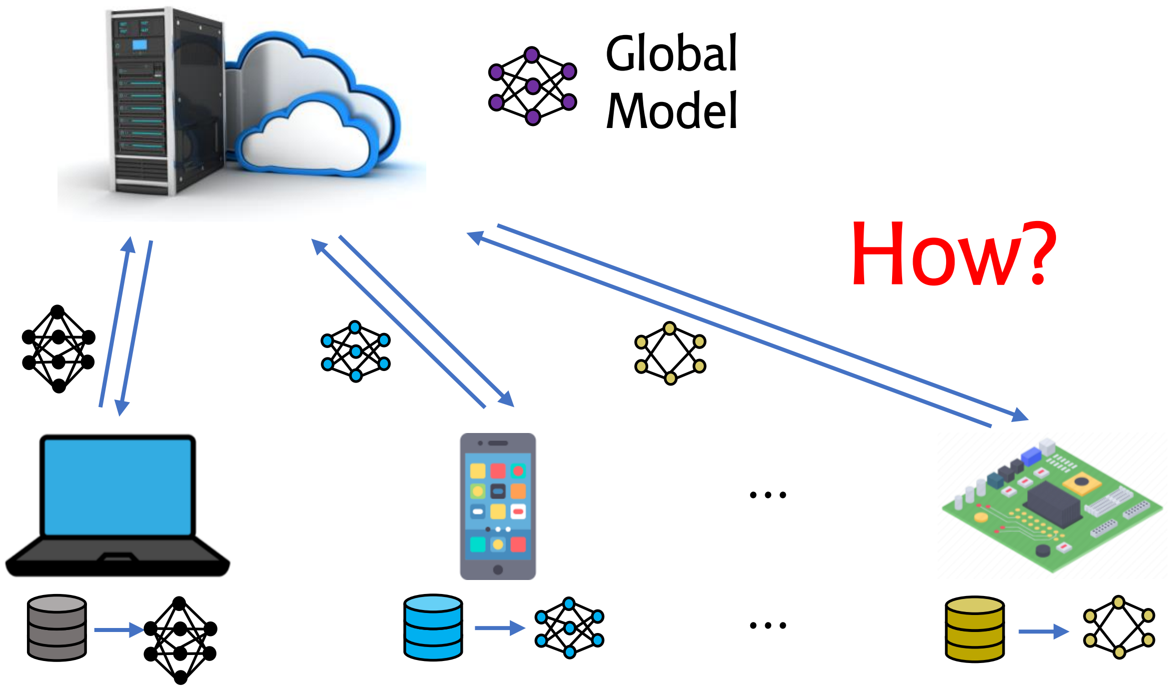
Federated Learning (FL)

- Collaboratively train a global model
- Without transmitting private data

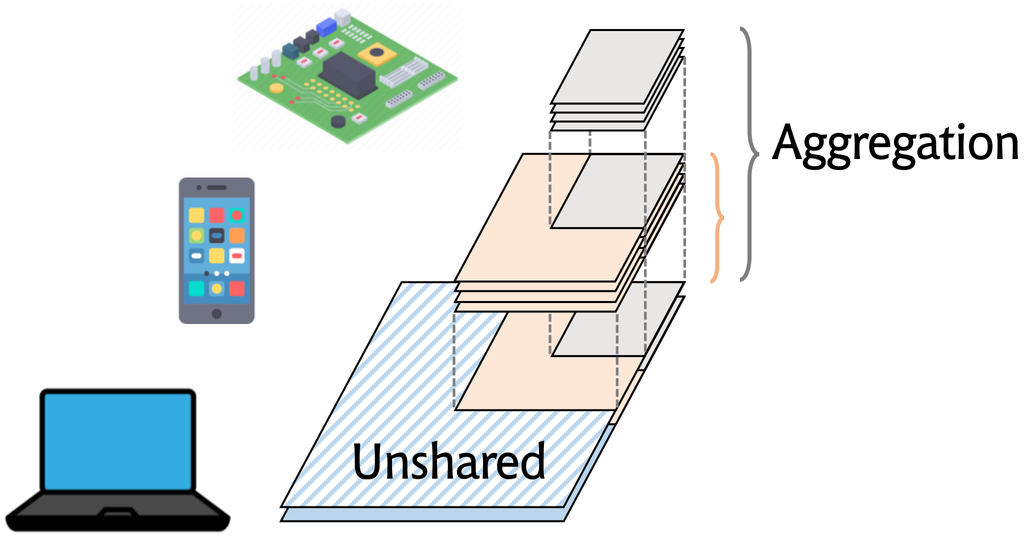


Model Heterogeneity in FL

- Mobile devices have diverse system resources.
- Smallest affordable model → performance ⬇

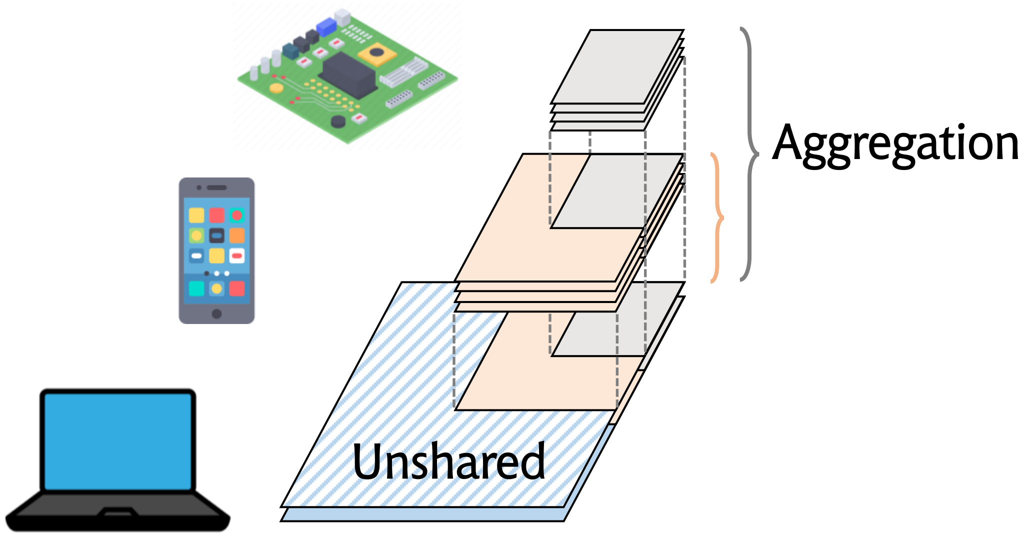


Existing Solution: Parameter Sharing



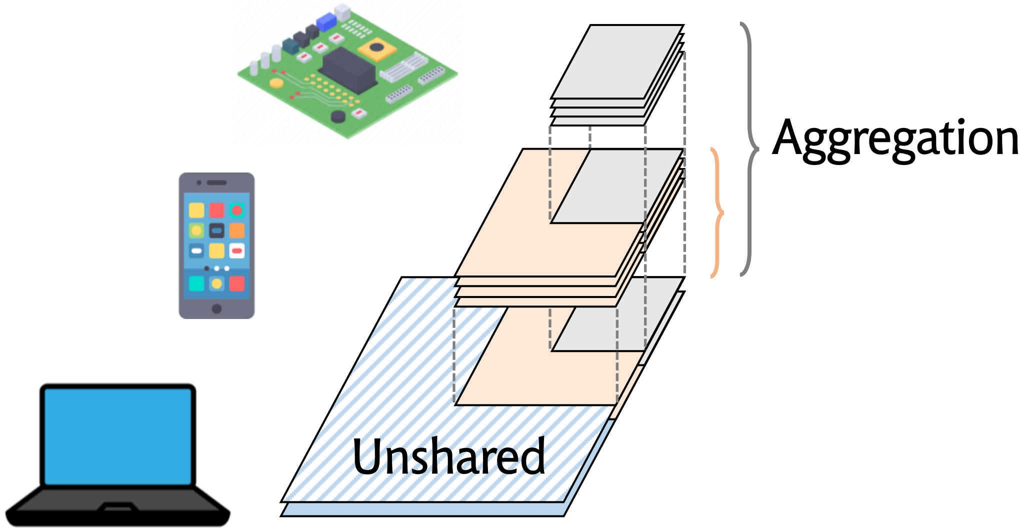
Existing Solution: Parameter Sharing

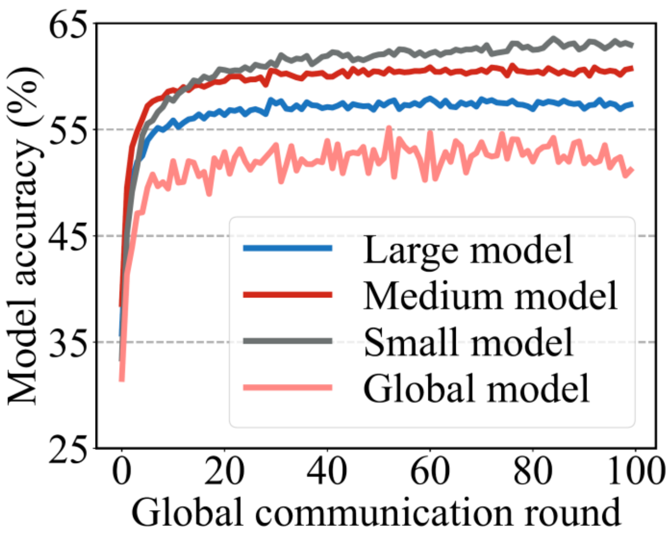
- Imbalanced Training (Fixed sharing portion)
  - Larger models miss the information from other clients.



Existing Solution: Parameter Sharing

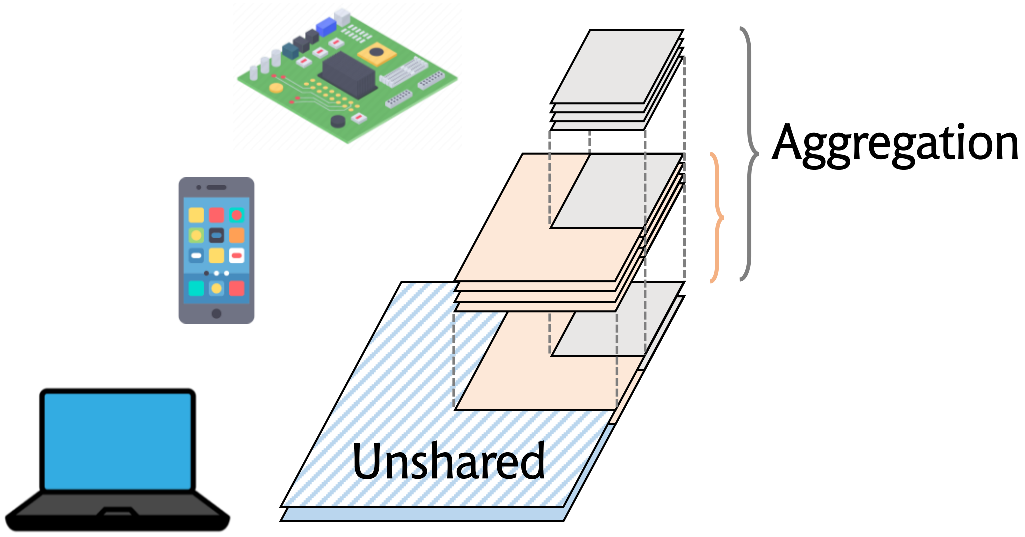
- Imbalanced Training (Fixed sharing portion)
  - Larger models miss the information from other clients.

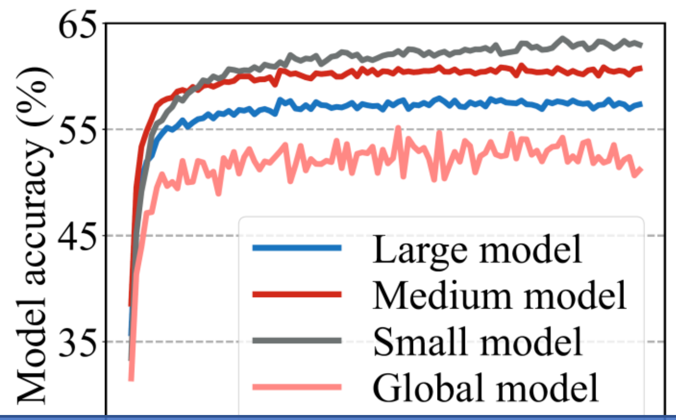




Existing Solution: Parameter Sharing

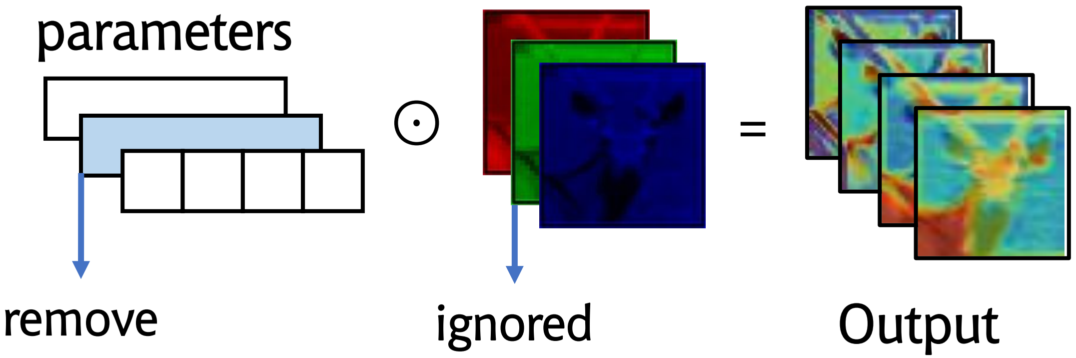
- Imbalanced Training (Fixed sharing portion)
  - Larger models miss the information from other clients.

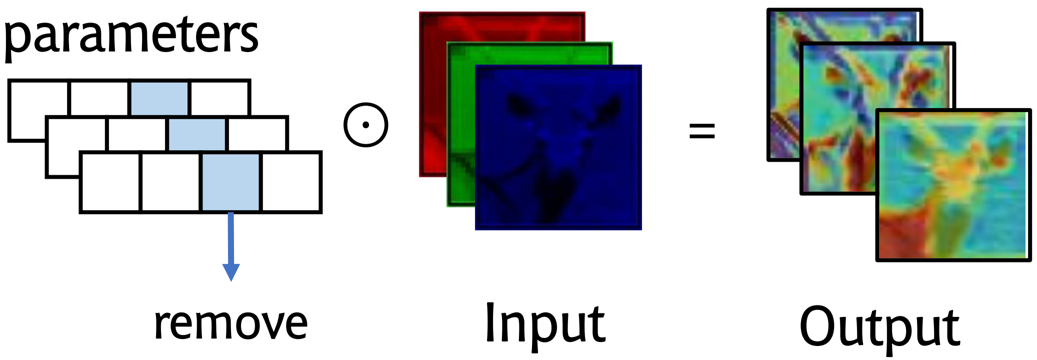




- Smaller models perform better
- The global model exhibits instability and even performs worse

Existing Solutions: Model Pruning





Channel-Level Pruning[1]

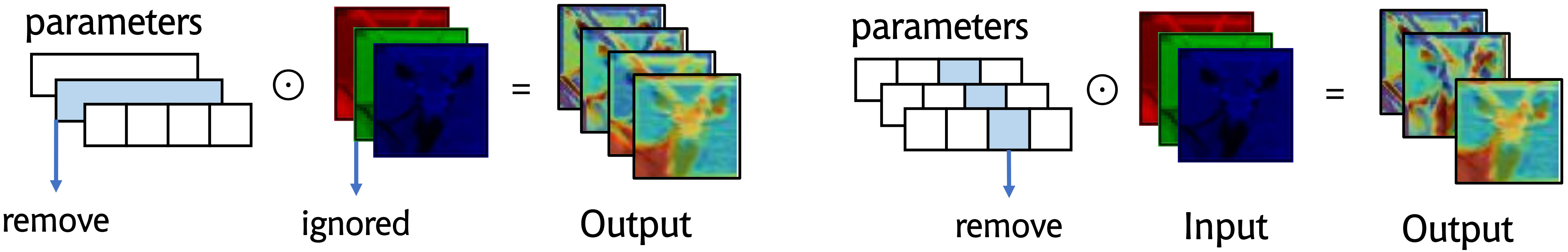
- Remove entire channels
- Less input data

Filter-Level Pruning[2]

- Remove entire filters
- Less output feature maps

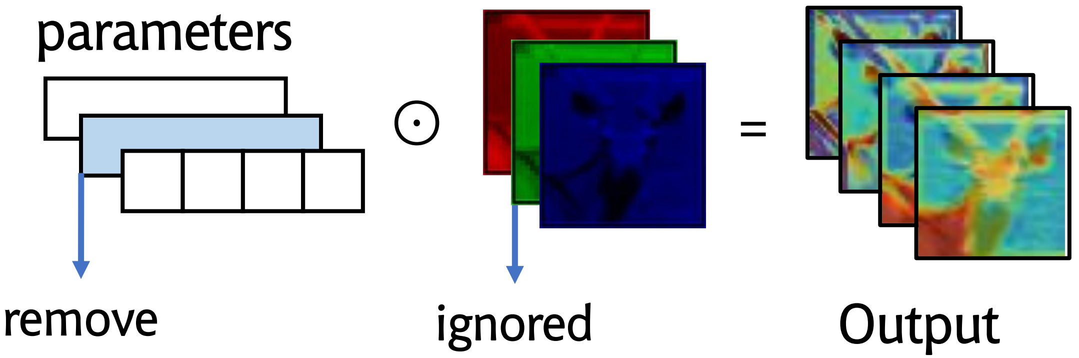
[1] Li, Ang, et al. "Hermes: an efficient federated learning framework for heterogeneous mobile clients." ACM MobiCOM. 2021.
[2] Deng, Yongheng, et al. "Tailorfl: Dual-personalized federated learning under system and data heterogeneity." ACM SenSys. 2022.

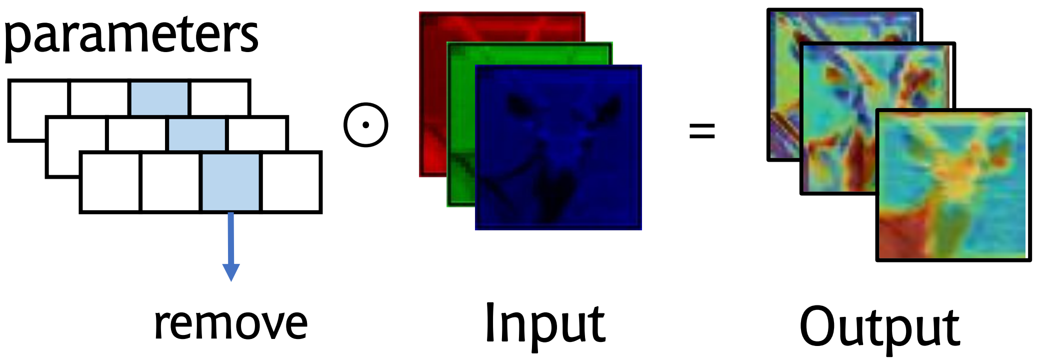
Existing Solutions: Model Pruning



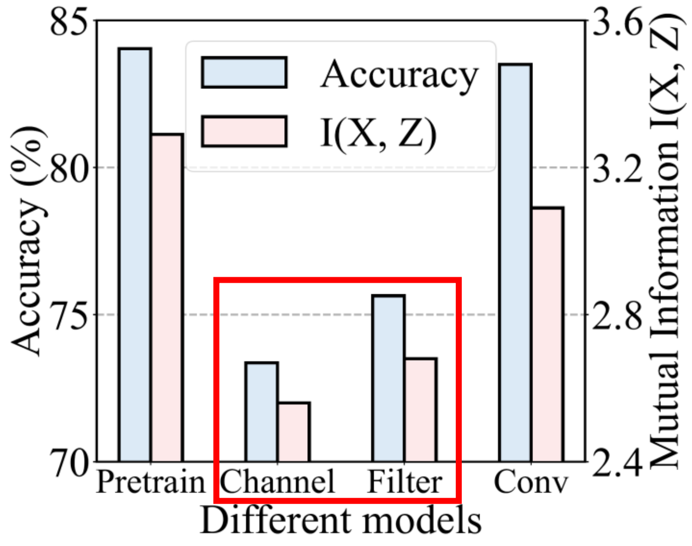
- Information Loss & Extra Overhead
  - Remove entire channels or filters
  - Pruning performed by the client

Existing Solutions: Model Pruning

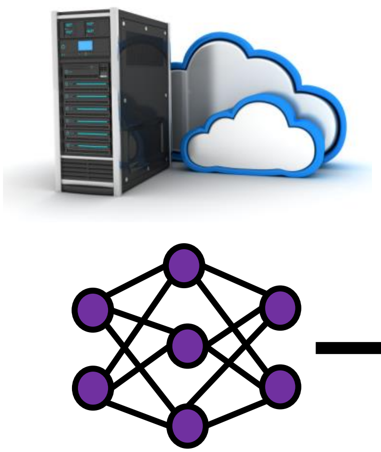




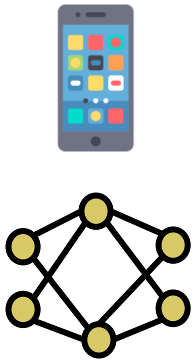
- Information Loss & Extra Overhead
  - Remove entire channels or filters
  - Pruning performed by the client



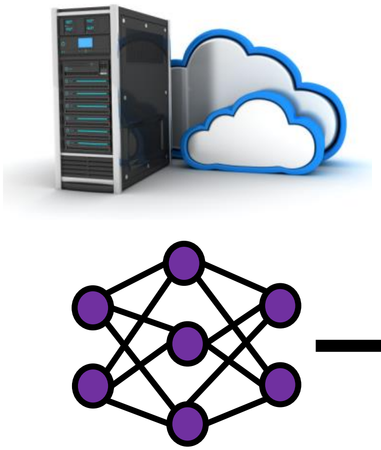
Ideally for Sub-model Generation…



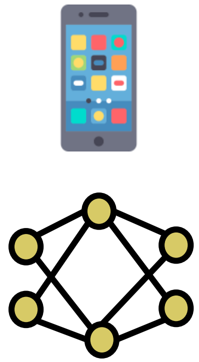
1. Minimize the information loss
2. Retain the performance
3. No extra overhead on clients



Ideally for Sub-model Generation…



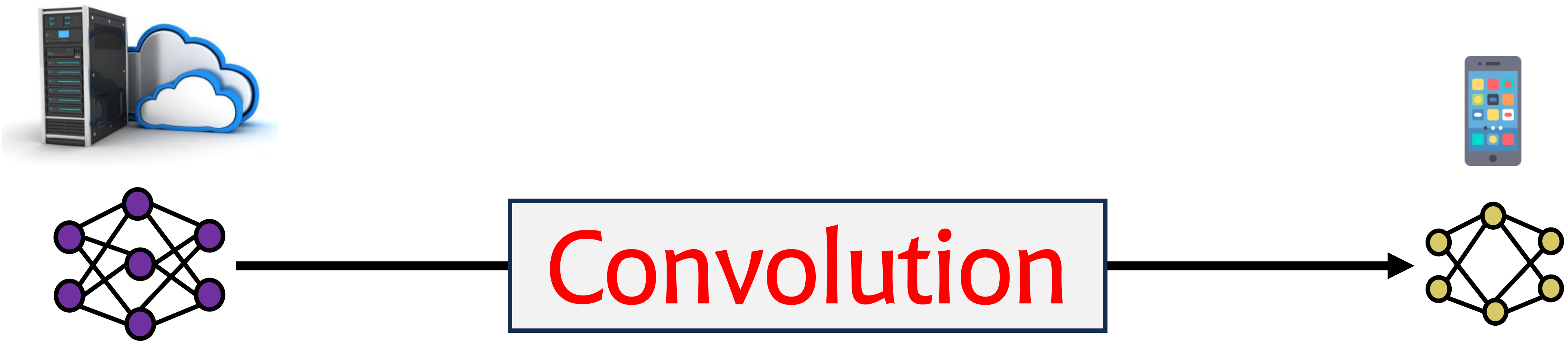
1. Minimize the information loss
2. Retain the performance
3. No extra overhead on clients



$\boxed{\text{Convolution}}$

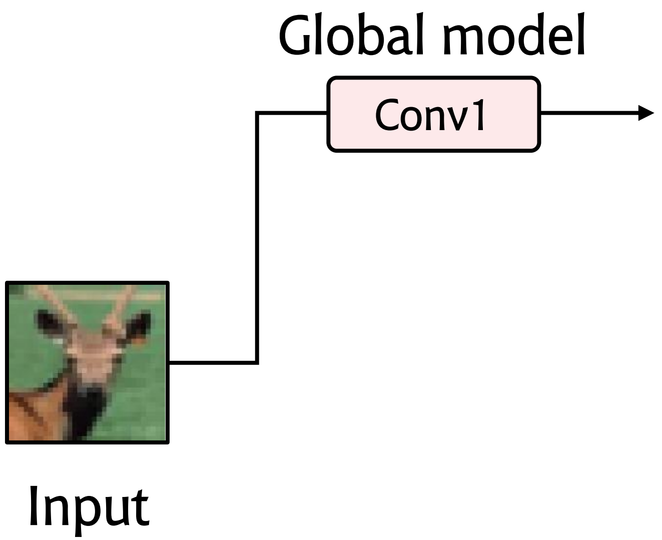
Insight

- Convolution can extract effective features from input images

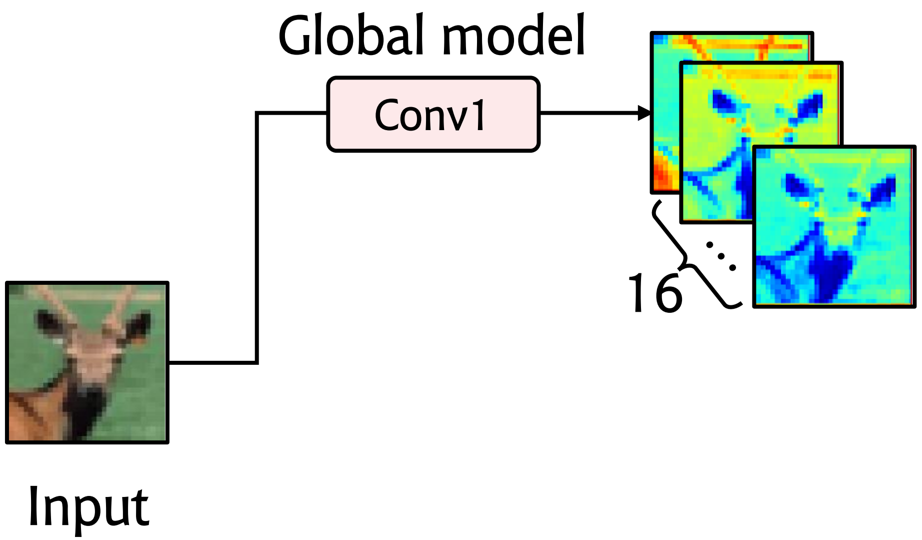


- We can also use it to extract crucial parameter information

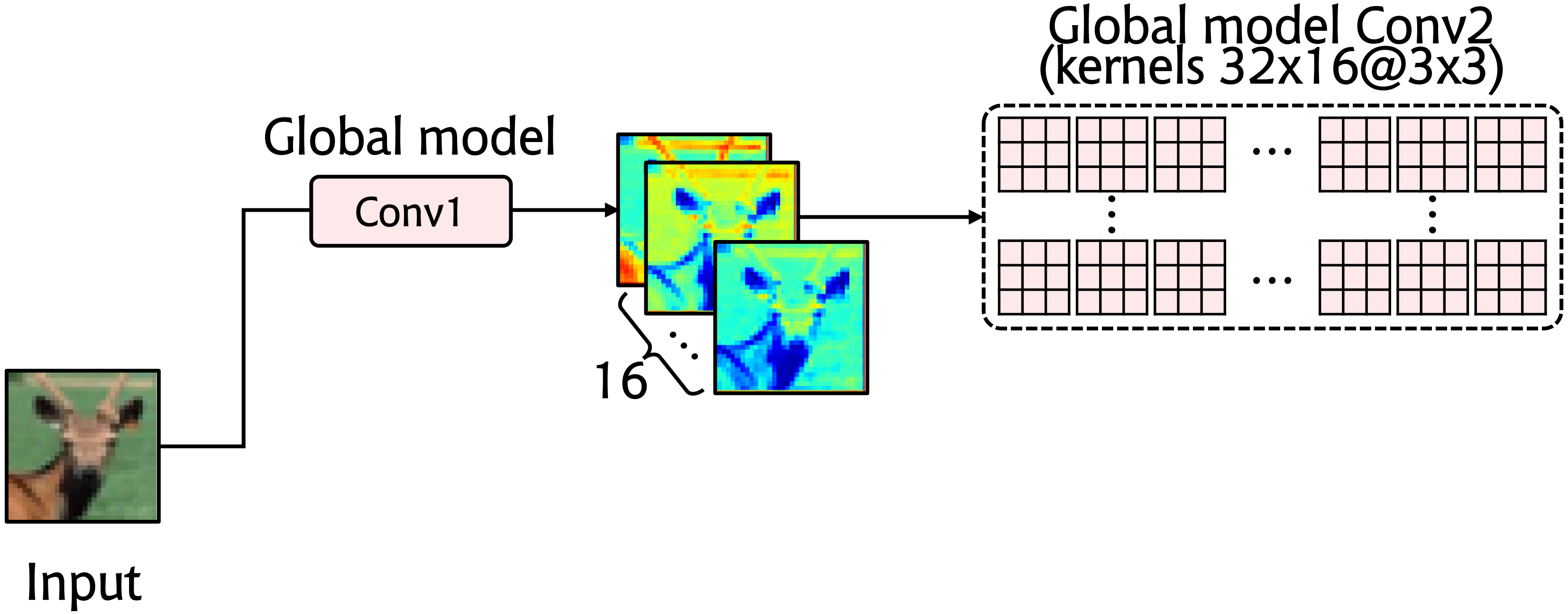
Convolutional Compression



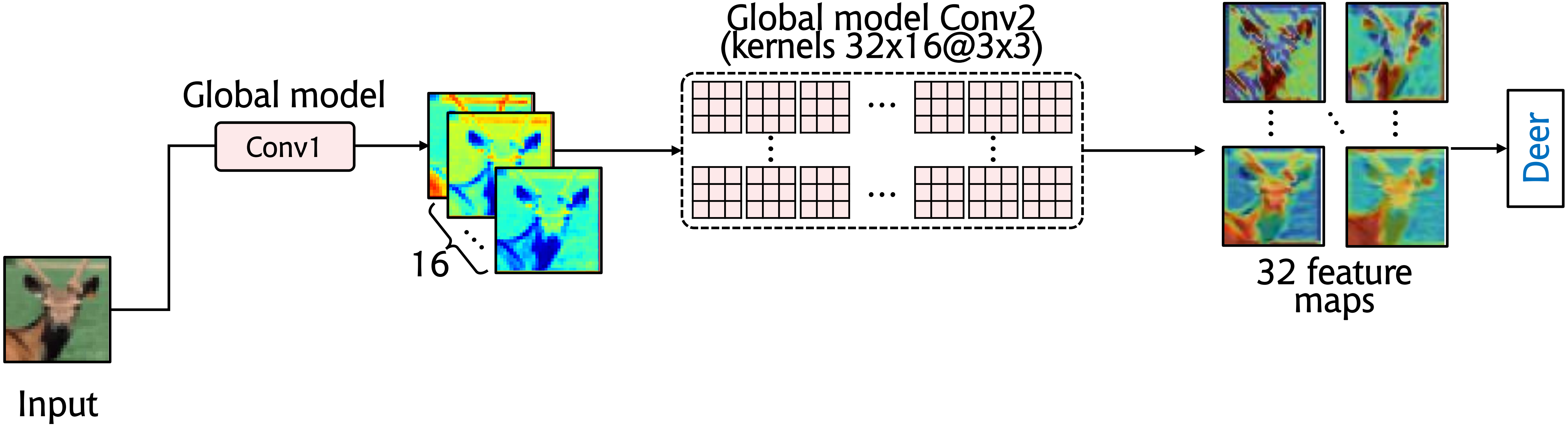
Convolutional Compression



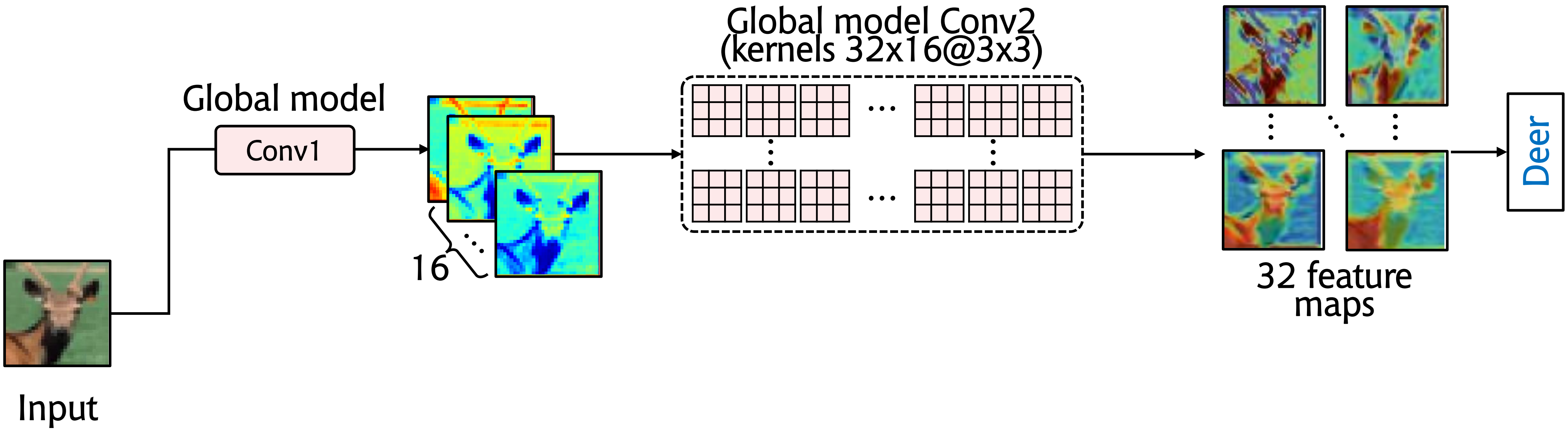
Convolutional Compression



Convolutional Compression

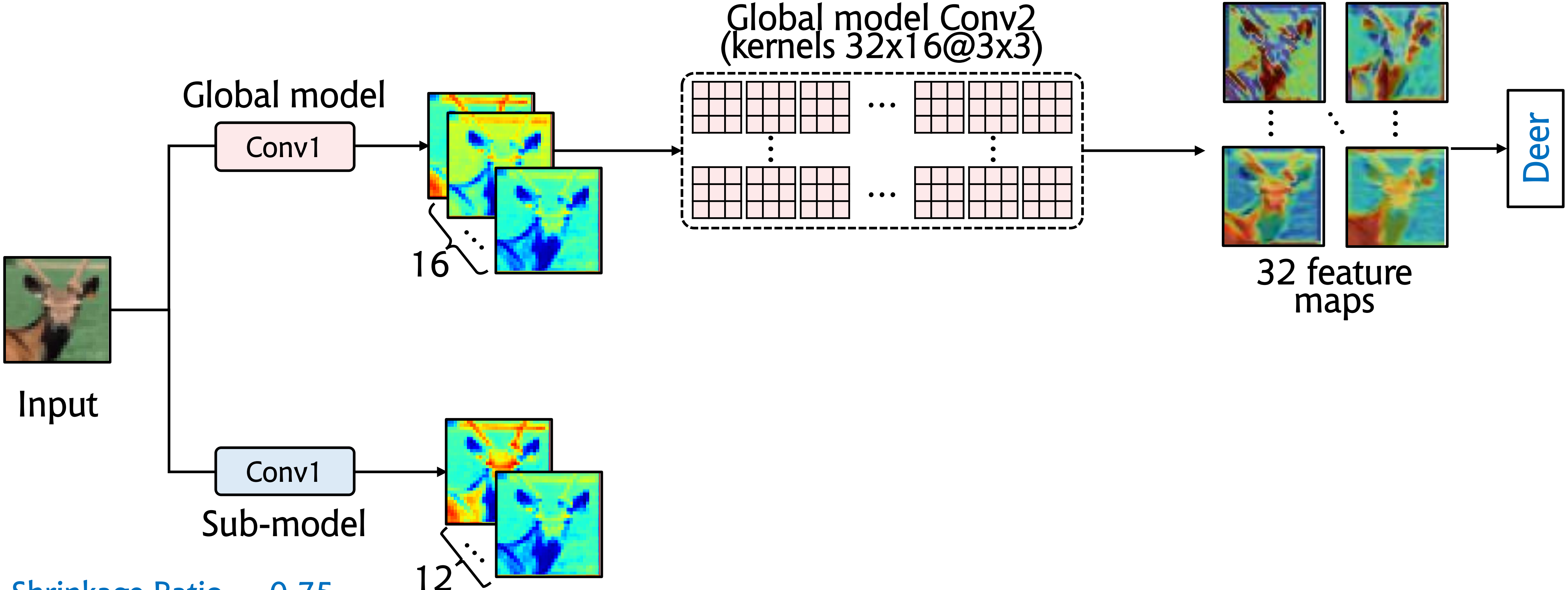


Convolutional Compression



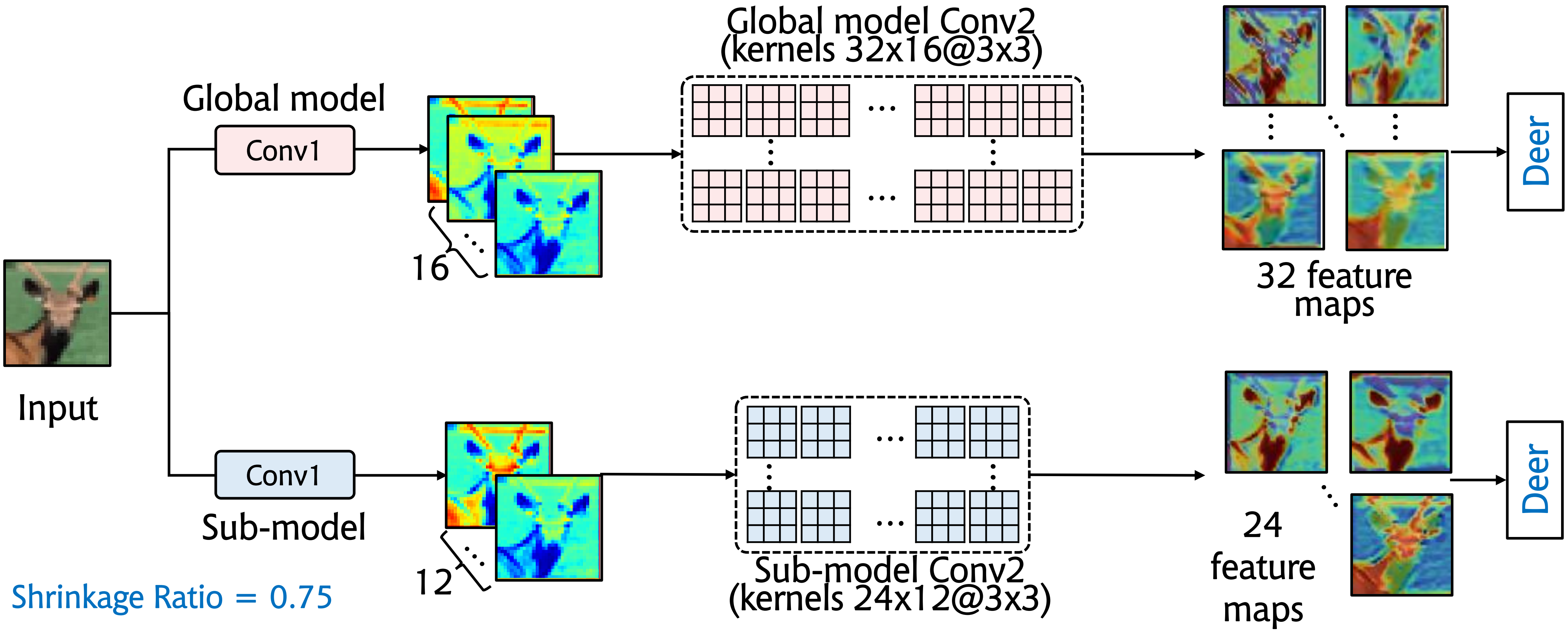
Shrinkage Ratio = 0.75

Convolutional Compression

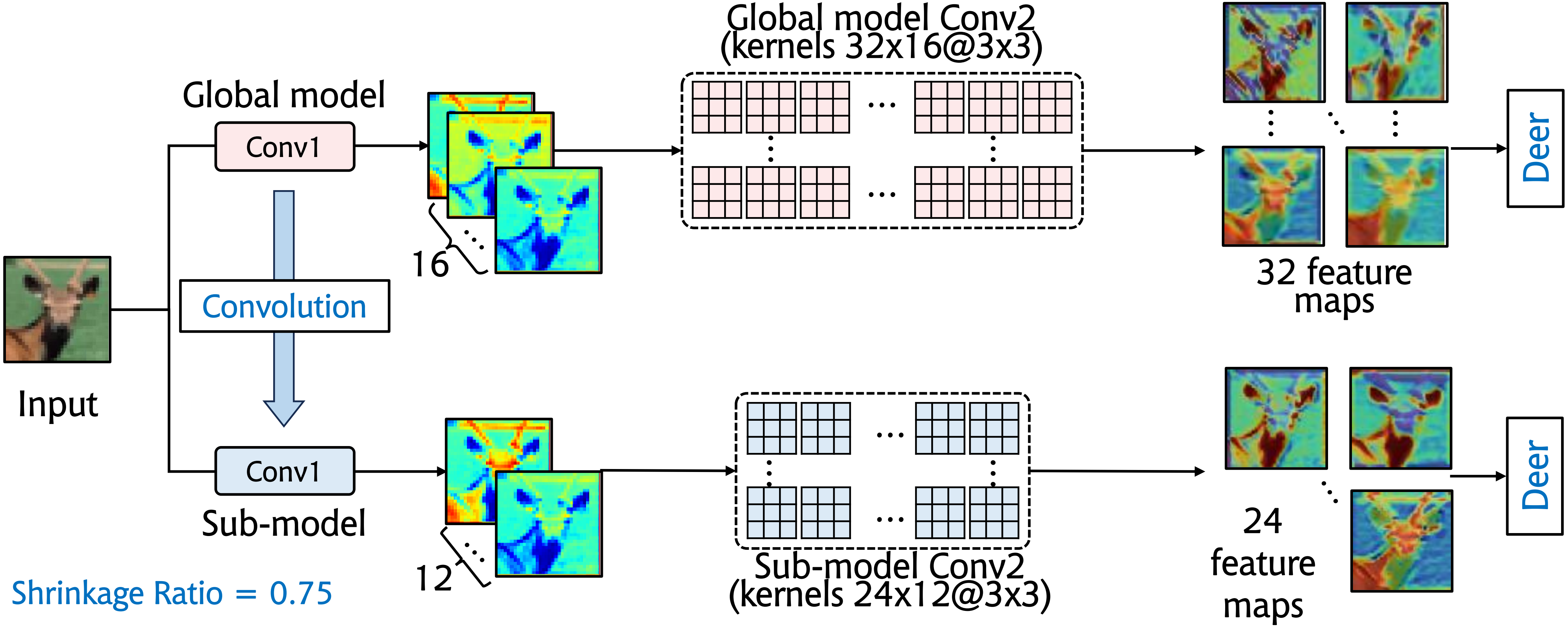


Shrinkage Ratio = 0.75

Convolutional Compression

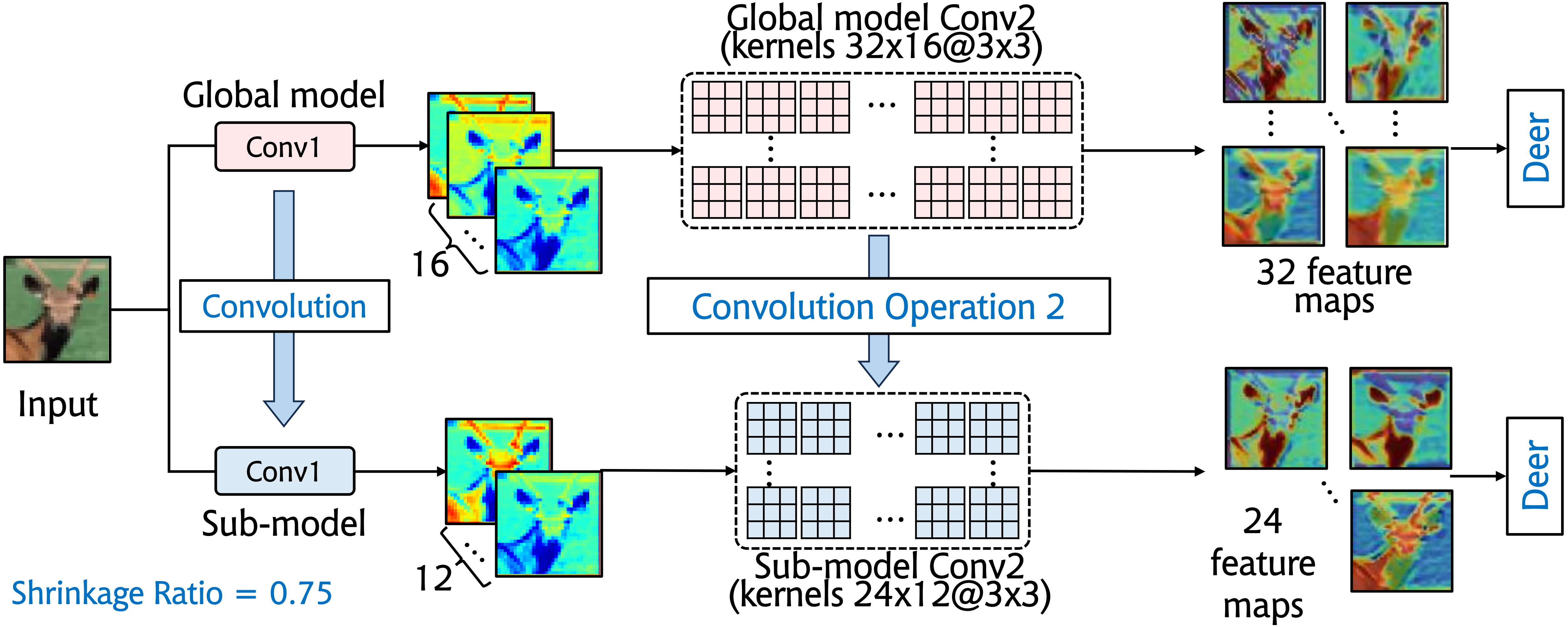


Convolutional Compression

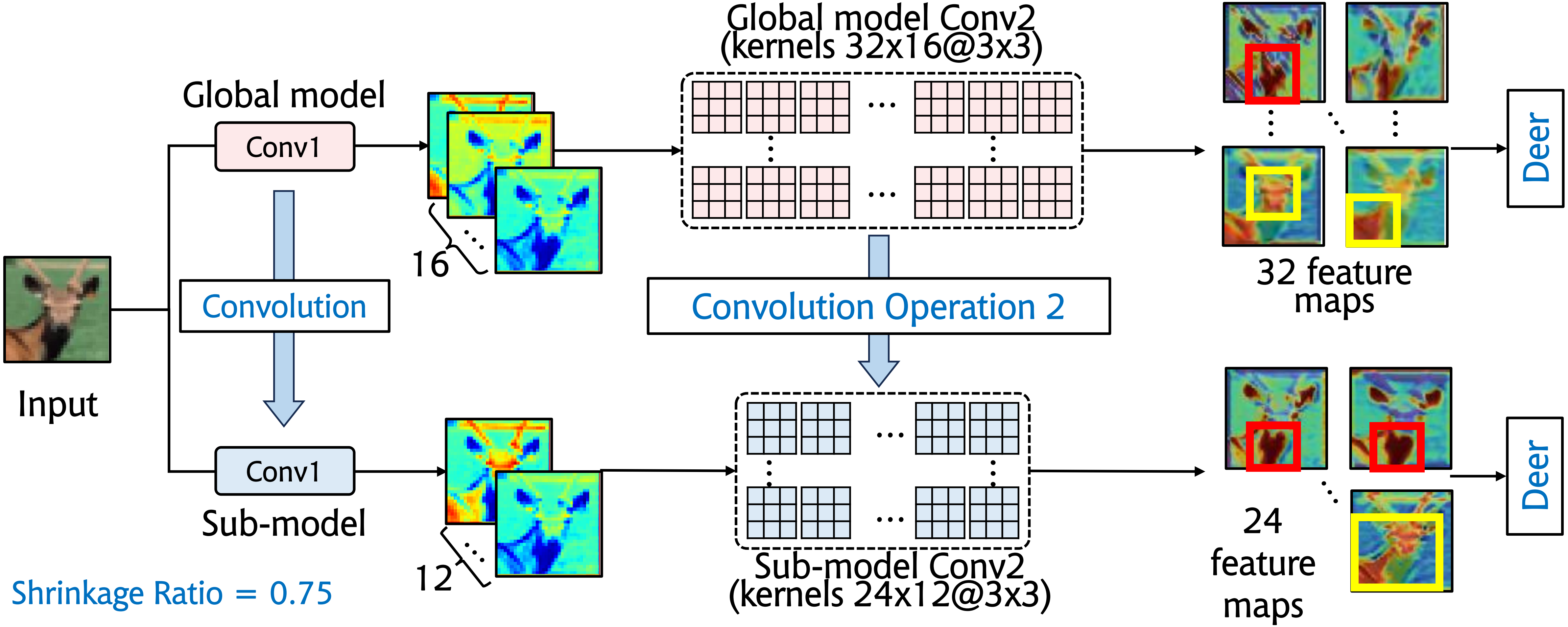


Shrinkage Ratio = 0.75

Convolutional Compression

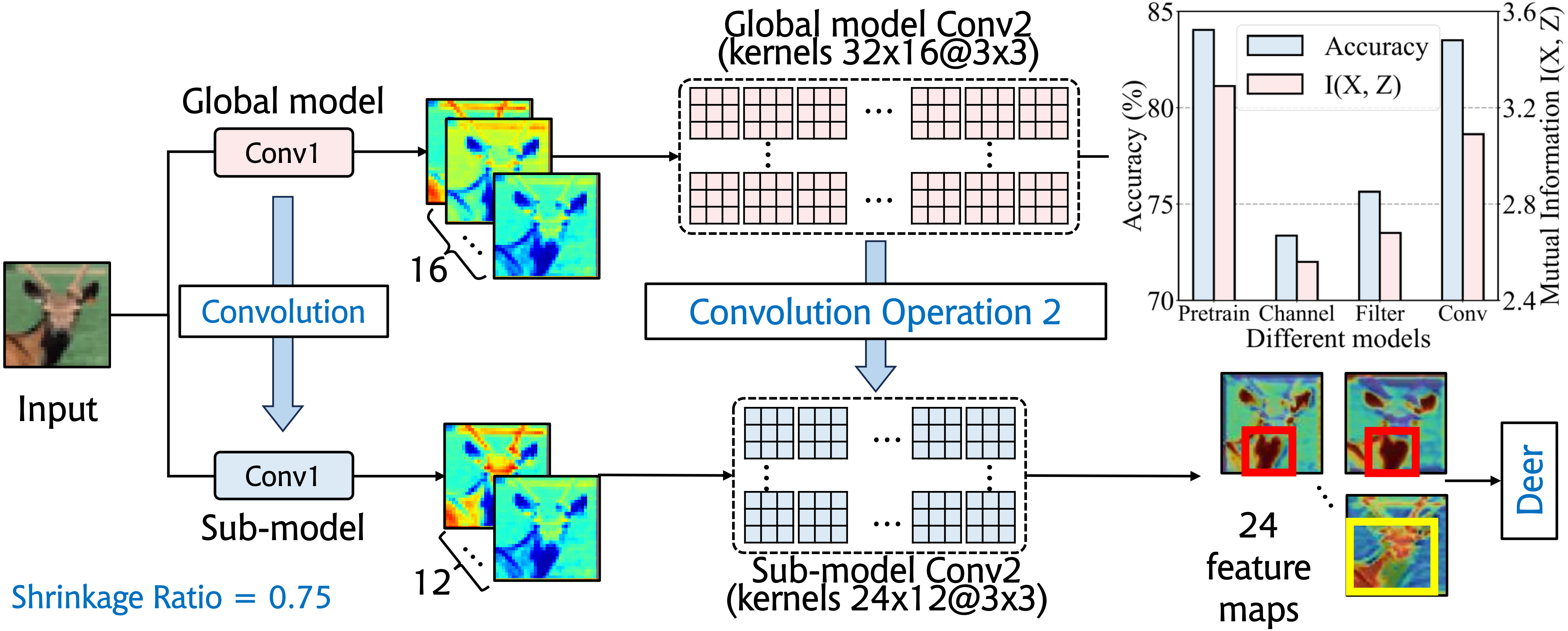


Convolutional Compression



Shrinkage Ratio = 0.75

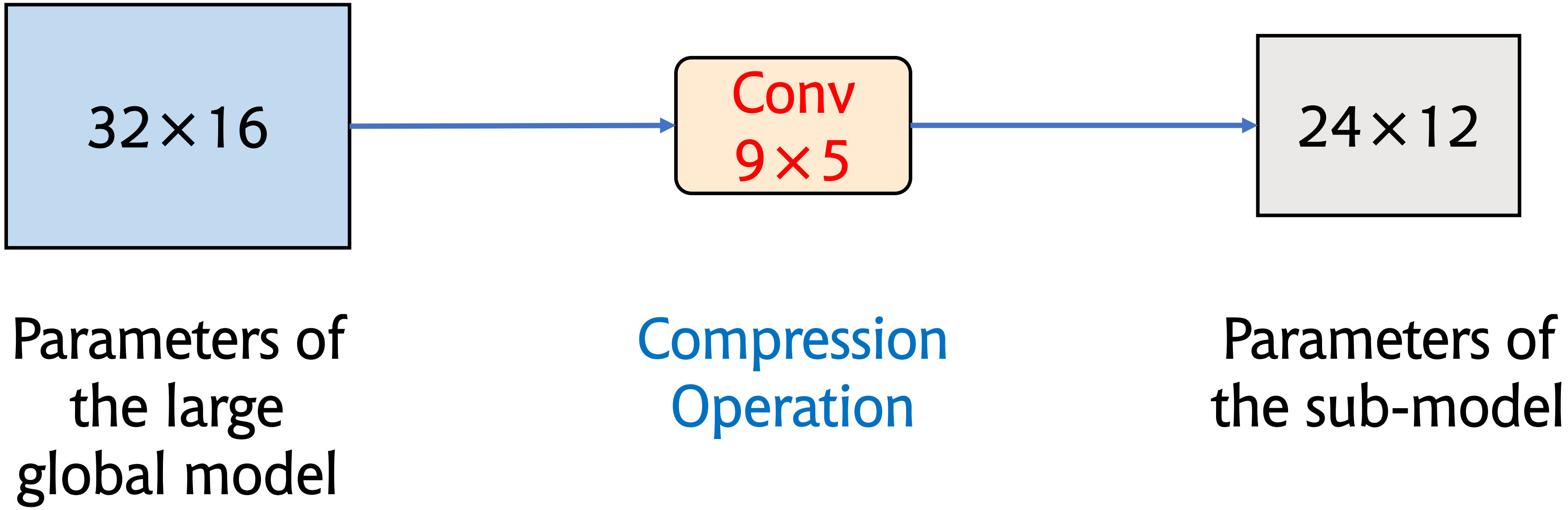
Convolutional Compression



Shrinkage Ratio = 0.75

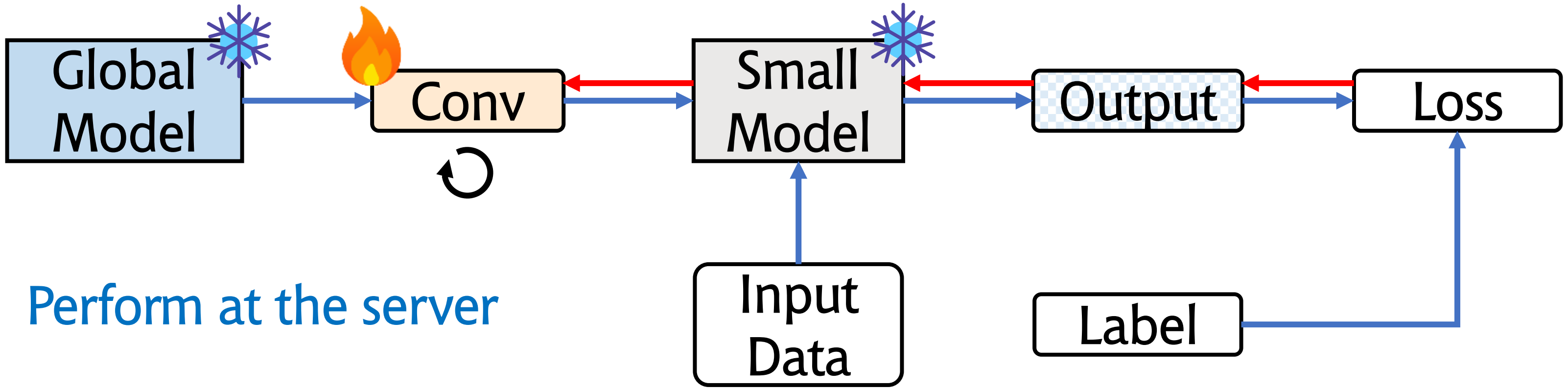
Convolutional Compression

- How to determine the size of the compressed model?
- Shrinkage Ratio $= 0.75$



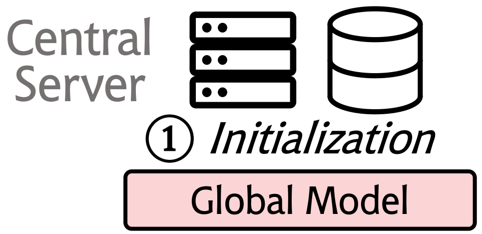
Convolutional Compression (Cont.)

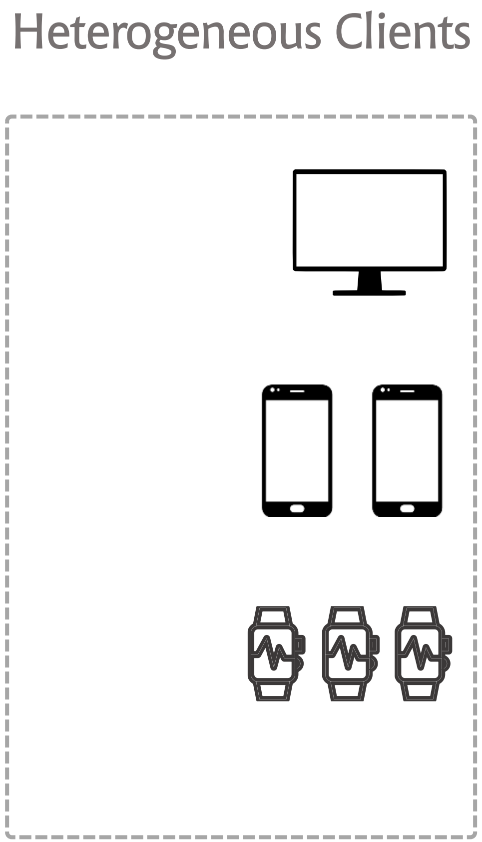
- How to retain performance?
- A learning-on-model paradigm



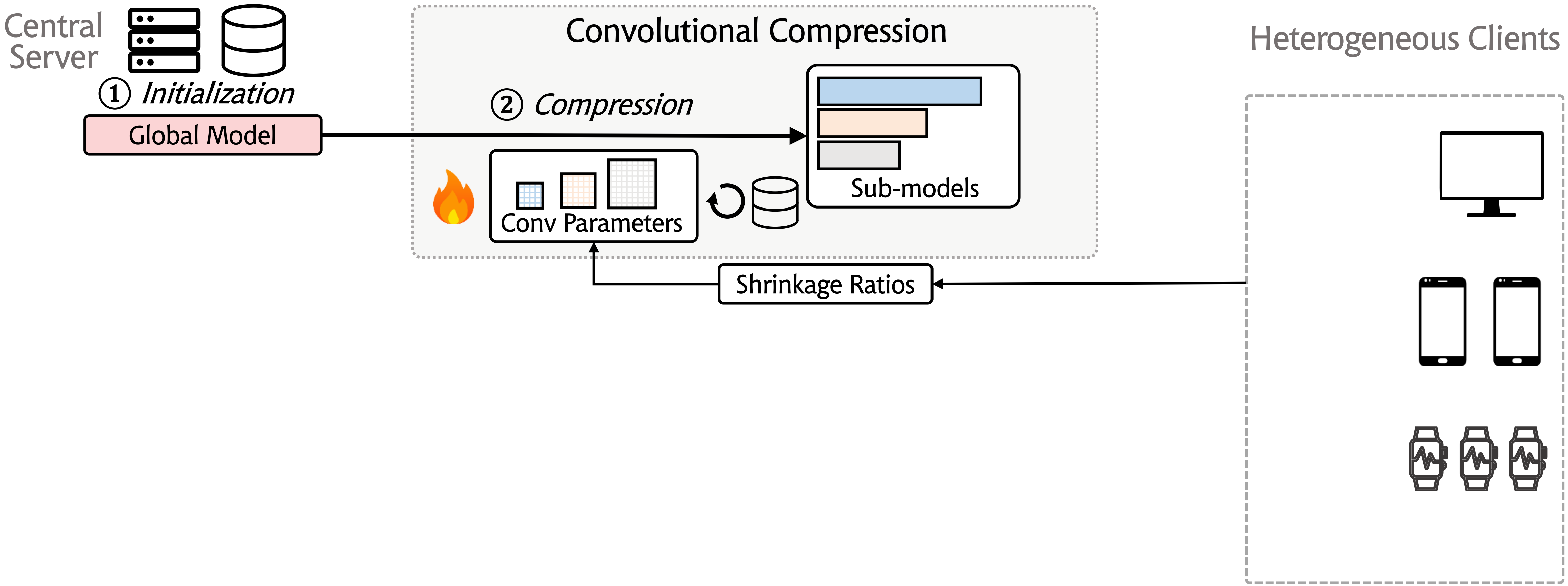
- Learning-on-data: raw data as input
- Learning-on-model: model parameters as input

System Overview – FedConv

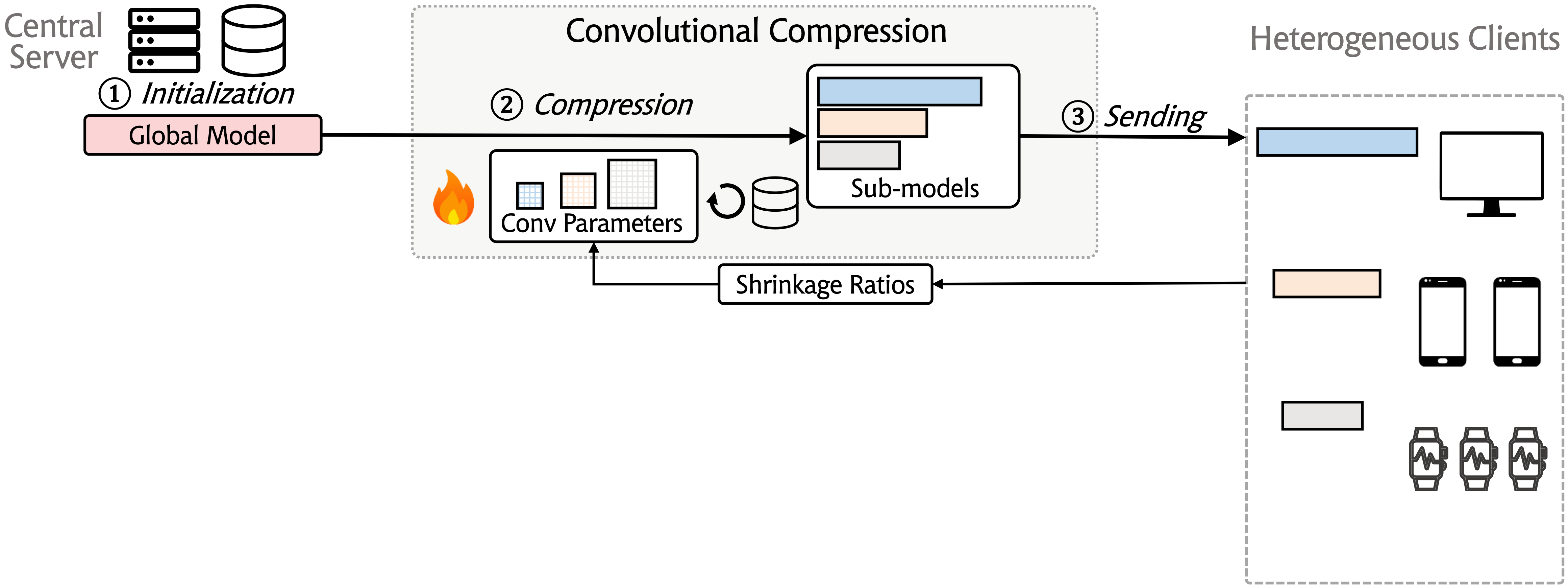




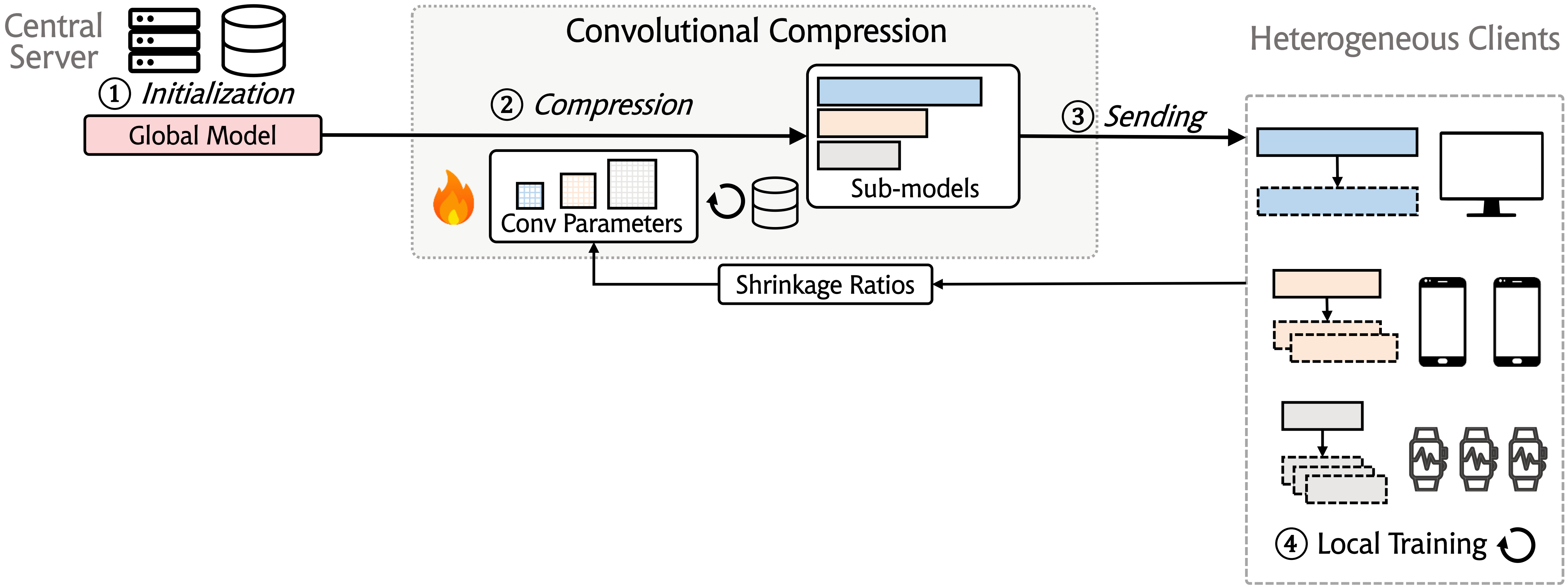
System Overview – FedConv



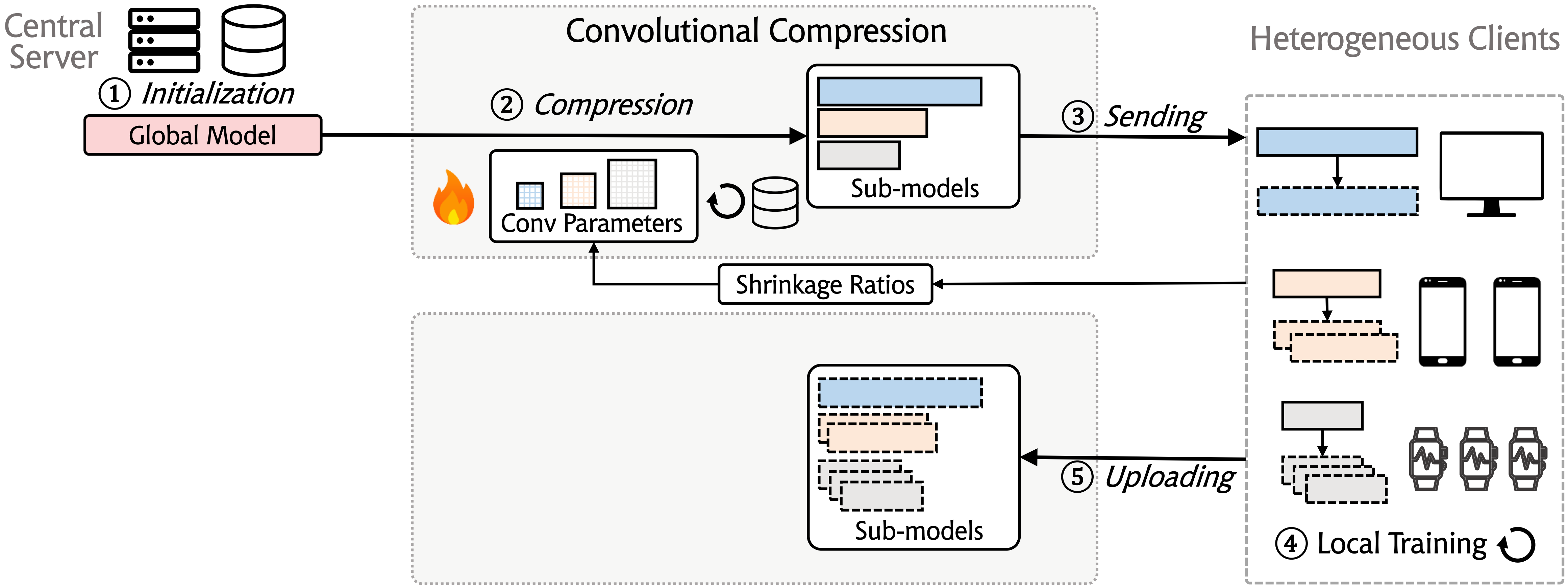
System Overview – FedConv



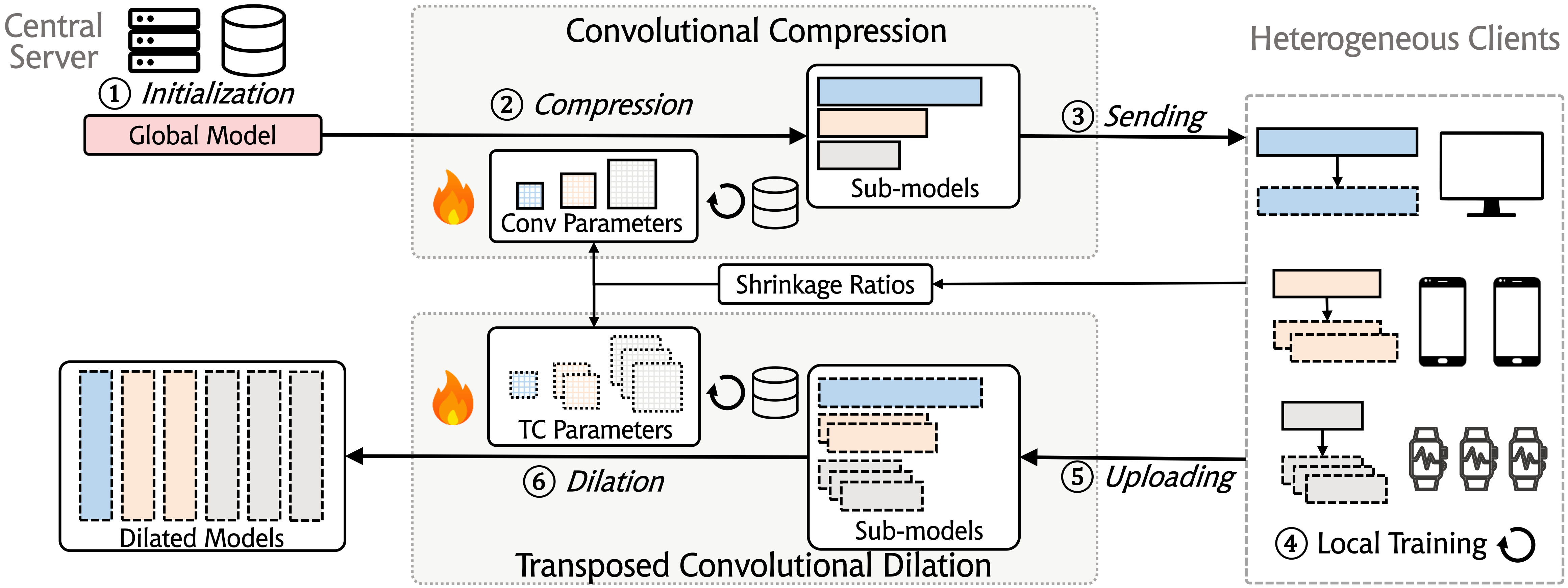
System Overview – FedConv



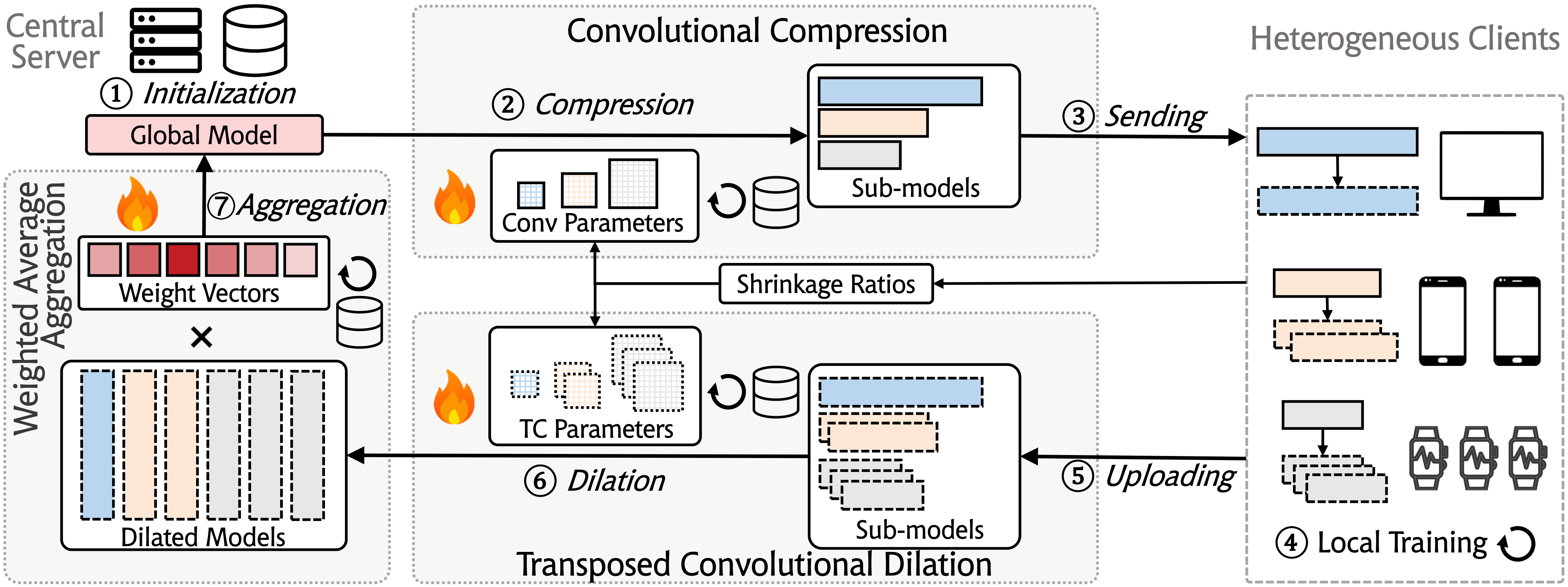
System Overview – FedConv



System Overview – FedConv



System Overview – FedConv



Experiment Setup

- Hardware

Type	Device Name	Number	CPU	RAM	GPU	GDDR	Network	SR
Server	ASUS W790-ACE Server	1	Intel Xeon Gold 6248R, 3.0GHz	640GB	NVIDIA A100	40GB	Ethernet	-
Router	Mi Router AX3000	1	Qualcomm IPQ5000 A53, 1.0GHz	256MB	-	-	Ethernet	-
PC	Supermicro X11SCA-F	2	Intel Xeon E-2236, 3.4GHz	32GB	NVIDIA RTX A4000	16GB	Ethernet	1.0
	Supermicro SYS-5038A-I	2	Intel Xeon E5-2620 v4, 2.10GHz	64GB	NVIDIA GeForce GTX 1080 Ti	12GB * 2	Wi-Fi	1.0
	ThinkPad P52s Laptop	4	Intel i5-8350U, 1.70GHz	32GB	NVIDIA Quadro P500	2GB	Wi-Fi	0.75
Board	NVIDIA Jetson TX2	4	Dual-Core NVIDIA Denver 2, 2GHz	8GB	256-core NVIDIA Pascal GPU	4GB	Wi-Fi	0.75
	NVIDIA Jetson Nano	4	ARM Cortex-A57 MPCore, 1.5 GHz	4GB	NVIDIA Maxwell architecture GPU	2GB	Wi-Fi	0.5
	Raspberry Pi 4	4	Quad core Cortex-A72, 1.8GHz	8GB	-	-	Wi-Fi	0.25

- Software
  - NN framework: PyTorch (we modify its package to enable back-propagation of the gradient to update convolution parameters)
  - FL framework: Flower

Experiment Setup (Cont.)

- Datasets & Models
  - Image Classification
    - MNIST: handwritten digits ---- CNN
    - CIFAR10: color images ---- ResNet18
    - CINIC10: color images ---- GoogLeNet
  - Human Activity Recognition (HAR) ---- CNN
    - WiAR: WIFI CSI data
    - Depth camera dataset: gray-scale depth images
    - HARBox: 9-axis IMU data

Experiment Setup (Cont.)

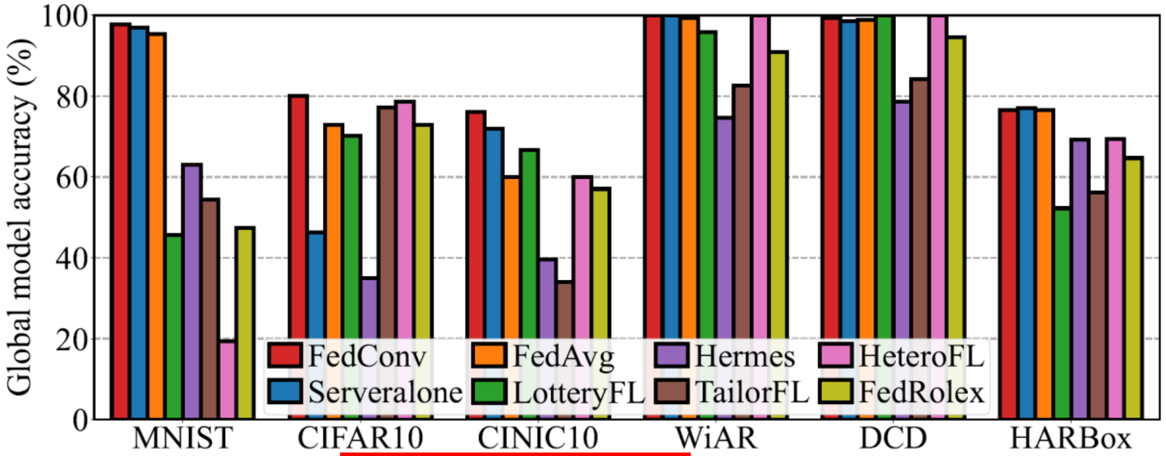
- Baselines
  - Serveralone: trains one model with only server-side data
  - Standalone: each client separately trains their local models
  - FedAvg: averages the model parameters
  - FedMD: a knowledge distillation-based method
  - LotterFL: uses Lottery Ticket hypothesis to generate heterogeneous models
  - Hermes: applies channel-level pruning
  - TailorFL: applies filter-level pruning
  - HeteroFL: static parameter sharing scheme
  - FedRolex: dynamic parameter sharing scheme

Evaluation – Metrics

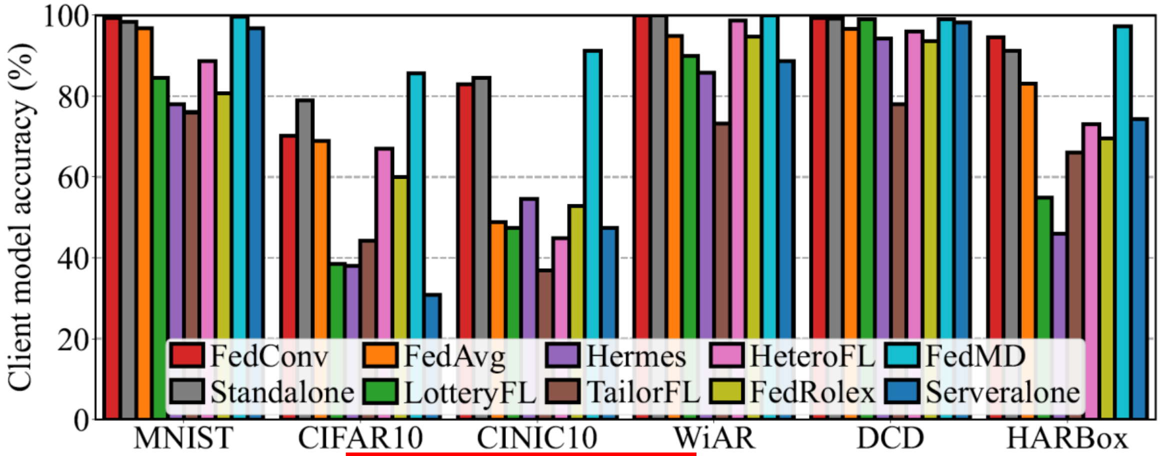
- Training Performance
  - Inference accuracy
    - Generalization: global model accuracy on global dataset
    - Personalization: client model accuracy on client dataset
  - Communication cost
- Runtime Performance
  - Memory footprint: CPU + GPU memory usage
  - Wall-clock time: total execution time of each client

Evaluation – Overall Performance

- Global model & client model performance



(a) Global model accuracy comparison

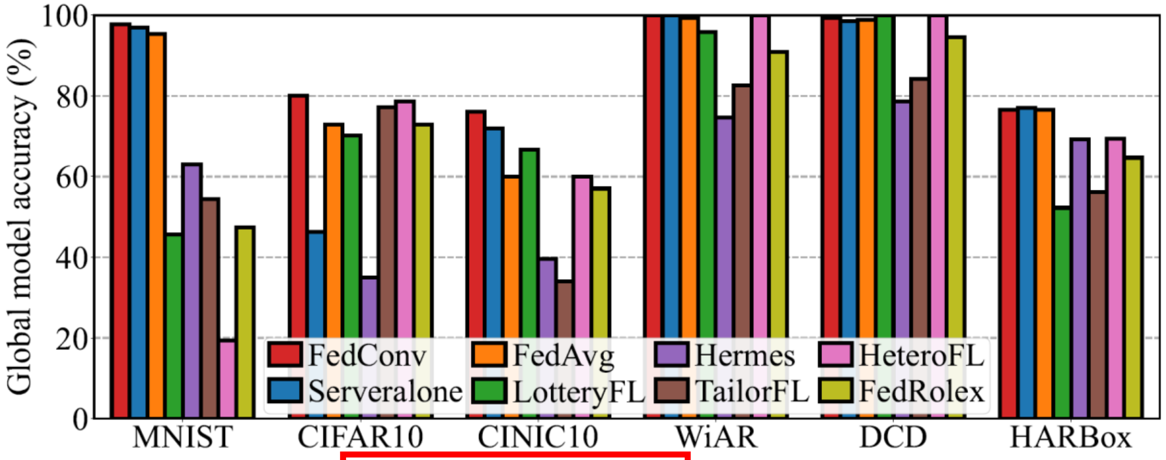


(b) Client model accuracy comparison

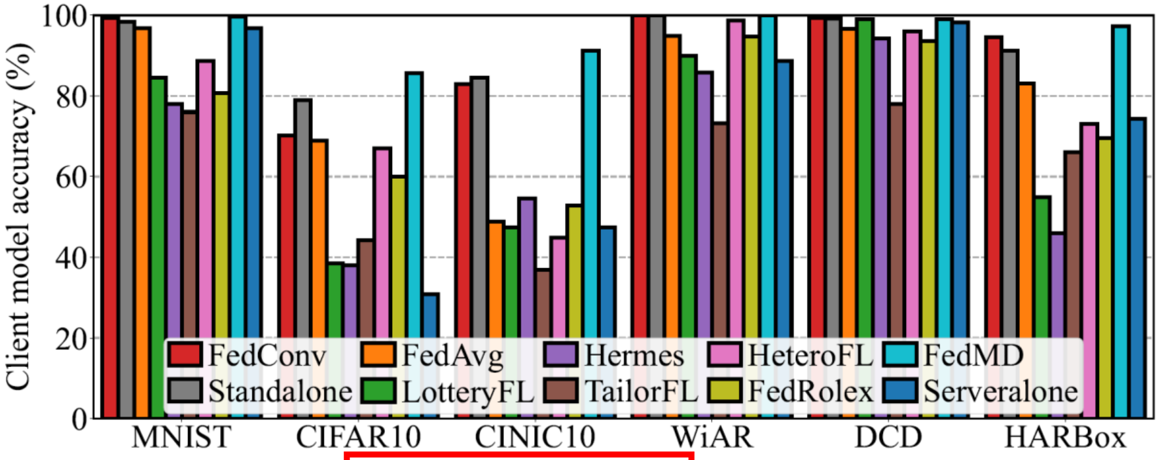
THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

Evaluation – Overall Performance

- Global model & client model performance



(a) Global model accuracy comparison



(b) Client model accuracy comparison

The superior generalization performance of FedConv

The personalization performance of FedConv

Evaluation – Overall Performance

- Global model & client model performance (Cont.)

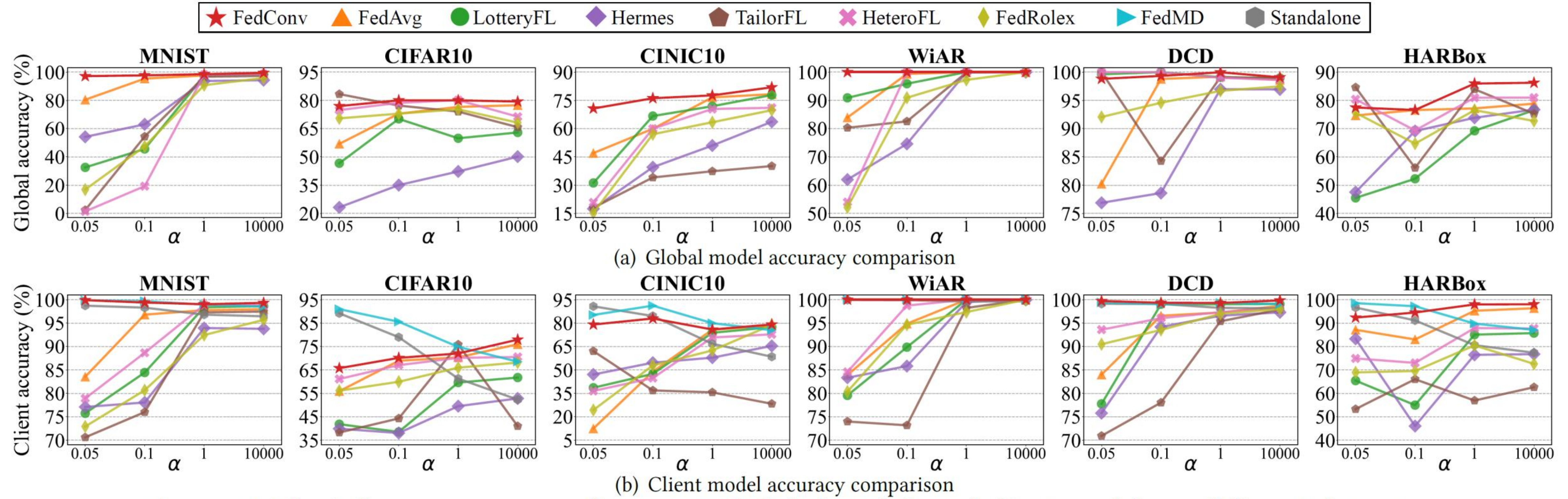
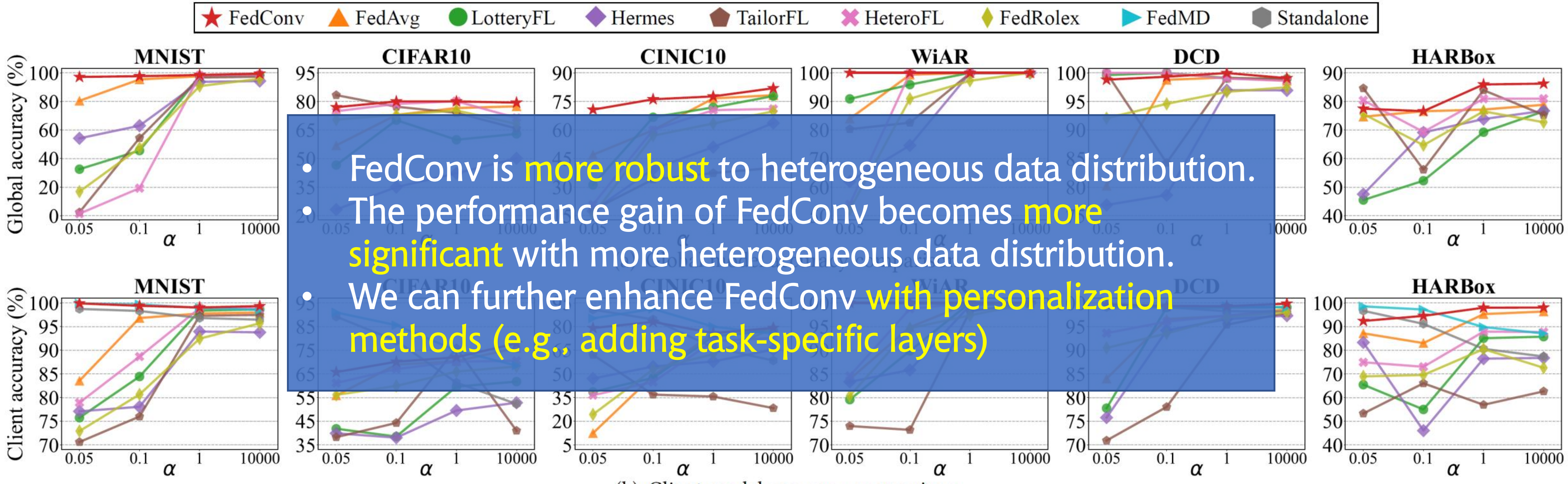


Figure 10: The inference accuracy of aggregated global models and client models on different datasets.

Evaluation – Overall Performance

- Global model & client model performance (Cont.)



- FedConv is more robust to heterogeneous data distribution.
- The performance gain of FedConv becomes more significant with more heterogeneous data distribution.
- We can further enhance FedConv with personalization methods (e.g., adding task-specific layers)

(b) Client model accuracy comparison

Figure 10: The inference accuracy of aggregated global models and client models on different datasets.

Evaluation – Overall Performance (Cont.)

- System Overhead

Table 2: System resource overhead.

Metric	System	Heterogeneous Data (α = 0.05)						Homogeneous Data (α = 10000)					
		MNIST	CIFAR10	CINIC10	WiAR	DCD	HARBox	MNIST	CIFAR10	CINIC10	WiAR	DCD	HARBox
Memory Footprint CPU + GPU (GB)	Standalone	2.14	3.51	4.07	3.95	2.24	2.19	2.13	3.47	4.47	4.03	2.21	2.17
	FedAvg	1.90	2.40	3.31	2.39	1.98	2.01	1.90	2.51	2.79	2.36	1.88	2.08
	FedMD	2.71	3.65	7.51	4.71	2.99	2.79	2.71	3.65	7.93	4.58	2.99	2.81
	LotteryFL	2.62	3.51	4.30	3.23	2.69	2.67	2.63	3.49	4.36	3.27	2.70	2.66
	Hermes	2.64	3.45	6.07	3.28	2.73	2.69	2.64	3.35	6.13	3.32	2.72	2.68
	TailorFL	2.75	3.61	5.09	3.41	2.79	2.71	2.75	3.47	7.52	3.16	2.77	2.70
	HeteroFL	2.63	3.31	4.15	3.25	2.73	2.67	2.63	3.45	4.10	3.08	2.73	2.67
	FedRolex	2.63	3.21	4.15	3.25	2.72	2.67	2.60	3.54	4.16	3.16	2.68	2.69
	FedConv	**2.52**	**3.21**	**4.15**	**3.02**	**2.60**	**2.67**	**2.52**	**3.35**	**4.10**	**3.14**	**2.62**	**2.67**
Wall-clock Time (s)	Standalone	3.87	24.65	279.62	8.05	5.91	3.54	9.38	52.38	273.52	7.60	6.14	3.56
	FedAvg	7.05	39.19	285.30	10.62	10.19	10.09	13.75	97.95	1711.34	20.79	43.67	26.98
	FedMD	44.34	437.14	5370.83	55.03	75.25	32.92	45.17	475.42	6700.17	64.43	79.10	34.53
	LotteryFL	9.18	147.98	699.35	8.89	8.61	5.69	17.59	235.89	1829.33	19.77	22.06	10.92
	Hermes	43.22	714.00	5580.71	103.90	169.97	104.53	43.84	937.82	7621.38	117.85	217.97	115.31
	TailorFL	6.98	62.89	393.46	14.44	12.72	10.11	13.61	99.60	813.94	25.53	13.96	13.27
	HeteroFL	6.96	42.56	641.21	10.78	10.03	5.10	13.56	82.07	1310.81	22.26	23.90	10.98
	FedRolex	6.92	45.98	602.48	11.57	12.34	4.87	12.46	84.25	1389.41	23.64	20.14	11.26
	FedConv	**5.96**	**40.68**	**264.30**	**12.96**	**10.15**	**4.40**	**10.33**	**71.26**	**1406.87**	**21.79**	**17.22**	**9.89**

Evaluation – Overall Performance (Cont.)

- System Overhead – Communication Cost

Table 3: Communication overhead comparison (GB).

System	MNIST	CIFAR10	CINIC10	WiAR	DCD	HARBox
FedAvg	14.80	4815.84	2697.85	28.24	13.45	8.87
FedMD	19.99	5126.46	2859.79	40.91	19.94	16.24
LotteryFL	11.11	4713.91	2623.93	23.01	10.05	8.55
Hermes	16.34	7099.66	2848.83	36.63	15.02	12.95
TailorFL	11.40	4787.18	2686.15	24.30	10.32	8.82
HeteroFL	11.11	4713.91	2623.93	23.01	10.05	8.55
FedRolex	11.11	4713.91	2623.93	23.01	10.05	8.55
FedConv	**11.11**	**4713.91**	**2623.93**	**23.01**	**10.05**	**8.55**

Conclusion

- We propose FedConv, a client-friendly federated learning framework for heterogeneous clients, aiming to minimize the system overhead on resource-constrained mobile devices.
- FedConv features three key technical modules: convolutional compression, TC dilation, and weighted average aggregation.
- We believe the proposed learning-on-model paradigm is worthy of further exploration (*e.g.*, configuration optimization).

Thanks for Listening!

- FedConv: A Learning-on-Model Paradigm for Heterogeneous Federated Clients

- Leming Shen, Qiang Yang, Kaiyan Cui, Yuanqing Zheng, Xiao-Yong Wei, Jianwei Liu, Jinsong Han