Scientific Journal of Impact Factor (SJIF): 5.71

E-ISSN (O): 2348-4470
P-ISSN (P): 2348-6406

International Journal of Advance Engineering and Research Development

Volume 8,Issue 11,November 2021

Large-scale Video Classification with Convolutional Neural Networks

Prathamesh Kshirsagar[1], Pooja Nagawade[2]

[1]*Computer Science, Bharati Vidyapeeth College Of Engineering, Lavale, Pune.*
[2]*Computer Science, Maharashtra Institute Of Technology, Pune.*

Abstract: Convolutional Neural Networks (CNNs) have acquired a strong reputation as an image recognition model class. As a result of these findings, we give a comprehensive empirical evaluation of CNNs for large-scale video classification using a fresh dataset of 1 million YouTube videos classified into 487 classes. We study a range of strategies for extending the time domain connectivity of a CNN in order to use local spatio-temporal information. We discuss the limitations of current training methods and propose a multiresolution, foveated architecture as a possible technique of expediting training. When compared to strong feature-based networks, our top spatio-temporal networks outperform them significantly. when compared to single-frame models (59.3 percent), however this is only a marginal improvement (55.3 percent to 63.9 percent). 60.9 percent in total). We delve deeper on the generalisation performance. Retrain our best model's top layers on the UCF101 Action Recognition dataset and observe considerable performance improvements over the UCF-101 baseline. prototype (63.3 percent up from 43.9 percent).

Keywords- Acoustic Event Detection, Acoustic Scene Classification, Convolutional Neural Networks, Deep Neural Networks, Video Classification.

I. INTRODUCTION

The prevalence of images and videos on the internet has prompted the creation of algorithms that can analyse their semantic information for a variety of purposes, including search and summarization. Convolutional Neural Networks (CNNs) [15] have recently been shown to be an effective class of models for comprehending visual content, providing state-of-the-art results on image recognition, segmentation, detection, and retrieval [11, 3, 2, 20, 9, 18]. Techniques for scaling up networks to tens of millions of parameters and enormous labelled datasets that can help the learning process were significant enabling aspects for these outcomes. CNNs have been proven to learn powerful and interpretable visual characteristics under these settings [28]. We investigate the effectiveness of CNNs in large-scale video classification, where the networks have access to not only the appearance information provided in single, static photos, but also their complicated temporal evolution, as a result of positive findings in the domain of images. Extending and implementing CNNs in this context presents a number of issues. Because videos are substantially more difficult to gather, annotate, and store, there are currently no video classification standards that match the scope and variety of existing picture datasets. To get enough data to train our CNN architectures, we created the Sports-1M dataset, which contains 1 million YouTube videos organised into 487 sports taxonomies. Sports-1M is made available to the research community in order to support future work in this field. We're looking for answers to the following questions from a modelling standpoint: What pattern of temporal connectivity in a CNN architecture is ideal for exploiting local motion information in the video? How does the added motion information affect a CNN's predictions, and how much does it increase overall performance? We test these hypotheses empirically by comparing numerous CNN designs, each of which takes a distinct approach to mixing data across time. In terms of computation, CNNs necessitate a significant amount of training time in order to successfully optimise the millions of parameters that define the model. This problem is exacerbated when the architecture's connectivity is extended in time, because the network must handle not just one image, but several frames of video at once. To address this problem, we show that changing the architecture of CNNs to include two separate streams of handling can greatly enhance their runtime performance: a context webcast that learns characteristics on low-resolution frames and a high-resolution fovea stream that only operates on the middle portion of the frame.

II. MOTIVATION

This dataset is the largest dataset accessible in the field of video classification in terms of the number of videos. Efficient Video Classification on a Large Scale Video Researchers are putting in a lot of effort because of the success of CNNs

III. PROBLEM STATEMENT

We look at a variety of methods for extending a CNN's connectivity in the time domain to take advantage of local spatio-temporal information, and we recommend a multiresolution, foveated design as a viable option to speed up training.

IV. LITERATURE SURVEY

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis," cnn architectures for large-scale audio classification" [1] Convolutional Neural Networks (CNNs) have been demonstrated to be extremely good at image classification, and they also demonstrate promise in audio. We use several CNN architectures to classify the soundtracks of 70 million training videos (5.24 million hours) using 30,871 video-level labels. AlexNet [1, 2], VGG [2, 3], Inception [3, 4], and ResNet [5] are all Deep Neural Networks that are fully connected (DNNs). We conduct experiments with various training sets and label vocabulary sizes and discover that analogues of CNNs used in image classification perform well on our audio classification task, and that larger training and label sets improve the task up to a point. On the Audio Set [5] Acoustic Event Detection (AED) classification test, a model constructed from these classifiers' embeddings outperforms raw features considerably.

JINZHUO WANG; WENMIN WANG; WEN GAO "MULTISCALE DEEP ALTERNATIVE NEURAL NETWORK FOR LARGE-SCALE VIDEO CLASSIFICATION"[2], VIDEO CATEGORIZATION HAS BECOME A DEMANDING AND COMPLEX STUDY AREA DUE TO THE RAPID EXPANSION IN THE AMOUNT OF MULTIMEDIA DATA. WHEN COMPARED TO IMAGE CLASSIFICATION, VIDEO CLASSIFICATION NECESSITATES MAPPING HUNDREDS OF FRAMES TO SEMANTIC TAGS, WHICH PRESENTS NUMEROUS HURDLES FOR ADVANCED MODELS DEVELOPED FOR IMAGE-ORIENTED TASKS. CONTINUOUS FRAMES IN A VIDEO, ON THE OTHER HAND, PROVIDE US WITH MORE VISUAL INFORMATION THAT WE MAY USE TO IMPROVE CLASSIFICATION. THE CONTEXT IN THE SPATIOTEMPORAL DOMAIN IS ONE OF THE MOST CRUCIAL CLUES. WE DESCRIBE THE MULTISCALE DEEP ALTERNATIVE NEURAL NETWORK (DANN) IN THIS RESEARCH, AN UNIQUE ARCHITECTURE THAT COMBINES THE BENEFITS OF BOTH CONVOLUTIONAL AND RECURRENT NEURAL NETWORKS TO CREATE A DEEP NETWORK CAPABLE OF COLLECTING RICH CONTEXT HIERARCHIES FOR VIDEO CATEGORIZATION.

QITING YE; ZHAO LUO; XIAOBING XIAO; SHIMING GE," GELOGO: DETECTING TV LOGOS FROM WEB-SCALE VIDEOS",[3] DETECTING SOME SPECIFIC TRADEMARKS IS AN EFFICIENT APPROACH TO DETERMINE THE SOURCE OF A WEB VIDEO (E.G. A TV LOGO OR A BRAND). IN THIS STUDY, WE PROPOSE THE "GELOGO" LOGO DETECTING SYSTEM FOR WEB-SCALE FILMS, WHICH CONSISTS OF FOUR PRIMARY ELEMENTS. WITH SHOT SEGMENT DETECTION, THE KEYFRAME MODULE PULLS MANY FRAMES INITIALLY. THE DATA CAN BE DRASTICALLY REDUCED IN THIS WAY. THE PROPOSAL MODULE THEN USES AN ENSEMBLE OF PRETRAINED LOGO DETECTORS TO LOCATE CANDIDATE LOGOS FROM THE RETRIEVED FRAMES. THE GEOGRAPHIC VERIFICATION MODULE THEN TAKES THE LOGO PROPOSALS AS INPUT AND PERFORMS THE CLASSIFICATION TASK WITHIN A RESNET NETWORK TO DETERMINE WHETHER THEY ARE REAL LOGOS OR NOT. FINALLY, THE TEMPORAL VERIFICATION MODULE CHECKS THE TEMPORAL CONSISTENCY OF LOGO LOCATIONS TO IDENTIFY THE DETECTION FINDINGS. EXPERIMENTS ON A LARGE-SCALE TV LOGO VIDEO DATASET REVEAL THAT THE SUGGESTED APPROACH CAN RECOGNISE LOGOS IN VIDEOS WITH A DETECTION ACCURACY OF 99.1% AND A SPEED OF 25 FRAMES PER SECOND.

DONGDONG ZENG; MING ZHU "MULTISCALE FULLY CONVOLUTIONAL NETWORK FOR FOREGROUND OBJECT DETECTION IN INFRARED VIDEOS",[4] DUE TO ITS IMPORTANCE IN IR TARGET RECOGNITION, IR PRECISION NAVIGATION, IR VIDEO SURVEILLANCE, AND OTHER APPLICATIONS, ACCURATE AND RAPID INFRARED (IR) FOREGROUND OBJECT DETECTION IS ONE OF THE MOST CRITICAL TOPICS TO BE SOLVED. BACKGROUND SUBTRACTION, WHICH SEEKS TO DISCOVER FOREGROUND OBJECTS BY BACKGROUND MODELLING, IS A COMMON SOLUTION FOR SUCH TASKS. MANY BACKGROUND SUBTRACTION APPROACHES HAVE BEEN PROPOSED SO FAR, AND THEY HAVE ALL PERFORMED WELL. HOWEVER, DUE TO THE UNIQUE PROPERTIES OF IR IMAGES, ONLY A FEW METHODS ARE ADEQUATE FOR DETECTING IR FOREGROUND OBJECTS. MANY VISION TASKS, SUCH AS CLASSIFICATION AND IDENTIFICATION, HAVE RECENTLY SHOWN TREMENDOUS EFFECTIVENESS USING FEATURES ACQUIRED FROM CONVOLUTIONAL NEURAL NETWORKS (CNNS).

YANYAN FANG; BIYUN ZHAN; WANDI CAI; SHENGHUA GAO; BO HU " LOCALITY-CONSTRAINED SPATIAL TRANSFORMER NETWORK FOR VIDEO CROWD COUNTING"[5] WHEN COMPARED TO SINGLE IMAGE-BASED CROWD COUNTING, VIDEO PROVIDES SPATIAL-TEMPORAL INFORMATION ABOUT THE CROWD, WHICH HELPS IMPROVE THE ROBUSTNESS OF CROWD COUNTING. HOWEVER, AS HUMANS ARE TRANSLATED, ROTATED, OR SCALED, THE DENSITY MAP OF THEIR HEADS CHANGES BETWEEN NEIGHBOURING FRAMES. MEANWHILE, CHANGES IN HEAD COUNTS OCCUR AS A RESULT OF INDIVIDUALS ENTERING/EXITING OR BEING OCCLUDED IN DYNAMIC SCENARIOS. TO SOLVE THESE ISSUES IN VIDEO CROWD COUNTING, A LOCALITY-CONSTRAINED SPATIAL TRANSFORMER NETWORK (LSTN) IS DESIGNED. TO BE MORE PRECISE, WE BEGIN BY ESTIMATING THE DENSITY MAP FOR EACH FRAME USING CONVOLUTIONAL NEURAL NETWORKS.

V. SYSTEM ARCHITECTURE

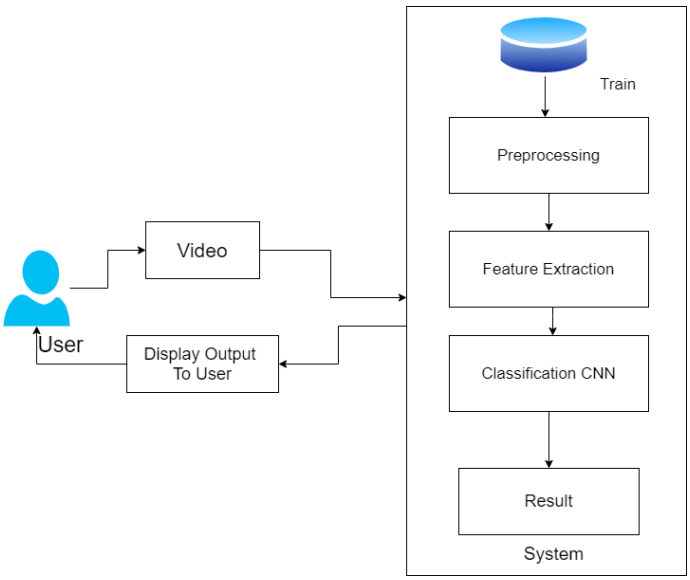

Fig. system architecture

- **Modules :**

Pre-processing : - Although geometric changes of images (such as rotation, scaling, and translation) are known as pre-processing methods, the purpose of pre-processing is to improve the image data by suppressing undesired distortions or improving particular image attributes that are required for future processing. Possessiveness is a term used to describe someone who is possessive about something. The use of a digital computer to perform an algorithm on digital photographs is known as image processing. As a subsection or area of digital logic, image processing provides a number of advantages over analogue image processing.

1. Read the Image

2. Image Resized (220,220, 3)/Resized (width, height, no. RGB channels)

3. Conversion of RGB to Grayscale

4. Identification of segmentation edges The Gaussian filter is used to remove noise.

Segmentation : It entails segmenting a visual input to facilitate picture analysis. We can segment the image and analyse it further if we wish to eliminate or identify something from the remainder of the image, for example, detecting an object in the background. This technique is referred to as segmentation. Segments are composed of "super-pixels," which are clusters of pixels that represent objects or segments of objects.

Feature Extraction : Complex structures, such as points, edges, or objects, can be found in an image. Feature extraction is a technique for reducing the amount of features in a dataset by creating new ones from old ones (and then discarding the original features). The new, condensed set of features should be able to summarise the bulk of the information in the old set. Feature extraction begins with a set of measured data and creates derived values (features) that are intended to be useful and non-redundant, easing the learning and generalisation phases and, in some situations, resulting in superior human interpretations. Dimensionality reduction is linked to feature extraction.

Classification : CNNs are utilised for picture detection and recognition because of their high precision. The classification convolutional neural network is three-dimensional, with each group of neurons analysing a different area or "function" of the image. In a CNN, each group of neurons concentrates on a distinct aspect of the image. Smaller sections of the photos are examined by the algorithm. The end result is a probabilistic vector that predicts the likelihood of each feature in the image belonging to a class or group.

VII.Algorithm

CNN (Convolutional Neural Network): CNN (convolutional neural network) is a type of neural network for deep learning. In a word, consider CNN to be a machine learning algorithm capable of taking an input image, assigning significance (learnable weights and biases) to various aspects/objects inside it, and discriminating between them. CNN works by pulling information from videos. A CNN is composed of the following elements:

1. The input layer is a grayscale image.

2. The output layer, which may be binary or multi-class in nature.

3. The hidden layers consist of convolutional layers, ReLU (rectified linear unit) layers, pooling layers, and a fully connected Neural Network. It is crucial to understand that ANNs, or Artificial Neural Networks, are incapable of extracting characteristics from images since they are composed of numerous neurons.

This is where the combination of convolution and pooling layers comes into play. Likewise, the convolution and pooling layers are unable to do classification, necessitating the usage of a fully connected Neural Network. Before we delve deeper into the principles, let us first attempt to comprehend these particular pieces.

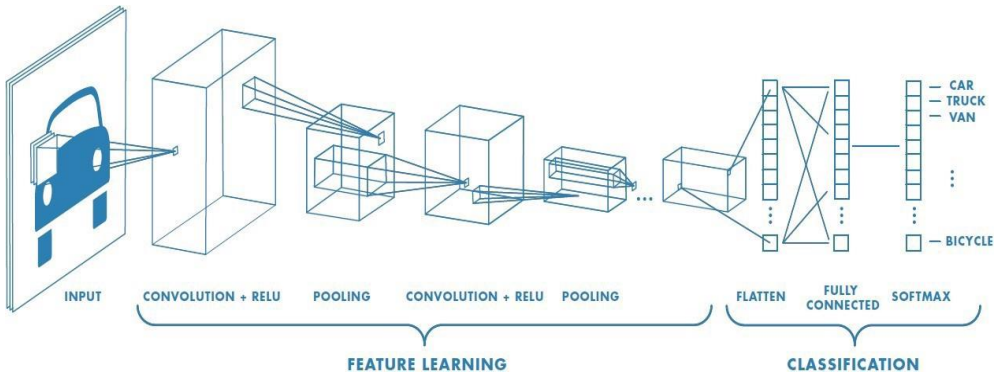

Fig - Illustrates the CNN process from input to Output Data

VI. CONCLUSION

We examined the performance in a broad video classification of convolutional neural networks. We discovered that CNN architectures have the potential to obtain power from slightly labelled data, which outperforms functional approaches, and that the changes in design connections throughout the period are surprisingly resilient. Qualitative network output analysis and confusion matrices disclose interpretable defects. Our findings show that although the architectural information of the time dependent connectivity is not crucial, an early and late fusion alternative is consistently outperformed by the Slow Fusion model. Surprisingly, we show that a single frame model performs well, which reveals that even for a dynamic dataset such as Sports, local movement indications may be less relevant than before. Another option is to have better control over camera motion (e.g. by removing the features of a tracked point in the local coordinate system, as proven in [25]), but this will need substantial changes in the CNN architecture, which we will leave for further research. In addition, the combination of a low-resolution context with a fovea stream of high-resolution was found to be a viable method for accelerating CNNs without compromising precision.

REFERENCES

[1] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In Human Behavior Understanding, pages 29– 39. Springer, 2011. 2, 3

[2] D. Ciresan, A. Giusti, J. Schmidhuber, et al. Deep neural net works segment neuronal membranes in electron microscopy images. In NIPS, 2012. 1

[3] L. N. Clement Farabet, Camille Couprie and Y. LeCun. Learning hierarchical features for scene labeling. PAMI, 35(8), 2013. 1, 2

[4] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor semantic segmentation using depth information. Internatinal Conference on Learning Representation, 2013. 2

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In CVPR, volume 1, 2005. 5

[6] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Y. Ng. Large scale distributed deep networks. In NIPS, 2012. 4

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei Fei. Imagenet: A large-scale hierarchical image database. In CVPR, 2009. 2

[8] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behav- ´ ior recognition via sparse spatio-temporal features. In Inter national Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2, 5

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich fea ture hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014. 1, 2

[10] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. PAMI, 35(1):221– 231, 2013. 2, 3

