How to Make a BLT Sandwich? Learning VQA towards Understanding Web Instructional Videos

Shaojie Wang*
Washington University in St. Louis
joss@wustl.edu

Wentian Zhao*
Adobe
wezhao@adobe.com

Ziyi Kou*
University of Notre Dame
zkou@nd.edu

Jing Shi
University of Rochester
j.shi@rochester.edu

Chenliang Xu
University of Rochester
chenliang.xu@rochester.edu

Abstract

Understanding web instructional videos is an essential branch of video understanding in two aspects. First, most existing video methods focus on short-term actions for a-few-second-long video clips; these methods are not directly applicable to long videos. Second, unlike unconstrained long videos, e.g., movies, instructional videos are more structured in that they have step-by-step procedures constraining the understanding task. In this work, we study problem-solving on instructional videos via Visual Question Answering (VQA). Surprisingly, it has not been an emphasis for the video community despite its rich applications. We thereby introduce YouCookQA, an annotated QA dataset for instructional videos based on YouCook2 [27]. The questions in YouCookQA are not limited to cues on a single frame but relations among multiple frames in the temporal dimension. Observing the lack of effective representations for modeling long videos, we propose a set of carefully designed models including a Recurrent Graph Convolutional Network (RGCN) that captures both temporal order and relational information. Furthermore, we study multiple modalities including descriptions and transcripts for the purpose of boosting video understanding. Extensive experiments on YouCookQA suggest that RGCN performs the best in terms of QA accuracy and better performance is gained by introducing human-annotated descriptions. YouCookQA dataset is available at https://github.com/Jossome/YoucookQA.

1. Introduction

Humans can acquire knowledge by watching instructional videos online. A typical situation is that people confused by specific problems try to look for solutions in related instructional videos. For example, while learning to cook new dishes, they may wonder *how a particular ingredient is added*, and *what happens between the two procedures*. Watching instructional videos can often clarify these questions and hence, guide humans in accomplishing tasks. The underlying process goes beyond the simple recognition of objects and actions. Still, it requires more complicated spatiotemporal and commonsense reasoning, which imposes a set of new challenges for existing AI algorithms. We hereby propose the question: can machines also understand instructional videos as humans do? As a proxy to answer this question, we study Visual-Question-Answering (VQA) for web instructional videos.

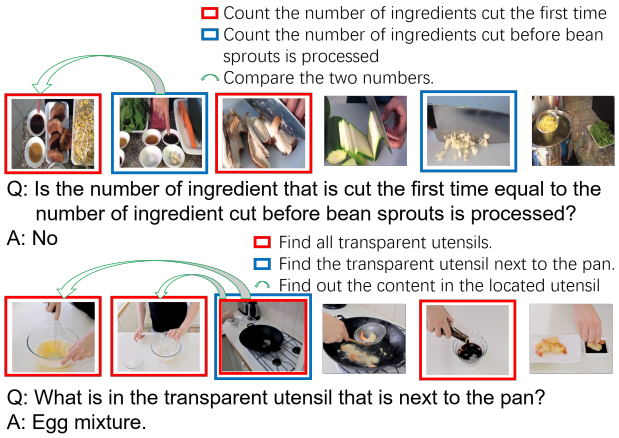


Figure 1: Demonstration of YouCookQA dataset. Colored boxes and arrows represent different steps required to answer the given questions. Red boxes denote the first step, blue boxes denote the second, and green arrows are for the final step. Better view zoomed in and with color.

Current instructional video understanding studies focus

*This work was done while they were at University of Rochester.

on various tasks e.g., reference resolution [7], procedure localization [27, 2], dense captioning [28, 18], activity detection [15, 12] and visual grounding [8, 20]. Despite its potential rich applications, VQA for instructional videos is less well-developed. Yet, it may act as a unified testbed for the above collective tasks. Previous works, e.g., ImageQA [3, 13, 11], and VideoQA [17, 25], also leverage the QA task as an automatic evaluation for image and short-video understanding. Still, QA on instructional videos has never been tackled before.

Observing the lack of suitable dataset on instructional videos, we propose YouCook Question Answering (YouCookQA) dataset based on YouCook2 [27], a recent popular instructional video dataset. We employ question-answering as intuitive interpretations for various styles of problem-solving. Figure 1 presents two exemplar QA pairs in our dataset along with the corresponding example human problem-solving steps involved to answer the questions. YouCookQA dataset contains 15,355 manually-collected QA pairs that are divided into different categories regarding different problem-solving styles, e.g., counting, ordering, comparison, and changing of properties.

Upon the newly built dataset, we explore in two directions. The first one concerns effective representations of modeling instructional videos. The videos in our consideration have an average length of 5.27 min and as instructional videos, they are structured and have step-by-step procedure constraining the understanding task. By modeling the temporal relations among different procedures, we are expecting valuable information to be extracted from the instructional videos, for which we study various model structures, including Recurrent Graph Convolutional Network (RGCN). The RGCN deals with complex information exchange by message passing in the graph, but also maintains the sequential ordering information by a supporting RNN. In this design, graph and RNN can boost each other since the information can be swapped between the two pathways.

Second, we explore the use of different modalities in video modeling. Apart from visual information, temporal boundaries, descriptions for each procedure, and transcripts are explored. In this direction, we want to test the effect of combining various types of available annotations with our developed video models on understanding instructional videos. Given that modeling instructional videos from vision alone is hard, combining such information approximates better the human learning experiences and it, in turn, guides better models for machine intelligence.

We conduct extensive experiments on the YouCookQA dataset. In the ablation study, we find that attention mechanism helps boost the performance. Our proposed RGCN model outperforms all other models in overall accuracy, even without attention. From the multi-modality perspective, modeling instructional videos using temporal boundaries together with descriptions can help dig more valuable information from videos. We also conduct human quiz on the QAs in our dataset. Results show that machines still have a large gap to human performance in that even without visual information, humans still can answer some questions correctly using life experience, or common sense, which hints us that incorporating the external knowledge with video models will be helpful for future works.

Our main contributions are summarized as follows.

(1) We propose YouCookQA dataset, a problem-solving-oriented dataset for understanding instructional videos.

(2) We perform solid evaluations with both graph and temporal models with various structures for video modeling. The RGCN model outperforms all other models even without attention.

(3) We incorporate multi-modal information to perform extensive experiments on YouCookQA showing that description can boost the video understanding capability, while transcripts could not.

The rest of the paper is organized as the following. We first discuss some related works in Sec. 2, and introduce the proposed YouCookQA dataset in Sec. 3. Then in Sec. 4, we set up series of baseline models for the dataset. In Sec. 5, we demonstrate and discuss the experiment results. Conclusions are drawn in Sec. 6.

2. Related Work

Instructional Video Understanding: Instructional video understanding has received much attention recently. Alayrac et al. [2] and Kuehne et al. [12] both leverage the natural language annotation of the videos to learn the instructional procedure in videos. Zhou et al. [27], however, propose to learn the temporal boundaries of different steps in a supervised manner without textual information. Dense captioning is also posed on instructional videos in [28], which aims at localizing temporal events from a video, and describing them with natural language sentences. Visual-linguistic ambiguities is a common problem in instructional videos with narratives. Huang et al. [7] focus on ambiguities caused by the changing in visual appearance and referring expression, and aim to resolve references with no supervision. Huang et al. [8] perform visual grounding task in instructional videos, also coping with visual-linguistic ambiguities. Yet, none of these works have tackled the QA problem on instructional videos, despite the uniqueness for instructional videos.

Video Question Answering: People are gaining interests in video question answering (VideoQA) in recent years. Most of the current VideoQA tasks are focusing on direct facts in short videos [22, 26, 25, 21, 29]. They all automatically generate QA pairs using a state-of-the-art question generation algorithm proposed in [5]. However, such auto-generation mechanism often generates QA pairs with

poor quality and low diversity, though grammatically correct. Worse still, auto-generated QA pairs rarely involve temporal relations among multiple frames. MovieQA [17] use human annotated QA pairs on movies to evaluate automatic story comprehension. SVQA [16], following the step of [10], extend the CLEVR dataset to the video version. Yet, it still focuses on short-term relations, and does not fit natural settings.

An other branch for VQA is neural-symbolic reasoning [24, 14, 23] which mainly focus on synthesized images and questions with composed logic and the paired reasoning program. While YouCookQA has real videos with multi-hop logic, combining symbolic reasoning would be an interesting future direction.

3. YouCookQA Dataset

To validate the proposed task of problem-solving on instructional video, we introduce YouCookQA dataset, a video question answering dataset based on YouCook2 dataset. The dataset contains 15,355 question-answer (QA) pairs in total. Tailored for our dataset, we annotate the QA pairs with six different tags, where each QA pair could be labeled with more than one tag. We show example QA pairs for each tag described below in appendix.

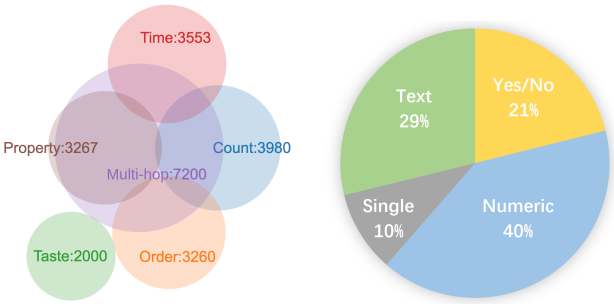
Counting: This tag annotates a QA pair that involves counting. One may count the occurrence time of certain actions or the number of certain ingredients. E.g., "How many white ingredients are used in the recipe?" Apart from counting, we also need to find out the target ingredients according to their colors.

Time: Time is a distinguishing feature in videos compared to images. This category of questions are mainly about timing and duration. A typical example is, "Which one is faster: adding water or adding salt?". To answer this question, we not only need to know how long it takes for both actions, but also need to compare the duration.

Order: Long-term temporal order is a unique feature for instructional videos, because instructional videos come with step-by-step procedures, and the order information matters. E.g., in YouCook2, the ordering of procedure is critical to the success of one recipe. Therefore, we stress out questions related to *action orders*, e.g., "What happens before/after/between ...?", and "Does it matter to change the order of ... and ...?"

Taste: YouCook2 is an instructional cooking video dataset, so we bring up with the taste questions. This type of QA pairs is about the flavor and the texture of the dish. Taste can be a unique type of question in this dataset that one can infer the taste from the ingredients used, and the texture from the cooking methods applied. Note that we avoid questions that are subjective such as "Is this burger tasty?", which involves too much subjective inspection.

Multi-hop: This tag presents a broader concept than all other tags above. By "multi-hop", we emphasize a question involving more than 2 sets of frames. This type of questions overlaps with all other types.



(a) Distribution on categories. (b) Different answer tags.

Figure 2: Statistics for our dataset.

Property: Cooking usually involves changes of ingredients. The properties of ingredients, e.g. their shape, color, size, location, etc., may vary at different time points as the cooking procedure goes on. This type of questions is different from "order" questions since we are asking about certain ingredients rather than actions.

In Tab. 1, we contrast our dataset to some other VideoQA datasets. Our dataset is unique in that we not only build the dataset based on instructional videos, but also focus on relations among multiple frames.

3.1. QA collection

Many existing VideoQA datasets [25, 21, 26, 22, 29] adopt an automatic question-answer (QA) generation technique proposed by [5] to generate QA pairs from texts. However, QA pairs obtained via this method suffer from extremely low diversity. Also, automatic methods can hardly generate questions involving multiple frames, which goes against our goal of constructing the dataset. Therefore, we apply Amazon Mechanical Turk (AMT) to collect question and answer pairs. For details about the collection of QA and multiple choice alternatives, please refer to appendix.

3.2. Statistics

In Fig. 2a, we show the statistics of six different categories of questions. We have 7,200 multi-hop QA pairs, consisting nearly half of our dataset. Other questions involve fewer frames, but still cannot be answered by direct observation from the videos. On average, we have 1.478 tags per QA pair, 2.289 words per answer, and 7.678 QA pairs per video.

To illustrate our dataset better, we split the QA pairs into four categories with respect to answer types, namely "Yes/No" for answers containing yes or no; "Numeric" for answers containing numbers, mostly related to counting and time; "Single word" for answers with only one word, excluding QA pairs in "Yes/No" and "Numeric"; "Text"

Table 1: Comparison among different video question answering datasets. The first four columns are: "Inst." for whether it is based on instructional videos; "Natural" for whether videos are of natural world settings; "Reason" for whether questions are related to reasoning; "Human" for whether QA pairs are collected through human labor.

	Inst.	Natural	Reason	Human	# of QA	Per video length	Answering form
VTW [25]	✗	✓	✗	✗	174955	1.5 min	Open-ended
Xu et al. [21]	✗	✓	✗	✗	294185	14.07 sec	K-Space
Zhu et al. [29]	✗	✓	✗	✗	390744	>33 sec	Fill in blank
Zhao et al. [26]	✗	✓	✗	✗	54146	3.10 sec	Open-ended
SVQA [16]	✗	✗	✓	✗	118680	-	K-Space
MovieQA [17]	✗	✓	✓	✓	6462	200 sec	Multiple choice
YouCookQA (Ours)	✓	✓	✓	✓	15355	5.27 min	Multiple choices and K-Space

for answers with multiple words, excluding QA pairs in "Yes/No" and "Numeric". Fig. 2b shows the distribution of four different types of answers in our dataset.

4. VQA on Instructional Videos

With the newly collected YouCookQA dataset, we perform VQA tasks by answering questions on instructional videos. We first formally define our problem in Sec. 4.1. Then in Sec. 4.2, based on attention mechanism, we design sequential model (SEQ-SA) and graph convolutional model (GCN-SA). We also propose Recurrent Graph Convolutional Network (RGCN) which captures both temporal order and multi-hop relations to overcome the limitation of SEQ-SA and GCN-SA. In Sec. 4.3, additional modalities such as description and transcripts are added to the multi-modal model to help gain better performance.

4.1. Problem Formalization

Multiple Choice: Since the questions in the YouCookQA dataset have alternative choices, we can use a three-way score function $f(v, q, a)$ to evaluate each alternative and choose the one with the highest score as correct answer:

$$j^* = \arg\max_{j=1,\dots,M} f(v, q, a_j) \ , \tag{1}$$

where $M = 5$ in our case, and v, q, a represent the feature of video, question and answer respectively. In this work, q and a are the final hidden states by encoding the question and answer via RNNs. Here, $f(\cdot, \cdot, \cdot)$ denotes a MLP whose input is the concatenation of v, q, and a and output is a single neuron classifying how likely the given answer a is the correct one.

K-Space: Similar to other VQA problems, our task can also be formulated as a classification problem on the answer space. Then the alternative (negative) answers are all other answers in the training set. Here, K types of distinct answers are assigned to K categories $\{A_i\}_{i=1}^K$. A MLP with K output neurons is tasked to predict the correct answer A^* by taking in v and q:

$$A^* = \arg\max_{j=1,\dots,K} g_j(v, q) \ , \tag{2}$$

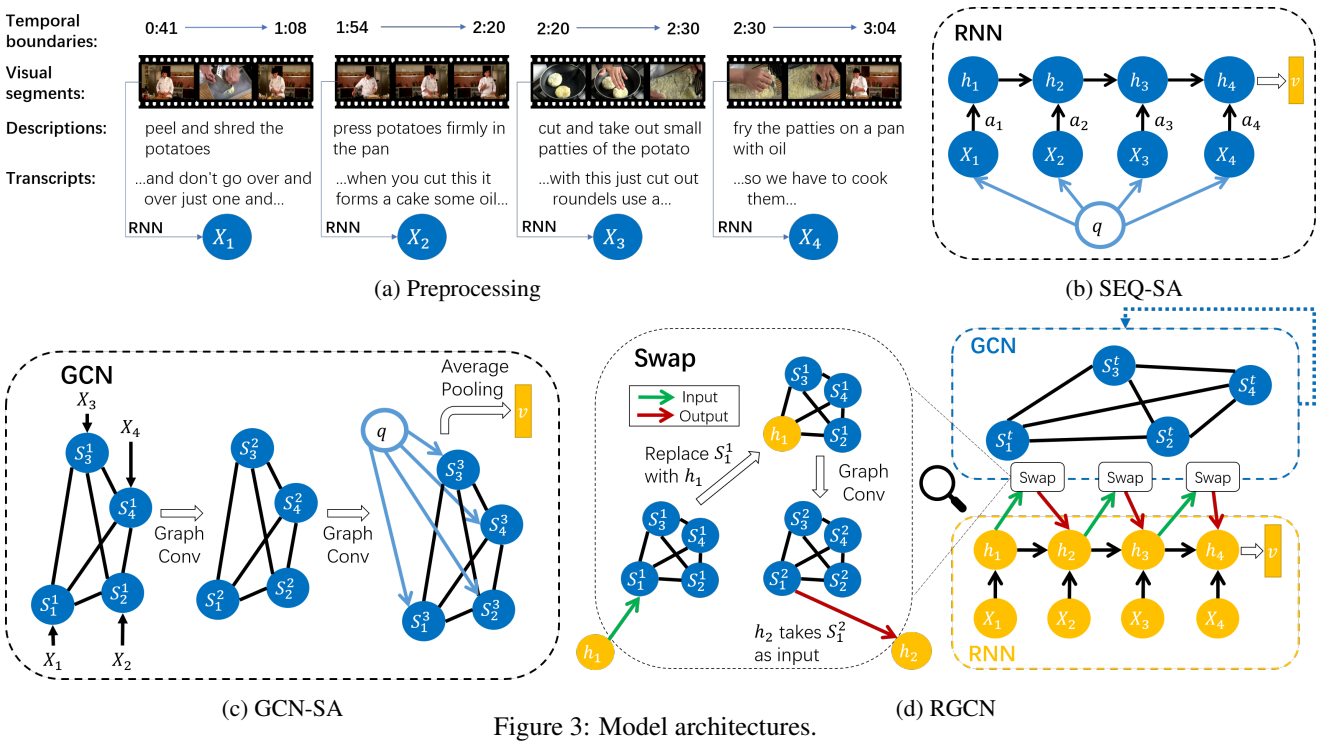
where g_j denotes the output score of the j-th neuron.

4.2. Models

In this section, we mainly focus on the design of video models that can capture procedure relations in instructional events. Their generated video feature v will be used for question answering. First, we describe how we pre-process the videos. Then, we introduce the architecture of proposed models that are suitable for VideoQA. Especially, we leverage RGCN architecture that can perform message passing between two paths: RNN and GCN, in order to capture both time series and global properties for modeling videos.

In Fig. 3, we present all the model architectures we use in this paper. In (a), we demonstrate the pre-processing procedure. We show an example video on how to make hash brown potatoes (YouTube ID: kj5y_71bsJM). It demonstrates the basic concepts of instructional videos in YouCook2 dataset. *Temporal boundaries* means the human annotated start/end time stamp of a *procedure*, which is well defined in [27]. Video are segmented into several *segments* (procedures) by the temporal boundaries. Descriptions are also annotated by human, corresponding to each procedure. Transcripts are auto-generated by speech recognition on YouTube. An example QA pair for the video in (a) is, Q:"*How many actions involving physical changes to potatoes are done before adding salt?*" A:"2.". In (b) and (c), we have question feature attending on each segment. In (d), we illustrate the structure of our proposed RGCN model, where GCN interacts with RNN via "swap" operation which takes in RNN's hidden state h_{t-1} and outputs the graph node S_{t-1}^t of GCN. We zoom in the first swap operation to provide an intuitive visualization.

Pre-processing: The videos in our consideration have an average length of 5.27 minutes, which requires us to process the videos into more tractable representations before any sophisticated modelings. Following [27], we define *procedure* as the sequence of necessary steps comprising



(a) Preprocessing (b) SEQ-SA

(c) GCN-SA (d) RGCN

Figure 3: Model architectures.

a complex instructional event and segment a video into N procedure segments (see Fig.3a). To directly benchmark the problem-solving ability, we use the ground truth provided by [27] instead to avoid any errors caused by intermediate processing. Note that one can apply method developed in [27] for automatically segmentation. The frames within each segment are sampled, of which the features are then extracted by ResNet [4] and encoded by a RNN model. Therefore, we can obtain the features of the procedure segments $\{X_i\}_{i=1}^N \in \mathbb{R}^d$ and use them for relation modeling.

SEQ-SA: We first propose an attention-based RNN model (see Fig. 3b for an example of $N = 4$) to model video representation v, where the encoded question feature is used to attend all video features at different time steps. The similarity a_i between question feature q and segment feature X_i is computed by taking the dot product of q and X_i: followed by a soft-max normalization: $a_i = \frac{exp(q^T X_i)}{\sum exp(q^T X_i)}$. Then we multiply each X_i by a_i to obtain the question-attended video feature $X_i^{'}$: $X_i^{'} = a_i X_i$. Finally, we feed $X_i^{'}$ into an RNN model of which the final hidden state h_N of RNN is taken as the video feature representation v.

GCN-SA: We consider a fully-connected graph (see Fig. 3c) to model multi-hop relations among the procedure segments. Although the time dependencies defined by the original video are omitted, different edges in the graph can mine different relations for various problem-solving tasks. We use a multi-layer GCN model for this purpose. We define $\{S_i^j\}_{i=1}^N{}_{j=1}^M$, where $S_i^j \in \mathbb{R}^d$, as the graph nodes, where N is the number of nodes within one layer, M is the number of layers. We first initialize nodes $\{S_i^1\}_{i=1}^N$ in the first layer by segment features $\{X_i\}_{i=1}^N$ correspondingly. We adopt the same GCN structure as described in [19]: $\mathbf{Z} = \text{ReLU}\{\mathbf{GSW}\}$, where $\mathbf{G} \in \mathbb{R}^{N\times N}$ represents the adjacency graph, $\mathbf{S} \in \mathbb{R}^{N\times d}$ denotes the concatenation of all node features $\{S_i\}_{i=1}^N$ in one arbitrary layer, and $\mathbf{W} \in \mathbb{R}^{d\times d}$ is the weight matrix which is different for each layer. Each element G_{ij} in $\mathbf{G}$ is the dot product similarity $S_i^T S_j$. Three GCN layers are used, where the output of the previous layer serves as the input of the next layer.

To apply the attention mechanism, we add an additional node in the last layer of the GCN to represent the question feature q, and this question node is connected with all other graph nodes $\{S_i^M\}_{i=1}^N$ through N edges. Question node attends to each graph node through different weights on the edges. Similar to SEQ-SA, the weights between q and $\{S_i^M\}_{i=1}^N$ are the dot products of corresponding node pairs, followed by a soft-max normalization. Finally, we use an average pooling operation to compress the output of the last layer $\mathbf{Z} \in \mathbb{R}^{N\times d}$ to $v \in \mathbb{R}^d$.

RGCN: Since the aforementioned GCN-SA is unable to capture the temporal order of video features [19], and SEQ-SA cannot model the relations between segments with long time spans, we construct Recurrent Graph Convolutional Network (RGCN) architecture (see Fig. 3d) to overcome such limitation. The RGCN is a recurrent model that consists of two pathways: RNN and GCN. RNN interacts with GCN mainly through a *swap* operation (see Fig. 3d). On one hand, the swap operation enables the temporal informa-

tion flows from the RNN path to GCN path, which enhances the representation power of GCN. On the other hand, graph convolution operation in GCN provides structural representation of different time steps of the video like humans do, which, in turn, provides more information for RNN model. The details are as follows.

The RNN pathway with N time steps takes in the segment features X_i at each time step. The GCN pathway has N layers, each of which contains N graph nodes. Note that the GCN has the same number of layers as the time steps in RNN pathway. We adopt the same GCN architecture as described in GCN-SA model except that a recurrent computation paradigm is applied here, where the weights $\mathbf{W}$ is shared among all layers. The computation within the RNN memory cell at each time step and the computation of each GCN layer are performed alternatively. For each time step t, we first concatenate together the segment feature X_t and the feature of node S_{t-1}^t in GCN, which is then used as the input to RNN memory cell at the t-th time step. Following [6], we update the hidden state h_t of RNN: $h_t = \text{RNN}\{[X_t, S_{t-1}^t], h_{t-1}\}$.

Then we replace GCN's graph node S_t^t with the updated hidden state h_t of RNN. This *swap* operation act as a bridge between RNN and GCN for message passing. Finally, the $(t+1)$-th GCN layer takes all $\{S_i^t\}_{i=1}^N$ as input to compute the response $\{S_i^{t+1}\}_{i=1}^N$: $\mathbf{Z_{t+1}} = \text{ReLU}\{\mathbf{G Z_t W}\}$, where $\mathbf{Z_t}$ is the concatenation of $\{S_i^t\}_{i=1}^N$. We take the final hidden state h_N of RNN as the video representation v.

Additionally, we extend the proposed RGCN with attention mechanism. The two pathways corresponds to the SEQ and GCN model, so we simply adopt how attention is cast on both pathways, and obtain RGCN-SA.

4.3. Multiple modalities

Besides videos and questions, we further investigate how much benefit we can obtain from other modalities such as narratives, which is very common in instructional videos. We are interested in two types of narratives, namely transcripts and descriptions.

Transcripts: The audio signal is an important modality for videos. In our dataset, the valuable audio information in videos is all chefs speaking. Therefore, we substitute audio with auto-generated transcripts on YouTube. Transcripts, which can be seen as describing the corresponding procedures, are highly unstructured, noisy, and misaligned narratives [7] in that chefs may talk about things not related to the cooking procedure, or that the speech recognition on YouTube may generate some unexpected sentences. Nevertheless, it can provide extra information to solve visual ambiguities, e.g., distinguishing water from white vinegar.

Descriptions: In YouCook2 dataset, each procedure in a video corresponds to a sentence of language description annotated by a human. Different from transcripts, descriptions are much less dense with respect to time, and can be seen as highly constrained narratives because human labor is applied to extract the essence of the corresponding procedures. Each piece of description is associated with the procedure it describes because they are highly related semantically .

For each individual modality (which can be description or transcripts), we aim to model a feature representation m, then fuse it with v and q to predict the answer A^*. To achieve this goal, we make use of a hierarchical RNN structure: a lower-level RNN models the natural language words within each segment, and a higher level RNN models the gloabal feature of the video.

5. Experiments

First, we introduce the implementation details of the training process. Then some baseline models are described, followed by results analysis. Also, we explored the benefit introduced by other modalities such as description and transcripts. All experiments conducted in this work are evaluated on both multiple choice and K-Space evaluation metrics. In Tab. 2, only multiple choice accuracy is provided for discussion. All other results on K-Space are in appendix.

5.1. Implementation details

Our codes are based on PyTorch deep learning framework. ResNet is used to extract visual features of 500 frames in each video, producing a 512-d vector. By using embedding layers, the question words are transformed into 300-d vectors which are optimized during the training process. For all models involving RNNs in this work, we apply single direction LSTMs [6] (an improved version of vanilla RNN) with 512 hidden units. Adam optimizer is used with the learning rate of 0.0001.

We split the training/testing set according to the original YouCook2 dataset. All videos in the YouCook2 training set are used as training videos in our dataset. Therefore, there are 10,179 QA pairs in our training set, and the rest are treated as testing set.

5.2. Baselines

We set up some baseline models which takes no instructional information, i.e. only the original video is presented to the models without temporal boundaries or descriptions.

Bare QA: First, we build the QA model which predicts answers based on questions only (without videos). Then for multiple choice, the answer is predicted by a similar way as Eq. 1: $j^* = \arg\max_{j=1...M} f(q, a)$. For K-Space, we adopt a similar formula as Eq. 2: $A^* = \arg\max_{j=1,...,K} g_j(q)$.

Naive RNN: RNN is a base of most state-of-the-art ImageQA [3, 13, 11] and VideoQA[21, 25] models. Instead of applying the segmentation pre-processing which we introduced in Sec. 4.2, Naive RNN takes in the ResNet feature

Table 2: Results on different model architectures.

	Count	Order	Taste	Time	Multi-hop	Property	All
Common sense	0.535	0.432	0.654	0.485	0.511	0.588	0.528
Bare QA	0.435	0.321	0.466	0.239	0.292	0.438	0.348
Naive RNN	0.434	0.330	0.467	0.234	0.283	0.449	0.347
MAC	0.438	0.331	0.462	0.229	0.294	0.437	0.348
SEQ	0.452	0.337	0.476	0.230	0.288	0.449	0.352
GCN	0.452	0.341	0.464	0.224	0.282	0.427	0.346
RGCN	**0.522**	**0.371**	0.478	**0.277**	**0.329**	**0.490**	**0.392**
SEQ-SA	0.473	0.355	**0.483**	0.256	0.316	0.465	0.373
GCN-SA	0.477	0.343	**0.487**	0.229	0.311	0.446	0.365
RGCN-SA	**0.545**	**0.367**	0.481	**0.279**	**0.316**	**0.486**	**0.403**

of sampled video frames directly. Similar to other models discussed previously, we take the final hidden state of the RNN as the video feature v. Then we evaluate the model performance based on Eq. 1 and Eq. 2.

MAC: MAC [9] is currently the state-of-the-art model on CLEVR dataset. Since our proposed YouCookQA dataset shares similar question style with CLEVR dataset, we adopt MAC as another alternative model. To extend MAC, which is designed for spatial relations, to the temporal relations in our work, we replace the input image features $\{I_i\}_{i=1}^{L}$, where $I_i \in \mathbb{R}^d$ (L is the number of spatial dimension of an image), with video frame features $\{X_i\}_{i=1}^{N}$, where $X_i \in \mathbb{R}^d$ (N is the number of sampled frames).

Human quiz: Apart from using deep learning models to complete VideoQA tasks, we also conduct human test with ten workers. First, they are asked to answer the questions without any other information, but by guessing or using common sense. Second, they are allowed to watch the videos without audio. Finally, audio is also turned on to match transcripts. Details of the setting are in appendix.

5.3. Results Analysis

Tab. 2 shows the experiment results on all models and baselines. We start with the comparison among baseline models that are without temporal boundary information (i.e., Bare QA, Naive RNN and MAC). As we can see from row 2 to row 4 of Tab. 2 that the three baselines have very close overall accuracy. Though Naive RNN take in the video stream, it cannot achieve better results than the bare QA. Therefore, we claim that as the base of most state-of-the-art VQA models, RNN fails to extract meaningful visual information for problem-solving on instructional video. The reason is that it is difficult for RNN to model multi-hop relations due to its sequential structure. Another reason is that RNN cannot capture long time dependencies of videos due to the memory limitation, even for RNN variants such as LSTM and GRU. As the best model on CLEVR, MAC achieves the same overall accuracy with Bare QA on YouCookQA, which demonstrates the special difficulty of video understanding compared with ImageQA. Besides, the high performance of bare QA suggests that we can use counterfactual modeling [1] to reduce bias.

Then we analyze the performance of models proposed in Sec 4.2, which incorporate temporal boundary information of instructional videos to boost the performance. Recall that the temporal boundaries are provided by the ground truth in [27]. First, to evaluate the improvement introduced by attention mechanism, we remove the question attention operation to formulate the models: SEQ, GCN, RGCN, the results of which are shown in row 5 to row 7 of Tab. 2. We can see from row 5 to row 10 of Tab. 2 that the margins gained by introducing attention are from 1.1% to 2.1%, which demonstrates that question can guide the models to extract more meaningful features, and all these models outperform baselines by a big margin. Especially, RGCN-SA achieves the highest overall accuracy of 40.3%, 5.5% higher than MAC, and SEQ-SA ranks second among the attention based models with an overall accuracy of 37.3%. This demonstrates that the procedure segmentation helps models make better use of video streams.

Finally, we investigate the performance of attention based models on various question categories. The comparison between SEQ-SA and GCN-SA shows that GCN-SA achieves higher accuracy scores on "count" and "taste" questions, while SEQ-SA performs better on all other categories. Intuitively, "order", "property" questions require temporal order information to be answered, because the questions usually contain sequence-related keywords, e.g., "before/after/between". Graph structure can hardly capture such ordering information. Nevertheless, the capability of modeling relations gives graph structure a reasonably good performance, especially on "count" and "taste" questions which challenge less on ordering. Since both sequence and graph models show advantages on different categories of questions, we take the advantages of both two models to build RGCN-SA, which is capable of passing messages between the two different pathways. Results show that graph and sequence can boost each others' performance on most

Table 3: Results on multiple modalities, where V stands for visual information, CC for transcripts, and D for descriptions.

	SEQ		**SEQ-SA**		**GCN**		**GCN-SA**		**RGCN**		**RGCN-SA**	
	MC	KS	MC	KS	MC	KS	MC	KS	MC	KS	MC	KS
Visual	0.352	**0.160**	0.373	0.164	0.346	0.150	0.365	0.164	0.392	0.179	0.403	0.182
CC	0.346	0.159	0.353	0.152	0.343	0.143	0.346	0.150	0.361	0.152	0.366	0.144
Description	**0.353**	0.158	0.365	0.156	**0.352**	**0.157**	0.347	0.153	0.385	0.163	0.389	0.162
V+CC	0.347	0.151	0.375	0.167	0.348	0.150	0.375	0.177	0.390	0.173	0.393	0.180
V+D	0.351	**0.160**	**0.379**	**0.173**	0.349	0.148	**0.383**	**0.183**	**0.413**	**0.194**	**0.416**	**0.203**

question types except for "taste".

5.4. Multimodalities

Based on temporal boundary annotations, we further explore other modalities. As described in Sec. 4.3, we experiment on two types of narratives, unconstrained transcripts and concentrated descriptions. Descriptions are already associated with video segments in the YouCook2 dataset, so we only need to align the transcripts with segments by selecting transcripts that lay between the temporal boundaries. Results are shown in Tab. 3.

As for different modalities, we first compare visual information, transcripts and description separately. Although descriptions are human annotated, highly refined reconstruction of the content of instructional videos, mere description seems not helpful when compared with visual information. Transcripts, to be worse, always decrease the performance. However, when narratives and visual information are combined together, we can see a significant increase in accuracy scores. SEQ-SA, GCN-SA, RGCN and RGCN-SA all achieve highest multiple choice accuracy when trained with both visual features and descriptions. SEQ with visual and description information also gets the highest K-Space accuracy compared to SEQ models trained on other modalities. However, transcripts still fail to provide as much valuable information as descriptions on videos, since the performance of models with visual and transcript information is worse than visual plus description. Transcripts even have a negative effect on SEQ and RGCN in that multiple choice accuracy is dropped when transcripts are added to visual information. Possible reasons are that the transcripts are too dense, and the quality of auto-generated transcripts are uncontrollable. As for different structures, we can see that our RGCN-SA still achieves the highest performance, while all attention models provides reasonable results.

5.5. Human quiz

In the human quiz part, participants are asked to do three sets of tests, namely guessing with common sense, with visual information, and with both visual and audio information. The results of the guessing step are shown at the top row in Tab. 2. As we can see, even without any video information, human can achieve an accuracy as high as 52.8%. An interesting fact here is that human participants did a good job on the "when" questions, which is unexpected because one cannot know the exact time point of what is going to happen without watching the video. The reason is that humans have an intuition of which ingredients is more likely to be added first, or which step is less likely to happen at the beginning, owing to their common sense or life experience. Another support for the power of common sense is the high accuracy score for "taste" questions. For machines, the taste can only possibly be learned from the relations between ingredients and correct answers. However, for human beings, the tastes of different ingredients is already known in daily life. Given visual information, the human performance becomes almost perfect (97.0%), which is not provided in the form of tables. This is reasonable because human has a powerful visual understanding and comprehending system. Given that the accuracy is already very high and that the dataset is collected without audio information, the improvement is minor (97.7%) after adding audio information. It is worth mention that RGCN-SA outperforms the human baseline on "count" questions, yet there is still a long way to go in VQA tasks on instructional videos.

6. Conclusion

In this paper, we emphasize problem-solving on instructional videos. We construct YouCook Question Answering (YouCookQA) dataset, and three models with sequence (SEQ), graph (GCN), and fused (recurrent graph convolutional network, RGCN) structures are proposed to explore the instructional information. Attention mechanism is applied on the proposed models to boost performance, and RGCN-SA achieves the best accuracy on both multiple choice and K-Space evaluation metrics. Experiments show that RGCN successfully fuse the order and relation information together for modeling instructional videos. Also, multiple modalities for instructional videos are analyzed, showing that human annotated temporal boundaries and descriptions are critical for instructional video understanding.

Acknowledgement

This work was supported in part by NSF 1813709 and 1741472. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

[1] Ehsan Abbasnejad, Damien Teney, Amin Parvaneh, Javen Shi, and Anton van den Hengel. Counterfactual vision and language learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10044–10054, 2020.

[2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4575–4583, 2016.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. In *ICCV*, pages 2425–2433. IEEE Computer Society, 2015.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[5] Michael Heilman and Noah A. Smith. Good question! statistical ranking for question generation. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA*, pages 609–617, 2010.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[7] De-An Huang, Joseph J. Lim, Li Fei-Fei, and Juan Carlos Niebles. Unsupervised visual-linguistic reference resolution in instructional videos. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1032–1041, 2017.

[8] De-An Huang, Shyamal Buch, Lucio M. Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. Finding " it " : Weakly-supervised reference-aware visual grounding in instructional videos. 2018.

[9] Drew A. Hudson and Christopher D. Manning. Compositional attention networks for machine reasoning. *CoRR*, abs/1803.03067, 2018.

[10] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997, 2017.

[11] Kushal Kafle and Christopher Kanan. Answer-type prediction for visual question answering. In *CVPR*, pages 4976–4984. IEEE Computer Society, 2016.

[12] Hilde Kuehne, Ali Bilgin Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 780–787, 2014.

[13] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1682–1690, 2014.

[14] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[15] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. A database for fine grained activity detection of cooking activities. In *CVPR*, pages 1194–1201. IEEE Computer Society, 2012.

[16] Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. Explore multi-step reasoning in video question answering. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, pages 239–247, 2018.

[17] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4631–4640, 2016.

[18] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence - video to text. In *ICCV*, pages 4534–4542. IEEE Computer Society, 2015.

[19] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *ECCV (5)*, volume 11209 of *Lecture Notes in Computer Science*, pages 413–431. Springer, 2018.

[20] Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. Weakly-supervised visual grounding of phrases with linguistic structures. In *CVPR*, pages 5253–5262. IEEE Computer Society, 2017.

[21] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1645–1653, 2017.

[22] Hongyang Xue, Zhou Zhao, and Deng Cai. Unifying the video and question attentions for open-ended video question answering. *IEEE Trans. Image Processing*, 26(12):5656–5666, 2017.

[23] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[24] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic VQA: disentangling reasoning from vision and language understanding. In Samy Bengio, Hanna M. Wallach, Hugo

Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 1039–1050, 2018.

[25] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 4334–4340, 2017.

[26] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 3518–3524, 2017.

[27] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7590–7598, 2018.

[28] Luowei Zhou, Yingbo Zhou, Jason J. Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. *CoRR*, abs/1804.00819, 2018.

[29] Linchao Zhu, Zhongwen Xu, Yi Yang, and Alexander G. Hauptmann. Uncovering temporal context for video question and answering. *CoRR*, abs/1511.04670, 2015.