Assignment 3: Shikra

The objective of this assignment is to gain a comprehensive understanding of Shikra. One of the tasks is to replicate the primary experimental outcomes in Shikra to become familiar with the model and its setup. The subsequent task involves expanding the current evaluations to encompass more datasets and tasks.

For all tasks, use **Shikra-7B** model.

Source Code: https://github.com/shikras/shikra
Model: https://huggingface.co/shikras/shikra-7b-delta-v1-0708

Note: Follow the instructions in the Shikra Readme [https://github.com/shikras/shikra?tab=readme-ov-file#checkpoint] to recover the full weights. This requires applying the delta to LLaMA weights in order to get the full weights. Please follow the steps exactly to get the correct weights. You may have to setup huggingface account and request base LLaMA weights to get access to it.

Tasks:

1. Reproduce [20 points]

a. REC task Generalist VL SOTAs [10 points]

Table 3 RefCOCO, RefCOCO+, RefCOCOg

This task uses MSCOCO images available at /datasets/MSCOCO on newton. The annotations are available here: https://huggingface.co/datasets/AoZhang/nextchat-annotation/tree/main

b. LookTwice-QA val set Table 5 [10 points]

This uses images from /datasets/VG on newton. Annotations and data files can be found here: https://huggingface.co/datasets/AoZhang/nextchat-annotation/tree/main

2. Benchmark: #Objects Complexity [20 points]

You will benchmark the Shikra model for number of objects complexity on LVIS dataset (images available at /datasets/MSCOCO17 on newton, annotations: https://dl.fbaipublicfiles.com/LVIS/lvis_v1_val.json.zip). You must generate a table and plot for this benchmark comparing performance (Object detection: mAP) against number of objects per image. Include #samples for each row in the table.

Show results for following #objects per image : [1, 2, 3, 4, 5, 6-10, 11-20, 20+]

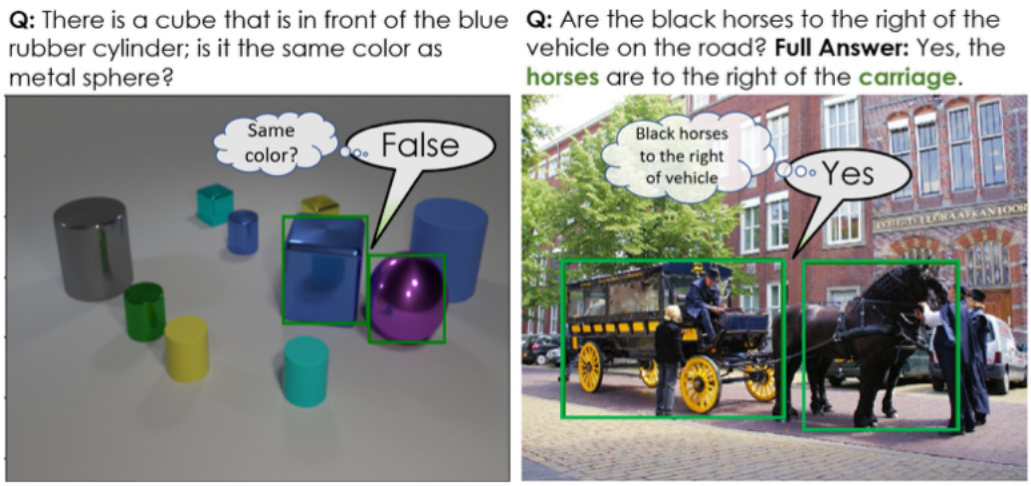
3. Benchmark: Novel Objects [20 points]

You will benchmark the Shikra model for rare vs common objects in LVIS dataset (images available at /datasets/MSCOCO17 on newton, annotations: https://dl.fbaipublicfiles.com/LVIS/lvis_v1_val.json.zip). You must generate a table and plot for this benchmark comparing performance (Object detection: mAP) for `rare` and `common` classes. For this task, you are also expected to show some qualitative results comparing ground truth and predictions along with the original image.

4. New Task: Evaluate on Grounding VQA [40 points]

The goal is to perform VQA with grounding, i.e., you are expected to ground the relevant visual entities. See examples below. You will evaluate on GQA dataset (images available at /datasets/GQA on newton and annotations: https://downloads.cs.stanford.edu/nlp/data/gqa/questions1.2.zip) and report performance metrics IoU, Overlap (refer to https://openaccess.thecvf.com/content/CVPR2021/papers/Urooj_Found_a_Reason_for_me_Weakly-supervised_Grounded_Visual_Question_Answering_CVPR_2021_paper.pdf).

For this task, you are expected to show some qualitative results comparing ground truth and predictions as shown below.



Note: This is an individual assignment, with each student working separately and submitting the necessary deliverables. No training required for this assignment. Submit the report in the expected format.

Data:

Datasets will be available on newton soon; you may have to download annotations as per the instructions.

Expected Deliverables (zip file):

- A report detailing the tasks and their corresponding results (use the provided format for report).
- A recorded video of the assignment, demonstrating sample inferences for each task.
- Source code.