Classification for Retrieval in Image Geolocalization

Hesam Khanjani
Politecnico di Torino
hesam.khanjanikakroodi@studenti.polito.it

Sepehr Alemzadeh
Politecnico di Torino
s314315@studenti.polito.it

Solmaz Verdiyeva
Politecnico di Torino
s289491@studenti.polito.it

Abstract

Geolocation, commonly referred to as geo-localization, pertains to the precise determination and representation of the geographical position of a device, user, or object in the physical realm. This involves assigning specific coordinates (latitude and longitude) or a distinct address to a particular entity. The primary focus of this paper is to extend and evaluate the efficacy of the CosPlace Model [3] on a smaller-sized dataset, thereby showcasing its versatility across various scenarios. Despite the existence of multiple models in the domain of visual geo-localization tasks, the CosPlace method stands out for its robustness, facilitated by the introduction of a new large-scale dataset and a classification-based approach for training models to extract discriminative descriptors without reliance on negative examples [3]. To enhance the effectiveness of the CosPlace method, several crucial aspects have been addressed in this research. Augmentation techniques, in conjunction with the backbone architecture, have been carefully modified. Prominent convolutional neural network (CNN) models, including ResNets and VGG16, have been evaluated as potential backbone architectures, considering different output dimensions. Furthermore, the optimization process has been optimized by examining various optimizers across different hyperparameters. Throughout the project, extensive experiments have been conducted for each of the three main components. The most favorable outcomes achieved from each individual experiment have been amalgamated to yield the final best result. This paper offers a comprehensive elucidation of each step and experiment, accompanied by graphical representations illustrating the obtained results. The transformed CosPlace and associated code can be accessed through our GitHub repository at https://github.com/hesamkh1/CosPlace_XS

1. Introduction

geo-localization, plays a vital role in our increasingly connected world by accurately determining and representing the geographic location of devices, users, and objects. It involves assigning coordinates or specific addresses to enable precise identification of locations. This information forms the foundation for various applications, such as navigation systems, location-based services, and mapping tools. With the ability to track and analyze objects or individuals based on their geographical positions, geo-localization provides real-time monitoring and enhances our spatial understanding. This research paper aims to extend and assess the effectiveness of the CosPlace Model in the context of visual geo-localization tasks, specifically on a smaller-sized dataset. While several models exist for this purpose, the CosPlace method demonstrates its versatility and robustness. This is achieved through the utilization of a novel large-scale dataset, as well as a unique classification-based training approach that eliminates the need for negative examples in extracting discriminative descriptors [9]. By exploring these advancements, this study aims to enhance the performance and applicability of visual geo-localization models in diverse scenarios, showcasing the potential of the CosPlace method.However, in this project we face a limitation.

small dataset size. A small dataset may result in reduced coverage of geographic locations, limited diversity in environmental conditions, and a restricted range of visual variations. Hence, the model may struggle to learn robust and accurate representations, leading to lower accuracy and recall. as shown in Tab.1 the size of training and testing data set considerably reduced.

Low resolution. Low resolution poses significant challenges and adversely impacts the recall in geo-localization. When comparing the original resolution of the SF-XL dataset (512 * 512) with the resolution of the new dataset SF-XS (224 * 224), several challenges arise. Firstly, the

	SF=XL (Train)	SF-XS (Train)	SF-XL (Val)	SF-XS (Val)	SF-XL test v1 (Test)	SF-XL test v2 (Test)	SF-XS (Test)
Database	41.2M	57K	8K	8K	2.8M	2.8M	27K
Queries			8K	8K	1000	598	1000

Table 1. Comparison between size of SF-XL and SF-XS

reduction in resolution results in a loss of detail, making it harder to extract distinctive features and accurately match them with reference images. Secondly, feature discrimination suffers as low-resolution images lack the level of detail required for precise localization. Additionally, the limited spatial context provided by low-resolution images hinders the accurate estimation of location. These challenges collectively contribute to decreased accuracy and recall in geo-localization tasks.

Contributions. In this paper,we address this limitation with following contributions:

- augmentation techniques, such as rotation adjustment and cutout, have been employed. The color Jitter component, including brightness, contrast, saturation, and hue adjustments, has been fine-tuned.
- The backbone architecture has been modified, testing popular CNN models like ResNet (Resnet-18, Resnet-50) and VGGNet (VGG16) with different output dimensions.
- The choice of optimizer has also been optimized, exploring options like Adam, AdamW and ASGD with different hyperparameters.Extensive experiments have been conducted for each component, and the best results from individual experiments are combined for the final outcome. This report provides detailed explanations, graphical representations of the results, and conclusions based on the findings.

2. Related works

Geolocation plays a crucial role in various applications, and place recognition is a fundamental task within this domain. In this context, NetVLAD [1], an influential CNN architecture, has made a significant contribution to the field. It addresses place recognition tasks by utilizing contrastive learning techniques, employing a triplet loss, and leveraging geo-tagged database images for weak supervision. NetVLAD combines deep learning with the Vector of Locally Aggregated Descriptors (VLAD) method, extracting local features and generating global descriptors. With its superior performance on large-scale place recognition datasets, NetVLAD demonstrates its potential for accurate and efficient geolocation, positioning it as a notable approach in the field.

In related works, some studies like [5] prioritize the preservation of high-dimensional embeddings, typically employing a dimensionality of 4096. This choice allows for the retention of information in a large feature space, providing potential benefits specific to their respective tasks or objectives. Conversely, other studies explore alternative approaches by foregoing the use of NetVLAD and instead employ pooling techniques to construct smaller embeddings. By utilizing pooling operations, these approaches achieve dimensionality reduction while still capturing relevant information for their specific purposes.

The Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization (SAR-ELSI) is an advanced method employed to determine the geographic origin [9]. By employing attraction and repulsion forces, SAR-ELSI organizes images such that similar ones are in close proximity while dissimilar ones are distanced apart . This iterative process continues until an optimal arrangement is achieved. SAR-ELSI is particularly advantageous in managing vast collections of images and holds potential for accurately estimating the location of photographs within cities or even on a global scale. Through the utilization of SAR-ELSI, precise geo-localization of images can be accomplished, offering valuable insights for various applications. Beside all previous methods Cosplace has shown potent result and performance which we use it in our project.

3. Proposed Method

The proposed method in this paper aims to compensate for the limitations of a small dataset in geo-localization. for this purpose we integrate augmentation modification [4], backbone fine-tuning, and optimizer selection. we choose best Augmentation techniques such as rotation, scaling, and random erasing to simulate images with nigh light, sunny light or images with different degree of capturing, and also when simulate when some object change the location that we can see. Simultaneously, the backbone architecture is adjusted and fine-tuned to extract and represent relevant features effectively. Furthermore, the choice of optimizer is carefully considered and optimized to enhance the learning process and model performance. By combining these three techniques, the project seeks to improve the recall and robustness of CosPlace model.

3.1. Augmentation

Random rotation is a crucial augmentation technique in geolocalization tasks as it enhances model robustness by introducing variations in object orientations. By training models on images with diverse rotations, the generalization capability of the models improves, enabling them to recognize objects from different viewpoints during testing [6]. The application of random rotation in geo-localization projects has been shown to enhance model performance and

yield more accurate and reliable results. We tried random Rotation technique with different angles. the input of Random Rotation is the degree of rotating image in backward or forward. this method can simulate the query images by phone with different angles. So choosing the best degree is crucial for this task.

Cutout also known as erasing is a dropout-based technique that applies occlusion to input images instead of feature maps, making it more resilient to noise compared to traditional noise-based dropout methods. Cutout offers two significant advantages. By employing Cutout, we can simulate scenarios where the subject of interest is partially occluded, enhancing the model's ability to handle such situations effectively. It encourages the model to utilize a broader range of content within the image for classification, preventing the network from fixating solely on salient areas and reducing the risk of overfitting. This integration of Cutout enhances both the robustness and generalization capability of the model in classification tasks. We used probability of 0.5 as one of the input for this method and with other input determine different size of cutouts.

colorJitter is the final step we tune the hyper parameter of colorJitter which was very important step of augmentation. in geolocalization tasks, where images are captured from different locations and at different times of the day, color jittering can help simulate different lighting conditions. By increasing the brightness, the augmented images can resemble images taken during the day or under strong sunlight. Conversely, by decreasing the brightness, the augmented images can simulate images taken during nighttime or in low-light conditions. Similarly, color jittering can also adjust contrast and saturation, allowing the augmented images to represent scenes with different levels of vibrancy, color richness, or atmospheric conditions. By manipulating hue, the augmented images can simulate variations in color tones, such as warm or cool lighting environments.By incorporating color jittering techniques in geolocalization tasks, the models can learn to generalize better and become more robust to diverse lighting and environmental conditions. This enables the models to perform well when applied to real-world geolocation scenarios, where lighting conditions and environmental factors can vary significantly.

3.2. Backbone

The backbone in a geo-localization model refers to the underlying architecture or feature extraction network that processes the input images and extracts meaningful representations. The choice of backbone can have a significant impact on the performance of geo-localization results. In this project Resnet-18,Resnet-50 and VGG16 have been tested along four different output layer dimensions. In the context of Resnet, when we increase the number of layers, feature vectors produced by the geo-localization model will have a higher dimensionality. This means that each feature vector will contain more elements or channels, representing more diverse and detailed information about the input images.Also the change in the output dimension layer affects the model's complexity and capacity. Increasing the dimensionality adds more parameters to the model, which can make it more expressive and capable of learning complex relationships. However, it also increases the risk of overfitting, especially when the available training data is limited. Regularization techniques such as weight decay may be necessary to reduce overfitting effects [14].The limited dataset size can prevent ResNet50 from fully exploring and leveraging the expressive power of its deep layers. Deep layers in convolutional neural networks are designed to learn more abstract and high-level representations, capturing complex features and patterns. However, in the absence of sufficient training examples, the deep layers may not be able to learn effectively, leading to suboptimal performance.

3.3. Optimizer

optimizer plays a vital role in updating the parameters of a machine learning model during training to minimize the loss function and enhance performance. In geo-localization, By determining the model's learning rate, update strategies, and momentum values, different optimizers affect the speed and efficiency with which the model converges to an optimal solution. In this project, we tested Adam,AdamW and ASGD with different learning rates. The performance of optimization algorithms in geo-localization tasks can vary depending on factors such as the specific dataset, model architecture, and training settings. However, in many deep learning applications, including geo-localization, Adam and its variants, such as AdamW, have demonstrated competitive performance [15]. These optimizers are often preferred due to their adaptive learning rate mechanisms. The results of our experiments support this claim, as Adam and AdamW yielded favorable outcomes in the context of geo-localization.

4. Experiments

In the initial phase of our experimentation, we utilized the Cosplace framework in conjunction with ResNet-18 as the underlying architecture, employing an output dimension of 512. This configuration yielded a R@1 result of 21 on the SF-XS test set after 10 epochs. Subsequently, we also explored the use of ResNet-50 with 10 epochs and an output dimension of 512. However, the R@1 performance on the SF-XS test set decreased to 5.2.

The decline in recall performance observed while transitioning from ResNet-18 to ResNet-50 can be partially ascribed to the limited size of our training dataset. The process of training deep learning models with relatively small

datasets presents inherent challenges in learning intricate representations. Due to its greater capacity and increased number of parameters, ResNet-50 is more prone to overfitting when trained with a limited amount of data. This susceptibility to overfitting can impede the model's ability to generalize to new examples, resulting in diminished recall rates. It is worth noting that the diversity and size of the training dataset exert a significant influence on feature learning within deep neural networks [6]. In geo-localization, where the task involves recognizing and localizing objects in diverse environments, a restricted training dataset may hinder ResNet-50's capacity to capture the intricate spatial and contextual information necessary for accurate recall [2].

To address the limitations imposed by the limited training dataset, we introduced data augmentation techniques to enhance the diversity of our dataset. Our initial approach involved the utilization of random rotation with varying degrees. As indicated in Tab.2, degrees of +10 and -10 yielded the most favorable results. Subsequently, we explored the implementation of random cutout, also known as random erasing. Fig.1 illustrates that random erasing can introduce variations into images, potentially obstructing the visualization of specific objects or impeding accurate localization. To determine the optimal size of random cutout, we experimented with different sizes while maintaining a 50% probability. The selection of the cutout size was dependent on the image resolution, ultimately leading to improved results.

Figure 1. A sample of transformation of random rotation with 10 degree of freedom in right image and random erasing in left image

Moreover, we observed significant changes in the augmented images by modifying the parameters of ColorJitter. Increasing the brightness parameter resulted in augmented images with heightened illumination, allowing for improved visibility [13]. By expanding the brightness range from 0.7 to 0.8, we observed that images could simulate varying lighting conditions, such as a night scene with reduced brightness or a sunny day with intensified illumination as depicted in Fig.2 . Decreasing the contrast parameter led to augmented images with reduced visual distinctness, mitigating the impact of high contrast variations [7]. Setting the hue parameter to 0.0 eliminated any hue shifts in the augmented images, resulting in images with consistent base colors.

Augmentation Method	SF-XS		Tokyo-XS	
	R@1	R@5	R@1	R@5
Rotation(-15,15)	18.9	32.1	27.6	50.5
Rotation(-45,45)	17.4	29.7	27.3	47.0
Rotation(-25,25)	17.2	31.1	32.4	49.5
Rotation(-10,10)	19.1	32.1	30.8	50.8
Cutout(16,0.5,0)	19.2	31.0	33.3	55.6
Cutout(16,0.8,0)	17.6	31.6	31.1	51.1
Cutout(16,0.3,0)	18.1	31.6	31.4	54.6
Cutout(32,0.5,0)	18.7	31.5	34.0	57.1
ColorJitter (brightness=0.8 , contrast=0.5 ,saturation=0.7, hue=0.5)	19.4	31.4	32.4	54.6
ColorJitter (brightness=0.5 , contrast=0.7 ,saturation=0.7, hue=0.5)	18.9	32.3	33.7	54.6
ColorJitter (brightness=0.8 , contrast=0.3 ,saturation=0.7, hue=0.0)	23.9	38.5	36.2	59.0

Table 2. Testing Augmentation methods with backbone resnet-18 on 1 epoch with output Dimentsion 512 train data on SF-XS training database. In each step, we choose the best hyperparameter for previous method and implement next method.

Figure 2. A sample of transformation of colorjitter daytime photos into nighttime photo and sunny and more contrast photo

These adjustments in augmentation parameters carry significant implications for downstream tasks involving trained models. Increasing brightness enhances visibility in darker images, although caution should be exercised to prevent overexposure. Decreasing contrast enhances the model's robustness to high contrast variations, but excessively low contrast may result in flattened images lacking depth [11]. Setting the hue parameter to 0.0 ensures consistent color representations, particularly when color information is not

	Adam		AdamW		ASGD	
Learning rate	1e-05	0.001	1e-05	0.001	1e-05	0.001
weight decay	0.001	0.001	0.001	0.001	0.001	0.001
R@1	30.3	25	29.7	23.3	29.1	23.1

Table 3. R@1 result for two elected experiments for each optimizer with different hyperparameters having weight decay.

critical for the task.We also explored the inclusion of Gaussian noise [10], horizontal flipping, and vertical flipping as additional augmentation techniques aimed at improving the dataset and training the model. However, the results yielded no considerable improvements for this specific task and dataset. To further analyze the impact of augmentation, we conducted testing with resnet-18 with one epoch, and the results are presented in Tab.2 . So far, we have achieved the best results in the section of data augmentation. In addition to that, we conducted experiments using different optimizers to determine if they could improve the overall R@1 performance. The default optimizer used was Adam, but we replaced it with other optimizers such as ASGD and AdamW to compare their performance against Adam. The experiments were conducted in 5 epochs, The new augmentation techniques, which were found to be the best among other settings, were chosen. The default backbone architecture used was Resnet-18. for ASGD, we conducted experiments with different hyperparameters, including weight-decay and lambda. The best R@1 achieved using ASGD was 29.1. For AdamW, the best R@1 calculated was 29.7. Additional hyperparameters such as weight-decay and lambda were also adjusted for this optimizer. Despite conducting numerous experiments throughout this process, the results indicate that Adam itself yields the best performance among all the optimizers. The adjustment resulting in the best performance was a learning rate of 1e-05.The default values for classifiers_lr, classifiers_wd, and classifiers_lambda are set to 0.001 for all classifiers. Overall, the experiments revealed that Adam with a learning rate of 1e-05 provided the most favorable results which was 30.3. The results are all shown in the Tab.3 .

Following the modifications made to the optimizer and the incorporation of dataset augmentation techniques, we proceeded to evaluate the performance of three different backbone architectures: ResNet-18, ResNet-50, and VGG16. ResNet-18 demonstrated favorable speed characteristics and exhibited promising results when varying the output dimension of descriptors. Notably, the modifications applied to the model architecture, along with the augmented dataset, resulted in improved performance for ResNet-50. This suggests that deeper layers of ResNet-50 have the potential to capture more complex descriptors, enabling enhanced predictive capabilities [8, 12]. Specifically, our experimentation revealed that ResNet-50 with an output layer of dimension 1024 achieved the most favorable results in practical query prediction. Furthermore, we explored the utilization of the VGG16 backbone architecture. The incorporation of VGG16 led to a substantial improvement in the recall at R@1, achieving a noteworthy R@1 score of 42 on the final SF-XS test set. This improvement is highly significant and underscores the efficacy of VGG16 in the given context. The enhanced performance of the models can be attributed to several factors. Firstly, the optimization of the model architecture and hyperparameters, along with the utilization of an augmented dataset, contributed to the models' ability to learn more discriminative features. The augmentation techniques employed, such as random rotation, random cutout, and modified ColorJitter parameters, enhanced the diversity and richness of the training data. This, in turn, facilitated better generalization and improved the models' capacity to handle various image conditions and variations. Moreover, the choice of backbone architecture played a crucial role in the models' performance. While ResNet-18 exhibited efficient speed characteristics, ResNet-50 proved to be more effective in capturing complex descriptors, especially when using a higher output dimension. The deeper layers of ResNet-50 enabled the model to extract more intricate and meaningful features, resulting in enhanced predictive capabilities. Similarly, VGG16 demonstrated remarkable performance, achieving a substantial increase in R@1 score. The inherent architecture and design choices of VGG16, such as its deep convolutional layers and max pooling, contributed to its success in image classification tasks. We tested these three backbones with different output layers as shown in Fig.3 and selected the optimal size for each backbone to run for 10 epochs. The observed improvements in the models' performance carry significant implications for practical applications and downstream tasks and the best results are shown in Tab.4 .

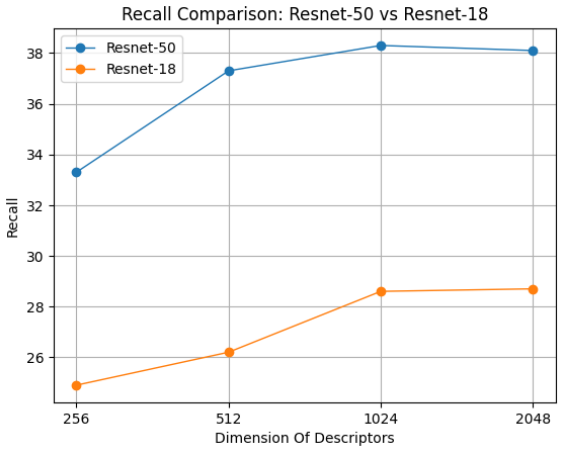

Figure 3. Result of Resnet-50 and Resnet-18 with Adam on SF-XS test

	SF-XS		Tokyo-XS	
Backbone	R@1	R@5	R@1	R@5
Resnet-18	29.6	46.1	39.0	56.8
Resnet-50	39.2	53.8	47.9	64.4
VGG16	41.2	55.6	53.3	68.9

Table 4. final result with 10 epochs

5. Conclusion

In conclusion, this study highlights the importance of improving datasets through augmentation techniques in the domains of geo-localization and image retrieval. By incorporating various augmentation methods such as image rotations, random erasing, color adjustments, and contrast modifications, the dataset becomes enriched with diverse and representative samples. This augmentation process plays a vital role in enhancing the robustness and generalization ability of models, enabling them to perform effectively in real-world scenarios. Moreover, the selection of an optimal backbone architecture is critical for extracting discriminative features from the input images. A well-chosen backbone architecture serves as the foundation for accurate and reliable geo-localization and image retrieval. It allows the models to capture and leverage important spatial and semantic information, leading to improved performance. Furthermore, the careful tuning of hyperparameters, including the learning rate, regularization techniques, and optimizer settings, is crucial for achieving the best possible results. The learning rate determines the rate of convergence and the balance between exploration and exploitation during training. Fine-tuning these hyperparameters ensures efficient optimization and convergence, leading to enhanced performance and accuracy in geo-localization and image retrieval tasks. By combining these key elements—augmented datasets, optimal backbone architectures, and fine-tuned hyperparameters—we can significantly improve the performance and reliability of geo-localization and image retrieval systems. This, in turn, leads to more accurate and precise query predictions, benefiting a wide range of applications that rely on effective information retrieval. Accurate and efficient query prediction holds great importance in real-world scenarios where quick and precise retrieval of relevant information is vital. The advancements achieved in recall performance can contribute to enhanced user experiences, improved search functionality, and increased productivity.

References

[1] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition, 2016. 2

[2] Mathieu Aubry and Bryan Russell. Understanding deep features with computer-generated imagery, 2015. 4

[3] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications, 2022. 1

[4] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark, 2023. 2

[5] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 369–386. Springer, 2020. 2

[6] Tong He, Zhi Zhang, Hang Zhang, Zhongyue Zhang, Junyuan Xie, and Mu Li. Bag of tricks for image classification with convolutional neural networks, 2018. 2, 4

[7] Philip TG Jackson, Amir Atapour Abarghouei, Stephen Bonner, Toby P Breckon, and Boguslaw Obara. Style augmentation: data augmentation via style randomization. In *CVPR workshops*, volume 6, pages 10–11, 2019. 4

[8] Brett Koonce and Brett Koonce. Resnet 50. *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*, pages 63–72, 2021. 5

[9] Liu Liu, Hongdong Li, and Yuchao Dai. Stochastic attraction-repulsion embedding for large scale image localization, 2019. 1, 2

[10] Florian Luisier, Thierry Blu, and Michael Unser. Image denoising in mixed poisson–gaussian noise. *IEEE Transactions on image processing*, 20(3):696–708, 2010. 5

[11] Maria Ines Meyer, Ezequiel de la Rosa, Nuno Pedrosa de Barros, Roberto Paolella, Koen Van Leemput, and Diana M Sima. A contrast augmentation approach to improve multiscanner generalization in mri. *Frontiers in neuroscience*, 15:708196, 2021. 4

[12] Edmar Rezende, Guilherme Ruppert, Tiago Carvalho, Fabio Ramos, and Paulo De Geus. Malicious software classification using transfer learning of resnet-50 deep neural network. In *2017 16th IEEE international conference on machine learning and applications (ICMLA)*, pages 1011–1014. IEEE, 2017. 5

[13] Ryo Takahashi, Takashi Matsubara, and Kuniaki Uehara. Ricap: Random image cropping and patching data augmentation for deep cnns. In *Asian conference on machine learning*, pages 786–798. PMLR, 2018. 4

[14] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. *arXiv preprint arXiv:1810.12281*, 2018. 3

[15] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank J Reddi, Sanjiv Kumar, and Suvrit Sra. Why {adam} beats {sgd} for attention models, 2020. 3