*International Journal of Engineering Applied Sciences and Technology, 2021*
**Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 217-221**
Published Online March 2021 in IJEAST (http://www.ijeast.com)

# LANDMARK-BASED VISUAL PLACE RECOGNITION

Swapnali Gavali
Department of CSE
WCE, Sangli, Maharashtra, India

Dr. Bashirahamad Momin
Department of CSE
WCE, Sangli, Maharashtra, India

***Abstract*— Landmark recognition is one type of problem of object recognition that has not been well solved. The classical techniques used for object recognition can not be applied directly due to a large number of landmarks and a highly imbalanced dataset. This paper presents the application of a triplet network for large scale landmark-based visual place recognition. By fine-tuning pre-trained convolutional neural network (CNN) and minimizing triplet loss, the triplet network can learn appropriate metrics so that most similar images can be retrieved through algorithms for the k-nearest neighbor (KNN). The performance of the proposed method is evaluated on a data set for the recognition of real-world landmarks.**

*Keywords*— **Landmark recognition, Triplet network, Deep learning, Metric learning**

## I. INTRODUCTION

I Over the past few years, people have made huge progress in computer vision with the rapid development of deep learning techniques, particularly convolutionary neural networks (CNN). Different kinds of network architectures have been proposed. Some widely used architectures include VGG, ResNet, Inception Network and so on. Some algorithms have achieved greater accuracy than humans on the famous ImageNet Large Scale Visual Recognition Challenge (ILSVRC). CNN's have proven to be one of the best solutions for computer vision tasks, widely used in various fields, such as autonomous cars and automatic face recognition. The rapid development of the technologies of deep learning and computer vision has transformed the lifestyle of people. Classical activities in object recognition typically require a large number of learning objects, though. For example, in 10 different classes, the data set includes 60,000 training objects. ImageNet dataset contains 14,191,122 images in 1,000 classes. One of the key factors in ensuring CNN's success is a large number of training images. Moreover, object recognition systems usually use a fully-connected layer as the structure for the final one or several layers, typically requiring a large number of parameters. However, there exists a special kind of problem called one-shot learning that cannot be easily solved with classical methods. The goal of one-shot learning is to learn information from one or only a few images of training. And there might be a huge number of classes. New solutions are needed for these two factors. There are several well-known examples of one-shot learning, such as face recognition/verification and famous street-to-shop systems. Over the past several years, a series of solutions have been proposed to solve one-shot learning problems. Of example, of various issues, including authorship and image recognition, the Siamese network has been introduced. Recognition of face or facial verification was well studied. For audio and image recovery problems, triplet-based networks are tested. This paper focuses on the recognition of images as attractions. There are thousands of landmarks. Some of them are very popular and others are less so.

## II. LITERATURE SURVEY

According to Karen Simonyan and Andrew Zisserman, in [1] end-to-end Visual Recommendations and e-commerce search solutions on a large scale. We shared the details of our deep Convolution Neural Network model's architecture and training, achieving state-of-the-art results for image retrieval on the Street2Shop dataset that is publicly available. As well as described the challenges involved in the deployment of a large-scale Visual Recommendation Engine and various trade-offs. Our business impact figures show that a visual recommendation engine is an important weapon in any retailer's arsenal. Kaiming He et al. proposed that [2], very interesting findings and the fact that it works so well is somewhat surprising. In fact, developing small networks that can run on a mobile phone and are compliant with a broader server-side system would be interesting. Christian Szegedy et al. proposed in [3]a new profound ranking system to learn similarity models with the fine-grained objects. The deep ranking model uses a triplet-based loss ranking method to classify relationships of fine-grained image similarity, and multi-scale architecture of the neural network to capture both global visual and semic image properties. We also deliver an effective online triplet sampling method that allows us to learn from a very large amount of training data for deep ranking models. The empirical assessment shows that the model of deep ranking achieves much better performance than the state-of-the-art hand-crafted features based models and deep classification models. Models of image similarity can be applied to many other applications for computer vision, such as object recognition/detection based on examples and image deduplication. We are going to explore these ways.Alex


Krizhevsky and Geo Rey Hinton have presented in [4] a neural network-based approach to matching a consumer image in online shopping sites to precisely the same item. To mitigate the effect of tag interference and prevent overfitting caused by some (visually different) positive pairs and robust contrastive failure which eliminates these learning samples in the network training process automatically. A multitask approach was also proposed to exploit additional ImageNet information for improved results with a softmax failure. J. Deng et al. [5] It has been shown that, contrary to intuition, useful aspects of a new category of objects can be learned from a single (or just a few) training examples. The key insight utilized is that with fewer learning experiences, the categories we have already mastered provide us with information that helps us learn new categories. A Bayesian learning framework based on representing categories of objects with probabilistic models was developed to pursue this idea. Past information from previously learned groups was represented on the parameters of their equations with an appropriate prior probability density function. Such previous models were revised with the few available learning examples to generate posteriors that, in effect, can be used for both identification and discrimination. Experiments carried out on images from 101 categories were promising in that they indicate that very few (1 to 5) learning examples yield models capable of achieving detection output of around 70-95 percentiles. Li Fei-Fei, Rob Fergus, and Pietro Perona presented in [6] the Deep CNN framework in a recommendation system for quick clothing recovery. To simplify quest, we add a latent layer to the binary code learning network that can be used to quickly identify a pool of object candidates for eventual refinement. Experimental results on a large clothing dataset show that our hierarchical, coarse-to-fine search results in a retrieval speed of 10x and 50x compared to a comprehensive search using CNN and hand-crafted baseline features. In the future, apart from category labels, we plan to include attributes in order to provide a more comprehensive description of the collected clothing images and to test our method on the images with attribute annotations. Si Liu et al. [7] In this study, examined Siamese convolutional neural network architectures to verify authorship of handwritten text. We first examined the length of text required to make an accurate prediction and discovered that lines of words performed much better than single words. Semantically, this allowed us to determine how much is writing is necessary to create an accurate writer verification system. In this case, each line contains 5 to 10 words on average. In the future, the next logical step will be to expand the use of the complete images of pages rather than single lines. This has been dependent on the available memory on the Google Cloud machine. Next, we examined different architectures to determine which architecture provided the most informative encoding that was able to differentiate between authors. Beyond our baseline model, we explore VGGNet, GoogLeNet and different sizes of ResNet. The best performing model was a shortened version of ResNet which we coined TinyResNet. A single TinyResNet model trained on 25,000 samples was able to achieve 92.08% accuracy on a held-out test set. Ensembling 5 TinyResNet models trained on 10,000 datapoints each did not significantly improve the model. Thus, in the future, it would be interesting to continue training on even larger datasets not limited by memory restrictions. Visualizing the saliency maps for the Siamese network trained with TinyResNet shows that the networks unsurprisingly examines the handwritten text in each image. Strokes that go below or above the line of text typically have high importance. The end and beginning of strokes typically have high importance as well. Visualizing the filters for the network ultimately produces somewhat noisy filters despite the network being well-trained. These filters do show some semblance of curvature. However, perhaps most importantly, these filters which are trained from scratch perform extremely well on the given task. These filters look extremely different from those typically trained on image classification tasks like ImageNet. This could indicate that the type of filter necessary to achieve good results on images of handwriting is different than those for image classification. This means that pretrained weights would theoretically perform poorly on this task. An interesting exploration in the future would be to use pretrained weights from ImageNet on this task in order to validate our hypothesis. Future improvements of this model could incorporate the word embeddings for each of the words included in the image as well as mathematical representations of vectorized Handwriting. Generative Adversarial Nets (GANs) to construct counterfeit pieces of writing. The goal of this sub-problem would be to train a network that would be able to generate handwriting similar to a given input. Ideally, the goal will be: given a piece of handwriting, can we generate a counterfeit handwritten piece? However, we first will need to confirm that a GAN on handwriting generates handwriting with a proper alphabet rather than random lines. Thus, in this paper, we have shown that using Siamese convolutional neural networks are able to perform well on the handwriting verification task. This network could have large impacts on forensic analysis, historical text verification, and signature verification. Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov presented in [8] a one-shot identification technique by first studying strongly coevolutionary Siamese neural validation networks. Comparing the output of our networks with an existing state-of-the-art classifier built for the Omniglot data set, presented new results. The networks surpass with a significant margin all available baselines and come close to the best numbers obtained by previous researchers. We argued that the strong performance of these networks in this challenge not only suggests that human-level precision is feasible with our metric learning approach, but that this approach could apply to one-stage learning tasks in other fields, in particular with regard to object classification. Florian Schro, Dmitry Kalenichenko, and James Philbin in [9] the new mission, Exact Street to Shop, was presented and a


new database was launched. Using this dataset, we tested three street-to-shop retrieval approaches, including our Similar to Request Similar to object approach to learning correlation measurements between street and market domains. Lastly, we carried out quantitative and human evaluations of our results, demonstrating good accuracy for this challenging task of recovery. Such approaches provide an initial step in helping online retailers to reliably retrieve clothing items. Future work involves designing approaches to match street and shop products more accurately to boost recovery quality. Jingtuo Liu et al. [10] demonstrated that the batch hard triplet approach significantly outperforms conventional triplet approaches, even when conventional approaches use hard mining. Batch hard approaches have more stable networks and converge much faster than conventional Approaches. However, the accuracy for queried results even with the batch hard approach, on test split (unknown classes) is only 10% to 20%, which still has a lot of room for improvement. The high intraclass variance and low inter-class variance is one of the challenges with fine-grained classes and cluster in and querying images from unknown classes. To account for the variance, we experimented with local positive sampling and clustering with more than 1 cluster per class. While we did not see huge improvements with local positive sampling for batch hard approach, we did notice an improvement in accuracy when we use clustering accuracy for 3 clusters per class to drive the model selection, vs. 1 cluster per class. One important issue with training triplet networks, that we noticed, is that clustering and querying accuracy can improve with extended training, even if the loss does not change much, as the network continues to learn from hard triplets, pushing hard negatives apart. This was also reported in, which recommends continuing training triplet networks even if the loss flattens. It should be possible to design new sampling methods or loss functions or add a classification loss using cluster centroids obtained using K-means clustering, to the total loss term, to drive the network to learn embeddings that map to extended manifolds rather than spheres, and capture intraclass variance, similar to, even for batch hard networks. It does not make sense to learn centroids or key points since classes from test data are not known at training time. Computing cluster centroids, however, is expensive, and that limits how frequently we can update these centroids for the classification loss term. Another interesting area of future work would be exploring other batch global loss functions, rather than pairs, triplets or quadruplets within a batch. Batch hard triplets already outperform conventional triplets, and if we are paying the cost of going over all triplets within a batch, it should be possible to design alternative loss functions over the entire batch that performs better. It may even be possible to design more efficient batch global loss formulations, using matrix operations. This is the idea behind the lifted structured loss, and exploring alternatives is an interesting future direction. Even though forward and backward propagation for each batch becomes more expensive in these cases, networks converge with far fewer iterations, making these approaches more appealing than conventional contrastive or triplet approaches. Jiang Wang et al.[11] suggested a two-stage face recognition system that blends deep CNN with metric learning. Our method can handle cases with variant poses, occlusions and expressions well, benefiting from multi-patch features. The performance improves correspondingly as the number of identities and faces per identity in learning information decreases. The proposed method outperforms state-of-the-art methods on LFW under key protocols and when the FAR is rather small, it achieves a quite high confirmation level. As the algorithm continues to improve, we hope that face recognition technology can ultimately be used widely in more challenging real-world conditions. Kevin Lin et al.[12] suggested an optimal facial classifier that would reliably recognize faces that are matched only by humans. For pose, light, voice, and image quality, the underlying face descriptor should be invariant. It should also be universal in the sense that, if any, it can be extended to different populations with few changes. Therefore, simple descriptors and, if possible, minimal elements are preferred. Fast computing time, of course, is also a problem. We believe that this work, which starts from the recent trend of using more features and using a more powerful method of metric learning, has addressed this challenge, closing the vast majority of this performance gap. Their work shows that integrating a 3D model-based integration with high-capacity feedforward models can easily benefit from many instances to address the drawbacks and shortcomings of previous methods. The ability to show a marked improvement in face recognition testifies to the promise of such a pairing in other fields of vision as well.

## III. METHODOLOGY

Recently, Google released a large landmark dataset on Kaggle. It contains more than 1 million images labeled into 14,951 different landmarks. Some famous landmarks contain a huge number of training images. However, for the not-so-famous landmarks, there are less labeled images. A large number of classes and a few numbers of training examples per class increase the complexity of the problem. The given dataset only contains the URL and landmark id for each image. All the images need to be downloaded from online and different images have different sizes and resolutions. A large number of training images, preprocessing steps, and a large number of images to analyze, make this problem challenging for both algorithms and computation power. Triplet loss is a loss function of artificial neural networks where baseline (anchor) input is compared to positive (true) input and negative (false) feedback. The variance from the baseline input (anchor) to the positive input (truthy) is minimized and the distance from the baseline input (anchor) to the negative input (false) is maximized. It is often used to learn similarities for the purpose of learning embeddings, such as word embeddings and even

*International Journal of Engineering Applied Sciences and Technology, 2021*
**Vol. 5, Issue 11, ISSN No. 2455-2143, Pages 217-221**
Published Online March 2021 in IJEAST (http://www.ijeast.com)

thought vectors, and metric learning. A Euclidean distance function can be used to define the loss function.

$$\mathcal{L}(A,P,N) = \max\left(\|\mathrm{f}(A) - \mathrm{f}(P)\|^2 - \|\mathrm{f}(A) - \mathrm{f}(N)\|^2 + \alpha, 0\right)$$

Where A is an anchor input, Pis a positive input of the same class as A.N is a negative input of a different class from A,α is a margin between positive and negative pairs, and f is an embedding. This can then be used in a cost function, which is the sum of all losses, which can then be used to minimize the optimization problem.

$$\mathcal{J} = \sum_{i=1}^{M} \mathcal{L}\left(A^{(i)}, P^{(i)}, N^{(i)}\right)$$

The indices are given as a triplet for the individual input vectors. The triplet consists of drawing an anchor input, a positive input describing the same entity as the anchor entity, and a negative input not describing the same entity as the anchor entity. Then these inputs are run through the network, and the loss function uses the outputs.
In computer vision, a prevailing belief has been that the triplet loss is inferior to using surrogate losses followed by separate metric learning steps. This method showed that for scratch-trained models as well as pre-trained models, a special version of triplet loss performs deep metric learning end-to-end.

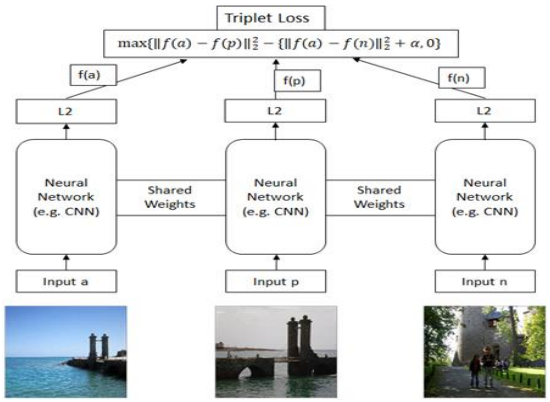


Fig. 1. Triplet Network

Works have been well trained with the large training set, sufficient training time and advanced hardware and fine-tuning with pre-trainedVGG16, InceptionV3, and ResNet models. Usually, lower layers encode more generic, reusable features that encode more specialized features with higher layers. Freeze lower layers and train only multiple layers at the top. Given a new image, prediction with landmark K-Nearest Neighbors (KNN) the following Fig. 2.

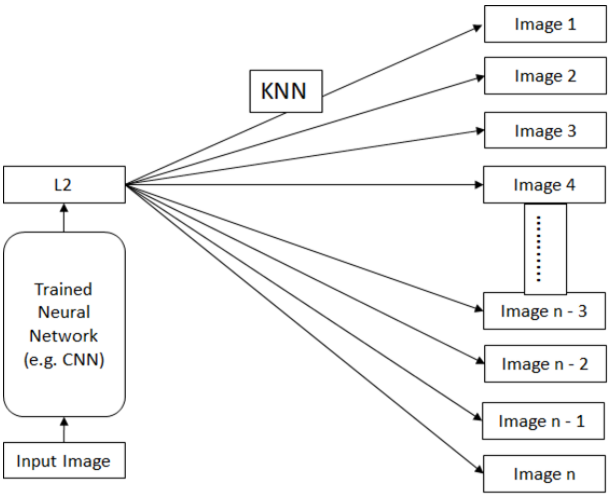


Fig. 2. Illustrate Prediction with K-Nearest Neighbors (KNN)

**Result and Analysis-**

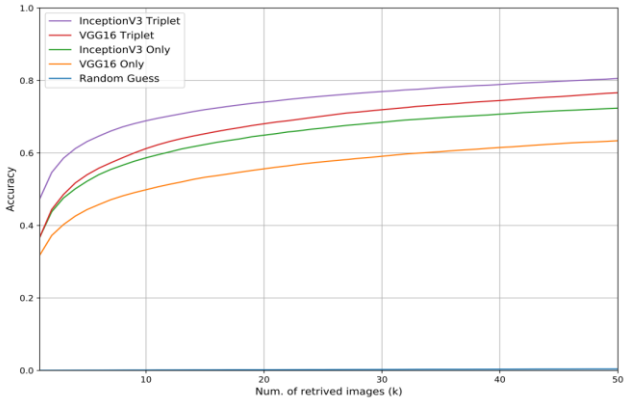


Fig. 3. Top-k image retrieval accuracy for different numbers of retrieved images.

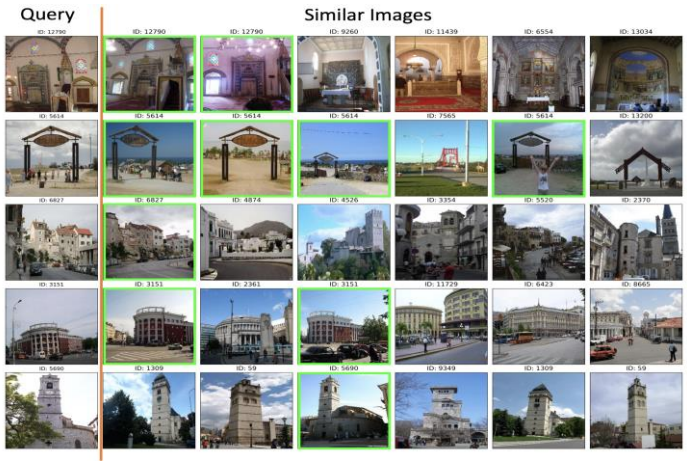


Fig. 4. Result of Multiple similar Landmarks.


## IV. CONCLUSION

Proposed method expands the CNN based landmark recognition from two viewpoints, i.e. series query indexing of landmarks. Extensive tests were carried out to check the validity of the query sequence combination and the indexing of KNN. Such analytical findings on four demanding reference datasets demonstrate that the situation and perspective adjustments improve the reliability of the Landmark Recognition using Deep Learning Models. The conventional landmark-based neural network, some of our pipeline's work is done offline, restricting our robot applications function. Develop a real-time model of future work; this approach can, therefore, be used in problems of robotics. To enhance the Deep Learning landmark-based in the future, conduct research on the use of semantic landmarks and leverage further prior information in pictures.

## V. REFERENCE

[1] Karen Simonyan, and Andrew Zisserman (2014)'. Very deep convolutional networks for large-scale image recognition. (Pg 1309-1332).

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016)'. Deep residual learning for image recognition, IEEE conference (Pg 143-154).

[3] Christian Szegedy, Vincent Vanhoucke, Sergey Ioe, Jon Shlens, and Zbigniew Wojna (2016) .Rethinking the inception architecture for computer vision, IEEE. (Pg 17).

[4] Alex Krizhevsky, and Geo Rey Hinton (2009). Learning multiple layers of features from tiny images.(Pg 3-7).

[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei.(2009). ImageNet: A Large-Scale Hierarchical Image Database. (Pg 1097-1105).

[6] Li Fei-Fei, Rob Fergus, and Pietro Perona (2006). One-shot learning of object categories, IEEE transactions.

[7] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan (2012). Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In Computer Vision and Pattern Recognition (CVPR), IEEE Conference. IEEE.(Pg 1097-1105).

[8] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov (2015), Siamese neural networks for one-shot image recognition. In ICML Deep Learning Workshop, volume 2. (Pg 2238-2245).

[9] Florian Schro, Dmitry Kalenichenko, and James Philbin (2015). Facet: A unified embedding for face recognition and clustering. IEEE (Pg 4297-4304).

[10] Jingtuo Liu, Yafeng Deng, Tao Bai, Zhengping Wei, and Chang Huang (2015). Targeting ultimate accuracy: Face recognition via deep embedding. (Pg 23-36).

[11] Jiang Wang, Thomas Leung, Chuck Rosenberg, Jinbin Wang, James Philbin, Bo Chen, Ying Wu, et al. (2014) Learning engrained image similarity with deep ranking.(Pg 21)

[12] Kevin Lin, Huei-Fang Yang, Kuan-Hsien Liu, Jen-Hao Hsiao, and Chu-Song Chen (2015). Rapid clothing retrieval via deep learning of binary codes and hierarchical search, ACM. (Pg 26-30).