Once for All: Train One Network and Specialize it for Efficient Deployment

Han Cai, Chuang Gan, Tianzhe Wang, Zhekai Zhang, Song Han

Massachusetts Institute of Technology
Challenge: Efficient Inference on Diverse Hardware Platforms

- Cloud AI
  - Memory: 32GB
  - Computation: $10^{12}$ FLOPS
- Mobile AI
  - Memory: 4GB
  - Computation: $10^9$ FLOPS
- Tiny AI (AIoT)
  - Memory: 100 KB
  - Computation: $<10^6$ FLOPS

Different hardware platforms have different resource constraints. We need to customize our models for each platform to achieve the best accuracy-efficiency trade-off, especially on resource-constrained edge devices.
The design cost is calculated under the assumption of using MnasNet.
Challenge: Efficient Inference on Diverse Hardware Platforms

Diverse Hardware Platforms

- 2019
- 2017
- 2015
- 2013

Design Cost (GPU hours)

- 40K
- 160K

The design cost is calculated under the assumption of using MnasNet.
Challenge: Efficient Inference on Diverse Hardware Platforms

Diverse Hardware Platforms

Cloud AI (10^{12} FLOPS)  Mobile AI (10^{9} FLOPS)  Tiny AI (10^{6} FLOPS)

The design cost is calculated under the assumption of using MnasNet.
Challenge: Efficient Inference on Diverse Hardware Platforms

Diverse Hardware Platforms

Cloud AI ($10^{12}$ FLOPS)  
Mobile AI ($10^9$ FLOPS)  
Tiny AI ($10^6$ FLOPS)

Design Cost (GPU hours)

- 40K → 11.4k lbs CO$_2$ emission
- 160K → 45.4k lbs CO$_2$ emission
- 1600K → 454.4k lbs CO$_2$ emission

1 GPU hour translates to 0.284 lbs CO$_2$ emission according to Strubell, Emma, et al. "Energy and policy considerations for deep learning in NLP.” ACL. 2019.
Challenge: Efficient Inference on Diverse Hardware Platforms

Diverse Hardware Platforms

Cloud AI ($10^{12}$ FLOPS)  Mobile AI ($10^9$ FLOPS)  Tiny AI ($10^6$ FLOPS)

Design Cost (GPU hours)

- 40K → 11.4k lbs CO$_2$ emission
- 160K → 45.4k lbs CO$_2$ emission
- 1600K → 454.4k lbs CO$_2$ emission

1 GPU hour translates to 0.284 lbs CO$_2$ emission according to Strubell, Emma, et al. "Energy and policy considerations for deep learning in NLP." ACL. 2019.
Challenge: Efficient Inference on Diverse Hardware Platforms

Diverse Hardware Platforms

Cloud AI ($10^{12}$ FLOPS)  Mobile AI ($10^9$ FLOPS)  Tiny AI ($10^6$ FLOPS)

Design Cost (GPU hours)

- 40K $\rightarrow$ 11.4k lbs CO$_2$ emission
- 160K $\rightarrow$ 45.4k lbs CO$_2$ emission
- 1600K $\rightarrow$ 454.4k lbs CO$_2$ emission

Once-for-All Network

1 GPU hour translates to 0.284 lbs CO$_2$ emission according to Strubell, Emma, et al. "Energy and policy considerations for deep learning in NLP." ACL. 2019.
Once-for-All Network: Decouple Model Training and Architecture Design
Once-for-All Network: Decouple Model Training and Architecture Design
Once-for-All Network: Decouple Model Training and Architecture Design
Once-for-All Network: Decouple Model Training and Architecture Design
Progressive Shrinking for Training OFA Networks

• More than $10^{19}$ different sub-networks in a single once-for-all network, covering 4 different dimensions: resolution, kernel size, depth, width.

• Directly optimizing the once-for-all network from scratch is much more challenging than training a normal neural network given so many sub-networks to support.
• More than $10^{19}$ different sub-networks in a single once-for-all network, covering 4 different dimensions: resolution, kernel size, depth, width.

• Directly optimizing the once-for-all network from scratch is much more challenging than training a normal neural network given so many sub-networks to support.

• Small sub-networks are nested in large sub-networks.

• Cast the training process of the once-for-all network as a progressive shrinking and joint fine-tuning process.
Connection to Network Pruning

- **Network Pruning**
  - Train the full model
  - Shrink the model (only width)
  - Fine-tune the small net
  - Single pruned network

- **Progressive Shrinking**
  - Train the full model
  - Shrink the model (4 dimensions)
  - Fine-tune both large and small sub-nets
  - Once-for-all network

- Progressive shrinking can be viewed as a generalized network pruning with much higher flexibility across 4 dimensions.
Randomly sample input image size for each batch

Progressive Shrinking

- Elastic Resolution
- Elastic Kernel Size
- Elastic Depth
- Elastic Width
Progressive Shrinking

Start with full kernel size
Smaller kernel takes centered weights via a transformation matrix
Progressive Shrinking

Gradually allow later layers in each unit to be skipped to reduce the depth.
Progressive Shrinking

Gradually shrink the width

Keep the most important channels when shrinking via channel sorting
Performances of Sub-networks on ImageNet

Sub-networks under various architecture configurations
  D: depth, W: width, K: kernel size

- Progressive shrinking consistently improves accuracy of sub-networks on ImageNet.
OFA: 80% Top-1 Accuracy on ImageNet

- Once-for-all sets a new state-of-the-art **80% ImageNet top-1 accuracy** under the mobile setting (< 600M MACs).
Comparison with EfficientNet and MobileNetV3

- Once-for-all is **2.6x faster** than EfficientNet and **1.5x faster** than MobileNetV3 on Google Pixel1 without loss of accuracy.
OFA for Fast Specialization on Diverse Hardware Platforms

- **OFA**
- **MobileNetV3**
- **MobileNetV2**

### Samsung S7 Edge

- **Latency (ms):**
  - 25: 67.4, 70.5, 73.1, 74.7, 76.3
  - 100: 75.2
- **Top-1 ImageNet Acc (%):**
  - 25: 67, 70, 73, 74, 76
  - 100: 77

### Google Pixel2

- **Latency (ms):**
  - 23: 67, 69, 71, 73, 75
  - 68: 77
- **Top-1 ImageNet Acc (%):**
  - 23: 70, 73, 74, 75, 76
  - 68: 77

### LG G8

- **Latency (ms):**
  - 7: 67, 69, 71, 73, 75
  - 25: 77
- **Top-1 ImageNet Acc (%):**
  - 7: 70, 73, 74, 75, 76
  - 25: 77

### NVIDIA 1080Ti

- **Latency (ms):**
  - 10: 60.3, 65.4, 69.8, 72.6
  - 30: 77
- **Batch Size = 64**

### Intel Xeon CPU

- **Latency (ms):**
  - 9: 60.3, 65.4, 69.8, 71.1
  - 19: 77
- **Batch Size = 1**

### Xilinx ZU3EG FPGA

- **Latency (ms):**
  - 3.0: 59.1, 63.3, 69.6
  - 8.0: 77
- **Batch Size = 1 (Quantized)**
OFA Saves Orders of Magnitude Design Cost

- Geen AI is important. The computation cost of OFA stays constant with #hardware platforms, reducing the carbon footprint by 1,335x compared to MnasNet under 40 platforms.
OFA for FPGA Accelerators

Measured results on XILINX FPGA

• Non-specialized neural networks do not fully utilize the hardware resource. There is a large room for improvement via neural network specialization.
Summary

• We introduce once-for-all network for **efficient inference on diverse hardware platforms**.
• We present an effective **progressive shrinking** approach for training once-for-all networks.

**Progressive Shrinking**

- Train the full model
- Shrink the model in 4 dimensions
- Fine-tune both large and small sub-nets
- Once-for-all network

• Once-for-all network surpasses MobileNetV3 and EfficientNet by a large margin under all scenarios, setting a new state-of-the-art **80% ImageNet Top1-accuracy** under the mobile setting (< 600M MACs).
  • **First place** in the 3rd Low-Power Computer Vision Challenge, DSP track @ICCV’19
  • **First place** in the 4th Low-Power Computer Vision Challenge @NeurIPS’19

• Released **50 different pre-trained OFA models** on diverse hardware platforms (CPU/GPU/FPGA/DSP).
  ```python
  net, image_size = ofa_specialized(net_id, pretrained=True)
  ```

• Released the **training code & pre-trained OFA network** that provides diverse sub-networks without training.
  ```python
  ofa_network = ofa_net(net_id, pretrained=True)
  ```

Project Page: [https://ofa.mit.edu](https://ofa.mit.edu)