Sequence to Sequence – Video to Text (S2VT)
Subhashini Venugopalan1, Marcus Rohrbach2, Jeff Donahue2, Raymond Mooney1, Trevor Darrell2, Kate Saenko3 
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.

Raw Frames
CNN - Object pretrained
CNN Outputs
LSTMs
Our LSTM network is connected to a CNN for RGB frames or a CNN for optical flow images.

Flow images
CNN - Action pretrained
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
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 clips.

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

The YouTube video 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 corpora - the Montreal (M-VAD) and MPII movie description corpora.

RESULTS - YOUTUBE DATASET
Results on the MSVD corpus of YouTube video clips.

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

RESULTS - YOUTUBE DATASET
Results on the MSVD corpus of YouTube video clips.

QUALITATIVE RESULTS - YOUTUBE
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) description of the clip.

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) description of the clip.

LINKS
Code: https://github.com/vsubhashini/s2vt.html
CAFFE: https://github.com/vsubhashini/caffe/tree/recurrent/examples/s2vt

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

S2VT (Durs): (1) Now, the van pulls out a window and a tall brick facade of tall trees. (2) A figure stands at a curb.
(3) They drive off the passenger car and drives off.
(4) They drive off.
(5) They drive.
(6) Someone sits in the doorway and stares at her oasis motel.

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.