On Train-Test Class Overlap and Detection for Image Retrieval

Chull Hwan Song[1], Jooyoung Yoon[1], Taebaek Hwang[1], Shunghyun Choi[1], Yeong Hyeon Gu[2*], Yannis Avrithis[3]

[1]Dealicious Inc. [2]Sejong University [3]Institute of Advanced Research on Artificial Intelligence (IARAI)

Introduction

Background
► **Importance of Image Retrieval:** Crucial for applications like search engines, e-commerce, and digital libraries.

Problem Statement
► **Non-overlapping Training and Evaluation Sets:** Essential to prevent inflated performance metrics and ensure true generalization.

Motivation
► **Dataset Issues:** Google Landmarks v2 (GLDv2) dataset has class overlaps with evaluation sets, compromising evaluation robustness.
► **Complex Methods:** Current methods involve complex two-stage processes, increasing computational overhead.

Objective
► **RGLDv2-clean:** A new version of GLDv2, cleaned to remove class overlaps.
► **CiDeR:** An end-to-end, single-stage image retrieval pipeline that simplifies the process and requires no location supervision.
► **Robust Evaluation:** Ensuring accurate performance evaluation by removing class overlaps and enhancing retrieval efficiency.

Contribution 1

Data Cleaning Process
► Analysis of the Google Landmarks v2 dataset: Identify and analyze duplicated images.
► Removal of overlapping landmarks with Revisited Oxford and Paris datasets using metadata and visual similarity.
► Verification: Confirm that duplicates are removed accurately.

Flowchart of the GLDv2-clean data cleaning process, including feature extraction, indexing, ranking, and verification steps.

EVAL	#EVAL IMG	#DUPL EVAL	#DUPL GLDV2 GID	#DUPL GLDV2 IMG
$\mathcal{R}$Par	70	36 (51%)	11	1,227
$\mathcal{R}$Oxf	70	38 (54%)	6	315
TEXT		1		23
TOTAL	140	74	18	1,565

Statistical information about duplicate images/categories with ($\mathcal{R}$Oxf, $\mathcal{R}$Par) and GLDV2.

TRAINING SET	#IMAGES	#CATEGORIES
NC-clean	27,965	581
SfM-120k	117,369	713
GLDv2-clean	1,580,470	81,313
$\mathcal{R}$GLDv2-clean (ours)	1,578,905	81,295

Statistics of clean landmark training sets for image retrieval.

Overlapping landmark images.

Contribution 2

Model Development and CiDeR Pipeline
► **Comparison of two-stage and single-stage pipelines:** The CiDeR model's single-stage pipeline integrates instance detection and embedding extraction.
► **Attentional Localization component:** This process uses spatial attention to detect objects of interest and fuse the resulting tensors into a single, localized feature tensor.

(a) Two-stage (b) One-stage

Important Equations

Spatial Attention Map:
$$A = \eta(\zeta(f^\ell(\mathbf{F}))) \in \mathbb{R}^{w \times h}$$
► The spatial attention map equation calculates the attention map for an input feature tensor, normalizing it to a $[0, 1]$ range.

Linear Normalization:
$$\eta(X) := \frac{X - \min X}{\max X - \min X} \in \mathbb{R}^{w \times h}$$
► The linear normalization equation transforms the input matrix to a normalized scale.

Mask Generation:
$$M_i(\mathbf{p}) = \begin{cases} \beta, & \text{if } A(\mathbf{p}) < \tau_i \\ 1, & \text{otherwise} \end{cases}$$
► The mask generation equation creates a binary mask based on the attention map values, applying a threshold.

Masked Feature Tensor:
$$\mathbf{F}^\ell = \mathcal{H}(M_1 \odot \mathbf{F}, \ldots, M_T \odot \mathbf{F}) \in \mathbb{R}^{w \times h \times d}$$
► The masked feature tensor equation combines multiple feature maps using Hadamard products with the masks.

Convex Combination for Fusion:
$$\mathcal{H}(\mathbf{F}_1, \ldots, \mathbf{F}_T) = \frac{w_1 \mathbf{F}_1 + \cdots + w_T \mathbf{F}_T}{w_1 + \cdots + w_T}$$
► The convex combination for fusion equation calculates the weighted sum of the masked feature maps.

Visualization

Visual Results
► Examples of top-5 ranking images retrieved by our CiDeR model from evaluation sets Ox5k/Par6k and associated spatial attention map
► The spatial attention map focuses exclusively on the object of interest as specified by the cropped area provided by the evaluation set, essentially ignoring the background.

Top-5 retrieved images with spatial attention maps.

Experiments

METHOD	TRAIN SET	BASE Ox5k mAP	Par6k mAP	MEDIUM $\mathcal{R}$Oxf mAP	$\mathcal{R}$Oxf mP@10	$\mathcal{R}$Par mAP	$\mathcal{R}$Par mP@10	HARD $\mathcal{R}$Oxf mAP	$\mathcal{R}$Oxf mP@10	$\mathcal{R}$Par mAP	$\mathcal{R}$Par mP@10	MEAN	DIFF
Yokoo et al. [46]	GLDv2-clean	91.9	94.5	72.8	86.7	84.2	95.9	49.9	62.1	69.7	88.4	79.5	-5.4
Yokoo et al. [60]†	$\mathcal{R}$GLDv2-clean	86.1	93.9	64.5	81.0	84.1	95.4	35.6	51.5	68.7	86.4	74.1	
SOLAR [58]	GLDv2-clean	–	–	79.7	–	88.6	–	60.0	–	75.3	–	75.9	-8
SOLAR [27]†	$\mathcal{R}$GLDv2-clean	90.6	94.4	70.8	84.6	84.1	95.4	48.0	62.3	68.7	86.4	67.9	
GLAM [46]	GLDv2-clean	94.2	95.6	78.6	88.2	88.5	97.0	60.2	72.9	76.8	93.4	83.4	-4.1
GLAM [46]‡	$\mathcal{R}$GLDv2-clean	90.9	94.1	72.2	84.7	83.0	95.0	49.6	61.6	65.6	87.6	79.3	
DOLG [47]	GLDv2-clean	–	–	78.8	–	87.8	–	58.0	–	74.1	–	74.7	-7.4
DOLG [59]†	$\mathcal{R}$GLDv2-clean	88.3	93.9	70.8	85.3	83.2	95.4	47.4	60.0	67.9	87.4	67.3	
Token [58]	GLDv2-clean	–	–	82.3	–	75.6	–	66.6	–	78.6	–	75.8	-18.2
Token [58]†	$\mathcal{R}$GLDv2-clean	84.3	90.0	61.4	76.4	75.8	94.0	36.9	55.2	54.4	81.0	57.6	

► Comparison of $\mathcal{R}$GLDv2-clean and GLDv2-clean datasets.

METHOD	TRAIN SET	NET	POOLING	LOSS	FT	E2E	SELF	DIM	BASE Oxf5k	Par6k	$\mathcal{R}$MEDIUM $\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$HARD $\mathcal{R}$Oxf	$\mathcal{R}$Par	MEAN
LOCAL DESCRIPTORS															
HesAff-rSIFT-ASMK*+SP [34]	SfM-120k	R50	–	CLS	✓	–	–	–	60.6	61.4	36.7	35.0	–		
DELF-ASMK*+SP [34]	SfM-120k	R50	–	CLS	✓	–	–	–	67.8	76.9	43.1	55.4	–		
LOCAL DESCRIPTORS+D2R															
R-ASMK* [48]	NC-clean	R50	–	CLS,LOCAL	✓	–	–	–	–	–	69.9	78.7	45.6	57.7	–
R-ASMK*+SP [48]	NC-clean	R50	–	CLS,LOCAL	✓	–	–	–	–	–	71.9	78.0	48.5	54.0	–
GLOBAL DESCRIPTORS															
DIR [47]	SfM-120k	R101	RMAC	TP	✓	–	–	2048	79.0	86.3	53.5	68.3	25.5	42.4	59.2
Radenovic et al. [36, 34]	SfM-120k	R101	GeM	SIA	–	–	–	2048	87.8	92.7	64.7	77.2	38.5	56.3	69.5
AGeM [9]	SfM-120k	R101	GeM	SIA	–	–	–	2048	–	–	67.0	78.1	40.7	57.3	–
SOLAR [47]	SfM-120k	R101	GeM	TP,SOS	✓	–	–	2048	78.5	86.3	52.5	70.9	27.1	46.7	60.3
GLAM [46]	SfM-120k	R101	GeM	AF	–	–	–	512	89.7	91.1	66.2	79.3	41.6	54.3	69.7
DOLG [47]	SfM-120k	R101	GeM,GAP	AF	–	–	–	512	72.8	74.5	44.6	56.6	18.1	26.6	49.2
GLOBAL DESCRIPTORS+D2R															
Mei et al. [26]	[O]	R101	FC	CLS	–	–	–	4096	38.4	–	–	–	–	–	–
Salvador et al. [43]	Pascal VOC	V16	GSP	CLS,LOCAL	✓	–	–	512	67.9	72.9	–	–	–	–	–
Chen et al. [5]	OpenImageV4 [17]	R50	MAC	MSE	–	–	–	2048	50.2	65.2	–	–	–	–	–
Liao et al. [22]	Oxford,Paris	A,V16	CroW	CLS,LOCAL	–	–	–	768	80.1	90.3	–	–	–	–	–
DIR+RPN [5]	NC-clean	R101	RMAC	TP	✓	–	–	2048	85.2	94.0	–	–	–	–	–
CiDeR (Ours)	SfM-120k	R101	GeM	AF	✓	✓	✓	2048	89.9	92.0	67.3	79.4	42.4	57.5	71.4
CiDeR-FT (Ours)	SfM-120k	R101	GeM	AF	✓	✓	✓	2048	92.6	95.1	76.2	84.5	58.9	68.9	79.4

► Performance Comparison of State-of-the-Art Methods on Existing Clean Datasets.

METHOD	BASE Ox5k mAP	Par6k mAP	MEDIUM $\mathcal{R}$Oxf mAP	$\mathcal{R}$Oxf +$\mathcal{R}$1M mP@10	$\mathcal{R}$Par mAP	$\mathcal{R}$Par +$\mathcal{R}$1M mP@10	HARD $\mathcal{R}$Oxf mAP	$\mathcal{R}$Oxf +$\mathcal{R}$1M mP@10	$\mathcal{R}$Par mAP	$\mathcal{R}$Par +$\mathcal{R}$1M mP@10
GLOBAL DESCRIPTORS (SFM-120K)										
DIR [47]	79.0	86.3	53.5	76.9	–	–	68.3	97.7	–	–	25.5	42.0	–	–	42.4	83.6	–	–
Filip et al. [36, 34]	87.8	92.7	64.7	84.7	45.2	71.7	77.2	98.1	52.3	95.3	38.5	53.0	19.9	34.9	56.3	89.1	24.7	73.3
AGeM [9]	–	–	67.0	–	–	–	78.1	–	–	–	40.7	–	–	–	57.3	–	–	–
SOLAR [47]	78.5	86.3	52.5	73.6	–	–	70.9	98.1	–	–	27.1	41.4	–	–	46.7	83.6	–	–
GeM [47]	79.0	82.6	54.0	72.5	–	–	64.3	92.6	–	–	25.8	42.2	–	–	36.6	67.6	–	–
GLAM [47]	89.7	91.1	66.2	–	–	–	77.5	–	–	–	39.5	–	–	–	54.3	–	–	–
DOLG [47]	72.8	74.5	46.4	66.8	–	–	56.6	91.1	–	–	18.1	27.9	–	–	26.6	62.6	–	–
CiDeR (Ours)	89.9	92.0	67.3	85.1	50.3	75.5	79.4	97.9	51.4	95.7	42.4	56.4	22.4	35.9	57.5	87.1	22.4	69.4
CiDeR-FT (Ours)	92.6	95.1	76.2	87.3	60.5	78.6	84.5	98.0	56.9	95.9	58.9	71.1	36.8	55.7	68.9	91.3	30.1	73.9
GLOBAL DESCRIPTORS ($\mathcal{R}$GLDV2-CLEAN)										
Yokoo et al. [60]† (Base)	86.1	93.9	64.5	81.0	51.3	72.1	84.1	95.4	54.2	90.3	35.6	51.5	22.2	42.9	68.7	86.4	27.4	66.9
SOLAR [27]†	90.6	94.4	70.8	84.6	55.8	76.1	80.3	94.6	57.6	92.0	48.0	62.3	30.3	45.3	61.8	83.9	30.7	71.6
GLAM [46]‡	90.9	94.1	72.2	84.7	58.6	76.1	83.0	95.0	58.6	91.7	49.6	61.6	34.1	50.9	65.6	87.6	33.3	72.1
DOLG [59]†	88.3	93.9	70.8	85.3	57.3	76.8	83.2	95.4	57.3	92.0	47.4	60.0	29.5	46.2	67.9	87.4	32.7	72.4
Token [58]†	81.2	93.9	60.8	77.7	44.0	60.9	75.8	94.3	44.1	86.9	37.3	54.1	23.2	37.7	54.8	81.3	19.7	54.4
CiDeR (Ours)	89.8	94.6	73.7	85.5	58.6	76.3	84.6	96.7	59.0	95.1	54.9	66.6	34.6	54.7	68.5	89.1	33.5	76.9
CiDeR-FT (Ours)	90.9	96.1	77.8	88.0	61.8	78.0	87.4	97.0	61.6	94.3	61.9	70.4	39.4	56.8	75.3	90.0	35.8	72.7

► Large-Scale Performance Comparison on $\mathcal{R}$GLDv2-clean and SfM-120k.

Conclusion

► Emphasize the critical nature of non-overlapping training and evaluation sets.
► Demonstrates the effectiveness of our data cleaning process.
► Summary of CiDeR's advantages and performance improvements.