Exploring Deep Learning Models for Video Captioning

Haithem Afli, Feiyan Hu, Jinhua Du, Daniel Cosgrove, Kevin McGuinness, Noel E. O'Connor, Eric Arazo Sanchez, Jiang Zhou and Alan F. Smeaton.

Engaging Content
Engaging People

Insight
Centre for Data Analytics

Introduction

The Insight and ADAPT research centres at Dublin City University collaborated for the video captioning task proposed in the TRECVid 2017 competition. Four runs were submitted following the three approaches showed below. In the first two approaches keyframes are extracted from the videos and provided to a image-to-caption CNN-RNN model [1] that generates natural language descriptions from each image.

In these two first approaches a machine translation combination system [2] based on statistical methods is used to recombine the captions for each keyframe in a single caption for the whole video. The third approach consist on an end-to-end Deep Learning model [3] that learns to generate captions directly from the videos.

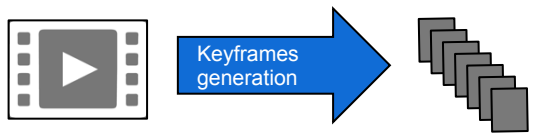
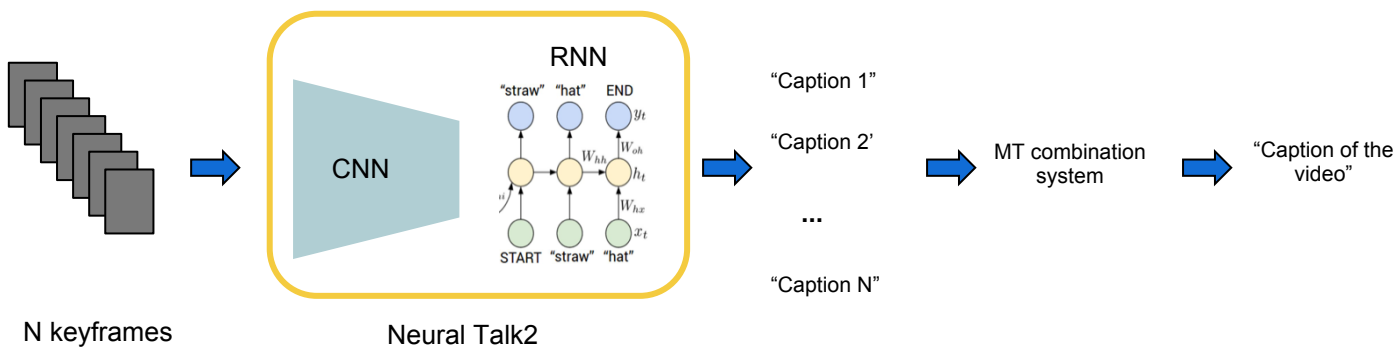


Figure 1. N Keyframes extracted from each video.

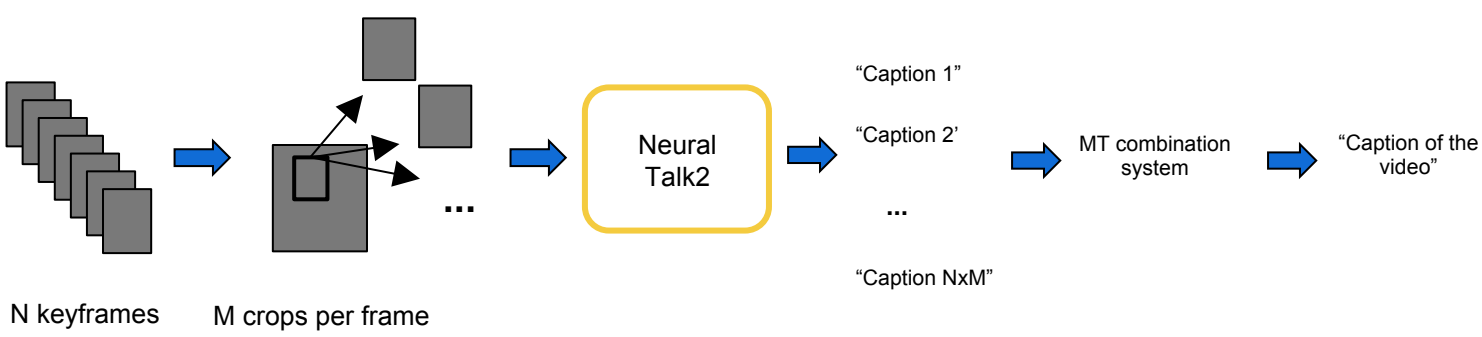
Approach 1 (run 1 and 2)

- One caption is generated for each keyframe.
- One run implemented in Pytorch and the other in Tensorflow.



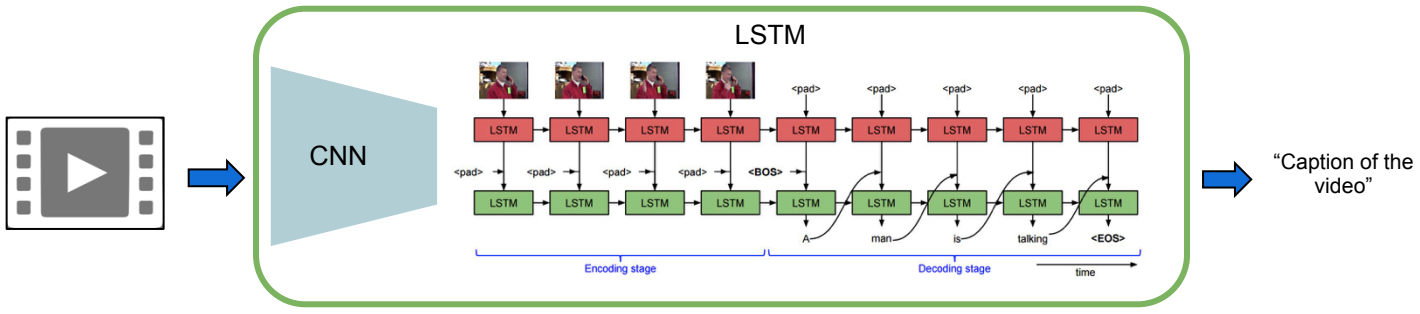
Approach 2 (run 3)

- M crops are extracted for each keyframe.
- Crops based on salient regions from the image.
- One caption is generated for each crop.



Approach 3 (run 4)

- Features generated with a CNN and yielded to a 2 LSTM stack.
- The LSTM's encode the features and decode them into natural language descriptions.



Sequence to Sequence - Video to Text (S2VT)

Results and Performance

- Very similar results in run 1 and 2 (both from approach 1).
- The results from the run 3 show that approach 2 needs some refinement to use the salient information from the frames.
- Run 4 is far from the other results, which might be due to the time consuming nature of the experiments.

	CIDEr	CIDEr-D
run 1	**0.184**	**0.122**
run 2	0.183	**0.122**
run 3	0.146	0.093
run 4	0.073	0.041

Table 1. Results for the four submissions in terms of CIDEr and CIDEr-D.

References:
1. Karpathy et al. "Deep visual-semantic alignments for generating image descriptions." CVPR, 2015.
2. Du et al. MaTrEx: The DCU MT System for WMT 2009. WMT, 2009.
3. Venugopalan et al. Sequence to Sequence – Video to Text. ICCV, 2015.

A World Leading SFI Research Centre

This project has been funded by Science Foundation Ireland under grant numbers SFI/12/RC/2289 (Insight Centre), SFI/13/RC/2106 (ADAPT Centre) and SFI/15/SIRG/3283 (SIRG)

EUROPEAN UNION
European Regional Development Fund

DCU
NUI Galway OÉ Gaillimh
UCC
UCD
Science Foundation Ireland sfi For what's next