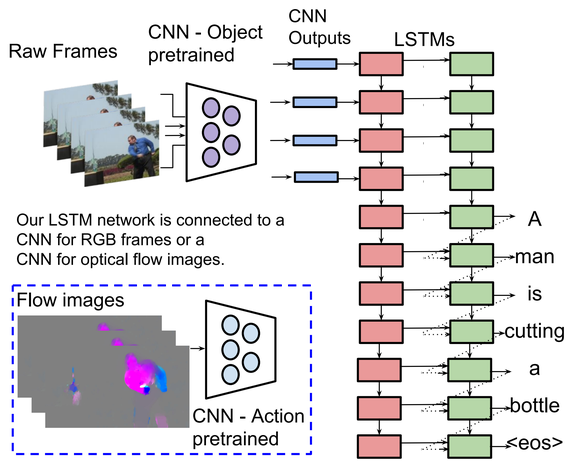
Sequence to Sequence – Video to Text (S2VT)

Subhashini Venugopalan[1], Marcus Rohrbach[2], Jeff Donahue[2],
Raymond Mooney[1], Trevor Darrell[2], Kate Saenko[3]
[1] UT-Austin [2] UC-Berkeley [3] UMass-Lowell

Goals

Given a video clip, output a natural language sentence that describes the event depicted in the video.

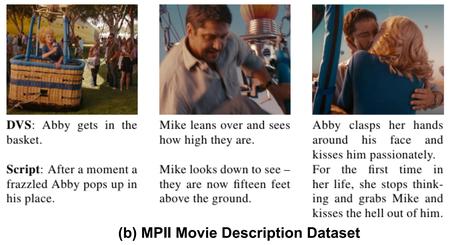


Our model uses a CNN-RNN based encoder-decoder approach and learns to associate a sequence of video frames to a sequence of words in order to generate a description of the event in the video clip. The model is trained on raw RGB frames as well as optical flow images.

Datasets

We demonstrate our approach on large, realistic collections of YouTube videos and Hollywood movie clips.

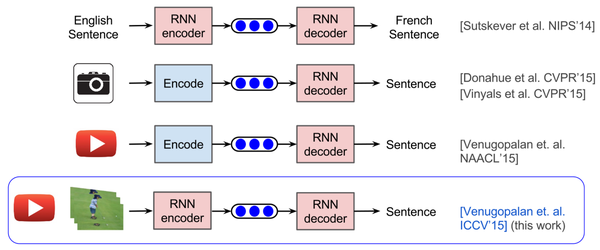
(a) YouTube Video corpus



(b) MPII Movie Description Dataset

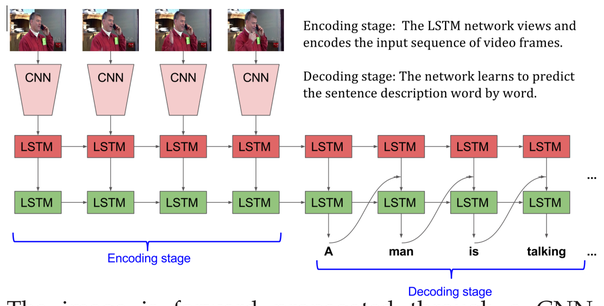
The YouTube dataset, collected by Chen and Dolan (ACL 2011) consists of 1970 videos, where each video is accompanied by about 41 human descriptions (sentences), see (a) above. We also show results on two large movie description corpora - the Montreal (M-VAD) and MPII movie description datasets consisting of nearly 200 Hollywood movies with DVS sentences, see (b) above.

Insight



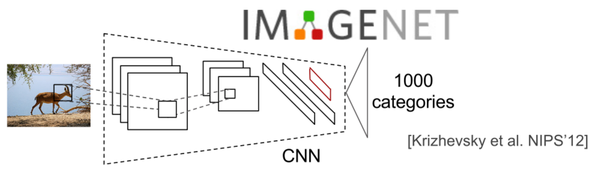
The broad idea of our approach is to encode a video frame sequence and decode it to a sequence of english words (sentence) using LSTMs.

Overview



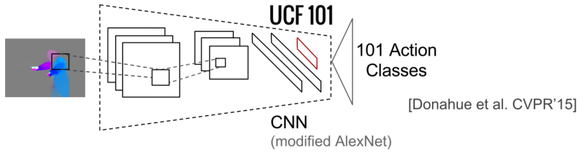
The image is forward propagated through a CNN. The activations of the fully connected layer just before classification forms the input to the LSTM network.

Raw RGB Frames: CNN is initialized with weights from a model trained on the ImageNet classification task.



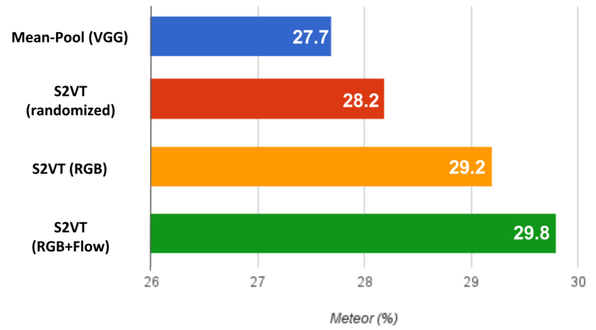
Optical Flow Frames: Additionally, our model incorporates activity information using optical flow.

Optical flow is extracted by considering two consecutive frames in the video (above). The flow CNN in the S2VT network is initialized with weights from a model trained on the UCF101 activity recognition task (below).



Results - Youtube Dataset

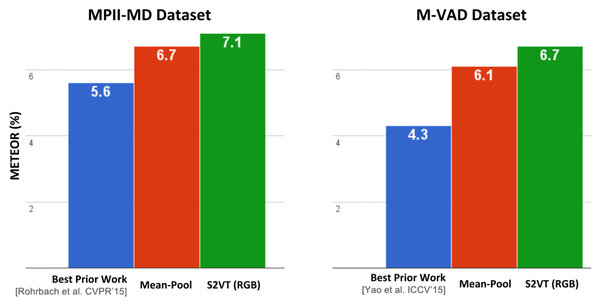
Results on the MSVD corpus of Youtube video clips.



We use the machine translation metric METEOR to compare the quality of the generated description against the multiple ground truth reference sentences.

Results - Movie Description

Results on the MPII-MD and M-VAD movie corpora.

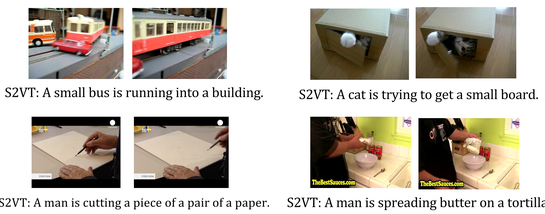


We use METEOR to compare the predicted sentence against the ground truth DVS description of the clip.

Qualitative Results - Youtube

Correct descriptions.

Relevant but incorrect descriptions.



Irrelevant descriptions.

Links

Project Page:
`http://vsubhashini.github.io/s2vt.html`
Code: `https://github.com/vsubhashini/caffe/tree/recurrent/examples/s2vt`

Qualitative Results - Movie Description

Representative frames of 6 contiguous clips from the movie "Big Mommas: Like Father, Like Son". Below are descriptions generated by prior art (Temporal Attention), and our model (S2VT), as well as the groundtruth (DVS) sentences.

Temporal Attention (GNet+3D-conv$_{att}$): (1) At night , SOMEONE and SOMEONE step into the parking lot. (2) Now the van drives away. (3) They drive away. (4) They drive off. (5) They drive off. (6) At the end of the street , SOMEONE sits with his eyes closed.

S2VT (Ours): (1) Now, the van pulls out a window and a tall brick facade of tall trees . a figure stands at a curb. (2) Someone drives off the passenger car and drives off. (3) They drive off the street. (4) They drive off a suburban road and parks in a dirt neighborhood. (5) They drive off a suburban road and parks on a street. (6) Someone sits in the doorway and stares at her with a furrowed brow.

DVS: (1) Now , at night , our view glides over a highway its lanes glittering from the lights of traffic below. (2) Someone's suv cruises down a quiet road. (3) Then turn into a parking lot . (4) A neon palm tree glows on a sign that reads oasis motel. (5) Someone parks his suv in front of some rooms. (6) He climbs out with his briefcase , sweeping his cautious gaze around the area.