DramaQA: Character-Centered Video Story Understanding with Hierarchical QA

Seongho Choi,[1] **Kyoung-Woon On,**[1] **Yu-Jung Heo,**[1]
Ahjeong Seo,[1] **Youwon Jang,**[1] **Minsu Lee,**[1] **Byoung-Tak Zhang**[1,2]
[1] Seoul National University
[2] AI Institute (AIIS)
{shchoi,kwon,yjheo,ajseo,ywjang,mslee,btzhang}@bi.snu.ac.kr

Abstract

Despite recent progress on computer vision and natural language processing, developing a machine that can understand video story is still hard to achieve due to the intrinsic difficulty of video story. Moreover, researches on how to evaluate the degree of video understanding based on human cognitive process have not progressed as yet. In this paper, we propose a novel video question answering (Video QA) task, DramaQA, for a comprehensive understanding of the video story. The DramaQA focuses on two perspectives: 1) Hierarchical QAs as an evaluation metric based on the cognitive developmental stages of human intelligence. 2) Character-centered video annotations to model local coherence of the story. Our dataset is built upon the TV drama "Another Miss Oh"[1] and it contains 17,983 QA pairs from 23,928 various length video clips, with each QA pair belonging to one of four difficulty levels. We provide 217,308 annotated images with rich character-centered annotations, including visual bounding boxes, behaviors and emotions of main characters, and coreference resolved scripts. Additionally, we suggest Multi-level Context Matching model which hierarchically understands character-centered representations of video to answer questions. We release our dataset and model publicly for research purposes[2], and we expect our work to provide a new perspective on video story understanding research.

Introduction

Stories have existed for a long time with the history of mankind, and always fascinated humans with enriching multimodal effects from novels to cartoons, plays, and films. The story understanding ability is a crucial part of human intelligence that sets humans apart from others (Szilas 1999; Winston 2011).

To take a step towards human-level AI, *drama*, typically in the form of video, is considered as proper mediums because it is one of the best ways to covey a story. Also, the components of drama including image shots, dialogues, sound effects, and textual information can be used to build artificial ability to see, listen, talk, and respond like humans. Since drama closesly describes our everyday life, the contents of drama also help to learn realistic models and patterns of humans' behaviors and conversations. However, the causal and temporal relationships between events in drama are usually complex and often implicit (Riedl 2016). Moreover, the multimodal characteristics of the video make the problem trickier. Therefore, video story understanding has been considered as a challenging machine learning task.

One way to enable a machine to understand a video story is to train the machine to answer questions about the video story (Schank and Abelson 2013; Mueller 2004). Recently, several video question answering (Video QA) datasets (Tapaswi et al. 2016; Kim et al. 2017; Mun et al. 2017; Jang et al. 2017; Lei et al. 2018) have been released publicly. These datasets encourage inspiring works in this domain, but they do not give sufficiently careful consideration of some important aspects of video story understanding. Video QA datasets can be used not only for developing video story understanding models but also for evaluating the degree of intelligence of the models. Therefore, QAs should be collected considering difficulty levels of the questions to evaluate the degree of story understanding intelligence (Collis 1975). However, the collected QAs in the previous studies are highly-biased and lack of variance in the levels of question difficulty. Furthermore, while focalizing on characters within a story is important to form local story coherence (Riedl and Young 2010; Grosz, Weinstein, and Joshi 1995), previous works did not provide any consistent annotations for characters to model this coherence.

In this work, we propose a new Video QA task, DramaQA, for a more comprehensive understanding of the video story. 1) We focus on the *understanding* with hierarchical QAs used as a hierarchical evaluation metric based on the cognitive-developmental stages of human intelligence. We define the level of understanding in conjunction with Piaget's theory (Collis 1975) and collect QAs accordingly. In accordance with (Heo et al. 2019), we collect questions along with one of four hierarchical difficulty levels, based on two criteria; memory capacity (MC) and logical complexity (LC). With these hierarchical QAs, we offer a more sophisticated evaluation metric to measure understanding levels of Video QA models. 2) We focus on the *story* with character-centered video annotations. To learn character-centered video representations, the DramaQA provides rich

[1]We have received an official permission to use these episodes for research purposes from the content provider.

[2]https://dramaqa.snu.ac.kr

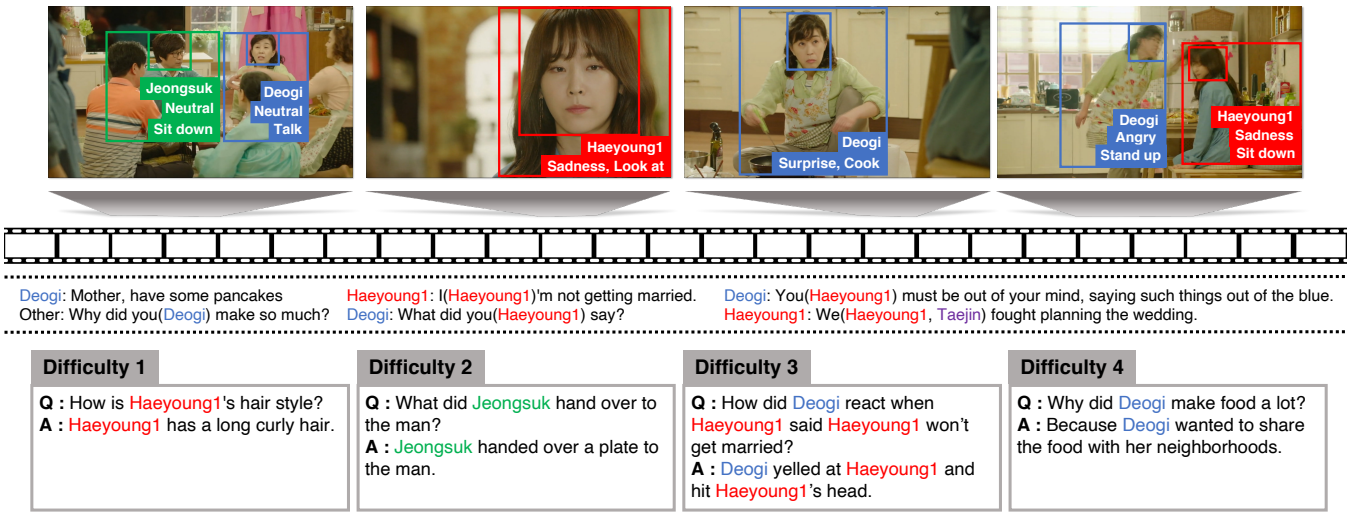

Figure 1: An example of DramaQA dataset which contains video clips, scripts, and QA pairs with levels of difficulty. A pair of QA corresponds to either a shot or a scene, and each QA is assigned one out of possible four stages of difficulty (details in Section DramaQA Dataset). A video clip consists of a sequence of images with visual annotations centering the main characters.

annotations for main characters such as visual bounding boxes, behaviors and emotions of main characters and also coreference resolved scripts. By sharing character names for all the annotations including QAs, the model can have a coherent view of characters in the video story. 3) We provide Multi-level Context Matching model to answer the questions for the multimodal story by utilizing the character-centered annotations. Using representations of two different abstraction levels, our model hierarchically learns underlying correlations between the video clips, QAs, and characters.

Related Work

This section introduces Question and Answering about Story and Cognitive Developmental Stages of Humans. Because of the page limit, we inroduce Video Understanding in appendix A.

Question and Answering about Story

Question and answering (QA) has been commonly used to evaluate reading comprehension ability of textual story. (Hermann et al. 2015; Trischler et al. 2016) introduced QAs dataset about news articles or daily emails, and (Richardson, Burges, and Renshaw 2013; Hill et al. 2016) dealt with QAs built on children's book stories. Especially, NarrativeQA suggested by (Kočiskỳ et al. 2018) aims to understand the underlying narrative about the events and relations across the whole story in book and movie scripts, not the fragmentary event. (Mostafazadeh et al. 2016) established ROCStories capturing a set of causal and temporal relations between daily events, and suggested a new commonsense reasoning framework 'Story Cloze Test' for evaluating story understanding.

Over the past years, increasing attention has focused on understanding of multimodal story, not a textual story. By exploiting multimodalities, the story delivers the more richer semantics without ambiguity. (Tapaswi et al. 2016; Kim et al. 2017; Mun et al. 2017; Jang et al. 2017; Lei et al. 2018) considered the video story QA task as an effective tool for multimodal story understanding and built video QA datsets. The more details on the comparison of those datasets with the proposed dataset, DramaQA, are described in the section titled 'Comparison with Other Video QA Datasets.'

Cognitive Developmental Stages of Humans

We briefly review the cognitive development of humans based on Piaget's theory (Piaget 1972; Collis 1975) that is a theoretical basis of our proposed hierarchical evaluation metric for video story understanding. Piaget's theory explains in detail the developmental process of human cognitive ability in conjunction with information processing.

At *Pre-Operational Stage* (4 to 6 years), a child thinks at a symbolic level, but is not yet using cognitive operations. The child can not transform, combine or separate ideas. Thinking at this stage is not logical and often unreasonable. At *Early Concrete Stage* (7 to 9 years), a child can utilize only one relevant operation. Thinking at this stage has become detached from instant impressions and is structured around a single mental operation, which is a first step towards logical thinking. At *Middle Concrete Stage* (10 to 12 years), a child can think by utilizing more than two relevant cognitive operations and acquire the facts of dialogues. This is regarded as the foundation of proper logical functioning. However, the child at this stage lacks own ability to identify general fact that integrates relevant facts into coherent one. At *Concrete Genelization Stage* (13 to 15 years), a child can think abstractly, but just generalize only from personal and concrete experiences. The child does not have own ability to hypothesize possible concepts or knowledge that is quite abstract. *Formal Stage* (16 years onward) is characterized purely by abstract thought. Rules can be integrated to obtain novel results that are beyond the individual's own personal experi-

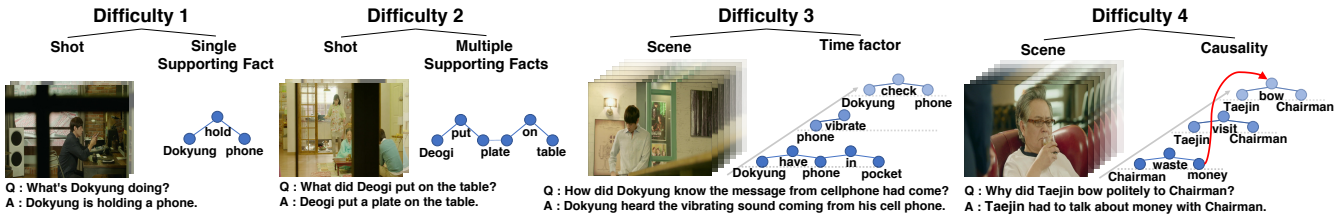

Figure 2: Four examples of different QA level. Difficulty 1 and 2 target shot-length videos. Difficulty 1 requires single supporting fact to answer, and Difficulty 2 requires multiple supporting facts to answer. Difficulty 3 and 4 require a time factor to answer and target scene-length videos. Especially, Difficulty 4 requires causality between supporting facts from different time.

ences. In this paper, we carefully design the evaluation metric for video story understanding with the question-answer hierarchy for levels of difficulty based on the cognitive developmental stages of humans.

DramaQA Dataset

We collect the dataset on a popular Korean drama *Another Miss Oh*, which has 18 episodes, 20.5 hours in total. DramaQA dataset contains 23,928 various length video clips which consist of sequences of video frames (3 frames per second) and 17,983 multiple choice QA pairs with hierarchical difficulty levels. Furthermore, it includes rich character-centered annotations such as visual bounding boxes, behaviors and emotions of main characters, and coreference resolved scripts. Figure 1 illustrates the DramaQA dataset. Also, detailed information of the dataset including various attributes, statistics and collecting procedure can be found in Appendix B.

Question-Answer Hierarchy for Levels of Difficulty

To collect question-answer pairs with levels of difficulty, we propose two criteria: Memory capacity and logical complexity. Memory capacity (MC) is defined as the required length of the video clip to answer the question, and corresponds to working memory in human cognitive process. Logical complexity (LC) is defined by the number of logical reasoning steps required to answer the question, which is in line with the hierarchical stages of human development (Seol, Sharp, and Kim 2011).

Criterion 1: Memory Capacity The capacity of working memory increases gradually over childhood, as does cognitive and reasoning ability required for higher level responses (Case 1980; Mclaughlin 1963; Pascual-Leone 1969). In the Video QA problem, the longer video story to answer a question requires, the harder to reason the answer from the video story is. Here, we consider two levels of memory capacity; shot and scene. Detailed definitions of each level are below:

- **Level 1 (shot):** The questions for this level are based on video clips mostly less than about 10 seconds long, shot from a single camera angle. This set of questions can contain atomic or functional/meaningful action in the video. Many Video QA datasets belong to this level (Jang et al. 2017; Maharaj et al. 2017; Mun et al. 2017; Kim et al. 2017).
- **Level 2 (scene):** The questions for this level are based on clips that are about 1-10 minutes long without changing location. Videos at this level contain sequences of actions, which augment the shots from Level 1. We consider this level as the "story" level according to our working definition of story. MovieQA (Tapaswi et al. 2016) and TVQA+ (Lei et al. 2019) are the only datasets which belong to this level.

Criterion 2: Logical Complexity Complicated questions often require more (or higher) logical reasoning steps than simple questions. In a similar vein, if a question needs only a single supporting fact with single relevant datum, we regard this question as having low logical complexity. Here, we define four levels of logical complexity from simple recall to high-level reasoning, similar to hierarchical stages of human development (Seol, Sharp, and Kim 2011).

- **Level 1 (Simple recall on one cue):** The questions at this level can be answered using simple recall; requiring only one supporting fact. Supporting facts are represented as triplets in form of {*subject-relationship-object*} such as {*person-hold-cup*}.
- **Level 2 (Simple analysis on multiple cues):** These questions require recall of multiple supporting facts, which trigger simple inference. For example, two supporting facts {*tom-in-kitchen*} and {*tom-grab-tissue*} are referenced to answer "Where does Tom grab the tissue?".
- **Level 3 (Intermediate cognition on dependent multiple cues):** The questions at this level require multiple supporting facts with time factor to answer. Accordingly, the questions at this level cover how situations have changed and subjects have acted.
- **Level 4 (High-level reasoning for causality):** The questions at this level cover reasoning for causality which can begin with "Why". Reasoning for causality is the process of identifying causality, which is the relationship between cause and effect from actions or situations.

Hierarchical Difficulties of QA aligned with Cognitive Developmental Stages From the two criteria, we define four hierarchical difficulties for QA which are consistent with cognitive developmental stages of Piaget's theory (Piaget 1972; Collis 1975). Questions at level 1 in MC and LC belong to Difficulty 1 which is available from *Pre-Operational Stage* where a child thinks at a symbolic level,

Kyungsu : Yes. Yes, that's right. Something came up. I(Kyungsu)'m sorry. I(Kyungsu)'m really sorry.

Deogi : … Are you(Heayoung1) even a human being? Still smiling after you(Heayoung1) called off the wedding?

Haeyoung1 : … We (Heayoung1,Taejin) wouldn't have been able to live together for a long time anyway!

(a)

(b)

Figure 3: Examples of character-centered video annotations: (a) coreference resolved scripts and (b) visual metadata which contains the main characters' bounding box, name, behavior, and emotion. All annotations for characters in script and visual metadata can be co-referred by unique character's name.

	# Q	# Annotated Images	Avg. Video len. (s)	Textual metadata	Visual metadata	Q. lev
TGIF-QA (Jang et al. 2017)	165,165	-	3.1	-	-	-
MarioQA (Mun et al. 2017)	187,757	-	< 6	-	-	-
PororoQA (Kim et al. 2017)	8,913	-	1.4	Description, Subtitle	-	-
MovieQA (Tapaswi et al. 2016)	6,462	-	202.7	Plot, DVS, Subtitle	-	-
TVQA (Lei et al. 2018)	152,545	-	76.2	Script	-	-
TVQA+ (Lei et al. 2019)	29,383	148,468	61.49	Script	Char./Obj. Bbox**	
DramaQA	17,983	217,308	3.7[a] 91.3[b]	Script*	Char. Bbox, Behavior, Emotion	✓

[a] Avg. video length for shot [b] Avg. video length for scene * Coreference resolved script ** Only mentioned in QAs

Table 1: Comparison between video story QA datasets. Only DramaQA dataset provides hierarchical QAs from shot-level and scene-level videos and character-centered visual metadata (bounding box, name, behavior, and emotion).

but is not yet using cognitive operations. Questions at level 1 in MC and level 2 in LC belong to Difficulty 2 which is also available from *Early Concrete Stage* where a child can utilize a relevant operation between multiple supporting facts. Questions at level 2 in MC and level 3 in LC belong to Difficulty 3 which is available from *Middle Concrete Stage* where a child can think by utilizing more than two relevant cognitive operations and utilize dependent multiple supporting facts across time. Questions at level 2 in MC and level 4 in LC belong to Difficulty 4 which is available from *Concrete Generalization Stage* where a child can just generalize only from personal and concrete experience and have a higher thought on causality in relation to "Why". Examples for each Difficulty are illustrated in Figure 2.

Character-Centered Video Annotations

As the characters are primary components of stories, we provide rich annotations for the main characters in the video contents. As visual metadata, main characters are localized in the appeared image frames sampled in video clips and annotated with not only the character names but also behavior and emotion states. Also, all coreferences (e.g. he/she/they) of the main characters in scripts are resolved to give a consistent view of the characters. Figure 3 shows the examples of visual metadata and coreference resolved scripts.

Visual Metadata

- **Bounding Box:** In each image frame, bounding boxes of both a face rectangle and a full-body rectangle for the main characters are annotated with their name. In total, 20 main characters are annotated with their unique name.
- **Behavior & Emotion:** Along with bounding boxes, behaviors and emotions of the characters shown in the image frames are annotated. Including `none` behavior, total 28 behavioral verbs, such as `drink`, `hold`, `cook`, are used for behavior expression. Also, we present characters' emotion with 7 emotional adjectives; `anger`, `disgust`, `fear`, `happiness`, `sadness`, `surprise`, and `neutral`.

Coreference Resolved Scripts To understand video stories, especially drama, it is crucial to understand the dialogue between the characters. Notably, the information such as "*Who* is talking to *whom* about *what* did what?" is significant in order to understand whole stories. In DramaQA, we provide this information by resolving all coreferences for main characters in scripts. As shown in Figure 3(a), we annotate the characters' names to all personal pronouns for characters, such as I, you, we, him, etc. By doing so, characters in scripts can be matched with those in visual metadata and QAs.

Comparison with Other Video QA Datasets

We also present a comparison of our dataset to some recently proposed video QA datasets (Table 1). TGIF-QA and

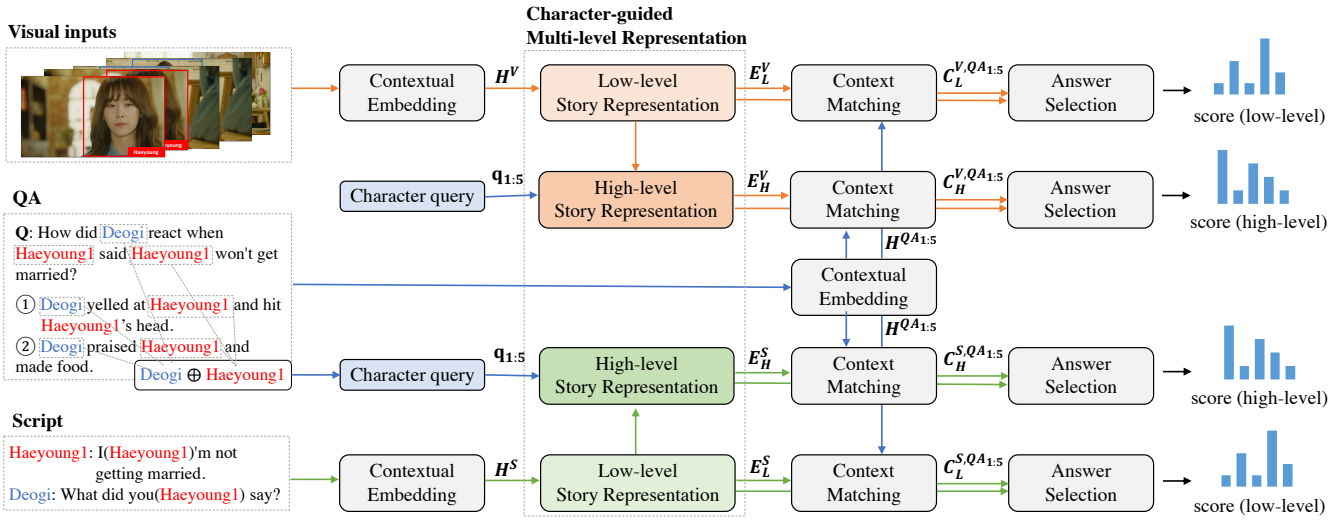

Figure 4: Our Multi-level Context Matching model learns underlying correlations between the video clips, QAs, and characters using low-level and high-level representations. Final score for answer selection is the sum of each input stream's output score.

MarioQA (Jang et al. 2017; Mun et al. 2017) only dealt with a sequence of images not textual metadata, and focused on spatio-temporal reasoning tasks about short video clips. PororoQA (Kim et al. 2017) was created using animation videos that include simple stories that happened in a small closed world. Since most of the questions in PororoQA are very relevent to subtitles and descriptions, most answers can be solved only using the textual information. MovieQA (Tapaswi et al. 2016) contains movie clips and various textual metadata such as plots, DVS, and subtitles. However, since the QA pairs were created based on plot synopsis without watching the video, collected questions are not grounded well to the video contents. TVQA+ (Lei et al. 2019), a sequel to the TVQA (Lei et al. 2018) particularly included annotated images with bounding boxes linked with characters and objects only mentioned in QAs. Although TVQA+ provides spatial and temporal annotations for answering a given question, most of their questions are aligned to relatively short moments (less than 15 seconds). Among the datasets, only the DramaQA 1) provides difficulty levels of the questions and rich information of characters including visual metadata and coreference resolved scripts and 2) tackles both shot-level and scene-level video clips.

Model

We propose Multi-level Context Matching model which grounds evidence in coherent characters to answer questions about the video. Our main goal is to build a QA model that hierarchically understands the multimodal story, by utilizing the character-centered annotations. The proposed model consists of two streams (for vision and textual modality) and multi-level (low and high) for each stream. The low-level representations imply the context of the input stream with annotations related to main characters. From low-level representations, we get high-level representations using character query appeared in QA. Then we use Context Matching module to get a QA-aware sequence for each level. Outputs of these squences are converted to a score for each answer candidate to select the most appropriate answer. Figure 4 shows our network architecture.[3]

Contextual Embedding Module

An input into our model consists of a question, a set of five candidate answers, and two types of streams related to video context which are coreference resolved scripts and visual metadata. Each question is concatenated to its five corresponding answer candidates. We denote a QA pair as $\text{QA}_i \in \mathbb{R}^{(T_Q+T_{A_i})\times D_W}$, where T_Q and T_{A_i} are the length of each sentence and D_W is the word embedding dimension. We denote the input stream from the script $S \in \mathbb{R}^{T_{\text{sent}}\times T_{\text{word}}\times D_W}$ where T_{sent} is the number of sentences and T_{word} is the maximum number of words per a sentence. Behavior and emotion are converted to word embedding and concatenated to each bounding box feature. We denote the visual metadata stream $V \in \mathbb{R}^{T_{\text{shot}}\times T_{\text{frame}}\times (D_V+2*D_W)}$ where T_{shot} is the number of shots in clips, T_{frame} is the number of frames per a shot, and D_V is the feature dimension of each bounding box.

In order to capture the coherence of characters, we also use a speaker of script and a character's name annotated in bounding box. Both pieces of character information are converted to one-hot vector and concatenated to input streams respectively. Then, we use bi-directional LSTM to get streams with temporal context from input streams, and we get $H^{\text{QA}_i} \in \mathbb{R}^{(T_Q+T_{A_i})\times D}$, $H^S \in \mathbb{R}^{T_{\text{sent}}\times T_{\text{word}}\times D}$ and $H^V \in \mathbb{R}^{T_{\text{shot}}\times T_{\text{frame}}\times D}$ for each stream, respectively.

Character-guided Multi-level Representation

Under the assumption that there is background knowledge that covers the entire video clip, such as the characteristics of each of the main characters, we have global representations

[3] https://github.com/liveseongho/DramaQA

Model	Diff. 1	Diff. 2	Diff. 3	Diff. 4	Overall	Diff. Avg.
QA Similarity	30.64	27.20	26.16	22.25	28.27	26.56
S.Only−Coref	54.43	51.19	49.71	52.89	52.89	52.06
S.Only	62.03	63.58	56.15	55.58	60.95	59.34
V.Only−V.Meta	63.28	56.86	49.88	54.44	59.06	56.11
V.Only	74.82	70.61	54.60	56.48	69.22	64.13
Our−High	75.68	72.53	54.52	55.66	70.03	64.60
Our−Low	74.49	72.37	55.26	56.89	69.60	64.75
Our (Full)	75.96	74.65	57.36	56.63	71.14	66.15

Table 2: Quantitative result for our model on test split. Last two columns show the performance of overall test split and the average performance of each set. `S.Only` and `V.Only` indicate our model only with script and visual inputs respectively. `S.Only−Coref.` and `V.Only−V.Meta` are `S.Only` with removed coreference and speaker annotation and `V.Only` with removed visual metadata. `Our(Full)` contains all elements of our model. `Our−High` and `Our−Low` are with removed high-level representations and with remove low-level representations from `Our(Full)`.

for each character name $\mathbf{m} \in \mathbb{R}^d$, where d is a dimension of each character representation. In our case d is same with the dimension of each contextual embedding. We use characters in question and i-th candidate answer pair to get character query $\mathbf{q}_i = \sum_j \mathbf{m}_j$.

Using this $\mathbf{q}_i$ as a query, we obtain character-guided high-level story representations for each stream E_H^V and E_H^S from low-level contextual embeddings by using attention mechanism:

$$E_H^V[j] = \text{softmax}(\mathbf{q}_i H^V[j]^\top) H^V[j] \quad (1)$$

$$E_H^S[j] = \text{softmax}(\mathbf{q}_i H^S[j]^\top) H^S[j] \quad (2)$$

We note that $E_H^V[j]$ and $E_H^S[j]$ represent sentence-level embedding for script and shot-level embedding for visual inputs, respectively. For the low-level story representations, we flatten H^S and H^V to 2-D matrices, so that $E_L^S \in \mathbb{R}^{(T_{\text{sent}} * T_{\text{word}}) \times D}$ and $E_L^V \in \mathbb{R}^{(T_{\text{shot}} * T_{\text{frame}}) \times D}$ is obtained.

Context Matching Module

The context matching module converts each input sequence to a query-aware context by using the question and answers as a query. This approach was taken from attention flow layer in (Seo et al. 2016; Lei et al. 2018). Context vectors are updated with a weighted sum of query sequences based on the similarity score between each query timestep and its corresponding context vector. We can get C^{S,QA_i} from E^S and C^{V,QA_i} from E^V.

Answer Selection Module

For embeddings of each level from script and visual inputs, we concatenate E^S, C^{S,QA_i}, and $E^S \odot C^{S,\text{QA}_i}$, where $\odot$ is the element-wise multiplication. We also concatenate boolean flag f which is `TRUE` when the speaker or the character name in script and visual metadata appears in the question and answer pair.

$$X_L^{S_i} = [E_L^S; C_L^{S,\text{QA}_i}; E_L^S \odot C_L^{S,\text{QA}_i}; f] \quad (3)$$

$$X_H^{S_i} = [E_H^S; C_H^{S,\text{QA}_i}; E_H^S \odot C_H^{S,\text{QA}_i}; f] \quad (4)$$

where we can get $X_L^{V_i}$ and $X_H^{V_i}$ in the same manner.

For each stream $X_L^{S_i}$, $X_H^{S_i}$, $X_L^{V_i}$, $X_H^{V_i}$, we apply 1-D convolution filters with various kernel sizes and concatenate the outputs of the kernels. Applying max-pool over time and linear layer, we calculate scalar score for i-th candidate answer. The final output score is simply the sum of output scores from the four different streams, and the model selects the answer candidate with the largest final output score as the correct answer.

Results

Quantitative Results

Here, we discuss an ablation study to analyze the model's characterstics profoundly. Table 2 shows the quantitative results of the ablation study for our model, and we described our experimental settings and implementation details in the Appendix C. `QA Similarity` is a simple baseline model designed to choose the highest score on the cosine similarity between the average of question's word embeddings and the average of candidate answer's word embeddings. The overall test accuracy of `Our(Full)` was 71.14% but the performance of each difficulty level varies. The tendency of poor performance as the level of difficulty increases shows that the proposed evaluation criteria considering the cognitive developmental stages are designed properly.

To confirm the utilization of multi-level architecture is effective, we compare the performance of our full model `Our(Full)` with those of the model excluding the high-level story representation module `Our−High` and the model excluding the low-level story representation module `Our−Low`. We can see that performances on Diff. 3 and 4 are more degraded in `Our−High` than `Our−Low`, whereas performances on Diff. 1 and 2 are more degraded `Our−Low` than `Our−High`. These experimental results indicate that the high-level representation module helps to handle difficult questions whereas the low-level representation module is useful to model easy questions.

Note that both script and visual input streams are helpful to infer a correct answer. `S.Only` uses only the script as the input and shows a sharp decline for Diff. 1 and 2. Since about 50% of QAs at Diff. 1 and 2 has a (shot-level) target video without a script, such questions need to be answered only with visual information. `V.Only` uses only visual input and shows decent performance on the overall difficulties.

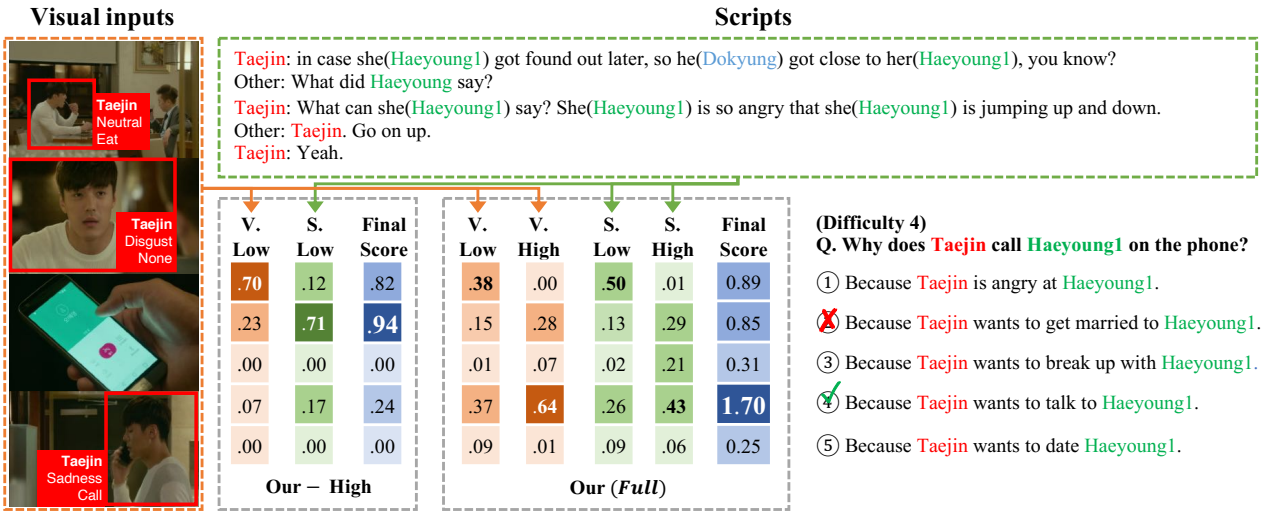

Figure 5: An example of correct prediction case to answer the question in Difficulty 4. To see the effectiveness of multi-level representation, we present the results of `Our(Full)` and `Our−High` in parallel. Scores of visual inputs are colored in orange and scores of scripts are colored in green. We colored final scores in blue. Prediction of `Our(Full)` is indicated by green checkmark which is ground truth answer, and prediction of `Our−High` is indicated by red crossmark.

Especially, the results show that the rich visual information is dominantly useful to answer the question at Diff. 1 and 2.

To check the effectiveness of character-centered annotation, we experimented with two cases: `V.Only−V.Meta` and `S.Only−Coref`. Here, `V.Only−V.Meta` only includes the visual feature of the corresponding frame by excluding visual metadata (bounding box, behavior, and emotion) of the main characters. Since it is hard to exactly match between characters of QA and video frames, the performance of `V.Only−V.Meta` was strictly decreased. For the same reason, `S.Only−Coref`, which removed coreferences and speakers from the `S.Only`, showed low performance in overall. These results show the effect of the proposed approach on character-centered story understanding.

We also compared our model with recently proposed methods for other video QA datasets. Due to the space limitation, the results are described in the Appendix D.

Qualitative Results

In this section, we demonstrate how each module of the proposed model works to answer questions. As shown in Figure 5, our model successfully predicts an answer by matching the context from candidate answers with the context from each input source. Especially, it shows that high-level representations help to infer a more appropriate answer from the context. In `Our(Full)`, Low-level scores from scripts of our model confused the answer with the first candidate including the word *angry*, but high-level scores from scripts chose the ground truth answer. Also, low-level scores from visual inputs inferred the first candidate answer to be correct based on the visual metadata *disgust, sadness*, but high-level scores from visual inputs gave more weight to the fourth candidate answer. As we discussed, character-guided high-level representations help to answer the question which requires complex reasoning. Without the high-level representations (shown in the results of `Our−High`), the model cannot fully understand the story and focuses on the low-level details. More examples including failures are provided in the Appendix E.

Conclusion

To develop video story understanding intelligence, we propose DramaQA dataset. Our dataset has cognitive-based difficulty levels for QA as a hierarchical evaluation metric. Also, it provides coreference resolved script and rich visual metadata for character-centered video. We suggest a Multi-level Context Matching model to verify the usefulness of multi-level modeling and character-centered annotation. Using both low-level and high-level representations, our model efficiently learns underlying correlations between the video clips, QAs and characters.

The application area of the proposed DramaQA dataset is not limited to QA based video story understanding. Our DramaQA dataset with enriched metadata can be utilized as a good resource for video-related researches including emotion or behavior analysis of characters, automatic coreference identification from scripts, and coreference resolution for visual-linguistic domain. Also, our model can be utilized as a fine starting point for resolving the intrinsic challenges in the video story understanding such as the integrated multimodal data analysis.

As future work, we will extend the two criteria of hierarchical QA so that the dataset can deal with longer and more complex video story along with expanding the coverage of evaluation metric. Also, we plan to provide hierarchical character-centered story descriptions, objects, and places. We expect that our work can encourage inspiring works in the video story understanding domain.

Acknowledgements

This work was partly supported by the Institute for Information & Communications Technology Promotion (2015-0-00310-SW.StarLab, 2017-0-01772-VTT, 2018-0-00622-RMI, 2019-0-01367-BabyMind) and Korea Institute for Advancement Technology (P0006720-GENKO) grant funded by the Korea government.

References

Case, R. 1980. Implications of neo-Piagetian theory for improving the design of instruction. *Cognition, development, and instruction* 161–186.

Collis, K. F. 1975. *A Study of Concrete and Formal Operations in School Mathematics: A Piagetian Viewpoint.* Hawthorn Vic : Australian Council for Educational Research.

Grosz, B. J.; Weinstein, S.; and Joshi, A. K. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational linguistics* 21(2): 203–225.

Heo, Y.; On, K.; Choi, S.; Lim, J.; Kim, J.; Ryu, J.; Bae, B.; and Zhang, B. 2019. Constructing Hierarchical Q&A Datasets for Video Story Understanding. *CoRR* abs/1904.00623. URL http://arxiv.org/abs/1904.00623.

Hermann, K. M.; Kocisky, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, 1693–1701.

Hill, F.; Bordes, A.; Chopra, S.; and Weston, J. 2016. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. In Bengio, Y.; and LeCun, Y., eds., *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. URL http://arxiv.org/abs/1511.02301.

Jang, Y.; Song, Y.; Yu, Y.; Kim, Y.; and Kim, G. 2017. TGIF-QA: Toward Spatio-Temporal Reasoning in Visual Question Answering. In *CVPR*.

Kim, K. M.; Heo, M. O.; Choi, S. H.; and Zhang, B. T. 2017. Deepstory: Video story QA by deep embedded memory networks. *IJCAI International Joint Conference on Artificial Intelligence* 2016–2022. ISSN 10450823.

Kočiskỳ, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K. M.; Melis, G.; and Grefenstette, E. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics* 6: 317–328.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. L. 2018. TVQA: Localized, Compositional Video Question Answering. In *EMNLP*.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2019. TVQA+: Spatio-Temporal Grounding for Video Question Answering. *CoRR* abs/1904.11574. URL http://arxiv.org/abs/1904.11574.

Maharaj, T.; Ballas, N.; Rohrbach, A.; Courville, A.; and Pal, C. 2017. A Dataset and Exploration of Models for Understanding Video Data Through Fill-In-The-Blank Question-Answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mclaughlin, G. H. 1963. Psycho-logic: A possible alternative to Piaget's formulation. *British Journal of Educational Psychology* 33(1): 61–67.

Mostafazadeh, N.; Chambers, N.; He, X.; Parikh, D.; Batra, D.; Vanderwende, L.; Kohli, P.; and Allen, J. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 839–849.

Mueller, E. T. 2004. Understanding script-based stories using commonsense reasoning. *Cognitive Systems Research* 5(4): 307–340.

Mun, J.; Seo, P. H.; Jung, I.; and Han, B. 2017. MarioQA: Answering Questions by Watching Gameplay Videos. In *ICCV*.

Pascual-Leone, J. 1969. *Cognitive development and cognitive style : a general psychological integration*. Toronto. Microfiche positive.

Piaget, J. 1972. Intellectual evolution from adolescence to adulthood. *Human development* 15(1): 1–12.

Richardson, M.; Burges, C. J.; and Renshaw, E. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 193–203.

Riedl, M. O. 2016. Computational narrative intelligence: A human-centered goal for artificial intelligence. *arXiv preprint arXiv:1602.06484* .

Riedl, M. O.; and Young, R. M. 2010. Narrative planning: Balancing plot and character. *Journal of Artificial Intelligence Research* 39: 217–268.

Schank, R. C.; and Abelson, R. P. 2013. *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.

Seo, M.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016. Bidirectional Attention Flow for Machine Comprehension. *ArXiv* abs/1611.01603.

Seol, S.; Sharp, A.; and Kim, P. 2011. Stanford Mobile Inquiry-based Learning Environment (SMILE): using mobile phones to promote student inquires in the elementary classroom. In *Proceedings of the International Conference on Frontiers in Education: Computer Science and Computer Engineering (FECS)*, 1.

Szilas, N. 1999. Interactive drama on computer: beyond linear narrative. In *Proceedings of the AAAI fall symposium on narrative intelligence*, 150–156.

Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; and Fidler, S. 2016. MovieQA: Understanding

Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Trischler, A.; Wang, T.; Yuan, X.; Harris, J.; Sordoni, A.; Bachman, P.; and Suleman, K. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830* .

Winston, P. H. 2011. The strong story hypothesis and the directed perception hypothesis. In *2011 AAAI Fall Symposium Series*.