Jointly Generating Image Captions to Aid Visual Question Answering

Raymond Mooney, University of Texas at Austin

Visual question answering (VQA) and image captioning require a shared body of general knowledge connecting language and vision. We present an approach that exploits this connection by jointly generating captions that are targeted to help answer a specific visual question. The model is trained using an existing caption dataset by automatically determining question-relevant captions using an online gradient-based method. Experimental results on the VQA v2 challenge demonstrates that our approach obtains state-of-the-art VQA performance by simultaneously generating question-relevant captions.