# KnowledgeNet: Disaggregated and Distributed Training and Serving of Deep Neural Networks

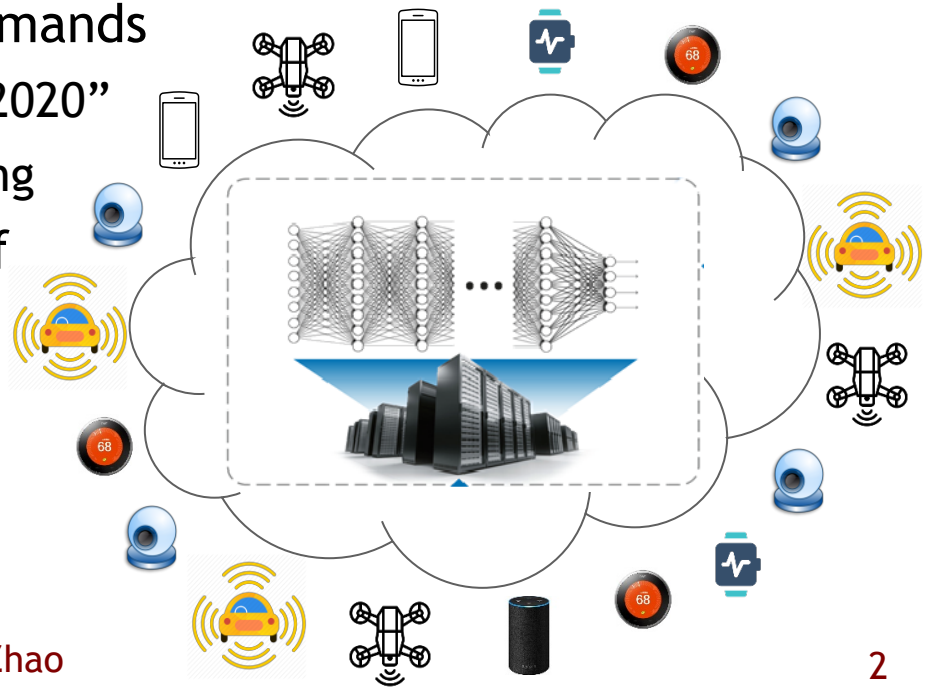
Saman Biookaghazadeh, Yitao Chen, Kaiqi Zhao,
***Ming Zhao***

Arizona State University

http://visa.lab.asu.edu

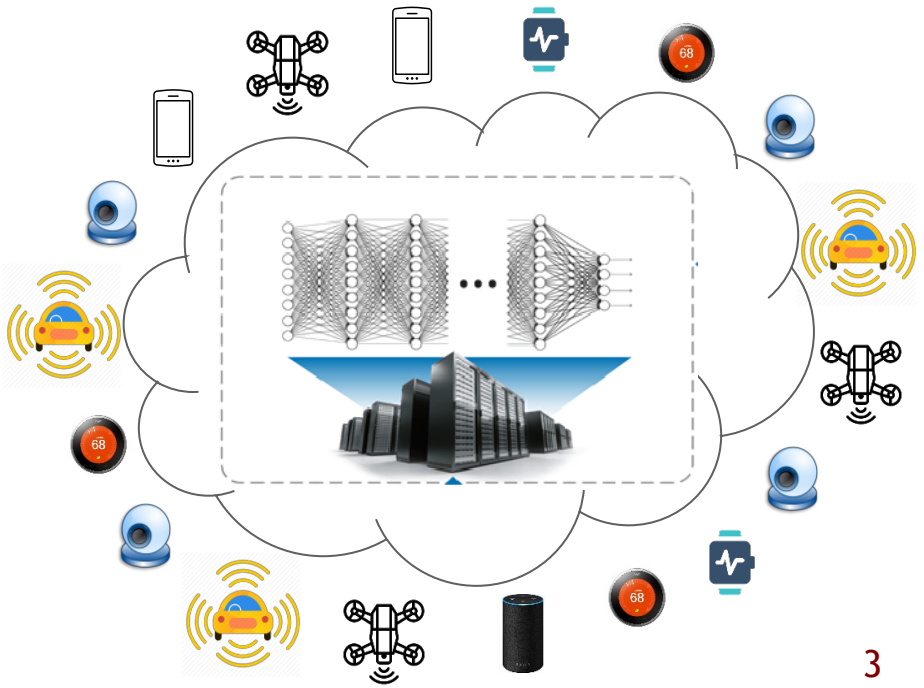
# Background

- Deep neural networks (DNNs) haven shown great potential
  - Solving many challenging problems
  - Relying on large systems to learn on big data
- Limitations of centralized learning
  - Cannot scale to handle future demands
    - “20 billion connected devices by 2020”
    - Real-time learning/decision making
  - Cannot exploit the capabilities of the increasingly powerful edge
    - Multiprocessing, accelerators



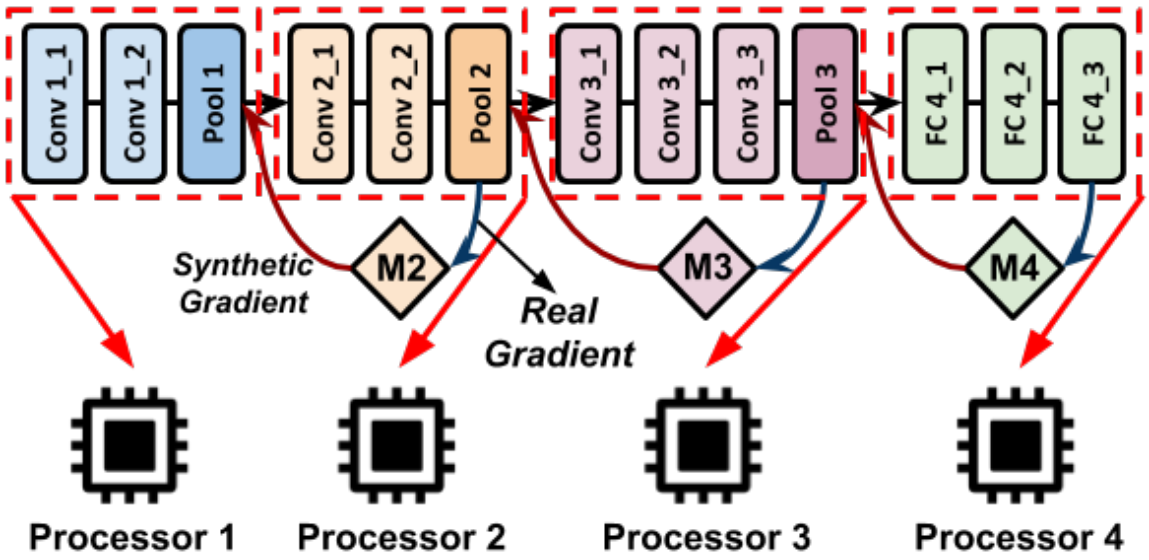
# Objectives

- Distributed learning across heterogeneous resources
  - Including diverse and potentially weak resources
  - Across edge devices and cloud resources
- Limitations of existing solutions
  - Data parallelism does not work for heterogeneous resources
  - Conventional backpropagation based training makes model parallelism difficult
  - Edge devices cannot train large models
  - Model compression works only for inference, not training



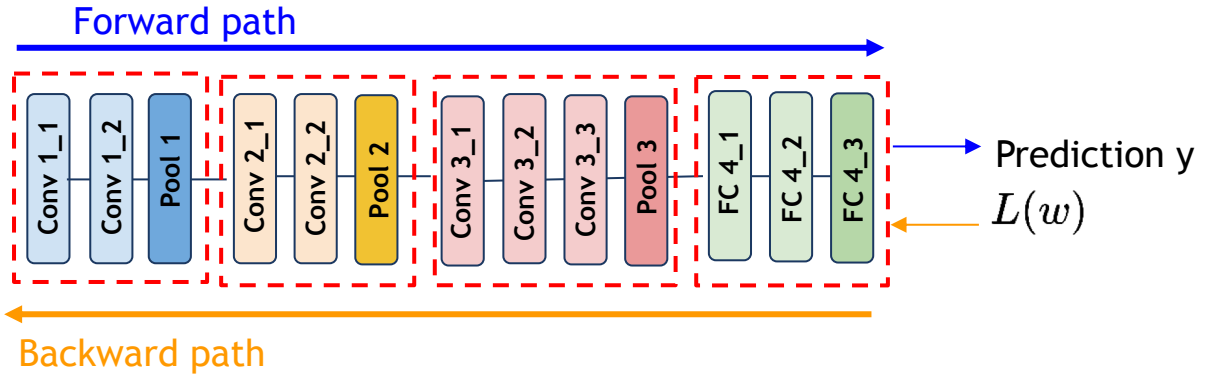
# Proposed approach

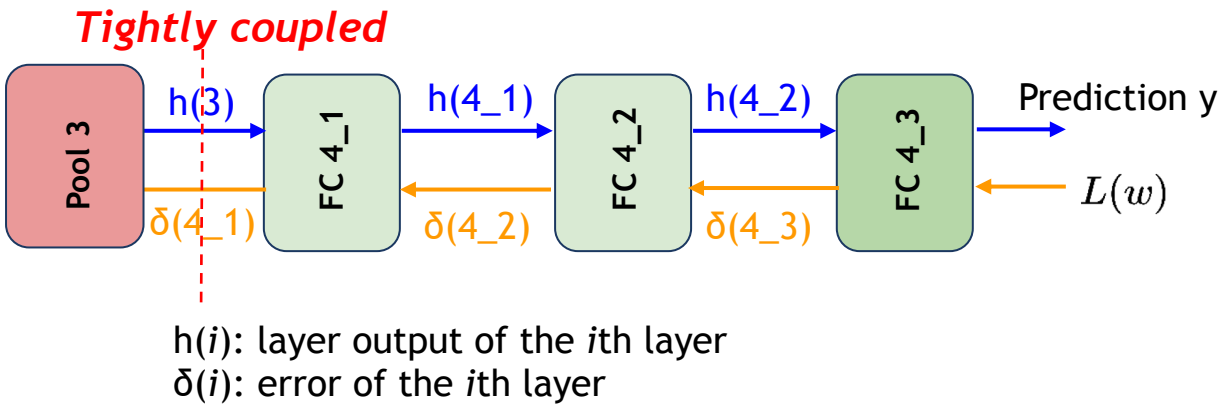
- Using synthetic gradients to achieve model parallelism
  - Split a large DNN model into many small models
  - Deploy small models on distributed and potentially weak resources
  - Use synthetic gradients to decouple the training of the small models



# Conventional training

- To training a neural network model
  - forward path: calculate loss of the class prediction
  - backward path: calculate the gradients of loss w.r.t the parameters



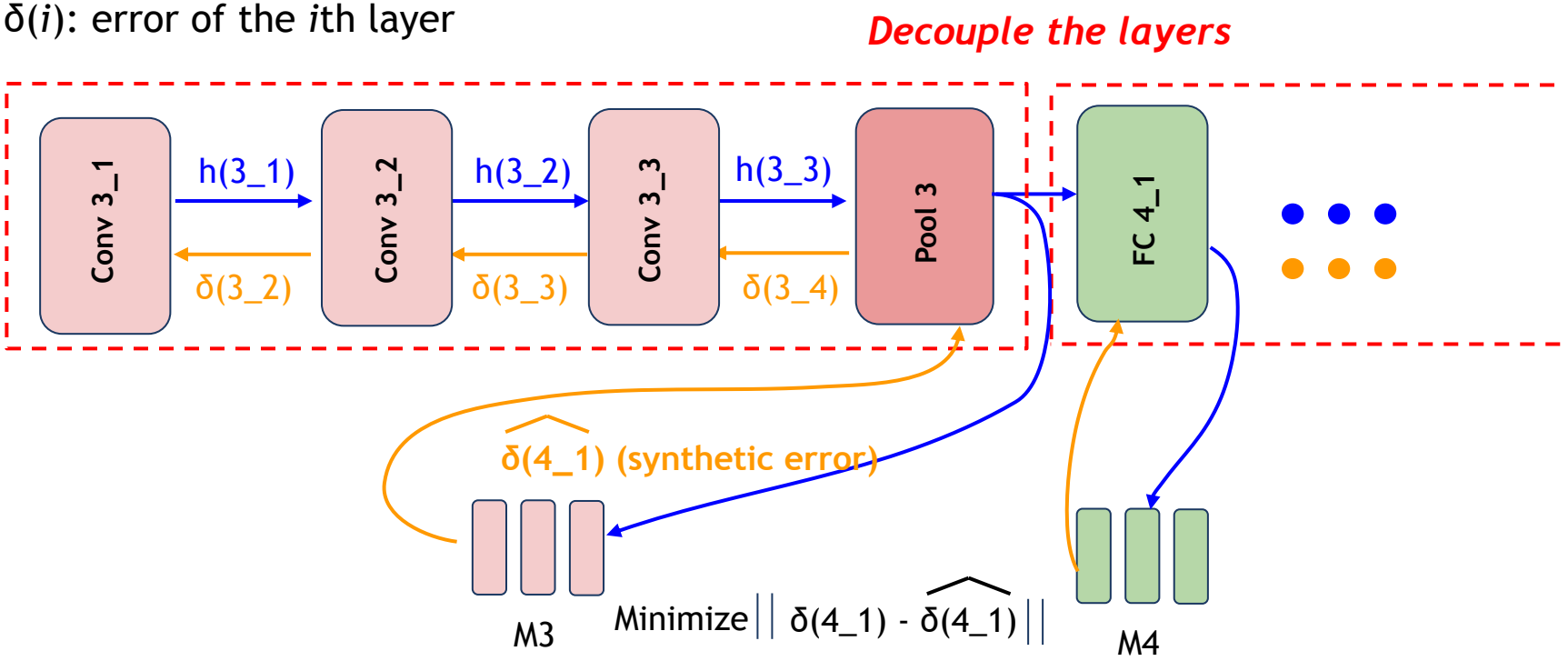




- Need tens of thousands iterations to achieve a high accuracy
- Layers are coupled and training are sequential
- Limited scalability in training

VISA

# Training with synthetic gradients

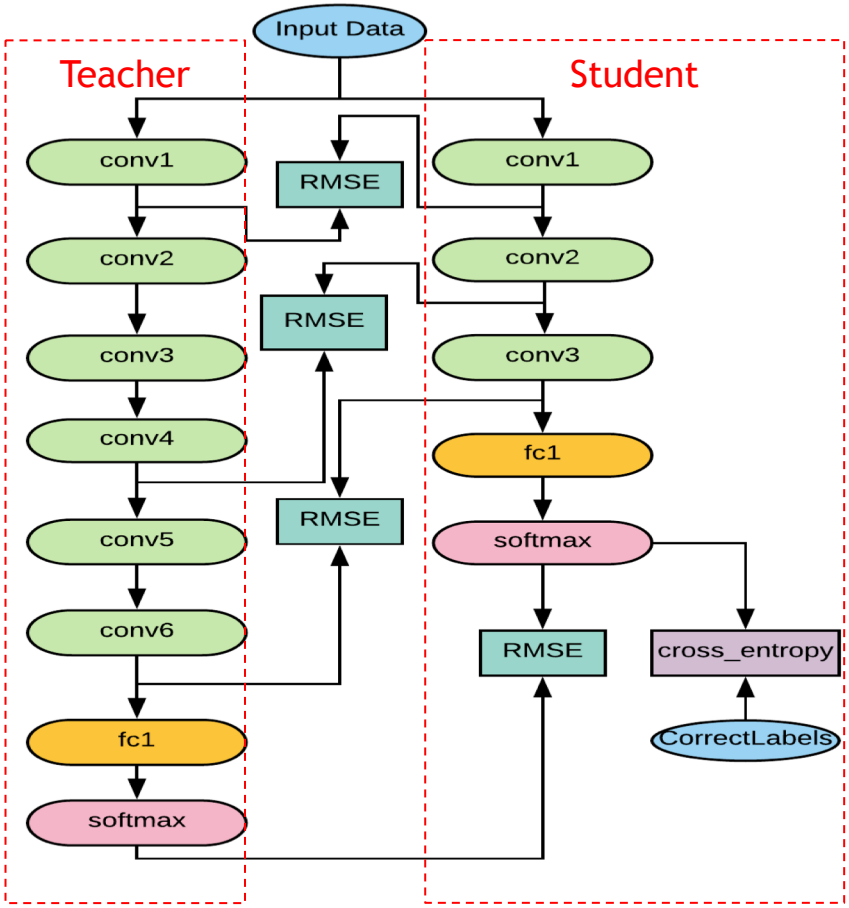
h(*i*): layer output of the *i*th layer
δ(*i*): error of the *i*th layer



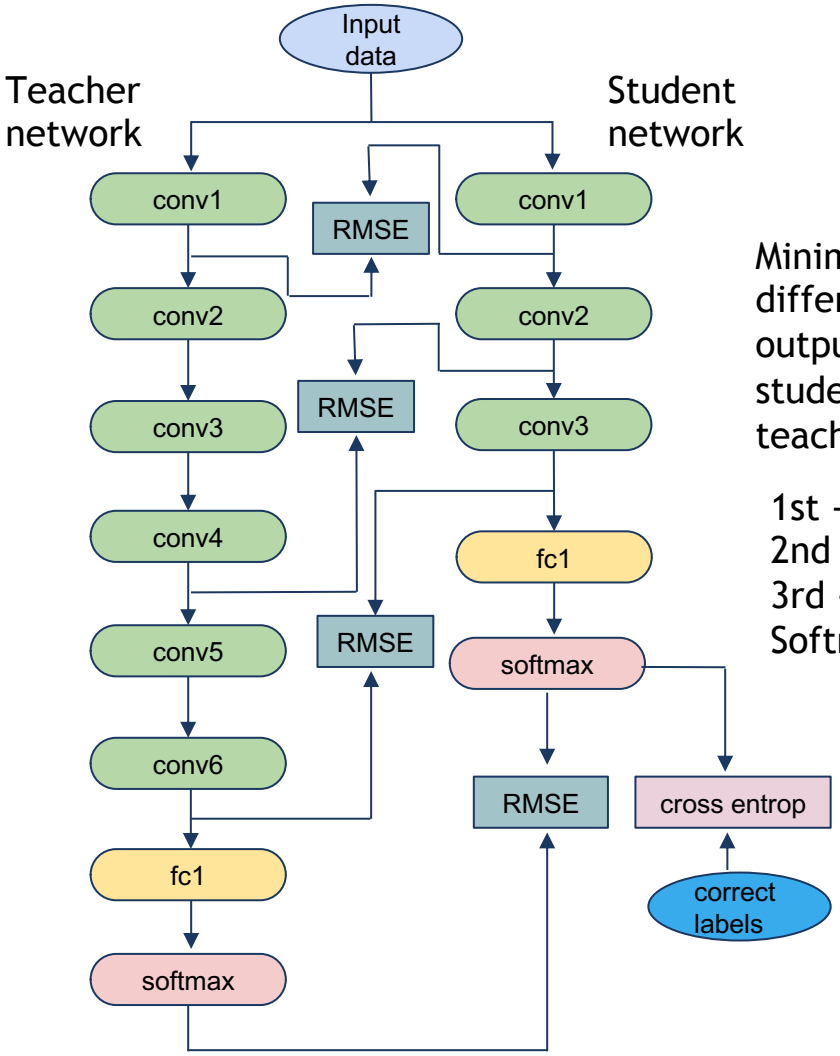

- Use another small neural network to predict the error
  →synthetic gradients
- Use synthetic gradients to train the small model instead of waiting for the actual gradients from backpropagation

# Proposed approach

- Using knowledge transfer to improve training on weak resources
  - Train a large, teacher model in the cloud
  - Train many small, student models on the devices
  - Exploit the knowledge of the teacher to train the students
  - Improve the accuracy and convergence speed of the students



# Knowledge transfer




Minimize squared difference between layer outputs (RMSE) to allow student to learning from teacher representations

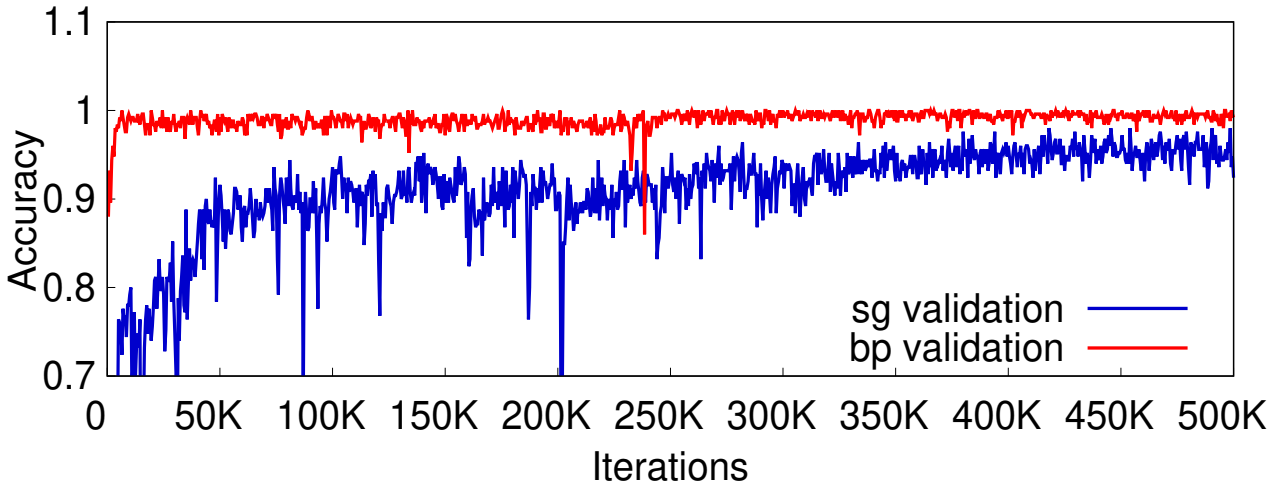
1st -> 1st
2nd -> 4th
3rd -> 6th
Softmax -> Softmax

# Preliminary results

- Using synthetic gradients to achieve model parallelism
  - Model: a simple 4-Layer
    - One convolution layer and three fully-connected
  - MNIST Dataset
  - 500K Iteration of training

**Training accuracy**

| | Back propagation | Synthetic gradients |
|---|---|---|
| **4-layer model** | 0.984 | 0.977 |
| **8-layer model** | 0.992 | 0.924 |
| **VGG16** | 0.994 | 0.845 |



# Preliminary results

- Using knowledge transfer to improve training on weak resources
  - Teacher Model: VGG-16 (8.5M parameters)
  - Student Model: miniaturized VGG-16 (3.2M parameters)
  - Training student for 100 Epochs

| | Accuracy |
|---|---|
| **Teacher** | 76.88% |
| **Student (dependent)** | 74.12% |
| **Student (independent)** | 61.24% |

"Are Existing Knowledge Transfer Techniques Effective For Deep Learning on Edge Devices?"
R. Sharma, S. Biookaghazadeh, and M. Zhao, EDGE, 2018

# Conclusions and future work

- Rethink DNN platforms to handle future learning needs
  - Utilize heterogeneous resources to train DNNs
  - Distribute learning across edge and cloud
- Achieve efficient model parallelism
  - Use synthetic gradients to decouple the training of distributed models
  - Need to further improve its accuracy for complex models
- Overcome the resource constraints of edge devices
  - Use knowledge transfer to help train on-device models
  - Need to further study its effectiveness under scenarios with limited data and limited supervision

# Acknowledgement


- National Science Foundation
  - GEARS project: CNS-1629888
  - CNS-1619653, CNS-1562837, CNS-1629888, CMMI-1610282, IIS-1633381


GEARS
ENERGY-EFFICIENT BIG-DATA RESEARCH SYSTEM

- VISA Lab @ SU
  - Saman, Yitao, Kaiqi, and others


- Thank you!
  - *Questions and suggestions?*
  - *Come to our poster at Happy Hour!*