Extending Long Short-Term Memory for Multi-View Structured Learning

Shyam Sundar Rajagopalan[1](✉), Louis-Philippe Morency[2],
Tadas Baltrušaitis[2], and Roland Goecke[1]

[1] Vision and Sensing, Human-Centred Technology Research Centre,
University of Canberra, Canberra, Australia
Shyam.Rajagopalan@canberra.edu.au, roland.goecke@ieee.org

[2] Language Technologies Institute, School of Computer Science,
Carnegie Mellon University, Pittsburgh, USA
{morency,tbaltrus}@cs.cmu.edu

Abstract. Long Short-Term Memory (LSTM) networks have been successfully applied to a number of sequence learning problems but they lack the design flexibility to model multiple view interactions, limiting their ability to exploit multi-view relationships. In this paper, we propose a Multi-View LSTM (MV-LSTM), which explicitly models the view-specific and cross-view interactions over time or structured outputs. We evaluate the MV-LSTM model on four publicly available datasets spanning two very different structured learning problems: multimodal behaviour recognition and image captioning. The experimental results show competitive performance on all four datasets when compared with state-of-the-art models.

Keywords: Long Short-Term Memory · Multi-View Learning · Behaviour recognition · Image Caption

1 Introduction

There is a need for computational approaches that can model multimodal structured and sequential data. This is important for modelling human actions, caption generation and other sequence analysis problems. The integration of multimodal or multi-view data can occur in different stages. We use a general definition of views as "a particular way of observing a phenomena". For example, in image captioning, views are from the image and its text caption. For child engagement level prediction from videos, the views are defined by three visual descriptors: Head pose, HOG and HOF. Two ways of fusing multi-view data are early and late fusion techniques [19]. However, these techniques do not take advantage of complex view relationships that may exist in the input data. Structured multi-view learning is aimed at capturing view interactions, thereby exploiting their relationships for effective learning. The key challenge to multi-view structured learning is to model both the *view-specific* and *cross-view*

B. Leibe et al. (Eds.): ECCV 2016, Part VII, LNCS 9911, pp. 338–353, 2016.
DOI: 10.1007/978-3-319-46478-7_21

dynamics. The *view-specific* dynamics capture the interaction between hidden outputs from the same view, while *cross-view* captures the interactions between hidden outputs of other views. These dynamics enable learning of subtle view relationships for better representation learning. The notion of capturing view-specific and cross-view dynamics is application specific and, hence, a need exists for flexibility in the design to model such dynamics.

We propose Multi-View LSTM (MV-LSTM), an extension to LSTM, designed to model both view-specific and cross-view dynamics by partitioning internal representations to mirror the multiple input views (see Fig. 1). We define a new family of activation functions (shown as MV-sigmoid and MV-tanh), which update the MV-LSTM internal memory partitions with three main factors: (1) input observations for the same view, (2) the hidden outputs from the same view (for view-specific dynamics), and (3) the hidden outputs from other views (for cross-view dynamics). Figure 2 shows example topologies of these different update factors. We evaluate the MV-LSTM model on four publicly available datasets spanning two different research problems: multimodal behaviour recognition and image caption generation.

2 Related Work

We first discuss related work in deep multi-view learning models and then present prior work related to two structured learning problems: multimodal behaviour recognition and image captioning.

Deep Multi-view Learning Models. Broadly, current approaches to multi-view learning can be grouped into three categories: (a) co-training, (b) multiple kernel learning, and (c) subspace learning [25]. Co-training algorithms train alternatively on different views to maximize the mutual agreement between the views. Multiple kernel learning involves learning linear/non-linear combinations of view-specific kernels. Subspace learning assumes that the views are generated from a latent subspace and the goal is to learn the latent subspace. Recently, the processing of multimodal inputs in LSTM networks is explored in image caption generation [22] and speaker recognition tasks [17]. In these models, multi-view learning is done by presenting all modalities either at the beginning or at all time steps of the LSTM network.

Wang *et al.* [23] have compared several deep multi-view representation learning models and proposed a new variant combining Canonical Correlation Analysis (CCA) and an Autoencoder. However, the applicability to sequence learning problems has not been explored. Extensions have been proposed using conventional LSTM for early fusion of language and images during decoding [7,26]. In the image caption generation task involving image and text as two modalities, Vinyals *et al.* [22] have used LSTM in the decoder module to generate image sentence representations. In their model, the image modality is shown only at the beginning of the decoding process and the text modality at all times.

Ren *et al.* [17] have proposed a multimodal LSTM for the task of speaker identification. In their design, all view representations from previous time step are used for view-specific gate updates by sharing weights across all modalities. Even though all modalities are present at all times, there is no flexibility to design various types of view-specific and cross-view interactions, nor in using only a portion of the views. To the best of our knowledge, the proposed MV-LSTM is the first multi-view structured LSTM to offer design flexibility to construct different network topologies for modelling both view-specific and cross-view interactions.

Behaviour Recognition. The development of computational models to understand the social-interactive behaviours of children is a relatively new area of study, facilitated by the recent public release of an annotated Multimodal Dyadic Behaviour Dataset (MMDB) dataset [16]. Presti *et al.* [14] proposed a variable Time-Shift Hidden Markov Model for learning and modelling pairs of correlation streams and validated their formulation for predicting the engagement level of a child using the MMDB dataset. The electrodermal activity (EDA) of the children, obtained from wearable sensors, has been used to predict the engagement level of the child [4]. Finally, acoustic signals have also been evaluated in models aiming to predict child engagement in the MMDB dataset [3].

Image Captioning. Inspired by recent successes in sequence generation in machine translation [20], automatic generation of natural sentences for images is gaining significant momentum. Jia *et al.* have proposed the gLSTM model [7], where a semantic representation together with text inputs was used as LSTM inputs at each time step. Karpathy and Fei-Fei [8] proposed a model that generates image region descriptions using the full image and their associated sentences. The image regions were obtained using Region Convolutional Neural Networks(CNN). More recently, Xu *et al.* [26] computed a context vector from salient regions of an image and used it on the decoder side.

3 Background: Long Short-Term Memory

The Long Short-Term Memory represents a class of Recurrent Neural Networks successfully applied to sequence learning problems [2] such as image caption generation. In the image caption generation task, the goal is to generate an appropriate sentence (Y) given an image (X) and LSTMs are commonly used as language generators. LSTMs are designed to address the exploding and vanishing gradients problems that may occur in Recurrent Neural Networks [6]. At the heart of an LSTM unit is a memory cell, C, that remembers inputs it has seen so far. The memory cell contents are explicitly controlled by sigmoidal gates that enable the network to decide when to read (i_t), write (o_t) and clear the memory contents (f_t). The input non-linearity is applied through the update term (g_t).

The LSTM update operations are given by

$$i_t = sigm(W_{ix}x_t + W_{ih}h_{t-1}) \quad (1)$$

$$f_t = sigm(W_{fx}x_t + W_{fh}h_{t-1}) \quad (2)$$

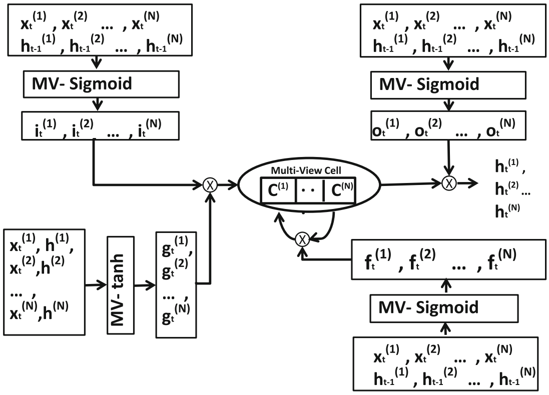
$$o_t = sigm(W_{ox}x_t + W_{oh}h_{t-1}) \quad (3)$$

$$g_t = tanh(W_{gx}x_t + W_{gh}h_{t-1}) \quad (4)$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t \quad (5)$$

$$h_t = o_t \odot c_t \quad (6)$$

where x_t is the input representation vector at time t. h_{t-1} is the LSTM output from previous time step. $sigm$, $tanh$ represent sigmoid and tanh non-linear transfer functions. W are the model parameters. y_t can be inferred at each time step by adding a softmax layer from the LSTM output h_t and selecting the label (e.g. word) with highest probability.



x_t^(1), x_t^(2) ... , x_t^(N)
h_t-1^(1), h_t-1^(2) ... , h_t-1^(N)
MV- Sigmoid
i_t^(1), i_t^(2) ... , i_t^(N)
x_t^(1), x_t^(2) ... , x_t^(N)
h_t-1^(1), h_t-1^(2) ... , h_t-1^(N)
MV- Sigmoid
o_t^(1), o_t^(2) ... , o_t^(N)
Multi-View Cell
C^(1) · · C^(N)
h_t^(1),
h_t^(2)...
h_t^(N)
x_t^(1), h^(1),
x_t^(2), h^(2)
... ,
x_t^(N), h^(N)
MV- tanh
g_t^(1),
g_t^(2),
... ,
g_t^(N)
f_t^(1), f_t^(2) ... , f_t^(N)
MV- Sigmoid
x_t^(1), x_t^(2) ... , x_t^(N)
h_t-1^(1), h_t-1^(2) ... , h_t-1^(N)

Fig. 1. The proposed Multi-View LSTM. $X_t^{(k)}$ represents k-th view input at time step 't' and $h_{t-1}^{(k)}$ is the MV-LSTM output from time step $t-1$ corresponding to the k-th view. N is the total number of views. The multi-view sigmoid and tanh gate functions are defined in Eqs. 7–13.

4 Multi-View LSTM

The Multi-View LSTM partitions the memory cell and the gates into regions corresponding to multiple modalities or views. The proposed MV-LSTM model brings two novel ideas, the second idea being the most important: (1) A view has its own internal dynamic: The MV-LSTM model keeps one memory partition (referred to as "region" in Fig. 1) for each input view. E.g. when modelling engagement, the MV-LSTM will be partitioned in three memory partitions, one for each of the three input views. (2) The memory partition of a specific view should be flexible in how it integrates information from other views: MV-LSTM allows four types of memory cells: (a) View-specific cells: affected by a hidden state from the same view (orange in Fig. 2), (b) Coupled cells: affected by a hidden state from other views (green in Fig. 2), (c) Fully-connected cells: affected by both same-view and other-view hidden states (brown in Fig. 2), (d) Input-oriented cells: not affected by either the same-view or other view hidden states (yellow in Fig. 2).

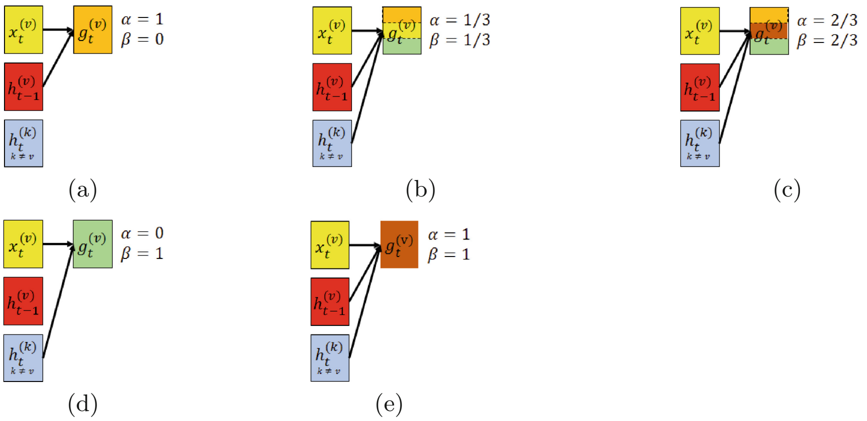


Fig. 2. MV-LSTM topologies. The input update term is represented by g with the superscript indicating the view. (a) View-specific: each view at time t is interacting with the corresponding view representations from time $t-1$. (b) Hybrid topology: a portion of view-specific and cross-views defined by the hyper-parameters α and β at time '$t-1$' is connected at time step t. (c) Hybrid topology: another configuration with different view proportions defined by the hyper-parameters α and β. (d) Coupled topology: each view at time t is interacting with other view representations from time $t-1$. (e) Fully connected topology: all views from time $t-1$ interact with each view at time t. (Color figure online)

4.1 Multi-View Interactions

The view-specific and cross-view interactions are very common in many problems. For example, in a group meeting scenario, it is often the case where a person's utterance at time t is influenced by her utterances and the responses of other people at time $t-1$. In this situation, all modalities are *fully connected* between adjacent time steps. Another type of view relationships that are very common is *dependency relationships*. For example, a child's response at time t will be based on an adult's prompt at time $t-1$ in adult-child interactions. In other example, the adult asking the name of the picture on a book page to a child at time $t-1$, followed by the child's response at time t. Such situations mandate modelling cross-view interactions between adjacent time steps in a *coupled topology*. The other interesting situation is a hybrid scenario, where only a certain portion of views will be interacting with other views between adjacent time steps. For example, in a classical classroom scenario, the teacher has to remember only key highlights or portions of last day's lecture to continue his lecture for today. In such situations, one needs to construct a *hybrid topology* to capture only a portion of corresponding or cross-view data from previous time step to update the view at current time step. Clearly, there is a need to design different topologies to model view-specific and cross-view interactions.

4.2 Model Definition

Figure 1 shows a schematic representation of our MV-LSTM model. Our MV-LSTM is defined by the following update operations for gates and cells[1].

$$i_t^{(v)} = sigm(W_{ix}^{(v)} x_t^{(v)} + W_{ih}^{(v)} A h_{t-1}^{(v)} + \sum_{\substack{k=1 \\ k \neq v}}^{N} W_{ih}^{(k)} B h_{t-1}^{(k)}) \quad {\scriptstyle v \in 1,2,\ldots,N} \tag{7}$$

$$f_t^{(v)} = sigm(W_{fx}^{(v)} x_t^{(v)} + W_{fh}^{(v)} A h_{t-1}^{(v)} + \sum_{\substack{k=1 \\ k \neq v}}^{N} W_{fh}^{(k)} B h_{t-1}^{(k)}) \quad {\scriptstyle v \in 1,2,\ldots,N} \tag{8}$$

$$o_t^{(v)} = sigm(W_{ox}^{(v)} x_t^{(v)} + W_{oh}^{(v)} A h_{t-1}^{(v)} + \sum_{\substack{k=1 \\ k \neq v}}^{N} W_{oh}^{(k)} B h_{t-1}^{(k)}) \quad {\scriptstyle v \in 1,2,\ldots,N} \tag{9}$$

$$g_t^{(v)} = tanh(W_{gx}^{(v)} x_t^{(v)} + W_{gh}^{(v)} A h_{t-1}^{(v)} + \sum_{\substack{k=1 \\ k \neq v}}^{N} W_{gh}^{(k)} B h_{t-1}^{(k)}) \quad {\scriptstyle v \in 1,2,\ldots,N} \tag{10}$$

$$c_t^{(v)} = f_t^{(v)} \odot c_{t-1}^{(v)} + i_t^{(v)} \odot g_t^{(v)} \quad v \in 1, 2, ..., N \tag{11}$$

$$h_t^{(v)} = o_t^{(v)} \odot c_t^{(v)} \quad v \in 1, 2, ..., N \tag{12}$$

$$p_{t+1} = softmax(Z(h_t^{(v)})) \tag{13}$$

where $x_t^{(v)}$ is the input representation at time t for view v. $A \in \mathbb{R}^{c \times d}$ where c is the view gate size and d is view memory cell size. $B \in \mathbb{R}^{c \times d}$ where c is the view gate size and d is the view memory cell size. $h_{t-1}^{(v)}$ is the output from the previous MV-LSTM unit for view v. N is the total number of views. W's are the model parameters. Notice that all gates $(i, f$ and $o)$ and the input update term (g) explicitly model the view-specific and cross-view interactions: $W_{ih}^{(v)} A h_{t-1}^{(v)}$ term models the view-specific and $W_{ih}^{(k)} B h_{t-1}^{(k)}$ models cross-view interactions. Z is a transformation function that concatenates h_t of all views. The symbol $\odot$ denotes an element-wise multiplication of the variables.

[1] We present the update function for chain-like structured output but our derivation can be easily extended to any tree structure.

The two matrices A and B are central to defining the four types of memory cells mentioned above. They are parametrised by the α and β hyper-parameters illustrated in Fig. 2. Formally, matrices A and B are defined as:

$$A[i,i] = 1; i <= \alpha \times d \tag{14}$$

$$A[i,j] = 0; otherwise \tag{15}$$

$$B[i,i] = 1; i >= (1-\beta) \times d \tag{16}$$

$$B[i,j] = 0; otherwise \tag{17}$$

where d represents the memory size of this specific view. When $\alpha = 1/3$ and $\beta = 1/3$, the memory will contain three types of cells: view-specific (shown in orange in Fig. 2), input-oriented (shown in yellow in Fig. 2) and coupled (shown in green in Fig. 2)). $\alpha = 1/3$ means that only a third of the cells will be affected by the same-view hidden state h^v.

The MV-LSTM is different from early fusion of modalities in that it allows four types of interactions between views/modalities. In early fusion, if one of the modalities has strong dynamics, it may overwhelm other modalities during the gate updates. If a modality is negatively influencing the model performance, there is no design flexibility to minimize its effect. MV-LSTM allows flexible integration of modality-specific and cross-modality dynamics.

4.3 Learning

The MV-LSTM parameters are learned using backpropagation. The gradient with respect to all parameters needs to ensure view correspondences and cross-view term updates. Due to space constraints, we provide the gradient computation procedure for a single parameter to demonstrate the changes needed for MV-LSTM and a similar procedure is adopted for all other parameters. The gradient computation for the parameter $W_{ix}^{(v)}$ for the input gate $i_t^{(v)}$ is given by:

$$\partial h_t^{(v)} = \partial y_t W_d \tag{18}$$

$$o_{ft}^{(v)} = sigm(o_t^{(v)}) \tag{19}$$

$$\partial c_t^{(v)} = \partial c_t^{(v)} + o_{ft}^{(v)} \partial h_t^{(v)} \tag{20}$$

$$g_{ft}^{(v)} = tanh(g_t^{(v)}) \tag{21}$$

$$\partial i_{ft}^{(v)} = g_{ft}^{(v)} \partial c_t^{(v)} \tag{22}$$

$$i_{ft}^{(v)} = sigm(i_t^{(v)}) \tag{23}$$

$$\partial i_t^{(v)} = i_{ft}^{(v)}(1 - i_{ft}^{(v)}) \partial i_{ft}^{(v)} \tag{24}$$

$$\partial W_{ix}^{(v)} = \partial W_{ix}^{(v)} + x_t^{(v)} \partial i_t^{(v)} \tag{25}$$

where W_d are the decoder weights and ∂y_t is the output error at time t. All *sigm* operations are computed during the forward procedure. y_t can be inferred at each time step from the LSTM output h_t by selecting the label (e.g. word) with highest probability. These computations ensure view correspondences for a parameter. A similar procedure is applied to weight parameters corresponding to output (o), forget (f) gates and the update term (g).

The topology structure enables view-specific and cross-view interactions between views at adjacent time steps. Hence, the gradient updates for the $h_{t-1}^{(v)}$ term have to be carefully computed using the view-specific connection proportion and cross-view term outputs. The gradient computation for $h_{t-1}^{(v)}$ is given by:

$$\partial h_{t-1}^{(v)} = \partial h_{t-1}^{(v)} + A \partial h_{t}^{(v)} + \sum_{\substack{k=1 \\ k \neq v}}^{N} B \partial h_{t}^{(k)} \tag{26}$$

where a, b, k, v, A, B and N are the same as described earlier.

5 Experiments and Results

The goal of our experiments is threefold: (1) Study the effect of topologies on a multimodal sequence problem that has dynamic interaction between modalities at all times. (2) Compare the results with prior work. (3) Study the MV-LSTM topology when one of the modalities is static.

The following sub-section describes an evaluation of MV-LSTM topologies for multimodal behaviour recognition task. We also compare the results with prior work and analyze the effect of varying α and β values providing a discussion on our findings. Finally, we evaluate our model on the image caption generation task, which adds new challenges.

5.1 Child Engagement Level Prediction Model

The Multimodal Dyadic Behavior Dataset [16] was used in the experiments for predicting the engagement level of a child in a social interaction. In this dataset, an examiner engages a child in five structured play activities or stages. The stages are: greeting the child by saying hello (Greeting), rolling a ball back and forth (Ball), looking through pictures in a book (Book), placing the book on your head to pretend it is a hat (Hat), and gentle tickling (Tickle). Each activity is designed to elicit various behaviours from the child, including common social-communicative behaviours observed in toddlers. In addition, for each stage, the examiner rates how easy or difficult it was to engage the child in the activity, as follows: *0 = Easy to Engage*, *1 = Requires Some Effort to Engage*, and *2 = Requires Extensive Effort to Engage*. The engagement level distribution is biased (>75 %) towards *Easy to Engage* for all stages except for the *Book* stage. Hence, the robustness of the computational model can be validated most effectively for this stage and all our experiments were done only on the *Book* stage. In order to have a balanced dataset, the labels 1 (*Requires Some Effort*) and 2 (*Requires Extensive Effort*) are combined to form a single label, resulting in a binary classification problem.

We employ child's head poses tracked across the video, Histogram of Oriented Gradients (HOG) and Histogram of Oriented Flow (HOF) around the child's upper body region as 3-views for the model. The videos are partitioned into multiple clips and each clip becomes an instance for either training or testing the MV-LSTM networks. The single video label indicating the child's engagement level is propagated to all frames. The video partitioning strategy is similar to the one used by Sharma *et al.* [18].

The HOG, HOF and Headposes are mapped to a common embedding space using linear embedding matrices. The output of this linear transformation using embedded matrices is the final view representation vectors and used as inputs to the MV-LSTM at each time step t. The MV-LSTM cell and gates are partitioned into three equal sized regions corresponding to three input views. The three MV-LSTM topologies (see Fig. 2) are constructed to enable multiple view interactions at each time step. A softmax layer computes the probability distribution of class labels from the MV-LSTM outputs $h_t^{(v)}$. During training, at each time step, the probability, $p(y_t|x_t)$, of obtaining the class label y_t is maximized, given three views of a frame (x_t). During testing, a frame label y_t is predicted at each time step and a video clip label Y is obtained by max-pooling the frame labels $y_{t-1}, y_t, y_{t+1}, ...y_T$. T is the number of frames in each video clip and corresponds to the number of times steps in the MV-LSTM network. Other strategies such as selecting the LSTM output from last time step as the predicted label and averaging the labels over all times are investigated by Sharma *et al.* [18] and found negligible performance difference among strategies. So, a max pooling strategy is used in our work.

Experiment Methodology. The HOG and HOF features are computed around the spatio-temporal interest points [10] in each frame. The child's upper body region is detected using the method proposed by Hoai and Zisserman [5]. The HOF features are mapped to a visual vocabulary built using the HOF features of all frames. The visual word representing the maximum number of interest points is taken as the representative feature for a frame. A similar technique is applied for the HOG features. In addition to these two views, we have used head poses as a third view. The 3 degrees of freedom of a head pose – *Pitch*, *Yaw* and *Roll* – angles were obtained by tracking the child's face using the IntraFace tracker library [24] and used them as features for the third view. All 3-views were employed as inputs at each time step of a MV-LSTM network. For MV-LSTM networks, the input view sizes were set to 32 and size of the memory cell was 96. The learning rate was initialized to 1e-4, dropout to 0.5 and the batch size was 100 in the experiments. Leave-One-Out testing is performed on 59 videos and the precision, recall and F1-scores are computed. The modified version of Neuraltalk [8] codebase from Jia *et al.* [7] was modified for the classification problem and used in the experiments.

Results. To understand the impact of different topologies, we compare a baseline model constructed by early fusion of all modalities, i.e. all modalities are presented at all time steps in a MV-LSTM network. We call this model LSTM (Early fusion). The experiments are conducted for baseline and models with full, coupled and hybrid topologies by configuring α and β parameters. The results are presented in Table 1. The proposed multi-view learning model in a hybrid topology shows improved performance over the baseline model for both engagement levels.

Table 1. The child's engagement level prediction scores using 3-views in MV-LSTM networks for different topologies. In a fully connected topology, all views from time $t-1$ interact with each view at time t (see Fig. 2(e)). In a coupled topology, all views other than the corresponding view at time $t-1$ interact with each view at time t (see Fig. 2(d)). This topology models the cross-view interactions. In a hybrid topology, a portion of corresponding view and all other views from time $t-1$ interact with each view at time t (see Fig. 2(b)). The portion of view-specific connection between adjacent time steps is controlled by a hyperparameter α. The results in this table correspond to $\alpha = 0.1$ and $\beta = 1$. The hybrid topology has performed significantly better for both engagement levels as compared to the LSTM (Early fusion) model, indicating the strength of view interactions in MV-LSTM networks.

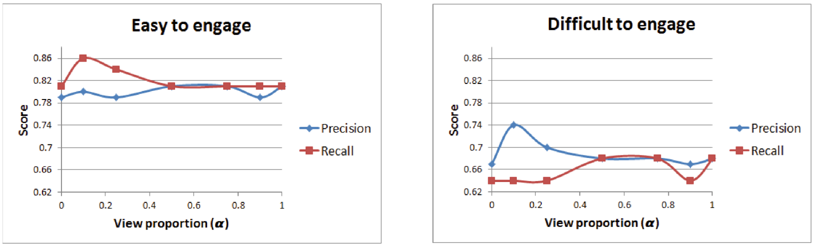
Class labels	Model	Precision	Recall	F1
Easy to engage	LSTM (Early fusion)	0.75	0.81	0.78
	MV-LSTM Full	0.81	0.81	0.81
	MV-LSTM Coupled	0.79	0.81	0.80
	MV-LSTM Hybrid	**0.80**	**0.86**	**0.83**
Difficult to engage	LSTM (Early fusion)	0.63	0.55	0.59
	MV-LSTM Full	0.68	0.68	0.68
	MV-LSTM Coupled	0.67	0.64	0.65
	MV-LSTM Hybrid	**0.74**	**0.64**	**0.68**

Studies on predicting the engagement level of a child in adult-child interactions using the MMDB dataset are relatively new and limited. Rehg *et al.* [16] developed a computational model using object and head trajectories together with audio features to predict engagement ratings. Presti *et al.* [14] proposed a variable Time-Shift Hidden Markov Model for learning and modelling pairs of correlation streams and validated their formulation for predicting the engagement level of a child using the MMDB dataset. Hernandez *et al.* [4] have used the electrodermal activity of the children, obtained from wearable sensors to predict the engagement level of the child. Gupta *et al.* [3] have used the acoustic signals in their models to predict child engagement level. Rajagopalan *et al.* [15] have used the low-level vision features and proposed a two-stage model to predict the engagement level. In all these studies, the set of videos used in their experiments, the experiment methodology and the result metrics all vary and hence no standard benchmark has been established yet. Hence, direct comparison with prior work is not possible. However, we computed the commonly reported accuracy metric for MV-LSTM and presented it in Table 2 along with prior results. The MV-LSTM accuracy outperforms all previous approaches with the exception of Hernandez *et al.* that captured interaction synchrony using child and adult (not used in our work) EDA features. This resulted in better performance on an "easier or harder to engage" binary task.

Table 2. Reported results on child's engagement level prediction accuracies. The MV-LSTM accuracy outperforms all previous approaches with the exception of Hernandez *et al.* [4], however, direct comparison is not possible due a lack of a standard experiment methodology.

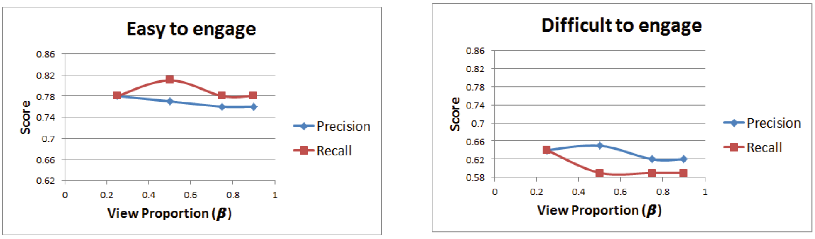
Model	Accuracy
Rehg *et al.* [16]	73.3 %
Presti *et al.* [14]	76.7 %
Hernandez *et al.* [4]	81.0 %
Gupta *et al.* [3]	62.9 %
Rajagopalan *et al.* [15]	74.4 %
MV-LSTM	**77.9 %**

Model Analysis. An interesting design choice with MV-LSTM are the tunable hyper-parameters α and β to control the view-specific and cross-view interactions. We have investigated the model performance for different values of α and β and the results are shown in Figs. 3 and 4. The model performance varies as α changes with a potential to reach a maximum at a certain value. In our experiments, we have found a maximum performance at the $\alpha = 0.1$ or 10 % and $\beta = 1$. This way the view interactions can be fine tuned for a better model performance.



(a) (b)

Fig. 3. The graph showing the change in precision and recall values as the hyperparameter α is tuned. $\beta = 1$ in this experiment. The maximum performance is observed for a hybrid topology with $\alpha = 0.1$ for both engagement levels.



(a) (b)

Fig. 4. The graph showing the change in precision and recall values as the hyperparameter β is tuned. $\alpha = 1$ in this experiment.

A set of statistical significance tests are performed to compare the baseline and the proposed MV-LSTM models. The Asymptotic and Mid-p-value variants of McNemar hypothesis tests [12] are performed. The predicted labels from the baseline and MV-LSTM hybrid topology model at $\alpha = 0.1$ and $\beta = 1.0$ are compared for the model performance. The *null hypothesis* of "MV-LSTM models are more accurate than baseline models" is used in the tests. The p-value obtained was 0.97 and 0.96 for the asymptotic and mid-p variants, respectively, at the 0.05 significance level. The high p-value confirms the null hypothesis that MV-LSTM models are more accurate than baseline models.

5.2 Image Caption Generation

The goal of this task is to generate a rich sentence description for a given image. We used an encoder-decoder pipeline similar to gLSTM [7] and NIC [22] models to validate our MV-LSTM topology. In this pipeline, a vision based deep CNN is used as an encoder module to compute the image representation. This serves as input to the language generator module that uses a recurrent network architecture to generate corresponding natural language descriptions. The language generator module is also referred to as a decoder. A common approach for language generation is the Long Short-Term Memory [6] network that is capable of remembering long range temporal dependencies. In this task, since the image modality remains constant over time, we did not investigate studying different types of view relationships as was done for the children behaviour recognition problem. For this task, to capture the image and text relationships at all times, we have applied the MV-LSTM in a coupled topology structure.

The MV-LSTM memory cell and gates are partitioned into two equal sized regions corresponding to image and text modalities. The inputs to the MV-LSTM are embedded image, text representations and a global semantic context at all time steps. The global semantic context is computed by projecting the CNN image feature representation into a learnt shared representation space using a normalized Canonical Correlation Analysis (CCA) [7]. The MV-LSTM outputs representing the memory cell contents are fed to MV-LSTM gates in a coupled connection, i.e. the output from the memory cell region corresponding to image modality from time step $t-1$ is used to update the gate region corresponding to the text modality at time step t and vice versa. The same process is applied to all MV-LSTM gates (*input*, $forget$ and *output*) and to the input update term g. The memory cell regions $c^{(v)}$ are updated using the corresponding gates and input g. The coupled connection enables interaction between image and text modalities at each time step and the memory cell regions are updated with a joint representation. Finally, the MV-LSTM output corresponding to text region from the last time step is passed to the softmax layer to compute the probability distribution of words in the vocabulary.

The MV-LSTM update operations for the proposed image caption generation model using a coupled topology is given by Eqs. 7–13 for two views, i.e. $N = 2$. The Z transformation function extracts previous output of the text modality. The semantic context information is added to the LSTM update operations as defined by Jia *et al.* [7].

Experiment Methodology. The performance of the model is studied on the Flickr8k, Flickr30k and MS COCO benchmarking datasets. The publicly available splits from Karpathy and Fei-Fei [8] are used in the experiments. The modified version of Neuraltalk [8] codebase from Jia *et al.* [7] is used in the experiments. The hidden layer size, word and image embedding sizes are initialized to 256. The semantic context dimension is set to 200. The learning rate is initialized to 1e-4 and the batch size of 100 is used in the experiments. The beamsize is set to 10, 20 and 10 for the Flickr8k, Flickr30k and MS COCO datasets, respectively. The Gaussian length normalization strategy adopted in Jia *et al.* [7] is used in the experiments. The BLEU [13], METEOR [1], CIDEr [21] and ROUGE [11] metrics are used to evaluate the performance of our model. BLEU is a precision metric that computes the precision of word n-grams between generated and ground truth sentences. BLEU-n is a geometric average of precisions over 1- to n-grams. METEOR considers precision, recall and alignment while computing a score for a generated sentence. The recent CIDEr considers precision, recall, grammar and saliency to compare the sentence similarities.

Results and Discussion The results of our experiments are shown in Table 3. The proposed method achieves state-of-the-art performance on all three datasets. In prior models [7,22], the image modality is presented only at the first time step, which makes it challenging for longer captions where the images would be helpful later in the caption generation process (gLSTM [7] try to prevent this influence loss with a "semantic context" applied at all time steps). The obvious solution of applying the image at each time step was shown to underperform by Vinyals *et al.* [22]. The MV-LSTM manages to integrate the image modality at each time step in a coupled topology where only a portion of the memory cells is influenced by the image modality. This flexible integration results in improved performance over prior models, especially for longer sentences as seen in the BLEU-3 and BLEU-4 scores.

Table 3. Comparison of the proposed model with state-of-the-art methods (higher value is better in each column). Note that our model achieves especially good results on the BLEU-3 and BLEU-4 metrics, indicating its strength when generating long sentences.

Dataset	Model	BLEU 1	BLEU 2	BLEU 3	BLEU 4	METEOR	CIDEr	ROUGE_L
Flickr8K	Log bilinear [9]	65.6	42.4	27.7	17.7	17.3	-	-
	NIC [22]	63.0	41.0	27.0	-	-	-	-
	BRNN [8]	57.9	38.3	24.5	16.0	16.7	31.8	-
	Soft attention [26]	67.0	44.8	29.9	19.5	18.9	-	-
	Hard attention [26]	67.0	45.7	31.4	21.3	20.3	-	-
	gLSTM [7]	64.7	45.9	31.8	21.6	20.1	-	-
	MV-LSTM	65.7	**46.9**	**32.6**	**22.2**	19.9	**53.7**	**46**
Flickr30K	Log bilinear [9]	60.0	38.0	25.4	17.1	16.8	-	-
	NIC [22]	66.3	42.3	27.7	18.3	-	-	-
	BRNN [8]	57.3	36.9	24.0	15.7	15.3	24.7	-
	Soft attention [26]	66.7	43.4	28.8	19.1	18.4	-	-
	Hard attention [26]	66.9	43.9	29.6	19.9	18.4	-	-
	gLSTM [7]	64.6	44.6	30.5	20.6	17.9	-	-
	MV-LSTM	64.5	**44.6**	**31.1**	**21.2**	17.4	**42.0**	**42.2**
MS COCO	Log bilinear [9]	70.8	48.9	34.4	24.3	20.0	-	-
	NIC [22]	66.6	46.1	32.9	24.6	-	-	-
	BRNN [8]	62.5	45.0	32.1	23.0	19.5	66.0	-
	Soft attention [26]	70.7	49.2	34.4	24.3	23.9	-	-
	Hard attention [26]	71.8	50.4	35.7	25.0	23.0	-	-
	gLSTM [7]	67.0	49.1	35.8	26.4	22.7	81.2	-
	MV-LSTM	69.1	**51.5**	**37.7**	**27.6**	22.3	80.2	**49.6**

6 Conclusions

We have extended the LSTM to enable designing different topologies to capture multiple view relationships. The proposed Multi-View LSTM (MV-LSTM) partitions memory cells and gates into multiple regions corresponding to different views. To validate its ability to do multi-view learning and its generalizability to different problem domains, we have constructed topology of MV-LSTM networks and applied them to behaviour recognition and image caption generation problems. Our model has led to better performance due to cross-view learning on both the problems. We have observed that for behaviour recognition problems a simple fusion of multiple modalities may yield a sub-optimal performance, while a multi-view learning can provide better performance by exploiting view relationships. For the image caption generation problem, the proposed model integrating both modalities at all time steps allowed for better longer sentence generation. In future, we plan to apply MV-LSTM to other problem domains.

References

1. Denkowski, M., Lavie, A.: Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the Ninth Workshop on Statistical Machine Translation, pp. 376–380. Association for Computational Linguistics, Baltimore, June 2014. http://www.aclweb.org/anthology/W14-3348

2. Graves, A.: Generating sequences with recurrent neural networks. arXiv:1308.0850 (2013)
3. Gupta, R., Lee, C.C., Bone, D., Rozga, A., Lee, S., Narayanan, S.S.: Acoustical analysis of engagement behavior in children. In: Proceedings of the Workshop on Child, Computer and Interaction, Portland, USA, September 2012
4. Hernandez, J., Riobo, I., Rozga, A., Abowd, G.D., Picard, R.W.: Using electrodermal activity to recognize ease of engagement in children during social interactions. In: Proceedings of the International Conference on Ubiquitous Computing, Seattle, USA, pp. 301–317, September 2014
5. Hoai, M., Zisserman, A.: Talking heads: detecting humans and recognizing their interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Ohio, USA, pp. 875–882, June 2014
6. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
7. Jia, X., Gavves, E., Fernando, B., Tuytelaars, T.: Guiding the long-short term memory model for image caption generation. In: Proceedings of the International Conference on Computer Vision, pp. 2407–2415 (2015)
8. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3128–3137 (2015)
9. Kiros, R., Zemel, R.S., Salakhutdinov, R.: Multimodal neural language models. In: Proceedings of the 31st International Conference on Machine Learning, pp. 595–603 (2014)
10. Laptev, I.: On space-time interest points. Int. J. Comput. Vis. **64**(2–3), 107–123 (2005)
11. Lin, C.Y.: Rouge: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, vol. 8 (2004)
12. McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika **12**(2), 153–157 (1947)
13. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 311–318 (2002)
14. Presti, L.L., Sclaroff, S., Rozga, A.: Joint alignment and modeling of correlated behavior streams. In: Proceedings of the IEEE ICCV Workshop on Decoding Subtle Cues from Social Interactions, Sydney, Australia, pp. 730–737, December 2013
15. Rajagopalan, S.S., Murthy, O.R., Goecke, R., Rozga, A.: Play with me - measuring a childs engagement in a social interaction. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Ljubljana, Slovenia, May 2015
16. Rehg, J.M., Abowd, G.D., Rozga, A., Romero, M., Clements, M.A., Sclaroff, S., Essa, I., Ousley, O.Y., Li, Y., Kim, C., Rao, H., Kim, J.C., Presti, L.L., Zhang, J., Lantsman, D., Bidwell, J., Ye, Z.: Decoding children's social behavior. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, pp. 3414–3421, June 2013
17. Ren, J., Hu, Y., Tai, Y.W., Wang, C., Xu, L., Sun, W., Yan, Q.: Look, listen and learn - a multimodal LSTM for speaker identification. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence (2016)
18. Sharma, S., Kiros, R., Salakhutdinov, R.: Action recognition using visual attention. In: Proceedings of the International Conference on Learning Representations Workshops (2016)

19. Snoek, C.G., Worring, M., Smeulders, A.W.: Early versus late fusion in semantic video analysis. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, pp. 399–402 (2005)
20. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 3104–3112 (2014)
21. Vedantam, R., Zitnick, C.L., Parikh, D.: CIDEr: Consensus-based Image Description Evaluation. arXiv:1411.5726 (2015)
22. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
23. Wang, W., Arora, R., Livescu, K., Bilmes, J.: On deep multi-view representation learning. In: Proceedings of the 32nd International Conference on Machine Learning, pp. 1083–1092 (2015)
24. Xiong, X., de la Torre, F.: Supervised descent method and its application to face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Oregon, USA, pp. 532–539, June 2013
25. Xu, C., Tao, D., Xu, C.: A survey on multi-view learning, April 2013. arXiv:1304.5634. Accessed 15 June 2016
26. Xu, K., Ba, J.L., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning (2015)