About me

• Research interests:
  – Dense prediction tasks
  – Efficient model training
  – Self-supervise/unsupervised training
  – Robust models in the wild
Content

• The power of large model
  – Increased model size
  – Increased labeled training dataset
  – Multimodality

• Efficient model training
  – Knowledge distillation
  – Network pruning/ Quantization
Deploying highly efficient, compact models on edge devices (e.g., AIoT)
Content

- The power of large model
  - Increased model size
  - Increased labeled training dataset
  - Multimodality

- Efficient model training
  - Knowledge distillation
  - Network pruning/Quantization
Knowledge Distillation

- Knowledge distillation for classification
  - Soften output
  - Compact model (student) learns from large models (teacher)
Knowledge Distillation

- Knowledge distillation for classification
  - Soften output
  - Compact model (student) learns from large models (teacher)

Hard label: 1 0 0
Soft target: 0.8 0.19 0.01
Knowledge distillation for semantic segmentation

Baseline: applying KD to each pixel on the logits

$$D_{KL}(P \| Q) = \sum_{x \in \mathcal{X}} P(x) \log \left( \frac{P(x)}{Q(x)} \right).$$
Structural Knowledge Distillation

- Ours: Knowledge distillation considering structural correlations

  Idea 1: Learn from correlations among spatial locations
  ✓ Pair-wise
  ✓ Holistic

Structured knowledge distillation for semantic segmentation, Liu, Yifan and Chen, Ke and Liu, Chris and Qin, Zengchang and Luo, Zhenbo and Wang, Jingdong, 2019, CVPR

Structured Knowledge Distillation for Dense Prediction, Liu, Yifan and Shu, Changyong and Wang, Jingdong and Shen, Chunhua, TPAMI, 2020
Generator: try to generate fake distributions which is similar to the real ones, to fool the discriminator

Discriminator: try distinguish between the real distribution and the fake distribution

First proposed in image generation tasks
Analyzing and Improving the Image Quality of StyleGAN, Tero Karras et al. 2019
External: GAN
 Structural Knowledge Distillation

• Ours: Knowledge distillation considering structural correlations
  Idea1: Learn from correlations among spatial locations
  ✔ Pair-wise
  ✔ Holistic
Spatial distillation

• Mimic
  Minimize the L2 similarity among features
  A 1 x 1 convolution is employed to align the channel of the feature

• Attention transfer
  Get an attention map with one channel from the feature map.
  Merging all the channels into one channel.
Channel-wise Distillation

- Ours: Knowledge distillation considering the information in the channels.

Channel-wise Knowledge Distillation for Dense Prediction, Yifan Liu*, Changyong Shu*, Jianfei Gao, Zheng Yan, Chunhua Shen, ICCV, 2021
structure_knowledge_distillation

The official code for the paper 'Structured Knowledge Distillation for Semantic Segmentation'. (CVPR 2019 ORAL) and extension to other tasks.
Knowledge distillation on video frames

- Core idea:
  - Considering the correlations among frames during training, and **inference on single frames**:
    - Learning from a large **temporally consistent model**
    - Learning the correlations from a large **optical flow model**

Efficient semantic video segmentation with per-frame inference, Liu, Yifan et al, ECCV, 2020
Content

• The power of large model
  – Increased model size
  – Increased labeled training dataset
  – Multimodality

• Efficient model training
  – Knowledge distillation
  – Network pruning/ Quantization
Pruning Happens in Human Brain

50 Trillion Synapses → 1000 Trillion Synapses → 500 Trillion Synapses

Newborn → 1 year old → Adult
Pruning

1. Prune weights
   • setting individual parameters to zero and making the network sparse.

2. Remove entire nodes from the network
   • make the network architecture itself smaller, while aiming to keep the accuracy of the initial larger network
Channel pruning
Network Quantization

- 0.33323134411 → float point → range(0,1)
- 01011010 → int8 → range(-127,128)
- First, we normalize the weight of the network into the range (-127,128)
Network Quantization

• If the output of the network is \((X1,X2)\),
• For a weight \(x\), we can use

\[
\text{new}_w = \text{round}\left(\frac{(X2 - X1)}{255} \cdot x\right)
\]
Network Quantization

• 1. Training
• 2. Quantization
• 3. Retraining
Network Quantization

Challenges:

• Non-differentiable quantization functions (e.g., round, sign).
• Quantized structure needs to be re-designed.
• Large gap between theory and reality.
## Network Quantization

### How efficient?

NVIDIA INT8: >3x speedup vs. 32-bit

<table>
<thead>
<tr>
<th></th>
<th>CAFFE</th>
<th>TENSORRT FP32</th>
<th>TENSORRT INT8</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Runtime (ms)</strong></td>
<td>242</td>
<td>170</td>
<td>50</td>
</tr>
<tr>
<td><strong>Images/sec</strong></td>
<td>4</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td><strong>Class IoU</strong></td>
<td>48.4</td>
<td>48.4</td>
<td>48.1</td>
</tr>
<tr>
<td><strong>Category IoU</strong></td>
<td>76.9</td>
<td>76.9</td>
<td>76.8</td>
</tr>
</tbody>
</table>

Batch Size = 1, Input/Output Resolution = 512 x 1024
Summary

• Large model:
  – Powerful
  – Expensive
  – Inefficient
Summary

• Small model:
  – Hard to train
    • Knowledge distillation: Improve the performance
    • Pruning: change the model structure to reduce the size
    • Quantization: keep the structure and change the type of the weights of the network
Thanks!