Efficient Movie Scene Detection using State-Space Transformers

Md Mohaiminul Islam, Mahmudul Hasan, Kishan Shamsundar Athrey,

Tony Braskich, Gedas Bertasius

Presentation Date: June 22, 2023

Tag: THU-AM-217

Summary

We propose an efficient model (TranS4mer) for long-range movie scene detection.

Applications:

- Understanding the storyline of the movie
- Content-driven video search
- Preview generation

Challenges

Short-Range Video Understanding

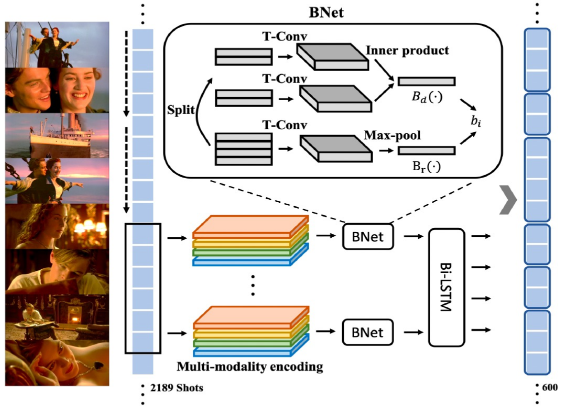
Action, object, etc.

Scene Detection

Both short-range and long-range understanding

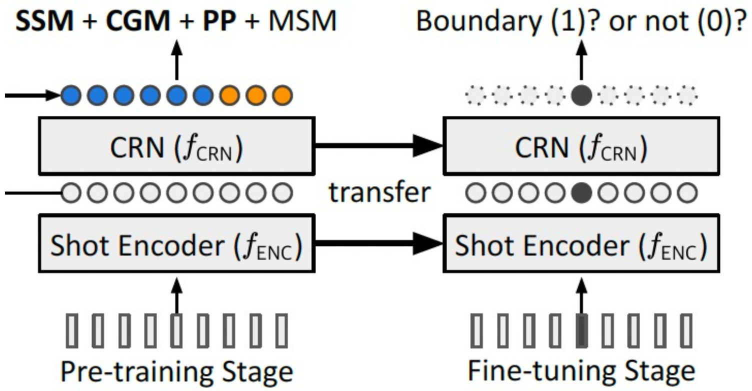
Prior Works

Most works are CNN based which is inherently designed for short-range modeling.



[Rao et al. 2020]

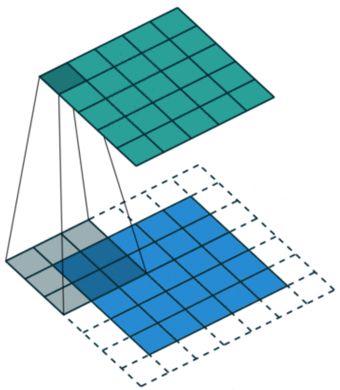
Pre-extracted CNN features



[Mun et al. 2022]

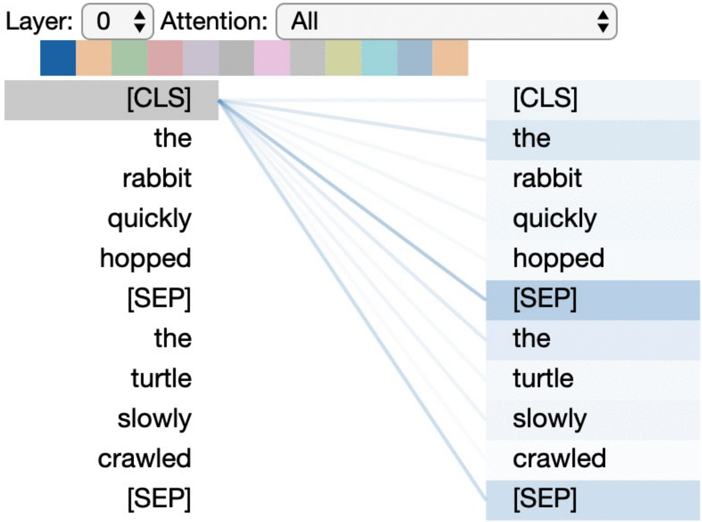
Small transformer on top of CNN encoder

Convolution



Small receptive field

Transformers

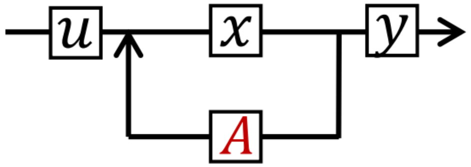


Global operation: long-range dependencies

14*14*100 = 19,600 tokens

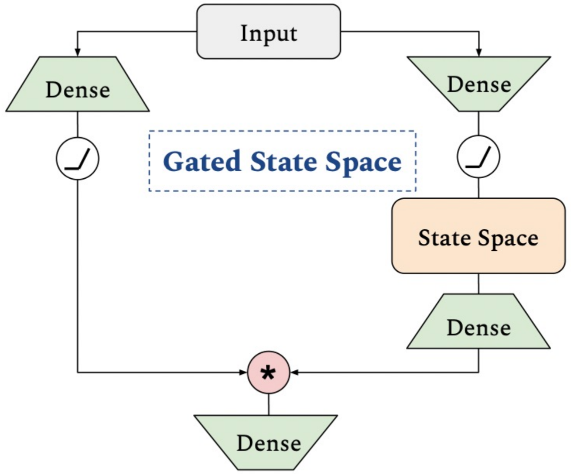
384 Million pair-wise comparison for self-attention operation

Structured State-Space Model (S4 model)



$$x_k = Ax_{k-1} + Bu_k$$
$$y_k = Cx_k$$

	Attention	State-Space
Run-time	$L^2H + H^2L$	**H^2**
Memory	$B(L^2 + HL)$	**BLH**



✓ No pair-wise comparison

✓ Linear runtime and memory

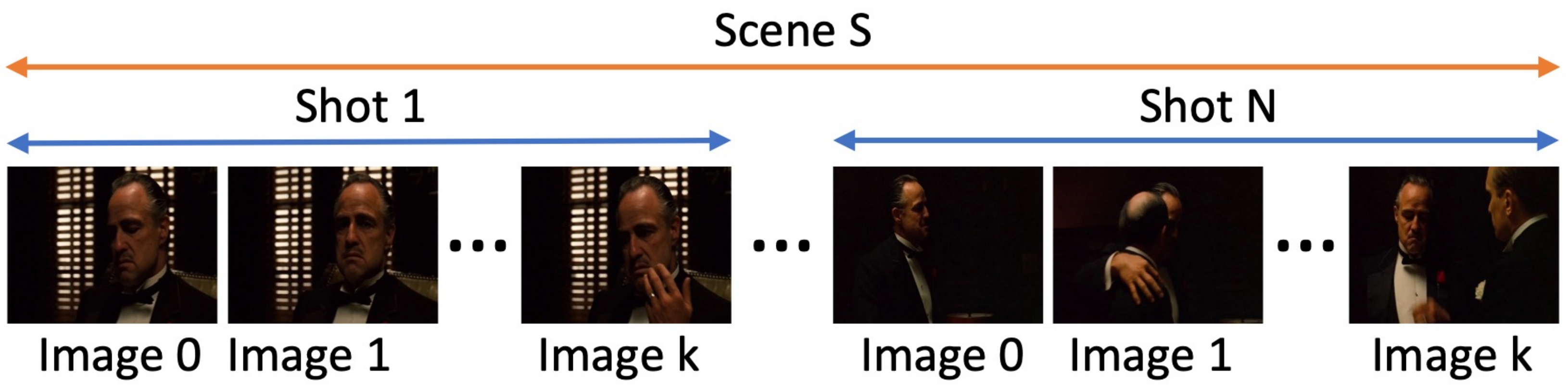
✓ Gating further improves efficiency

Efficiently Modeling Long Sequences with Structured State Spaces, Gu *et al*., ICLR 2022

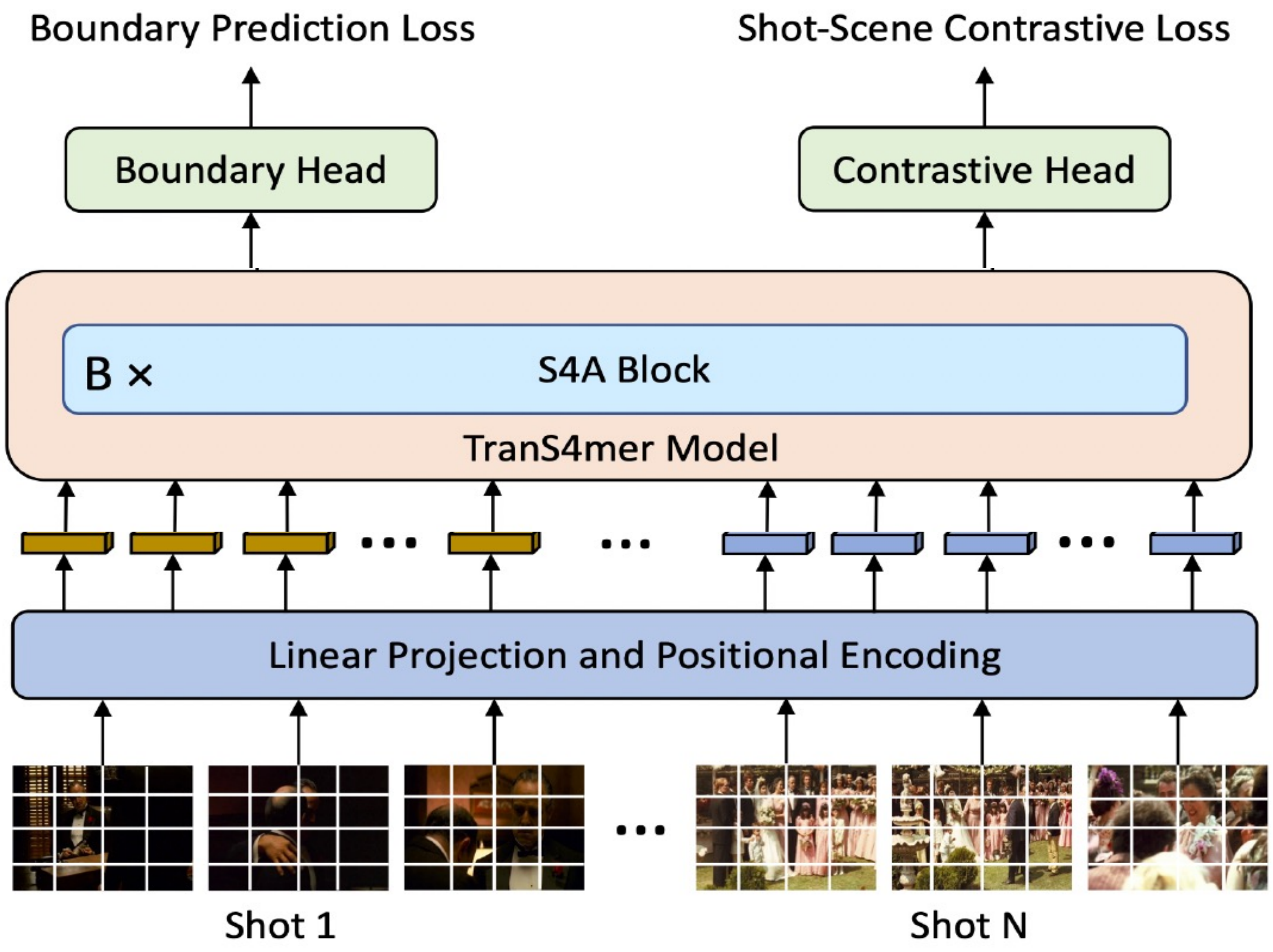
Long Range Language Modeling via Gated State Spaces, Mehta et al. 2022

Hierarchical Input Structure

1. Multiple frames taken from the same camera position constitute a shot.
2. Multiple shots capturing a semantically high-level event is a scene.
3. A movie is composed of a collection of scenes.

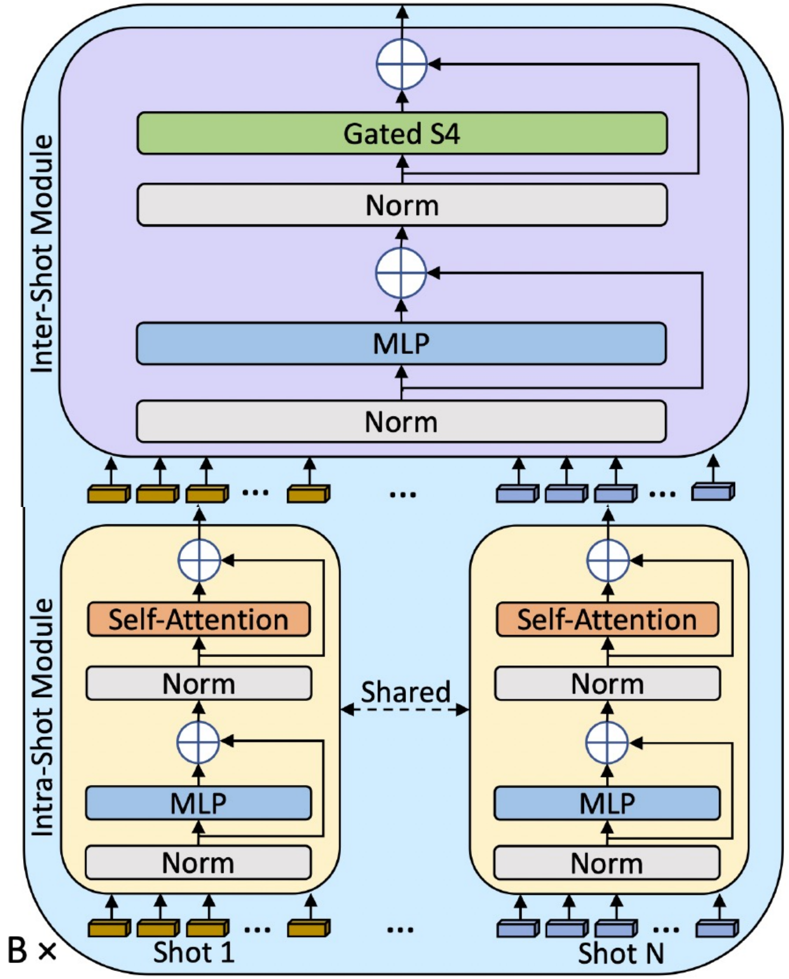


TranS4mer Model



S4A Block

S4A Block:
- Intra-Shot Module: Self-Attention
- Inter-Shot Module: Gated State-Space



Experimental Setup

Datasets

1. **MovieNet** a large-scale dataset containing 1100 movies with 1.6 million shots.
2. **BBC** contains 11 episodes from the BBC TV series Planet Earth.
3. **OVSD** contains 21 short films with an average duration of 30 minutes.

Our Implemented Baselines

1. **Transformer:** self-attention
2. **TimeSformer:** divided space-time attention
3. **Vanilla S4**: vanilla S4 layers

Main Results on MovieNet

Method	AP (↑)	mIoU (↑)	AUC-ROC (↑)	F1 (↑)	Memory (GB) (↓)	Samples/s (↑)
Siamese [3]	35.80	39.60	-	-	-	-
MS-LSTM [25]	46.50	46.20	-	-	-	-
LGSS [38]	47.10	48.80	-	-	-	-
ViS4mer [34]	55.13	48.27	88.74	46.15		
ShotCoL [11]	53.40	-	-	-	34.28	0.96
BaSSL [35]	57.40	50.69	90.54	47.02	34.28	0.96
Transformer [16]	58.81	51.21	90.84	47.88	30.28	1.27
TimeSformer [6]	59.62	50.75	90.66	48.02	28.12	1.47
Vanilla S4 [18]	59.71	51.32	90.96	47.85	15.62	1.83
TranS4mer	**60.78**	**51.91**	**91.89**	**48.36**	**10.13**	**2.57**

- Our model achieves the state-of-the-art performance in all metrics.
- TranS4mer is 2.5x faster and 3x memory efficient than prior best method BaSSL.
- Trans4ormer is also 2x faster and 3x memory efficient than self-attention based baseline while achieving better performance than all other baselines.

Results on BBC and OVSD

Method	AP (↑)
BaSSL [35]	39.98
Transformer [16]	41.86
TimeSformer [6]	42.23
Vanilla S4 [18]	42.56
TranS4mer	**43.64**

(a) Performance on BBC [3].

Method	AP (↑)
BaSSL [35]	28.68
Transformer [16]	33.12
TimeSformer [6]	33.87
Vanilla S4 [18]	34.21
TranS4mer	**36.04**

(b) Performance on OVSD [40].

TranS4mer outperforms the prior best method (BaSSL) by a large margin of 4.66% AP on BBC and 7.36% AP on OVSD datasets.

Ablation Studies

Intra-Shot	Inter-Shot	Ap (↑)
✓	✗	57.80
✗	✓	55.59
✓	✓	**60.78**

(a) TranS4mer modules.

S4 layers	AP (↑)
1-6	58.31
7-12	58.82
1-12	**60.78**

(b) S4 in different layers.

S4 layers	AP (↑)
every 2nd	59.82
every 4th	58.01
all	**60.78**

(c) S4 in every k^{th} layer.

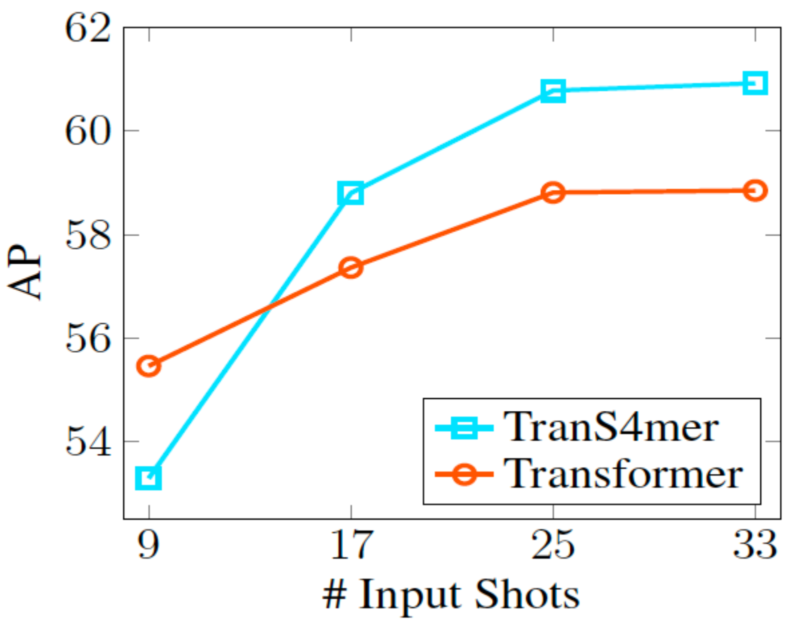
S4 layers	AP (↑)
S4	59.71
DS4	60.13
GS4	**60.78**

(d) Different S4 variants.

Intra-shot and inter-shot modules are complementary

TranS4mer with Gated S4 layers at every S4A block yields the best performance.

Temporal Extent Ablation



(a) **Performance** ↑ TranS4mer achieves much better performance for longer videos.

Generalization to other Tasks

Long Movie Clip Classification: TranS4mer achieves best performance in 5 out of 7 movie clip classification tasks in LVU benchmark.

Method	Relation	Speak	Scene	Director	Genre	Writer	Year
ObjTrans. [48]	53.10	39.40	56.90	51.20	54.60	34.50	39.10
ViS4mer [34]	57.14	**40.79**	67.44	62.61	54.71	**48.80**	44.75
TranS4mer	**59.52**	39.21	**70.93**	**63.86**	**55.85**	46.93	**45.45**

(a) Performance on LVU [48]

Procedural Activity Classification: it performs best on the Breakfast and second-best on the COIN datasets, while using significantly less pretraining data.

Model	#Data(↓)	Acc.(↑)
GHRM [51]	306K	75.50
Dist.Sup. [31]	**136M**	89.90
ViS4mer [34]	495K	88.17
TranS4mer	495K	**90.27**

(b) Performance on Breakfast [30]

Model	#Data(↓)	Acc.(↑)
TSN [44]	306K	73.40
Dist.Sup. [31]	**136M**	**90.00**
ViS4mer [34]	495K	88.41
TranS4mer	495K	89.23

(c) Performance on COIN [43]

Thank You