陳駿丞老師 - CAC = MA + GL

出自

ICPR Pattern Recognize

ACCV

題目

A Recipe for CAC: Mosaic-based Generalized Loss for Improved Class-Agnostic Counting

1. Mosaic Data Augmentation : 將照片合併

2. Generalized Loss : 考慮 Wasserstein Distance 的 Loss
3. Class-Agnostic Counting :
   class-agnostic ⇒ zero shot 的 class (ex: training data中沒有藍鯨)
   要去計算圖片中的物件數量，連 class-agnostic 也要一起算

Introduction

背景

1. 可以應用在 visual surveillance 監視器
2. CAC 任務中， input 是 K 張參考照片，照片裡面有很多物體，這些物體就是希望 CAC 去數的。接著才會 input query image 讓 CAC 去找

動機

1. Reference Oversight : 不看 Refference 照片就胡亂 count 物體 ⇒ 提出了 multi-class mosaic evaluation dataset
2. FSC-147 評判標準有問題，Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) 讓模型只對高數量的任務作對 ⇒ 提出 Normalized Absolute Error (NAE) and Squared Relative Error (SRE)
3. pixel-wise losses (e.g., MSE) 沒辦法準確呈現 target objects 的偏移大小

為什麼沒有一步到位的做法 ⇒ 一次給兩種以上的 Refference 讓 Model 可以標示出這幾種的數量

優點

1. 快速
2. 兩種之間可以互斥排除

目的

相關研究

1. Class-specific Object Counting
    a. detection-based ：使用 object detectors ⇒ 不容易作對 overlapped, occluded, or crowded. 的 case
    b. regression-based : predict input images 的 density map (每個 pixel 上都有信心分數)
    ground truth : 將 point annotations (人工將物體位置標註 1) 經過 Gaussian Kernel 卷積後得出

2. CAC：主要都是 regression-based

    GMNet : 用 Resnet 抽 Refference & Query Image Feature，然後做 regression

    FamNet : convolving the reference feature maps across query feature maps

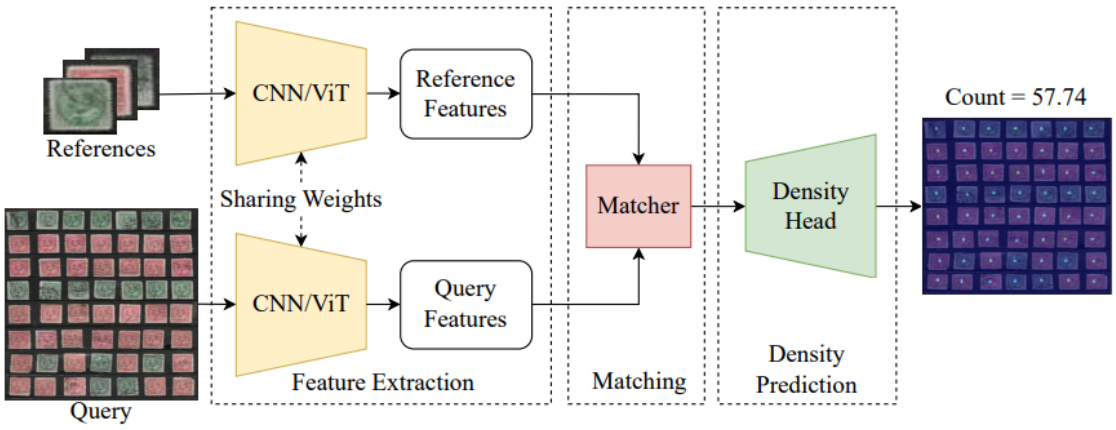
    CFOCNet : multi-scale mapping 與 self-attention
    BMNet : bilinear similarity

    SPDCN：scale awareness，利用 Scale-Prior Deformable Convolution and Scale-Sensitive Generalized Loss

    CounTR： transformer-based ， 利用 attention between the query and the references

    LOCA：使用object prototype extraction module (OPE) 來找出 strong object prototypes 協助定位 ， 使用 cross-attention 以及利用 shape information and object appearances
    New Models : adopt CLIP，利用語意資訊來 count

3. CAC Framework = feature extractor + cross-matching module + density map estimator



$$\text{query image } X \in \mathbb{R}^{H_X \times W_X \times 3}$$

refference image

$$\overline{Z = \{z_i\}_{i=1}^{K}, z_i \in \mathbb{R}^{H_{Z_i} \times W_{Z_i} \times 3}};$$

ground truth (density map)

$$\text{ground truth label } Y \in \mathbb{R}^{H_X \times W_X}$$

Syi : Feature Extractor

Phi : Similarity

Theta : density map

$$\min_{\theta,\phi,\psi} \mathbb{E}_{(X,Z,Y)}[\mathcal{L}(R_\theta(S_\phi(F_\psi(X), F_\psi(Z)), Y)],$$

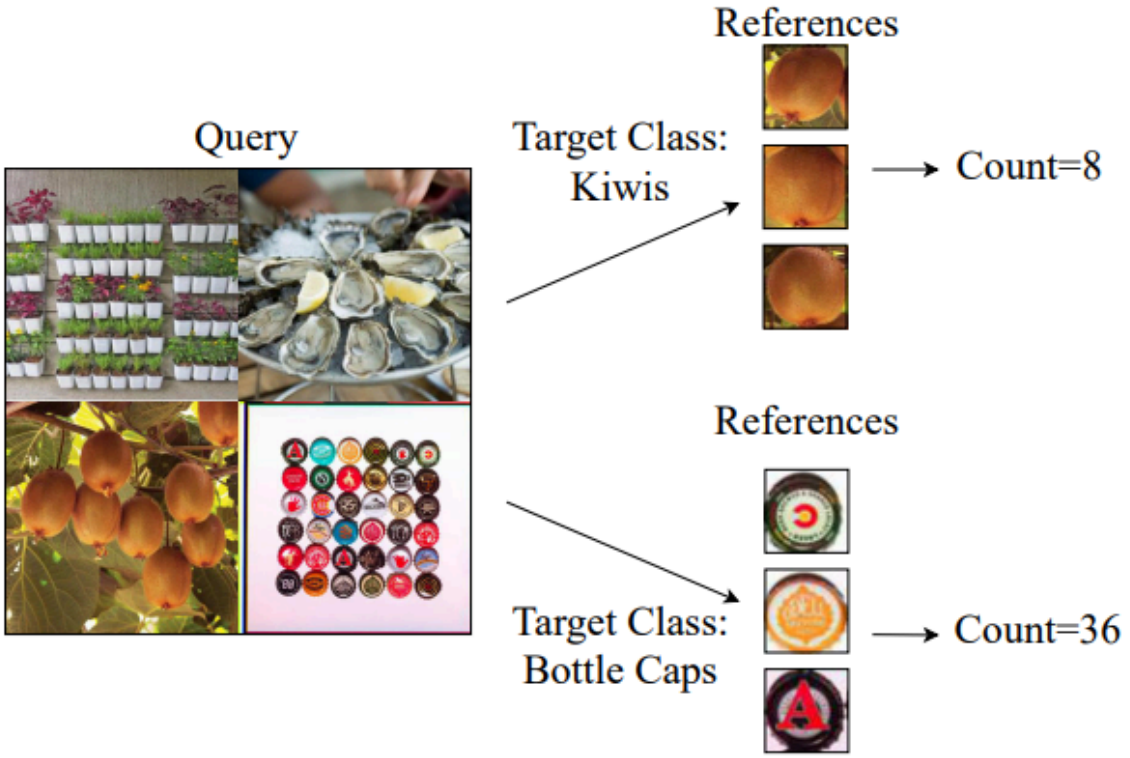
物體大小是否會影響判別 ⇒ CFOCNet 已經解決

物體型變是否會影響判別 ⇒ LOCA 已經解決

CAC 可不可以做在影片上，並且利用 Spatial 與 Time 的特性

方法

1. multi-class counting datasets FSC-Mosaic

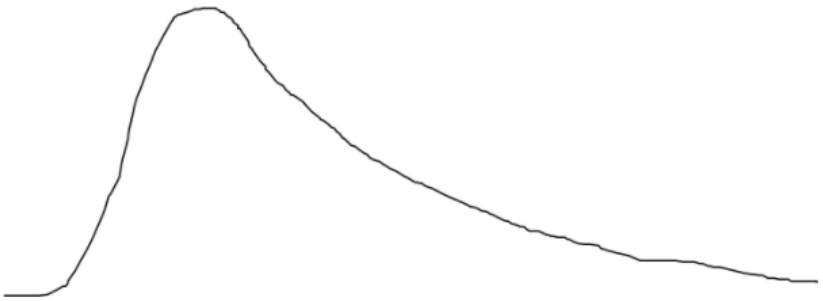


COCO dataset 雖然一個 query 中有多個種類，但照片中的物件數量太少，所以改調 FSC17

2. Metrics

   MSE 直接去加總做錯的數量會被長尾分布影響，特別會受到 large-count query images 的影響

count c_l and predicted count $\tilde{c}_l$ for the lth image in the dataset with L images, $\text{MAE} = \frac{1}{L}\sum_{i=1}^{L}|\tilde{c}_l - c_l|$ and $\text{RMSE} = \sqrt{\frac{1}{L}\sum_{i=1}^{L}(\tilde{c}_l - c_l)^2}$ are usually used. The



將 large counts 移除之後，原本 Model 的 MAE 與 RMSE 效果會好很多，代表很受極端值影響，因此將這兩個改成 NAE 與 SRE

$$\mathrm{NAE} = \frac{1}{L}\sum_{l=1}^{L}\frac{|\tilde{c}_l - c_l|}{c_l}$$

$$\mathrm{SRE} = \sqrt{\frac{1}{L}\sum_{i=1}^{L}\frac{(\tilde{c}_l - c_l)^2}{c_l}}$$

3. Generalized Loss (GL)：localization-aware

L2 無論 density map 上錯的距離是遠是近， L2 的懲罰都是一樣的。

uses unbalanced optimal transport (OT) to measure the
transport cost between the predicted density map and the ground truth point

GL penalizes more when the predictions are farther from the ground truth.

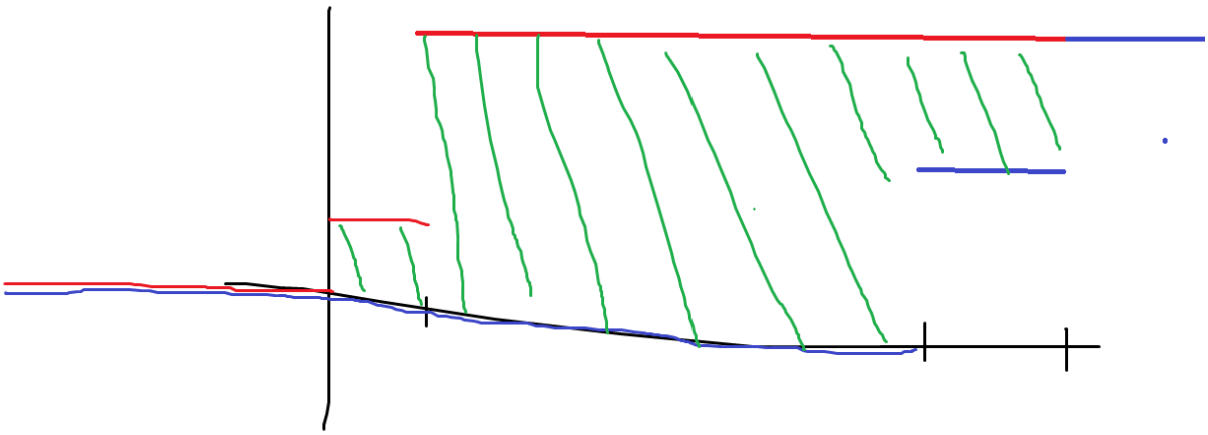
考慮了分佈之間的移動成本

P 和 Q 的總質量碰巧都是 1.0

Wasserstein Distance (1D - 平衡情況)

$$W_1(P, Q) = \int_{-\infty}^{\infty} |F_P(x) - F_Q(x)| \, dx$$

其中 $F_P(x)$ 和 $F_Q(x)$ 分別是分佈 P 和 Q 的**累積分佈函數 (Cumulative Distribution Function, CDF)**。



$$L_C^\tau = \min_P \langle C, P \rangle - \varepsilon H(P) + \tau \|P\mathbf{1}_m - a\|_2^2 + \tau \|P^\top \mathbf{1}_n - b\|_1, \tag{2}$$

where C is the transport cost of moving predicted density to ground truth dot annotation, P is the corresponding transport plan, $H(\cdot)$ is the entropic regularization term, n is the number of pixels, m is the number of annotation points, a is the predicted density map, and b is the ground truth dot map. Furthermore, to better encode perspective information, it proposes the Perspective-Guided Transport Cost:

$$C_{ij} = \exp\left(\frac{1}{\eta(x_i, y_i)} \|x_i - y_j\|_2\right), \tag{3}$$

where $\eta(x_i, y_i)$ is a adaptive perspective factor. We simply choose a fixed η for our experiment.

Experiments

FSC147 dataset：147 categories and 6,135 images across different scales，每張 query 有 3 張 refference image (圖中的 refference image)

CARPK： 1,448 photos cars in parking lots from a bird-view

如何使用資料集：

1. 用 FSC147 dataset 來 train model
2. 用 FSC-Mosaic dataset 來 validate
3. 用 CARPK 來 validate

使用 metric：MAE and RMSE NAE and SRE

model：MixFormer [5] architecture + multi-scale density head (來判斷 multi-scale)

與 BMNet+, CounTR, and LOCA(SOTA) 相比

data 會隨機 flip 與 隨機 crop成 384 * 384

還會 crop 成 192 * 192 讓 4 張照片組成 mosaic

our performance yields an improvement of 45.4% on VAL
NAE, 1.9% on VAL SRE, and 15.7% on TEST NAE

用 FSC-147 來 Validate 與 Test

Table 2: Comparisons with SOTA CAC models on the FSC-147 dataset. Notation '∗' indicates that we further adopt the test-time normalization [2] during inference.

Method	VAL (validation)				TEST			
	MAE↓	RMSE↓	NAE↓	SRE↓	MAE↓	RMSE↓	NAE↓	SRE↓
GMN [13]	29.66	89.81	—	—	25.52	124.57	—	—
FamNet [17]	23.75	69.07	0.51	4.24	22.08	99.54	0.44	6.45
CFOCNet [23]	21.19	61.41	—	—	22.10	112.71	—	—
BMNet+ [18]	15.74	58.53	0.25	2.73	14.62	91.83	0.27	6.20
CounTR [2]	13.13	49.83	0.23	2.59	11.95	91.23	0.23	7.44
SAFECount [24]	15.28	48.20	0.26	3.99	14.32	85.54	0.25	6.37
SPDCN [12]	14.59	49.97	0.22	2.79	13.51	96.80	0.22	6.70
LOCA [21]	10.24	**32.56**	0.22	2.09	10.79	56.97	0.19	**2.19**
MGCAC (Ours)	11.00	51.42	**0.12**	2.05	10.46	96.60	**0.16**	6.17
MGCAC* (Ours)	**9.93**	41.08	**0.12**	**1.91**	**9.10**	**54.21**	**0.16**	5.68

MGCAC * : 特別針對 count 多的 image 做訓練，可以當作 upper bound

其中他使用了 TTN :

計算 reference feature maps 的 mean 和 std

對 query images feature maps 做 normalization : (Q - R_mean) / R_sigma

再將調整完的 query images' 丟入模型

Test-Time Normalization（TTN）僅在 inference 階段使用。

如果 training 時使用了 TTN， 那有可能造成 model 過度依賴 TTN
具體來說 : 如果說好死不死 Inference 時遇到了一個 refference image 與 query image 差很多的 case， 就有可能會造成結果很糟

TTN 是一種 CAC 任務中，降低 Overfitting 的方法

FamNet 為什麼沒有補 MA, GL 的版本呢?

SPDCN 為什麼沒有補 MA, GL 的版本呢?

MA	GL	Model	MAE	RMSE	NAE	SRE
×	×	FamNet [17]	37.09	66.63	1.65	10.57
×	✓	SPDCN [12]	40.60	85.38	0.99	6.36

Abalation Study :

可以學習的地方

問題