Learning From Web data: learning all or learning good?

Presenter：Shuyang Sun
Date：2018/06/18

Team Member

Shengju Qian[1] Qing Lian[1] Wayne Wu[2, 3] Chen Qian[3]

Fumin Shen[1] Heng Tao SHENG[1]

[1]University of Electronic Science and Technology of China,

[2]Tsinghua University,

[3]SenseTime Research.

01 Webvision Challenge Dataset’s Analysis and Overview

02 Existing Methodology and Our Approach

03 Experiments and Results

01
PART

Analysis

Webvision Dataset Overview

1 Huge

1. Huge: with 5000 classes and 16M images

Way larger than ImageNet --> Strategies to pick data.

2 Noisy

1. With Large number of annotations

Intra class bias for each class is very high.

See some samples…

2. Data imbalance:

Number of Samples in different labels has large variance: 10 to a thousand. --→ Class-Weighted Loss

3. Lots of mistaken labeled images

Samples from Webvision Dataset:








Pumpkin Ash (a tree!)

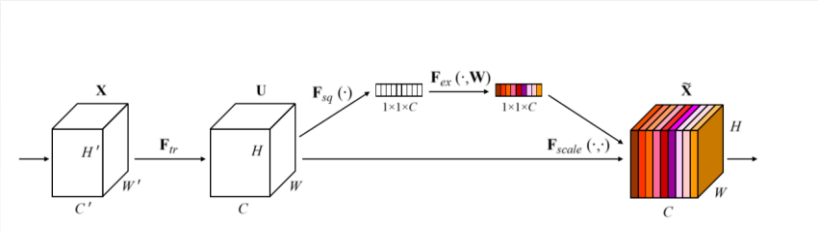
WIFI

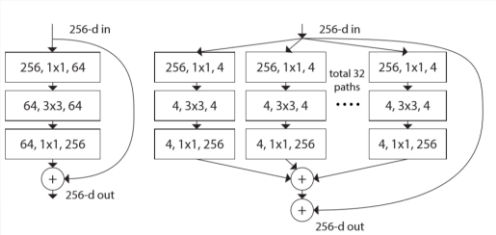
02
PART

Methodology

Related works in learning from large scale datasets

1 More Robust Models





1. Emergence of novel models for image classification
2. SENet, ResNeXt and so on….
3. Ensemble different models

2 Better Learning Strategies

1. Curriculum Learning
   (last year’ s winning entry)
2. Knowledge Distillation

   Li, Yuncheng, Yang, Jianchao, Song, Yale, Cao, Liangliang, Luo, Jiebo, and Li, Jia. <Learning from noisy labels with distillation>. ICCV 2017
3. Gated Back propagation

   (Our approach)

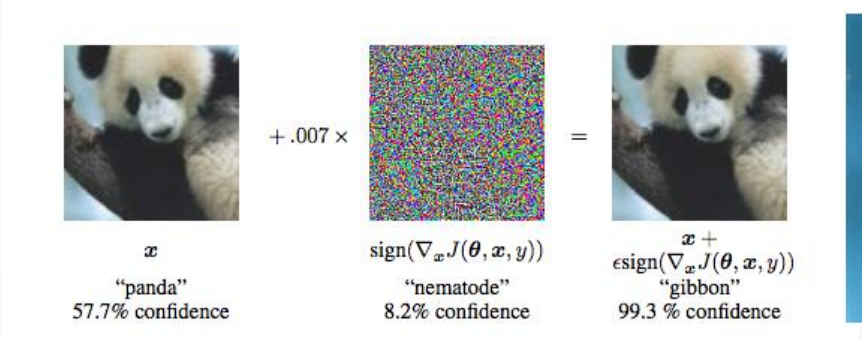
Thoughts on Learning from Web Data: learning which part ?

Compared to ImageNet dataset, Webvision has much larger diversity and semantic labels, the problem is: how to classify objects from those noisy labels.

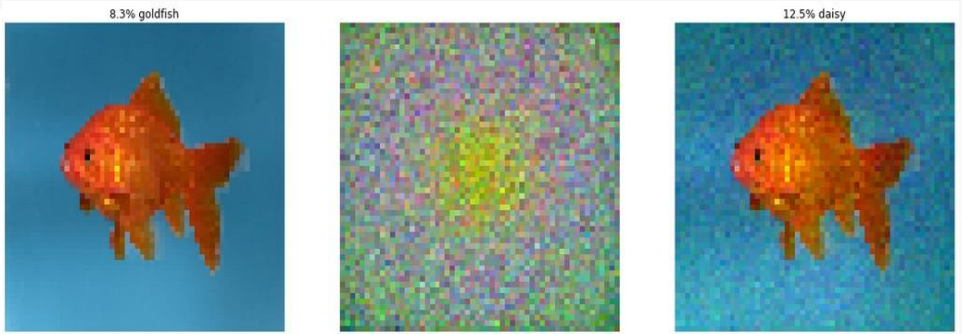
	Learn all	Learn good
Advantages	Robust to "good" noise	Fast, easy to converge
Disadvantages	Slow, sometime too noisy	Hard to select good part

Thoughts on Learning from Web Data: learning which part?

1. Different from ImageNet, Webvision’ s labels are not human annotated.
2. Adding “noise” : Between “good and bad” noise (Gaussian or Attack) ?



Panda to Gibbon

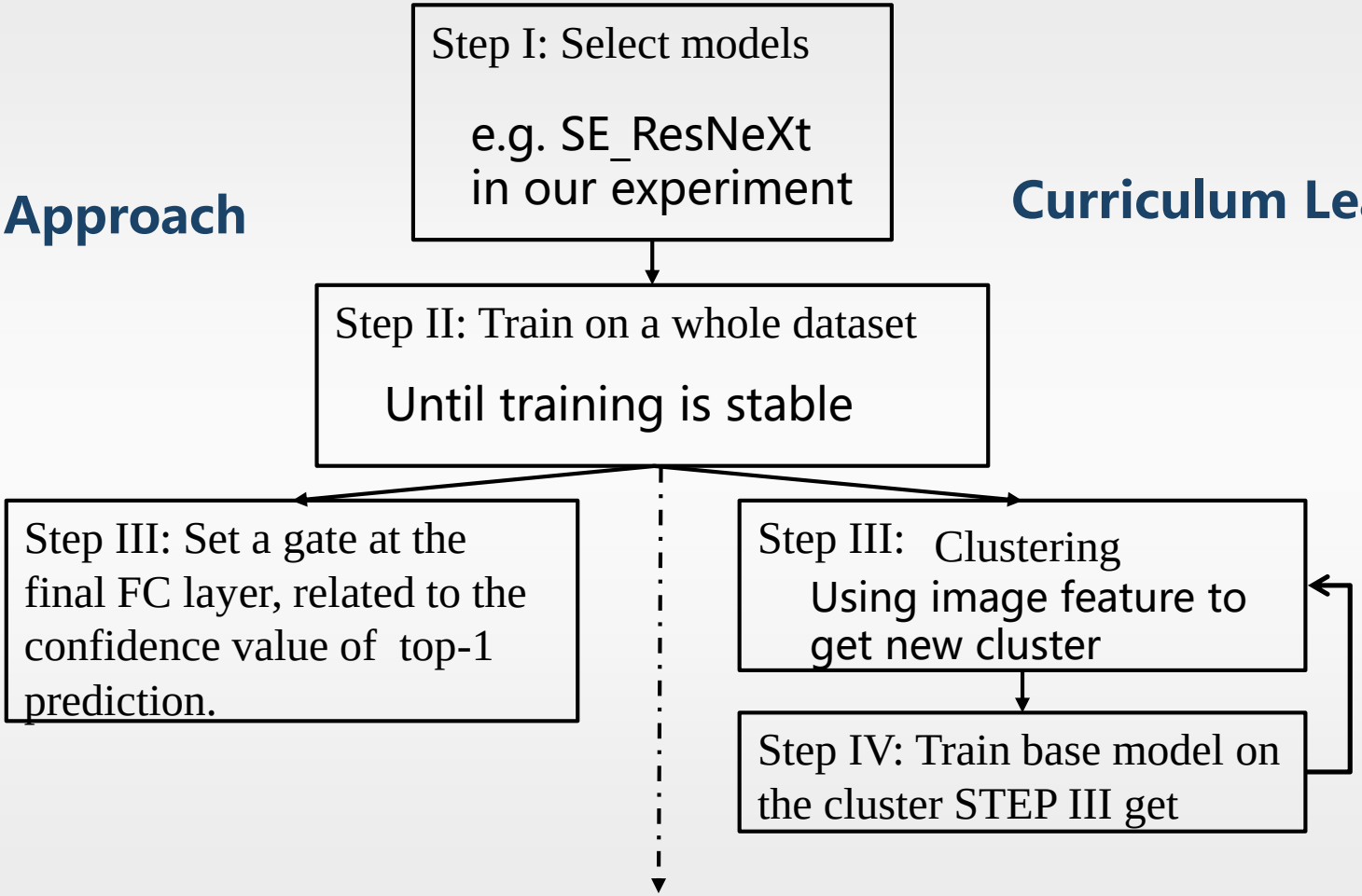


Goldfish to Daisy

3. “noise” in Webvision Dataset: **Bad** noise

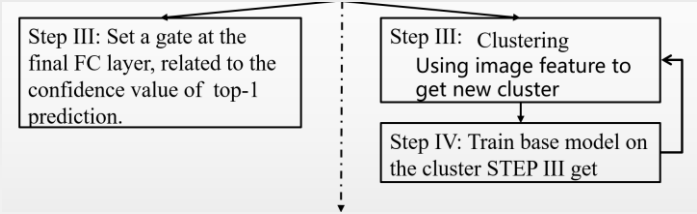
Pipeline

Our Approach



Curriculum Learning

Difference from Curriculum Learning



1 Faster:

More images and classes are **HARD** to do clustering.

2 More "violent" :

By **manually** setting the threshold in gated operation

3 Smoother.:

By adjusting the threshold, for example, by decreasing it,

We can let the network learns from easy cases to hard cases.

03
PART

Experiments

Experiment Details: Training

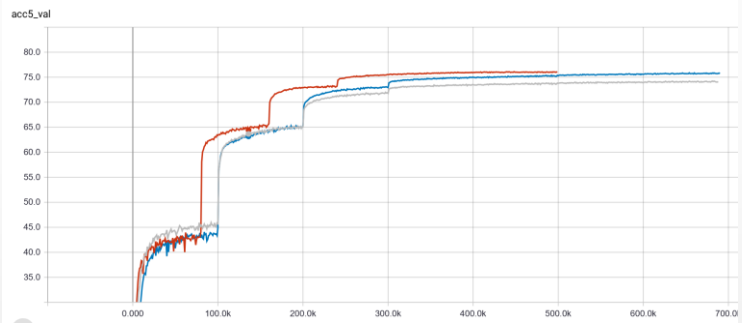
1. During training, We trained about twelve Based Models firstly on the whole dataset, including:
   - 3 ResNet-based models,
   - 3 DenseNet-based models,
   - 3 Inception-based models,
   - 2 'Squeeze and extraction' SE-ResNet based models. (Final best model is based on SE)

2. A batch size of 1024 and step learning rate scheduler with lr Gamma of 0.1 are applied, all these based models are trained on 32 GPUs for about 400k iterations.

Experiment Details: Training

2. After the loss becoming stable, we apply the "gated BP" operation during training. Due to limitation of GPU resources, we manually set the gate value after several trails and decrease the value to 0 during training to let the model learn from easy to hard.

3. After the gate value becomes 0, the model is again trained on the whole dataset for about 10k iterations with a lower learning rate.



At this stage, different learning rates to conv layers and final FC layer are applied, which show better performance during testing.

Experiment Details: Testing

We apply center crop during testing. Ensembling over several models with different weights has been tried in our setting while we didn’ t spend too much time on ensembling.

(Note: Our Final score(#5) is using **single model** and **single crop**)

Rank	Team name	Top-5 Accuracy (%)
1	**Vibranium**	**79.25**
2	**Overfit**	**75.30**
3	**ACRV_ANU**	**69.56**
4	EBD_birds	69.44
5	INFIMIND	68.74
6	CMIC	61.14

(The reason we didn’ t use multi-crop is that we found multi-crop strategies face fluctuation in our sub-test set.

03
PART

Summary

Summary

- We adopt a simple gate operation during training to filter clean and noisy images, which can achieve a fair score with only single model and single crop.

- It seems larger batch size can help smooth the noise and lead to better result

- Learning from web-supervision data has become very important in deep learning, we are still working on our algorithm design to a more general and robust framework. Hope you can join!

THANKS

For Further question
thesouthfrog@gmail.com