VideoABC: A Real-World Video Dataset for Abductive Visual Commonsense Reasoning

Anonymous ECCV submission

Paper ID 2011

Abstract. Deep learning techniques have demonstrated great potentials in solving basic vision tasks involving perception such as object classification and detection. However, it is still difficult for artificial vision systems to reason more deeply about the complex and dynamic scenes and situations, which requires higher-level cognition and commonsense knowledge about the physical world. In this paper, we conceptualize a new task of abductive visual commonsense reasoning that requires the vision systems to infer the most plausible sequence of steps given two observations that describe the initial configuration and desired goal. Different previous attempts that study visual reasoning on static images or synthesized scenes, we explore long-term reasoning based on instructional videos that contain a wealth of detailed information about the physical world. We introduce a new dataset, VideoABC, that consists of 46,354 unique steps derived from 11,827 videos, formulated as 13,522 multiple choice questions with an average reasoning duration of 47s. With an adversarial hard choice mining algorithm, non-trivial and high-quality problems are generated efficiently and effectively. To move towards human-level reasoning, we propose a Hierarchical Dual Reasoning Network (HDRNet) to capture the long-term dependencies among steps and observations. Our method achieves 83% accuracy on the long-term reasoning benchmark, which significantly narrows the gap between humans (over 92%) and state-of-the-art deep video architectures (∼76%), but still leaves substantial room for further research.

Keywords: Visual Reasoning, Abductive Reasoning, Dataset

1 Introduction

Recent years have witnessed the great success achieved by deep neural networks for various computer vision tasks, such as image recognition [23,15], object detection [31,11], semantics segmentation [14,26], action recognition [34,44] and many others. While the artificial vision systems are gaining the human-level **perception** ability for many real-world applications, they are still limited in the higher-level **cognition** and commonsense knowledge, which becomes a major obstacle to reasoning more deeply about the complex and dynamic scenes of the real-world.

To address this issue, the computer vision community is paying growing attention to the problem of visual reasoning in the past few years, and a variety

Fig. 1: **Examples of abductive visual commonsense reasoning.** Given two observations of the visual world, the abductive visual commonsense reasoning task is to infer the most plausible sequence of steps between these two observations. We show the examples of short-term reasoning in (a) and (b) and more challenging long-term reasoning in (c) and (d). While humans find it easy, this task remains difficult for neural networks, especially when the ability of long-term reasoning is required.

of works have been proposed on this emerging topic. Compared with performing visual reasoning on image-based data [21,50], it is more challenging to explore the commonsense knowledge of video-based data because of the more complex structure in the temporal domain. Recently, several video-based datasets have been collected to facilitate the research on temporal visual reasoning, such as CATER [10], MovieQA [40] and VideoQA [49]. However, the CATER dataset is comprised of synthesized data, which makes it far away from the real-world scenarios. The models developed on the MovieQA and VideoQA datasets heavily rely on the natural languages, which limits in mining the intrinsic temporal relation of the video. Towards the goal to explore the pure visual reasoning, we borrow the merit from instructional videos [2,55,39], which demonstrate visual examples for people to accomplish a certain task. The different steps in the instructional videos are with high dependency, providing a resource to perform abductive visual reasoning.

Abductive reasoning is a form of logical inference that seeks to find the most plausible explanation for incomplete observations [29,1]. Given two observations of an event $\mathcal{O}_1$ and $\mathcal{O}_2$, people can hypothesize different potential explanations $\mathcal{H}$ and reason about which is the most likely based on commonsense knowledge about the world. For example (Fig. 1), given two frames of a video, we can easily understand the temporal relation between these two states of the visual world and infer what has happened between the two frames. Abductive reasoning is viewed as the only logical operation that introduces new ideas [29,4] and has long been considered to be at the core of reasoning about everyday situations [30] and text understanding [4]. Despite the broad recognition of its importance, the problem of abductive reasoning has rarely been investigated in computer vision research.

As a step toward human-level visual understanding, we present the first study to explore the abductive visual reasoning. We propose a new task formulation, **Abductive Visual Commonsense Reasoning**, which requires the vision systems to infer the most plausible explanation given two observations that describe the initial configuration and desired goal. Given images of these two observations, a model is expected to find a sequence of steps that lead to the desired goal in real-life tasks such as "change the car tire". A step is presented as a video clip of an atomic action, such as "unscrew the screws", "jack up the car", "put on the tire", *etc.* Solving the problem requires both commonsense knowledge about the physical world and a thorough understanding of the temporal relation between the two states of the visual world evoked by the images, which makes our task very challenging. To support easy and reliable automatic evaluation, all our tasks are formulated as a multiple-choice task.

We also introduce a real-world video dataset, **VideoABC**[1], that consists of 46,354 unique steps derived from 11,827 videos, formulated as 13,522 multiple choice questions with an average reasoning duration of 47s. Our dataset is built on the annotations of the COIN dataset [39], which contains instructional videos of 180 real-life tasks. An adversarial hard choice mining algorithm is also proposed to generate non-trivial and high-quality problems efficiently and effectively. The key characteristics of our dataset are: (1) the formulation of visual reasoning on real-world videos instead of static images or synthesized scenes; (2) long-term reasoning with more complex temporal dependencies; (3) diverse scenes covering various daily activities. These characteristics make our dataset unique compared to existing datasets for visual reasoning and video understanding.

As another contribution, we propose a Hierarchical Dual Reasoning Network (HDRNet) to capture the long-term dependencies among steps and observations. Our method achieves 83.1% accuracy on the long-term reasoning benchmark, which outperforms state-of-the-art video understanding methods (R(2+1)D [43] with Non-Local blocks [46], 76.0%) by a large margin. However, the task and dataset are far from solved: humans can easily achieve over 92% accuracy. We also provide extensive experimental results to establish a benchmark on our dataset and detailed analysis to point to interesting avenues for future research. The dataset and code for our model will be made publicly available.

2 Related Work

Visual Reasoning. The interest in high-level vision tasks is growing rapidly in recent years. A variety of new tasks are proposed to evaluate the reasoning ability of artificial vision systems. Visual Question Answering (VQA) [3] frame visual understanding as a QA task, where a dataset with questions about COCO images [25] is developed. This line of work is also extended to video-based QA [40,49]. However, the reasoning models developed on VQA datasets largely

[1] **VideoABC**: a **Video** dataset for **AB**ductive visual **C**ommonsense reasoning

rely on the natural language questions [13] and requires relatively less background knowledge about how the world works. In contrast, our dataset provides a pure vision benchmark for visual reasoning based on commonsense knowledge about the physical world. There are also some efforts to design unbiased datasets for visual reasoning. For example, CLEVR [20] and CATER [10] are synthesized datasets for reasoning on static images and dynamic scenes. RAVEN [51] and V-PROM [41] are diagnostic datasets based on Raven's Progressive Matrices (RPM). Different from these works, we focus on visual reasoning in real-world scenarios. Recently, [48] proposes a new task to explore the commonsense knowledge on image-based data. In this paper, we study abductive visual reasoning on dynamic scenes, which has not been visited yet.

Video Understanding. Convolutional neural networks (CNN) have dominated many computer vision tasks since AlexNet [24]. To exploit the temporal information in videos, a straightforward way is to process each video frame independently using state-of-the-art 2D CNN models, then fuse features from different frames. For example, Karpathy *et al.* [22] propose to use LSTM to model temporal relation among frames. Simonyan *et al.* [35] use a two-stream CNN model to process RGB input (spatial stream) and optical flow input (temporal stream) respectively. Wang *et al.* [45] design the Temporal Segment Networks (TSN) to fuse features from stridden sampled frames via average pooling. These methods learn spatial and temporal features separately and perform temporal reasoning on the extracted feature vectors, thus cannot infer more complicated spatio-temporal relationships. Although many efforts have been made to further integrate temporal module into 2D CNN [42,5,43], simple 2D models like TSN can still achieve very competitive performance on widely used video recognition benchmarks such as UCF101 [36] and Kinectis [5]. We show the temporal information in videos is crucial in our dataset, where the performance of 2D models is close to random guessing (see Section 5).

Instructional Video Analysis. People from all over the world can acquire expert knowledge to complete a certain task by watching instructional videos. In recent years, there have been rapidly increasing instructional videos uploaded to the Internet. While the community is paying more and more attention to this emerging field, growing efforts are being devoted to different tasks for instructional video analysis, *e.g.,* step localization [2,55], procedure localization [54], action segmentation, video caption [6], visual grounding [17,18,53], action prediction [9,33], skill determination [7,8], representation learning [37,28,27] and many others. More recently, Tang *et al.* [39] proposed a large-scale dataset called "COIN" for comprehensive Instructional analysis. Though a great number of step segments have been annotated with the corresponding labels and temporal boundaries, they have not explored the dependency of these steps, which is one of the core issues on this topic. To address this issue, we re-establish the COIN dataset under the visual reasoning paradigm. To our best knowledge, this is the first attempt to study visual reasoning task for instructional video analysis.

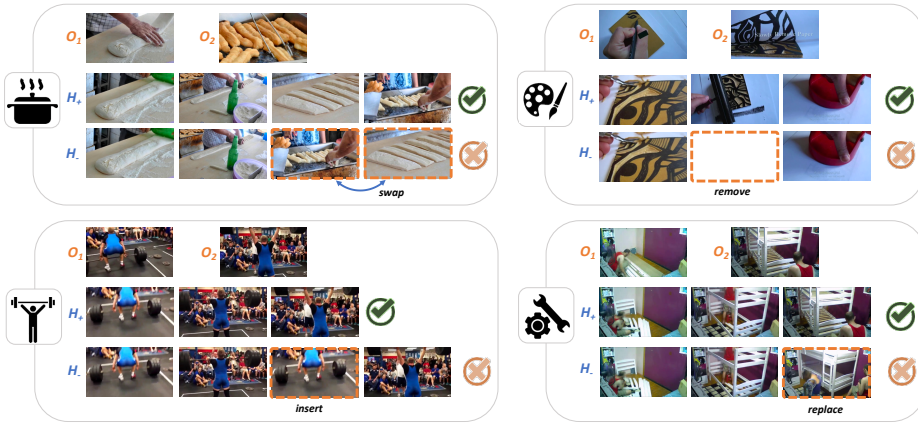

Fig. 2: **Some illustrative examples from the VideoABC dataset.** We show the pair of observations $\mathcal{O}_1$ and $\mathcal{O}_2$ as well as ground-truth choice ($\mathcal{H}_+$) and one of the incorrect choices ($\mathcal{H}_-$) as examples.

3 The VideoABC Dataset

In this section, we present VideoABC, a real-world video dataset for abductive visual commonsense reasoning. To our best knowledge, it is the first large-scale benchmark dataset for studying abductive reasoning in computer vision. We will first introduce the task definition, then describe how we construct the dataset, and provide the basic statistics of our dataset. Some illustrative examples are presented in Fig. 2

Task Definition. We formulate abductive visual reasoning as multiple choice problems consisting of a pair of observations as context and several hypothesis choices. Each instance in our dataset is defined as follows:

- $\mathcal{O}_1$: The observation at time t_1.
- $\mathcal{O}_2$: The observation at time $t_2 > t_1$.
- $\{\mathcal{H}\}$: A set of hypothesis choices that includes a plausible hypothesis $\mathcal{H}_+$ that explains what takes place between two observations $\mathcal{O}_1$ and $\mathcal{O}_2$, and several implausible hypotheses $\{\mathcal{H}_-\}$.

Specifically, we use the images of the initial configuration and the final goal in instructional videos as $\mathcal{O}_1$ and $\mathcal{O}_2$. A hypothesis choice is a sequence of video clips, where each clip contains an atomic action in instructional videos. We construct $N = 4$ hypothesis choices for each instance. Given the observations and a set of hypotheses choices, the goal of our task is to select the most plausible hypothesis choice.

To construct an abductive visual reasoning dataset that consists of diverse and high-quality problems, we choose to re-build the COIN dataset [39], because: (1) a great amount of step segments from diverse tasks have been annotated in COIN and can be used as the atomic actions in abductive reasoning; (2) the steps

in the instructional videos are with high temporal dependency, which provides a resource for abductive reasoning; (3) solving the abductive reasoning problem in instructional videos requires thorough understanding of the physical world.

Observations. In the original COIN dataset, we already have the label of categories and temporal boundaries of each step. However, we find that in many videos, the same task is completed many times, sometimes even by different performers or in different scenes. For example, in the video of "PlayCurling", there are multiple rounds played by different teams. These lead to discontinuity in the reasoning process and bring unnecessary difficulties to the reasoning models if we simply use the whole video to construct our dataset. Therefore, we first annotate all the cutting points at the end of each task, especially where the performers or scenes change. In the example of "PlayCurling", we place a cutting point at the end of each round of the game. In order to improve the annotation efficiency, we developed an annotation tool (see Supplementary Material for more details). For each step, we extract 16 frames uniformly. We show the last frame of the current step and the first from of the next step at the same time, along with the label of categories of the steps. Thus, the annotator can easily decide whether to place a cutting point between the two steps. After annotating cutting points, we obtained several segments for each video, with each segment containing a series of continuous steps. The pair of observations can be obtained by extracting the beginning and the ending frames of an arbitrary subsequence of these continuous steps. To avoid too long or too short sequence, we set the minimum and maximum number of steps in each question as 2 and 6 respectively. We also analyzed how the difficulty of our dataset is influenced by the sequence length in our experiments (see Section 5).

Choices. The sequence of all the steps between the pair of observations is naturally selected as the ground-truth choice. Based on the correct choice, we generate incorrect choices of various types by (1) removing one step (**remove**). (2) swapping two steps (**swap**). (3) inserting an extra step (**insert**). (4) replacing one step with another step outside the sequence (**replace**). Since the first and the last frames of a step may be identical to frames of the neighboring steps or the observation, to avoid trivial solutions, we used the middle 50% frames instead of the whole video clip for each step in the choices.

Adversarial Hard Choice Mining. A crucial challenge in constructing the dataset is how to generating high-quality distractors. To tackle this problem, we propose an adversarial hard mining algorithm to filter the trivial choices and obtain high-quality problems. Firstly, we generate the incorrect choices randomly using the aforementioned method. Secondly, we train several baseline models on the dataset and selected the best model. Thirdly, we re-generate at most 15 incorrect choices of each type as candidates and calculated the scores of them using our pre-trained model. The three incorrect choices with the highest scores are picked to form a new dataset. We repeat this process until a satisfactory dataset is obtained.

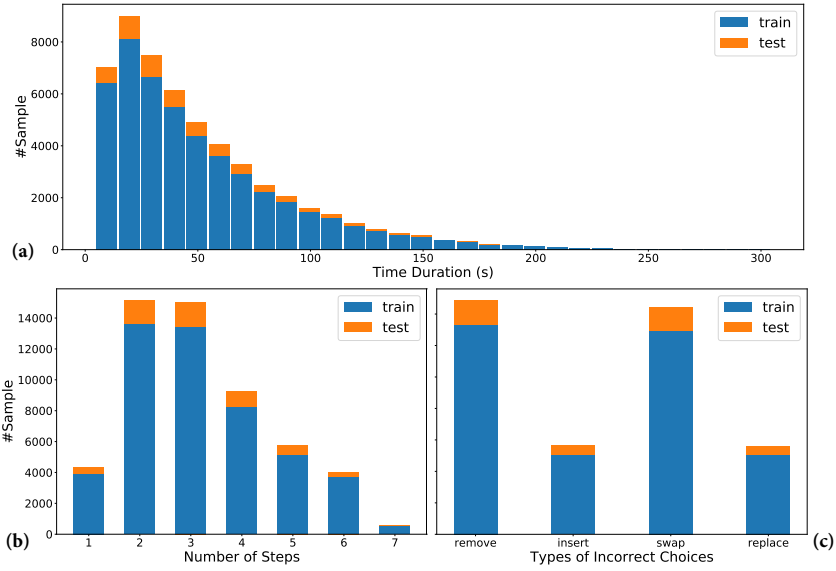

Fig. 3: The choice statistics of the our dataset. We show the distributions of (a) duration of the choices, (b) number of steps of the choices, and (c) types of distractors.

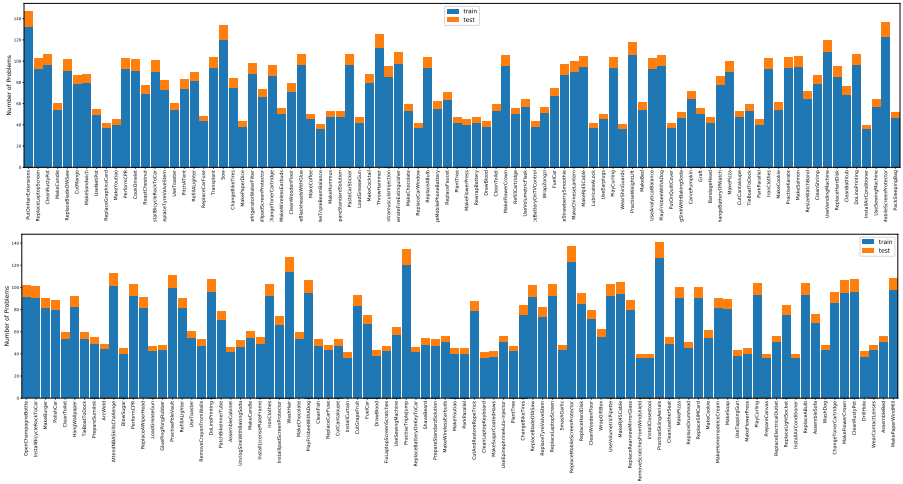

Fig. 4: Numbers of problems per class. The blue bars and the orange bars indicate the number of training and testing problems in each class respectively.

Statistics. The final version of our dataset consists of are 13,522 problems in total, covering 180 real-life tasks. We further divide these problems into train and test sets and ensure that the clips from the same video only exist in the same split, resulting in 12,086 problems for training and 1,436 for testing. There are 46,354 unique steps belonging to 778 different action categories in our dataset. More detailed statistics of our dataset is illustrated in Fig. 3 and Fig. 4.

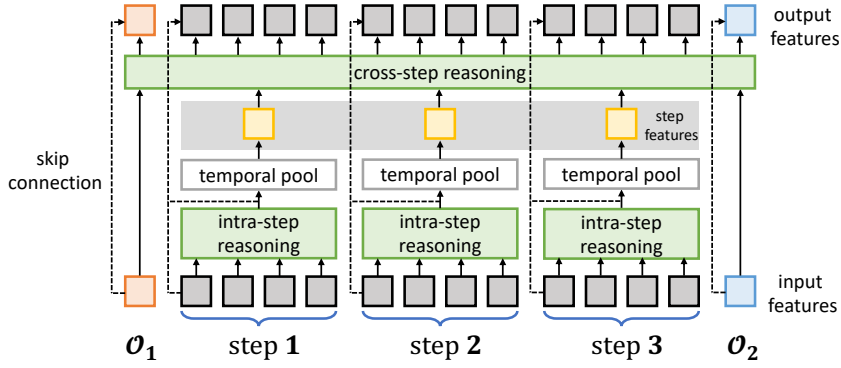

Fig. 5: The architecture of our **Hierarchical Dual Reasoning Block**. We extend the vanilla CNN with an intra-step reasoning module and a cross-step reasoning module that capture both short-term and long-term dependencies for temporal visual reasoning. Our model is able to integrate information over a long time interval, which makes abductive reasoning for a video lasting minutes possible. The proposed block can be used as a plug-and-play module in off-the-shelf 2D and 3D CNN models.

4 Hierarchical Dual Reasoning Network

The long-term temporal relation between steps is the key to solving abductive reasoning problems in our dataset. In our experiments, we found that state-of-the-art video understanding models failed to achieve satisfying performance in long-term reasoning tasks. Therefore, we propose a novel hierarchical dual reasoning network (HDRNet) that explicitly performs intra-step and cross-step reasoning hierarchically. In this section, we will first demonstrate the two essential components of the HDRNet: intra-step and cross-step reasoning module, and then describe our hierarchical architecture. The architecture of the proposed model is presented in Fig. 5.

Intra-Step Reasoning. We use a bi-directional recurrent network (RNN) to perform intra-step reasoning, which takes as input a sequence of features of all the frames in a step:

$$^{\text{intra}}S_i^{(k)} = [f_{i,1}^{(k)}, f_{i,2}^{(k)}, \ldots, f_{i,L}^{(k)}], \tag{1}$$

where L is the number of frames used in each step, $f_{i,j}^{(k)}$ denotes the feature of the j-th frame of the i-th step at the k-th layer. To reduce the computational cost, the RNN cell simply consists of several $1 \times 1 \times 1$ convolutional layers:

$$\begin{aligned} \mathbf{h}_j &= \text{ReLU}(\text{norm}(\text{conv}(\mathbf{x}_j))) + \text{ReLU}(\text{norm}(\text{conv}(\mathbf{h}_{j-1}))), \\ \mathbf{y}_j &= \text{ReLU}(\text{norm}(\text{conv}(\mathbf{h}_j))), \end{aligned} \tag{2}$$

where $\mathbf{x}_j, \mathbf{h}_j, \mathbf{y}_j$ are the input feature, hidden feature, and output feature at timestep j, respectively. Since the batch size is relatively small during training

due to the limitation of GPU memory, we use group normalization [47] as the normalization operation. Note that the number of channels is not changed when performing the convolution. We collect the outputs of all the RNN cells and average across the direction axis. Therefore, the output of our intra-step reasoning module ${}^{\text{intra}}R_i^{(k)}$ has the exact same shape as the input S_i^k.

Cross-Step Reasoning. The cross-step reasoning aims to explore the relationship among all the steps in a choice and the observations. We first pool the output of intra-step reasoning module ${}^{\text{intra}}R_i^{(k)}$ across the frames to obtain $p_i^{(k)}$. Along with the features of the questions, the input sequence of the cross-step reasoning module is :

$$ {}^{\text{cross}}S^{(k)} = [o_1^{(k)}, p_1^{(k)}, p_2^{(k)}, \ldots, p_N^{(k)}, o_2^{(k)}], \tag{3} $$

where $o_1^{(k)}$ and $o_2^{(k)}$ are the features of the two observations and N is the number of steps in the choice. Same as the intra-step reasoning, we feed this sequence into a bi-directional RNN and obtain the outputs ${}^{\text{cross}}_{\text{c}}R_i^{(k)}$ (for choice, $1 \le i \le N$) and ${}^{\text{cross}}_{\text{o}}R_i^{(k)}$ (for observations, $i = 1, 2$). Finally, we are able to update the features of observations and choice by:

$$ \begin{aligned} f_{i,j}^{(k)} &\leftarrow f_{i,j}^{(k)} + {}^{\text{intra}}R_{i,j}^{(k)} + {}^{\text{cross}}_{\text{c}}R_i^{(k)}, \quad (1 \le i \le N) \\ o_i^{(k)} &\leftarrow o_i^{(k)} + {}^{\text{cross}}_{\text{o}}R_i^{(k)}. \qquad\qquad (i = 1, 2) \end{aligned} \tag{4} $$

HDRNet. In our implementation, we use a 2D ResNet [16] to extract spatial features. After each stage in ResNet (except the first stage), we plug an intra-step reasoning module and a cross-step reasoning module to leverage the temporal information and update the visual features (i.e., 4 HDR blocks are inserted in total). Therefore, the reasoning process is performed at each spatial resolution. Finally, we use a fully connected layer to map the features to the score of the choice. The choice with the highest score is the predicted answer of the model. By combining the spatial and temporal information captured by ResNet and the proposed module respectively, HDRNet is able to learn both short-term and long-term temporal relation and integrate information over a long time interval, which makes abductive reasoning for videos lasting minutes possible. Note that although we use spatial CNN as our backbone for efficiency, our model is also compatible with spatiotemporal networks, which can serve as a plug-and-play module to enhance temporal features.

5 Experiments

In order to provide a benchmark for our VideoABC dataset, we conducted experiments with several state-of-the-art video understanding and reasoning models. We compare the proposed HDRNet with these baseline methods and provide the

human performance as reference. Besides, we use a probabilistic framework to verify that the proposed task requires a thorough understanding of both observations and hypothesis choices. We also provide extensive analysis and ablation study on our method and the dataset. The following describes the details of our experiments and results.

5.1 Benchmark on VideoABC

To establish a benchmark for our VideoABC dataset, we evaluate the following baseline methods:

- **Random Selection**: An answer is randomly selected from the alternatives.
- **TSN** [44]: Temporal Segment Networks (TSN) is a simple yet very competitive 2D model, which achieves top performance on many video understanding benchmakrs such as UCF101 [36] and Kinectis [5].
- **R3D** [43]: R3D is a 3D video understanding model built upon 3D resnet, which is effective in many video-based tasks and achieve the state-of-the-art accuracy in action classification datasets.
- **R(2+1)D** [43]: R(2+1)D further factorizes 3D convolutional filters into separate spatial and temporal components, which is prominent for its efficiency and accuracy to extract spatiotemporal features from the videos.
- **TRN** [52]: Temporal Relation Network (TRN) is an effective and interpretable model. TRN is designed to help CNN capture temporal relations in videos, which outperforms many two-stream and 3D convolution methods. In our experiments, we use the multi-scale TRN as our baseline.
- **Non-Local Networks** [46]: Non-local networks are capable of capturing long-range dependencies by adding spatiotemporal interaction layers. In our experiments, we add the non-local blocks to the R3D and R(2+1)D model to evaluate the performance on our dataset.

Implementation Details. For TSN and TRN, we use BNInception [38] as the backbone network following their original implementation [44] and evaluate the model under different consensus modules: `max` and `average`. For R3D and R(2+1)D, we use the 18-layer ResNet as the backbone network following their original implementation [43]. For a fair comparison, we use the spatial ResNet-18 as the backbone network for HDRNet. In all of our experiments, we use Group Normalization [47] instead of Batch Normalization [19] as the normalization method due to the small batch size. We extract 8 frames for each step and resize the frame to 112×112 in all the experiments. For each choice, we concatenate the first observation, the sequence of steps and the second observation along the temporal dimension to form the input for video understanding models and use a fully connected layer to map the final feature of the input to a scalar score. During training, we use the cross-entropy loss on the scores of the four choices as the supervision signal. During inference, we select the choice with the highest score as the answer.

Human Performance. To show the human performance on our dataset, four volunteers are asked to choose the most plausible hypothesis choice from the

Table 1: **Experimental results on VideoABC.** We report the results of several baseline methods, HDRNet and the human volunteers. Our method significantly outperform state-of-the-art video understanding models. Still, all models underperform human accuracy on our task.

Method	Accuracy (%)
Random	25.0
TSN-`average` [44]	16.4
TSN-`max` [44]	28.6
TRN [52]	30.2
R3D [43]	73.1
R(2+1)D [43]	74.5
R3D + Non-Local [46]	74.3
R(2+1)D + Non-Local [46]	76.0
HDRNet	**83.1**
Human	92.4

four candidate choices. We use the average accuracy of the four volunteers as the human performance. We developed a website for online testing. More details are provided in the Supplementary Material.

Results. We report the performance of baseline models, HDRNet and the human volunteers in Table 1. We see the 2D models including TSN and TRN struggles to make a right choice on our dataset. Although TRN can improve the performance of TSN by 1.6%, the performance of 2D models is very close to random choice, which indicates that the temporal dependencies, especially long-term dependencies is crucial in our task. The 3D models R3D and R(2+1)D are able to understand the problems in our dataset, but still face a dilemma and get confused in many cases. By adding non-local, the performance can be further improved by 1.2% and 1.5% for R3D and R(2+1)D, respectively. The proposed HDRNet outperforms all other models by a large margin. Using the same backbone network, HDRNet can improve the performance of R3D and R(2+1)D by 10.0% and 8.6% respectively and established a strong baseline on our dataset. Although HDRNet significantly narrows the performance gap between video understanding models and humans, there is still substantial headroom remaining.

5.2 Probabilistic Framework for Abductive Visual Reasoning

A key challenge in our reasoning task is that it requires a thorough understanding of all the observations so that our dataset is not solvable by trivial patterns. Following [4], we use a probabilistic framework to test our dataset. The multiple choice task is to choose the most plausible hypothesis choice $\mathcal{H}^*$ given the pair of observations:

$$\mathcal{H}^* = \arg\max_{\mathcal{H}^i} P(\mathcal{H}^i|\mathcal{O}_1, \mathcal{O}_2). \tag{5}$$

Table 2: Input formats of different probabilistic models.

Probabilistic Model	Input Format
Choice Only	$\mathcal{H}$
Choice and the First Observation	concat($\mathcal{H}$, $\mathcal{O}_1$)
Choice and the Second Observation	concat($\mathcal{H}$, $\mathcal{O}_2$)
Linear Chain	concat($\mathcal{H}$, $\mathcal{O}_1$); concat($\mathcal{H}$, $\mathcal{O}_1$)
Fully Connected	concat($\mathcal{H}$, $\mathcal{O}_1$, $\mathcal{O}_2$)

Table 3: The prediction accuracy (%) of the R3D baseline models under different independence assumptions on the short-term reasoning (STR) and long-term reasoning (LTR) dataset.

Probabilistic Model	STR	LTR
Choice Only	25.2	56.6
Choice and the First Observation	70.1	67.3
Choice and the Second Observation	79.1	69.9
Linear Chain	82.9	71.9
Fully Connected	84.7	73.1

In order to evaluate the independence between observations and choices, we test the reasoning model under various independence assumptions:

- **Choice Only:** Observations and choices are independent from each other. The model can select the best choice without the observations, i.e., the model aims to maximize $P(\mathcal{H}_+)$.
- **Choice and Part of Observation:** Given the choice and the first (or second) part of the observations, we evaluate the model with a weaker assumption that option $\mathcal{H} \perp \mathcal{O}_1$ or $\mathcal{H} \perp \mathcal{O}_2$.
- **Linear Chain:** The linear chain model considers each observation's influence on the choice separately. Specifically, we train two reasoning models ϕ_1 and ϕ_2 simultaneously and optimize:

$$\mathcal{H}^* = \arg\max_{\mathcal{H}^i} \ \phi_1(\mathcal{H}^i, \mathcal{O}_1) + \phi_2(\mathcal{H}^i, \mathcal{O}_2). \tag{6}$$

- **Fully Connected:** The full connected model is defined by Equation 5, which implies a close relationship between the observations and the choices.

To test these five types of assumptions in the experiment, we trained five R3D baseline models using different input formats. We summarize these input formats in Table 2. The experiments are conducted on the standard dataset and a short-term reasoning variant, where only single-step reasoning is required. More details about several variants of our dataset are provided in the Supplementary Material. Table 3 shows the performance of the R3D baseline models under different independence assumptions on our long-term and short-term reasoning dataset. With limited observations, the choices become more ambiguous and

Table 4: **Effects of the number of steps**. We show the prediction accuracy of the R3D baseline models on different variants of our dataset. With the maximum number of steps M increasing, our dataset becomes more challenging.

Method	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$	$M = 6$
Random	25.0	25.0	25.0	25.0	25.0	25.0
TSN-`average`	30.2	18.3	15.3	17.1	14.6	16.4
TSN-`max`	24.1	26.4	29.3	27.6	26.9	28.6
R3D	84.7	85.5	79.1	74.4	73.9	73.1
R(2+1)D	86.9	86.8	81.6	74.4	74.4	74.5

Table 5: **Effects of adversarial hard choice mining**. We shows the prediction accuracy of R3D on the dataset before and after hard mining. The large margin verifies that our algorithm can pick out the most deceptive distractors and effectively improve the quality of problems.

Method	$M = 1$	$M = 2$	$M = 3$	$M = 4$	$M = 5$	$M = 6$
Random	90.0	94.0	86.1	82.3	83.9	78.7
Hard Mining	84.7	85.5	79.1	74.4	73.9	73.1

the model will face a dilemma to make a prediction. When the model observes only the choices, for the short dataset, the prediction degenerates to random choice with only 25.15% accuracy. Although the model is able to capture useful information from the choices of the long-term reasoning dataset, the margin between `Fully Connected` and other probabilistic models shows that our task requires full observations to make a satisfactory prediction.

5.3 Analysis

Effects of the number of steps. We first investigate how the maximum number of steps affects the difficulty of our dataset. We generate six versions of our dataset with the maximum number of steps M ranging from 1 to 6 by using the adversarial hard choice mining algorithm mentioned above. When the number of steps rises, the relation between steps becomes more complex and intricate. Table 4 shows that with M increasing, our dataset becomes more challenging, where the prediction accuracy of R3D drops from 84.72% to 73.11%, while the performance of TSN is very close to random choice under all the settings.

Effects of adversarial hard choice mining. The adversarial hard choice mining aims to generate more challenging distractors, which is crucial in constructing a high-quality dataset. Table 5 shows the prediction accuracy of R3D on the dataset before and after hard mining. The large margin verifies that our

Table 6: **Effects of pre-trained models**. We shows the prediction accuracy various models on the dataset with/without pre-training.

Method	Pre-training	Accuracy (%)
TRN	Scratch	30.2
	Something-Something	39.9
R3D	Scratch	73.1
	Kinetics	73.4
HDRNet	Scratch	83.1
	ImageNet	83.2

Table 7: **Comparison of different modifications on R3D**. We shows the prediction accuracy of different variants of R3D on VideoABC.

Modification	R3D	R(2+1)D	Non-Local	Deeper	HDRNet
Accuracy (%)	73.1	74.5	74.3	74.9	83.1

algorithm can pick out the most deceptive distractors and effectively improve the quality of problems.

Effects of pre-trained models. We investigate different pre-trained models on Something-Something [12], Kinetics [5] and ImageNet [32] in Table 6. We see in general pre-trained models will not significantly improve the performance on our dataset, which demonstrates the knowledge learned on existing video and image understanding datasets is not very useful to solve the proposed abductive reasoning problem. Temporal relation learned on Something-Something is more useful than classification ability learned on Kinetics or ImageNet.

Effects of different modifications on R3D. We compared the effects of several modifications on R3D in Table 7. We tested R(2+1)D [43], Non-Local blocks [46] and the proposed HDRNet, where the same ResNet-18 backbone is used. We also report the result of a deeper ResNet-34 backbone network. We see using deeper or more complex models like the Non-Local network will not significantly improve the baseline model in our dataset, which indicates designing models that can capture hierarchical relation among steps and observations is the key to solve abductive reasoning problems.

6 Conclusions

In this paper, we have introduced a new video-based dataset called VideoABC for our newly conceptualized task on abductive visual commonsense reasoning. We have devised an adversarial hard choice mining algorithm to generate non-trivial and high-quality problems. We have established a benchmark and developed a bi-directional temporal reasoning module on our dataset. Experimental results

have shown the superiority of our proposed approach and demonstrated its effectiveness to explore the long-term dependencies among steps and observations. We will release our dataset and source code for the community to promote this emerging field.

References

1. Abductive reasoning — Wikipedia, https://en.wikipedia.org/wiki/Abductive_reasoning
2. Alayrac, J., Bojanowski, P., Agrawal, N., Sivic, J., Laptev, I., Lacoste-Julien, S.: Learning from narrated instruction videos. TPAMI **40**(9), 2194–2208 (2018)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: Vqa: Visual question answering. In: ICCV. pp. 2425–2433 (2015)
4. Bhagavatula, C., Bras, R.L., Malaviya, C., Sakaguchi, K., Holtzman, A., Rashkin, H., Downey, D., Yih, S.W.t., Choi, Y.: Abductive commonsense reasoning. ICLR (2019)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
6. Das, P., Xu, C., Doell, R.F., Corso, J.J.: A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In: CVPR. pp. 2634–2641 (2013)
7. Doughty, H., Damen, D., Mayol-Cuevas, W.W.: Who's better? who's best? pairwise deep ranking for skill determination. In: CVPR. pp. 6057–6066 (2018)
8. Doughty, H., Mayol-Cuevas, W.W., Damen, D.: The pros and cons: Rank-aware temporal attention for skill determination in long videos. In: CVPR (2019)
9. Farha, Y.A., Richard, A., Gall, J.: When will you do what? - anticipating temporal occurrences of activities. In: CVPR. pp. 5343–5352 (2018)
10. Girdhar, R., Ramanan, D.: CATER: A diagnostic dataset for compositional actions and temporal reasoning. In: ICLR (2020)
11. Girshick, R.B.: Fast R-CNN. In: ICCV. pp. 1440–1448 (2015)
12. Goyal, R., Kahou, S.E., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al.: The" something something" video database for learning and evaluating visual common sense. In: ICCV. vol. 1, p. 5 (2017)
13. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: CVPR. pp. 6904–6913 (2017)
14. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. In: ICCV. pp. 2980–2988 (2017)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
17. Huang, D.A., Buch, S., Dery, L., Garg, A., Fei-Fei, L., Carlos Niebles, J.: Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos. In: CVPR. pp. 5948–5957 (2018)
18. Huang, D., Lim, J.J., Fei-Fei, L., Niebles, J.C.: Unsupervised visual-linguistic reference resolution in instructional videos. In: CVPR. pp. 1032–1041 (2017)
19. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
20. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR. pp. 2901–2910 (2017)
21. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (2017)

22. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR. pp. 1725–1732 (2014)
23. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NeurIPS. pp. 1106–1114 (2012)
24. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. pp. 1097–1105 (2012)
25. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV. pp. 740–755. Springer (2014)
26. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. pp. 3431–3440 (2015)
27. Miech, A., Alayrac, J., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. CoRR **abs/1912.06430** (2019)
28. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In: ICCV (2019)
29. Peirce, C.S.: Collected papers of charles sanders peirce, volume 5. Harvard University Press (1965)
30. Peirce, C.S.: Pragmatism and pragmaticism, vol. 5. Belknap Press of Harvard University Press (1965)
31. Ren, S., He, K., Girshick, R.B., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NeurIPS. pp. 91–99 (2015)
32. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. IJCV **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
33. Sener, F., Yao, A.: Zero-shot anticipation for instructional activities. In: ICCV (2019)
34. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NeurIPS. pp. 568–576 (2014)
35. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: NeurIPS. pp. 568–576 (2014)
36. Soomro, K., Zamir, A., Shah., M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. Technical Report CRCV-TR-12-01, University of Central Florida (2012)
37. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: ICCV (2019)
38. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016)
39. Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., Zhou, J.: COIN: A large-scale dataset for comprehensive instructional video analysis. In: CVPR. pp. 1207–1216 (2019)
40. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: CVPR. pp. 4631–4640 (2016)
41. Teney, D., Wang, P., Cao, J., Liu, L., Shen, C., Hengel, A.v.d.: V-prom: A benchmark for visual reasoning using visual progressive matrices. arXiv preprint arXiv:1907.12271 (2019)

42. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV. pp. 4489–4497 (2015)
43. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR. pp. 6450–6459 (2018)
44. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Gool, L.V.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV. pp. 20–36 (2016)
45. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Val Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV (2016)
46. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. arXiv preprint arXiv:1711.07971 **10** (2017)
47. Wu, Y., He, K.: Group normalization. In: ECCV. pp. 3–19 (2018)
48. Zellers, R., Bisk, Y., Farhadi, A., Choi, Y.: From recognition to cognition: Visual commonsense reasoning. In: CVPR. pp. 6720–6731 (2019)
49. Zeng, K., Chen, T., Chuang, C., Liao, Y., Niebles, J.C., Sun, M.: Leveraging video descriptions to learn video question answering. In: AAAI. pp. 4334–4340 (2017)
50. Zhang, C., Gao, F., Jia, B., Zhu, Y., Zhu, S.C.: Raven: A dataset for relational and analogical visual reasoning. In: CVPR (2019)
51. Zhang, C., Gao, F., Jia, B., Zhu, Y., Zhu, S.C.: Raven: A dataset for relational and analogical visual reasoning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5317–5327 (2019)
52. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: ECCV. pp. 803–818 (2018)
53. Zhou, L., Louis, N., Corso, J.J.: Weakly-supervised video object grounding from text by loss weighting and object interaction. In: BMVC. p. 50 (2018)
54. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: AAAI. pp. 7590–7598 (2018)
55. Zhukov, D., Alayrac, J., Cinbis, R.G., Fouhey, D.F., Laptev, I., Sivic, J.: Cross-task weakly supervised learning from instructional videos. In: CVPR (2019)