Hypothesis - Memorization in the ACT Model

Information Flow in ACT Model

Input: Images from two cameras: $I_{\text{top}}, I_{\text{lat}} \in \mathbb{R}^{480\times640\times3}$ and joint angles $q \in \mathbb{R}^6$. Images are processed by ResNet:

$$V_{\text{top}} = \text{ResNet}(I_{\text{top}}), \quad V_{\text{lat}} = \text{ResNet}(I_{\text{lat}}), \quad V_{\text{top}}, V_{\text{lat}} \in \mathbb{R}^{300\times512}.$$

Tokenization: Flattened camera features (600 "pixel tokens" - 300 for each camera), arm state token[1] ($q \in \mathbb{R}^{512}$), and condition token [2]($c \in \mathbb{R}^{512}$) form 602 tokens of dimension 512.

Transformer Encoding: The 602 tokens are processed through 3 Transformer encoder layers, enabling global information flow:

$$602 \times 512 \to 602 \times 512.$$

Action Vector Update: A zero-initialized action vector ($a \in \mathbb{R}^{512}$) is updated using cross-attention with encoder output and an MLP:

$a \in \mathbb{R}^{512} \to$ Cross-attention with the 602 tokens from the encoder $\to$ MLP $\to \mathbb{R}^6$ (robot movement).

Hypothesis: Memorization in ACT

Hypothesis: The MLP layer in the decoder part of the ACT (Zhao et al. (2023)) model memorize exact actions instead of generalizing. During inference, **the action vector after cross-attention [3] contains complete environmental information (e.g., cube location, gripper state) even of unseen training examples**. However, the MLP acts as a "lookup table" of actions rather than applying general rules for novel scenarios. (This is based on observationa from Geva et al. (2020); Meng et al. (2022) that the mlp layers in transformers is where memories are stored.)

[1]Why not create six tokens, one for each joint angle? This might help the model better understand the arm's configuration.

[2]We focuses here only on the model's inference process. During training, an additional encoder processes samples from the teleoperation setup, producing a latent variable z. This variable serves as a condition for the decoder, which takes the inputs described here. The model's objective is to reconstruct teleoperation actions — a non-trivial choice, as the CVAE framework was used instead of a more direct approach to predict the next steps.

[3]Observation: Attention maps suggest only a subset of tokens (5-20) receive high attention scores, varying across different input images. This indicates that during encoding, environmental information is represented by a small subset of the 602 tokens. Interestingly, the arm state token is not among those receiving high weights, implying that pixel tokens absorb and convey necessary arm state information to the action vector.

Validation Plan

1.**Probing Action Vector:** Train linear classifiers on the action vector to verify if it encodes all relevant environmental details.

2. **Generalization to Novel Scenes:** Test whether the action vector generalizes to new, unseen cube positions.

3. **Visualization:** Adapt techniques from Toker et al. (2024) ("Diffusion lens) to visualize hidden states of the action vector and analyze stored information.

Future Steps: Investigate whether the MLP memorizes steps in a specific environment representation or if it applies generalizable rules.

References

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2020. Transformer Feed-Forward Layers Are Key-Value Memories. *arXiv*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. *arXiv*.

Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. 2024. Diffusion Lens: Interpreting Text Encoders in Text-to-Image Pipelines. *arXiv*.

Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. 2023. Learning Fine-Grained Bimanual Manipulation with Low-Cost Hardware. *arXiv*.