Token Turing Machines

Ryoo, Gopalakrishnan, Kahatapitiya, Xiao, Rao, Stone, Lu, Ibarz & Arnab

Google

THU-AM-247

A sequential, autoregressive model

A sequential, autoregressive model

with **external memory**

A sequential, autoregressive model

with **external memory**

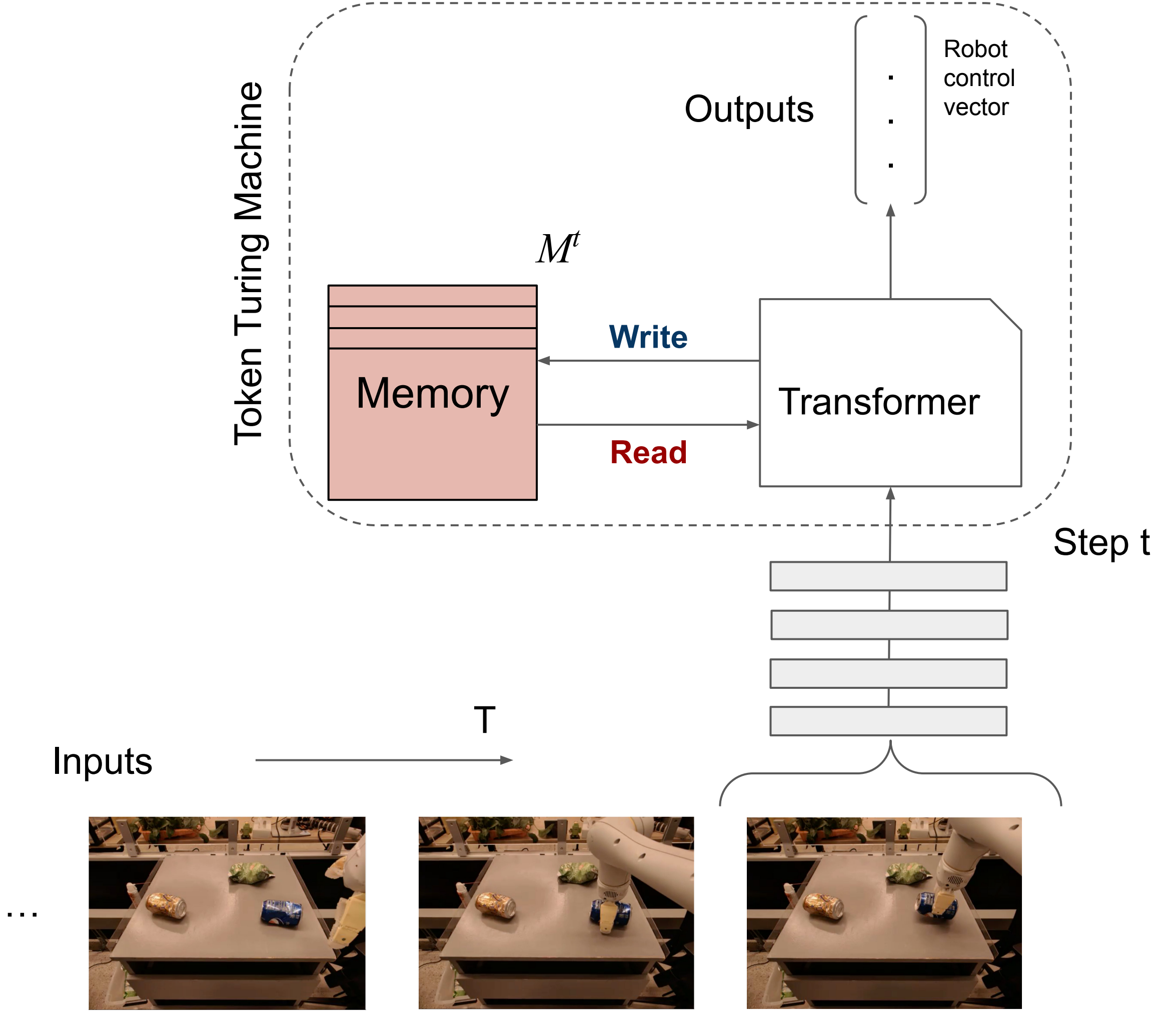
designed for streaming visual data

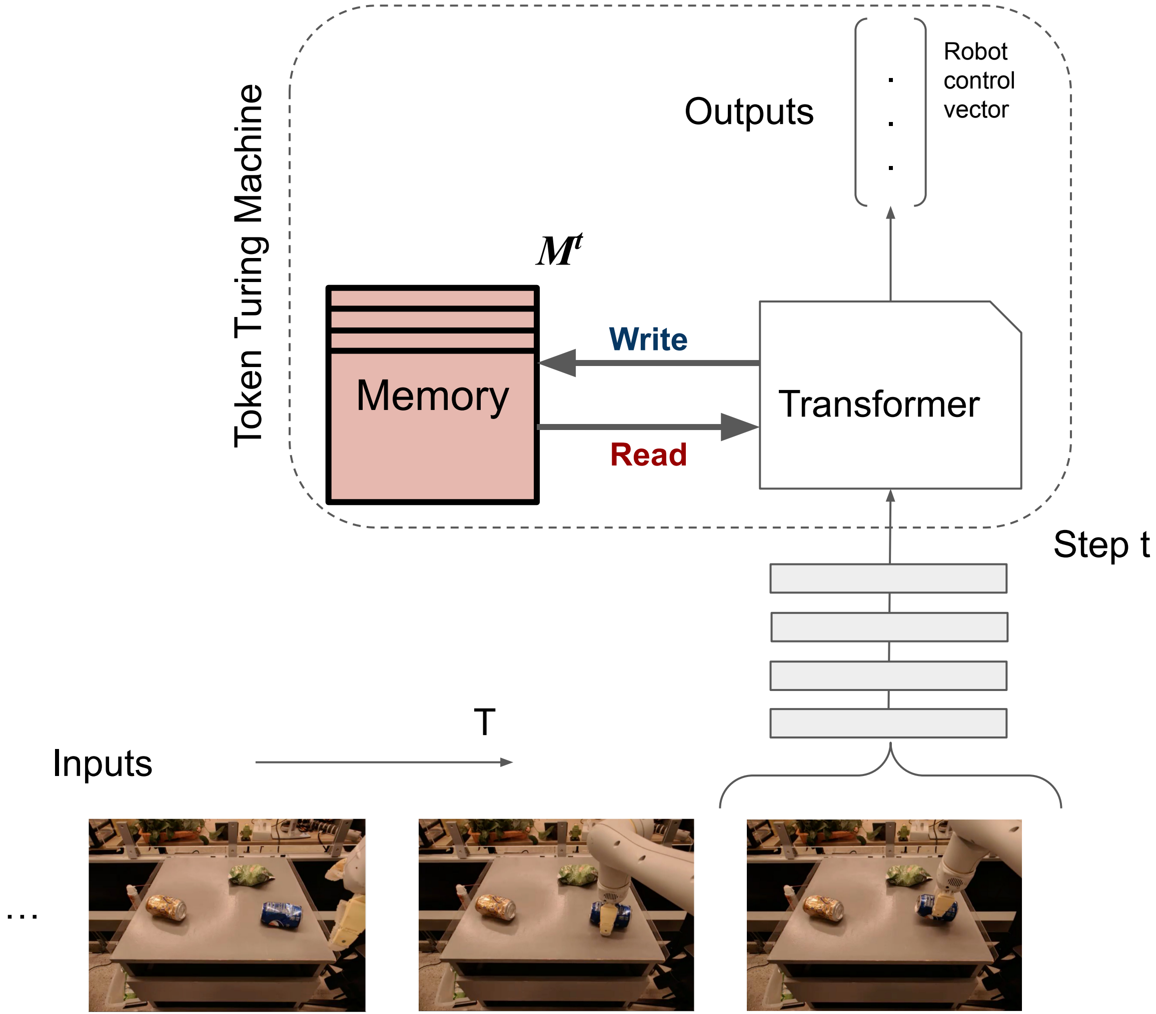
Video representation learning

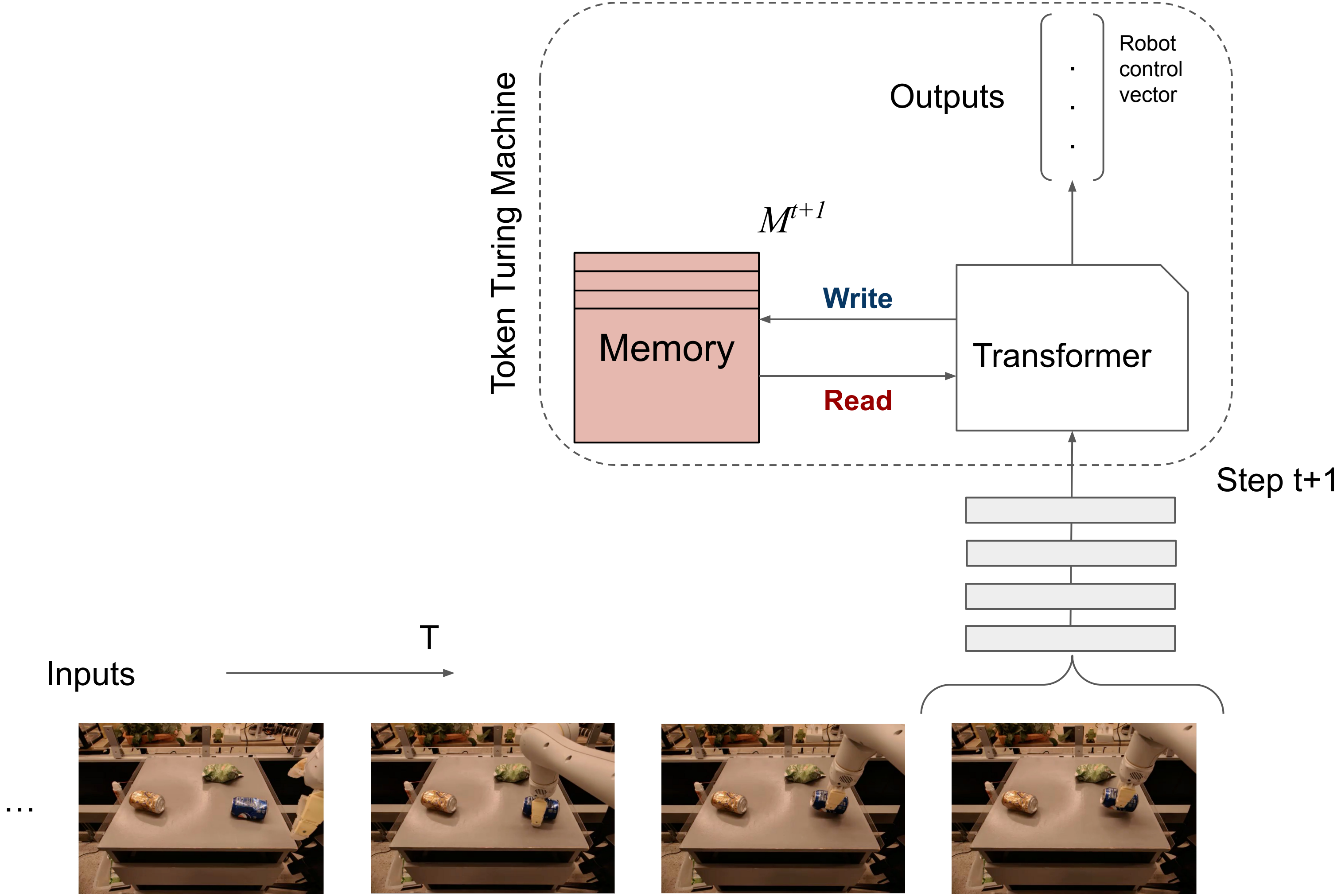
[Charades dataset, ECCV 2016]

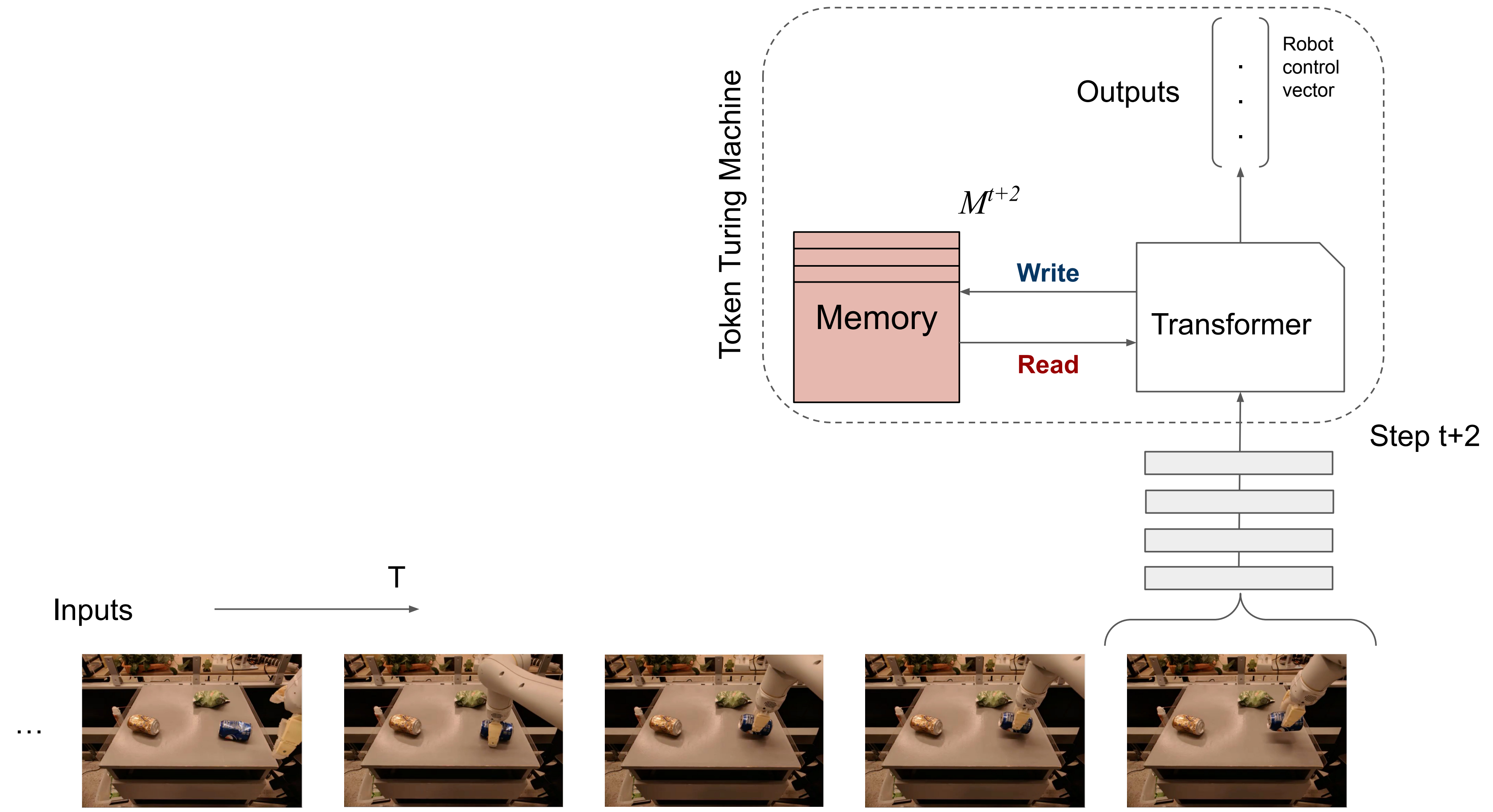
Video representation learning

[Robotics Transformer-1, RSS 2023]


Token Turing Machine
Outputs
.
.
.
Robot
control
vector
M^t
Memory
Write
Read
Transformer
Step t
T
Inputs
...

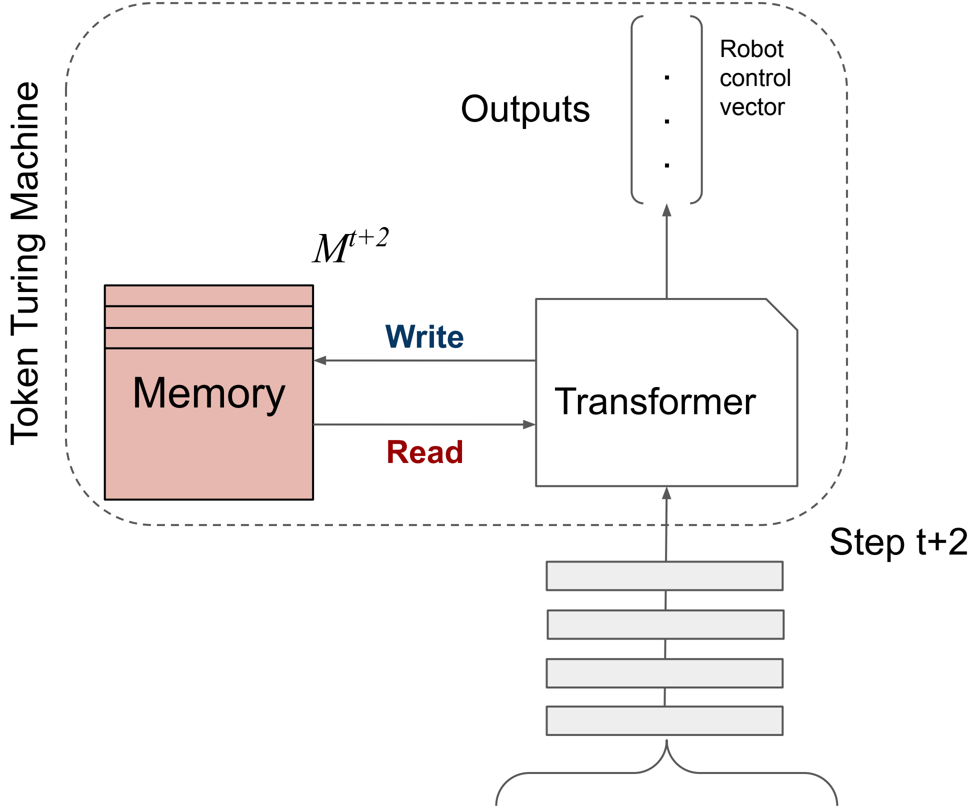

Token Turing Machine
Outputs
Robot
control
vector
M^t
Memory
Write
Read
Transformer
Step t
T
Inputs
...

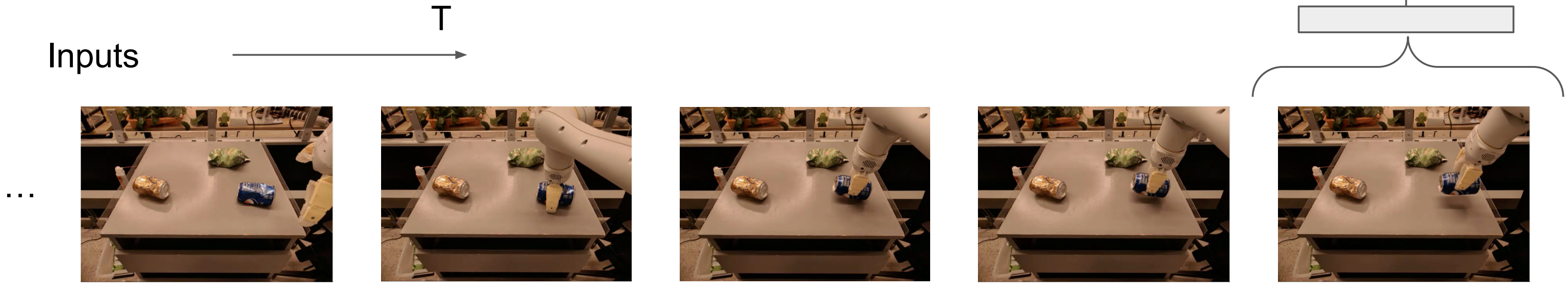

Token Turing Machine
Outputs
Robot
control
vector
M^{t+1}
Memory
Write
Read
Transformer
Step t+1
T
Inputs
...



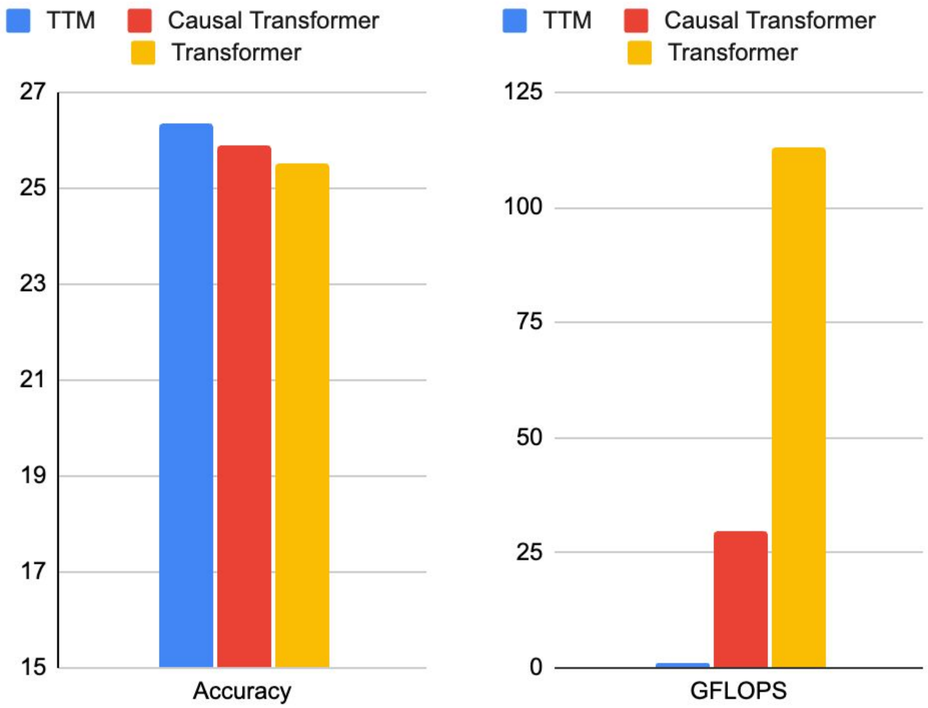
Token Turing Machine
Outputs
Robot
control
vector
M^{t+2}
Memory
Write
Read
Transformer
Step t+2
T
Inputs
...

Time complexity: O(1), instead of O(T) or O(T^2)





Better performance, while requiring significantly less compute

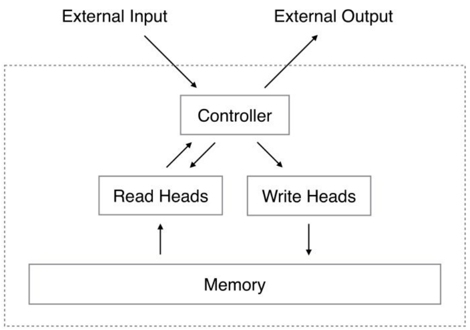


Charades (temporal) Localization task

Token Turing Machine?

It is a modernization of Neural Turing Machines (NTM) [Graves et al., 2014]

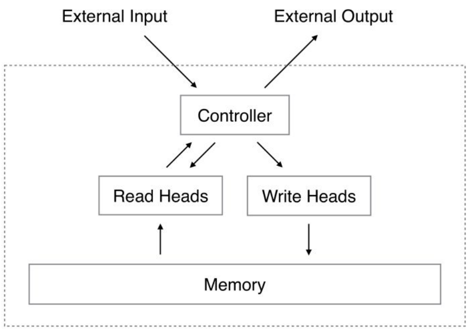
- Transformer as a controller + memory tokens



Token Turing Machine?

It is a modernization of Neural Turing Machines (NTM) [Graves et al., 2014]

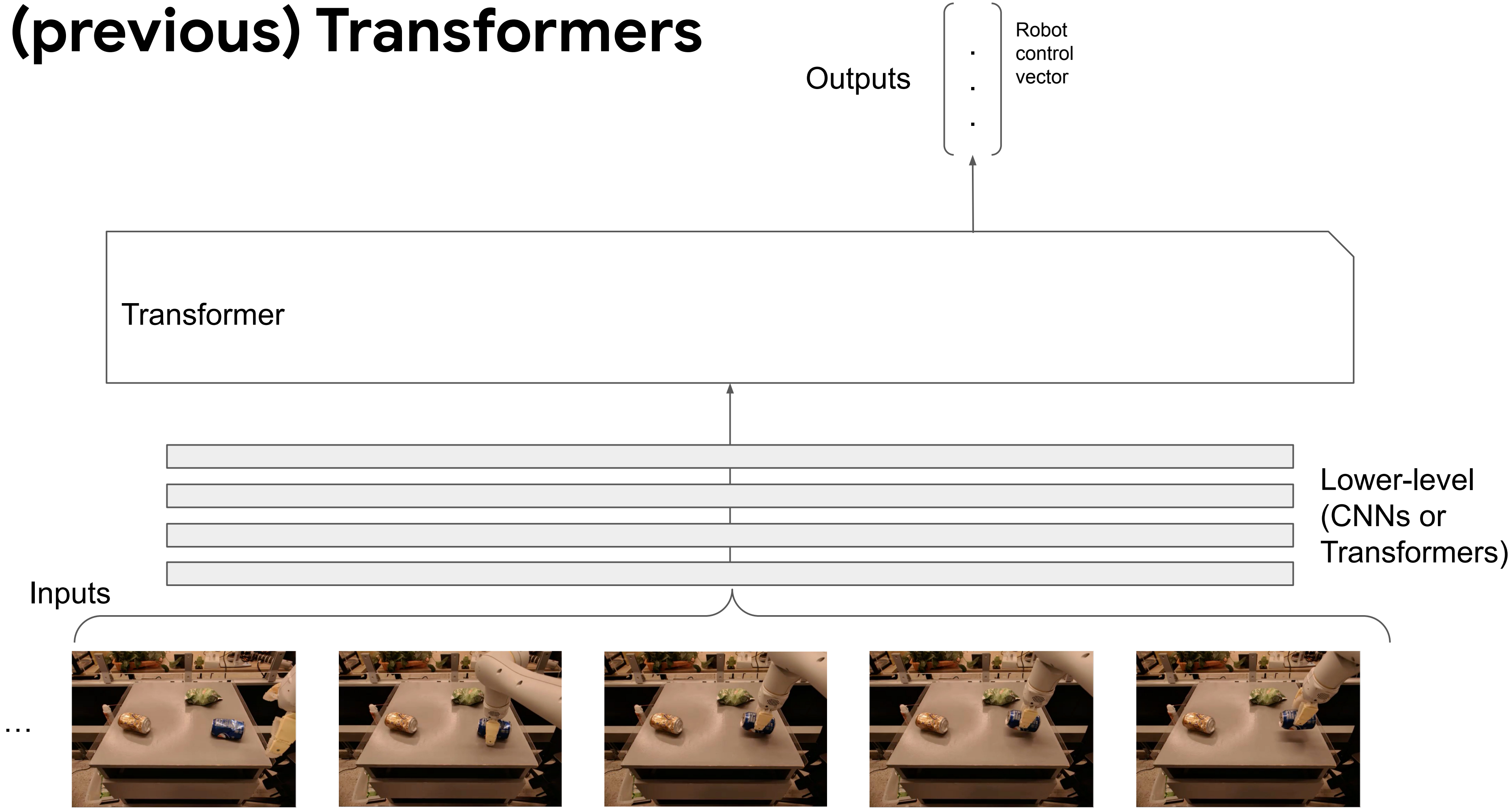
- Transformer as a controller + memory tokens

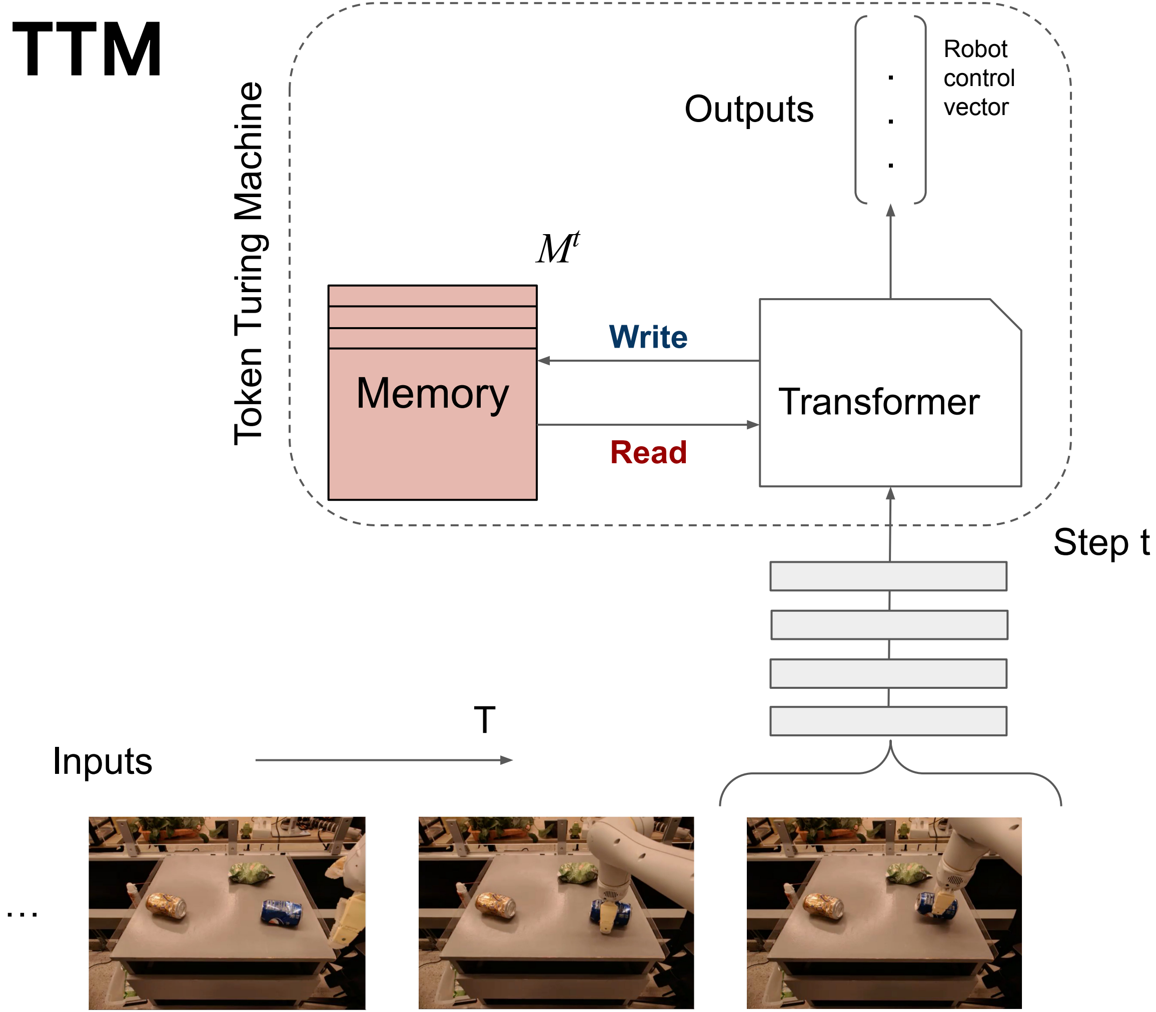


A recurrent neural network, where all its components are implemented with token-based operations.

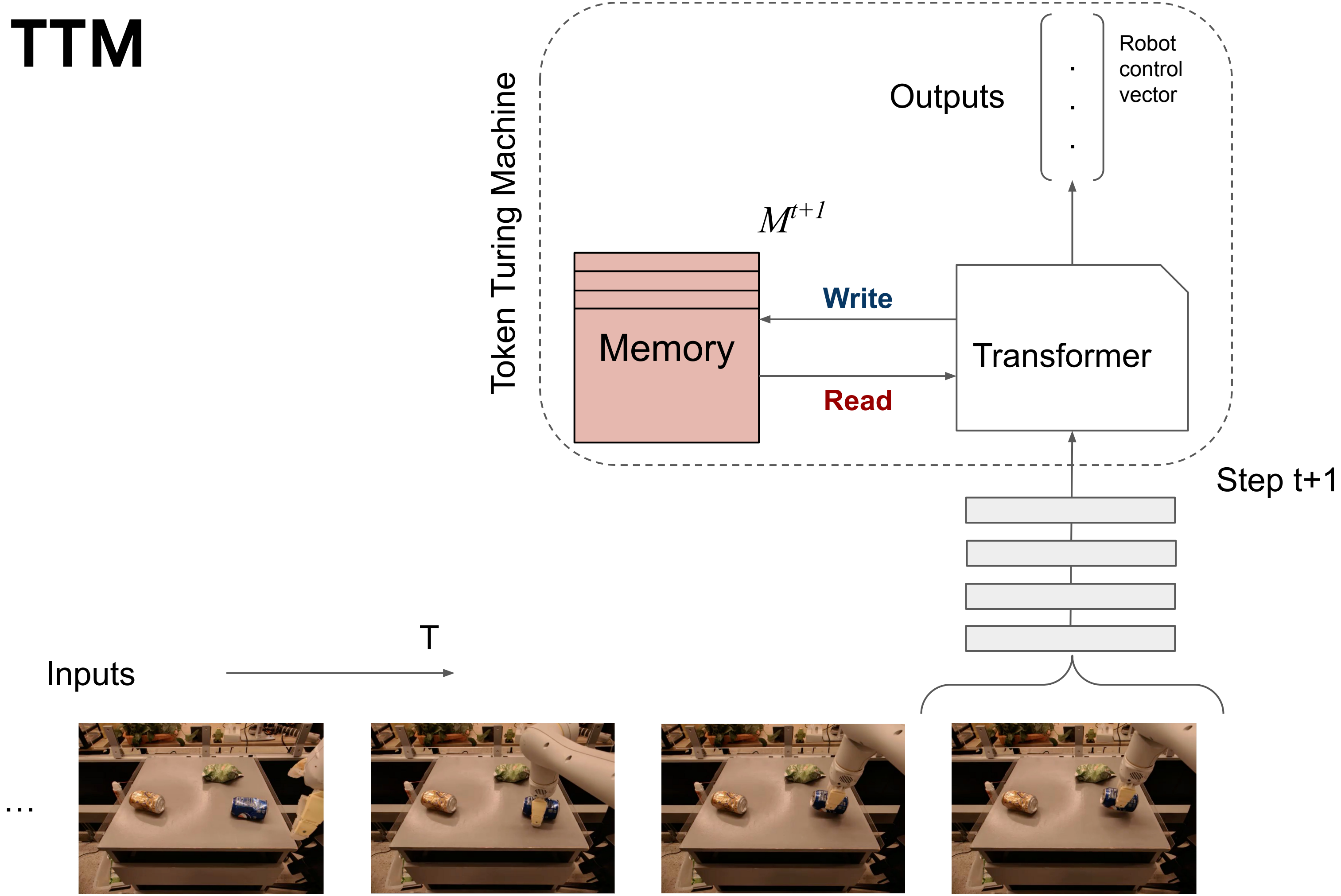
- RNN with Transformer + TokenLearner

(previous) Transformers

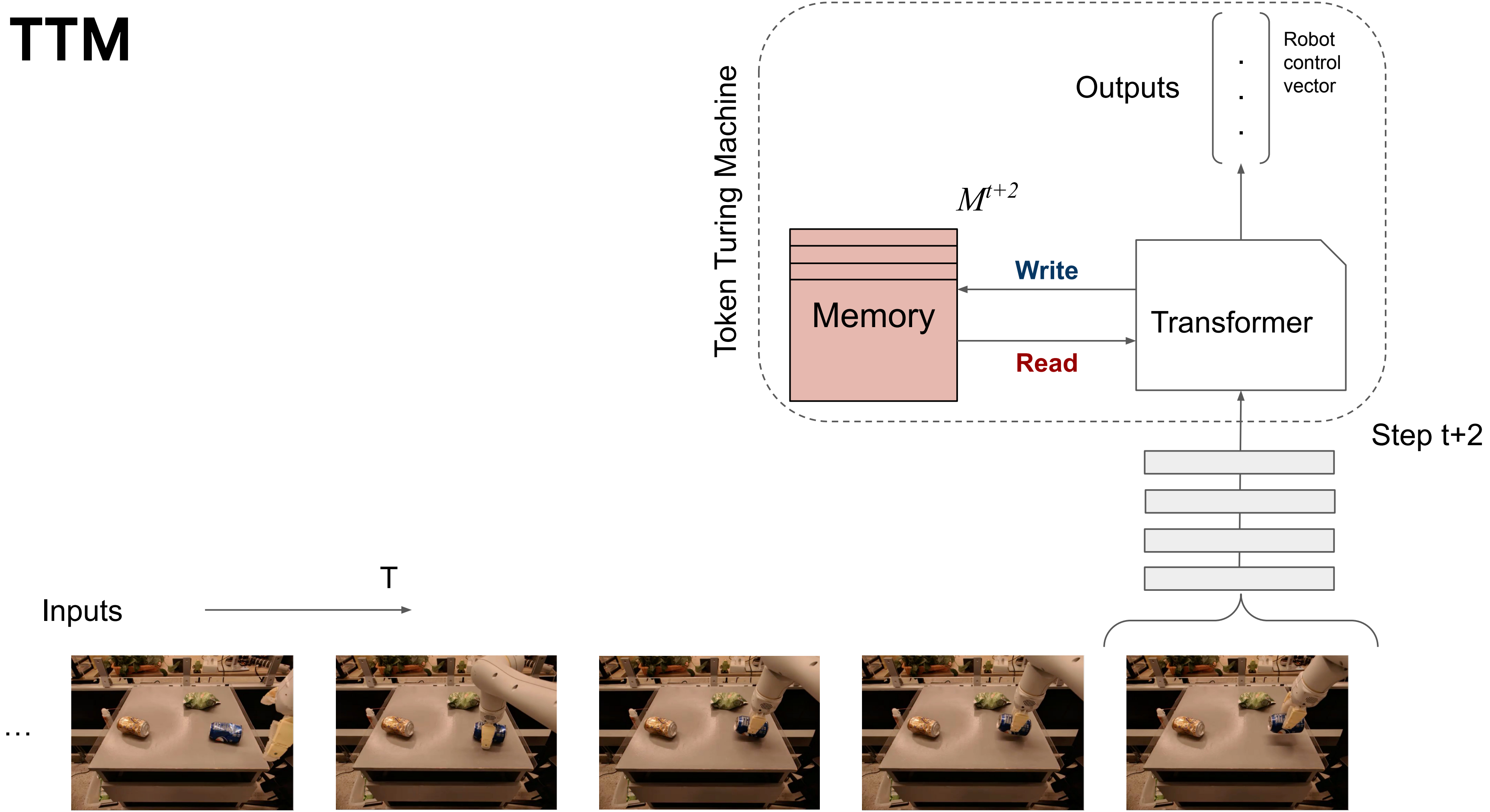



TTM
Token Turing Machine
Outputs
Robot control vector
M^t
Memory
Write
Read
Transformer
Step t
T
Inputs
...

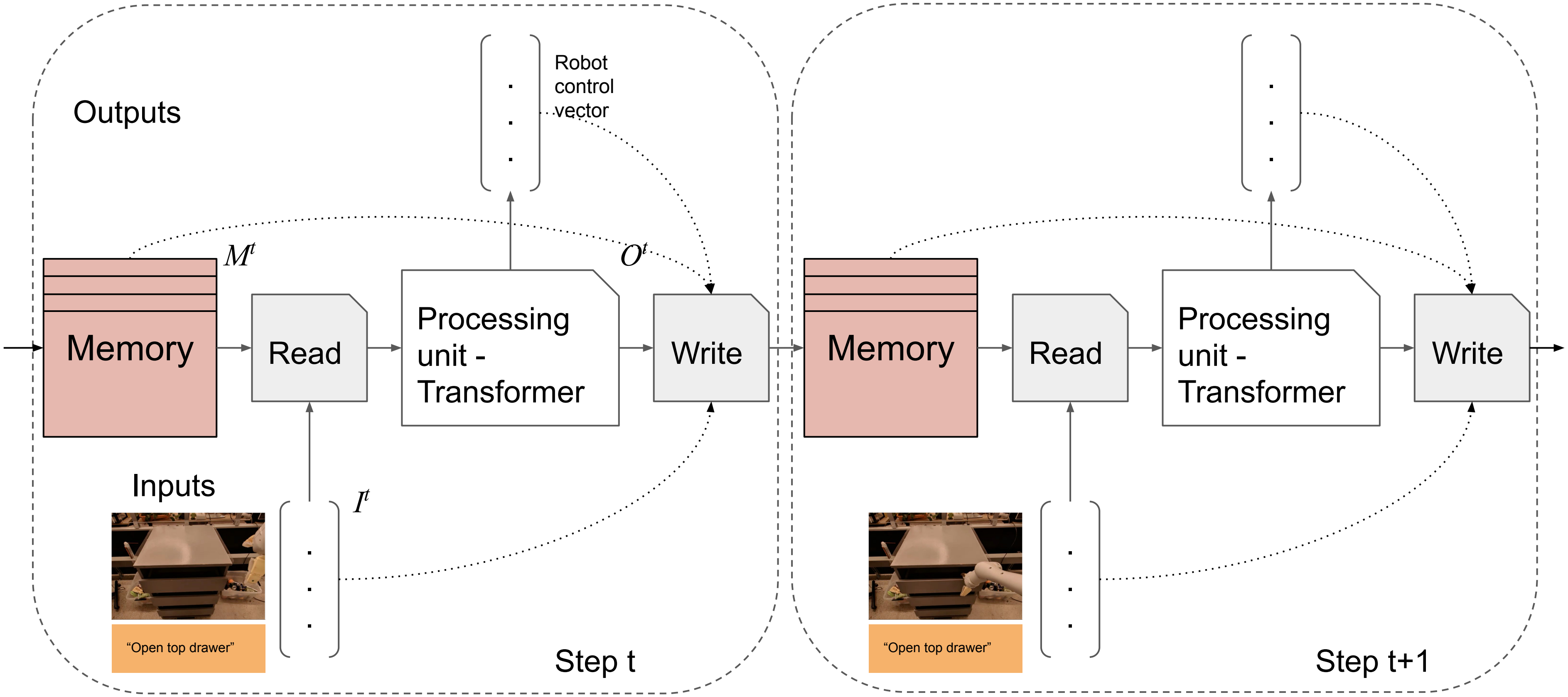
TTM


Token Turing Machine
Outputs
Robot
control
vector
M^{t+1}
Memory
Write
Read
Transformer
Step t+1
T
Inputs
...

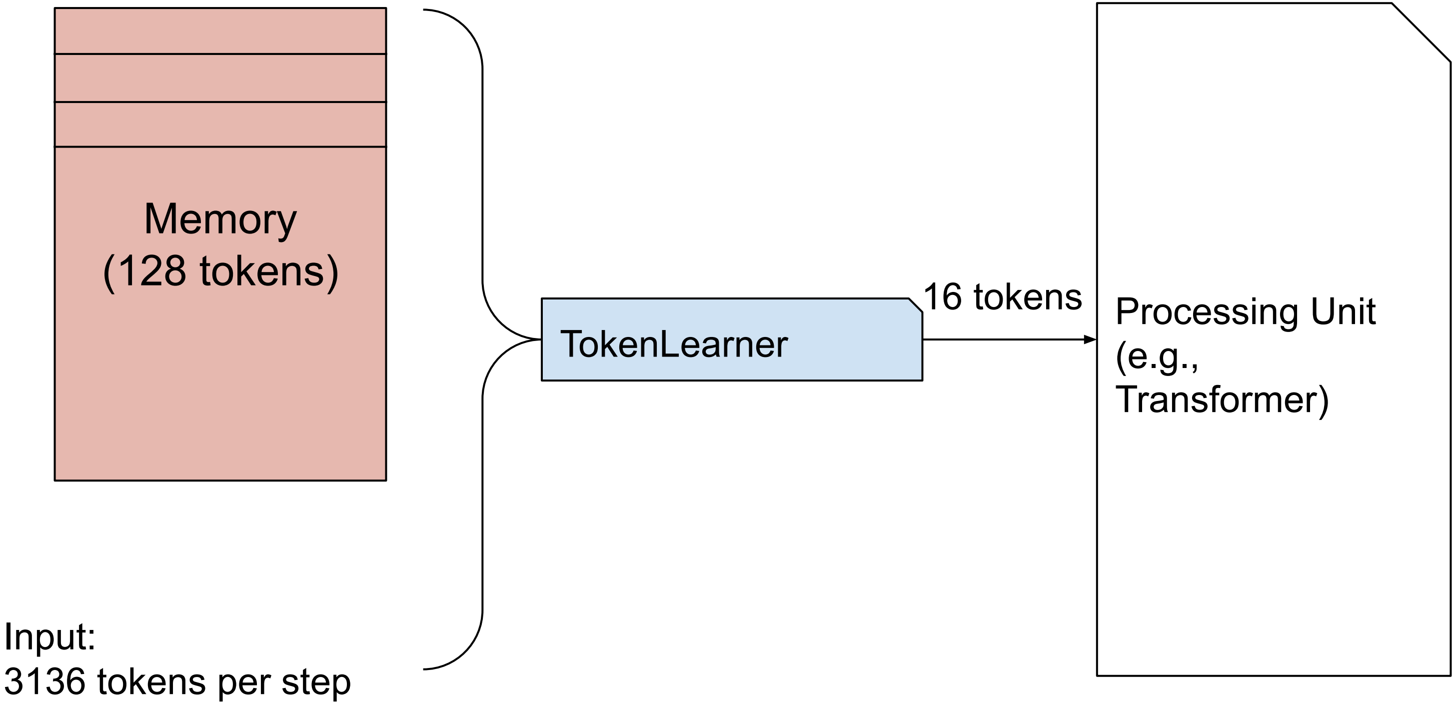
TTM


Token Turing Machine
Outputs
Robot control vector
M^{t+2}
Memory
Write
Read
Transformer
Step t+2
T
Inputs
...

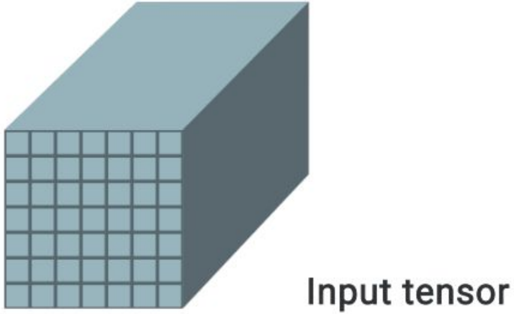
Token Turing Machines - Architecture



Token Turing Machines - Read

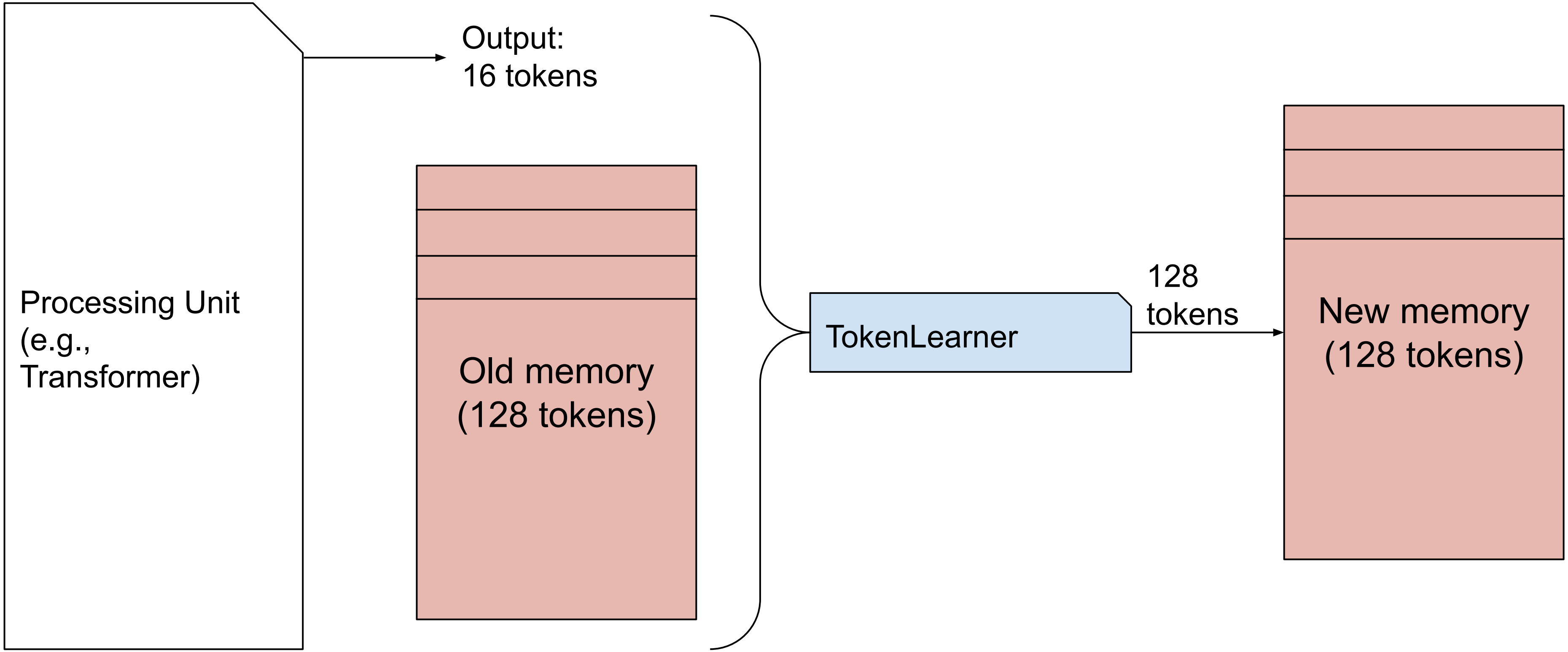


TokenLearner



[Ryoo et al., NeurIPS 2021]

Token Turing Machines - Write



Could be interpreted as a recurrent neural network with an explicit memory, where all its components are implemented with token-based operations.

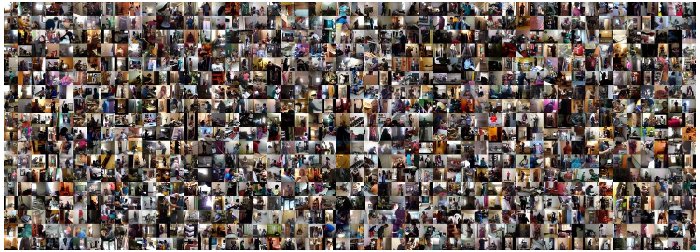
$$
\begin{aligned}
Z^t &= \mathrm{Read}(I^t, M^t) \\
O^t &= \mathrm{Process}(Z^t) \\
M^{t+1} &= \mathrm{Write}(M^t, O^t, I^t) \\
Y^t &= \mathrm{Output}(O^t)
\end{aligned}
$$

Temporal activity detection

Comparison against SOTA

Method	Setting	modality	mAP
I3D + super-events (Piergiovanni & Ryoo, 2018)	offline	RGB + Flow	19.41
I3D + super-events + TGM (Piergiovanni & Ryoo, 2019)	offline	RGB + Flow	22.30
I3D + STGCN (Ghosh et al., 2020)	offline	RGB + Flow	19.09
I3D + biGRU + VS-ST-MPNN (Mavroudi et al., 2020)	offline	RGB + Object	23.7
Coarse-Fine (w/ X3D) (Kahatapitiya & Ryoo, 2021)	offline	RGB	25.1
I3D + CTRN (Dai et al., 2021a)	offline	RGB	25.3
I3D + MS-TCT (Dai et al., 2022)	offline	RGB	25.4
I3D + PDAN (Dai et al., 2021b)	offline	RGB + Flow	26.5
I3D + CTRN (Dai et al., 2021a)	offline	RGB + Flow	27.8
I3D (Carreira & Zisserman, 2017)	online	RGB + Flow	17.22
X3D (Feichtenhofer, 2020)	online	RGB	18.87
ViViT-B (Arnab et al., 2021)	online	RGB	23.18
ViViT-B + TTM (ours)	online	RGB	26.34
ViViT-L (Arnab et al., 2021)	online	RGB	26.01
ViViT-L + TTM (ours)	online	RGB	28.79

Charades dataset

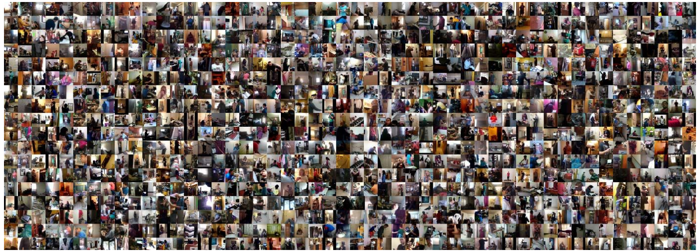


Temporal activity detection

Comparison against regular Transformers

Method	mAP	GFLOPS
ViViT only	23.18	-
Alternative temporal models		
Temporal MLPMixer (tokens=96)	24.41	0.382
Causal Transformer (tokens=96)	25.85	0.523
Temporal Transformer (tokens=96)	25.61	1.269
Temporal MLPMixer (tokens=3360)	24.26	13.317
Causal Transformer (tokens=3360)	25.88	29.695
Temporal Transformer (tokens=3360)	25.53	112.836
Alternative recurrent networks		
LSTM	23.96	0.107
Recurrent Transformer (tokens=16+16)	25.97	0.410
Recurrent Transformer (tokens=3136+16)	25.97	17.10
Token Turing Machines		
TTM-Mixer ($n = 16$)	25.83	0.089
TTM-Transformer ($n = 16$)	26.24	0.228
TTM-Mixer ($n = 3136$)	26.14	0.704
TTM-Transformer ($n = 3136$)	26.34	0.842

Charades dataset



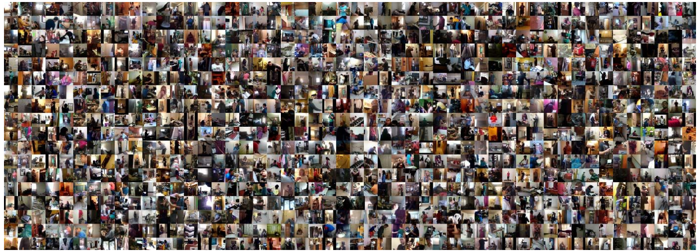
Time complexity is much lower:
O(T) vs. O(1).

Temporal activity detection

Comparison against regular Transformers

Method	mAP	GFLOPS
ViViT only	23.18	-
Alternative temporal models		
Temporal MLPMixer (tokens=96)	24.41	0.382
Causal Transformer (tokens=96)	25.85	0.523
Temporal Transformer (tokens=96)	25.61	1.269
Temporal MLPMixer (tokens=3360)	24.26	13.317
Causal Transformer (tokens=3360)	25.88	29.695
Temporal Transformer (tokens=3360)	25.53	112.836
Alternative recurrent networks		
LSTM	23.96	0.107
Recurrent Transformer (tokens=16+16)	25.97	0.410
Recurrent Transformer (tokens=3136+16)	25.97	17.10
Token Turing Machines		
TTM-Mixer ($n = 16$)	25.83	0.089
TTM-Transformer ($n = 16$)	26.24	0.228
TTM-Mixer ($n = 3136$)	26.14	0.704
TTM-Transformer ($n = 3136$)	26.34	0.842

Charades dataset



Time complexity is much lower: O(T) vs. O(1).

Spatio-temporal activity detection

Comparison against MeMViT, which is a memory + MViT.

Model	mAP	+GFLOPS
MViT	26.2	-
+ memory (i.e., MeMViT [66])	28.5 (+2.3)	1.3
ViViT-B	25.2	-
+ TTM per video	27.9 (+2.7)	0.8
+ TTM per box (# layers=1)	31.3 (+6.1)	1.0
+ TTM per box (# layers=4)	31.5 (+6.3)	2.0

On AVA v2.2, with K400 pre-training

AVA v2.2 dataset

Left: Sit, Talk to, Watch; Right: Crouch/Kneel, Listen to, Watch

Left: Stand, Carry/Hold, Listen to; Middle: Stan Carry/Hold, Talk to; Right: Sit, Write

Left: Sit, Ride, Talk to; Right: Sit, Drive, Listen to

Left: Stand, Watch; Middle: Stand, Play instrument; Right: Sit, Play instrument

Robot action policy learning

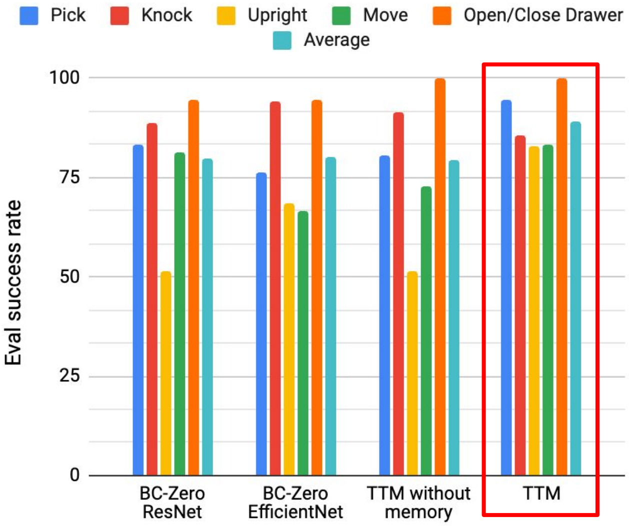
Data used in Google’s Robotics Transformer-1 (RT-1)

Data: 130k episodes of over 700+ tasks, collected using 13 robots over 17 months

Goal: robot control

[RT-1, RSS 2023]

Robot action policy learning



Data used in Google's Robotics Transformer-1 (RT-1)

[Robotics Transformer-1, RSS 2023]

Summary

Token Turing Machines

- Represent and process a sequence of **many** tokens

It is a generic framework - a recurrent Transformer with token-based **memory**

Contact: mryoo@google.com

https://github.com/google-research/scenic/tree/main/scenic/projects/token_turing