# VEQA: Visual Question Answering from the lens of Visual Entailment

CS 5824: Advanced Machine Learning (Fall 2022)

Meghana Holla
Surendrabikram Thapa

# Visual Question Answering

- Given an image and a natural language question, can our machine arrive at an answer?
- Vision+Language Task
- Demands Recognition+Reasoning


Image Source: Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In Conference on Computer Vision and Pattern Recognition (CVPR), 2016

# Entailment - Roots and Applications

***Roots in Logic:***

Premise    Hypothesis

Given a pair of statements, the first statement entails the second if, when the first statement is true, there is enough evidence to conclude that the second one is true too.

***Popular Adaptation:***

NLP - Natural Language Inference

**Premise:** A woman is talking on the phone while standing next to a dog

**Hypotheses**

A woman is on the phone

A woman is walking her dog

A woman is sleeping

Source: Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounded112
language learning. arXiv preprint arXiv:1811.10582, 2018.

# Visual Entailment

Premise → Image
Hypothesis → Textual description

Does the image provide enough evidence to conclude what the text coveys?

*Two woman are holding packages.*
*The sisters are hugging goodbye while holding to go packages after just eating lunch.*
*The men are fighting outside a deli.*

# Motivation

- Learn the datasets not the task: VQA approaches are prone to learning biases in the datasets
- Language-bias: VQA may have the tendency to answer using the language data alone - co-occurrence patterns
- Visual Entailment, similar to textual entailment can enforce semantic inference capabilities.

How can we best leverage Visual Entailment for Visual QA?

In other words,

Can we reformulate VQA into an entailment task?

# VEQA

- ***VEQA***: **V**isual **E**ntailment for Visual **Q**uestion **A**nswering - VQA Framework for multiple choice QA
- Reformulates VQA to a VE task

***How to obtain premise and hypothesis?***
Image - Premise
Question + Answer Choice - Hypothesis

***How to choose the right answer?***
Hypothesis with highest *entailment score* predicted as correct answer

# VEQA - Proposed Approach

Two Components:

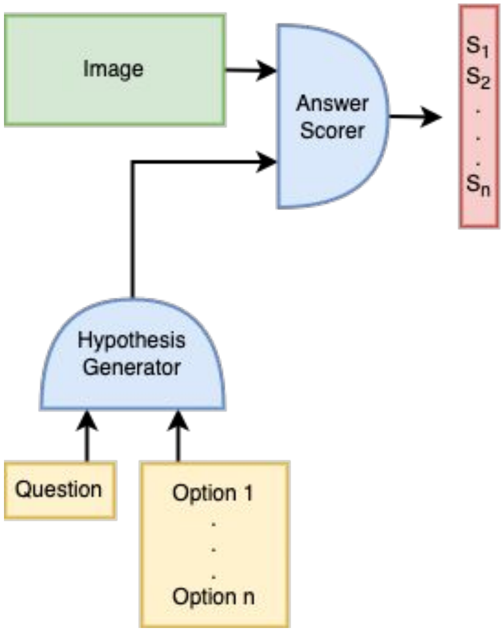
1. Hypothesis Generator
    - Responsible for generating a natural language hypothesis, given question + answer choice

      ***Question + Choice$_x$ → Hypothesis$_x$***

2. Answer Scorer
    - Predict the score of a given answer choice conditioned on the image and question
    - Visual Entailment model under the hood

      ***Image + Hypothesis → Answer Score***

    - Answer score is the entailment probability of hypothesis H



# Hypothesis Generator

Hypothesis Generator: Rule-based hypothesis generator
Generation based on ***question types***
Eg:
Question type: **“What ____ is the”**
Hypothesis: **The ____ is +** <Answer Choice i>

Question Type: **“What color is the”**
Answer Structure: **The color is + <Answer Choice>**

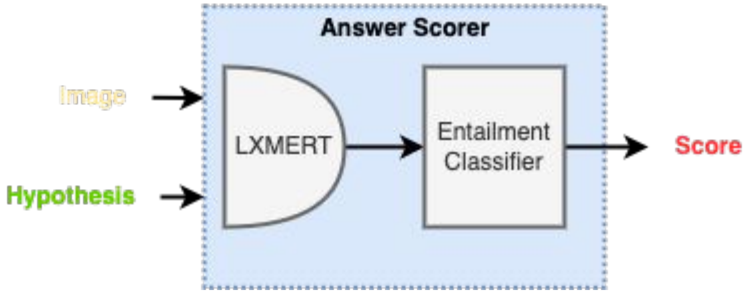
A. The color is red.
B. The color is blue.
C. The color is yellow.
D. The color is green.

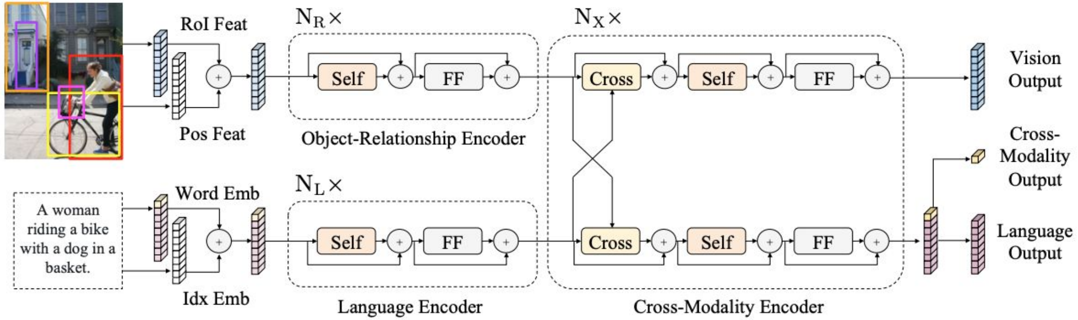
What is the color of the bird?

A. Red
B. Blue
C. Yellow
D. Green

# Answer Scorer

- Leverage Vision-Language models for answer scoring
- Finetune pre-trained LXMERT [1] with a classifier head
- LXMERT tokenizer + LXMERT
- Image Features + Bounding box data - Faster RCNN




Source: Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from116 transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language117 Processing, 2019.

# Current Experimental Configurations

**Hypothesis Generator**

Influence of sentence structure on ***VEQA***

**Concatenation of Question-Answer choice (CQ):**
Hypothesis = Question + Answer

**Natural Language Hypothesis (HG):**
Hypothesis = $H(Q, A_i)$

**Answer Scorer**

Influence of source training data on ***VEQA***

**Training with SNLI-VE ($AS_{SNLI\text{-}VE}$):**
- Equal class distribution
- Appropriate representation of entailment/contradiction

**Training with VQA v1 ($AS_{VQA}$):**
- Same training and testing distribution

# Results

| *Configuration* | *Accuracy @ Top 1* | *Accuracy @ Top 2* | *Recall* | *Precision* |
|---|---|---|---|---|
| **HG + $AS_{SNLI-VE}$** | 37.971% | 55.128% | 36.381% | 36.597% |
| **QA + $AS_{SNLI-VE}$** | 39.616% | **55.720%** | 38.215% | 38.167% |
| **HG + $AS_{VQA}$** | **50.927%** | 51.612% | **51.065%** | **50.914%** |

# Analysis and Future Directions

**Analysis**

1. HG + $AS_{VQA}$ highest top-1 accuracy - Least difference between top-1 and top-2 accuracies.
2. Between QA and HG hypothesis generations: QA performs better  - Inherent structure may not be as important as we think - information is important.

**Future Directions**

1. Zero-Shot/Self supervised VQA
2. VE for Open-ended QA?
3. Visual Question entailment

# References

[1] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from116 transformers. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language117 Processing, 2019.
[2] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and Yang: Balancing and answering binary visual questions. In Conference on Computer Vision and Pattern Recognition (CVPR), 2016
[3] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment task for visually-grounde language learning. arXiv preprint arXiv:1811.10582, 2018