Object Detection in Complex Food Images

A Study Based on Google Open Image Dataset

Yaxin Fang (PUID:0031073186) Yung-Ching Chen(PUID:0030867342)

PURDUE UNIVERSITY

Problem Description

- Problem: object detection model for 64 categories of food
- Labels: fast food, snacks, desserts, baked goods, seafood, vegetables, fruits, dairy and many other
- Model output: bounded box coordinates and label confidence

Dataset Description

- Source: Google Open Images Dataset
- Number of training data: 15K images and 50K bounded boxes
- Data properties: unbalanced distribution of labels and some boxes contain a group of objects

Motivation

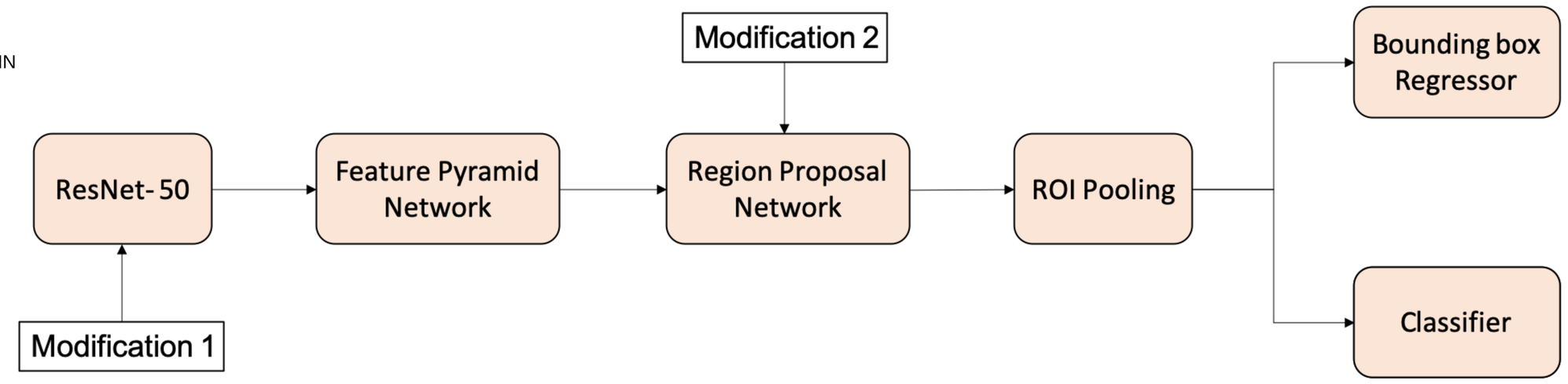
- "Cameras eat first" trend provides a big amount of food image data and potential applications

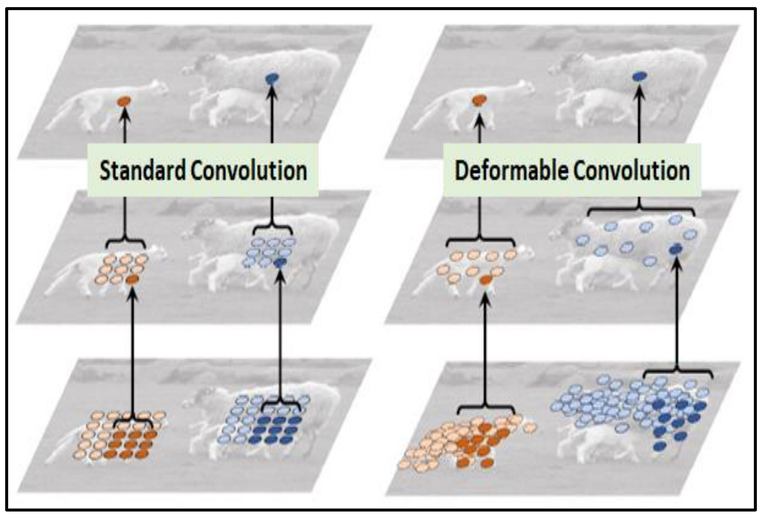
State-of-the-art Models

- Two-stage Detectors
  - First stage: a sparse set of candidate proposals with high probabilities of containing objects are generated
  - Second stage: object classification and coordinate regression
  - Models: R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN
- One-stage Detectors
  - Propose predicted boxes from input images directly without region proposal step, thus they are time efficient and can be used for real-time devices
  - Models: YOLOv3, RetinaNet

Model Structure

Based on structure of Faster R-CNN





Modification 1: Improve recognition of geometric variations by modulated deformable module

- Modulated deformable convolutions are applied in all the 3 * 3 conv layers in stages conv3, conv4, and conv5 in pretrained ResNet-50.
- Modulated Deformable Convolution

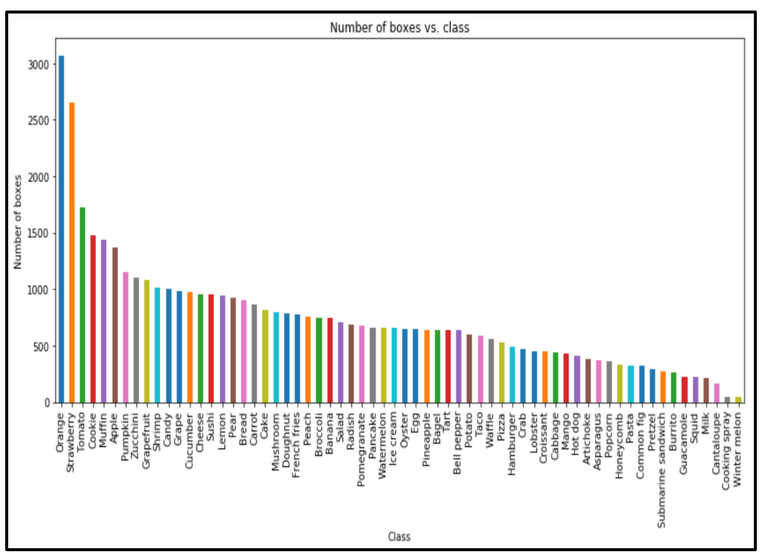
  - Standard Convolution: operates on a pre-defined rectangular grid (determined by filter size) from input image or feature maps.

  - Deformable Convolution [1]: operates on the same regular grid as standard convolution does, but with each grid point augmented by a learnable offset. By picking values at different locations for convolution from input feature maps, deformable convolution can better deal with geometric variations.

  - Modulated Deformable Convolution [2]: adds learnable modulation scalar to deformable convolution to decrease the redundant context created by uncontrolled offsets, hence reduce detection error.

[1] J.Dai, H. Qi, Y. Xiong, et, al. (2017). Deformable Convolutional Networks

[2] X. Zhu, H. Hu, S. Lin, et, al. (2018). Deformable ConvNets v2: More Deformable, Better Results



Modification 2: Mitigate class imbalance by focal loss and manual adjustment

- Foreground-Background class imbalance: Apply focal loss to RPN [1]

  - Over-represented class is background (regions that don't include any objects); under-represented class is foreground(regions that contain objects)

  - Focal loss: $FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$

    p_t: probability of the ground truth class; $\alpha_t \in [0,1]$, re-weighting factor to balance positive and negative samples; $\gamma \geq 0$, hyperparameter

  - Replace cross entropy loss with focal loss in Region Proposal Network (RPN): $L_{RPN-FL} = \frac{\lambda_{fl}}{N}\sum_i FL(p_i^t) + \frac{1}{N}\sum_i I(t_i^t)L_{reg}(t_i, t_i^t)$

    N: number of object samples; λ_{fl}: balancing weight; $I(t_i^t)$: indicator function of ground truth; $L_{reg}(t_i, t_i^t)$: regression loss (smooth L1 loss)

- Foreground-Foreground class imbalance: Manual adjustment

  - Imbalance of object classes: some object classes contain dramatically more images in training set than other object classes.

  - We manually limit the number of images in each object class to be no more than 300 in training set.

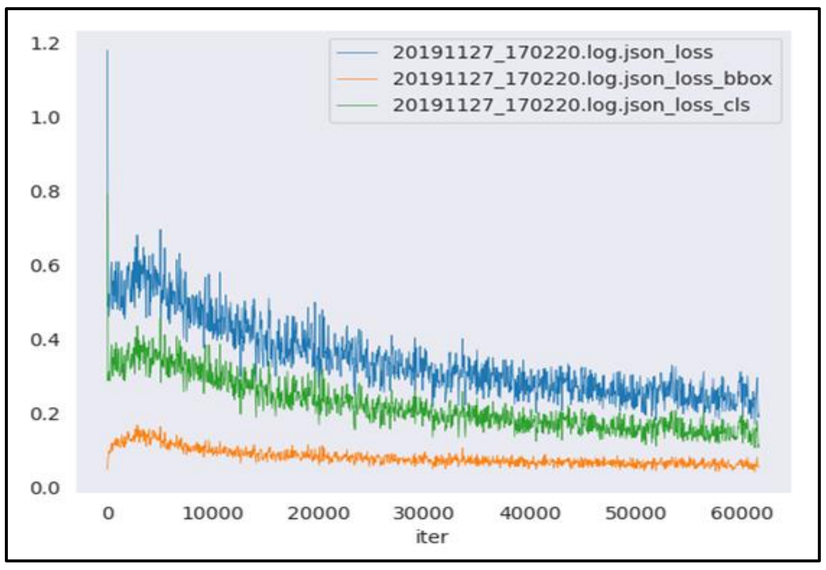
[1] C. Chen, X. Song and S. Jiang, (2018). Focal Loss for Region Proposal Network

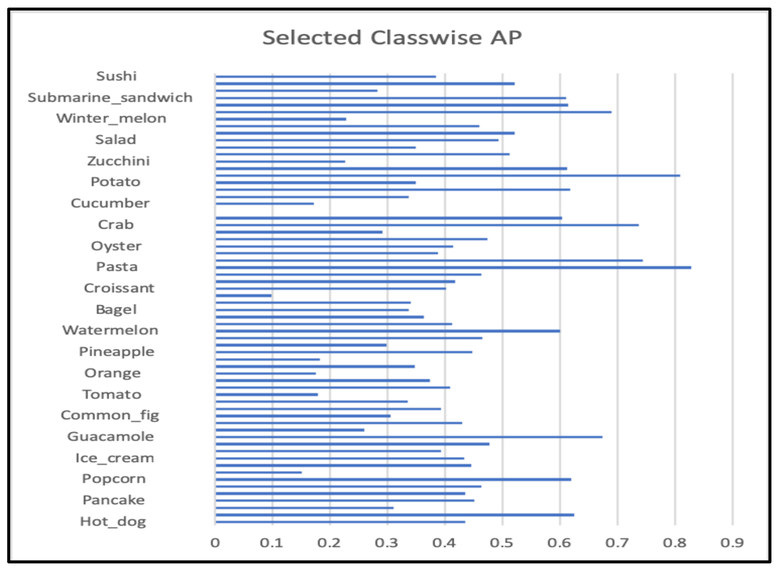
Training

- Implemented the model using mmdetection toolbox
- Transfer learning on pretrained ResNet-50
- Optimizer: SGD with momentum (momentum = 0.9)
  learning rate = 0.001, learning rate decay = 0.0001
- Baseline models: Faster R-CNN, YOLOv3
- Trained all three models for 8 epochs.

Results

	Our model	Faster R-CNN	YOLOv3
mAP	0.43	0.40	0.42





IE 590 Deep Learning in Machine Vision (Fall 2019)