

# A Experimental Setup

Our experiments are on the TVQA dataset and we use and adapt the code provided by the authors<sup>1</sup>. Due to their size, the regional features are unavailable for download and we extract them ourselves following the author’s instructions. The models are trained on an RTX 2080 Ti GPU with batch size 32 and a rectified-Adam solver [10]. We use a pretrained, non-finetuned BERT embedding layer using the uncased base tokenizer<sup>2</sup>. When using regional features we use the top 20 detections per video segment. All further settings are as described in TVQA, most notably: We use 6B-300d GloVe embeddings and all word embedding layers are frozen during training. We use the timestamps annotations and train the model until improvements on the validation set accuracy is not made for 3 epochs. We check validation and training set accuracies every 400 iterations, except for the models that include regional features where we check every 800 iterations as these run significantly slower. In this study we control for the modality used in order to isolate its influence on the performance of the overall model. Details of the different variations are evaluated and their associated results are discussed in the next section.

# B Model Similarities

We provide the IoU scores between the GloVe embedding model variations. The IoU scores in Figure 6 are similar to their BERT counterparts shown in Figure 3b. As an alternative set

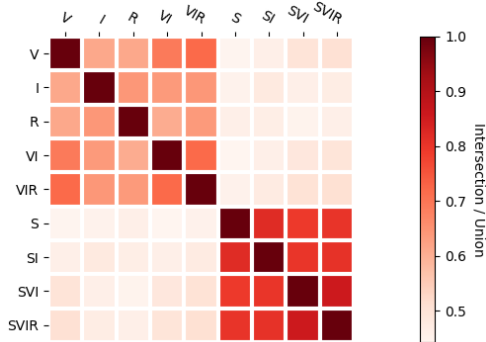


Figure 6: Intersection / Union (IoU) score for correct predictions in the validation set between GloVe models.

comparison measure, we consider the proportion of questions in the validation set that each pair of models answer the same, regardless if the answer is correct or incorrect. We find the non-subtitle models with GloVe embeddings (Figure 7) agree *slightly* more than those with BERT embeddings (Figure 8).

<sup>1</sup><https://github.com/jayleicn/TVQA>  
<sup>2</sup><https://github.com/huggingface/transformers>

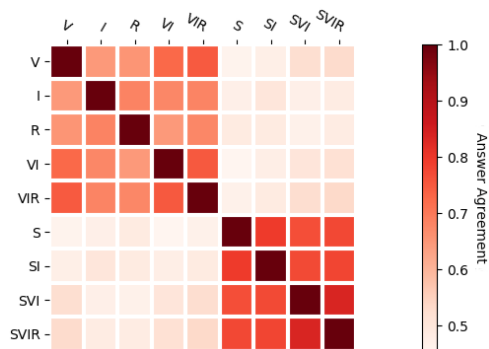


Figure 7: Proportion of the validation set that GloVe models answer the same.

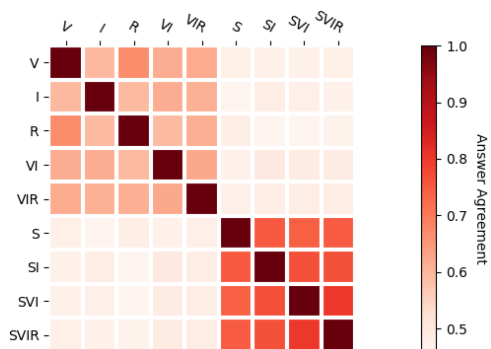


Figure 8: Proportion of the validation set that BERT models answer the same.

## C Question Type Analysis

Question ‘Type’	Example ‘Other’ Questions
Spelling Variation	‘ <i>Whom</i> did Roger say was following him after he made the drop?’
Typo	‘ <b>tWhat</b> was the reason House said they should do a brain biopsy when they were discussing options of what to do?’
<i>Did/Does</i>	‘ <i>Did</i> Joey walk into the room before or after Chandler?’
Double ‘When’ Question	‘ <i>When</i> did Lucas say he made the video <i>when</i> he was showing to Beckett and Castle?’

Table 3: Example questions from ‘other’ question type category. The ‘other’ category makes up 1.1% of the validation set.

## D Feature Contributions

To complement the true positive and false positive vote contributions analysed in Figure 2, we present the answer vote contributions of true negative and false negative answers between VIR and SVIR trained models with both BERT and GloVe embeddings.

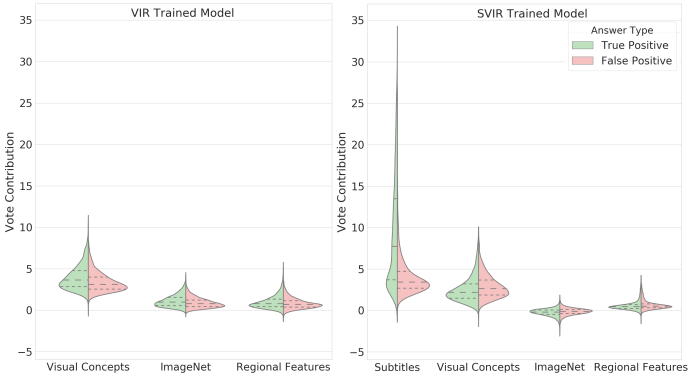


Figure 9: Pre-softmax vote contributions for answers in the validation set for the VIR (left) and SVIR (right) trained models with GloVe embeddings. This is the GloVe embedding counterpart to Figure 2.



Figure 10: Pre-softmax vote contributions for answers in the validation set for the VIR (left) and SVIR (right) trained models with GloVe embeddings.



Figure 11: Pre-softmax vote contributions for answers in the validation set for the VIR (left) and SVIR (right) trained models with BERT embeddings.

## E Training Set Inclusion-Exclusion

Group A	Group B	BERT Models	GloVe Models
<i>All</i>	-	96.77%	94.54%
<i>All</i>	<i>Non-Subtitle</i>	14.32%	14.56%
<i>All</i>	SVIR	15.19%	14.47%
<i>Subtitle</i>	-	94.80%	89.91%
<i>Subtitle</i>	<i>Non-Subtitle</i>	14.32%	14.56%
<i>Non-Subtitle</i>	-	82.45%	79.99%
<i>Non-Subtitle</i>	<i>Subtitle</i>	1.96%	4.63%
<i>Non-Subtitle</i>	S	12.34%	15.97%
S, V, I, R	-	91.11%	90.52%
S, V, I, R	SVIR	12.15%	12.03%
SVIR	-	81.58%	80.07%
SVIR	S, V, I, R	2.62%	1.58%
S	-	80.77%	76.41%
S	<i>Non-Subtitle</i>	10.67%	12.39%

Table 4: The percentages of the *training* set that are correctly answered by models in Group A, but incorrectly answered by Group B. *Subtitle models* = {S, SI, SVI, SVIR}, *Non-Subtitle models* = {V, I, R, VI, VIR}. *All models* = *Subtitle* + *Non-Subtitle*. Though considering responses of the training set is inherently flawed due to training bias, it provides a reasonable starting point and considerable size boost to our initially proposed IEM subsets.

## F RUBi Learning Strategy

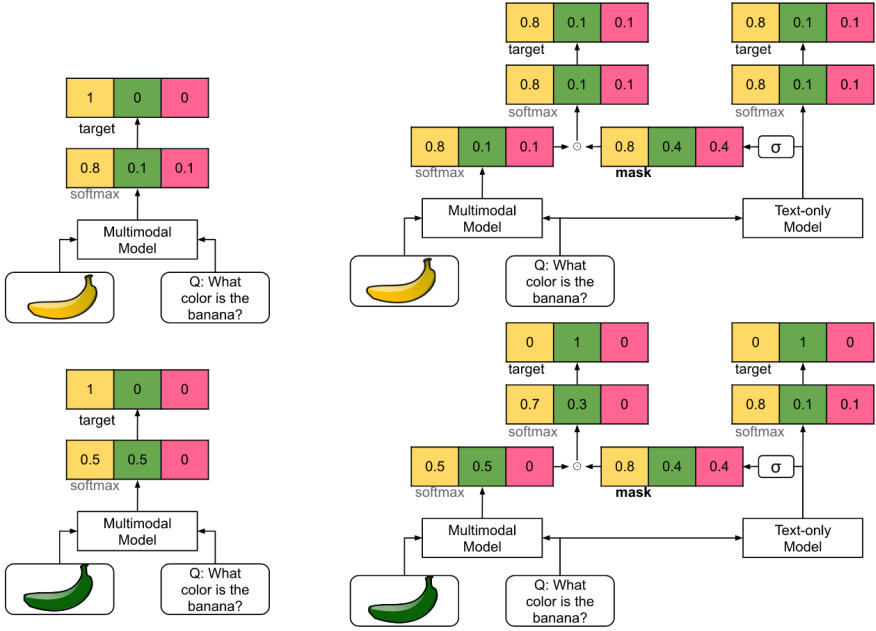


Figure 12: The RUBi (reducing unimodal bias) learning strategy used in VQA. The model-agnostic RUBi strategy [1] uses a text-only variant of a model during training to reduce (increase) the loss, and therefore importance, of highly-biased (visually dependent and difficult) training samples.

## References

- [1] Rémi Cadène, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering. In *NeurIPS*, 2019.
- [2] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.