Toward a Human-Level Video Understanding Intelligence

Yu-Jung Heo[1*], Minsu Lee[1,2*], Seongho Choi [1], Woo Suk Choi [1], Minjung Shin [3], Minjoon Jung [1], Jeh-Kwang Ryu [3] and Byoung-Tak Zhang [1,2]

[1] Seoul National University, [2] AI Institute (AIIS), Seoul National University, [3] Dongguk University
{yjheo,mslee,shchoi,wschoi,mjjung,btzhang}@bi.snu.ac.kr, {shinymj,ryujk}@dongguk.edu,

Abstract

We aim to develop an AI agent that can watch video clips and have a conversation with human about the video story. Developing video understanding intelligence is a significantly challenging task, and evaluation methods for adequately measuring and analyzing the progress of AI agent are lacking as well. In this paper, we propose the Video Turing Test to provide effective and practical assessments of video understanding intelligence as well as human-likeness evaluation of AI agents. We define a general format and procedure of the Video Turing Test and present a case study to confirm the effectiveness and usefulness of the proposed test.

Introduction

AI agents that can watch videos with people and share empathy for video content through various conversations are promising AI applications that people expect. To this end, AI agents have to accurately perceive and recognize contents of a video and have a natural muti-turn conversation with people based on understanding of the contents.

Recently, researches on text-to-video retrieval, video captioning, and video question answering (videoQA) have been actively conducted to improve video understanding intelligence. In addition, large-scale datasets have been built and publicly available to facilitate the researches (Alamri et al. 2019; Lei et al. 2018, 2020; Choi et al. 2021). Studies using these datasets usually apply automatic evaluation metrics to measure the performances of AI agents. For videoQA task, multiple-choice QA typically uses overall accuracy, and open-ended QA applies evaluation metrics frequently used in natural language generation tasks (e.g., BLEU (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015)).

These automatic evaluation metrics are convenient to apply, but they also have limitations. For example, the overall accuracy, while being intuitive and easy to calculate, does not take into account the difficulty of questions or required cognitive components. In addition, the scores of evaluation metrics for language generation models cannot determine whether the content is a correct answer to the question.

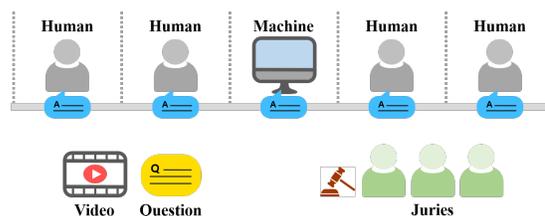Conversation about video story after watching a video clip

Figure 1: An illustration of the Video Turing Test (VTT). Detailed procedure of VTT is described in 'Video Turing Test' section.

Therefore, these metrics are hard to apply to measure the video understanding ability and the human-likeness of an AI agent. Furthermore, since story understanding task usually leads to multiple interpretations, it is essential to evaluate how persuasive the answers of AI are for people.

In this paper, we propose a novel Video Turing Test (VTT) as a method for evaluating the video understanding ability and the human-likeness of an AI agent (Figure 1). In VTT, all players see a given video clip consisting of multimodality, then have question and answering about the video story. We also exploit a new evaluation metric, CogME, established on the story elements and thinking strategies. We can analyze intelligence in the aspect of humans' understanding process through the metric.

Related Work

Turing Test

In 1950, Alan Turing proposed a variation of imitation game, now called Turing Test, where a computational machine and a human player communicate with an interrogator using text while invisible to each other (Turing 1950). The objective of the machine is to ensure that the interrogator cannot distinguish the machine from the human player. If the machine achieves the objective, we can regard the machine intelligence can be considered human-like. The Turing Test has been established as a novel measurement, but has been criticized for the following limitations. First, a machine can pass the test based on 'trickery or guile' (Weizenbaum 1966;

Shieber 1994; Boden 2006). Second, the Turing Test focused primarily on language capability without considering additional components of human intelligence (e.g., visual understanding). Third, the passing criteria for the Turing Test, as well as the interrogator's judgement, are quite subjective.

To complement the limitations, several new suites of tests are suggested by extending the original Turing Test. To provide a more quantitative measure of intelligence preventing passing the test by trickery or guile, (McKinstry 1997; Geman et al. 2015) considered only boolean type questions about a series of facts or an image scene, respectively. To avoid reliance on conversational ability, (Olague et al. 2021) suggested a purely visual processing-based Turing test which aims to imitate the emotional interpretation of humans. On the other hand, (Adiwardana et al. 2020) proposed a novel evaluation metric, Sensibleness and Specificity Average to capture human-likeness of conversational ability of a machine. (Adams, Banavar, and Campbell 2016) suggested I-athlon as a multidimensional Turing Test. The authors aim to evaluate a wide variety of intelligence behavior such as video understanding, not only the conversational ability of machines. Lastly, (Zhang 2009) proposed a Multimodal Memory Game where human teachers watch a video with an AI agent and teach the agent to generate text from a given image (video frame) or generate an image from a given text. This work shares similar motivations and experimental settings with ours, however, we focus on developing and evaluating the video understanding intelligence via question answering about video story.

Video Turing Test

This section describes a novel Turing Test, called Video Turing Test (VTT), for measuring video understanding intelligence. Video understanding ability includes many dimensions of intelligence, such as linguistic ability, visual understanding, story understanding, and reasoning. To circumvent the well-known limitations of the Turing Test described in the previous section, the VTT is designed as a standardized test with an effective and practical assessment of many aspects of video understanding intelligence.

Procedure for Video Turing Test

The general VTT procedure is illustrated in Figure 1. All participants are separated from one another. Especially the jury can not see all players. A set of paired video clips and questions is prepared. Then, the rest of the procedures are as follows: (i) Players and the jury watch a video clip. (ii) A question about the video story is given to both the players and the jury. (iii) Each player submits an answer to the question. (iv) The submitted answers are presented to the jury. (v) After checking all the answers, the jury guesses who the AI agent is and votes for their prediction.

Format of Video Turing Test

Open-ended multi-turn conversations about a given video clip among players and an interrogator may be regarded as the most appropriate approach to confirm the human-likeness of AI agents. However, since the open-ended conversation about the video is analogous to the Turing Test,

the resulting conversation scripts naturally inherit the drawbacks of the Turing Test. Therefore, it is difficult to qualitatively and quantitatively measure the degree of video understanding intelligence. VTT should be easy to evaluate but hard to solve. The test must not be solvable using trickery or guile. To meet these requirements, we argue that a question and answer is more suitable for the VTT. The QA format allows us to thoroughly examine the video understanding intelligence by carefully selecting the set of intended questions and reviewing the players' answers. Unlike conversation tasks where golden answer does not exist, it is relatively apparent to evaluate answers to questions.

We can consider three QA types according to the way of expressing the answer to the question. The most definite QA type is to ask multiple-choice questions. Since multiple-choice question requires to pick an answer among several candidate answers, the evaluation process becomes simple. Another more challenging option is to leave the answer open-ended. For short-answer questions, assessing the correctness of the answers is straightforward. Short-answer questions are frequently used in quiz shows because this type of QA is simple and clear. Full-sentence-answer questions are the most appropriate method for examining human-level video understanding intelligence. The full-sentence answer represents players' linguistic competence and thinking perspectives as well as correctness of the answer. Juries can get great hints about the human-likeness of players from their fluency. However, in order to evaluate whether each full-sentence answer is correct or not, it requires an additional human scoring work based on the conversational human-likeness (sensibleness) and correctness of the answer (specificity) (Adiwardana et al. 2020). Organizers of the VTT can select an appropriate QA format according to the circumstances such as the characteristics of the players, the target audience and the test environment.

Evaluation metrics for Video Turing Test

We aim to evaluate whether an AI agent understands video, and to analyze the test results so that future research directions can be suggested based on the current strengths and weaknesses of the AI agent. To this end, it is important to prepare carefully selected a set of questions that can evaluate the players' video understanding ability in a systematic way. In addition, by checking the answers to the question, it should be possible to evaluate each player's video understanding intelligence specifically. By using a systematic evaluation metric for composing a set of questions and analyzing the level of intelligence of players, we can improve the reliability and usefulness of the VTT.

For a systematic evaluation of video understanding intelligence, we develop Cognitive Modules for Evaluation (CogME) which is a top-down evaluation system for videoQA based on the cognitive process of human and story elements (Shin et al. 2021). CogME is composed of three cognitive modules: `Target`, `Content`, and `Thinking`. The interaction among the modules in the understanding procedure can be expressed as "I understand the `Content` of the `Target` through a way of `Thinking`". Each module has corresponding story elements as follows.

Question: What did Dokyung do in this scene?	
Story elements: Character and Event in Target, Identity in Content, Recognition in Thinking	
AI agent	Answer: Dokyung was sitting on the ground. (✗)
Pre-operational stage	Answer: Dokyung is looking at Haeyoung. (✗)
Middle concrete stage	Answer: He is sitting down next to Haeyoung. (✓)
Concrete generalization stage	Answer: He was sat on the hospital bed. (✓)
Formal stage	Answer: Dokyung was sitting in the bench in the hospital within Haeyoung's sight. (✓)

Table 1: A question and answer example sampled in the case study of the Video Turing Test (VTT). The check-mark (✓) and x-mark (✗) represent correct and wrong answers, respectively.

- `Target`: Character, Object, Place, Conversation, Behavior, Event, Emotion, Commonsense
- `Content`: Identity, Feature, Relationship, Means, Context, Sequence, Causality, Motivation
- `Thinking`: Recall, Recognize, Reasoning

The story elements can be assigned on the questions to specify required information, knowledge, or thinking ability by each question. Also, the correct answer rate for each story element is calculated by combining results from the answers. Through this analysis, we can identify the strengths and weaknesses of the video understanding intelligence of each player very specifically. The set of questions for the VTT can be organized so that CogME attributes appear evenly across the questions. This means that all story elements related to video understanding intelligence can be assessed with a set of questions.

The composition of human players

Human intelligence is not uniform, and there are significant differences in the cognitive aspect by age group according to their developmental stages and individual characteristics. Each person has a different pattern of strengths and weaknesses in their performances and related components of cognition. For example, children are generally good at recall but clumsy in reasoning, while some adults usually focus on the main character in a video clip and ignore surroundings. Therefore, judging the human-likeness of an AI agent can be influenced by the diverse characteristics of the interrogator or the human player. Ultimately, it becomes difficult to generalize the results of the test. To evaluate the human-likeness of an AI agent, VTT conducts tests with a group of human players rather than a single human player. Depending on the characteristics of the human player group, the test can derive more concrete and specified results. For example, if test organizers want to evaluate an AI agent based on age groups, human player groups can be organized based on the developmental stages of Piaget's theory (Piaget 1972).

Case Study

Here, we provide a case study of the VTT to confirm the feasibility and effectiveness of the proposed test. We explain detailed implementations of the case study in this section.

Video selection. We utilize shot and scene clips sampled from a Korean popular TV drama "Another Miss Oh".

Question selection. We carefully select 30 questions for the case study considering i) the distribution of story elements,

ii) question type (i.e., 15 multiple-choice QA vs. 15 open-ended QA), and iii) associated video length for each question (i.e., 17 QA for shot clip and 13 QA for scene clip).

Human players. We organize a human player group that consists of four people from the different cognitive developmental stages of humans (Heo et al. 2019). The age of the four players is 4 years (pre-operational stage), 10 years (middle concrete stage), 14 years (concrete generalization stage) and 20+ years (formal stage), respectively.

AI agent. As an AI agent, we employ two video QA algorithms: (Choi et al. 2021) for multiple-choice QA and (Lee et al. 2021) for open-ended QA.

Juries. As juries for the test, we invite 16 people of ages from the 20s to the 50s and both sexes. Six people of juries are experts for AI and others are non-expert audiences.

Miscellaneous. We divide 30 questions into 5 rounds of 6 questions. All participants are randomized after each round. The juries are provided with a judgment sheet to record one's thoughts of each QA. After each round consisting of 6 QAs, the juries have to guess who is the AI, based on the judgment sheet as a cue.

Results

Question answering and juries' decision

First, we present examples of question answering conducted in the case study. In Table 1, given questions, story elements associated with the questions, and answers generated by players in different developmental stages are described. As shown in the table, our AI agent or few humans fails to answer a given question. At each round after conducting six question and answering pairs, the juries guess who is the AI among the players. As a result, juries perceive the AI around 62.5% and 75% at the first and second rounds, respectively. On the contrary, in the third round and fifth round, juries vote one of the human players as an AI around 62.5% and 87.5%, respectively. Interestingly, in the fourth round, the juries are quite confused and the voting results were a close match. The juries guessing the AI agent as an AI around 55.6% (correct guess), and guessing one of the human players as an AI around 45.4% (incorrect guess). In 4 rounds among a total of 5 rounds, one player receives more than half of the votes with a consistent result[1].

---

[1]The juries guess the AI agent as an AI around 48.6% on average for 5 rounds.
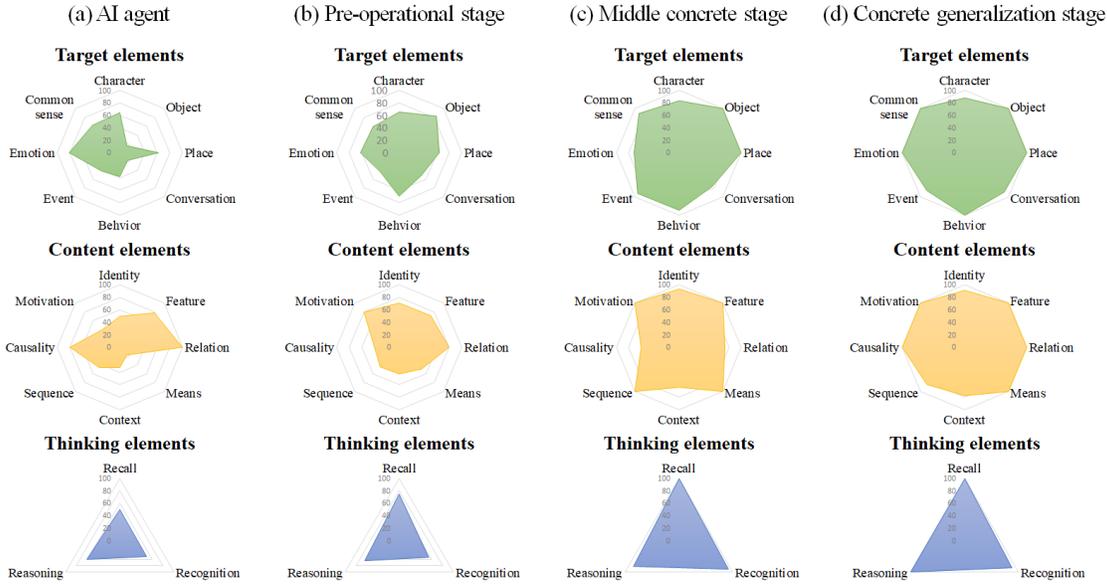
Figure 2: The accuracy-based profiles for each player participated in the case study of the Video Turing Test (VTT). To evaluate each player's video story intelligence, we apply CogME based on cognitive process of human.

Analysis on cognitive modules

We analyze each player's performance based on CogME for specifying their video understanding intelligence. We calculate the correct answer rate (accuracy) for each story element associated with the questions used in the case study. The accuracy-based profiles for each player on story elements in the three cognitive modules (i.e., target, content, thinking) are shown in Figure 2. Here, we highlight two key observations as follows. First, as the developmental stages of human progress, the accuracy of QA across all three modules is gradually improved. This means that some cognitive processes mature in a way that is reflected in the chosen metrics of this test. This result validates the use of these measures to evaluate video understanding intelligence. Second, when comparing the profiles between the AI agent and the human players, we notice that the performance of each component revealed more evenly in human players than the AI agent. Quantitatively, the standard deviation of the accuracy of each story element is 13.78 for a child in the pre-operational stage and 23.13 for the AI agent. In particular, the AI shows deficient performance in Object and Conversation in Target elements, Means and Motivation in Context elements, and Recall in Thinking elements. It is expected for two reasons that the AI agent's profile differs from the previous paper (Shin et al. 2021). This study considers two types of QA, multiple-choice and open-ended, while the previous work use only the former. Furthermore, we employ two distinct video QA algorithms as our AI agents to answer multiple-choice and open-ended QA.

Discussion and Conclusion

In this paper, we introduced the Video Turing Test (VTT), a novel measurement to evaluate a human-likeness of video understanding intelligence. We defined a general format and procedure of VTT and conducted a case study to confirm the feasibility and effectiveness of the proposed test. The case study provided new insight into the association of video understanding intelligence between the AI agent and human players from the different developmental stages.

While the case study suggested a new perspective of measurement for video understanding intelligence, we still need to discuss several aspects. First, a 4-years child was included in the test as a human player for comparing the video understanding ability between the AI agent and a human from the pre-operational stage. However, several requirements (e.g., writing as a full sentence, choosing an answer among five answer candidates) were not familiar for the child so that it is hard to interpret the answers as totally reflecting the child's video understanding ability. Furthermore, we did not strictly specify the criteria to pass the VTT. The passing criteria can be defined by considering the detailed design of the VTT such as the composition of human players, juries, and the selected question set. As future work, we expect that the additional case study and its interpretation validates the clearness of VTT. For more objective and analytical evaluation for VTT, we plan to conduct various case studies and suggest appropriate guidelines including the composition of participants and the arrangement of question set.

Acknowledgements

References

Adams, S. S.; Banavar, G.; and Campbell, M. 2016. I-athlon: Towards a multidimensional turing test. *AI Magazine* 37(1): 78–84.

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977* .

Alamri, H.; Cartillier, V.; Das, A.; Wang, J.; Cherian, A.; Essa, I.; Batra, D.; Marks, T. K.; Hori, C.; Anderson, P.; et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7558–7567.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Boden, M. A. 2006. *Mind as Machine: A History of Cognitive Science*, volume 1. Oxford University Press.

Choi, S.; On, K.-W.; Heo, Y.-J.; Seo, A.; Jang, Y.; Lee, M.; and Zhang, B.-T. 2021. DramaQA: Character-Centered Video Story Understanding with Hierarchical QA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 1166–1174.

Geman, D.; Geman, S.; Hallonquist, N.; and Younes, L. 2015. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences* 112(12): 3618–3623.

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.

Lee, D.; Choi, S.; Jang, Y.; and Zhang, B.-T. 2021. Mounting Video Metadata on Transformer-based Language Model for Open-ended Video Question Answering. *CoRR* abs/2108.05158. URL https://arxiv.org/abs/2108.05158.

Lei, J.; Yu, L.; Bansal, M.; and Berg, T. 2018. TVQA: Localized, Compositional Video Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1369–1379.

Lei, J.; Yu, L.; Berg, T. L.; and Bansal, M. 2020. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, 447–463. Springer.

McKinstry, C. 1997. Minimum intelligent signal test: an objective Turing Test. *Canadian Artificial Intelligence* 17–18.

Olague, G.; Olague, M.; Jacobo-Lopez, A. R.; and Ibarra-Vazquez, G. 2021. Less Is More: Pursuing the Visual Turing Test With the Kuleshov Effect. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1553–1561.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Piaget, J. 1972. Intellectual Evolution from Adolescence to Adulthood. *Human Development* 15(1): 1–12.

Shieber, S. M. 1994. Lessons from a restricted Turing test. *Communications of the Association for Computing Machinery* 37.

Shin, M.; Kim, J.; Choi, S.; Heo, Y.; Kim, D.; Lee, M. S.; Zhang, B.; and Ryu, J. 2021. CogME: A Novel Evaluation Metric for Video Understanding Intelligence. *CoRR* abs/2107.09847. URL https://arxiv.org/abs/2107.09847.

Turing, A. 1950. Computing Machinery and Intelligence. *Mind* 59(236): 433–460.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4566–4575.

Weizenbaum, J. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Communications of the ACM* 9(1): 36–45.

Zhang, B.-T. 2009. Teaching an Agent by Playing a Multimodal Memory Game: Challenges for Machine Learners and Human Teachers. In *AAAI Spring Symposium: Agents that Learn from Human Teachers*, 144.