Coupled Mamba: Enhanced Multi-modal Fusion with Coupled State Space Model

华中科技大学 HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

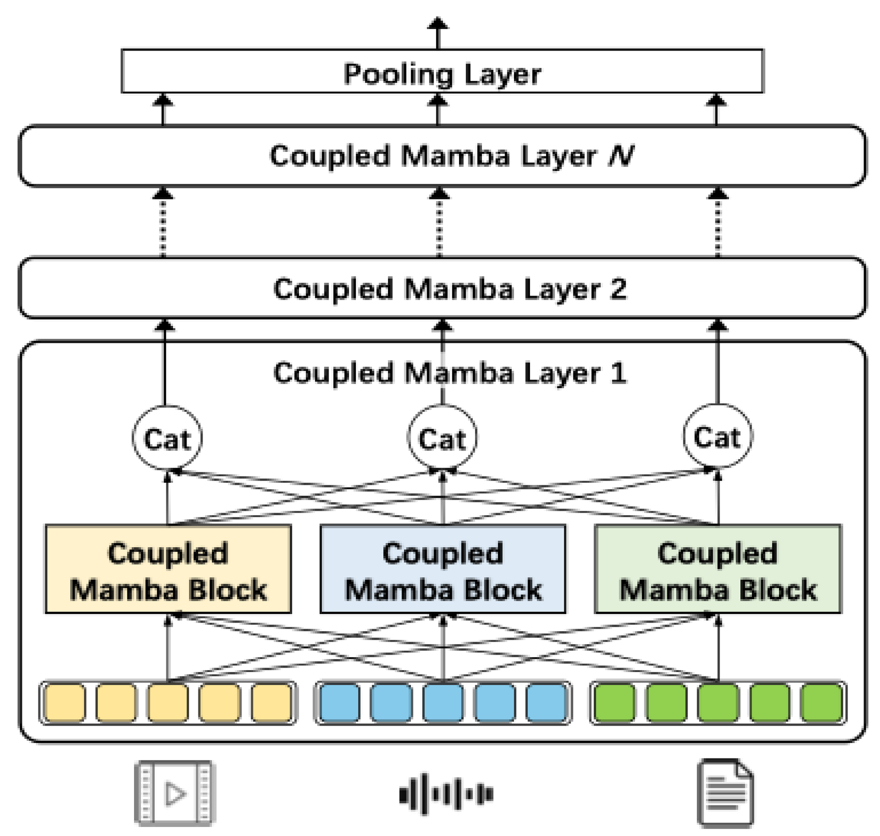
Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, Wei Yang

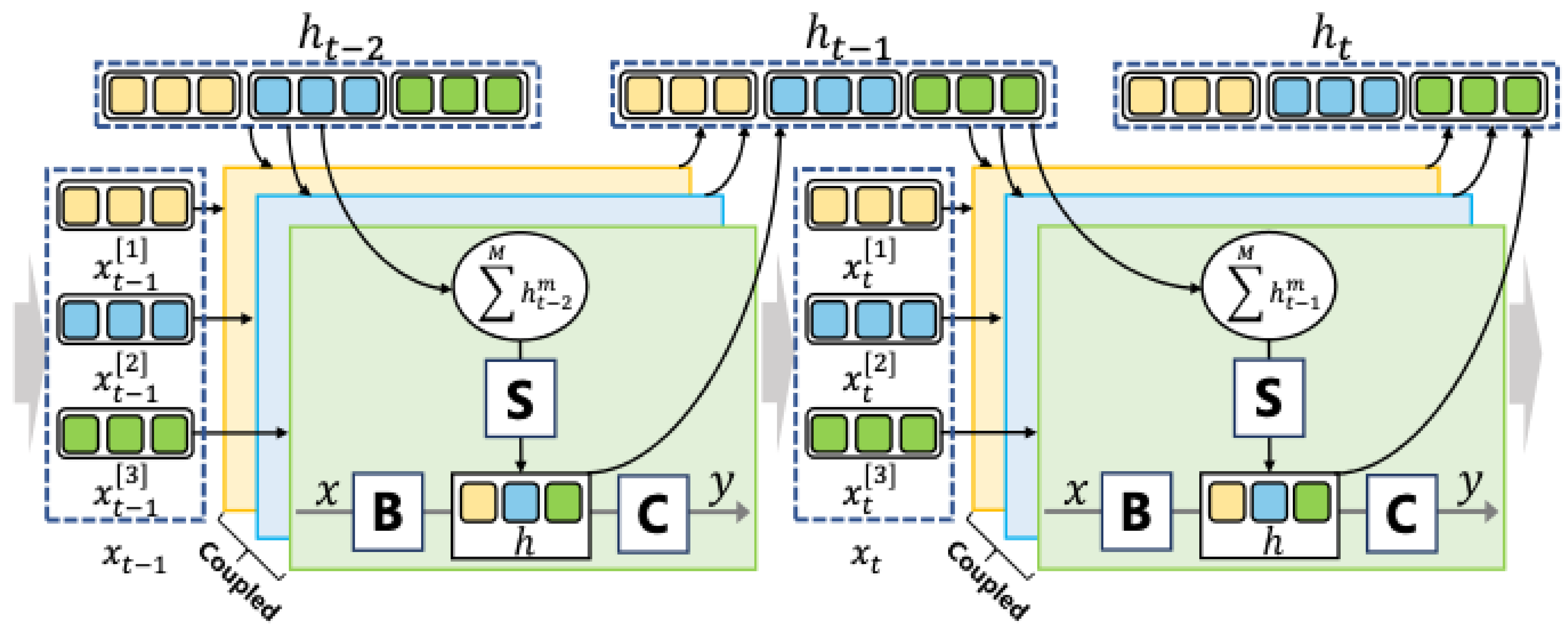
NEURAL INFORMATION PROCESSING SYSTEMS

Abstract

- *Background.* The core of multi-modal fusion is to leverage the complementary information from different modalities. Existing methods often rely on traditional neural architectures, which struggle to capture complex interactions between modalities. Recent advances in State Space Models (SSMs), such as the Mamba model, have shown promise for stronger fusion. Despite this, SSMs face challenges in fusing multiple modalities due to hardware parallelism constraints.
- *Our method*, we propose the Coupled SSM model, which links the state chains of multiple modalities while keeping intra-modality processes independent. Our model includes an inter-modal state transition mechanism, where the current state depends on both its own chain and neighboring chains' states from the previous time step. We also develop a global convolution kernel to support hardware parallelism.
- *Experiment.* Our extensive experiments on the *CMU-MOSEI, CH-SIMS, CH-SIM SV2, MM-IMDB, and BRCA* datasets demonstrate that our proposed method e ffectively enhances multi-modal information integration with faster inference speeds and reduced memory consumption compared to current state-of-the-ar t methods.

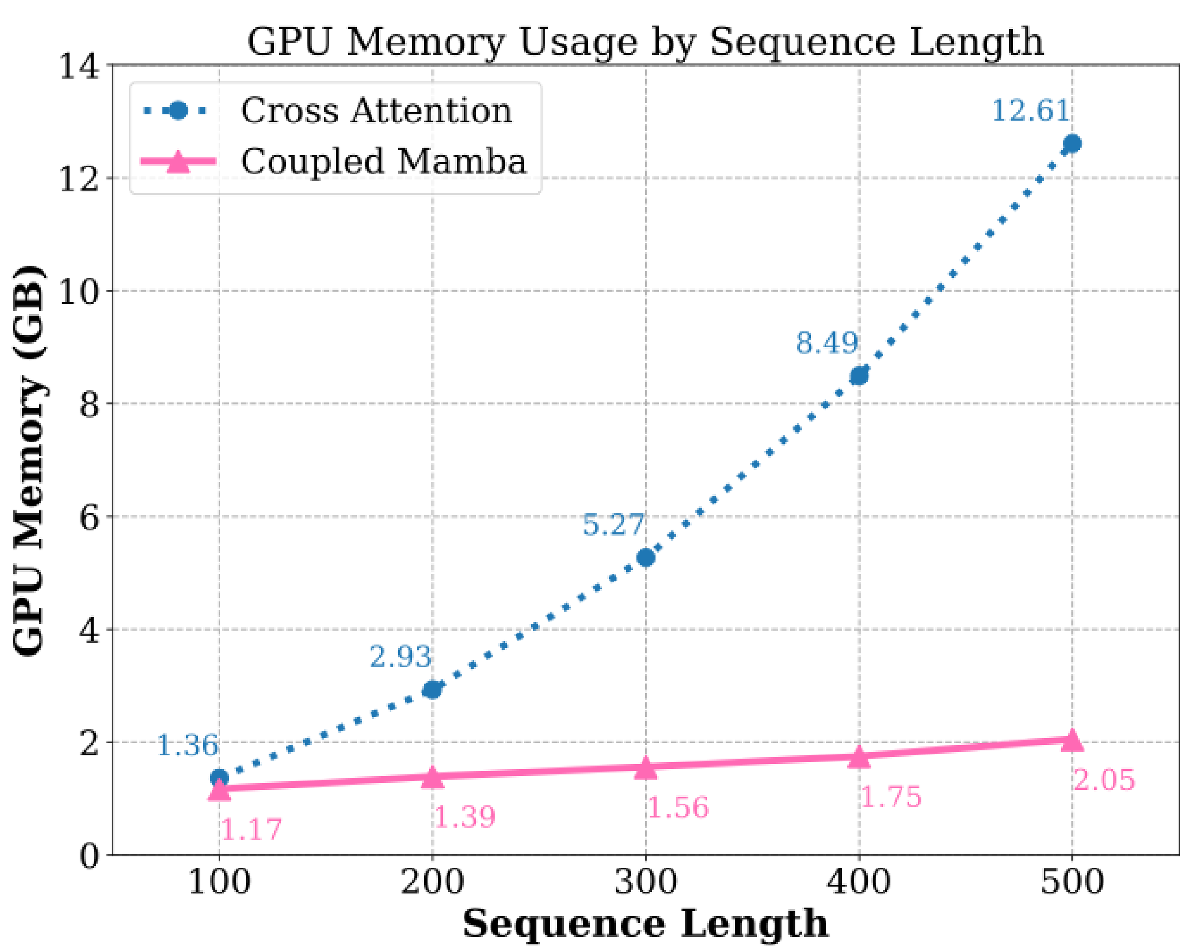
Proposed Architecture





- Coupled Mamba receives input x_{t-1}.
- Performs internal state switching and output through three key parameter matrices, where B,C and S are respectively represented as the input, output and state transfer matrix.
- The hidden states are summed across modalities and used for state transition input to generate next time states. The state is propagated sequentially in time.

Experiment



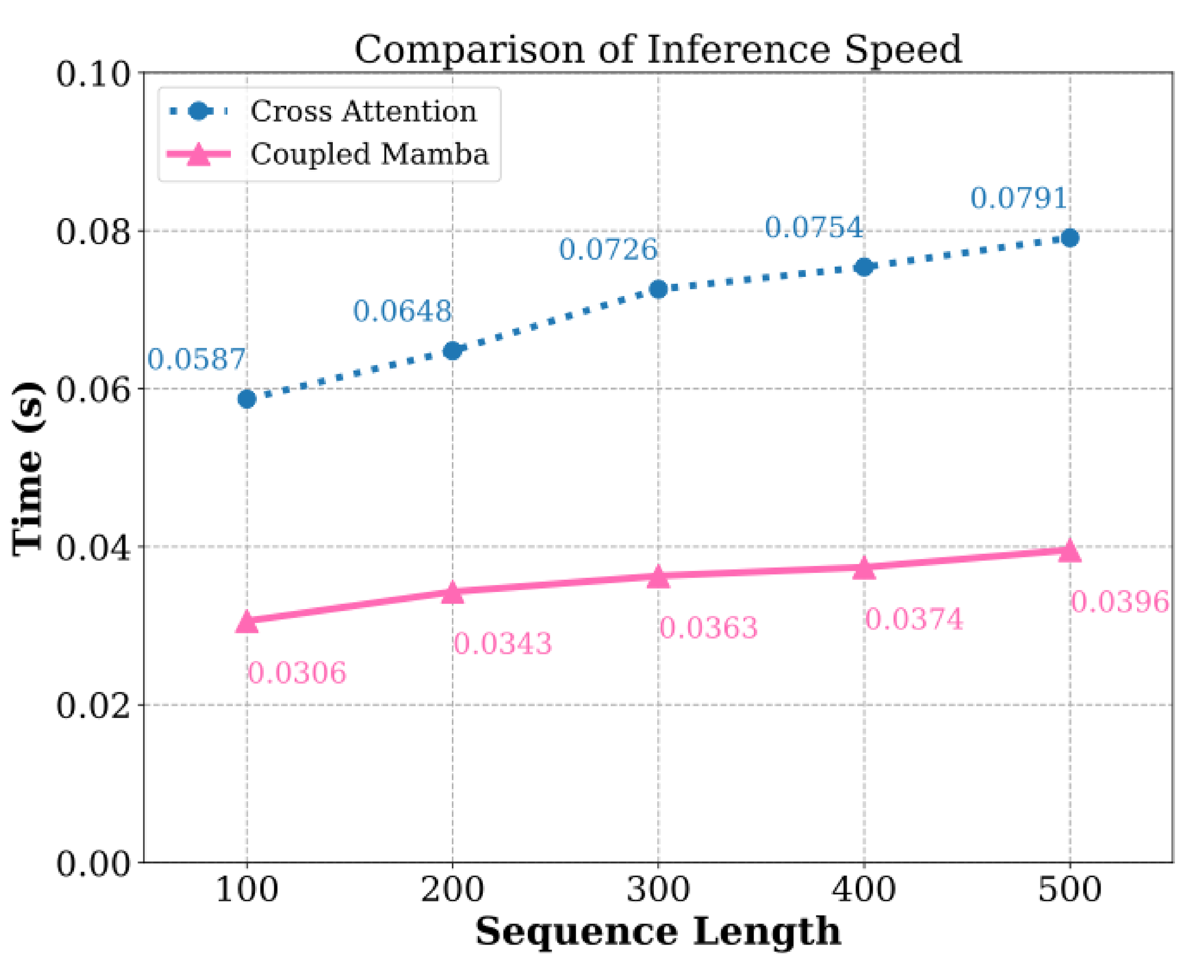


Table 6: Result on the MM-IMDB benchmark. **I** and **T** denote image and text respectively. The best results are in bold.

	Modality	MicroF1(%)↑	MacroF1(%)↑
LRMF [57]	I+T	58.95	50.73
MFM [46]	I+T	56.44	48.53
MI-Matrix [58]	I+T	55.87	46.77
RMFE [59]	I+T	58.67	49.82
CCA [60]	I+T	60.31	50.45
RefNet [61]	I+T	59.45	51.51
DynMM [62]	I+T	60.35	51.60
Coupled Mamba (Ours)	I+T	**62.41**	**52.58**

Table 1: Results on CMU-MOSEI. All models are based on language features extracted by BERT. The one with $*$ indicates that the model reproduces under the same conditions.

Model	CMU-MOSEI				Data Setting
	$MAE\downarrow$	$Corr\uparrow$	$Acc-2\uparrow$	$F1-Score\uparrow$	
TFN [9]	0.593	0.700	82.5	82.1	Unaligned
LMF [30]	0.623	0.677	82.0	82.1	Unaligned
MFN [10]	-	-	76.0	76.0	Aligned
MFM [46]	0.568	0.717	84.4	84.3	Aligned
MulT [27]	0.580	0.703	82.5	82.3	Aligned
MAG-BERT [47]	-	-	84.7	84.5	Aligned
ICCN [48]	0.565	0.713	84.2	84.2	Aligned
MISA [11]	0.555	0.756	85.5	85.3	Aligned
TETFN [45]	0.551	0.748	85.1	85.2	Unaligned
DMD [44]	-	-	84.8	84.7	Unaligned
IMDer3 [43]	-	-	85.1	85.1	Unaligned
MAG-BERT* [47]	0.549	0.753	85.2	85.1	Aligned
Coupled Mamba (Ours)	**0.547**	**0.756**	**85.6**	**85.5**	Unaligned
Coupled Mamba (Ours)	**0.547**	**0.758**	**85.7**	**85.6**	Aligned

Table 2: Results on CH-SIMS (Chinese). All models are based on language features extracted by BERT, and the results are compared on unaligned data. Acc-N represents N-level accuracy.

Model	CH-SIMS				
	$Acc-2\uparrow$	$Acc-3\uparrow$	$Acc-5\uparrow$	$F1-Score\uparrow$	$MAE\downarrow$
TFN [9]	78.4	65.1	39.3	78.6	0.432
LMF [30]	77.8	64.7	40.5	77.9	0.411
MFN [10]	77.9	65.7	39.5	77.9	0.435
MulT [27]	78.6	64.8	37.9	79.7	0.453
Self-MM [8]	80.0	65.5	41.5	80.4	0.425
TETFN [45]	81.2	63.2	41.8	80.2	0.420
IMDer [43]	76.3	-	**50.7**	76.4	-
Coupled Mamba (Ours)	**81.8**	**68.7**	42.1	**81.3**	**0.409**

Ablation

Table 8: All things being equal, replacing Coupled Mamba with Cross attention, we execute it five times and report the average results.

Method	CMU-MOSEI				Data Setting
	$MAE\downarrow$	$Corr\uparrow$	$Acc-2\uparrow$	$F1-Score\uparrow$	
Cross Attention	55.9	73.3	84.6	84.5	Unaligned
Coupled Mamba (Ours)	**54.7**	**75.6**	**85.6**	**85.5**	Unaligned

Table 12: Performance of Coupled Mamba on CMU-MOSEI dataset when data is missing. Other baselines are from [63]

MR	DCCA [64]	DCCAE [65]	MCTN [66]	MMIN [67]	GCNET [68]	Coupled Mamba
0.0	80.7/80.9	81.2/81.2	84.2/84.2	84.3/84.2	85.2/85.1	**85.5/85.6**
0.1	77.4/77.3	78.4/78.3	81.8/81.6	81.9/81.3	82.3/82.1	**82.6/82.7**
0.2	73.8/74.0	75.5/75.4	79.0/78.7	79.8/78.8	80.3/79.9	**81.1/80.9**
0.3	71.1/71.2	72.3/72.2	76.9/76.2	77.2/75.5	77.5/76.8	**81.0/81.0**
0.4	69.5/69.4	70.3/70.0	74.3/74.1	75.2/72.6	76.0/74.9	**78.4/78.5**
0.5	67.5/65.4	69.2/66.4	73.6/72.6	73.9/70.7	74.9/73.2	**77.4/77.7**
0.6	66.2/63.1	67.6/63.2	73.2/71.1	73.2/70.3	74.1/72.1	**75.1/75.4**
0.7	65.6/61.0	66.6/62.6	72.7/70.5	73.1/69.5	73.2/70.4	**74.1/74.2**
Average	70.3/71.2	72.6/71.2	77.0/76.1	77.3/75.4	77.9/76.8	**79.4/79.5**

Table 11: Comparison of fusion methods

Model	CMU-MOSEI			
	$MAE\downarrow$	$Corr\uparrow$	$Acc-2\uparrow$	$F1-Score\uparrow$
Average Fusion	56.4	73.6	84.2	84.1
Concat Fusion	56.2	72.8	84.8	84.5
Mamba Fusion	55.3	74.9	85.3	85.3
Coupled Fusion	**54.7**	**75.6**	**85.6**	**85.5**

Table 9: Performance on CMU-MOSEI with different timescale Δ

Δ	CMU-MOSEI		
	Corr↑	Acc-2↑	F1-Score↑
$dstate/16$	75.3	85.2	85.0
$dstate/8$	**75.6**	**85.6**	**85.5**
$dstate/4$	74.2	85.0	84.9

Table 10: Performance on CMU-MOSEI with different $dstate$

$dstate$	CMU-MOSEI		
	Corr↑	Acc-2↑	F1-Score↑
128	74.1	84.2	84.1
64	**75.6**	**85.6**	**85.5**
32	75.0	84.9	84.9

Conclusions

In this paper, we introduce Coupled Mamba, a novel approach to enhance multi-modal fusion by leveraging state evolution chains within state space. Our method integrates intermediate information from various modalities, capturing dynamic multi-modal interactions over time. This addresses challenges in parallel SSM with multiple inputs. Both quantitative and qualitative experiments confirm the effectiveness of Coupled Mamba.

Wechat **Paper**

If you have any questions, you are welcomed to add my WeChat to discuss them.