Artificial Intelligence Review (2023) 56:13619–13661
https://doi.org/10.1007/s10462-023-10488-2

A comprehensive survey on image captioning: from handcrafted to deep learning-based techniques, a taxonomy and open research issues

Himanshu Sharma[1] · **Devanand Padha**[1]

Published online: 17 April 2023

Abstract
Image captioning is a pretty modern area of the convergence of computer vision and natural language processing and is widely used in a range of applications such as multi-modal search, robotics, security, remote sensing, medical, and visual aid. The image captioning techniques have witnessed a paradigm shift from classical machine-learning-based approaches to the most contemporary deep learning-based techniques. We present an in-depth investigation of image captioning methodologies in this survey using our proposed taxonomy. Furthermore, the study investigates several eras of image captioning advancements, including template-based, retrieval-based, and encoder-decoder-based models. We also explore captioning in languages other than English. A thorough investigation of benchmark image captioning datasets and assessment measures is also discussed. The effectiveness of real-time image captioning is a severe barrier that prevents its use in sensitive applications such as visual aid, security, and medicine. Another observation from our research is the scarcity of personalized domain datasets that limits its adoption into more advanced issues. Despite influential contributions from several academics, further efforts are required to construct substantially robust and reliable image captioning models.

Keywords Attention-based image captioning · Encoder-decoder architecture · Image captioning · Multimodal embedding

1 Introduction

One of the fundamental abilities of humans is the potential to detect and comprehend a succinct description of the prominent components of an image using natural language. With such powerful abilities, it is incredibly simple to predict, and correlate a description with every image we encounter. However, making machines imitate such human expertise with reliable precision level is the most challenging research problem. A sentence that concisely describes the contents of an image is called the image caption as shown in Fig. 1.

✉ Himanshu Sharma
himanshusharma.csit@gmail.com

[1] Department of Computer Science and Information Technology, Central University of Jammu, Jammu, Jammu & Kashmir 181124, India