International Journal of Future Generation Communication and Networking

Vol. 13, No. 3s, (2020), pp. 196–204

A Novice Approach of Hybrid Transfer Learning for Video Classification

Sourav Joshi, Ameya Karhadkar, Niranjan Thatte, Kunwar Chopra, Tanaji Khadtare

Computer Engineering Department, Final Year Engineering SavitriBai Phule University

sourav.intellectual@gmail.com

ameya.karhadkar@yahoo.com

thatteniranjan0@gmail.com

kunwar.satara11@gmail.com

tanajikhadtare@gmail.com

Abstract

One of the very interesting data modalities is video. From a dimensionality and size perspective, videos are one of the most interesting and intuitive data types which enable fast and easy object recognition and learning. Video classification is an important task for archiving digital contents for various video service providers. Video uploading platforms such as YouTube are collecting enormous datasets, empowering Deep Learning research. Video being an important source to recognize any activity by the humans, video classification becomes an important and critical job for video service providers. The survey paper studies various deep learning, transfer learning and hybrid model approaches.

I. INTRODUCTION

Over a long period of time, Researchers applied many approaches to analyse scene context and classify the visual information. Recently, deep learning-based models have become increasingly popular for complex tasks like these. There are various drawbacks associated with the traditional techniques but with the availability of tools and technologies like Transfer Learning, Millions of videos flood the data servers daily including online Television as well as other sources.

Various Methods and approaches to classify a video have been discussed in the survey paper. Since a video sequence is analysed to classify it, this work is related to the analysis of the context of a scene; thus, it can be further applied to advance high-level understanding of video scenes in machines. Deep neural networks, deep belief networks and recurrent neural networks have been applied to fields such as computer vision, speech recognition, natural language processing, audio recognition, social network filtering, machine translation, and bioinformatics where they produced results comparable to and in some cases better than human experts have. Large datasets like Youtube-8M [1] have been considered using CNN for video classification. Hybrid Models which use RNN along with CNN have been discussed in the survey paper. The paper also discuss the 3D CNN classification in brief.

II. LITERATURE SURVEY

The previous chapter talked about the insights of the basics of sports classification. This chapter gives better insights on the project via the analysis done on various research papers related to sports classification.

ISSN: 2233-7857 IJFGCN

2.1. Deep Learning Video Classification

Andrej Karparthy et.al. [1] proposes the use of CNN to classify videos. For this purpose, author picked up 1M YouTube videos belonging to 487 classes of sports. The author talks about temporal activity pattern in CNN architecture to take advantage of local motion information present in video and how does this additional motion information Influence the prediction. The architecture processes inputs at two spatial resolution a low-resolution context stream and high-resolution stream as a promising way of improving the runtime performance of CNN at no cost in accuracy. The network is capable of learning powerful features from weakly labelled data and also on dynamic data set.

YouTube 8M is discussed by Nisarg Kothari et.al. [2], A dataset specifically created for video classification. With YouTube-8M, the goal is to advance the field of video understanding, similarly to what image datasets which are largest in size have done for image understanding. The paper talks about challenges faced during video classification like the problem of quality labels and also the computational barriers which are proposed to be solved by providing a state-of-the-art frame level feature. The strategy used is that each video is decoded at one-frame-per-second using Deep CNN [3] pre-trained on ImageNet to extract the layer prior to the classification layer. Hence this paper talks about Youtube-8M, a manually created dataset and uses Deep CNN for classification. However, the scope of the dataset is huge and acquiring the results on a single machine using a single framework is difficult.

A model that is proposed by Mohmmad Ashraf et.al. [7] is a combination of Convolutional and Recurrent network. The convolutional part has four layers: the first layer consists of 32 features map with 3*3 respective field and is followed by max pooling with stride the next three layers are 3 different kinds of dilated convolutional layers which we can describe as a context module. To solve the sport classification problem, author uses relation between human action sequence and their surrounding environmental context. The dataset used here contains of total 300 video sequence where each sequence contains 64 frames. 80 percent of dataset was chosen for training and 20 per cent for testing randomly. Improved CNN and deep learning techniques is proposed by Ou Ae et.al. [8] for scene identification for complex scenery of coal mine with increasing the depth of CNN. The restructured model consists of 10 Layers of neurons, including 7 coiling layers, 5 pool layers, 3 full connection of mine video scene. It uses ALexNet model to structure and get a kind of mine video classification algorithm which is suitable for complex background. The convolutional layer is used to extract features of the mine.

H.Wei et.al. [19] discuss about a video classification technique which can detect and classify persons in a video data that is captured from several miles away. Two approaches are discussed in this namely histograms of oriented gradients classifier and a support vector machine classifier along with an Adaboost Classifier. For classification purpose, a Convolution Neural Network with transfer learning is used to detect the person of interest. The scope of the video frame discussed is a 3 mile distance taking into account various environmental factors like wind, dust, etc. A deep learning framework in a weakly supervised learning environment is discussed by Jiajun Wu. et.al. [17]. While much has been talked about deep learning on supervised tasks, a very little has been discussed about deep learning on weak instances of supervision. Human labels play a very important role in this framework. A valid point is discussed that most of the data available on world web doesn't have labels, hence these weakly supervised methods can be preferred for video classification than the traditional methods. Various search engines like Google, Bing can perform classification in an unsupervised way; however the keywords of the images might not be necessarily accurate. Hence, the three fold approach is used in the proposed framework.

ISSN: 2233-7857 IJFGCN

Mounira Hmyad et.al.[13] propose a model for automatic identification of the TV stream for multimedia information. The approach used is a spatiotemporal approach to identify the programs in a TV stream using deep learning as a two-step approach. The database used for the purpose is a video of visual jingles that is handcrafted for training. While video classification the same jingles are used to identify the programs in TV stream. The underlying theme of the approach is the use of the sparse auto-encoder as a feature extractor of visual jingles for training a video category classifier. This feature extraction governs the structuring and detection of the generics (i.e. the starting points) of programs in a TV stream. Basically, an auto-encoder is simply a multi-layer feed forward neural network trained to represent the input with back-propagation. By applying back-propagation, the auto encoder tries to decrease the discrepancy as much as possible between input and reconstruction by learning an encoder and a decoder.

Jing Li et.al. [11] propose a training model for a 3-Dimensional CNN for complex classification tasks with respect to videos. The method is relatable to exclusion principle of the human judgement. It is a combination of two-class classifiers providing the ability to recognise unknown 3D model to the classifier. The video-set is pre-processed after downscaling at 8 Hz, converting RGB to grey image and resizing to 60*80 pixels. For training, three videos are selected randomly from each class as the training class for the training videos. Each training video is down-sampled into several sub-videos and each sub-video has the same number of frames. The testing videos are made up of the remaining videos in each class and 120 videos with no class and thus down-sampling them.

2.2 Hybrid Models

The paper focuses on recognition, detection, segmentation and retrieval solution for image recognition and then the processing of the image. It uses Convolutional Neural Networks (CNN) as its base. In the paper, Joe Ng et. al. [3] proposed and evaluated several deep neural network architectures to combine image information over longer time periods than previously attempted versions. For this, some methods have been proposed which includes Convolutional temporal feature pooling architecture and focuses on examining the various design choices which are needed when adapting a CNN for this task and the Long Short-Term Memory (LSTM) cells architecture which proposes and explicitly models the video as an ordered sequence of frames, the only disadvantage being the slow and tardy processing speed.

Videos serve to be one of the most efficient and easiest approaches for understanding different things as stated by Tian H et. al. [10] in IEEE Conference in China. These may include food, storm, and animals. The authors of the paper present a multimodal deep learning framework to improve video concept classification by combining various platforms together with the recent advances in transfer learning and sequential deep learning models. Long Short Term Memory (LSTM) Recurrent Neural Networks (RNN) models are then used to improve efficiency. The proposed framework is applied to a disaster-related video dataset that includes not only disaster scenes, but also the activities that took place during the disaster event. The experimental results show the effectiveness of the proposed framework. An improvement technique called the Convolutional Drift Network (CDN) is used. CDNs produce per-frame appearance features from video using a deep CNN, and those features are pushed into a randomly initialized network as proposed by Dillon Graham et. al. [4]. Compressed videos bit stream which are processed by video decoder are used for the purpose of classification rather than general approach of using images.

The main idea proposed by Aaron Chaddha et.al. [20] is to use the Illumination changes that occur during object motion In pixel and the way they are stored in stream. Two stream CNN is used, one stream recognises textures in image and other computes motion vector which deals with

ISSN: 2233-7857 IJFGCN

correspondence between two images and output of these two is fused together to predict final result. Author states that the proposed system is faster and more cost efficient in terms of using computing resource than systems based upon optical flow estimations.

Another approach for low complexity video classification is proposed by Ifat Abramovich et.al. [18]. This methodology gives simple yet effective way to classify videos. InceptionV3 is used to extract necessary features from images and then RNN is used in classifying task. Dataset here is subset of YouTube 8M dataset which is widely available for free. After successful implementation of the proposed model it is seen that it works fine on dataset with less categories but its performance starts to dive low when categories are increased. Model is trained with 5 categories and 1000 videos out of available dataset and remaining is used to fine tuning. It uses 2D CNN to learn spatial features and RNN for temporal features, LSTM (Long Short-Term Memory) is used to store feature vectors and access them rapidly. This paper proposes a model which works on images and audio to extract features from dataset and use their correlation in classification. Inception V3 is used to extract feature from images and MFCC to extract features from audio, these features are stored in feature vectors and used as input together to model. Two different datasets are integrated together one is YIL-MED (Multimedia event detection) dataset and other is curated from YouTube by author Junghon lee et.al. [14] itself.

2.3 Transfer Learning Approaches

Sorin Jurj et.al. [5] propose a design for identifying and classifying the Romanian traditional motifs found on 4 different categories by training a Convolutional Neural Network (CNN) model derived from the Residual Network(ResNet-50) architecture. The author claims to have implemented a system which can detect and identify through a webcam if the object in front of it contains a learned motif. The author talks about various DNN techniques have been using it traditionally. Various existing methods of CNNs with regard to clothing detection and classification are implemented. It also introduces a system design for identifying and classifying the Romanian traditional motifs as discussed earlier. The system inherits the advantages of the ResNet-50 architecture which includes higher accuracy and faster training performance regarding image classification. The model is trained using Keras framework, a Tensorflow high-level API written in Python and integrated in the proposed detection and identification system. For training and processing the features detected in the hidden CONV layers, a CPU as well as a high-performance GPU are used. Using a webcam, these detected features (motifs) are identified by the proposed CNN. The proposed model on a widely known academically dataset called ImageNet using a modified ResNet-50 architecture.

Similarly another paper stated by Sai Bharadwaj Reddy et.al. [2019][6] talks about using Transfer Learning using ResNet-50(1) for Malaria Cell Image Classification. The author talks about Malaria disease which is caused by single-celled parasite of plasmodium group. It also reveals the statistics which shows the worldwide deaths caused by Female Anopheles mosquito. In spite of many advanced evaluation techniques for identifying the infection there is a chance that wrong diagnostic decisions are at times taken. According to the author these Deep-learning based classification of cell images can prevent this. The Dataset used is from the official website of National Library of Medicine (NLM).

ISSN: 2233-7857 IJFGCN

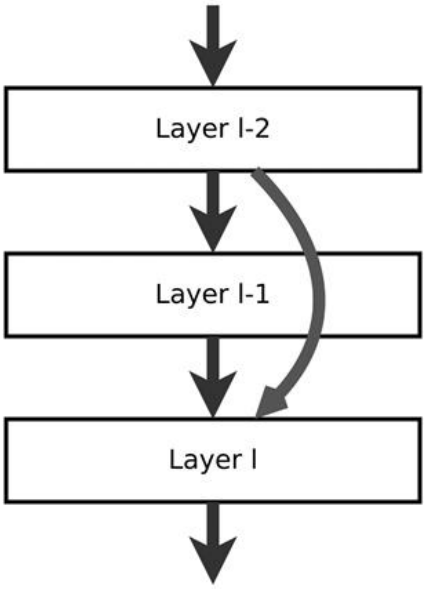

Fig 1- Resnet-50 Architecture

The author takes into account 3 models Oxford VGG model, Google Inception Model and Microsoft Resnet-50 Model and then proposes a model which takes a RGB image as an input. The image will enter the ResNet50 Layer and passes through its multiple layers and provides us the results. Hence the authors claim the use of Transfer learning for malaria image classification has brought good results even without using any modern hardware such as GPU's or Tensor Processing Unit's (TPU's). Use of this modern hardware might increase the accuracy and can bring down run time to a great extent. Other than the ResNet50 the models discussed such as Google Inception or Oxford's VGG should also be used on the malaria cell image dataset.

Ling Shao et al. [2015][9] propose that Machine learning and deep learning scientists as well as the researchers while model building and studying the data for future analysis make a big assumption at the start-'The future data are within the same feature space or have the same distribution as the training data itself'. However the reference to the same doesn't guarantee that the over fitting problem. In the real world applications along with target future reference data related data can also be included to expand the over-all project scope. This paper using transfer learning addresses such cross-domain learning problems by extracting useful information from data in a related domain and using them in target tasks. Some typical problems, e.g., view divergence in action recognition tasks and concept drifting in image classification tasks, can be efficiently solved. The survey paper claims state-of-the-art transfer learning algorithms in visual categorization applications, such as object recognition, image classification, and human action recognition.

Inad Aljarrah and Duaa Mohammad et.al. [16] talk about video analysis by using convolutional neural networks. Reviewing videos recorded by video surveillance systems is an area where video content analysis can be handy. The authors put forth the idea of instead of rewinding over hours of recorded video to spot an action, an automated content analysis system that produces a searchable text file that summarizes the video content. Video Content Analysis (VCA) is the method of automatically analyzing video streams to detect and determine temporal and spatial events. The paper introduces an intelligent video search technique -the VGGNet convolution neural network. Yuxi Hong et.al.[15] and his team propose an efficient method for video scene and frame event classification on the basis of transfer learning. The sport they have chosen is soccer(football as is called in many countries) where the author has carried out soccer frames semantic analysis using CNN model. The author proposes methods to solve these two tasks separately. In order to solve two tasks at the same time and improve

ISSN: 2233-7857 IJFGCN

the efficiency of video processing, he treats them as one end-to-end classification task. He further introduces a new Soccer Video Scene and Event Dataset (SVSED) for the project. 3D visualization show enhancement in accuracy of the classification of the dataset.

The author Mohammad Ashraf Russo et. al. [12] narrates the importance of sports video classification in the sci-fi modern world era. Archiving of the digital content is a task of immense importance to the video broadcasters. As a result, accuracy and security are the two major goals of the service providers. The author uses deep learning along with recurrent networks imbibing transfer learning approach models to classify over 15 sports. Later, transfer learning is applied with the VGG-16 model which was able to achieve 94% and 92% test accuracy for 10 and 15 sports classes respectively. By reaching 94% test accuracy it is seen that using transfer learning with deep convolutional networks like VGG is able to achieve greater levels of accuracy.

III. PROPOSED METHODOLOGY

To overcome the shortcomings of all the techniques and methodologies we have proposed a new methodology for Video Classification which is based on Resnet50 and Transfer Learning approach which is discussed for malaria cell [6]. Video Classification if implemented only through CNN gives a problem of prediction flickering i.e. For any sport the output is not stable and doesn't necessarily give correct label output. So, we have proposed a methodology where we can pass each and every frame through CNN in a loop. The model will be pre-trained with the images of the sports category which are hand-curated specifically to categorize the dataset on these 22 classes of sports on Resnet-50 and the model's output is expected to be more stable after fine-tuning with ResNet-50.

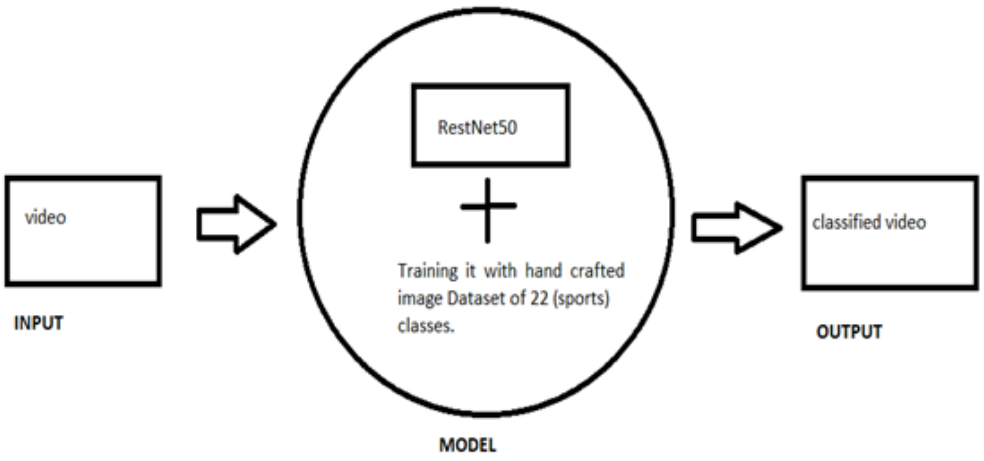

Fig 2- Block diagram of model architecture

IV. IMPLEMENTATION SPECIFICATION

We have taken the images as a loop over the frames in the video file. Then this frame is passed through our CNN and independent classification of the video is done according to the labels of the classes of sports. The images as described above are trained on the ResNet-50 architecture and then the output of the frame is written on the disk. The graph and the sample output generated are shown in the figure:

ISSN: 2233-7857 IJFGCN

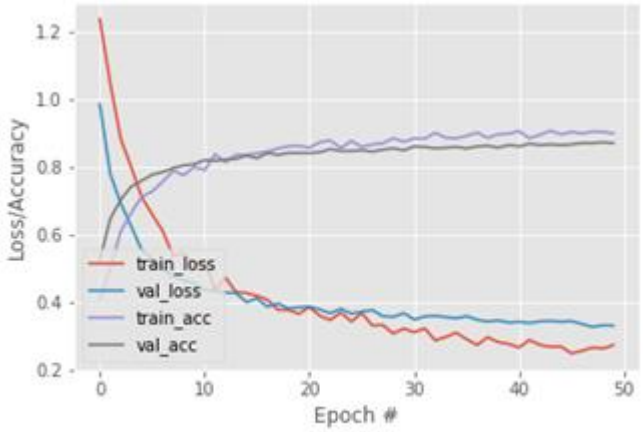

Fig 3- Training Loss and Accuracy on Dataset

V. JUSTIFICATION PROPOSAL FOR EVALUATION

Evaluation at any stage plays an important role in improving and enhancing the development, implementation and analysis of a project. As mentioned above, we have proposed a hybrid model comprising of Resnet-50 model incorporating transfer learning. Traditionally, videos were examined and analysed as the main source for training of the dataset. This enclosed a complex temporal activity pattern in CNN resulting in flickering and less accurate output model. Our model on the other hand is proposed to be pre-trained on the basis of a hand-scripted image dataset. We primarily focus on sports classes at the early stage of the project. The dataset comprises of 22 sports classes comprising of around 15000 images.

Also, it can be inferred that, some methodologies suggest video understanding through image datasets using deep learning approaches as base. Rensnet-50 is a Convolutional Neural Network which is 50 layers deep and is a complex network trained on more than a million images. It is based on the ImageNet Database. Hence, Resnet-50 can be evaluated as one of the most powerful tools in image classification. In accordance with transfer learning research problem, that focuses on storing and implementing of knowledge gained while solving previous problem to different related problems, Resnet-50 has reduced the time and cost required for the designing, training and implementation of the model. Accuracy in terms of f1-score can be evaluated as 0.87 whereas avg to recall is 0.86. Thus, results obtained of our proposed model are as follows:

	Precision	Recall	F1-score	Support
Accuracy			0.87	518
Macro avg	0.89	0.86	0.86	518
Weighted avg	0.88	0.87	0.87	518

Classification Results

VI. CONCLUDING REMARKS

After analysing a number of papers, we conclude that video classification carries immense importance for video service providers. Over the years, researchers have applied many approaches to analyse scene context and classifying visual information. Recently according to these video classification approaches, hybrid models incorporating deep learning-based models have become increasingly popular for complex tasks. This paper thus provides a critical analysis of video classification techniques of deep learning, transfer learning and hybrid models.

ISSN: 2233-7857 IJFGCN

REFERENCES

[1] Andrej Karpathy , George Toderici, Sanketh Shetty *Large - Scale video classification with Convolutional Neural Network (CNN)*" : 2018, Stanford University

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, Sudheendra Vijayanarasimhan "*YouTube-8M: A Large-Scale Video Classification Benchmark*", 2016

[3] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, George Toderici "*Beyond Short Snippets: Deep Networks for Video Classification*", 2016

[4] Dillon Graham, Seyed Hamed Fatemi Langroudi, Christopher Kanan, Dhireesha Kudithipudi "*Convolutional Drift Networks for Video Classification*", 2017, Published in IEEE rebooting computing

[5] Sorin Liviu Jurj, Flavius Opritoiu, Mircea Vladutiu, ``*Identification of Traditional Motifs using Convolutional Neural Networks*", 2018 IEEE 24th International Symposium for Design and Technology in ElectronicPackaging (SIITME)

[6] A. Sai Bharadwaj Reddy and D. Sujitha Juliet , ``*Transfer Learning with ResNet-50 for Malaria Cell-Image Classification.*", 2019, International Conference on Communication and Signal Processing,

[7] Mohammad Ashraf Russo, Alexander Filonenko, Kang-Hyun Jo, ``*Sport Classification in Sequential Frames Using CNN and RNN*", 2018, Graduate School of Electrical Engineering, University of Ulsan, Ulsan, Republic of Korea

[8] Ou Ye, Yao Li, Guimin Li, Zhanli Li, Tong Gao, Tian Ma, ``*Video scene classification with complex background algorithm based on improved CNNs.*", 2018, School of Computer Science and Technology Xi'an University of Science and Technology, Xi'an , China

[9] Ling Shao, Senior Member, IEEE, Fan Zhu, Student Member, IEEE, and Xuelong Li, ``*Transfer Learning for Visual Categorization: A Survey*", 2015, IEEE Conference

[10] Tian, H., Cen Zheng, H., & Chen, S.-C, ``*Sequential Deep Learning for Disaster-Related Video Classification*", 2018, IEEE conference

[11] Jing Li, ``*Parallel Two-Class 3D-CNN Classifiers for Video Classification*", 2017, International Symposium on Intelligent Signal Processing and Communication System, Shandong Management University, Jinan, China

[12] Mohammad Ashraf Russo, Laksono Kurnianggoro, & Kang-Hyon Jo, `` *Classification of sports videos with combination of deep learning models and transfer learning*", 2019, International conference on Electrical, Computer and Communication Engineering

[13] Mounira Hmyada., Ridha Ejbali & Mourad Zaied, ``*Program Classification in a stream TV using Deep Learning*", 2017, International conference on Parallel and Distributed Computing, Application and Technologies

ISSN: 2233-7857 IJFGCN

[14] Jungheon Lee, Youngsan Koh & Jihoon Yang, ``*A Deep Learning based Video Classification System using Multimodality Correlation Approach*", 2017, in Sogang University, South Korea

[15] Yuxi Hong, Chen Ling, & Zuochng Ye, ``*End-to-End Soccer Video Scene and Event Classification with Deep Transfer Learning*", 2018, Tsinghua University, China

[16] Inad Aljarrah & Duaa Mohammad, ``*Video Content Analysis using CNN*", 2018, Jordan University of Science and Technology

[17] Jiajun Wu, Yinan Wu, & Kai Yu, ``*Deep Multiple Instance Learning for Image Classification and Auto-annotation*", 2015, Massachussets Institute of Technology

[18] Ifat Abramovich, Tomer Ben-Yehuda & Rami Cohen ``*Low Complexity Video Classification using RNNs*", 2018, Israel Institute of Technology

[19]H.wei, M. Laszewski & N Kehtarnavaz ``*Deep Learning based Person Detection and Classification for Far Field Video Sveillance*", 2018, University of Texas at Dallas

[20]Aaron Chaddha, Alhabib Abbas & Yiannis Andreopoulos ``*Video Classification with CNNs:Using the Codec as a Spatio-Temoral Activity Sensor*", 2019, IEEE transactions on Circuits and Systems for Video Technolgies

ISSN: 2233-7857 IJFGCN