Pre-proposal for Multimodal Machine Learning Project

Research Problem

Improving the understanding of a visual scene and the relationships between objects for multimodal question answering.

Datasets

- Natural Language for Visual Reasoning (NLVR2)
  - It is a language grounding dataset containing natural language sentences grounded in images. The task is to determine whether a sentence is true about a visual input. The data was collected through crowd sourcing, and solving the task requires reasoning about sets of objects, comparisons, and spatial relations.
  - NLVR2 contains 107,292 examples of human-written English sentences grounded in pairs of photographs. NLVR2 retains the linguistic diversity of NLVR, while including much more visually complex images.
- GQA
  - A new dataset for real-world visual reasoning and compositional question answering, seeking to address key shortcomings of previous VQA datasets. The authors have developed a strong and robust question engine that leverages Visual Genome scene graph structures to create 22M diverse reasoning questions, which all come with functional programs that represent their semantics. They also use the programs to gain tight control over the answer distribution and present a new tunable smoothing technique to mitigate question biases.
  - GQA contains 22M questions across 113k images. Vocabulary size is 3097 words with a total of 1878 possible answers.

Modalities Involved

- We will be using visual and textual representations for question answering tasks.
- We are planning to use image representations, object representations and relations between them, textual descriptions of the image

Multimodal Challenges

- **Representation** - Using representations for the image/textual features that store more relevant information and make it easier to process.
- **Fusion** - Learning a joint representation for the unimodal features and fusing them to the same semantic space
- **Alignment** - Connecting locations in images with features that the questions will be asking about

- **Co-learning** - Pretraining model on a richer dataset and using it for a downstream task. This often occurs as there Question Answering datasets are less common than scene labeling datasets, which can be relied upon for learning better representations.
- **Translation** - Extracting textual information from the images using image captioning, object detection and appending them with the question-answer text, which will remove dependency on the heavy multimodal pretrained model.

Evaluation Metrics

These metrics are taken from the GQA paper.

- **Accuracy** - Standard QA accuracy for each question-answer pair, give 1 point if predicted answer matches and 0 otherwise.
- **Consistency** - A metric for the level of consistency in responses across different questions.
- **Validity**: Measures whether the model gives valid answers, ones that can be theoretically correct for the question.
- **Plausibility** - Measures whether the model responses are reasonable in the real world or not making sense.
- **Grounding** - For attention models only. Measures whether the model attends to regions in the image that are relevant for the question.
- **Distribution Scores** - Measures the overall match between the true answer distribution and the model predicted distribution.

Baselines

Models	Code Available	Accuracy		
		VQA	NLVR2	GQA
LXMERT	Yes	72.5	76.2	60.3
UNITER	Yes	73.8	79.1	-
MAC	Yes	68.3	-	54.1
Visual BERT	Yes	70.8	67.4	-

Extension of Prior Work

- As there are multiple VQA datasets, we leverage the techniques used in different datasets and co-learning to improve results across the different datasets.
- We will visualize current models to analyze where they fail, and improve the model's performance for those cases. Also, we will work towards visualizing parameters of our model to make it more interpretable.

- Feature engineering by extracting object representations and relations that might be useful for the model on the multimodal QA task. Further, looking at the use of keyphrase extraction models on the input text. Overall our idea is to augment input modalities to learn better representations.
- Lastly, we want our approach to learn commonsense by developing better understanding between the images and questions and choose answers that maximizes their alignment.

Team Members

- Andrew Lyubovsky (alyubovs@andrew.cmu.edu)
- Feng Xiang (fxx@andrew.cmu.edu)
- Sumit Agarwal (sumita@andrew.cmu.edu)
- Suraj Tripathi (surajt@andrew.cmu.edu)
- Syeda Nahida Akter (sakter@andrew.cmu.edu)

Each team member will go through a different approach and address one challenge each.

Timeline

- **Sept 15** - Submit pre-proposal
- **Sept 26** - Research Literature
- **Oct 10** - Experiment with unimodal representations
- **Nov 1** - Implement and evaluate state-of-the-art models
- **Dec 5** - Implement and evaluate new research ideas and work on the multimodal challenges

GPU/CPU information

Yes, we are interested in using Amazon Web Services or Google Cloud Platform as part of our project.