Model

- The person detection model based on SSD architecture
- There are 3 blocks: Backbone, Extralayers, Detection head

Backbone

- The extractor is used: Mobilenet
- The original Mobilenet-V2 can achhive high accuracy (use pre-trained from VOC dataset). However the running time also high
- The tiny Mobilenet-V2 is customed Mobilenet-V2 (vertical, horizontal).

```
The backbone can be customed at: ./module/mobilent_v2.py
```

Extra layers

- Multi-scale feature maps for detection

- The SSD architecture uses multiple layers (multi-scale feature maps) to detect objects independently. As CNN reduces the spatial dimension gradually, the resolution of the feature maps also decrease. SSD uses lower resolution layers to detect larger scale objects and vice versa. For example, the 4× 4 feature maps are used for larger scale object.

- Because the model person detection is used for surveillance camera (detect samll, medium bojects) so that the small feature maps do not contains lots of information. Thus, this model is eliminated 2 last feature maps

- To the person dection use feature maps: 38, 19, 8, 4 . The resolution of feature map depending on the the input of network.

- Note that: The high resolution of input network is not synonymous with the good results

```
The Extra Layers can be customed at: ./model/mb_ssd_lite_f38.py and
./module/ssd.py
```

Detection head

- Regression head (location) and classification header (classification)
- The most impotant factor in this component is anchor boxes (whith 3 parametert can obtimize scale, ratio, number of anchor per gid cell)
- The anchor boxes are defined in [] based on the COCO, VOC dataset with many objects as well as ratio, size...
- In [], the author proposed formula to generate anchors boxes. However, this formula is designed for many object catagory , size... For instance, the proposed scale factor in [] are [0.2, 0.9] which can be

suitable for COCO dataset. However, this factor does not give good results... As mentioned above, the objects in surveillance applivation distribute small, medium and rarely large. Therefor the scale should have distribution at smaller than 0.5

- Ratio: the ration between height/width of objects. statistics to get this factor. Code is available here

```
Anchor boxes: ./model/config/mb_ssd_lite_f38_config.py and ./module/ssd.py
```

Loss

- Localization oss: Use smooth_l1_loss
- Classification loss: Use focal loss instead of CE loss to to address the issue of the class imbalance problem (person/background)

Dataset

- There are 2 main dataset: Crowd-Human- 15k and Wider-face- 16k, brainwash-11k
- SCUTB (valid)

Requirements

- anaconda
- Python-3.6
- Pytorch-1.2
- Torchvision-0.4
- opencv-python-
- pandas

Training

- Optimizer: SGD, with weight decay: 5e-4, batch size: 32, Number of echop:150
- Data augmentation:

```
python train.py, type_network 'mb2-ssd-lite_f38'
```

Testing

```
Folder image: python detect_imgs.py --net_type <model_path> --test_path
<path_dir_image>
video: python live_demo.py --model_path <path_network>
```

ModelParameter and results

- Input: (300, 300)
- Feature maps: 38-38, 19-19, 10-10, 5-5
- Step (shrinkage): 8, 16, 32, 64
- Scale (box size): (16-32), (32-64),(64,128),(128,256)
- Ratios: 1.7

Pytorch model (mb_ssd_lite_f38_150_193_14)

input network	parameter	FLOPs	Miss rate	mAP	Running time	Model weight
300x300	2.7 M	1.2 G	7%	90.02%	39ms	Weight

dlc model:

- dlc and quantized dlc
- Total parameters: 2750864
- Memory Needed to Run: 345.9 MiB
- Total MACs per inference: 618M (100%)