Efficient Solution to Large-scale Image Classification

Presenter: Chenhao Lin
Team: BigVideo

16/06/2019
Team: BigVideo

Team Member:

Huabin Zheng  Litong Feng  Yuming Chen  Weirong Chen

Zhe Huang  Zhanbo Sun  Wayne Zhang
Results

Validation Top5

<table>
<thead>
<tr>
<th></th>
<th>2018 Winner final ensemble</th>
<th>Our single model</th>
<th>Our final ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>Validation Top5</td>
<td>79.8</td>
<td>80.79</td>
<td>81.91</td>
</tr>
</tbody>
</table>

Test Top5

<table>
<thead>
<tr>
<th></th>
<th>2018 Winner final ensemble</th>
<th>Our final ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Top5</td>
<td>79.25</td>
<td>82.05</td>
</tr>
</tbody>
</table>
Overview

Challenge:

Limited GPU resources
VS

Large-scale data
Idea Validation

Pipeline:

Many-model Ensemble

Model Selection
Efficient & Powerful Network Architectures

Fine-tuning
Large Input Size
Self-Supervised Loss
Using Description Text

Final Ensemble of Three Models

Multi-Crop Testing

Model Training (Starting from our in-house image classification tool)

ImageNet-style Training
# Efficient & Powerful Networks

<table>
<thead>
<tr>
<th>Network (Input Size)</th>
<th>ImageNet Top1</th>
<th>Estimated Training Time on WebVision*</th>
</tr>
</thead>
<tbody>
<tr>
<td>NASNet-A (331)</td>
<td>82.70</td>
<td>64 GPUs 67 days</td>
</tr>
<tr>
<td>PNASNet-5 (331)</td>
<td>82.90</td>
<td>64 GPUs 61 days</td>
</tr>
<tr>
<td>SENet154 (224)</td>
<td>81.32</td>
<td>64 GPUs 18 days</td>
</tr>
<tr>
<td><strong>ResNeXt152 variant (224)</strong> (Our Primary Model)</td>
<td><strong>81.53</strong></td>
<td><strong>64 GPUs 12 days</strong></td>
</tr>
<tr>
<td>Inception-ResNet-v2 (299)</td>
<td>80.10</td>
<td>64 GPUs 12 days</td>
</tr>
<tr>
<td>DPN98(224)</td>
<td>79.80</td>
<td>64 GPUs 11 days</td>
</tr>
<tr>
<td>SEResNet152(224)</td>
<td>78.43</td>
<td>64 GPUs 9 days</td>
</tr>
</tbody>
</table>

*Estimated training time for Webvision 150 epochs on TITANXp
Fine-tuning with Expanded Input Size

- Experience from ImageNet:
  - Larger input size performs better.
  - Due to limited resources, we fine-tune with large input sizes only.
- Generalized-Mean (GeM) pooling [1] adapts with large inputs better than global average pooling.

On-the-fly Self-supervised Loss

- After supervised training converges, pseudo labels from network itself are more reliable than noisy ground-truth labels.
Using Description Text

- Select samples by semantic similarity between embeddings of sample description text and label description text.
Despite of visually appealing selection, we found training from scratch with the selected partial training set did not perform as well as with the full training set. Nevertheless, partial-set model contributes to the final ensemble's performance.
Ensemble

- Single Model 1
- Multi Models

- Top5
- model 1
- model 1 + 2
- model 1 + 2 + 3

- from scratch
- large input size
- self-supervised
- gem
- multi-crop
Take-home Message

- Fundamental improvements of image classification bring large gains.
  - Efficient network with large capacity
  - Expanded input size + GeM pooling
  - On-the-fly self-supervised loss

- Side information may bring gains, however we did not have enough time and GPUs to explore them.
  - Description text based sample selection using BERT

- De-noising tricks are hard to tune well.
  - GHM
  - Focal loss
BigVideo Research Team of SenseTime

Dedicated to research on deep understanding of Internet photos & videos

- Holistic Semantic Understanding
  - People, Scene, Action, Event

- Big Data
  - 1 billion Images/Frames processed per day

- High Accuracy
  - 90% recall @ 1 / 1,000,000 FAR

- High Performance
  - 3000 QPS single GPU

50+ Researchers, 8 PhDs, 100+ Publications
Thank You!