Image quality classification algorithm based on InceptionV3 and SVM

Yu Li[1], and *Lizhuang* Liu[2,*]

[1]School of Mechanical and Electrical Engineering and Automation, Shanghai University, Shanghai, China
[2]Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai, China

Abstract. In this work we investigate the use of deep learning for image quality classification problem. We use a pre-trained Convolutional Neural Network (CNN) for image description, and the Support Vector Machine (SVM) model is trained as an image quality classifier whose inputs are normalized features extracted by the CNN model. We report on different design choices, ranging from the use of various CNN architectures to the use of features extracted from different layers of a CNN model. To cope with the problem of a lack of adequate amounts of distorted picture data, a novel training strategy of multi-scale training, which is selecting a new image size for training after several batches, combined with data augmentation is introduced. The experimental results tested on the actual monitoring video images shows that the proposed model can accurately classify distorted images.

1 Introduction

A large number of surveillance cameras are distributed in the public areas of our daily life. The digital images captured by the camera are transmitted through a wired bandwidth channel, and are susceptible to human and natural factors so as to distort images after a series of processes such as acquisition, compression, processing, transmission and display. The detection of the distortion type of the video image is of great significance to the maintenance of video surveillance system. The current detection methods are mostly based on appropriate handcraft features for specific distortion types. Wang et al proposed a connected-domain pixel ratio algorithm for signal loss anomaly detection and a detection algorithm of Definition of Exception (DOE) [1]. Cheng designed an average gradient detecting algorithm based on modified focus window method to detect blur images and a covariance correlation function based color cast detecting algorithm [2]. Xia et al. presented a quality diagnosis algorithm based on median filtering and frame differential method using the image features of grayscale and gradient value [3].

Convolutional Neural Network is one of the most effective methods to solve image classification problems. Unlike traditional machine learning methods which require handcraft features, CNN can automatically learn the characteristics of data, greatly reducing the dependence of feature selection. Wu et al. introduce a CNN based image classifier

* Corresponding author: liulz@sari.ac.cn

which took small patches segmented from video images as input and introduced the positive and negative sample equalization and an adaptive learning rate to reduce overfitting. Meanwhile, the class of an image is decided using the majority voting rule [4]. In this work we focus on the more complex image quality classification problem in which more distortion types are considered and of which the model is tested directly on monitoring images collected in daily life. The following aspects are studied: 1) Introduce a new image quality classification algorithm combined with convolutional neural network and SVM; 2) Comparing the ability of different deep CNN architectures to extract image quality features; 3) Analyze the features extracted from different processing units in deep convolutional neural networks; 4) Propose a training strategy of multi-scale training to improve the generalization ability.

2 Image quality classification algorithm

The framework of the proposed algorithm for distortion image detection problem is shown in figure 1. We use the pre-trained CNN model on the ImageNet dataset as the base model, and fine-tune it with the surveillance image dataset. The fine-tuned CNN model has two purposes: 1) The model without the final layer is employed as an image describer, from which the features extracted is used as the training data for a SVM model, with Bayesian algorithm to optimize the characteristics. The trained SVM model is used as the image quality classifier; 2) The model with the softmax layer as the final layer is directly trained as the classifier for comparison.

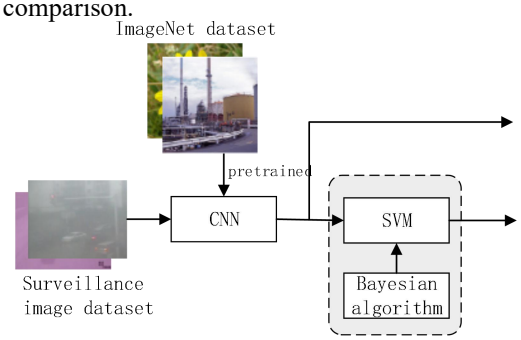


Fig. 1. Flow chart of the proposed image quality classification algorithm.

2.1 Image description using CNN

The generic descriptors extracted from the CNN model are of remarkable generalization capabilities when the model is trained on large scale labeled datasets, and outperforms descriptors of hand crafted, state-of-the-art systems in many image classification tasks. The classic CNN models such as VGGNet [5], GoogLeNet [6], ResNet [7], etc. have shown great classification effects. In this paper, we adopt the InceptionV3 model as the feature extractor, because of its outstanding performance and good generalization ability.

In order to improve the performance of deep CNN model, the common strategy is to increase the depth and width of the network, but this will increase the number of network parameters, which will lead to the over-fitting problem and requires a large amount of computing resources. To solve this problem, the GooLeNet model introduces the Inception architecture, of which the main idea is to approximate the optimal local sparse structure through dense components, so as to maintain the sparseness of the network structure and the high computational performance of the dense matrix. The core of the Inception architecture is three Inception modules, the structure of which is shown in figure 2. The procedures of the Inception modules are as follow: dimensions (or the input channel

numbers) of the feature map of the previous layer are reduced by the first 1x1 layers; then, extract the features through the 1x1, 1xn, 3x3 convolutional layers; the last layer is the Filter Concat layer, consisting of an LRN (Local Response Normalization) layer and a DepthConcat layer, wherein the DepthConcat layer fuses the features extracted by convolutional layers. The number of feature channels will increase as the network deepens, enlarging the proportion of convolution kernels in the Inception architecture. If large convolutional kernels such as 3x3,5x5 are selected, the amount of computation will increase significantly. Therefore, the convolutional layers are mostly small convolutional kernels such as 1x1,1xn, the effective receptive fields of which may enlarge to capture global picture characteristics by stacking multiple convolutional layer, meanwhile which will reduce the compute. The Inception architecture also uses an average pooling layer instead of a fully connected layer, and add two auxiliary softmax layers for the forward propagation to avoid vanishing gradient problem.

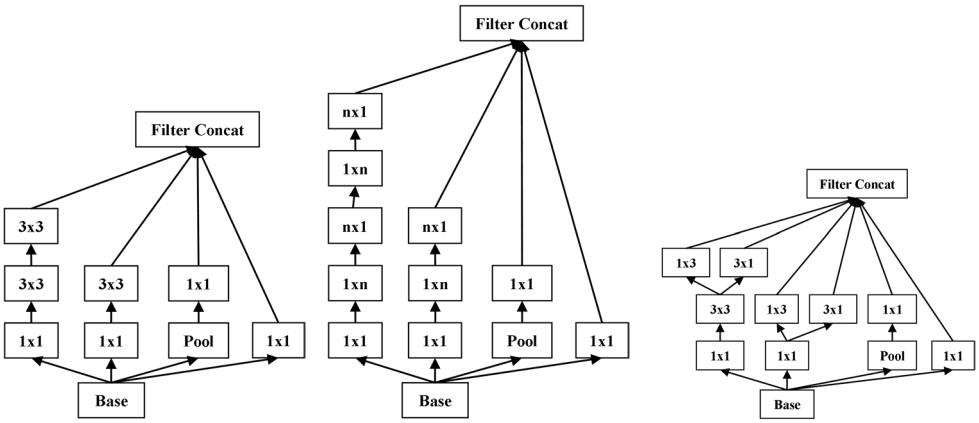


Fig. 2. The Structure of three Inception modules.

The very deep CNN models trained in large datasets offer a powerful classification capability, however, the effect will be limited by a lack of adequate data such as distorted surveillance images [8]. In this paper, the fine-tuning method is introduced to copy with the problem [9]. Fine-tuning has been applied in a variety of different medical imaging modality [10] and image quality evaluation tasks [11], the result of which shows fine-tuning is as effective as training a CNN from scratch while being more robust to the small training data. Compared with directly training a CNN model with small datasets, the fine-tuned model is more robust and is more easy to train. The CNN model we used were pre-trained on the ImageNet natural image dataset. We adapted the CNNs to our problem by replacing the top layers from the last global average pooling layer to the softmax layer with a new layer for the 11 classes in our dataset. The initial CNN weights derived from the ImageNet dataset were fine-tuned through back-propagation so that they can better reflect the modalities in the surveillance image dataset. We used mini-batch stochastic gradient descent to find the optimal weights.

2.2 Multi-scale training

The performance of a deep CNN model generally depends heavily on the size of the available training datasets. However, the collected available surveillance image data is insufficient. Common strategies for attacking this problem are data augmentation technique, which seek to multiply the dataset by rotating, cropping, reflection, and so on. While

generating more picture content is simple, ensuring adequate distortion diversity and realism is the harder thing which should be taken into account.

We also resize pictures to generate different size of images not only to augment the dataset but also to make the model robust to running on images of different sizes. Instead of fixing the input image size, we randomly choose a new image dimension size to train the CNN model after several iterations. This regime forces the network to learn to predict well across a variety of input dimensions, making the model capable of predicting at different resolution.

2.3 Support vector machine

Support vector machine is a machine learning method proposed based on the structural risk minimization principle by Vapnik et al [12], which has good generalization ability and is suitable for small sample classification problems. The idea of the SVM is to find the optimal margin separating hyperplane with the maximum classification boundary. In this paper, the SVM is applied to classify the features extracted by the CNN model. In this paper, there are multiple image distortion type need to be detected. The SVM model is adopted with the one-against-one method for the multiple classification problem [13]. The idea of the method is to train a binary classifier between every two kind of image samples. While testing, the label of the test sample is given by these classifiers and the sample is classified into the class with the most votes. Therefore, $k(k-1)/2$ classifiers are constructed where k is the number of classes. For training data from the i th and the j th classes, we solve the following binary classification problem:

$$\begin{aligned} \min_{w^{ij},b^{ij},\xi^{ij}} \quad & \frac{1}{2}\| w^{ij} \|^2 + C\sum_t \xi_t^{ij} \\ s.t. \quad & (w^{ij})^T \phi(x_t) + b^{ij} \geq 1 - \xi_t^{ij}, \ if \ y_t = i \\ & (w^{ij})^T \phi(x_t) + b^{ij} \leq -1 + \xi_t^{ij}, \ if \ y_t = j \\ & \xi_t^{ij} \geq 0. \end{aligned} \tag{1}$$

where ξ is the slack variable, C is the penalty parameter, and $\phi(x_t)$ is kernel function, which can implicitly map the input feature space to a high-dimensional space, making the features easier to classify without increasing the computational complexity, and effectively avoiding the dimensional disaster. Commonly used kernel functions are as follows: linear kernel, polynomial kernel, radial basis function (RBF) kernel and sigmoid kernel function. Then the decision function is

$$f_n = \mathrm{sgn}((w^{ij})^T \phi(x) + b^{ij})) \tag{2}$$

where f_n is the result of the n th classifier. Input the test sample X_{test} to these classifiers and calculate the votes. Given the numbers of votes of s th class is $\mathrm{D}_s(\mathrm{X}_{test})$, the final classification result is

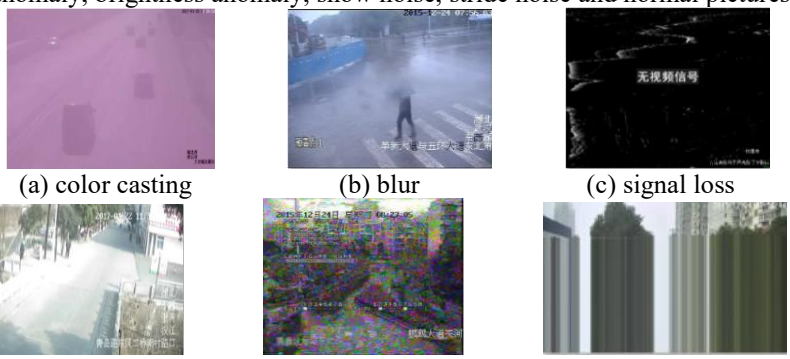
$$f = \arg\max_{s=1,\ldots,\mathrm{k}(\mathrm{k}-1)/2} \mathrm{D}_s(\mathrm{X}_{test}) \tag{3}$$

We select the RBF function as the kernel function with two hyper-parameters ξ , C . The Bayesian optimization algorithm based on Gaussian process is applied to find the optimal hyper-parameters, where the Gaussian process(GP) is used to model the distribution of the objective function and the Expected Improvement(EI) function is used as the acquisition function to trade off the mean and the variance, the optima are located where the uncertainty in the surrogate model is large and/or where the model prediction is high [14]. Bayesian optimization then select the next query point by maximizing such

acquisition functions. By iterating this process, the GP and the hyper-parameters are updated until the optima are found.

3 Experiments

In the experiments, the surveillance image dataset collected in daily life is utilized, consisting of 7 kinds of natural images shown in figure 3, such as color casting, blur, luminance anomaly, brightness anomaly, snow noise, stride noise and normal pictures.



(a) color casting (b) blur (c) signal loss

(d) luminance anomaly (e) snow noise (f) stride noise

Fig. 3. Distorted surveillance images.

The dataset contains 2355 distorted and normal images with different resolutions and is randomly divided into two parts, 80% of which is the training set and 20% as the test set to evaluate the proposed image quality classification model. We adopt three indices as the criteria for evaluating the performance of the surveillance image metrics, which are accuracy, missed diagnosis rate(MDR) and wrong diagnosis rate(WDR). The consistent result means that the classification result of a picture matches its label. If a distorted picture is recognized as a normal image, the classification result is recorded as miss diagnosis. If the type of a distorted picture is misjudged or the normal picture is recognized as a distorted image, the classification result is recorded as misdiagnosis. Given the consistent number a, missed diagnosis number m and wrong diagnosis number w, the accuracy r_a, missed diagnosis rate r_m and wrong diagnosis rate r_w can be expressed as

$$r_a = a / (a + m + w) \tag{4}$$

$$r_m = m / (a + m + w) \tag{5}$$

$$r_w = w / (a + m + w) \tag{6}$$

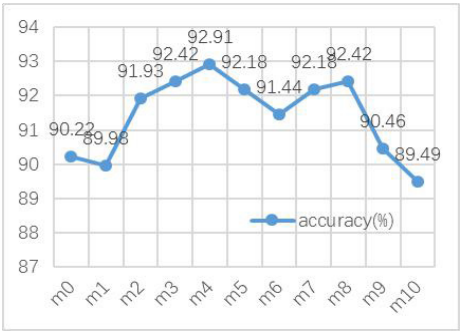
4 Results

Three CNN architecture of VGG16, Inceptionv3, and ResNet50 are all tested in the experiments, and they are trained in the original dataset and the data-augmented dataset respectively. The performance of these models are shown in table 1. Due to the number of the original dataset is limited, the models trained on this dataset cannot effectively extract the image feature, thus, the accuracy of these models are all worse than that of the model trained in the data-augmented dataset. Specifically, the accuracy of the augVgg16 classifier is improved by 13.24%, the augInceptionV3 classifier improved by 14.18 and the augResNet50 improved by 13%.

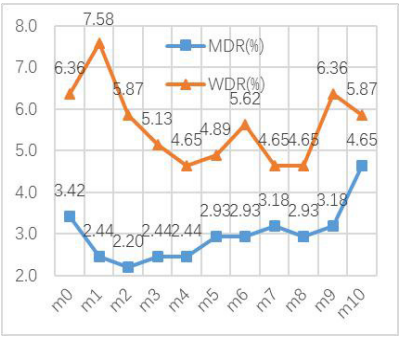
Table 1. Test results of CNN classifiers.

model	accuracy (%)	MDR(%)	WDR(%)
Vgg16	74.29	17.92	7.79
InceptionV3	75.26	16.63	8.31
ResNet50	74.29	16.10	9.61
augVgg16	87.53	4.89	7.58
augInceptionV3	89.24	4.40	6.36
augResNet50	87.29	5.87	6.85

The above results also show that the augInceptionV3 classifier has the best performance among all the CNN models, the features extracted by which can best describe the characterization of the surveillance image dataset. Therefore, we select the augInceptionV3 model as the base model of the proposed image quality classifier.



(a) accuracy



(b) missed/wrong diagnosis rates

Fig. 4. Results of classifiers with features extracted by different Inception modules.

There are 46 layers in the InceptionV3 model, consisting of 11 Inception modules. In order to analyze the influence of different processing units on the performance of the image quality classifier, we output features extracted by every Inception module to train the SVM model. The classification result of different features is shown in figure 4. The results show that the performance of the higher layer is the worst and the middle layers produce the best accuracy. The accuracy of the 5th module which has the best performance is 3.42% higher than that of the 11st module. This might due to the fact that the lower layers of the CNN model focus on the texture features of an image such as color, edge and corner point, and the higher layers captures higher abstraction of image content which is not sensitive to the image quality. On the other hand, as is shown in the results, the classifier combined the InceptionV3 model with the SVM performs better than the simple CNN model.

5 Conclusion

The results testing on the surveillance images yield a solid evidence that the proposal model based on InceptionV3 and SVM is suitable for the image quality classification task, which has very high classification accuracy, low missed and wrong diagnosis rate. Multiscale training and data augmentation technique can greatly improve the performance of the CNN model. We also analyze the effect of features extracted by different Inception modules, the experimental shows that the middle module especially the 5th module can learn the image quality feature better.

References

1. Wang W H, Liu Y S. Image anomalies analysis and detection algorithm for surveillance video[J]. Journal of Central China Normal University (Natural Sciences), 2015, 49(05): 692-695.
2. Cheng C S. Design and implementation of the real-time detection system for surveillance video quality[D]. Zhejiang University of Technology, 2014.
3. Xia Y J, Sun H. Anomaly detection and quality diagnosis of surveillance video[J]. Computer Applications and Software, 2016, 33(06): 163-167+211.
4. Wu M Y, Chen L, Tian J. Video image distortion detection and classification based on CNN[J]. Application Research of Computers, 2016, 33(09): 2827-2830.
5. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
6. Szegedy C, Liu W, Jia Y, et al. & RABINOVICH, A. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2015:1-9.
7. 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.
8. Shin H C, Roth H R, Gao M, et al. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning[J]. IEEE transactions on medical imaging, 2016, 35(5): 1285-1298.
9. Donahue J, Jia Y, Vinyals O, et al. Decaf: A deep convolutional activation feature for generic visual recognition[C]//International conference on machine learning. 2014: 647-655.
10. Tajbakhsh N, Shin J Y, GURUDU S R, et al. Convolutional neural networks for medical image analysis: Full training or fine tuning?[J]. IEEE transactions on medical imaging, 2016, 35(5): 1299-1312.
11. Gao F, Wang Y, Li P, et al. DeepSim: Deep similarity for image quality assessment[J]. Neurocomputing, 2017, 257: 104-114.
12. Vapnik V. Statistical learning theory[M]. New York: Wiely, 1998.
13. Hsu C W, Lin C J. A comparison of methods for multiclass support vector machines[J]. IEEE Transactions on Neural Networks, 2002, 13(4):1026.
14. Shahriari B, SWERSKY K, Wang Z, et al. Taking the human out of the loop: A review of bayesian optimization[J]. Proceedings of the IEEE, 2016, 104(1): 148-175.