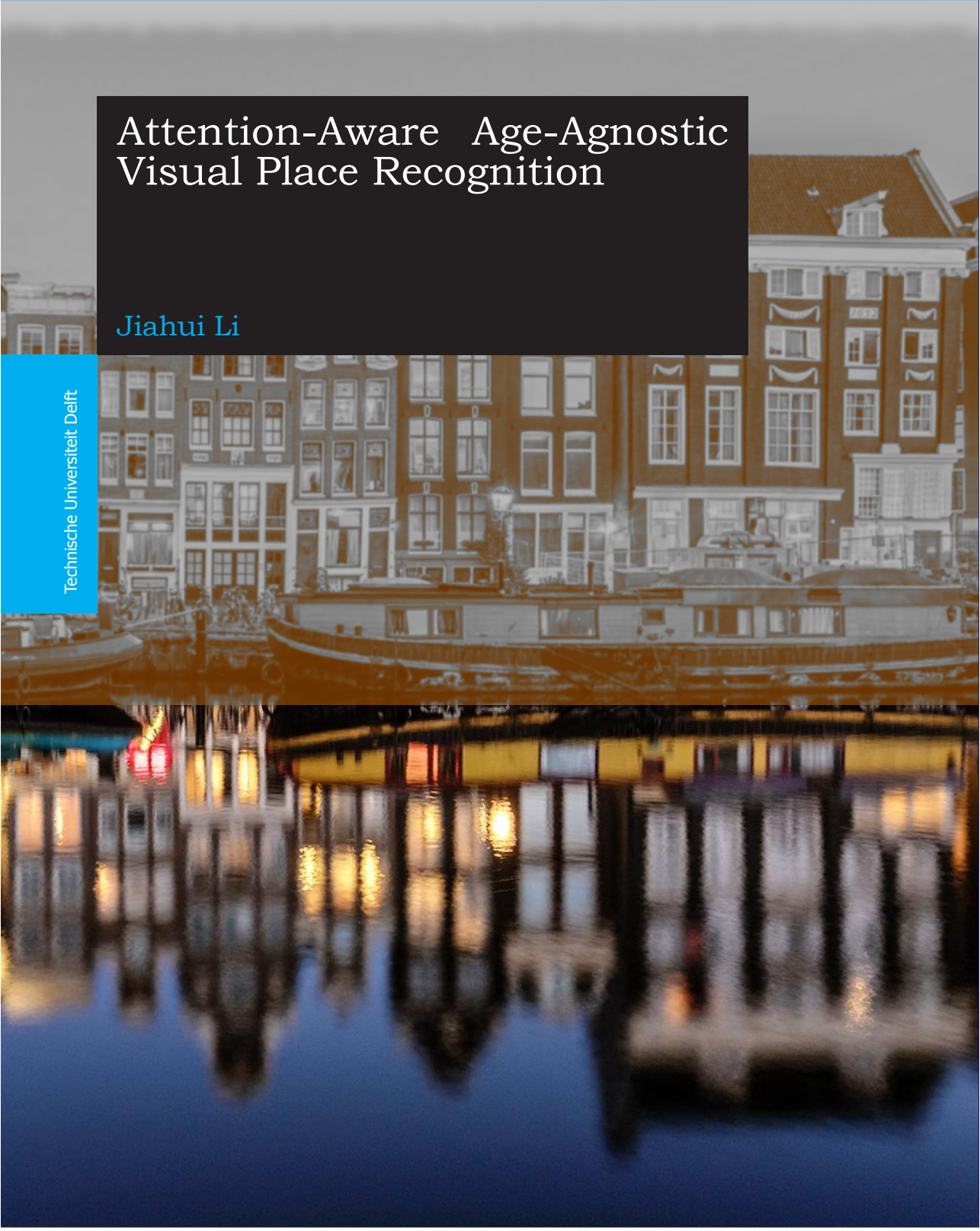
Attention-Aware Age-Agnostic Visual Place Recognition

Jiahui Li

Technische Universiteit Delft



TU Delft Delft University of Technology

Challenge the future

. Attention-Aware Age-Agnostic Visual Place Recognition

by

Jiahui Li

to obtain the degree of Master of Science
at the Delft University of Technology,
to be defended publicly on Wednesday August 28, 2019 at 13:00

Student number:	4734769	
Project duration:	November 2018 - August 2019	
Thesis committee:	Prof. dr. M. J. T. Reinders,	TU Delft, chair
	Dr. J. C. van Gemert,	TU Delft, supervisor
	Dr. L. Nan,	TU Delft, committee member

An electronic version of this dissertation is available at
http://repository.tudelft.nl/.

TU Delft Delft University of Technology

Preface

The thesis report documents the results of my master's thesis project. First we will show the main content, the scientific paper, including motivation, related work, methodology, experiments and results of this research project. The latter chapters introduce some related knowledge needed to understand my thesis work.

Throughout the thesis work I have received a great deal of support and guidance. I would like to first thank my supervisor Dr. J. C. van Gemert, whose formulate the research topic and steered me in the right the direction during my research through each stage of the process.

I would like to acknowledge my daily supervisors, Dr. S. Khademi and Z. Wang for their wonderful help. Dr. S. Khademi first helped me with formulating the basic methodology for the research problem, and was always keen to know what I was doing and how I was proceeding. Z. Wang always supported me in verifying my own ideas and provided valuable insight. The scientific paper would not have been completed without them..

I would also like to thank my family and friends, whose encouragement and support have made my accomplishments possible.

Jiahui Li
Delft, August 2019

Contents

1

Scientific Paper

Attention-Aware Age-Agnostic Visual Place Recognition

Jiahui Li

Delft University of Technology
Delft, The Netherlands
jiahuili67210@gmail.com

Abstract

A cross-domain visual place recognition (VPR) task is proposed in this work, i.e., matching images of the same architectures depicted in different domains. VPR is commonly treated as an image retrieval task, where a query image from an unknown location is matched with relevant instances from geo-tagged gallery database. Different from conventional VPR settings where the query images and gallery images come from the same domain, we propose a more common but challenging setup where the query images are collected under a new unseen condition. The two domains involved in this work are contemporary street view images of Amsterdam from the Mapillary dataset (source domain) and historical images of the same city from Beeldbank dataset (target domain). We tailored an age-invariant feature learning CNN that can focus on domain invariant objects and learn to match images based on a weakly supervised ranking loss. We propose an attention aggregation module that is robust to domain discrepancy between the train and the test data. Further, a multi-kernel maximum mean discrepancy (MK-MMD) domain adaptation loss is adopted to improve the cross-domain ranking performance. Both attention and adaptation modules are unsupervised while the ranking loss uses weak supervision. Visual inspection shows that the attention module focuses on built forms while the dramatically changing environment are less weighed. Our proposed CNN achieves state of the art results (99% *accuracy) on the single-domain VPR task and* 20% *accuracy at its best on the cross-domain VPR task, revealing the difficulty of age-invariant VPR.*

1. Introduction

Recently, there has been interest among the computer vision researchers to solve the visual place recognition (VPR) task in the form of image retrieval [3, 12, 20, 24, 27, 41, 48]. In [38], the discriminative visual cues learned for visual place classification task are investigated. Interestingly, CNN filters learn human-like discriminative visual cues to recognize a place, including built forms, signs or vegetation. Among these discriminative attributes, buildings are the most robust that remain, more or less, invariant during the changes in day and night lighting, different seasons and even years. However, CNNs are still influenced by irrelevant objects like roads and the sky. In this work, we introduce a CNN model with attention aggregation module to focus on domain invariant features, i.e. buildings, for the cross-domain VPR task. We will demonstrate that our work can be further combined with multi-kernel Maximum Mean Discrepancy (MK-MMD) loss to obtain better domain adaptation results. The images from the two domains with a large time lag are depicted in Fig.1, being historical images (queries) and current street view images (gallery) of Amsterdam.

The VPR task is commonly formulated as content based image retrieval (CBIR), i.e., sorting the geo-tagged gallery images by their distances to the unknown query image. The query is then labeled based on its best matching image in the gallery. Deep image representation learning is currently state of the art for almost all CBIR settings. Among the deep feature learning methods, distance learning CNNs are the most popular ones [13, 19]. Nevertheless, supervised deep distance learning requires similar and dissimilar image pairs for training. In this work, image pair labels are not available and we only have access to geo-tagged images from the *Mapillary* street view imagery and thus a weakly supervised deep feature learning is used, similar to the work of NetVLAD[3].

Different from [3], our queries are historical images which are not geo-tagged and exhibit a domain discrepancy between training data and test data. Age-agnostic place recognition that is addressed in this paper is a more challenging problem firstly due to the lack of image pair labels for training, secondly due to the domain shift between the

Figure 1: Correctly retrieved images with our proposed method. The top row illustrate the general place recognition in the same domain: both the query (left) and gallery image (right) are from the same dataset. The bottom row shows the cross-domain place recognition task where the query (left) is from the *Beeldbank* dataset and the galley image (right) is from the Mapillary dataset.

gallery and query images caused by the change of scenery over a large time gap and thirdly due to the outliers in target domain. Different technologies of photography, equipment and processes used in the production of photos in the past also contribute to this domain shift. Fig1 shows the general and the age-agnostic place recognition task.

We are inspired by [31], which introduces an attention module into NetVLAD for the classification task to address the unequal importance of local features in VLAD feature aggregation layer. In our work, a new attention aggregation technique is proposed to weigh both global VLAD descriptors and local descriptors. A domain adaptation loss based on MK-MMD is additionally introduced to achieve better cross-domain performance. Note that both the attention and the domain adaption modules are unsupervised and thus no labels are required.

Our attention-aware architecture is depicted in Fig.2 which consists of three modules and a shared convolutional neural network for feature extraction (AlexNet cropped before $conv5$). The attention module is a single convolutional layer followed by softplus activation function, transforming the feature map to a heatmap. This heatmap contains attention scores for the deep features. The VLAD module aggregates deep features in the attention-aware scheme by assigning attention scores to both local and global descriptors. The unsupervised domain adaptation module is additionally used to learn domain-invariant features. Our oblation studies show that both modules are important to reach state of the art results. Our speculation is that MK-MMD loss aligns the photo styles while attention module focuses on domain invariant contents.

Our contributions are summarized as:

- To the best of our knowledge, this is the first large scale (40k) image database for age-invariant visual place recognition task. We manually annotated 104 matching pairs between historical images and current street view images only for evaluation purpose.
- A new attention aggregation scheme is proposed to combine both the local and global image descriptors (Section 3.2).
- We combined the MK-MMD domain adaptation loss with the ranking loss to learn domain-invariant features for cross-domain VPR task. (Section 3.3).

We tested our proposed model on conventional VPR task and our experiments show the state of the art results on Mapillary dataset compared to other competitors. Detailed results and ablation studies will be presented in Section 4.5. The comparison of single-domain and cross-domain results reveals the difficulty of age-agnostic place recognition task.

2. Related Work

The performance of VPR as an image retrieval problem depends on the ranking accuracy w.r.t. a similarity metric. The query location is suggested based on the top M similar images (annotated with geo-tags). To extract good features for indexing, traditional works focus on handcrafted features such as SIFT[28]) and SURF[8]. Some other efficient methods are based on the aggregation of local gradient-based descriptors like Fisher Vectors [34] and VLAD[21]. [24] is a SURF based model which improves the performance by detecting and removing 'confusing objects'. [41] uses SIFT to detect the repetitive patterns in the image which is representative for buildings. [30] focuses on matching images that have large view point changes by generating artificial views of a scene for the training process.

Recent works suggest that a CNN trained on a large scale dataset as a feature extractor outperforms hand crafted features on various tasks [11, 12, 32, 36]. In turn, [6] shows that features in the early layers of a CNN trained for image classification can be effectively used as visual descriptors for image retrieval. LIFT [47] is a learning pipeline for feature extraction which introduces an end-to-end unified network for detection, orientation estimation, and feature description. [40] proposes a global image representation by the regional maximum activation of convolutional layers (R-MAC) well-suited for place recognition. [12] proposes

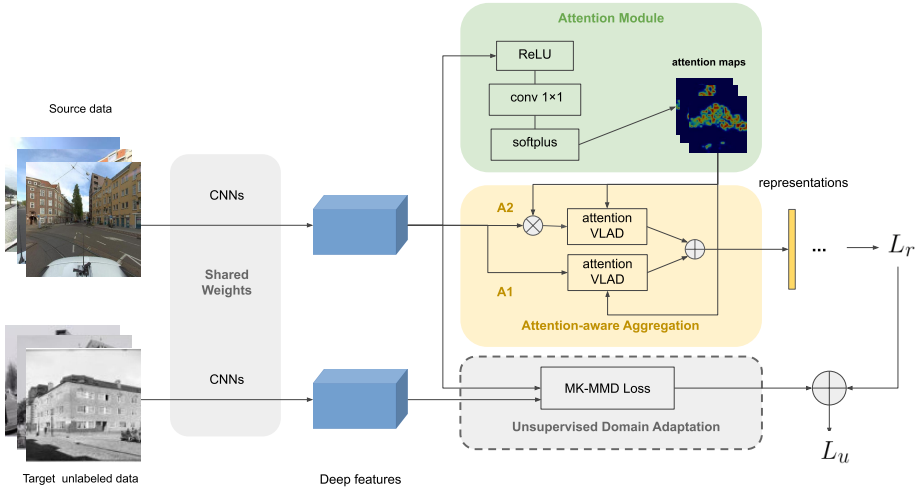


Figure 2: Our proposed CNN model include three modules: an attention module, am attention-aware VLAD module and a domain adaptation module. The attention-aware VLAD module uses the attention scores to weigh both the deep features and the global descriptors with two streams, A_1 and A_2, which are explained in Section 3.2.

novel CNN-based features designed for place recognition by detecting salient regions and extracting regional representations as descriptors. NetVLAD [4] introduces a novel triplet ranking loss together with a VLAD aggregation layer that can learn powerful representations for the VPR task in an end-to end manner. A known disadvantage of NetVLAD lies in its global feature aggregation. [16] proposes a region proposal network to learn which regions should be pooled to form the final global descriptor. Similar to [4] , we use current geo-location tags for weakly supervised feature learning using triplet distance learning network. However, we do not have access to matched image pairs from the two domains for supervised training, i.e., matched historical and contemporary images. To address this domain mismatch between the test and train data, we need to promote domain-invariant feature learning.

We tailor an attention aggregation model that can boost the cross-domain performance for our specific task, age-agnostic urban scene matching. Attention model is broadly used in natural language processing [14, 43] and computer vision tasks [18, 26, 32, 37, 39, 48, 45]. [23] shows that attention model can also be adopted to benefit metric learning. [32] proposes an attention mechanism to select key points for matching. Attention model is considered to be effective for domain adaptation as well [22, 44]. Our attention model is implemented in an unsupervised way which means no ground truth score maps are available for training. The learning process of the attention module is guided by the image retrieval ranking loss.

Given two different domains, unsupervised deep domain adaptation schemes [25, 42] are mostly used to enhance the performance of CNNs on target domain by using labels only from the source domain. Among the vast amount of literature on deep domain adaption for classification tasks, the Maximum Mean Discrepancy (MMD) loss is introduced by [10] to minimize the domain discrepancy by projecting data into a kernel space. Later [17] proposed multi-kernel MMD (MK-MMD) which uses linear combination of multiple kernels. We adopt MK-MMD loss as an additional domain adaptation module for our attention aggregation model. Similarly, we feed untagged historical images of Amsterdam to the adaptation layer in an unsupervised manner.

3. Method

Our proposed model consists of three modules for feature extraction, namely a weakly supervised image retrieval module with a triplet ranking loss (Section 3.1), an attention aggregation module(Section 3.2) and an unsupervised domain adaptation module with MK-MMD loss (Section 3.3). MK-MMD loss constrains the feature maps after the last convolution layer ($conv\_5$). The final loss function for training, L_u, can be expressed as:

$$L_u = L_r + \alpha M(\mathcal{D}_s, \mathcal{D}_t) \quad (1)$$

where $M(\mathcal{D}_s, \mathcal{D}_t)$ is the MK-MMD loss term, $\mathcal{D}_s$ and $\mathcal{D}_t$ denote the source domain and target domain, L_r is the triplet ranking loss used in NetVLAD [3], α is the weight that trades off the image retrieval loss and the domain adaptation loss.

3.1. Image retrieval with weak supervision

We use NetVLAD [3] as our baseline model which tackles the weakly supervised image retrieval task with a triplet ranking loss. NetVLAD considers the generated $H \times W \times D$ feature maps as a set of $N(H \times W) \times D$ local descriptors where N is the number of local descriptors and D is the dimension. Latter, a soft clustering is used to store the residual information contained in the descriptors to form $K \times D$ final descriptors denoted as V where K is the number of cluster centers. $V(j,k)$ can be expressed as:

$$V(j,k) = \sum_{i=1}^{N} a_k(x_i)(x_i(j) - c_k(j)), \quad (2)$$

where $j \in \{1, \ldots, D\}$ is the j-th dimension of a descriptor $\{x_i\}$, $k \in \{1, \ldots, K\}$ is the k-th cluster center, and $a_k(x_i)$ is the soft assignment of the descriptor x_i to k-th cluster center c_k. In Eq.1, A weakly supervised triplet ranking loss L_r is used to govern the learning process of descriptors that ensures the Euclidean distance between the query image and the best potential positive images are smaller than the Euclidean distance between the query image and all the negative pairs (based on geo-tags).

$$L_r = \sum_j l(\min_i d_\theta^2(q, p_i^q) + m - d_\theta^2(q, n_j^q)), \quad (3)$$

where, q denotes the query image and p_i^q are potential positive images. $\min_i d_\theta^2(q, p_i^q)$ denotes the best matching pair with shortest distance d_θ. In turn, n_j^q are all negative image pairs and m is the distance margin to be maintained. The function l is the hinge loss which penalizes the pairs that violate the margin.

3.2. Attention module

The triplet network for image retrieval task produces feature maps with the dimension of $H \times W \times D$. The inserted attention module consists of a 1×1 convolutional layer with coefficients $\mathbf{w}_a \in \mathbb{R}^{D \times 1}$ and a softplus activation function. This convolutional layer will produce an attention score map H_a with spatial size $H \times W$, which could be interpreted as the weight $\{w_i\}$ for each descriptor $\{x_i\}$. [31] proposed an attention aware aggregation scheme A_1 as:

$$V(j,k)_{A_1} = \sum_{i=1}^{N} w_i a_k(x_i)(x_i(j) - c_k(j)), \quad (4)$$

where $w_i \in \mathbf{w}_a$. Note that the VLAD module first assigns the local descriptors $\{x_i\}$ to K cluster centers $\{c_k\}$, then computes the residuals of each descriptor $x_i - c_k$ to its cluster center and assigns the weight a_k of descriptor x_i to cluster c_k proportional to their proximity.

In Eq.4, the global descriptors (residuals) are weighed after clustering. However, the VLAD descriptor is very sensitive to cluster centers [5] since it defines the origin of coordinates system to a cluster. Under this circumstance, we propose to weigh the local descriptors according to attention scores before performing clustering. The soft-assignment term a_k is re-calculated based on the newly weighed descriptors. Our proposed aggregation scheme A_2 can be formulated as

$$V(j,k)_{A_2} = \sum_{i=1}^{N} w_i a_k(x_i w_i)(x_i(j) w_i - c_k(j)). \quad (5)$$

The difference between A_1 and A_2 is that A_1 assigns the attention scores after clustering the descriptors to multiple centers so the attention scores are only used to weigh the residuals but A_2 first uses the attention scores to filter out uninteresting regions in the individual local descriptors and then performs the same step as A_1. Considering that the reweighing of individual descriptors may remove information that are useful for global descriptor generation, we aggregate the two attention schemes linearly:

$$V(j,k)_{our} = V(j,k)_{A_1} + V(j,k)_{A_2}. \quad (6)$$

3.3. Domain adaptation module

We use MK-MMD loss [17] with five Gaussian kernels of different bandwidths for unsupervised domain adaptation. The loss minimizes the distance between the expectation of the kernel mappings $\phi(.)$ of the descriptors in the source domain x_i^s and the target domain x_i^t.

$$M(\mathcal{D}_s, \mathcal{D}_t) = \sum_{i}^{N} ||\mathbb{E}(\phi(x_i^s)) - \mathbb{E}(\phi(x_i^t))||_2. \quad (7)$$

The MK-MMD loss guides the CNN to learn a latent space where the two domains are not distinctive, i.e., the gap between the statistical means of these two domains are closed in the reproducing kernel Hilbert space (RKHS).

4. Experiment

4.1. Dataset

We construct a cross-domain dataset with two sources of data to evaluate our proposed method, namely the street view panorama images of Amsterdam city from the *Mapillary* dataset[2] and the *Beeldbank* dataset[1] containing historical images from Amsterdam city archives.

Dataset		Gallery	Query
Source	*Mapillary40k*-train	20,884	2,320
	single-domain-test	18,980 (**M**)	2,108 (**M**)
Target	*Beeldbank*-train	29,726	-
	cross-domain-test	2,469 (**M**)	104 (**B**)

Table 1: *Mapillary40k→Beeldbank dataset*, the source domain is *Mapillary40k* and the target domain is *Beeldbank* denoted by **M** and **B**, respectively. *Beeldbank-train* is only used for unsupervised domain adaptation. cross-domain VPR requires matching query images from *Beeldbank* to gallery images from *Mapillary40k*.

Mapillary40k is a subset of *Mapillary250k* dataset collected from the source domain. The source domain contains panoramic images with high resolution collected from the *Mapillary*, Amsterdam area. Each image is annotated with a geotag. The cylindrical panorama is converted to 6 cubmaps (all share the same geotag): 'top', 'down', 'left', 'right', 'front' and 'back' textures with 512 $\times$ 512 resolution. The 'top' and 'down' textures are discarded since they usually contain sky and the vehicle that carries the camera. 40k gallery images and 4k query images are collected in total which are then divided into two roughly equal parts for training and testing when tested for single-domain VPR task, each containing around 20k gallery images and 2k queries. The two sub-datasets are geographically disjoint.

Beeldbank, the target domain, contains historical images of Amsterdam with low resolution and random size (height and weight are around 100 pixels). This dataset not only depicts Amsterdam street view in the past but also contains outliers including people, sketches and indoor scenes.

Mapillary40k - Beeldbank dataset is introduced in this work for the cross-domain VPR task (Table 1). The cross-domain test set contains 104 labeled queries from the target domain and 2,469 gallery images from the source domain. In the cross-domain test set, each target query has around 10 corresponding matched images in source domain. 30k unlabeled *Beeldbank* images are used during training for domain adaptation.

4.2. Single-domain and Cross-domain VPR tasks

Single-domain VPR task ($S \rightarrow S$) In the single-domain VPR setup, we train the network with only weakly labeled source domain images. The network is tested on the test set of the same domain. *Mapillary40k* is used for this single-domain VPR experiment as shown in Tab.1. This is the common setting for VPR task as there is no domain mismatch between train and test data. We use single-domain VPR as a pilot experiment to evaluate the performance of the proposed model on conventional VPR task.

Cross-domain VPR task ($S \rightarrow T$) The domain discrepancy between train and test data makes the cross-domain VPR task more challenging. This is the core of our experiments in this work which aims at labeling the images from beeldbank dataset with correct geo-location. We train the MK-MMD layer with weakly labeled source data and unlabeled target data for the cross-domain VPR task. Labeled data with matching pairs from beeldbank and Mapilary dataset is only used for evaluation of the model. We use queries from the target domain to retrieve relevant gallery image(s) collected from the source domain.

We made the hypothesis that our attention module itself can improve the cross-domain VPR task to some extent without the MK-MMD loss compared to vanilla NetVLAD. An experiment was carried out to examine the function of the attention module later in Section 3.2. Further ablation study of the attention module and the MK-MMD loss will be presented in section 4.5.1 and 4.5.2.

Baseline work We compare our attention-aware framework with 'off-the-shelf' CNNs for both single-domain VPR and cross-domain VPR tasks. The baseline work used AlexNet pretrained on ImageNet cropped before $conv5$ as feature extractor. Features are then sub-sampled by either max pooling (f_{max}), average pooling (f_{avg}), vanilla VLAD pooling without attention(f_{VLAD}) and VLAD with attention-aware A_1 method ($f_{A1-VLAD}$) [31].

4.3. Evaluation metrics

We follow the standard place recognition evaluation metric in [3] where the query image is considered as correctly matched if at least one of the retrieved top N images is located within 25 meters away from the ground truth query location. The Recall@N evaluates the percentage of correctly localized queries at different N matching levels. For cross-domain place recognition, since the *Beeldbank* dataset contains labeled positive pairs, the Recall@N will be directly calculated using these labels.

4.4. Implementation Details

The attention module starts with a ReLU activation, followed by a 1×1 convolutional layer and softplus activation to produce attention scores. In the VLAD layer, the number of cluster centers used is $K = 64$. *Mapillary* images were cropped with a random proportion to the original size between (0.3 1.0) for data augmentation before training.

We froze the layers before $conv4$ and fine-tuned the weights of all the other layers afterwards with the optimizer ADAM. We used the following hyperparameters: learning rate $lr = 1e\text{-}5$, batch size = 2 tuples (each tuple contains 24 images, including query, positive and negative pairs), epochs = 25. The hard negatives mining uses the same technique as NetVLAD[3]: it first caches all the training queries and gallery images for a time and then randomly selects 1000 negatives (image away from 25 meters). It keeps

Figure 3: Correctly retrieved top1 image from our framework trained with unsupervised domain adaptation, (top) queries are from the *Beeldbank* dataset, (bottom) retrieved images are from the *Mapillary* dataset. Our model can retrieve images not only depicting a similarly scene (a.), but also images from a different perspective (c.) and images captured further away from the query (b., d.). (b.) is correctly retrieved by matching the features of the building like the window and the unique shape of the door on the right side.

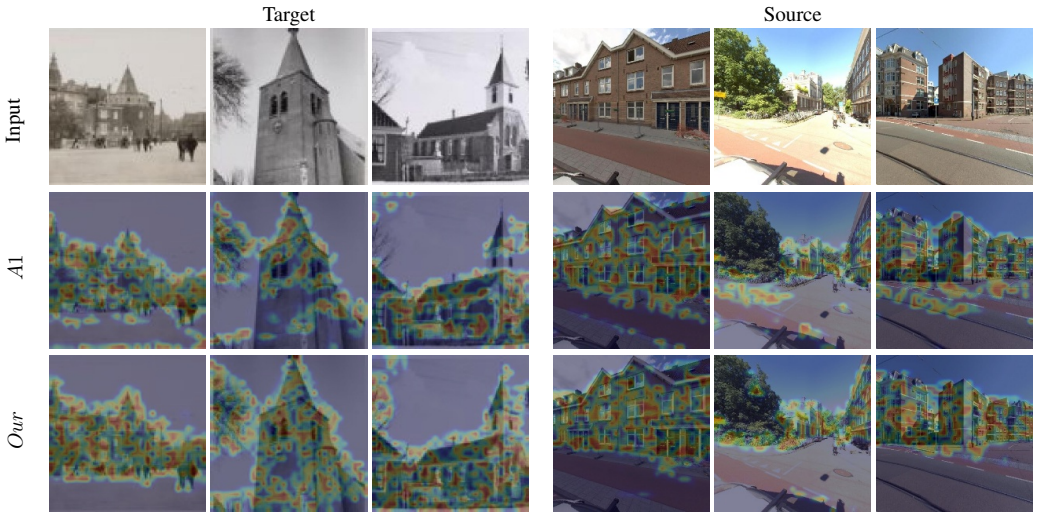


Figure 4: Visualization of attention score maps for source and target images. The top row shows the input images. The middle row is the heatmaps obtained by using [31], defined in Eq.4. The bottom row presents the results from our proposed method defined in Eq.6. It shows that our proposed attention module can generate accurate attention score maps with higher density on domain invariant objects for both source images and target images.

the top 10 hardest negatives from the cached gallery image features. The cache is updated every 1000 training queries.

We center cropped and reshaped all target *Beeldbank* images to 512×512 pixels in the cross-domain VPR experiment. The MK-MMD loss is calculated after $conv5$. The weight α in Eq.1 is 0.99. The margin m in Eq.3 is set as 0.1.

4.5. Results

This section presents the results of the experiments with a detailed ablation study for the attention module (Section 4.5.1) and the domain adaptation module (Section 4.5.2) separately on both single-domain and cross-domain VPR tasks. Visual inspection of retrieval results and attention heatmaps are shown in Fig.3 and Fig.4.

4.5.1 Attention module

To evaluate the performance of our attention aggregation module on both single and cross-domain VPR tasks, we first trained the model on the source domain (*Mapillary40k*) and directly tested it on the source test set and the target test set without MK-MMD loss. Tab.2 shows the retrieval results where our attention aggregation method consistently outperforms the model without attention on both $S \rightarrow T$ and $S \rightarrow S$ tasks. A possible explanation could be that the VLAD descriptors are easily affected by the irrelevant objects. By not focusing on representative details that describe unique features of each building, it may retrieve an image that has a similar road or sky etc.

To inspect whether our attention module can produce reasonable attention scores for each descriptor, we visualize the attention maps of different attention-aware schemes in Fig.4. Our attention aggregation method generates heatmaps with higher densities on representative features and better robustness against irrelevant objects. Most attention is assigned to the architectures and less attention is assigned to non representative regions such as road and sky as expected. Note that in Tab.2, the performance of $f_{A1\text{-}VLAD}$ is worse than f_{VLAD} and $f_{our\text{-}VLAD}$ achieves the best results. We conclude that an insufficient attention map will deteriorate the performance.

4.5.2 Domain adaptation module

The additional domain adaptation loss (MK-MMD) is added to our model and all baseline works in this section. The MK-MMD loss is adopted to further minimize the domain discrepancy in this experiment. We applied it on the vanilla NetVLAD (f_{VLAD}-DA), A_1 attention model ($f_{A1\text{-}VLAD\text{-}DA}$) and our attention aggregation model ($f_{our-VLAD}$-DA). The performance of different models with and without MK-MMD loss are examined on both source and target test test. The results are visualized at different recall rates in Fig.5.

When trained with the MK-MMD loss for the $S \rightarrow T$ cross-domain VPR task, both f_{VLAD}-DA and $f_{our\text{-}VLAD}$-DA benefit from domain adaptation, while no significant improvement of $f_{A1\text{-}VLAD\text{-}DA}$ is observed. Detailed results are presented in Tab.3.

In addition, we also examined the performance of the model trained for the cross-domain VPR task on the source domain due to the reason that extra data from the target doamin may also help with retrieval in source domain if the model is robust to the outliers in the *Beeldbank* dataset. f_{VLAD}-DA does not show much power in the original source domain compared to f_{VLAD}. The retrieval accuracy of $f_{A1\text{-}VLAD\text{-}DA}$ in the source domain decreases after domain adaptation. Our proposed model gets better retrieval result even on the source domain as shown in Fig.5. This experiment proves that the domain specific features and outliers are reduced while more domain invariant features are captured by our proposed attention aggregation model which further facilitates the domain adaptation procedure.

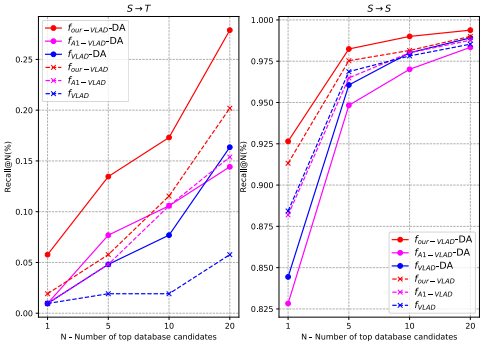


Figure 5: Comparison of the models trained with or without the MK-MMD loss on both single-domain ($S \rightarrow S$) and cross-domain ($S \rightarrow T$) tasks. DA denotes that the MK-MMD loss is added during training

Overall, we show that our attention aggregation model can achieve more accurate retrieval results on both single-domain $S \rightarrow S$ and cross-domain $S \rightarrow T$ VPR tasks even without domain adaptation and it can further facilitate unsupervised domain adaptation to achieve better performance on both source and target test sets.

5. Discussion

Usually we assume that the training data and test data are sampled from an identical distribution which is violated in our cross-domain setting. We designed an attention-aware

	$S \rightarrow T$				$S \rightarrow S$			
	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
$f_{max}{}^{+}$	0.0096	**0.0577**	0.0769	0.1058	0.6347	0.8226	0.8800	0.9203
$f_{avg}{}^{+}$	0.0000	0.0096	0.0481	0.0769	0.7884	0.9284	0.9535	0.9730
f_{max}	0.0000	0.0000	0.0577	0.1250	0.7410	0.9108	0.9431	0.9639
f_{avg}	0.0096	0.0192	0.0481	0.0577	0.7984	0.9269	0.9564	0.9725
f_{VLAD}	0.0096	0.0192	0.0192	0.0577	0.8843	0.9687	0.9782	0.9853
$f_{A1\text{-}VLAD}$	0.0096	0.0481	0.1058	0.1538	0.8819	0.9649	0.9801	0.9877
$f_{our\text{-}VLAD}$	**0.0192**	**0.0577**	**0.1154**	**0.2019**	**0.9132**	**0.9753**	**0.9815**	**0.9900**

Table 2: The $^{+}$ denotes that the 'off-the shelf' model is pretrained on ImageNet[15] for classification task. The others are trained on *Mapillary40k* for place recognition from scratch, and directly tested on the cross-domain dataset.

	with DA				without DA			
	R@1	R@5	R@10	R@20	R@1	R@5	R@10	R@20
f_{VLAD}	0.0096	0.0481	0.0769	0.1635	0.0096	0.0192	0.0192	0.0577
$f_{A1\text{-}VLAD}$	0.0096	0.0769	0.1058	0.1442	0.0096	0.0481	0.1058	0.1538
$f_{our\text{-}VLAD}$	**0.0577**	**0.1346**	**0.1731**	**0.2788**	**0.0192**	**0.0577**	**0.1154**	**0.2019**

Table 3: Comparison of different models' performance on the cross-domain $S \rightarrow T$ VPR task under two conditions: with or without domain adaptation using the MK-MMD loss. DA stands for domain adaptation using MK-MMD loss.

adaptive network to tackle the existing distribution shift. The results indicate that both the attention and adaptation modules contribute to the accurate retrieval of visual information. We speculate that the attention module mainly helps with focusing on domain invariant objects and the domain adaptation module aligns the depiction styles between the two different domains. Our dual experiments on both conventional and cross-domain VPR tasks admit the difficulty of learning age-invariant features when there is no cross-domain pairing labels available for directly training CNNs.

Besides the large domain shift, our *Beeldbank* target dataset contains various classes of images like people, indoor scenes, sketches and ground plans of buildings. These outliers are not contained in source dataset *Mapillary40k* rendering the task more difficult. Domain adaptation with more classes or outliers in the target domain compared to the source domain can be considered as open-set domain adaptation problem [7, 9, 33, 35]. Some other works refer to this as outlier detection problem [46, 29]. We speculate that the attention module can filter out the outliers by weighing them less with the heatmaps.

6. Conclusion

We proposed a specially-designed CNN for automatic annotation of historical images with their location. This is helpful specifically for museum curators and historians to retrieve the location information of a historical urban scene or architecture. This task is more challenging than single-domain (conventional) location retrieval due to the domain discrepancy caused by the large time lag between depicted scenes. A cross-domain dataset is collected accordingly with *Mapillary40k* used as source domain and *Beedldbank*, as target domain. To tackle this challenge, an attention aggregation module with a domain adaptation layer is designed, the performance of which is demonstrated by detailed experiments and ablation studies. Our attention aggregation model achieves state of the art results on both single and cross-domain VPR tasks by focusing more on domain invariant objects. It can be further combined with an extra domain adaptation module using the MK-MMD loss to achieve higher retrieval accuracy not only on the target domain but also on the source domain. Moreover, we believe our methods can achieve promising results on open-set domain adaptation tasks where unseen classes or outliers are not involved during training.

References

[1] https://beeldbank.amsterdam.nl/beeldbank. 4

[2] https://www.mapillary.com. 4

[3] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016. 1, 4, 5

[4] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):14371451, Jun 2018. 3

[5] R. Arandjelovic and A. Zisserman. All about vlad. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1578–1585, 2013. 4

[6] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky. Neural codes for image retrieval. *Lecture Notes in Computer Science*, page 584599, 2014. 2

[7] M. Baktashmotlagh, M. Faraki, and T. Drummond. Open-set domain adaptation. 2019. 8

[8] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006. 2

[9] A. Bendale and T. E. Boult. Towards open set deep networks. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. 8

[10] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. 3

[11] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford. Deep learning features at scale for visual place recognition. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3223–3230. IEEE, 2017. 2

[12] Z. Chen, F. Maffra, I. Sa, and M. Chli. Only look once, mining distinctive landmarks from convnet for visual place recognition. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9–16. IEEE, 2017. 1, 2

[13] S. Chopra, R. Hadsell, Y. LeCun, et al. Learning a similarity metric discriminatively, with application to face verification. In *CVPR (1)*, pages 539–546, 2005. 1

[14] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015. 3

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 8

[16] A. Gordo, J. Almazán, J. Revaud, and D. Larlus. Deep image retrieval: Learning global representations for image search. In *European Conference on Computer Vision*, pages 241–257. Springer, 2016. 3

[17] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur. Optimal kernel choice for large-scale two-sample tests. In *Advances in neural information processing systems*, pages 1205–1213, 2012. 3, 4

[18] Y. Gu, C. Li, and J. Xie. Attention-aware generalized mean pooling for image retrieval, 2018. 3

[19] E. Hoffer and N. Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015. 1

[20] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum. Panorama to panorama matching for location recognition. *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval - ICMR 17*, 2017. 1

[21] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010. 2

[22] G. Kang, L. Zheng, Y. Yan, and Y. Yang. Deep adversarial attention alignment for unsupervised domain adaptation: The benefit of target expectation maximization. *Lecture Notes in Computer Science*, page 420436, 2018. 3

[23] W. Kim, B. Goyal, K. Chawla, J. Lee, and K. Kwon. Attention-based ensemble for deep metric learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 736–751, 2018. 3

[24] J. Knopp, J. Sivic, and T. Pajdla. Avoiding confusing features in place recognition. In *European Conference on Computer Vision*, pages 748–761. Springer. 1, 2

[25] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015. 3

[26] X. Long, C. Gan, G. de Melo, J. Wu, X. Liu, and S. Wen. Attention clusters: Purely attention based local feature integration for video classification. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018. 3

[27] M. Lopez-Antequera, R. Gomez-Ojeda, N. Petkov, and J. Gonzalez-Jimenez. Appearance-invariant place recognition by discriminatively training a convolutional neural network. *Pattern Recognition Letters*, 92:89–95, 2017. 1

[28] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, Nov 2004. 2

[29] L. Luo, L. Chen, S. Hu, et al. Discriminative label consistent domain adaptation. *arXiv preprint arXiv:1802.08077*, 2018. 8

[30] A. L. Majdik, Y. Albers-Schoenberg, and D. Scaramuzza. Mav urban localization from google street view data. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 3979–3986. IEEE. 2

[31] K. K. Nakka and M. Salzmann. Deep attentional structured representation learning for visual recognition. *arXiv preprint arXiv:1805.05389*, 2018. 2, 4, 5, 6

[32] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3456–3465, 2017. 2, 3

[33] P. Oza and V. M. Patel. Deep cnn-based multi-task learning for open-set recognition, 2019. 8

[34] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer, 2010. 2

[35] K. Saito, S. Yamamoto, Y. Ushiku, and T. Harada. Open set domain adaptation by backpropagation. *Lecture Notes in Computer Science*, page 156171, 2018. 8

[36] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 2

[37] Z. Seymour, K. Sikka, H.-P. Chiu, S. Samarasekera, and R. Kumar. Semantically-aware attentive neural embeddings for image-based visual localization, 2018. 3

[38] X. Shi, S. Khademi, and J. van Gemert. Deep visual city recognition visualization, 2019. 1

[39] J. Song, Q. Yu, Y.-Z. Song, T. Xiang, and T. M. Hospedales. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5551–5560, 2017. 3

[40] G. Tolias, R. Sicre, and H. Jgou. Particular object retrieval with integral max-pooling of cnn activations, 2015. 2

[41] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi. Visual place recognition with repetitive structures. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 883–890. 1, 2

[42] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014. 3

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[44] X. Wang, L. Li, W. Ye, M. Long, and J. Wang. Transferable attention for domain adaptation. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019. 3

[45] T. Xiao, Y. Xu, K. Yang, J. Zhang, Y. Peng, and Z. Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015. 3

[46] M. Yamaguchi, Y. Koizumi, and N. Harada. Adaflow: Domain-adaptive density estimator with application to anomaly detection and unpaired cross-domain translation. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3647–3651. IEEE, 2019. 8

[47] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. Lift: Learned invariant feature transform. *Lecture Notes in Computer Science*, page 467483, 2016. 2

[48] Y. Zhu, J. Wang, L. Xie, and L. Zheng. Attention-based pyramid aggregation network for visual place recognition. *2018 ACM Multimedia Conference on Multimedia Conference - MM 18*, 2018. 1, 3

2

Introduction

The ability to find the place depicted in an image has a wide range of exciting applications such as: (i) automatic registration of uploaded consumer photos with maps (ii) transferring landmark information to the query image, or (iii) accurate visual localization for robotics. Visual place recognition(VPR) task often cast as image retrieval problem[1–7], the standard procedure mainly consists of two steps: First, extracting and storing feature vectors of gallery images. Then, extracting the feature vector of the query and computing the similarity between the query and all the gallery images. Finally, ranking all the gallery images based on the similarities and get the top K similar images (annotated with geo-tag) suggesting the location of the query. Conventional place recognition methods mostly based on local hand crafted features such as SIFT[8]) and SURF[9], but recent works suggest that Convolutional Neural Networks(CNNs) trained on large scale dataset as feature extractor outperforms hand crafted features on various tasks. Motivated by theses achievements, CNN based place recognition has emerged[1, 2, 10–12] and reached the state-of-the-art performance.

The VPR in single or similar domain has been widely studied, however, a more challenging problem, cross domain place recognition, has been rarely studied. The cross domain aims to match images of the same architectures taken in a wide time range, e.g. historical images and current street view images.

2.1. Motivation

A cross-domain VPR task is proposed in this work. VPR is commonly treated as an image retrieval task, where a query image from an unknown location is matched with relevant instances from geo-tagged gallery database. Different from conventional VPR settings where the query images and gallery images come from the same domain, we propose a more common but challenging setup where the query images are collected under a new unseen condition. The two domains involved in this work are contemporary street view images of Amsterdam (source domain) and histori-

cal images of the same city (target domain). The cross domain place recognition problem is challenging due to lack of labels and large domain disparity caused by large time gap, change of environment, also different technology of photography, equipment, techniques, and processes used in the production of photos in the past. Fig.2.1 shows the example of the single domain and cross domain VPR.

Figure 2.1: The top row illustrate the general place recognition in the same domain: both the query and gallery images are from the same dataset,contemporary street view images of Amsterdam(source). The bottom row shows the cross-domain place recognition task where the query is the historical images of the same city(target) and the galley image is from source domain. The green boxes denote the matching images in the gallery.

2.2. Research objectives

The goal of this paper is to find the place depicted in a historical image(target domain) in a current street view image gallery(source domain) based on visual similarity. The main challenges includes, general image matching problem, changes of view point and illumination, another challenge is the domain shift between the source and target. To overcome these challenges, our contributions are:

1. A new attention aggregation scheme is proposed to CNN to combine both the local and global image descriptors hence reduce the effect of 'confusing objects'.
2. We combined the MK-MMD domain adaptation loss with the triplet ranking loss to learn domain-invariant features for cross-domain VPR task.

2.3. Outline

The rest of the thesis report is organized as follows: Chapter 3 gives an background introduction to general convolutional neural networks. Chapter 4 introduces the concept of domain adaptation. Chapter 5 explains the feature aggregation method(VLAD) for image retrieval. Chapter 6 introduces the attention mechanism used in visual tasks.

References

[1] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, *Netvlad: Cnn architecture for weakly supervised place recognition,* in *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition (2016) pp. 5297–5307.

[2] Z. Chen, F. Maffra, I. Sa, and M. Chli, *Only look once, mining distinctive landmarks from convnet for visual place recognition,* in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2017) pp. 9–16.

[3] A. Iscen, G. Tolias, Y. Avrithis, T. Furon, and O. Chum, *Panorama to panorama matching for location recognition,* Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval - ICMR '17 (2017), 10.1145/3078971.3079033.

[4] J. Knopp, J. Sivic, and T. Pajdla, *Avoiding confusing features in place recognition,* in *European Conference on Computer Vision* (Springer) pp. 748–761.

[5] M. Lopez-Antequera, R. Gomez-Ojeda, N. Petkov, and J. Gonzalez-Jimenez, *Appearance-invariant place recognition by discriminatively training a convolutional neural network,* Pattern Recognition Letters **92**, 89 (2017).

[6] A. Torii, J. Sivic, T. Pajdla, and M. Okutomi, *Visual place recognition with repetitive structures,* in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 883–890.

[7] Y. Zhu, J. Wang, L. Xie, and L. Zheng, *Attention-based pyramid aggregation network for visual place recognition,* 2018 ACM Multimedia Conference on Multimedia Conference - MM '18 (2018), 10.1145/3240508.3240525.

[8] D. G. Lowe, *Distinctive image features from scale-invariant keypoints,* International Journal of Computer Vision **60**, 91 (2004).

[9] H. Bay, T. Tuytelaars, and L. Van Gool, *Surf: Speeded up robust features,* in *European conference on computer vision* (Springer, 2006) pp. 404–417.

[10] Z. Chen, A. Jacobson, N. Sünderhauf, B. Upcroft, L. Liu, C. Shen, I. Reid, and M. Milford, *Deep learning features at scale for visual place recognition,* in *2017 IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2017) pp. 3223–3230.

[11] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, *Large-scale image retrieval with attentive deep local features,* in *Proceedings of the IEEE International Conference on Computer Vision* (2017) pp. 3456–3465.

[12] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, *From coarse to fine: Robust hierarchical localization at large scale,* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019) pp. 12716–12725.

3

Background on Deep Learning

3.1. Introduction to Deep Learning

Deep learning[1] also referred as Neural Network(NNs) is a subset of machine learning, has achieved great achievements in image, video and audio related fields. Below we will introduce some basic knowledge of NNs.

Neuron The neuron is the basic unit of neural network. To describe the neurons statistically, let $\mathbf{x} = \{x_1, \ldots, x_n\}^T$ be the input, y be the output, Fig. 3.1 shows a mathematical model of the neuron. The final output of the neuron is the sum of weighted input **wx** with bias b

$$u = \sum_{i=1}^{n} w_i x_i + b, \tag{3.1}$$

and then pass it through a nonlinear activation function f, so the final output is

$$y = f(u), \tag{3.2}$$

where the weights **w** and bias b are trainable parameters.

Neural Networks Neural Networks consist of a collection of neurons that are connected in an acyclic graph. Fig. 3.2 illustrates a simple neural networks consisting input layer, hidden layer and output layer(bias dose not show in this figure). Neurons in two neighbor layers are fully pairwise connected, but neurons within a

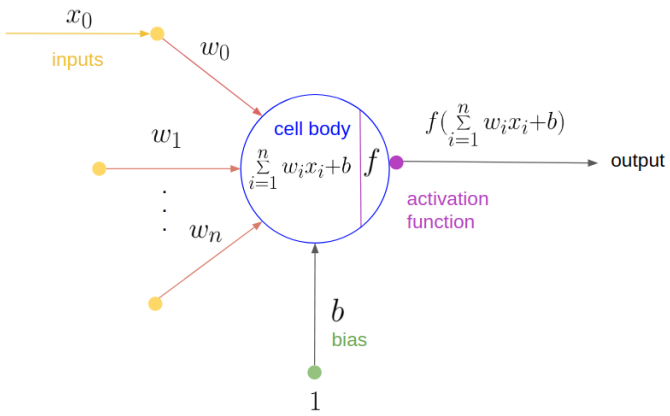


Figure 3.1: A mathematical model of a single neuron, consist of inputs $\{x_1, \ldots, x_n\}$, weights $\mathbf{w} = \{w_1, \ldots, w_i\}$, a bias b and an activation function f.

single layer share no connections. The output $\vec{y}$ is calculate as:

$$\vec{a} = W^{(1)}\vec{x} + b^{(1)} \tag{3.3}$$

$$\vec{y} = W^{(2)}\vec{a} + b^{(2)} \tag{3.4}$$

which can be easily implemented through matrix operations.

Activation function activation function is an nonlinear mapping from the weighted input to output. And it is usually but not necessary monotonic and differentiable, we need it to be this way so as to perform backpropagation optimization strategy, Fig. 3.3 shows some commonly used activation functions. The activation function introduces nonlinear properties to neural networks, and enable the network to learn and model other complicated, high dimensional types of data such as image, video, audio etc..

Training The weights and bias in neural networks are the trainable parameters, are usually randomly initialized and updated during training by minimizing the loss function. The loss is usually defined as the sum of the squared difference between the target values and the network output. This is a classic optimization problem which could be solved by gradient decent method, hence we can optimize all parameters based on the gradient. But when the networks gets 'deeper', due to the complexity of its structure and huge amount of parameters, it is impractical to calculate all gradients. Backpropagation[2] makes it possible to solve this problem. In this way we can propagate the total loss back into the neural network to know how much of the loss every neuron is responsible for, and subsequently update the weights in such a way that minimizes the loss.

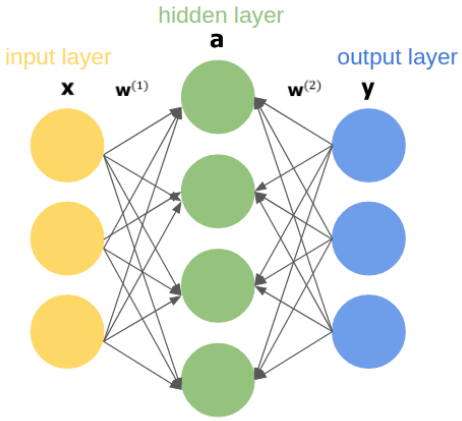


Figure 3.2: A simple neural network with input layer, one hidden layer and output layer

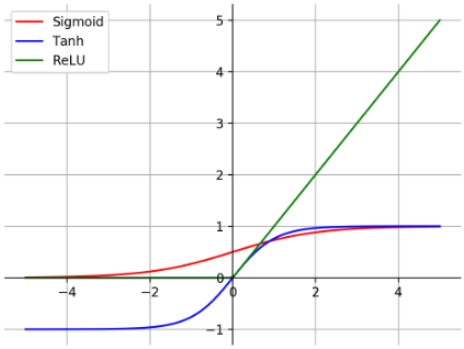


Figure 3.3: Some commonly used activation functions: Sigmoid, Tanh and Rectified Linear Unit(ReLU)

3.2. Convolutional Neural Network

Convolutional Neural Networks(ConvNets) is built upon neural networks, it assumes the inputs are images and builds an efficient architecture to extract certain visual features. The previous 'hidden layer' is replaced with three main types pf layers, including 'convolutional layer', 'pooling layer' and 'fully connected layer'(the same layer type as 'hidden layer'). Networks built by these types of layers are more efficient and the amount of parameters can be reduced.

Convolutional layer Given a high-dimensional input, such as image, it is impractical to connect all the neurons in the way as in Fig.3.2. The convolutional(CONV) layers consist of a series of filters with size $W \times H \times D$, these filters connect each neuron to only a local region of the input, the D is always the same with the D of input volume, Fig.3.4 is an example of the CONV layer operation. When sliding

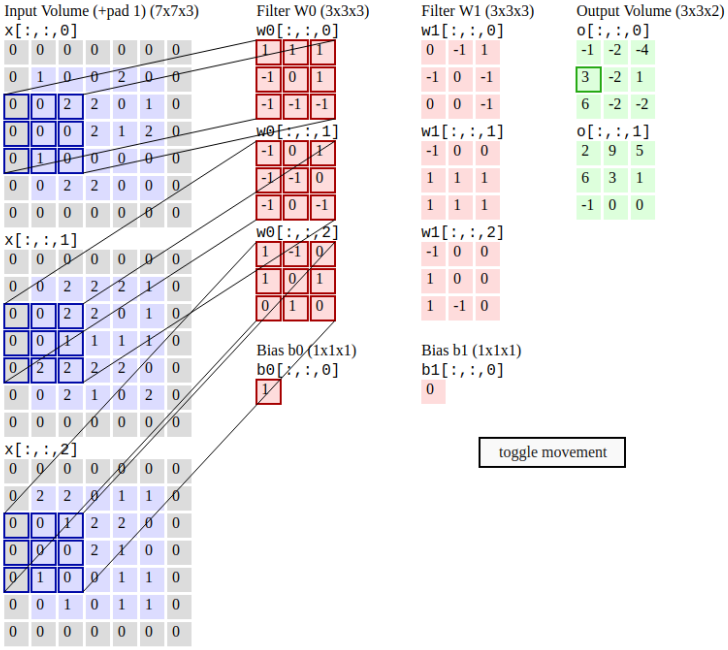


Figure 3.4: Example of CONV layer operation[3], the filters interact with regional input with dot products. One element of output(green) is produced with element-wise multiplication of the regional input(highlighted blue) and filter(red), then summing it up and adding the bias.

the input window, the regional input changes, while the parameters of the filter is the same, this is the parameter sharing scheme used in CONV layers to reduce the number of parameters. The basic idea behind it is that a useful feature extracted at a certain spatial position may also be useful to extract at another position.

Pooling layer is used to reduce the number of parameters also to prevent overfitting. It downsamples the input spatially at each depth slice with max operation, and the depth of output remains unchanged. Fig.3.5 illustrates the process of max pooling operation.

Fully-connected layer connects each neuron to all neurons in the previous layer, as seen in Fig.3.2. In AlexNet[4], CONV layers and pooling layers serve as feature extractor, and fully-connected layers is followed to combined these features together. Finally, an activation function will classify the output as different classes.

AlexNet is used as base network in our work. AlexNet won the championship in the 2012 ImageNet LSVRC competition, also popularized CNN in computer vision. Fig.3.6 illustrates the architecture of AlexNet, it consist of 5 convolutional

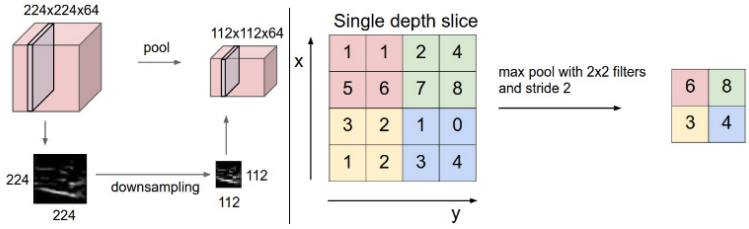


Figure 3.5: [3] The pooling layer downsamples the spatial size of the input independently in each depth slice of the input volume(left). A general pooling layer is pooled with filter size 2×2, stride 2, to preserve the maximum value in the 2×2 window.

layers, 3 max pooling layers and 3 fully connected layer.

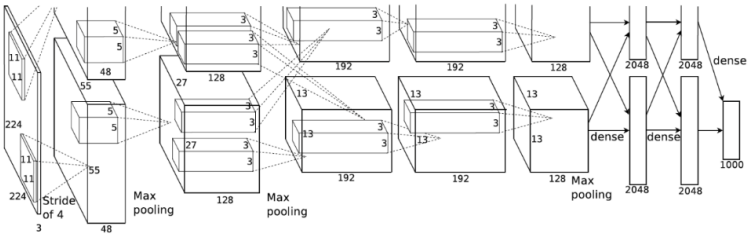


Figure 3.6: AlexNet architecture resided on 2 GPUs, the network consist of 5 convolutional layers, and 3 max pooling layers to downsample the feature map, and 3 fully connected layer at the end. The output is the 1000-class softmax encoding class information.

3.3. Triplet Network

CNN architecture has gained great achievements in classification tasks, however, the number of classes may be variant or very large, for example, we want to compare two unknown faces and say whether they are from the same person or not. To solve this, triplet network is a good method by learning deep face embeddings through distance comparisons, has been widely used for face recognition.[5]. Triplet network is inspired from Siamese Network[6], which contains three feed-forward networks with shared parameters, and three inputs: an anchor a, a positive from the same class as anchor p, a negative from different class n, like in Fig3.7. To bring the positive pair (a, p) together, the triplet loss is introduced, the learning objects are (i) for samples with the same labels, their distance in the embedding space should be close and form separable clusters, and (ii) otherwise should be far away. Formally, for the distance metric d in the embedding space, the loss $\mathcal{L}$ for an input tuple (a, p, n) is:

$$\mathcal{L} = \max(d(a,p) - d(a,n) + \text{margin}, 0). \tag{3.5}$$

We minimize the loss by forcing the distance of negative pair $d(a,n)$ to be greater by a margin than the distance of positive pair $d(a,p)$. The loss is zero when

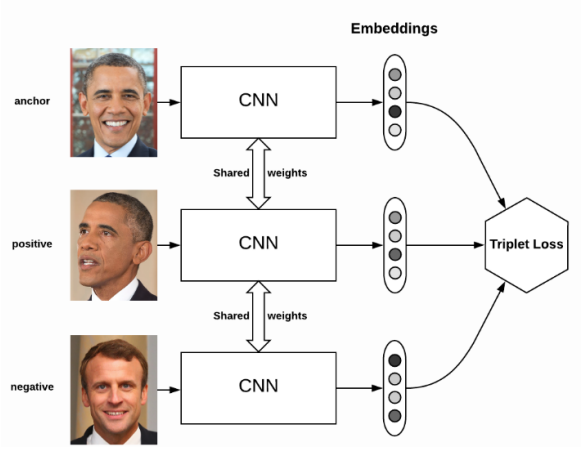


Figure 3.7: Triplet network architecture

the distance of positive pair is less than the distance of negative pair by a certain margin.

3.3.1. Triplet mining

Based on the definition of triplets, here are three categories:

1. easy triplet: $d(a,p) + m < d(a,n)$, the loss is zero.
2. hard triplet: $d(a,p) > d(a,n)$
3. semi-hard triplet: $d(a,p) + m > d(a,n) > d(a,p)$

The choice of triplet has great impact on the result. Paper[7] shows that selecting the hardest positive(biggest $d(a,p)$), and hardest negative(smallest $d(a,n)$), yields the best performance.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, *Deep learning,* nature **521**, 436 (2015).

[2] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, *Backpropagation applied to handwritten zip code recognition,* Neural computation **1**, 541 (1989).

[3] *http://cs231n.github.io/convolutional-networks/,* .

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks,* in *Advances in neural information processing systems* (2012) pp. 1097–1105.

[5] F. Schroff, D. Kalenichenko, and J. Philbin, *Facenet: A unified embedding for face recognition and clustering,* 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015), 10.1109/cvpr.2015.7298682.

[6] S. Chopra, R. Hadsell, Y. LeCun, *et al.*, *Learning a similarity metric discriminatively, with application to face verification,* in *CVPR (1)* (2005) pp. 539–546.

[7] A. Hermans, L. Beyer, and B. Leibe, *In defense of the triplet loss for person re-identification,* (2017), arXiv:1703.07737 [cs.CV] .

4

Domain Adaptation

This chapter aims to introduce basic definition of domain adaptation, and deep domain adaptation method that has been investigated in our research.

4.1. Introduction to Domain Adaptation

Machine learning aims to learn knowledge from data relying on the features and inference. The successful applications of many machine learning algorithms are built on the premise that the training and future data are in the same feature space and follow the same distribution. However, this assumption may not hold for real world problem. For example, when have an interested domain with insufficient training data and another domain with sufficient training data, but two domains are in different feature spaces or have different data distribution. In such cases, if knowledge learned form the latter domain could be flexibly adapt to the interested domain, it would greatly improve the performance in the interested domain without expensive data labeling efforts.

In recent years, domain adaptation as a subproblem of transfer learning, has emerged as a new learning framework to address the domain shift problem, the key of domain adaptation is learning the shared feature in both domains. Specifically, it applies the distribution adaptation on the source domain to the target domain by making use of the similarity of the data.

Formally, a domain D can be described by a feature space $\mathcal{X}$ and a marginal distribution $P(X)$, where $X = \{x_i, \dots, x_n\} \in \mathcal{X}$ is a set of samples from the same feature space $\mathcal{X}$. Given a domain $\mathcal{D} = \{\mathcal{X}, P(X)\}$, a task $\mathcal{T}$ is consist of label $Y = \{y_1, \dots, y_n \in \mathcal{Y}\}$ from space $\mathcal{Y}$ and an objective predictive function $f(\cdot)$, which from statistic prospective, can be viewed as a conditional probability distribution $P(Y|X)$. So we can learn the predictive function$f(\cdot)$ by making use of labeled data $\{x_i, y_i\}$, where $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$.

Given a labeled source domain $D_s = \{x_i, y_i\}_{i=1}^{n}$, and an unlabeled target domain $D_t = \{x_j\}_{j=n+1}^{n+m}$. Traditional machine learning assumes $D_s = D_t$ and $\mathcal{T}_s = \mathcal{T}_t$,

paper[1] summarized that the domain shift can be caused by domain divergence $D_s \neq D_t$, including distribution shift $P(X)$ or feature space difference $\mathcal{X}$ and task difference $\mathcal{T}_s \neq \mathcal{T}_t$, like conditional distribution $P(Y|X)$ shift or label space $\mathcal{Y}$ difference. In our case, we consider the first situation, domain divergence, so our goal is to use labeled source domain data $D_s = \{x_i, y_i\}_{i=1}^n$ and unlabeled target sample $D_s = \{x_i, y_i\}_{i=1}^n$ to learn an objective predictive function $f : x_t \rightarrow y_t$ to predict target domain label $y_t \in y_t$.

4.2. Deep Domain Adaptation

Compared with traditional domain adaptation methods, deep domain adaptation based methods boost the performance on many tasks[2, 3]. It directly learns from origin data, automatically extracts more representative features in an end-to-end manner.

Many deep approaches [3, 4] introduced the adaptation layer to minimize the domain disparity. The basic idea of deep adaptation network is to (i) find out which layer is adaptable, then (ii) choose a metric to perform adaptation within these layers. Compared to traditional domain adaptation, the deep one defines loss as:

$$l = l_c(D_s, Y_s) + \lambda l_A(D_s, D_t),$$

where $l_c(D_s, Y_s)$ is the task loss (e.g. classification loss) on labeled data (usually the source domain), $l_A(D_s, D_t)$ is the adaptation loss, and $\lambda > 0$ determines how strongly we would like to confuse the domains. With this loss, we can simultaneously minimize the classification loss and maximize the domain confusion, so that the domain invariant and discriminative representations can be learned, Figure 4.1 gives an explicit image interpretation.

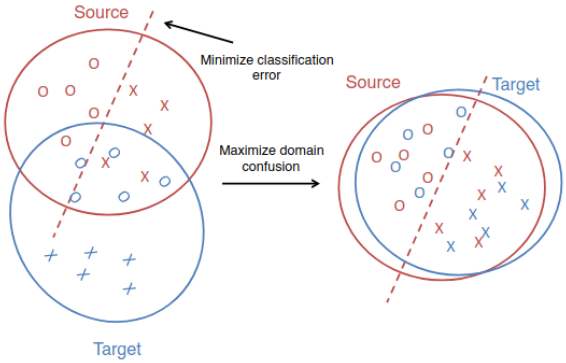


Figure 4.1: [4]. Classifier trained on source domain do not ensure good performance on target domain due to domain shift (left), by maximizing domain confusion, classifier can be well adapted to target domain (right).

One classic deep domain adaptation method[4] is DDC(Deep Domain Confusion). It uses the AlexNet[5] pretrained on ImageNet, Fig. 4.2 shows how the DDC

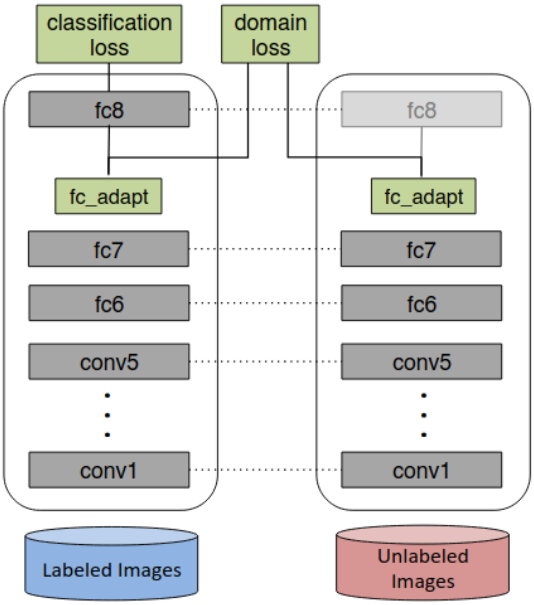


Figure 4.2: DDC architecture[4].

works In this architecture, there are a source and a target network which share the weights to ensure domain invariant features. Layers before $fc8$ all stay intact, serve as the feature extractor, and followed with an adaptation layer which evaluates the domain disparity with MMD (Maximum mean discrepancy) metric. MMD is defined by the distances between mean embeddings of the features. It maps the features to the reproducing kernel Hilbert space [6], denoted as $\mathcal{H}$.

$$MMD^2(X,Y) = \left\| \sum_{i=1}^{n1} \phi(x_i) - \sum_{j=1}^{n2} \phi(y_i) \right\|_{\mathcal{H}}^2,$$

where ϕ maps the feature to $\mathcal{H}$. so the MMD is the distance between the means of the two distributions in space $\mathcal{H}$.

Different from MMD based on single kernel transformation, multiple-kernel MMD[7](MK-MMD) assumes the optimum kernel can be described by the linear combination of multiple kernels and has been widely used in latter research[3], which introduces multiple adaptation layers with MK-MMD metric.

References

[1] S. J. Pan and Q. Yang, *A survey on transfer learning,* IEEE Transactions on knowledge and data engineering **22**, 1345 (2009).

[2] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, *Learning and transferring mid-level image representations using convolutional neural networks,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2014) pp. 1717–1724.

[3] M. Long, Y. Cao, J. Wang, and M. I. Jordan, *Learning transferable features with deep adaptation networks,* arXiv preprint arXiv:1502.02791 (2015).

[4] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, *Deep domain confusion: Maximizing for domain invariance,* (2014), arXiv:1412.3474 [cs.CV] .

[5] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, *Decaf: A deep convolutional activation feature for generic visual recognition,* in *International conference on machine learning* (2014) pp. 647–655.

[6] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, *Integrating structured biological data by kernel maximum mean discrepancy,* Bioinformatics **22**, e49 (2006).

[7] A. Gretton, D. Sejdinovic, H. Strathmann, S. Balakrishnan, M. Pontil, K. Fukumizu, and B. K. Sriperumbudur, *Optimal kernel choice for large-scale two-sample tests,* in *Advances in neural information processing systems* (2012) pp. 1205–1213.

5

VLAD: vector of locally aggregated descriptors

In this section, some image representations for image retrieval will be introduced. The way to embed the VLAD in the CNN architecture will be shown as well.

5.1. Image vector representation

The task large scale image search/retrieval has been traditional addressed with bag-of-visual-words (BOV)[1] which assigns the image descriptor (eg. SIFT descriptor) to its closest visual word in a 'visual vocabulary': a codebook $C = \{c_1, \dots, c_k\}$ of k visual word which learned with k-means, and this will yield a fixed size high-dimensional histogram representation.

Given the descriptors of an image $\{x_i\}(i = 1, \dots, N)$ of D dimension, the idea of VLAD[2] is to aggregate, for each cluster center $c_k(k = 1, \dots, K)$, the residuals $x - c_i$ of the descriptors. The size of the final VLAD representation V will be $D \times K$, so one element of V is computed as:

$$V(j,k) = \sum_{i=1}^{N} a_k(x_i)(x_i(j) - c_k(j)), \tag{5.1}$$

where j, k respectively denote the $j^{th}(j = 1, \dots, D)$ component of the descriptor and the index of the clusters. $x_i(j)$ and $c_k(j)$ are the j^{th} component of the i^{th} descriptor and k^{th} cluster. The $a_k(x_i)$ denotes the membership of x_i to the k^{th} cluster with hard assignment, that is, for every x_i whose closest center is c_k, the corresponding $a_k(j)$ is 1, otherwise is 0.

Inspired from VLAD, paper [3] proposed a new architecture named NetVLAD by plugging the VLAD layer into CNN architecture. The non differentiable part of VLAD comes from the hard assignment $a_k(x_i)$, to make this operation differentiable,

Paper[3] replaces the assignment of the descriptor x_i to k cluster center c_k with a soft one:

$$\overline{a}_k(\mathbf{x}_i) = \frac{e^{-\alpha\|\mathbf{x}_i-\mathbf{c}_k\|^2}}{\sum_{k'} e^{-\alpha\|\mathbf{x}_i-\mathbf{c}_{k'}\|^2}}, \tag{5.2}$$

which assigns the descriptor x_i to cluster c_k with a weight according to the distance between them.
By expanding the squares in (5.2), it will result in the following form:

$$\overline{a}_k(\mathbf{x}_i) = \frac{e^{\mathbf{w}_k^T\mathbf{x}_i+b_k}}{\sum_{k'} e^{\mathbf{w}_{k'}^T\mathbf{x}_i+b_{k'}}}, \tag{5.3}$$

where vector $w_k = 2\alpha c_k$, scalar $b_k = -\alpha\|\mathbf{c}_k\|^2$, so the final VLAD representation which is able to train via backpropagation will be:

$$V(j,k) = \sum_{i=1}^{N} \frac{e^{\mathbf{w}_k^T\mathbf{x}_i+b_k}}{\sum_{k'} \mathbf{w}_{k'}^T\mathbf{x}_i + b_{k'}} (x_i(j) - c_k(j)), \tag{5.4}$$

where the $\{w_k\},\{c_k\},\{b_k\}$ are trainable parameters, and K as the hyper-parameter, so compared to the origin VLAD only trains $\{c_k\}$, this enables more flexibility and adaptability when facing more complex examples.

5.2. Deep architecture with VLAD

This section describes how the CNN architecture is combined with the VLAD layer. Image retrieval tasks usually follow this pipeline: (i) extract local image descriptors, then (ii) perform spatial pooling over the feature map. To learn the representation end-to-end, paper[3] introduced a CNN architecture called NetVLAD, it follows the same image retrieval pipeline: (i) crop the CNN at the last convolutional layer, and treat it as the local descriptor extractor, so the $H \times W \times D$ output can be viewed as D dimensional descriptors at $H \times W (H \times W = N)$ spatial locations. (ii) use the differentialble VLAD layer mentioned in Section 5.1 to pool the descriptors into a fixed size $D \times K$ image representation. The parameters from VLAD layer are trainable via backpropageation, this new architecture is called 'NetVLAD'. Figure 5.1 shows the whole pipeline and its inner operation of NetVLAD.

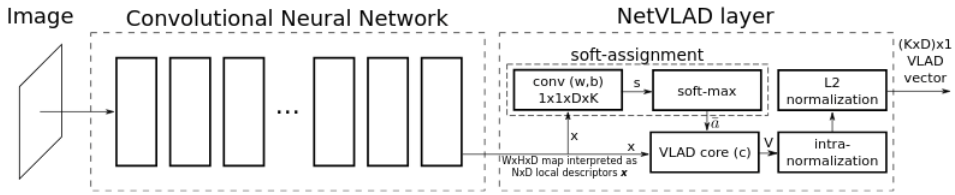


Figure 5.1: CNN architecture with VLAD layer[3]. The VLAD layer can be naturally implement through normal CNN operations: convolution, soft-max, L2-normalization. And the aggregation layer ('VLAD core') is also very easy to implement by following equation (5.4)

References

[1] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, *Object retrieval with large vocabularies and fast spatial matching,* in *2007 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2007) pp. 1–8.

[2] H. Jégou, M. Douze, C. Schmid, and P. Pérez, *Aggregating local descriptors into a compact image representation,* in *CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition* (IEEE Computer Society, 2010) pp. 3304–3311.

[3] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, *Netvlad: Cnn architecture for weakly supervised place recognition,* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016) pp. 5297–5307.

6

Attention Module

This chapter will introduce the general attention module for visual recognition, and how the attention module combines with NetVLAD, also will show the geometric interpretation of attention for VLAD.

6.1. Introduction to Attention Module

The attention module has been widely studied in visual recognition domain. CNN has been proven to be able to localize the image areas related to the label [1, 2]. The general idea of the attention module is inspired from how human eye works: Our retina is capable of receiving a wide view of the scene, but our visual processing mechanism is prone to "glance" the overview and only focus on a particular region of the view, while the rest is 'blurred out'. The visual attention model is trying to leverage on this idea, by paying more attention on a selected informative region, and less attention elsewhere. Fig.6.1 shows some examples of visual attention mechanism.

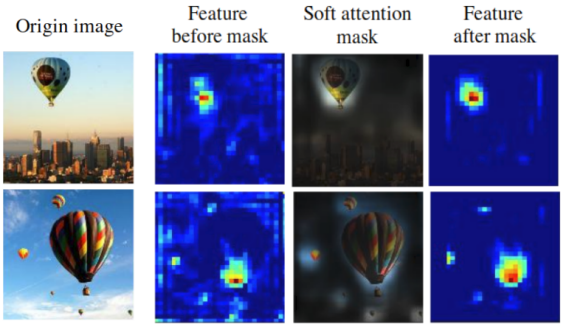


Figure 6.1: [3] The soft attention serves as a mask to filter out the uninterested region(sky) and highlights the parts that are important for classification.

6.2. Attention-aware VLAD Layer

From chapter 5 we show that the VLAD representation is highly effective for complex visual recognition tasks by aggregating local features from the entire image. Paper [4] addresses that while it is effective, it ignores the fact that some local descriptors may not be equally important, e.q. for descriptors locate at the sky region are less important than those locate at building region. Based on this, [4] introduces an attentional structured representation learning frame-work that incorporates an image-specific attention mechanism within the aggregation process to ignore non-relevant background information.

Formally, let $W \times H \times D$ be the final convolution layer output. The attention module usually consists of one 1×1 convolutional layer which will produce one heatmap with size $W \times H$. To incorporate the attention map (heatmap) into the VLAD layer, we can use the heatmap $w(x_i)$ to reweight the descriptors, since they share the same spatial size. Specifically, we can rewrite Eq 5.1 into

$$V(j,k) = \sum_{i=1}^{N} w(x_i) a_k(x_i)(x_i(j) - c_k(j)). \tag{6.1}$$

6.3. Geometric Interpretation of Attention

Given the descriptors extracted from two images from the same class but have different background, and the descriptors are assigned to the same cluster center, illustrated in Fig. 6.2 the original aggregation scheme would yield resid-

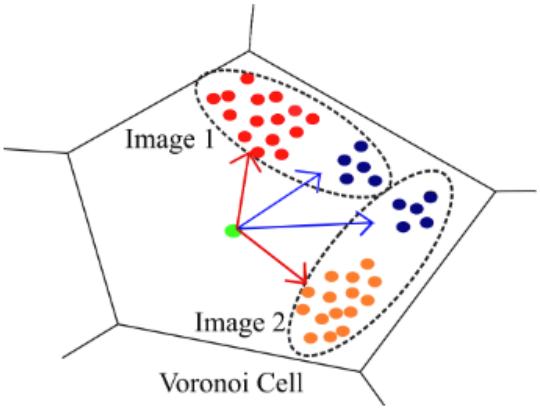


Figure 6.2: **Geometric interpretation of attention[4]**, the Voronoi cell shows the clustering used in VLAD descriptor construction, the green dot represents the cluster center, and the dots grouped into two ellipses represent local descriptors extracted from two different images, the descriptors with high attention are show in blue, and those with low attention are show in red and orange, while red arrows correspond to the sum of their residuals, the blue arrows correspond to the weighted sum of residuals.

ual vectors(red arrows) pointing in almost opposite directions, while the resulted

attention weighted aggregation vectors(blue arrows) are with high cosine similarity.Attention allows us to ignore confusing objects and improve the discriminative power of VLAD.

References

[1] A. Jimenez, J. M. Alvarez, and X. Giro-i Nieto, *Class-weighted convolutional features for visual instance search,* arXiv preprint arXiv:1707.02581 (2017).

[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, *Learning deep features for discriminative localization,* in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 2921–2929.

[3] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, *Residual attention network for image classification,* 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017), 10.1109/cvpr.2017.683.

[4] K. K. Nakka and M. Salzmann, *Deep attentional structured representation learning for visual recognition,* arXiv preprint arXiv:1805.05389 (2018).