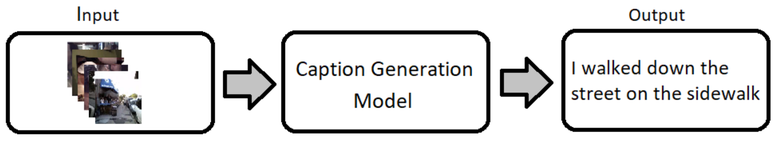
UCF | Center for Research in Computer Vision
UNIVERSITY OF CENTRAL FLORIDA

Video Caption Generation

Ngoc Ta, Aidean Sharghi, Dr. Niels Lobo and Dr. Mubarak Shah

Introduction

Given a video, we would like to automatically translate all possible events in the video into natural language sentences.



Dataset

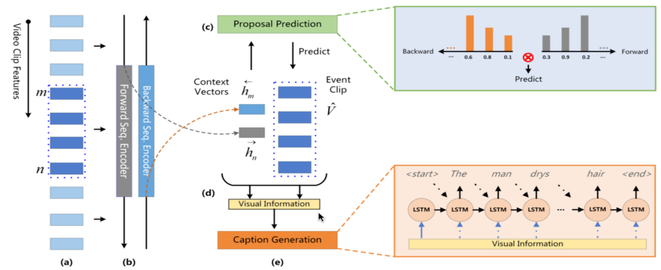
ActivityNet Captions dataset

- PCA-based C3D features
- 20k YouTube untrimmed videos from real life
- 120 seconds long on average
- Most contain over 3 annotated events
- Number of videos in train/validation/test split is 10024/4926/5044 [1]

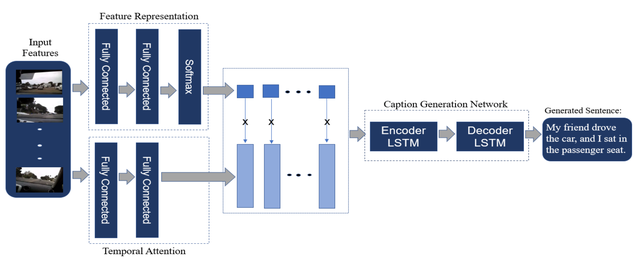
UT Ego dataset

- GoogLeNet architecture
- 4 videos captured from head-mounted cameras.
- Each video is about 3-5 hours long, captured in a natural, uncontrolled setting.

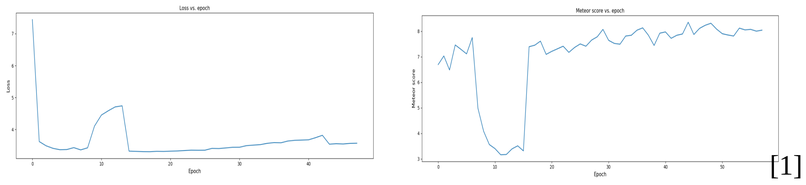
Architecture



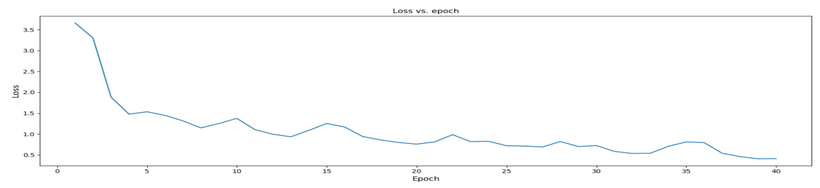
[1]



Graphs



[1]



Results

A man is seen speaking to the camera while holding a large exercise stick and leads into him moving on a mat.

A man is seen standing on a mat and moving himself around while looking to the camera.

A man is seen standing on a mat and leads into him moving on a piece of exercise equipment. [1]

My friend and I ate pizza together.

I sat at the table and took a drink.

My friend and I played with the lego.

References

We would like to thank the National Science Foundation for funding this project.
NSF Grand No. CNS-1757858

References

[1] J. Wang, W. Jiang, L. Ma, W. Liu, and Y. Xu. Bidirectional attentive fusion with context gating for dense video captioning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7190–7198, 2018.