Spatial Consistency Loss for Training Multi-Label Classifiers from Single-Label Annotations

Supplementary material

A. Data-augmentation settings

We use the following data-augmentation pipeline during trainings:

MS-COCO 2014, Pascal VOC 2012, NUS-WIDE, CUB-Birds 200-2011

Train

- Resize to square image of resolution 672×672
- Random square crop with cropped area uniformly varying between 0.25 and 1 (torchvision [38] RandomResizedCrop implementation), resized to 448×448
- Random horizontal flip

Test

- Resize to square image of size 448×448

Imagenet-1k ILSVRC2012

Train

- Random square crop with cropped area uniformly varying between 0.08 and 1 and aspect ratio between 3/4 and 4/3 (torchvision [38] RandomResizedCrop implementation with default arguments, same as [19]), resized to size 224×224
- Random horizontal flip

Test

- Resize smallest image side to 256
- Center crop of 224×224 pixels

B. Comparison with Hill/SPLC

Zhang *et al*. [59] use different splits on MS-COCO [33] to evaluate training from a single positive label. In addition, they perform experiments on the partial label settings where 75% and 40% of the positive labels are annotated, and no annotated negatives. We evaluate our method on their dataset with our setup as described in section 4. Table B.1 shows that out results surpass those of [59] in all scenarios.

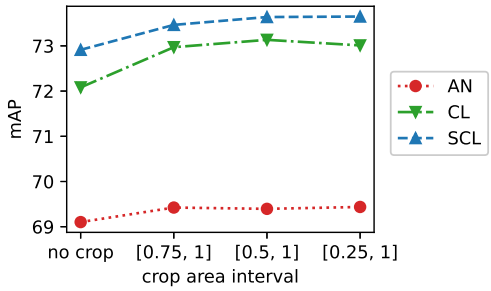

Figure B.1. Best MS-COCO validation mAP obtained when training with different data-augmentation crop area. The cropped area size, compared to the full image area, is randomly and uniformly sampled from the interval.

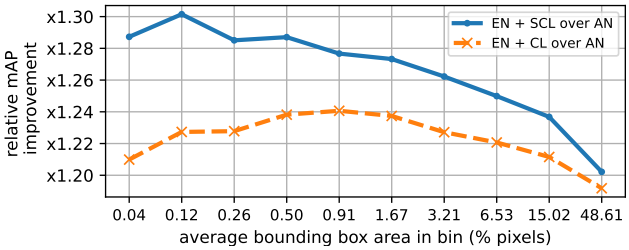

Figure D.1. Relative improvement per object size.

C. Ablation on the crop parameters

Figure B.1 shows the accuracies obtained with AN, and CL/SCL (with EN), when varying the random interval for the area of the crop data-augmentation. We see that CL and SCL are able to benefit more from the crop data-augmentation, compared to AN. This is consistent with our intuition that the crop data-augmentation can lead to incorrect supervision due to the single annotated objects being possibly partially or entirely cropped out. Moreover, SCL's improvements over CL are consistent over the different data-augmentation parameters.

D. Analysis over object sizes

We check the impact of object size by splitting the positive annotations of the COCO *val* split into equally-sized bins, grouped by relative area of the ground truth bounding box. Then, for each bin we compute the mAP using the positive labels within that bin, and negatives over the whole *val* split since negatives have no object size. Figure D.1 shows that the usage of consistency loss (CL) and spatial consistency loss (SCL) both improve mAP for all object sizes, compared to the AN baseline. Interestingly, SCL

Table B.1. Comparison with Hill/SPLC [59] with ResNet-50 [19] on MS-COCO [33]. Results with † are reported by [59]

		75% labels	40% labels	1 label
	BCE (fully annotated) †	80.32	80.32	80.32
Baselines	AN †	76.81	70.49	68.57
	WAN †	77.25	72.05	70.17
	BCE-LS †	78.27	73.13	70.53
Loss re-weighting	Focal [32] †	76.95	71.66	70.19
	ASL [42] †	77.97	72.70	71.67
	Hill [59] †	78.84	75.15	73.17
Loss correction	BCE + pseudo label †	77.05	71.46	69.77
	ROLE [9] †	78.43	73.67	70.90
	Focal margin + SPLC [59] †	78.44	75.69	73.18
Ours	BCE (fully annotated)	80.2	80.2	80.2
	AN	76.8	71.7	69.8
	EN + CL	77.6	75.8	74.1
	EN + SCL	**79.3**	**75.9**	**74.7**

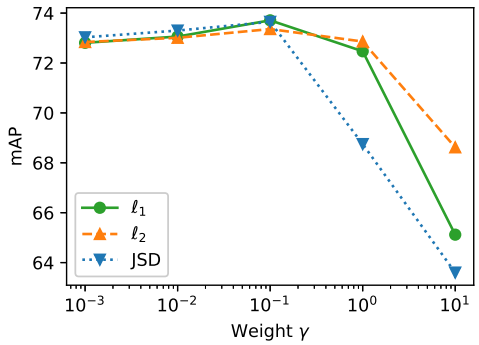

Figure E.1. Analysis of distance functions with ℓ_1 norm, ℓ_2 norm, Jensen–Shannon divergence (JSD) over weights γ.

yields higher mAP gains for smaller object sizes. Our hypothesis is that smaller objects are more likely to be cropped out, which is handled by the SCL. In addition, the crop augmentation zooms in on small objects, and those soft labels are recorded in the heatmaps as supervision.

E. Ablation of distance functions and weights

Figure E.1 compares different distance functions to measure the difference between exponential moving averages and predictions for (spatial) consistency losses.

F. Score distributions

Figure F.1 shows the distributions of the top-4 scores over all validation images. In contrast to the fully annotated baseline, the single-positive dataset in combination with AN loss leads to low-scoring predictions. SCL with EN loss (eq. (8)) reduces the amount of false negative labels and leads to a distribution more akin to the fully annotated case.

G. Details on heatmaps computation

We store heatmaps on 2 times the resolution of the feature maps (e.g. input resolution of 448×448 results in feature maps of 14×14 is stored in heatmaps of 28×28 pixels.). Heatmaps are stored in 8-bit unsigned integer format.

For ImageNet-1K [10] (section 4.3), we reduce the memory load by only keeping heatmaps for the top-k classes. The selection is based on the per-class EMA scores s_n^t computed as described in eq. (4), after the 5 epochs of pretraining the linear layer. In our experiments, we select the 10 highest-scoring classes per image based on s_{ni}^5. Heatmaps of other classes are assumed to be uniformly 0 in the SCL. Given 1.3 million training images, heatmaps of 14×14 and 1000 classes stored in `uint8`, this optimization reduces the required memory from approximately 250 GB to 2.5 GB.

H. Impact of SCL on heatmaps

A comparison of the heatmaps generated with and without SCL is given in fig. H.1, as an extra example in addition to fig. 4.

I. Uncurated heatmap examples

Figures K.1 and K.2 show the heatmaps corresponding to the samples with lowest COCO image id having suitable licenses for reproduction in the paper. In agreement with the observations in section 4.2, we see that the SCL tends to improve the object localization in the heatmaps, especially

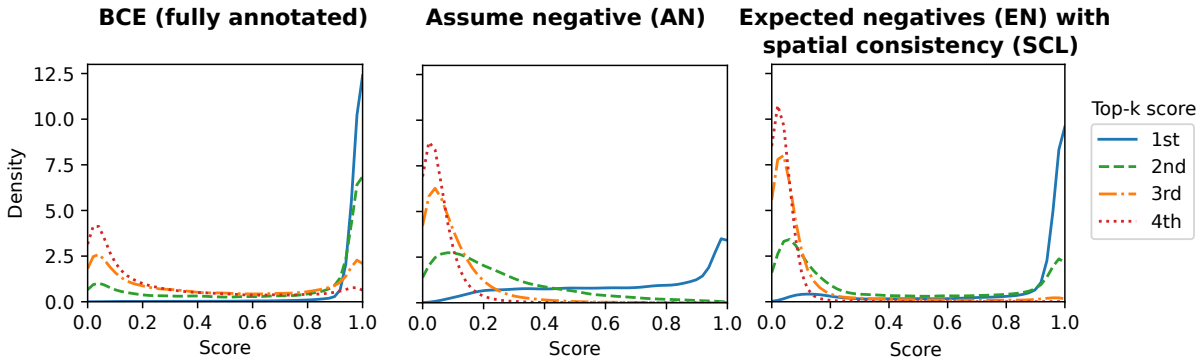

Figure F.1. Score distribution over all MS-COCO validation images, for 1st, 2nd, 3rd and 4th highest predicted scores per image. The BCE method is a fully annotated baseline. Training with AN and a single-positive label leads to a bias towards single positive predictions. With EN and SCL, the network more confidently predicts multiple positives.

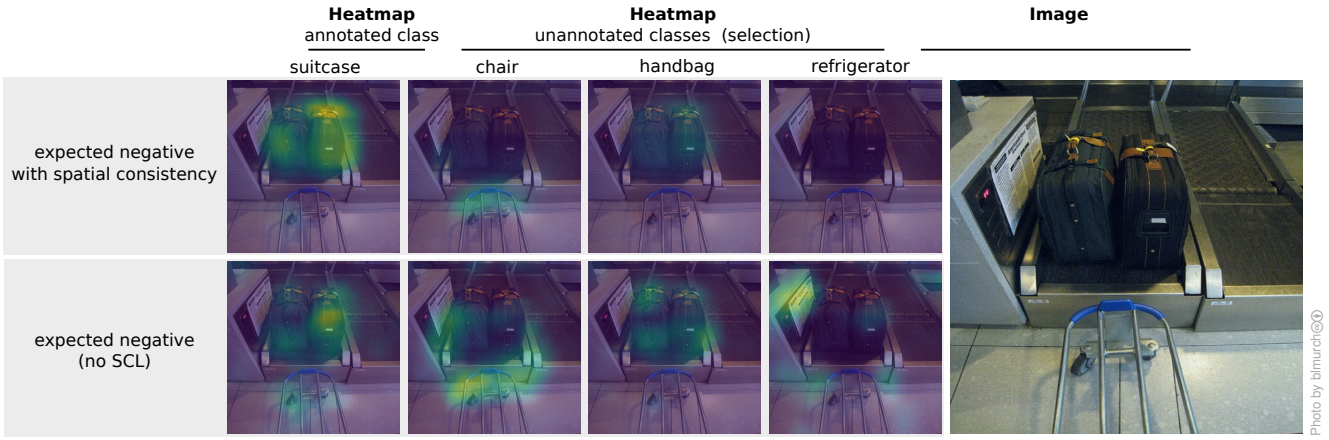

Figure H.1. Comparison of heatmaps generated in the final training epoch with and without spatial consistency loss (second example).

when looking at the negative classes which tend to be more present when using the EN alone.

J. Distribute property of final pooling and linear layer

To obtain predictions for each spatial position, we flip the order of the average pooling layer and the final linear classification layer. The linear layer can be executed as a 1×1 convolution over the feature map, resulting in class-wise predictions per spatial position. While this introduces extra computations at training time, the inference time is not impacted. Due to the distributive property, the order of the average pooling and 1×1 convolutions can be reversed at inference time without affecting the network outputs. Indeed, denoting by ϕ the $G \times G \times M$ network output before average pooling and 1×1 convolution, and by A the $M \times L$ matrix representing the 1×1 convolution, it can be seen that

$$\frac{1}{G^2} \sum_{g,g'=1}^{G} \sum_{m=1}^{M} A_{ml} \phi_{gg'm} = \sum_{m=1}^{M} A_{ml} \frac{1}{G^2} \sum_{g,g'=1}^{G} \phi_{gg'm}$$

for all l. That is, convolving and then average pooling is equal to average pooling and then convolving.

K. Dataset statistics

Table K.1 lists some statistics on the datasets used in the paper, as well as the value of the hyperparameter K computed on the validation set based on these statistics. Tables K.2 and K.3 show detailed breakdown of positive annotations per class in the MS-COCO and Pascal datasets using the splits of [9].

Figure K.1. Heatmaps and scores of the top-5 scoring classes in the last epoch training with EN+SCL, along with the corresponding heatmaps for EN alone.

Figure K.2. Heatmaps and scores of the top-5 scoring classes in the last epoch training with EN+SCL, along with the corresponding heatmaps for EN alone.

Table K.1. Dataset statistics. For COCO, VOC, NUS and CUB we use the train/val/test splits from [9]. For ImageNet-1K we report both the original [10] and multi-label ReaL [2] validation sets. K is the average number of positives per image on the validation set.

Dataset	Num. classes	Number of images			Number of annotations			K
		train	val	test	train	val	test	
MS-COCO 2014 [33]	80	65,665	16,416	40,137	193078	47957	116592	2.9
Pascal VOC 2012 [14]	20	4574	1143	5823	6665	1143	5823	1.5
NUS-WIDE [8]	81	120000	30000	60260	226833	57778	113418	1.9
CUB-200-2011 [49]	312	4795	1199	5794	150551	37792	182704	31.5
ImageNet-1K [10]	1000	1,281,167	50,000/46,837	-	1,281,167	50,000/46,837	-	1/1.2

Table K.2. Annotation statistics on MS-COCO [33]. For each class, we show the total amount of annotations in the original MS-COCO annotations (*total*), as well as the percentage of single-positive annotations selected for that class in the splits of [9].

Class	**# train**		**# val**		**# test**
	total	single-pos	total	single-pos	total
all classes	193078	34%	47957	34%	116592
person	36192	34%	8982	34%	21634
chair	7138	22%	1812	21%	4404
car	6895	30%	1711	30%	4180
dining table	6701	21%	1677	21%	3960
cup	5219	20%	1299	19%	3061
bottle	4790	20%	1178	21%	2912
bowl	4042	21%	986	22%	2397
handbag	3927	23%	934	20%	2272
truck	3447	33%	874	31%	2056
backpack	3109	25%	815	25%	1832
bench	3078	34%	766	35%	1961
book	2994	22%	740	23%	1828
cell phone	2644	29%	678	30%	1695
sink	2640	33%	651	34%	1574
tv	2525	23%	666	24%	1577
couch	2515	22%	655	22%	1448
clock	2506	50%	653	47%	1704
potted plant	2497	24%	587	23%	1540
knife	2491	20%	606	19%	1410
dog	2428	39%	613	39%	1521
sports ball	2401	30%	585	29%	1445
traffic light	2292	37%	601	36%	1437
cat	2267	43%	551	45%	1480
bus	2240	33%	551	34%	1350
umbrella	2183	30%	566	32%	1393
tie	2132	34%	535	34%	1288
fork	2058	18%	479	17%	1173
bed	2054	38%	485	39%	1292
vase	2025	35%	505	36%	1200
skateboard	2021	40%	490	40%	1092
spoon	2005	18%	488	21%	1189
motorcycle	1961	37%	481	38%	1219
train	1958	58%	506	62%	1281
laptop	1943	24%	532	24%	1232
tennis racket	1903	35%	465	37%	1193
surfboard	1876	44%	467	47%	1292
bicycle	1847	26%	440	30%	1114
toilet	1842	58%	475	59%	1185
airplane	1797	68%	446	69%	840
bird	1784	64%	457	64%	1121
skis	1775	44%	434	43%	993
remote	1750	25%	430	23%	1041
pizza	1734	37%	468	37%	1117
boat	1708	47%	390	43%	1048
cake	1670	30%	410	29%	969
horse	1668	52%	400	48%	1001
oven	1584	26%	419	28%	989
baseball glove	1519	30%	365	32%	845
baseball bat	1467	31%	337	30%	799
wine glass	1428	20%	343	18%	872
giraffe	1426	80%	372	82%	849
sandwich	1359	30%	286	31%	818
refrigerator	1344	27%	327	24%	790
banana	1316	40%	302	40%	728
suitcase	1313	34%	318	35%	876
kite	1286	42%	339	47%	727
elephant	1226	68%	292	65%	714
teddy bear	1219	47%	291	47%	724
frisbee	1215	43%	296	46%	757
keyboard	1161	21%	310	25%	750
cow	1124	67%	265	70%	666
broccoli	1080	41%	260	44%	670
zebra	1065	86%	259	88%	677
mouse	1008	23%	282	20%	674
orange	1003	34%	213	32%	568
stop sign	969	53%	245	52%	589
carrot	968	31%	218	35%	578
fire hydrant	954	52%	251	47%	592
apple	942	28%	229	31%	491
snowboard	936	41%	234	42%	533
donut	865	41%	197	40%	523
sheep	856	73%	249	75%	489
microwave	853	23%	236	25%	512
hot dog	661	38%	160	41%	452
toothbrush	570	36%	130	49%	341
scissors	535	44%	138	42%	302
bear	531	88%	137	86%	341
parking meter	395	42%	86	50%	261
toaster	125	28%	26	23%	74
hair drier	103	27%	25	28%	70

Table K.3. Annotation statistics on Pascal VOC 2012 [14]. For each class, we show the total amount of annotations in the original MS-COCO annotations (*total*), as well as the percentage of single-positive annotations selected for that class in the splits of [9].

Class	**# train**		**# val**		**# test**
	total	single-pos	total	single-pos	total
all classes	6665	68%	1666	68%	8351
person	1584	59%	410	66%	2093
dog	504	83%	128	82%	654
car	474	68%	116	60%	571
chair	459	49%	107	39%	553
cat	436	90%	103	87%	541
bird	310	93%	85	98%	370
bottle	294	51%	71	52%	341
aeroplane	264	95%	63	95%	343
tvmonitor	233	61%	57	63%	285
diningtable	221	45%	48	43%	269
train	220	85%	53	83%	271
pottedplant	214	47%	55	61%	258
boat	210	82%	50	72%	248
motorbike	206	65%	59	50%	261
sofa	201	53%	56	58%	250
bicycle	200	64%	68	66%	284
horse	195	69%	42	66%	245
bus	176	67%	37	64%	208
sheep	135	90%	36	86%	154
cow	129	86%	22	90%	152