Supplementary

This file provides some additional information from two perspective: compared semi-supervised methods and structures of used crowd counting networks, which correspond to the Section 6 in the paper.

Introduction of Compared Methods

Learning to Rank (L2R): is proposed in [1]. This method collects a large amount of unlabeled extra images from the Internet for density estimation and produces a serial of sub-images $I_1, I_2, ..., I_M$ and ensures the smaller sub-image I_t is covered in the larger sub-image I_s. The number of pedestrians in the smaller sub-image is not greater than the larger one, though the exact person counts is unknown. It can be used for crowd counting model training as shown in Eq. 1.

$$\mathcal{L}_r = \sum_{s=1}^{M} \sum_{t=1}^{s} max(0, \hat{C}_t - \hat{C}_s), \tag{1}$$

where $\hat{C}_t$ and $\hat{C}_s$ are the estimated count value in the $t-th$ and $s-th$ sub-image. Note that the $t-th$ sub-image is contained in the $s-th$ sub-image.

Unsupervised Data Augmentation (UDA): is a classical consistency-based semi-supervised learning method for classification. The core of this method is to create an augmented version of the original unlabeled image and reduce the diversity of predicted density maps directly. It can be formulated as:

$$\mathcal{L}_{UDA} = \sum_{(i,j)} |\hat{D}_A(i,j) - \hat{D}(i,j)|^2, \tag{2}$$

where $\hat{D}_A$ is the estimated density map from augmented version of unlabeled images and $\hat{D}$ is the predicted density map from original unlabeled images.

Mean Teacher (MT): is widely used in semi-supervised learning tasks. This method defines a teacher model and student model, these two models have the same structure. For unlabeled images, the student model uses the original image to produce a predicted density map while the teacher model leverage an augmented version of the unlabeled image to produce another version of the predicted density map. The process of updating the parameters of the student model can be formulated as:

$$\mathcal{L}_{MT_S} = \sum_{(i,j)} |\hat{D}_S(i,j) - \hat{D}_T(i,j)|^2, \tag{3}$$

where $\hat{D}_S$ is the predicted density map from the student model, $\hat{D}_T$ is the predicted density map from the teacher model. Besides, unlike the student model, the update process of teacher model averages model weights which is formulated as:

$$\theta_t^T = \alpha\theta_{t-1}^T + (1-\alpha)\theta_t^S, \tag{4}$$

where θ_t^T is the parameter of the teacher model at the $t-th$ steps, α is a smoothing coefficient hyper parameter, θ_{t-1}^T is the parameter of the teacher model at the $(t-1)-th$ steps and θ_t^S is the parameter of the student model currently.

Interpolation Consistency Training (ICT): is a semi-supervised learning method for classification. The core of this method is using the original unlabeled image and the augmented version to train the crowd counting model. The training loss is formulated as:

$$\mathcal{L}_{ICT} = \sum_{(i,j)} |\hat{D}_\lambda(i,j) - \lambda(\hat{D}_A(i,j), \hat{D}(i,j)|^2, \tag{5}$$

where λ is the mix-up function. D_λ is the predicted density map produced by the mix-up version of the unlabeled image which is $I_\lambda = \lambda I_A + (1-\lambda)I$. The $\lambda(\hat{D}_A(i,j), \hat{D}(i,j)$ is the mix-up version of predicted density maps which is equal to $\lambda D_A + (1-\lambda)D$.

Introduction of Network structures

In the paper, the crowd counting model includes the feature extractor, the density regressor and the segmentation predictor. In our experiment, we use the CSRNet [2] and the SPN [3]. The details of these structures are shown in Table. 1. The segmentation predictor has a similar structure with the corresponding density regressor expect the output of the last layer contains two channels.

Table 1. The structure of the CSRNet and the SPN. The k is the convolutional kernel size, c means the number of output channel, s means stride, p means padding size and d indicates the dilation rate.

	CSRNET	SPN
Feature Extractor	k(3,3)-c64-s1-p1-d1	k(3,3)-c64-s1-p1-d1
	k(3,3)-c64-s1-p1-d1	k(3,3)-c64-s1-p1-d1
	maxpooling(2,2)	maxpooling(2,2)
	k(3,3)-c128-s1-p1-d1	k(3,3)-c128-s1-p1-d1
	k(3,3)-c128-s1-p1-d1	k(3,3)-c128-s1-p1-d1
	maxpooling(2,2)	maxpooling(2,2)
	k(3,3)-c256-s1-p1-d1	k(3,3)-c256-s1-p1-d1
	k(3,3)-c256-s1-p1-d1	k(3,3)-c256-s1-p1-d1
	k(3,3)-c256-s1-p1-d1	k(3,3)-c256-s1-p1-d1
	maxpooling(2,2)	maxpooling(2,2)
	k(3,3)-c512-s1-p1-d1	k(3,3)-c512-s1-p1-d1
	k(3,3)-c512-s1-p1-d1	k(3,3)-c512-s1-p1-d1
	k(3,3)-c512-s1-p1-d1	k(3,3)-c512-s1-p1-d1
	k(3,3)-c512-s1-p1-d2	[k(3,3)-c512-s1-p2-d2, k(3,3)-c512-s1-p4-d4, k(3,3)-c512-s1-p8-d8, k(3,3)-c512-s1-p12-d12]
	k(3,3)-c512-s1-p2-d2	k(3,3)-c512-s1-p1-d1
	k(3,3)-c512-s1-p2-d2	k(3,3)-c256-s1-p1-d1
	k(3,3)-c256-s1-p2-d2	
Density regressor	k(3,3)-c128-s1-p2-d2	k(3,3)-c128-s1-p1-d1
	k(3,3)-c64-s1-p2-d2	k(3,3)-c64-s1-p1-d1
	k(1,1)-c1-s1-p0-d0	k(1,1)-c1-s1-p0-d0

References

1. Xialei Liu, Joost van de Weijer, and Andrew D. Bagdanov. Leveraging unlabeled data for crowd counting by learning to rank. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7661–7669, 2018.
2. Yuhong Li, Xiaofan Zhang, and Deming Chen. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1091–1100, 2018.
3. Xinya Chen, Yanrui Bin, Nong Sang, and Changxin Gao. Scale pyramid network for crowd counting. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1941–1950, 2019.