Memory-Based Sequential Attention

Jason Stock[1] and Charles Anderson

Department of Computer Science, Colorado State University, 2023

Introduction

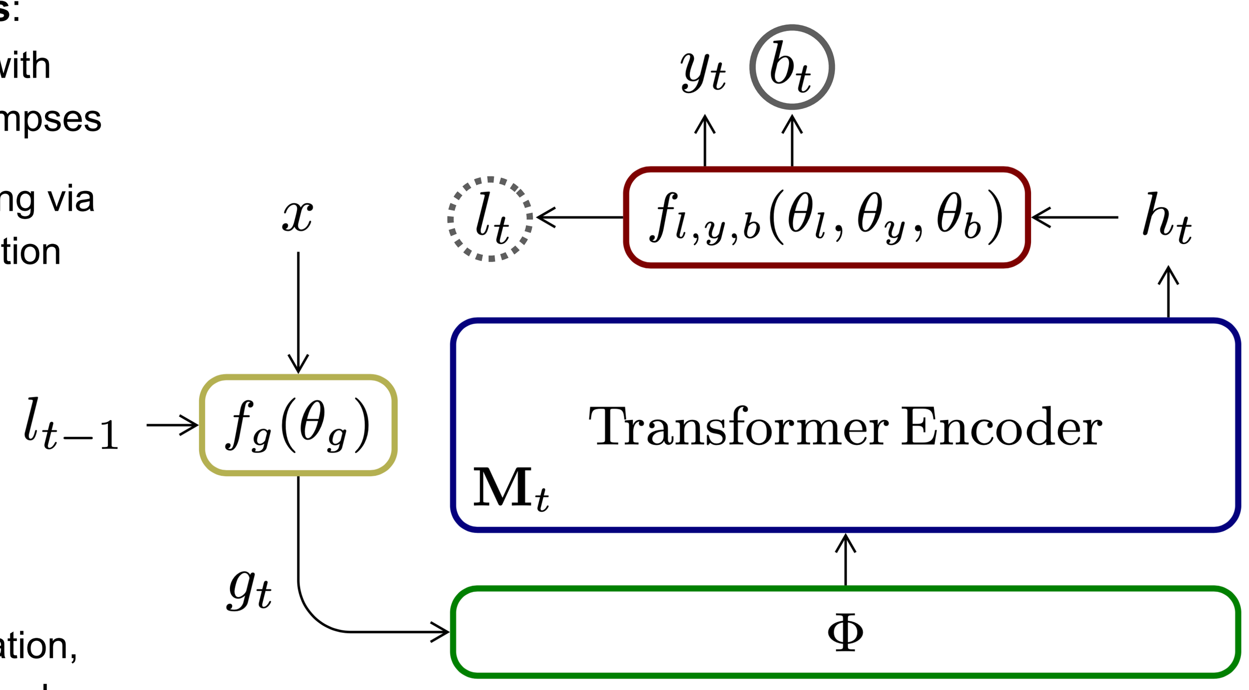
Problem Motivation: sequential attention actively samples glimpses of information from a sensory scene over multiple time steps. Prior computational methods involving recurrent neural networks are limited as they **(a)** may lose information over accumulated glimpses and **(b)** are unable to dynamically reweigh glimpses at individual steps.

Contributions: a biologically-inspired model of sequential attention for classification using a transformer-based memory module, improving performance and interpretability.

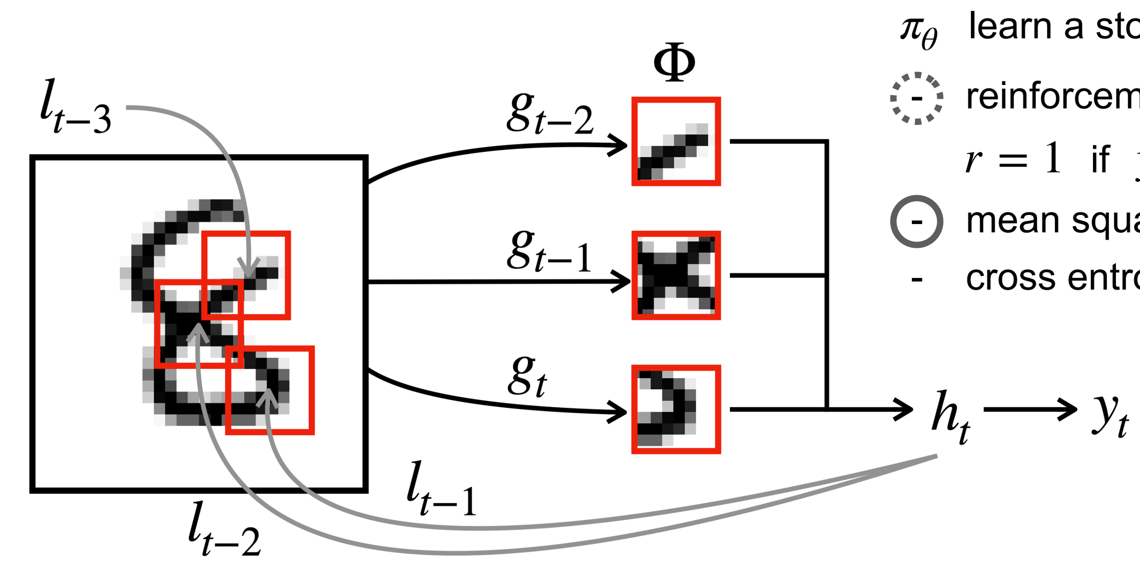
Methodology

Network Components:

- Φ memory module with history of prior glimpses
- $\mathbf{M}_t$ dynamic reweighing via masked self-attention
- f_g glimpse network
- $f_{l,y,b}$ location, classification, and baseline networks



(a) Network Architecture



(b) Flow Diagram

Training Procedure:

- ? where to look next: $(t-1) \xrightarrow{?} (t)$
- π_θ learn a stochastic policy
- reinforcement learning (policy gradient)
- $r = 1$ if $y = t$, 0 otherwise
- mean squared error (stabilize variance)
- cross entropy (all other components)

Figure 1: High-level overview of our proposed method.

Experimental Results

Table 1: Classification error where S is the glimpse scale, H is the number of attention heads, and K is the number of sequential glimpses.

Model	Error
FC, 2 layers [256, 256]	2.20
CNN, 2 layers [16, 32]	1.17
ViT, 7 × 7, 4H	3.48
RAM, 6 K, 8 × 8, 1 S	**0.94**
Ours, 6 K, 8 × 8, 1 S, 1 H	1.29
Ours, 6 K, 8 × 8, 1 S, 2 H	1.11
Ours, 6 K, 8 × 8, 1 S, 4 H	**1.05**

(a) MNIST

Model	Error
FC, 2 layers [256, 256]	56.82
CNN, 4 layers [8, 16, 32, 64]	6.71
ViT, 12 × 12, 4H	29.48
RAM, 6 K, 12 × 12, 3 S	6.43
Ours, 6 K, 12 × 12, 3 S, 1 H	7.89
Ours, 6 K, 12 × 12, 3 S, 2 H	7.47
Ours, 6 K, 12 × 12, 3 S, 4 H	**6.20**

(b) Cluttered

Datasets: MNIST and Cluttered and Translated MNIST.
Baseline Models: Recurrent Model of Sequential Attention (RAM) and common networks of comparable size with increased complexity.

Sample Trajectories (Figure 2): our learned policy finds meaningful locations without observing the entire image or digit.
Self-Attention Weights: task irrelevant locations have little to no positional significance, whereas conspicuous locations are largely attended to.

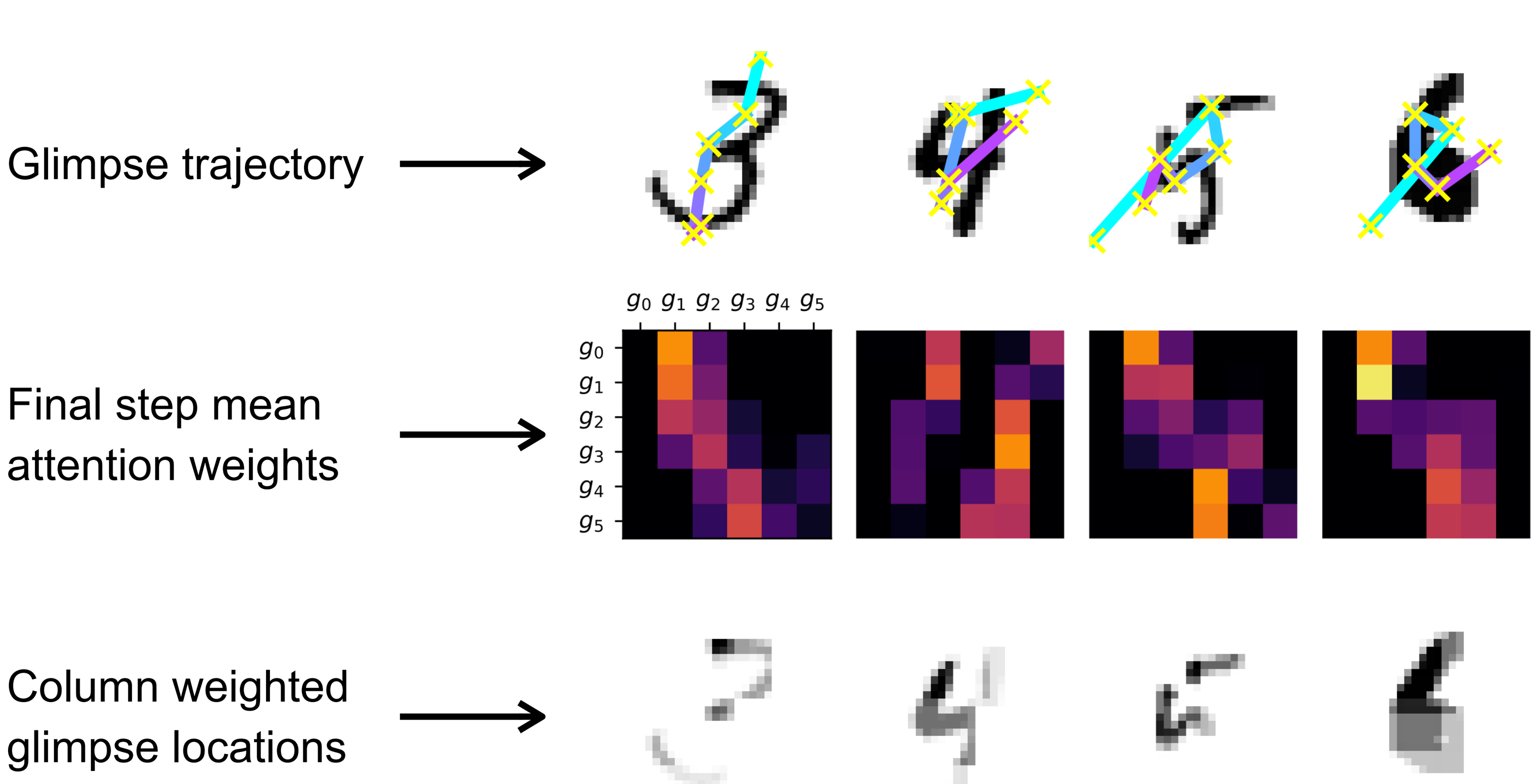


Figure 2: Attention weighted glimpse locations on MNIST.

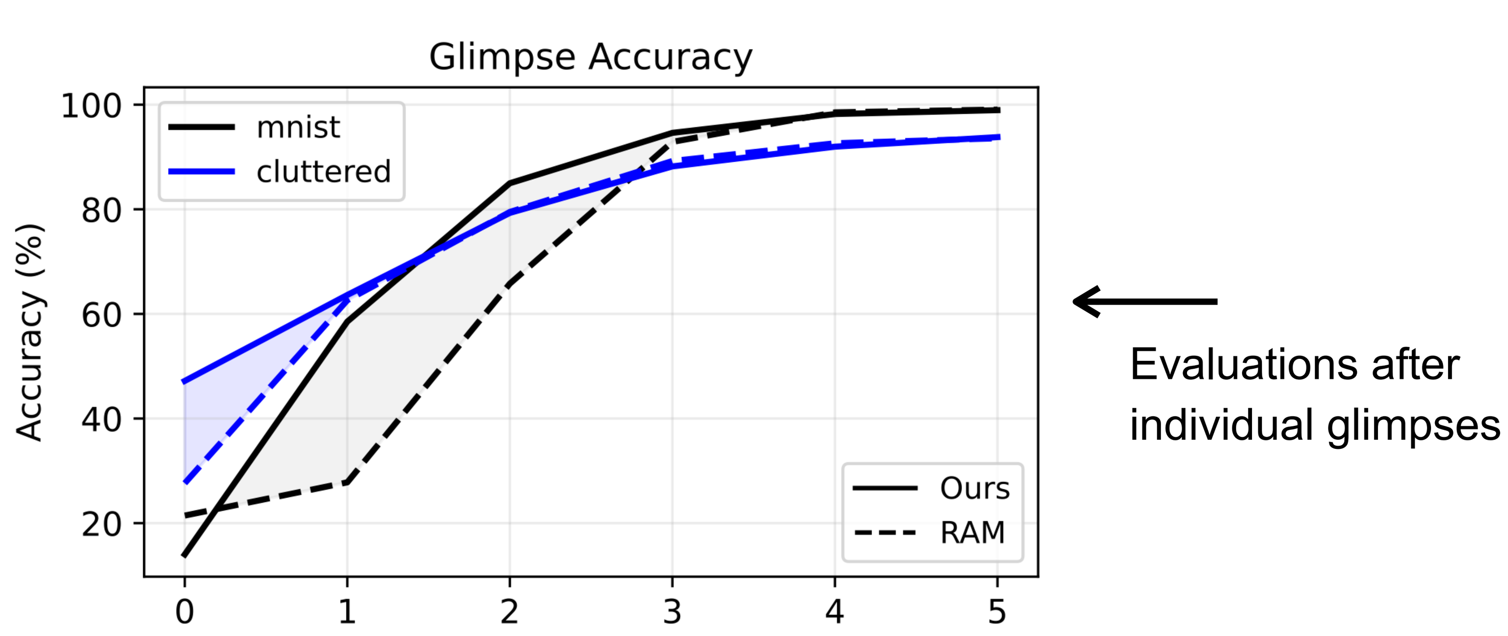


Figure 4: Test accuracy achieved after each glimpse.

Permuting Initial Locations

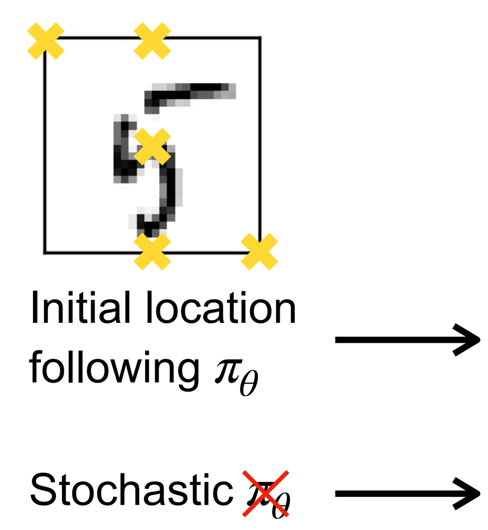


Table 2: Classification error (mean ± std) at different starting locations, w/ & w/o π.

Position	MNIST	Cluttered
Random	**1.12 ± 0.03**	6.76 ± 0.07
Top-Middle	1.13 ± 0.05	7.13 ± 0.11
Top-Left	1.16 ± 0.03	7.02 ± 0.22
Center	1.20 ± 0.05	**6.36 ± 0.19**
Bottom-Middle	1.20 ± 0.07	7.32 ± 0.15
Bottom-Right	1.12 ± 0.05	6.74 ± 0.18
Random Policy	29.49 ± 0.28	25.66 ± 0.03

Additional Interpretations

Class Specific Trajectories (Figure 3): digit structure is learned by the policy for the individual classes of MNIST.

Sequential Performance (Figure 4): model performance increases as additional glimpses are made; outperforming RAM on early time steps.

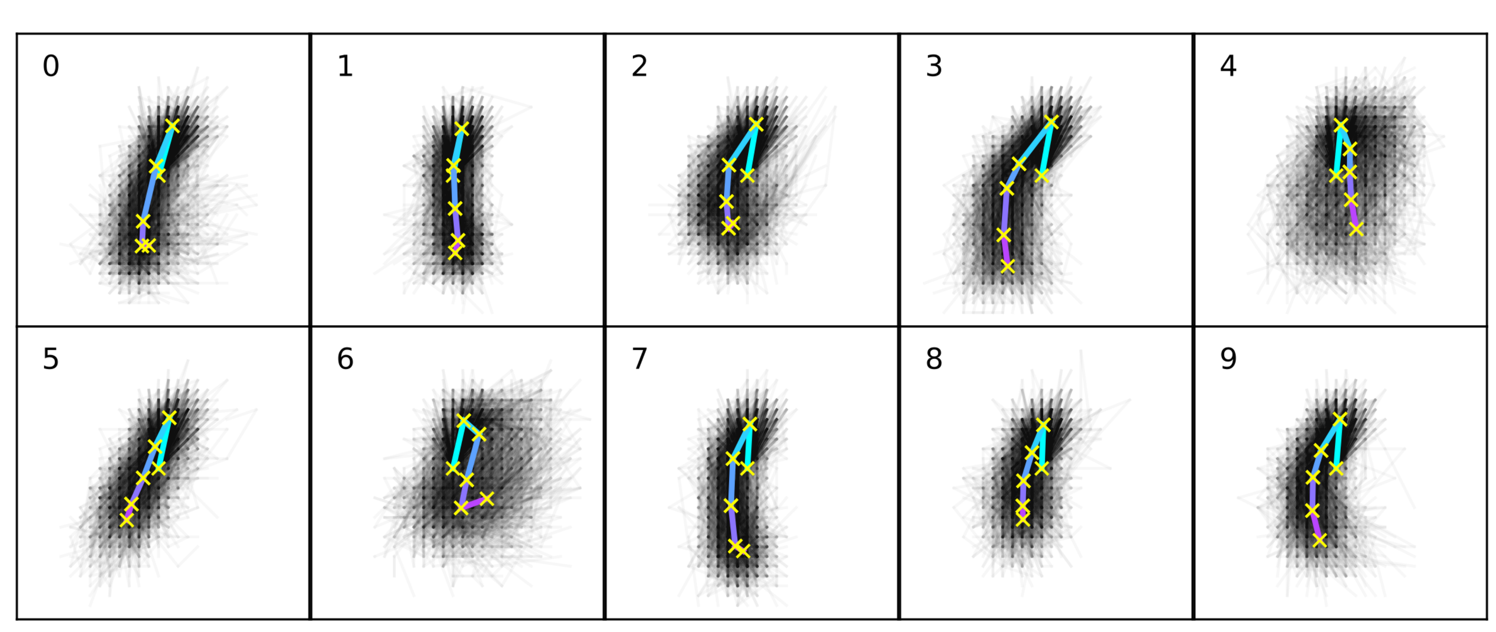


Figure 3: Class specific trajectories of all MNIST test samples and mean trajectory.

Conclusions

Overview: we contextualize the relationships of sequential glimpses using a transformer with a history of previous locations, yielding dynamic interactions and an interpretable policy for selecting image regions.

Takeaways: able to scale to arbitrary input sizes, while leveraging transformer properties with a shorter sequence length and inherent translation invariance.

Other Datasets (?): applications domains such as medical diagnoses and forecasting weather and identifying indicators of climate change.

Future Work: saliency measures to guide initial locations and disentangle the contribution of attention weights for location and output predictions.

[1] stock@rams.colostate.edu | cs.colostate.edu/~stock