ABCDE: An Agent-Based Cognitive Development Environment

Jieyi Ye
Nanyang Technological University (NTU)
ye0003yi@e.ntu.edu.sg

Jiafei Duan
Institute for Infocomm Research, A*STAR
duan_jiafei@i2r.a-star.edu.sg

Samson Yu
Centre for Frontier AI Research
samson_yu@ihpc.a-star.edu.sg

Bihan Wen
NTU
bihan.wen@ntu.edu.sg

Cheston Tan
Centre for Frontier AI Research
cheston-tan@i2r.a-star.edu.sg

Abstract

Children's cognitive abilities are sometimes cited as AI benchmarks. How can the most common 1,000 concepts (89% of everyday use) be learnt in a naturalistic children's setting? Cognitive development in children is about quality, and new concepts can be conveyed via simple examples. Our approach of knowledge scaffolding uses simple objects and actions to convey concepts, like how children are taught. We introduce ABCDE, an interactive 3D environment modeled after a typical playroom for children. It comes with 300+ unique 3D object assets (mostly toys), and a large action space for child and parent agents to interact with objects and each other. ABCDE is the first environment aimed at mimicking a naturalistic setting for cognitive development in children; no other environment focuses on high-level concept learning through learner-teacher interactions. The simulator can be found at https://pypi.org/project/ABCDESim/1.0.0/

1. Introduction

The cognitive abilities of children are sometimes held up as informal benchmarks for AI. For example, Bengio in 2019 said that "I don't think we're anywhere close today to the level of intelligence of a two-year-old child" [13]. Typical 5-year-olds have probably experienced between 0.5M to 50M[1] "training examples" (including many repetitions), which are not too far off from the size of current AI datasets, e.g. 22M images [3] and 8M videos [1].

However, the characteristics are very different. 5-year-olds are estimated to recognize 5K to 10K words [8] – far fewer than the 20K+ object classes in ImageNet and 100K+ synsets in WordNet. Given a hypothetical "budget" of 5M training examples, children would experience something

---

[1]5.3M = 5 years x 365 days x 12 waking hours x 240 "learning events" (each 15 secs) per hour. Then allow for factor of 10 more or fewer.

like 5K classes x 20 unique instances x 50 variations, rather than a typical AI dataset of say 20K classes x 250 unique instances. Of course, children eventually attain adult-level vocabularies, typically 25K to 50K English words [8].

Cognitive development is about quality before quantity. The current AI paradigm of "training" is akin to animal training. However, conceptual knowledge can be conveyed via simplified, abstracted examples [16]. Children are taught, not trained, and educational interactions are designed to convey ideas with clarity and illustrativeness.

In this work, we introduce ABCDE, an interactive 3D environment modeled after a typical playroom for children, along with child and parent avatars. It comes with 300+ unique 3D object assets, most being toys with various physical attributes. There is a large action space for child and parent agents to interact with objects and each other. ABCDE is the first interactive environment that mimics a naturalistic setting for cognitive development in children; no other environment [6, 7, 9–11] focuses on high-level concept learning through learner-teacher interactions.

2. Objective and Approach

The objective of developing ABCDE is to start addressing this challenge: Can common concepts be learnt in a grounded, naturalistic children's setting? The most common 1,000 concepts cover 89% of everyday usage [8]. They fall into a variety of word classes; not just nouns, verb and adjectives, but also other important classes such as determiners, prepositions, etc. In contrast, AI tasks have mostly been for nouns (objects), verbs (actions) and adjectives (attributes), and these have mostly been tackled in isolation.

Our proposed approach is to use simple objects and actions to convey concepts, similar to how young children are taught in a playroom. This approach uses knowledge scaffolding, e.g. using tangible nouns and verbs to convey abstract prepositions and determiners, etc. The focus is on common concepts, not necessarily activities of adult daily
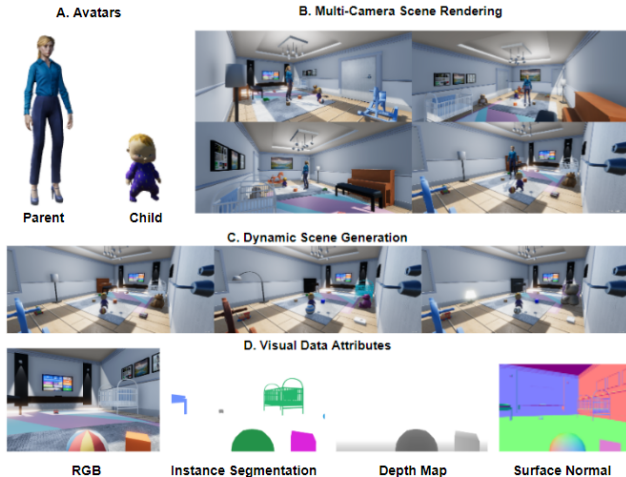
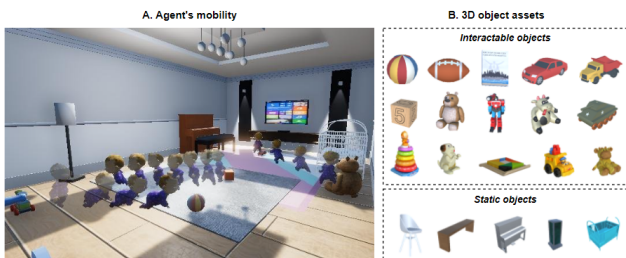Figure 1. Overview of the various components of ABCDE.

Figure 2. A) Time-lapse visualization depicting some child actions. B) The two categories of 3D object assets.

living (e.g. cooking and general household tasks) [2, 5, 12]. A parent avatar is used for realism (objects can't move by themselves), embodiment, referencing (e.g. pointing to an object being mentioned verbally) and for imitation learning (i.e. for child agents to observe body motions). ABCDE can be used for both training and evaluation of learning agents.

3. ABCDE Simulator

The Agent-Based Cognitive Development Environment (ABCDE) is designed with 3 key aspects. **1)** A parent (teacher) agent, which can interact with the child (learner) agent in several ways, including passively "speaking" (generating natural language text) when the child is playing, and also actively demonstrating to the child. **2)** The natural language text "spoken" by the parent can be generated at different levels of grammatical complexity. **3)** The parent can be algorithmically programmed to carry out high-level actions to demonstrate concepts, meaning that labelled video examples can be generated for free. For instance, one way to demonstrate the "put-on" concept is to pick up object A, then lower it over object B until they are in contact (e.g. "put the cup on the table"). This approach allows for tangible demonstration for grounding of natural language con-

cepts that can sometimes be abstract, e.g. "**only** the red ball is in the box, not the green balls".

3.1. Scene Structure

ABCDE contains a single scene that allows for an unlimited amount of room configuration through the random sampling of interactable objects at potential locations on the floor grid. Figure 1C shows the dynamic generation of scene. The interactable objects will be spawned with a separation buffer of +/-0.5 units between objects, so as to prevent overlaps. More object classes can be easily imported and spawned at available locations on the grid. Dynamic generation could also be initialized via the Python module. All the objects' positions and metadata are logged.

3.2. Actions

The simulator allows both the child and parent agents to interact with and move around the environment via specific actions. For navigation, the available actions available are [*NavigateTo, Run, Crawl, WalkForward, WalkBackwards, TurnLeft, TurnRight, Wander*]. For interactions, the actions are [*Grab, LookAt, PutBack, LookAround, Touch, Rotate*]. The interactions with the object assets within the environment are also influenced by the internal proprieties of the 3D objects. The 3D objects in the environment are split into two categories: [*Static, Interactable*]. All interactions use the Nvidia PhysX Engine to enable realistic physics.

3.3. Observations

ABCDE supports a wide range of visual data attributes, from RGB, depth and instance segmentation to surface normal data (Figure 1D). We provide four stationary cameras to provide full visual coverage of the environment and render all four cameras simultaneously (Figure 1B). All observations are output at customizable frame rates. We also provide a GUI for easier control of the scene and agents.

3.4. 3D Object Assets

We use 332 3D object assets from public marketplaces and the Google Scanned Object dataset [4], falling into two categories: 89 static objects and 243 interactable objects.

3.5. Tasks

Following the nature of different types of knowledge and the corresponding organization of human memory [14, 15], the tasks cover both declarative and procedural concepts. Declarative concepts include "static" concepts such as nouns, adjectives, etc. Procedural concepts are typically actions or action sequences. These concepts could potentially be tested in various ways, including telling the child agent to demonstrate the concept (e.g. "**keep** the red ball in the box"), Q&A (e.g. "is the toy car **on** the table or **under** it?"), fill-in-the-blank ("the toy car is ___ the table"), etc.

References

[1] https://research.google.com/youtube8m/. 1

[2] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The EPIC-KITCHENS dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 1

[4] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Michael Hickman, Krista Reymann, Thomas B. McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3D scanned household items. 2022. 2

[5] Jiafei Duan, Samson Yu, Hui Li Tan, and Cheston Tan. Actionet: An interactive end-to-end platform for task-based data collection and augmentation in 3D environment. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 1566–1570. IEEE, 2020. 2

[6] Jiafei Duan, Samson Yu, Hui Li Tan, Hongyuan Zhu, and Cheston Tan. A survey of embodied ai: From simulators to research tasks. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2022. 1

[7] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, et al. ThreeD-World: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*, 2020. 1

[8] Nickee De Leon Huld. How many words does the average person know? https://wordcounter.io/blog/how-many-words-does-the-average-person-know/, May 2022. 1

[9] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. AI2-Thor: An interactive 3D environment for visual AI. *arXiv preprint arXiv:1712.05474*, 2017. 1

[10] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. VirtualHome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8494–8502, 2018. 1

[11] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9339–9347, 2019. 1

[12] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10740–10749, 2020. 2

[13] Eliza Strickland. AI experts speak: Memorable quotes from Spectrum's AI coverage. https://spectrum.ieee.org/artificial-intelligence-quotes, Sept. 2021. 1

[14] Cheston Tan, Stephane Lallee, and Bappaditya Mandal. Vision and memory: Looking beyond immediate visual perception. In *Computational and Cognitive Neuroscience of Vision*, pages 195–219. Springer, 2017. 2

[15] Hui Li Tan, Hongyuan Zhu, Joo-Hwee Lim, and Cheston Tan. A comprehensive survey of procedural video datasets. *Computer Vision and Image Understanding*, 202:103107, 2021. 2

[16] C Lawrence Zitnick and Devi Parikh. Bringing semantics into focus using visual abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3009–3016, 2013. 1