RISE RESEARCH INNOVATION SCHOLARSHIP ENTREPRENEURSHIP

Northeastern University

VODKA: Voting Over Dissociated Knowledge Associations

Parijat Chatterjee, Pranav Kompally

Background

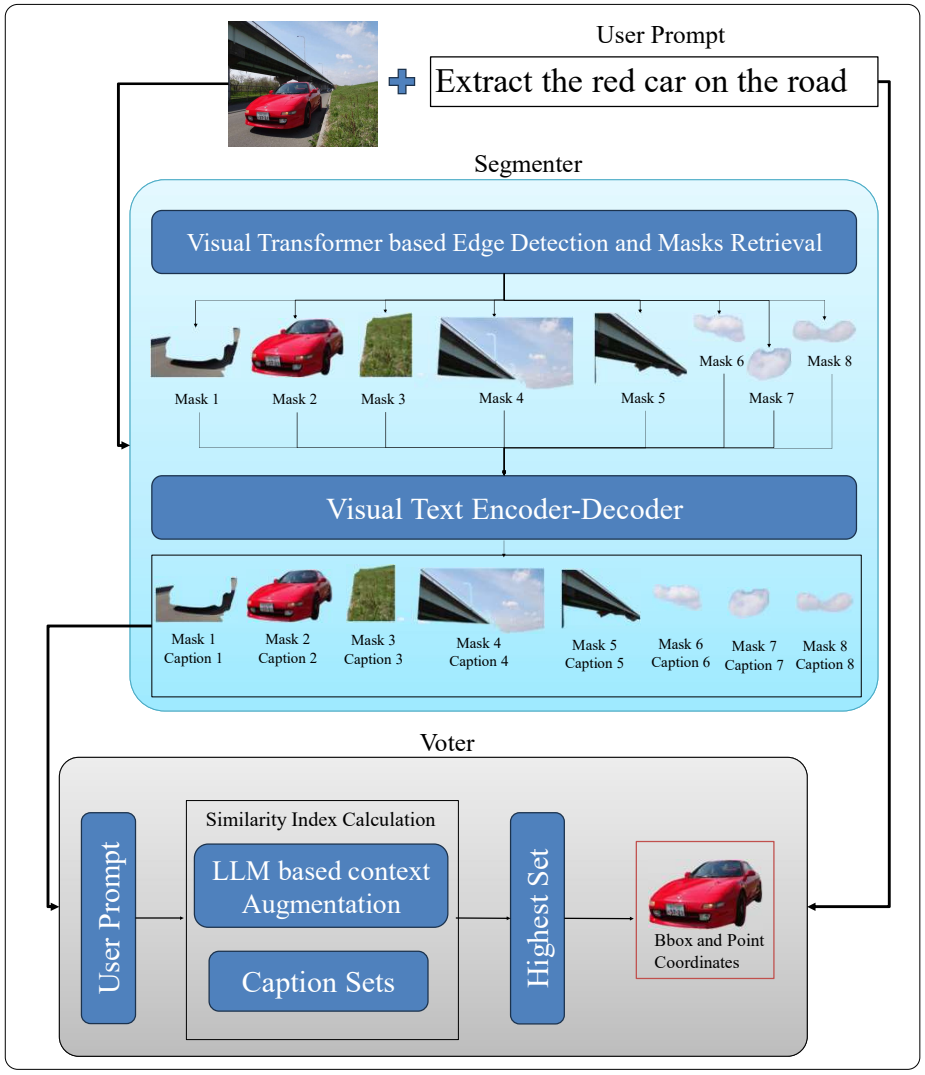
- Visual data Annotation is a daunting and time-consuming task due to the effort and manual labor required to label visual data precisely.
- The time taken per annotation also increases exponentially as a scene gets more complex. Thus, making the task harder for annotators.
- Many companies rely on expensive manual data annotation tools and services to have their data captioned and prepared for training.

Motivation and Aim

- With VoDKA, we aim to reduce the time it takes to annotate multiple objects over millions of images.
- By leveraging the recent advancements in Vision Transformers and Large Language Models. VoDKA, can annotate the objects using a simple text prompt.

Approach and Methodology

- The VoDKA pipeline has two main sections: the “Segmenter” and then the “Voter”.
- The Segmenter splits/masks out all the objects present in an image and then generate text captions for each mask.
- The Voter then makes sense of these masks and caption pairings by calculating a similarity between the user prompt and the caption generated from the mask.
- The masks that receives the highest similarity (i.e. vote), is then returned to the user as the correct annotation.



Segmenter Details

- The image that the user provides is first fed into SAM (Segment Anything Model) to generate the top-k masks.
- The segments/masks are then passed to an image captioning model (BLIP) to generate meaningful text captions sets.
- The Masks and their relevant captions are then used by the Voter.

Voter Details

- Voter consists of a Large Language Model that generates three versions of the user prompt, called "clues".
- The voter then calculates a similarity score between the clues and the captions.
- The mask associated with the caption that records the highest similarity is returned to the user as the valid annotation.

Results and Findings

- Our experiments were based on the benchmarking dataset ‘TextCaps’.
- Out of the images we experimented with, we recorded a very high segmentation accuracy.
- On average it takes less than 5 minutes to annotate an image using our pipeline running purely on a CPU. Which is a lot faster than human annotations.