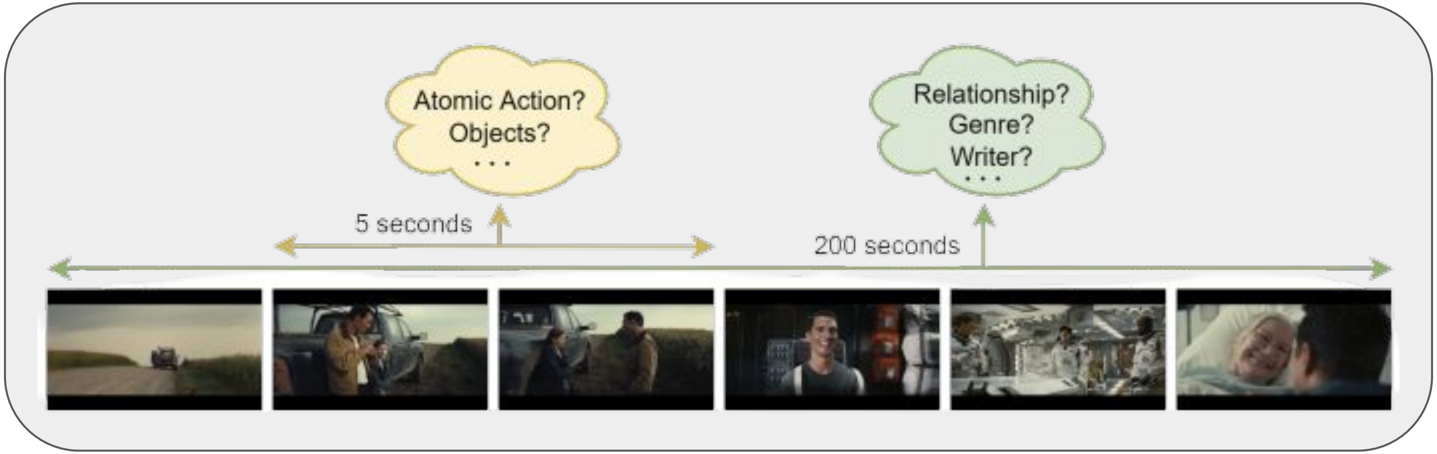
Long Movie Clip Classification with State-Space Video Models

Md Mohaiminul Islam, Gedas Bertasius

Presented by Vish Ravichandran

Introduction

Local vs. Long Prediction Tasks

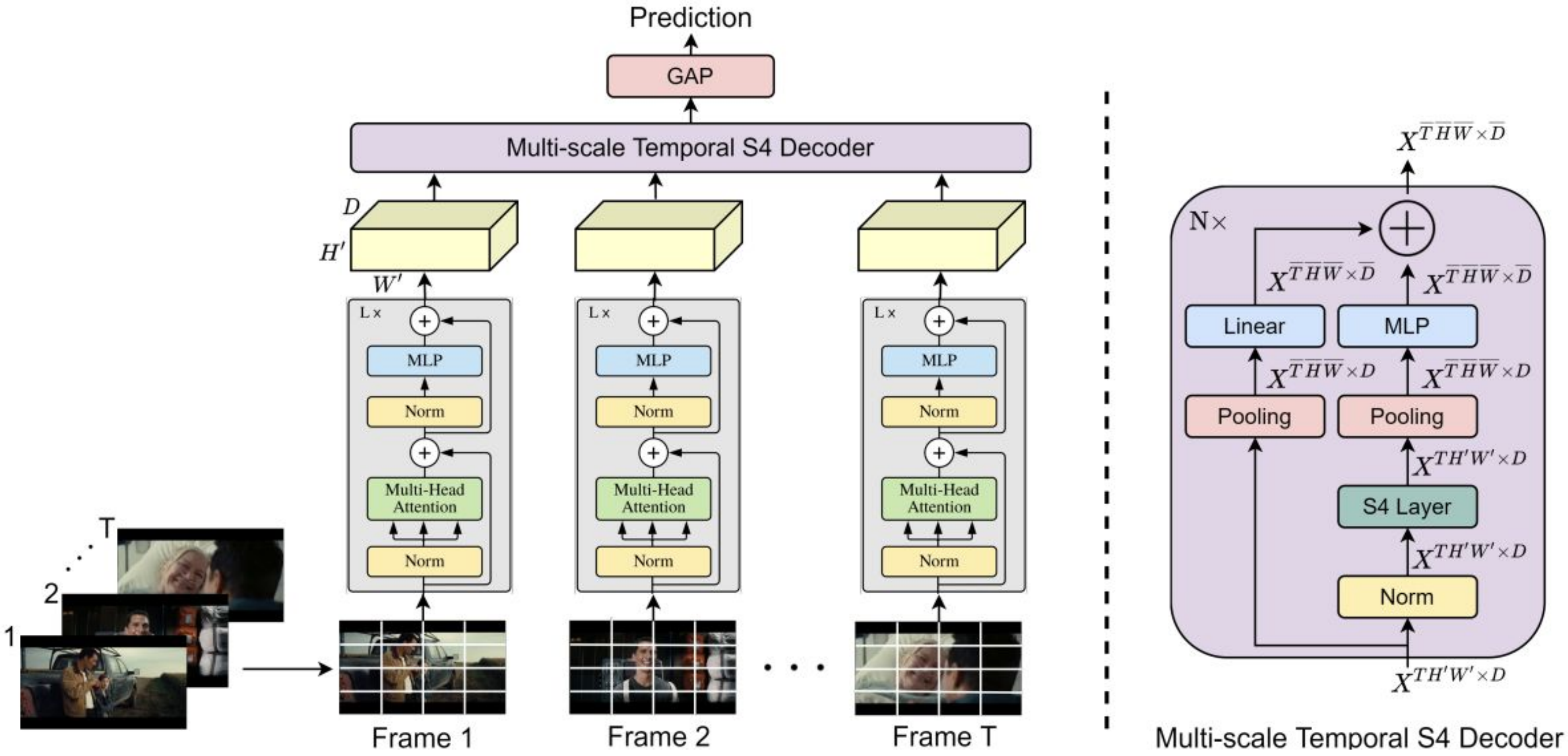


Technical Challenges & Related Work

- CNN’s
  - Fail to capture necessary fine-detail for genre, director, etc. understanding
- Transformers
  - Quadratic cost of self-attention makes only feasible for short clips
- Past work
  - Structured state-space sequence (S4) by Gu *et al.*
    - Improvement on state-space modelling via HiPPo theory (i.e. clever math)
  - Long-form Video Understanding benchmark by Wu *et al.*
    - Series of nine long-range video classification tasks

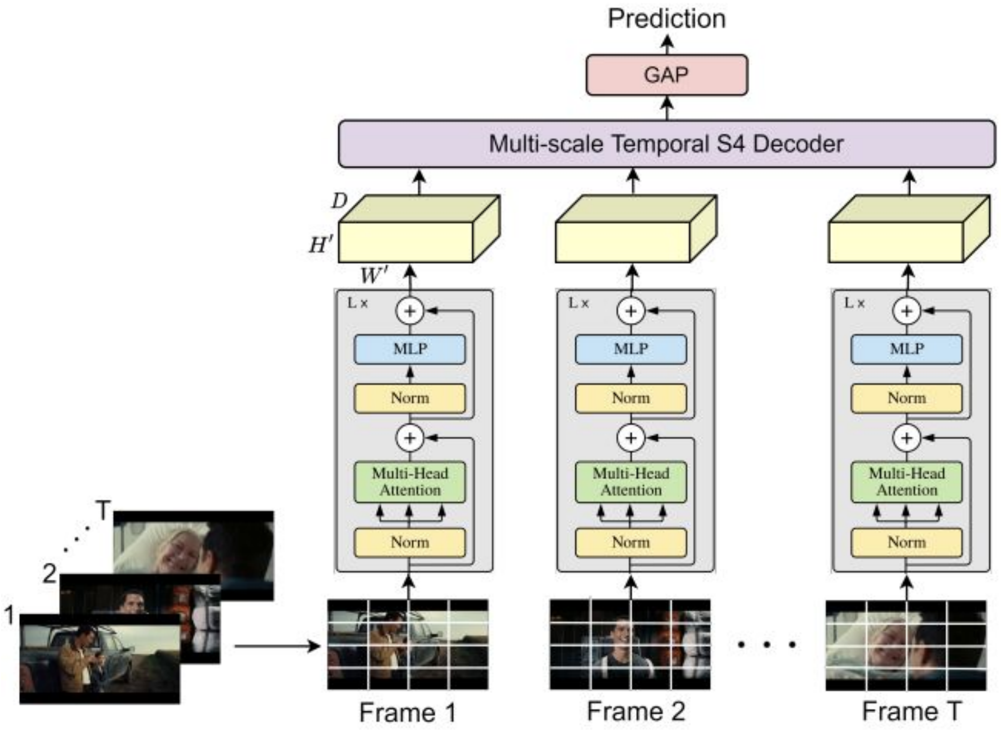
ViS4mer Model

High-level Architecture



Encoder

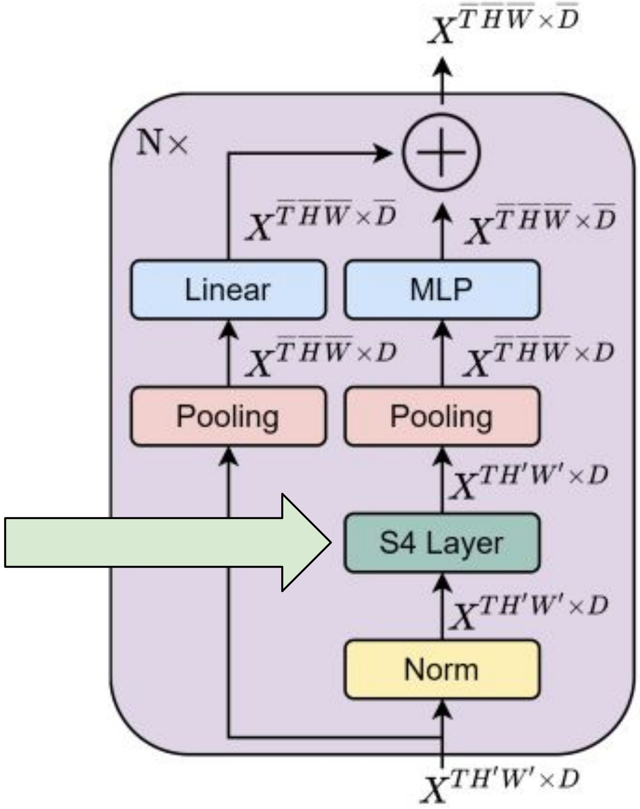
- Video V $\in R^{T \times H \times W \times 3}$
- Split into non-overlapping patches P x P
- Patches projected into latent dimension D added to positional embedding E $\in R^{N \times D}$
- $\mathbf{z'} = \mathrm{MHA}(\mathrm{LN}(\mathbf{z}_{\mathbf{in}})) + \mathbf{z}_{\mathbf{in}}$
- $\mathbf{z}_{\mathbf{out}} = \mathrm{MLP}(\mathrm{LN}(\mathbf{z'})) + \mathbf{z'}$
- Transformer outputs are concatenated X $\in R^{T \times H' \times W' \times D}$
  - W' = W/P, H' = W/P



Structured State Space Sequence vs Self-attention

- H = hidden dimension
- B = batch size
- L = sequence length
- Tilde (~) represents log

	Self-attention	State-space
Parameters	H^2	H^2
Memory	$B(L^2 + HL)$	BLH
Training	$B(L^2H + LH^2)$	$BH(\tilde{H} + \tilde{L}) + B\tilde{L}H$
Inference	$L^2H + LH^2$	H^2



SSM layer to Structured state space sequence (S4) layer

- u(t) = 1-dimension input signal
- x(t) = N-dimension hidden state
- y(t) = 1-dimension output signal

Simple SSM

$$x'(t) = Ax(t) + Bu(t)$$
$$y(t) = Cx(t) + Du(t)$$

Expensive
Gradient issues

Add constraints

S4 by Gu *et al.*

$$A_{nk} = \begin{cases} (2n+1)^{1/2}(2k+1)1/2 & \text{if } n > k \\ n+1 & \text{if } n = k \\ 0 & \text{if } n < k \end{cases}$$

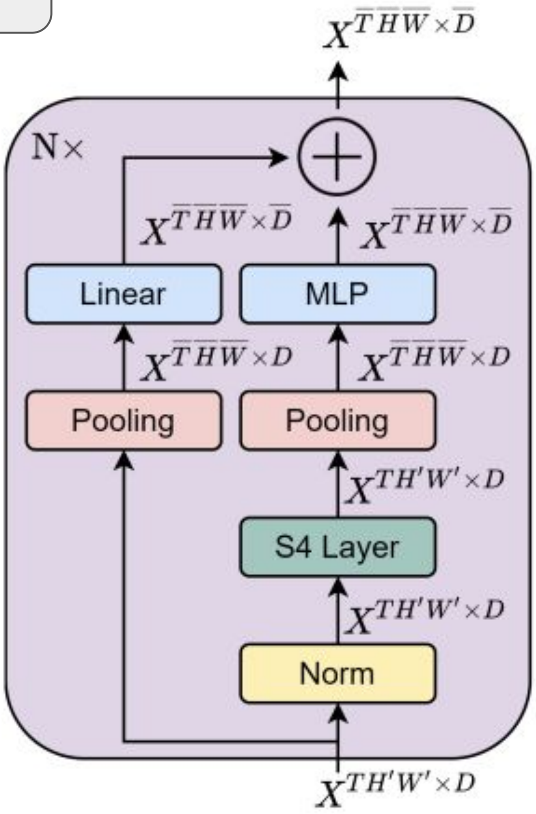
Theoretical guarantees for LRD
Reduced computation

S4

Pooling + MLP

Skip

- Multi-scale decoder with N blocks
- Gradually decreases spatio-temporal features to prevent overfitting
- Input tensor X is flattened to $\mathbf{x}_{in} = (x_1 \ldots, x_L)$
  - $L = T * H' * W'$
  - $x_i \in R^D$
  - $\mathbf{x}_{in} \in R^{L \times D}$

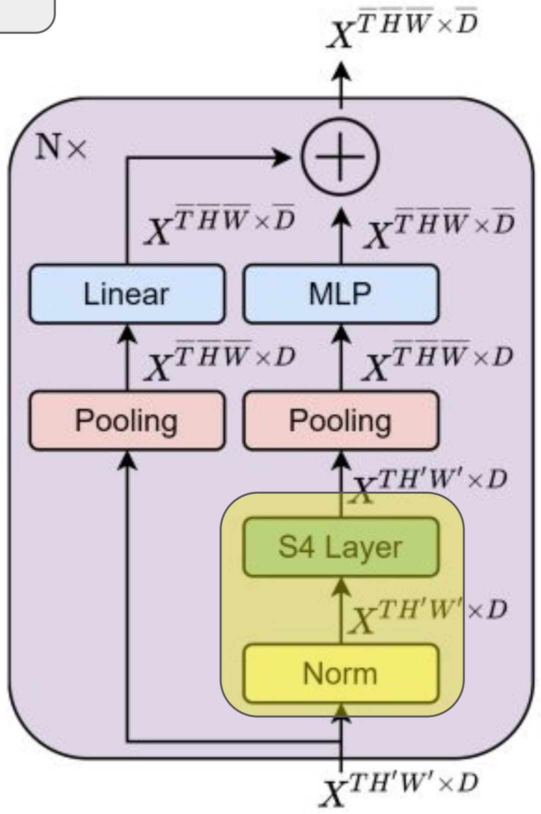


S4

Pooling + MLP

Skip

- Pass through layer normalization and S4 layer
- $\mathbf{x}_{s4} = \mathrm{S4}(\mathrm{LN}(\mathbf{x}_{in})) \in \mathrm{R}^{L \times D}$

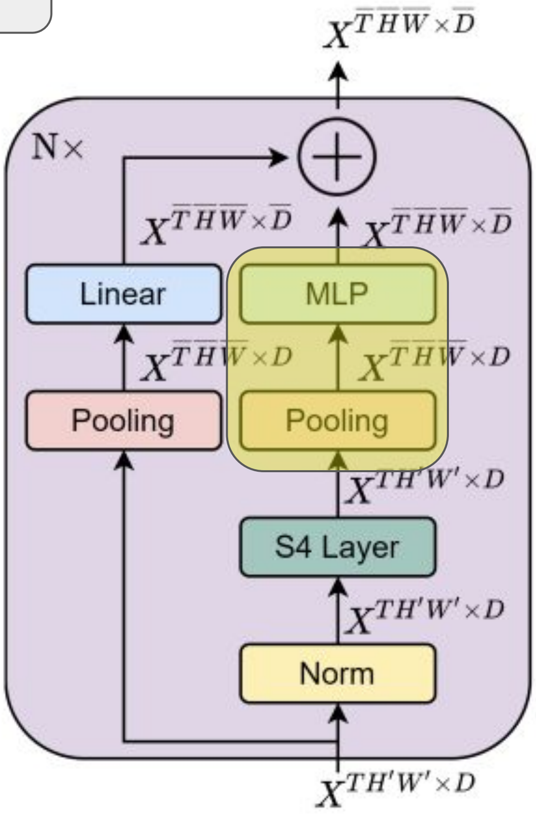


S4

Pooling + MLP

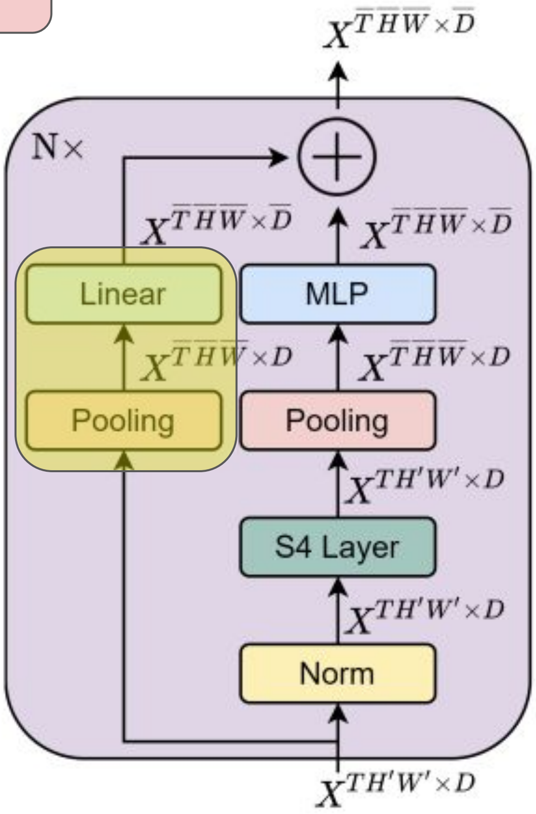
Skip

- Pooling layer to reduce spatiotemporal resolution
    - Reduces computation
- MLP layer to reduce channel resolution
    - Prevents overfitting
- $\mathbf{x}_{\mathbf{mlp}}$ = MLP(Pooling($\mathbf{x}_{\mathbf{s4}}$))



S4 | Pooling + MLP | Skip

- Skip connection uses pooling to reduce spatiotemporal resolution
- Linear layer to reduce channel dimension
- $\mathbf{x}_{skip}$ = Pooling(Linear($\mathbf{x}_{in}$))
- $\mathbf{x}_{out} = \mathbf{x}_{skip} + \mathbf{x}_{mlp}$



Loss Functions

- B = Batch Size
- K = # Classes
- y = label
- x = input
- F = model
- θ = params

Cross-entropy for Classification

$$L_{ce}(\mathcal{F}_{\mathcal{C}}(\theta)) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{K} y_j^i \log(\mathcal{F}_{\mathcal{C}}(\theta; x^i)_j)$$

MSE for Regression

$$L_{mse}(\mathcal{F}_{\mathcal{R}}(\theta)) = -\frac{1}{B}\sum_{i=1}^{B}(y^i - \mathcal{F}_{\mathcal{R}}(\theta; x^i))^2$$

Experiments

Implementation Details

- Frame size: H x W = 224 x 224
- Patch size: P x P = 16 x 16
- Encoder: L = 24 block transformer architecture pretrained on ImageNet
  - Hidden Dimension: D = 1024
- Decoder: N = 3 block S4 architecture
  - Pooling kernel 1 x 2 x 2
  - Stride 1 x 2 x 2
  - Padding 1 x 1 x 1
  - MLP layer reduces channel dimension by 2
- Optimizer: Adam
  - Learning rate: 10^{-3}
  - Weight decay: 0.01
- Batch Size: 16

Long-form Video Understanding (LVU) Benchmarks

- Made from MovieClip dataset containing ~30K 1-3min clips from ~3K movies
    - ViTrained on 60s clips
- **Content understanding**—predicting…
    - Relationship
    - Speaking style
    - Scene/place
- **Metadata prediction**—predicting…
    - Director
    - Genre
    - Writer
- **User engagement**—predicting…
    - YouTube like ratio
    - YouTube popularity

Results

- Content & Metadata use Top-1 Accuracy
- User uses MSE

	Sequence Model	Content (↑)			Metadata (↑)				User (↓)	
		Relation	Speak	Scene	Director	Genre	Writer	Year	Like	Views
SlowFast+NL [16, 51]	non-local	52.40	35.80	54.70	44.90	53.00	36.30	**52.50**	0.38	3.77
VideoBERT [44]	self-attention	52.80	37.90	54.90	47.30	51.90	38.50	36.10	0.32	4.46
Obj. Transformer [53]	self-attention	53.10	39.40	56.90	51.20	54.60	34.50	39.10	**0.23**	**3.55**
Long Seq. Transformer	self-attention	52.38	37.31	62.79	56.07	52.70	42.26	39.16	0.31	3.83
ViS4mer	state-space	**57.14**	**40.79**	**67.44**	**62.61**	**54.71**	**48.8**	44.75	0.26	3.63

Performance on Breakfast and COIN datasets

- Breakfast: ~1.7k videos with 10 cooking activities
- COIN: ~11.8k videos with 180 tasks
- Distant Supervision requires HowTo100M pretraining

(a) Long-range procedural activity classification on the Breakfast [30] dataset.

Model	Pretraining Dataset	Pretraining Samples	Accuracy(↑)
VideoGraph [24]	Kinetics-400	306K	69.50
Timeception [23]	Kinetics-400	306K	71.30
GHRM [59]	Kinetics-400	306K	75.50
Distant Supervision [33]	HowTo100M	**136M**	**89.90**
ViS4mer	Kinetics-600	495K	88.17

(b) Long-range procedural activity classification on the COIN [45] dataset.

Model	Pretraining Dataset	Pretraining Samples	Accuracy(↑)
TSN [46]	Kinetics-400	306K	73.40
Distant Supervision [33]	HowTo100M	**136M**	**90.00**
ViS4mer	Kinetics-600	495K	88.41

Qualitative Results

(a) **Task**: *'Relationship'*, **Ground Truth Label**: *'friends'*, **Our Prediction**: *'friends'*

(b) **Task**: *'Relationship'*, **Ground Truth Label**: *'boyfriend-girlfriend'*, **Our Prediction**: *'ex_boyfriend-ex_girlfriend'*

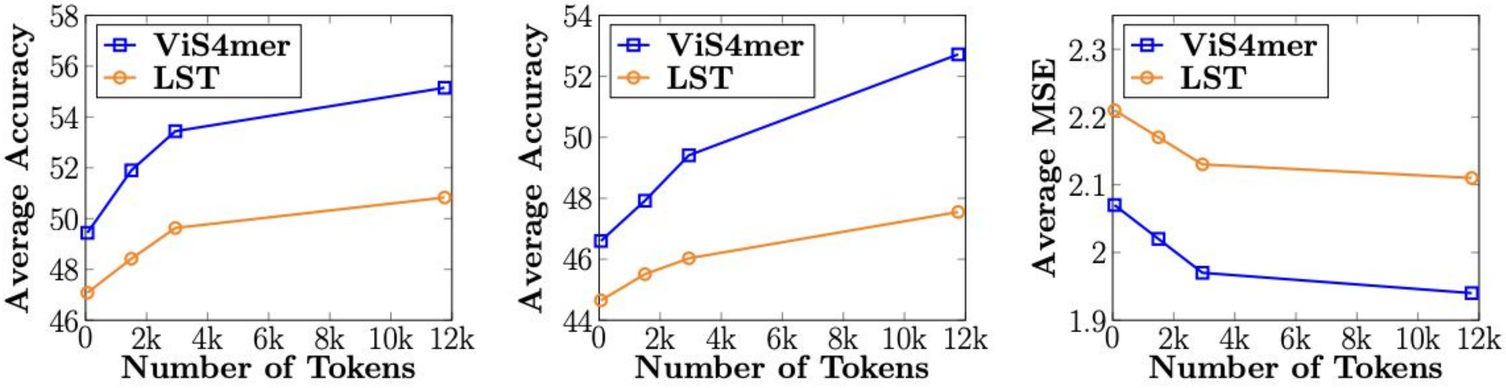
(c) **Task**: *'Genre'*, **Ground Truth Label**: *'Action/Crime/Adventure'*, **Our Prediction**: *'Action/Crime/Adventure'*

(d) **Task**: *'Genre'*, **Ground Truth Label**: *'Comedy'*, **Our Prediction**: *'Romance'*

Ablations

Accuracy by Number of Tokens

- S4 layer is superior to self-attention layers
- Gap increases as token size increases



(a) Content Understanding ↑ (b) Metadata Prediction ↑ (c) User Engagement Prediction ↓

Training Speed & Memory Utilization by Token Size

- Overall S4 requires 8x less memory and 2.63x faster than self-attention on long videos (i.e. 11,760 tokens)
- Gap grows significantly with greater token sizes

# of Tokens	Samples/s (↑)		GPU Memory (GB)(↓)	
	ViS4mer	LST	ViS4mer	LST
60	**12.46**	8.85	**2.23**	2.45
1,500	**8.27**	6.31	**3.61**	3.99
2,940	**6.25**	4.47	**3.67**	5.43
11,760	**4.95**	1.88	**5.15**	41.38

Comparison to Non-quadratic w.r.t. Input Length Methods

- Replace S4 Layer with other self-attention alternatives
- State-space still outperforms in almost LVU benchmarks

	Content (↑)			Metadata (↑)				User (↓)			
	Relation	Speak	Scene	Director	Genre	Writer	Year	Like	Views	Sam./s (↑)	Mem (↓)
Self-attention	52.38	37.31	62.79	56.07	52.70	42.26	39.16	0.31	3.83	1.88	41.38
Performer	50.00	38.80	60.46	58.87	49.45	48.21	41.25	0.31	3.93	4.67	5.93
Orthoformer	50.00	39.30	66.27	55.14	**55.79**	47.02	43.35	0.29	3.86	4.85	5.56
State-space	**57.14**	**40.79**	**67.44**	**62.61**	54.71	**48.8**	**44.75**	**0.26**	**3.63**	**4.95**	**5.15**

Significance of Dimension Reduction

- Multi-scale decoder provides best performance on LVU benchmarks
  - MLP/Linear layer for channel reduction
  - Pooling layers for spatiotemporal reduction
- ViS4mer beats vanilla S4 by wide margin

Pooling	Scaling	Content(↑)	Metadata(↑)	User(↓)	Samples/s(↑)	Memory(GB)(↓)
✗	✗	49.53	49.26	2.30	2.25	7.27
✓	✗	48.96	49.77	2.10	3.98	5.96
✗	✓	52.25	48.79	2.09	4.12	5.95
✓	✓	**55.12**	**52.72**	**1.94**	**4.95**	**5.15**

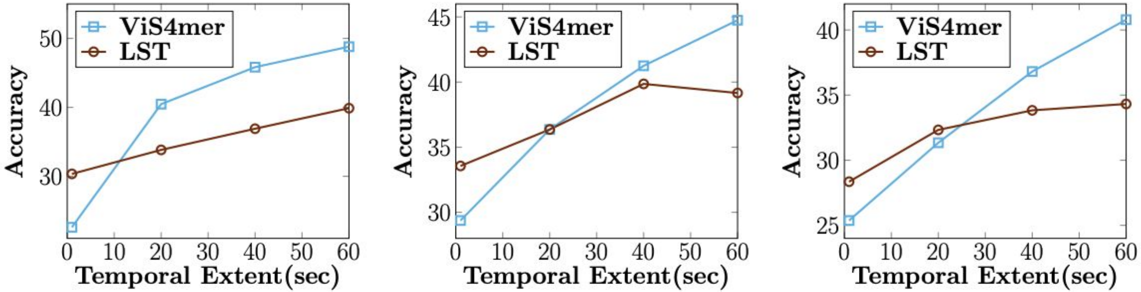
Short-range Encoders

- ViS4mer with ViT encoder outperforms in 6/9 tasks

Model	Encoder	Content (↑)			Metadata (↑)				User (↓)	
		Relation	Speak	Scene	Director	Genre	Writer	Year	Like	Views
Obj. Trans. [53]	SlowFast [16]	53.10	39.40	56.90	51.20	54.60	34.50	39.10	**0.23**	**3.55**
	ViT [13]	54.76	33.17	52.94	47.66	52.74	36.30	37.76	0.30	3.68
ViS4mer	SlowFast [16]	**59.52**	40.29	60.46	53.27	52.74	42.85	39.86	0.27	3.70
	ConvNeXt [37]	**59.52**	38.30	62.79	57.00	54.40	45.83	42.65	0.30	3.74
	Swin [35]	54.76	37.31	61.62	56.07	49.45	47.61	39.86	0.31	3.56
	ViT [13]	57.14	**40.79**	**67.44**	**62.61**	**54.71**	**48.8**	**44.75**	0.26	3.63

Varying Training Video Length

- ViS4mer yields sharp gains in longer training videos due to more effective long-range temporal reasoning



(a) Writer Prediction. (b) Year Prediction. (c) Speaking Style Prediction.