Санкт-Петербургский государственный университет

Программная инженерия

Ван Тяньцзин

Проектирование и реализация системы сжатия изображений на основе свёрточных нейронных сетей

A Design and Implementation of Image Compression on Convolution Neural Network

Отчёт по учебной (ознакомительной) практике

Научный руководитель:

К.ф.-м.н., доцент Луцив Д.В.

Санкт-Петербург

2021 год

1. Research background and significance

1) Research background

With the development and progress of science and technology, the popularization of multimedia application and the inevitable trend of digitalization, people have a high demand and strict requirements for the network transmission of images or video in production and life. The speed should be as fast as possible, and the image quality should be as high as possible, but the digital images occupy considerable resources. Undoubtedly, this great demand and strict requirements have great challenges to the past compression technology under the conditions of the existing network resources. Therefore, it is particularly important to improve the compression ratio of the image and improve the ability of decoding the image.

Images is a similar and vivid description of objective things and a more intuitive way of representing objective objects. It contains information about the described object and is the main source of information for people. According to statistics, about 75% of the information a person comes from vision. Entering the information age, people will more and more rely on computers to obtain and use information, and After digitization, multimedia information has massive data, and there is a big gap between the computer storage resources and network bandwidth provided by the current hardware technology. In this way, it creates great difficulties in the storage and transmission of information, and becomes a bottleneck in preventing people's effective access and use of information. As the most important resource on the computer, the effective compression processing of image information will undoubtedly bring great benefits to people.

2) Research significance

The compression of digital image information will bring many practical meanings to people: save computer storage and computing resources, save network resources, and the compression of digital image information can also bring economic benefits to people, such as:

a. Fast transmission speed, less time consuming bandwidth resources;

b. Improve the bandwidth utilization rate, parallel transmission of video, voice, fax and other information resources;

c. Reduce storage resources and save storage overhead.

Traditional image compression methods, such as JPEG, JPEG2000, and BPG, enabling the unprecedented development of the theory and technology of digital image processing. Digital image data volume is very large, but between each pixel is related, there is a lot of redundant information between adjacent rows and between adjacent pixels in video sequence, namely spatial correlation and temporal correlation, and information entropy redundancy, structural redundancy, visual redundancy, etc., redundancy is the entry point of image compression, can be used to remove the redundancy as much as possible, reduce the amount of image date.

In 2012, Convolutional Neural Network (CNN) saw a historic breakthrough — AlexNet, with the top-5 error rate down ten percentage points in ImageNet from the previous year, and far surpassing the second place in that year. CNN reputation, deep learning has developed in full swing in the image field, image classification, video tracking, target detection, image compression and so on can be realized through deep learning. Advantages of the image compression method based on deep learning: first, the image compression ratio, compared with the traditional image compression method, the image compression; and the traditional image quality, under the same compression ratio, the deep learning based image compression method is higher than the traditional image compression method.Therefore, it is of great significance to study deep learning-based image compression theory and methods in both academic value and practical application.At the same time, the research of deep learning in the image field is still in its infancy, providing researchers with more difficult challenges and broader research space.

2. Research status at home and abroad

In recent years, deep learning, such as convolutional neural networks, has achieved great success in image processing and computer vision, especially for in high-level visual applications such as recognition and understanding. These networks tend to outperform state-of-the-art engineering codecs, such as BPG, WebP, and JPEG2000 in perceptual metrics. In addition to achieving higher compression ratios on natural images, they can also easily be adapted to specific target domains, such as stereoscopic or medical images, and promise to process and index directly from the compressed representation.This makes great progress in image compression technology.

The research work of image compression encoding technology has had a history of more than 60 years since the digitization of TV signals was proposed in 1948. Kunt proposed the concept of first-generation data compression coding. He called the encoding methods based on removing redundancy developed in the 1940s the first generation of coding. Such as PCM, DPCM, subsampling coding method, transformation coding method, etc. Until the 1950s and 1960s, image compression technology was only stuck in the research of predictive encoding, subsampling, and interpolation restoration technologies, but it was very immature. Huffman first published his paper "Construction methods of minimum redundancy code" in 1952, and data compression began to be implemented in commercial programs and applied in many technical fields. The second generation of data compression coding began in the 1990s, and mathematicians aren't satisfied with some fatal weaknesses in Huffman coding and decided to design another, more accurate, more close to the "entropy" limit in information theory --- "Arithmetic coding", On the basis of arithmetic coding, transformation coding was developed, such as Pyramid coding, Fractal coding. In the late 1970s and early 1980s, people gradually realize that for many grayscale or color images and even sound files, all their information is not faithfully retained. Under the condition of allowing certain loss of accuracy, more effective compression methods can be achieved. By the late 1980s, many had gained gains in the field, designing a number of sound and image compression algorithms that are amazing in terms of compression.

The third generation of data compression coding technology is mainly present from the 1990s, and the main achievements of image compression technology is wavelet coding, fnactal Coding, etc. Vector quantitative coding technology has also been greatly developed. Based on the establishment of wavelet transformation theory, fractal theory, artificial neural network theory, and visual simulation theory, people began to break through the traditional source coding theory, for example no longer assuming that the image is stationary random fields. Scientific and technological achievements and papers on image coding technology are also increasing day by day, and image coding technology begins to flourish.

Deep neural networks have developed into a state-of-the-art technology for computer vision tasks. Although these neural networks are powerful, large number of weights consume considerable storage and memory bandwidth. The storage and computation of neural network models on embedded devices remains a huge challenge.

At present, there are four main directions for the design of lightweight neural network model in industrial and academic circles: (1) artificial design of lightweight neural network model, (2) automatic design of neural network based on Neural Architecture Search (NAS); (3) CNN model compression; (4) automatic model compression --- based on AutoML. Traditional deep network models mainly consists of modules stacked such as convolutional layer, nonlinear activation layer, downsampling layer and fully connected layer. Conconvolutional layer is characterized by local connection and weight sharing. Although few parameters need to be trained, the time consuming of one forward is large. In contrast, although full connection layer can reach more than 80% of all parameters of the network, it doesn’t occupy much time of forward. Classical deep network models can refer to network models such as AlexNet or VGG originally used for image recognition. These modules can be roughly divided into two categories: One is the module containing training parameters in convolution layer and full connection layer, etc., in which the number of parameters is often artificially set; and those without no training parameters in non-linear activation layer and downsampling layer. The model parameters represent the model complexity to some extent, and also determine how much space the model takes up to some extent. The artificial number of parameters is often calculated after repeated experiments in the laboratory, and such locally optimal hyperparameters don’t represent the "real needs" of the network. They have both a degree of redundancy nor weigh the relationship between cost and effect. Therefore, one direction of network compression is to reduce the complexity of the model by squeezing the number of parameters of the model. The main goal of model compression is to improve the model compression ratio and improve the model prediction speed while satisfying the accuracy loss requirements. Model compression is the basis of the engineering implementation of the embedded deep learning, which can optimize the model volume and computational amount at the algorithm level, and on the basis of the embedded architecture design and the scheduling optimization of the model and data.

3. Research content of thesis

This research is a deep learning based convolutional neural network compression of digital image, and then the convolutional neural network model compression method. It mainly includes the construction of image compression convolutional network model, model compression method (model pruning, model reconstruction).

Specifically includes the following aspects:

a. In the field of image processing, convolution operation can be used to scale and filter the image;

b. Build the image compression convolutional neural network model according to the characteristics of the image, including the convolutional network layer structure, loss function, convolutional kernel size, activation function, and so on;

c. Neural network model compression method, pruning, quantification, lightweight network, etc., performs the compression processing of the image compression convolutional neural network model through the model compression method;

d. According to the comparison of traditional image compression method and convolutional neural network image compression, the advantages of convolutional neural network on image compression are verified;

e. The comparison between the convolutional neural network model and the light model after the model is compressed shows that the design method proposed in this paper has great advantages.

4. Organizational structure of the thesis

This thesis consists of six chapters, each summarized as follows:

Chapter 1, for the introduction chapter, this chapter mainly introduces the background knowledge of image compression convolutional neural network and the significance of researching this topic, and the development of image compression technology and the research situation of image compression algorithm at home and abroad, and also points out that the current academic circle is also a popular research direction for using deep learning convolutional neural network for image compression.

Chapter 2, for the basic introduction chapter of convolutional neural network, this chapter first introduces the image processing of neural network and convolutional neural network, and then the introduction of automatically generated data use GAN (Generative Adversarial Networks, GAN) in Deep Learning. Finally, this thesis briefly expounds the compression technology of neural network model.

Chapter 3, for the design scheme of image compression convolutional neural network, the first section of this chapter mainly introduces the implementation of convolutional neural network for image information compression, and the second part mainly introduces the restoration of compressed image by GAN. Through the joint training of convolutional neural network and adversarial generative network, the deep learning image compression method is realized.

Chapter 4 is the design scheme of model compression. The first section of this chapter mainly describes the research of model pruning in model compression method on the convolutional neural network of image compression, the second section mainly describes the lightweight network in the model compression method to research of image compression convolutional neural network. The expected results and effects can be obtained by both methods.

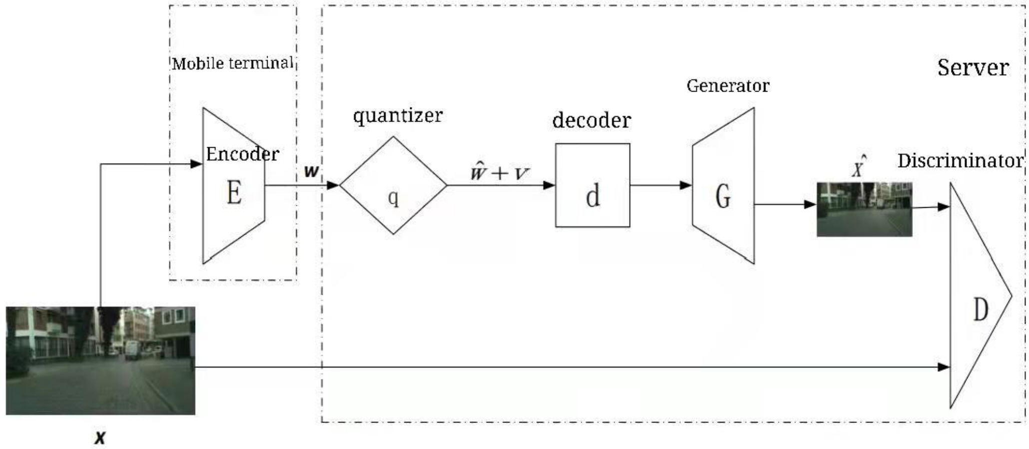
Chapter 5, this chapter mainly through the analysis and comparison of the experimental results, and compare to other image compression methods, concluded that the image compression method proposed in this thesis has advantages.

Chapter 6 is the summarization and prospect of the thesis. this chapter mainly summarizes the research content and research results, and points out the shortcomings in this research and the need to be improved, and finally makes prospects for the future.

5. Here mainly introduces chapter 3: a scheme based on the DCGAN network model architecture

1) End-to-End image compression method

The proposed end-to-end deep convolutional generative adversarial network (Deep Convolution Generative Adversarial Networks, DCGAN). Schematic diagram of the image compression network structure as shown in the figure below, Original image x for City Street View plot, E is an encoder composed of a convolutional network, Image x encoded by encoder E produces w, The w was then quantified by the quantizer q and then will get $\hat{w}$, Combining random noise v, enters the adversarial generative network (GANs) after the decoder d. G is generating network (generator), generating compressed image $\hat{x}$, and D is discriminant network (discriminator), in order to performing error judgment of original image x and compressed image $\hat{x}$.



End-to-End DCGAN image compression network structure diagram

2) The Design of the Encoder E---The Image downsampling Method

This thesis mainly proposes the use of convolutional neural networks to replace image coding operations in traditional image compression methods. The design of the encoder E is based on the basic principles of convolutional neural networks, and meets the basic requirements of image compression. Its convolutional calculation can keep the information of the image while narrowing the image. The network structure is relatively small, which is the first modularization of the whole network structure. The encoder can save the tedious steps of traditional methods and improve the compression effect.

In this thesis, a convolutional neural network convolutions the images to achieve the effect of downsampling. As shown below, assuming that the size of a layer of the input image is a=9*9. Each time convolution is carried out through the convolution kernel, the pixel value of a convolution kernel size area of an image can be downsampled into a pixel value. After the convolution kernel slides on the image for many times until the convolution of the whole image is completed, realizing the purpose of one downsampling. In order to capture the edge information of the image, the pixel value is usually supplemented at the edge of the image, which can be a pixel value that mirrors the edge pixel, or can be a pixel value that is symmetric with the edge pixel, or directly supplement the 0 pixel value, etc.

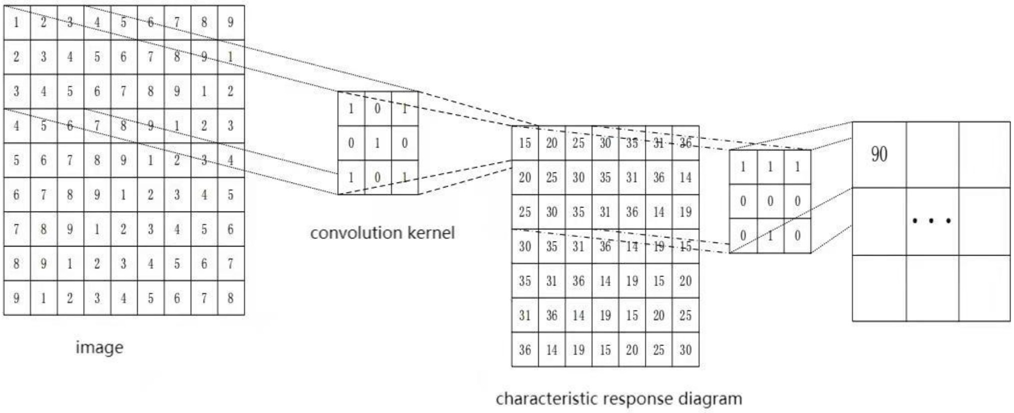
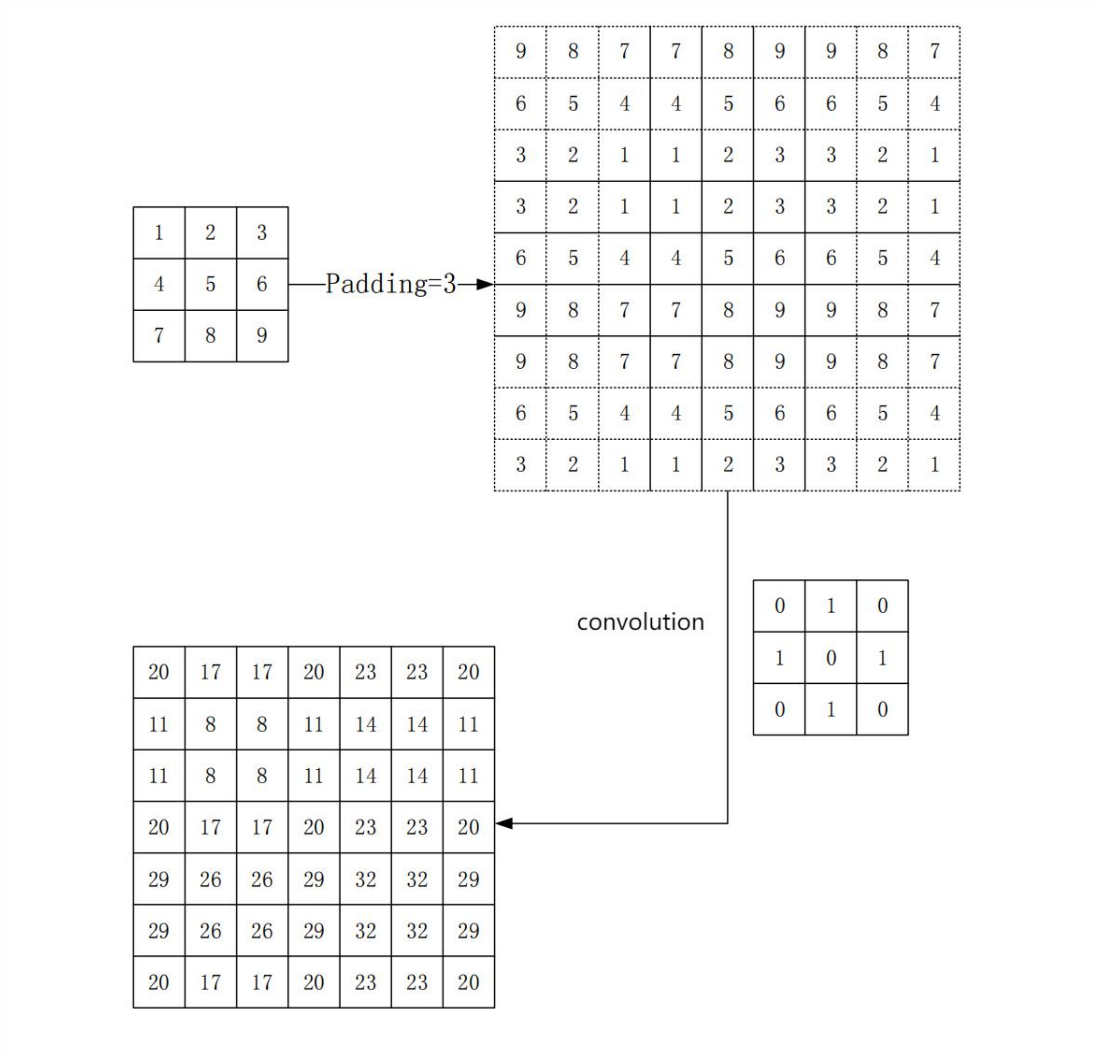


Image subsampling method based on CNN

3) Design of the decoder d---The Image upsampling Method

The image decoding unit presented in this thesis is different from the traditional decoding unit, which uses deep convolutional neural network model to realize decoding the image coding results of the convolutional neural network. Using convolutional neural network to decode the image can not only restore the image information, but also get clearer and higher quality images than the traditional image decoder.

In contrast to downsampling, upsampling is the operation of enlarging the reduced image. Under normal circumstances, image convolution operation will reduce or keep the size of the image unchanged. In this thesis, convolution is used to upsample the image. Another effect of convolution on the image is amplification. Because it can to do the padding during convolution, and it's the padding that makes it possible for convolution to upsample the image. As shown below, it is a schematic diagram of upsampling operation on image by convolution.



1	2	3
4	5	6
7	8	9

9	8	7	7	8	9	9	8	7
6	5	4	4	5	6	6	5	4
3	2	1	1	2	3	3	2	1
3	2	1	1	2	3	3	2	1
6	5	4	4	5	6	6	5	4
9	8	7	7	8	9	9	8	7
9	8	7	7	8	9	9	8	7
6	5	4	4	5	6	6	5	4
3	2	1	1	2	3	3	2	1

0	1	0
1	0	1
0	1	0

20	17	17	20	23	23	20
11	8	8	11	14	14	11
11	8	8	11	14	14	11
20	17	17	20	23	23	20
29	26	26	29	32	32	29
29	26	26	29	32	32	29
20	17	17	20	23	23	20

Image upsampling method

6. JPEG statdard

JPEG standards use two entropy codes: Hoffman coding and Arithmetic coding, which can further reduce the amount of data.

The basic idea of the Hoffman encoding method is based on data statistics, which provides the best method of assigning variable-length codewords to symbolic letters with known probability distributions. It represents letter symbols by variable code length, depending on their occurrence probability (the more symbols, the shorter the code assigned to them). Assuming all exact code squares with a source symbol probability of 1 / 2, the shortest mean code length can be obtained by Hoffman encoding. The Hoffman encoding algorithm can be described as follows:

Step 1: Lists the probabilities of the various gray scales (source symbols) in the image. These probabilities are arranged in descending order, with the highest probability at the top and the lowest probability at the bottom. The junction set is generated by making these probabilities the leaves of the binary tree.

Step 2: Remove two knots with the lowest probability from the set and generate a new probability representing the sum of the two probabilities. The order of probabilities was reorganized in descending order for processing.

Step 3: Create a parent node with a new probability, and mark the branch of the top (or left) subjunction as 1, and mark the branch of the bottom (or right) subjunction as 0.

Step 4: Update the junction set by replacing the two nodes with the two lowest probabilities of the newly generated node. Exit if the junction set contains only one node, exit. Otherwise, go to step 2.

This is the best prefix code generated from the junction set, a notion reduces the average length of the coding and the overall size of the compressed data becomes smaller compared to the original data. The Hoffman algorithm is the first to provide a solution for building encoding with less redundancy.

Arithmetic coding is a variable length coding process similar to Hoffman coding and also aims to reduce coding redundancy and is the best encoding only if an integer power with all sign probabilities of 1 / 2. Glen and Langdon detail the arithmetic coding in the literature. The basic idea of arithmetic encoding is to map the input data between 0 and 1 and then divide it into multiple smaller intervals that correspond to the probability of the input symbol, interval left closed and right open. The next input symbol selects one of these intervals, then repeats the process and, and finally, the message can be represented using any number within the final interval. The arithmetic encoding uses the probability of the symbol and its encoding interval, which determines the output value of the compressed symbol.

JPEG2000, similar to JPEG, also contains three parts: transformation, quantification, and entropy encoding, except with different methods. The transformation adopted by JPEG2000 is a wavelet transform, transforming the image into four different subbands, concentrating most of the energy on low frequency subbands, then quantidifferent quantification parameters according to different choices of the subbands, using EBCOT coding method to obtain the bit flow. JPEG2000 supports progressive coding while supporting losssy and lossless compression. At the high compression ratio, the JPEG2000 is reconstructed better than the JPEG method, but the advantage of the JPEG2000 is not obvious at the low compression ratio (less than 10 times), and even in some cases, the JPEG method works better than the JPEG2000. Because of the loss of high-frequency information of the image, the distortion of JPEG2000 is mainly fuzzy distortion, but because the method has high compression performance, the high spectral image compression mostly adopts the combination of JPEG2000 and other transformations.

7. References

[1] Wen Tao, Feng Jiang, Shengping Zhang et al:An end-to-end compression framework based on convolutional neural networks[C]. Data Compression Conference(DCC). Snowbird, UT, USA, 2017: 463

[2] Bellard, F.: BPG Image format. https://bellard.org/bpg/

[3] WebP Image format. https://developers.google.com/speed/webp/

[4] Taubman, D.S., Marcellin, M.W.: JPEG 2000: Image Compression Fundamentals, Standards and Practice[C]. Kluwer Academic Publishers, Norwell, MA, USA, 2001.

[5] Krizhevsky, Alex, Sutskever et al. Imagenet classification with deep convolutional neural networks[C]. In NIPS, pp. 1097–1105, 2012.

[6] Simonyan, Karen and Zisserman, Andrew. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[7] Song Han, Huizi Mao, William J. Dally: Deep Compression: Compressing Deep Neural Network with Pruning, Trained Quantization and Huffman Coding[J]. International Conference on Learning Representations (ICLR), 2016.

[8] Sercu T, Puhrsch C, Kingsbury B et al. Very de multilingnal convolutional neural networks for LVCSR[C]. In: Proe. of the Acoustics, Speech and Signal Processing(ICASSP). Shanghai: IEEE, 2016. 4955—4959.

[9] He K, Zhang X, Ren S et al. Deep residual learning for image recognition[J]. In: Proc. of the IEEE Cone on Computer Vision and Pattem Recognition(CVPR). Las Vegas: IEEE, 2016. 770—778.

[10] Langdon G G. An introduction to arithmetic coding[J]. IBM Journal of Research and Development, 1984, 28(2): 135-149