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

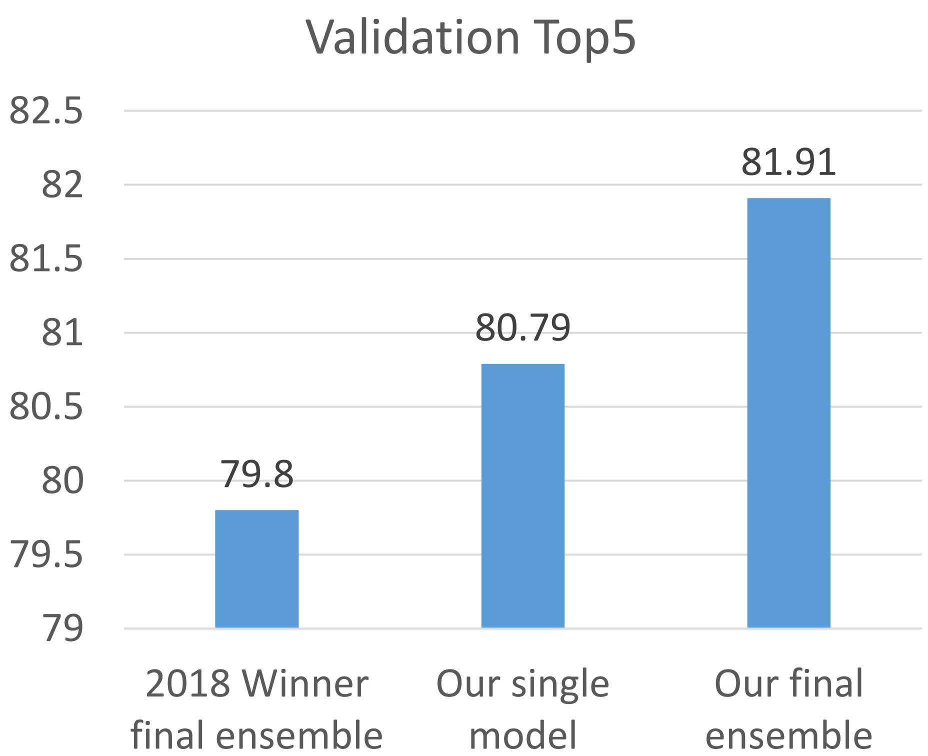
Weirong Chen

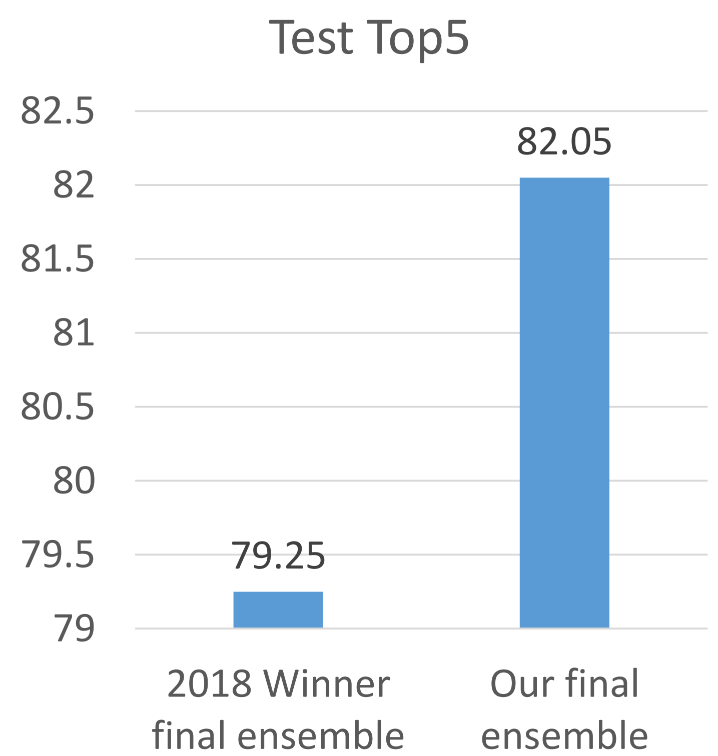
Zhe Huang

Zhanbo Sun

Wayne Zhang

Results


Validation Top5
82.5
82
81.5
81
80.5
80
79.5
79
79.8
80.79
81.91
2018 Winner final ensemble
Our single model
Our final ensemble


Test Top5
82.5
82
81.5
81
80.5
80
79.5
79
79.25
82.05
2018 Winner final ensemble
Our final ensemble

Overview

Challenge: Limited GPU resources

VS

Large-scale data | Idea Validation | Many-model Ensemble

Pipeline:

Model Selection

Efficient & Powerful Network Architectures

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

ImageNet-style Training

Fine-tuning

Large Input Size

Self-Supervised Loss

Using Description Text

Final Ensemble of Three Models

Multi-Crop Testing

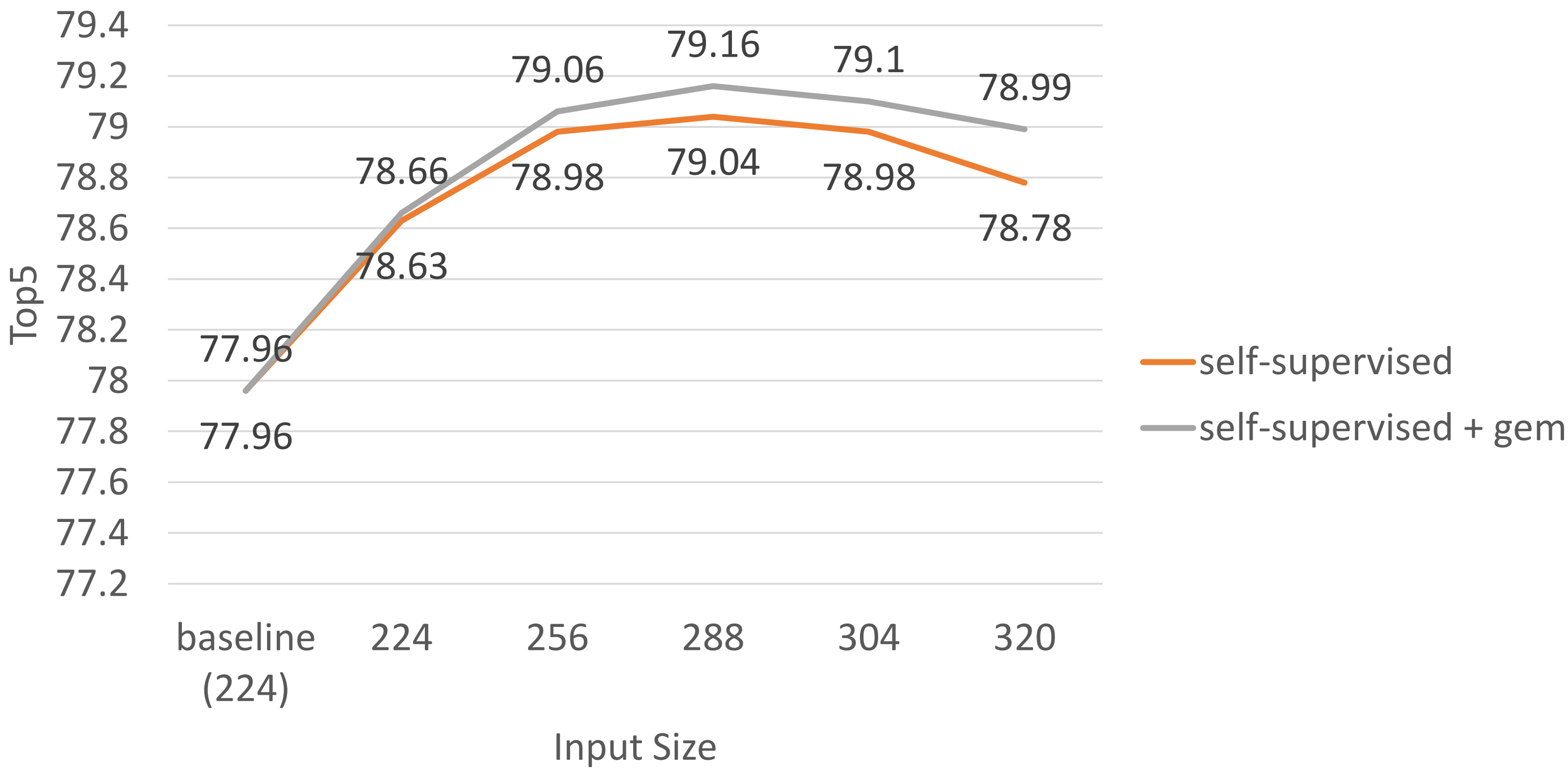
Efficient & Powerful Networks

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

*Estimated training time for Webvision 150 epochs on TITANXp

Fine-tuning with Expanded Input Size

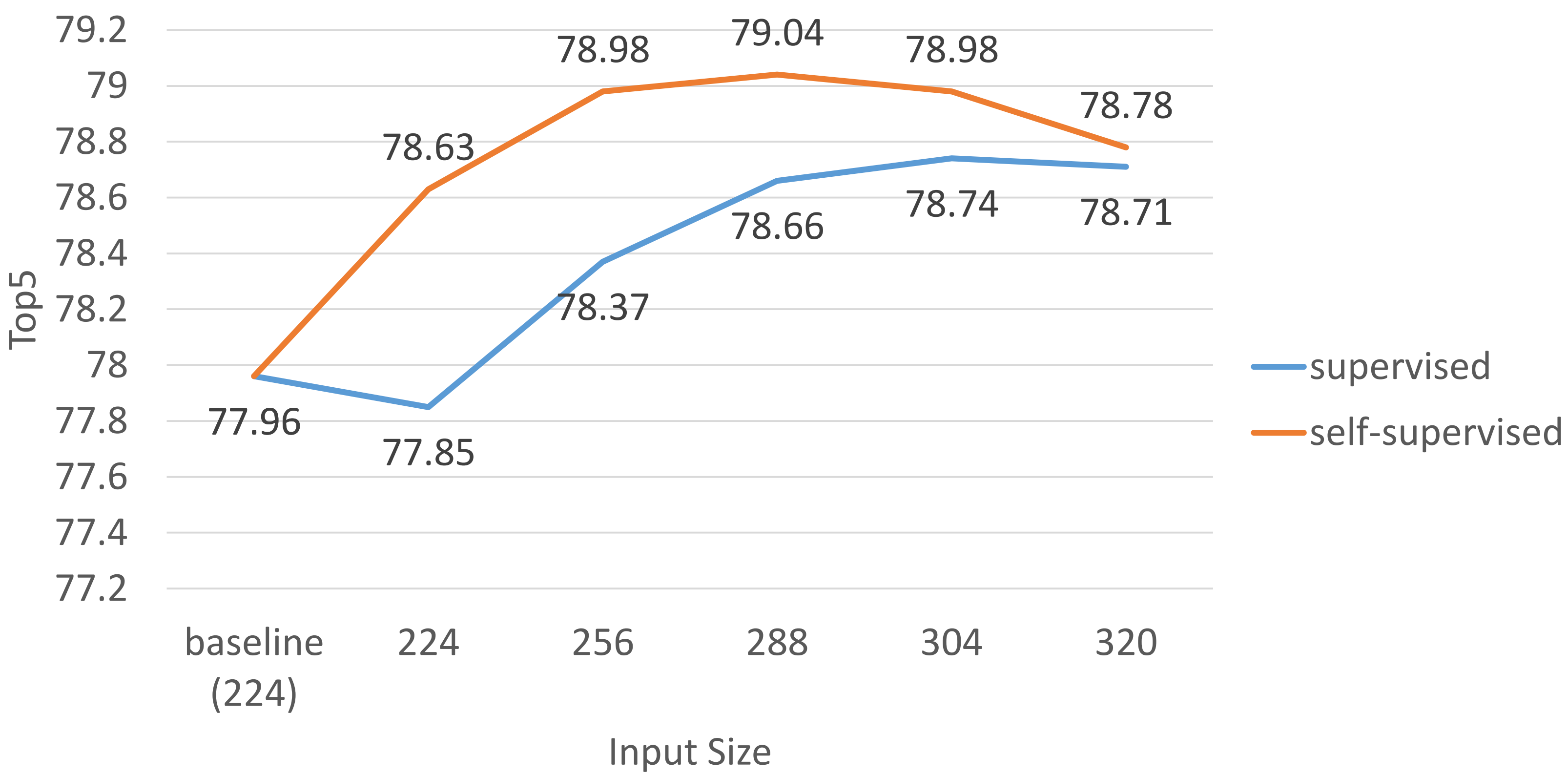
- ❑ Experience from ImageNet:
  - ❑ Larger input size performs better.
  - ❑ Due to limited resources, we fine-tune with large input sizes only.
- ❑ Generalized-Mean (GeM) pooling [1] adapts with large inputs better than global average pooling.



[1] Berman, Maxim, et al. "MultiGrain: a unified image embedding for classes and instances."

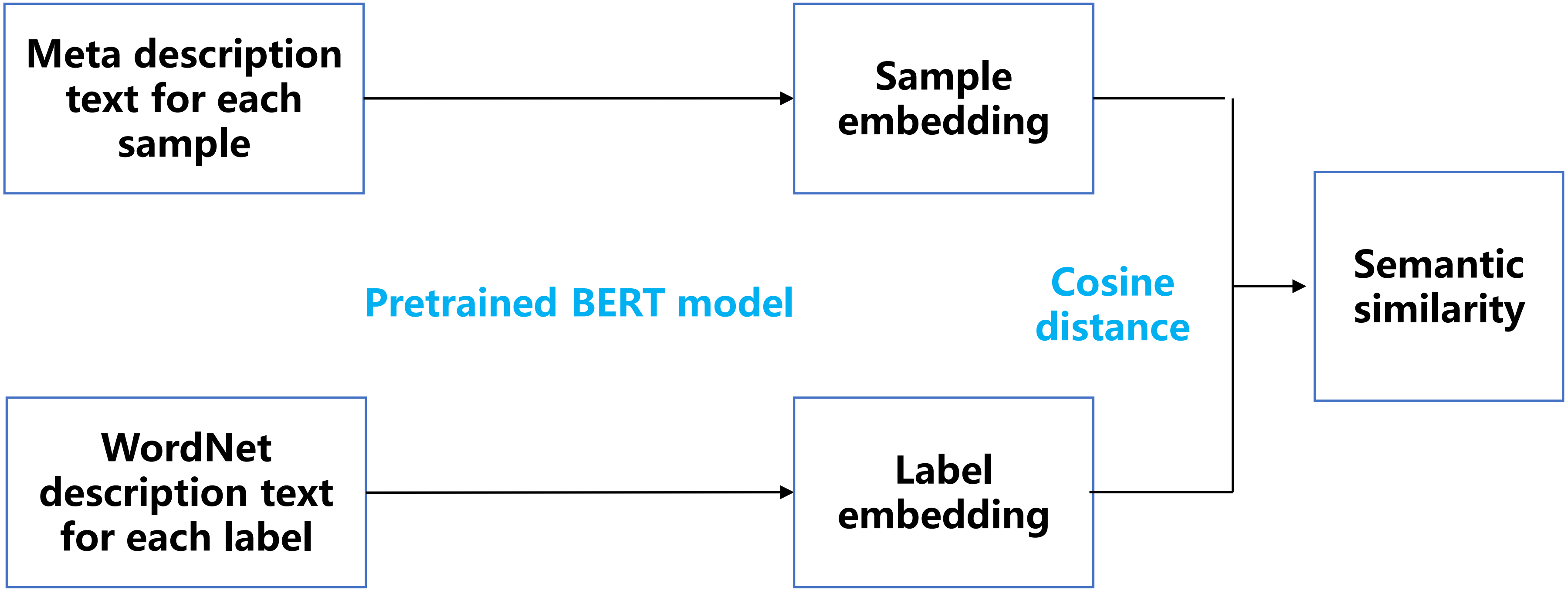
On-the-fly Self-supervised Loss

- 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.



Using Description Text

Tag: Yardbird

random selection

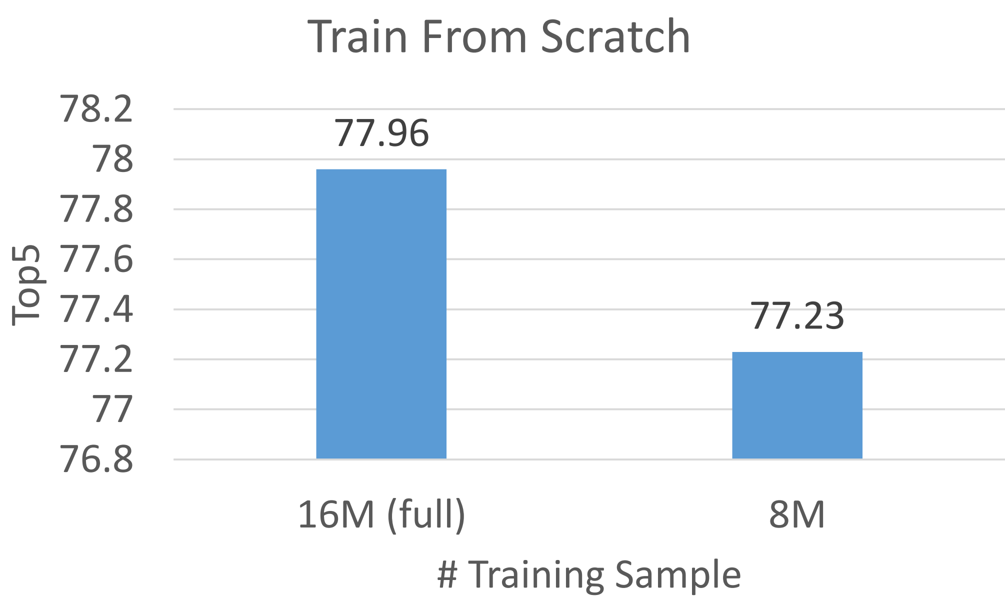
top 10

middle 10

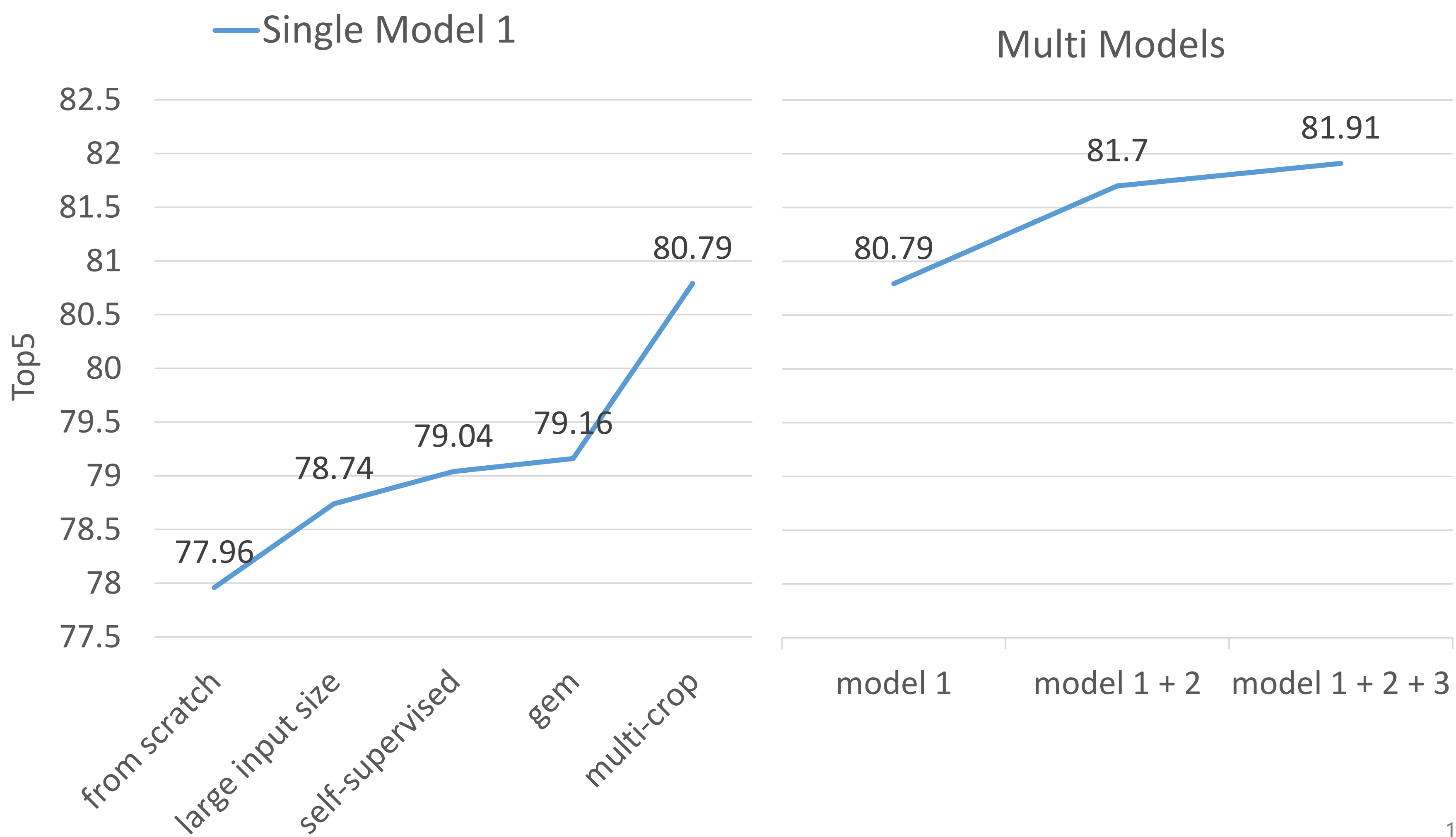
bottom 10

Using 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
82.5
82
81.5
81
80.5
80
79.5
79
78.5
78
77.5
77.96
78.74
79.04
79.16
80.79
80.79
81.7
81.91
from scratch
large input size
self-supervised
gem
multi-crop
model 1
model 1 + 2
model 1 + 2 + 3

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!