DeepMind

Solving Reasoning Tasks with a Slot Transformer

Ryan Faulkner, Daniel Zoran

Introduction & Model Architecture

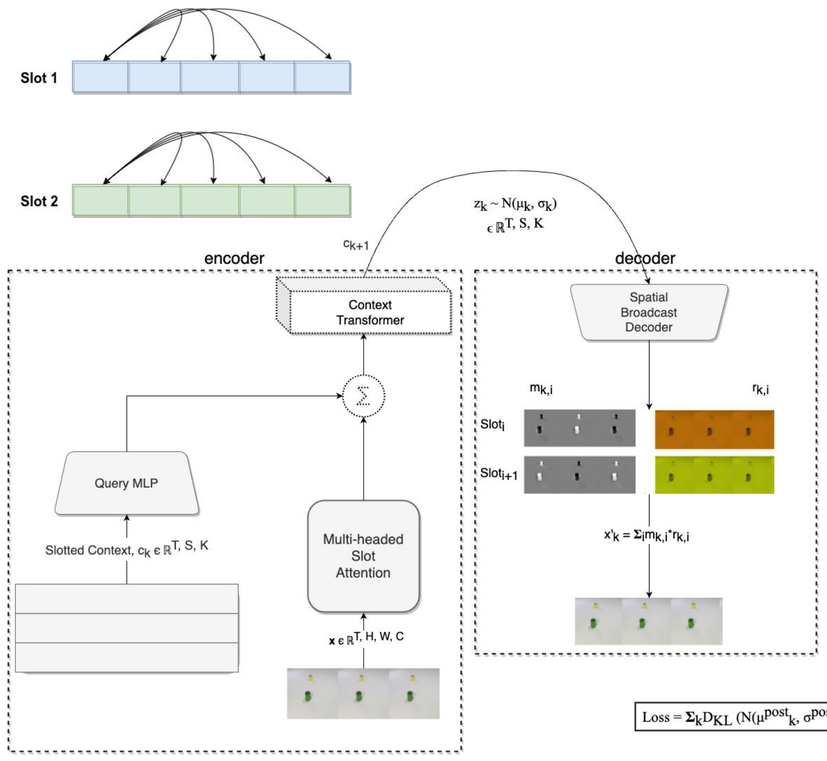
Our goal is to train a model that can be induced to learn to understand visual scenes that play out over time. Also to generalise across these kinds of domains which may contain complex visual interactions and temporal dynamics.

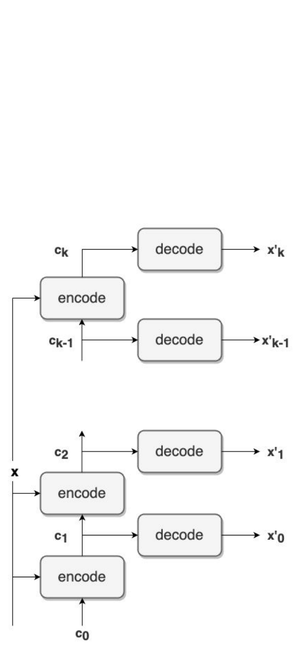
1. **Slotted Representation** for multi-object scenes.

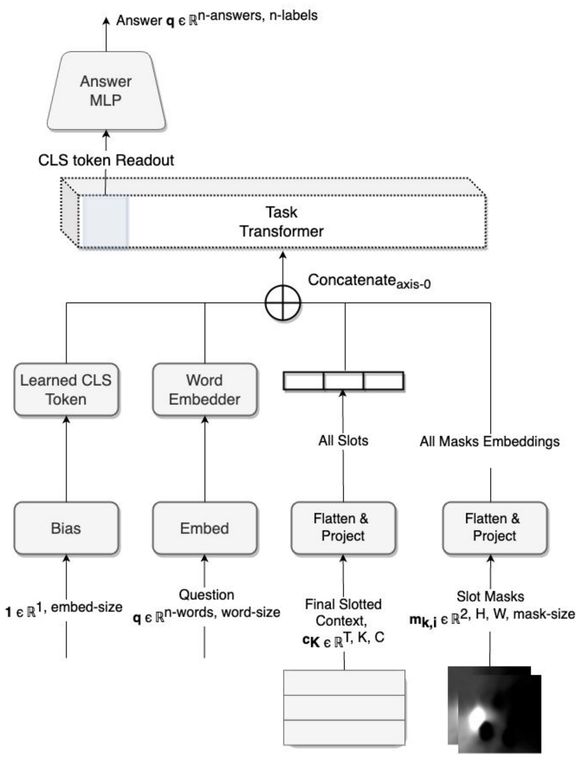
2. The **Context Transformer** allows relationships among objects to be learned over many timesteps.

3. An **Iterative Model** to enable reasoning across steps.

4. **Generative Model** for improved generalisation over visual reasoning domains.

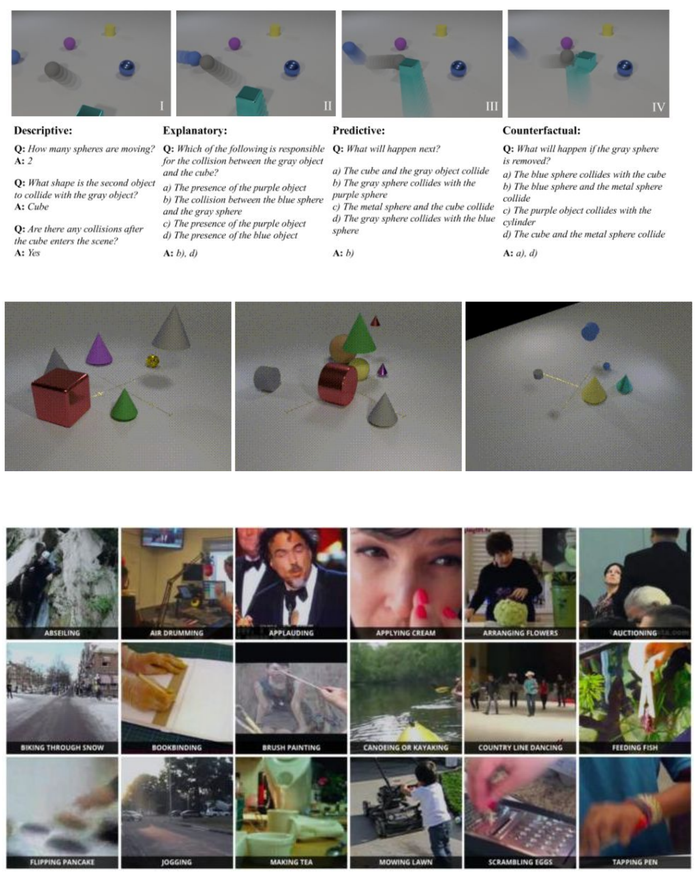


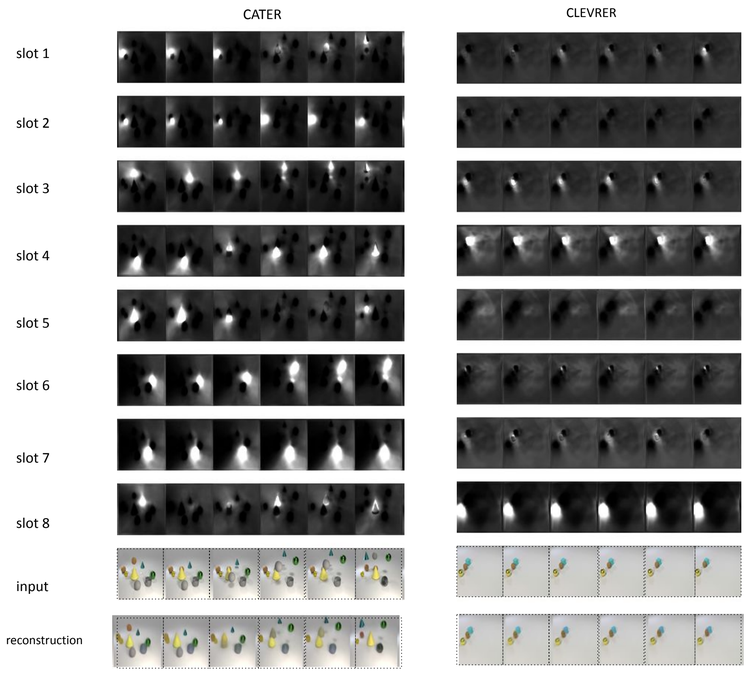




Experiments & Results

- CLEVRER: QA & Spatial Reasoning
- CATER: QA & Spatial Reasoning
- Kinetics 600: Video Classification





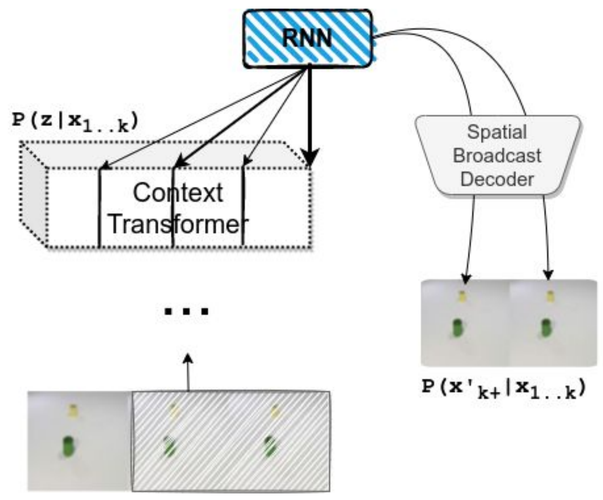


Table 1: CLEVRER per Question Accuracies (Chen et al. [2021])

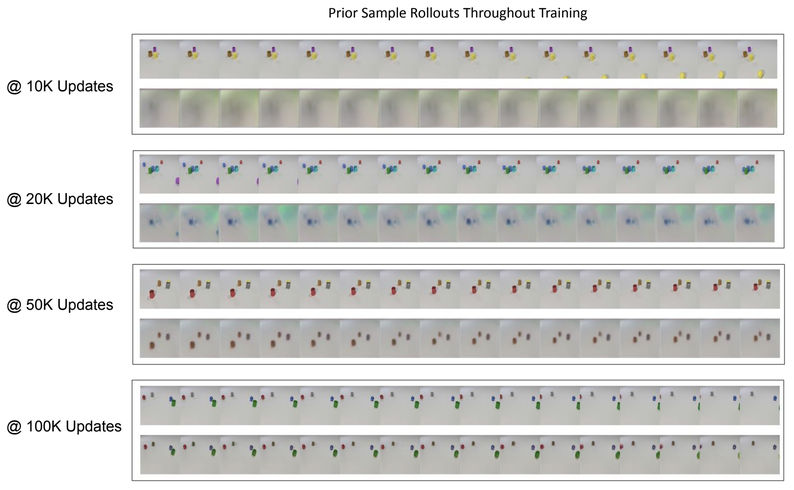
Methods	Extra Labels Attr.	Extra Labels Prog.	Descriptive	Predictive	Explanatory	Counterfactual
CNN+MLP			48.4	18.3	13.2	9.0
CNN+LSTM			51.8	17.5	31.6	14.7
Memory	No	No	54.7	13.9	33.1	7.0
HCRN			55.7	21.0	21.0	11.5
MAC (V)			85.6	12.5	16.5	13.7
Slot Transformer (Ours)			**87.4**	**48.3**	**65.3**	**45.2**
TVQA+	Yes	No	72.0	**23.7**	**48.9**	4.1
MAC (V+)			**86.4**	22.3	42.9	**25.1**
IEP (V)			52.8	14.5	9.7	3.8
TbD-net (V)	No	Yes	79.5	3.8	6.5	4.4
DCL			**90.7**	**82.8**	**82.0**	**46.5**
NS-DR			88.1	79.6	68.7	42.4
NS-DR (NE)	Yes	Yes	85.8	74.3	54.1	42.0
DCL-Oracle			**91.4**	**82.0**	**82.1**	**46.9**

Methods	Top-1	Top-5
TimeSformer (+pre-train)	**82.2**	**95.6**
I3D (+pre-train)	71.7	90.4
SimCLR (+pre-train)	51.6	-
ImageNet (+pre-train)	54.7	-
CVRL (+pre-train)	72.9	-
Slot Transformer (Ours)	68.2	87.6

Table 2: CATER Results on Task 3: "Localization"

Methods	Static Camera Top-1	Static Camera Top-5	Static Camera L1	Moving Camera Top-1	Moving Camera Top-5	Moving Camera L1
Random	2.8	13.8	3.9	-	-	-
R3D LSTM	60.2	81.8	1.2	28.6	63.3	1.7
R3D + NL LSTM	46.2	69.9	1.5	38.6	70.2	1.5
Aloe Ding et al. [2020]	**74.0**	**94.0**	**0.44**	**59.7**	**90.1**	**0.69**
Slot Transformer (Ours)	62.9	84.7	0.86	35.8	60.5	2.4

Prior Sample Rollouts Throughout Training



Discussion & Conclusions

- Scene and sequence representation.
  - Masks + slot ablation show reasonable segmentation.
- More Iteration has a positive effect on the model's ability to generalize
- Auxiliary Losses help with both CLEVRER & CATER test performance