|  |  |  |  |
| --- | --- | --- | --- |
| Title | Att-BiL-SL: Attention-Based Bi-LSTM and Sequential LSTM for Describing Video in the Textual Formation | | |
| Author(s) Name | Shakil Ahmed, A F M Saifuddin Saif, Md Imtiaz Hanif, Md Mostofa Nurannabi Shakil, Md Mostofa Jaman, Md Mazid Ul Haque, Siam Bin Shawkat, Jahid Hasan, Borshan Sarker Sonok, Farzad Rahman, and Hasan Muhommod Sabbir | | |
| Contact Email(s) | mazid@aiub.edu | | |
| Published Journal Name | Applied Sciences | | |
| Type of Publication | Journal | | |
| Volume | 12 | Issue | 1 |
| Publisher | MDPI | | |
| Publication Date | Dec 29, 2021 | | |
| ISSN | 2076-3417 | | |
| DOI | https://doi.org/10.3390/app12010317 | | |
| URL | https://www.mdpi.com/2076-3417/12/1/317 | | |
| Other Related Info. | Article No: 317 | | |
|  | | | |

|  |  |
| --- | --- |
| Abstract |  |
| With the advancement of the technological field, day by day, people from around the world are having easier access to internet abled devices, and as a result, video data is growing rapidly. The increase of portable devices such as various action cameras, mobile cameras, motion cameras, etc., can also be considered for the faster growth of video data. Data from these multiple sources need more maintenance to process for various usages according to the needs. By considering these enormous amounts of video data, it cannot be navigated fully by the end-users. Throughout recent times, many research works have been done to generate descriptions from the images or visual scene recordings to address the mentioned issue. This description generation, also known as video captioning, is more complex than single image captioning. Various advanced neural networks have been used in various studies to perform video captioning. In this paper, we propose an attention-based Bi-LSTM and sequential LSTM (Att-BiL-SL) encoder-decoder model for describing the video in textual format. The model consists of two-layer attention-based bi-LSTM and one-layer sequential LSTM for video captioning. The model also extracts the universal and native temporal features from the video frames for smooth sentence generation from optical frames. This paper includes the word embedding with a soft attention mechanism and a beam search optimization algorithm to generate qualitative results. It is found that the architecture proposed in this paper performs better than various existing state of the art models. | |