Reassessing Evaluation Practices in Visual Question Answering: A Case Study on Out-of-Distribution Generalization

Aishwarya Agrawal*‡
Ivana Kajić*
Emanuele Bugliarello*
Elnaz Davoodi△
Anita Gergely△
Phil Blunsom
Aida Nematzadeh*‡

EACL 2023

* equal first-author contribution  ‡ equal senior contribution
△ equal contribution
Visual Question Answering (VQA)

Question: What is this truck?
Answer: Fire truck

Example from: https://visualqa.org/vqa_v2_teaser.html
Progress on VQA\textsuperscript{v2} (Goyal et al., 2017)

Best performing models:*  
1. PaLI: 84.3\%  
   (Chen et al., 2022)  
2. BEiT-3: 84.0\%  
   (Wang et al., 2022)  
3. Flamingo: 82.0\%  
   (Alayrac et al., 2022)

* As of March 2023

- Is the VQA challenge solved?  
  - No, we need to better evaluate our models  
  - Are models learning to solve the task of VQA or the dataset?
### Experimental Setup

#### Datasets

| Dataset     | Reference        |
|-------------|------------------|
| **VQA v2**  | (Goyal et al., 2017) |
| **VG**      | (Krishna et al., 2017) |
| **GQA**     | (Hudson and Manning, 2019) |
| **VizWiz**  | (Gurari et al., 2018) |

#### Example Questions and Answers

| Q | A1 | A2 | A3 |
|---|----|----|----|
| Q: What is the color of the hydrant? | orange | yellow | orange |
|   | A1: orange | A2: yellow | A3: orange |
| Q: What are these zebras doing? | Eating |
|   | A: Eating |
| Q: What is the large container made of? | cardboard |
|   | A: cardboard |
| Q: Please fully describe what you see in this image, thank you. | bird cage bottles paper towels |
|   | A1: bird cage bottles paper towels | A2: birdcage cleaning supplies | A3: unanswerable |
|   | [...] | [...] | [...] |
Experimental Setup [cont’d]

Models

- Two representative, widely-used pretrained models achieving strong performance in V&L tasks:
  - ALBEF (Li et al., 2021)
  - ViLBERT (Lu et al., 2019)

- Total: 128 experiments
Research Questions

1. How well do current VQA models generalize under out-of-distribution (OOD) settings?

2. Are generative models more robust to OOD generalization than discriminative ones?

3. Does multimodal pretraining help with OOD generalization?

4. Are current automatic VQA evaluation metrics suitable for OOD evaluation?
IID vs OOD (out-of-distribution) performance
ALBEF generative
IID vs OOD (out-of-distribution) performance

ALBEF generative

![Graph showing VQA Accuracy for different datasets and conditions.]

- **GQA**
- **VQAv2**
- **VG**
- **VizWiz**

**Fine-tuning dataset**
- GQA
- VQAv2
- VG
- VizWiz

**Evaluation**
- IID
- OOD

Accuracy values range from 0% to 80%.
How well do current VQA models generalize under OOD settings? Poorly
Generative vs Discriminative Evaluation

- Discriminative models are bounded by the top-\(k\) answer sets
- This limitation does not apply to generative evaluation

Are generative models more robust to OOD generalization than discriminative ones? Yes, in most cases
The Case for Multimodal Pretraining
ALBEF generative

Is multimodal pretraining helpful?
Yes, in most cases

- More effective in the generative setting
- Least helpful in OOD VizWiz
OOD Evaluation of VQA Systems

- **Generative** models are more robust
- Multimodal **pretraining** is often helpful
- Yet current models perform **poorly**...

...or do they?

**VG Question:** When was this photo taken?
OOD Evaluation of VQA Systems: Human Evaluation
ViLBERT generative

Are current automatic VQA evaluation metrics suitable? Not really

- Human evaluation is more helpful in the **generative** setting
- Human evaluation is more helpful in the **OOD** settings

![Graph showing ΔVQA Accuracy (human - auto) for different datasets](image)
OOD Evaluation of VQA Systems: Human Evaluation [cont’d]

ViLBERT generative

Does human evaluation close the OOD gap? **Not really**

Even after human evaluation, models still exhibit poor OOD generalization.
Takeaways
Summary

**OOD Generalization** as a more rigorous and representative *evaluation protocol*

- How well do current VQA models generalize under OOD settings? *Poorly*
- Are generative models more robust to OOD generalization than discriminative ones? *Yes, in most cases*
- Does multimodal pretraining help with OOD generalization? *Yes, in most cases*
- Are current automatic VQA evaluation metrics suitable for OOD evaluation? *Not really*
Next Steps

- **Evaluation Metric**: need more robust automatic metric or scalable human evaluation

- **Modelling**: improve reasoning and overfitting to spurious correlations

  - Poor reasoning skills (logical, spatial, compositional)
    - E.g., “Is the cheese to the right or to the left of the empty plate?”

  - Overfitting to answer priors
    - E.g., “What is the skateboarder wearing to protect his head?” → “helmet”

  - Overfitting to question format
    - E.g., “What animal ... ?”, “What kind of animal ... ?” (GQA)
    - 45% accuracy drop
      - “Who is ... ?”, “What is ... ?” (VG)