An Image-Text Matching Method for Multi-Modal Robots

Ke Zheng, Hunan Biological and Electromechanical Polytechnic, China

Zhou Li, Hunan Biological and Electromechanical Polytechnic, China*

ABSTRACT

With the rapid development of artificial intelligence and deep learning, image-text matching has gradually become an important research topic in cross-modal fields. Achieving correct image-text matching requires a strong understanding of the correspondence between visual and textual information. In recent years, deep learning-based image-text matching methods have achieved significant success. However, image-text matching requires a deep understanding of intra-modal information and the exploration of fine-grained alignment between image regions and textual words. How to integrate these two aspects into a single model remains a challenge. Additionally, reducing the internal complexity of the model and effectively constructing and utilizing prior knowledge are also areas worth exploring, therefore addressing the issues of excessive computational complexity in existing fine-grained matching methods and the lack of multi-perspective matching.

KEYWORDS

Image-Text Matching, Multi-View Matching, Transformer

INTRODUCTION

With the continuous advancement of technology, robotics has made significant progress in various fields. Especially with the fusion of multimodal perception and artificial intelligence, robots have evolved from simple tools for task automation into partners with multisensory capabilities and intelligent interactions (B·Hme et al., 2012; Zhang et al., 2022; Paolanti et al., 2017). For example, tour guide robots, as prominent representatives of robotics technology, have garnered widespread interest in the tourism and cultural heritage sectors. In this challenging domain, multimodal robots with multi-view image-text matching capabilities are emerging, providing richer and more precise ways of information exchange for tour guide robots. Robots typically interact with their environment and humans through visual and textual data. Understanding images enables robots to interpret the physical world, while comprehending text helps them communicate with humans and access information on the internet. A deep understanding of both modalities allows robots to have a comprehensive perception of their surroundings, combining visual and textual information to make sense of complex situations. However, images are a form of visual data, while text is linguistic data, and they represent information with inherent differences. To bridge the gap between images and text, image-text matching technology

DOI: 10.4018/JOEUC.334701

*Corresponding Author

for robots requires a deep understanding of both modalities and their seamless integration, which adds complexity to the task of feature extraction (Russell et al., 2002; Yang et al., March 2019). Furthermore, reducing the model's complexity while enhancing its representation capabilities and interpretability is a significant challenge in this context (Paolanti et al., 2019). For the task of image-text matching, traditional methods mainly relied on manually annotating images and then comparing the text words with the manually assigned image labels (Changet al., 1981; Li et al., 2016). These methods involve fixed extraction of features from images and text words followed by matching, making them highly dependent on the quality of manually labeled images. These traditional methods also suffer from several disadvantages: weak feature extraction capabilities, poor noise resistance due to noise in manual annotations, mostly linear structures leading to weak generalization abilities. These drawbacks limited their applicability in real-world scenarios. Subsequently, researchers started to explore more sophisticated learning-based approaches for achieving image-text matching. For instance, Rasiwasia et al. used scale-invariant feature transform algorithms and document topic generation models to represent images and text, and then applied Canonical Correlation Analysis to learn the cross-modal correlations (Rasiwasia et al., 2010). Zhuang et al. leveraged commonality in multimodal data to construct a unified cross-modal association graph, which helped explore the connections between visual and textual data (Zhuang et al., 2008). Yang et al. established a cross-modal index space by mining heterogeneous multimodal data, subsequently generating a semi-semantic graph for cross-modal retrieval (Yang et al., 2010). While these methods have provided valuable insights and made significant progress in image-text matching research, they are often limited to specific small datasets. They may have excellent performance on those datasets but struggle to generalize to broader applications and different domains.

With the rapid advancement of deep learning, cross-modal research has become a popular field (Ma et al., 2022). In the feature learning of multimodal data, deep learning has the capability to nonlinearly map low-level features of multimodal data into high-level abstract representations (Salman et al., 2022). Image-text matching tasks, as a fundamental task in cross-modal research, have garnered extensive attention from scholars. In terms of images, early models utilized Convolutional Neural Networks (CNNs) to extract image features, often pre-trained on image classification tasks. For example, models such as 2WayNet (Eisenschtat et al., 2017), sm-LSTM (Huang et al., 2017), and SAN (Ji et al., 2019) used pre-trained VGG (Simonyan et al., 2014) networks to extract image features, while other models like VSE++ (Fartash et al., 2018), DPC (Zheng et al., 2020), and SCO (Huang et al., 2018) employed deep residual networks (ResNet) (He et al., 2016) pre-trained on the ImageNet (Deng et al., 2009) dataset for image feature extraction. On the text side, early models like m-RNN (Mao et al., 2014) and LRCN (Donahue et al., 2017) used recurrent neural networks (RNNs) or Long Short-Term Memory (LSTM) (Hochreiter, et al. 1997) networks to represent textual information and mapped each text sequence into a feature vector. However, these methods typically represented images and texts as global vectors, resulting in the loss of fine-grained information among image regions and text words, leading to relatively lower accuracy.

Due to the limitations of the aforementioned pre-trained networks, many researchers have attempted to use finer-grained feature representations: dividing images into multiple regions and sentences into words and phrases for representation. Regarding images, SCAN (Lee et al., 2018) was the first to use the object detection pre-trained model Faster-RCNN (Ren et al., 2015) to extract image features. SCAN utilized Faster-RCNN to extract 36 salient regions from an image, encoding each region into a feature vector. This approach allowed for the inclusion of detailed image features within these 36 salient regions. For text, VSE++ introduced the use of Bidirectional Gated Recurrent Units (Bi-GRU (Schuster et al., 1997)) to extract text feature vectors. Bi-GRU consists of both a forward and a backward gated unit, aggregating information from both directions of the sentence words to represent the word's features. Following this, almost all image-text matching models have used Faster-RCNN and Bi-GRU for extracting image and text features. Models such as CAMP (Wang et al., 2019), IMRAM (Chen et al., 2020), VSRN (Li et al., 2019), and others employed these pre-trained models

Related Content