Frame-Subtitle Self-Supervision for Multi-Modal Video Question Answering

Jiong Wang, Zhou Zhao, Weike Jin
College of Computer Science, Zhejiang University, Hangzhou, China
liubinggunzu, zhaozhou, weikejin@zju.edu.cn

Abstract

Multi-modal video question answering aims to predict correct answer and localize the temporal boundary relevant to the question. The temporal annotations of questions improve QA performance and interpretability of recent works, but they are usually empirical and costly. To avoid the temporal annotations, we devise a weakly supervised question grounding (WSQG) setting, where only QA annotations are used and the relevant temporal boundaries are generated according to the temporal attention scores. To substitute the temporal annotations, we transform the correspondence between frames and subtitles to Frame-Subtitle (FS) self-supervision, which helps to optimize the temporal attention scores and hence improve the video-language understanding in VideoQA model. The extensive experiments on TVQA and TVQA+ datasets demonstrate that the proposed WSQG strategy gets comparable performance on question grounding, and the FS self-supervision helps improve the question answering and grounding performance on both QA-supervision only and full-supervision settings.

Note that this paper was rejected by AAAI 2021 and we submit this version only for possible inspirations. We're not planning to release the codes.

Introduction

Video question answering (VideoQA) task (Jang et al. 2017; Kim et al. 2017) has consistently attracted considerable attention in recent years and it is challenging by requiring the understanding and reasoning on video and text modalities. Multi-modal VideoQA (Tapaswi et al. 2016; Lei et al. 2018, 2020) provides video frames and associate subtitles, which requires both visual and language understanding in videos to answer corresponding questions. Multi-modal VideoQA is general because the subtitles or dialogs are accessible from videos with the help of the automatic speech recognition (ASR) system to convert speech into text (Miech et al. 2019; Sun et al. 2019).

Recent multi-modal VideoQA works (Lei et al. 2018, 2020) extend VideoQA with spatio-temporal annotations and require intelligent systems to simultaneously retrieve relevant moments and detect referenced visual concepts to answer questions. The temporal annotations of questions improve the QA performance and interpretability of recent works (Kim, Tang, and Bansal 2020; Kim et al. 2020), but these annotations are usually empirical and imprecise. Even though (Lei et al. 2020) refined the original imprecise temporal annotations, they are still costly to obtain. In this way, a Weakly-Supervised Question Grounding (WSQG) setting with only question-answer annotations used is necessary to avoid costly annotation. The proposed WSQG strategy generates temporal boundaries from the temporal attention scores in VideoQA model, which is inspired by the recent weakly supervised temporal localization works (Nguyen et al. 2018; Shou et al. 2018). Considering the unique properties of question grounding, we additionally devise a proposal selection strategy to get complete temporal proposals.

To merge the multi-modal video inputs, most of existing works adopt two-stream structures, which separately encode contextual inputs (subtitles, frames) with QA pairs, then late-fusion with a classifier predicts the final answers. In each stream, (Lei et al. 2018) adopts bi-directional LSTM (BiLSTM) to encode both textual and visual sequences, then

---

Figure 1: Illustration of the frame-subtitle self-supervision. The adjacent subtitle pairs are connected to subject proposals, which are projected to temporal dimension of frames to get corresponding temporal boundaries.
Our contributions are threefold. Supervision to avoid costly temporal question annotations. A model hence guides video-language understanding in VideoQA. In training phrase, the temporal supervision is used to optimize the temporal attention scores and hence guides video-language understanding in VideoQA model.

In summary, we devise a weakly supervised question grounding (WSQG) strategy and adopt the FS self-supervision to avoid costly temporal question annotations. Our contributions are threefold:

- We propose a WSQG strategy, where question grounding supervision is not used and the relevant temporal boundaries are generated according to the temporal attention scores. The WSQG strategy can be applied to existing VideoQA models and directly reflects the interpretability.
- We transform the correspondence between video frames and subtitles to FS self-supervision for optimizing the temporal attention scores, thus guides video-language understanding in VideoQA model.
- Extensive experiments on two multi-modal VideoQA datasets show that the proposed WSQG method gets comparable performance on question grounding, and the FS self-supervision consistently improves question answering and grounding performance on QA-supervision only and full-supervision setting, and achieves new state-of-the-art performance.

**Related Works**

**Video Question Answering**

Video question answering (VideoQA) is an emerging research field and there are rapid development on datasets and techniques in recent years. Yapawi et al. (2016) proposed the MovieQA dataset and extend memory networks to VideoQA domain, following works (Na et al. 2017; Kim et al. 2019b) devise well-designed memory architectures to enable better interaction on different modalities. Jang et al. (2017) proposed the TGIF-QA dataset and combine deep appearance and motion features with spatial and temporal attention for accurate question answering.

Recent multi-modal VideoQA works (Lei et al. 2018, 2020) extend VideoQA with spatio-temporal annotations and require intelligent systems to simultaneously retrieve relevant moments and detect referenced visual concepts. These works prove that question temporal annotations improve accuracy and interpretability of QA models. Following works (Kim, Tang, and Bansal 2020; Kim et al. 2020) use caption as additional input sources and devise modality weighting strategy for better question answering performance. However, these works ignore the correspondence between frame and subtitles, which can serve as supervision to guide the language and video understanding in VideoQA model.

**Weakly-Supervised Temporal Action Localization**

Weakly supervised temporal action localization aims to localize multiple actions in videos with only video-level action labels given. The prevailing methods can be grouped into two categories: Top-down (Wang et al. 2017; Paul, Roy, and Roy-Chowdhury 2018) and Bottom-up (Nguyen et al. 2018; Shou et al. 2018) approaches. Most of existing bottom-up works follow the Temporal Class Activation Sequence (TCAS) baseline (Nguyen et al. 2018), where the class-specific attention score is predicted for each frame and consecutively salient attention scores are summarized to temporal boundaries. Recent works explore background modeling (Lee, Uh, and Byun 2020) and generative attention modeling (Shi et al. 2020). The proposed WSQG strategy is in a similar way where only QA annotation given and we summarize the temporal attention scores to temporal proposals. Considering the unique properties of question grounding, we devise a proposal selection strategy to avoid the local maximums in attention scores, and hence generates complete temporal proposals.

**Visual-Text Self-Supervision**

Recently, there emerge many works (Lu et al. 2019; Li et al. 2019; Su et al. 2020) that extend Bert (Devlin et al. 2019) in vision-language field as a powerful pre-trained model. These works adopt the image-caption self-supervision from the Conceptual Captions (Sharma et al. 2018) or COCO captions (Chen et al. 2015) dataset. As a representative of these works, ViLBERT (Lu et al. 2019) is pretrained on the Conceptual Caption dataset with masked multi-modal reconstruction and multi-modal alignment prediction.

In video domain, recent works resort to the frame-dialog self-supervision in narrated instructional videos, where the spoken words are more likely to refer to video content. HowTo100M (Miech et al. 2019) learns powerful video-language embedding with 1.36 million video clip-transcribed narration pairs. VideoBERT and ActBERT (Sun et al. 2019; Zhu and Yang 2020) build on the BERT model to learn bidirectional joint distributions over sequences of visual and linguistic tokens. These works pre-train powerful video-language joint models by matching video contents to corresponding dialogs, and show excellent generalization capability on downstream video-language tasks.

In this paper, we transform the correspondence between video frames and subtitles to the frame-subtitle self-supervision, which temporally grounds the subtitles to video frames. Different from the above works adopt visual-text self-supervision for pre-training, the proposed video FS self-supervision for pre-training, the proposed video FS self-supervision consistently improves question answering performance. However, these works ignore the correspondence between frame and subtitles, which can serve as supervision to guide the language and video understanding in VideoQA model.
Supervision is adopted to complement the QA supervision for accurate question answering and grounding.

**Proposed Method**

**Preliminary**

In multi-modal VideoQA, an input instance includes a question $q$ with 5 candidate answers $\{a_k\}_{k=1}^5$, video frames and associated subtitles with timestamps. The goal is to predict correct answer and ground the QA pair temporally. The question and answers are concatenated to 5 QA pairs (hypothesis) $h_k = [q, a_k]$.

For each frame, Faster R-CNN (Ren et al. 2017) pre-trained on Visual Genome (Krishna et al. 2017) is used to detect objects and extract their region embeddings as the video frame features. The top-20 object proposals are preserved and dimension-reduced from 2048 to 300 by PCA. BERT word embeddings with 768 dimensions are used to embed QA pairs and subtitle sentences. For each frame, we pair it with two neighboring subtitle sentences based on the subtitle timestamps. In this way, the resulted aligned subtitles are convenient for late fusion with video frame features. We pair adjacent subtitles to subtitle proposals as illustrated in Figure 1.

Both the object-level embeddings and the word-level embeddings are then projected into a 128-dimensional space using a linear layer with ReLU activation. We adopt convolutional encoder with positional encoding to get the encoded frame, subtitle and QA features with shape preserving following (Lei et al. 2020; Kim, Tang, and Bansal 2020). We denote the number of words in each hypothesis as $L_h$, the aligned subtitles as $L_s$ and the subtitle proposals as $L_{sp}$ respectively. We use $N_o$ as the number of object regions in a frame and $T$, $T_{sp}$ as the frame number, subtitle proposal number respectively. Thus we get encoded QA features $H \in \mathbb{R}^{5 \times L_h \times 128}$, video frame features $V \in \mathbb{R}^{T \times N_o \times 128}$, subtitle features $S \in \mathbb{R}^{T \times L_s \times 128}$ and subtitle proposal features $S_{sp} \in \mathbb{R}^{T_{sp} \times L_{sp} \times 128}$.

**Model**

The pipeline of our model is inspired by recent multi-modal VideoQA works (Lei et al. 2020; Kim, Tang, and Bansal 2020) with well-designed cross-attention mechanism. We take one QA pair as an example for clarity as shown in Figure 2. The upper part illustrates the baseline method where the two-stream architecture with late fusion is adopted for question answering and grounding, which is supervised by the QA labels and question temporal labels ($T$-label).

In the bottom part, the subtitles are transformed to subtitle proposals with corresponding temporal labels (Sub prop label), which optimize the attention scores in temporal attention module. These two parts share some weights which are optimized by three kinds of supervision.

Concretely in each stream of the upper part, the inputs are query (QA or subtitle proposal) and context (video and subtitle) features. The dual-level attention module is adopted to capture the bi-directional feature interaction on word-level and frame-level.

**Word-level attention.** Taking one encoded video frame feature $v \in \mathbb{R}^{T \times N_o \times 128}$ and one QA pair feature $h \in \mathbb{R}^{L_h \times 128}$ as an example, we apply word-level attention where a similarity matrix is first calculated for all word-region pairs as

$$Sim_v = hv^T, Sim_v \in \mathbb{R}^{L_h \times N_o}.$$  \hspace{1cm} (1)

Then $Sim_v$ and $Sim_v^T$ are respectively multiplied on $v$ and $h$, and max pooling is applied on word or region-level to reduce dimension as follow.

$$v_{att} = \max(Sim_v v), v_{att} \in \mathbb{R}^{128}. \hspace{1cm} (2)$$

$$h_{att} = \max(Sim_v^T h), h_{att} \in \mathbb{R}^{128}. \hspace{1cm} (3)$$
The resulted attentive features are fused together to get the attentive frame feature \( v_f \in \mathbb{R}^{128} \) as follow.

\[
v_f = (\{v_{att}; h_{att}; v_{att} \odot h_{att}; v_{att} + h_{att}\})W_1 + b_1,
\]

where \( W_1 \) and \( b_1 \) are parameters to reduce the 512-d fused features to 128-d. The whole attentive frame features are \( V_f \in \mathbb{R}^{T \times 128} \). In similar way, we get the attentive subtitle features \( S_f \in \mathbb{R}^{T \times 128} \) in another stream.

**Frame-level attention.** After the word-level attention, we get the attentive frame feature \( V_f \) and subtitle features \( S_f \). Similar to word-level attention, a similarity matrix \( Sim_f \in \mathbb{R}^{T \times T} \) is calculated on frame-level.

\[
Sim_f = S_f V_f^T, Sim_f \in \mathbb{R}^{T \times T}.
\]

Then \( Sim_f \) and \( Sim_f^T \) are separately multiplied on \( S_f \) and \( V_f \) as follow.

\[
V_{fatt} = Sim_f V_f, V_{fatt} \in \mathbb{R}^{T \times 128}.
\]

\[
S_{fatt} = Sim_f S_f, S_{fatt} \in \mathbb{R}^{T \times 128}.
\]

The resulted frame-wise attentive features are fused to get the fused features \( F \in \mathbb{R}^{T \times 128} \) as follow.

\[
F = (\{V_{fatt}; S_{fatt}; V_{fatt} \odot S_{fatt}; V_{fatt} + S_{fatt}\})W_2 + b_2,
\]

where \( W_2 \) and \( b_2 \) are parameters for dimension reduction to 128-d.

**Temporal attention module.** Getting the fused features \( F \), a temporal attention (T-Att) module, consisting of a fully-connected layer and sigmoid function, is applied to get the attention scores \( A \in \mathbb{R}^T \) which identify which frames are relevant to questions. Then we get the attentive fused features by multiplying \( A \) on \( F \). Note that these attention scores are used to generate temporal proposals in the proposed WSQG strategy when testing.

Finally, the attentive fused features are feed to answer predictor and span predictor to get the answer score and temporal boundary, respectively. Concretely, the span predictor predicts the start and end probabilities on temporal dimension, and is optimized by the temporal question supervision as \( \text{Lei et al. 2020} \) do. Given the softmax normalized start and end probabilities \( p_{st} \) and \( p_{ed} \), we use cross-entropy loss as follow.

\[
loss_{span} = -\frac{1}{2} (\log p_{st} + \log p_{ed}).
\]

The QA loss is calculated as

\[
loss_{qa} = -\log \left( \frac{e^{p_a}}{\sum_{k=1}^{5} e^{p_k}} \right),
\]

where \( p_y \) is predicted answer score for the ground-truth index.

**Video Frame-Subtitle Self-Supervision**

As illustrated in Figure 2, the adjacent subtitles are connected to subtitle proposals. The reason is only the start-times of subtitle are given in the meta-data, so we connect adjacent subtitles to get their rough start-end time-span.

In the bottom part of Figure 2, we replace the QA features with subtitle proposals and feed them to dual-level attention, fusion and T-Att modules with contextual features. The predicted temporal scores \( A \in \mathbb{R}^T \) are directly supervised by the subtitle proposal labels (Sub prop label).

Note that the word-level attention in the bottom part is slightly different from the upper part, the reason is we don’t want to introduce more computations on the baseline method.

Before the dual-level attention, we first sum the word and region-level subtitle, video and subtitle proposal features to frame-level features. So the word-level attention module actually capture coarsely frame-level interaction between subtitle proposals and contextual features. The follow-up frame-level attention is same as the upper part.

We found this modification doesn’t cause performance drop when testing, and only slightly increases the computations on baseline part because the computation of word-level attention is much larger than the frame-level counterpart.

The predicted temporal attention scores \( A \in \mathbb{R}^T \) reflect which frames are relevant to the subtitle proposals. To calculate the loss between the A and Sub prop label, a ranking loss and a Binary Cross-Entropy (BCE) loss are complemented for optimization.

The ranking loss forces the average attention scores inside the temporal boundary \( (A_{in}) \) to be larger than the outside counterparts \( (A_{out}) \) as follow.

\[
loss_{rank} = 1 + \text{avg}(A_{out}) - \text{avg}(A_{in}).
\]

The BCE loss enforces the attention scores of frames inside the temporal boundary to be 1 and outside to be 0, which is defined as follow.

\[
loss_{bce} = -\frac{1}{T_{in}} \sum_{i=1}^{T_{in}} (\log A_{in}^i) - \frac{1}{T_{out}} \sum_{j=1}^{T_{out}} (\log(1 - A_{out}^j)),
\]

where \( T_{in} \) and \( T_{out} \) are the number of frames inside and outside the temporal boundary. The total self-supervision loss is defined as

\[
loss_{self} = loss_{ro} + loss_{bce}.
\]

In this way, the FS self-supervision guides the temporal attention module to generate reasonable attention scores for language queries and helps the VideoQA model better understand the video-language interaction. In practice, it improves the performance of answer predictor ans span predictor in the upper part, demonstrating it is complementary to other two kinds of supervision.

**Weakly-Supervised Question Grounding (WSQG)**

To generate consecutive proposals from temporal attention scores, we propose a WSQG strategy which is inspired by the weakly-supervised temporal action localization works \( \text{Nguyen et al. 2018} \), \( \text{Shou et al. 2018} \). But different from these works which localize multiple actions with multiple proposals, we aim to find the most relevant temporal proposal to the question. So we further devise a proposal selection method, which refines the proposal scores with a length-
Figure 3: Proposal generation and refinement process from attention scores. The red dashed line denotes threshold.

We expect the proposal with highest score to be the temporal grounding result.

In practice we found that the proposal scores will be dominate by some local maximums with quite short temporal boundary, as illustrates by the third proposal of Figure 3. So we devise length-aware score refinement strategy where the proposal scores are refined by multiplying a proposal length-aware term as follow:

\[ A_P' = A_P \times (ed - st)^\alpha, \]

where \( \alpha \) is a parameter to control the effect of proposal length, and we set it to 0.5 in our configuration. Then the proposal with highest refined score is selected as final prediction.

Note that the temporal attention module is a plug-to-play module and is applicable to existing VideoQA models, so the proposed WSQG strategy is suit for existing VideoQA models for question grounding to reflect the interpretability.

Four Level of Supervision Settings

Together with the proposed FS self-supervision, there are three kinds of supervision for optimizing the VideoQA model. As shown in Table 1, we incrementally summarize four level of supervision settings, including QA supervision only (QA only), QA supervision and self-supervision (QA + self), QA and question grounding (QG) supervision (Full), full supervision and self-supervision (Full + self). The QA supervision is adopted without exception and question answering is consistent in training and testing phrase for four settings, while question grounding is not. In QA only and QA + self settings, where the temporal supervision is not adopted, the proposed WSQG strategy is used to get temporal proposals from attention scores. In Full and Full + self settings, the temporal proposals are generated from the span predictor in Figure 2. Specifically in the strongest level, the Full + self setting, three kinds of supervision are adopted and the total training loss is

\[ \text{loss} = \text{loss}_{qa} + \lambda_1 \text{loss}_{span} + \lambda_2 \text{loss}_{self}. \]

where \( \lambda_1 \) and \( \lambda_2 \) are the hyper-parameters to control the balance of three losses.

Experiments

Datasets

TVQA. TVQA (Lei et al. 2018) is a large-scale multi-modal video QA dataset based on 6 popular TV shows spanning 3 genres: medical dramas, sitcoms, and crime shows. It consists of 152.5K QA pairs from 21.8K video clips, spanning over 460 hours of video. The training set consists of 122,039 QAs from 17,435 clips, the validation set 15,252 QAs from 2,179 clips and test set consist of 7,623 QAs from 1,089 clips. The video contents consist of video frames and subtitles. The multiple-choices form questions-answer pairs are human-annotated and designed to be compositional. TVQA dataset requiring systems to jointly comprehend subtitles-based dialogue to correctly answer questions and localize relevant moments within video frames.

TVQA+. TVQA+ (Lei et al. 2020) is a subset of the TVQA dataset and augments TVQA dataset with frame-level bounding box annotations, which links video objects to visual concepts in questions and answers. TVQA+ additionally refine the temporal boundary so the temporal annotations are more precise than TVQA. TVQA+ is also the first dataset that combines moment localization, object grounding and question answering. In this paper, we only consider the temporal grounding on TVQA+ dataset.

Implementation Details

The training and testing pipeline on TVQA and TVQA+ datasets follow previous works (Lei et al. 2018, 2020). The frames are extracted at 0.5 FPS and the questions are grounded in frames’ temporal dimension. RoBERTa (Liu et al. 2019) is used to embed words in QA pairs and subtitle on TVQA (Kim, Tang, and Bansal 2020) and BERT word embeddings are used on TVQA+ (Lei et al. 2020). The Adam optimizer is used with batch size of 16, initial
Table 2: Comparison with state-of-the-art works on TVQA validation set in QA supervision and full supervision setting. * denotes the results are implemented by us.

| Method               | QA supervision | T.mIoU | ASA |
|----------------------|----------------|--------|-----|
| Two-stream (Lei et al. 2018) | 67.70          | -      | -   |
| PAMN (Kim et al. 2019b) | 66.22          | -      | -   |
| Multi-task (Kim et al. 2019a) | 66.38         | -      | -   |
| QA only*             | 69.13          | 29.28  | 20.78 |
| QA + Self*           | 72.13          | 33.60  | 25.71 |

| Method               | QA supervision | T.mIoU | ASA |
|----------------------|----------------|--------|-----|
| STAGE (Lei et al. 2020) | 70.50          | -      | -   |
| MSAN (Kim et al. 2020)  | 71.13          | 30     | -   |
| DenseCap (Kim, Tang et al. 2020) | 73.34        | -      | -   |
| Full*                 | 73.59          | 39.80  | 29.66 |
| Full + Self*          | 74.20          | 40.49  | 30.88 |

Table 3: Comparison with state-of-the-art works on TVQA+ validation set in QA supervision and full supervision setting.

| Method               | QA supervision | T.mIoU | ASA |
|----------------------|----------------|--------|-----|
| Two-stream (Lei et al. 2018) | 62.28          | -      | -   |
| STAGE (Lei et al. 2020)  | 68.31          | -      | -   |
| QA only*             | 67.88          | 30.30  | 21.98 |
| QA + Self*           | 69.47          | 32.03  | 23.57 |

Table 4: Ablation of WSQG strategy for question grounding with different settings on TVQA val set.

| Method               | R@1, IoU = 0.3 | 0.5  | 0.7  |
|----------------------|----------------|------|------|
| WSQG w/o refinement | 43.26          | 24.17 | 9.55 |
| WSQG (QA only)      | 45.41          | 24.40 | 9.57 |
| WSQG (QA + Self)    | 53.26          | 30.67 | 11.54 |

Table 5: Comparison with state-of-the-art works on TVQA+ val set.

Evaluation Metric
The experimental evaluation follows previous works (Lei et al. 2018, 2020). The question answering performance is measured by classification accuracy (Acc.). Same as language-guided video moment retrieval works (Gao et al. 2017), Temporal mean Intersection-over-Union (T. mIoU) is adopted to evaluate question grounding performance. We also report the R@1, IoU=m score, which measures the percentage of the top-1 predicted temporal boundaries having IoU with ground-truth larger than m. The Answer-Span joint Accuracy (ASA) is used to jointly evaluate both answer prediction and span prediction, where a prediction is considered correct only when the predicted span has an IoU ≥ 0.5 with the GT span and the answer prediction is correct.

Comparison with State-of-the-Arts
Even though TVQA and TVQA+ datasets require both question answering (QA) and grounding (QG), some recent works only consider question answering. The most comparable works are MSAN (Kim et al. 2020) and STAGE (Lei et al. 2020) which consider both QA and QG. In this subsection, we give comprehensive comparison with these works.

TVQA. In Table 2 we compare with existing works with different settings on the TVQA dataset. It is clear that our “QA only” setting outperforms existing counterpart works, the reason is we adopt dual-level attention mechanism and RoBERTa word embeddings following (Kim, Tang, and Bansal 2020). When paired with the proposed FS self-supervision, the “QA + Self” setting gets considerable increments on QA and QG performance, even outperforms MSAN which adopts full supervision.

In the bottom part, we can see that our full supervision setting slightly outperforms the State-of-the-art works on QA and largely outperforms them on QG task. The “Full + Self” setting achieves new state-of-the-art performance on QA and QG, demonstrating that the proposed FS self-supervision is complementary with the original question answering and grounding supervision. It can also be seen that the increments on QA only setting are more distinct, so we expect the FS self-supervision will be a substitute to the costly temporal annotations, and be intensively explored in the future works.

TVQA+. In Table 3 we compare with existing works in different settings on the TVQA+ dataset. It can be seen that our QA performance on QA only setting is slightly low, but the QG performance with the proposed WSQG strategy used is comparable to STAGE with full supervision. When paired with self-supervision, the “QA + Self” setting gets consistent improvement on QA and QG accuracy, surpassing counterpart works.

In the bottom part, the Full setting largely exceeds STAGE on QG performance. When paired with FS self-supervision, “Full + Self” setting gets consistent increments and achieves new state-of-the-art results on three metrics.

Ablation Study
In this subsection, we give ablation study on the proposed WSQG strategy and the FS self-supervision. In Table 4 the QG performances of WSQG in different settings are reported. As can be seen, the proposed proposal score refinement strategy helps WSQG improve the R@1, IoU=m and T.mIoU scores. It can also be seen that the QA + Self setting gets largely increments on the grounding performance, which is benefited from the FS self-supervision.

Note that we connect adjacent two subtitles to subtitle proposals, and it is natural to explore proposals with longer scale (more than two subtitles in one proposal) or multi-scale. We give ablation study on different scale of subtitle
Figure 4: Qualitative results on TVQA val set. The question-answer pairs and relevant subtitles are provided in the red boxes. The frame inside ground-truth (GT) span is highlighted with red box bounded. The attention scores, grounding results of WSQG with QA only and QA + Self supervision settings are provided for comparison with GT.

Table 5: Comparison of different scale of subtitle proposals on TVQA validation set. Scale (3) denotes three adjacent subtitles are connected to one subtitle proposal.

| Scale     | Acc.  | T.mIoU | ASA  |
|-----------|-------|--------|------|
| Scale (2) | 72.13 | 33.60  | 25.71|
| Scale (3) | 71.76 | 33.42  | 24.75|
| Scale (4) | 71.25 | 33.26  | 24.43|
| Scale (2, 3, 4) | 71.78 | 33.31  | 24.54|

First, the attention scores of QA only setting sometimes response on different time-spans, in which some may be falsely matched (Example 1, 3). The scores of QA + Self, by contrast, are more concentrated on the relevant frames. We suppose the reason is training with Self-supervision forces the VideoQA model to focus on the most relevant frames while suppress the irrelevant ones, which benefits both question answering and grounding. Second, we found some questions are ambiguous to determine their temporal boundaries, and the annotators may empirically focus on some contents. As shown in example 4, given the question, it is not clear enough whether to localize the time-span where “Sheldon is angry” or “Sheldon is talking with Penny”. The VideoQA models tend to localize the temporal spans where “Sheldon is talking with Penny” and “Sheldon is angry”, while the GT only cover the span where “Penny say it is a little idea”. It reflects that the temporal annotations may be empirical to different annotators who focus on different contents.

Qualitative Analysis

Figure 4 illustrates qualitative results of the proposed WSQG strategy and the FS self-supervision on TVQA dataset. The proposed WSQG strategy helps to summarize temporal attention scores to temporal proposals, and there are two observations can be summarized.
Conclusion

In this paper, we consider question answering and grounding in multi-modal VideoQA. We devise the frame-subtitle self-supervision and a weakly supervised question grounding strategy in multi-modal VideoQA, to alleviate the empirical and costly temporal annotations. The WSQG strategy is applicable to existing VideoQA models to reflect the interpretability. The FS self-supervision helps to regular the temporal attention scores, thus guide video-language understanding in VideoQA models. It improves question answering and grounding performance in both QA only and full supervision setting, and achieves new state-of-the-art results.

References

Chen, X.; Fang, H.; Lin, T.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. CoRR abs/1504.03325.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT (1), 4171–4186. ACL.

Gao, J.; Sun, C.; Yang, Z.; and Nevatia, R. 2017. Tall: Temporal activity localization via language query. In ICCV, 5267–5275.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In CVPR, 2758–2766.

Kim, H.; Tang, Z.; and Bansal, M. 2020. Dense-Caption Matching and Frame-Selection Gating for Temporal Localization in VideoQA. ACL.

Kim, J.; Ma, M.; Kim, K.; Kim, S.; and Yoo, C. D. 2019a. Gaining Extra Supervision via Multi-task learning for Multi-Modal Video Question Answering. In IJCNII, 1–8. IEEE.

Kim, J.; Ma, M.; Kim, K.; Kim, S.; and Yoo, C. D. 2019b. Progressive attention memory network for movie story question answering. In CVPR, 8337–8346.

Kim, J.; Ma, M.; Pham, T. X.; Kim, K.; and Yoo, C. D. 2020. Modality Shifting Attention Network for Multi-Modal Video Question Answering. In CVPR. IEEE.

Kim, K.; Heo, M.; Choi, S.; and Zhang, B. 2017. DeepStory: Video Story QA by Deep Embedded Memory Networks. In IJCAI, 2016–2022. ijcai.org.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. Int. J. Comput. Vis. 123(1): 32–73.

Lee, P.; Uh, Y.; and Byun, H. 2020. Background Suppression Network for Weakly-Supervised Temporal Action Localization. In AAAI, 11320–11327. AAAI Press.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. Tvqa: Localized, compositional video question answering. EMNLP.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020. TVQA+: Spatio-temporal grounding for video question answering. ACL.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.; and Chang, K. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. CoRR abs/1908.03557.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR abs/1907.11692.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In NeurIPS, 13–23.

Miech, A.; Zhukov, D.; Alayrac, J.; Tapaswi, M.; Laptev, I.; and Sivic, J. 2019. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In ICCV, 2630–2640. IEEE.

Na, S.; Lee, S.; Kim, J.; and Kim, G. 2017. A Read-Write Memory Network for Movie Story Understanding. In IJCNN, 677–685. IEEE Computer Society.

Nguyen, P.; Liu, T.; Prasad, G.; and Han, B. 2018. Weakly Supervised Action Localization by Sparse Temporal Pooling Network. In CVPR, 6752–6761. IEEE Computer Society.

Paul, S.; Roy, S.; and Roy-Chowdhury, A. K. 2018. W-TALC: Weakly-Supervised Temporal Activity Localization and Classification. In ECCV, volume 11208, 588–607.

Ren, S.; He, K.; Girshick, R. B.; and Sun, J. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6): 1137–1149.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual Captions: A Cleaned, Hypervuned, Image Alt-text Dataset For Automatic Image Captioning. In ACL, 2556–2565. ACL.

Shi, B.; Dai, Q.; Mu; Y.; and Wang, J. 2020. Weakly-Supervised Action Localization by Generative Attention Modeling. In CVPR, 1006–1016. IEEE.

Shou, Z.; Gao, H.; Zhang, L.; Miyazawa, K.; and Chang, S. 2018. AutoLoc: Weakly-Supervised Temporal Action Localization in Untrimmed Videos. In ECCV, volume 11220, 162–179.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In ICLR. OpenReview.net.

Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; and Schmid, C. 2019. VideoBERT: A Joint Model for Video and Language Representation Learning. In IJCV, 7463–7472. IEEE.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. Movieqa: Understanding stories in movies through question-answering. In CVPR, 4631–4640.

Wang, L.; Xiong, Y.; Lin, D.; and Gool, L. V. 2017. UntrimmedNets for Weakly Supervised Action Recognition and Detection. In CVPR, 6402–6411. IEEE Computer Society.

Zhu, L.; and Yang, Y. 2020. ActBERT: Learning Global-Local Video-Text Representations. In CVPR, 8743–8752. IEEE.