11-777 Report 1: Dataset Proposal and Analysis by Team-Samosa

Bhavuk Sharma* **Debayan Ghosh*** **Rituraj Singh***
Saahithya Yendluri* **Wallace Dalmet***
{bhavuks, debayang, riturajs, syendlur, wdalmet}@andrew.cmu.edu

1 Problem Definition and Dataset Choice

Our team has decided to work on the TVQA [1] / TVQA+[2] dataset. TVQA [1] is a large scale localized video question-answering dataset. The goal of the dataset is to develop multi-modal models that can employ visual concepts and text embeddings to answer questions based on the contents of a video.

Our objectives are to build baselines for Questions-Answers (QA) on full videos, develop a mechanism that can evaluate interpretability of a model's output and to find the relation between the dominance of a particular modality over the other for answering certain question types. For example, to answer the question "What was Sheldon holding when Penny was painting?", a well generalized model is expected to gather sufficient information from visual cues over textual embeddings.

1.1 What phenomena or task does this dataset help address?

The dataset can be used for three major tasks: QA on the grounded clip, QA on full clip and question-driven moment localization. Our focus is on the second task i.e. QA on full clip which can be formulated as follows; given a video V, subtitles S, a question q, possible answers $\{a_i\}_0^4$ and a set of bounding boxes B , the model should pick the correct answer [2].

1.2 What about this task is fundamentally multimodal?

The task of QA on a full clip requires a 60-90 second video clip q, a subtitle S, a question q, and five candidate answers $\{a_i\}_0^4$ as input to the model. The visual signal as input from the video represents one modality, the word embeddings generated from the written natural language i.e. subtitle, question and answers form another modality that are fed into the model. These different modalities as inputs to the model form a multimodal system which shares cross-modal information capturing correspondences between modalities.

1.3 Hypothesis

The three places cross-modal information can be used are:

1. *Natural Answer Generation:* Given a question, visual features and text transcripts, generate natural answers to the questions asked. This is in contrast to choosing a correct answer from a predefined set of answers.

2. *Modality Contribution Metric:* Given multiple modalities, build a framework which can assess how essential each modality is for a given task.

3. *Video Segment Retrieval:* Given a Natural Language Query, retrieve a segment from the video that best describes the input query.

1.4 Expertise

We have the following expertise in the underlying modalities required by this task:

1. Bhavuk Sharma: Took Intro to ML (10-601). Currently taking Advanced DL (10-707), Machine Learning with Large Datasets (10-605) and Intro to NLP (11-611).

2. Debayan Ghosh: Taking Intro to DL (11-785) currently. Took Intro to ML (18-661), ML for Large Datasets (10-605) and Speech Recognition (11-751).

3. Rituraj Singh: Took Intro to DL (11-785), Natural Language Processing (11-611) and Intro to Machine Learning (18-661).

*Everyone Contributed Equally – Alphabetical order

4. Saahithya Yendluri: Took Intro to DL (11-785), Computer Vision (16-720) and Machine Learning for Large Datasets (10-605).

5. Wallace Dalmet: Took Intro to DL (11-785), Computer Vision (16-720) and Machine Learning for Signal Processing (11-755)

2 Dataset Analysis

There are two datasets in the project; TVQA [1] and TVQA+ [2]. The latter is an augmented version of a subset of TVQA compiled by adding bounding boxes to relevant objects in each frame. This section presents the nature of the two datasets.

2.1 Dataset Properties

The TVQA [1] dataset is a compilation of clips from 6 popular TV shows across 925 episodes spanning a total of 460 hours. There are 21,793 clips that are coupled with 152,545 QA pairs.

Each clip is 60 to 90 seconds long. The subtitles contain character names and dialogues with their start and end timestamps. Each video has 7 corresponding multiple choice questions with 5 possible answers, one of which is correct. The questions in the dataset are carefully curated and have two parts: a main question which requests the answer to "what"/"why"/"how"/"where" and a moment localization indicator that points to the relevant segment in the video that specifies "when" that event occurred. For example, "*What* was in Sheldon's hand *when* he was playing tennis with Penny?". For majority of the videos, the localized segment length spans between 5-10 seconds. Figure 1 indicates the distribution of the various types of questions.

The TVQA+ [2] dataset is developed from a subset of TVQA. This subset contains 4198 clips with 29,383 questions. The primary difference between TVQA and TVQA+ is the introduction of 310k frame level bounding boxes across 2500 categories. The idea behind TVQA+ is to use "referenced visual concepts" i.e. objects and people that can potentially bolster the performance of a multi modal question answering model. While QA pairs are temporally localized, detected objects add spatial localization.

TVQA and TVQA+ are both split into 80% train, 10% validation and 10% test sets.

2.2 Compute Requirements

1. **Files**
   Table 1 shows the file size requirements for the TVQA+ dataset.

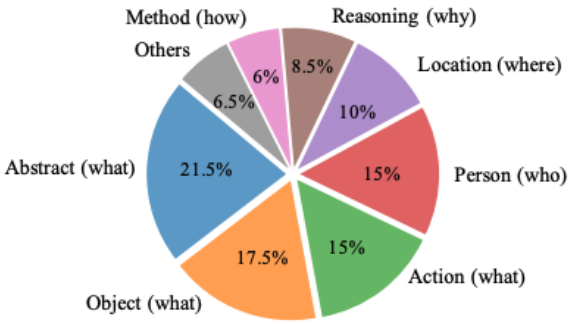

Figure 1: Distribution of question types [1]

File Type	Size
Video Frames	43 GB
ImageNet Features	34 GB
Visual Concept Features	97 MB
Annotations	7 MB
Subtitles	5 MB

Table 1: File Sizes for TVQA+.

2. **Models**
   On average, each clip has 230 frames or 3 frames per second. The total size of the dataset is 43 GB. Further, an Imagenet pretrained Resnet model was used to extract $N \times 2048$ dimension features, where $N \in (0, 300)$ is the number of frames in the clip. The total size of the extracted features is 34GB. Google Colab Pro+ and Amazon Web Services (AWS) g4dn.xlarge instance should be sufficient to store this data.

2.3 Modality analysis

1. **Lexical diversity:** The subtitles in the TVQA+ dataset without the <eos> tags have a total of 648,729 tokens with 24,825 of them being unique. Next, the questions in the TVQA+ dataset has 297,790 words with 8453 being unique. Finally, the answers in the TVQA+ dataset has 567,643 words with 16,063 being unique.

2. **Sentence length:** The subtitles for a particular clip have several sentences of the format: "Howard : What if they 're women ? <eos>". On average each sentence uttered by a character is approximately 7 words and has the

Method	VQA Acc.
iPerceive	**0.769**
HERO	0.742
STAGE	0.705
PAMN	0.667
Multi-task learning	0.662

Table 2: Baseline models for TVQA.

maximum length of 38 words without including the character's name.

3. **Bounding boxes:** The TVQA+ dataset contains approximately 2.09 boxes per image and 10.58 boxes per question. The bounding boxes cover over 2527 categories.

2.4 Metrics used

The TVQA+ [2] paper mentions four metrics for evaluating QA performance namely, classification accuracy for quantifying if correct answer choice was selected, mean Intersection over Union (mIoU) for evaluating span predictions, Answer-Span joint Accuracy (ASA) for jointly evaluating answer and span predictions and mean Average Precision (mAP) for evaluating object grounding performance.

2.5 Baselines

Building video question answering models is a challenging task. The human QA accuracy on the TVQA dataset is 90.46 %. As of 2022, iPerceive by Chadha et. al holds the best QA accuracy of 76.96 % on TVQA. Table 2 shows recently developed methods that are competitive with the STAGE baseline provided by the authors of TVQA+.

1. *iPerceive [3]:* The authors have propose a framework that enables common-sense learning for videos to infer causal relationships between objects in a video by building a knowledge base using contextual cues leading to improved performance on the TVQA dataset. The train/val/test splits for TVQA were 80%/10%/10% with a validation accuracy of 76.9% on the VideoQA task.

2. *PAMN [4]:* The authors propose a progressive attention memory network that provides a multi-step reasoning approach to localize necessary information from question, answers and visual features as opposed to traditional architectures which employ single-step reasoning. The question answering task introduced in the paper achieved 66.77% accuracy on the TVQA dataset.

3. *HERO [5]:* In this method, the authors use a hierarchical architecture, consisting of cross-modal transformer to capture local context from videos and their corresponding text embeddings and a temporal transformer to capture the global video context in the form of video frame embeddings. This method achieves 74.24 % accuracy on the TVQA dataset.

4. *Multi-task learning for Multi-Modal Video Question Answering [6]:* Here the authors proposed a multi-task learning approach for VQA. Their model simultaneously learned QA, modality alignment and temporal localization, using hierarchically shared layers. This method showed an improvement in performance, as their model was able to outperform the state-of-the-art at the time, with a test accuracy of 67.05%.

3 Team Member Contributions

Bhavuk Sharma Performed baseline analysis, assisted with creating hypothesis.

Debayan Ghosh Analyzed the PAMN dataset as a baseline, and performed modality analysis.

Rituraj Singh Analyzed dataset properties, added modality contribution hypothesis and evaluated HERO as a baseline.

Saahithya Yendluri Performed modality analysis, added metrics used, assisted for the paper HERO as a baseline.

Wallace Dalmet Assisted in cross-modal hypothesis and evaluated iPerceive as a baseline.

References

[1] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018.

[2] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Tech Report, arXiv*, 2019.

[3] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. *CoRR*, abs/2011.07735, 2020.

[4] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D. Yoo. Progressive attention memory network for movie story question answering. *CoRR*, abs/1904.08607, 2019.

[5] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: hierarchical encoder for video+language omni-representation pre-training. *CoRR*, abs/2005.00200, 2020.

[6] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D. Yoo. Gaining extra supervision via multi-task learning for multi-modal video question answering. *CoRR*, abs/1905.13540, 2019.