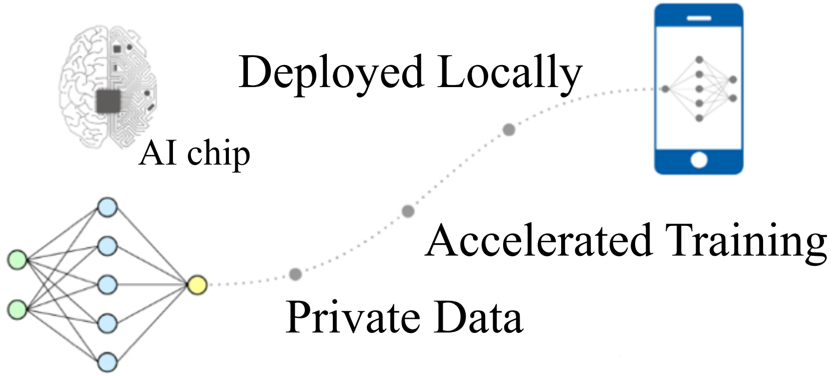
Towards Context-aware Distributed Learning for CNN in Mobile Applications

Zhuwei Qin
Assistant Professor
School of Engineering
San Francisco State University
zwqin@sfsu.edu

Hao Jiang
Associate Professor
School of Engineering
San Francisco State University
jianghao@sfsu.edu

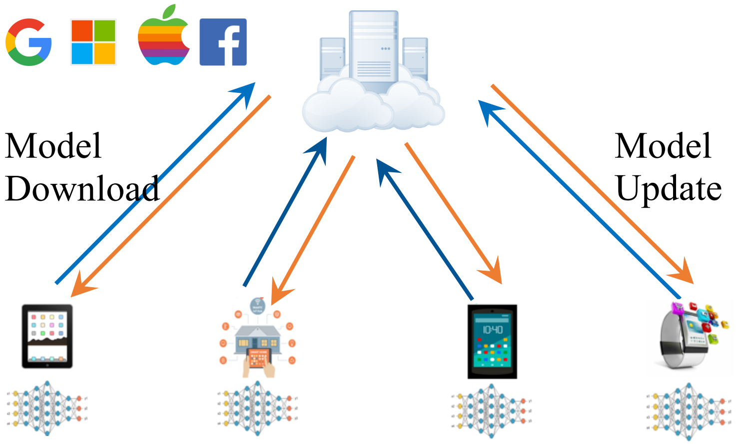
Distributed Learning on Mobile Devices

Massive Amount of Mobile Devices



- Featured with an AI chip
- Underutilized Processing Capacity

Large Scale System Deployment

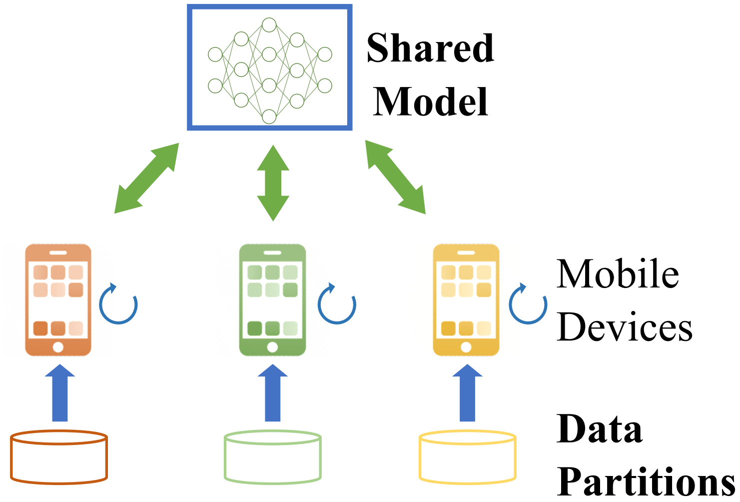


- Efficient Computation and Communication
- Scalable Number of Devices
- Accurate Distributed Aggregation

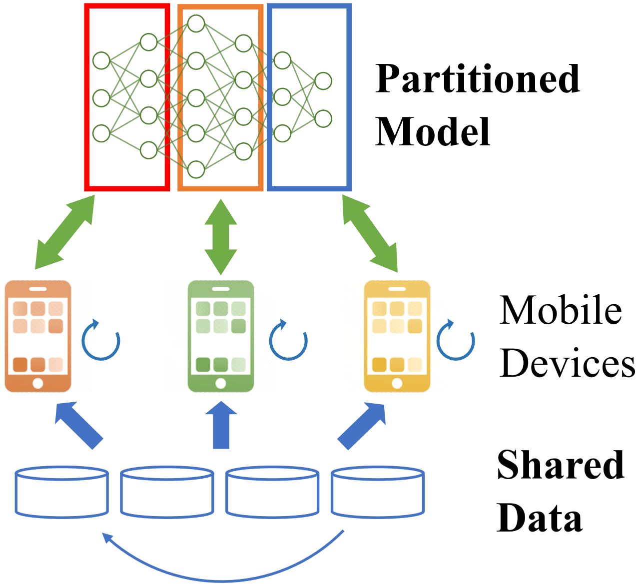
Traditional Distributed Learning Framework

Parallelism in Distributed Neural Network Training

Data Parallelism



Model Parallelism



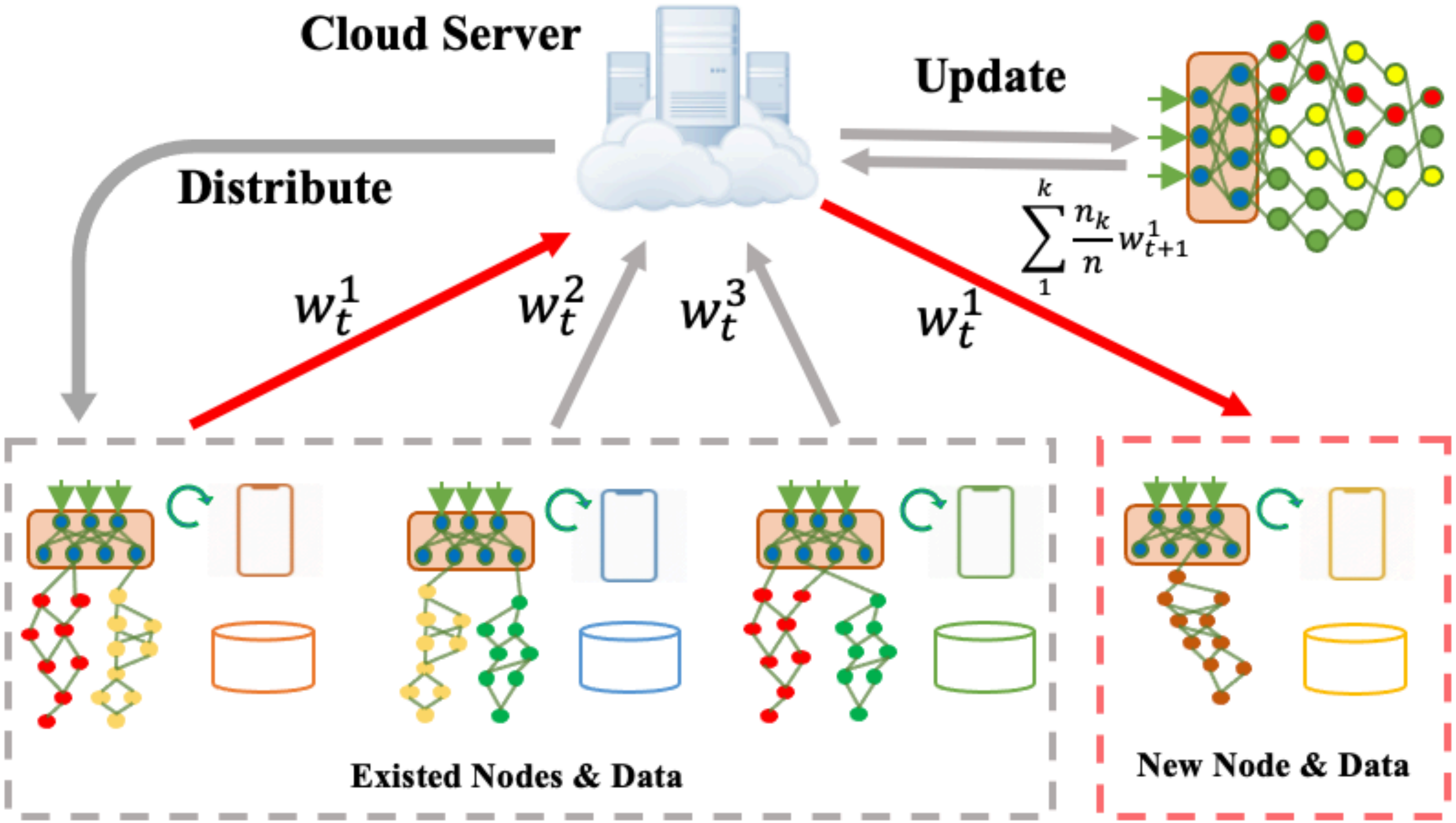
- High Data Throughput
- Heavy Computation Load

- Massive Node Communication
- Unscalable Collaboration Topology

Works	M	D	Works	M	D
Federated Learning	×	✓	DistBelief	✓	×
Parameter Sever	×	✓	COTS	✓	×
1-bit SGD	×	✓	SpecTrain	✓	×
ParallelSGD	×	✓	Ours	✓	✓

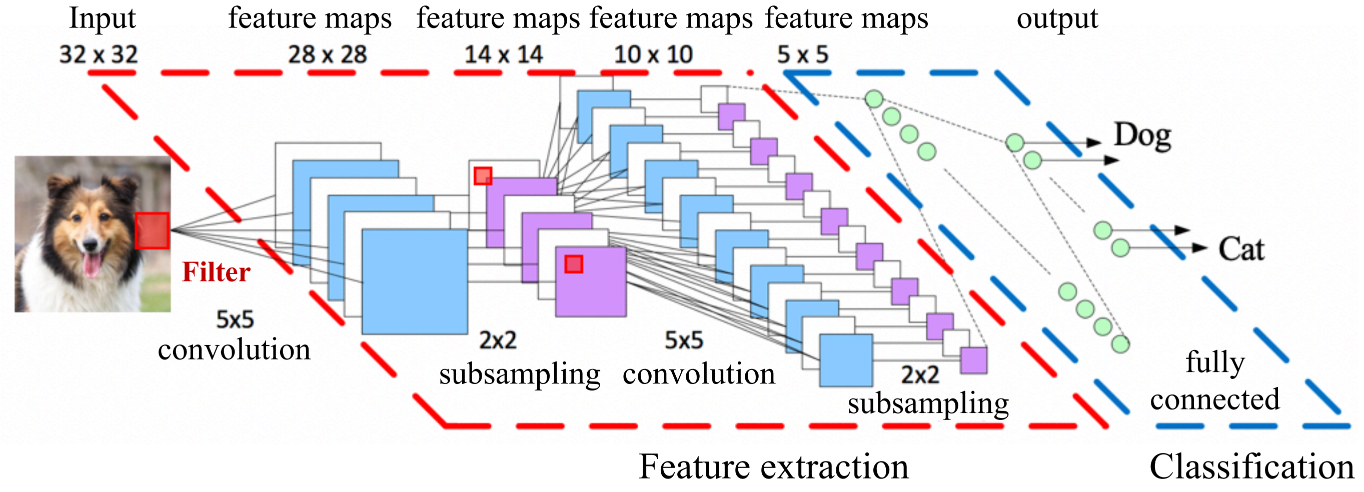
Ours: Context-aware Distributed Learning Framework

Framework Overview

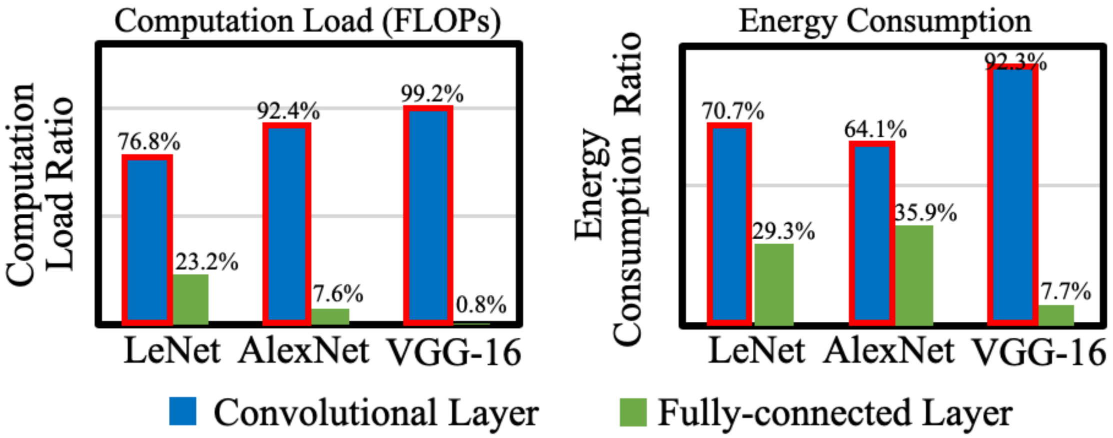


Neural Network Computation

Convolutional Neural Network (CNN) for Image Classification



Intensive Computation Load for Convolutional Operation



Optimization Target: Convolutional Layers

Convolutional Filter Visualization

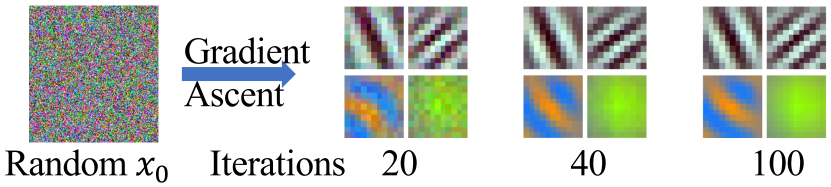
Filter Visualization by Activation Maximization

$$I^* = \underset{I}{\operatorname{argmax}} \, (A_i(x) - R_\theta(x))$$

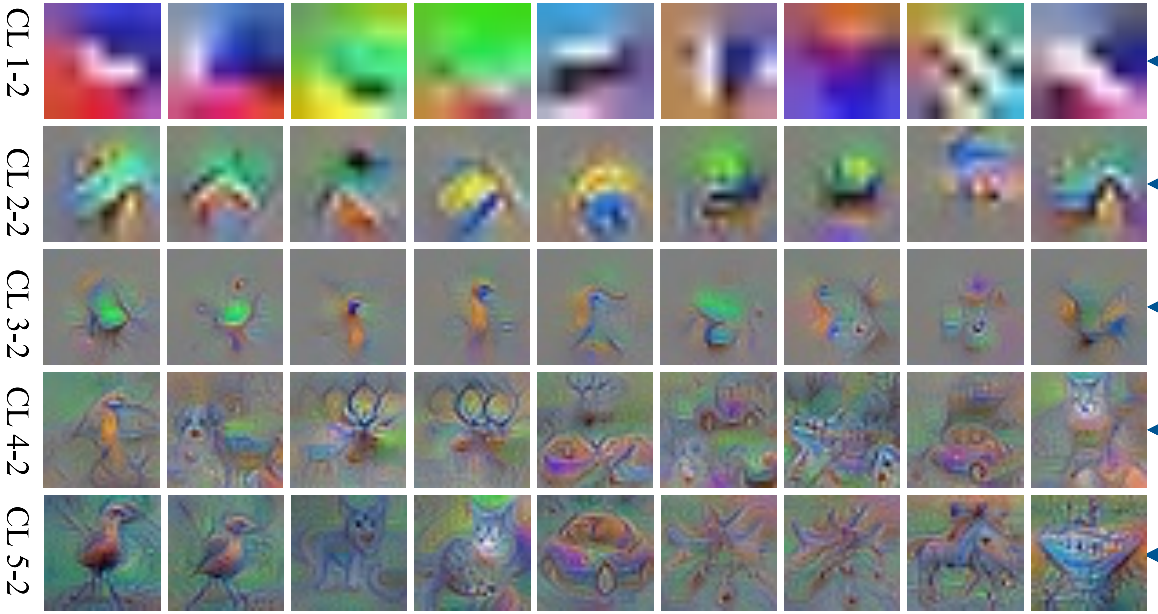
$$x \leftarrow r_\theta \left(x + \eta \frac{\partial A_i(x)}{\partial x} \right)$$

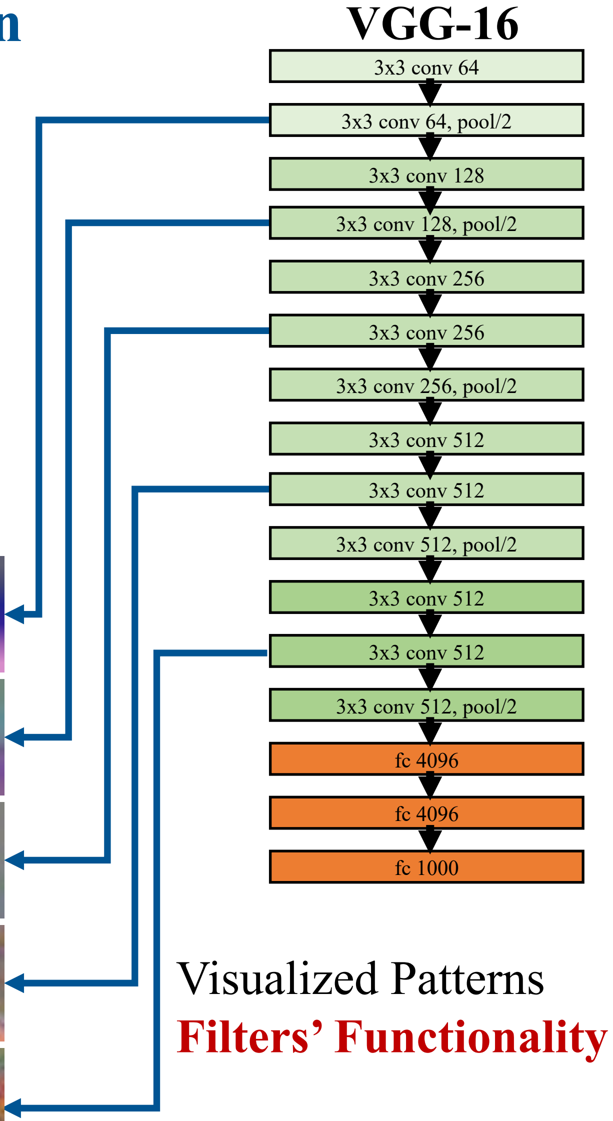
$A_i(x)$: Activation (feature map)

$R_\theta(x)$: Parameterized Regularization



Interpret Filters in Arbitrary Layers



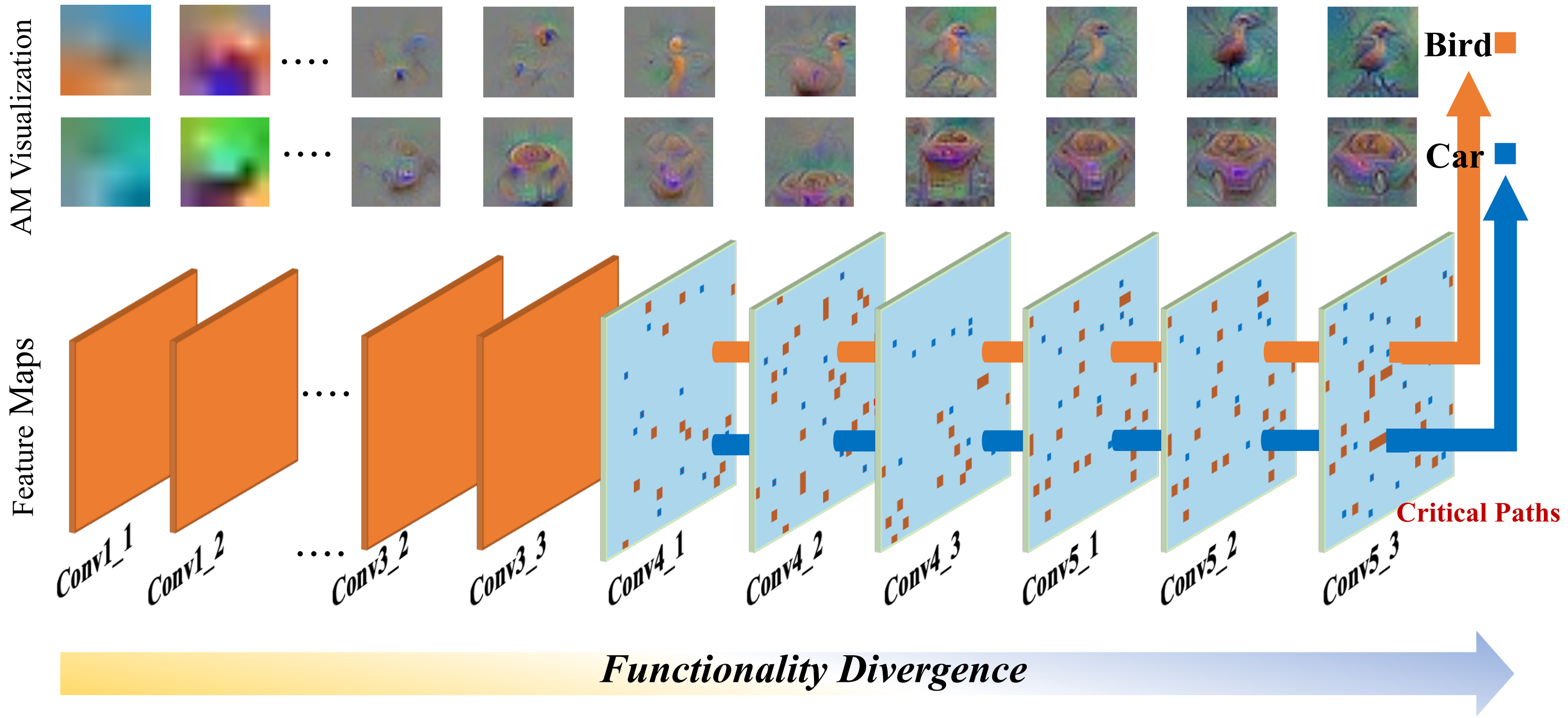


Visualized Patterns

Filters' Functionality

Class-specific “Critical Paths” of Filters

Filter Functionality Divergence

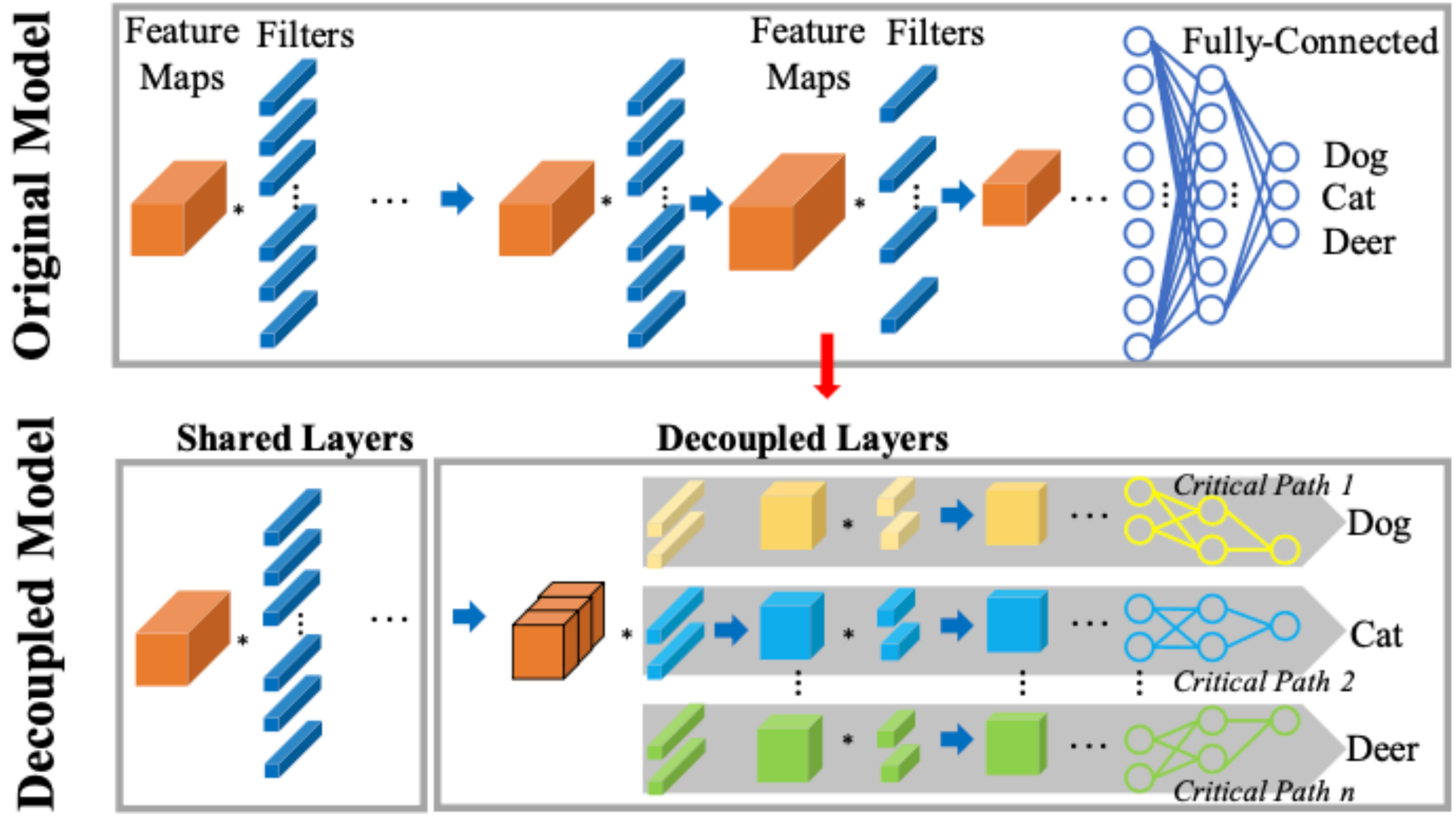


- **Shallow layers extract basic features.**
- **Deeper layers extract class-specific features.**
- **Class-specific “Critical Paths” exist in deep layers !**

Heterogenous Model Initialization by Model Decoupling

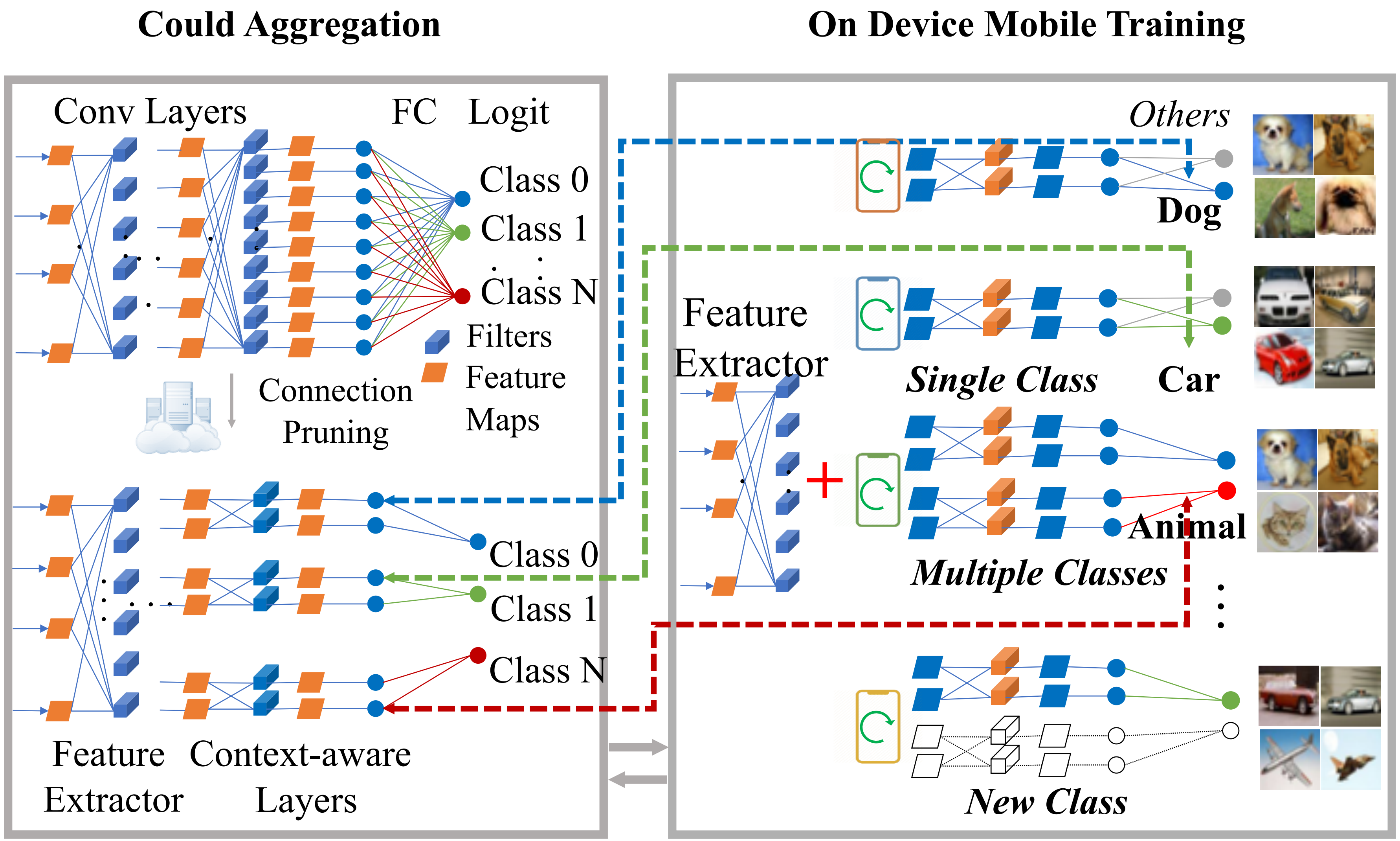
Neural Network Decoupling

$C = \operatorname{Arg}\max_{C} \sum_{i} A_i^l(x_i^c).$ • **Per-class Activation Preference for Each Filter**



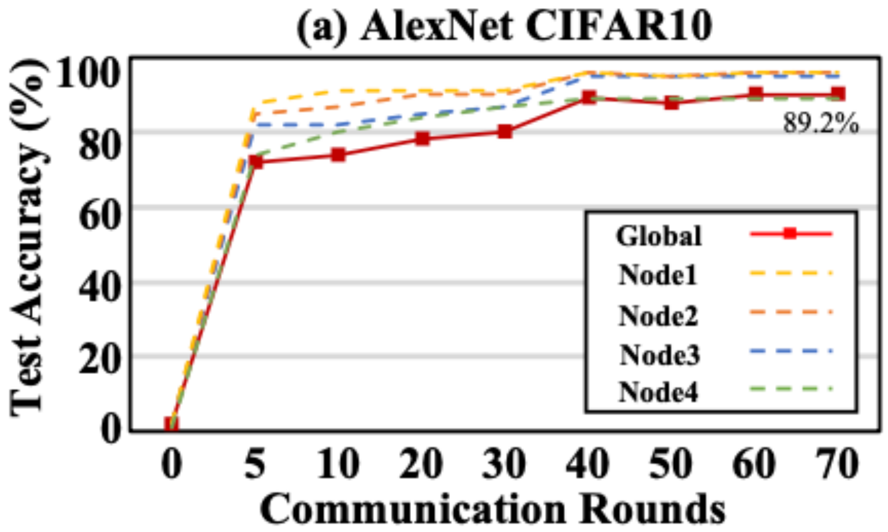
CAD: Context-aware Distributed Learning Framework

Distributed Learning

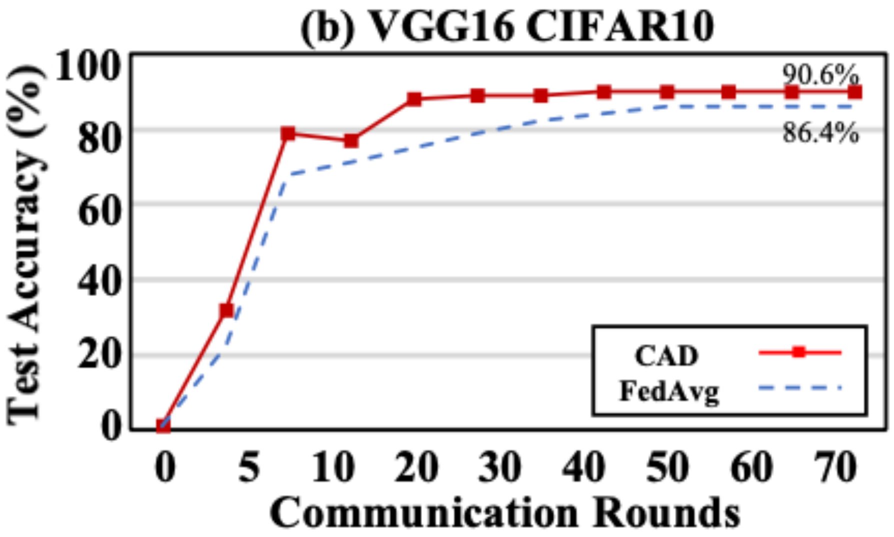


Preliminary Results

Convergence Evaluation



- Randomly select 5 classes from CIFAR10 dataset
- 1 Epoch Local Training



- High Convergence Speed
- High Convergence Accuracy

SAN FRANCISCO STATE UNIVERSITY

Conclusion

- We explored the context-aware distributed learning for CNN in mobile applications.

- We can re-configure the pre-defined CNN models into context-adaptive sub-models for on-device training.

- Experiment results show that our framework can achieve higher convergence speed as well as higher convergence accuracy.


Thank You!