Video To Text Analysis : Deep Learning

Akshay Khatter[a] , Dr. Namita Gupta[b] , Keshav Issar[c] , Nishant Sardana[d]
[a]Maharaja Agrasen Institute Of Technology, Rohini, Delhi 110086 , India
[b]Professor at Maharaja Agrasen Institute Of Technology, Rohini, Delhi 110086 , India
[c]Maharaja Agrasen Institute Of Technology, Rohini, Delhi 110086 , India
[d]Maharaja Agrasen Institute Of Technology, Rohini, Delhi 110086 , India

***Abstract*— Using Deep learning techniques, proposing a new approach that analyses a video and then present it in understandable language using NLP techniques. For most people, watching a brief video and describing what happened (in words) is an easy task. For machines, extracting the meaning from video pixels and generating natural-sounding language is a very complex problem. Solutions have been proposed for narrow domains with a small set of known actions and objects. Video captioning is the method of generating a natural language sentence that explains the content of the input video. This paper proposes a deep neural network model for effectively captioning the video. Apart from visual features, our proposed model additionally describes the video content effectively such that we can understand the easily video and in lesser time. This paper consists of methods, structure and design explored by us, to understand videos. This paper shows results of implementation of frame by frame captioning, YOLO and also proposes a hybrid approach combining YOLO, image captioning and a double convolutional neural networks approach. We propose the models to do both classification and caption also, using hybrid of the discussed models.**

***Keywords*— Video-Classification; Video-Captioning; Action-Recognition; CNN; LSTM; Video-To-Text**

I. INTRODUCTION

Need of the Study Because this problem is a million-dollar problem at present and is being researched in many of the premier institutions of the world. Solving the visual symbol grounding problem has long been a goal of artificial intelligence. The field appears to be advancing closer to this goal with recent breakthroughs in deep learning for natural language grounding in static images. It is proposed in this paper to translate videos directly to sentences using a unified deep neural network with both convolutional and recurrent structure.

Recognition of activities in videos is a challenging task which has received a significant amount of attention in recent years. Compared to image classification, the temporal component of videos provides an additional and important clue for recognition, as a number of actions can be reliably recognised based on the motion information. This paper is organised as follows:

- Frame by Frame Image Captioning.
- YOLO for recording activity related data.
- Hybrid Model.
- Proposed framework with Dual-CNN along with above techniques.

II. OBJECTIVE

Generating captions to an image automatically shows how computers understands the image, which is a fundamental objective of intelligence. For a caption model, it is needed to find which objects are present in the image and their connection in a language such as English. We can also utilize natural language processing technologies & understand the world in images. Besides, we added attention mechanism, which is able to acknowledge what a word alludes to in the image, and thus summarize the connection between objects in the image. This will be a strong tool to utilize the substantial unformatted data of image, which in turn can help us decode video data as well.

This has a variety of applications like helping the blind, faster processing of CCTV feeds, automated monitoring of meeting halls, schools etc.

III. SCOPE

- There is a few captioning online.
- Only Some high-profile projects are readily obtainable with captioning.
- Quantities of captioning are increasing gradually.
- It is feasible to closed-caption online video in all major formats, and all video formats can be open-captioned.
- Industry requirements for video captioning are growing; in those sectors, captioning is not carefully aware of all circumstances, resulting in more video with captions.
- Transcripts created from captions preferably with valid structure are a useful resource for searching and archiving; they give accessible video a second life.

IV. APPROACH AND RESULTS

- **Frame by Frame Image Captioning.**

Dataset used here was We have used Flickr 8k dataset. This dataset contains 8000 images each with 5 captions. Recognizing actions from videos is done by capturing information from each frame. Action recognition task involves the identification of different actions from video clips is like a natural extension of image classification tasks to multiple frames and then aggregating predictions from

each frame. Despite the success of deep learning architectures in image classification (ImageNet), progress in architectures for video classification has been slower.

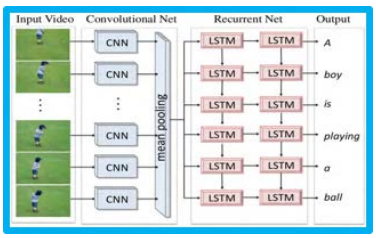

Fig. 1. Captioning Architecture.

Various real time problem, like understanding what's in front of a car for autonomous driving applications can be solve using Continuous classification. We have used a basic classification approach in this paper which gives us a straightforward approach to solving our problem and it forms the basis for more advanced approached proposed in this paper.

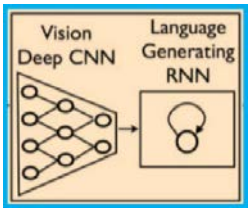

Fig. 2. Caption Generation.

This model has following main components:

Photo Feature Extractor: The Photo Feature Extractor model inputs photo to be a vector of 4,096 elements. These images are processed by a Dense layer to produce a 256 element representation of the image. Photos were pre-processed with the VGG model (without the output layer) and use the extracted features predicted by this model as input.

Sequence Processor: This is a word embed-ding layer for work on the text input, followed by a Long Short-Term Memory (LSTM) recurrent neural network layer.

Separator Model: The feature extractor and sequence processor outputs a fixed-length vector. These are then merged together and processed by a Dense layer to make a final prediction. The Separator model adds the vectors from both input models. This output is then fed to a Dense 256 neuron layer and then to an output Dense layer that makes a pre-diction over the entire output vocabulary for the next word in the sequence. Here we use Softmax regression that is a form of logistic regression which normalizes an input value into a vector of values that follows a probability distribution whose total sums up to 1.

Fig. 3. Caption Generation Output.

• **YOLO for recording activity related data and creating derived measures.**

You only look once (YOLO) is a real-time object detection system. On a Pascal Titan X it process images at 30 FPS and has a mAP of 57.9% on COCO test-dev.

The object detection task consists in determining the location on the image where objects are present, as well as classify those objects. Various other methods for this, like R-CNN and its variations, used a pipeline to perform this task in multiple steps. This can be slow to run and also hard to optimize, because each individual component must be trained separately. YOLO, works with a single neural network.

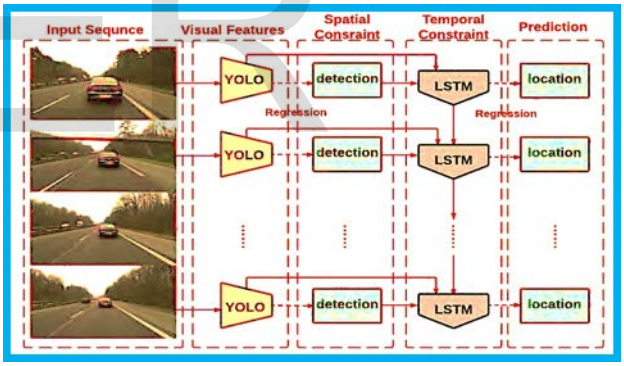

Fig. 4. YOLO Architecture.

YOLOv3 is very fast and accurate. In mAP measured at .5 IOU YOLOv3 is at par with Focal Loss and about 4x faster. Moreover, it can easily tradeoff between speed and accuracy simply by changing the size of the model without retraining.

LSTM:

Another approach for classifying video is to use LSTM. Similar to temporal feature pooling, LSTM networks

operate on frame-level CNN activations as well as integrate information over time. LSTM out-puts a hidden vector for each activation frame.

LSTM is the combination of 3 functions input gate, output gate and forget gate in which the input gate takes input and when another input is taken then previous input get into forget gate and the network generate its output according to the new data and previously learned data.

LSTM layers accepts h_{t-1} and x_t as inputs. The input enters four gates after dot production with weights (W). Each gate has separate functions.

LSTM cell can be expressed as the following equations

$$\begin{pmatrix} i \\ f \\ o \\ g \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} W \begin{pmatrix} h_{t-1} \\ x_t \end{pmatrix}$$

$$c_t = f \odot c_{t-1} + i \odot g$$

$$h_t = o \odot \tanh(c_t)$$

where σ represents sigmoid function, and DOT element is element wise multiplication. LSTM has uninterrupted gradient flow, which makes it easier to back-propagate. LSTM is also more stable without gradient exploding and vanishing. The first layer of LSTM accept the inputs from the pretrained CNN model, and then the second layer of LSTM receives sequential outputs from the previous LSTM layer. At last, we apply the output into fully-connected layer and calculate the final score.

Fig. 5. YOLO Output Screenshot From a Clip.

Hybrid Model expected outcome:

Basic Idea: For detection of objects we use a pre trained model name Yolo for object detection then after detection of an object we track the position of that object frame by frame and according to the position in every frame we detect some part of nature of the contents present inside the video like A car is moving, horse is running etc. by using the NLP we arrange it on a semantic arrangements of words.

While much progress has been achieved on above methodologies, a still left task is video understanding - analyse a video frame segment and explain what's happening inside it. Despite some earlier progress on solving video understanding, contemporary algorithms are still far from human-level results.

Here we are proposing a novel method to analyze the contents of a video frame segment, achieving state-of-the-art results. In the method we have use two parallel convolution neural networks (CNNs) on the same video segment--- Both on a Background Segment and also a Subject Segment.

We observe that frames in video usually contain two distinct parts - static areas, which don't change at all or change slowly, and dynamic area which indicates something important that is currently going on. For instance, a video of a dog running will include a relatively static background like grass with a dynamic object (the dog) moving quickly in the scene. Accordingly we can use a better resolution CNN to analyze the static content of a video while running in parallel a fast, low-definition CNN whose goal is to analyze the action content of a video.

Along with the above results we can also add further meaning to our captions by merging our outcomes from yolo to improve on parts of sentence, which ensures that we use the right name for the object(s) present in the video, and also create measures to track both inter-object movements and object-background movements. All these measures will ensure that we can get an output very close to the human-interpreted version.

V. OBESERVATIONS & CONCLUSIONS

- **Limitations**:

1. Huge Computational Cost A simple convolution net for two dimensional images classifying 101 classes has huge number of parameters which makes search difficult and likely for overfitting .

2. Capturing long context Action recognition involves capturing spatiotemporal context across frames. Additionally, the spatial information captured has to be compensated for camera movements.

3. Designing classification architectures Designing architectures that can capture spatiotemporal information involve multiple options which are non-trivial and expensive to evaluate.

4. No standard benchmark.

- **Other conclusions:**

Fast and accurate processing. Predictions are made from one network and can be trained end-to-end to improve accuracy. YOLO accesses to the complete image in predicting boundaries. With all the additional context, YOLO shows fewer false positives in background areas.

Many of the problems that we faced while training our models had to do with overfitting. Complete supervised approaches require large amounts of data, but the datasets we used had 8k images. The task of giving a description is much harder than object classification and using datasets like ImageNet have only recently become dominant.

As a result, we can see that, even with the results we obtained are commendable, the advantage of our method versus most current human-engineered approaches will only increase in the coming years as training set sizes will

grow as and when a standard framework is developed where we can have open source contributions.

VI. ACKNOWLEDGEMENT(S)

We have taken efforts in this project. However, it would have not been possible without the kind support and help of many individuals and researchers doing related work. I would like to extend my sincere thanks to all of them. We are highly indebted to Dr. Namita Gupta ma'am for her guidance and constant supervision as well as providing quintessential information regarding the project and also for their support in completing the project. We would like to express our gratitude towards our parents and members of CSE Department Maharaja Agrasen Institute of Technology for their co-operation and enlivenment which helped me in completion of this project.

REFERENCES

[1] Jiajun Sun, Jing Wang,Ting-chun Yeh. Video Understanding: From Video Classification to Captioning. In ACL, 2010.

[2] D. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). *Large-scale video classification with convolutional neural networks.* In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1725-1732).

[3] K. Cho, B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. *Learning phrase representations using RNN encoder-decoder for statistical machine translation*, In EMNLP, 2014.

[4] J. Wu, Z., Yao, T., Fu, Y., & Jiang, Y. G. (2016). *Deep Learning for Video Classification and Captioning*. arXiv preprint arXiv:1609.06782.

[5] Subhashini Venugopalan, Marcus Rohrbach, Jeff Donahue, Raymond Mooney, Trevor Darrell, Kate Saenko. Sequence to Sequence – Video to Text, In IEEE Explore, 2016.

[6] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. *Every pic ture tells a story: Generating sentences from images*. In ECCV, 2010.

[7] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In ECCV, 2014.

[8] A. Graves. Generating sequences with recurrent neural networks. arXiv:1308.0850, 2013.

[9] S. Hochreiter and J. Schmidhuber. *Long short-term memory.* Neural Computation, 9(8), 1997.

[10] R. Kiros and R. Z. R. Salakhutdinov. *Multimodal neural language models*. In NIPS Deep Learning Workshop, 2013.

[11] G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: *Understanding and generating simple image descriptions.* In CVPR, 2011.

[12] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In ACL, 2012.

[13] M. Mitchell, X. Han, J. Dodge, A. Mensch, A. Goyal, A. C. Berg, K. Yamaguchi, T. L. Berg, K. Stratos, and H. D. III. Midge: *Generating image descriptions from computer vision detections*. In EACL, 2012.

[14] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In NIPS, 2011.

[15] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. *Collecting image annotations using amazon's mechanical turk.* In NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 139– 147, 2010.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. *ImageNet Large Scale* Visual Recognition Challenge, 2014.

[17] R. Socher, A. Karpathy, Q. V. Le, C. Manning, and A. Y. Ng. *Grounded compositional semantics for finding and describ-ing images with sentences*, In ACL, 2014.