CNN Based Image Classification

Xiaoling Long
SIST
ShanghaiTech Univ.
longxl@shanghaitech.edu.cn

Hongyu Chen
SIST
ShanghaiTech Univ.
chenhy3@shanghaitech.edu.cn

Second Author
SIST
ShanghaiTech Univ.
secondauthor@i2.org

Abstract

The ABSTRACT is to be in fully-justified italicized text, at the top of the left-hand column, below the author and affiliation information. Use the word "Abstract" as the title, in 12-point Times, boldface type, centered relative to the column, initially capitalized. The abstract is to be in 10-point, single-spaced type. Leave two blank lines after the Abstract, then begin the main text. Look at previous CVPR abstracts to get a feel for style and length.

1. Introduction

Image classification is a fundamental challenge in computer vision [5]. Consider the problem of detecting objects from a category, such as people or cars, in static images. This is a difficult problem because objects in each category can vary greatly in appearance. Variations arise from changes in illumination, viewpoint, and intra-class variability of shape and other visual properties among object instances. For example, people wear different clothes and take a variety of poses while cars come in various shapes and colors.Classification between the objects is easy task for humans but it has proved to be a complex problem for machines.Image classification refers to the labeling of images into one of a number of predefined categories. Image classification is an important and challenging task in various application domains, including biomedical imaging, biometry, video surveillance, vehicle navigation, industrial visual inspection, robot navigation, and remote sensing [9].

Image classification uses both supervised and unsupervised Traditional methods of computer vision and machine learning cannot match human performance on tasks such as the recognition of handwritten digits or traffic signs. [3]

Nonparametric classifiers such as decision tree classifier, neural network, SVM classifier and knowledge based classification are also very common in image classification.

Disadvantages about the deep learning.For the deep learning based algorithm required the ability of labeled samples for training. The collection of labeled data is a time consuming process as well as costly.

2. CNN Based Image classification

The ImageNet Large Scale Visual Recognition Challenge(ILSRC) is a benchmark in object category classification and detection on 1000-classes and millions of images. After AlexNet achieved hugu success in ILSRC-2012, there are various variations of AlexNet [10] and many other types of ConvNet for image classification. Since that, ConvNet is widely used for image classification. [2] illustrate beriefly what is ConvNet, the components of ConvNet, the activation function in ConvNet, from LeNet to ResNet bunch of successful ConvNet and some open issues on CNN based image classification.

AlexNet [10] brought Convolutional neural network into ILSRC. In this implementation, it contains 8 layers 5 convolutional and 3 fully-conneted. The main features of this network's architecture are ReLU [13] as activation function, overlapping pooling and skills for reducing overfitting. Based on ReLU and overlapping pooling, the network err rate has more or less reduction, and ReLU network learn several times faster than other saturating acitvation function such as tanh neurons network. Overfitting is common issue for machine learning, it uses data augmentation and dropout to avoid overfit. Dropout is a skill to reduce argument or increase hypothesis. There are many discuss about this. As to data augmentation, it enlarge the dataset by cropping 224×224 patched from the original image as well as these patchs' horizontal refliction. At the end averaging the predications as final socre.

[4] explored the generalization ability of ConvNet features, releasing DeCAF. Traditional image classifition pipeline is extracting feature, building bag of feature then put into classifier. [25] propose a CNN based feature extractor. This is an unsurpervising learning ConvNet or in other words, input image is also the kind of ground truth. After feature extracting, the final result can classify by any classifier. This strategy is used as a tool for visualizing and undestanding how ConvNet works [24].ConvNet have an impressive classification performance. However there is no

clear understanding for why this work. [24] propose a architecture for visualizing and undestanding how ConvNet works.

OverFeat [18] is a integrated recognition, localization and detection. This network uses CNN extract feature from image and then perform classification and localization and detection. Multi-scale classification brought up in [18] to increase accuracy.

“Networ In Network” [11] propsed a new deep network structure. Different from conventional convolution layer, it brings up a new Mlpconv layer. This Mlpconv layer consist of sliding multilayer perceptron(MLP) window. In stead of fully-connected layer at the top of network, global average pooling is used to produce the resulting vector fed directly into the softmax layer. Verified by experiments, this NIN structure indeed works well on some benchmark datasets, and global average pooling can be regarded as regularizer. THis glolbal average pooling has no parameter. This stratety is used widely afterwards. 1×1 convolution conception proposed in [11] is used in GoogLeNet for dimension reduction.

AlexNet make a great success in image classification. Afterwards many various Network appear. GoogLeNet [22] proposed by google is a new level of oganization in the form of the “ Inception module”. This is a multi-scale arcthitecture. With the limitation of computational resource, it perform a 1×1 convolution to dimension reduction. Auxiliary classifier is also a brilliant strategy.This smart design makes a great success in ILSRC-2014. At the same time, the widely used ConvNet architecture VGGNet [19] won the first place in *Classification* + *Localization competition*. It adds the number of layers up to $16 - 19$. Instead of 7×7 convolution filter in [19], it uses 3×3 as convolution filter. After multiple layers, it can get similar effect as 7×7 one. This design significantly reduces the parameters, and then reduces the overfitting. It also means the number of layers significantly increases. Altering convolutional layers and poolint layer became a common used Network architecture.

As the depth of ConvNet increasing, training gets more and more difficult. The training of very deep network becomes a open issue in CNN. Highway Networks [20, 21] propose *information highways* which allow unimpeded information flow across several layers. The *transform gate* $T(x, W_T)$ and the *carry gate* $C(x, W_C)$ proposed for decided how much flow pass through to output. The new model given by

$$y_{output} = H(x, W_H) \cdot T(x, W_T) + x \cdot C(x, W_C) \quad (1)$$

For simplicty, [8] set $C = 1 - T$. This design make training hundreds of layers be possible and the err rate just has slightly increase. This architecture promote the success of ResNet [8]. ResNet has similar structure as deep plain network stacked by dozens of convolution layers followed by global pooling layer and 1 fully-connected layer. except shortcup connection. This design has a residual representation which called deep residual learning. This archicture keeps parameter less than VGG-19 model even the network has 152 layers. This smart design make ResNet won first place in ILSRC-15.

There are several works based on region based image classification.Such as [7], [6] [23], [1]. Region based methods are computationally expensively.

R-CNN and fast R-CNN is regioned based convolutional network method for object detection and image classification. In [6],they use the selective search to select those propose regions. Features are extracted from each proposal region. Then a SVM classifier is used for the category classification. This kind of algorithm need a lot of time to process each image. About 2000 regions are proposed from each image where there maybe several objects in the image.

Compared to R-CNN [6], Fast R-CNN [7] employs several innovations to improve training and testing speed while also increasing detection accuracy. the fast R-CNN has several advantages: (1) higher detection quality than R-CNN 2 using a multi task loss ,predict the object and its confidence. No disk storage is required for additional feature catching.The computation is shared during training. Fast achieves a near real time rates uses a very deep network.

Proposal based image classification also contains some great work. Such as [17] [14] [15] [16].

Faster R-CNN [17] is another convolutional network which based on the region proposal methods. In [17],The author show that an algorithmic change computing proposals with a deep convolutional neural network leads to an elegant and effective solution where proposal computation is nearly cost-free given the detection networks computation. They removed the select search algorithm replaced by a region proposal network. The high quality proposal is used by the Fast R-CNN network for detection and classification. For the very deep VGG-16 model the detection system has a frame rate of 5fps (including all steps) on a GPU, while achieving state-of-the-art object detection accuracy.

Fast and Faster R-CNN focus on speeding up the R-CNN framework by sharing computation and using neural networks to propose regions instead of Selective Search. While they offer speed and accuracy improvements over R-CNN, both still fall short of real-time performance. Fast R-CNN speeds up the classification stage of R-CNN but it still relies on selective search which can take around 2 seconds per image to generate bounding box proposals.

In [14], the author achieves a real time detection and classification in 45 frames per second. A smaller version of their network can be achieved by 155 frames per second. The problem is view as a regression task. The output of the net contains five parameters. $(x, y, w, h, confidence)$ where x, y, w, h means x location , y location, the width and the height of the target, the probabilistic of the cell contains an object respectively. YOLO shares some similarities with R-CNN. Each grid cell proposes a potential bounding boxes and scores those boxes using convolutional features.

[12],the author propose an algorithm which can runs real time object detect and classification. A key feature of the SSD algorithm is that multi-scale of the convolutional bounding boxes outputs are attached to different feature maps.SSD is faster than R-CNN and its variants.

Real-Time Detectors	Train	mAP	FPS
100Hz DPM	2007	16.0	100
30Hz DPM	2007	26.1	30
Fast YOLO	2007+2012	52.7	**155**
YOLO	2007+2012	**63.4**	45
Less Than Real-Time			
Fastest DPM	2007	30.4	15
R-CNN Minus R	2007	53.5	6
Fast R-CNN	2007+2012	70.0	0.5
Faster R-CNN VGG-16	2007+2012	73.2	7
Faster R-CNN ZF	2007+2012	62.1	18

Figure 1. Real-Time Systems on P ASCAL VOC 2007.Comparing the performance and speed of fast detectors. Fast YOLO is the fastest detector on record for P ASCAL VOC detection and is still twice as accurate as any other real-time detector. YOLO is 10 mAP more accurate than the fast version while still well above real-time in speed.

Figure1 shows the most famous network on image classification. The accuracy and the speed is shown on the table.

3. conclusion

In this paper we have discussed about the different types of image classification techniques. So this paper will help us in selecting an appropriate classification technique among all the available techniques.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2189–2202, 2012.

[2] N. Aloysius and M. Geetha. A review on deep convolutional neural networks. In *Communication and Signal Processing (ICCSP), 2017 International Conference on*, pages 0588–0592. IEEE, 2017.

[3] D. Ciregan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. *computer vision and pattern recognition*, pages 3642–3649, 2012.

[4] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International conference on machine learning*, pages 647–655, 2014.

[5] P. F. Felzenszwalb, R. B. Girshick, D. A. Mcallester, and D. Ramanan. Visual object detection with deformable part models. *Communications of The ACM*, 56(9):97–105, 2013.

[6] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, June 2014.

[7] R. B. Girshick. Fast r-cnn. *international conference on computer vision*, pages 1440–1448, 2015.

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] P. Kamavisdar, S. Saluja, and S. Agrawal. A survey on image classification approaches and techniques. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(1):1005–1009, 2013.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] M. Lin, Q. Chen, and S. Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[12] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. Ssd: Single shot multibox detector. *european conference on computer vision*, pages 21–37, 2016.

[13] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[14] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *computer vision and pattern recognition*, pages 779–788, 2016.

[15] J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger. pages 6517–6525, 2016.

[16] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018.

[17] S. Ren, K. He, R. B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal net-

works. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.

[18] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

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

[20] R. K. Srivastava, K. Greff, and J. Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015.

[21] R. K. Srivastava, K. Greff, and J. Schmidhuber. Training very deep networks. In *Advances in neural information processing systems*, pages 2377–2385, 2015.

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, et al. Going deeper with convolutions. Cvpr, 2015.

[23] J. R. R. Uijlings, K. E. A. V. De Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171, 2013.

[24] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[25] M. D. Zeiler, G. W. Taylor, and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2018–2025. IEEE, 2011.