Advancing Web-Based Visual Question Answering with Efficient Image-Text Alignment

UNIVERSITY OF CALIFORNIA SANTA CRUZ

Shubham Gaur

Baskin Engineering

Motivation

Challenges associated with VQA (Visual Question Answering) benchmark in open-domain environment

- VLP scale - Parameter Inefficiency
- Complex Queries
- Dependence on textual information

Research Goals:

- Image-Text Matching Loss
- Efficient Encoders
- Text-Image alignment
- Knowledge Distillation

Example: Given an image of a dog sitting next to a child with a ball, if the question is, "What is the child holding?" the model incorrectly respond with "the dog,"

Example: If a question asks, "How many people are wearing hats in the image?" and there are multiple people, but only some are wearing hats, the model respond with an incorrect count, such as "All of them are wearing hats," ..

Methodology

Dataset:

- Utilise the **WebQA Dataset**
- 36766 training questions
- 4966 validation questions
- 7540 test questions

Baseline Models:

- Visual Language Pretraining (VLP)
- RoBERTa

Key Findings:

- RoBERTa baseline outperformed VLP baseline with F1 of 75.73 v/s 67.96
- VLP performed better on text-based questions (F1 score of 70.07) compared to image-based questions (F1 score of 67.88)
- VLP achieved full accuracy on 41.38% of questions
- RoBERTa performed superior in multi-hop reasoning tasks compared to VLP model
- Both models struggled with image-based questions (often selected too few or too many sources), VLP model failed to select correct number of sources for 18.52% questions while 34.63% had too many predicted sources

Conclusions

New Research Ideas:

- Image-Text Match Loss
- Detection-Free Visual Encoders
- Multimodal Knowledge Distillation
- Exploration of Multimodal Cross-Attention Modules
- More exploration with BERT architectures

Formulas

Cross-Modal Attention

$$Y_\alpha = CM_{\beta\to\alpha}(X_\alpha, X_\beta) = \text{softmax}\left(\frac{Q_\alpha K_\beta}{\sqrt{d_k}}\right)V$$

Loss for Source Retrieval

$$\text{Loss}_{\text{retrieval}} = \sum_{s_i \in G} \log p(s_i) + \sum_{s_i \in D} \log(1 - p(s_i))$$

Results

Table 4 Percentage of fully correct question predictions in different question categories

Category	VLP (%)	RoBERTa (%)
Choose	40.50	42.23
Shape	50.00	45.95
YesNo	33.67	35.99
Others	43.12	46.19
Color	44.12	49.16
Number	59.27	62.16

Table 2 Overall classification metrics of the baselines

VLP baseline	F1	Precision	Recall
Val text-only	70.07	–	–
Val image-only	67.88	65.20	78.48
Test	67.96	–	–
RoBERTa baseline	F1	Precision	Recall
Val text-only	76.22	67.86	86.95
Val image-only	65.75	52.50	87.92
Test	75.73	–	–

Table 5 Percentage of fully correct question predictions in different question categories

Sources	VLP (%)	RoBERTa (%)
Single	51.49	54.91
Multi	28.73	30.11

Evaluation Criteria

$$\text{FL}(c, R) = \max\left\{\min\left(1, \frac{\text{BARTScore}(r, c)}{\text{BARTScore}(r, r)}\right)\right\}_{r \in R}$$

$$\text{acc} = \begin{cases} \text{F1} & \text{if QType} \in [\text{color, shape, number, } Y/N] \\ \text{Recall} & \text{otherwise} \end{cases}$$