Video SemNet: Memory-Augmented Video Semantic Network

Prashanth Vijayaraghavan
MIT Media Lab
pralav@mit.edu

Deb Roy
MIT Media Lab
dkroy@media.mit.edu

Abstract

Stories are a very compelling medium to convey ideas, experiences, social and cultural values. Narrative is a specific manifestation of the story that turns it into knowledge for the audience. In this paper, we propose a machine learning approach to capture the narrative elements in movies by bridging the gap between the low-level data representations and semantic aspects of the visual medium. We present a Memory-Augmented Video Semantic Network, called Video SemNet, to encode the semantic descriptors and learn an embedding for the video. The model employs two main components: (i) a neural semantic learner that learns latent embeddings of semantic descriptors and (ii) a memory module that retains and memorizes specific semantic patterns from the video. We evaluate the video representations obtained from variants of our model on two tasks: (a) genre prediction and (b) IMDB Rating prediction. We demonstrate that our model is able to predict genres and IMDB ratings with a weighted F-1 score of 0.72 and 0.63 respectively. The results are indicative of the representational power of our model and the ability of such representations to measure audience engagement.

1 Introduction

The ubiquity of videos in this digital age and their ability to captivate the viewers have made them one of the strongest mediums of storytelling. However, not every video garners equal attention due to several factors like duration, content, target audience, etc. But narrative elements such as plot and structure play a crucial role in defining how well it will resonate with its audience. Research in video analysis [12, 11, 3, 15] revolving around narrative analysis, video semantic indexing and retrieval, summarization, etc. have focused on mapping low-level engineered features for understanding the video semantics. Most of these techniques use shot boundary detection, dominant color region detection or other shot detection methods to understand the aspects of the narrative content. Many times these low level features fail to capture the story or its narrative features.

The recent advancements in the field of machine learning and deep learning have paved way for large-scale analysis of videos that include video classification [9, 29], text captioning [16, 27, 20] or summarization [13, 5, 30], visual question answering [1, 4], unsupervised video representation learning [22, 18, 26], etc. In this paper, we incorporate some of the ideas from previous studies and propose a new architecture that can learn latent semantic descriptors and remember some of the semantic patterns in a video using a memory module. The latent descriptors obtained from the video can provide a semantic summary of the video by a simple weighted summation. The memory module not only retains specific patterns in videos but also promotes transfer of information across various time steps. We leverage the ability of descriptors and memory module to generate a rich semantic representation for a given video. We train our models on a movie corpus (explained in section 3) and experiment the representations produced by our model on genre and IMDB rating prediction tasks. These evaluations enable us to fathom how well our representations can (a) capture the video

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

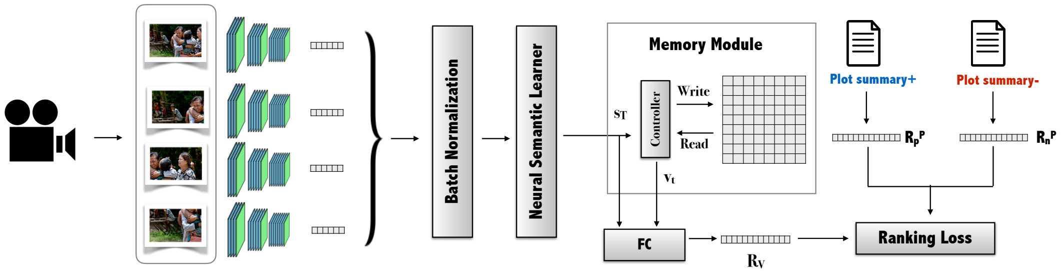

Figure 1: Illustration of Video SemNet

semantics, and (b) predict audience engagement. Our model can, therefore, bolster the efforts in improving the analysis around narrative aspects of a video.

2 Video SemNet

Overview: Given a collection of movie videos, our goal is to leverage the structure and content of the videos to automatically learn a semantic representation that can encapsulate the narrative elements (like plot structure, setting, story, etc.) of the video and eventually help us predict audience engagement. The learning is driven by the positive and negative samples of plot summaries for each video. We introduce a memory-augmented video semantic network, henceforth referred to as Video SemNet, for this purpose. It comprises of two main components: (i) Neural Semantic Learner, and (ii) Memory Module. Figure 1 illustrates the entire model. Before we go into the details of various components of our model, we briefly explain below the features and the loss function used to train our model. Video SemNet takes movie video as input and produces a video representation, R_V. Given both positive and negative samples of plot summaries related to the video, we train the model using a raking loss objective.

Video Feature Extraction: Motivated by Temporal Segment Network [25], our model enables learning over the entire movie by dividing it into L equal segments and operating on a sequence of snippets (usually an RGB frame) sparsely sampled from the movie segments. We apply a pretrained convolutional architecture over those L short snippets to extract rich latent features from the videos. A lot of literature [21, 28] in the past have demonstrated the capabilities of GoogleNet [24], Batch normalized-Inception [10], ResNet-101 [7], etc. to extract features from still images. We use ResNet-101 on RGB frame snippets to get a 2048-dimensional feature vector g_t. Thus, the input feature vector can be represented as $x_t \in \mathbb{R}^{|g_t|}$ i.e $x_t \in \mathbb{R}^{2048}$. We complete this feature extraction step by a batch normalization component to deal with the problem of covariate shift.

Modeling Plot summaries: To compute the vector representation of positive and negative samples of plot summaries, we can utilize any of the encoding schemes [14, 17]. We adopt a two level positional encoding strategy. First, we encode every sentence in the plot summary individually using positional encoding. The positional encoding technique is applied on the resulting sentence encodings to produce an embedding for the entire plot summary. We implemented the positional encoding described in [23] instead of simple averaging of word vectors to prevent loss of sequence order in the plot summary. Formally, positional encoding for a sequence (f_i) is given by $f_i = \sum_{j=1}^{L_s} l_j \cdot w^j$ where $\cdot$ is element-wise multiplication and l_j is a column vector with structure $l_{jk} = (1 - j/L_s) - (k/d)(1 - 2j/L_s)$, where k is the embedding index, d is the dimension of the embedding, w^j is an embedding for j^{th} element (word or sentence) in the sequence and L_s is the length of the sequence. The embedding for the plot summary is denoted as R^P.

Loss Function: We train our model to learn video representations using a ranking loss objective which measures their relevance to plot summaries. We want to maximize the similarity of video representation with its corresponding plot summary and minimize its similarity with negative examples. The objective function of the triplet ranking loss is defined as

$$L = max(0, s(R_V, R_n^P) - s(R_V, R_p^P) + \alpha) \tag{1}$$

where α is the margin parameter and $s(\cdot, \cdot)$ is the cosine similarity between video representation (R_V), positive (R_p^P) and negative (R_n^P) examples.

2.1 Neural Semantic Learner

Neural Semantic Learner builds a latent embedding for descriptors and outputs a semantic summary of the video. In order to the understand the narrative plot or structure from short snippets, it is important to track the variations between such sampled snippets. This can be accomplished by training these sequences of snippets through a recurrent neural network. Consider that we sample L snippets $S \in \{S_1, S_2, ..., S_L\}$ from their corresponding segments in the video. The features extracted for those sampled snippets are denoted as $X = x_1, x_2, ..., x_L$, where $X \in \mathbb{R}^{L \times 2048}$. The higher the value of L, the more computationally intensive the model turns out. Moreover, deeper LSTM do not always contribute to good performance as can be seen in [28]. Hence, we introduce our input X through temporal convolution layers with different filter sizes followed by a semantic descriptor learner.

Temporal Convolutional Layers: Let L be the number of layers, F_l is the number of convolutional filters at the layer l, and K_l is the total time steps at layer l. We define a series of filters in each layer as $W = \{W^{(1)}, W^{(2)}, ..., W^{(F_l)}\}$. With the output from previous layer $H^{(l-1)}$, we compute activations H^l using $H^l = \Phi(W * H^{(l-1)} + b)$ where b is the the corresponding bias, Φ is the activation function. We perform a max pool operation that reduces the time step by half i.e. $K_l = K_{l-1}/2$. Finally, the output from the temporal convolutional layers will be $H^L \in \mathbb{R}^{T_r \times d}$, T is the resulting time steps and d represents the dimensions of abstract video features.

Semantic Descriptor Learner: Semantic descriptor learner learns latent embeddings for each of the descriptors using a matrix $D \in \mathbb{R}^{N_D \times d}$ where N_D is the number of descriptors, d is the dimensions of video features. The input h_t to the learner refers a particular time step from H^L. Given h_t, we compute a weight vector r_t over N_D descriptors to measure its importance for that video. We want to have a function that can project h_t to N_D dimensions such that it models the temporal aspect of the video. Hence, we add a recurrence, $r_t = softmax(W_D[h_t; r_{t-1}])$ where $W_D \in \mathbb{R}^{N_D \times (d+N_D)}$. After computing r_t, we use it along with descriptor matrix D to generate a semantic summarizer s_t until time step t as,

$$s_t = D^T r_t \tag{2}$$

Finally, we use the semantic summary at time step T (s_T) as input to our memory module.

2.2 Memory Module

Recently, memory-augmented neural network [19] has been applied in areas which require fast adaptation and memorizing events. An end-to-end differentiable memory network is used to memorize facts that are relevant. In our scenario of video semantic analysis, we take inspiration from previous work associated with memory networks [6, 8] as it helps reason with the limited data and improve our ability to represent critical, unseen instances of the video. Our memory module comprises of (i) an external memory matrix $M \in \mathbb{R}^{m \times d}$, were m is the size of the memory and d is the dimensionality of each location in the memory (ii) a neural controller which uses a read head w_t^r and age vector A_t ($w_t^r, A_t \in \mathbb{R}^m$) to perform memory operations at a particular time step t. The various memory operations are summarized below:

Memory Read: The read head w_t^r is computed from semantic summary (s_T), memory state (M_t) and controller context C_t via, $w_t^r(i) = softmax(\theta_t(i))$ where $\theta_t(i) = tanh(W_s^\alpha s_T + W_c^\alpha C_t + W_m^\alpha M_t(i))$ and $W_s^\alpha, W_c^\alpha, W_m^\alpha, W_h^\alpha$ are trainable parameters, C_t is the controller context vector that summarizes the controller operations until time step t. We get the memory read vector using $v_t = M_t(j)$ where $j = argmax_i(w_t^r(i))$. After the read step, we perform two operations: (1) a new controller context vector (C_{t+1}) is calculated using a non-linear combination of current context vector (C_t), current read vector v_t and our semantic summary s_T via $C_{t+1} = tanh(W_c^\beta C_t + W_v^\beta v_t + W_s^\beta s_T)$, where $W_s^\beta, W_c^\beta, W_h^\beta$ are learnable parameters, and (2) update age vector A_t by incrementing the age of all indices by 1, i.e., $A_{t+1} = A_t + 1$.

Models	IMDB Rating Prediction	Genre Prediction
SSM	0.48	0.54
SLM	0.58	0.62
Video SemNet	**0.63**	**0.72**

Table 1: Evaluation of different variants of our model on two tasks: Genre Prediction and IMDB Rating Prediction. We show that Video SemNet outperforms all the other models based on weighted F1-score.

Memory Update: The memory update is a linear projection of the updated context vector to a particular memory location p, given by, $M_{t+1}(p) = W_m^{\gamma} C_{t+1}$, where W_m^{γ} is a learnable projection matrix. We find memory location p by selecting the index with maximum age in age vector A_t with some randomness. Formally, we choose the memory location p to update via $p = argmax_i(A_{t+1}(i)) + r$ where $r \ll m$ is a random number so as to avoid race conditions in asynchronous training. Finally, we refresh the age vector by setting $A_{t+1}(p)$ to 0. Our update mechanism, that conditions controller's context vector C_t with read vector v_t, promotes the flow of extracted semantic information across various time steps. During an update, whenever there is a need to remember the contents of memory location that is being overwritten, the model training will enforce the controller to copy its read vector to the corresponding memory location.

The final representation that encodes the semantic representation for the video is computed using a combination of both summary vector s_T and memory read vector v_t, formally given by,

$$R_V = tanh(W_{sv}[s_T; v_t]) \tag{3}$$

where $W_{sv} \in \mathbb{R}^{d \times 2d}$ represents the trainable projection matrix.

3 Experiments

Since our model learns the representations from plot summaries, the choice of our dataset is movies. The movies dataset comprises of 1436 Hollywood films available on DVD. We performed data augmentation as mentioned in [25] for our training. We train different variants of our model on this dataset: (i) Simple Standard Model (SSM): model without neural semantic learner and memory module (R_v is a projection over mean of extracted video features), (ii) Semantic Learner Model (SLM): model includes neural semantic learner but no memory module, (iii) Video SemNet: model includes both semantic learner and memory module. In order to evaluate the video representations, we use them for two tasks (a) movie Genre Prediction, (b) IMDB Rating Prediction. Each movie in IMDB website [1] is associated with genres and ratings. For genre prediction task, we shortlisted $\sim$100 movies for each of the five most frequent genres: action, comedy, drama, horror and romance. We used 80% of data for training our model using the learned representations and tested it on the remaining 20%. For IMDB rating prediction, we round off the ratings associated with movies and classify them across 10 different classes (Ratings 1-10) with baseline accuracy of 30%. Table 1 shows the results of our evaluation. We find that our Video SemNet model outperforms the other models on both the tasks.

4 Conclusion

We have proposed a new model to compute a semantic embedding for videos that can better represent the narrative aspects of a video. The representations obtained from our model are able to predict movie genres and IMDB ratings successfully, proving such representations can be generalized across different tasks. Though the current model showcases the power of our representations, it would be interesting to experiment with incremental improvements to the model like adding optical flow features and combining emotional arc analysis [2] with our semantic analysis. Beyond audience engagement analysis, we can also use such representations to generate summaries, predict propagation patterns on social media platforms, perform video retrieval and a lot more.

[1] http://www.imdb.com/

Acknowledgments

The authors would like to thank Eric Chu for his help with data collection for this project.

References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2425–2433, 2015.

[2] Eric Chu and Deb Roy. Audio-visual sentiment analysis for learning emotional arcs in movies. In *Data mining (ICDM), 2015 IEEE international conference*. IEEE, 2017.

[3] Ahmet Ekin, A Murat Tekalp, and Rajiv Mehrotra. Automatic soccer video analysis and summarization. *IEEE Transactions on Image processing*, 12(7):796–807, 2003.

[4] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[5] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems*, pages 2069–2077, 2014.

[6] Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint arXiv:1410.5401*, 2014.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[8] Łukasz Kaiser, Ofir Nachum, Aurko Roy, and Samy Bengio. Learning to remember rare events. *arXiv preprint arXiv:1703.03129*, 2017.

[9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li 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.

[10] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] Ying Li, Shrikanth Narayanan, and C-C Jay Kuo. Content-based movie analysis and indexing based on audiovisual cues. *IEEE transactions on circuits and systems for video technology*, 14(8):1073–1085, 2004.

[12] Rainer W Lienhart. Comparison of automatic shot boundary detection algorithms. In *Storage and Retrieval for Image and Video Databases VII*, volume 3656, pages 290–302. International Society for Optics and Photonics, 1998.

[13] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2714–2721, 2013.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[15] M Ramesh Naphade, Igor V Kozintsev, and Thomas S Huang. Factor graph framework for semantic video indexing. *IEEE Transactions on circuits and systems for video technology*, 12(1):40–52, 2002.

[16] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1029–1038, 2016.

[17] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[18] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[19] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*, 2016.

[20] Rakshith Shetty and Jorma Laaksonen. Video captioning with recurrent networks based on frame-and video-level features and visual content classification. *arXiv preprint arXiv:1512.02949*, 2015.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International Conference on Machine Learning*, pages 843–852, 2015.

[23] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448, 2015.

[24] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[25] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016.

[26] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2015.

[27] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei 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.

[28] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George 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.

[29] Shengxin Zha, Florian Luisier, Walter Andrews, Nitish Srivastava, and Ruslan Salakhutdinov. Exploiting image-trained cnn architectures for unconstrained video classification. *arXiv preprint arXiv:1503.04144*, 2015.

[30] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European Conference on Computer Vision*, pages 766–782. Springer, 2016.