EXTENDING LONG SHORT-TERM MEMORY FOR MULTI-VIEW STRUCTURED LEARNING

UNIVERSITY OF CANBERRA

Carnegie Mellon University
Language Technologies Institute

Shyam Sundar Rajagopalan, Louis-Philippe Morency, Tadas Baltrušaitis and Roland Goecke

Shyam.Rajagopalan@canberra.edu.au morency@cs.cmu.edu tbaltrus@cs.cmu.edu roland.goecke@ieee.org

Multi-View Learning

1. **View** – a particular way of observing a phenomena. Eg. (a) Image and text for image captioning, (b) Headpose, HOG, HOF for videos
2. **Multi-View Interactions** – View-specific dynamics capture interactions between hidden outputs from the same view, while cross-view dynamics capture interactions between hidden outputs of different views. Both interactions are very common in many problems.

Related Work

Deep Multi-View Representation Learning Models - CCA + Autoencoder, LSTM as language decoder integrating images and text, Multimodal LSTM for speaker identification. Models are applied to behaviour recognition and image captioning problems.

Existing models lack flexibility in designing multiple topologies to model view-specific and cross-view interactions.

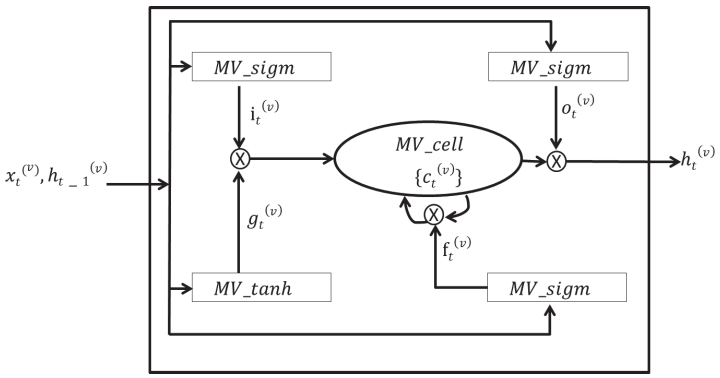
Contributions

Multi-View LSTM (MV-LSTM), an extension to LSTM, is designed to model view-specific and cross-view interactions.

The LSTM internal representations are partitioned to mirror multiple input views.

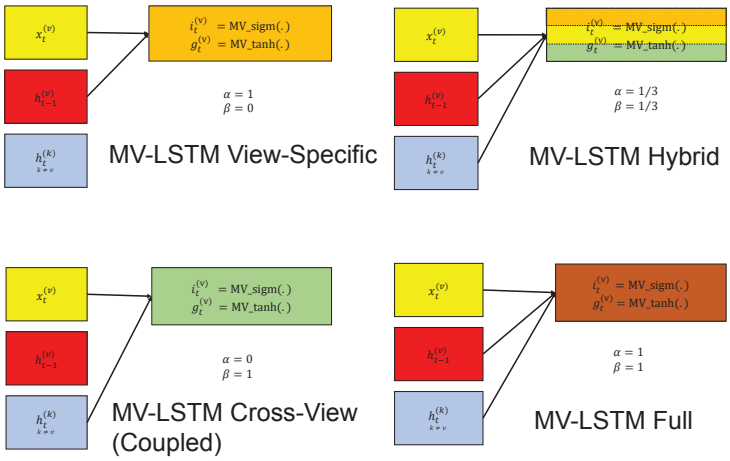
Family of activation functions, MV-Sigmoid and MV-Tanh to update MV-LSTM internal memory partitions.

Multi-View LSTM (MV-LSTM)



$x_t^{(v)}$ represents v -th view input at time step 't' and $h_{t-1}^{(v)}$ is the MV-LSTM output from time step 't-1' corresponding to the v -th view.

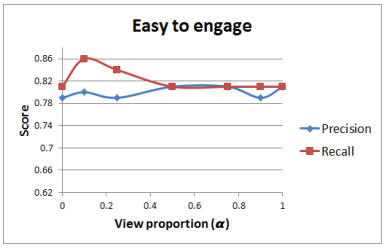
Multi-View Interaction Topologies

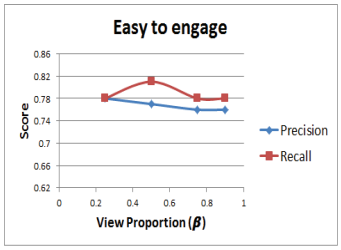


Experiment Results – Child's Engagement Level Prediction – Multimodal Dyadic Behavior Dataset

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**

Model Analysis





Experiment Results – Image Caption Generation

Dataset	Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Flickr8K	Log Bilinear	65.6	42.4	27.7	17.7
	NIC	63.0	41.0	27.0	-
	BRNN	57.9	38.3	24.5	16.0
	Soft Attention	67.0	44.8	29.9	19.5
	Hard Attention	67.0	45.7	31.4	21.3
	gLSTM	64.7	45.9	31.8	21.6
	MV-LSTM	65.7	**46.9**	**32.6**	**22.2**
Flickr30K	Log Bilinear	60.0	38.0	25.4	17.1
	NIC	66.3	42.3	27.7	18.3
	BRNN	57.3	36.9	24.0	15.7
	Soft Attention	66.7	43.4	28.8	19.1
	Hard Attention	66.9	43.9	29.6	19.9
	gLSTM	64.6	44.6	30.5	20.6
	MV-LSTM	64.5	**44.6**	**31.1**	**21.2**
MS-COCO	Log Bilinear	70.8	48.9	34.4	24.3
	NIC	66.6	46.1	32.9	24.6
	BRNN	62.5	45.0	32.1	23.0
	Soft Attention	70.7	49.2	34.4	24.3
	Hard Attention	71.8	50.4	35.7	25.0
	gLSTM	67.0	49.1	35.8	26.4
	MV-LSTM	69.1	**51.5**	**37.7**	**27.6**

Conclusion

1. Extended LSTM for designing multiple topologies to model view relationships

2. Cross-view learning using the proposed model helps in better representation for behaviour recognition and image captioning

European Conference on Computer Vision (ECCV 2016), Oct 8-16, 2016, Amsterdam, Netherlands.