Designing Energy-Efficient Convolutional Neural Networks using Energy-Aware Pruning

Tien-Ju Yang, Yu-Hsin Chen, Vivienne Sze
Massachusetts Institute of Technology

CNN – Accurate but High Energy Consumption
- High energy consumption hinders CNN deployment on battery-powered devices
- We propose an energy-aware pruning algorithm for CNNs that directly targets energy rather than number of weights
- We perform energy analysis on various CNNs and provide insights

Energy Estimation Methodology
- Energy consumption: a combination of # of memory accesses and # of MACs (energy model from hardware measurements*)
- Energy model considers bit-width and sparsity, as well as all data types (weights and feature maps)

Energy Consumption Analysis of CNNs
- Number of weights is not a good estimator of energy
  - Example 1: CONV layers consume more energy than FC layers
  - Example 2: deeper CNNs with fewer weights do not necessarily consume less energy than shallower CNNs with more weights

Energy-Aware Pruning Algorithm
- Reduce the energy consumption by pruning a network
- Focus on minimizing the output error instead of the filter error

Reducing Number of Target Classes
- Key observation: by reducing the number of target classes on AlexNet, the model size is greatly reduced but the energy reduction is limited

Energy Consumption Analysis of CNNs
- Number of weights is not a good estimator of energy
  - Example 1: CONV layers consume more energy than FC layers
  - Example 2: deeper CNNs with fewer weights do not necessarily consume less energy than shallower CNNs with more weights

Network Comparison (Energy vs. Accuracy)
- Our pruned networks achieve better accuracy-energy trade-off
- Feature maps need to be factored in when estimating energy

Pruning Results for AlexNet and GoogLeNet
- 3.7x and 1.6x energy reduction
- 10.6x and 3.0x number of non-zero weights reduction
- 6.6x and 3.4x number of non-skipped MACs reduction

Network Comparison (Energy vs. Accuracy)
- Our pruned networks achieve better accuracy-energy trade-off
- Feature maps need to be factored in when estimating energy

Eyeriss
- The tool is available on the website
- CNN Shape Configuration (# of channels, # of filters, etc.)
- CNN Weights and Input Data (9, 5, 64, 97, 1, 1, ...)

* Eyeriss (CNN accelerator)

Energy Analysis of CNNs
- Example 1: 50x less # of MACs

CNN Energy Consumption
- Hardware Energy Costs of each MAC and Memory Access
- Mem. Hierarchy Example
- CNN Energy Consumption

Replacing FC with CONV
- CONV layers consume less energy than FC layers

Reducing Number of Non-Zero Weights
- Zero weights reduction
- Skipped MACs reduction

Accesses
- # of MACs
- # acc. at mem. level
- # acc. at mem. level
- # of MACs
- # of MACs

Pruning Results for AlexNet and GoogLeNet
- Model
- Top-5 Accuracy
- # of Non-zero Weights
- # of Non-skipped MACs
- Normalized Energy

Network Comparison (Energy vs. Accuracy)
- Our pruned networks achieve better accuracy-energy trade-off
- Feature maps need to be factored in when estimating energy

Normalized Energy Consumption
- Original
- DC
- Energy-aware pruning (This Work)