The 6th Place Solution for the Open Images 2019 Object Detection Track

Team 'Schwert', Hiroto Honda (solo)

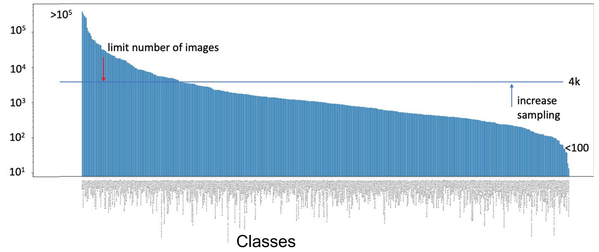
DeNA Co., Ltd.

Abstract

- Class balancing by limiting non-rare classes and re-sampling rare classes.
- FPN [1] detector with strong ResNeXt152-DCNv2 [2] backbone achieves 73.2 val AP and 56.4 AP at private LB, which outperforms last year's winners [3][4].
- Ensembling of eight models and enhanced NMS boost AP to 60.23 at private leaderboard. (6th place)

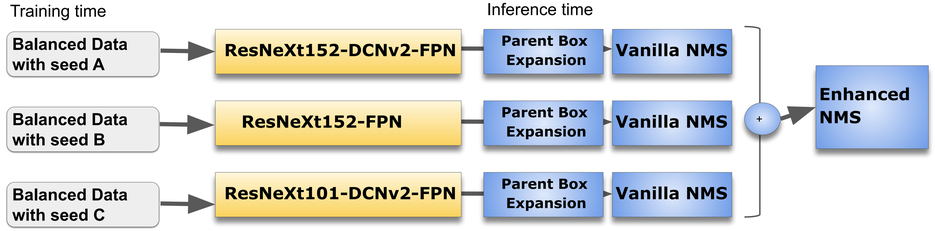
Dataset

- No external dataset.
  Only ImageNet pretrained weights are used for initialization.
- Class balancing.
  Equal probability for a model to encounter a certain class.
  - Rare classes: increase sampling rate.
  - Non-rare classes: limit number of images.
  - Total number of images: 4k x 500 (2M) → efficient training



Class balancing method.

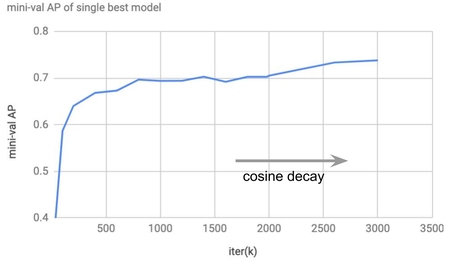
Pipeline and Models



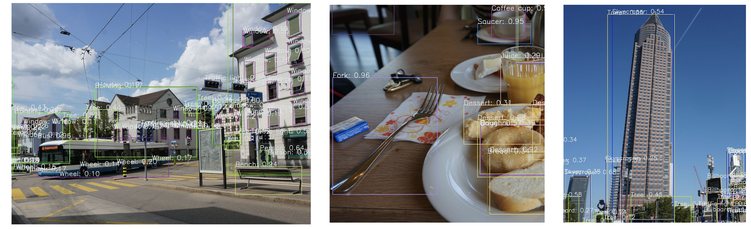
- Parent class expansion.
  Parent boxes are added after inference, which achieves empirically better AP than multi-class training.
- Ensembling with Enhanced NMS.
  Scores of box pairs with higher overlap than the threshold are added together.

Training

- Single GPU training.
  The training conditions are optimized for single GPU (V100) training.
  The baseline model has been trained for 3 million iterations and cosine decay is scheduled for the last 1.2 million iterations. Batch size is 1 and loss is accumulated for 4 batches.
- Mini-validation.
  A subset of the validation dataset consisting of 5,700 images is used.



Results



Inference examples of our best single model.

Backbone	Deformable Convolutions	Parent Expansion	Data Size	val AP	private LB
ResNeXt101	None	Inference Time	4k per class	69.8	
ResNeXt101	DCN v2	Inference Time	4k per class	72.2 (+2.4)	
ResNeXt152	None	Inference Time	4k per class	72.2 (+2.4)	
ResNeXt152	None	Inference Time	16k per class	72.4 (+2.6)	
ResNeXt152	DCN v2	Inference Time	4k per class	73.2 (+3.4)	**56.4 (best single model)**
ResNeXt152	DCN v2	Training Time	4k per class	70.7 (+0.9)*	
Ensemble					**60.23**

AP comparison between different models.

* The training condition is different from the others - number of iteration is 12M, not 30M.

References

[1] Tsung-Yi Lin et al., "Feature Pyramid Networks for Object Detection", CVPR 2017
[2] Xizhou Zhu et al., "Deformable ConvNets v2: More Deformable, Better Results", CVPR 2019
[3] Takuya Akiba et al., "PFDet: 2nd Place Solution to Open Images Challenge 2018 Object Detection Track", arXiv:1809.00778
[4] Yuan Gao et al., "Solution for Large-Scale Hierarchical Object Detection Datasets with Incomplete Annotation and Data Imbalance", arXiv:1810.06208