A generalizable VQA model should answer similar multi-modal inputs are involved; answer space differ vastly across datasets.

We thank NSF Award #1718262, Google/Amazon/Adobe Gifts, and Pitt Momentum Funds.

Domain Gap Analysis

Introduction

Modern computer vision methods suffer from overfitting to dataset specifics, which calls for domain adaptation techniques to increase robustness and practically.

A generalizable VQA model should answer similar questions from a different, domain-shifted dataset.

However, domain adaptation is challenging in VQA:
– multi-modal inputs are involved;
– complex optimization over diverse modules;
– answer space differ vastly across datasets.

In this work, we share our explorations about domain robustness over multiple popular datasets and several most recent mainstream VQA approaches.

Inspired by neuro-symbolic models, we propose two-stage training to disentangle representation and reasoning for more effective domain adaptation.

VQA Abstract

We calculate MMD over nine popular VQA datasets using ResNet feature for images and BERT embedding for textual domain shifts.

To estimate the semantic gaps across datasets, we also analyze syntax/appearance domain gaps (our paper Tab 1/2).

We observe several interesting patterns across datasets, e.g., VQA Abstract is unique from other datasets in images, but very similar in questions with VQA v2.

We apply image style transfer and question paraphrasing to VQA datasets, so we can precisely control domain shifts to occur in individual modality.

– Style transfer creates semantically similar but stylistically shifted images.

In this work, we share our explorations about domain robustness over multiple popular datasets and several most recent mainstream VQA approaches.

– Paraphrase generation creates similar questions in different writing styles.

We apply two-stage training to disentangle representation and reasoning for more effective domain adaptation.

Robustness of VQA Models

We choose most recent VQA models and evaluate their domain robustness on CLEVR with synthetic domain shifts. Specifically, we analyze three families: classic two-stream (CL), neuro-symbolic (NS) and transformer variants (TR).

Similar with previous work, we find neuro-symbolic models are more robust to visual shifts.

Presumably due to the extensive pre-training, we find transformer models are more robust to textual domain shifts.

We also directly test the domain robustness of some models on real datasets:
– To mitigate discrepancy in answer space, we keep the top-1000 most frequent answers across all datasets, and evaluate cross-dataset accuracies.

– Since source/target datasets have different upper bounds, we normalize the transferred accuracy and illustrate relative performance with shading intensity.

Conclusions

We find performance degradation generally aligns with our domain gap analysis.

We apply domain adaptation to VQA, and find two-stage training advantageous.

We find disentangled compositional models are promising in domain robustness.

Domain Adaptation for VQA

We arrive at one-stage VQA domain adaptation, and find how two-stage can bring an extra boost.

We also directly test the domain robustness of some models on real datasets.

– To mitigate discrepancy in answer space, we keep the top-1000 most frequent answers across all datasets, and evaluate cross-dataset accuracies.

– Since source/target datasets have different upper bounds, we normalize the transferred accuracy and illustrate relative performance with shading intensity.