Chapter 88
Cross-Modal Learning for Free-Text Video Search

Damianos Galanopoulos
http://orcid.org/0000-0002-1852-4545
CERTH-ITI, Greece

Vasileios Mezaris
http://orcid.org/0000-0002-0121-4364
CERTH-ITI, Greece

ABSTRACT

This article focuses on cross-modal video retrieval, a technology with wide-ranging applications across media networks, security organizations, and even individuals managing large personal video collections. The authors discuss the concept of cross-modal video learning and offer an overview of deep neural network architectures in the literature, focusing on methods combining visual and textual representations for cross-modal video retrieval. They also examine the impact of vision transformers, a learning paradigm significantly improving cross-modal learning performance. Also, they present a novel cross-modal network architecture for free-text video retrieval called T×V+Objects. This method extends an existing state-of-the-art network by incorporating object-based video encoding using transformers. It leverages multiple latent spaces and combines detected objects with textual features, creating a joint embedding space for improved text-video similarity.

INTRODUCTION

The explosive growth of media collections in the media industry and Web and social media platforms requires techniques for compelling video searching in the wild. Free-text video search is a type of video search in which no restrictions are imposed on the user with respect to how their searching needs are expressed; the only limitation is the user's imagination. It allows associating parts of a video with text without any time-consuming manual labelling or annotation with closed sets of pre-defined labels (visual concepts or events); the users can search for videos by simply forming their questions in natural-language text. This contrasts with annotating videos with labels (using e.g. concept or event detectors), as this inevitably involves defining a closed set of labels that restricts the way that a query can be formulated.

DOI: 10.4018/978-1-6684-7366-5.ch088

Moreover, introducing new labels often means having to process the entire media collection again, which can be overwhelming. On the other hand, free-text video search solves these issues because it is based on training a model that can represent both text and videos in a joint feature space; when a new textual query is formulated, a quick inference process can transform this as well in a similar representation, making search and retrieval a fast and efficient process.

Given the above, this chapter focuses on the cross-modal video search task that could significantly advantage the productivity workflow of a variety of organizations, such as media networks and security organizations, as well as for individuals with large personal collections. We start by discussing and defining cross-modal video learning and providing a broad overview of relevant deep neural network architectures found in the literature. Next, our literature survey will focus on cross-modal video retrieval, putting emphasis on methods that rely on combining multiple visual and textual representations. Moreover, we will discuss recent works that rely on vision transformers, a learning paradigm that has generally boosted the performance of cross-modal learning. Based on this survey, we will make some general remarks reflecting on how the cross-modal learning field has evolved over the last years and identify open challenges for further advancing the performance and explanation capacity of cross-modal learning technologies. Next, we will present in detail a state-of-the-art network for cross-modal video search called T×V+*Objects*. This method effectively combines various textual and visual encoders, as well as an object detector, a Vision Transformer (ViT) backbone, and a transformer-based mechanism to extract objects and object-based visual features. These mechanisms automatically extract rich spatial information from the video frames, providing clues for associating video parts with text parts.

BACKGROUND

Cross-modal learning refers to associating information derived from multiple modalities (e.g., video, image, audio, language) using advanced techniques such as deep neural networks. It is currently a hot and active research topic due to the recent breakthroughs in deep learning methods, which have demonstrated remarkable capabilities in handling complex multimodal data. Many different tasks could be included under the umbrella of cross-modal learning, such as image/video captioning (Liming Xu, 2023), retrieval (Cunjuan Zhu, 2023), visual question answering (Siyu Lu, 2023), and audio-visual captioning. These tasks benefit significantly from the flexibility and scalability offered by deep neural networks in capturing intricate patterns across modalities.

Cross-modal video retrieval has gained much attention over the last year due to the advantages of deep learning methods and explosion. However, the automated association of textual and visual content, e.g., images or video, is not new, and several approaches have been proposed to bridge the gap between textual and visual information. These early approaches (J. Lu, 2016), (Markatopoulou, Galanopoulos, Mezaris, & Patras, 2017) (Habibian, Mensink, & Snoek, 2014), (Liu, Albanie, Nagrani, & Zisserman, 2019) are mainly focusing on associating different modalities by trying to detect a set of concepts that can occur in the video using sets of pre-defined concepts as a stepping stone. Due to computational limitations, these works deal with relatively small datasets and the available concept sets were limited to a few and very specific scenarios. Moreover, due to the fact they rely on pre-trained visual and language "experts," they cannot be optimized end-to-end, leading to poor performance.