Abstract

Visual Question Answering (VQA) is an exciting field of research that involves answering natural language questions asked about an image. This multimodal task requires models to understand the syntax and semantics of the question, interact with the relevant objects in the image, and infer the answer using both image and text semantics. Due to its complex behavior, VQA has gained considerable attention from both vision and natural language research community.

Most contributions to VQA focus on improving model performance by developing better mechanisms for attaining question and image representations that facilitate interaction between the two. However, despite the progress made, there is still room for improvement in terms of the accuracy of inferred answers. To address this, various methods have been introduced, such as attention mechanisms, that enable effective interaction between the two input modalities.

In this context, this work contributes to the ongoing efforts to improve VQA model performance. Specifically, novel VQA models are proposed that break down the problem into smaller components, making it easier to predict the answer. The focus is given on improving the attention mechanism for the two modalities, resulting in a richer and more accurate feature representation. This work demonstrates that improving VQA model performance can be achieved through multiple avenues, and by combining these approaches, we can achieve even better results. These findings have the potential to enhance the performance of VQA models and contribute to the development of more advanced AI systems that can accurately understand and respond to natural language questions about images.

The first model (ACA-VQA), Aggregated Co-attention based Visual Question Answering, aims to improve VQA performance by exploiting cross modality attention in multiple stages. The attention is aggregated at each stage to preserve the cues obtained from multiple stages. This proposal is benchmarked on the TDIUC and VQA2.0 dataset against state-of-the-art approaches. The experimental results demonstrated the efficacy of multistage co-attention mechanism.

The second model (CSCA-VQA) has an attention block containing both self-attention and co-attention on image and text. The self-attention modules provide contextual information of objects (for an image) and words (for a question) crucial for inferring an answer. On the other hand, cross-modal attention aids the interaction of image and text. To obtain fine-grained information from the two modalities, dense attention blocks are cascaded multiple times. Benchmarking on the widely used VQA2.0 and TDIUC datasets demonstrates the efficacy of key components of the model and the stacking of attention modules.

The third contribution (DAQC-VQA) addresses two important issues in VQA: answer prediction in a large output answer space and obtaining enriched representation through cross-modality interactions. The DAQC-VQA system consists of three main network modules. The first module is a dual attention mechanism that helps in obtaining an enriched cross-domain representation of the two modalities. The second module is a question classifier subsystem that identifies input question category, that helps reduce the answer search space. The third module predicts the answer depending on the question category. All component networks of DAQC-VQA are trained in an end-to-end manner with a joint loss function.

Overall, this work contributes to the ongoing efforts to improve the accuracy of VQA models and enhance their ability to accurately understand and respond to natural language questions about images.