Domain-robust VQA with diverse datasets and methods but no target labels

Mingda Zhang, Tristan Maidment, Ahmad Diab, Adriana Kovashka, Rebecca Hwa

University of Pittsburgh

CVPR VIRTUAL JUNE 19-25

Introduction

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


What foods are placed on the table?


How many people will dine over there?

Domain Gap Analysis

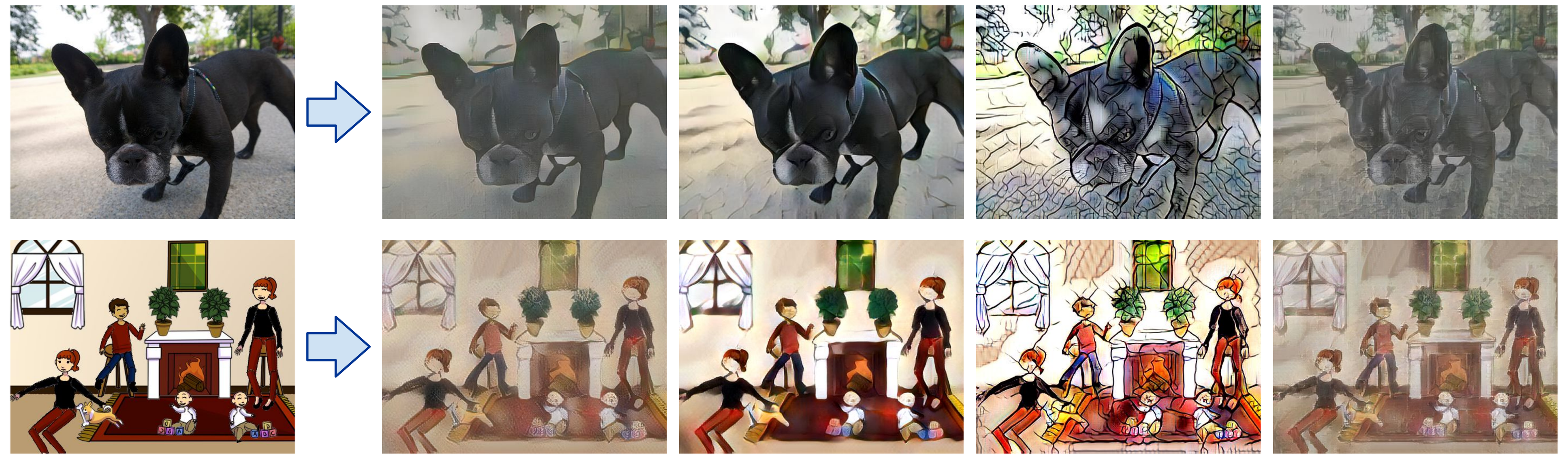
	Visual 7W	Visual Genome	VQA v1	VQA v2	COCO QA	CLEVR	VQA Abstract	GQA	VizWiz
Visual 7W	—	0.01	0.06	0.06	0.20	0.22	0.06	0.10	0.06
Visual Genome	0.01	—	0.07	0.07	0.20	0.22	0.06	0.11	0.06
VQA v1	0.02	0.02	—	0.00	0.15	0.17	0.02	0.06	0.10
VQA v2	0.03	0.02	0.01	—	0.15	0.17	0.02	0.06	0.10
COCO QA	0.04	0.04	0.03	0.03	—	0.19	0.15	0.15	0.23
CLEVR	0.54	0.54	0.54	0.54	0.54	—	0.19	0.13	0.22
VQA Abstract	0.36	0.36	0.36	0.36	0.36	0.59	—	0.07	0.10
GQA	0.03	0.03	0.03	0.03	0.04	0.54	0.36	—	0.12
VizWiz	0.22	0.22	0.21	0.21	0.21	0.52	0.42	0.22	—

- We calculate MMD over nine popular VQA datasets using BERT embedding for questions and ResNet feature for images, 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 thank NSF Award #1718262, Google/Amazon/Adobe Gifts, and Pitt Momentum Funds.

Synthetic Domain Shifts

- 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.



  - Paraphrase generation creates *similar questions* in *different writing styles*.

Original Question	Paraphrased Question
What is written on the white square on the bus?	What does the white square say on the bus?
What shape is the bench seat?	What is the shape of the bench?
What number of red spheres are behind the shiny object that is on the left of the tiny matte cylinder?	How many red spheres are behind the shiny object on the left side of the small dull cylinder?

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.

Methods	Original	Image Shift	Question Shift	Both Shift
NSCL (NS)	98.0	**71.0**	—	—
MAC (NS/CL)	93.4	45.9	52.2	28.1
TbD (NS/CL)	**99.1**	55.7	**52.9**	**36.1**
RelNet (CL)	93.7	20.5	49.6	19.1
LXMERT (TR)	94.8	50.6	**53.4**	**36.6**

- 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.

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

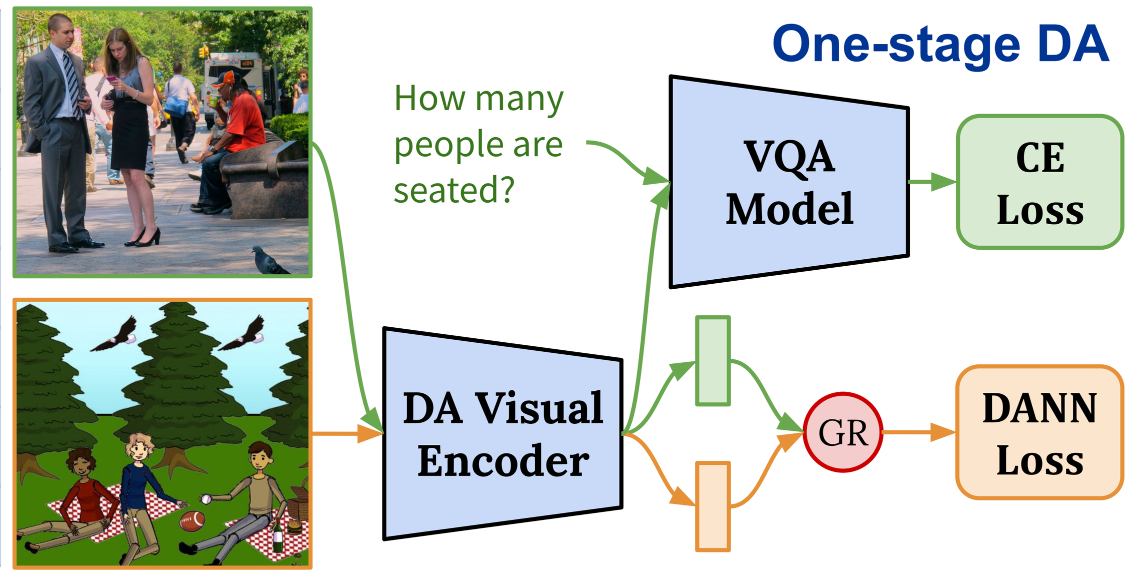
Datasets A	Datasets B	Image App.	Image Sem.	Question Syn.	Question Sem.
VQA v2	CLEVR	High (0.10)	High (0.54)	High (0.81)	High (0.17)
VQA v2	GQA	Low (0.01)	Low (0.03)	Med H (0.35)	Medium (0.06)
VQA v2	VQA Abstract	Med H (0.08)	Med H (0.36)	Low (0.03)	Low (0.02)
VQA v2	Visual Genome	Low (0.01)	Low (0.02)	Med L (0.16)	Medium (0.07)

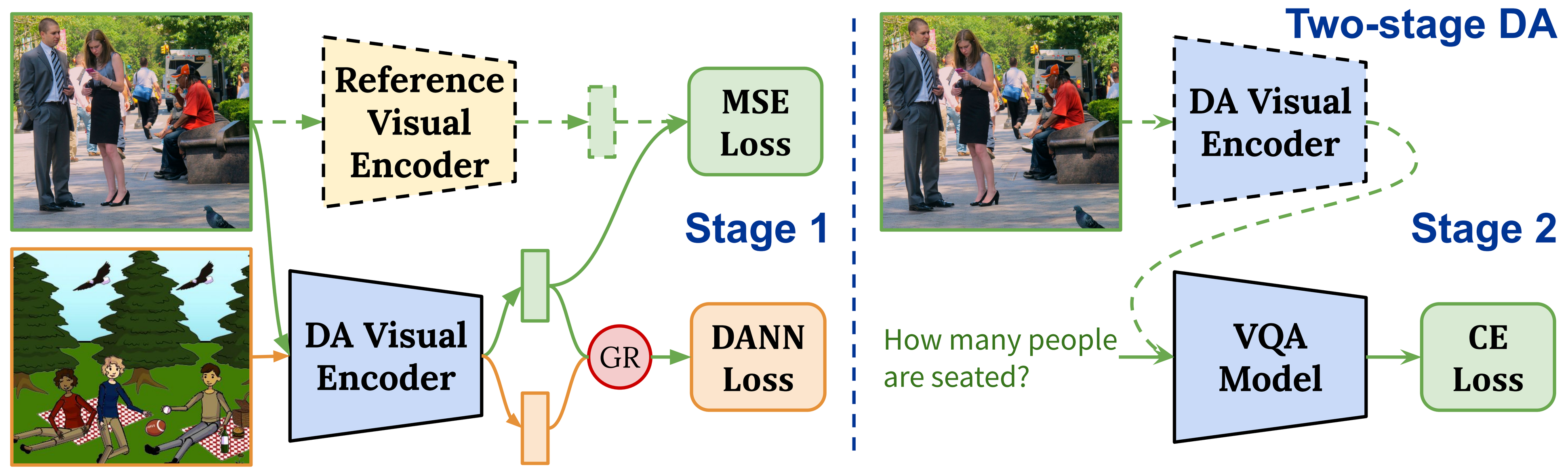
Model	Accuracy (%) A → A	B → B	A → B	B → A
MAC (NS/CL)	53.3	95.9	29.8	18.7
		44.4	32.0	35.6
		48.3	33.6	31.7
		33.3	26.2	23.1
LXMERT (TR)	67.6	84.9	31.6	34.8
		58.2	50.5	51.5
		56.3	34.3	34.6
		41.0	36.7	31.4

Domain Adaptation for VQA

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

	VQA v2	CLEVR	GQA
Source Acc.	54.0	95.8	44.6
Target (Direct)	41.0	45.9	37.3
1-stage DANN	42.2	45.7	37.4
1-stage MM	42.6	46.6	**38.6**
2-stage DANN	**42.8**	**46.7**	38.5
Target (Full)	49.1	90.0	42.1





Conclusions

- Domain gaps in VQA datasets can come from either visual or linguistic space, and each can affect the cross dataset generalizability of the model. Different model families also have varied sensitivity towards domain shifts.
- We find disentangled compositional models are promising in domain robustness.