THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

Empirical Study on Vanilla CNN Model for Mobile Visual Search

STUDENT NAME: XiaoWen LI

STUDENT NUMBER: 20543541

SUPERVISOR: James SHE

Date:2019.04

Empirical Study on Vanilla CNN Model for Mobile Visual Search

Abstract

Vanilla convolutional neural networks have showed very good performance on visual recognition tasks. AlexNet, VGG, Inception, ResNet are some of the popular networks. In this paper, I introduce the CNN network which is used for mobile visual search and show the performance in Visual Arts search based on a content-based image retrieval system. The dataset is visual arts images. The empirical study focuses on the performance of CNN network on the realistic condition. And the result shows that CNN model has good performance in popular standard image dataset, e.g. ImageNet, but for the images that were taken with rotation, from different angles or with other variations, the performance has been badly affected.

Keywords: CNN network, mobile visual search, ResNet50 model

CONTENTS

1 INTRODUCTION

The rapid development of mobile devices has led to many new mobile visual search applications that allow users to perceive their environment through their mobile devices. While it seems to be easier for human being to find similar images in the image database, it is quite challenging for machines because it requires the model to understand the image content and find connections with other images. As a result, content based image retrieval is not an easy task in mobile visual search.

For example, visual arts online search may be a very demanding application. When people visit an exhibition in a museum or art gallery, they may wonder if there are other similar works of the art they are looking at. But this is even more challenging for a content-based image retrieval (CBIR) system because the arts images are captured by people. In this scenario, the arts may have rotation, blur, shot from different angles or have different illumination. All this changes may affect the precision of image retrieval.

One important techniques is known as “deep learning” [1], which can model senior abstractions in data using a deep architecture made up of of multiple nonlinear transformations. Convolutional neural network (CNNs) is one of the main categories to do images recognition, images classifications. Image retrieval, target detection and face recognition are among the most widely used fields of CNNs. For CBIR system of mobile visual search, the mainstream method consists of two stages: (i) the CNN model is trained from a large number of training data and (ii) applying this model to learn the feature expression of CBIR tasks.

Fig.1 shows the image retrieval results in Google between the high-definition image and the image taken by visitors. The first line is the retrieval result of high-definition images, and the second line is the query images from the captured image of this art prints. It shows that in addition to the painting I want to retrieve, the picture also includes the wall background that is not related to the content. Therefore, shooting this behavior may result in many changes in the presentation image of the artwork with the same content with the consequences that precision of image retrieval decrease significantly.

In this paper, I attempt to explore CNN architecture with application to CBIR tasks in mobile visual search. Although there are much attention of applying CNN networks for image retrieval, The focus on the original data set remains limited. I aim to investigate the following questions:

(i) Are CNN network effective for feature representations of images to settle the CBIR tasks?

(ii) How the shooting behavior affect the image retrieval precision on the CBIR system based on CNN networks?

In order to answer these questions, I firstly introduce and explain the CNN architecture and the classical model such as VGG and ResNet. Then I propose a CBIR system to finish empirical investigation of the issues that may affect the visual arts

retrieval precision. The CBIR system includes two steps: Firstly extract the feature by employing the ResNet architecture to represent the contents of visual arts. Secondly compute the similarity between the features of query image and the database to find the top-k most similar images. As a summary, the major contributions are as follows.

(i) Introduce the CNN framework and some classical models for mobile visual search.

(ii) Based on the CNN model, an empirical investigation is implemented to inspect the issues that may affect the performance of mobile visual search.

The rest of this paper is organized as follows: section 2 discusses related work of this paper. Section3 introduce the CNN network and the classical models for mobile visual search. Section 4 implement a CBIR system and present the empirical investigation. Additional conclusions are given in section5.

Fig.1 the Top-5 image retrieval results in Google between the high-definition image and the image taken by visitors.

2. RELATED WORK

In this section, I review previous work that is related to mobile visual search, content-based image retrieval and image feature extraction.

2.1 Mobile Visual Search

Mobile devices have blossomed out into having strong images and video processing capabilities. And smart mobile devices are also equipped with high-resolution cameras, fast hardware and intelligent algorithm. All this make a new class of applications possible to use the phone to launch search queries about objects the user is visually close to.[3]. Mobile visual search allows an image used for a search query and returned results are related to this image. Such applications are promising and can be used for identifying objects, online shopping, real estate or artworks. And there are already some

commercial applications, such as Google Images Mobile, Mobile Tao Bao, Bing Mobile and so on. However, this kind of application are usually migrated from desktop to mobile devices while the query images captured by mobile may need specific optimization for mobile devices. To improve the mobile visual search, some compression and coding design have been proposed in paper [4] and [5]. In this way, the amount of data transferred from mobile devices to the search server can be reduced. So that it can provide more efficient indexing and searching on the search server. The method encodes directly the images taken by mobile and ignore the issues appeared in captured photos which may have significantly influence on the performance.

2.2 Content-based Image Retrieval

Content-based image retrieval (CBIR) is a process of retrieving similar images based on content similarity for a given query image. Image content refers to its visual characteristics and the mathematical representation of digital image. Image retrieval task mainly rests with image feature extraction and similarity measurement between feature vectors [6]. And the performance of the task not only depends on the optimum characteristics extraction, but also on the right selection of similarity matching (distance measures). CBIR system has been widely studied for decades. It meets new challenges due to swift growth in image details and complexity.

2.3 Image Feature Extraction

Among CBIR development phases, feature extractions play an important role. Common image features include color feature, texture feature, shape feature, spatial relationship feature such as SIFT[7], DoG[8], and SUFT[9]. Another important image feature is machine-learned feature. For example, Convolutional Neural Network (CNN) shows good performance due to the ability of deep learning.[10]. The paper[11][12] use CNN networks features to complete arts image classification and retrieval. These methods assume the input images are perfect with high quality and low distortion. For distortion images, data augmentation can improve CNNs ability, but it also reduce the network feature representation performance and take up some of the network capacity. A spatial transform network may solve the rotation issues, but it can only handle the rotation input problem. SIFT features has good performance to deal with scale, illumination and scale issues, but it is not enough for captured images.

In summary, for CBIR system based on CNNs in mobile visual search, it is important to investigate how the issues from captured images may affect the performance of image retrieval. We may be able to get some inspirations from the results and propose some solutions.

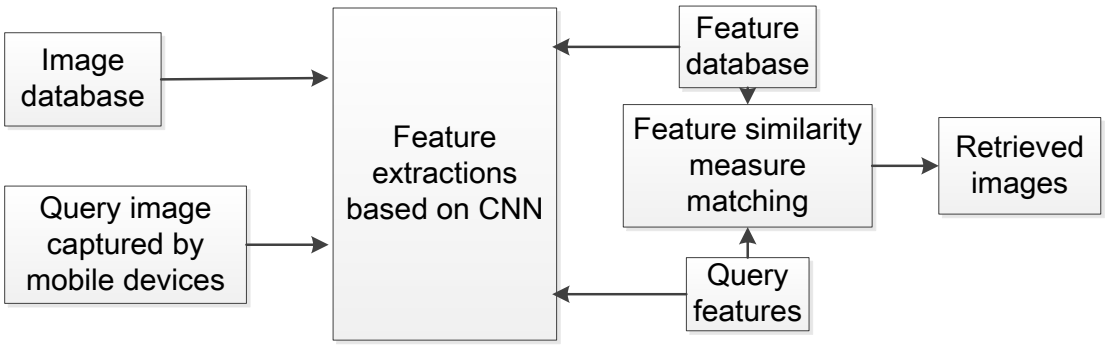


Fig.2 mobile visual search framework

3. MODEL ARCHITECTURE

To illustrate the merit of vanilla CNN model, I introduce the overall framework of CNN model and the popular outstanding models in this section. Then I state two important parts of CBIR system.

3.1 Overall Framework

The mobile visual search framework is shown in Fig2. The core of this framework consist of feature extraction and feature similarity measure matching. In this section, I firstly introduce the concepts of CNNs and then focus on the feature representation based on the trained deep CNN models. Then I talk about the similarity measure matching.

3.1.1 The concepts of CNNs model

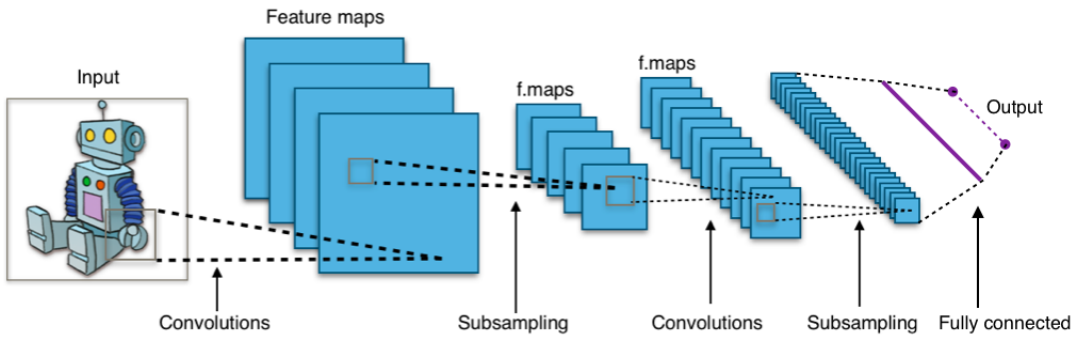


Fig.3 typical CNN architecture [23]

Convolution Neural Network (CNN) is a kind of feedforward Neural network, within range of its artificial neuron can be part of the response around the unit. CNNs are suitable for processing huge image dataset and have really surprising performance. CNN is very similar with normal Neural network, which are made by a study of weights

and bias of constants (biases) neurons every neurons receive some input and do some dot product calculation. The default output is each classification marks. The default input of CNNs is image which can make us code the nature of the specific into the network structure. It makes our feedforward function more effective and reduce lots of arguments. Fig3 shows a typical CNN architecture. The CNNs usually include several layers: Convolutional layer, Rectified Linear Units layer (ReLU layer), Pooling layer and Fully-connected layer.

(1).Convolutional layer. Each layer of the convolutional layer is composed of multiple convolution units, and the parameters of each convolution unit are optimized by the back-propagation algorithm. The purpose is to obtain various features of input. The first convolutional layer may only obtain certain low-level features, such as edges, lines and angles, etc. The multi-layer network can repeatedly obtain more complicated features from low-level features.

(2) ReLU layer. This layer neural activation function using Rectified Linear Units (ReLU). The nonlinearity of the decision function and the whole network is supplemented without influence on the receiver domain of the convolutional layer. [13]

(3) Pooling layer: pooling or down-sampling is used to reduce feature graphs. The pooling operation is independent for each depth slice, and the size is usually 2 times 2. Compared with the convolution layer, the operations in the Pooling layer are generally as follows: 1. Max Pooling is to take the maximum value of 4 points, which is the most popular pooling method. 2. Mean Pooling is to take the average value of 4 points.

(4) Fully-connected layer. All the local features are combined into a global feature, which is used to calculate the score mark of the last category.

3.1.2 CNNs Models

Various architectures of convolutional networks were developed over these years. For example: AlexNet, VGGNet and ResNet.

1. AlexNet

AlexNet is an early application of deep network on ImageNet, and its accuracy is greatly improved compared with the traditional method[14]. AlexNet uses the ReLU activation function instead of the Tanh or Sigmoid activation functions used in the early days of traditional neural networks. The advantage of ReLU over Sigmoid is that its training speed is faster, because the derivative of Sigmoid is very small in the stable region, so the weight is basically not updated. This is the gradient disappearance problem. Therefore, AlexNet uses ReLU behind the convolutional layer and the full connection layer. Another feature of AlexNet is that it can reduce the overfitting problem of the model by adding Dropout layer after each full connection layer. The dropout is effective because neurons are randomly selected. And the interdependence between neurons can be reduced to ensure the extraction of important features that are mutually independent.

2. VGG16

VGG16 was proposed by Oxford University's VGG group.[15] Compared with AlexNet, the improvement of VGG16 is to use several consecutive 3 x3 convolution

kernels instead of the larger AlexNet convolution kernels (11 x11, 5 x5). For a given acceptance field (partial size of input images related to the output). The accumulation of small convolution kernels is better than using convolution kernels, because multiple nonlinear layer can increase the depth of the network to guarantee learning more complex patterns. And the price is relatively less. The top-5 accuracy on ImageNet is 92.3%. Although VGG perform well on ImageNet, VGG has a high memory and time requirement. And VGG is not efficient due to the great deal of of channels in the convolutional layer.

3. ResNet

As the deepness of the network increases, the precision of the network should improve synchronously, overfitting problem should be noticed. However one of the problems with increasing network deepness is that these additional layers are signals of parameter updates. The gradient is transmitted from back to front, when you increase the deepness of the network, the gradient of the higher layer will be very small, which means that the learning of these layers is basically stagnant, and this is the gradient disappearance problem. The second issue of deep networks is training. The deeper the network, the larger the parameter space and the more difficult it is to optimize the problem. Therefore, merely increasing the network deepness will lead to higher training errors. The residual network[16] designs a residual module that allows us to train deeper networks. Residual learning can solve the degradation problem. For an overlay structure, when the input is x, the characteristic it learns is H(x), and now it is expected to learn the residual characteristic F(x)=H(x)-x. So in fact the original learning feature is F(x)+x and the residual learning is easier than learning the original features directly. When the residual error is 0, at this time, the additive layer only does identity mapping, at least the network performance will not be degraded. In fact, the residual error will not be zero, which will also make the additive layer learn new features based on the input features, so as to have better performance. Fig4 shows the residual learning architecture, which is a shortcut connection. Fig5 shows that ResNet has a very good performance in ImageNet. However, how well is it doing in Mobile Visual Search?

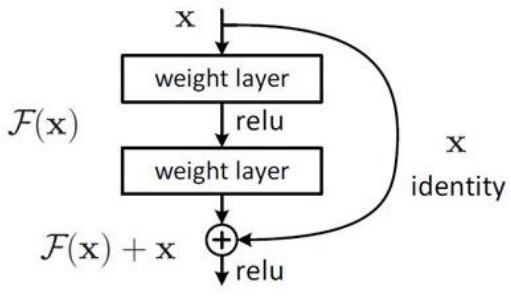


Fig.4 residual learning: a building block

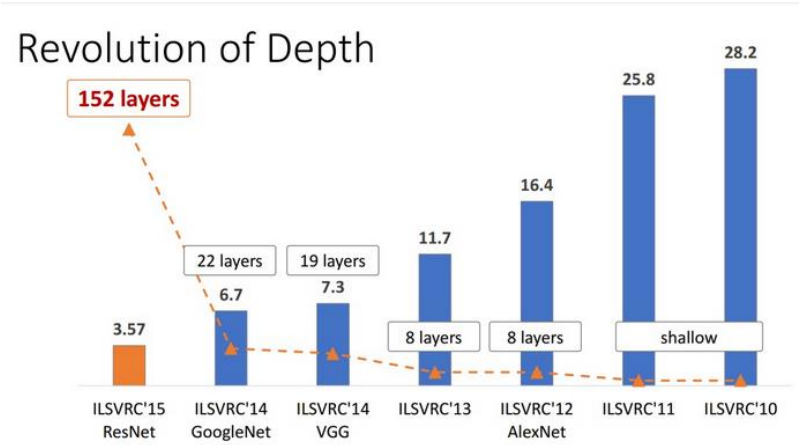


Fig5 ImageNet classification top-5 error @ ILSVRC 2015 competition

3.2 Feature Extraction Model

The content representations presented are generated on the ResNet50 network in this paper. And after the processing in ResNet50, the returned feature of each image includes 2048 dimensions as shown in Fig6. This network can be trained to perform object recognition. When this network is trained to do object identification, the output represents the image content features. And the content features are more accurate with the deeper layers.

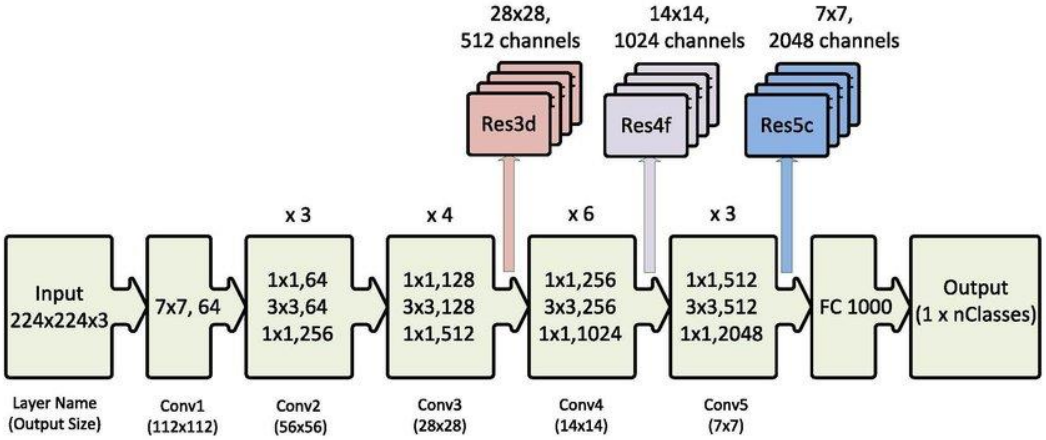


Fig.6 ResNet50 architecture [24]

3.3 Similarity Measure Matching

The performance of the CBIR process not merely depends on the optimal features obtained from the input, but also on the correct selection of similarity matching (distance measures). In common, distance metrics includes the statistical distance metrics such as Cosine Similarity, Chi-square, Geometrical distances such as Manhattan and Chebyshev, Cumulative statistical metrics and so on. Among all these similarity measures, statistical distance metrics shows good mean Average Precision (mAP) scores for all the types of query images. Due to complexity and efficiency consideration, this paper adopt Cosine Similarity to perform similarity measures

matching.

4 EMPIRICAL INVESTIGATION

In this section, I conduct several experiments on the art dataset to evaluate the performance of different issues that may occur when people take photos. To evaluate the mobile visual search I firstly introduce the implementation details about this empirical investigation. Then classify the problem to different issues. Finally based on evaluation methods, I conduct the experiments on the different issues and compare their performance based on CNNs.

4.1 Implementation Details

In the following experiments, A CBIR system is built for mobile visual search based on a pretrained ResNet50 model to compute the content features and use Cosine Similarity measure the distance between query image and the image database. The ResNet50 is pretrained on ImageNet which is a large dataset with over 15 million tagged high-resolution images in approximately 22,000 categories.
The dataset that used for the query dataset include 43,453 visual arts images. And the input query images are taken by the mobile with various issues that may affect the performance of image retrieval based on CNNs.

(a) Illumination (b) Blur (c) Rotation (d) perspective (e) Scale

Fig.7 common issues: (a) The illumination on art works (b) The motion blur in images (c) The visual art is rotated in images (d) The viewing angle is different (e) The area ratio of the art work is too small

4.2 Issues

Arts and paintings are always hung on the walls of gallery, museum or public spaces. When people want to query the relative information about the visual arts, the quality of this image are likely influenced by environment or the capture process

behavior. In this investigation, my partner and I collect large of numbers of photos of art paintings. After observation and thinking, I summarize five issues people may meet when they take such photos. As shown in Fig.7, 7(a) and 7(b) shows that the image may be influenced by environment illumination and the image may be blurred if the mobile devices are hold unstably. Fig.7(c) shows that the image may experience rotation with different angles. And when people take photos from their view, the image may have different perspective as shown in Fig.7(d). The artworks may also have different size in images because of the distance from people to walls or focal length of camera.

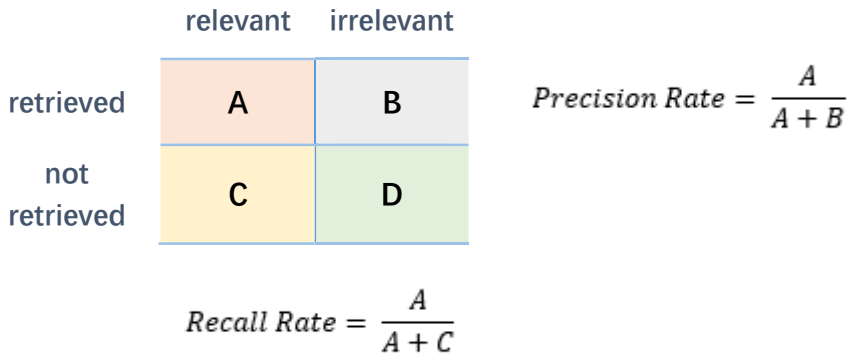


Fig.8 the definition of Precision and Recall

4.3 Evaluation Method

The evaluation metrics in this experiment mainly include two metrics: Precision-Recall curve and mAP. Fig.8 shows the definition of Precision and Recall rate. mAP is to solve the limitation of single point value of Precision and Recall. And it can reflect the global performance. To evaluate the performance of different issues, I will first calculate the Precision-Recall curve on the Top-50 retrieved returned results. Then I conduct experiments on the mAP at Top8 retrieved images on different issues.

4.4 Quatitative Analysis Results

In this part, I give a quantitative analysis about how the five issues and their levels may impact the performance of mobile visual search. The dataset, as stated above, contains 43,453 visual arts images. And their content feature are extracted based on ResNet50 for the retrieval process. To simulate real condition when people use mobile visual search, my partners and I take photos of printed visual arts hung on the wall and categorize the images into five issues with different extent. For illumination issue, I focus on the illumination on the object because it can directly influence the important content on the art images. For the blur issue, I try to take the images with motion blur. In practice, image rotation may not be so severe, so I pick 20, 40 60 and 90 angles to simulate rotation issue. And when people take images, the angle of view cause the perspective issue, and we categorize it into 3 levels. As for scale, I choose area ratio 15%, 30% 50% and 90% to study. The feature extraction method is based on the

pretrained ResNet50 model, and the similarity measure use Cosine Similarity. The performance is measured by mAP which is calculate based on Top-8 returned images about random selection 600 images in each category. The original image is retrieved and its representation is treated as the ground truth (GT), then I conduct a search using images from different category respectively and calculate the mAP based on GT. The result is shown in Fig.9.

Fig.9(a) shows that the area ratio of images has a positive relation with the performance. When the area ratio reaches 90%, the mAP is close to GT which means a quite good performance. It is hard to define the level of blur issue, so the mAP value is calculated through all the blurred images. As shown in Fig.9(b), it can be observed that motion blur has severe impact on the performance of image retrieval. It means the feature of blurred image extracted by the CNNs has large variation. Therefore the blur issue is a critical problem in mobile visual search. In Fig.9(d), it can be observed that the effect of illumination on performance is usually insignificant. The reason may be that the illumination has small influence on the content features. For rotation issue shown in Fig.9(c), it is observed that the rotation issue has obvious negative impact on the performance. The large angle of rotation increase may cause the dramatically decrease of mAP value. For the perspective issue, there are 3 levels of viewing angle. As shown in Fig.9(e), we can find the heavier the degree, the worse the effect.

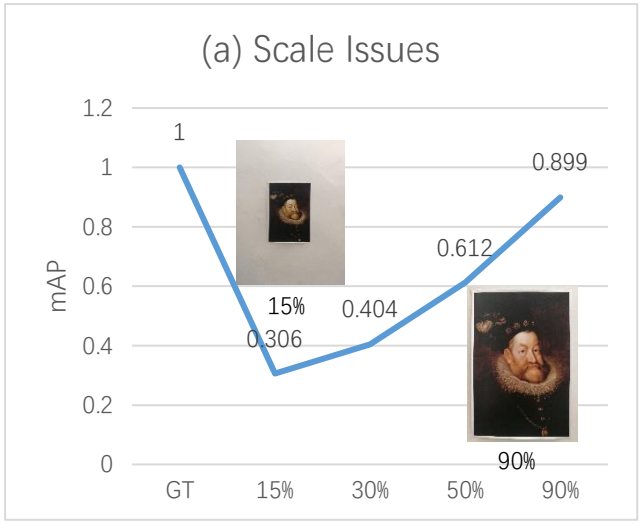
In Fig.11, and the precision-recall curve shown in Fig.10 It can be observed that the comparison between the five issues visually. We found that illumination and scale of 90% area ratio are not particularly serious. Rotation issue, blur issue and perspective issue has much influence on the performance.

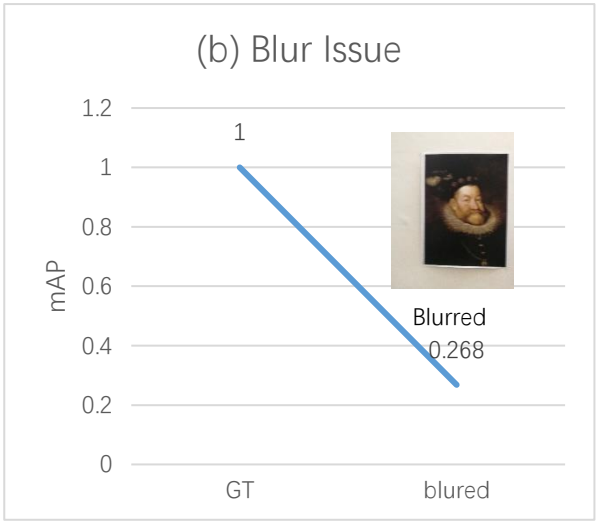
4.5 Discussion

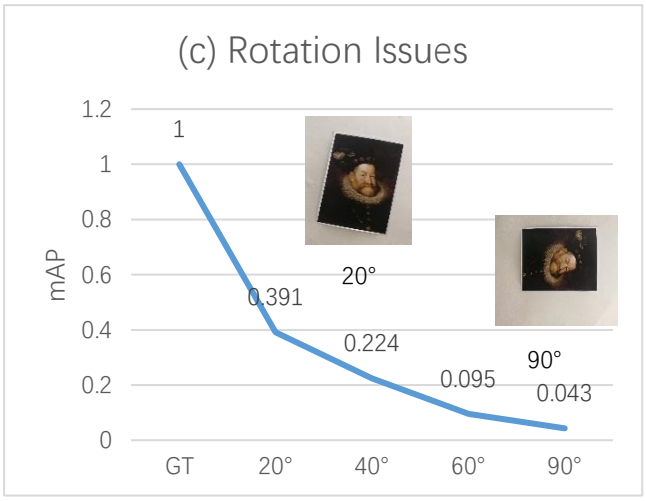
From the quantitative analysis above, I have illuminated how the issues affect the performance on mobile visual search. In this subsection, I try to analyze why this issues based on CNN model have such influence and how to improve.

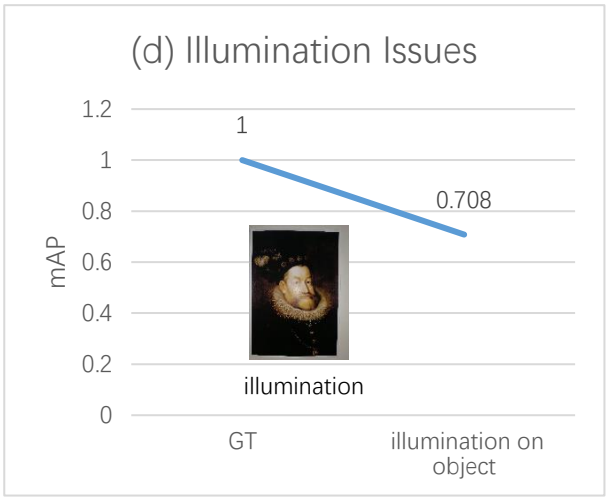
(1) For the scale issue, it is easy to understand that the lower the proportion of visual arts in the image, the worse the retrieval accuracy will be. And it can be solved by adding a locator based on edge detection.

(2) For rotation issue, CNN has translation invariance due to the global Shared weight and pool operation. Input the translated image and the result will generate the corresponding feature map, but it is not so easy for the rotated image. Input a rotated image into CNN, the feature vector may not rotate in a meaningful or predictable way because the pooling layer inherent structure cannot effectively handle excessive shape changes and rotation. To solve this issue, the spatial transformation network is proposed in paper[18] which can be inserted into the existing convolutional architecture to enable the neural network to actively transform feature









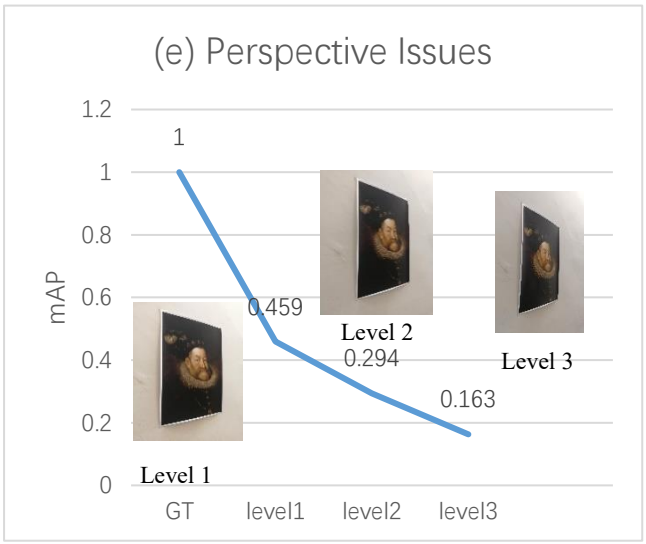


Fig.9 The mAP value of different issues (a) The area ratio from 15% to 90%; (b) Blur issue; (c) Rotation angle from 20° to 90°; (d) Illumination on object; (e) Viewing angle changes within 3 levels.

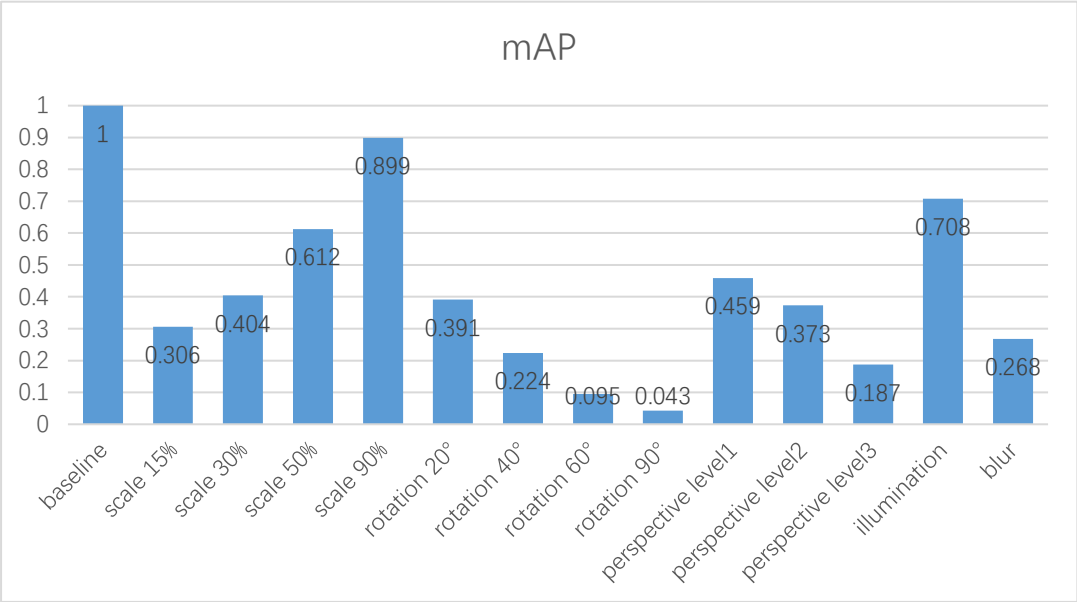


Fig.10 mAP value of all issues and their different degrees

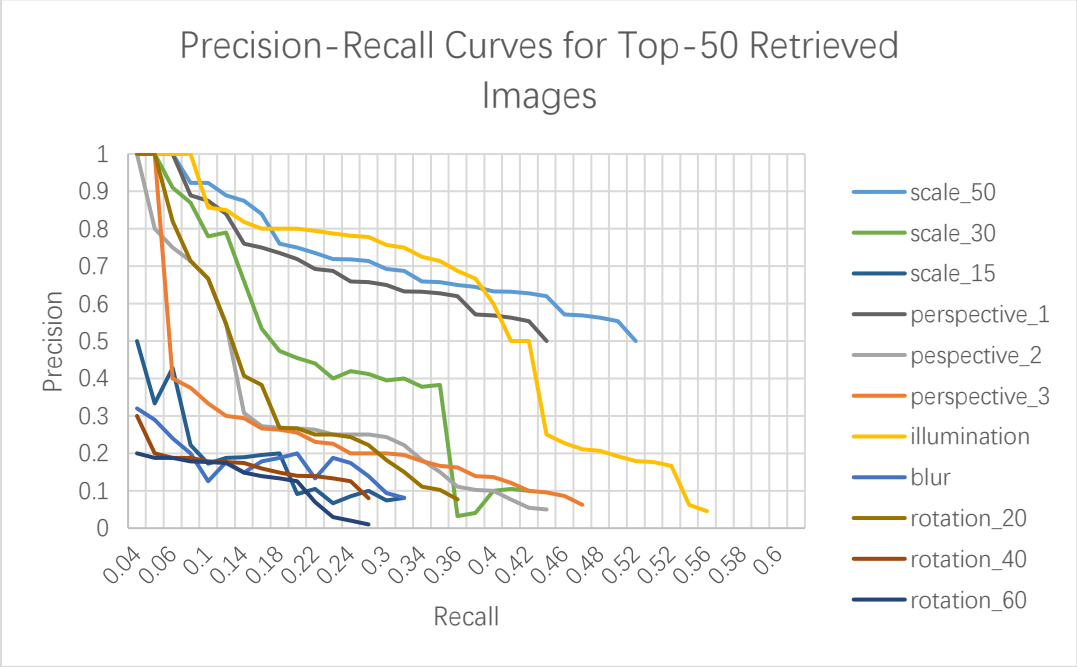


Fig.11 Precision-Recall curves for Top-50 retrieved images

maps in space. And Harmonic Networks, an improved neural network, can clearly identify the graphics and boundaries of the translation and rotation. The equivariance is much better than traditional CNN. The idea is replacing the original CNN filter with a controllable circular harmonic.

(3) For perspective issue, The modern CNN architecture does not follow the classical sampling theorem and cannot guarantee the universality. So the viewing angle issue can be solved in user interface design.

(4) Illumination issue actually has small influence on the performance.

(5) The mAP of blurred images decrease drastically, I think we can recover the blurred image first before input them to CNNs. For example, scale-recurrent network for

deep image deblurring[20] and blind image deblurring using dark channel prior[21]. In paper[22], the author proposed a deblurring generator and trained it to rectify the motion blur.

In theory, when the image is shifted, scaled and deformed, it has no effect on feature extraction. However, modern CNN generally includes subsampling operation, which is the commonly known as the down sampling layer or pooling layer. Its original intention is to improve the translation invariance of the image and reduce the parameters, but its performance is really very general. Another factor is that Cifar-10 and ImageNet have a large number of "photographer bias" in their images. Therefore, the performance on vanilla CNN model for mobile visual search is ordinary or even poor on some issues.

5 CONCLUSION

In this paper, I conduct a CBIR system based on ResNet50 of CNN model to explore the issues that occur on the mobile visual search. The investigation shows that in mobile visual search, the performance on the CNNs is not as good as we though. The captured image has severe influence on the performance. Among these issues, rotation, motion blur, small area ratio and high viewing angle all bring the negative influence in different degrees. Future research directions will go towards improving the existing CNN network to make it more applicable to different data sets, such as the data collected by people themselves.

ACKNOWLEDGEMENTS

Upon the completion of this report, I am grateful to those who have offered many couragement and support during the course of my study. First, special acknowledgment is given to my respectable supervisor James SHE, whose patient instruction and constructive suggestions are beneficial to me a lot. Second, particular thanks go to Pc. Ng, Hui Mao and Ching Hong Lam who have taught me for their instruction and generous support during this course.

REFERENCES

[1] Deng L. A tutorial survey of architectures, algorithms, and applications for deep learning[J]. APSIPA Transactions on Signal and Information Processing, 2014, 3.

[2] Wan J, Wang D, Hoi S C H, et al. Deep learning for content-based image retrieval: A comprehensive study[C]//Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 157-166.

[3] Girod B, Chandrasekhar V, Chen D M, et al. Mobile visual search[J]. IEEE signal processing magazine, 2011, 28(4): 61-76.

[4] Zhang Z, Li L, Li Z, et al. Mobile Visual Search Compression with Grassmann Manifold Embedding[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2018.

[5] Hsiao J H, Li J. Mobile visual search using deep variant coding: U.S. Patent Application 14/715,246[P]. 2016-11-24

[6] Pasumarthi N, Malleswari L. An empirical study and comparative analysis of Content Based Image Retrieval (CBIR) techniques with various similarity measures[J]. 2016.

[7] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International journal of computer vision, 2004, 60(2): 91-110.

[8] Pei S C, Chen L H. Image quality assessment using human visual DOG model fused with random forest[J]. IEEE Transactions on Image Processing, 2015, 24(11): 3282-3292.

[9] Huynh-Kha T, Le-Tien T, Ha-Viet-Uyen S, et al. The efficiency of applying DWT and feature extraction into copy-move images detection[C]//2015 International Conference on Advanced Technologies for Communications (ATC). IEEE, 2015: 44-49.

[10] Shah A, Naseem R, Iqbal S, et al. Improving CBIR accuracy using convolutional neural network for feature extraction[C]//2017 13th International Conference on Emerging Technologies (ICET). IEEE, 2017: 1-5.

[11] Crowley E J, Zisserman A. In search of art[C]//European Conference on Computer Vision. Springer, Cham, 2014: 54-70.

[12] Matsuo S, Yanai K. CNN-based style vector for style image retrieval[C]//Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval. ACM, 2016: 309-312.

[13] Romanuke, Vadim (2017). "Appropriate number and allocation of ReLUs in convolutional neural networks" (PDF). Research Bulletin of NTUU "Kyiv Polytechnic Institute". 1: 69–78. doi:10.20535/1810-0546.2017.1.88156. Retrieved 17 February 2019.

[14] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[15] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[16] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[17] Pasumarthi N, Malleswari L. An empirical study and comparative analysis of Content Based Image Retrieval (CBIR) techniques with various similarity measures[J]. 2016.

[18] Jaderberg M, Simonyan K, Zisserman A. Spatial transformer networks[C]//Advances in neural information processing systems. 2015: 2017-2025.
[19] Worrall D E, Garbin S J, Turmukhambetov D, et al. Harmonic networks: Deep translation and rotation equivariance[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 5028-5037.
[20] Tao X, Gao H, Shen X, et al. Scale-recurrent network for deep image deblurring[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8174-8182.
[21] Pan J, Sun D, Pfister H, et al. Blind image deblurring using dark channel prior[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1628-1636.
[22] Mao H, Cheung M, She J. Deepart: Learning joint representations of visual arts[C]//Proceedings of the 25th ACM international conference on Multimedia. ACM, 2017: 1183-1191.
[23] Typical cnn.png. [Online] Available: https://commons.wikimedia.org/wiki/File:Typical_cnn.png
[24] Mahmood A, Bennamoun M, An S, et al. Resfeats: Residual network based features for image classification[C]//2017 IEEE International Conference on Image Processing (ICIP). IEEE, 2017: 1597-1601.