23rd Feb 2023

RMT: Recurrent Memory Transformer

```
@inproceedings{bulatovrecurrent,
  title={Recurrent Memory Transformer},
  author={Bulatov, Aydar and Kuratov, Yuri and Burtsev, Mikhail},
  booktitle={Advances in Neural Information Processing Systems (NeurIPS)},
  year={2022}
}
```


Course Instructor

Prof. C. Krishna Mohan

Presenter

Rahul Vigneswaran K*

CS23MTECH02002

Assigned TA

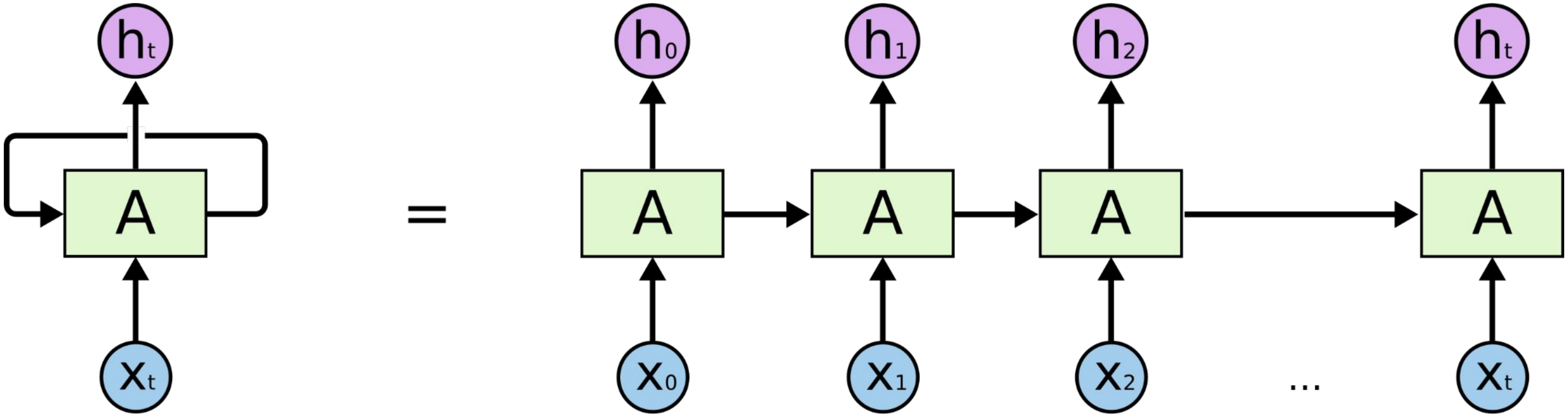
Peketi Divya

*Presented as part of the coursework for Visual Computing (CS6450)

Table of Contents

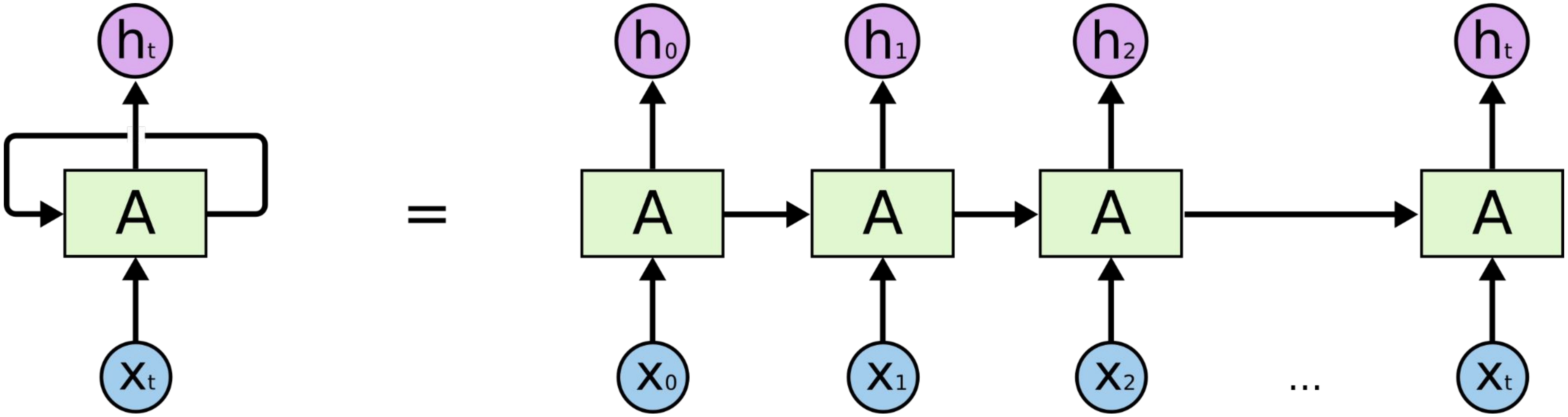
Introduction: What problem are we trying to solve?

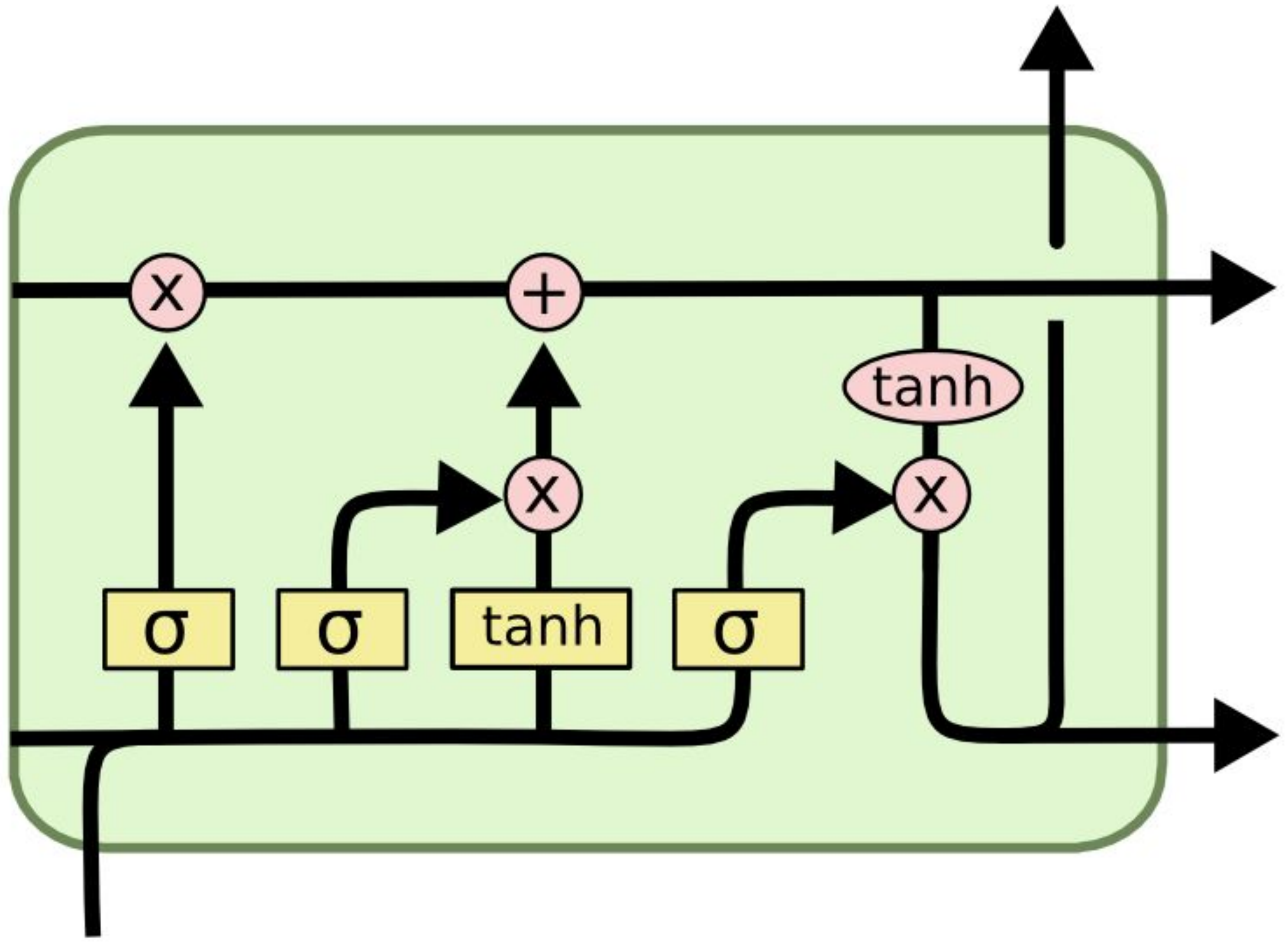
Introduction: What problem are we trying to solve?



“Understanding LSTM Networks.” *Understanding LSTM Networks -- Colah's Blog*, https://colah.github.io/posts/2015-08-Understanding-LSTMs/ . (Blog)

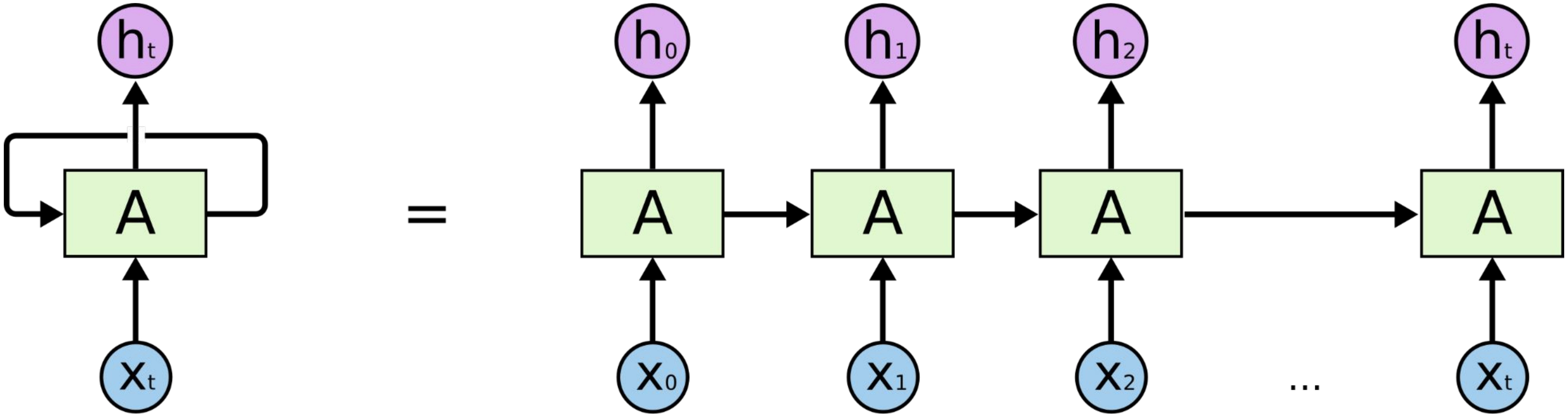
Introduction: What problem are we trying to solve?

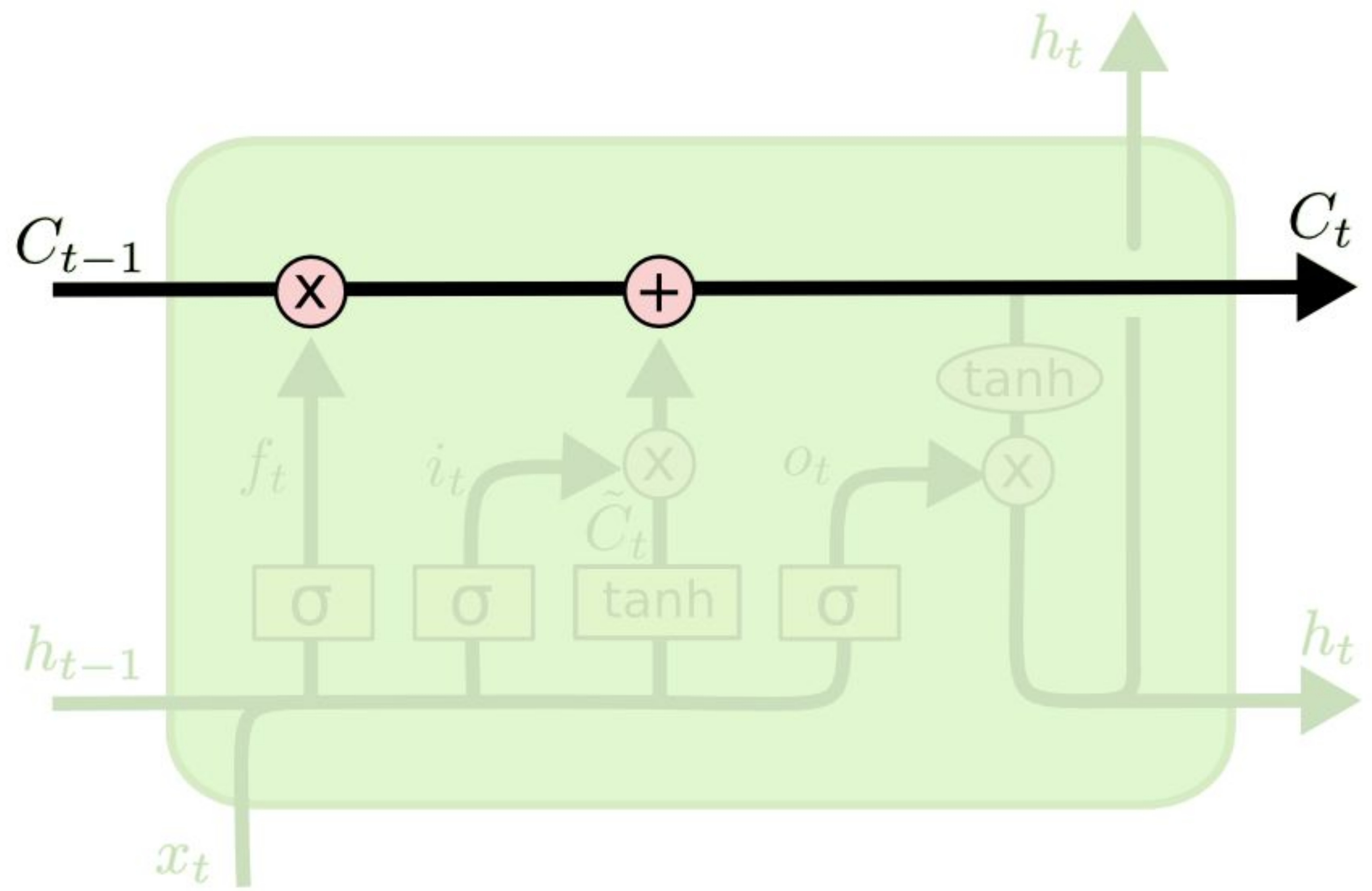




“Understanding LSTM Networks.” *Understanding LSTM Networks -- Colah's Blog*, https://colah.github.io/posts/2015-08-Understanding-LSTMs/ . (Blog)

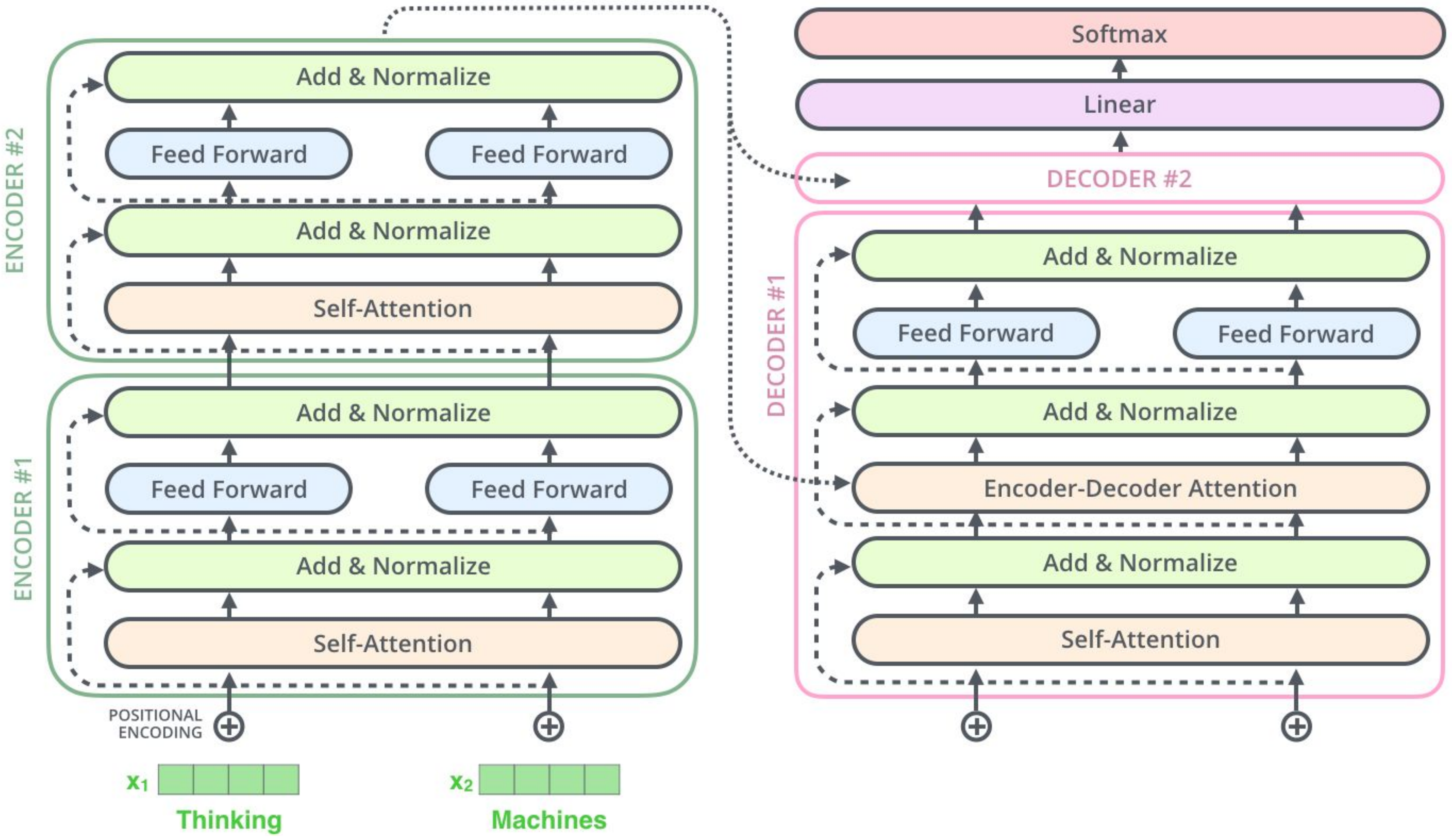
Introduction: What problem are we trying to solve?





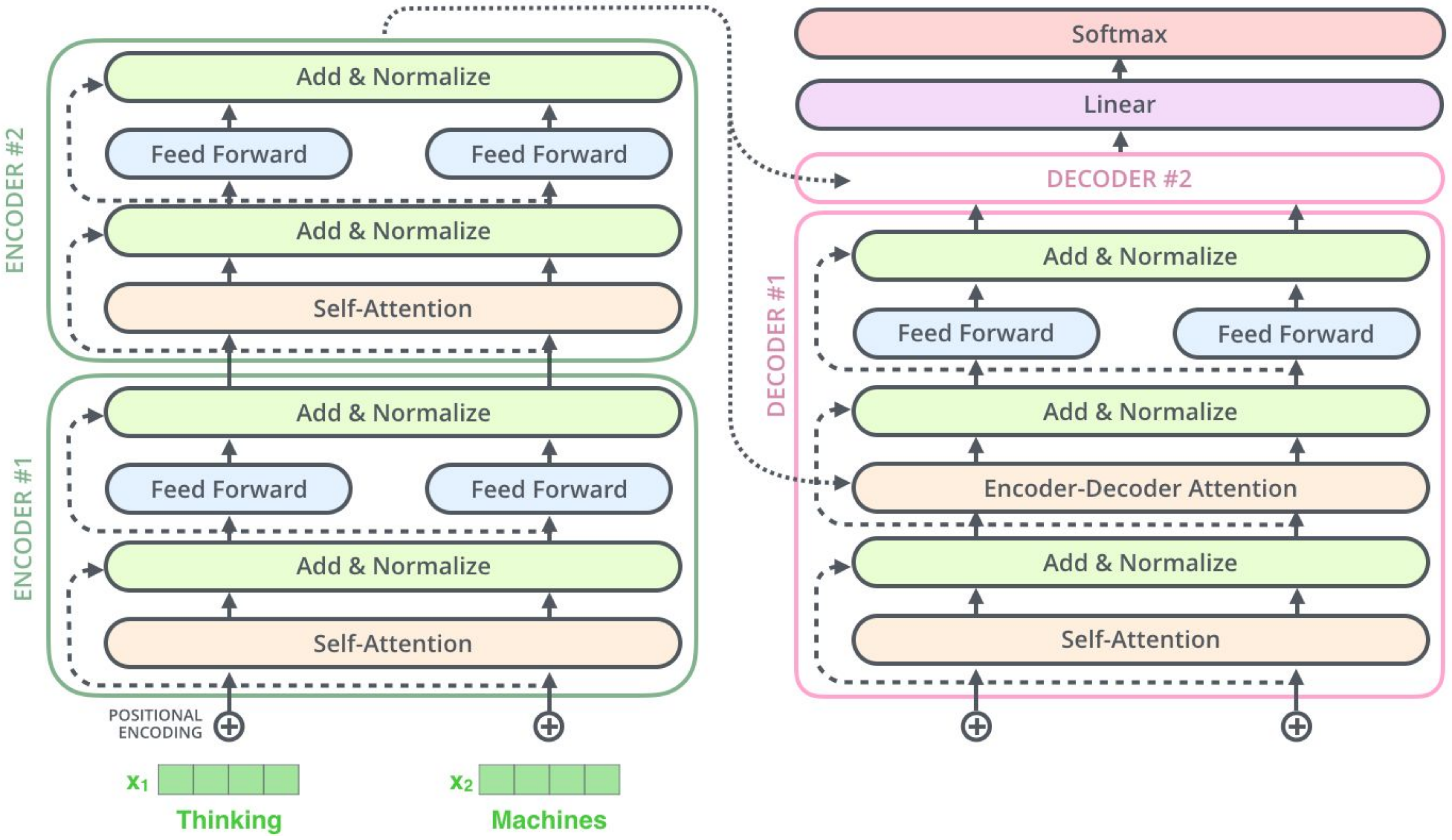
"Understanding LSTM Networks." *Understanding LSTM Networks -- Colah's Blog*, https://colah.github.io/posts/2015-08-Understanding-LSTMs/ . (Blog)

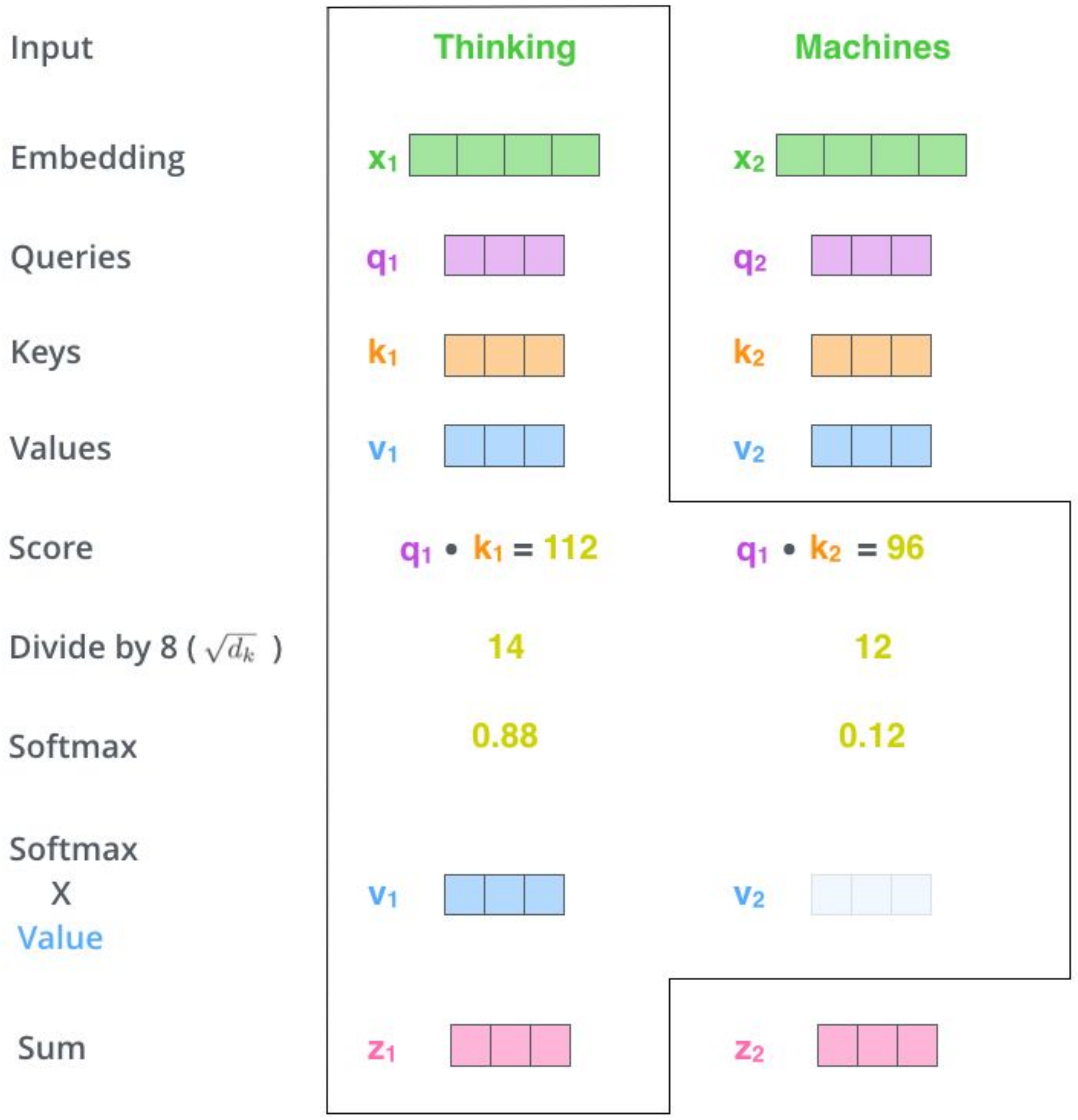
Introduction: What problem are we trying to solve?



Alammar, J. (n.d.). The illustrated transformer. The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. Retrieved February 22, 2023, from https://jalammar.github.io/illustrated-transformer/ (Blog)

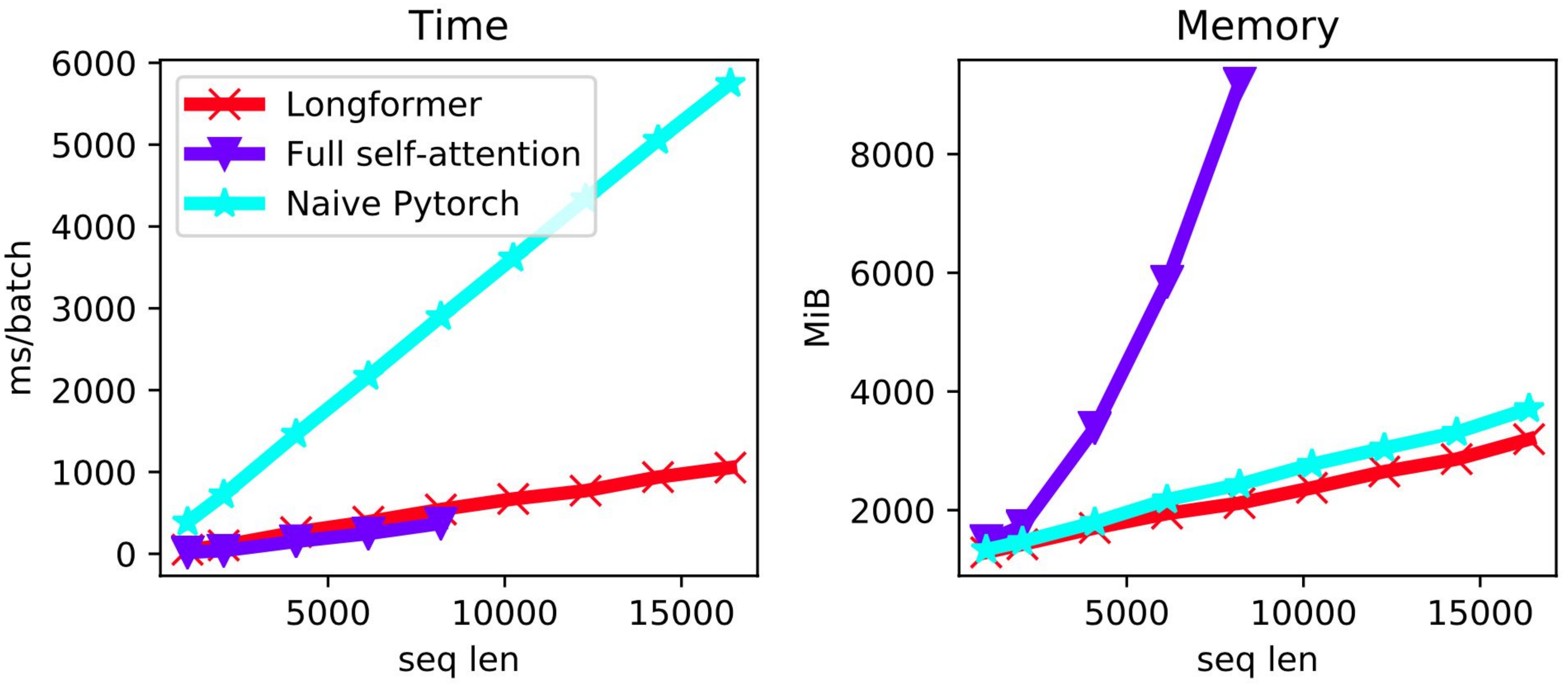
Introduction: What problem are we trying to solve?





Alammar, J. (n.d.). The illustrated transformer. The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. Retrieved February 22, 2023, from https://jalammar.github.io/illustrated-transformer/ (Blog)

Introduction: What problem are we trying to solve?

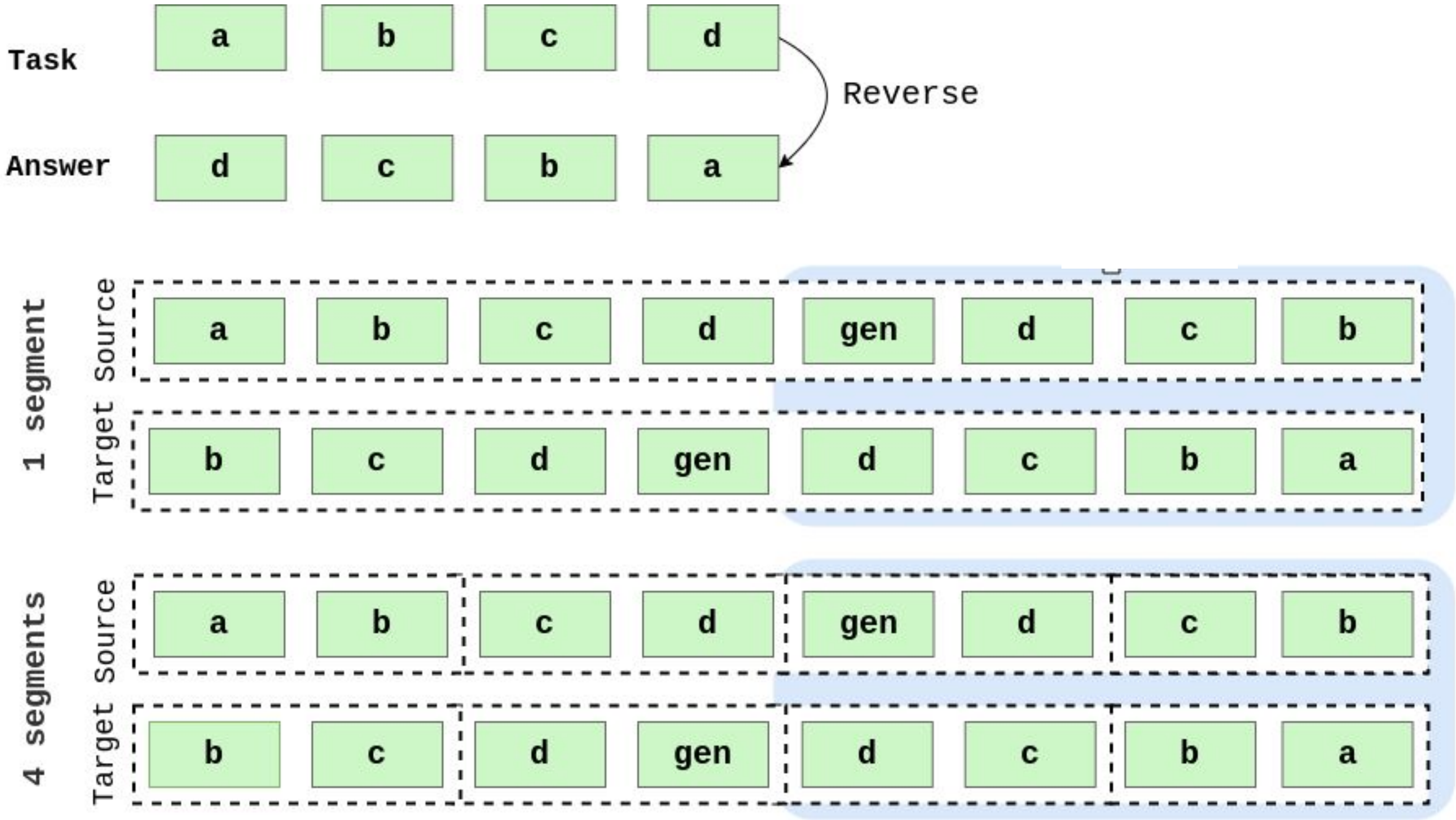


Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The Long-Document Transformer." (Arxiv 2020)

01

Algorithmic Task

- Reverse Task
- Copy Task

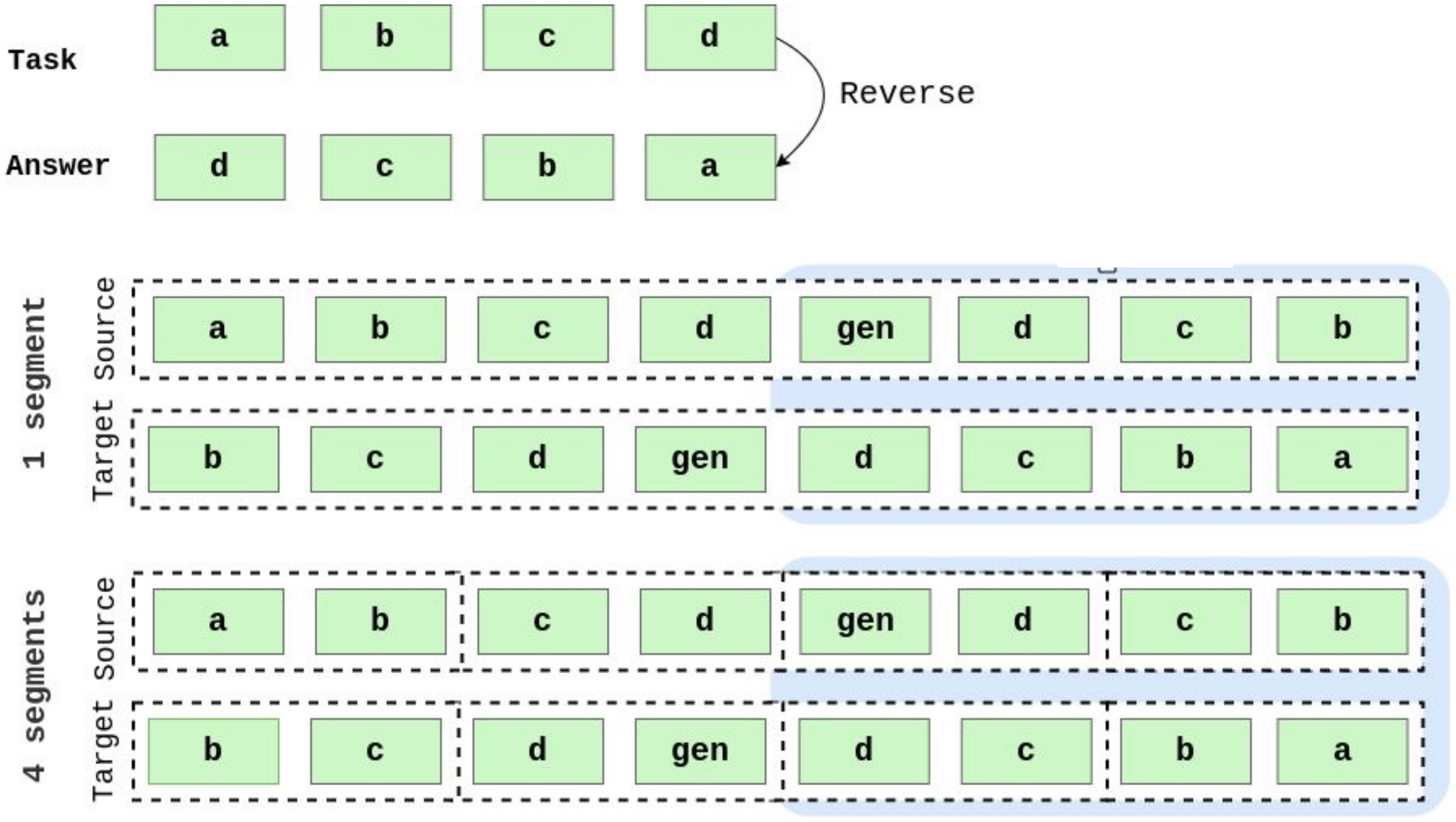


Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

01

Algorithmic Task

- Reverse Task
- Copy Task
- Associative Retrieval Task
- Quadratic Equations



Example equation string:

*-4*x^2+392*x-2208=0,*

solution string:

*x^2-98*x+552=0;D=98^2-4*1*552=7396=86^2;x=(98-86)/2=6;x=(98+86)/2=92*

Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

01

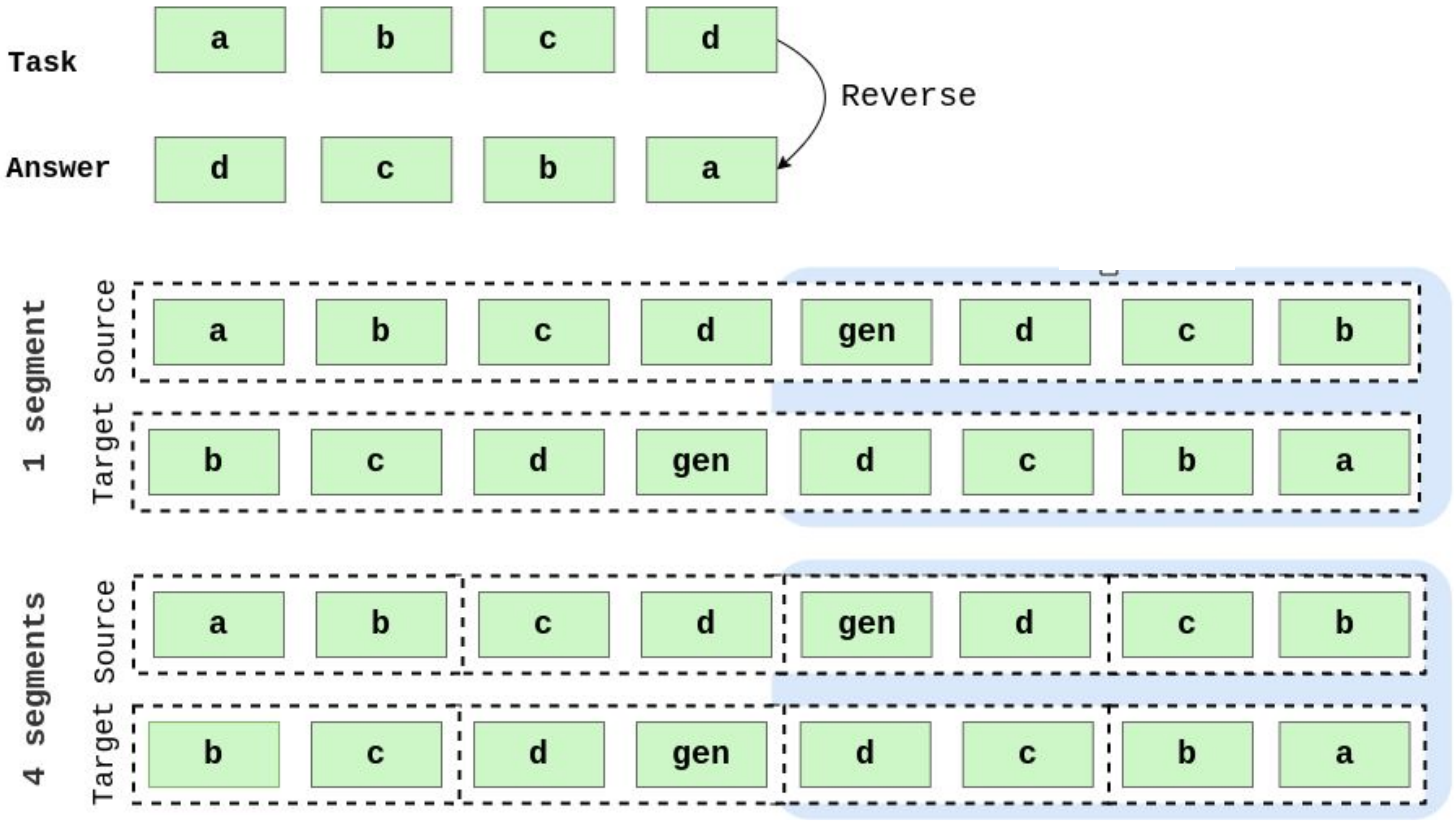
Algorithmic Task

- Reverse Task
- Copy Task
- Associative Retrieval Task
- Quadratic Equations

02

Language Modelling & NLP Tasks

- WikiText103 (word level)
- enwik8 (char level)
- Hyperpartisan news



Example equation string:

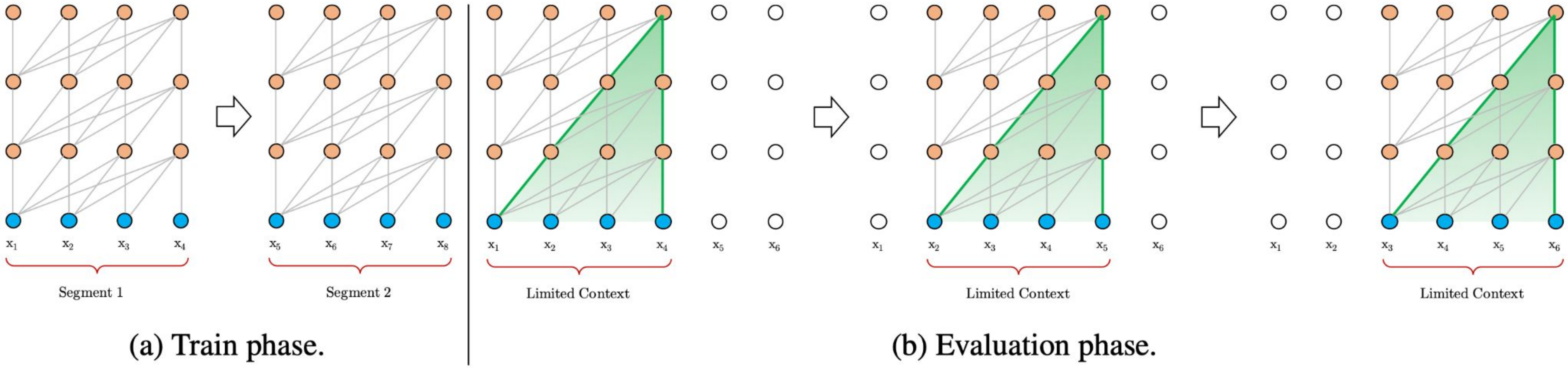
*-4*x^2+392*x-2208=0,*

solution string:

*x^2-98*x+552=0;D=98^2-4*1*552=7396=86^2;x=(98-86)/2=6;x=(98+86)/2=92*

Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Related Work: Transformer-XL



(a) Train phase.

(b) Evaluation phase.

Figure 1: Illustration of the vanilla model with a segment length 4.

Dai, Zihang, et al. "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)

Related Work: Transformer-XL

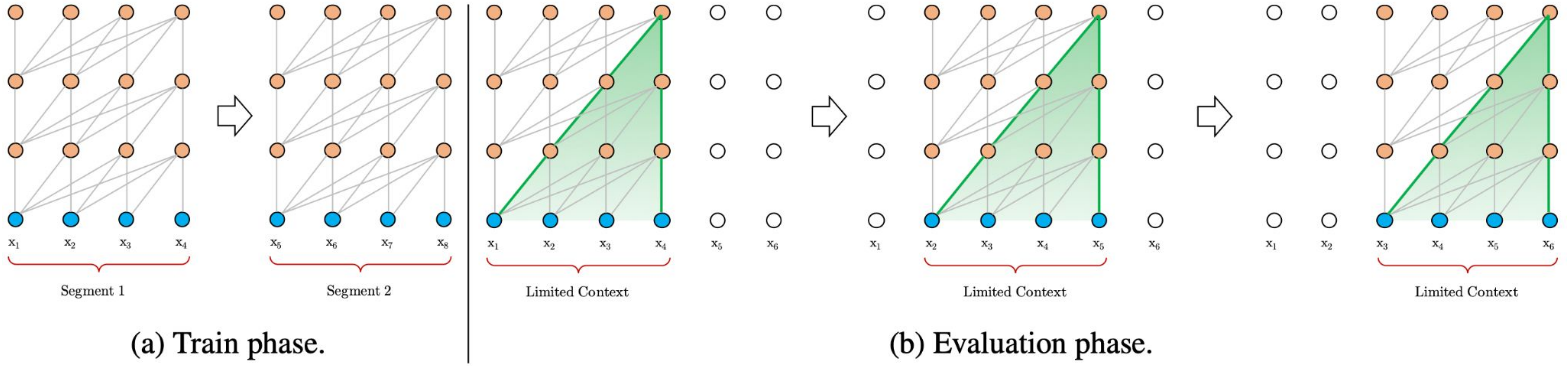


Figure 1: Illustration of the vanilla model with a segment length 4.

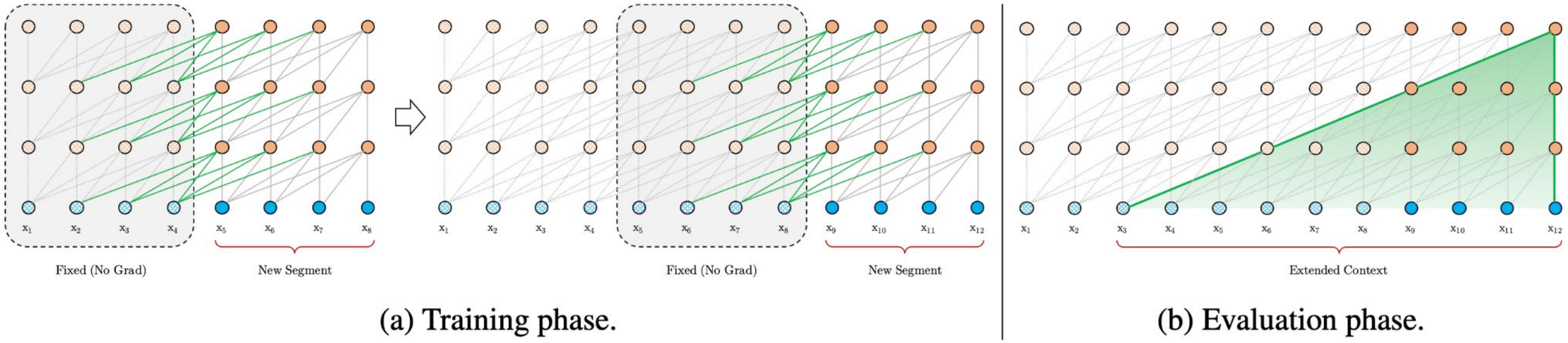
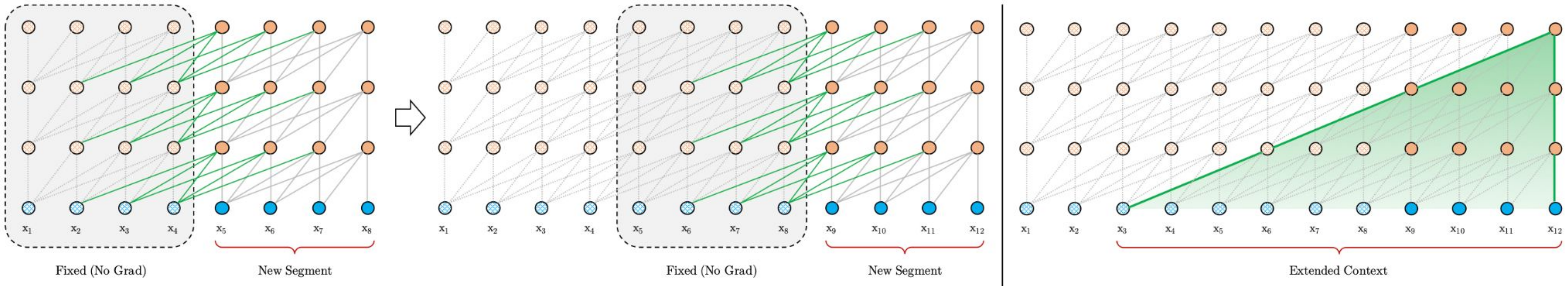


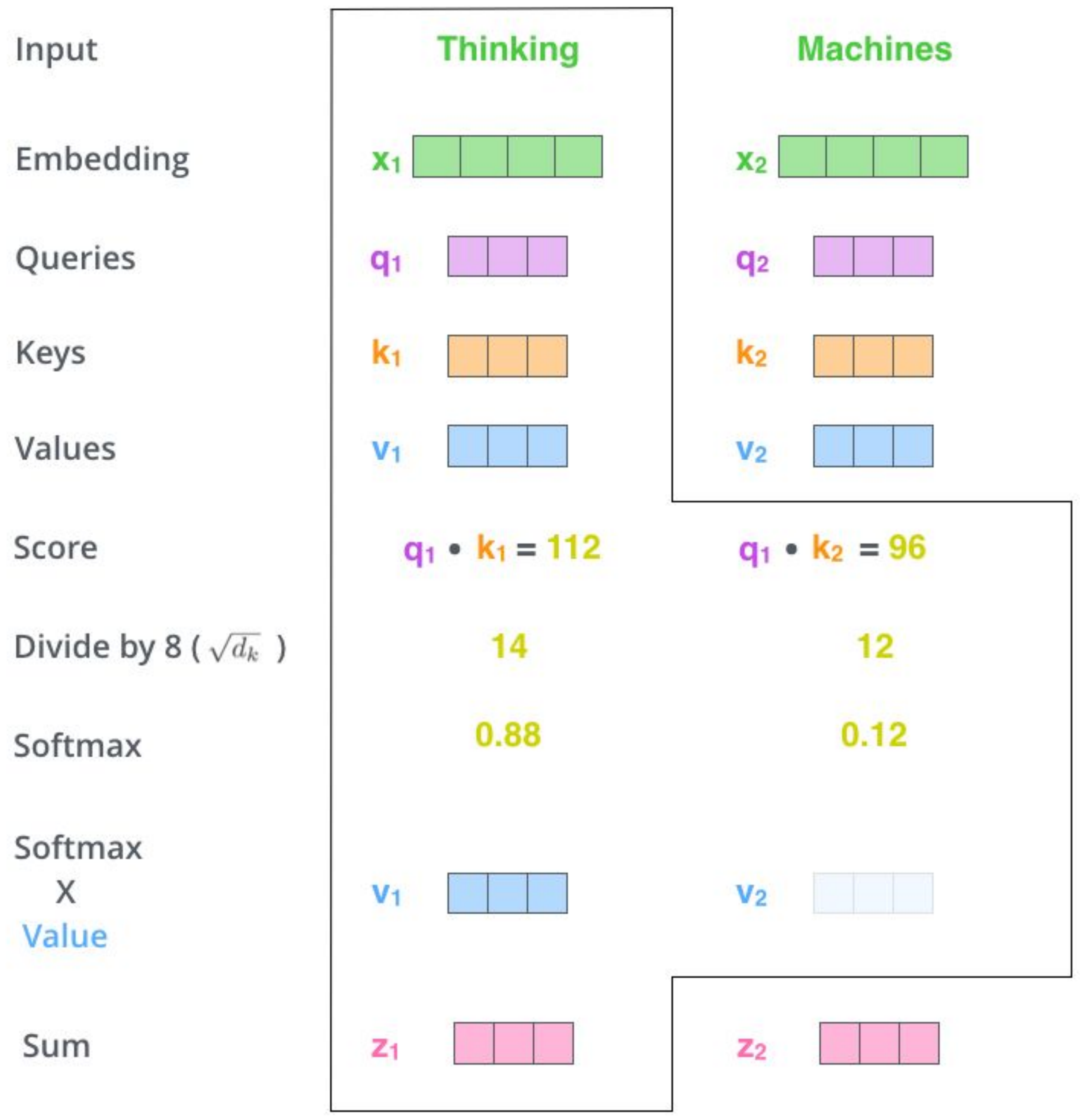
Figure 2: Illustration of the Transformer-XL model with a segment length 4.

Dai, Zihang, et al. "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)

Related Work: Transformer-XL



(a) Training phase. (b) Evaluation phase.



$$\widetilde{\mathbf{h}}_{\tau+1}^{n-1} = \left[\mathrm{SG}(\mathbf{h}_{\tau}^{n-1}) \circ \mathbf{h}_{\tau+1}^{n-1}\right],$$

$$\mathbf{q}_{\tau+1}^{n}, \mathbf{k}_{\tau+1}^{n}, \mathbf{v}_{\tau+1}^{n} = \mathbf{h}_{\tau+1}^{n-1}\mathbf{W}_q^{\top}, \widetilde{\mathbf{h}}_{\tau+1}^{n-1}\mathbf{W}_k^{\top}, \widetilde{\mathbf{h}}_{\tau+1}^{n-1}\mathbf{W}_v^{\top},$$

$$\mathbf{h}_{\tau+1}^{n} = \text{Transformer-Layer}\left(\mathbf{q}_{\tau+1}^{n}, \mathbf{k}_{\tau+1}^{n}, \mathbf{v}_{\tau+1}^{n}\right).$$

Dai, Zihang, et al. "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)

Related Work: Transformer-XL - Potential drawbacks

Recurrence in the RMT is different compared to the Transformer-XL because the former stores only m memory vectors per segment. On the other hand, the Transformer-XL stores $m \times N$ vectors per segment. Also, in the RMT model memory representations from the previous segment are processed by Transformer layers together with the current segment tokens. This makes memory part of RMT effectively deeper in a number of applied Transformer layers $\tau \times N$. Additionally, we allow all memory tokens in the read/write block to access all other tokens in the same block. The causal attention mask is applied only to tokens of the input sequence (Figure $\boxed{6}$(d)).

We train the RMT with Backpropagation Through Time (BPTT). During backward pass, unlike in Transformer-XL, memory gradients are not stopped between segments. The number of previous

Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Related Work: Transformer-XL - Potential drawbacks

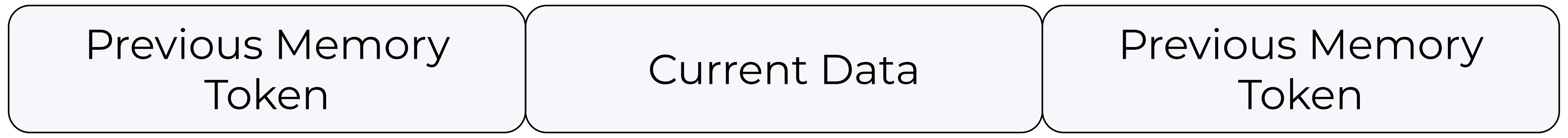
Recurrence in the RMT is different compared to the Transformer-XL because the former stores only m memory vectors per segment. On the other hand, the Transformer-XL stores $m \times N$ vectors per segment. Also, in the RMT model memory representations from the previous segment are processed by Transformer layers together with the current segment tokens. This makes memory part of RMT effectively deeper in a number of applied Transformer layers $\tau \times N$. Additionally, we allow all memory tokens in the read/write block to access all other tokens in the same block. The causal attention mask is applied only to tokens of the input sequence (Figure 6(d)).

We train the RMT with Backpropagation Through Time (BPTT). During backward pass, unlike in Transformer-XL, memory gradients are not stopped between segments. The number of previous

Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

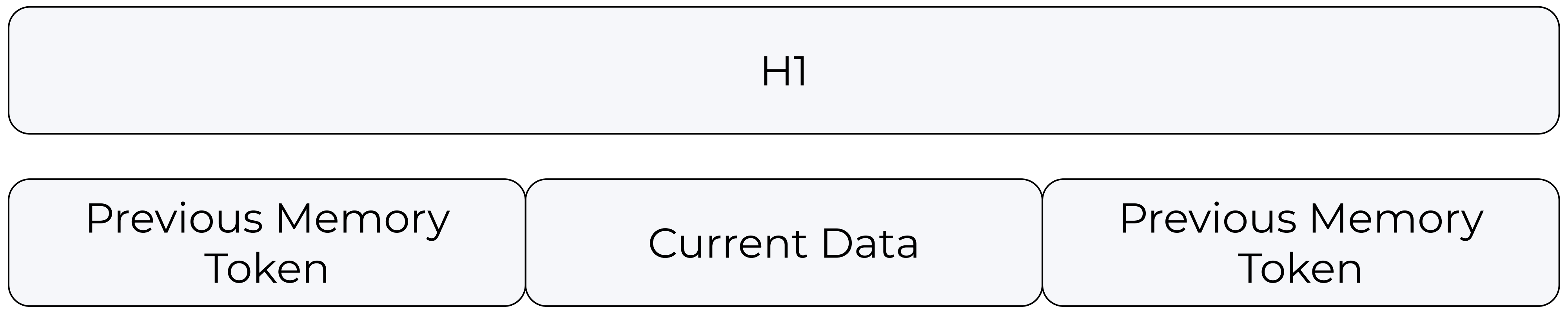
Current Data

Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)



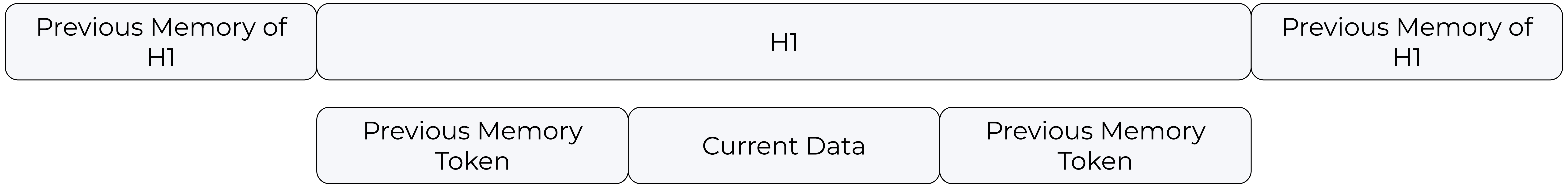
Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Methodology



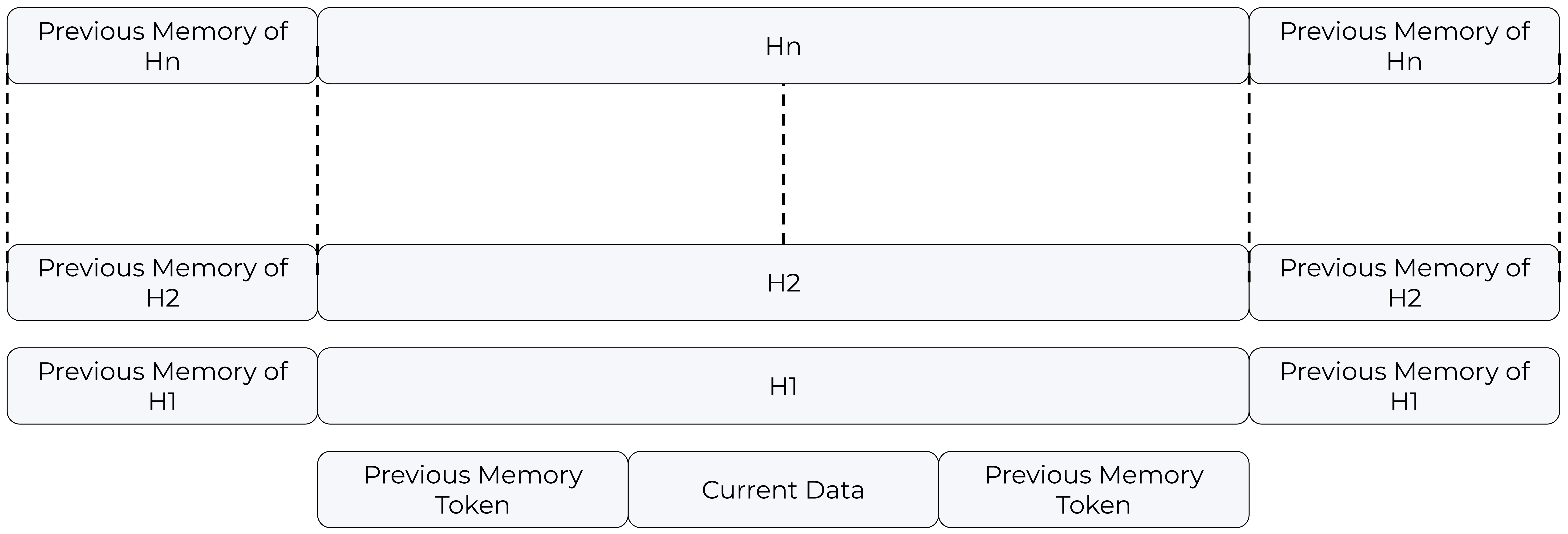
Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Methodology



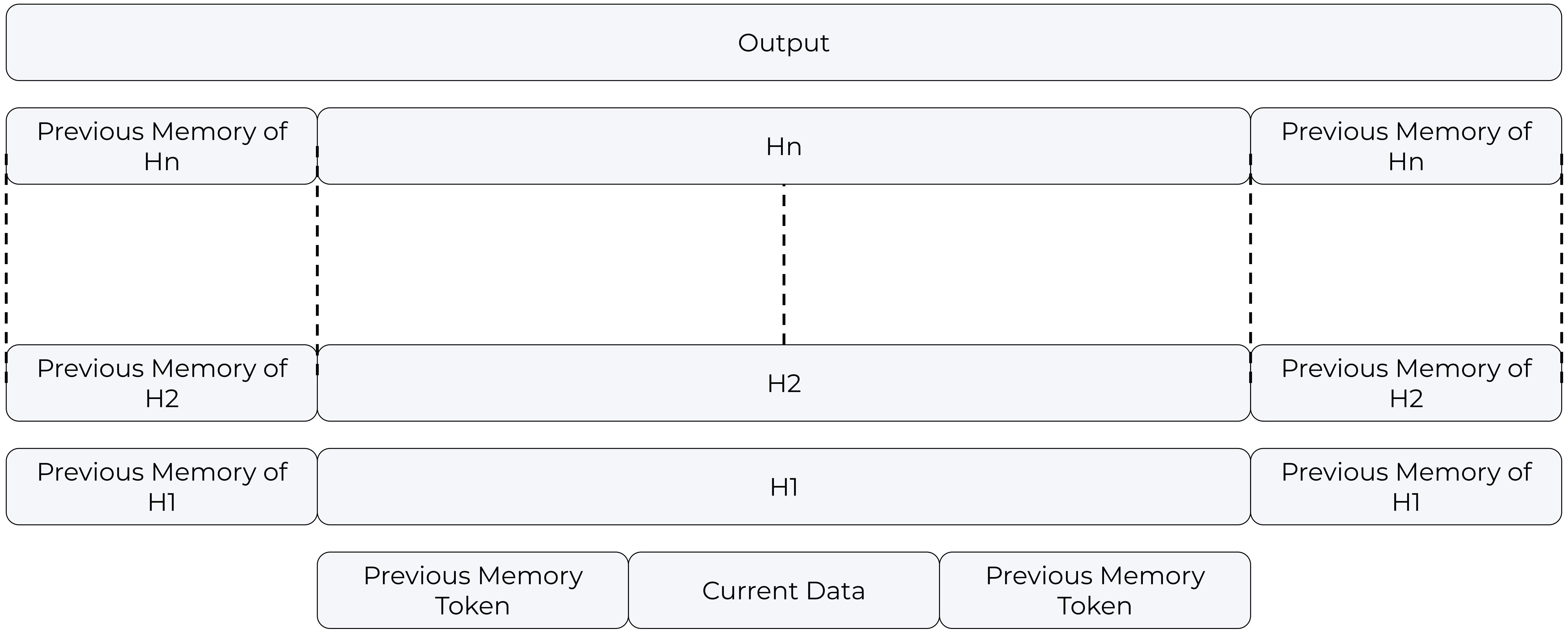
Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Methodology



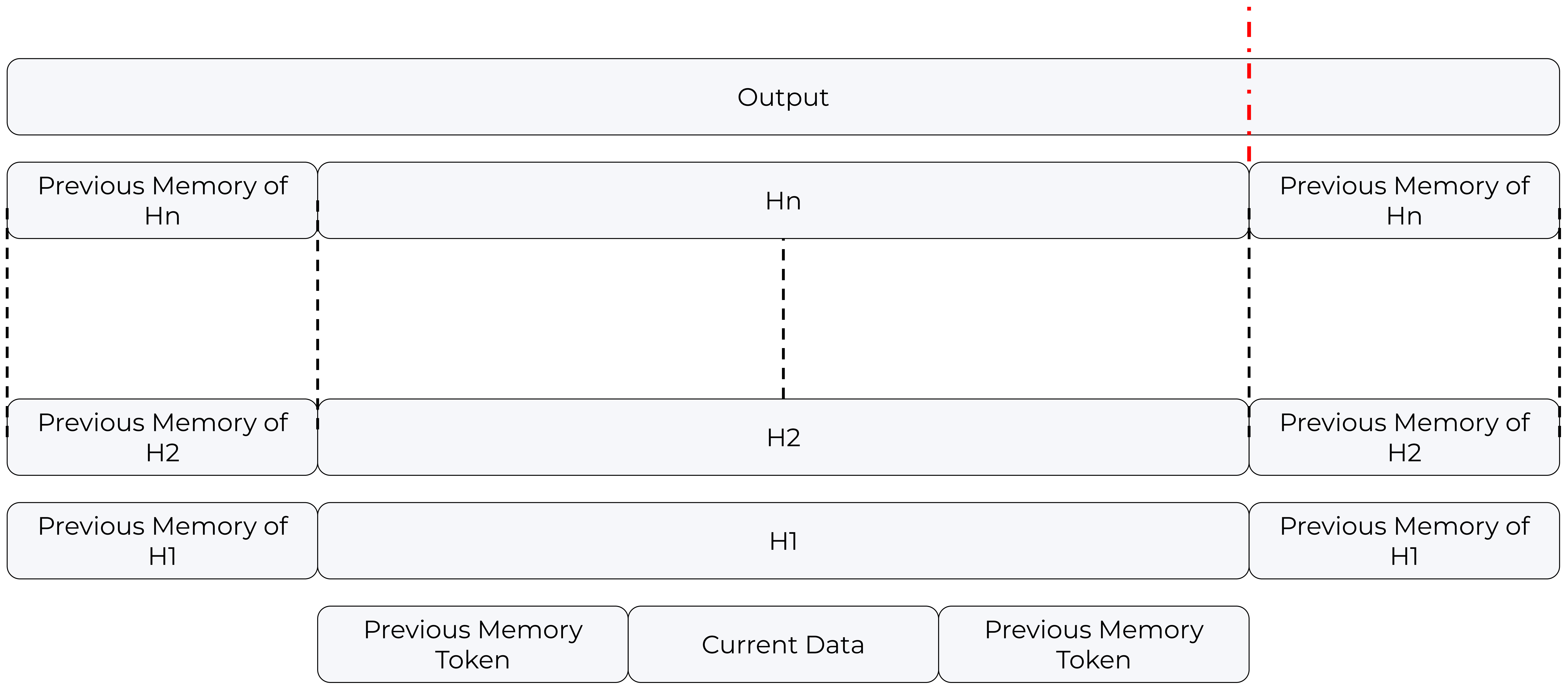
Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Methodology

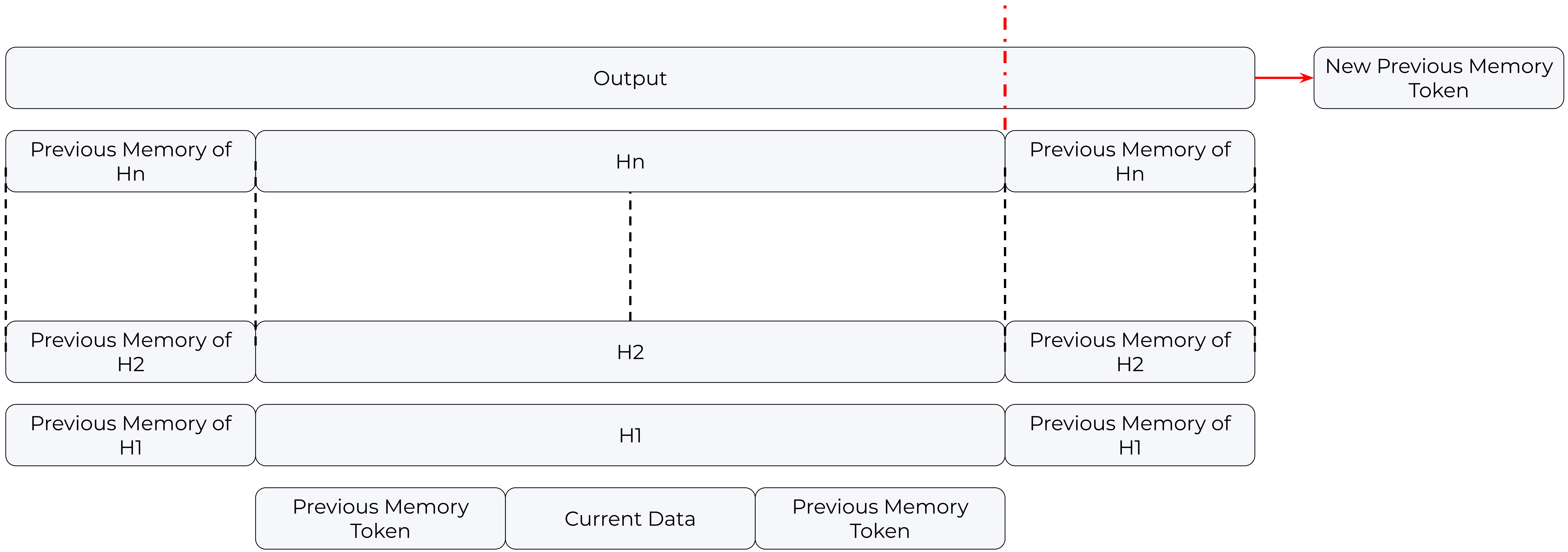


Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Methodology

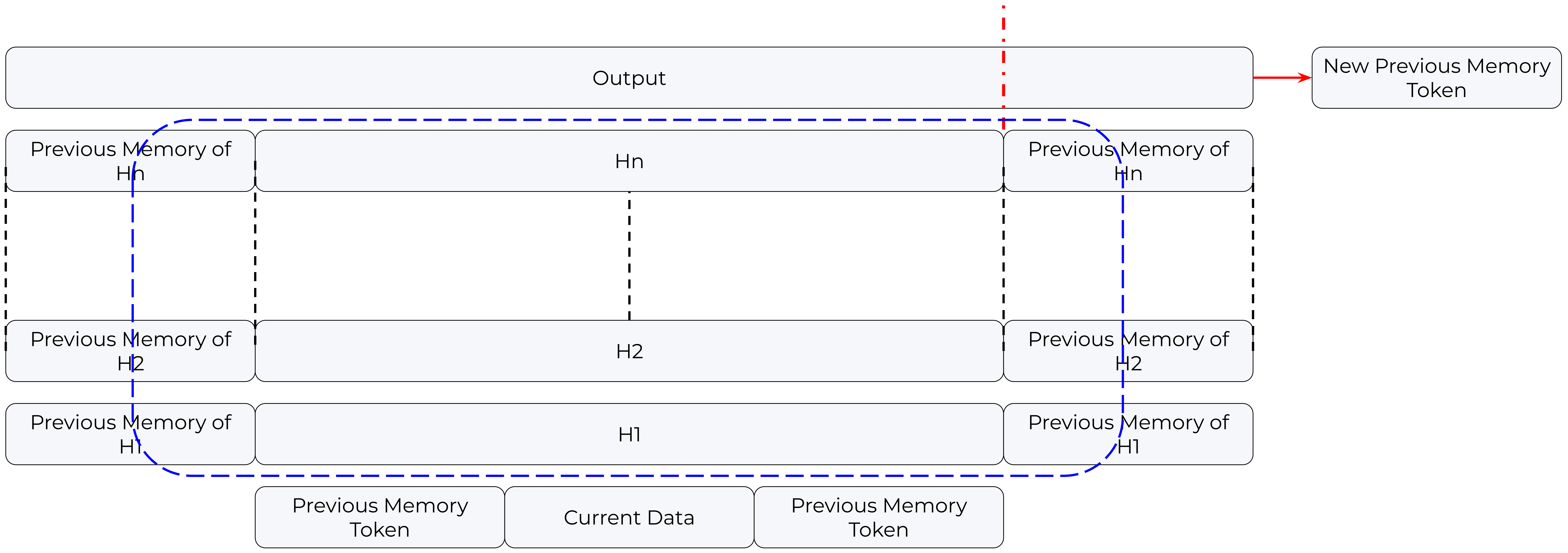


Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)



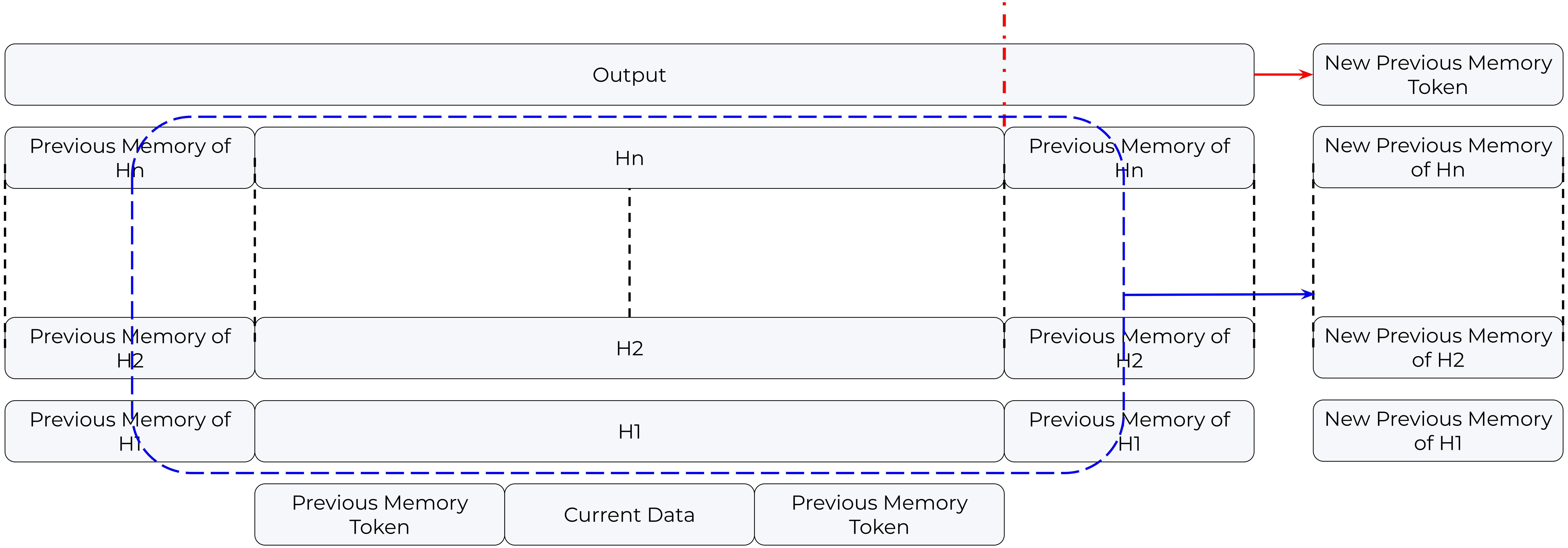
Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Methodology



Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Methodology

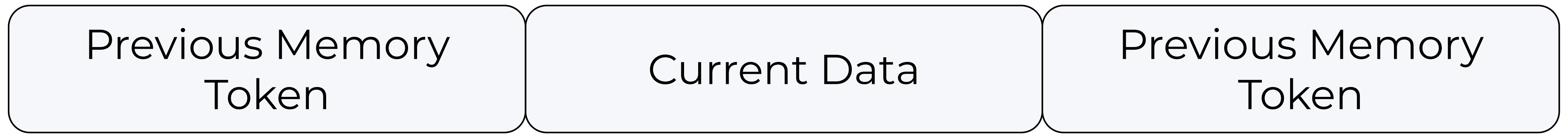


Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Methodology

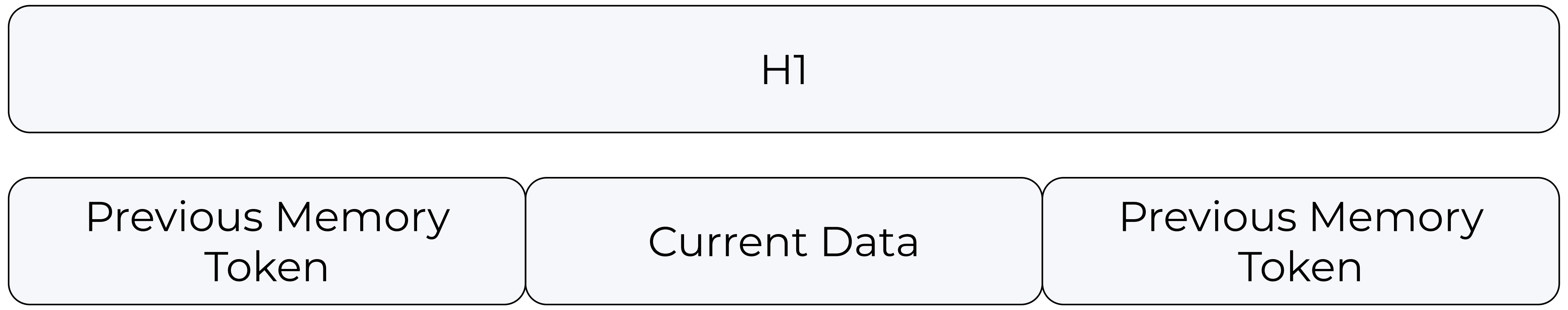
Current Data

Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)



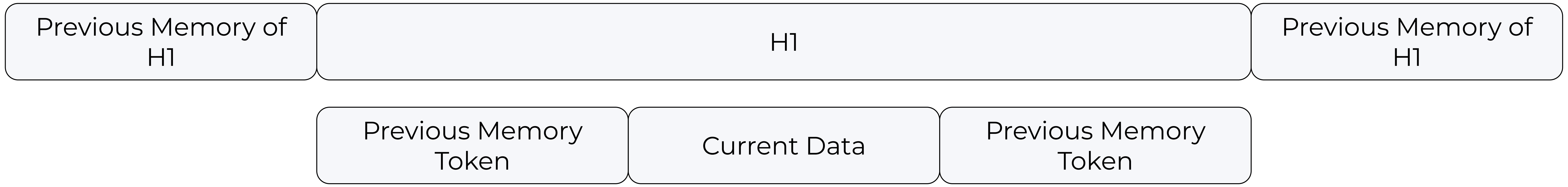
Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Methodology

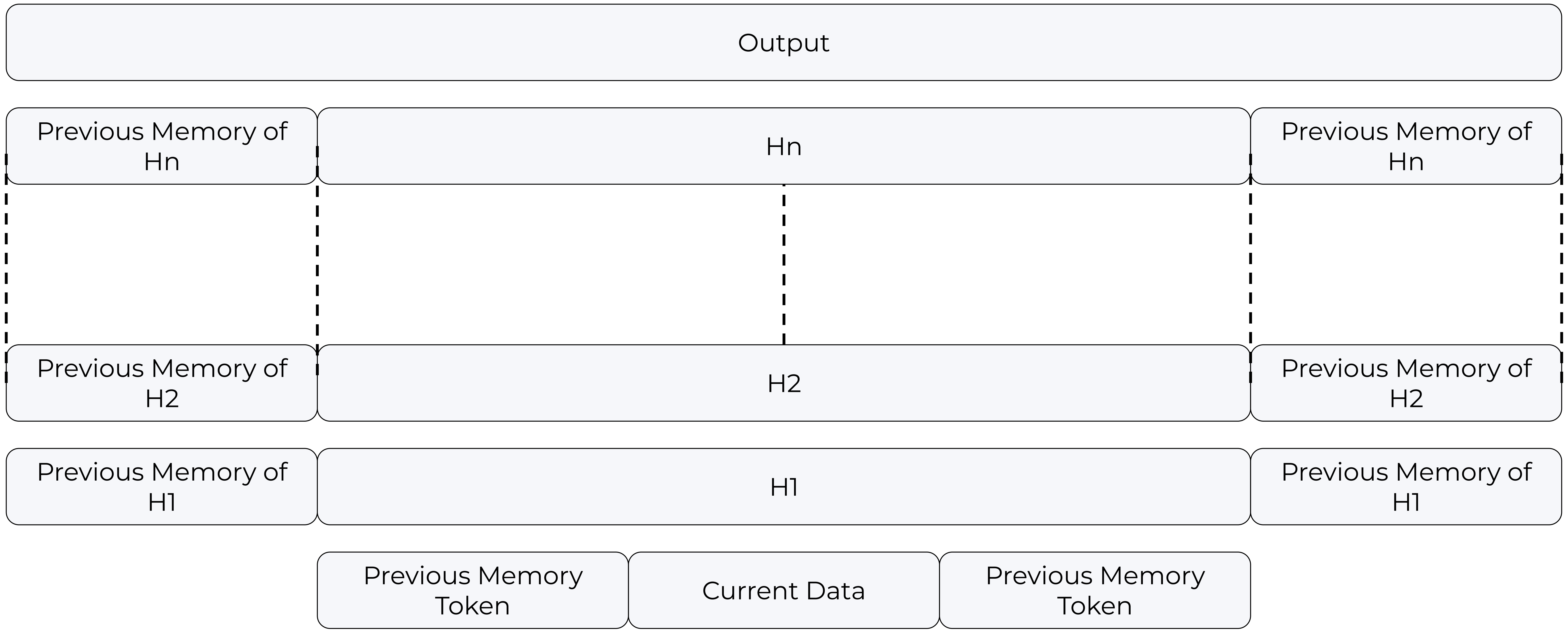


Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

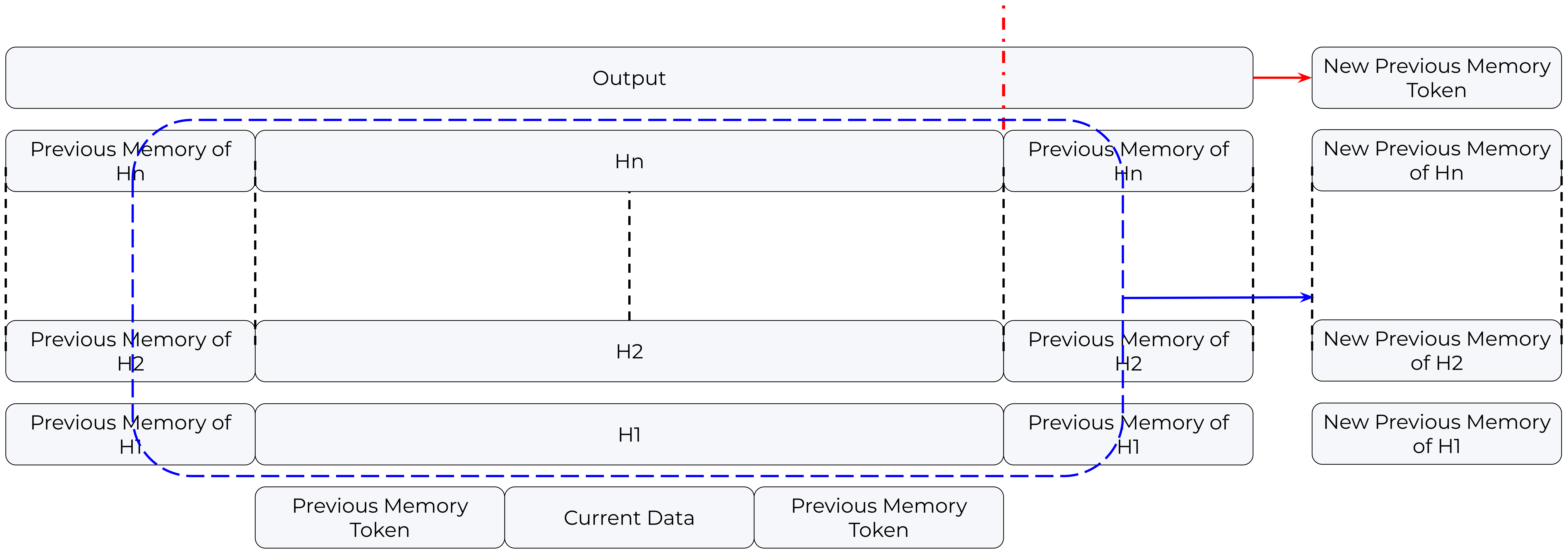
Methodology



Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

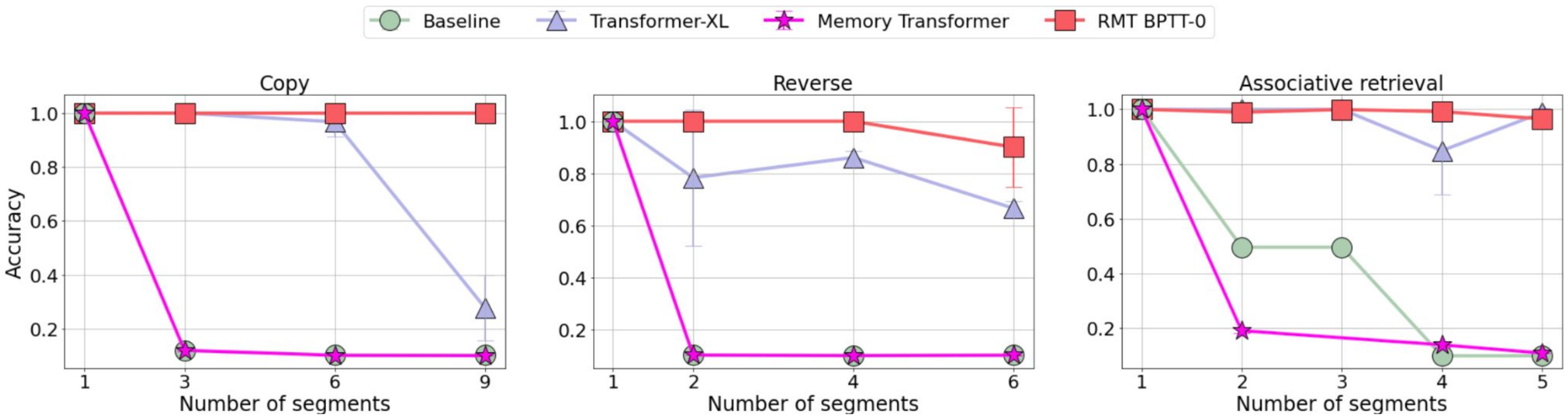


Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

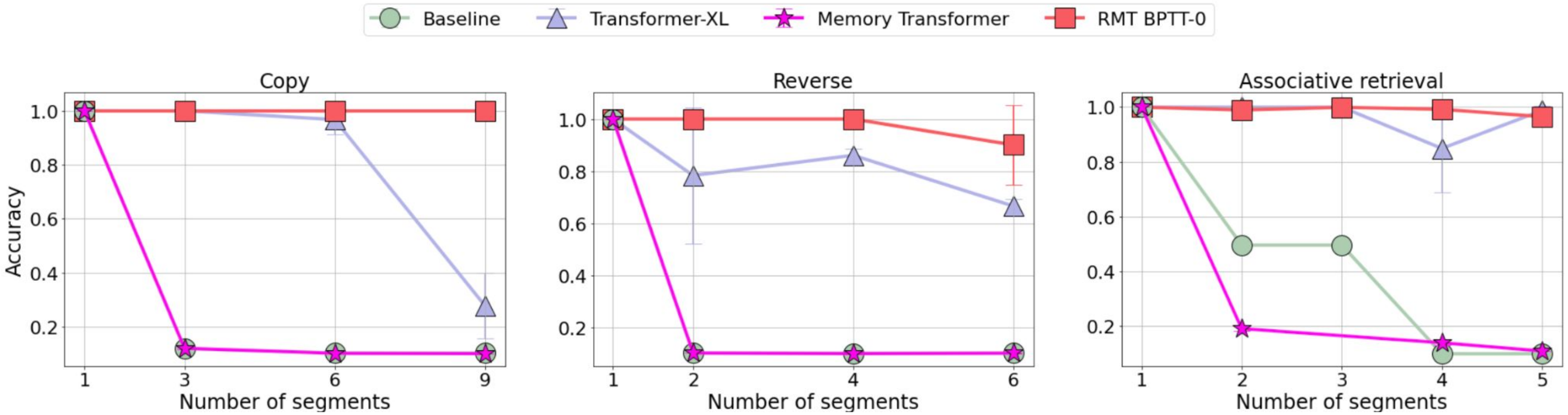


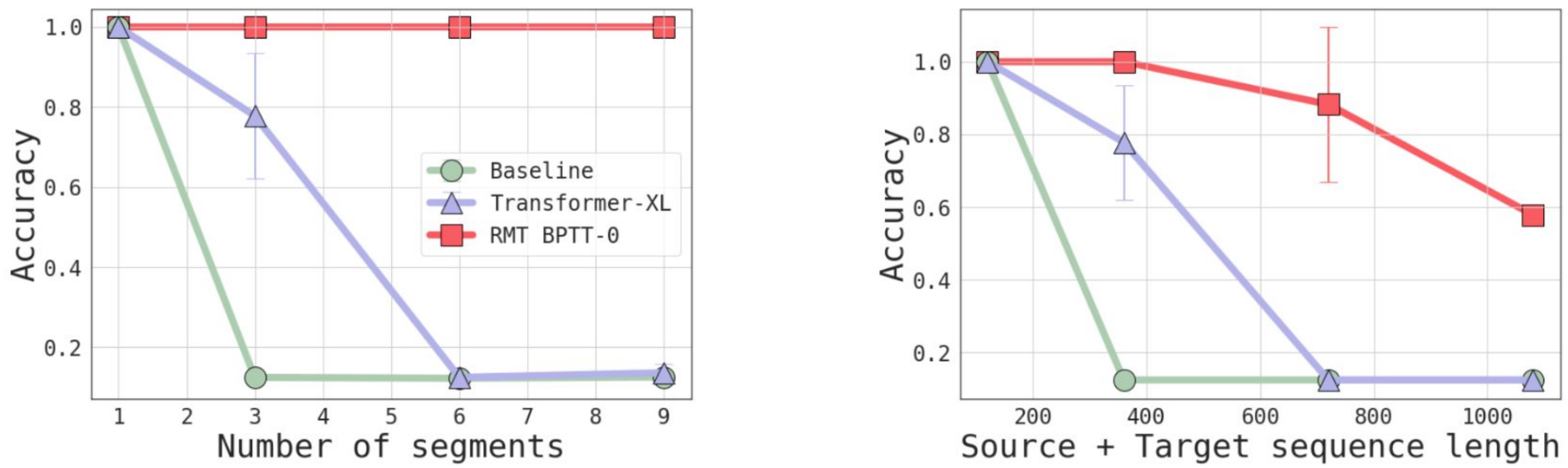
Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Results



Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)





Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

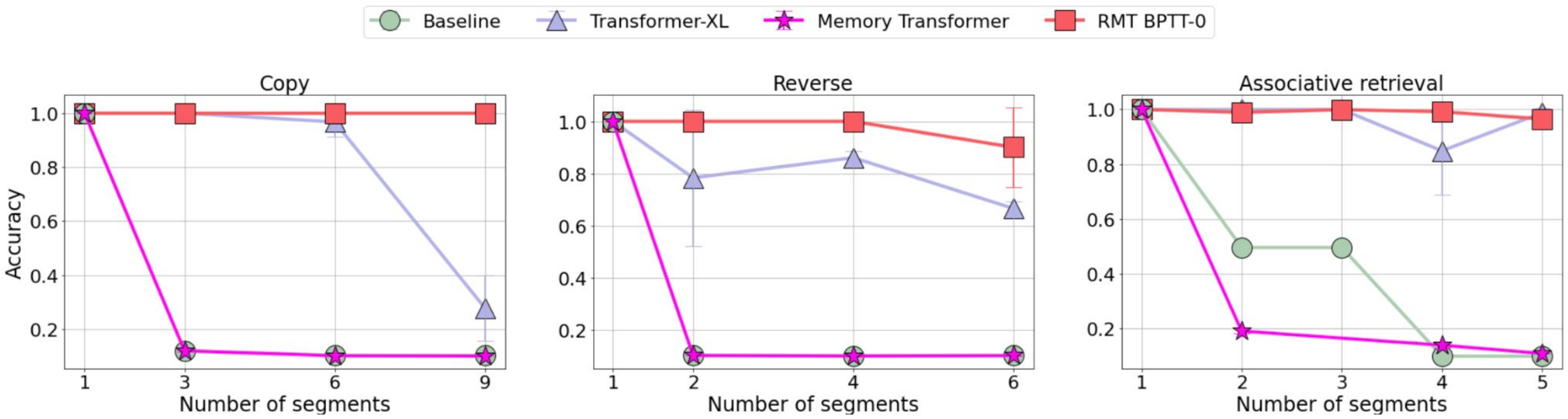
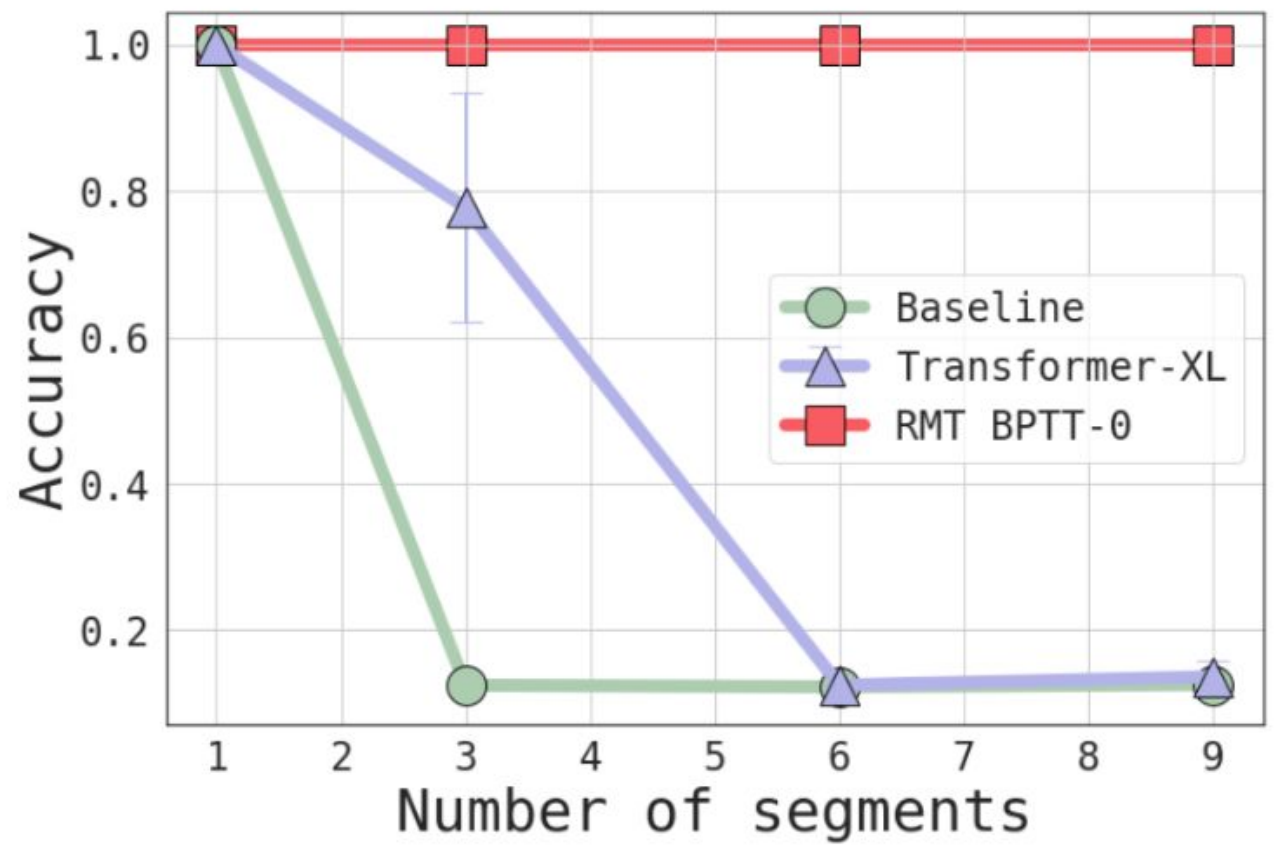
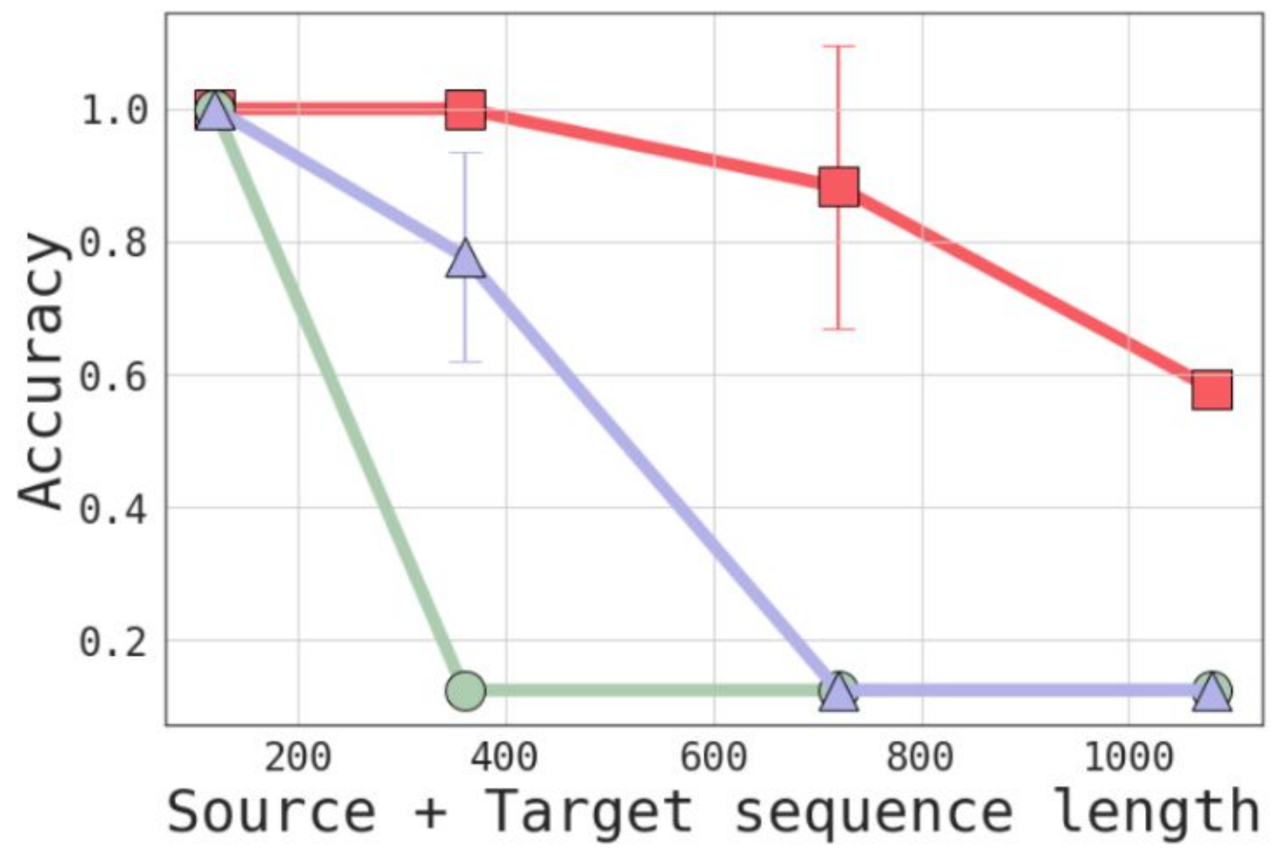


Table 1: **Quadratic equations task.** Sequence of 180 tokens consists of quadratic equation, a solution, and an answer. It is split into a number of segments with an answer in the last segment. Accuracy equals 1.0 if the full answer is predicted correctly.

MODEL	MEMORY	SEGMENTS	ACC$_{\pm \text{STD}}$
BASELINE	0	1	0.99 ± 0.01
TRANSFORMER-XL	30	6	0.93 ± 0.02
RMT	30	6	0.99 ± 0.002





Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

Table 2: **Language modeling on WikiText-103.** Average perplexity for the best performed variations of RMT models reported (see full results in Appendix A.5). Underlined values show Tr-XL and RMT models with close results. RMT models with smaller memory sizes achieve similar scores to Tr-XL models with larger memory. Combination of cache with recurrent memory (Tr-XL + RMT) shows the best performance.

MODEL	MEMORY	SEGMENT LEN	PPL$\pm$STD
TR-XL (PAPER)	150	150	24.0
BASELINE	0	150	29.95 $\pm$ 0.15
MEMTR	10	150	29.63 $\pm$ 0.06
TR-XL (OURS)	150	150	24.12 $\pm$ 0.05
TR-XL	25	150	25.57 $\pm$ 0.02
TR-XL	75	150	24.68 $\pm$ 0.01
RMT BPTT-3	10	150	25.04 $\pm$ 0.07
RMT BPTT-2	25	150	24.85 $\pm$ 0.31
TR-XL + RMT	75+5	150	24.47 $\pm$ 0.05
TR-XL + RMT	150+10	150	**23.99** $\pm$ 0.09
BASELINE	0	50	39.05 $\pm$ 0.01
TR-XL	100	50	**25.66** $\pm$ 0.01
TR-XL	50	50	26.54 $\pm$ 0.01
TR-XL	25	50	27.57 $\pm$ 0.09
TR-XL	10	50	28.98 $\pm$ 0.11
RMT BPTT-1	1	50	28.71 $\pm$ 0.03
RMT BPTT-3	10	50	26.37 $\pm$ 0.01

Table 3: **Hyperpartisan news detection.** Models starting with RMT are taken from HuggingFace Transformers and augmented with 10 memory tokens and recurrence before fine-tuning. Train/valid/test split as in (Beltagy et al., 2020) and metric is F1.

MODEL [INPUT SIZE]	NUMBER OF SEGMENTS 1	2	3	4
BIG BIRD [4096] (ZAHEER ET AL., 2020)	92.20			
LONGFORMER [4096] (BELTAGY ET AL., 2020)	94.80			
GRAPH-ROBERTA [512X100] (XU ET AL., 2021)	96.15			
ERNIE-DOC-LARGE [640] (DING ET AL., 2021)	96.60			
ERNIE-SPARSE [4096] (LIU ET AL., 2022)	92.81			
RMT BERT-BASE-CASE [512]	91.60	94.12	93.06	94.34
RMT ROBERTA-BASE [512]	94.87	**97.20**	**96.72**	**98.11**
RMT DEBERTA-V3-BASE [512]	94.17	96.78	94.80	94.80
RMT T5-BASE [512]	**94.99**	95.32	96.12	97.20

Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)

My Thoughts

- I felt like there was not enough motivation provided in the paper.

My Thoughts

- I felt like there was not enough motivation provided in the paper.
    - They assume that this is well established.
    - I had to dig into several previous papers to get the context.
    - Easy to read only for that specific community members.

My Thoughts

- I felt like there was not enough motivation provided in the paper.
  - They assume that this is well established.
  - I had to dig into several previous papers to get the context.
  - Easy to read only for that specific community members.
- The implementation is not explained well.

My Thoughts

- I felt like there was not enough motivation provided in the paper.
  - They assume that this is well established.
  - I had to dig into several previous papers to get the context.
  - Easy to read only for that specific community members.
- The implementation is not explained well.
  - I had to look into the code to understand it.
  - Dedicated only 2 paragraphs for it.
  - The equations provided in the paper does not convey the method properly.
  - They claim they don't do SG but they still do it in the code anyway.

My Thoughts

- I felt like there was not enough motivation provided in the paper.
  - They assume that this is well established.
  - I had to dig into several previous papers to get the context.
  - Easy to read only for that specific community members.
- The implementation is not explained well.
  - I had to look into the code to understand it.
  - Dedicated only 2 paragraphs for it.
  - The equations provided in the paper does not convey the method properly.
  - They claim they don't do SG but they still do it in the code anyway.
- RMT feels like step backwards from T-XL.

My Thoughts

- I felt like there was not enough motivation provided in the paper.
  - They assume that this is well established.
  - I had to dig into several previous papers to get the context.
  - Easy to read only for that specific community members.
- The implementation is not explained well.
  - I had to look into the code to understand it.
  - Dedicated only 2 paragraphs for it.
  - The equations provided in the paper does not convey the method properly.
  - They claim they don't do SG but they still do it in the code anyway.
- RMT feels like step backwards from T-XL.
- Should have compared with Memory Transformer paper instead of T-XL.

My Thoughts

- I felt like there was not enough motivation provided in the paper.
  - They assume that this is well established.
  - I had to dig into several previous papers to get the context.
  - Easy to read only for that specific community members.
- The implementation is not explained well.
  - I had to look into the code to understand it.
  - Dedicated only 2 paragraphs for it.
  - The equations provided in the paper does not convey the method properly.
  - They claim they don't do SG but they still do it in the code anyway.
- RMT feels like step backwards from T-XL.
- Should have compared with Memory Transformer paper instead of T-XL.

[−] **Official Review of Paper11788 by Reviewer ien2**

NeurIPS 2022 Conference Paper11788 Reviewer ien2

11 Jul 2022 (modified: 01 Aug 2022) NeurIPS 2022 Conference Paper11788 Official Review Readers: Everyone

Rating: 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

[−] **Official Review of Paper11788 by Reviewer VBgN**

NeurIPS 2022 Conference Paper11788 Reviewer VBgN

10 Jul 2022 NeurIPS 2022 Conference Paper11788 Official Review Readers: Everyone

Rating: 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

[−] **Official Review of Paper11788 by Reviewer hj1Y**

NeurIPS 2022 Conference Paper11788 Reviewer hj1Y

09 Jul 2022 (modified: 09 Jul 2022) NeurIPS 2022 Conference Paper11788 Official Review Readers: Everyone

Rating: 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

My Thoughts

- I felt like there was not enough motivation provided in the paper.
  - They assume that this is well established.
  - I had to dig into several previous papers to get the context.
  - Easy to read only for that specific community members.
- The implementation is not explained well.
  - I had to look into the code to understand it.
  - Dedicated only 2 paragraphs for it.
  - The equations provided in the paper does not convey the method properly.
  - They claim they don't do SG but they still do it in the code anyway.
- RMT feels like step backwards from T-XL.
- Should have compared with Memory Transformer paper instead of T-XL.
- Potential future directions

My Thoughts

- I felt like there was not enough motivation provided in the paper.
  - They assume that this is well established.
  - I had to dig into several previous papers to get the context.
  - Easy to read only for that specific community members.
- The implementation is not explained well.
  - I had to look into the code to understand it.
  - Dedicated only 2 paragraphs for it.
  - The equations provided in the paper does not convey the method properly.
  - They claim they don't do SG but they still do it in the code anyway.
- RMT feels like step backwards from T-XL.
- Should have compared with Memory Transformer paper instead of T-XL.
- Potential future directions
  - What does long term memory mean in terms ViTs?

My Thoughts

- I felt like there was not enough motivation provided in the paper.
  - They assume that this is well established.
  - I had to dig into several previous papers to get the context.
  - Easy to read only for that specific community members.
- The implementation is not explained well.
  - I had to look into the code to understand it.
  - Dedicated only 2 paragraphs for it.
  - The equations provided in the paper does not convey the method properly.
  - They claim they don't do SG but they still do it in the code anyway.
- RMT feels like step backwards from T-XL.
- Should have compared with Memory Transformer paper instead of T-XL.
- Potential future directions
  - What does long term memory mean in terms ViTs?
  - Can we add this to video transformers in some way?
    - Some synthetic tasks could be maze solving.

My Thoughts

- I felt like there was not enough motivation provided in the paper.
  - They assume that this is well established.
  - I had to dig into several previous papers to get the context.
  - Easy to read only for that specific community members.
- The implementation is not explained well.
  - I had to look into the code to understand it.
  - Dedicated only 2 paragraphs for it.
  - The equations provided in the paper does not convey the method properly.
  - They claim they don't do SG but they still do it in the code anyway.
- RMT feels like step backwards from T-XL.
- Should have compared with Memory Transformer paper instead of T-XL.
- Potential future directions
  - What does long term memory mean in terms ViTs?
  - Can we add this to video transformers in some way?
    - Some synthetic tasks could be maze solving.

Questions?

Citations

- "Understanding LSTM Networks." *Understanding LSTM Networks -- Colah's Blog*, https://colah.github.io/posts/2015-08-Understanding-LSTMs/ . (Blog)
- Alammar, J. (n.d.). The illustrated transformer. The Illustrated Transformer – Jay Alammar – Visualizing machine learning one concept at a time. Retrieved February 22, 2023, from https://jalammar.github.io/illustrated-transformer/ (Blog)
- Beltagy, Iz, Matthew E. Peters, and Arman Cohan. "Longformer: The Long-Document Transformer." (Arxiv 2020)
- Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)
- Dai, Zihang, et al. "Transformer-XL: Attentive Language Models beyond a Fixed-Length Context." Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. (ACL 2019)
- Burtsev, Mikhail S., et al. "Memory transformer." arXiv preprint arXiv:2006.11527 (2020).

Why the paper is the way it is?

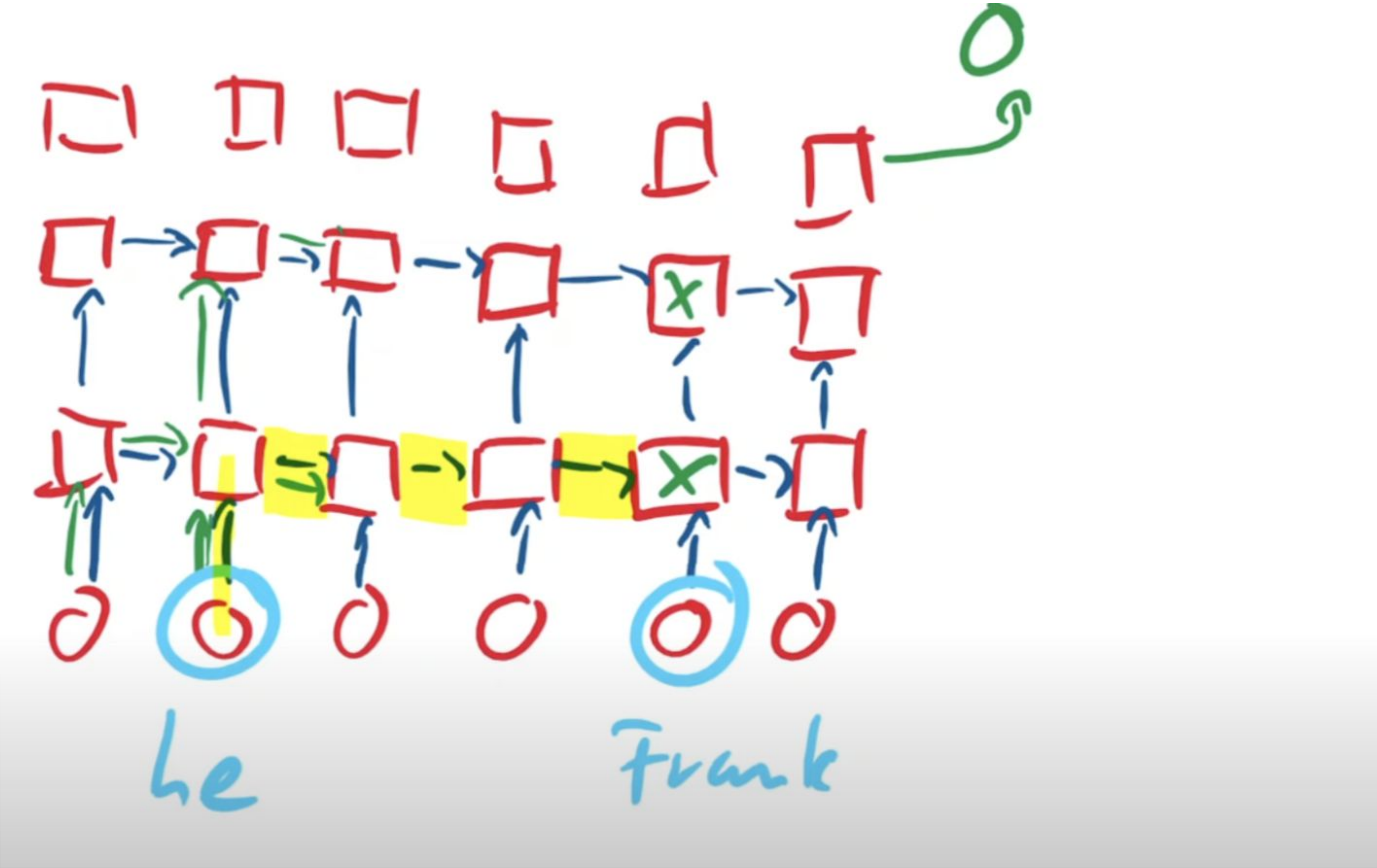
- Abstract : This is like short glimpse, it is meant to attract the target audience as fast and precise as possible. If you dont get this right, the probability of your paper reaching a wider audience is very low as most people wont even read the paper.
  - Start with the main task at hand
  - Point out the issues
  - Say that our proposed method solves
  - (optionally) short explanation on how it solves the problem
  - Say that the proposed method outperform existing latest method on a, b, c datasets by x%.
  - What additional things do you bring to the table
    - Conduct detailed analysis
    - Reasoning for why it works
    - etc

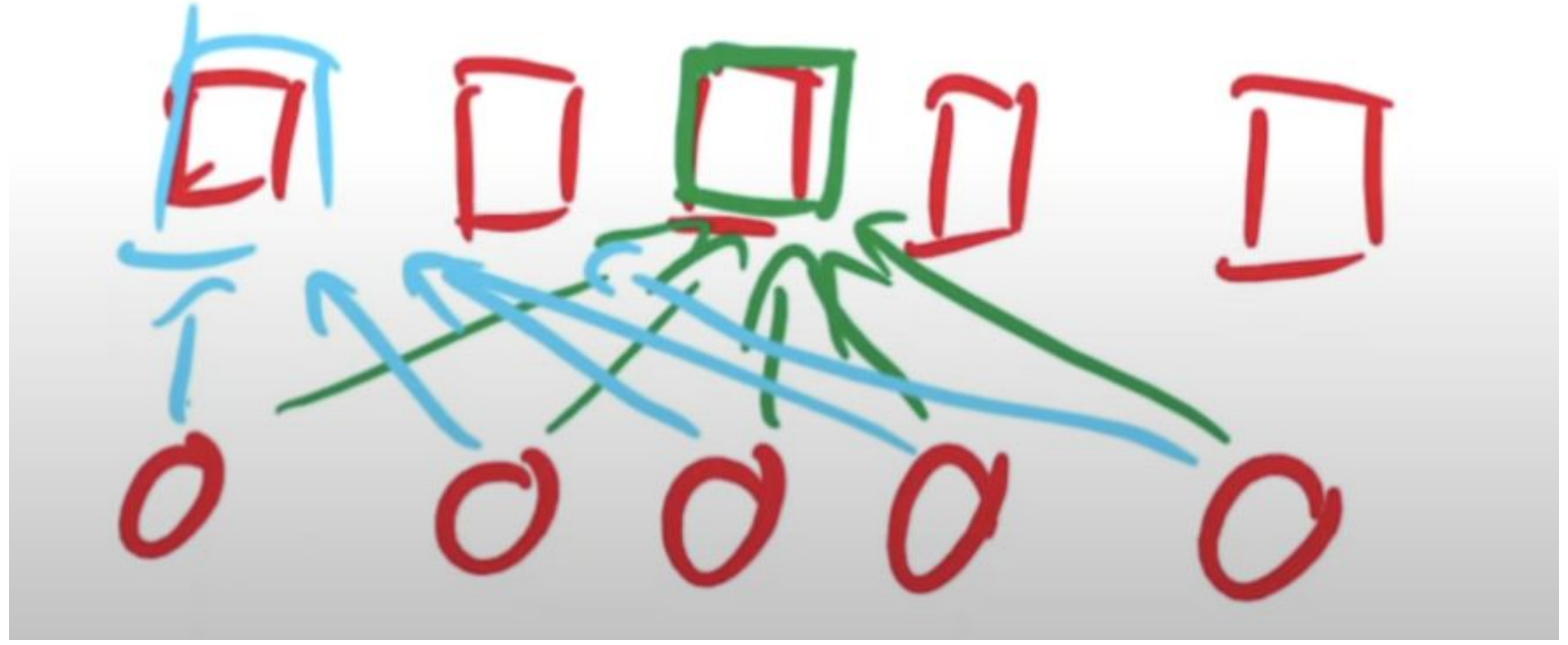
- Missing

- Introduction : Provide context for the new readers. prime the readers for what to come in the paper
  - Re-iterate on the task at hand but with a bit more detail
  - Talk a bit more about the issue
  - Talk about how you tackle it with you method and explain your method a bit more detailed than earlier.
  - Point out key contributions (Generally under a subsubheading)
- Related works : Inform the readers on the existing directions in the field, are you gonna iterate on a existing direction or going to do something orthogonal
  - Introduce the seminal papers in the field, how various papers have tackled this issue that your are trying to solve.
  - If possible, categorize the strategies
  - Now mentioned the methods that are very close to yours and explain how you stand apart from them and what novelty do you bring in when compared to them.

Why the paper is the way it is?

- Methodology : This is the part that should go indepth into you method. If anyone wants to understand your method end to end this is where they will come. Should explain every detail about your method. Dont keep anything for later.
  - Introduce your setting (generally in a mathematical way).
  - Introduce appropriate variabele necessary for explaining your method.
  - Explain the method. Full working details.
  - Sometime a much more elaborate version is provided in the appendix
  - Sometime will include a pseudo code for the method
- Experiments : details About how you conducted the experiments
  - Introduce all the datasets used in details
    - How you decide train, val, test.
    - Cite if you are using a already existing setup
  - Explain how the training and inference are done.
  - How many GPUs are used, batch size, all the hyperparameters necessary for implementing your method from scratch
- Results : Show, discuss, compare
  - Show results of your experiments in the form of tables and graphs and compare the results with other existing methods.
  - Point out the nuances in the results. Say sometime an unexpected existing method outperforms your method. Speculate why it is like that and provide possible reasons.
  - If available provide ablations of important hyperparameters
- Conclusion : Structured results
  - Most of the time, the conclusion is made in the results itself.
  - This section is mostly used to rephrase and put strong emphasis on those conclusions
- Future work : Things that we thought of doing but couldn't do it before deadline
  - Generally there will be like a 100 different ideas going through your head when you are trying to make you method work, but you wont be able to try all those out before deadline, should mention those here.
  - Speculate on how the community as a hole will or should proceed in future.
- References


he
Frank



Results

Table 2: **Language modeling on WikiText-103.** Average perplexity for the best performed variations of RMT models reported (see full results in Appendix A.5). Underlined values show Tr-XL and RMT models with close results. RMT models with smaller memory sizes achieve similar scores to Tr-XL models with larger memory. Combination of cache with recurrent memory (Tr-XL + RMT) shows the best performance.

MODEL	MEMORY	SEGMENT LEN	PPL$\pm$STD
TR-XL (PAPER)	150	150	24.0
BASELINE	0	150	29.95 $\pm$ 0.15
MEMTR	10	150	29.63 $\pm$ 0.06
TR-XL (OURS)	150	150	24.12 $\pm$ 0.05
TR-XL	25	150	25.57 $\pm$ 0.02
TR-XL	75	150	24.68 $\pm$ 0.01
RMT BPTT-3	10	150	25.04 $\pm$ 0.07
RMT BPTT-2	25	150	24.85 $\pm$ 0.31
TR-XL + RMT	75+5	150	24.47 $\pm$ 0.05
TR-XL + RMT	150+10	150	**23.99** $\pm$ 0.09
BASELINE	0	50	39.05 $\pm$ 0.01
TR-XL	100	50	**25.66** $\pm$ 0.01
TR-XL	50	50	26.54 $\pm$ 0.01
TR-XL	25	50	27.57 $\pm$ 0.09
TR-XL	10	50	28.98 $\pm$ 0.11
RMT BPTT-1	1	50	28.71 $\pm$ 0.03
RMT BPTT-3	10	50	26.37 $\pm$ 0.01

Table 4: Test set bits-per-character on enwik8. Our experimental setup shows similar scores to the original paper (Dai et al., 2019) with segment length 512.

MODEL	MEMORY	SEGMENT LEN	BPC$\pm$STD
TR-XL (DAI ET AL., 2019)	512	512	1.06
TR-XL (OURS)	512	512	1.071
TR-XL	200	128	1.140
TR-XL	100	128	1.178
TR-XL	75	128	1.196
TR-XL	40	128	1.230 $\pm$ 0.001
TR-XL	20	128	1.261
TR-XL	10	128	1.283 $\pm$ 0.001
RMT BPTT-1	5	128	1.241 $\pm$ 0.002
RMT BPTT-2	5	128	1.231 $\pm$ 0.002
RMT BPTT-1	10	128	1.240 $\pm$ 0.006
RMT BPTT-2	10	128	1.228 $\pm$ 0.003
RMT BPTT-0	20	128	1.301
RMT BPTT-1	20	128	1.229
RMT BPTT-2	20	128	1.222

Bulatov, Aydar, Yuri Kuratov, and Mikhail Burtsev. "Recurrent Memory Transformer." Advances in Neural Information Processing Systems. (NeurIPS 2022)