

# 11-777 Report 3: Baselines and Analysis by Team-Samosa

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

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

## 1 Models

The task of Visual Question Answering requires models to intelligently retrieve data from multiple modalities namely, video, subtitle, and question simultaneously to predict the correct answer. The objective of a VQA system is defined as follows; given a video  $V$ , subtitles  $S$ , a question  $q$  and 5 possible answers  $\{a_i\}_0^4$ , the model should pick the correct answer. In this report, we describe several uni-modal and multi-modal baselines and replicate prior work results to analyze the performance of various state-of-the-art approaches and effect of different modalities on answer prediction. Moreover, we report our observations on choosing different approaches and the trade-offs to build an efficient architecture. The following models are implemented on the TVQA(Lei et al., 2018)/TVQA+(Lei et al., 2019) datasets. Section 1 describes the 6 models we implemented and 3 prior works. Section 2 reports the accuracy for all the 9 models that were tested, and the changes in training and validation accuracies, the loss and the learning rates with respect to the training steps. In section 3, we perform an analysis of two intrinsic metrics relevant to the task of VQA and illustrate example outputs for our implementations. Section 4 highlights individual contributions to the project.

### 1.1 Unimodal Baselines

The foundation of a robust multimodal solution lies with strong analysis of unimodal experiments and gathering evidence to create a multimodal hypothesis. In this section, we present three unimodal architectures and their respective analysis. At a bird's eye view, we built 3 models, each having one primary modality; video, subtitles or just the question-answer (QA) pairs.

#### 1. Video Unimodal: For the Video Unimodal

\*Everyone Contributed Equally – Alphabetical order

baseline in figure 1, given a video frame  $V$ , a question  $Q$  and candidate answer choices  $\{a_i\}_0^4$ , we first get the BERT (Devlin et al., 2018) embeddings for the question  $Q \in \mathbb{R}^{seq\_len_Q \times 768}$  and answer choices  $\{a_i\}_0^4 \in \mathbb{R}^{seq\_len_{a_i} \times 768}$  from a BERT model pretrained on the Toronto Book and Wikipedia corpus. The video features  $V \in \mathbb{R}^{n\_frames \times 2048}$  are extracted using the ImageNet pretrained ResNet-101(He et al., 2015) model (at pool5 layer) on the TVQA(Lei et al., 2018) dataset. The BERT embeddings of the question and answer choices are projected to a lower 256 dimension using a Linear layer with GELU(Hendrycks and Gimpel, 2016) activation to get hidden representations as  $H_Q \in \mathbb{R}^{seq\_len_Q \times 256}$  and  $H_{a_i} \in \mathbb{R}^{seq\_len_{a_i} \times 256}$ . 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. The 2048 dimensional video representations are also projected down to 256 dimensions by passing them through a convolution layer with batch-normalization and LeakyRELU activation and finally applying a Maxpool operation in the last two dimensions to get hidden video representations as  $H_V \in \mathbb{R}^{\frac{n\_frames}{2} \times 256}$ . The hidden representations are fused by concatenating them for each answer choices to get fused features as  $G_i \in \mathbb{R}^{seq\_fused \times 256}$ , for each answer choice  $a_i$ . We finally use a classification layer comprised of linear layers and GELU(Hendrycks and Gimpel, 2016) activations to convert fused features into answer probabilities for all the five answers.

#### 2. Subtitles Unimodal: The Subtitle Unimodal baseline in figure 2 uses a subtitle $S$ , a question $Q$ and candidate answer choices

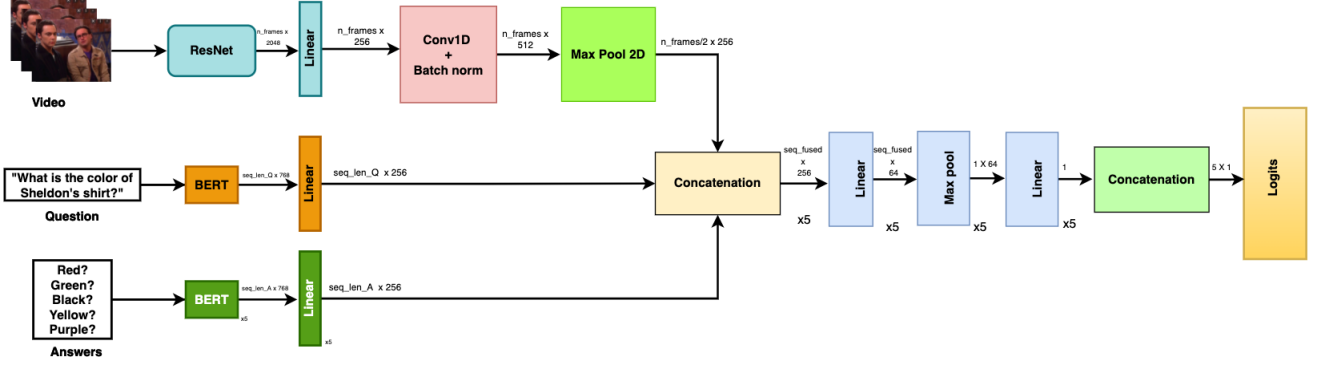


Figure 1: Proposed model for video+QA uni-modal baseline

$\{a_i\}_0^4$  with their 768 dimensional embeddings obtained from a BERT(Devlin et al., 2018) model pretrained on the Toronto and Wikipedia corpus as  $S \in \mathbb{R}^{seq\_len_S \times 768}$ ,  $Q \in \mathbb{R}^{seq\_len_Q \times 768}$  and  $\{a_i\}_0^4 \in \mathbb{R}^{seq\_len_{a_i} \times 768}$ . These embeddings are projected to a lower 256 dimension and concatenated along the first dimension for each of the answer choices to get hidden representations as  $H_i \in \mathbb{R}^{seq\_fused \times 256}$ . 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 by using a Linear layer with their dimensions as  $q_i \in \mathbb{R}^{seq\_fused \times 64}$ ,  $k_i \in \mathbb{R}^{seq\_fused \times 64}$  and  $v_i \in \mathbb{R}^{seq\_fused \times 64}$ . We then apply self-attention mechanism using the query, key and value triplets to get attended feature representations of dimension  $G_i \in \mathbb{R}^{seq\_fused \times 64}$ . Next, we take the maximum values across the first dimension and finally use a linear layer to obtain the answer probabilities for all the five answers as logits.

3. **QA Unimodal:** Our third unimodal baseline in figure 3, consisted simply of a question  $Q$  and candidate answers  $\{a_i\}_0^4$ . Similar, to the first two approaches we obtained BERT(Devlin et al., 2018) embeddings for the question  $Q \in \mathbb{R}^{seq\_len_Q \times 768}$  and answer choices  $\{a_i\}_0^4 \in \mathbb{R}^{seq\_len_{a_i} \times 768}$  which were projected down to 256 dimension and fused by concatenating to get hidden representations as  $H_i \in \mathbb{R}^{seq\_fused \times 256}$ . We then took the maximum values across the first dimension and finally used a classification layer comprised of

linear layers and GELU(Hendrycks and Gimpel, 2016) activations to get probabilities over the candidate answer choices.

## 1.2 Simple Multimodal Baselines

We created 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 implemented are as follows:

1. **BERT + RESNET with Self-Attention:** In this baseline implementation as seen in figure 4, we first obtained the BERT embeddings for the question  $Q \in \mathbb{R}^{seq\_len_Q \times 768}$ , subtitle  $S \in \mathbb{R}^{seq\_len_S \times 768}$  and candidate answer choices  $\{a_i\}_0^4 \in \mathbb{R}^{seq\_len_{a_i} \times 768}$  as well as extracted the video features  $V \in \mathbb{R}^{n\_frames \times 2048}$  using ImageNet pretrained ResNet-101(He et al., 2015) model. 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 sequence length dimension with each answer choice, to obtain fused features as  $G_i \in \mathbb{R}^{seq\_qas \times 256}$  (where  $seq\_qas = seq\_len_Q + seq\_len_{a_i} + seq\_len_S$ ). Similarly, we get fused features for each answer choice with question and video representations as  $F_i \in \mathbb{R}^{seq\_qav \times 256}$  (where  $seq\_qav = seq\_len_Q + seq\_len_{a_i} + n\_frames$ ). Next, we calculate the query  $qs_i \in \mathbb{R}^{seq\_qas \times 64}$ , key  $ks_i \in \mathbb{R}^{seq\_qas \times 64}$  and value  $vs_i \in \mathbb{R}^{seq\_qas \times 64}$  projections by using a Linear layer for the subtitle fused features

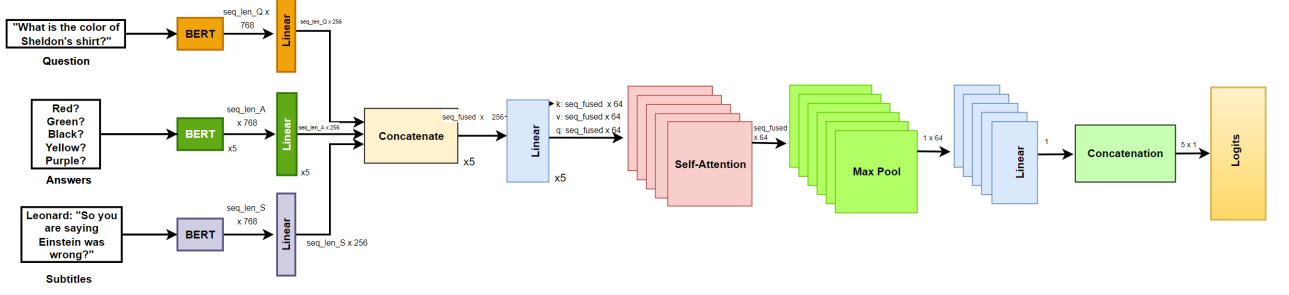


Figure 2: Proposed model for Subtitle+QA uni-modal baseline

as well as for the video fused features as  $qv_i \in \mathbb{R}^{seq\_qav \times 64}$ ,  $and  $vv_i \in \mathbb{R}^{seq\_qav \times 64}$ . Using these query, key and value triplets we apply self-attention to attend regions in the subtitle and video modality necessary for the answer prediction and take the maximum values along the first dimension which are then finally classified using a Linear layer to obtain the answer probabilities.$

## 2. BERT + RESNET without Self-Attention:

As seen in figure 5, similar to the first multi-modal baseline, here we obtain the fused features as  $G_i \in \mathbb{R}^{seq\_qas \times 256}$  (where  $seq\_qas = seq\_len_Q + seq\_len_{a_i} + seq\_len_S$ ) and  $F_i \in \mathbb{R}^{seq\_qav \times 256}$  (where  $seq\_qav = seq\_len_Q + seq\_len_{a_i} + n_{frames}$ ) for the subtitle and video representations respectively with the question and answer choices. However, instead of self-attention we simply use linear layers with GELU(Hendrycks and Gimpel, 2016) activations to further learn the representations and extract the maximum values along the first 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 get the logits with answer probabilities.

- GloVe + RESNET with Self-Attention:** In our third multimodal baseline as seen in figure 6, instead of using BERT we used GloVe(Pennington et al., 2014) to get the word embeddings for the question  $Q \in \mathbb{R}^{seq\_len_Q \times 50}$ , subtitle  $S \in \mathbb{R}^{seq\_len_S \times 50}$  and candidate answer choices  $\{a_i\}_0^4 \in \mathbb{R}^{seq\_len_{a_i} \times 50}$ . The video features  $V \in \mathbb{R}^{n_{frames} \times 2048}$  are extracted in a similar manner to previous methods. The subtitle, ques-

tion and answers embeddings are projected to a higher 256 dimension and video features to lower 256 dimensions with the Linear layers. We then get the fused features from subtitle and question answer set as  $G_i \in \mathbb{R}^{seq\_qas \times 256}$  (where  $seq\_qas = seq\_len_Q + seq\_len_{a_i} + seq\_len_S$ ) and video, question, answer set as  $F_i \in \mathbb{R}^{seq\_qav \times 256}$  (where  $seq\_qav = seq\_len_Q + seq\_len_{a_i} + n_{frames}$ ). Next, we apply self-attention mechanism by calculating and using the query, key and value triplet for attending different regions in the subtitle and video modalities. Finally, to obtain answer probabilities we apply Linear layer on the maximum value obtained from fusing the subtitle and video set representations for each candidate answer.

## 1.3 Competitive Baselines

We performed analysis of the following methods:

- TVQA(Lei et al., 2018):** Lei et al, in their paper implement a multi-stream end-to-end trainable neural network for VQA. The video features are extracted using RCNN whereas textual features are extracted using GloVe(Pennington et al., 2014). Following the feature extraction stage, the model has several bi-directional LSTM's and performs context matching to jointly model the contextual inputs (video/subtitle) with the query (QA pair). Softmax is finally used to get answer scores. The paper reports an accuracy of 66.46% and we were able to replicate these results with a small error to get 64.71%.
- iPerceive(Chadha et al., 2020):** The authors propose a framework that enables common-sense learning to infer causal relationships be-

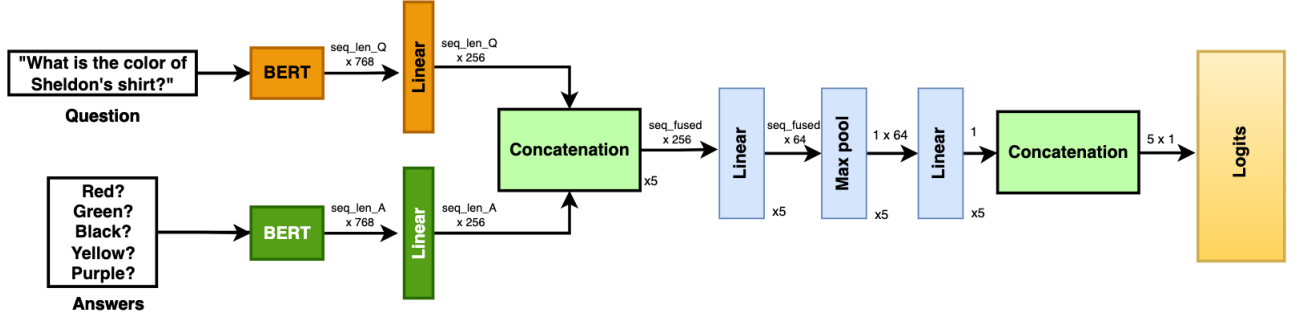


Figure 3: Proposed model for QA only uni-modal baseline

tween objects in a video by building a knowledge base using contextual cues. The model has 3 stages: temporally localize a set of events in a video using CNN and BiLSTM, build a knowledge base for common-sense reasoning via self-supervised representation learning and finally produce a textual description using audio, visual, and speech cues for each event using a transformer based architecture. The reported accuracy of the model is 76.97% but we got an accuracy of 25.32% because the pre-trained model provided by the authors didn't contain the common-sense features and replicating the results would require that the model be trained from scratch.

3. **HERO**(Li et al., 2020): In this paper, Li et al propose HERO (Hierarchical EncoderR for Omni- representation learning) which takes as input a sequence of video clip frames and their accompanying subtitle sentences. Instead of adopting a flat BERT-like encoder, HERO encodes multi-modal inputs in a hierarchical fashion, with a Cross-modal Transformer to fuse a subtitle sentence and its accompanying local video followed by a Temporal transformer to get sequential video embeddings using all surrounding frames as global context. This method achieves 74.24 % accuracy on the TVQA dataset but using the pretrained TVR(Lei et al., 2020) weights given we get an accuracy of 29%.

## 2 Results

Table 1 shows the performance of our unimodal and multimodal models. In addition, we report the performance of open sourced models from prior literature trained on the TVQA dataset. We started

Methods	Validation	
	Acc (%)	Loss
QA	41.22	1.09
QA + V	43.47	0.64
<b>QA + S</b>	<b>46.16</b>	<b>0.12</b>
GloVE + ResNet + SA	33.71	1.23
BERT + ResNet + SA	45.19	0.74
BERT + ResNet + Linear	42.85	0.93
iPerceive	25.32	-
HERO	29	-
<b>TVQA</b>	<b>64.71</b>	-

Table 1: Model Analysis

the unimodal analysis by first using just the question and answers pairs and obtained a validation accuracy of 41.16 %. On enabling the video modality, we observed a 2 % improvement in the accuracy. Interestingly, enabling just the subtitles modality gave the highest validation accuracy of 46.16 % . This indicates a strong influence of subtitles on the model's performance.

The validation accuracy and loss curves are shown in Figures [7-10].

We experimented with 3 multimodal architectures. While we employed ResNet101 video feature extraction pipeline in all three, the primary difference in the models lies in the text embedding and modality concatenation pipelines. The GloVe+ResNet+Self Attention model gave a validation accuracy of 33.71 %. However, on switching GloVe with a BERT encoder, we noted an accuracy improvement of almost 12 %. Our understanding is that unlike GloVe, the BERT text encoder generates context rich embeddings thereby, giving the model a significant boost in performance. To fuse the modality representations, we experimented with a simple linear layer versus a self attention block. Using self attention, we observed an im-

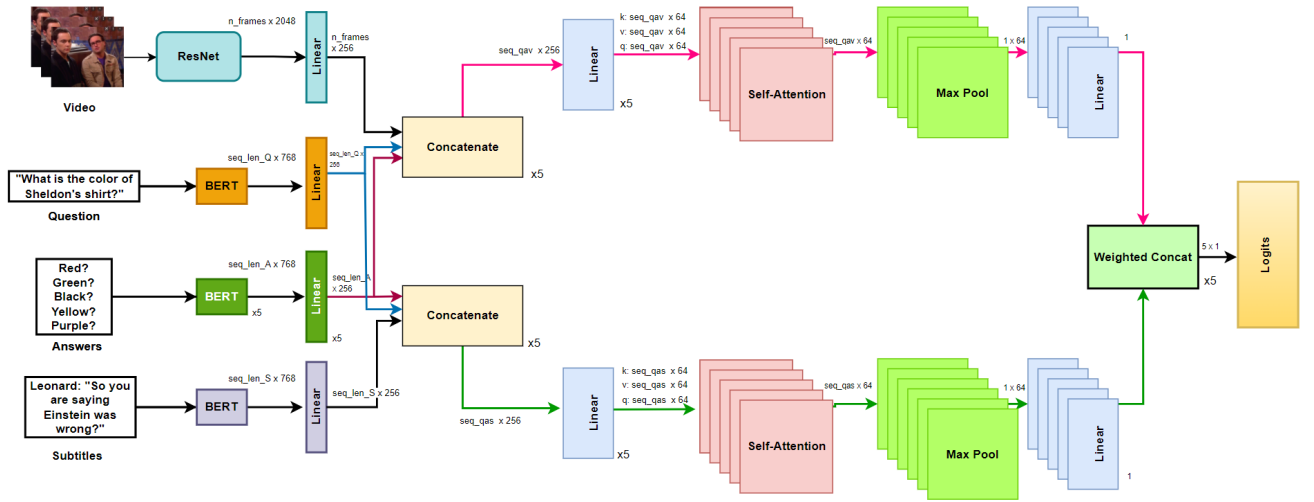


Figure 4: BERT + ResNet + Self Attention Model multimodal baseline

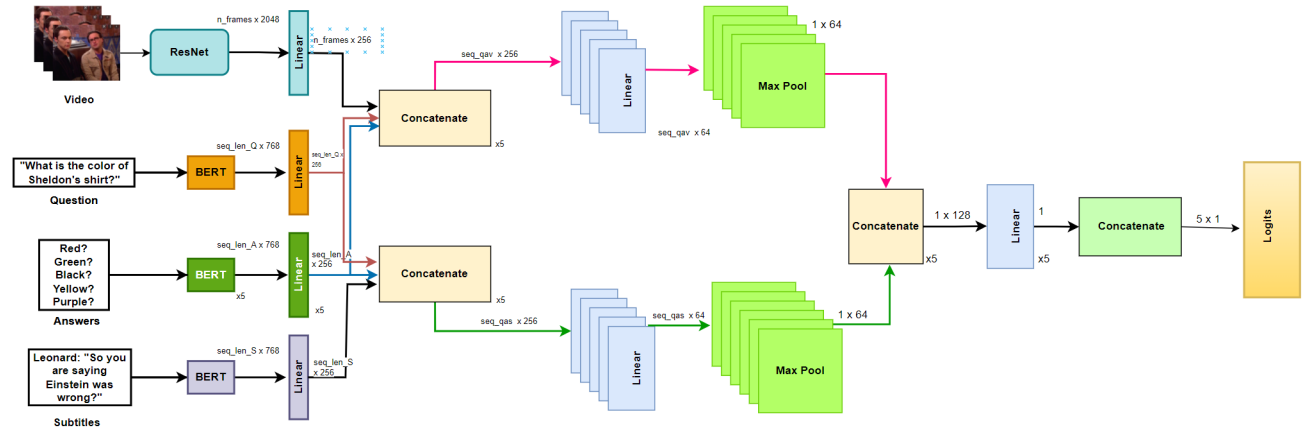


Figure 5: BERT + ResNet + Linear Layer Model multimodal baseline

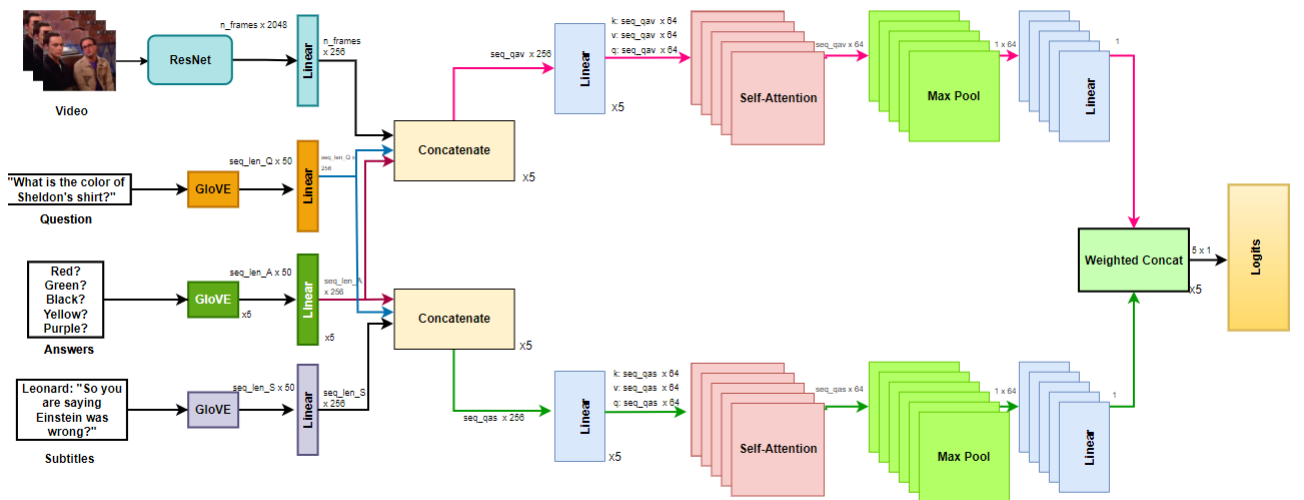


Figure 6: GloVe + ResNet + Self Attention Model multimodal baseline



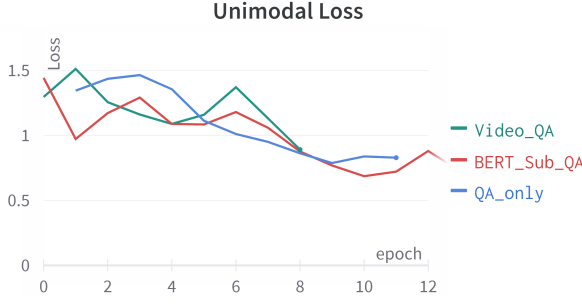


Figure 7: Unimodal Models Loss Curves

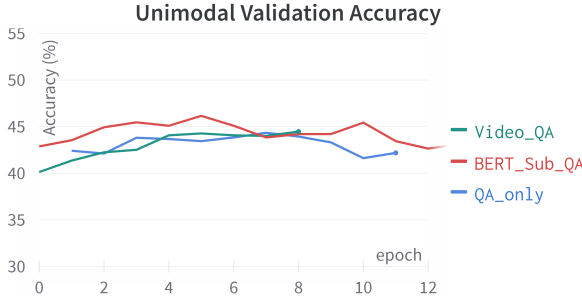


Figure 8: Unimodal Models Validation Accuracy

provement of 3 % over a linear layer.

Finally, we performed inference using open sourced pretrained models found in recent literature. The baseline model proposed by the authors of the TVQA dataset performed exceedingly well with an accuracy of 64.71 %. However, running iPerceive and HERO gave disappointing results. Surprisingly, their performance fell short of our expectations and was only slightly better than random selection of a candidate answer. We suspect that unresolved issues in the source of these open sourced projects are the reason for their poor performance.

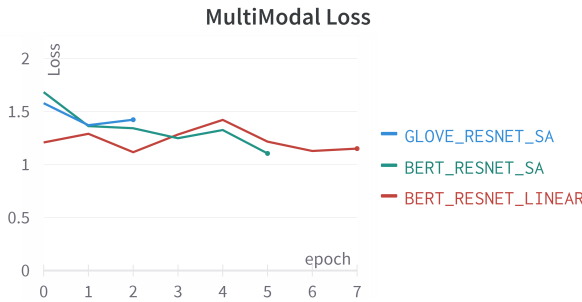


Figure 9: Multimodal Models Loss Curves

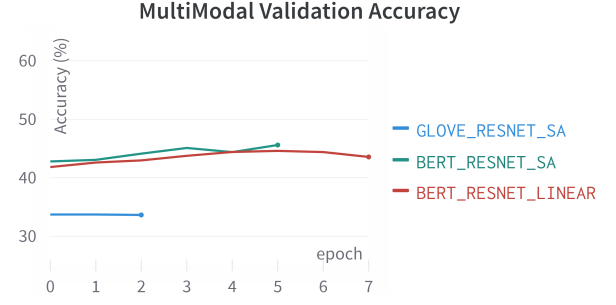


Figure 10: Multimodal Models Validation Accuracy

### 3 Analysis

We observed that in the current implementation of our unimodal and multimodal designs, text is a stronger modality compared to visual inputs. We used Resnet101(He et al., 2015) feature extraction to capture semantics of each frame. Initially, we plugged in a ImageNet pretrained ResNet101 block that took each frame in the video and generated a feature of length 2048. However, extracting features and stacking them across time proved to be very inefficient. The estimated time for one epoch was in the order of tens of hours. To work around this, we used the pre-trained features provided on the TVQA webpage. This resulted in faster video feature extraction and cut down the epoch completion time to 45 minutes. Moreover, we integrated a FasterRCNN(Ren et al., 2015) object detection module to get the list of objects in a frame and use the GloVe embedding of the list as a “visual concept”. While interesting, we omitted this module for efficiency and will reintegrate it on acquiring more compute.

At the moment, we are using a self-attention block to fuse multiple modality representations. An interesting observation in case of video unimodal analysis was that self-attention was detrimental to the model’s performance. Naively using self-attention with video modality resulted in just 22 % validation accuracy. However, replacing self-attention with a Convolution and 2D Maxpool layer gave an accuracy of 43 %. We infer that self-attention in itself is insufficient in capturing frame level understanding of the video and will require more sophistication and data to be useful.

#### 3.1 Intrinsic Metrics

In this section we analyze intrinsic metrics to gather evidence of subproblems that are solved by our models. Table 2 shows the accuracy on various

Model	Overall Acc (%)	% of “wh” questions answered correctly					
		“who”	“where”	“how”	“what”	“why”	“which”
QA only	41.22	34.62	38.01	48.71	44.74	54.16	<b>50</b>
QA + V	43.47	28.47	40.06	48.71	43.42	53.47	30
QA + S	<b>46.16</b>	<b>35.27</b>	39.72	46.88	<b>47.48</b>	<b>57.63</b>	20
GloVE + ResNet + SA	33.71	24.91	27.39	39.19	33.84	49.65	20
BERT + ResNet + SA	45.19	33.65	<b>40.75</b>	<b>50.91</b>	45.67	56.59	20
BERT + ResNet + Linear	42.85	27.83	37.67	46.15	43.75	55.2	10

Table 2: Model Analysis on wh questions

Model	Overall Acc.	% 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	<b>46.16</b>	32.76	<b>51.48</b>	23.76	<b>4.01</b>
GloVE + ResNet + SA	33.71	18.97	39.24	<b>30.46</b>	0
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

Table 3: Model Analysis on predefined attributes

“wh” questions for unimodal and multimodal models. “Who” questions ask about what someone “did”; an action. To answer a “who” question, we believe that subtitles and video, both are essential. In our experiments, we observed that QA + S model had the highest validation accuracy for “who” questions. “Where” questions request for a location i.e. visual cues are needed to answer this question. The BERT + ResNet + Self Attention multimodal model had the highest accuracy for where questions. Our hypothesis is that this model is capable of capturing frame level object locations relative to other objects. In addition, we believe that self attention is effective in answering “how” questions, as depicted in Table 2. Currently, we used limited compute to get our baselines up and running, however, this experiment compels one to think in the direction of adding more attention layers to strengthen how the model comprehends different modalities. QA + S model is the winner for “what” and “why” questions which further establishes the conclusion that subtitles have a considerable effect on the model’s performance.

Certain questions require heavy reliance on a particular modality. For example, questions that request the “color” of an object must be addressed by the video parsing pipeline. Similarly, questions pertaining to discourse between people require sufficient information in text; or subtitles in this case.

To measure the model’s performance for these two intrinsic metrics, we consider “color” and “dialog” related questions. Table 3 shows the percentage of questions with correct answer predictions that fall under “color” or “dialog”. We expected our multimodal models to have the highest accuracy for “color” questions. Surprisingly, it was the QA only model that had the best score. We infer that the video modality needs further investigation as to how we can capture frame level and temporal semantics. For dialog questions, the QA+S baseline surpasses all other models, primarily because dialog is explicitly dependent on text i.e. subtitles.

The authors of the TVQA highlighted that it is possible that a model learns a dependence between the correct option and the lengths of candidate answers. We explored this aspect for our 3 unimodal and 3 multimodal models. It was found that of the total number of questions wherein the correct answer is the longest in length, 30.46 % of such questions were correctly answered by Glove + ResNet(He et al., 2015) + Self Attention model. On the other hand, shortest candidate answers are seldom correct. The maximum reported validation accuracy for QA + S is poor. Only 4 % of such questions were answered correctly. The inference is that all 6 baseline models perform poorly on short answers which are correct.

Owing to its memory requirements, video data

is always challenging to pre-process, maintain and efficiently pass through a model training mechanism. The TVQA dataset is more than 150 GiB. Storing it on Google drive (shared by the authors of TVQA) and accessing it was extremely slow. To solve this latency issue, we un-tarred the data files on an EC2 instance and synced its data to an S3 bucket. This led to a massive improvement in fetching frames for a video clip. We also introduced a “skip” parameter that determines the number of skips while fetching frames. Our experiments were performed on a skip value of 10. A low skip values corresponds to more training data at the expense of computation resources. However, a relatively high skip value may result in diminished performance but with the added advantage of saving resources.

### 3.2 Qualitative Analysis and Examples

Figure 11 shows a few example inputs and the predictions of our unimodal and multimodal models. In the first example, the question “Where is Leonard sitting when Lesley talks about the kiss?” requires understanding of the frame. Video unimodal and Multimodal BERT + ResNet + Linear model correctly predict the ground truth as desired. However, the other four models fail to capture Leonard’s position and all of them predict that he is sitting “behind” the desk. Introducing a visual concept feature i.e. a list of detected objects in each frame should potentially solve the problem of relative positioning of objects.

In the second example, the question asks what Howard showed to Bernadette after he does a magic trick. The frames depict Howard holding two purple vests. Four out of six models correctly predict the label. At first glance, one might fall under the misconception that it is the rich video features which may have resulted in most of the models predicting the right answer. However, taking a closer look at the subtitles, we see that the phrase “two vests” is also present in four of the candidate answers as separate words. In previous sections, we highlighted how subtitles have a stronger influence on the prediction compared to visual cues. This inference implies that possibly, it is not the video features that are responsible for the desired performance, but the subtitles which have a high similarity with the candidate answers.

In the third and fourth examples, the question requests the position of a character and an object respectively. In these examples, we expect the model

to detect objects and their positioning (front, back, over, on, behind etc) with other objects. Similar to the first, these examples can be accurately predicted if we introduce a visual concepts pipeline using FasterRCNN. However, the challenge is to efficiently integrate such a pipeline.

## 4 Team Member Contributions

**Bhavuk Sharma** Subtitles unimodal architecture, integrating subtitles for multimodal models, analysis on predictions, modularized run scripts

**Debayan Ghosh** Subtitles unimodal architecture, integrating subtitles for multimodal models, analysis on predictions, integrated visualization, modularized run scripts

**Rituraj Singh** Video unimodal architecture, integrating video for multimodal models, analysis on predictions, efficient data storage/fetch pipeline

**Saahithya Yendluri** Subtitles unimodal architecture, integrating subtitles for multimodal models, analysis on predictions, prior work analysis

**Wallace Dalmet** Video unimodal architecture, integrating video for multimodal models, analysis on predictions, prior work analysis






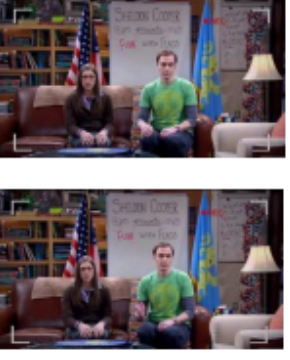
	<p>Lesley : Fair enough . &lt;eos&gt;  Lesley : On the plus side , it was a good kiss , reasonable technique . &lt;eos&gt; Lesley : . no extraneous spittle . &lt;eos&gt;  Lesley : On the other hand , no arousal . &lt;eos&gt; UNKNAME : None ? None . &lt;eos&gt; Leonard : Ah . &lt;eos&gt; Leonard : Well , thank you for your time .</p>	<p>Question: Where is Leonard sitting when Lesley talks about the kiss ?  a0: Leonard is sitting on the sink .  <b>a1: Leonard is sitting on the barstool .</b>  a2: Leonard is sitting on the floor .  a3: Leonard is sitting in Lesley 's lap .  a4: Leonard is sitting behind the desk .</p> <p>Predictions:  <b>UNI_VIDEO : Leonard is sitting on the barstool .</b>  UNI_SUB : Leonard is sitting behind the desk .  UNI_QA : Leonard is sitting behind the desk .  <b>MM_LIN : Leonard is sitting on the barstool .</b>  MM_GLOVE : Leonard is sitting behind the desk .  MM_SA : Leonard is sitting behind the desk .</p>
	<p>Howard : . the only wand that ever saw any action was this one . &lt;eos&gt; UNKNAME : Oh ! Look what my mom made us for the act . &lt;eos&gt; Howard : Well , she cut up one of her old bathing suits . &lt;eos&gt; Howard : She made these two vests and half a dozen napkins .</p>	<p>Question: What does Howard show Bernadette after he does a magic trick ?  a0: His old box of magic tricks .  a1: Half a dozen napkins .  a2: His mom 's favorite dress .  <b>a3: Two purple vests .</b>  a4: A pink robe he bought her .</p> <p>Predictions:  <b>UNI_VIDEO : Two purple vests .</b>  UNI_SUB : A pink robe he bought her .  <b>UNI_QA : Two purple vests .</b>  <b>MM_LIN : Two purple vests .</b>  MM_GLOVE : His old box of magic tricks .  <b>MM_SA : Two purple vests .</b></p>
	<p>Raj : Oh , another two hours to go . &lt;eos&gt; Raj : The wait is killing me . &lt;eos&gt; Howard : I know . I get it . When I was in the Soyuz capsule &lt;eos&gt; Howard : returning from the space station , plummeting toward Earth &lt;eos&gt; Howard : at 17 , 000 miles per hour . Before you finish , &lt;eos&gt; Raj : is this a story about patience and waiting</p>	<p>Question: Where sat Howard when talking to Raj ?  a0: Rug .  a1: Chair .  a2: Stool .  <b>a3: Couch .</b>  a4: Table .</p> <p>Predictions:  UNI_VIDEO : Chair .  <b>UNI_SUB : Couch .</b>  <b>UNI_QA : Couch .</b>  MM_LIN : Chair .  MM_GLOVE : Chair .  MM_SA : Chair .</p>
	<p>Sheldon : What 's the only non rectangular flag ? &lt;eos&gt;  Sheldon : What animal appears most often on flags ? &lt;eos&gt;  Sheldon : What animal appears second most often on flags ? &lt;eos&gt; Sheldon : And more . &lt;eos&gt; Amy : Sweet . &lt;eos&gt;  Sheldon : Why are you waving a white flag ? &lt;eos&gt; Amy : I 'm surrendering to fun .</p>	<p>Question: What was behind Amy and Sheldon when they were recording themselves ?  a0: Lifesized superheros  a1: Windows  a2: A dog  <b>a3: Flags</b>  a4: A mirror</p> <p>Predictions:  UNI_VIDEO : Lifesized superheros  <b>UNI_SUB : Flags</b>  <b>UNI_QA : Flags</b>  MM_LIN : Lifesized superheros  MM_GLOVE : Lifesized superheros  <b>MM_SA : Flags</b></p>

Figure 11: Example predictions from proposed models

## References

- Aman Chadha, Gurneet Arora, and Navpreet Kaloty. 2020. [iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering](#). *CoRR*, abs/2011.07735.
- 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.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. [Deep residual learning for image recognition](#). *CoRR*, abs/1512.03385.
- Dan Hendrycks and Kevin Gimpel. 2016. [Bridging non-linearities and stochastic regularizers with gaussian error linear units](#). *CoRR*, abs/1606.08415.
- Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2018. [TVQA: localized, compositional video question answering](#). *CoRR*, abs/1809.01696.
- Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2019. [TVQA+: spatio-temporal grounding for video question answering](#). *CoRR*, abs/1904.11574.
- Jie Lei, Licheng Yu, Tamara L. Berg, and Mohit Bansal. 2020. [TVR: A large-scale dataset for video-subtitle moment retrieval](#). *CoRR*, abs/2001.09099.
- Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. [HERO: hierarchical encoder for video+language omni-representation pre-training](#). *CoRR*, abs/2005.00200.
- 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.
- Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. [Faster R-CNN: towards real-time object detection with region proposal networks](#). *CoRR*, abs/1506.01497.