International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.(02), March 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

Semantic Content Extraction using Modularised Tensor Flow Models

S.Sundar1, S.Yeshwant[2], T.Manonmani[3]

[1,2,3]*Computer Science and Engineering,*

Kamaraj College of Engineering and Technology,

Anna University, Tamil Nadu, (India)

ABSTRACT

A semantic object is a delineation of an accumulation of attributes to infer or depict an identifiable thing in a user environment. By utilizing semantic objects, we can bundle applications that mirror a particular situation. This paper examines a modularized work process with multiple TensorFlow models for extracting the semantic objects from images or edges of recordings. The work process incorporates the way towards preparing and extracting semantic contents by modularizing the whole procedure and utilizing them in parallel to foresee the inference. The trained objects are segregated into primary objects and first level objects. For category prediction, primary objects are utilized and first level objects are used for inference and scene prediction. The inference is predicted with the help of Ontology, based on the parameters defined for each inference. By considering a gathering of continuous edges in a video, a scene can be anticipated by defining parameters that have been satisfied by every scene. It is used for tagging the pictures or recordings, without the human help, in a faster and compelling way. By modularizing the training and extraction process, the whole computation time is lessened radically up to half.

Keywords— Deep Learning; Ontology; Parallel Computing; Scene Prediction; TensorFlow

I.INTRODUCTION

Images and videos make the most of the internet. Due to the rapid development in digital photography, storage and sharing technologies developed rapidly over the last ten years. So semantic based classification has emerged as an important field. In the future, more and more information will be stored in the digital form.

From microblogs to social media, every websites have huge number of photos and videos in them but they are classified only based on the description provided by the uploader. They are unstructured data that are stored in the original form without any processing. The images or videos need not be the same as described.

Semantic Content Extraction is the process of extracting an inference from the objects present in an image. Manual labelling of images and videos is an expensive and time consuming process. So with automated extraction of semantic content, we can label them in a faster and efficient manner.

TensorFlow is an open source library for numerical computation using data flow graphs. It allows to deploy computation to one or more CPUs or GPUs in any device. It was developed by Google Brain Team for Deep Neural Networks, but was used in many other domains. TensorFlow Object Detection API is a framework built

on TensorFlow that can be used to construct, train and deploy object detection models. It also offers stable models for feature extraction.

Single Shot Detection(SSD) is a bounding box framework used to detect objects in high fps video feeds. SSD framework determines all the bounding box probabilities in a single shot. When combined with MobileNet neural network architecture, it forms a powerful and efficient model.

The proposed work is to train and extract semantic content with the help of multiple TensorFlow models, so each model that is trained with unique set of objects are used to extract the objects in parallel to reduce the computation time. Based on the semantic content, the collection of digital photos and videos could be classified automatically. Predicting the scene in an effective and simplified manner is also discussed in this proposed work.

The rest of the research paper will unfold as follows: Section two will present related work and section three focuses on the approach, training, extracting, category predicting, inference predicting and scene predicting methodologies. The section four contains environmental setup and the evaluation results of the work carried out and section five concludes the research.

II. RELATED WORK

There are several proposed ways to reduce the computation power and time required to extract the objects. In [1], Hyoung Lee and Byung Lee proposed a middle output layer exit algorithm to reduce the computation time and power. But they achieved the time reduction by neglecting some training parameters, thus the output may not be accurate at all times. In[2], Chaitanya used colours and textures to retrieve the objects from images but partial objects are not detected at all times and it was implemented using MATLAB which is expensive to use. In[3], Kithmi Ashangani proposed a video search method by duplicating the shots and predicting. Duplicating is an expensive operation and the memory requirement is high. In[4], Junwei proposed prediction using the visual attention objects which we will use it in a further enhanced manner. In[5], proposed a face recognition system using SVM in MATLAB which detects only face. This can be further enhanced by adding different objects. In[6], Maryam surveyed techniques for detecting the action and gestures of humans which can be used for player identification. In[7], Li Wong surveyed deep learning techniques for a true smart city. In[8], Swapnali reviewed Ontology based content extraction. Ontology is the basis of inference prediction and the parameters should be defined properly to predict inference. In[9], Jiajun proposes video captioning by training the system with small videos of respective categories. We can achieve the same result by using image dataset instead of video set.

The above proposed ways may have some kind of drawbacks like either they are not accurate at all times or they do only a single part in entire semantic content extraction process. Most of the above proposed works use MATLAB. When they are improved and combined together, they form a powerful and efficient system.

III. METHODOLOGY

We will use the TensorFlow Object Detection API provided by [10] and the SSD MobileNet model. To prove, we are considering sports category. In sports category, we have taken objects related to football, basketball and volleyball games.

3.1 Approach

The first important step in semantic content extraction is training all the needed objects from the image dataset. The more efficient we train the system the more accurate the extraction. To modularise the training step, we will split the set of objects into small subsets and train them separately.

We have separated the superset of all objects into two subsets. Set A, which is the superset, consists of Players, Goal Post, Basketball, Football, Volleyball while Set B consists of Players, Goal Post and Set C consists of Basketball, Football and Volleyball, which are subsets of A.

SET A = SET B + SET C

After we split the objects into subsets, the images containing the respective objects are gathered and labelled. The system is trained with labelled images of each object in each set until TOTAL LOSS was less than 1.0 continuously. To prove the modularity, we even trained the SUPERSET of SET B and C, that is SET A, with all the labelled images. The total time to train and achieve the TOTAL LOSS less than 1.0 for superset was so much higher than the combined time of training subsets individually. This was achieved because of lesser objects in each SET needed to train at once.

Now the inference graphs for each training data is generated and used to extract the objects. We can filter the first level objects by specifying a threshold value and predict the category and inference from them. We have initialised 0.8 as threshold value for higher accuracy in prediction.

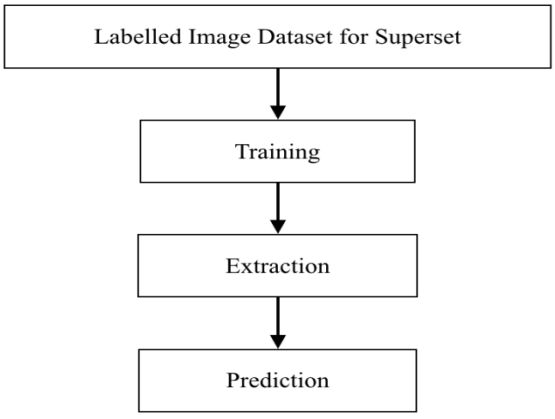

Fig 1

Existing Method

3.2 Extraction

Extraction is the most important part after the training. To extract the objects, we have taken all three graphs from all the training sets, to prove the approach. First we used the SET A's inference graph to extract the objects. It extracted all the trained objects in a specific time. Then we used SET B's and SET C's graphs in serial and parallel to extract the objects. We achieved an overall reduction of extraction time in parallel method up to 25%.

3.3 Category Prediction

Every images can be categorised but the parameters needed to categorise are the once that need to be defined properly. All the extracted objects are not needed for further processing. So we set a threshold value of 0.8. If the objects are lesser than the threshold, then they are discarded and the higher threshold objects are alone extracted. Images can be categorized by the visible objects in it. For example a crime scene can be detected if there is a gun or knife with blood, etc. We are using this approach by defining parameters for each category. For sports, there should be a ball or players or referees which act as the primary object in categorising. So if the image contains any one of the primary objects defined for a category, then we can categorise efficiently.

For example, we should set parameters for the sports category as balls, bats, goal post, players, referees. If any image contains anyone of these objects, then it can be classified as sports category.

3.4 Inference Prediction

Objects with threshold value higher than 0.8 are known as the first level objects. For predicting the inference, only the first level objects are used. All the needed inferences should be defined with accurate parameters so the inference is just like human prediction. The list of inferences can grow, so that the system can scale from 1 to *n* inferences with the help of human defined parameters.

Table 1Example Of Inference

Football	**Players**	**Inference**
1	2	Playing Football
>2	3	Practising Football

Table 1 consists of sample inference for a football image. We can predict that if the image consists of 1 football and 2 players, then our system can tell that they are playing football. If there are 2 balls and players, then it is Football practising image.

3.5 Scene Prediction

Scenes are predefined with human help. Each scene contains certain primary objects to be inferred. Those objects are visually available in the frames. For each scene, frames at 2 second interval are considered. Then the primary objects are extracted from each frames. To prove a scene, the primary object must occur in at least ¾ of the total frames considered.

For crime scene, there should be either a gun, knife with blood, dead body or blood. In our proposed work, frames at 2 second interval are taken and the objects are extracted. If one of the defined primary objects defined occurs at ¾ of the total frames, then it is a CRIME scene.

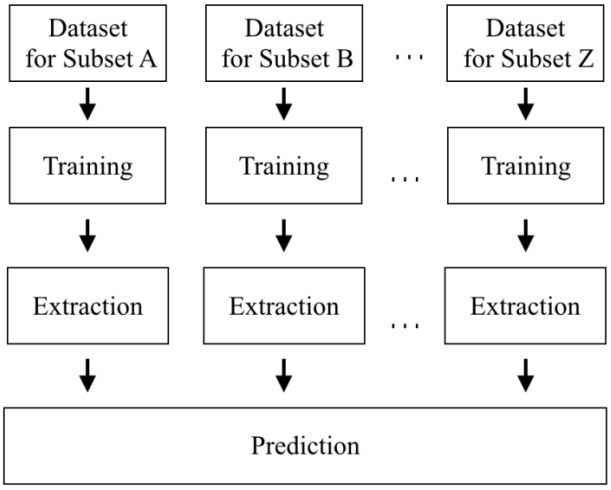

Fig 2

Proposed Method

3.6. Issues

We have modularised the training method thus saving the computing power and time but the images are processed with each models in parallel. With each TensorFlow session running in parallel for each model, requires high computing power thus the machine should be of high configuration.

IV. EVALUATION

4.1 Experiment Setting

To examine the effectiveness of our approach, we are considering sports category as mentioned above. We will be training three sets of objects, SUPERSET A, SET B and SET C to prove our approach. Training and extraction will be done using all three sets to prove the approach. In this evaluation, time reduction in the process of training, extracting and predicting inference will be proved. Set A is a superset of Set B and Set C. Three sets were trained separately with the SSD_MobileNet model. The machine is a non-GPU, 8vCPUs and 8GB RAM machine. The system was trained until the Total Loss was under 1.0.

4.2 Training

The Table 2 shows that training an entire superset A is costlier than the combined training time of Subset A and B. We reduced the overall training time by 33%. Inference graphs for each set of objects were generated.

Table 2Set of Objects

Set	Objects	Training Time
A	Players, Goal Post, Basketball, Football, Volleyball	12 hrs
B	Players, Goal Post	5 hrs
C	Basketball, Football, Volleyball	3 hrs

4.3 Testing

Testing was also done on the same machine. Three approaches were followed to evaluate. First approach was to extract objects by using the SET A model. Second approach was to use the models one by one in serial and the third approach was to use the models in parallel.

Table 3Testing Approaches

S No	Approach	Time
1.	Single	38 seconds
2.	Serial	35 seconds
3.	Parallel	26 seconds

Table 3 shows, parallel approach is 25% faster in extracting, segregating the objects that had high threshold, predict the category and inference.

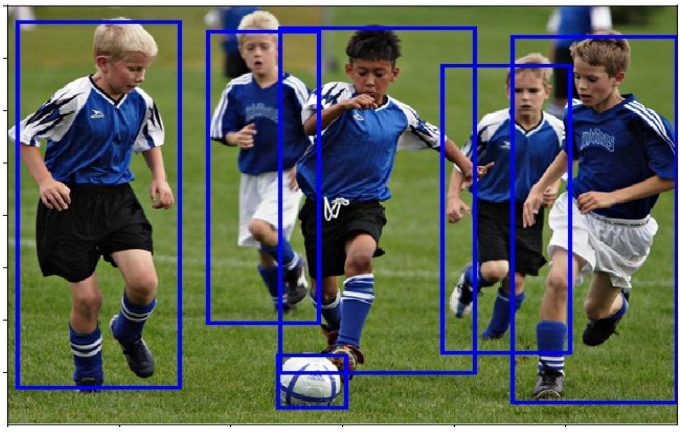

Image 1

Image 1 is the output of our system. The first level objects are the 5 players and a football and the second level objects are the two players behind. Here the system considers the football as a primary object and predicts category as “SPORTS”. The inference is predicted as “5 players playing football”.

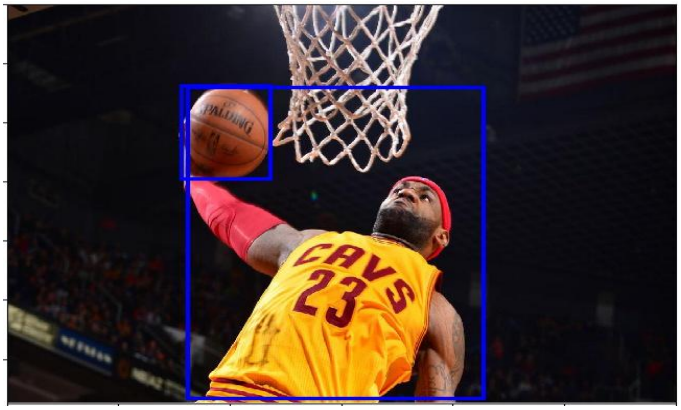

Image 2

In Image 2, the first level objects are the basketball and a player. Here, basketball is a primary object and it helps in predicting the category as “SPORTS”. The inference is predicted as “Playing Basketball”.

Image 3

Image 3 represents the frames of a video that have a 2 second interval between each of them. In 5 of the 6 frames, basketball is visible and satisfies the condition of being present in ¾ of the total frames. Thus, the category is predicted as “SPORTS”.

IV. CONCLUSION

Handling growing set of objects and achieving output in lesser time is the key to semantic content extraction. Through our approach we proved that there is no need to train the system with all objects at once and retrain again when new objects are added. Experimental results confirm the approach by saving the computation time upto 30% in overall process.

REFERENCES

[1] HyunYong Lee and Byung-Tak Lee, Selective Inference for Accelerating Deep Learning-based Image Classification, Information and Communication Technology Convergence (ICTC), International Conference, 2016.

[2] Chaitanya Vijaykumar Mahamuni and Neha Balasaheb Wagh, Study of CBIR Methods for Retrieval of Digital Images based on Colour and Texture Extraction, International Conference on Computer Communication and Informatics (ICCCI -2017), 2017.

[3] Kithmi Ashangani , Wickramasinghe K.U, De Silva D.W.N, Gamwara V.M, Anupiya Nugaliyadde and Yashas Mallawarachchi, Semantic Video Search by Automatic Video Annotation using Tensorflow, Manufacturing & Industrial Engineering Symposium, 2016.

[4] Junwei Han, King N. Ngan, Mingjing Li, and Hong-Jiang Zhang, Unsupervised Extraction of Visual Attention Objects in Color Images, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, 2006.

[5] Salah NASR, Muhammad Shoaib, Kais BOUALLEGUE, Hassen, Face Recognition System Using Bag of Features And Multi-Class SVM For Robot Applications, MEKKI,ICCAD'17, Hammamet – Tunisia, 2017.

[6] Maryam Asadi -Aghbolaghi, Albert Clape´s , Marco Bellantonio , Hugo Jair Escalante , V´ıctor Ponce - Lo´pez, Xavier Baro, Isabelle Guyon, Shohreh Kasaei, Sergio Escalera , A survey on deep learning based approaches for action and gesture recognition in image sequences, Automatic Face& Gesture Recognition, 12th IEEE International Conference, 2017.

[7] Li Wang , Deep Learning Algorithms with Applications to Video Analytics for A Smart City: A Survey, Computer Vision and Pattern Recognition, 2015.

[8] Swapnali. N.Tambe, Prof. D.B Kshirsagar, Devyani Bhamare, A Review paper on Semantic Content Extraction in Video Using ontology Based Fuzzy model, in IJEDR, 2015.

[9] Jiajun Sun, Jing Wang, Ting-chun Yeh, Video Understanding: From Video Classification to Captioning, Stanford 2017.

[10] Huang J, Rathod V, Sun C, Zhu M, Korattikara A, Fathi A, Fischer I, Wojna Z,Song Y, Guadarrama S, Murphy K, Speed/accuracy trade-offs for modern convolutional object detectors, CVPR, 2017.