Video Classification using Recurrent Neural Networks

Misbah Munir
Oxford Brookes University
Oxford, UK
15059148@brookes.ac.uk

Fabio Cuzzolin
Oxford Brookes University
Oxford, UK
fabio.cuzzolin@brookes.ac.uk

Abstract

Video Classification is an important and widely researched problem in the field of Computer Vision. Machine Learning models have proved to perform significantly well in order to generate good results. Existing methods focus on traditional Convolutional Neural Network (CNN) models to analyse visual data. Since the dimensionality of visual data is very high, this makes the classification problem complex in nature. Most learning algorithms are restricted in the way they process the information, due to the restriction on having a set number of parameters to process similar type of data in order to generate similar results which leads to the risk of losing the useful information being significantly high. Usually the video clips are of variable length but they are required to be mapped onto a smaller and fixed dimensional vector representations of features in order to be processed by CNNs. This also creates an overhead of information pre-processing before the results can be achieved. In recent years, Recurrent Neural Networks (RNNs) have proven their strength to for analysing sequence-based data without losing much information. In this research, we aim to investigate the strengths of RNN models with a focus on Long-term Recurrent Neural Network (LRCN) model to classify the videos. Since visual data is sequential in nature and high complex, therefore, theoretically RNNs can be a great fit for its analysis. In these experiments we aim to analyse the results we achieve on video classification problem using LRCN.

1. Introduction

There has been a dramatic increase in the amount of visual data being captured and shared online. With the availability of platforms like YouTube[1] and Facebook[2], which provides the facility to create and share videos with other users within the network or globally. In addition to this, the availability of cameras in hand-held devices like mobile phones and tablets, the trend has only picked up pace. The videos available online have a particular set of audience based on the content and context of the videos. Currently, the system to automatically classify the videos into a particular category is imperfect and hugely rely on human input. This leads to a higher chance of error and deliberate mistakes because human input is restricted by the restricted knowledge and other factors. On a global perspective, there is a need to find a way to automatically classify the videos in order to provide the best possible results to the consumers who are looking for a particular type of video.

Computer Vision research aims to achieve this task by taking advantage of Machine Learning techniques in order to perform Video Classification. Classification task is complicated in nature due to the high dimentionality of data. So far, the research in the field of video classification has taken advantage of different approaches including, image analysis, audio processing, text analysis and more[13], but more recently, deep learning approach to analyze the visual data has proven to be most successful[8].

The machine learning algorithms that have proved to be more successful are very restricted in many ways including the size of data stream being provided to the algorithm need to be constant and the mapping is done to only finite number of classes. Visual data is sequential in nature, mapping the information on finite set cause the loss of important information in the data. Since, machine learning aim to replicate the results generated by human brain, the recurrence mechanism embedded in the brain has been used to map the sequence of data to generate great results for example, speech processing, text analysis, etc by using the type of learning algorithms - RNNs. Due to the sequential nature of visual data, similar approach can be used to improve the results of video analysis and specially classification.

2. Background

Less than a decade ago, the methods of video classification followed a three-modular architecture for generating classification results where visual data needed to be heavily

[1] https://www.youtube.com
[2] https://www.facebook.com

down sampled and processed before a model representation could be initiated. In most applications a linear classifier would create a classification model[8].In 2014, the trend for using deep methods for machine learning started getting attraction by the research community after the CNN based video classification proposed by Karpathy[8] which generated great results. The inspiration to replace the handcrafted models with deep CNNs came from the advancements in image-based classification, automatic segmentation and labelling[10, 4, 2].

Recurrent Neural Networks are special type of deep learning networks which provides promising direction in the research of sequence based learning. These models have been biologically inspired by human brain[11]. In technical terms, humans process information using temporal information as a major factor for developing understanding. Many machine learning methods, including most deep networks lack this ability which generate a semantic gap for proper data processing.

The introduction of recurrence in the research of sequential data processing have generated some *magical*[3] results specially in the field of signal processing for speech[5] and text[14]. Speech signals can be represented as a waveform in time domain[4], hence can be visualised as two dimensional data sequence. In contrast, video is also a sequence of data (images) with much higher dimension[3] as described in Chapter 4. There are many ways of generating feature hierarchies based on the

In case of modelling visual data using RNNs, image captioning have achieved some great results. It models a single input vector to a sequence of outputs generated to describe an image[12, 17, 11]. The main strength of RNN that needs to be exploited is sequence modelling which can be achieved by performing sequence learning on visual data. The results achieved by modelling video sequence to generate captions for videos as a sequence of outputs has also shown some promising results in the direcion of adopting RNNs[18, 19, 15]. So far, sequence generation seems like a major part of learning through RNN. RNNs can also map a sequence of input to a static output.

2.1. Long Short-Term Memory

Originally proposed in 1997[6], to overcome the long-term dependency problem in RNNs which was a major factor negatively affecting the adaptability of RNNs.

More recently there has been an increased interest in utilising the strength of LSTM-RNNs for sequence to static mapping has generated some interesting results. Long Recurrent Convolution Network (LRCN) is a simple end-to-end model that classify a video based on deep features in order to classify visual data. The evaluations based on

[3]http://karpathy.github.io/2015/05/21/rnn-effectiveness/

[4]https://en.wikipedia.org/wiki/Audio_signal

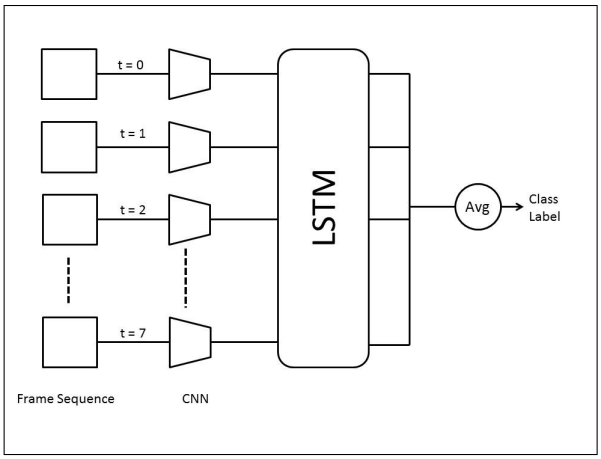

Figure 1. LRCN Architecture

the UCF-101 dataset have shown some great results[3]. In another research, fused Spatial and Motion features multi stream feature feed to LSTM network for categorise the videos in UCF-101 dataset have achieved similarly positive results[16]. Similarly, increasing the depth with added feature pooling achieved some interesting insights but did not prove that the stacked LSTM can have any significant performance gain for the system[20].

3. Long Recurrent Convolution Networks

We followed the video captioning and classification model initially proposed by Jeff Donahue[3]. They proposed a system for video captioning and classification by using the underlying strength of RNNs in order to achieve good results on visual data. The model is referred to as a Long-term Recurrent Convolution Networks (LRCN). The model architecture has been designed keeping end-to-end training in consideration. Figure 2 provides a graphical overview of the system architecture.

3.1. CNN - Feature Extractor

The system gets a sequence of image files as an input to the model. In case of LRCN, CaffeNet[7] has been used as a feature descriptor which is a variant of AlexNet[10]. Both feature descriptors shows similar performance therefore there is no significant performance concern whatsoever. However, based on the literature CaffeNet claims to be faster. We have not performed comparative analysis of both Networks therefore, it is difficult to judge about competitive efficiency of each approach.

3.2. Recurrent Neural Network

Recurrent Neural Networks (RNNs) are a class of machine learning algorithms that make use of the recurrence parameter and generate a sequence of values, that is the result of information processing via learning method. There

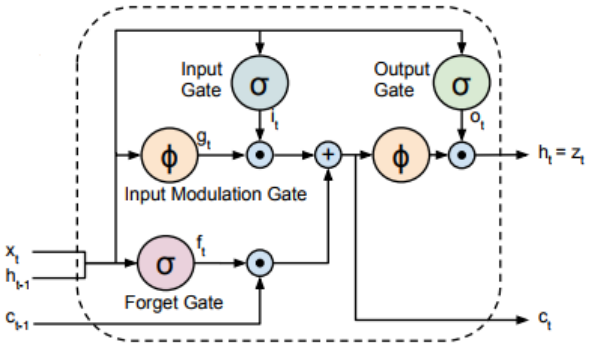

Figure 2. LSTM Unit[3]

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + b_f) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\
g_t &= \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot g_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}
$$

Figure 3. The Mathematical Model of LSTM[3]

are many variations of RNNs having shown promising results in sequence based information representations.

3.3. LSTM

Long Short-Term Memory[6] (LSTM) Models have shown very promising results towards sequence based learning. LSTMs processes the information sequence with a special property induced using a special variable - forget gate. The purpose of these forget gates is to make a decision about information propagation. As shown in Figure 2, the LSTM unit involves some specialized variables which keep the data in memory in order to efficiently utilize it as memory. The mathematical model has been explained using Figure 3.

3.4. Output Generation

LSTM unit will generate a class label for every sequence item generating *n* predictions. In order to assign a single label to the class, an average of these predictions was calculated to generate a final label prediction.

4. Experiments and Results

4.1. Dataset

The experiments were mainly performed on the HMDB-51 dataset which is a small and diverse dataset for human actions based on motion. There are 7000 videos included in the dataset divided into 51 action classes. The experiments considered a sub-sampled set of frames extracted from the videos where only 8 frames per video were selected based in time dimension based on the length of video at equal intervals.

4.2. Optimization

The Adam[9] optimizer was used instead of classical approach of Stochastic Gradient Descent[1] (SGD). Adam optimizer provides better computational efficiency with little to no negative effect on the model generalization.

4.3. Experiment

The proposed framework performs fundamentally well on more modest number of classes. The most noteworthy achievement was 92% accuracy on a small model containing only 2 classes. Including another class in the training set significantly decreased the performance.

Likewise, the precision is contrarily connected to the similitude of activities. Be that as it may, sometimes the likeness of activity was not a prime factor in dropping the performance of system. For instance, *brush hair* and *cart wheel* are totally unique activities according to human observation, yet for the framework, they were genuinely comparative which prompted just 67.2% accuracy. On the off chance that we observe a totally different set of classes, that is, *chew* and *catch*, the framework accurately distinguished the correct class precisely 80% of the time. In case of experiment on 13 distinctive activity classes which incorporates the ones stated above, the achieved precision was 67.8%. This prompted the conviction that there is a plausibility of the framework to require more data for a particular activity in order to enhance its learning capacity. In conclusion the framework was tried to recognize the names for cases from 51 activity classes. The performance of the framework dropped drastically and accomplished only 8.09% accuracy.

However, significantly great results were achieved by running experiments on subset of UCF-101 dataset. This small dataset was also sub-sampled on 8 frames per video but the key difference in the sub-sampling method was to use 8 consecutive frames rather than using the frames distributed in time space. A 4-class dataset was created to perform experiments in order to formulate a proof of concept. The experiments showed significant achievements where the classification accuracy achieved was 100%. Due to shortage of time, more detailed experiments could not be performed.

Based on these observations, it is speculated that there are many factors that might have had adverse effect on the results on HMDB51 dataset. First, the down-sampling method of data can be a significant factor. Though human subjects could correctly identify the action, there is still a significant loss in information that adds for the model for effective learning. In our research, this was a unique approach to reduce the size of input data. Secondly, limited

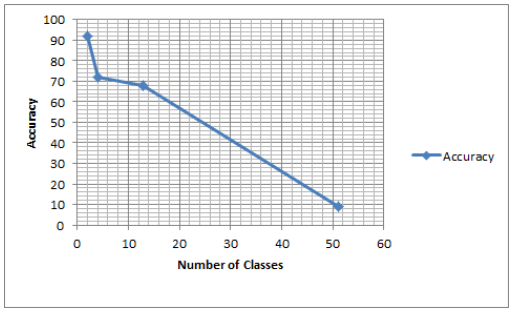

Figure 4. Classification accuracy graphical representation

Number of Classes	Maximum Accuracy
2	80%
4	71.9%
13	67.8%
51	8.9%

Table 1. Accuracy of Classification on HMDB-51 Dataset

time constraint was a major factor in running small scaled experiments. Detailed optimization of the system needs significantly more time in order to improve the results. Therefore, that was an out-of-scope task for this research.

5. Conclusion and Future Work

RNNs have proved to be a great fit for many sequential data types. However, in case of visual data, there is still a great deal to be analyses and understood in order to get the best results. First step towards the betterment of results can be making use of the motion-based information in order to detect an action in a better way. Another proposition is to make use of the variable length of video sequences since some actions can be performed in less time and some require more time to finish. In case of classification decision, rather than using an averaging technique, the system can use the voting method to assign a class label. Lastly, we can increase the depth of the network in order to further analyze the possibility of improving results.

References

[1] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006. 3

[2] C. Couprie, C. Farabet, L. Najman, and Y. LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013. 2

[3] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 2, 3

[4] C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2013. 2

[5] A. Graves, A.-r. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pages 6645–6649. IEEE, 2013. 2

[6] S. Hochreiter and J. Schmidhuber. Long short-term memory. 9:1735–80, 12 1997. 2, 3

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. B. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *CoRR*, abs/1408.5093, 2014. 2

[8] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 1, 2

[9] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 3

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012. 2

[11] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. 2

[12] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014. 2

[13] M. Roach, J. Mason, N. Evens, L. Xu, N. Evans, L. qun Xu, F. Stentiford, and M. Heath. Recent trends in video analysis: A taxonomy of video classification problems, 2002. 1

[14] I. Sutskever, J. Martens, and G. E. Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024, 2011. 2

[15] S. Venugopalan, H. Xu, J. Donahue, M. Rohrbach, R. Mooney, and K. Saenko. Translating videos to natural language using deep recurrent neural networks. *arXiv preprint arXiv:1412.4729*, 2014. 2

[16] Z. Wu, Y.-G. Jiang, X. Wang, H. Ye, and X. Xue. Multi-stream multi-class fusion of deep networks for video classification. In *Proceedings of the 2016 ACM on Multimedia Conference*, MM '16, pages 791–800, New York, NY, USA, 2016. ACM. 2

[17] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015. 2

[18] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*, pages 4507–4515, 2015. 2

[19] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593, 2016. 2

[20] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015. 2