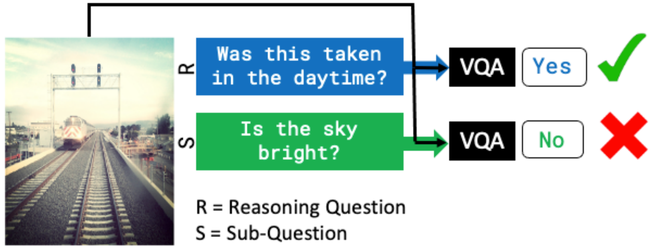
SOrT-ing VQA Models : Contrastive Gradient Learning for Improved Consistency

Sameer Dharur[1], Purva Tendulkar[1→3], Dhruv Batra[1,2], Devi Parikh[1,2], Ramprasaath R. Selvaraju[1→4]
[1]Georgia Institute of Technology, [2]Facebook AI Research, [3]University of California San Diego, [4]Salesforce Research

1 PROBLEM

- Current VQA models often struggle with consistency – they answer seemingly complex questions requiring higher-level reasoning correctly, but fail on associated lower-level perception questions.
- This indicates that the model likely answered the reasoning question correctly for the wrong reason(s).



An example of a commonly used VQA model, Pythia, exhibiting inconsistency – by answering a higher order reasoning question correctly, but failing on an associated perception sub-question.

2 CONTRIBUTIONS

We ask – can VQA models be made more **consistent** by learning to distinguish between **relevant and irrelevant perceptual concepts** for a reasoning question?

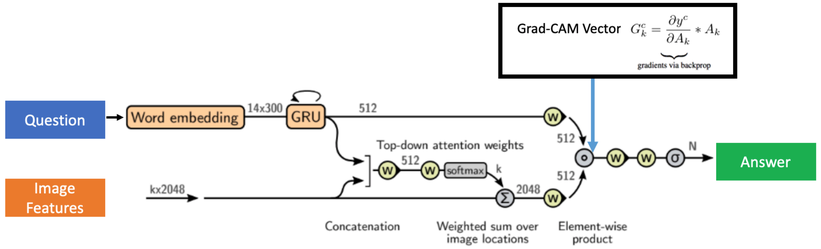
- We develop language-based interpretability metrics that **measure the relevance of a lower-level perception question** while answering a higher-level reasoning question.
- We find that state-of-the-art VQA models **often rank irrelevant questions higher** than relevant ones.
- To fix this, we introduce **Sub-question Oriented Tuning (SOrT)** to train VQA models to **rank sub-questions higher** than irrelevant questions for a reasoning question.
- This **improves model consistency and visual grounding** over baselines Pythia and SQuINT.

Paper

Code

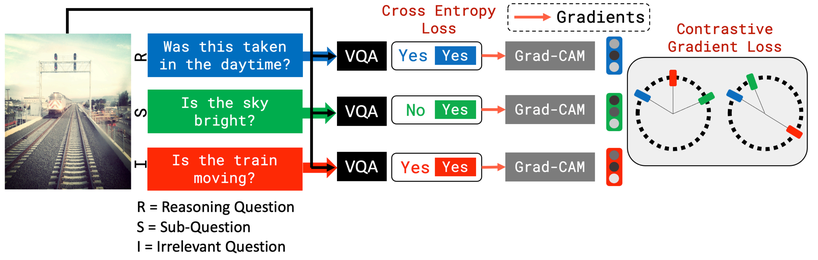
3 APPROACH

- We use **Grad-CAM vectors** to represent each question.
- This is a faithful function of the image, question, answer and the model's weights.
- Semantically, this represents the **most salient visual concepts** used to answer a question.



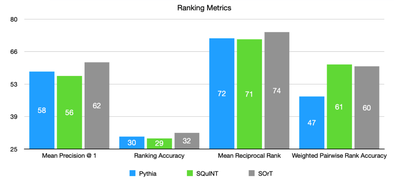
The architecture of our model is based on Pythia. The Grad-CAM vectors for each question are computed at the layer where the vision and language modalities are combined.

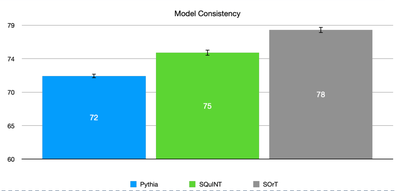
- We combine the VQA-Introspect and VQAv2 datasets to generate sets of **sub-questions** and **irrelevant questions** for a reasoning question.
- We **contrast sub-questions with irrelevant questions** for a reasoning question by using a **Contrastive Gradient Loss** w.r.t their Grad-CAM vectors.
- In addition, we use a **Cross Entropy Loss** for the questions to retain accuracy.



The reasoning question *Was this taken in the daytime?* has the sub-question *Is the sky bright?* and an irrelevant question *Is the train moving?* We tune the model with a Cross-Entropy Loss and a Contrastive Gradient Loss to align the reasoning question's Grad-CAM vector with its sub-question(s) and distance it from its irrelevant question(s).

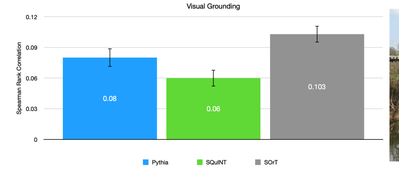
4 RESULTS AND ANALYSIS

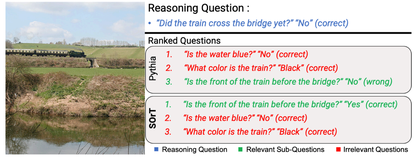




We find that our approach improves the ranking of relevant sub-questions across a range of metrics.

Our approach improves model consistency by 6.5% (absolute) over Pythia and 3.2% (absolute) over SQuINT.





Our approach also demonstrates statistically significant gains on visual grounding.

An example of improvement in consistency between Pythia and SOrT via better sub-question ranking.

A qualitative example of the improvement in visual grounding. SOrT's is the only heatmap that points to the essential visual components needed to answer the question.

5 CONCLUSION

- We seek to improve consistency in VQA models.
- We present Sub-question Oriented Tuning (SOrT), a contrastive gradient learning based approach for teaching VQA models to distinguish between relevant and irrelevant perceptual concepts while answering a reasoning question.
- Our approach improves ranking of sub-questions, model consistency and visual grounding.