DAP: Detection-Aware Pre-training with Weak Supervision

Yuanyi Zhong[1] Jianfeng Wang[2] Lijuan Wang[2] Jian Peng[1] Yu-Xiong Wang[1] Lei Zhang[2]
[1]University of Illinois at Urbana-Champaign [2]Microsoft

CVPR VIRTUAL JUNE 19-25

Motivation

- **Problem statement:** Better pre-train an object detector
- Existing common practice:
  Classification pre-training (on ImageNet); a *truly* pre-trained detector is not yet existing
- **Issues** with classification pre-training:
  - Tasks misaligned: Global classification vs. detection
  - Features not explicitly trained to do localization
  - Unable to pre-train detector-specific components (RPN, regressor, etc)
- **Two desired properties** in pre-training:
  - Classification should be done **locally** rather than globally
  - Features should be **capable of predicting bounding boxes** and can be easily adapted to any desired object categories after fine-tuning
- Our solution: **Detection-Aware Pre-training (DAP)**
  - A novel pre-training method with weak supervision (e.g. ImageNet)
  - Built on simple weakly-supervised object localization with class activation maps
  - Transfer better to downstream detection, especially in few-shot regimes

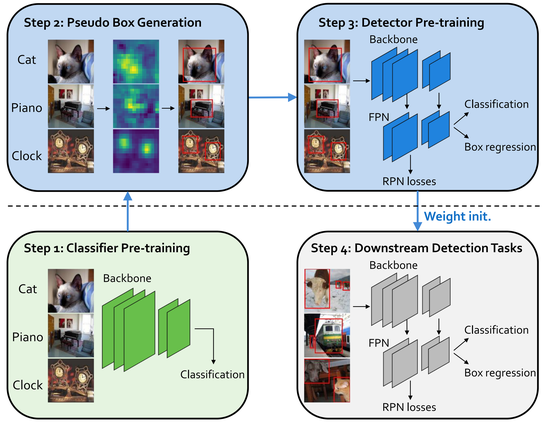
Experiment Setup: Pre-training

- **Classification-style datasets:**
  - ImageNet-1M: 1.28 million images | 1,000 classes
  - ImageNet-14M: 14 million images | 22K classes
- Only image-level labels
- Step 1 follows the standard classifier training
  Step 3 fine-tunes from the Step 1 weights with the pseudo-labeled data

Experiment Setup: Downstream Detection Fine-tuning

- **Detection datasets:**
  - VOC 2007: 5,011 trainval images | 4,952 test images | 20 classes
  - VOC 2007+2012: 11K training images (merge VOC 2007 with additional VOC 2012)
  - COCO: 118K training images | 5,000 val images | 80 classes
- **Evaluation metrics:** Average Precision (AP)
  For COCO experiments, take the mean APs at IoU thresholds $0.5 \ldots 0.95$
  For VOC experiments, report AP at IoU 0.5
- The network is initialized from either CLS or DAP pre-trained weights

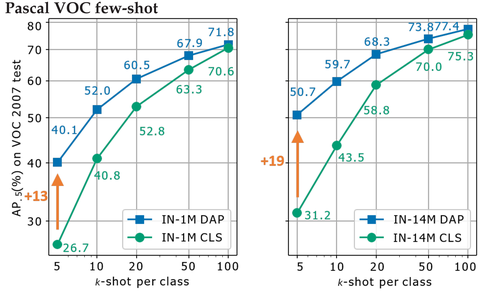
Detection-Aware Pre-training



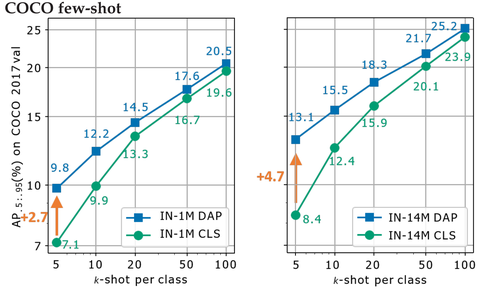
- **DAP steps:**
  - Step 1: Classifier pre-training
  - Step 2: Pseudo box generation with class activation maps (CAMs)
  - Step 3: Detector pre-training:
    The same setup as standard detector training but with the pseudo-labeled data.
  - Step 4: Downstream detector fine-tuning
- In comparison, the traditional Classification Pre-training goes directly from Step 1 to Step 4
- **DAP advantages:**
  - **Local** classification
  - **Features capable** of box localization
  - **Tasks aligned**: both are set up as detection training
  - **Architecture** (almost) the same:
    Be **able** to pre-train RPN and box regressor

Transfer to VOC and COCO Detection

- **Architecture:** Faster RCNN ResNet-50 FPN. **CLS:** Classification pre-training baseline. **DAP:** Our method
  DAP is consistently more effective than CLS, especially in few-shot regimes
- **Pascal VOC few-shot**



- **COCO few-shot**



- **VOC 2007 full**

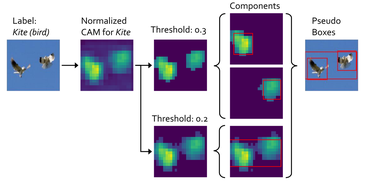
$AP_{.5}$	ImageNet-1M	ImageNet-14M
CLS	77.36	80.74
DAP (Ours)	79.93 (+2.57)	84.24 (+3.50)

- **COCO full**

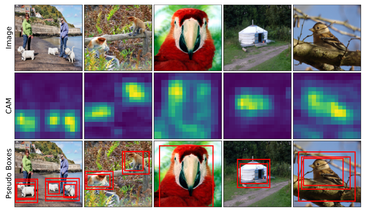
$AP_{.5:.95}$	ImageNet-1M	ImageNet-14M
CLS	36.73	38.87
DAP (Ours)	37.25 (+0.52)	39.57 (+0.70)

Pseudo Box Generation

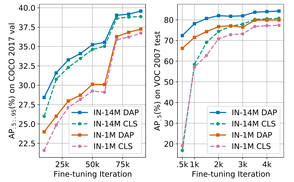
- **Technique:** Simple approach based on **Class Activation Maps (CAMs)**
  - Fit bounding boxes by matching mean and variance of the "blobs"
  - Easily scale to millions of images
  - Multi-threshold trick to improve the recall rates



- **ImageNet examples**
  - Input images
  - Class activation maps (CAMs)
  - Generated pseudo boxes:
    Grouped from multiple thresholds and locations



Additional Observations



- Better initialization leads to **faster convergence**
- **Scalability:** DAP benefits more from larger-scale pre-training datasets
- **Generalizability:** DAP also outperforms CLS with other detector architectures (RetinaNet) and backbones (ResNet-101)

Conclusion and Future Work

- Traditional classification pre-training is not optimal for detection
- Detection-aware pre-training (DAP) outperforms classification pre-training, especially in terms of few-shot detection results and convergence speed
- DAP scales to large-scale datasets and a larger-scale dataset is beneficial (ImageNet-14M vs. 1M)
- **Future work:** More advanced weakly supervised localization, leveraging mixed-labeled data in semi-supervised pre-training, extending to instance segmentation, etc
- Code: https://github.com/mikuhatsune/dap