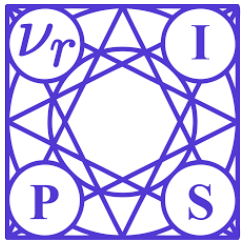
The University of Texas at Austin
Department of Computer Science
College of Natural Sciences

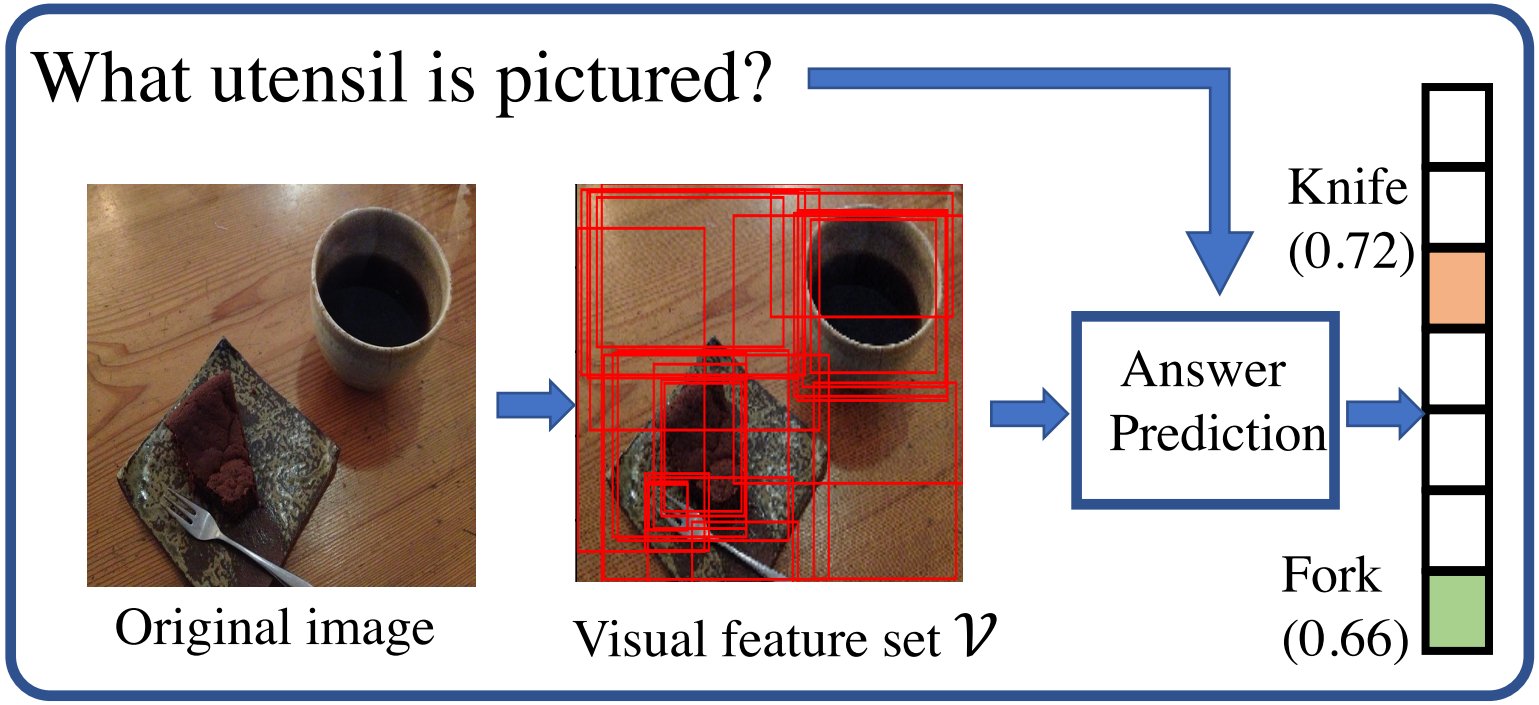


Self-Critical Reasoning for Robust Visual Question Answering

Jialin Wu and Raymond J. Mooney

Visual Question Answering (VQA)

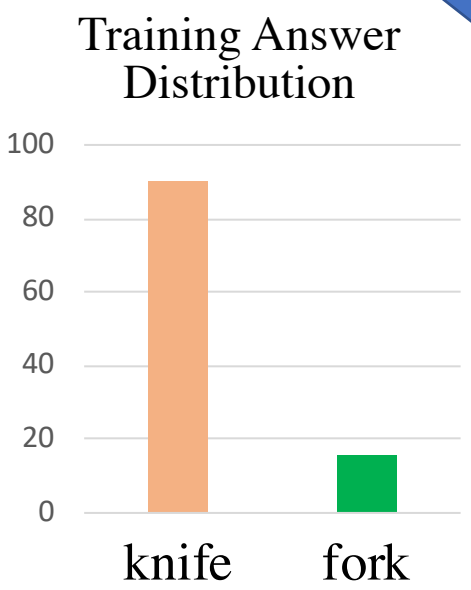
- Common VQA system



Capture superficial statistical correlations between QA pairs

What utensil is pictured?

Original image



I won't bother to look at the image, I can answer your question by just looking at the question

VQA system

Knife

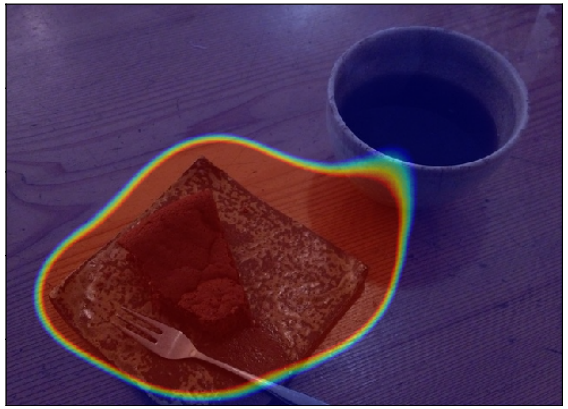
Force VQA to focus on what humans focus on

- Extract a proposal set of objects ($\mathcal{I}$) that humans focus on.

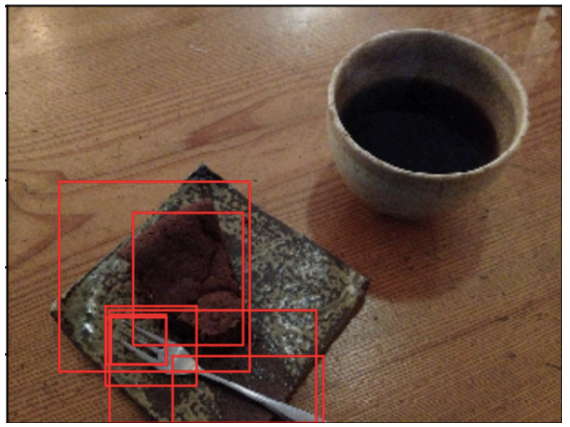
There is a fork near the cake.

Human textual explanation

OR



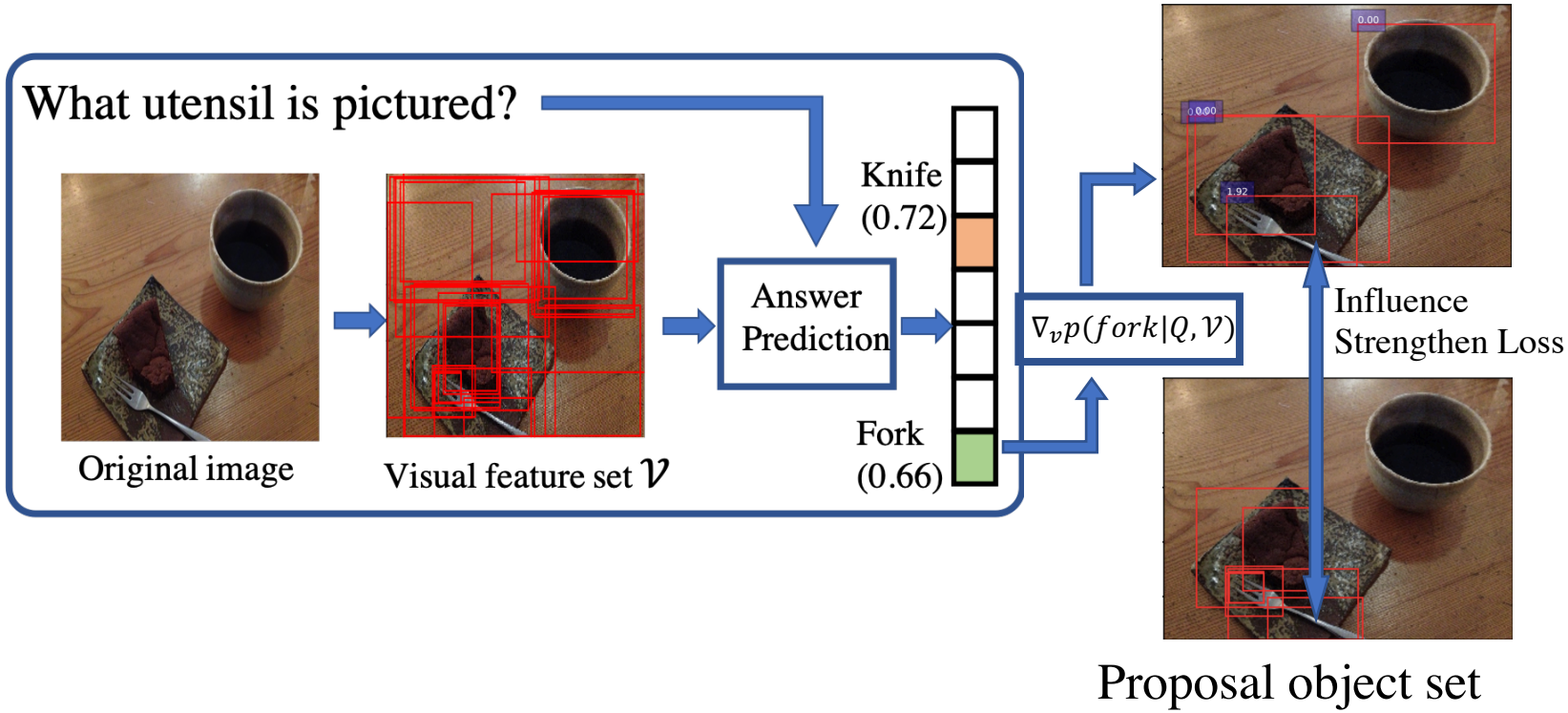
Human visual explanation



Proposal object set $\mathcal{I}$

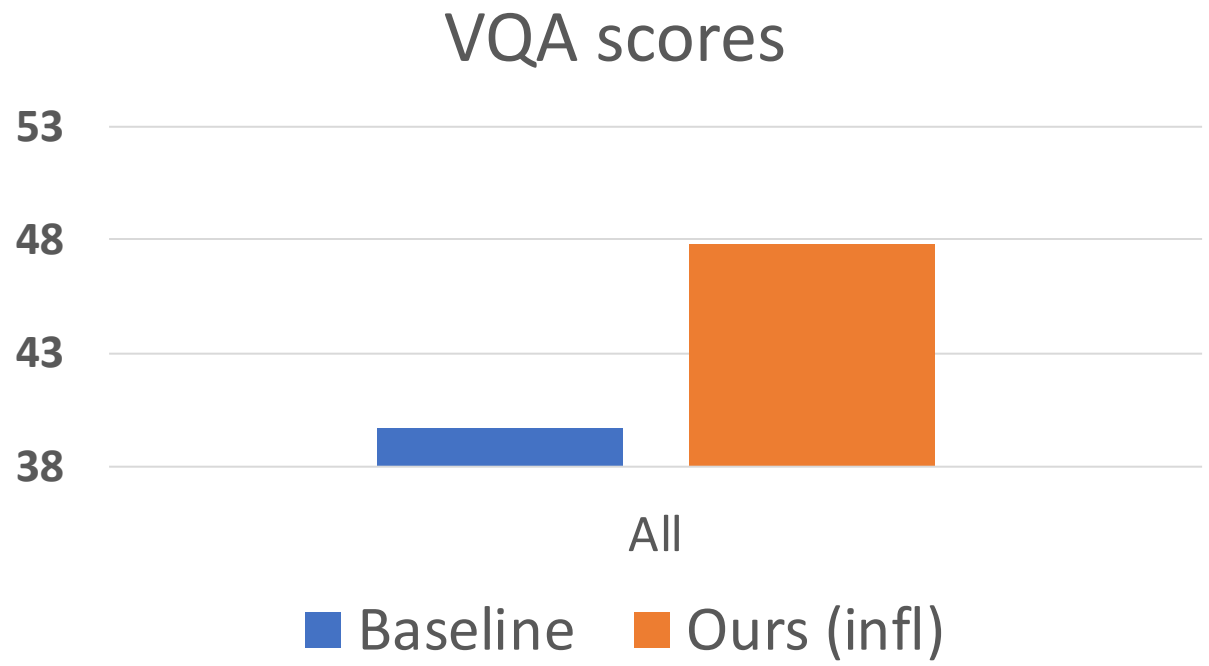
Force VQA to focus on what humans focus on

- Enforce the gradients for the correct answer to have the largest value for at least one of the extracted objects.



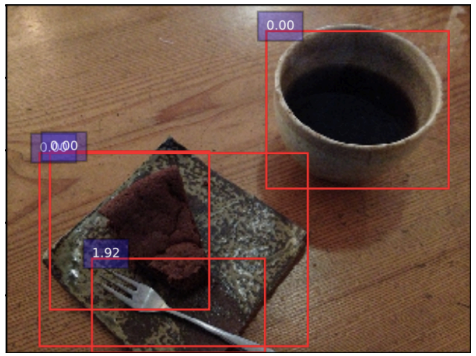
Results

- Compared to baseline model on VQA-CP dataset
- VQA-CP dataset manually set the train and test set in very different distribution

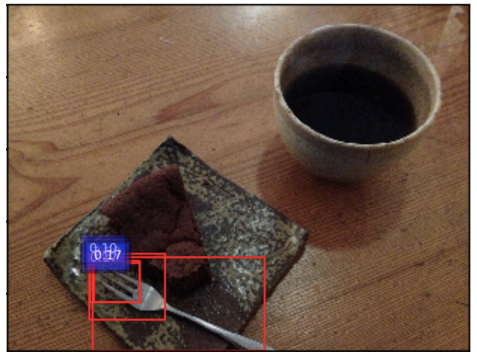


Over sensitivity to the most common objects

What utensil is pictured?



Focused objects
for answer “fork”



Focused objects
for answer “knife”

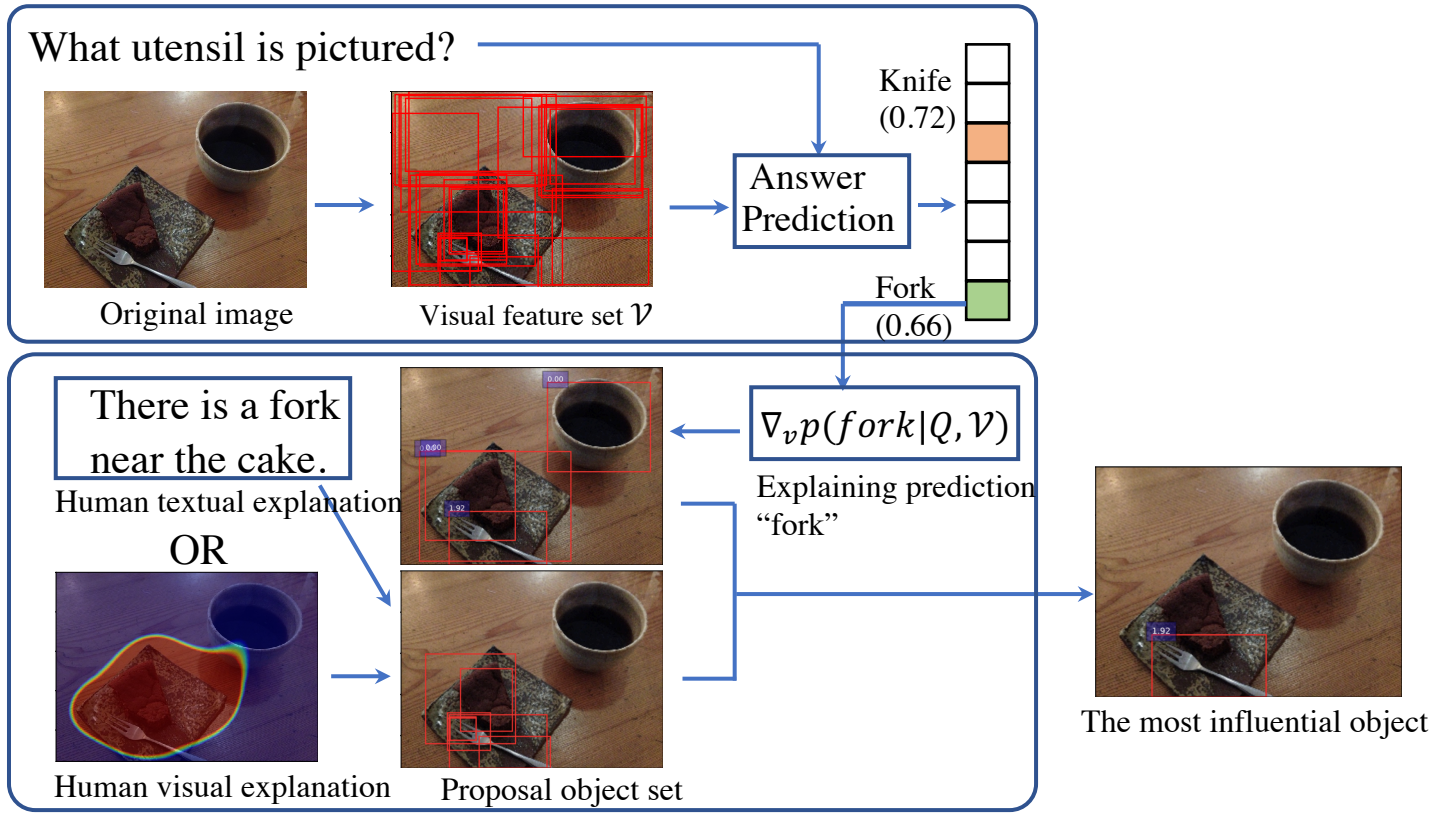
I can focus on the fork but I still think it is a knife

VQA system

Knife

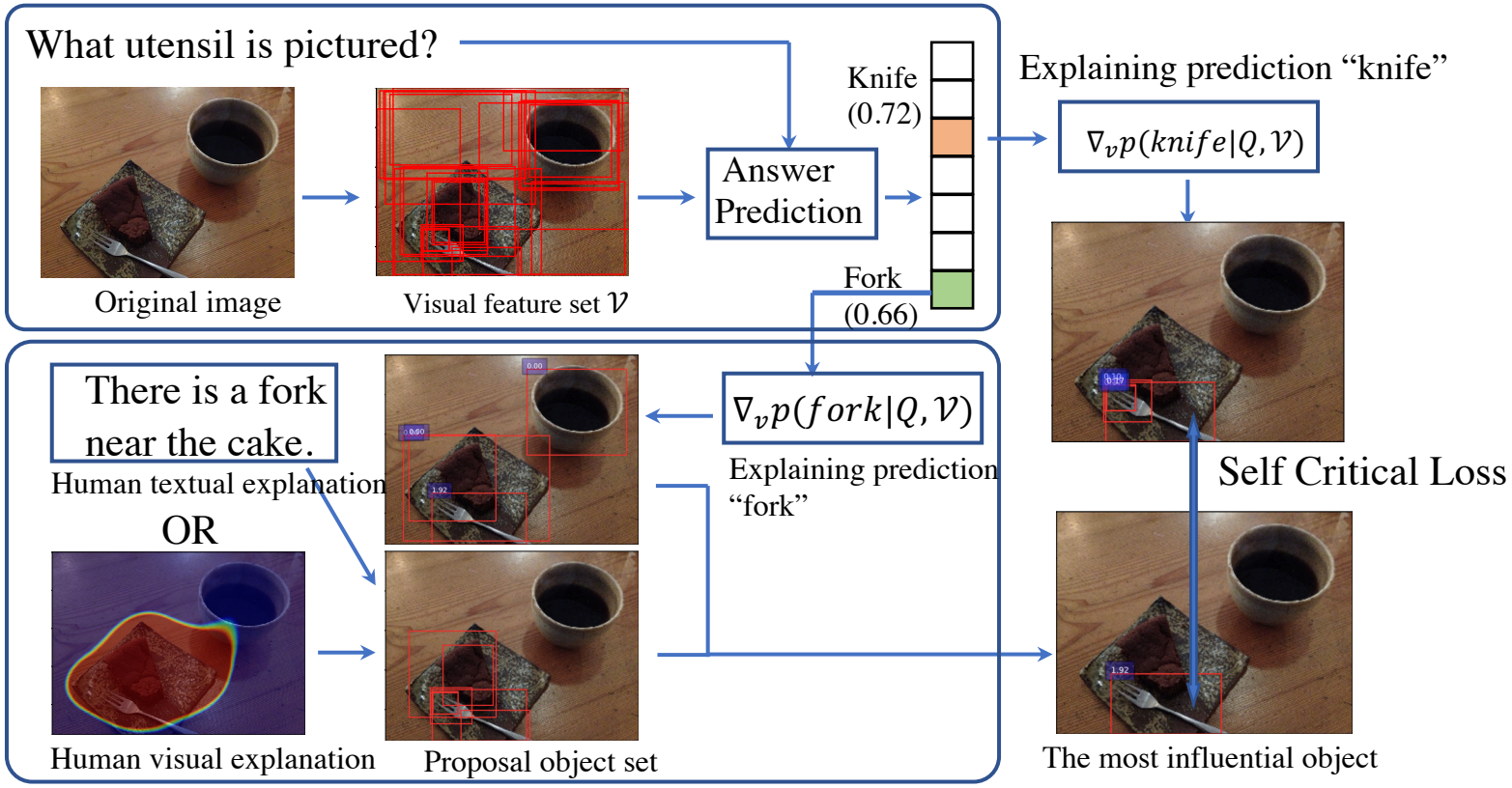
Criticizing the false influential object

- Find the most influential object for the correct answer using gradients



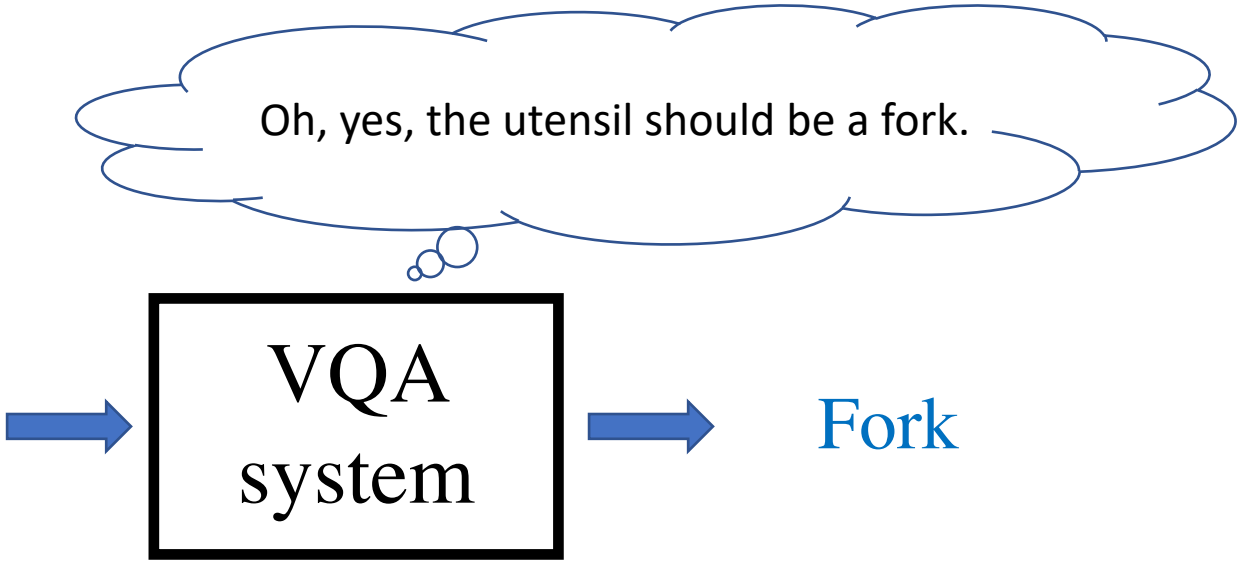
Criticizing the false influential object

- Force the object to contribute more to the correct answer.



Our self-critical approach

What utensil is pictured?



Results

- Compared to baseline model on VQA-CP dataset