IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 8 Issue 11, November 2021
ISSN (Online) 2348 – 7968 | Impact Factor (2020) – 6.72
www.ijiset.com

Revision of Video Based Soppy Examination Using CNN

Siddhant Vasantrao Wadmare
Researcher OPJS University Churu Rajasthan, Siddh.sw@gmail.com

ABSTRACT

Convolutional Neural Networks (CNNs) have been established as a powerful class of models for video, image recognition problems. Encouraged by these results, we provide an extensive empirical evaluation of CNNs on video classification using a new dataset of 1 million YouTube videos belonging to 487 classes. We study multiple approaches for extending the connectivity of a CNN in time domain to take advantage of local spatio-temporal information and suggest a multiresolution, foveated architecture as a promising way of speeding up the training. Videos have become ubiquitous on the internet, which has encouraged the development of algorithms that can analyze their semantic content for various applications, including search and summarization. Recently, Convolutional Neural Networks (CNNs) have been demonstrated as an effective class of models for understanding image content, giving state-of-the-art results on image recognition, segmentation, detection and retrieval. The key enabling factors behind these results were techniques for scaling up the networks to tens of millions of parameters and massive labeled datasets that can support the learning process. Under these conditions, CNNs have been shown to learn powerful and interpretable video features. Encouraged by positive results in domain of images, we study the performance of CNNs in video classification, where the networks have access to not only the appearance information present in single, static images, but also their complex temporal evolution.

KEYWORDS: Video, CNN, images, applications, algorithms, convolutional neural networks.

INTRODUCTION

The Convolutional Neural Networks (CNN) is used in a number of tasks which have a great performance in different applications. Hand-written digit recognition was one of the first applications for the successful implementation of CNN architecture. After the development of CNN, the networks with the invention of new layers and the involvement of various computer vision techniques have been steadily improving. CNN is often used for various combinations of datasets of sketches in the Image Net Challenge. Few researchers showed a link between the human subject and the ability of a qualified image data set detection network. The comparative tests have shown that the data set demonstrates an accuracy rate of 73.1% for humans, while a qualified network results indicate a precision rate of 64%. In the same way, the accuracy of Convolutional Neural Networks was 74.9 percent, which is higher than the human accuracy average. The techniques are primarily used to achieve a much better accuracy performance using the strokes. In various situations, studies are under way to understand the behavior of Deep Neural Network. This study demonstrates how the results of grouping are greatly influenced by small changes in an picture. Through this job, too, photos are presented that are totally unrecognized by people but are identified by the qualified networks at high accuracy levels.

Throughout the field of feature detectors and descriptors there have been many innovations and many algorithms and techniques for object classification and scene classification have been developed. The object detection and the scene recognition literature includes an abundance of work. In the multimedia community, the principle of creating specific object detectors for the basic interpretation of images is close to the research performed in the field of image and video annotations and semantine indexing using several "semantic principles". Each semanticized definition is trained by using video images or frames in literature relevant to our work. Therefore, with many unrestrained objects in the scene, the approach is difficult to use and understand. The previous methods concentrated on the identification and classification of individual objects based on human-defined feature collection. The methods proposed investigate the relation between objects in the classification of the level. Several techniques of scene classification are used to measure the usefulness of the object inventory. Many types of research have been conducted emphasizing their focus on low-level feature extraction for object recognition and classification, namely Histogram of oriented gradient (HOG), GIST, filter bank, and abag of feature (BoF) implemented though word vocabulary.

Our work’s principal objective is to understand the efficiency of both static and live video feed networks. The first step is to transfer data to the networks using image datasets. The prediction rate of the same object on static images and real-time video feeds is checked . This is followed by In further sections of the tables, different accuracy rates are observed and noted and presented. The third key criteria for performance assessment were to verify if the prediction precision varies among all the CNNs selected for the study. It should be noted that videos are not used as data set for training; they are used as datasets for testing. Therefore, we search for the best grade of photographs where the object is the key attribute for the scene rating. Different layers of the convolutional neural network used are:

- **Input Layer:** The first layer of each CNN used is 'input layer' which takes images, resize them for passing onto further layers for feature extraction.

- **Convolution Layer:** The next few layers are 'Convolution layers' which act as filters for images, hence finding out features from images and also used for calculating the match feature points during testing.

- **Pooling Layer:** The extracted feature sets are then passed to 'pooling layer'. This layer takes large images and shrinks them down while preserving the most important information in them. It keeps the maximum value from each window; it preserves the best fits of each feature within the window.

- **Rectified Linear Unit Layer:** The next 'Rectified Linear Unit' or ReLU layer swaps every negative number of the pooling layer with 0. This helps the CNN stay mathematically stable by keeping learned values from getting stuck near 0 or blowing up toward infinity.

- **Fully Connected Layer:** The final layer is the fully connected layers which takes the high-level filtered images and translate them into categories with labels.

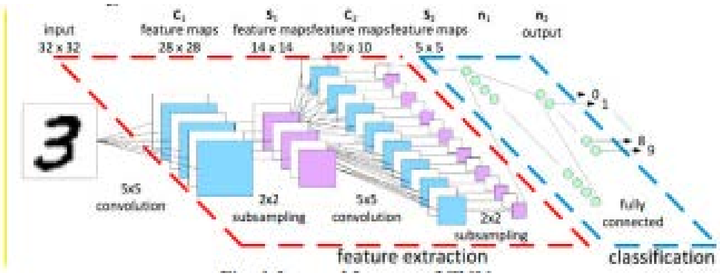

Fig. 1 Internal Layers of CNNs

The steps of proposed method are as follows:

1. Creating training and testing dataset: The super classes images used for training is resized [224,244] pixels for Alex Net and [227,227] pixels GoogLeNet and ResNet50, and the dataset is divided into two categories i.e. training and validation data sets.

2. Modifying CNNs network: Substitute a fully connected layer, a soft max layer and a classification output layer for the last three network layers. Set the final fully linked layer in the training data set to have the same size as the number of classes. Increase the learning rate of the fully connected network layer more rapidly.

3. Train the network: Set the training options, including learning rate, mini-batch size, and validation data according to GPU specification of the system. Train the network using the training data.

4. Test the accuracy of the network: Classify the validation images using the fine-tuned network, and calculate the classification accuracy. Similarly testing the fine tune network on real time video feeds for accurate results.

CONVOLUTIONAL NEURAL NETWORK (CNN)

The term CNN is used to describe a neural networks architecture based on spatially located neural input to two-dimensional arraies (usually images). The technique of shared weight or receptive areas has been identified also as this architecture. The concept of weight sharing is that of a set of neurons with the same weight in one layer. The use of shared weights means, in other words, that the image is combined with a kernel defined by weights, all of which detect the same feature but at various positions in the inbound picture (receptive fields). The weight sharing technique also decreases the difference from test error to training error, which is greatly beneficial in the field of image processing, but without the use of sufficient constraints the high dimensionality of the inputs usually contributes to an increase in the weight sharing technically. This means that weight sharing decreases the efficiency of the system. Processing systems are structured in a spatial context of similar weight vectors and local receptive areas, which establish architecture similarities to biological systems of viewing models. Neural networks in the field of speech and image analysis CNN with local weight sharing topology have gained considerable interest [7]. Our topology is similar to biological networks based on receptive fields and increases local distortions tolerance. Furthermore, weight sharing effectively reduces the strength of the model and the number of masses. It is an advantage if images with large-scale input vectors are introduced directly to the network instead of explicit extraction of features resulting in reductions that are usually implemented before classification. To order to minimize the number of weights, weight sharing can also be used as an alternative to weight reduction. In addition to fully connected feed forward networks, local topology network can be transferred more effectively to a locally connected parallel computer.

VIDEO FACE RECOGNITION

Real time frame processing of video image should be implemented in the period of 1/30 second (30 frame/sec) or less than this period. Detection and recognition should be reached as soon as possible in order to achieve this goal. The original Neocognitron is therefore updated, and in this work the MNEO is introduced to address this challenge. Face recognition is challenging visual classification task. Reasonable deviations from a 3D face must be recognized and the face must also be normalised in terms of size, position and orientation. The use of the neocognitron significantly simplifies this task, given that the neocognitron can correctly recognize stimulation pattern without being affected by position shifts or by affinity of scaling and rotating. The MNEO is trained to use images from different positions, scales, rotations and orientations in this work. In the reconnaissance cycle , the system should recognize if a trained image is applied to the system. The system can also recognize an image for the same class in a variant position, according to the abovementioned characteristics of Neocognitron (or MNEO). When you use more examples of instruction, you get more generalization. In the region of interest (segmented area) is performed the recognition. The MNEO takes into account variations in facial recognition problem. The detection process starts once the segmentation process of the colored frame is finished The block of face detection in the figure 3. In our proposed system, can be removed. The exercise samples of variant poses can be increased for each face class as shown in the last paragraph. Once different consequent facial index frames are found (see Figure 2), the facial identifiable can be regarded as a specific class according to the predefined statistical criteria. The known face is not a legitimate class if the threshold value is less than appropriate. This latter will occur if the region (the divided region) belongs to sections other than the regions in the face (may be hand or color near to the human skin).

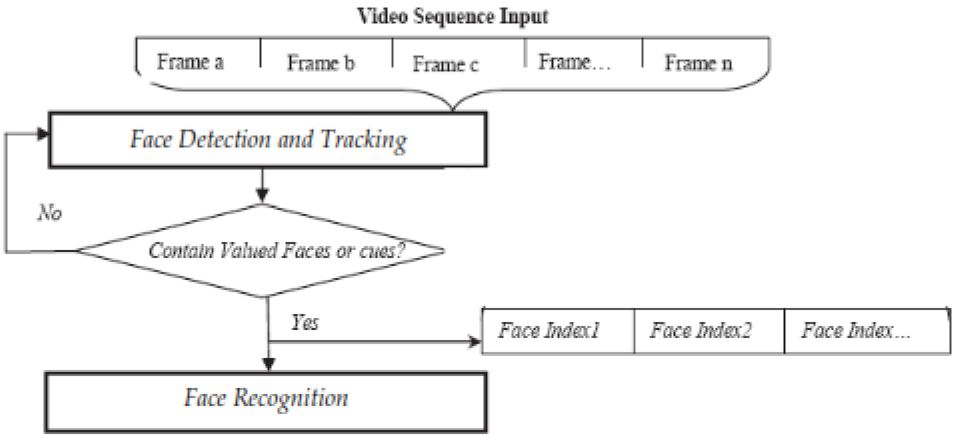

Fig. 2. Video based Face recognition system

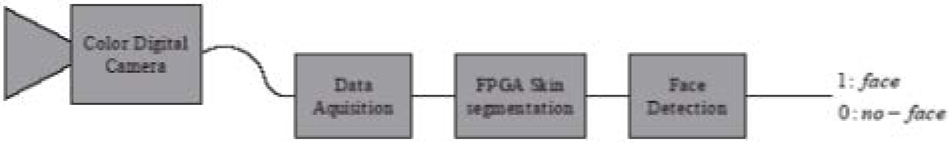

Fig. 3. Face Detector

MULTIRESOLUTION CNNs

CNNs usually take weeks of training on large datasets even on the most rapidly available GPUs, the efficiency of runtime is a critical part of our ability to experience different architecture and hyper parameter settings. This motivates approaches to speed up and maintain performance of the models. There are several fronts, including hardware improvements, weight measurement systems, better algorithms for optimization and initialization strategy, but we concentrate in this study on architectural changes that allow faster working times without sacrifiating efficiency. One way to accelerate networking is to reduce the number in each layer of layers and neurons, although we found that performance decreases consistently. We conducted further experiments with lower resolution pictures rather than reducing network size. However the high frequency detail in the pictures proved critical for good accuracy while it improved the operation time of the network.

Fovea and context streams

The suggested architecture of multi-resolutions attempts to achieve a consensus by providing 2 different data sources in two spatial resolutions (Figure 4). The Network Input consists of a 178 × 178 frame video clip. At half the original spatial resolution (89 x 89 pixels) the context stream receives the sampled frames, while the center of the fovea stream receives at the initial resolution 89×89. The total dimension of the input is thus reduced to half. This design especially benefits from the camera distortion in many online videos, as the object of interest always lies in the middle area.

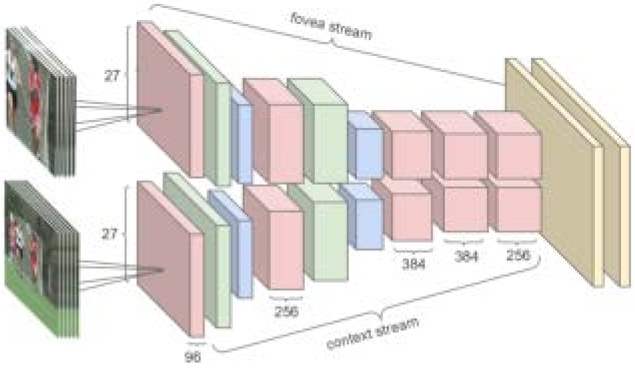

Figure 4: Multiresolution CNN architecture

Architecture changes

Both streams are managed by the same network as the full frame versions, but from 89 × 89 video clips. As the input is less than half the range size than the full frame models, we use the last pooling layer to ensure that the two streams still end in the 7×7×256 layer. The activations from both streams are linked and fed in with dense connections into the first fully connected layer.

CNN PERFORMANCE EVALUATION

The CNN SIMD FPGA framework is designed to test machine performance with the CNN SIMD FPGA framework since the program does not support learners. The most common rate of performance in the recall or execution process is CSC, which is specified as the number of operations multiply and accumulate per second. Among other FPGA systems, the speed performance of the FPGA system depends on the frequency of use of the FPGA model [11]. The operation frequency of the FPGA model is 50 MHz, the number of input vectors that processed in parallel are five, in each clock cycle, 5 input connections of (9bits≈1byte) are evaluated by 4 weight connections of the same bit precision, then the maximum CPS and CBS(= bytes(weight). bytes(input). CPS) achieved from the designed system are:

CPS=5×4×50×10^{6}=1GCPS

CBS=1×1×CPS =1GCBS

The above performance seems reasonable and comparable with the available neural network hardware.

CNN SYSTEM AND FACE RECOGNITION

The aim of the CNN system is to identify individuals, to allow them access to a group of people in real time and to deny everyone access. Many photographs are frequently available per person for training and recognition is needed in real time. With the same accuracy in the recognition of the image the CNN hardware system can recognize with the software version. This is because the efficient models are used, their parameters are set, functional approximations are used and the hardware deployment is based on the implementation of the competition. 12 separate groups are accepted by the program. If 60(12x5) training image and 60(12x5) test image were used, the recognition rate achieved for both soft and hardware versions was equivalent to 93 percent. Further improvement of recognition rates can be achieved through a fine tuning of the software CNN parameters during learning.

Speedup achieved in H/W CNN

From Fig. 5, one can see that the overall time required for processing one complete image on Xilinx Spartan-3 200,000 / Spartan 3E 500,000-gates Platform FPGAs of 50MHz is equal to (3.17)ms, while the same model needs (280) ms when implemented primarily in software, resulting a speedup of (88).

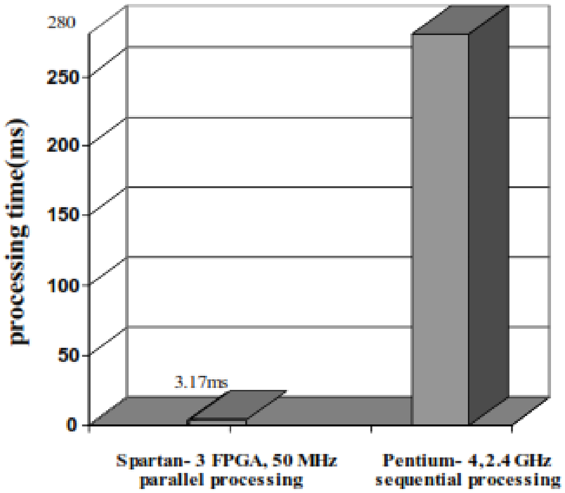

Fig. 5. An image processing time of parallel and sequential processors

Area consumption in H/W CNN

If the calculator word length (input value) is fixed at 9 bits and the accumulator word length at 13 bits, 1488 slices were needed in CNN hardware. This number accounts for 77 percent of the Spartan-3 200,000 FPGA and 30 percent of the Spartan-3E 500,000 FPGA slice. The CNN device can therefore be synthesized into a cheap FPGA chip. If a broad input system is required, the CNN system can also be synthesized in a reasonable cost chip.

CONCLUSION

Computational perspective, CNNs require extensively long periods of training time to effectively optimize the millions of parameters that parametrize the model. This difficulty is further compounded when extending the connectivity of the architecture in time because the network must process not just one image but several frames of video at a time. CNNs have been shown to learn powerful and interpretable image features. Encouraged by positive results in domain of images, we study the performance of CNNs in large-scale video classification, where the networks have access to not only the appearance information present in single, static images, but also their complex temporal evolution. There are several challenges to extending and applying CNNs in this setting. From a practical standpoint, there are currently no video classification benchmarks that match the scale and variety of existing image datasets because videos are significantly more difficult to collect, annotate and store. To obtain sufficient amount of data needed to train our CNN architectures.

REFERENCES

[1] Kou, F., Du, J., He, Y., & Ye, L. (2016) “Social Network Search Based on Semantic Analysis and Learning.” CAAI Transactions on Intelligence Technology.

[2] Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V.,& Garcia-Rodriguez, J. (2017) “A Review on Deep Learning Techniques Applied to Semantic Segmentation.”

[3] LSrinivas, S., Sarvadevabhatla, R. K., Mopuri, K. R., Prabhu, N., Kruthiventi, S. S., &Babu, R. V. (2016) “A taxonomy of deep convolutional neural nets for computer vision.”

[4] Yang, J., Jiang, Y. G., Hauptmann, A. G.,& Ngo, C. W. (2017) “Evaluating bag-of-visual-words representations in scene classification.” in Proceedings of the international workshop on Workshop on multimedia information retrieval.

[5] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ...&Rabinovich, A. (2015)“Going deeper with convolutions.” in Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[6] Bobić, V., Tadić, P., &Kvaščev, G. (2016) “Hand gesture recognition using neural network based techniques.” in Neural Networks and Applications (NEUREL), 2016 13th Symposium on (pp. 1-4). IEEE.

[7] Ballester, P., & deAraújo, R. M. (2016) “On the Performance of GoogLeNet and AlexNet Applied to Sketches.” in AAAI

[8] Dawwd Sh. (2019). High Performance Colored Image Segmentation System Based Neural Network, Al-Rafidain Engineering Journal, IRAQ, vol.17, No. 2, pages 1-10.

[9] Dawwd Sh., Mahmood B.(2019). A reconfigurable interconnected filter for face recognition based on convolution neural network, 4th IEEE international Workshop for Design and Test, Riyadh, Saudi Arabia, ISBN: 978-1-4244-5748-9.

[10] Wang H., Wang Y., Cao Y. (2019). Video-based Face Recognition: A Survey, World Academy of Science, Engineering and Technology, December 2009, Issue 60. 1307- 6892.

[11] B. Zhan, D. N. Monekosso, P. Remagnino, S. a. Velastin, and L.-Q. Xu, (2018) "Crowd analysis: a survey," Machine Vision and Applications, vol. 19, no. 5-6, pp. 345–357.

[12] J. Yim, J. Ju, H. Jung, and J. Kim, (2017) "Image Classification Using Convolutional Neural Networks With Multi-stage Feature," in Robot Intelligence Technology and Applications 3. Springer International Publishing, pp. 587–594.

[13] J. Shao, K. Kang, C. C. Loy, and X. Wang, (2018) "Deeply Learned Attributes for Crowded Scene Understanding," in Computer Vision and Pattern Recognition.

[14] P. Wang, Y. Cao, C. Shen, L. Liu, and H. T. Shen, (2015) "Temporal Pyramid Pooling Based Convolutional Neural Networks for Action Recognition," Arxiv, pp. 1–8.