In search of the embgram: forming episodic representations in a deep learning model

Chad DeChant (chad.dechant@columbia.edu)
Computer Science Department, Columbia University
New York, NY, United States

Iretiayo Akinola (iakinola@nvidia.com)
NVIDIA
Seattle, WA, United States

Daniel Bauer (bauer@cs.columbia.edu)
Computer Science Department, Columbia University
New York, NY, United States

Abstract

Enabling episodic memory in a robot or other AI agent would lead to better functioning AI as well as creating opportunities for modeling theories of human memory. An important step toward such an episodic memory system is developing representations that can be used to accurately recover information about past events. We introduce a method to obtain such representations with an artificial neural network model. To study episodes of realistic length, we utilize ego-centric video data from a dataset of an agent performing household chores in a simulated environment. In the first training phase, we use a transformer model to encode frames from this video into compact vector embeddings — *embgrams*. Next, using the embgrams as input, a second transformer model is fine-tuned to provide natural language descriptions and answers to questions about the episodes. Our results show that the embgrams facilitate retrieval of episode-related information. Importantly, we find that the usefulness of the embgrams as stores of information significantly depends on the task the encoding transformer performs during their generation.

Keywords: episodic memory; attention; representation learning

Introduction

While there has been a great deal of cognitive science research on developing models of episodic memory, the machine learning community has devoted much less attention to trying to incorporate something analogous to episodic memory in robotics or other ML systems. Cognitive science efforts have naturally prioritized models that can help further our understanding of episodic memory in humans or other animals; these often involve storing small amounts of data such as still images (Spens & Burgess, 2024). Machine learning models, by contrast, process much larger amounts of information but, with some exceptions, notably in reinforcement learning, have typically not included anything analogous to episodic memories (Nematzadeh, Ruder, & Yogatama, 2020; Lampinen, Chan, Banino, & Hill, 2021).

In an effort to begin bridging this divide, we introduce a method to form compact representations of long action sequences from simulated ego-centric video frames using a transformer-based model. We demonstrate that these representations can later be used in ways similar to episodic memory: recounting a natural language narrative of what happened; answering questions about what happened; and recovering what actions happened after any arbitrary point in the episode.

Given that the transformer model treats these vectors as embeddings, we refer to these embedding vectors as *embgrams*, following the name of the physical basis of human memories, engrams. We do not address the storage or retrieval of embgrams here, but we suggest that their structure will allow them to be more easily localized and understood than the notoriously difficult to identify engrams (Lashley, 1950; Eichenbaum, 2016).

What humans attend to affects what we remember (Aly & Turk-Browne, 2017). We observe something similar for models trained to produce embgrams. By varying the tasks used to train the embgram-generating models, we discover that the content the model is prompted to attend to has a marked effect on the usefulness of the resulting embgrams as stores of episodic information.

Methods

Data

We use episodes of action sequences from a dataset (Shridhar et al., 2020) containing ego-centric video of an agent in a virtual environment (Kolve et al., 2017) performing complicated, multi-step tasks, e.g. washing, slicing, and refrigerating an apple. We use a subset of frames (an average of 50 per episode, matched to each action) from the video along with a second dataset (DeChant, Akinola, & Bauer, 2023) of natural language questions and answers paired to each episode.

Model

A multimodal artificial neural network model including a T5 encoder-decoder transformer (Raffel et al., 2020) is trained to

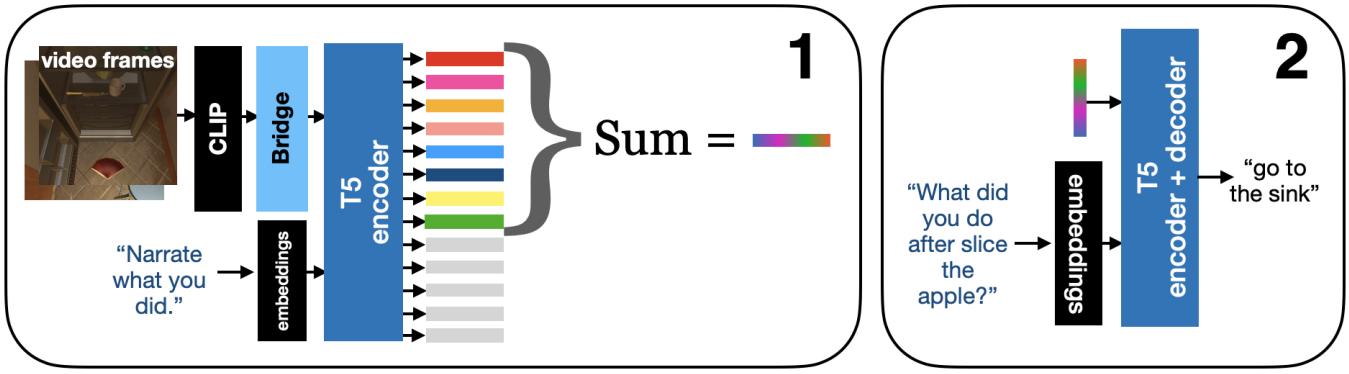

Figure 1: Depiction of the creation and use of an embgram. In step one, an episode is encoded into representations by the T5 encoder. The representations corresponding to word embeddings (gray) are discarded; those corresponding to the input images (multicolor) are summed into one embgram vector. In step two, this embgram is used as the input along with a question.

perform tasks based on video frames from each episode in the training set. Image frames are encoded by a CLIP encoder (Radford et al., 2021) and projected into the T5's encoding space by a small bridge neural network. The T5 model's encoder is fine-tuned to generate encodings of each frame conditioned on a natural language question. These encodings are then summed to form an embgram, a single vector representation of the episode. This embgram then serves as input to a second T5 transformer which is trained to answer natural language questions about the episodes without access to the original video frames. See Figure 1 for a schematic.

Experiments

We conduct two sets of experiments. In the first set, we obtain embgrams from a model trained to perform full narration of the episode (step 1 in Figure 1). We then train a new model on a variety of question answering tasks using the embgram vectors as input (step 2 in figure 1). For the second set of experiments, we instead generate embgrams by training a model to perform additional question answering tasks beyond narration. The resulting embgrams then serve as input to a model that we train to generate full narrations.

In addition to narration, the question types asked are: *Object either/or*: "was there an apple or CD?" and *Action just after*: "what did you do just after go to the refrigerator?" A final, fourth type of prompt is *Act from history*, which asks the model to output the next five actions the agent takes in the episode, given a set of image frames up to a particular point. In the encoding phase this trains a model to use the images to choose actions. In the second phase, the generated embgram is provided to the model which is trained to use it to recover the sequence of actions that took place.

Results

The first set of experiments shows that embgrams can be used to answer some questions about an episode at close to the same level of accuracy as a baseline model trained directly on the original video frames. Narratives are approximately 20% less accurate. Recovery of actions taken at a given point in the episode is more accurate than a baseline model which is not given an embgram but instead can only predict actions based on prior images. See Table 1. All results are reported from a validation set of previously unseen episodes and environments.

Table 1: Accuracy for question answering and narration from embgram vectors by question type, compared to a baseline accuracy of the model trained directly on the video frames. These embgrams were all generated by the narration model.

Question / prompt	Accuracy	% difference from baseline
Narration	37.4%	-21.3%
Action just after	77.2%	-5.8%
Object either/or	97.2%	+1%
Act from history	38.5%	+22.1%

The effect of attention on memory The second set of experiments demonstrates that the usefulness of embgrams as stores of episodic information is determined by the task the model performs during their generation (see Table 2). The "Action just after" questions result in embgrams which are nearly as useful in producing full narratives as the narration prompt is, while those produced by a model prompted to predict a sequence of actions ("Act from history") are slightly less than half as accurate. Object-focused questions result in embgrams with near zero accuracy in producing narratives.

Discussion

We found that a transformer-based model can be used to generate and use embgram vectors representing episodes of extended action. The usefulness of these embgrams is determined by the task the model performs while encoding the embgrams, which influences what parts of the input the encoder's self-attention mechanism attends to. We propose that

Table 2: Generating narrations from different types of embgrams: accuracy of producing narrations from a single embgram vector derived from models trained for the listed tasks.

Source of embgrams	Accuracy
Narration model	37.4%
Action just after question model	33.2%
Object either/or model	0.8%
Act from history only model	14.6%

this parallels the role attention plays in the formation of human memories. A model trained to answer questions about objects generates embgrams which are completely unable to recover a full narrative. More notably, a model which is trained to predict what actions to take during an episode does not produce very useful embgrams. Given that robots or other AI agents will likely focus their attention on aspects of the environment or other inputs which are useful when taking actions, this suggests that the creation of representations for episodic memory purposes might be challenging. A separate attention stream may be needed in order to form useful representations of episodes for later recall.

Using transformer models to create embgrams opens up a variety of ways to study episodic memory-like processing in a machine learning context, with potential applications to the study of human memory. Future work will address storage and retrieval, improved methods to compress episodic representations into embgrams, recall of visual stimuli, and investigate how information is encoded in an embgram.

References

Aly, M., & Turk-Browne, N. B. (2017). How hippocampal memory shapes, and is shaped by, attention. *The hippocampus from cells to systems: Structure, connectivity, and functional contributions to memory and flexible cognition*, 369–403.

DeChant, C., Akinola, I., & Bauer, D. (2023). Learning to summarize and answer questions about a virtual robot's past actions. *Autonomous Robots*, *47*(8), 1103–1118.

Eichenbaum, H. (2016). Still searching for the engram. *Learning & behavior*, *44*, 209–222.

Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., ... Farhadi, A. (2017). Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*.

Lampinen, A., Chan, S., Banino, A., & Hill, F. (2021). Towards mental time travel: a hierarchical memory for reinforcement learning agents. *Advances in Neural Information Processing Systems*, *34*, 28182–28195.

Lashley, K. S. (1950). In search of the engram.

Nematzadeh, A., Ruder, S., & Yogatama, D. (2020). On memory in human and artificial language processing systems. In *Proceedings of iclr workshop on bridging ai and cognitive science.*

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... others (2021). Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, *21*, 1–67.

Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., ... Fox, D. (2020). Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 10740–10749).

Spens, E., & Burgess, N. (2024). A generative model of memory construction and consolidation. *Nature Human Behaviour*, 1–18.