CERIAS
The Center for Education and Research in Information Assurance and Security

# Modular Neural Networks for Low-Power Computer Vision

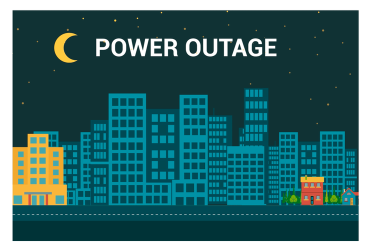
Abhinav Goel, Aniesh Chawla, Sara Aghajanzadeh, Caleb Tung, George K. Thiruvathukal, Yung-Hsiang Lu and Shuo-Han Chen

Text

## Motivation

- Need for mobile systems that can implement computer vision algorithms; drones, surveillance cameras.

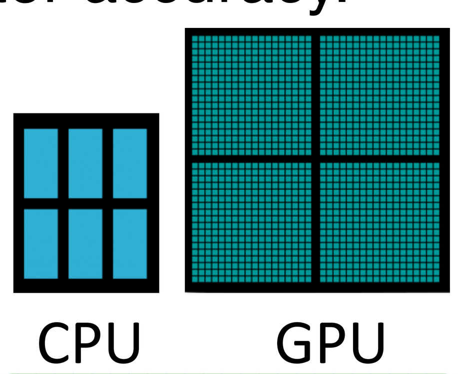


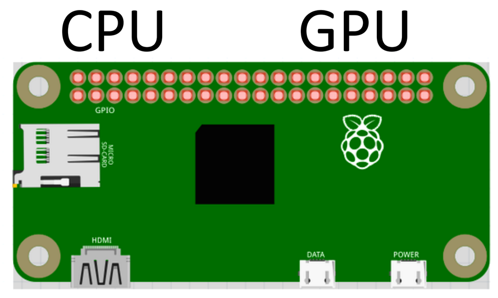

- Low-power solutions can be deployed away from electricity grid; critical in case of an infrastructure attack.

## Challenges

- Bigger Deep Neural Networks (DNNs) = Better accuracy.
- Big DNNs perform millions of operations: computation and memory accesses, need high power CPUs and GPUs.




- Embedded devices like Raspberry Pi Zero, cost only $5, with limited memory and compute capability. Can't run DNNs.

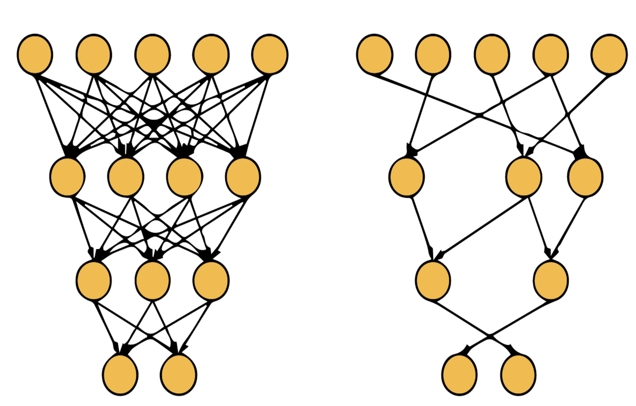


Raspberry Pi Zero

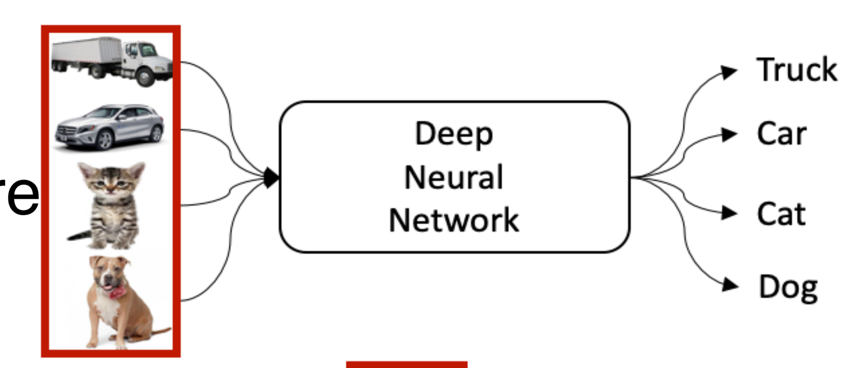
- DNNs are not designed for battery-powered devices.

## Where can we improve?

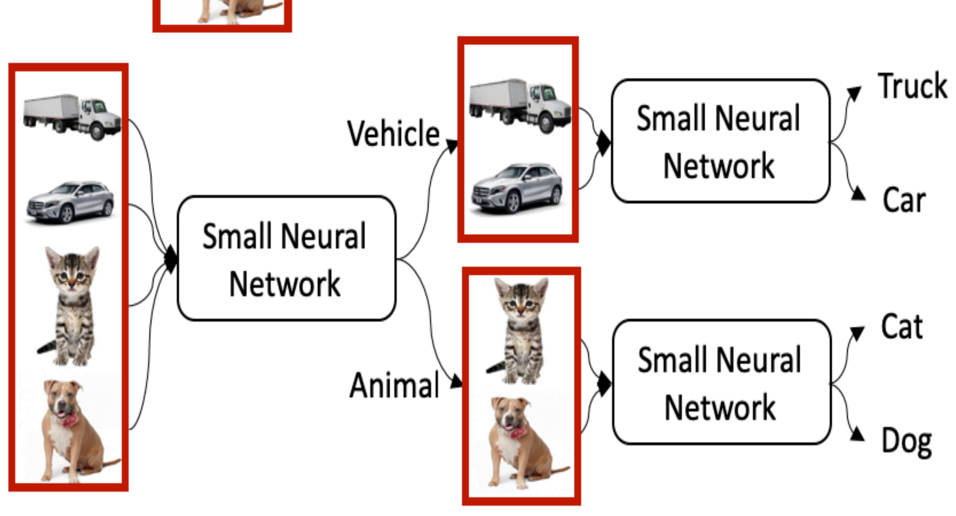
- Recent work has shown that DNNs have several redundancies!
- DNNs need millions of parameters to identify different operations.



- During inference only small subset of neurons are used.
- **Conventional DNN**: perform many different tasks which require a large number of neurons.

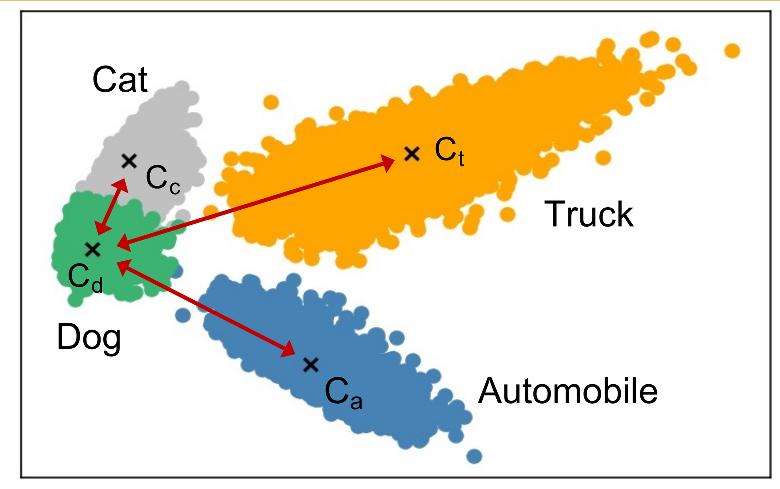



- **MNN-Tree**: The input image is processed incrementally using small DNNs. After detecting the type of images, finer classifications are made.




## Building the tree

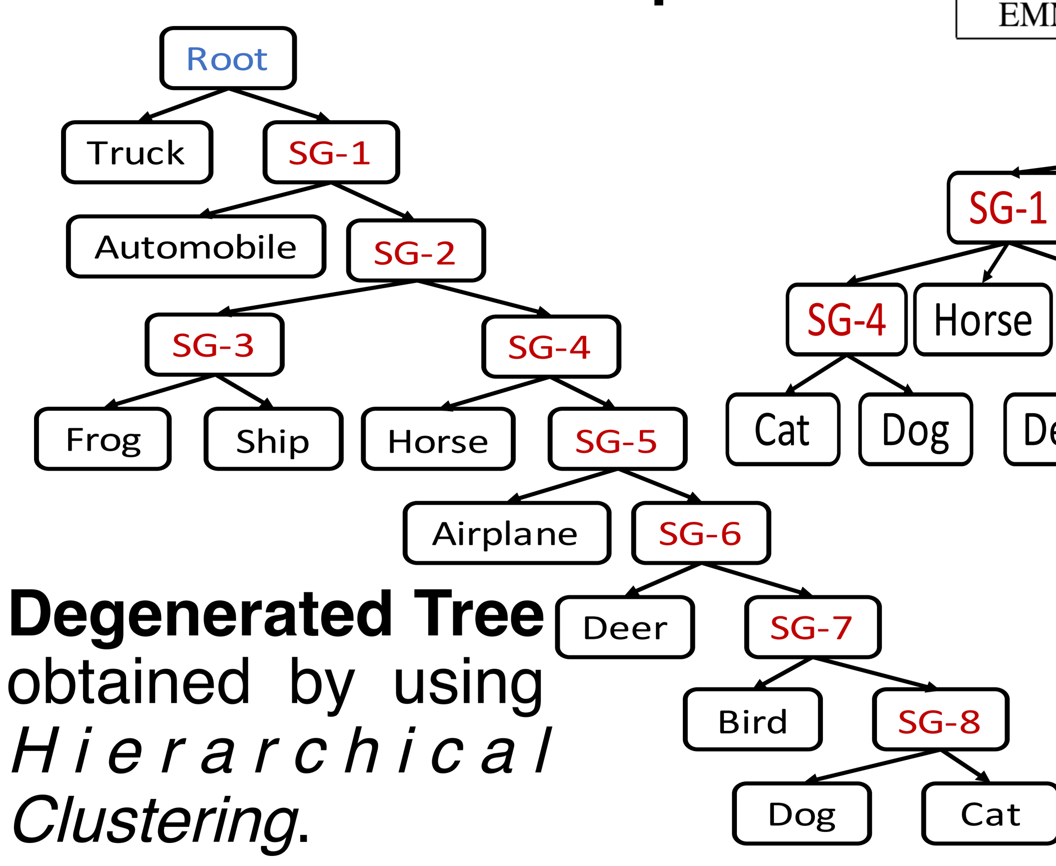
- Categories of the dataset are grouped together based on their similarity with each other.
- How to quantify the similarity?



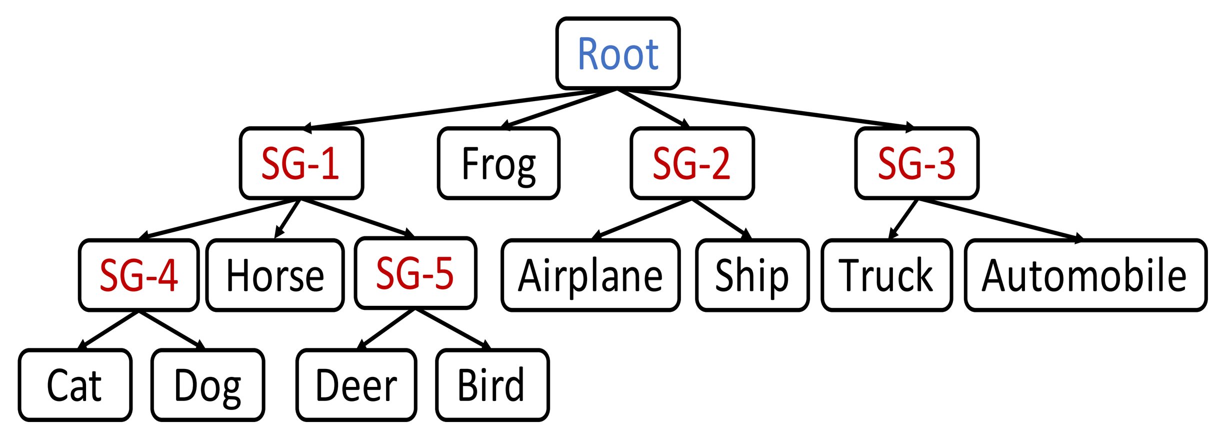

*Cat* and *Dog* are similar?

- Use distances between centroids of categories compared with a single threshold? **Not possible.**

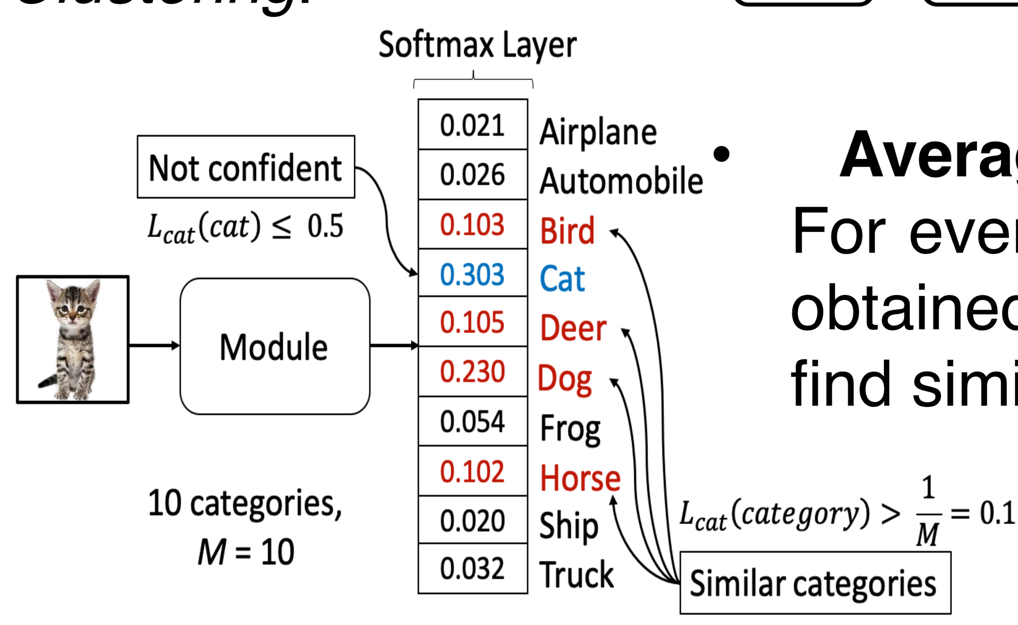
| Dataset | Avg. Distance Between Cluster | Min Distance Between Clusters | Max Distance Between Clusters |
|---|---|---|---|
| CIFAR-10 | 15.59 | 12.41 | 21.38 |
| CIFAR-100 | 15.77 | 6.12 | 21.74 |
| SVHN | 22.04 | 15.45 | 28.00 |
| EMNIST | 0.85 | 0.31 | 1.16 |




**Degenerated Tree** obtained by using *Hierarchical Clustering*.




Tree obtained by using the ***Averaged Softmax Likelihood***. This tree is the best configuration for reduced energy consumption.

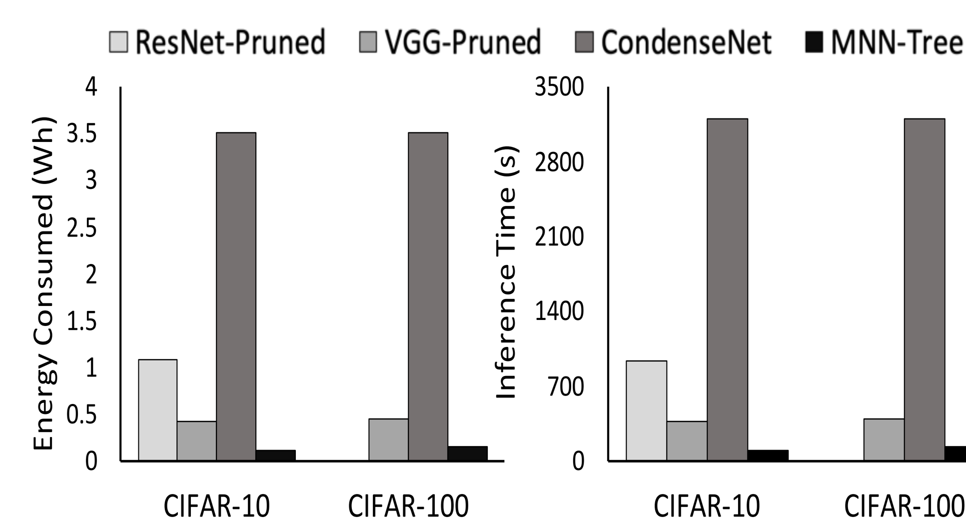



- **Averaged Softmax Likelihood:** For every category, use the values obtained at a DNN's output layer to find similarity.

$$L_A(B) = \frac{\sum_{i=0}^{|A|} softmax_A(B)}{|A|}$$

## Results

| Dataset | Technique | Model Size (KB) | Number of Operations | Val. Error |
|---|---|---|---|---|
| CIFAR-10 | VGG-Pruned | 17,000 | 2,060 M | 0.066 |
| | ResNet-Pruned | 3,400 | 1,120 M | 0.069 |
| | DenseNet | 4,200 | 9,388 M | 0.070 |
| | CondenseNet | 11,000 | 1,080 M | 0.034 |
| | Wide ResNet | 1,400 | 5,248 M | 0.040 |
| | **MNN-Tree** | 390 | 33 M | 0.079 |
| CIFAR-100 | VGG-Pruned | 17,010 | 2,060 M | 0.252 |
| | DenseNet | 4,200 | 9,388 M | 0.171 |
| | CondenseNet | 11,000 | 1,080 M | 0.184 |
| | Wide ResNet | 1,600 | 5,248 M | 0.192 |
| | **MNN-Tree** | 750 | 22 M | 0.209 |
| SVHN | DenseNet | 4,200 | 9,388 M | 0.017 |
| | Wide ResNet | 1,400 | 5,248 M | 0.016 |
| | **MNN-Tree** | 250 | 28 M | 0.019 |
| EMNIST | EDEN | - | - | 0.117 |
| | **MNN-Tree** | 460 | 58 M | 0.078 |




| Time | ResNet | VGG | CondenseNet | MNN-Tree |
|---|---|---|---|---|
| Load | 0.320 | 0.400 | 1.340 | 0.039 |
| Execution | 0.900 | 0.160 | 4.860 | 0.160 |

**Reduced** model size by **53%-97%**, energy by **67%-95%**, inference time by **66%-96%**, number of operations by **96%-99%**

PURDUE UNIVERSITY Discovery Park
PURDUE UNIVERSITY
AM²
LOYOLA UNIVERSITY CHICAGO
ACADEMIA SINICA
CERIAS