Using Pre-trained ViT to improve Decision Transformers

A project proposal by Akash

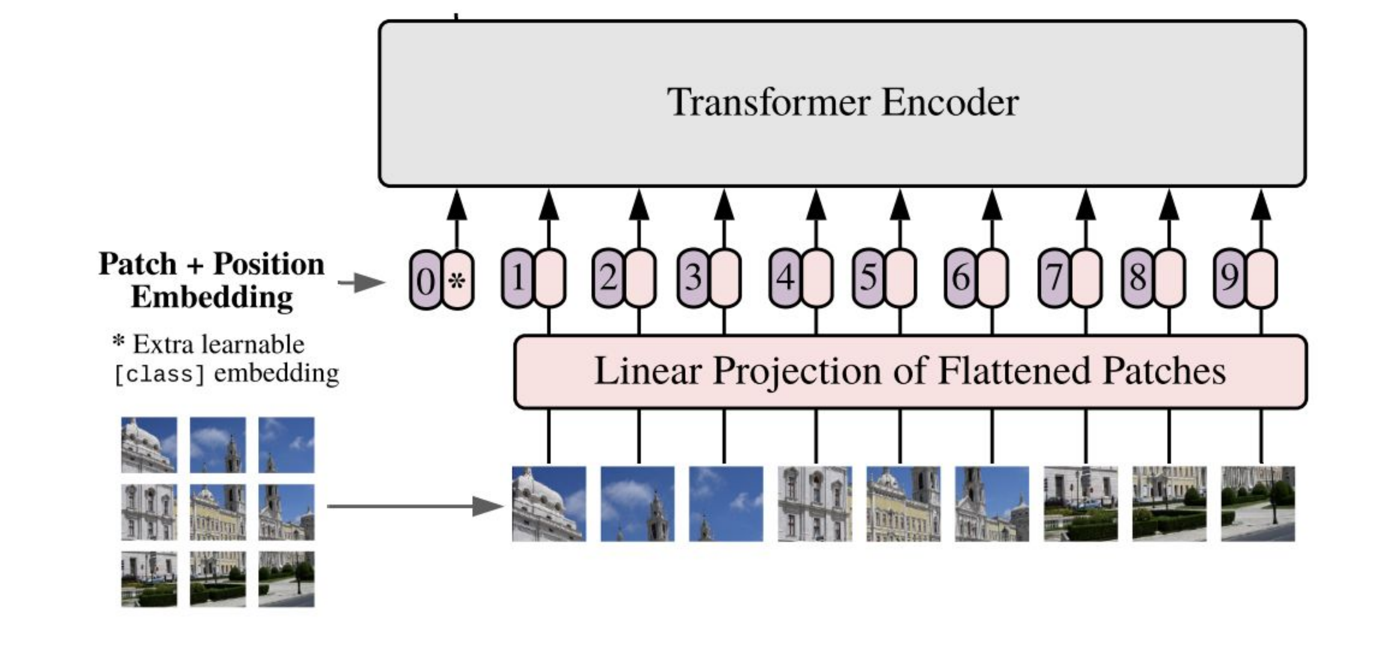
Existing System

- Decision Transformers is a novel architecture that aims to generate actions in an auto-regressive manner.
- Current DTs typically use CNNs to extract image features.
- Recent advancements like the StARformers architecture have introduced self-attention mechanism for improved performance, but it does not take advantage of pre-trained ViT models to improve the state representation.

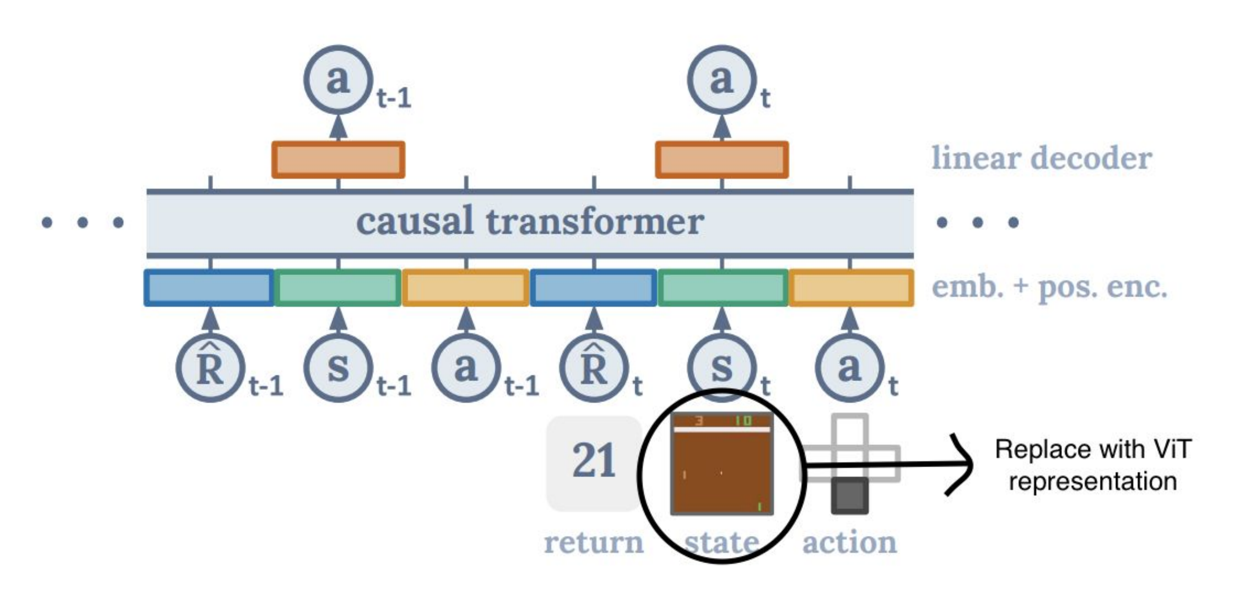
Proposed System

- The use of pre-trained Vision Transformer (ViT) models, such as the vit-base model introduced by Google Brain, has the potential to enhance the encoding of image/state representation in DTs.
- The hypothesis is that a pre-trained model, even one without an extensive domain-specific training, could enhance the policy learning process.

ViT-representation generation


Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

Where it fits in DT



References

- Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). Decision Transformer: Reinforcement Learning via Sequence Modeling. ArXiv. /abs/2106.01345
- J. Shang, X. Li, K. Kahatapitiya, Y. -C. Lee and M. S. Ryoo, "StARformer: Transformer with State-Action-Reward Representations for Robot Learning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, doi: 10.1109/TPAMI.2022.3204708.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv. /abs/2010.11929

Questions?