View-Invariant Change Captioning

Trevor Darrell, UC Berkeley

The ability to detect that something has changed in an environment is valuable, but often only if it can be accurately conveyed to a human operator. We introduce View-point Invariant Change Captioning, and develop models which can both localize and describe via natural language complex changes in an environment. Moreover, we distinguish between a change in a viewpoint and an actual scene change (e.g. a change of objects' attributes). To study this new problem, we collect a Viewpoint Invariant Change Captioning Dataset (VICC), building it off the CLEVR dataset and engine. We introduce 5 types of scene changes, including changes in attributes, positions, etc. To tackle this problem, we propose an approach that distinguishes a view-point change from an important scene change, localizes the change between "before" and "after" images, and dynamically attends to the relevant visual features when describing the change. We benchmark a number of baselines on our new dataset, and systematically study the different change types. We show the superiority of our proposed approach in terms of change captioning and localization. Finally, we also show that our approach is general and can be applied to real images and language on the recent Spot-the-diff dataset.