GSOC 2024 Project Proposal at RedHenLab

Archit Mangrulkar

April 2, 2024

Summary of the Proposal

In this project, I propose to develop a a system that utilizes multimodal vision models to process videos and images and return json output with annotation output. I have carried out extensive literature review in this domain covering techniques such as Memory, Transformer, Graph, Neural Modular Network, and Neural-Symbolic methods. I will evaluate the performance of our models using standard evaluation metrics such as and analyze the impact of different features on the performance of the models. I will also address ethical considerations such as bias and fairness due to spurious correlations in the dataset

Background

In the past, there has been significant work in the field of multimodal VQA. I have classified my literature survey into VQA for images and videoes-

- Visual QA for Videoes
  - In [5], a novel zero-shot (ZS) VQA algorithm incorporating external knowledge graphs (KGs) is proposed to address existing issues of error cascading and answer bias in traditional methods. The algorithm encodes image-question pairs and answers into three embedding spaces, leveraging semantic information and object classification to enhance answer prediction. Testing on the ZS-F-VQA dataset shows superior performance compared to state-of-the-art models, with GloVe embeddings yielding the best results.
  - The authors of [17] discuss language bias in VQA models and proposes a learning strategy to mitigate its impact while retaining its benefits. By jointly learning local content and global context, the model effectively reduces bias without underestimating individual instances' content. The proposed approach, evaluated on VQA v2 and VQA-CP v2 datasets, outperforms existing methods and leads models to focus on more relevant image areas during prediction.
  - Spatio-temporal Attention ([9]): This approach focuses on directing the model's attention to relevant parts of the video throughout its duration. This allows the model to not only identify objects in frames (spatial attention) but also understand how they move and interact over time (temporal attention).
  - Motion-appearance Memory ([7]): This method emphasizes the importance of remembering both an object's appearance and its motion history. This combined memory allows the model to answer questions that require understanding an object's past actions or interactions with other objects in the video.
  - Spatio-temporal or Hierarchical Graph Models ([6]; [16]): These approaches represent videos as graphs, where nodes represent objects or frames, and edges capture relationships between them. By incorporating both spatial (object proximity) and temporal (object interactions across frames) information, these models can effectively reason about complex video content and answer intricate questions.
- Visual QA for Images
  - In [14], a novel method is proposed for VQA, leveraging external textual sources like Wikipedia articles for relevant information retrieval. The approach encodes questions and images separately and retrieves pertinent passages from external sources based on this encoding. By incorporating external textual data, the method addresses the challenge of limited visual information, achieving state-of-the-art results on benchmark datasets. The authors stress the importance of selecting relevant textual sources and propose a method for source selection based on question-text similarity.

- The paper [10] introduces a weakly supervised grounding technique for VQA using vision-language transformers, eliminating the need for precise object localization during training. The approach relies on weak supervision to train the model in attending to relevant image regions based on questions. The proposed method, evaluated on various metrics and datasets, shows promise for weakly supervised grounding in VQA tasks.
- In [2], the VLMO architecture is presented, which resolves vision-language classification and image-text retrieval challenges by utilizing a Mixture-of-Modality-Experts (MoME) framework. VLMO integrates multiple modalities like audio, text, and image, improving model performance. Pre-training is conducted separately for image and text experts, facilitating better generalization. Experimental results demonstrate state-of-the-art performance on VQA and NLVR2 tasks.
- The authors of [13] propose a Coarse-to-Fine Reasoning (CFR) framework for VQA, aiming to bridge the semantic gap between images and questions. The framework processes images and questions through embedding modules and jointly learns features and predicates. Through intensive experiments on popular datasets, the proposed method achieves superior accuracy compared to existing techniques, providing an explainable decision-making process.
- In [11], Graphhopper is introduced as the first VQA method utilizing reinforcement learning for multi-hop reasoning on scene graphs. By combining computer vision, knowledge graph reasoning, and natural language processing techniques, it achieves accurate answers to free-form questions by associating semantic and linguistic understanding of objects in images. Experimental studies on the GQA dataset demonstrate Graphhopper's ability to reach human performance and outperform the Neural State Machine (NSM), a similar method.

Goal and Objectives

covering techniques such as Memory, Transformer, Graph, Neural Modular Network, and Neural-Symbolic methods. I will evaluate the performance of our models using standard evaluation metrics such as and analyze the impact of different features on the performance of the models

Goal 1 is creating a diverse and annotated dataset covering images, videoes from multiple scenes

Goal 2 will be to setup a baseline model training pipeline in a Torch environment

Goal 3 will be augmentation of external memory for improving QA

Goal 4 will be leveraging Knowledge Graph based representation of visual scenes for improving Answer quality

Goal 5 will be surveying Neural Symbolic methods for our use case

Methods

The overall approach to this project can be broken down into several steps:

1. Data Collection:
   The first step is to collect a large corpus of videoes & images for QA. Based on the research conducted by me, it is evident that among the most frequently utilized datasets for Visual Question Answering (VQA) are VQA v2 ([8]) and VQA-CP2 ([1]). BeiT-3 ([15]) emerged as the top-performing model on the VQA v2 dataset's teststd version and the PaLI ([4]) model on the test-dev version of VQA v2.

2. Annotation:
   Both the above datasets lack JSON schema input in the form that we need. To combat this, we can train a BERT model to identify value from the question, thus helping us restructure the input into question into description-value pairs.

3. Feature Extraction:
   Audio and visual information can be used to extract features that capture different aspects of the news program, such as the tone of the speaker's voice, facial expressions, and body language. The features extracted can be used as input to the machine learning models that will be trained to classify the stance of the news program. Since manual annotation of large amounts of data is expensive and time-consuming, we will utilize weak-supervised and unsupervised methods to reduce the annotation effort.

- Weakly Supervised Learning
  We will use a combination of manual annotation and weak supervision techniques to label a small sample of the videos. The weak supervision techniques include distant supervision and self-supervision. Distant supervision (Mintz et al., 2009) [12] involves using an external knowledge base (in this case, a set of manually annotated videos) to label a large corpus of unlabeled data. Self-supervision involves using the video itself to learn visual QA. We can perform self-training to iteratively improve model performance (Zhou & Li, 2005 [18]), where the model is first trained on the labeled data, and then used to classify the unlabeled data. The highly confident predictions are then added to the labeled data for retraining the model. Another technique that can be used is co-training, where multiple classifiers are trained on different sets of features, and each classifier labels a subset of the unlabeled data. The labeled data from each classifier is then combined to train the final model (Blum & Mitchell, 1998 [3]).

- Unsupervised Learning
  To perform unsupervised stance detection, we will use a variety of clustering and topic modeling techniques. We will use the same textual, visual, and acoustic features as in supervised learning. We will also use topic modeling techniques to identify topics and their associated stances.

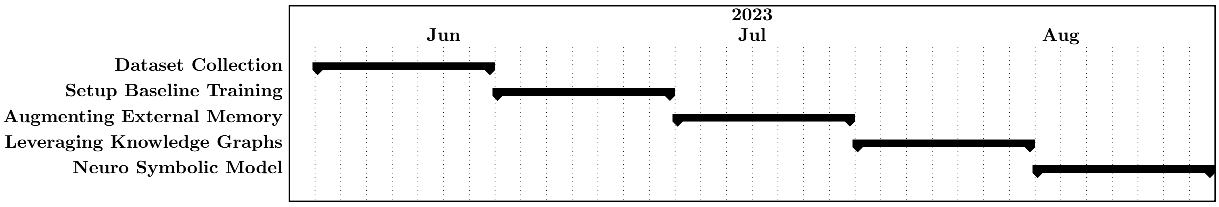
4. Evaluation: To evaluate the performance of our stance detection models, we will use standard evaluation metrics such as:

   - Accuracy
   - BLEU
   - ROUGE-L
   - CIDEr
   - Hit@K
   - Recall@K
   - Precision@K

   We will also perform cross-validation and test the models on different datasets to evaluate their generalizability. We will compare the performance of our supervised and unsupervised models and analyze the impact of different features on the performance of the models.

5. Analysis: Bias and fairness are important ethical considerations in this project. We will ensure that our annotated dataset is balanced and representative of different perspectives and viewpoints. We will also investigate the presence of bias in the data and the models and take steps to mitigate it.

Tentative Timeline



References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering, 2018.

[2] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts, 2022.

[3] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory (COLT '98)*, pages 92–100, 1998.

[4] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model, 2023.

[5] Zhuo Chen, Jiaoyan Chen, Yuxia Geng, Jeff Z. Pan, Zonggang Yuan, and Huajun Chen. Zero-shot visual question answering using knowledge graph, 2021.

[6] Anoop Cherian, Chiori Hori, Tim K. Marks, and Jonathan Le Roux. (2.5+1)d spatio-temporal scene graphs for video question answering, 2022.

[7] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering, 2018.

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering, 2017.

[9] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering, 2017.

[10] Aisha Urooj Khan, Hilde Kuehne, Chuang Gan, Niels Da Vitoria Lobo, and Mubarak Shah. Weakly supervised grounding for vqa in vision-language transformers, 2022.

[11] Rajat Koner, Hang Li, Marcel Hildebrandt, Deepan Das, Volker Tresp, and Stephan Günnemann. Graphhopper: Multi-hop scene graph reasoning for visual question answering, 2021.

[12] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP (ACL-IJCNLP)*, pages 1003–1011, 2009.

[13] Binh X. Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D. Tran, and Anh Nguyen. Coarse-to-fine reasoning for visual question answering, 2022.

[14] Chen Qu, Hamed Zamani, Liu Yang, W. Bruce Croft, and Erik Learned-Miller. Passage retrieval for outside-knowledge visual question answering. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21. ACM, July 2021.

[15] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. Image as a foreign language: Beit pretraining for all vision and vision-language tasks, 2022.

[16] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering, 2022.

[17] Chao Yang, Su Feng, Dongsheng Li, Huawei Shen, Guoqing Wang, and Bin Jiang. Learning content and context with language bias for visual question answering, 2020.

[18] Ding Zhou and Chong Li. Improving text categorization by using unlabeled documents. In *Proceedings of the 18th International Conference on Neural Information Processing Systems (NIPS'05)*, pages 1557–1564, 2005.