Learning Robust Scene Classification Model with Data Augmentation Based on Xception

Haiyan Chen, Yu Yang and Suning Zhang
State Grid Jiangsu Electric Power Limited Company, Suzhou Power Supply Branch
No. 555 Laodong Road, Suzhou Jiangsu 215004, China
Email: 13372175016@189.cn
Email: yyjady@qq.com
Email: 13646207157@163.com

Abstract. Scene classification technology based on computer vision has been widely applied in many fields. However, with the increasing complexity of images, many computer vision classification models are difficult to meet requirements of current scene classification tasks, as they not only require considering the object, background, spatial layout and other information, but also many relationships in the image. Based on the analysis of existing scene classification algorithms and Xception model, an approach that adds optimization from two aspects of data set processing is proposed to solve complicated scene classification tasks. Combined with the image enhancement technology, the serialized image enhancement method is used to expand the dataset and enhance the image features, and takes advantage of the Xception model to extract the image features to obtain the scene classification model with high robustness. The experimental results showed that Xception model was able to deal with scene classification efficiently by making up for the shortcomings of traditional Convolutional Neural Networks (CNN) models in feature extraction and generalization ability.

1. Introduction

With the opening of mobile Internet era, it is easy to get and share pictures, which have become an important medium for people to interact[1]. How to assign a semantic category to an image according to its visual content is not only the target of image scene classification, but also the basis of image retrieval, image content analysis and target recognition[2]. However, due to the diversity of image scale, angle, illumination and the complexity of scene definition, scene classification has always been a challenging problem in computer vision[3]. Therefore, how to train a more robust scene classification model[4] to solve the problem of angle, scale, and illumination diversity of images has practical research and application value.

At present, scene classification refers to distinguishing images with similar scene features from multiple images, and classifying these images correctly. Most of the scene classification methods based on vision are based on objects, and the classification is determined by identifying some representative features, referred as extracting image features[5-6], the key of scene classification. There are currently many classical methods, which can be mainly divided into three categories. The first category is to extract feature descriptors[7] directly from image; the second category is to continue feature extraction based on some underlying features extracted by image segmentation; the third category is through the training of depth network model, where the image features are automatically extracted. Although the first category of methods are simple in steps, they have higher requirements for the extracted feature descriptors due to the limitations of low-level features for the description of scene semantic information; the second category of methods have improved the
classification accuracy compared with the first method, but the processing will be more complex. In recent years, by using deep network is a rising method, because of its advantage that it is unnecessary to extract features manually, and after full training, the classification effect is outstanding.

In this paper, the Xception network model is applied to the feature extraction of high robustness scenes to achieve scene classification. In order to make up for the low recognition rate of traditional scene classification model in complex scene, we add optimization from two aspects of data set processing and model. During dataset processing, in order to deal with the complexity of the scene, a lot of image enhancement efforts will be done, so that the characteristics of each image can be more easily extracted by the model, to meet the data sufficiency problem in complex scene. In terms of model optimization, Xception is chosen as the optimized image classification model, which is trained many times, and the optimal model parameters are selected. The improved algorithm can effectively identify a variety of complex scenes.

2. Relate Works

2.1. Deep Learning

Simonyan et al. proposed Very Deep Convolutional Networks for Large-Scale Image Recognition (VGG)[8]. Unlike AlexNet[9], VGG net uses more, usually 16-19 layers, while AlexNet only has 8 layers. At the same time, all of the convolutionary layers of VGG net use the same size of the convolutionary filter, which is 3×3, making VGG nets more capable of learning features. Szegedy et al.[10] proposed Inception V1, whose structure is the main innovation point.

However, there are some problems in this structure. The filter parameters of each level of inception model are the sum of all branches. After the multi-level Inception, the parameters will be very large, which requires much computing resources. Ioffe et al.[11] proposed Inception V2 which adopts two 3×3 convolutions instead of 5×5 large convolutions, which reduces the parameters and establishes more nonlinear transformations, making Convolutional Neural Networks(CNN)[12] more capable of learning features. Szegedy et al.[13] proposed Inception V3, which decomposes convolution kernel into small convolutions. A larger two-dimensional convolution is divided into two smaller ones, as shown in the figure above. For example, it is more economical to split 7×7 into 1×7 and 7×1 into 3×3. Moreover, the nonlinearity is added to expand the model expression ability. This kind of asymmetric splitting is more effective than symmetrical splitting into several identical small convolution kernels, which can deal with more and richer spatial features. He et al.[14] proposed ResNet, the residual network[15] structure, by introducing a shortcut connection between the output and the input, rather than a simple stacking network, which can solve the problem that the gradient disappears due to a deep network.

2.2. Xception

Xception is another improvement of the Incrption V3 proposed by Google after the concept. It mainly uses the depthwise separate concept[16] to replace the convolution operation in the standard Inception V3. The illustration of the Incrption V3 is shown as Figure 1.

![Figure 1. A canonical Inception module (Inception V3).](image)
The intention [17] of Inception model is that feature extraction and transfer can be achieved by 1×1 convolution, 3×3 convolution, 5×5 convolution, pooling, etc. The concept structure can obtain the best feature extraction method through training, that is to say, provide an input to these several feature extraction methods at the same time, and then do the concatenation. The comparison [18] between Inception V3 and Inception V1 is mainly to replace 5×5 convolution with the superposition of two 3×3 convolutions. Figure 2 is a simplified concept structure from the Inception v3.

![Figure 2. A simplified Inception module.](image)

We can get Figure 3 by extending the network model of Figure 2. For an input, Figure 3 uses a unified 1×1 convolution kernel convolution, and then connects three 3×3 convolutions. These three convolution operations only take part of the previous 1×1 convolution results as their own input.

![Figure 3. A strictly equivalent reformulation of the simplified Inception module.](image)

Chollet et al. [19] proposed Xception to eliminate the correlation between channels from the spatial correlation. The convolution operation in the original concept-v3 is replaced by the separable solution (the extreme concept module) which accelerates the convergence process of Xception and achieves significantly higher accuracy. The potential problem is that although Depthwise Separable [20] revolution can improve the accuracy or reduce the theoretical calculation greatly, due to its scattered calculation process, the efficiency of the existing convolution neural network implementation is not high enough.

3. Method

3.1. Data Preprocessing
Data standardization is a basic work of data processing [21]. Different evaluation indexes often have different dimensions and dimension units, which will affect the results of data analysis. In order to eliminate the dimensional impact between indexes, data standardization is needed to solve the comparability between data indexes. After the original data is standardized, each index is in the same order of magnitude, which is suitable for comprehensive comparative evaluation.

Data standardization formula is shown as:
\[ x' = \frac{x - \mu}{\sigma} \]  

(1)

Data centre formula is shown as:

\[ x' = x - \mu \]  

(2)

In the above formula, \( \mu \) is the mean value, \( \sigma \) is the standard deviation. Through the centralized and standardized processing, the data with the mean value of 0 and the standard deviation of 1 are obtained. For example, in the process of training neural network, by standardizing the data, the convergence of weight parameters can be accelerated. The purpose of data pre-processing is to increase the orthogonality of base vectors.

3.2. Data Augmentation

The purpose of this paper is to train a highly robust scene classification model. According to the different space of the enhancement process[22], the general image enhancement methods can be divided into spatial and frequency-domain based methods. The method based on spatial domain directly processes the image; the method based on frequency domain modifies the transform coefficient of the image in some transform domain of the image, and then reversely transforms to the original spatial domain to get the enhanced image[23]. The main purposes are: first, to improve the visual effect of the image and improve the clarity of the image; second, to highlight some features of interest and suppress the features of no interest for a given image application, so as to expand the differences between the features of different objects in the image and meet the needs of some special analysis.

![Flow chart of Data Augmentation](image)

Figure 4. Flow chart of Data Augmentation

In Figure 4, the process of image enhancement is introduced in detail, which is an end-to-end process. In order to adapt to the complexity of the experimental scenes, the sequence of Data Augmentation is that Horizontally flip 50% of all images and crop some of the images by 0-10% of their height or width; apply affine transformations to some of the images; convert some images into their super-pixel representation; blur each image with varying strength using gaussian blur (sigma between 0 and 3.0); search in some images either for all edges or for directed edges. These edges are then marked in a black and white image and overlaid with the original image using an alpha of 0 to 0.7. Add Gaussian noise to some images. Either drop randomly 1 to 10% of all pixels. Invert each image's channel with 5% probability. Add a value of -10 to 10 to each pixel. Change brightness of images (50-150% of original value). Improve or worsen the contrast of images. Convert each image to grayscale and then overlay the result with the original with random alpha. In some images move pixels locally around (with random strengths). In some images distort local areas with varying strength.

3.3. Xception Model

Xception is completely based on the convolution neural network architecture of the separable revolution layer. In the feature mapping of convolutional neural network, the mapping of cross channel correlation and spatial correlation can be completely decoupled. Because this assumption is an
enhanced version of the assumption based on the assumption of the perception architecture, we name our proposed architecture Xception, which represents "extreme perception".

As an improvement of perception V3, Xception mainly introduces the depthwise separable revolution on the basis of perception V3, which improves the effect of the model without increasing the network complexity. The introduction of depthwise separable revolution can reduce the complexity of the network, because depthwise separable revolution is mainly designed to reduce the complexity of the network in MobileNet. Make the number of parameters similar to that of perception V3, and then compare the performance under this premise. So the purpose of Xception is not to compress the model, but to improve the performance.

The algorithm of Scene Classification Model with Data Augmentation Based on Xception is as follows.

Algorithm: Scene Classification Model with Data Augmentation Based on Xception

Input: Image M, data augmentation $S_{aug}$, Epoch, image size, Xception model

Return: The category of input image

1: divide the image M into n equal parts $M_n$
2: initialize Xception model
3: Repeat
4: $epoch + 1$
5: for $i=1$ to $n$ do
6: preprocessing image dataset by data augmentation

\[ Q_i = S_{aug}(M_i) \]
7: define loss function

\[ loss = different(label_{pred}(Q_i), label(Q_i)) \]
8: fit Xception model with input image $Q_i$
9: Until $epoch >= Epoch$

4. Experiments and Analysis

The dataset for this experiment selects 80000 pictures from 4 million Internet pictures. These pictures belong to 80 daily scene categories, such as terminal, football field, etc. Each scene category contains 600-1100 pictures.

Most deep learning methods deal with database by allocating the database to generate training set, verification set and test set in a certain proportion. In the experiment, the training set and test set are randomly generated according to the proportion of 8:2; in the training verification set, 80% of them are regarded as the training data set and the remaining 20% as the verification data set.

The scene classification model designed in this experiment includes ResNet, Inception, VGG16 and Xception. The parameters of these models can be seen in Table 1.

| Parameter     | Value | Description            |
|---------------|-------|------------------------|
| IMAGE_SIZE    | 299   | Size of input images   |
| BATCH_SIZE    | 32    | Number of pictures selected in one training |
| CLASSES       | 80    | Classes of dataset     |
| EPOCH         | 100   | Time of train all samples |

It can be seen from the experiment that the score of Xception in this dataset after data enhancement is better than that of the other three models in Table 2.
Table 2. The scores of scene classification model, including ResNet, Inception, VGG16 and Xception.

| Model   | Score |
|---------|-------|
| ResNet  | 0.040 |
| Inception | 0.043 |
| VGG16   | 0.033 |
| Xception | 0.051 |

It can be seen from Figure 5 that during the training process, the loss of Xception is significantly lower than that of other models.

Figure 5. The loss of scene classification model, include ResNet, Inception, VGG16 and Xception.

5. Conclusion
For classification tasks in complex scenes, data enhancement is used to increase the robustness of the model and reduce the interference of uncertain factors. The Xception model is used for research and analysis on dataset. Through experimental comparison, it is found that the score of scene classification based on Xception method is better, it has good generalization ability, and it has better robustness for complex scene classification. Because the differences between different image classes are more subtle, it is often only with the help of small local differences to distinguish different categories, so neural network learning these subtle features needs more data. In the future, we will continue to study the combination of data enhancement technology and deep vision neural network to reduce the need for data volume and further the accuracy of scene classification.

6. References
[1] HOSSAIN M D Z, SOHEL F, SHIRATUDDIN M F, et al. A comprehensive survey of deep learning for image captioning[J]. ACM Computing Surveys (CSUR), 2019, 51(6): 1-36.
[2] DOAN T, NGUYEN H, NGO D T, et al. Acoustic Scene Classification Using A Deeper Training Method for Convolution Neural Network[C]/2019 International Symposium on Electrical and Electronics Engineering (ISEE). IEEE, 2019: 63-67.
[3] YAO X, YANG L, CHENG G, et al. Scene Classification of High Resolution Remote Sensing Images Via Self-Paced Deep Learning[C]/IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium. IEEE, 2019: 521-524.
[4] MUN S, SHON S. Domain Mismatch Robust Acoustic Scene Classification using Channel Information Conversion[C]/ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 845-849.
[5] XIE J, ZHU M. Investigation of acoustic and visual features for acoustic scene classification[J]. Expert Systems with Applications, 2019, 126: 20-29.
[6] MINCIULLO L, PARKES M J, FELSON D T, et al. Comparing image analysis approaches versus expert readers: the relation of knee radiograph features to knee pain[J]. Annals of the rheumatic diseases, 2018, 77(11): 1606-1609.
[7] ZIJUN B, YUSHENG X. A method to extract instantaneous features of low frequency oscillation based on trajectory section eigenvalues[J]. Journal of Modern Power Systems and Clean Energy, 2019, 7(4): 753-766.

[8] Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale Visual Recognition Challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.

[9] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. et al. ImageNet classification with deep convolutional neural networks[J]. Communications of The ACM, 2017, 60(6): 84-90.

[10] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition, 2015: 1-9.

[11] IOFFE S, SZEGEDY C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[C]. International conference on machine learning, 2015: 448-456.

[12] HOU J, ADHIKARI B, CHENG J. DeepSF: deep convolutional neural network for mapping protein sequences to folds[J]. Bioinformatics, 2018, 34(8): 1295-1303.

[13] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the Inception Architecture for Computer Vision[C]. computer vision and pattern recognition, 2016: 2818-2826.

[14] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C]. computer vision and pattern recognition, 2016: 770-778.

[15] QIU Y, WANG R, TAO D, et al. Embedded Block Residual Network: A Recursive Restoration Model for Single-Image Super-Resolution[C]. international conference on computer vision, 2019: 4180-4189.

[16] FAN Q, CHEN C, KUEHNE H, et al. More Is Less: Learning Efficient Video Representations by Big-Little Network and Depthwise Temporal Aggregation[C]. neural information processing systems, 2019: 2261-2270.

[17] ZHENG K, GAO L, RAN Q, et al. Separable-spectral convolution and inception network for hyperspectral image super-resolution[J]. International Journal of Machine Learning and Cybernetics, 2019: 1-15.

[18] LIN C, LI L, LUO W, et al. Transfer Learning Based Traffic Sign Recognition Using Inception-v3 Model[J]. Periodica Polytechnica Transportation Engineering, 2019, 47(3): 242-250.

[19] CHOLLET F. Xception: Deep Learning with Depthwise Separable Convolutions[C]. computer vision and pattern recognition, 2017: 1800-1807.

[20] ALFASLY S, HU Y, LI H, et al. Multi-Label-Based Similarity Learning for Vehicle Re-Identification[J]. IEEE Access, 2019, 7: 162605-162616.

[21] GRANNIS S J, XU H, VEST J R, et al. Evaluating the effect of data standardization and validation on patient matching accuracy[J]. Journal of the American Medical Informatics Association, 2019, 26(5): 447-456.

[22] HU Z, TAN B, SALAKHUTDINOV R, et al. Learning Data Manipulation for Augmentation and Weighting[C]. neural information processing systems, 2019: 15738-15749.

[23] MASI I, TRĂN A T, HASSNER T, et al. Face-Specific Data Augmentation for Unconstrained Face Recognition[J]. International Journal of Computer Vision, 2019, 127(6-7): 642-667.