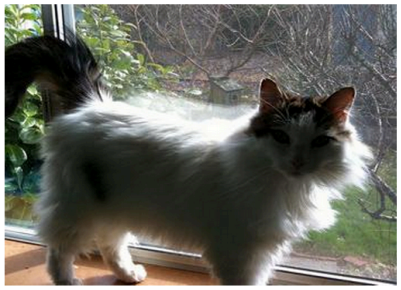
TWO IMPLEMENTATIONS OF RNNS IN IMAGE DESCRIPTION

ZHIYUAN GUO ZG2201

Explain Images with Multimodal Recurrent Neural Networks



a cat sitting on a bench in front of a window

a man riding a snowboard down a snow covered slope

a train traveling down a train track next to a forest

a man riding a wave on top of a surfboard

a clock tower with a clock on top of it

a person on skis is skiing down a hill

Explain Images with Multimodal Recurrent Neural Networks

- Motivation
- Current works can only label the query image with the sentence annotations of the images already existing in the datasets.
- Propose a m-RNN model to address both the task of generating novel sentences descriptions for images and task of image and sentence retrieval.

Explain Images with Multimodal Recurrent Neural Networks

- Related Work

1. Deep model for computer vision and natural language
2. Image-sentence retrieval
3. Generating novel sentence descriptions for images

- Creative point

Incorporate the RNN in a deep multimodal architecture

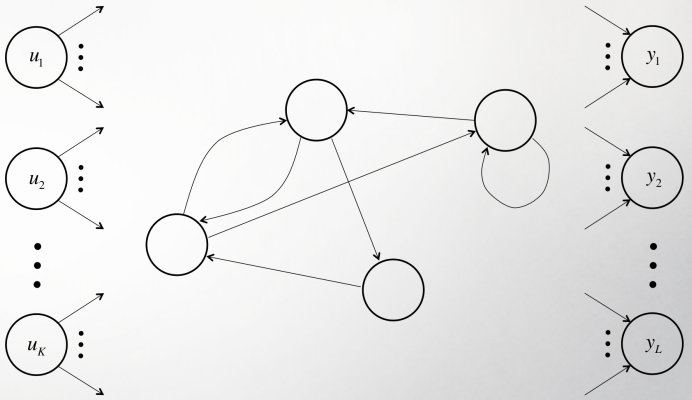
Explain Images with Multimodal Recurrent Neural Networks

- Simple RNN

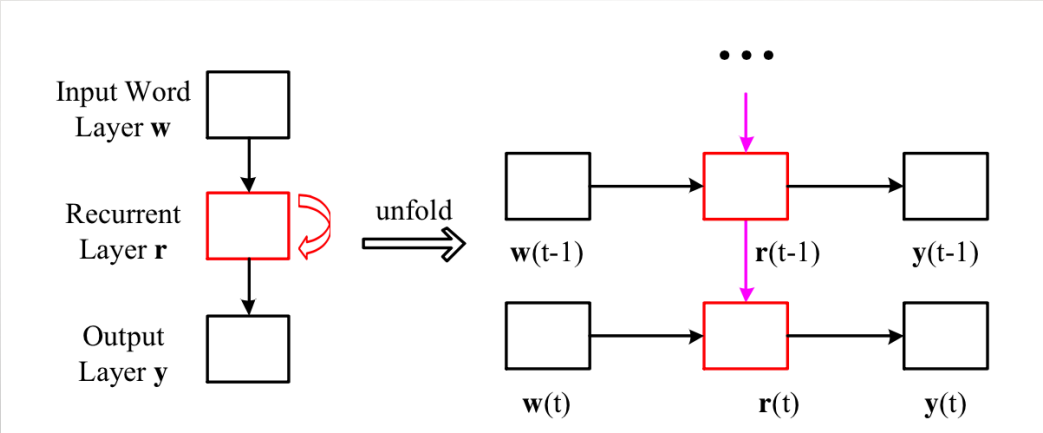
Contain at least one directed cycle.

Train using back propagation through time.

Gradient vanishing problem



Explain Images with Multimodal Recurrent Neural Networks

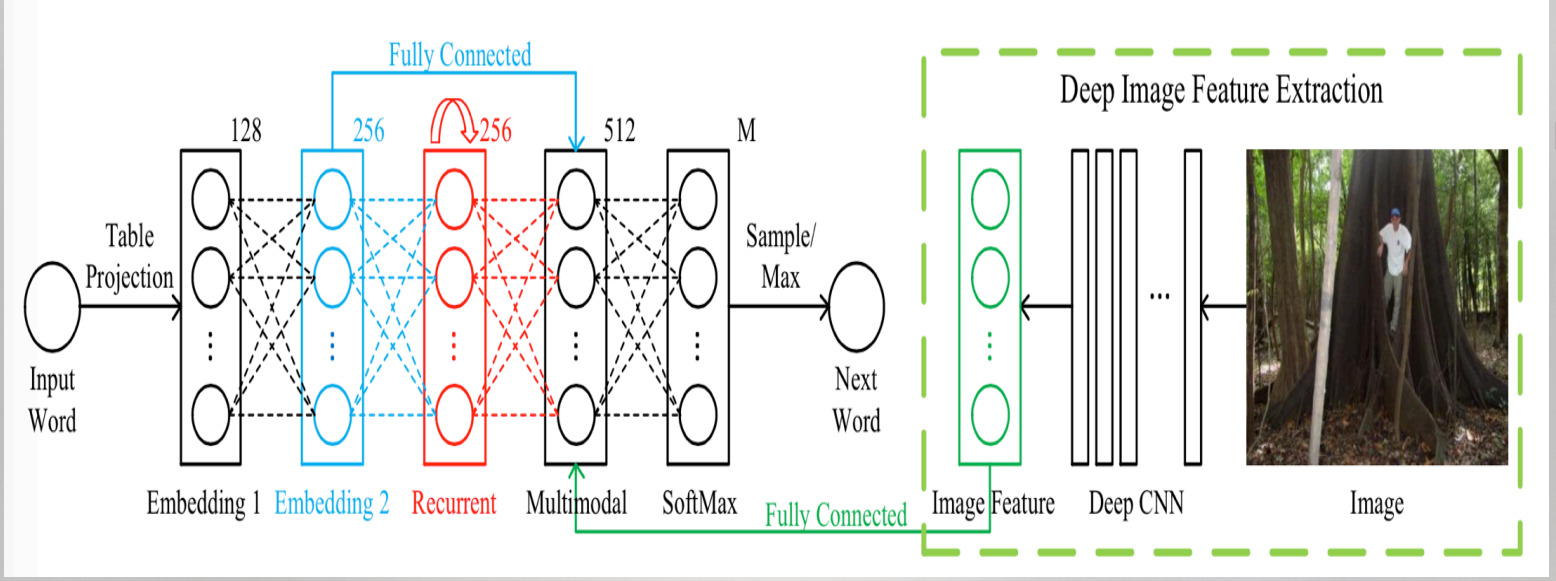


$$\mathbf{x}(t) = [\mathbf{w}(t) \;\; \mathbf{r}(t-1)]; \quad \mathbf{r}(t) = f_1(\mathbf{U} \cdot \mathbf{x}(t)); \quad \mathbf{y}(t) = g_1(\mathbf{V} \cdot \mathbf{r}(t));$$

Input is the one-hot representation, f and g are element-wised sigmoid and softmax function respectively. U and V are weights to learn.

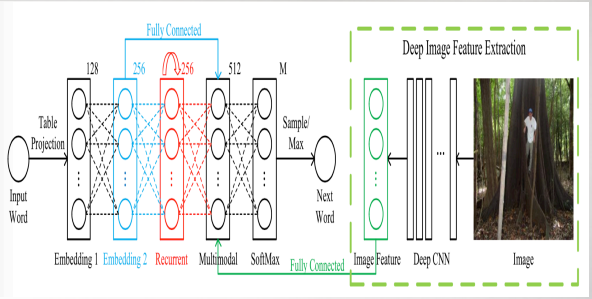
Explain Images with Multimodal Recurrent Neural Networks

m-RNN Model



Explain Images with Multimodal Recurrent Neural Networks

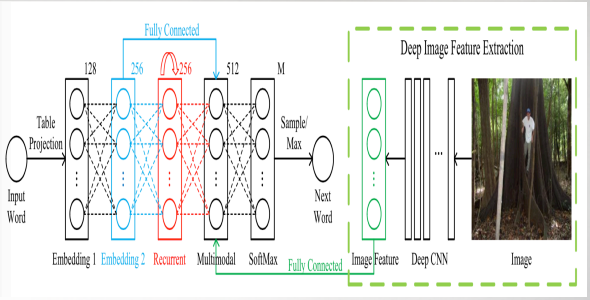
$$\mathbf{r}(t) = f_2(\mathbf{U}_r \cdot \mathbf{r}(t-1) + \mathbf{w}(t))$$



f2 function is set as the Rectified Linear Unit. ReLU is faster and harder to saturate or overfit the data than sigmoid function. This improvement is for vanishing gradient problem in RNN.

Explain Images with Multimodal Recurrent Neural Networks

Multimodal layer connect the language model part and the image part.



$$\mathbf{m}(t) = g_2(\mathbf{V}_w \cdot \mathbf{w}(t) + \mathbf{V}_r \cdot \mathbf{r}(t) + \mathbf{V}_I \cdot \mathbf{I})$$

M denotes the multimodal layer feature vector.

$$g_2(x) = 1.7159 \cdot \tanh(\frac{2}{3}x)$$

Explain Images with Multimodal Recurrent Neural Networks

- Training the m-RNN

Cost Function: perplexity of the sentences given corresponding image

$$\log_2 \mathcal{PPL}(w_{1:L}|\mathbf{I}) = -\frac{1}{L}\sum_{n=1}^{L} \log_2 P(w_n|w_{1:n-1}, \mathbf{I})$$

$$\mathcal{C} = \frac{1}{N}\sum_{i=1}^{N} L \cdot \log_2 \mathcal{PPL}(w_{1:L}^{(i)}|\mathbf{I}^{(i)}) + \|\theta\|_2^2$$

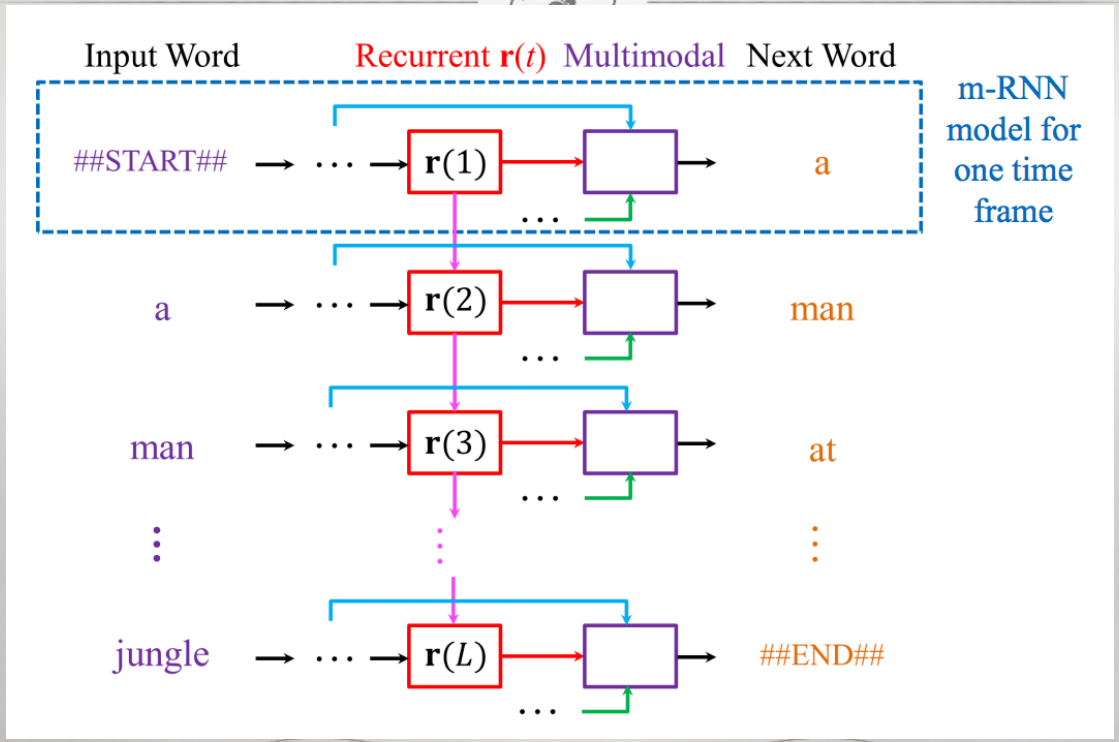
The cost function is the average log-likelihood of the words given their context words and corresponding images in the training sentences plus a regularization term. Training objective is to minimize cost function.

Explain Images with Multimodal Recurrent Neural Networks

- Utility

1. Sentence Generation: start from ##START## and end at ##END##
2. Image and Sentence Retrieval

Explain Images with Multimodal Recurrent Neural Networks



Explain Images with Multimodal Recurrent Neural Networks

	$\mathcal{PPL}$	B-1	B-2	B-3
BACK-OFF GT2	54.5	0.323	0.145	0.059
BACK-OFF GT3	55.6	0.312	0.131	0.059
LBL [26]	20.1	0.327	0.144	0.068
MLBL-B-DeCAF [16]	24.7	0.373	**0.187**	0.098
MLBL-F-DeCAF [16]	21.8	0.361	0.176	0.092
Gupta et al. [11]	/	0.15	0.06	0.01
Gupta & Mannem [10]	/	0.33	0.18	0.07
Ours-RNN-Base	7.77	0.3134	0.1168	0.0803
Ours-m-RNN	**6.92**	**0.3951**	0.1828	**0.1311**

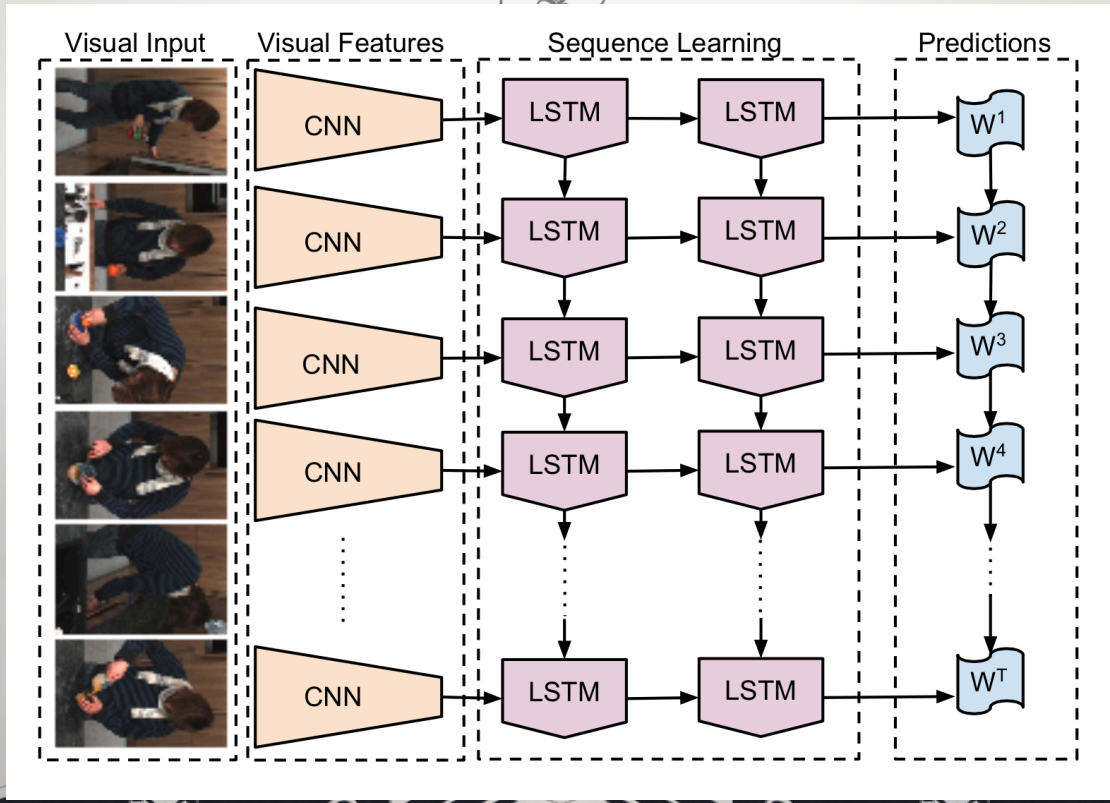
Table 1: Results of the sentence generation task on the IAPR TC-12 dataset. "B" is short for BLUE

Explain Images with Multimodal Recurrent Neural Networks

	Sentence Retrival (Image to Text)				Image Retrival (Text to Image)			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random	0.1	0.5	1.0	631	0.1	0.5	1.0	500
Socher-decaf [30]	4.5	18.0	28.6	32	6.1	18.5	29.0	29
Socher-avg-rcnn [30]	6.0	22.7	34.0	23	6.6	21.6	31.7	25
DeViSE-avg-rcnn [6]	4.8	16.5	27.3	28	5.9	20.1	29.6	29
DeepFE-decaf [15]	5.9	19.2	27.3	34	5.2	17.6	26.5	32
DeepFE-rcnn [15]	12.6	32.9	44.0	14	9.7	29.6	**42.5**	**15**
Ours-m-RNN-decaf	**14.5**	**37.2**	**48.5**	**11**	**11.5**	**31.0**	42.4	**15**

	Sentence Retrival (Image to Text)				Image Retrival (Text to Image)			
	R@1	R@5	R@10	Med r	R@1	R@5	R@10	Med r
Random	0.1	0.6	1.1	631	0.1	0.5	1.0	500
DeViSE-avg-rcnn [6]	4.8	16.5	27.3	28	5.9	20.1	29.6	29
DeepFE-rcnn [15]	16.4	**40.2**	**54.7**	**8**	10.3	**31.4**	**44.5**	**13**
Ours-m-RNN-decaf	**18.4**	**40.2**	50.9	10	**12.6**	31.2	41.5	16

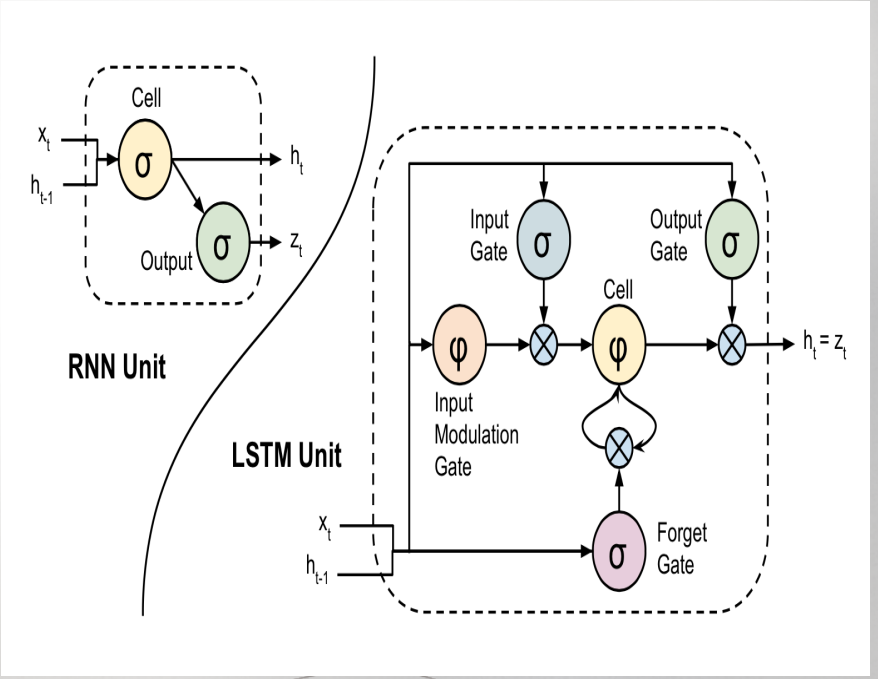
Long-term Recurrent Convolutional Networks



Long-term Recurrent Convolutional Networks

LSTM- long short term memory

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



Long-term Recurrent Convolutional Network

- LSTM

Due to vanishing gradients problem that can result from propagating the gradients down through many layers, RNNs could be difficult to train to be learned in long-term dynamics.

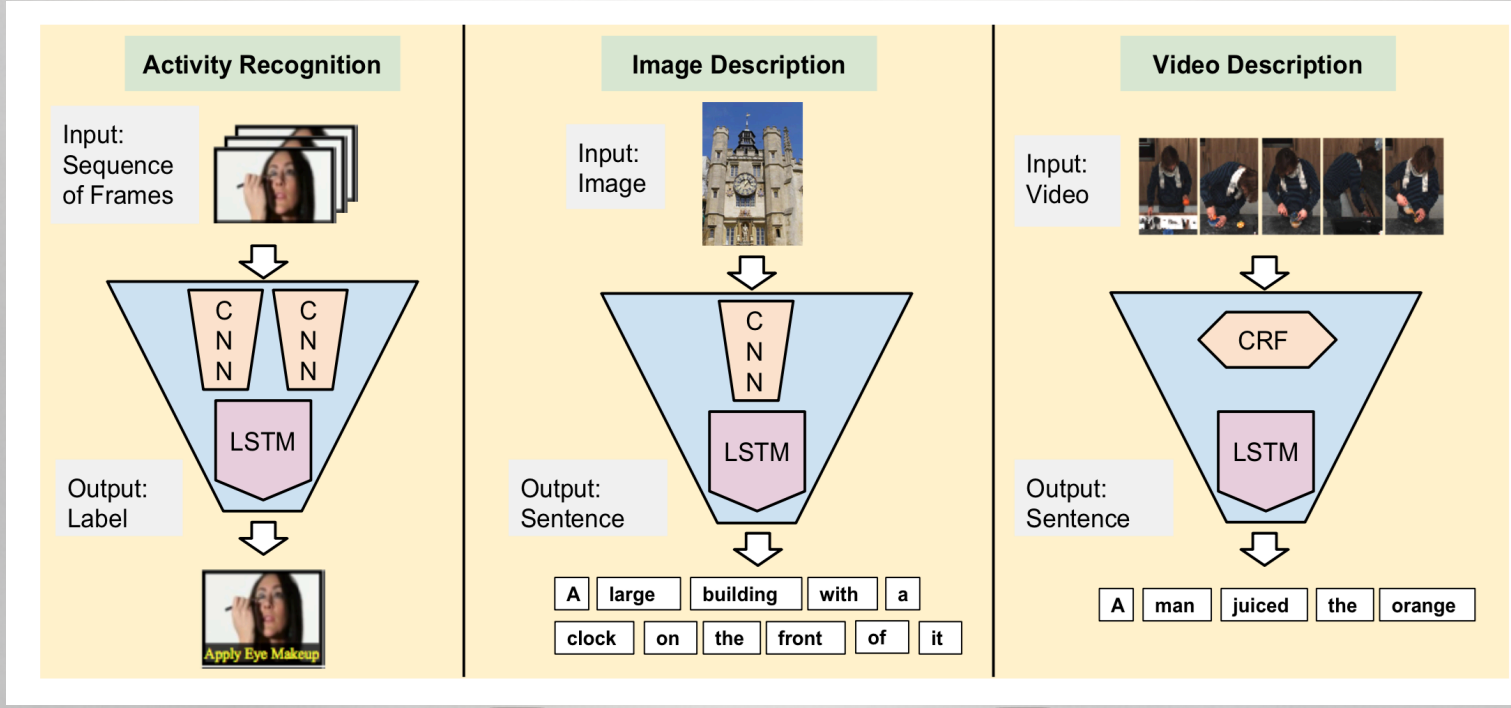
LSTM provide a solution by incorporating memory units that allow the network to learn when to forget previous hidden states and when to update hidden states given new information.

Long-term Recurrent Convolutional Network

- Procedure

1. Passing each visual input v through a feature transformation. Compute the feature-space representation.
2. Sequence model maps input and a previous timestep hidden state to an output.
3. The final step is taking a softmax over the outputs of the sequential model.

Long-term Recurrent Convolutional Network



Long-term Recurrent Convolutional Network

- Activity recognition

T individual frames are inputs into T convolutional networks which are then connected to a single-layer LSTM with 256 hidden units.

Long-term Recurrent Convolutional Network

Model	Input Type		Weighted Average	
	RGB	Flow	$1/2, 1/2$	$1/3, 2/3$
Single frame	65.40	53.20	–	–
Single frame (ave.)	69.00	72.20	75.71	79.04
LRCN-fc_6	**71.12**	**76.95**	81.97	**82.92**
LRCN-fc_7	70.68	69.36	–	–

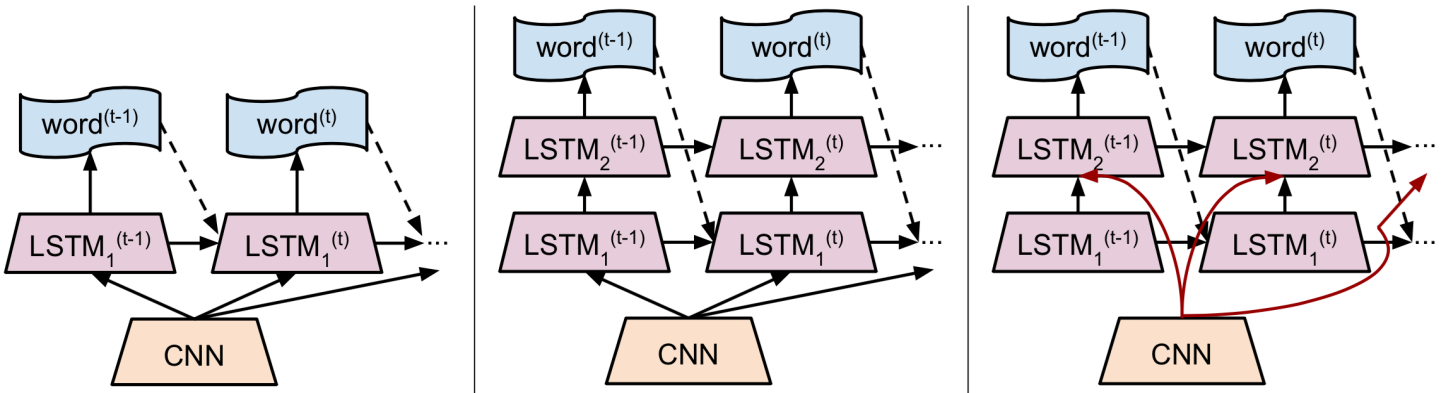
Long-term Recurrent Convolutional Network

- Image description

At each timestep, both the image features and the previous word are provided as inputs to the sequential model.

At timestep t, the input to the bottom-most LSTM is the embedded ground truth word from the previous timestep. The second LSTM fuses the outputs of the bottom-most LSTM with the image representation.

Long-term Recurrent Convolutional Network



Single Layer
LRCN$_{1u}$

Two Layers, Unfactored
LRCN$_{2u}$

Two Layers, Factored
LRCN$_{2f}$

Figure 4: Three variations of the LRCN image captioning architecture that we experimentally evaluate. We explore the effect of depth in the LSTM stack, and the effect of the "factorization" of the modalities (also explored in [19]).

Long-term Recurrent Convolutional Network

	R@1	R@5	R@10	Med*r*
DeViSE [8]	6.7	21.9	32.7	25
SDT-RNN [35]	8.9	29.8	41.1	16
DeFrag [15]	10.3	31.4	44.5	13
m-RNN [25]	12.6	31.2	41.5	16
ConvNet [18]	10.4	31.0	43.7	14
LRCN_{2f} (ours)	**17.5**	**40.3**	**50.8**	**9**

Long-term Recurrent Convolutional Network

	Flickr30k [28]			
	B-1	B-2	B-3	B-4
m-RNN [25]	54.79	23.92	19.52	-
1NN fc_8 base (ours)	37.34	18.66	9.39	4.88
1NN fc_7 base (ours)	38.81	20.16	10.37	5.54
LRCN (ours)	**58.72**	**39.06**	**25.12**	**16.46**

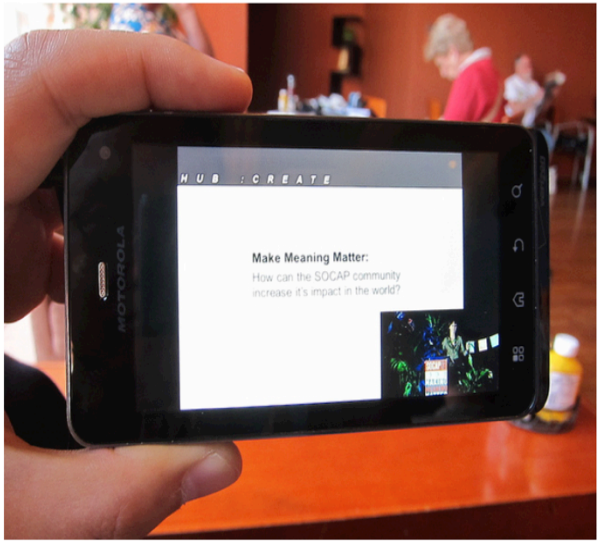
A female tennis player in action on the court.

A group of young men playing a game of soccer

A man riding a wave on top of a surf-board.

A baseball game in progress with the batter up to plate.

A brown bear standing on top of a lush green field.



A person holding a cell phone in their hand.