Title
Comparing Intuitions about Agents' Goals, Preferences and Actions in Human Infants and Video Transformers

Journal
Proceedings of the Annual Meeting of the Cognitive Science Society, 45(45)

Authors
Hein, Alice
Diepold, Klaus

Publication Date
2023

Peer reviewed

Comparing Intuitions about Agents' Goals, Preferences and Actions in Human Infants and Video Transformers

Alice Hein (alice.hein@tum.de)
Department of Computer Engineering, School of Computation, Information and Technology
Technical University of Munich, Germany

Klaus Diepold (kldi@tum.de)
Department of Computer Engineering, School of Computation, Information and Technology
Technical University of Munich, Germany

Abstract

Although AI has made large strides in recent years, state-of-the-art models still largely lack core components of social cognition which emerge early on in infant development. The Baby Intuitions Benchmark was explicitly designed to compare these "commonsense psychology" abilities in humans and machines. Recurrent neural network-based models previously applied to this dataset have been shown to not capture the desired knowledge. We here apply a different class of deep learning-based model, namely a video transformer, and show that it quantitatively more closely matches infant intuitions. However, qualitative error analyses show that model is prone to exploiting particularities of the training data for its decisions.

Keywords: intuitive psychology; machine learning; action understanding

Introduction

The foundations of "commonsense psychology" emerge early on in a human's development: Even pre-verbal infants have expectations about agents' goals, preferences and actions (Stojnić, Gandhi, Yasuda, Lake, & Dillon, 2023). Although deep learning (DL) has made tremendous progress in recent years, this core component of human cognition is still lacking in many state-of-the-art DL models (Lake, Ullman, Tenenbaum, & Gershman, 2017). When tested on the Baby Intuitions Benchmark (BIB), a dataset designed to compare the social cognitive abilities of infants and machines, behavioral cloning (BC) and video prediction models based on recurrent neural networks (RNNs) failed to show infant-like reasoning (Gandhi, Stojnic, Lake, & Dillon, 2021). We here evaluate a different class of DL model, namely a video transformer (VT), on the BIB dataset.

Recent years have seen the rise of transformers in various areas of AI, including tasks adjacent to social cognition, such as trajectory prediction for cars or pedestrians (Yuan, Weng, Ou, & Kitani, 2021; L. L. Li et al., 2020; Chen, Wang, & Sun, 2021; Sui, Zhou, Zhao, Chen, & Ni, 2021; Giuliari, Hasan, Cristani, & Galasso, 2021; Yu, Ma, Ren, Zhao, & Yi, 2020) and spatial goal navigation (Du, Yu, & Zheng, 2021; Chaplot, Pathak, & Malik, 2021; Fukushima, Ota, Kanezaki, Sasaki, & Yoshiyasu, 2022). As the transformer attention mechanism is based on computing pairwise interactions (C. Li & Liu, 2022), this family of models constitutes a promising approach for capturing the relations between, e.g., agents and goals in the BIB dataset. However, transformer-based video prediction models require many costly pairwise computations. They are usually trained and evaluated on datasets like *Kinetics-400* (Kay et al., 2017) or UCF101 (Soomro, Zamir, & Shah, 2012), where video clip lengths range from 7 to 10 seconds – much shorter than those used in BIB, which may be up to 2 minutes long. We therefore implement some modifications to allow a VT to process BIB episodes, and evaluate the resulting model. We find that the VT quantitatively more closely matches infant intuitions about agent's goal preferences and efficient actions than previously tested DL baselines. However, qualitative error analyses show that the model fails to generalize systematically on some of the test tasks when agent or environment dynamics differ slightly from background training observations.

Baby Intuitions Benchmark

BIB is a dataset designed to test whether machine learning systems can discern the goals, preferences, and actions of others (Gandhi et al., 2021). It consists of videos in the style of Heider and Simmel's animations (Heider & Simmel, 1944), where agents, represented by simple shapes, carry out actions in a 2D grid world. BIB follows the violation-of-expectation (VoE) paradigm, i.e., each video has a familiarization and a test phase. The familiarization phase consists of eight successive trials during which an agent consistently displays a certain behavior, allowing the observer to form an expectation of future actions. The test phase includes an expected outcome (perceptually similar to the previous trials, but involves a violation of expectation), and an unexpected outcome (perceptually less similar, but conceptually more plausible). BIB contains six types of test tasks, outlined in Table 1. It also contains background training episodes with four types of training tasks, which share the same structure as the test set. However, Gandhi et al. designed the BIB dataset such that only expected trials are provided in the background training episodes, and only isolated tasks are trained. Therefore, the systematic combination of acquired knowledge is needed to generalize to the test tasks. For examples and more details on BIB, see Gandhi et al.

Because BIB adopts its tasks and paradigm from developmental cognitive science and provides sufficient data to train DL-based models, it allows for the direct comparison of human and machine performance (Gandhi et al., 2021). A critical first step in this direction was taken by Stojnić et al., who collected infants' responses on a representative selection of BIB episodes and compared them with three state-of-the-art

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society.*

Table 1: Overview of BIB tasks.

	Familiarization trials	Test trial	Expected outcome	Unexpected outcome
Preference	Agent consistently chooses one of two goal objects and moves to it efficiently	Identical to a familiarization trial, but object positions are switched	Agent moves to preferred object at new location	Agent moves to nonpreferred object at familiar location
Multi-agent		New agent appears	New agent moves to object not preferred by familiar agent	Familiar agent moves to previously not preferred object
Inaccessible goal		Preferred goal becomes inaccessible	Agent moves to other goal	Agent moves to other goal, even though both are accessible
Efficient agent	Agent moves efficiently around a barrier towards goal	Barrier is removed	Agent moves efficiently	Agent moves inefficiently
Inefficient agent	One agent moves efficiently, one moves inefficiently	Both agents move inefficiently	Previously inefficient agent moves inefficiently	Previously efficient agent moves inefficiently
Instrumental action	Agent removes a green barrier (inserts key into lock), then moves to goal	Green barrier gone or inconsequential	Agent moves directly to goal	Agent still moves to key

DL models from two classes: Behavioral cloning (BC) and video modeling. Recently, Zhi-Xuan et al. proposed a principled alternative to DL approaches, based on a hierarchically Bayesian Theory of Mind (HBToM). Results from both works serve as comparisons in this paper. Note, however, that HBToM requires access to symbolic states and is specifically engineered to solve BIB-like social cognition tasks, whereas the data-driven baselines and VT model have weaker inductive biases in this regard.

Methods

Our model consists of a convolutional neural network (CNN) encoder, a transformer component, a CNN decoder, and a linear output layer. A schematic visualization is shown in Figure 1. The CNN encoder (Figure 1 A) has two convolutional layers and two max-pooling layers. For each $3 \times 84 \times 84$ input image, it produces a $30 \times 21 \times 21$ representation, which we concatenate with x- and y-position encodings, yielding $32 \times 21 \times 21$ patches. As attending over every pixel would be computationally prohibitive, the CNN encoder was designed to reduce the frame's resolution by extracting higher-level features, while retaining a sufficient level of spatial detail.

After encoding all the frames of an episode in this way, we extract the top-n patches per frame that display the highest change compared to the previous frame (Figure 1 B). This is done for each frame of the familiarization trials. The reason we only use n patches is that attending over every patch, frame, and trial would be extremely computationally expensive. N was set to 3, as using a higher number would have exceeded the memory resources in our training setup, even with our very small batch size. However, it is unlikely that a choice of $n > 3$ would have led to substantially better performance, as BIB trials are mostly static. The only movements stem from the agent and, in *instrumental-action* tasks, the green barrier. Therefore, there are seldom more than three patches that exhibit a change from one frame to the next.

The extracted patches are fed into the first of three blocks of the transformer component. Each block has 5 layers with 8 heads of input dimension 32 and hidden dimension 256. The number of heads and layers was chosen to strike a balance

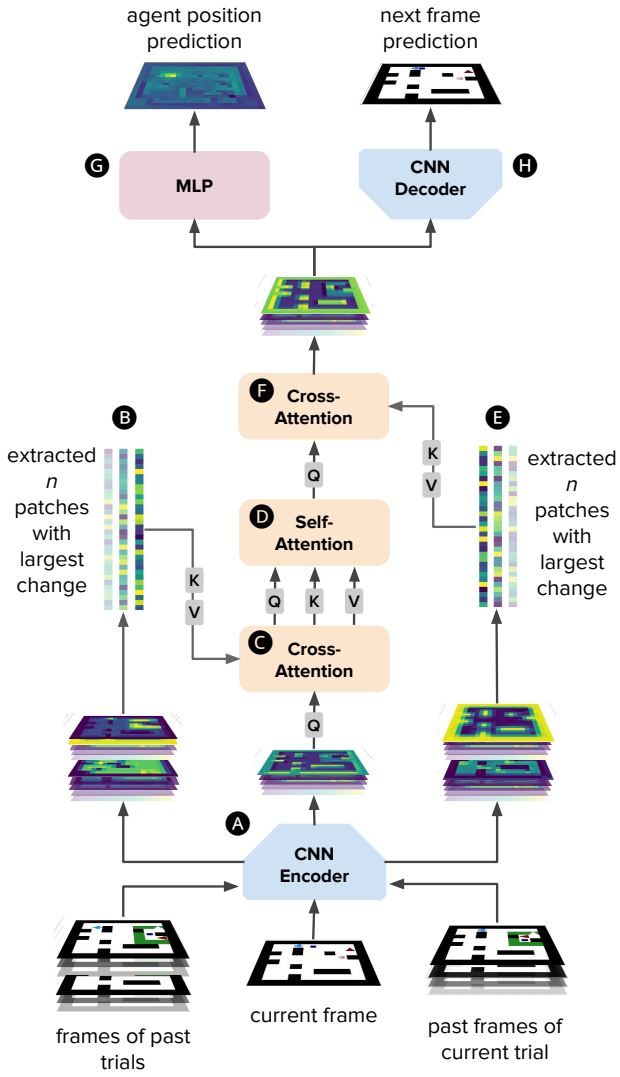


Figure 1: Schematic visualization of the VT architecture.

between performance and computational complexity. The first block (Figure 1 C) performs cross-attention over the test trial's encoded first frame and previous familiarization trials, effectively "priming" the model by calculating the influence of previous observations on the current input. The results of attending over each trial are averaged and passed through a self-attention block (Figure 1 D). We then extract n patches for each frame in the test trial (Figure 1 E) in the same way as we did for the familiarization trial frames. The patches serve as input to third attention block (Figure 1 F), which attends over past steps in the test trial. In a final step, the outputs of the transformer component are passed through an output layer (Figure 1 G), which produces a $1 \times 21 \times 21$ prediction of the agent's next position, and a CNN decoder (Figure 1 H), which produces a $3 \times 84 \times 84$ prediction of the next frame.

Given the model's two prediction targets, our loss function consisted of the sum of two terms. The first term was the binary cross-entropy (BCE) loss between the prediction of the agent's next step and the actual agent position. To address the imbalance between the "agent" and "no-agent" class, we employed a weighted version of the BCE loss, which is widely used in instance segmentation (Jadon, 2020). The second term was the mean squared error (MSE) between the prediction of the next frame and the actual next frame, upweighted by a constant factor so that both loss terms were scaled evenly. This second term was introduced because transformers may disregard agent identities unless incentivized otherwise (Yuan et al., 2021). For tasks like *preference*, which relies on the preservation of agent shapes and colors, we therefore found it improved performance to include an auxiliary reconstruction loss. During evaluation, only the main BCE loss was used.

As in Gandhi et al., the videos' frame rate was downsampled by a factor of 5. We used a maximum sequence length of 90. Frame rates of longer sequences were interpolated to fit the maximum length. Of the BIB background episodes, we used 80% for training, 15% for testing, and 5% for validation. Models were trained using the Adamax optimizer for a total of 6 epochs, after which point we saw no further improvement on background training tasks. The batch size was set to 6 because of the VT's high memory requirements, resulting in a total number of 7.373 training updates. We tested the models on the validation set in five evenly spaced intervals per epoch and saved the model with the lowest validation loss to avoid overfitting. The total number of trainable parameters in the VT is 772.162. For comparison, the two publicly available baseline BC methods by Gandhi et al. contain 925.666 and 986.306 trainable parameters, respectively. On a 16-Core AMD EPYC 7282 server with six GeForce RTX 2080 GPUs, training time was around 3 hours per epoch. Our code is available at `https://github.com/zero-k1/BIB-VT`.

Our model shares some commonalities with the BIB baseline DL models, but also differs in several aspects. Both the VT and baseline models use CNNs to encode frames and average embeddings across familiarisation trials to obtain context vectors. However, we use attention mechanisms to obtain these embeddings, whereas the baselines used RNNs or multilayered perceptrons (MLPs). In contrast with the BC baselines, we also do not pre-train our CNN encoder separately, and we do not add the agent's actions as inputs – only the video frames. Finally, we predict both the next frame and the agent's position, while Gandhi et al.'s video modeling approach predicted only the next frame, and their BC approach predicted only the agent's next action.

Table 2: Mean squared error (MSE) of the frame prediction and weighted binary cross-entropy loss (BCE) of the agent prediction on the test split of the BIB background training tasks, averaged over the five trained VT models.

Training task	**MSE**	**BCE**
Single object	7.05×10^{-4}	1.58×10^{-2}
Preference	7.07×10^{-4}	1.38×10^{-2}
Multi-agent	5.94×10^{-4}	1.32×10^{-2}
Instrumental actions	1.42×10^{-3}	1.33×10^{-2}

Results

In total, we trained five models on the BIB background training tasks, each with a different random weight initialization. We report the models' average performance on the test set of the background training tasks in Table 2 and the performance on BIB evaluation tasks in Table 3. The baseline DL models previously tested on BIB used the prediction error of the frame with the highest loss as their metric of "surprise", as this provided better results compared to the mean error over entire trials (Gandhi et al., 2021). In our case, the mean error yielded a higher performance on most tasks, which is why we here report both metrics. However, binary VoE accuracies include no information about the magnitude of the difference in surprisal scores between expected and unexpected trials. We therefore also show z-scored means of both the models' average prediction error and infants' looking times, as reported by Stojnić et al., in Figure 2.

Goal-directed

Preference In contrast to the DL-based baselines, the VT seems, at least to some degree, to associate agents with certain goal preferences in the *preference* task (see Figure 2). To investigate which parts of the familiarization trials the model relied on most, we performed a form of occlusion analysis. We used only one trial as the familiarization input (performance was almost identical when using one vs. the full eight trials), and dropped each of the patches fed into the first transformer block in turn. For each patch, we recorded the z-scored difference in prediction error between the expected and unexpected outcome. An example result is shown in Figure 3. Models tended to either rely on the agent's last or first step. Averaged over all models and episodes, the patch with the largest impact on the final prediction was part of the last two frames of the familiarization trial in 52.6% of cases.

Table 4: VoE Accuracy on BIB evaluation tasks. VoE Accuracy denotes whether model error is higher on expected trials than unexpected trials. VT (Mean) uses the avg. error over all test trial frames as the "surprise" metric, whereas VT (Max) uses the error for the frame with the highest loss. For the VT models, we report the average accuracy and standard deviation over five models trained on the same data, but with different random initialization. Baselines and Video Transformers are data-driven computer vision models, whereas hierarchically Bayesian Theory of Mind (HBToM) uses a principled Bayesian solution that requires access to symbolic states. Chance level accuracy is 50%.

		Baselines			Video Transformer (ours)	
Task	**HBToM**	**BC-MLP**	**BC-RNN**	**Video-RNN**	**VT (Mean)**	**VT (Max)**
Goal-directed						
Preference	99.7	26.3	48.3	47.6	82.1 ± 0.0	80.8 ± 0.0
Multi-agent	99.2	48.7	48.2	50.3	49.1 ± 0.0	49.2 ± 0.0
Inaccessible goal	99.7	76.9	81.6	74.0	89.8 ± 0.0	85.5 ± 0.0
Efficiency						
Efficient agent	95.8	96.0	95.3	99.5	98.3 ± 0.0	98.4 ± 0.0
Inefficient agent	96.6	73.8	56.5	50.1	29.5 ± 0.1	34.1 ± 0.1
Instrumental actions						
Instrumental action	98.5	67.0	77.9	79.9	92.6 ± 0.0	84.7 ± 0.0

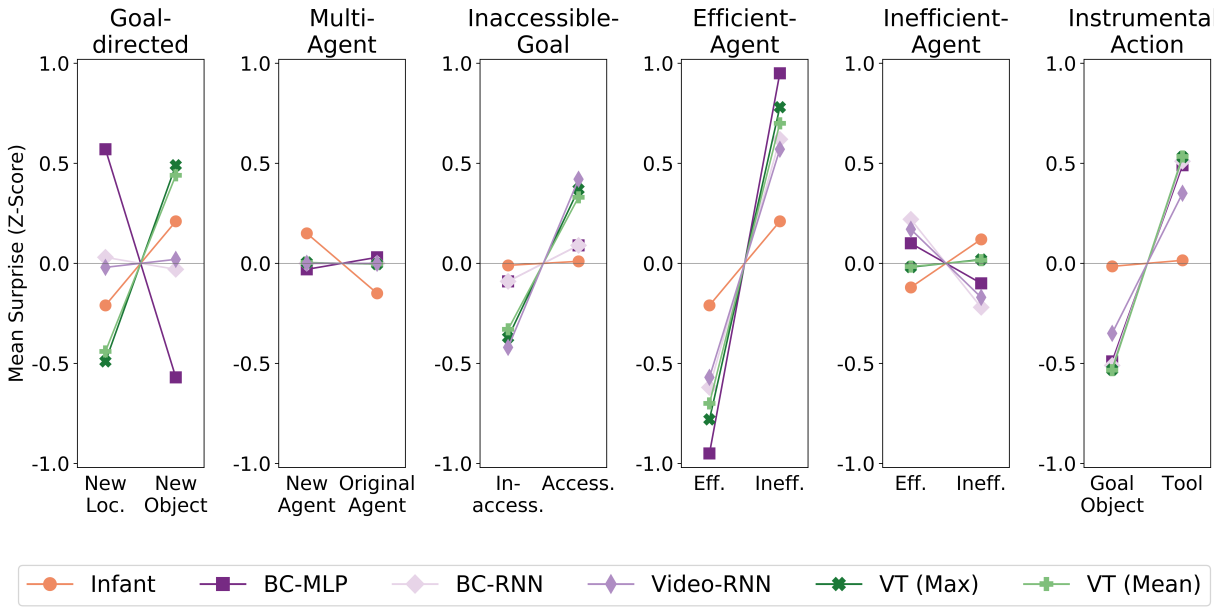


Figure 2: Z-scored means of the models' average surprisal scores and infants' looking times to the expected and unexpected outcomes in the BIB test episodes.

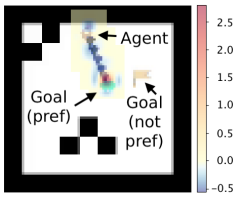


Figure 3: Z-scored impact of omitting a patch from the *preference* familiarization trial.

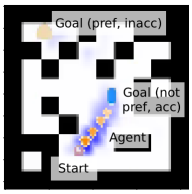


Figure 4: *Inaccessible goal* task. Predicted agent positions marked blue.

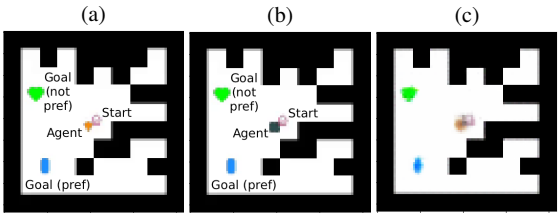


Figure 5: 5a: Unexpected *multi-agent* outcome (familiar agent). 5b: Expected outcome (new agent). 5c: Prediction for expected outcome.

Multi-Agent Similar to the other DL models, the VT does not acquire the desired knowledge from the *multi-agent* background training tasks, which feature both agents moving towards the same single goal across trials. Note that the infants tested on BIB were in fact more surprised at the supposedly "expected" trials (see Figure 2). Stojnić et al. hypothesize that this may be because of the increased novelty of the new agent. A closer look at the frame predictions produced by the VTs hints at some confusion regarding the agents' identity: In some cases, the model reconstructs the familiar agent in the unexpected trial, rather than the new agent present in the input (see Figure 5 for an example). Averaged over all models and episodes, this was the case 27.9% of the time.

Inaccessible In the *inaccessible-goal* task, the VT model achieves a higher accuracy than previous DL models. It exhibits a stronger deviation in surprise than the infants, who were indifferent on this task (see Figure 2). Stojnić et al. posit that infants may have considered the new barrier in the expected outcome as indicative of a new environment and not carried over any goal preference expectations from the familiarization trials. Although the VT has a lower prediction loss on the expected outcome in most cases, it is more "split" than in the single-object case (see Figure 4 for an example prediction). Averaged over all models and episodes, the entropy of the models' prediction on the test trial's last frame was 1.10 for the expected, and 1.47 for the unexpected outcome. For comparison, the average entropy for the last frame of the *single-object* background training task was only 0.58.

Efficiency

Similar to previous models, the VT's VoE accuracy on the *Efficient agent* tasks are nearly perfect – the model strongly expects agents to move towards their goal efficiently. This is

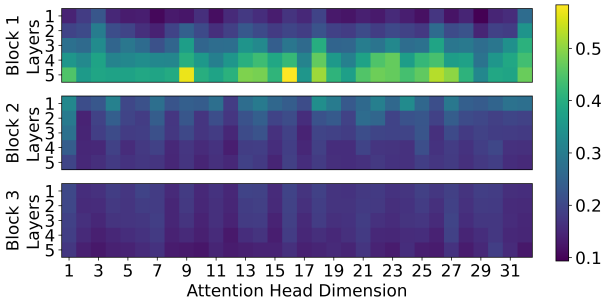
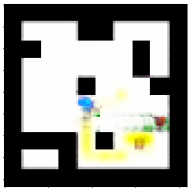


Figure 6: Avg. difference in the VT layers' activations when processing the episodes' unexpected vs. expected familiarization trials, featuring an efficient or an inefficient agent, respectively.

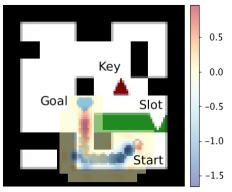
in accordance with infant's intuitions (see Figure 2). On the *inefficient-agent* task, the VT tends to be more "surprised" at the previously inefficient model moving inefficiently than at the previously efficient agent doing so. Although not necessarily a desired outcome, this is actually more in line with the intuitions of the infants tested on BIB, who attributed rational action both to previously efficient and inefficient agents in a new environment (see Figure 2). When we compare the impact of the familiarization trials featuring the efficient vs. inefficient agent on the VT model (see Figure 6), we see that a similar mechanism is at work: The lowest levels, which attend over past familiarization trials, show differences in activation. However, these differences all but disappear throughout the higher layers. This leads to the inefficient agent being treated in the same way as the efficient one, which explains the mean surprise score being almost the same in both cases. The slightly larger error for the inefficient agent most likely stems from the fact that inefficient agents are not seen during training, leading to higher prediction uncertainty.

Instrumental Actions

Compared with the other DL models, the VT performs similar on episodes with no barrier, and better on episodes with inconsequential or blocking barriers. Again, infants were indifferent on this task (see Figure 2). Stojnić et al. note that they may have failed to recognize the instrumental actions because they were causally opaque. Although the VT is correct in most cases in terms of VoE accuracy, it, too, seems to not have quite understood the causal mechanism. A look at the frame predictions shows that the model usually expects the disappearance of the key on the first step, even though the agent has not collected and inserted it. Averaged over all models and episodes, the VT at least partly predicts the key's position as the agent's first step in 47% of cases, even though the key is mostly far away from the agent. This is most likely because the key is always right next to the agent in the background *instrumental-action* tasks, and thus constitutes its first step. The VT also often predicts the disappearance of the green barrier towards the end of the episode, even



(a) Predicted last frame and agent trajectory (yellow).



(b) Z-scored impact of each test trial patch on final MSE error.

Figure 7: Prediction on an *instrumental-action* task.

though the key was not inserted. This is most likely because the green barrier has always disappeared by the time the agent reaches the goal in the background tasks. Occlusion analyses support this hypothesis: The parts of the test trial that most contribute to the z-scored MSE prediction error on expected *instrumental-action* outcomes were usually the agent's first and last steps (see Figure 7 for an example).

Decoding experiment

Inspired by probing analyses of pre-trained language models (Clark, Khandelwal, Levy, & Manning, 2019), we trained linear regression models to predict the current position of the agent, goal, and sub-goals (keys and locks), based on the activations in each layer of each VT block. Each linear model had an input dimension of 256 (8 attention heads per layer, each with dimension 32) and output dimension 4 (one for each prediction target). The models were trained with the Adam optimizer (Kingma & Ba, 2015) set to default parameters, using the same epoch number and batch size as the main experiments described in the Methods section. We used the background training set for optimization and display the results for the background validation set in Figure 8.

In general, we see errors decrease in the deeper layers of the attention blocks, indicating more focused attention heads. The heads in the first block, which attends over familiarization trials, do not display a large degree of specialization regarding the analysed categories. However, at least in the higher layers, the agent, key, and lock categories have a comparatively lower decoding error than the goal category. Note that the agent's position often corresponds with the key and lock position for long stretches of instrumental action trials, as the agent waits for the green barrier to disappear after having inserted the key into the lock. The second block, which self-attends over the test trial's first frame, has the lowest decoding error across categories and a particularly low error for the agent's current position. The third block shows a clear separation between categories, with locks and keys displaying a much lower decoding error than goals and agents. This is presumably because the third attention block autoregressively predicts the agent's next step, which, as mentioned, often coincides with the key and lock position while the agent is waiting in place for the barrier to disappear.

In summary, the VT seems to have learned to implicitly keep track of relevant semantic categories, such as agents, goals, and subgoals, which are usually modelled as explicit variables in Bayesian approaches.

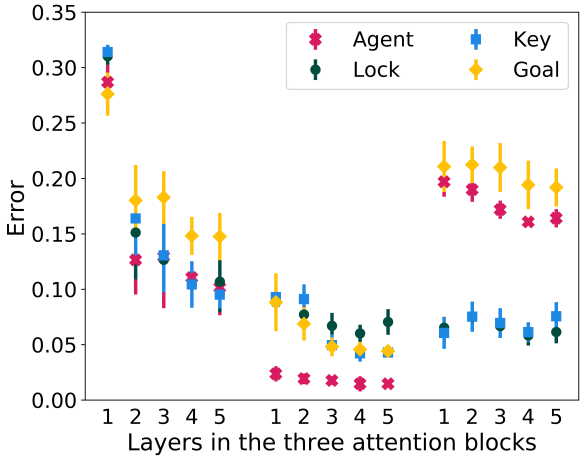


Figure 8: Weighted BCE loss of linear probes trained on decoding the current position of goals, agents, and sub-goals from attention head activations in each layer. Error bars indicate standard deviation across the five trained VT models.

Discussion and Conclusion

In conclusion, the VT model tested in this paper outperforms previous DL-based baselines on the *preference*, *inaccessible-goal*, and *instrumental-actions* BIB tasks in terms of VoE accuracy. Its surprisal scores are also more in line with infants' expectations than previous DL models, in that it tends to represent agents' actions as directed towards goals, rather than locations, and defaults to expecting rational actions. This suggests that the transformer's attention mechanism can be helpful in acquiring intuitions about agents' goals, preferences, and actions, purely from predicting the next step in videos. However, a qualitative analysis of the VT's errors also demonstrated the pitfalls of this approach: Models may exploit the particularities of a training dataset in an unintended way (Gardner et al., 2020; Geirhos et al., 2020), e.g. by associating the disappearance of the green barrier in the *instrumental-actions* task with the agent's first and last step rather than the key mechanism. This may be mitigated with a more realistic data setting, where models can gain experience with diverse agents and disambiguate causes and effects of instrumental mechanisms interactively, in a manner closer to human infants. The findings also support the benefit of investigating hybrid architectures that incorporate methods which explicitly model human intuitions, such as HBToM, to take advantage of both the flexibility of DL-based approaches and the data efficiency and robustness of principled Bayesian models.

References

Chaplot, D. S., Pathak, D., & Malik, J. (2021). Differentiable spatial planning using transformers. In *International conference on machine learning* (pp. 1484–1495).

Chen, W., Wang, F., & Sun, H. (2021). S2tnet: Spatio-temporal transformer networks for trajectory prediction in autonomous driving. In *Asian conference on machine learning* (pp. 454–469).

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019, August). What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 acl workshop blackboxnlp: Analyzing and interpreting neural networks for nlp* (pp. 276–286). Florence, Italy: Association for Computational Linguistics. Retrieved from https://aclanthology.org/W19-4828 doi: 10.18653/v1/W19-4828

Du, H., Yu, X., & Zheng, L. (2021). Vtnet: Visual transformer network for object goal navigation. *arXiv preprint arXiv:2105.09447*.

Fukushima, R., Ota, K., Kanezaki, A., Sasaki, Y., & Yoshiyasu, Y. (2022). Object memory transformer for object goal navigation. *arXiv preprint arXiv:2203.14708*.

Gandhi, K., Stojnic, G., Lake, B. M., & Dillon, M. R. (2021). Baby Intuitions Benchmark (BIB): Discerning the goals, preferences, and actions of others. *Advances in Neural Information Processing Systems*, *34*, 9963–9976.

Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., ... others (2020). Evaluating models' local decision boundaries via contrast sets. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 1307–1323).

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673.

Giuliari, F., Hasan, I., Cristani, M., & Galasso, F. (2021). Transformer networks for trajectory forecasting. In *2020 25th international conference on pattern recognition (icpr)* (pp. 10335–10342).

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, *57*(2), 243–259.

Jadon, S. (2020). A survey of loss functions for semantic segmentation. In *2020 ieee conference on computational intelligence in bioinformatics and computational biology (cibcb)* (pp. 1–7).

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... others (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *3rd international conference on learning representations, ICLR 2015, san diego, ca, usa, may 7-9, 2015, conference track proceedings*. Retrieved from http://arxiv.org/abs/1412.6980

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and brain sciences*, *40*.

Li, C., & Liu, Y. (2022). Rethinking query-key pairwise interactions in vision transformers. *arXiv preprint arXiv:2207.00188*.

Li, L. L., Yang, B., Liang, M., Zeng, W., Ren, M., Segal, S., & Urtasun, R. (2020). End-to-end contextual perception and prediction with interaction transformer. In *2020 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 5784–5791).

Soomro, K., Zamir, A. R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

Stojnić, G., Gandhi, K., Yasuda, S., Lake, B. M., & Dillon, M. R. (2023). Commonsense psychology in human infants and machines. *Cognition*, *235*, 105406.

Sui, Z., Zhou, Y., Zhao, X., Chen, A., & Ni, Y. (2021). Joint intention and trajectory prediction based on transformer. In *2021 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 7082–7088).

Yu, C., Ma, X., Ren, J., Zhao, H., & Yi, S. (2020). Spatio-temporal graph transformer networks for pedestrian trajectory prediction. In *European conference on computer vision* (pp. 507–523).

Yuan, Y., Weng, X., Ou, Y., & Kitani, K. M. (2021). Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the ieee/cvf international conference on computer vision* (pp. 9813–9823).

Zhi-Xuan, T., Gothoskar, N., Pollok, F., Gutfreund, D., Tenenbaum, J. B., & Mansinghka, V. K. (2022). Solving the baby intuitions benchmark with a hierarchically bayesian theory of mind. *arXiv preprint arXiv:2208.02914*.