Long-term Recurrent Convolutional Networks for Visual Recognition and Description

Jeff Donahue[1], Lisa Anne Hendricks[1] Sergio Guadarrama[1] Marcus Rohrbach[1,2] Subhashini Venugopalan[3] Kate Saenko[4] Trevor Darrell[1,2]
[1]UC Berkeley, Berkeley, CA. [2]ICSI, Berkeley, CA. [3]UT Austin, Austin, TX. [4]UMass Lowell, Lowell, MA.

Models based on deep convolutional networks have dominated recent image interpretation tasks; we investigate whether models which are also recurrent, or "temporally deep", are effective for tasks involving sequences, visual and otherwise. We develop a novel recurrent convolutional architecture suitable for large-scale visual learning which is end-to-end trainable, and demonstrate the value of these models on benchmark video recognition tasks, image description and retrieval problems, and video narration challenges. In contrast to current models which assume a fixed spatio-temporal receptive field or simple temporal averaging for sequential processing, recurrent convolutional models are "doubly deep" in that they can be compositional in spatial and temporal "layers". Such models may have advantages when target concepts are complex and/or training data are limited. Learning long-term dependencies is possible when nonlinearities are incorporated into the network state updates. Long-term RNN models are appealing in that they can directly map variable-length inputs (e.g., video frames) to variable length outputs (e.g., natural language text) and can model complex temporal dynamics; yet they can be optimized with backpropagation. Our recurrent long-term models are directly connected to modern visual convnet models and can be jointly trained to simultaneously learn temporal dynamics and convolutional perceptual representations. Our results show such models have distinct advantages over state-of-the-art models for recognition or generation which are separately defined and/or optimized.

Recognition and description of images and videos is a fundamental challenge of computer vision. Dramatic progress has been achieved by supervised convolutional models on image recognition tasks, and a number of extensions to process video have been recently proposed. Ideally, a video model should allow processing of variable length input sequences, and also provide for variable length outputs, including generation of full-length sentence descriptions that go beyond conventional one-versus-all prediction tasks. We propose *Long-term Recurrent Convolutional Networks* (LRCNs), a class of architectures leveraging the strengths of rapid progress in CNNs for visual recognition problem, and the growing desire to apply such models to time-varying inputs and outputs. LRCN processes the (possibly) variable-length visual input with a CNN (Figure 1 left), whose outputs are fed into a stack of LSTMs (Figure 1 middle), which finally produce a (possibly) variable-length prediction (Figure 1 right).

We show that long-term recurrent convolutional models are generally applicable to visual time-series modeling; we argue that in visual tasks where static or flat temporal models have previously been employed, long-term RNNs can provide significant improvement when ample training data are available to learn or refine the representation. Specifically, we show LSTM-type models provide for improved recognition on conventional video activity challenges and enable a novel end-to-end optimizable mapping from image pixels to sentence-level natural language descriptions. We also show that these models improve generation of descriptions from intermediate visual representations derived from conventional visual models.

We instantiate our proposed architecture in three experimental settings. First, we show that directly connecting a visual convolutional model to deep LSTM networks, we are able to train video recognition models that capture complex temporal state dependencies. We train our model with both RGB images and flow, and find that the best classification accuracy is achieved by averaging the outputs of a network trained with RGB and a network trained with flow. Our experiments demonstrate that learning temporal dynamics increases performance more when using flow as input as opposed to RGB. While existing labeled video activity datasets may not have actions or activities with extremely complex time dynamics, we nonetheless see improvements on the order of 4% on conventional benchmarks.

Second, we explore direct end-to-end trainable image to sentence mappings. Strong results for machine translation tasks have recently been reported [1, 4]; such models are encoder/decoder pairs based on LSTM networks. We propose a multimodal analog of this model, and describe an architecture which uses a visual convnet to encode a deep state vector, and an LSTM to decode the vector into an natural language string. The resulting model can be trained end-to-end on large-scale image and text datasets, and even with modest training provides competitive generation results compared to existing methods.

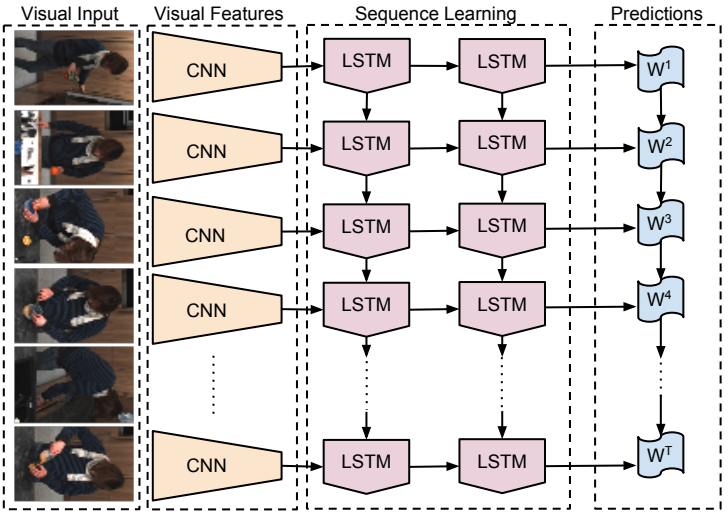

Figure 1: We propose *Long-term Recurrent Convolutional Networks* (LRCNs), a class of architectures leveraging the strengths of rapid progress in CNNs for visual recognition problem, and the growing desire to apply such models to time-varying inputs and outputs. LRCN processes the (possibly) variable-length visual input (left) with a CNN (middle-left), whose outputs are fed into a stack of recurrent sequence models (LSTMs, middle-right), which finally produce a variable-length prediction (right)

Finally, we show that LSTM decoders can be driven directly from conventional computer vision methods which predict higher-level discriminative labels, such as the semantic video role tuple predictors in [3]. While not end-to-end trainable, such models offer architectural and performance advantages over previous statistical machine translation-based approaches, as reported in our paper.

We have realized a generalized "LSTM"-style RNN model in the widely-adopted open source deep learning framework *Caffe* [2], incorporating the specific LSTM units of [1, 4, 5]. It is available at jeffdonahue.com/lrcn/.

[1] Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[2] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM MM*, 2014.

[3] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. Translating video content to natural language descriptions. In *ICCV*, 2013.

[4] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[5] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.