TAG: Boosting Text-VQA via Text-aware Visual Question-answer Generation

Jun Wang, Mingfei Gao, Yuqian Hu, Ramprasaath R. Selvaraju, Chetan Ramaiah, Ran Xu, Joseph F. JaJa, Larry S. Davis

Introduction

- Analyze issues of current Text-VQA models due to biases of data annotation and propose possible solutions to mitigate the issues.
- First architecture to explore new question-answer (QA) pairs generation for Text-VQA task without additional annotated data.
- We present consistently accurate results across two vanilla Text-VQA methods on two datasets.

Overall Architecture

The proposed Text-VQA framework. It consists of two parts: a text-aware visual question-answer generation module (TAG), followed by a downstream Text-VQA model. TAG takes a single image and text words (the answer) as input, and outputs a newly-generated question corresponding to the input answer. The generated QA pairs from TAG together with the originally labeled data are subsequently used to train Text-VQA models, leading to better Text-VQA performance.

Motivation - Biases of Data Annotation

More Generated QA Pairs

- Generated QA Pairs

- Image
- Text
- OCR
- Visual object
- Previous decoding

Inference

TAG Overview

- Training
- TAG Module
- Question
- Existing Q-A pairs
- Augmented Q-A pairs

Experiments on ST-VQA

- Each input modality matters
- Impact of answer selection

Ablation Study on TextVQA

- Answer Selection
- Val Acc

Statistics in TextVQA Training Set

- Number of Images
- QA Pair
- Scene Text Words

Generated QA Pairs

- Q: what is the first ingredient listed?
  A: water
- Q: what is the name of this drink?
  A: SAIGON
- Q: what type of store is this?
  A: duty free
- Q: what is the name of the store?
  A: URERTY
- Q: what is the van’s license plate number?
  A: ls09
- Q: what does the van say?
  A: CTYPOLICE
- Q: what kind of tickets can you buy?
  A: ferry
- Q: what letter is in the green?
  A: F
- Q: what is the name on the first bottle from the left?
  A: 100% satisfaction blend
- Q: what is written on the left side of the red label on the left?
  A: Sa4stacrie
- Q: what is the slogan printed on the “thanks” advertisement behind these rugby players?
  A: it’s how we connect
- Q: what does the sign say on the left?
  A: NK
- Q: what is the alcoholic content of the middle bottle?
  A: 9.2%
- Q: what is the brand of the wine?
  A: Trappistes
- Q: what 3 numbers come after “d”?
  A: 80 3
- Q: what is the title of this book?
  A: Oktett

- Experiments on ST-VQA

- Method
- OCR system
- Extra Data
- Val Acc
- Test Acc

- Method
- OCR system
- Extra Data
- Val ANLS
- Test ANLS

- Method
- OCR system
- Extra Data
- Val ACC
- Test ACC