

# 11-777 Final Report: Visual Concept Localizer based Transformer VQA System

Bhavuk Sharma\* Debayan Ghosh\* Rituraj Singh\*  
Saahithya Yendluri\* Wallace Dalmat\*

{bhavuks, debayang, riturajs, syendlur, wdalmat}@andrew.cmu.edu

## Abstract

Visual Question Answering (VQA) is one of the many open research problems in multi-modal environments. Enabling systems to form relationships between sequence of events and dialog to develop reasoning has been addressed in recent VQA architectures. We use the TVQA+ dataset which has videos, subtitles and natural language questions. More formally, given a video  $V$ , subtitles  $S$ , a question  $q$  and 5 possible answers  $\{a_i\}_0^4$ , the VQA model should pick the correct answer. In this work, we introduce Visual Concept Localizer (VCL), a fast and portable novel mechanism to retrieve relevant video frames needed to answer an input question. We observe that integrating the VCL with a Visual BERT transformer improves the performance over multimodal baseline models by 12 %. We employ early fusion of the text and visual modalities to enforce cross-modal learnings mapped by the transformer.

## 1 Introduction

Video Question Answering is one of the open challenges that falls under the pool of multi-modal problems. It requires models to utilise concepts learned from both the video and text modalities to answer natural-language questions. The high-level goal of any VQA task is to make machines infer and understand images/videos similar to how humans would and answer questions in natural language based on the inferences made. Video question answering has additional complexity over only Image-based Question Answering since it introduces a temporal aspect. The temporal nature of the problem warrants the use of attention and transformer based mechanisms which have been heavily adopted by state-of-the-art architectures recently.

Video based QA systems have several real-world applications as mentioned in (Barra et al., 2021),

the first of which is medical imaging. Building efficient VQA systems that can infer from medical scans and images will help support doctors in critical decision making. Another application is helping visually impaired people by answering questions about daily tasks. In addition to these two, VQA can be used for surveillance and for advertising.

A task like VQA which requires learning from multiple modalities brings forth various set of challenges as described by (Malinowski and Fritz, 2014). The 3 core challenges outlined here are: Effective representations of the visual and linguistic concepts such that they are scalable, address conceptual ambiguities, learn implicit attributes of objects and their relative position; The second challenge is that of learning common sense knowledge. An example mentioned is “What object on the table can be used for cutting?”. To answer this the model needs to understand the task of cutting and associate an object with it; The third challenge is defining a solid benchmark and quantifying performance of the models. We try to address the first and second challenges by incorporating visual concept features so as to push the model to learn more from the video modality instead of putting emphasis only on the textual modality.

In this work, we introduce a novel mechanism called Visual Concept Localizer (VCL) that identifies the set of video frames that are most relevant to answer a given input question. We integrate the VCL module with a Visual BERT pipeline and compare its performance improvement with unimodal and multimodal baselines. The primary objective of using a Visual BERT transformer is to encourage early fusion between text and visual modalities and ensure cross modal interaction such that the model learns the relationships between text and video frames. We run our experiments on the TVQA+ dataset which consists of clips from The Big Bang Theory.

---

\*Everyone Contributed Equally – Alphabetical order

The rest of the paper is organised as follows: Section 2 describes the related work in literature. Section 3 provides a detailed description of the dataset. The unimodal and multimodal baselines we introduced in our previous report is recapped in Section 4. In Section 5, our novel Visual Concept Localizer (VCL) module and the model architecture is introduced. Results and Analysis of our model performance are discussed in Section 6 and 7. Section 8 and 9 discusses possible extensions, limitations and ethical concerns. Finally, Section 10 lays out the team member contributions.

## 2 Related Work and Background

Visual QA datasets are generally composed of videos that are associated with descriptive text or dialog. Each video has a corresponding set of questions with either multiple choice, one-word or a single phrase answer. The TVQA (Lei et al., 2018) dataset has videos, subtitles and the QA pairs. The TVQA+ (Lei et al., 2019) has bounding boxes for objects in addition to the TVQA dataset. Unlike TVQA, the MovieQA (Tapaswi et al., 2015) dataset has significantly longer text as well as multiple sources of information like movie plots, subtitles, and scripts. In DiDeMo (Hendricks et al., 2017) (Distinct Describable Moments), there are 10,000 videos with 40,000 pairs of expressions that can help identify the moment in a video that holds the answer to a given question. In the TGIF-QA (Jang et al., 2017) dataset the authors propose 3 new tasks in addition to MCQ answering: “Repetition count”, “repeating action”, “state transition” which aids the model in performing better. In LifeQA (Castro et al., 2020), the videos capture real-life encounters unlike TVQA and MovieQA datasets which are scripted, directed and enacted by professionals. The Multimodal QA dataset (Talmor et al., 2021) has 35.7% of the questions which require cross-modal reasoning which helps evaluate a model’s ability to form inferences with cues from multiple modalities. In the ActivityNet dataset (Caba Heilbron et al., 2015) in addition to video and linguistic features has audio features which could be helpful in the VQA task. (Zhuang et al., 2020). The HowToVQA69M (Yang et al., 2020a) is a large-scale dataset for VQA generated automatically without manual annotations by employing a transformer based approach.

To observe the influence of each modality in a cross-modal VQA design, analyzing various uni-

modal baselines on the TVQA dataset is essential. (Kim et al., 2019a) shows that the subtitle features captures more information about the QA task in relation to the video features. (Geng et al., 2020) explored the idea of incorporating the relationship between different characters in the TVQA dataset which lead to a performance improvement on the unimodal and multimodal TVQA baselines. (Winterbottom et al., 2020) observed that the subtitle modality overshadowed the video modality in terms of information required to answer the questions and proposed using a subset of these samples to mitigate this bias. (Engin et al., 2021) uses dialog to summarize on a per scene and per episode basis to get a high level understanding of the video which brings forth an accuracy improvement of 5.4%. Further, we study the influence of the audio modality from (AlAmri et al., 2019) which shows that this addition of audio feature improves model performance by 0.13% when combined with dialog history and video features, indicating there is complementary knowledge between the video and audio signals.

Prior work for the task of VQA on TVQA and other related datasets encompasses several novel models as mentioned below. In (Kim et al., 2019b), the authors propose a dual memory structure that pinpoints to specific temporal parts of each modality and also developed a “Dynamic Modality Fusion” mechanism that uses soft-attention to adaptively fuse modalities conditioned on an input question; thereby determining the contribution of each modality. (Khan et al., 2020) achieves state-of-the-art performance on the TVQA dataset by heavily employing BERT encoders for video, subtitles, questions and answers. In (Guo et al., 2021), the idea of embedding features in the quaternion space is introduced to represent multimodal information; video, subtitles, questions and answers. The objective is to realize multimodal and structural reasoning through a quaternion hypergraph network (QHGN). In (Kim et al., 2020), the authors propose a Moment Proposal Network (MPN) that locates the most appropriate temporal moment from each modality (weighted) and a Heterogeneous Reasoning Network (HRN) that uses the localized moment to predict an answer using attention. (Kim et al., 2019a) proposes reusing weights from intermediate layers of the Video/Subtitle QA task for Modality Alignment and Temporal Localisation. The paper also describes a Multi-Task learning based

Multi-Task ratio scheduling method to learn the ‘easier’ task earlier, which allows us to set the inductive bias at the beginning of training. (Yang et al., 2020b) uses a dual-flow BERT system and observe that using timestamped subtitles and video information gave state-of-the-art performance for all variations. (Li et al., 2021) proposes a model which encodes the static and dynamic relationships and also the interactions between detected objects in the video by using GAT.

There are patterns as well as distinct approaches to how the problem of visual question answering can be solved. A common theme among these approaches is the use of visual features extraction using FasterRCNN as an example and text embeddings using BERT (Devlin et al., 2018). Depending on the model’s design, these representations are fused at a later stage after passing through a sophisticated attention mechanism. Finally, a combination of fully connected layers and softmax cross-entropy are used for the final output and loss. In this section, we summarize some of the relevant techniques that are potentially helpful in building a new VQA architecture. In (Lu et al., 2016), the authors devise “co-attention”, a novel mechanism that interlinks images and questions. The image representation are responsible for tending to a given question’s attention while the question representation tends to the image’s attention; thereby creating an effect of “natural symmetry”. (Gao et al., 2018) and (Zhuang et al., 2020) use models containing memory modules, that encode input sequence information and remember relevant information that could aid answer prediction.

### 3 Task Setup and Data

#### 3.1 Overview

We use the TVQA+ dataset (Lei et al., 2019) for our experiments. This dataset is a compilation of clips from the Big Bang Theory TV show, that contains 29,383 QA pairs from 4,198 videos, with 148,468 images annotated with 310,826 bounding boxes. Figure 1 describes the Question type distribution of the TVQA+ dataset.

#### 3.2 Data Format and Description

Each video 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

Question Type Distribution - TVQA+

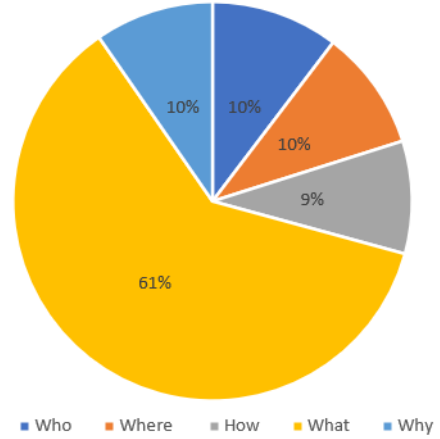


Figure 1: Question Type distribution for the TVQA+ (Lei et al., 2019) dataset

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.

In addition, TVQA+ has 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.

The TVQA+ dataset is split into 80% train, 10% validation and 10% test sets.

#### 3.3 Compute Requirements

##### 1. Files

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

##### 2. Pretrained 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

Table 1: File Sizes for TVQA+

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

be sufficient to meet the storage and compute requirements.

### 3.4 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 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.

### 3.5 Motivation

The TVQA+ dataset is set up in such a way that models are required to pick up cues from both the visual information and language information to efficaciously find an answer to a given question. This motivates us to build models that can effectively learn the interaction between multiple modalities for the Visual QA task.

## 4 Baselines

In this section, we briefly discuss the Unimodal and Multimodal baselines proposed in our previous report.

### 4.1 Unimodal Baselines

The foundation of a robust multimodal solution lies with strong analysis of unimodal experiments

and gathering evidence to create a multimodal hypothesis. We discuss the three unimodal architectures and their respective analysis. At a bird's eye view, each of the models discussed has one primary modality; video, subtitles or just the question-answer (QA) pairs.

1. **Video Unimodal:** Given a video frame  $V$ , a question  $Q$  and candidate answer choices  $\{a_i\}_0^4$ , we first get the 768 D BERT (Devlin et al., 2018) embeddings pretrained on the Toronto Book and Wikipedia corpus for the question and answer choices  $\{a_i\}_0^4$ . The 2048 D video features are extracted using the ImageNet pretrained ResNet-101 (He et al., 2015) model (at pool5 layer). The choice of BERT and ResNet as choice of feature embeddings were taken from (Khan et al., 2020) and (Lei et al., 2018). The BERT text embeddings and the ResNet-101 features are projected to a lower dimension (256 D). The projection to a lower dimension aims to capture as much information as possible from the original embeddings while reducing the computational complexity in the further layers. To compute this projection, the text features employ a Linear Projection, whereas the video features are downsampled using a Convolutional Neural Network. The question, answer and video hidden representations are fused by concatenating them across  $seq\_len$  for each answer. A Max-Pool across the  $seq\_len$  dimension is done to obtain one joint embedding per answer choice  $a_i$ . We finally use a classification layer comprised of linear layers and GELU activations to convert fused features into logit scores for all the five answers.
2. **Subtitles Unimodal:** Given a subtitle  $S$ , a question  $Q$  and candidate answer choices  $\{a_i\}_0^4$ , we compute 768 D embeddings using a pre-trained BERT (Devlin et al., 2018) model pretrained on the Toronto and Wikipedia corpus. These embeddings are projected to a lower 256 dimension and concatenated along the  $seq\_len$  for each of the answer choice. The aim of fusing the subtitle modality along with the question and individual answer choices is to capture the information in a single representation. For each of the hidden answer representations, we calculate the query, key and value projections and apply



self-attention mechanism using the query, key and value triplets to get the attended feature representations. Next, we take the maximum values across the  $seq\_len$  dimension to obtain a single cross-modal embedding per answer choice. We finally use a linear layer to obtain the logit scores for all the 5 answer choices.

3. **QA Unimodal:** This baseline consisted simply of a question  $Q$  and candidate answers  $\{a_i\}_0^4$ . Similar, to the first two approaches we obtained 768 D BERT(Devlin et al., 2018) embeddings for the question and answer choices  $\{a_i\}_0^4$ . These representations were projected down to 256 dimension and fused by concatenating along the  $seq\_len$  dimension. We then took the maximum values across the  $seq\_len$  dimension and finally used a classification layer comprised of linear layers and GELU activations to obtain the logit scores for all the 5 answer choices.

## 4.2 Simple Multimodal Baselines

In this section, we discuss simple multimodal models along the same architecture used for the unimodal baselines. The architectures can be distinguished by the type of embeddings used and the presence of self-attention blocks or Linear layer blocks. The three models discussed are as follows:

1. **BERT + RESNET with Self-Attention:** Given a subtitle  $S$ , Video Frame  $V$  a question  $Q$  and candidate answer choices  $\{a_i\}_0^4$ , we extract 768 D text embeddings for the Question, Answer Choices, and Subtitles using a pre-trained BERT module, and 2048 D Resnet features for the Video frames. All the embeddings and feature representations are projected down to 256 dimension using a Linear layer to reduce computation in the further layers while capturing as much information as possible from the original embeddings. The question and subtitle are concatenated along the  $seq\_len$  dimension with each answer choice, to obtain fused  $QAS$  features. Similarly, we get fused features for each answer choice with question and video representations  $QAV$ . Next, we pass both the fused  $QAS$  and  $QAV$  representations through a Self-Attention block to get the attended Video and Subtitle feature representations. These attended Video and Subtitle representations

are max-pooled across the  $seq\_len$  dimension and passed through a linear layer to get a single Video score  $Score_V$  and Subtitle score  $Score_S$  per answer choice. A weighted concatenation of  $Score_V$  and  $Score_S$  is used to obtain a single logit score per answer choice. The weighted concatenation scheme allows for flexibility in choosing which modality to focus on.

2. **BERT + RESNET without Self-Attention:** Similar to the first multimodal baseline, we obtain the fused features  $QAS$  and  $QAV$  representations for the subtitle and video modalities respectively. However, instead of self-attention we simply use linear layers with GELU activations to further learn the representations and extract the maximum values along the  $seq\_len$  dimension. The outputs from the subtitle and video sets are fused and finally a classification layer is applied comprising of Linear layers and GELU activation to obtain a single logit score per answer choice.
3. **GloVe + RESNET with Self-Attention:** Similar to the first multimodal baseline, in this model we employ GloVe (Pennington et al., 2014) to get a 50 D word embeddings for the question, subtitle, and candidate answer choices. The video features are still extracted using the Resnet-101 module (2048 D). The fused features  $QAS$  and  $QAV$  are obtained in a similar manner as described in the first multimodal baseline. Next, we pass both the fused  $QAS$  and  $QAV$  representations through a Self-Attention block to compute the attended Video and Subtitle feature representations. These attended Video and Subtitle representations are max-pooled across the  $seq\_len$  dimension and passed through a linear layer to get a single Video score  $Score_V$  and Subtitle score  $Score_S$  per answer choice. A weighted concatenation of  $Score_V$  and  $Score_S$  is used to obtain a single logit score per answer choice.

## 5 Proposed Model

### 5.1 Motivation

The goal of Visual QA is to facilitate Cross-Modal interaction such that the model is able to leverage both the vision and text modalities to efficaciously

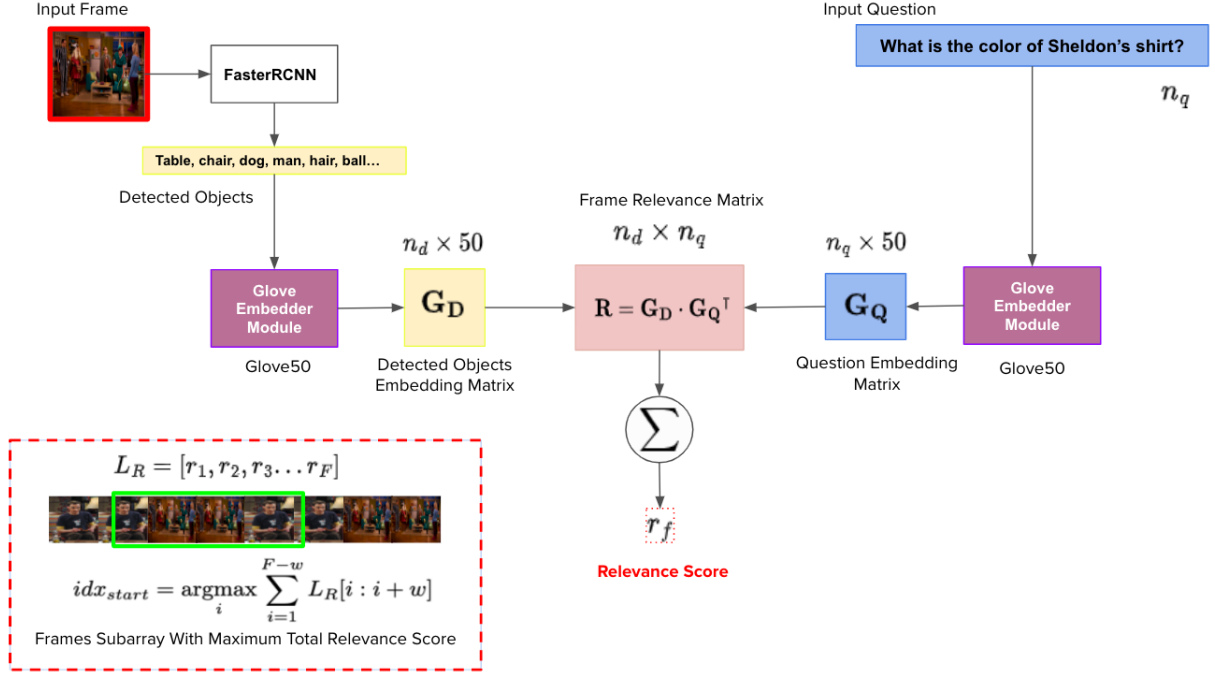


Figure 2: Visual Concept Localizer

find an answer to a given question. While humans have the natural ability to pick up multi-modal cues, it has been a challenge for machine learning models to replicate the same level of effectiveness. Motivated by this, in our prior work we built Multimodal baselines that employ multiple Self-Attention (SA) blocks in a step-by-step fashion to learn joint cross-modal representations from both vision and text.

We found that while the SA-based cross-modal representations were capable of answering questions that only required the text modality, it fell short in mapping the relationship between the question and the visual modality. This motivated us to examine how we can utilize the video modality more effectively.

In works such as UNITER (Chen et al., 2019) and VisualBERT (Li et al., 2019), the input features are visual regions of an image and textual tokens. The fusion of these two modalities is performed by a single multi-layer transformer block with the hypothesis that it will learn cross-modal contextualized embeddings by forming an implicit relationship between visual regions and text tokens. Inspired by their modality fusion techniques, we experimented with a VisualBERT block aided by a novel technique to localize moments in a video, called Visual Concept Localizer.

## 5.2 Model Components

### 5.2.1 Visual Concept Localizer

Humans have an impeccable ability to filter out noisy information in videos and focus on the most important parts. For a VQA task, a human can easily determine what part of a video is needed to answer a given input question. In our previous approach, we had a mechanism that used either all of the frames in the video or equally distributed frames based on the stride at which we sample the frames. Considering an example, if the question asks about the color of a character’s shirt, all frames except those having the shirt can be categorized as noise. Instead, we hypothesize that feeding only the frames which have the shirt into the VQA model can filter less relevant frames thereby allowing the visual pipeline to focus more on the relevant frames. Detecting objects in a frame is seminal in localizing moments in a video. The idea behind our novel Visual Concept Localizer (VCL) is that objects in a frame have a relationship with the words in the input question. In this work, we experimented with GloVe embeddings to compute such relationships. GloVe embeddings rely on the co-occurrence between pairs of words and can be represented as 25, 50, 100 or 200 dimensional vectors. In our model, we employed 50 dimensional GloVe vectors as it was observed that using lengthier dimensional vectors did not boost performance significantly, rather,

it added computational overhead.

Figure 2 shows the design of the VCL. For a given video and a question, the VCL takes one frame,  $f$ , and the given input question,  $q$ . The frame is passed through a FasterRCNN block which gives a list of detected objects. The top  $K$  detected objects are stored and in our empirical experiments we found that  $K = 10$  is a helpful choice. The list of detected objects  $D$  is fed into a GloVe embedder module which generates a matrix  $\mathbf{G}_D$  of size  $n_d \times 50$  where  $n_d$  is the number of detected objects. The question is tokenized and passed through the GloVe embedder module as well to get a matrix  $\mathbf{G}_Q$  of size  $n_q \times 50$  where  $n_q$  is the number of tokens in the question. Next, we compute the dot product of  $\mathbf{G}_D$  and  $\mathbf{G}_Q$  to get a frame relevance score matrix  $\mathbf{R}$ . The reasoning behind this stems from the fact that the dot product of a pair of GloVe vectors is directly proportional to the probability of the pair of words occurring in the same context (Pennington et al., 2014), (Link).

The relevance score,  $r_f$ , of frame  $f$  is represented by the sum of the matrix  $\mathbf{R}$ .

The relevance score computation pipeline is run for all the frames in the video that results in a relevance score list  $L_R$ . To obtain the most relevant frames, we slide a window across the relevance score list and retrieve the set of frames whose total relevance score is the maximum across all windows. The hypothesis is that the likelihood of the answer being in these set of frames is deemed to be high. Hence, these relevant frames are used as inputs to the transformer module of our architecture.

### 5.2.2 Inputs and Feature Extraction

The TVQA+ dataset consists of a single video clip per example. Each video clip from the TVQA+ dataset consists of at-most 300 frames. First, we obtain the video frames which are most relevant for answering the given question through the VCL module.

After obtaining a smaller subset of video frames using the VCL module, we perform ResNet feature extraction on these frames. The features are obtained by using a ResNet-101 model pretrained on the ImageNet dataset. The ResNet-101 model is given the video frame as input and it outputs a 2048-dimensional embedding, obtained from the pool5 layer.

We also used the visual concepts extracted using a FasterRCNN block which gives a list of detected objects per video frame. These visual concepts are

used in two parts of our system. They are passed to our visual concept localizer, which generates a smaller subset of video frames that contain objects similar to those in the question. We then retain the visual concepts from the smaller subset of video frames only. The first set of text features are obtained by concatenating the subtitle text with the visual concept features with a [SEP] token and passing the result through a BERT tokenizer.

As we are dealing with a multiple-choice answering task, every answer choice is concatenated with the question, separated by a [SEP] token, and treated as a separate input. This concatenated text is passed through a BERT tokenizer to obtain the second set of text features. Note that for each QA pair, our model runs forward propagation 5 times, once for every answer choice. Finally, the extracted Resnet-101 features, tokenized Subtitles + Visual concepts text, and the tokenized Question + Answer text is passed to the VisualBERT model transformer.

### 5.2.3 VisualBERT Transformer

We use a pre-trained VisualBert model (Li et al., 2019) from Hugging-Face as our Multi-Head Transformer Block. Except the 12<sup>th</sup> head and the output pooler layer, we freeze the weights of the VisualBert model. This is primarily done as we have limited GPU-RAM compute.

Then, we use the [CLS] vector output of the VisualBERT Transformer to obtain a single cross-modal embedding per answer choice. The [CLS] vector is chosen, as it gets computed using self-attention and retains the relevant information of the rest of the hidden states. In addition, our experiments show that the [CLS] vector gives better results, when compared to doing max-pooling across time and taking the VisualBERT-pooler layer output.

### 5.2.4 Feed Forward Network

The cross-modal embedding obtained from the output of the VisualBERT Transformer is fed through a feed-forward neural network to obtain a single scalar score per answer. The network had a depth of 2, with 256 neurons in the first layer, and 64 neurons in the second layer. Batch-normalization and Dropout were applied to help the model generalize well to the Test dataset. The GeLU activation was used in this module.

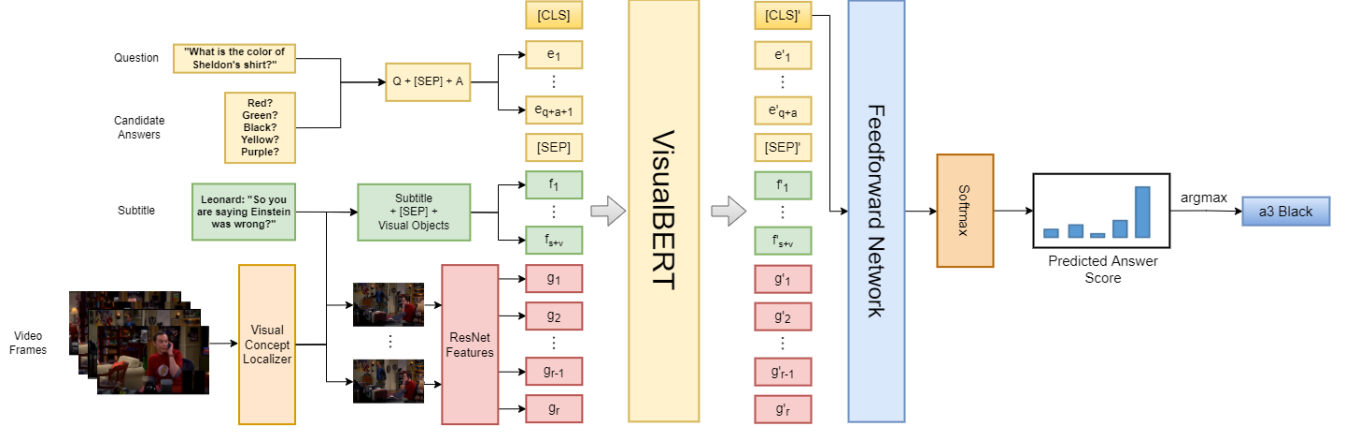


Figure 3: Visual Concept Localizer and VisBERT based VQA System

### 5.2.5 Loss Function

A single scalar score is obtained per answer choice after running the pipeline described above. We run the pipeline 5 times per answer choice to get a 5-D logit vector. The Logit-Vector is passed through a Cross-Entropy block to obtain the loss. The Softmax scores are used to determine the most likely answer choice.

### 5.2.6 Hyperparameters and their effects

We discuss the effect of hyperparameters on the proposed VisualBERT + VCL module. The results can be observed in Table 2.

Our default setup was the Visual BERT model with (a) VCL GloVe Dimension = 50, (b) VCL Window Size = 10, (c) The Adam Optimizer (d) No BatchNorm (e) No Dropout.

Unless specified, these settings were used in the discussion below.

- **VCL GloVe Dimension:** The 50 –  $D$  GloVe vector performed better than a 100 –  $D$  GloVe vector by 2.68%. Hence, we fixed the GloVe Dimension to 50 for the remaining experiments
- **Batch Normalization:** We observed that including Batch Normalization improved our performance by 1.82%. This is intuitive as batch normalization reduces internal covariate shift, and makes the model less volatile to hyperparameters like the initial learning rate. Hence, we decided to use Batch Normalization for the rest of our experiments.
- **VCL Window Size:** We experimented with window sizes of length 2, 4, and 10 with Batch Normalization. We observed a window size

of 10 gave the best performance. We hypothesise that a very small window will fail to capture important information that may have been detected otherwise. While we know a VCL window size of 10 performed better than the case where we do not employ VCL windowing, a higher VCL window size may prove to be more optimal than window size of 10, and we intend to tune this parameter further going forward. We use the VCL Size of 10 for the rest of our experiments

- **Optimizer:** We tried the Adam Optimizer and RAdam optimizer in our experiments with Batch Normalization. We see that the RAdam optimizer has a 2.58% gain over the Adam optimizer. RAdam is a variant of the Adam optimizer that seeks to tackle Adam’s convergence problem by introducing a term to rectify the variance of the adaptive learning rate. A notable benefit of using RAdam it makes the model performance independent of the initial learning rate, which would explain the better performance. Hence, we use the RAdam optimizer for further experiments.
- **Dropout:** For this experiment, we employ Batch Norm and use the RAdam optimizer. We added a dropout of 0.3 in the Feed Forward Neural Network. We note that including dropout reduced the accuracy by 0.66%. We reason this is the case as Batch Normalization already has a regularizing effect on the network, and hence adding dropout does not have a significant effect.



## 6 Results

Table 2 shows the performance of the TVQA baseline (Lei et al., 2018), our unimodal and multimodal baseline models along with the final proposed model and its variants. As shown in Table 2, our unimodal method of simply using the question and answer pairs achieves an accuracy of 41.22 % on the validation dataset. On enabling the video modality, we observed a 2.25 % improvement in the validation accuracy. Interestingly, on incorporating the subtitle modality along with the question-answer pairs improved the validation accuracy by 4.96 % which shows a strong influence of subtitles on the model’s performance. This indicates that in both the cases there is an improvement in accuracy when either the video or the subtitle modalities are included in addition to the question-answer pairs to the model. This motivated the idea of leveraging the information from both of these video and subtitle modalities into our final proposed model.

To determine how the information from these different modalities is to be incorporated, we experimented with three different simple multimodal architectures whose results are shown. While we employed ResNet101 video feature extraction pipeline in all three architectures, since the subtitle modality had a strong influence on the model, the primary focus here was to experiment with different text embeddings mechanisms to get the cross modal contextualized embeddings which would result in best validation accuracy. Thus, as seen in Table 2, the multimodal model with GLoVE word embeddings resulted in poor validation accuracy whereas on switching GloVe with a BERT encoder, we noticed an improvement of 11.48 % over GLoVE. Our understanding is that unlike GloVe, the BERT text encoder generates context rich embeddings thereby, giving the model a significant boost in performance.

In the baseline models, BERT was used purely as a feature extractor for text, with all the attention layers frozen. However, for our final model, we employ the VisualBERT Transformer, with unfrozen model parameters which enables it to tune itself to our task and learn the modality relationships.

Taking reference from works such as UNITER and VisualBERT, instead of finding the textual embeddings from BERT and visual representations from ResNet101 separately and then employing different fusion techniques for their alignment, we used a single Multi head transformer architecture

based on VisualBERT to learn the joint cross modal contextualized embeddings between the vision and text modality. This Vanilla VisualBERT approach surpassed all our previous results achieving an improvement of 6.76 % over our previous best multimodal model. The main benefit of this approach is that the self-attention mechanism within the Transformer implicitly aligns the elements of the input text and regions in the input image.

A key component behind the improvement over the baseline architectures is the introduction of the Visual Concept Localizer (VCL). The VCL module helps in retrieving the most relevant frames that are required for answering a question. This is predominantly effective in cases where visual cues are of primary importance. The VCL module uses Glove embeddings to compute the relevance between objects in a frame and the input question. We found that using higher dimensional Glove vectors did not lead to any considerable improvement in the validation accuracy, rather, it hurt the model efficiency. The VCL integration boosted the validation accuracy to 57.6 % leading to a 5.65 % improvement over the vanilla VisualBERT model.

We note that the TVQA baseline (Lei et al., 2018) has a better performance than our proposed model. The TVQA baseline model was designed specifically for the TVQA/TVQA+ VQA task, whereas our model applies transfer learning on a pre-trained VisualBert transformer for the VQA task. We hypothesize that with better fine-tuning, and implementation of our proposed extensions, we will potentially be able to match the TVQA baseline model accuracy.

Figures 4 and 5 show the Training and Validation Accuracy trends of the proposed model and its variants versus the number of epochs.

## 7 Analysis

### 7.1 Intrinsic Metrics

Problems that require cross modal interactions are associated with sub-problems that need to be addressed in order to tackle the main task at large. In this section, we represent our analysis on our final model’s performance on various scenarios for both successful and failure cases. The insights garnered from this analysis is the foundation for future work.

We investigate the performance of our model on all “wh” questions as shown in the Table 3. For all kinds of wh questions, our proposed model beats all the baselines we have run our experiments on.

Table 2: Validation Accuracies for different models on the TVQA+ dataset (Lei et al., 2019). The top 2 model performances are in bold. “-” indicates Not Applicable. **Legend- UB:** Unimodal Baseline, **QA:** Question Answer, **V:** Video, **S:** Subtitles, **MMB:** Multimodal Baseline, **ResNet:** ResNet video embeddings, **GloVe:** GloVe text embeddings, **BERT:** BERT text embeddings, **SA:** Self Attention, **Linear:** Feed Forward Network, **VB:** VisualBERT, **VCL:** Visual Concept Localizer, **BN:** Batch Normalization, **RAdam:** RAdam optimizer, **TVQA:** TVQA baseline model

Methods	VCL GloVe Dim	VCL Window Size	Val. Acc. %	Delta
UB - QA	-	-	41.22	-10.73
UB - QA + V	-	-	43.47	-8.48
UB - QA + S	-	-	46.16	-5.79
MMB - GloVe + ResNet + SA	-	-	33.71	-18.24
MMB - BERT + ResNet + SA	-	-	45.19	-6.76
MMB - BERT + ResNet + Linear	-	-	42.85	-9.10
Vanilla VB	-	-	51.95	0.00
VB VCL10	GloVe-50d	10	55.88	+3.93
	GloVe-100d		53.20	+1.25
VB BN VCLx	GloVe-50d	2	54.46	+2.51
		4	54.82	+2.87
		10	55.02	+3.07
<b>VB BN RAdam VCL10</b>	<b>GloVe-50d</b>	<b>10</b>	<b>57.60</b>	<b>+5.65</b>
VB BN RAdam VCL10 + dropout	GloVe-50d	10	56.94	+4.99
<b>TVQA (Lei et al., 2018)</b>	-	-	<b>66.39</b>	<b>+14.44</b>

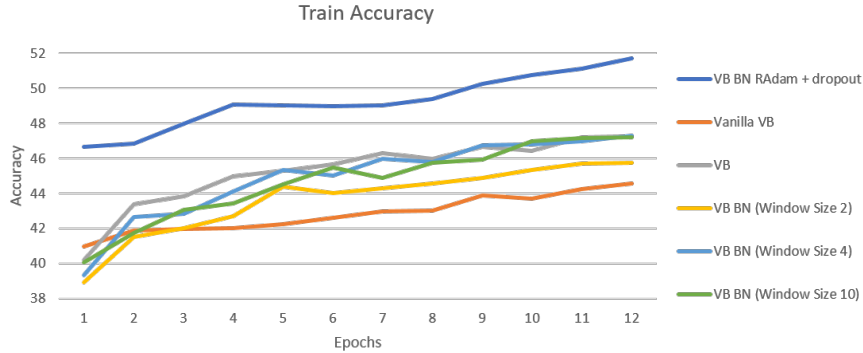


Figure 4: Train Accuracy V/S Epoch for different models. **Legend- VB:** VisualBERT, **RAdam:** RAdam optimizer, **BN:** Batch Normalization, **Window Size:** VCL Window Size

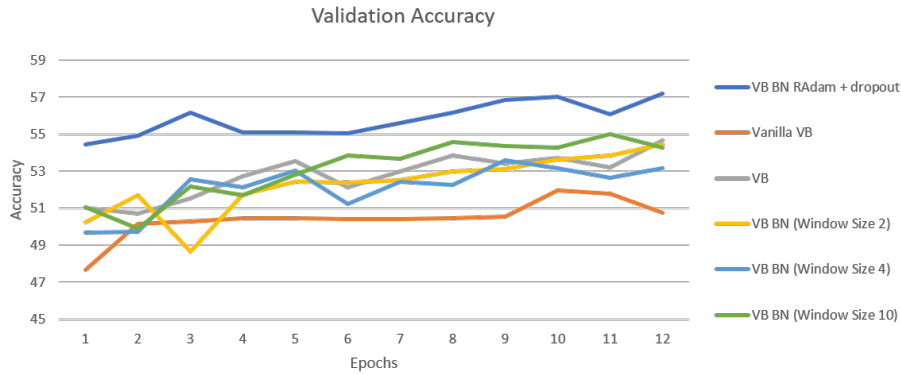


Figure 5: Validation Accuracy V/S Epoch for different models. **Legend- VB:** VisualBERT, **RAdam:** RAdam optimizer, **BN:** Batch Normalization, **Window Size:** VCL Window Size

Table 3: Model Analysis on **wh** questions on the TVQA+ dataset (Lei et al., 2019). The top 2 model performances are in bold. **Legend-** **QA:** Question Answer, **V:** Video, **S:** Subtitle, **GloVe:** GloVe embeddings, **BERT:** BERT embeddings, **ResNet:** ResNet embeddings, **SA:** Self Attention, **Linear:** Feed Forward Network, **VB:** VisualBERT, **VCL:** Visual Concept Localizer, **BN:** Batch Normalization, **RAdam:** RAdam optimizer, **TVQA:** TVQA baseline model

Model	Overall Val Acc %	% of “wh” questions answered correctly				
		“who”	“where”	“how”	“what”	“why”
QA only	41.22	34.62	38.01	48.71	44.74	54.16
QA + V	43.47	28.47	40.06	48.71	43.42	53.47
QA + S	46.16	35.27	39.72	46.88	47.48	57.63
GloVe + ResNet + SA	33.71	24.91	27.39	39.19	33.84	49.65
BERT + ResNet + SA	45.19	33.65	40.75	50.91	45.67	56.59
BERT + ResNet + Linear	42.85	27.83	37.67	46.15	43.75	55.2
Vanilla VB	52.07	42.39	42.28	54.94	52.84	59.02
<b>VB BN RAdam VCL10</b>	<b>57.60</b>	<b>48.54</b>	<b>52.39</b>	<b>55.67</b>	<b>58.21</b>	<b>71.87</b>
<b>TVQA (Lei et al., 2018)</b>	<b>66.39</b>	<b>65.04</b>	<b>52.05</b>	<b>67.39</b>	<b>66.15</b>	<b>82.98</b>

Table 4: Model Analysis on predefined attributes on the TVQA+ dataset (Lei et al., 2019). The top 2 model performances are in bold. **Legend-** **QA:** Question Answer, **V:** Video, **S:** Subtitle, **GloVe:** GloVe embeddings, **BERT:** BERT embeddings, **ResNet:** ResNet embeddings, **SA:** Self Attention, **Linear:** Feed Forward Network, **VB:** VisualBERT, **VCL:** Visual Concept Localizer, **BN:** Batch Normalization, **RAdam:** RAdam optimizer, **TVQA:** TVQA baseline model

Model	Overall Val Acc %	% Val Acc on Predefined Attributes			
		Color	Dialog	Longest ans	Shortest ans
QA only	41.22	<b>38.48</b>	46.48	21.51	3.94
QA + V	43.47	34.48	48.8	23.36	3.18
QA + S	46.16	32.76	51.48	23.76	4.01
GloVe + ResNet + SA	33.71	18.97	39.24	30.46	0.00
BERT + ResNet + SA	45.19	29.31	49.86	23.73	3.54
BERT + ResNet + Linear	42.85	32.76	47.82	23.93	2.71
Vanilla VB	52.07	32.76	58.86	22.80	5.23
<b>VB BN RAdam VCL10</b>	<b>57.60</b>	<b>32.76</b>	<b>65.68</b>	<b>25.48</b>	<b>6.26</b>
<b>TVQA (Lei et al., 2018)</b>	<b>66.39</b>	31.03	<b>76.51</b>	<b>5.50</b>	<b>1.50</b>

However, it is quite interesting to note the gradual progression in the accuracy of individual question cases with different configurations of the model. “Who” questions contribute to 10 % of the dataset. Switching to a transformer based approach gave an improvement of 6.76 % in the validation accuracy. In our final model, we execute early fusion of modalities using a Visual BERT encoder by discarding our previous technique that used late fusion. In addition, we also observed a 5.65 % improvement over vanilla Visual BERT encoder model by integrating the Visual Concept Localizer module. “Who” questions require the identification of people, thus, there is a strong reliance on visual features for such questions. For “where” questions, which comprise 10 % of the dataset, the understanding of objects and their positional semantics in the frame is necessary. For the unimodal baselines, we can see that the video only baseline has the highest validation accuracy for “where” questions among all the modalities. While Vanilla VB model brought an improvement of 2 % in where questions, it was the introduction of VCL that pushed the accuracy by 12 %. A similar trend is also evident for “what” questions. Interestingly, we observed that almost all the models had more than 50 % validation accuracy for “why” questions. In fact, our final model performed exceedingly well on “why” questions with 71.87 % accuracy, especially with the VCL integration. However, there was minute improvement over the VB baseline for “how” questions indicating that the model lacks in mapping the reasoning behind certain events mentioned in the question. A similar pattern of validation accuracies is observed for the TVQA model.

In Table 4, we analyze some specific attributes found in the input questions. In (Lei et al., 2018), it is stated that there is a possibility that the model learns to pick the longest answer candidate. This is an undesirable trait as this adds a bias towards the length of an answer candidate instead of attending to the semantics of the text and visual features. In our experiments, we found that our proposed model picked the longest option as the answer 25 % of the time. From the Table 4, it is evident that the Glove+ResNet+Self Attention model is greatly influenced by the answer length. Similarly, although rare, some models, including the final model, nevertheless get biased towards the shortest answer. Debiasing the models against the length of the candidate answer is important to protect the

model performance. However, it is interesting to note that the TVQA model is not influenced by the length of the candidate answers as only 5.5 % of the predictions comprise of the longest candidate answer. This is a limitation of our model that needs further investigation.

Our model performs poorly on questions that ask about an object’s color. From the Table 4, it is clear that the sophisticated models including the VCL integration fail to recognize color. However, the distribution of the data is such that it works against “color” based questions as only 1.9 % of the dataset is composed of “color” questions. This is a strong indicator for further investigation in the visual space with improvements in the visual feature extraction pipeline. Moreover, our model is not robust to questions that have negation. The distribution of the TVQA dataset is such that only 2 % of the questions have negation. In addition, negation questions heavily rely on the text modality. In fact, the VCL module might just add noise for such questions and this is a failure case of our model. A data augmentation strategy for this is to convert some questions into negation and re-labelling the answer options. We hope to integrate this technique in our model.

On the other hand, questions which seek what one character spoke are handled well by our model with a validation accuracy of 65.68 % out of all such questions. This stems from the fact that the subtitle modality is strong enough for the model to map the relationship between characters and their speech. In fact, the unimodal baseline with just the subtitle has a validation accuracy of 51.48 % for such questions. Introduction of visual BERT and VCL led to an improvement of 15 % over the QA+S baseline.

## 7.2 Qualitative Analysis and Examples

A detailed Qualitative Comparison between proposed Visual BERT + Visual Concept Localizer (VBVCL) model and the TVQA baseline (Lei et al., 2018) is described in Figure 6. The first two examples illustrate scenarios where the VBVCL model got the prediction right, whereas the TVQA baseline got the prediction wrong. Similarly, the last two examples describe situations where the TVQA model got the prediction right, but the VBVCL model got the prediction wrong. Finally, a detailed description detailing the reasoning of every outcome is also presented.







Frames	S, Q, A	Reasoning
 <p><b>Sheldon</b> : I 'm surprised you struck out with Penny . Apparently , she 's a big old five.</p>	<p>What does Penny have in her hands after she enters the apartment ?</p> <p>A0 a notebook A1 An egg A2 nothing <b>A3 Her laptop</b> <span style="background-color: yellow;">VBVCL</span> A4 an envelope <span style="background-color: lightblue;">TVQA</span></p>	<p>In this example, the VCL module is able to identify the most relevant set of frames. In the sample frame, a laptop is visible and is successfully detected by the VBVCL. On the other hand, the TVQA module confuses the object with an envelope.</p>
 <p><b>Penny</b> : Well , it was n't the first suggestion that came to mind .</p>	<p>What does Penny put on the couch after Sheldon is sitting on it ?</p> <p>A0 a soda A1 a couch cover <span style="background-color: lightblue;">TVQA</span> <b>A2 a pillow and blanket</b> <span style="background-color: yellow;">VBVCL</span> A3 a plate with a hamburger A4 a puppy</p>	<p>In the frame Penny is holding a pillow and a blanket. This moment is successfully localized by the VBVCL leading to the correct answer. However, it appears that TVQA deemed it to be a couch cover. Note that the word “couch” is present in the question as well.</p>
 <p><b>Amy</b> : You sure ? The nose does n't light up .</p>	<p>What is Amy holding when she suggest they go to the cadaver lab at UCLA ?</p> <p>A0 A green book <b>A1 A blue container</b> <span style="background-color: lightblue;">TVQA</span> <b>A2 A yellow bag</b> <span style="background-color: yellow;">VBVCL</span> A3 A white shirt A4 A brown mug</p>	<p>This is a failure case. In this example, the model needs to identify the moment when Amy “suggests” that they go to the lab. This is a case where moment localization depends heavily on the text modality. The VBVCL model fails to identify the most relevant frames and leads to an incorrect prediction. TVQA successfully localizes the moment based on text, and identifies the blue container as well.</p>
 <p><b>Leonard</b> : Well , who wants to stay in a hotel ?</p>	<p>What does Leonard pick up after Sheldon talks about scented candles with Elizabeth ?</p> <p><b>A0 A book</b> <span style="background-color: lightblue;">TVQA</span> A1 A poster A2 An action figure <b>A3 A laptop</b> <span style="background-color: yellow;">VBVCL</span> A4 A newspaper</p>	<p>This is a failure case. Similar to the previous example, TVQA does better by localizing the moment when Sheldon finishes talking to Elizabeth. It then correctly identifies Leonard and the book. The VBVCL model seems to have confused the book with a laptop.</p>

Figure 6: Qualitative Comparison between proposed VBVCL model and the TVQA baseline (Lei et al., 2018). **Legend:** (a) The **green** highlight indicates the right answer choice. (b) The **red** highlight indicates the **VBVCL** model got the prediction wrong. Note that nothing was highlighted if the TVQA baseline model got the prediction wrong. (c) **TVQA** indicates the prediction by the TVQA baseline. (d) **VBVCL**: VBVCL indicates the prediction by the proposed VBVCL model

## 8 Future work and Limitations

### 8.1 Limitations of proposed model

We evaluated variants of the VisualBERT model on the TVQA+ dataset. Unlike its predecessor, TVQA+ only contains visuals and dialogue from one TV show, as opposed to six shows in TVQA. This can be seen as a limitation of both the data and the model. Lack of diversity in the training data might affect the model’s ability to generalize on previously unobserved environments and text.

A problem with the question answering task is that it can be difficult to evaluate the model’s performance when the prediction does not match the answer exactly. Additionally, all answers choices are evaluated using the cross-entropy loss which looks at only binary outputs (whether the prediction belongs to the class or not).

We believe that the task can be improved by treating it as an answer generation problem. Using this approach, even if the model prediction does not match the correct answer, it does not get strictly penalized. Another improvement that can be made to the task is to improve the diversity of the dataset, and make it more visually dependent. As shown in (Lei et al., 2018), the TVQA dataset suffers from subtitle bias. Adding more visually dependent questions would help mitigate this issue.

Since the VCL looks at the aggregate score for a window of frames, it fails in cases where the frame with the maximum relevance is not present within the window with the highest collective score. This is possible if the camera quickly pans out of the relevant frame to a completely non-relevant frame for a short period.

### 8.2 Future Work

The Visual Concept Localizer has proved to be helpful for the vanilla Visual BERT model. The motivation behind the VCL is that the module should filter out frames that do not hold much relevance to the question being asked. VCL is helpful in cases where the question seeks the answers to objects in a set of frames. However, some questions solely rely on words spoken by the characters. In such a scenario, the VCL tends to diverge the model by feeding the frames which it perceives to be relevant but are actually noise to the model. In the next iterations of the VCL, the goal would be to introduce robustness to the module. While the VCL is fast, efficient and easily pluggable to any architecture, we believe that it can be improved by introducing an

auxiliary loss for it. The loss will be determined by comparing the set of output relevant frames and the actual set of frames that contains the answer to a given question. This additional loss would help the VCL learn a better mapping between the question and video frames. Weighing the frames and their corresponding relevance score is a future task. The current model caps the number of detected objects to 10. This is done in order to debias the network against the number of objects as it can result in a pseudo-frame-relevance effect where the model attends to noisy frames. However, frames having less number of objects can still get filtered out. Hence, it is logical to weight the frames based on the detected objects and words in the question as well. While VCL contributed to considerable improvement in the overall model, it would be interesting to see its effect on other VQA architectures.

It is also observed that visual question answering systems, including our model and the TVQA model, struggle to map faces to their names through text and visual features. This is a sub-task that can either be handled as a separate problem with another sub-module or can be integrated into the end-to-end pipeline using another auxiliary loss. The objective is for the model to learn the mapping between names and faces. For example, for the question “What did Sheldon pick after talking to Penny?” , the model needs to process two pieces of information; identifying Sheldon and Penny and the sequence of events i.e. focusing on Sheldon “after” he talks to Penny.

This brings the next drawback of our model wherein temporal information needs further work. The Visual BERT module has the potential of learning the temporal semantics in the question, subtitles and visual features, however, it needs further refinement in the form of unfreezing more layers, pre-training for sub-tasks and additional temporal feature extraction.

Another challenge faced by VQA models is that they tend to get biased towards the longest answer option which is evident in the Table 4. As shown in the table, attributes such as color and size are hard to identify, however, there are very few data points that are related to such attributes. For the model to learn this skill, additional data is required. The entire dataset has only 2 % of the questions which fall under negation. A data augmentation strategy to tackle this is to convert affirmative questions to negation and relabelling the candidate answers.

Finally, due to the fact that the model is trained on the The Big Bang Theory series, it is biased towards environments which are associated with US households and settings. Experimenting with more TV shows or movies from diverse environments is another task to be addressed in the future.

## 9 Ethical Concerns and Considerations

1. **Bias:** Working on VQA tasks involving people brings up concerns on the gender and racial bias present in the dataset and the model. Particularly for the TVQA dataset, it can be noted that all six TV shows are American and prominently featuring caucasian actors. Due to this bias in the data, the model may not perform as well when people of different races are present in the video, which can be problematic when deployed in real-world scenarios. It is also possible that the model may learn features that discriminate on the basis of genders. For example, it may learn to associate certain words and visual settings with certain genders, which could have a negative affect when it is selecting an answer.
2. **Transparency:** Another issue with using deep learning models such as VisualBERT is the lack of transparency. Explaining why particular answers were chosen by black-box models is a difficult task. While the attention layers help us understand what regions and words the models focus on, they only provide the reasoning for part of the whole architecture. Without an explanation mechanism, the model becomes unreliable, and it cannot be used for tasks where reliability is critical. This is true for visual question answering applications which involve assisting the visually-challenged.
3. **Mitigation** Bias in the machine learning model can only be mitigated through training on unbiased data. For our system, this can be done by adding samples into the model that can make the dataset less biased. We believe that adding in more TV shows with diverse genre would help in mitigation, but since genres themselves contain tropes, this problem may not be completely solvable for this dataset. Transparency in the model can be implemented by looking at attention weights in the self-attention blocks. The self attention

blocks would tell us which visual and text features contributed most towards the answer. Another proposed approach from us would be to have a modality-weighting module, which tells us how much each modality contributed towards the answer. This module could be trained alongside the main model and treated as an auxiliary loss.

## 10 Team member contributions

**Bhavuk Sharma** Experimented with different architectures, hyperparameter tuning, analyzed limitations of proposed model, ethical concerns, freezing/unfreezing visBERT weights

**Debayan Ghosh** Experimented with Visual BERT and its variants, integrated visualization and run scripts, hyperparameter experiments, formulated tokenization of inputs, freezing/unfreezing visual BERT weights

**Rituraj Singh** Ideated Visual Concept Localizer, integrated VCL with various model architectures, analyzed predictions, hyperparameter experiments, data storage and fetch pipeline, visualization scripts

**Saahithya Yendluri** Ideated Visual Concept Localizer, Modularized model pipelines, data preprocessing, hyperparameter tuning, analyzed predictions,

**Wallace Dalmat** Experimented with VisBERT, formulated tokenization of inputs, effect of freezing/unfreezing Visual BERT weights, analyzed predictions

## References

- Huda AlAmri, Vincent Cartillier, Abhishek Das, Jue Wang, Stefan Lee, Peter Anderson, Irfan Essa, Devi Parikh, Dhruv Batra, Anoop Cherian, Tim K. Marks, and Chiori Hori. 2019. [Audio-visual scene-aware dialog](#). *CoRR*, abs/1901.09107.
- Silvio Barra, Carmen Bisogni, Maria De Marsico, and Stefano Ricciardi. 2021. [Visual question answering: which investigated applications?](#)
- Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Nieves. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Santiago Castro, Mahmoud Azab, Jonathan Stroud, Cristina Noujaim, Ruoyao Wang, Jia Deng, and Rada



- Mihalcea. 2020. [LifeQA: A real-life dataset for video question answering](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4352–4358, Marseille, France. European Language Resources Association.
- Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. [UNITER: learning universal image-text representations](#). *CoRR*, abs/1909.11740.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. [BERT: pre-training of deep bidirectional transformers for language understanding](#). *CoRR*, abs/1810.04805.
- Deniz Engin, Yannis Avrithis, Ngoc Q. K. Duong, and François Schnitzler. 2021. [On the hidden treasure of dialog in video question answering](#). *CoRR*, abs/2103.14517.
- Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. [Motion-appearance co-memory networks for video question answering](#). *CoRR*, abs/1803.10906.
- Shijie Geng, Ji Zhang, Zuohui Fu, Peng Gao, Hang Zhang, and Gerard de Melo. 2020. [Character matters: Video story understanding with character-aware relations](#). *CoRR*, abs/2005.08646.
- Zhicheng Guo, Jiaxuan Zhao, Licheng Jiao, Xu Liu, and Fang Liu. 2021. [A universal quaternion hypergraph network for multimodal video question answering](#). *IEEE Transactions on Multimedia*, pages 1–1.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. [Deep residual learning for image recognition](#). *CoRR*, abs/1512.03385.
- Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. [Localizing moments in video with natural language](#). *CoRR*, abs/1708.01641.
- Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. [TGIF-QA: toward spatio-temporal reasoning in visual question answering](#). *CoRR*, abs/1704.04497.
- Aisha Urooj Khan, Amir Mazaheri, Niels da Victoria Lobo, and Mubarak Shah. 2020. [MMFT-BERT: multimodal fusion transformer with BERT encodings for visual question answering](#). *CoRR*, abs/2010.14095.
- Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D. Yoo. 2019a. [Gaining extra supervision via multi-task learning for multi-modal video question answering](#). *CoRR*, abs/1905.13540.
- Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang Dong Yoo. 2019b. [Progressive attention memory network for movie story question answering](#). *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8329–8338.
- Junyeong Kim, Minuk Ma, Trung X. Pham, Kyungsu Kim, and Chang D. Yoo. 2020. [Modality shifting attention network for multi-modal video question answering](#). *CoRR*, abs/2007.02036.
- Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. [Tvqa: Localized, compositional video question answering](#). In *EMNLP*.
- Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2019. [Tvqa+: Spatio-temporal grounding for video question answering](#). In *Tech Report, arXiv*.
- Fangtao Li, Ting Bai, Chenyu Cao, Zihe Liu, Chenghao Yan, and Bin Wu. 2021. [Relation-aware hierarchical attention framework for video question answering](#). *CoRR*, abs/2105.06160.
- Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. [Visualbert: A simple and performant baseline for vision and language](#). *CoRR*, abs/1908.03557.
- Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. [Hierarchical question-image co-attention for visual question answering](#). *CoRR*, abs/1606.00061.
- Mateusz Malinowski and Mario Fritz. 2014. [Towards a visual turing challenge](#).
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [Glove: Global vectors for word representation](#). In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hananeh Hajishirzi, and Jonathan Berant. 2021. [Multimodalqa: Complex question answering over text, tables and images](#). *CoRR*, abs/2104.06039.
- Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. [Movieqa: Understanding stories in movies through question-answering](#). *CoRR*, abs/1512.02902.
- Thomas Winterbottom, Sarah Xiao, Alistair McLean, and Noura Al Moubayed. 2020. [On modality bias in the TVQA dataset](#). *CoRR*, abs/2012.10210.
- Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2020a. [Just ask: Learning to answer questions from millions of narrated videos](#). *CoRR*, abs/2012.00451.
- Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. 2020b. [Bert representations for video question answering](#). In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1545–1554.
- Yueting Zhuang, Dejing Xu, Xin Yan, Wenzhuo Cheng, Zhou Zhao, Shiliang Pu, and Jun Xiao. 2020. [Multichannel attention refinement for video question answering](#). *ACM Trans. Multimedia Comput. Commun. Appl.*, 16(1s).