Visual Place Recognition Overview and Application to Postcards Search

Students: Ilias Eldarov, Sergey Dubovitsky

Thesis adviser: Vladislav Goncharenko

MIPT, Spring 2023

girafe ai

Outline

1. Problem statement
2. Datasets
3. Metrics
4. Classical approach
5. NN models
    a. NN scheme
    b. Loss functions
    c. HybridNet/AMOSNet
    d. NetVLAD
    e. Google Landmark Recognition Challenge
6. Results so far

Motivation: Postcards Collection

Input:

- 15 million unlabeled images dated 1900-1960
- humanly written description for each image
- images cover entire world

Objective:

make it a collection of geographical postcards

Tasks:

- Filter out duplicates, low quality, non-landmark, low interest postcards
- Determine landmarks
- Allow visual search by query picture
- Allow images search by landmark

girafe ai

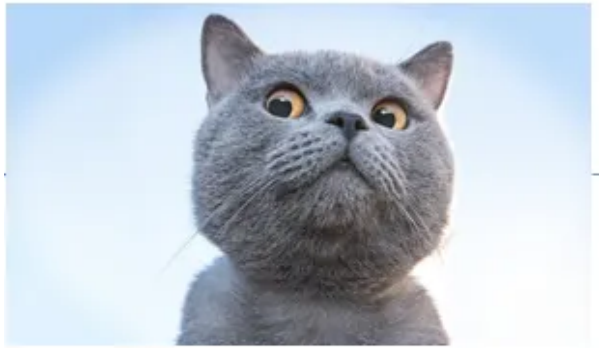
Visual Place Recognition Problem

Determine which place is depicted on the query image

girafe ai

Image Retrieval Problem

Given a query image find similar images from a database of known images



"Query" Image

Datasets

- Simultaneous Location And Mapping Datasets
- Landmarks Datasets
- Revisited London/Paris
- Google Landmarks V2
- Our extension to Google Landmarks V2
- KISA
- Good or Bad Image for Collection

SLAM-related Datasets

Dataset	Appearance Change	Setup	#Seq	GT
Alderley Day/Night	day, night/rain	car	2	ID
CMU	time, 4-seasons	car	16	raw GPS
Freiburg Across Seasons	summer/sunny, winter/low-sun, long-term change	car	3	ID
Freiburg-Bonn	summer-winter, morning-evening	car	2x2	ID
Gardenspoint Walking	day-night, viewpoint	hand-held	3	ID
NCLT	sunny/cloudy, time, seasons	robot	27	poses
Nordland	seasons, no viewpoint	train	4	ID
Oxford RobotCar	time, weather, seasons	car	>100	raw GPS
SFU Mountain Dataset	time, day-night, weather, seasons	robot	7	ID
South Bank Bicycle	day-night, viewpoint	bicycle	2	ID
StLucia	time of day	car	10	raw GPS
Symphony Lake	illumination, seasons	boat	121	poses
V4RL Urban Place	viewpoint	hand-held	3	ID
All day	day-night, time, dusk, dawn	car	6	raw GPS
City Center	dynamics, shadows	robot	1	ID
DIRD	time	car	2	not available
Ford Campus	sunny, overcast	car	2x1	raw GPS
Kelvin Grove Footpath	day-night	hand-held	2	ID
Kitti Odometry	dynamics	car	11	poses
Malaga Urban Dataset	dynamics	car	1	raw GPS
Mapilary Berlin	dynamics, viewpoint	bicycle, bus	2	raw GPS
New College	dynamics	robot	1	ID
Örebro Seasons	seasons, weather	robot	7	poses
PitOrlManh	dynamics, 3 cities	street view	3x1	poses
RAWSEEDS	time	robot	12+5	poses
StLucia Vision Dataset	none (no loop)	car	1	raw GPS
Surfers Paradise	day/rain-night	car	2	ID

Landmarks Datasets

Dataset name	Year	# Landmarks	# Test images	# Train images	# Index images	Annotation collection	Coverage	Stable
Oxford [41]	2007	11	55	-	5k	Manual	City	Y
Paris [42]	2008	11	55	-	6k	Manual	City	Y
Holidays [28]	2008	500	500	-	1.5k	Manual	Worldwide	Y
European Cities 50k [5]	2010	20	100	-	50k	Manual	Continent	Y
Geotagged StreetView [32]	2010	-	200	-	17k	StreetView	City	Y
Rome 16k [1]	2010	69	1k	-	15k	GeoTag + SfM	City	Y
San Francisco [14]	2011	-	80	-	**1.7M**	StreetView	City	Y
Landmarks-PointCloud [35]	2012	1k	10k	-	205k	Flickr label + SfM	Worldwide	Y
24/7 Tokyo [56]	2015	125	315	-	1k	Smartphone + Manual	City	Y
Paris500k [61]	2015	13k	3k	-	501k	Manual	City	N
Landmark URLs [7, 22]	2016	586	-	140k	-	Text query + Feature matching	Worldwide	N
Flickr-SfM [44]	2016	713	-	120k	-	Text query + SfM	Worldwide	Y
Google Landmarks [39]	2017	30k	**118k**	1.2M	1.1M	GPS + semi-automatic	Worldwide	N
Revisited Oxford [43]	2018	11	70	-	5k + 1M	Manual + semi-automatic	Worldwide	Y
Revisited Paris [43]	2018	11	70	-	6k + 1M	Manual + semi-automatic	Worldwide	Y
Google Landmarks Dataset v2	2019	**200k**	**118k**	**4.1M**	762k	Crowsourced + semi-automatic	Worldwide	Y

Revisited Paris/Oxford Datasets, 2018

girafe ai

Paper https://arxiv.org/pdf/1803.11285.pdf

Revisited Paris/Oxford Datasets, 2018

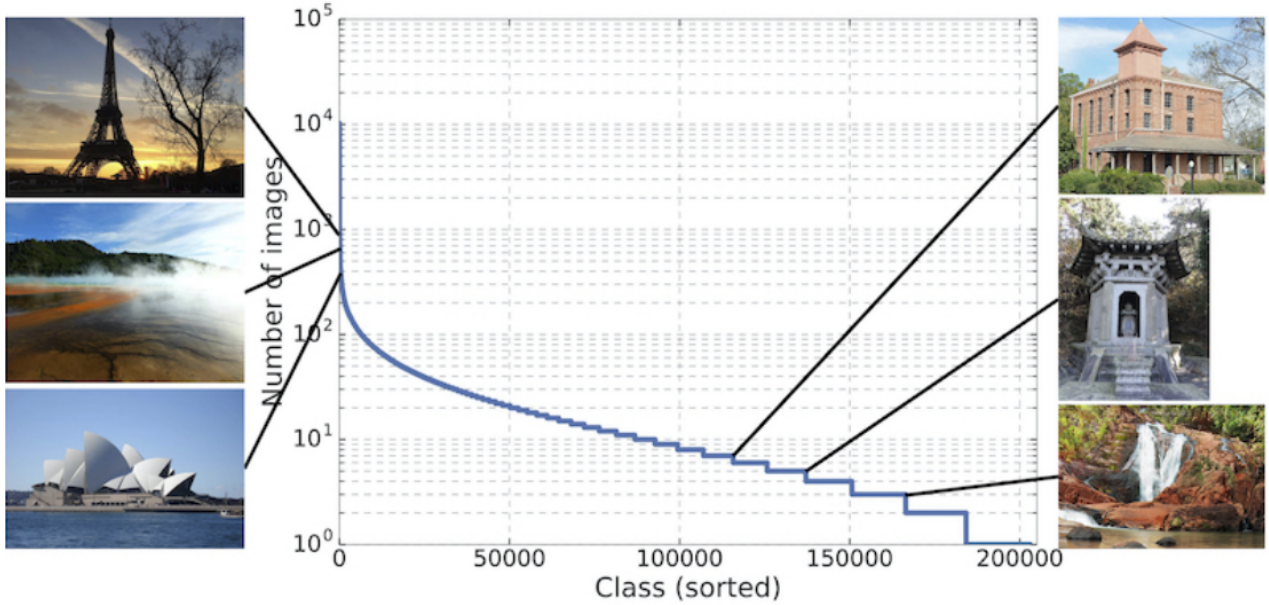
- One of the standard datasets to measure Image retrieval performance
- Image retrieval categories: Easy, Hard and Unclear, Negative

- 6K/7K positive pictures per city, 1 Million “distract images” from all over the world in the index set
- Update to Paris/Oxford datasets to fix annotation errors and make prediction harder

Google Landmarks v2 Dataset, 2019

Paper: https://arxiv.org/abs/2004.01804

Google Landmarks v2 Dataset, 2019



Class distribution is long tailed. ie classes are imbalanced

girafe ai

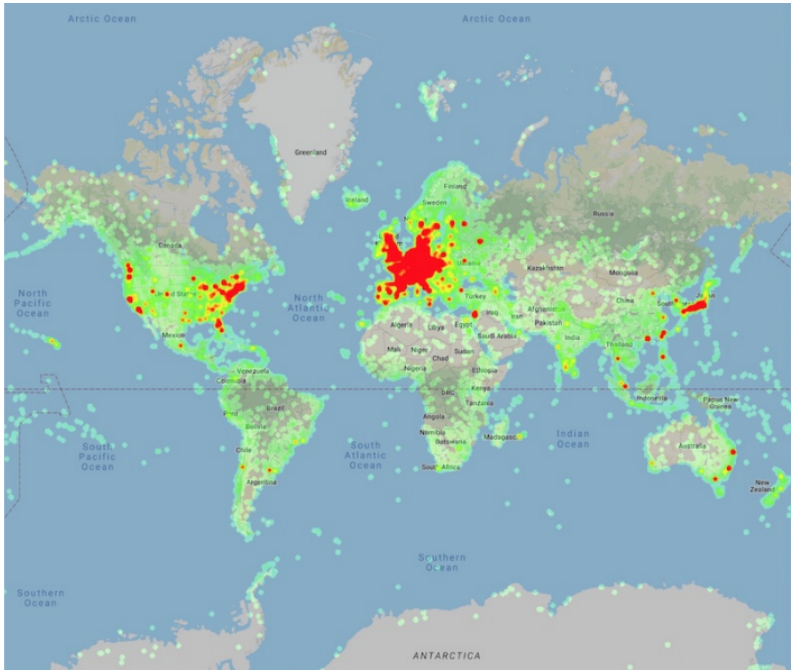
Paper: https://arxiv.org/abs/2004.01804

Google Landmarks v2 Dataset, 2019

- Suitable for VPR (203K landmarks) and IR (101K classes-landmarks) tasks

- landscape, outdoor, day/night, year-round, urban, rural
- Volume: ~4.1 M Train , 118K Test, 762K Index images
- “clean” version: 1.6 M Train, 81K VPR classes

- 99% of test images are out of domain
- Geography: entire world
- Origin: Wikipedia
- Foundation to a number of Kaggle competitions
- Successor to GLM v1 Dataset



Heatmap of image locations in GLMv2

Paper: https://arxiv.org/abs/2004.01804

Postcards Extension to GLM

- Intended for **VPR experiments**
- Scanned postcards, 1900-1960
- 2,7K classes, 7+ pictures per class
- 33K train images
- 5.5K test images
- Humanly created text annotation for each image
- OCR extracted for most of images

Classes assembled by IR,
identifying images similar to each other from entire collection
(cosine similarity 0.66 to 0.8)
and manually filtered from obvious errors and out of domain images

KISA (1K Image Similarity Assembly)

Objective: measure Image Retrieval performance on postcard type of images (1900-1960)

- 500 landmark postcard images, 101 landmark categories
- 500 non-landmark postcards, no similarity with each other and landmark images

girafe ai

DS Download: https://github.com/Iliasoft/VPR

Is it good for collection?

Objective: detect non-landmark, low quality, low interest images
Content:: 110K manually annotated images (20K good)

Classification mean: custom NN was created
Achieved Result: classification with 0.9 accuracy (Good/Bad)

Metrics

- Precision@K
- R-precision
- Recall@K
- mAP@K

Metrics

$$precision@k = \frac{REL_k}{k}$$

REL_k – number of relevant items in TOP-k results

k – fixed number of top positions in considerations

$$R\text{-}precision = \frac{REL_R}{R}$$

REL_R – number of relevant items in TOP-R results

R – number of all revelant items to a given query

girafe ai

https://habr.com/ru/articles/550604/

Metrics

$$recall@k = \frac{REL_k}{min(k, REL)}$$

REL_k – number of relevant items in TOP-k results

k – fixed number of top positions in considerations

REL – total number of relevant items to a given query

$mAP = \frac{1}{N} \sum_{n=1}^{N} AP@K_n$

where $AP@K = \frac{1}{r} \sum_{k=1}^{K} precision@k \cdot rel(k)$

N - number of Image Retrieval queries

K - items to be recommended

r - total relevant items

$rel(k)$ - indicator if that k-th item was relevant $rel(k) = 1$ or not $rel(k) = 0$

girafe ai

http://sdsawtelle.github.io/blog/output/mean-average-precision-MAP-for-recommender-systems.html

Classical Approach

Image Descriptors

- Edge boxes object proposal

- Hand-crafted local image descriptors

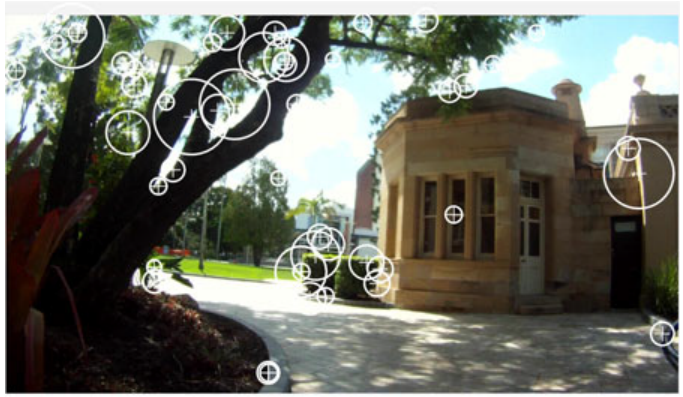
  SIFT, SURF, Root-SIFT, BRIEF, DELG,

  SuperPoint, etc.

- Global image descriptors: VLAD,

  Fisher vectors, ASMK, HOG, GIST

- VPR solved as IR + geotagged DB



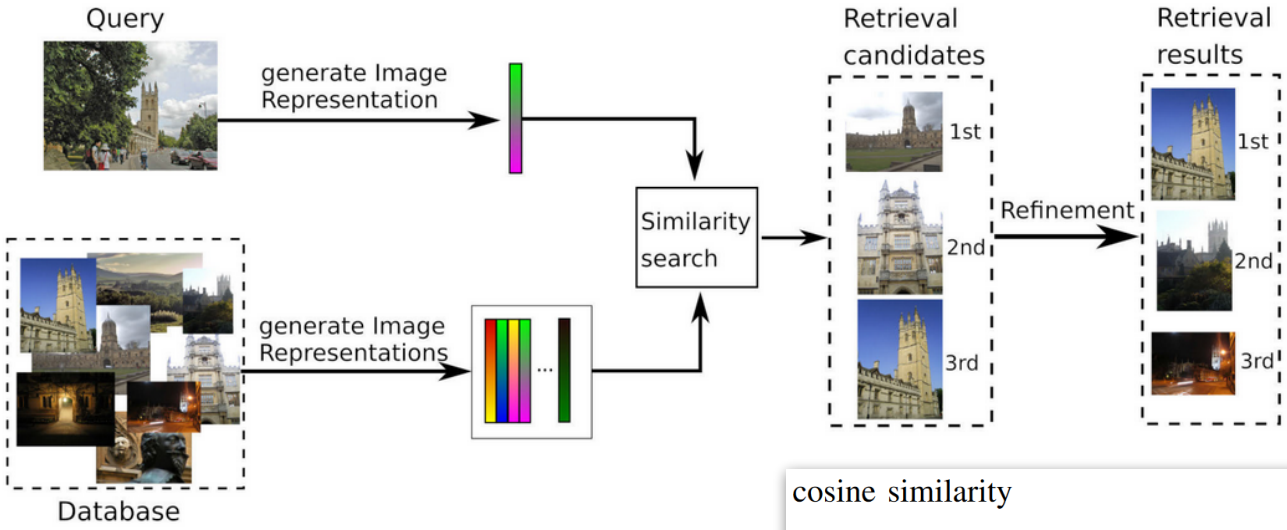
(a)

(b)

girafe ai

Image credit: https://www.researchgate.net/publication/284754832_Visual_Place_Recognition_A_Survey

Image Retrieval Approach



cosine similarity

$$s_{ij} = \frac{\mathbf{d}_i^T \cdot \mathbf{d}_j}{\|\mathbf{d}_i\| \cdot \|\mathbf{d}_j\|},$$

or the negative Euclidean distance

$$s_{ij} = -\|\mathbf{d}_i - \mathbf{d}_j\|.$$

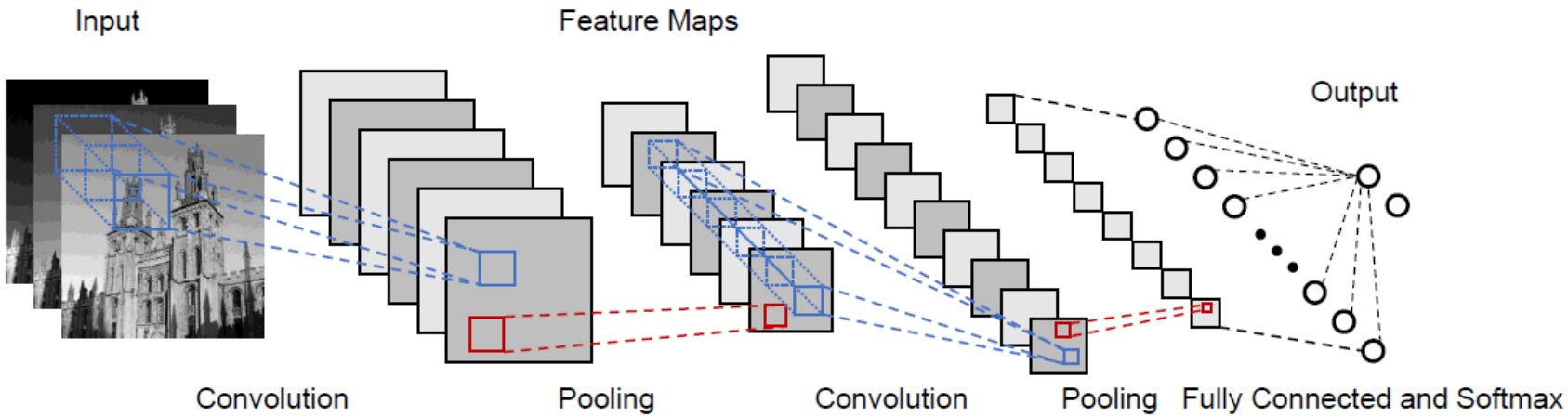
girafe ai

Image credit: https://www.researchgate.net/publication/348822040_A_Survey_on_Deep_Visual_Place_Recognition

NN Models

- NN scheme
- Loss functions
- HybridNet/AMOSNet
- NetVlad
- Google Landmark Recognition Challenge 2021

VPR via CNN-based classification

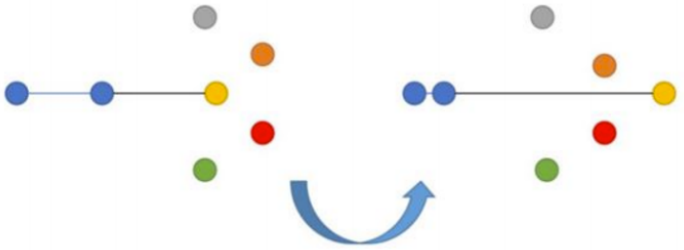


- Learn descriptors with CNNs
- Vectors for IR are obtained from the last feature layer preceding FC
- Train fully connected network as classifier (descriptor => place)

girafe ai

Image credit: https://paperswithcode.com/paper/visual-place-recognition-a-survey-from-deep

Triplet loss



$$L = max\{0, ||f(I_a) - f(I_p)||_2^2 - ||f(I_a) - f(I_n)||_2^2 + m\}$$

$I_a -$ anchor image

$I_p -$ positive image (similar to I_a)

$I_n -$ negative image (dissimilar to I_a)

$m -$ parameter to avoid convergence to trivial solutions

girafe ai

https://habr.com/ru/articles/550604

Contrastive loss

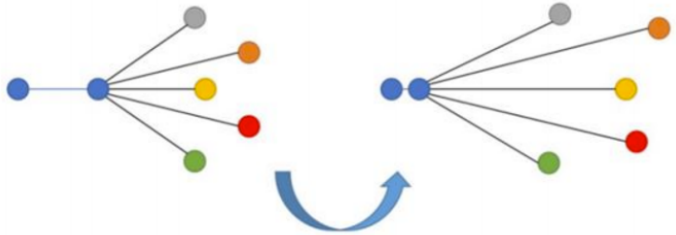
$$L = y * ||f(I_p) - f(I_q)||_2^2 + (1 - y) * max\{0, m - ||f(I_p) - f(I_q)||_2\}^2$$

$y = 1$ if images I_p, I_q are similar (positive pairs)

$y = 0$ if images I_p, I_q are dissimilar (negative pairs)

$m -$ margin which dissimilar samples should stay apart from

N-tupled loss (improved Triplet loss)



$$L = log[1 + \sum_{i=1}^{N} exp(||f(I_a) - f(I_p)||_2^2 - ||f(I_a) - f(I_n^i)||_2^2)]$$

I_a – anchor image

I_p – positive image (similar to I_a)

I_n^i – i-th negative image (dissimilar to I_a)

$N + 2$ – size of a tuplet (chosen empirically)

Angular Additive Margin (ArcFace)

1. Softmax loss:

$$W_j^T x_i = \|W_j\| \, \|x_i\| \cos\theta_j$$

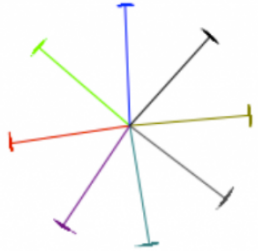
2. Transform

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s\,cos\theta_{y_i} + b_{y_i}}}{e^{s\,cos\theta_{y_i} + b_{y_i}} + \sum_{j=1, j\neq y_i}^{n} e^{s\,cos\theta_j + b_j}}$$

3. Additive margin:

$$L = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s\,cos(\theta_{y_i} + m) + b_{y_i}}}{e^{s\,cos(\theta_{y_i} + m) + b_{y_i}} + \sum_{j=1, j\neq y_i}^{n} e^{s\,cos\theta_j + b_j}}$$

(a) Softmax

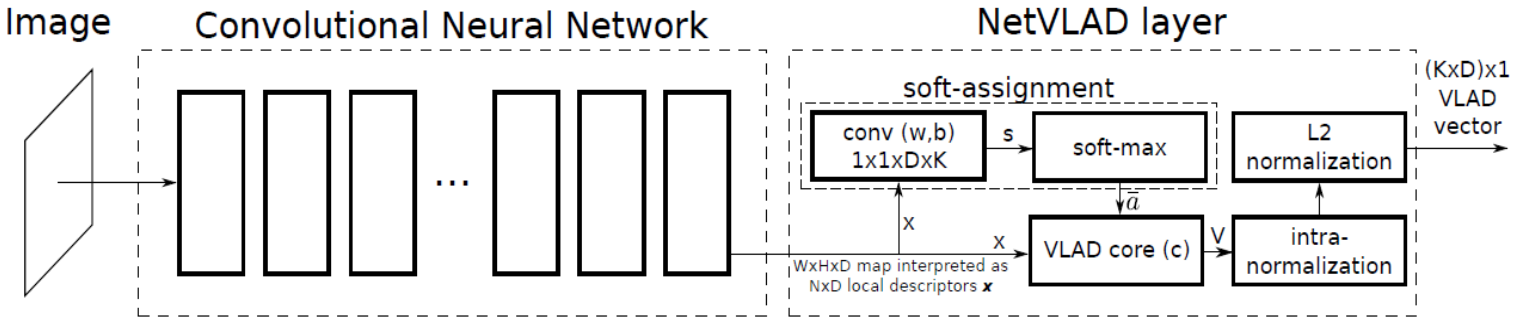


(b) ArcFace

girafe ai

https://habr.com/ru/articles/550604

NetVLAD: CNN architecture for weakly supervised place recognition, 2016

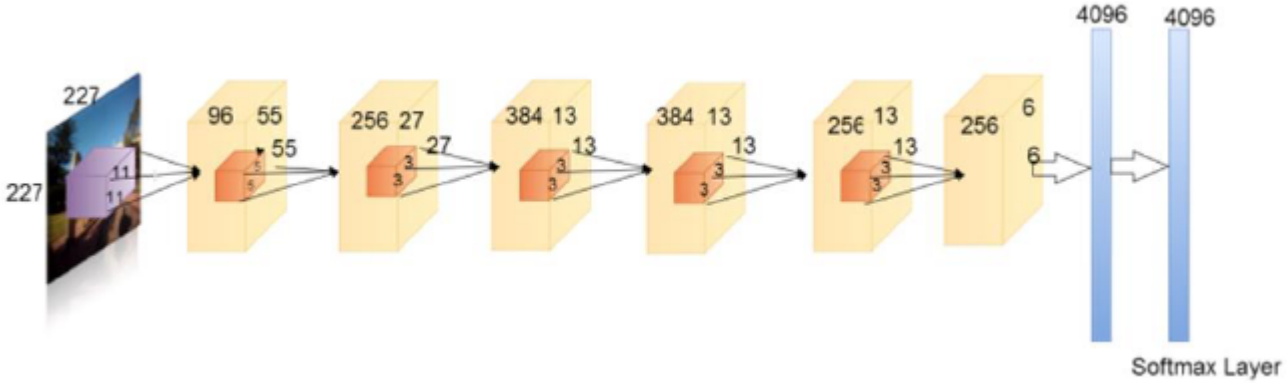


- V
- P
- VGG as backbone CNNs

girafe ai

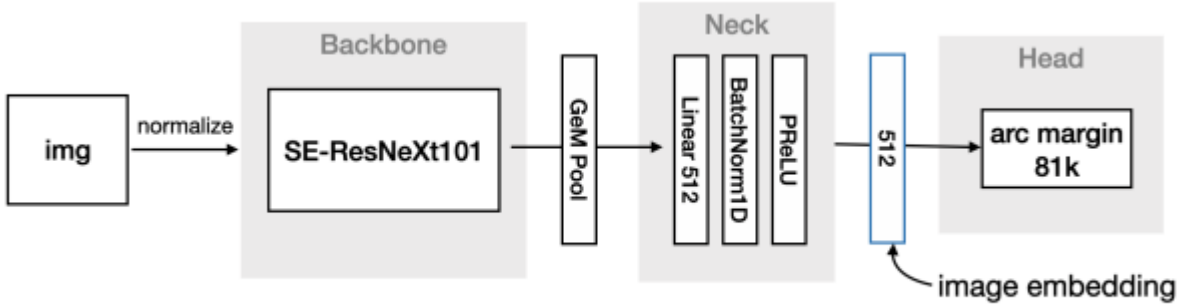
Paper: https://openaccess.thecvf.com/content_cvpr_2016/papers/Arandjelovic_NetVLAD_CNN_Architecture_CVPR_2016_paper.pdf

HybridNet/AMOSNet, 2017



- SPED Dataset (outdoor surveillance cameras)
- HybridNet shares same architecture with AMOSNet but uses CaffeNet weights for first layers, outperforms AMOSNet

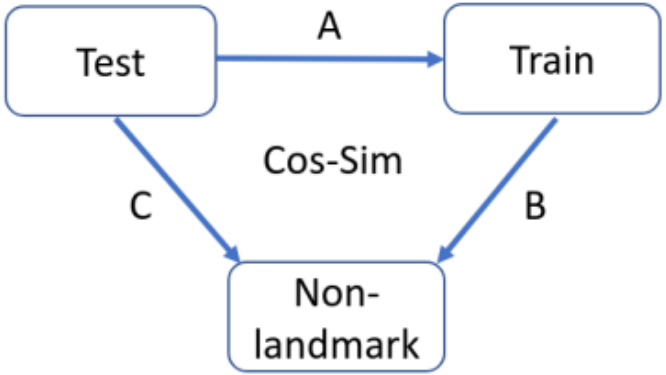
Google Landmark Recognition Challenge 2021



Model architecture with a SE-ResNeXt101 backbone

- VPR as classification + results refinement with IR/similarity
- Ensemble of 7 models based on SE-ResNeXt101, EfficientNet B3, ResNet152 and Res2Net101, trained on different image sizes/crop sizes, loss and pooling settings
- Trained on Google Landmarks “Clean” Dataset v2 (1.5 M images, 81K places)

VPR: Issue with non-landmarks in Test



99% of GLM Test images are non-landmark, these should yield no VPR prediction.

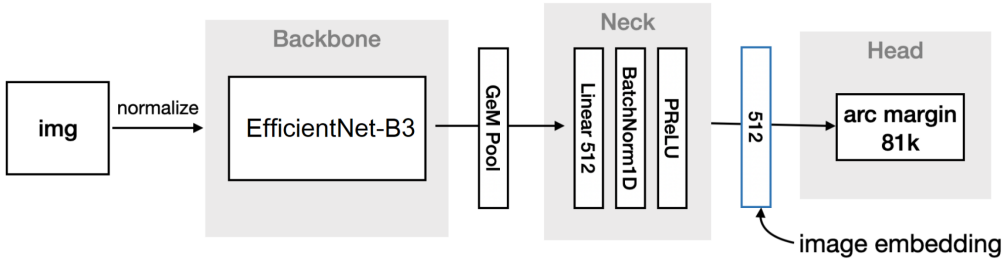
Possible ways to deal with non-landmark predictions:

- During training penalize model for wrong prediction
- Predict landmark status in addition to VPR class
- Introduce threshold for predictions, ignore prediction if model is not sure enough

Our Solution

- Baseline model
- Experiments and model derivations
- OCR Application

Solution: Model #1 (baseline)

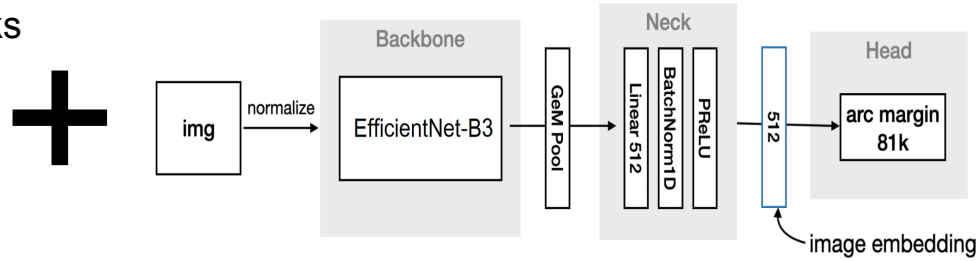


- EfficientNet-B3 was used as a backbone - a compromise between training time and VPR/IR performance
- Trainable parameters: 88 M, image embedding size: 512
- 448 x 448, ImageNet normalized images accepted as input
- Generalized mean (GeM) pooling was used to construct global image descriptors from CNN activations, ArcMargin as classifier and ArcFace as a loss function
- PyTorch lightning-based implementation, trained with Adam for 10 epochs, LR: 0.05

- Other backbones tested: SE-ResNeXt101, ResNet152 and Res2Net101

Model 1 trained on GLM

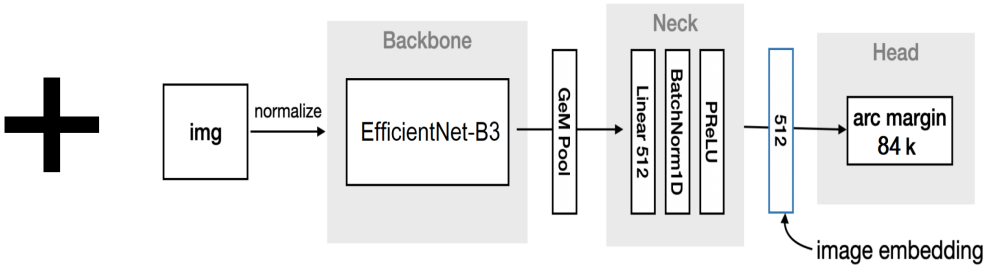
Google Landmarks Images



Model 2 trained on GLM & Postcards extension

Google Landmarks images

Scanned postcards



girafe ai

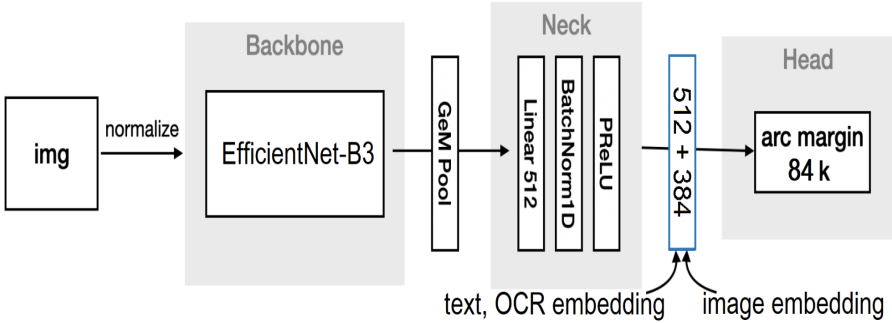
Download trained model weights: https://github.com/lliasoft/VPR

Models 3, 4 trained on GLM,Postcards & Text

Google Landmarks

Scanned postcards

+



OCR
+ human made annotation for scanned postcards

Model 3: GLM + Postcards + OCRs
Model 4: GLM + Postcards + OCRs + text annotations

All-MiniLM-L6-v2 as Sentence Embedder

girafe ai

Download trained model weights: https://github.com/Iliasoft/VPR

OCR Application

Postcards bear textual data which can be used to improve VPR accuracy

Easy OCR library was used to extract text from postcards

Easy OCR:

- is a deep learning solution based on PyTorch
- outperforms vanilla Tesseract in accuracy
- has multilanguage support
- can use GPU to speed up recognition
- written entirely on Python

Text cleanup was done after OCR in order to focus on geolocation specific words

girafe ai

https://github.com/JaidedAI/EasyOCR

Results Achieved

Experiment	Train data	Test data	Acc	Result Interpretation
Experiment 1.1	GLM Train, 1.5M	GLM Test, 101K	0.659	Reproduced GLM competition experiment (baseline model)
Experiment 1.2	GLM Train, 1.5M	GLM Test valid landmarks, 2K	0.880	Reproduced GLM competition experiment with valid LM only (baseline model)
Experiment 2.1	GLM Train, 1.5M + Postcards, 33K	GLM Test valid landmarks, 2K	0.885	Postcard extension to GLM did not decrease performance of the model
Experiment 2.2	GLM Train, 1.5M + Postcards, 33K	GLM Test valid landmarks, 2K + Postcards 2K	0.967	Performance of the model on mix of GLM images and extension (postcards) is good
Experiment 2.3	GLM Train, 1.5M + Postcards, 33K	Postcards 5.5K	0.983	Performance of the model on extension (postcards) only is better than on GLM images
Experiment 3	GLM Train, 1.5M + Postcards + OCR, 33K	Postcards 5.5K + OCR	0.986	Adding OCR info on the training phase increased model performance for postcards
Experiment 4.1	GLM Train, 1.5M + Postcards + Text annotations + OCR, 33K	GLM Test valid landmarks, 2K	0.890	Adding text annotation on the training phase on top of OCR did not decreased model performance for GLM images
Experiment 4.2	GLM Train, 1.5M + Postcards + Text annotations + OCR, 33K	Postcards 5.5K + OCR + Text annotations	0.992	Adding text annotation on the training phase on top of OCR increased model performance for postcards

IR on GLM and KISA

Metric	k=3 (GLM)	k=5 (GLM)	k=10 (GLM)	k=100 (GLM)	k=3 (KISA)	k=5 (KISA)
Precision@k	0.51	0.43	0.32	0.05	0.86	0.72
R-Precision	0.16	0.16	0.16	0.16	0.89	0.89
Mean Average Precision@k	0.46	0.41	0.36	0.3	0.93	0.93
Recall@k	0.52	0.47	0.41	0.4	0.65	0.57

IR tests were performed:

- on GLM Test (101K) against GLM Index 760K images
- on all landmark KISA images (500) against all other landmark KISA images.

Model was trained on GLM Train for both tests

IR: Near Identical Images Identification

IR: Similar Images Identification

IR: typical mistakes

The model has issues distinguishing

- multiple postcards on single image
- non-landmark images
- 2-sided postcards
- aerial images
- abstract drawings
- paintings
- faces
- animals
- cars, ships, trains etc.

Suggested reading

Approach on IR task

https://itnan.ru/post.php?c=1&p=550604&ysclid=lem22yatwc5432008 (Russian):

https://arxiv.org/abs/2101.11282 (English)

VPR model we used

https://www.kaggle.com/c/landmark-recognition-2020/leaderboard

https://arxiv.org/abs/2010.01650

https://github.com/psinger/kaggle-landmark-recognition-2020-1st-place

ArcFace Loss:

https://arxiv.org/abs/1801.07698

https://www.kaggle.com/code/josealways123/understanding-arcface-and-adacos

https://www.programmersought.com/article/26401469926/

GeM Pooling https://arxiv.org/abs/1711.02512

Google Landmark V2 dataset

https://github.com/cvdfoundation/google-landmark

https://arxiv.org/abs/2004.01804

KISA dataset https://drive.google.com/file/d/1eAiH5o32u8Ctt0UxcGX_PY3-S238wbT6

Landmark And Non-Landmark Domain Assembly dataset - tbd

Project github - https://github.com/lliasoft/VPR