Artificial Intelligence Review (2023) 56:13293–13372
https://doi.org/10.1007/s10462-023-10414-6

Video description: A comprehensive survey of deep learning approaches

Ghazala Rafiq[1] · **Muhammad Rafiq**[2] · **Gyu Sang Choi**[1]

Published online: 11 April 2023

Abstract
Video description refers to understanding visual content and transforming that acquired understanding into automatic textual narration. It bridges the key AI fields of computer vision and natural language processing in conjunction with real-time and practical applications. Deep learning-based approaches employed for video description have demonstrated enhanced results compared to conventional approaches. The current literature lacks a thorough interpretation of the recently developed and employed sequence to sequence techniques for video description. This paper fills that gap by focusing mainly on deep learning-enabled approaches to automatic caption generation. Sequence to sequence models follow an Encoder–Decoder architecture employing a specific composition of CNN, RNN, or the variants LSTM or GRU as an encoder and decoder block. This standard-architecture can be fused with an attention mechanism to focus on a specific distinctiveness, achieving high quality results. Reinforcement learning employed within the Encoder–Decoder structure can progressively deliver state-of-the-art captions by following exploration and exploitation strategies. The transformer mechanism is a modern and efficient transductive architecture for robust output. Free from recurrence, and solely based on self-attention, it allows parallelization along with training on a massive amount of data. It can fully utilize the available GPUs for most NLP tasks. Recently, with the emergence of several versions of transformers, long term dependency handling is not an issue anymore for researchers engaged in video processing for summarization and description, or for autonomous-vehicle, surveillance, and instructional purposes. They can get auspicious directions from this research.

Keywords Deep learning · Encoder–Decoder architecture · Text description · Video captioning techniques · Video description approaches · Video captioning · Vision to text

✉ Muhammad Rafiq
rafiq@kmu.ac.kr

✉ Gyu Sang Choi
castchoi@ynu.ac.kr

Extended author information available on the last page of the article