Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA

Hyounghun Kim **Zineng Tang** **Mohit Bansal**
UNC Chapel Hill
{hyounghk, terran, mbansal}@cs.unc.edu

arXiv:2005.06409v1 [cs.CL] 13 May 2020

Abstract

Videos convey rich information. Dynamic spatio-temporal relationships between people/objects, and diverse multimodal events are present in a video clip. Hence, it is important to develop automated models that can accurately extract such information from videos. Answering questions on videos is one of the tasks which can evaluate such AI abilities. In this paper, we propose a video question answering model which effectively integrates multi-modal input sources and finds the temporally relevant information to answer questions. Specifically, we first employ dense image captions to help identify objects and their detailed salient regions and actions, and hence give the model useful extra information (in explicit textual format to allow easier matching) for answering questions. Moreover, our model is also comprised of dual-level attention (word/object and frame level), multi-head self/cross-integration for different sources (video and dense captions), and gates which pass more relevant information to the classifier. Finally, we also cast the frame selection problem as a multi-label classification task and introduce two loss functions, In-and-Out Frame Score Margin (IOFSM) and Balanced Binary Cross-Entropy (BBCE), to better supervise the model with human importance annotations. We evaluate our model on the challenging TVQA dataset, where each of our model components provides significant gains, and our overall model outperforms the state-of-the-art by a large margin (74.09% versus 70.52%). We also present several word, object, and frame level visualization studies.[1]

1 Introduction

Recent years have witnessed a paradigm shift in the way we get our information, and a lot of it is related to watching and listening to videos that are shared in huge amounts via the internet and new high-speed networks. Videos convey a diverse breadth of rich information, such as dynamic spatio-temporal relationships between people/objects, as well as events. Hence, it has become important to develop automated models that can accurately extract such precise multimodal information from videos (Tapaswi et al., 2016; Maharaj et al., 2017; Kim et al., 2017; Jang et al., 2017; Gao et al., 2017; Anne Hendricks et al., 2017; Lei et al., 2018, 2020). Video question answering is a representative AI task through which we can evaluate such abilities of an AI agent to understand, retrieve, and return desired information from given video clips.

In this paper, we propose a model that effectively integrates multimodal information and locates the relevant frames from diverse, complex video clips such as those from the video+dialogue TVQA dataset (Lei et al., 2018), which contains questions that need both the video and the subtitles to answer. When given a video clip and a natural language question based on the video, naturally, the first step is to compare the question with the content (objects and keywords) of the video frames and subtitles, then combine information from different video frames and subtitles to answer the question. Analogous to this process, we apply dual-level attention in which a question and video/subtitle are aligned in word/object level, and then the aligned features from video and subtitle respectively are aligned the second time at the frame-level to integrate information for answering the question. Among the aligned frames (which contain aggregated video and subtitle information now), only those which contain relevant information for answering the question are needed. Hence, we also apply gating mechanisms to each frame feature to select the most informative frames before feeding them to the classifier.

[1] Our code is publicly available at: https://github.com/hyounghk/VideoQADenseCapFrameGate-ACL2020

Next, in order to make the frame selection more effective, we cast the frame selection sub-task as a multi-label classification task. To convert the time span annotation to the label for each frame, we assign a positive label ('1') to frames between the start and end points, and negative ('0') label to the others, then train them with the binary cross-entropy loss. Moreover, for enhanced supervision from the human importance annotation, we also introduce a new loss function, In-and-Out Frame Score Margin (IOFSM), which is the difference in average scores between in-frames (which are inside the time span) and out-frames (which are outside the time span). We empirically show that these two losses are complementary when they are used together. Also, we introduce a way of applying binary cross-entropy to the unbalanced dataset. As we see each frame as a training example (positive or negative), we have a more significant number of negative examples than positive ones. To balance the bias, we calculate normalized scores by averaging the loss separately for each label. This modification, which we call balanced binary cross-entropy (BBCE), helps adjust the imbalance and further improve the performance of our model.

Finally, we also employ dense captions to help further improve the temporal localization of our video-QA model. Captions have proven to be helpful for vision-language tasks (Wu et al., 2019; Li et al., 2019; Kim and Bansal, 2019) by providing additional, complementary information to the primary task in descriptive textual format. We employ dense captions as an extra input to our model since dense captions describe the diverse salient regions of an image in object-level detail, and hence they would give more useful clues for question answering than single, non-dense image captions.

Empirically, our first basic model (with dual-level attention and frame-selection gates) outperforms the state-of-the-art models on TVQA validation dataset (72.53% as compared to 71.13% previous state-of-the-art) and with the additional supervision via the two new loss functions and the employment of dense captions, our model gives further improved results (73.34% and 74.20% respectively). These improvements from each of our model components (i.e., new loss functions, dense captions) are statistically significant. Overall, our full model's test-public score substantially outperforms the state-of-the-art score by a large margin of 3.57% (74.09% as compared to 70.52%).[2] Also, our model's scores across all the 6 TV shows are more balanced than other models in the TVQA leaderboard[3], implying that our model should be more consistent and robust over different genres/domains that might have different characteristics from each other.

Our contributions are four-fold: (1) we present an effective model architecture for the video question answering task using dual-level attention and gates which fuse and select useful spatial-temporal information, (2) we employ dense captions as salient-region information and integrate it into a joint model to enhance the videoQA performance by locating proper information both spatially and temporally in rich textual semi-symbolic format, (3) we cast the frame selection sub-task as a multi-level classification task and introduce two new loss functions (IOFSM and BBCE) for enhanced supervision from human importance annotations (which could be also useful in other multi-label classification settings), and (4) our model's score on the test-public dataset is 74.09%, which is around 3.6% higher than the state-of-the-art result on the TVQA leaderboard (and our model's scores are more balanced/consistent across the diverse TV show genres). We also present several ablation and visualization analyses of our model components (e.g., the word/object-level and the frame-level attention).

2 Related Work

Visual/Video Question Answering Understanding visual information conditioned on language is an important ability for an agent who is supposed to have integrated intelligence. Many tasks have been proposed to evaluate such ability, and visual question answering is one of those tasks (Antol et al., 2015; Lu et al., 2016; Fukui et al., 2016; Xu and Saenko, 2016; Yang et al., 2016; Zhu et al., 2016; Goyal et al., 2017; Anderson et al., 2018). Recently, beyond question answering on a single image, attention to understanding and extracting information from a sequence of images, i.e., a video, is rising (Tapaswi et al., 2016; Maharaj et al., 2017; Kim et al., 2017; Jang et al., 2017; Lei et al., 2018; Zadeh et al., 2019; Lei et al., 2020; Garcia et al., 2020). Answering questions on videos requires an

[2] At the time of the ACL2020 submission deadline, the publicly visible rank-1 entry was 70.52%. Since then, there are some new entries, with results up to 71.48% (compared to our 74.09%).

[3] https://competitions.codalab.org/competitions/20415#results

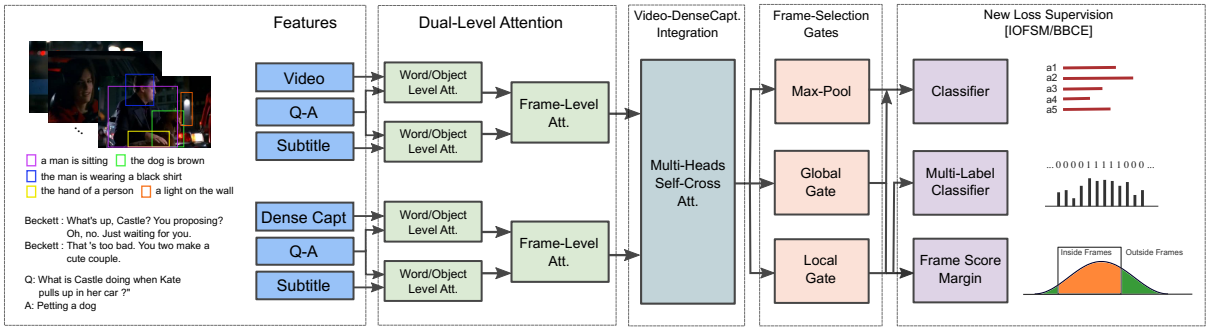


Figure 1: Our model consists of three parts: Dual-Level Attention, Video-DenseCapt Integration, and Frame-Selection Gates. The new loss functions (IOFSM/BBCE) also help improve the model with enhanced supervision.

understanding of temporal information as well as spatial information, so it is more challenging than a single image question answering.

Temporal Localization Temporal localization is a task that is widely explored in event/object detection in video context. There has been work that solely processes visual information to detect objects/actions/activity (Gaidon et al., 2013; Weinzaepfel et al., 2015; Shou et al., 2016; Dai et al., 2017; Shou et al., 2017). At the same time, work on natural language-related temporal localization task is less explored with recent work that focuses on the retrieval of a certain moment in a video by natural language (Anne Hendricks et al., 2017; Gao et al., 2017). With deliberately designed gating and attention mechanisms, our work, in general, will greatly contribute to the task of temporal localization, especially under natural language context and multimodal data.

Dense Image Captioning Image captioning is another direction of understanding visual and language information jointly. Single-sentence captions (Karpathy and Fei-Fei, 2015; Anderson et al., 2018) capture the main concept of an image to describe it in a single sentence. However, an image could contain multiple aspects that are important/useful in different ways. Dense captions (Johnson et al., 2016; Yang et al., 2017) and paragraph captions (Krause et al., 2017; Liang et al., 2017; Melas-Kyriazi et al., 2018) have been introduced to densely and broadly capture the diverse aspects and salient regions of an image. Especially, dense caption describes an image in object-level and gives useful salient regional information about objects such as attributes and actions. In this paper, we take advantage of this dense caption ability to help our video QA model understand an image better for answering questions.

3 Model

Our model consists of 2 parts: feature fusion and frame selection. For feature fusion, we introduce dual-level (word/object and frame level) attention, and we design the frame selection problem as a multi-label classification task and introduce 2 new loss functions for enhanced supervision (Figure 1).

3.1 Features

We follow the same approach of Lei et al. (2020)'s work to obtain features from video, question-answer pairs, and subtitle input and encode them. We sample frames at 0.5 fps and extract object features from each frame via Faster R-CNN (Girshick, 2015). Then we use PCA to get features of 300 dimension from top-20 object proposals. We also create five hypotheses by concatenating a question feature with each of five answer features, and we pair each visual frame feature with temporally neighboring subtitles. We encode all the features using convolutional encoder.

$$\phi_{en}(x): \begin{cases} x_0^0 = E_{pos}(x) \\ x_t^i = f_{i,t}(x_{t-1}^i) + x_{t-1}^i, \\ f_i(x_0^i) = g_n(x_L^i) \\ y = f_N \circ ... \circ f_1(x_0^1) \end{cases} \quad (1)$$

where E_{pos} denotes positional encoding, $f_{i,t}$ convolution preceded by Layer Normalization and followed by ReLU activation, and g_n the layer normalization. The encoder is composed of N blocks iterations. In each iteration, the encoded inputs are transformed L times of convolutions. The L is set to 2, and N to 1 in our experiment (Figure 2).

3.2 Dual-Level Attention

In dual-level attention, features are sequentially aligned in word/object-level and frame-level (Figure 3).

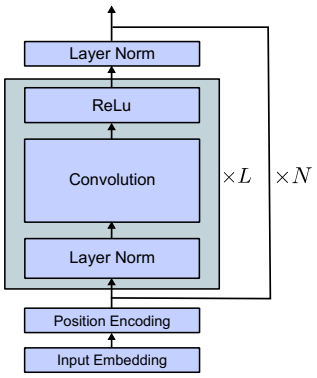


Figure 2: CNN encoder. We use this block to encode all the input features.

Word/Object-Level Attention The QA feature, $qa = \{qa_0, qa_1, .., qa_{T_{qa}}\}$, are combined with subtitle feature, $s_t = \{s_{t0}, s_{t1}, .., s_{tT_{s_t}}\}$, and visual feature, $v_t = \{v_{t0}, v_{t1}, .., v_{tT_{v_t}}\}$, of t-th frame respectively via word/object-level attention. To be specific, we calculate similarity matrices following Seo et al. (2017)'s approach, $S_t^v \in \mathbb{R}^{T_{qa} \times T_{s_t}}$ and $S_t^s \in \mathbb{R}^{T_{qa} \times T_{v_t}}$, from QA/subtitle and QA/visual features respectively. From the similarity matrices, attended subtitle features are obtained and combined with the QA features by concatenating and applying a transforming function. Then, max-pooling operation is applied word-wise to reduce the dimension.

$$(S_t^s)_{ij} = qa_i^\top s_{tj} \quad (2)$$
$$s_t^{att} = \text{softmax}(S_t^s) \cdot s_t \quad (3)$$
$$qa_s^m = \text{maxpool}(f_1([qa; s_t^{att}; qa \odot s_t^{att}])) \quad (4)$$

where f_1 is a fully-connected layer followed by ReLU non-linearity. The same process is applied to the QA features.

$$qa^{att} = \text{softmax}(S_t^{s\top}) \cdot qa \quad (5)$$
$$s_t^m = \text{maxpool}(f_1([s_t; qa^{att}; s_t \odot qa^{att}])) \quad (6)$$

The fused features from different directions are integrated by concatenating and being fed to a function as follows:

$$s_t^w = f_2([qa_s^m; s_t^m; qa_s^m \odot s_t^m; qa_s^m + s_t^m]) \quad (7)$$

where f_2 is the same function as f_1 with non-shared parameters. All this process is also applied to visual features to get word/object-level attended features.

$$v_t^w = f_2([qa_v^m; v_t^m; qa_v^m \odot v_t^m; qa_v^m + v_t^m]) \quad (8)$$

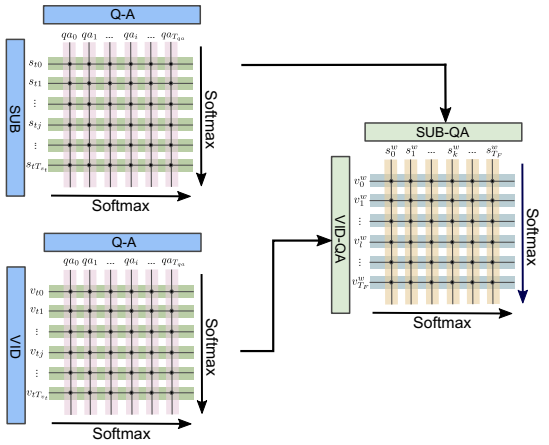


Figure 3: Dual-Level Attention. Our model performs two-level attentions (word/object and frame level) sequentially. In the word/object-level attention, each word/object is aligned to relevant words or objects. In the frame-level attention, each frame (which has integrated information from the word/object-level attention) is aligned to relevant frames.

Frame-Level Attention The fused features from word/object-level attention are integrated frame-wise via frame-level attention. Similar to the word/object-level attention, a similarity matrix $S \in \mathbb{R}^{T_F \times T_F}$, is calculated, where T_F is the number of frames. Also, from the similarity matrix, attended frame-level features are calculated as

$$(S)_{kl} = s_k^{w\top} v_l^w \quad (9)$$
$$s^{att} = \text{softmax}(S) \cdot s^w + s^w \quad (10)$$
$$\hat{v} = f_3([v^w; s^{att}; v^w \odot s^{att}; v^w + s^{att}]) \quad (11)$$
$$v^{att} = \text{softmax}(S^\top) \cdot v^w + v^w \quad (12)$$
$$\hat{s} = f_3([s^w; v^{att}; s^w \odot v^{att}; s^w + v^{att}]) \quad (13)$$

where f_3 is the same function as f_1 and f_2 with non-shared parameters. The frame-wise attended features are added to get an integrated feature.

$$u^{sv} = \hat{s} + \hat{v} \quad (14)$$

3.3 Video and Dense Caption Integration

We also employ dense captions to help further improve the temporal localization of our video QA model. They provide more diverse salient regional information (than the usual single non-dense image captions) about object-level details of image frames in a video clip, and also allow the model to explicitly (in textual/semi-symbolic form) match keywords/patterns between dense captions and questions to find relevant locations/frames.

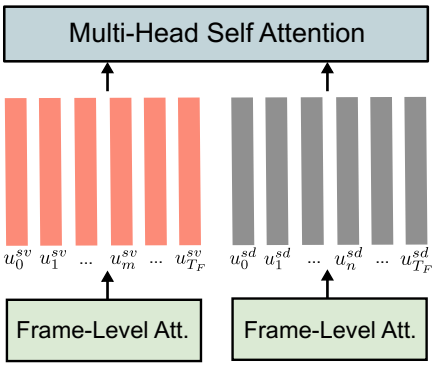
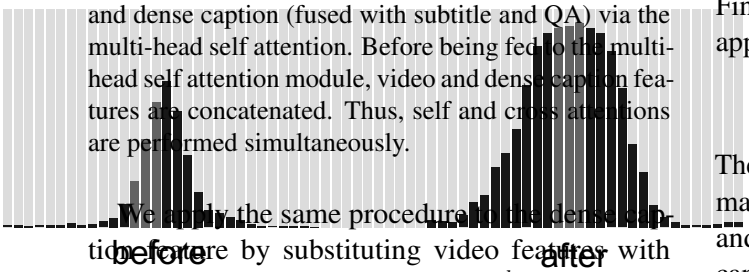


Figure 4: Self-Cross Attention. We combine information each from the video (fused with subtitle and QA) and dense caption (fused with subtitle and QA) via the multi-head self attention. Before being fed to the multi-head self attention module, video and dense caption features are concatenated. Thus, self and cross attentions are performed simultaneously.



We apply the same procedure to the dense caption feature by substituting video features with dense caption features to obtain u^{sd}. To integrate u^{sv} and u^{sd}, we employ multi-head self attention (Figure 4). To be specific, we concatenate u^{sv} and u^{sd} frame-wise then feed them to the self attention function.

$$\phi_{\text{self-att}}(x) \begin{cases} h_{\text{i}} = g_a(w_q^\top x_i, w_k^\top x_i, w_v^\top x_i) \\ y = w_m^\top [h_1; \ldots; h_{\text{k}}] \end{cases} \quad (15)$$

where g_a denotes self-attention.

$$u^{svd} = \phi_{\text{self-att}}([u^{sv}; u^{sd}]) \quad (16)$$

In this way, u^{sv} and u^{sd} attend to themselves while attending to each other simultaneously. We split the output, u^{svd} into the same shape as the input, then add the two.

$$z = u^{svd}[0 : T_F] + u^{svd}[T_F : 2T_F] \quad (17)$$

3.4 Frame-Selection Gates

To select appropriate information from the frame-length features, we employ max-pooling and gates. Features from the video-dense caption integration are fed to the CNN encoder. A fully-connected layer and sigmoid function are applied sequentially to the output feature to get frame scores that indicate how relevant each frame is for answering a given question. We get weighted features by multiplying the output feature from the CNN encoder with the scores.

$$\hat{z} = \phi_{\text{en2}}(z) \quad (18)$$

$$g^L = \text{sigmoid}(f^L(\hat{z})) \quad (19)$$

$$z^{gl} = \hat{z} \odot g^L \quad (20)$$

We calculate another frame scores with a different function f^G to get another weighted feature.

$$g^G = \text{sigmoid}(f^G(\hat{z})) \quad (21)$$

$$z^{gg} = \hat{z} \odot g^G \quad (22)$$

Finally, following Lei et al. (2020)'s work, we also apply frame-wise max-pooling.

$$z^{max} = \text{maxpool}(\hat{z}) \quad (23)$$

The three features (from local gate, global gate, and max-pooling, respectively), are then concatenated and fed to the classifier to give scores for each candidate answer.

$$logit = \text{clssifier}([z^{max}; z^{gg}; z^{gl}]) \quad (24)$$

We get the logits for the five candidate answers and choose the highest value as the predicted answer.

$$loss_{cls} = -\log(\frac{e^{s_g}}{\sum_k e^{s_k}}) \quad (25)$$

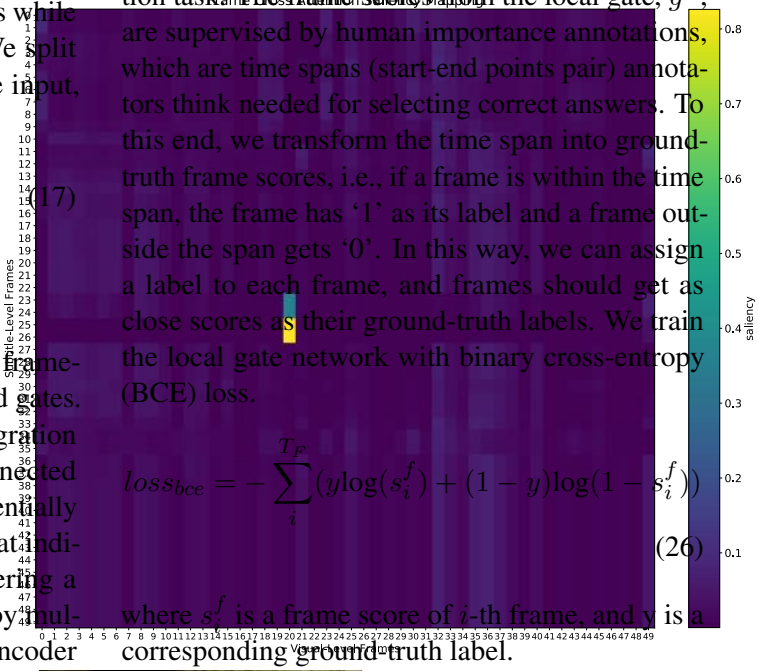
where s_g is the logit of ground-truth answer.

3.5 Novel Frame-Selection Supervision Loss Functions

We cast frame selection as a multi-label classification task. The frame scores from the local gate, g^L, are supervised by human importance annotations, which are time spans (start-end points pair) annotators think needed for selecting correct answers. To this end, we transform the time span into ground-truth frame scores, i.e., if a frame is within the time span, the frame has '1' as its label and a frame outside the span gets '0'. In this way, we can assign a label to each frame, and frames should get as close scores as their ground-truth labels. We train the local gate network with binary cross-entropy (BCE) loss.

$$loss_{bce} = -\sum_i^{T_F} (y\log(s_i^f) + (1-y)\log(1-s_i^f)) \quad (26)$$

where s_i^f is a frame score of i-th frame, and y is a corresponding ground-truth label.



Esposito : Upstairs. go.
Unkname : Carol!

Frame 20

Frame 25

In-and-Out Frame Score Margin For additional supervision other than the binary cross-entropy loss, we create a novel loss function, In-and-Out Frame Score Margin (IOFSM).

$$loss_{io} = 1 + \text{Avg}(\text{OFS}) - \text{Avg}(\text{IFS}) \quad (27)$$

where OFS (Out Frame Score) is scores of frames whose labels are '0' and IFS (In Frame Score) is scores of frames whose labels are '1'.

Balanced Binary Cross-Entropy In our multi-label classification setting, each frame can be considered as one training example. Thus, the total number of examples and the proportion between positive and negative examples vary for every instance. This variation can cause unbalanced training since negative examples usually dominate. To balance the unbalanced training, we apply a simple but effective modification to the original BCE, and we call it Balanced Binary Cross-Entropy (BBCE). To be specific, instead of summing or averaging through the entire frame examples, we divide the positive and negative examples and calculate the average cross-entropy scores separately, then sum them together.

$$\begin{aligned} loss_{bbce} = -\Big(& \sum_{i}^{T_{F_{in}}} \log(s_i^{f_{in}})/T_{F_{in}} \\ & + \sum_{j}^{T_{F_{out}}} \log(1 - s_j^{f_{out}})/T_{F_{out}} \Big) \end{aligned} \quad (28)$$

where $s_i^{f_{in}}$ and $s_j^{f_{out}}$ are i-th in-frame score and j-th out-frame score respectively, and $T_{F_{in}}$ and $T_{F_{out}}$ are the number of in-frames and out-frames respectively.

Thus, the total loss is:

$$loss = loss_{cls} + loss_{(b)bce} + loss_{io} \quad (29)$$

4 Experimental Setup

TVQA Dataset TVQA dataset (Lei et al., 2018) consists of video frames, subtitles, and question-answer pairs from 6 TV shows. The number of examples for train/validation/test-public dataset are 122,039/15,253/7,623. Each example has five candidate answers with one of them the ground-truth. So, TVQA is a classification task, in which models select one from the five candidate answers, and models can be evaluated on the accuracy metric.

Dense Captions We use Yang et al. (2017)'s pre-trained model to extract dense captions from each video frame. We extract the dense captions in advance and use them as extra input data to the model.[5]

Training Details We use GloVe (Pennington et al., 2014) word vectors with dimension size of 300 and RoBERTa (Liu et al., 2019) with 768 dimension. The dimension of the visual feature is 300, and the base hidden size of the whole model is 128. We use Adam (Kingma and Ba, 2015) as the optimizer. We set the initial learning rate to 0.001 and drop it to 0.0002 after running 10 epochs. For dropout, we use the probability of 0.1.

5 Results and Ablation Analysis

As seen from Table 1, our model outperforms the state-of-the-art models in the TVQA leaderboard. Especially our model gets balanced scores for all the TV shows while some other models have high variances across the shows. As seen from Table 2, the standard deviation and 'max-min' value over our model's scores for each TV show are 0.65 and 1.83, respectively, which are the lowest values among all models in the list. This low variance could mean that our model is more consistent and robust across all the TV shows.

Model Ablations As shown in Table 3, our basic dual-attention and frame selection gates model shows substantial improvement over the strong single attention and frame span baseline (row 4 vs 1: $p < 0.0001$), which is from the best published model (Lei et al., 2020). Each of our dual-attention and frame selection gates alone shows a small improvement in performance than the baseline (row 3 vs 1 and 2 vs 1, respectively).[6] However, when they are applied together, the model works much better. The reason why they are more effective when put together is that frame selection gates basically select frames based on useful information

[4] At the time of the ACL2020 submission deadline, the publicly visible rank-1 entry was 70.52%. Since then, two more entries have appeared in the leaderboard; however, our method still outperforms their scores by a large margin (71.48% and 71.13% versus 74.09%).

[5] This is less computationally expensive and dense captions from the separately trained model will be less biased towards the questions of TVQA dataset, and hence provide more diverse aspects of image frames of a video clip.

[6] Although the improvements are not much, but performing word/object level attention and then frame level attention is more intuitive and interpretable than a non-dual-attention method, allowing us to show how the model works: see visualization in Sec. 6.

	Model	Test-Public (%)							Val (%)
		all	bbt	friends	himym	grey	house	castle	
1	jacobssy (anonymous)	66.01	68.75	64.98	65.08	69.22	66.45	63.74	64.90
2	multi-stream (Lei et al., 2018)	66.46	70.25	65.78	64.02	67.20	66.84	63.96	65.85
3	PAMN (Kim et al., 2019b)	66.77	-	-	-	-	-	-	66.38
4	Multi-task (Kim et al., 2019a)	67.05	-	-	-	-	-	-	66.22
5	ZGF (anonymous)	68.77	-	-	-	-	-	-	68.90
6	STAGE (Lei et al., 2020)	70.23	-	-	-	-	-	-	70.50
7	akalsdnr (anonymous)	70.52	71.49	67.43	72.22	70.42	70.83	72.30	71.13
8	Ours (hstar)	**74.09**	**74.04**	**73.03**	**74.34**	**73.44**	**74.68**	**74.86**	**74.20**

Table 1: Our model outperforms the state-of-the-art models by a large margin. Moreover, the scores of our model across all the TV shows are more balanced than the scores from other models, which means our model is more consistent/robust and not biased to the dataset from specific TV shows.[4]

	Model	TV Show Score		
		avg.	std.	max-min
1	jacobssy (anonymous)	66.37	2.01	5.48
2	multi-stream (Lei et al., 2018)	66.34	2.15	6.29
3	akalsdnr (anonymous)	70.78	1.65	4.87
4	Ours	74.07	0.65	1.83

Table 2: Average and standard deviation of the test-public scores from each TV show (for this comparison, we only consider models that release the scores for each TV show).[8]

	Model	Val Score (%)
1	Single-Att + Frame-Span	69.86
2	Single-Att + Frame-Selection Gates	70.08
3	Dual-Att + Frame-Span	70.20
4	Dual-Att + Frame-Selection Gates (w/o NewLoss)	71.26
5	Dual-Att + Frame-Selection Gates	72.51
6	Dual-Att + Frame-Selection Gates (w/o NewLoss) + RoBERTa	72.53
7	Dual-Att + Frame-Selection Gates + RoBERTa	73.34
8	Dual-Att + Frame-Selection Gates + RoBERTa + DenseCapts	74.20

Table 3: Model Ablation: our dual-attention / frame-selection Gates, new loss functions, and dense captions help improve the model's performance (NewLoss: IOFSM+BBCE).

from each frame feature and our dual-attention can help this selection by getting more relevant information to each frame through the frame-level attention. Next, our new loss functions significantly help over the dual-attention and frame selection gates model by providing enhanced supervision (row 5 vs 4: $p < 0.0001$, row 7 vs 6: $p < 0.005$). Our RoBERTa version is also significantly better than the GloVe model (row 6 vs 4: $p < 0.0005$, row 7 vs 5: $p < 0.01$). Finally, employing dense captions further improves the performance via useful textual clue/keyword matching (row 8 vs 7: $p < 0.005$).[7]

[7]Statistical significance is computed using the bootstrap test (Efron and Tibshirani, 1994).

[8]Two more entries have appeared in the leaderboard since the ACL2020 submission deadline. However, our scores are still more balanced than their scores across all TV shows (std.: 2.11 and 2.40 versus our 0.65, max-min: 5.50 and 7.38 versus our 1.83).

	Loss	Val Score (%)	IFS		OFS	
			avg	std	avg	std
1	BCE	71.26	0.468	0.108	0.103	0.120
2	IOFSM	70.75	0.739	0.127	0.143	0.298
3	BCE+IOFSM	72.22	0.593	0.128	0.111	0.159
4	BBCE	72.27	0.759	0.089	0.230	0.231
5	BBCE+IOFSM	72.51	0.764	0.098	0.182	0.246

Table 4: IOFSM and BBCE help improve the model's performance by changing in and out-frame scores.

IOFSM and BCE Loss Functions Ablation and Analysis To see how In-and-Out Frame Score Margin (IOFSM) and Binary Cross-Entropy (BCE) loss affect the frame selection task, we compare the model's performance/behaviors according to the combination of IOFSM and BCE. As shown in Table 4, applying IOFSM on top of BCE gives a better result. When we compare row 1 and 3 in Table 4, the average in-frame score of BCE+IOFSM is higher than BCE's while the average out-frame scores of both are almost the same. This can mean two things: (1) IOFSM helps increase the scores of in-frames, and (2) increased in-frame scores help improve the model's performance. On the other hand, when we compare row 1 and 2, the average in-frame score of IOFSM is higher than BCE's. But, the average out-frame score of IOFSM is also much higher than BCE's. This can mean that out-frame scores have a large impact on the performance as well as in-frame scores. This is intuitively reasonable. Because information from out-frames also flows to the next layer (i.e., classifier) after being multiplied by the frame scores, the score for the 'negative' label also has a direct impact on the performance. So, making the scores as small as possible is also important. Also, when we compare the row 2 and others (2 vs. 1 and 3), the gap between in-frame scores is much larger than the gap between out-frame scores. But, considering the scores are average values, and the number of out-frames is usually much larger than in-frames,

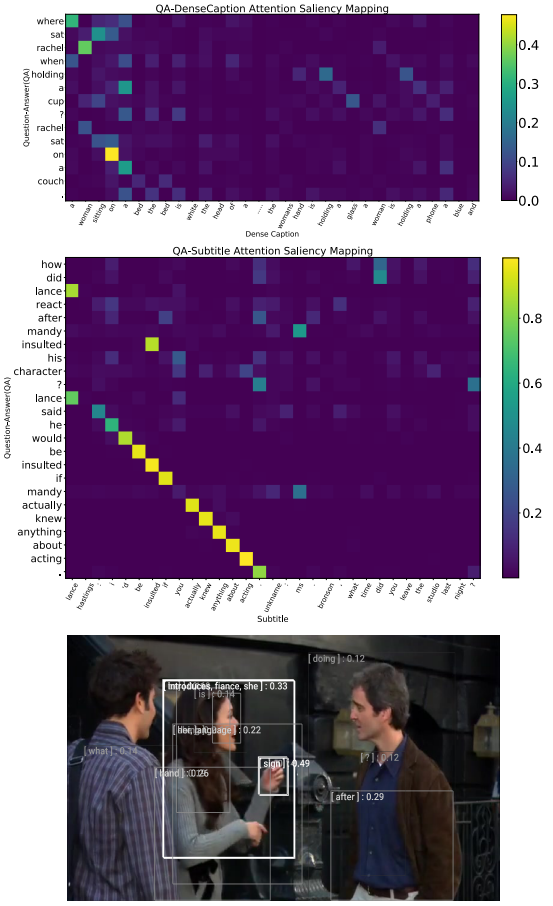
the difference between out-frame scores would affect more than the gap itself.

Balanced BCE Analysis We can see from row 1 and 4 of the Table 4 that BBCE shift the average scores of both in-frames and out-frames to higher values. This can show that scores from the BCE loss are biased to the negative examples, and BBCE can adjust the bias with the separate averaging. The score shift can help improve the model's performance. But, when comparing row 2 and 4, the out-frame scores of BBCE are higher than IOFSM, and this may imply that the result from BBCE should be worse than IOFSM since out-frame scores have a large impact on the performance. However, as we can see from row 2, the standard deviation of IOFSM's out-frame scores is larger than BBCE. This could mean that a model with IOFSM has an unstable scoring behavior, and it could affect the performance. As seen from row 5, applying BBCE and IOFSM together gives further improvement, possibly due to the increased in-frame scores and decreased out-frame scores while staying around at a similar standard deviation value.

6 Visualizations

In this section, we visualize the dual-level attention (word/object and frame level) and the frame score change by new losses application (for all these attention examples, our model predicts the correct answers).

Word/Object-Level Attention We visualize word-level attention in Figure 5. In the top example, the question and answer pair is "Where sat Rachel when holding a cup?" - "Rachel sat on a couch". Our word/object-level attention between QA pair and dense caption attend to a relevant description like 'holding a glass' to help answer the question. In the middle example, the question and answer pair is, "How did Lance react after Mandy insulted his character?" - "Lance said he would be insulted if Mandy actually knew anything about acting". Our word/object-level attention between QA pair and subtitle properly attend to the most relevant words such as 'insulted', 'knew', and 'acting' to answer the question. In the bottom example, the question and answer pair is, "What is Cathy doing with her hand after she introduces her fiance to Ted?" - "She is doing sign language". From the score of our word/object-level attention, the model aligns the word 'sign' to the woman's hand



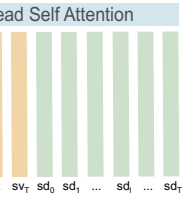
Q: What is Cathy doing with her hand after she introduces her fiance to Ted?
A: She is doing sign language.

Figure 5: Visualization of word/object level attention. Top: words from a question-answer pair to words from dense captions alignment. Middle: words from a question-answer pair to words from subtitles alignment. Bottom: words from a question-answer pair to regions (boxes) from an image (only boxes with top 1 scores from each word are shown).

to answer the question.

Frame-Level Attention As shown in Figure 6, our frame-level attention can align relevant frames from different features. In the example, the question and answer pair is "Where did Esposito search after he searched Carol's house downstairs?" - "Upstairs". To answer this question, the model needs to find a frame in which 'he (Esposito) searched Carol's house downstairs', then find a frame which has a clue for 'where did Esposito search'. Our frame-level attention can properly align the information fragments from different features (Frame 20 and 25) to help answer questions.

Frame Score Enhancement by New Losses As seen in Figure 7, applying our new losses (IOFSM+BBCE) changes the score distribution



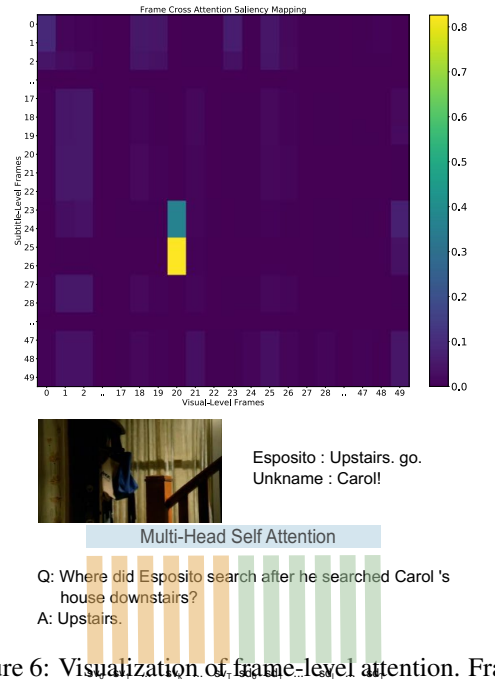


Figure 6: Visualization of frame-level attention. Frame 25 (which c[illegible] f[illegible]btitle features and frame 2[illegible]rs' by banister upward) from visual features are aligned.

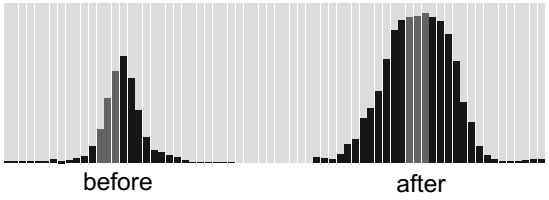
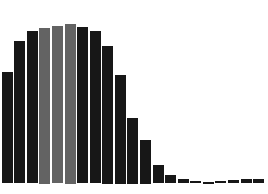


Figure 7: Visualization of distribution change in frame selection scores. Left: the score distribution before applying new losses (IOFSM+BBEC). Right: the score distribution after applying the losses. Scores neighboring in-frame (gray) are increased. For this example, the model does not predict the right answer before applying the losses, but after training with the losses, the model chooses the correct answer.

over all frames. Before applying our losses (left figure), overall scores are relatively low. After using the losses, overall scores increased, and especially, scores around in-frames get much higher.

7 Conclusion

We presented our dual-level attention and frame-selection gates model and novel losses for more effective frame-selection. Furthermore, we employed dense captions to help the model better find clues from salient regions for answering questions. Each component added to our base model architecture (proposed loss functions and the adoption of dense captions) significantly improves the model's performance. Overall, our model outperforms the state-of-the-art models on the TVQA leaderboard, while showing more balanced scores on the diverse TV show genres.



Acknowledgments

We thank the reviewers for their helpful comments. This work was supported by NSF Award 1840131, ARO-YIP Award W911NF-18-1-0336, DARPA KAIROS Grant FA8750-19-2-1004, and awards from Google and Facebook. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the funding agency.

References

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5803–5812.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen. 2017. Temporal context network for activity localization in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5793–5802.

Bradley Efron and Robert J Tibshirani. 1994. *An introduction to the bootstrap*. CRC press.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. Multimodal compact bilinear pooling for visual question answering and visual grounding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 457–468.

Adrien Gaidon, Zaid Harchaoui, and Cordelia Schmid. 2013. Temporal localization of actions with actoms. *IEEE transactions on pattern analysis and machine intelligence*, 35(11):2782–2795.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5267–5275.

Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Hyounghun Kim and Mohit Bansal. 2019. Improving visual question answering by referring to generated paragraph captions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. 2019b. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8337–8346.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2016–2022. AAAI Press.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015,San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. 2017. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 3337–3345. IEEE.

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. 2020. Tvqa+: Spatio-temporal grounding for video question answering. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Hui Li, Peng Wang, Chunhua Shen, and Anton van den Hengel. 2019. Visual question answering as reading comprehension. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. 2017. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3362–3371.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297.

Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893.

Luke Melas-Kyriazi, Alexander Rush, and George Han. 2018. Training for diversity in image paragraph captioning. *EMNLP*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *ICLR*.

Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. 2017. Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5734–5743.

Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. 2015. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE international conference on computer vision*, pages 3164–3172.

Jialin Wu, Zeyuan Hu, and Raymond Mooney. 2019. Generating question relevant captions to aid visual question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3585–3594.

Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer.

Linjie Yang, Kevin Tang, Jianchao Yang, and Li-Jia Li. 2017. Dense captioning with joint inference and visual context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2193–2202.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4995–5004.