Video Token Merging for Long Video Understanding

Seon-Ho Lee, Jue Wang, Zhikang Zhang, David Fan, Xinyu Li

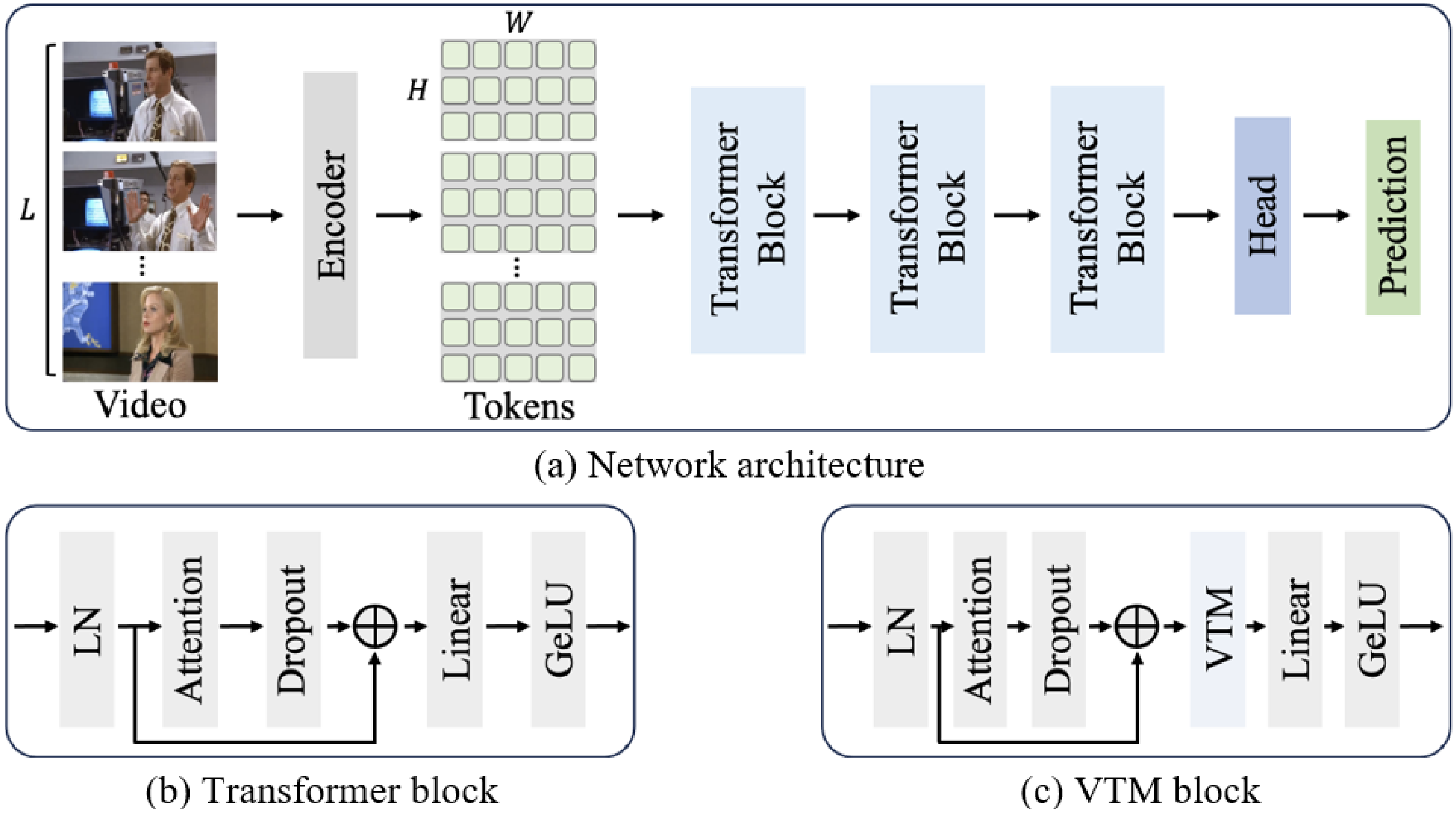
Motivation & Goal

- Transformer based networks have shown great results
- However, for long video input, it requires high computation cost

- Goal
  - Explore various token merging methods for long video understanding
  - Find effective token merging method for long videos

Naïve Video Token Merging

- Combine the standard token merging with the baseline network as intact as possible



Naïve Video Token Merging

- Combine the standard token merging with the baseline network as intact as possible

	Content (↑)			Meta data (↑)				User (↓)	
Algorithm	Relationship	Speaking	Scene	Director	Genre	Writer	Year	Like	View
Baseline	57.14	36.68	69.76	62.61	56.73	49.40	39.86	0.28	4.18
Naïve	<u>61.90</u>	36.18	72.09	<u>67.28</u>	55.12	51.19	44.75	0.28	**4.01**

Region-Concentrated Video Token Merging

- A video contains redundant spatiotemporal tokens
- Some tokens are more important than others
- Naïve VTM: uniform token partitioning

Naïve VTM

Center-concentrated VTM

Region-Concentrated Video Token Merging

- A video contains redundant spatiotemporal tokens
- Some tokens are more important than others
- Naïve VTM: uniform token partitioning

	Content (↑)			Meta data (↑)				User (↓)	
Algorithm	Relationship	Speaking	Scene	Director	Genre	Writer	Year	Like	View
Baseline	57.14	36.68	69.76	62.61	56.73	49.40	39.86	0.28	4.18
Naïve	<u>61.90</u>	36.18	72.09	<u>67.28</u>	55.12	51.19	44.75	0.28	**4.01**
Boundary	59.52	37.18	69.76	61.68	57.21	50.0	<u>47.55</u>	0.26	4.16
Center	<u>61.90</u>	<u>40.20</u>	<u>74.41</u>	62.61	<u>58.81</u>	51.19	44.05	0.25	<u>4.11</u>

Motion-Based Video Token Merging

- Moving objects carry important cues in general
- Assign higher sampling probability to tokens with large movement

Motion-based VTM

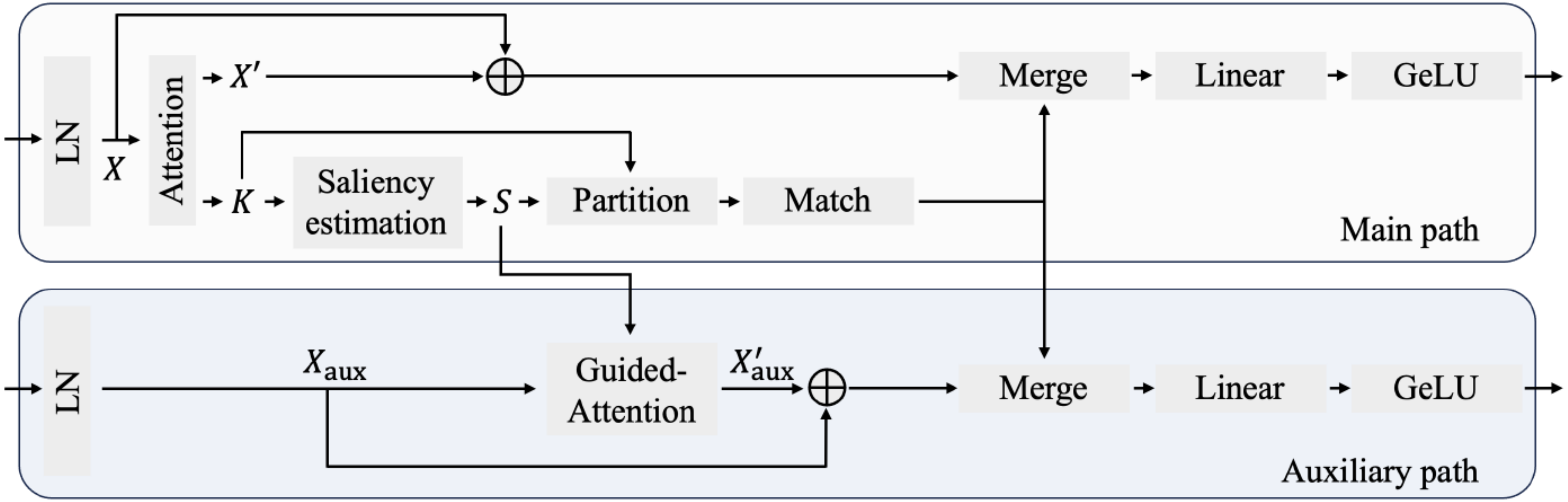
Motion-Based Video Token Merging

- Moving objects carry important cues in general
- Assign higher sampling probability to tokens with large movement

	Content (↑)			Meta data (↑)				User (↓)	
Algorithm	Relationship	Speaking	Scene	Director	Genre	Writer	Year	Like	View
Baseline	57.14	36.68	69.76	62.61	56.73	49.40	39.86	0.28	4.18
Naïve	<u>61.90</u>	36.18	72.09	<u>67.28</u>	55.12	51.19	44.75	0.28	**4.01**
Boundary	59.52	37.18	69.76	61.68	57.21	50.0	<u>47.55</u>	0.26	4.16
Center	<u>61.90</u>	<u>40.20</u>	<u>74.41</u>	62.61	<u>58.81</u>	51.19	44.05	0.25	<u>4.11</u>
Motion	**64.28**	37.68	<u>74.41</u>	64.48	58.49	**55.95**	<u>47.55</u>	<u>0.24</u>	4.13

Learnable Video Token Merging

- Predict the saliency score of each token
- Sample target tokens based on the estimated saliency



Architecture of learnable VTM block

Learnable Video Token Merging

- Predict the saliency score of each token
- Sample target tokens based on the estimated saliency

	Content (↑)			Meta data (↑)				User (↓)	
Algorithm	Relationship	Speaking	Scene	Director	Genre	Writer	Year	Like	View
Baseline	57.14	36.68	69.76	62.61	56.73	49.40	39.86	0.28	4.18
Naïve	61.90	36.18	72.09	67.28	55.12	51.19	44.75	0.28	**4.01**
Boundary	59.52	37.18	69.76	61.68	57.21	50.0	47.55	0.26	4.16
Center	61.90	40.20	74.41	62.61	58.81	51.19	44.05	0.25	4.11
Motion	**64.28**	37.68	74.41	64.48	58.49	**55.95**	47.55	0.24	4.13
Learnable	**64.28**	**42.12**	**75.58**	**70.09**	**59.77**	53.57	**48.55**	**0.21**	**4.01**

Experiments

- Comparison on the LVU dataset

	Content (↑)			Meta data (↑)				User (↓)			
Algorithm	Relationship	Speaking	Scene	Director	Genre	Writer	Year	Like	View	GPU	Throughput
Obj. T4mer (Wu & Krähenbühl, 2021)	54.76	33.17	52.94	47.66	52.74	36.30	37.76	0.30	3.68	-	-
VideoBERT (Sun et al., 2019)	52.80	37.90	54.90	47.30	51.90	38.50	36.10	0.32	4.46	-	-
Performer (Choromanski et al., 2021)	50.00	38.80	60.46	58.87	49.45	48.21	41.25	0.31	3.93	5.93GB	-
Orthoformer (Patrick et al., 2021)	50.00	38.30	66.27	55.14	55.79	47.02	43.35	0.29	3.86	5.56GB	-
LST (Islam & Bertasius, 2022)	52.38	37.31	62.79	56.07	52.70	42.26	39.16	0.31	3.83	41.38GB	-
ViS4mer (Islam & Bertasius, 2022)	57.14	40.79	67.44	62.61	54.71	48.80	44.75	0.26	3.63	5.15GB	25.64
S5 (Wang et al., 2023)	61.98	41.75	69.88	66.40	58.80	50.60	47.70	0.25	**3.51**	3.85GB	25.0
S5+LSMCL (Wang et al., 2023)	61.98	41.75	72.53	66.40	**61.34**	50.60	47.70	0.24	**3.51**	3.85GB	25.0
Learnable VTM	**64.28**	**42.12**	**75.58**	**70.09**	59.77	**53.57**	**48.55**	**0.21**	4.01	1.60GB	44.94

Experiments

- Results on the COIN and Breakfast dataset

Algorithm	PT Dataset	#PT Samples	Accuracy
TSN (Tang et al., 2020)	Kinetics-400	306K	73.40
D-sprv (Lin et al., 2022)	HowTo100M	136M	90.00
ViS4mer (Islam & Bertasius, 2022)	Kinetics-600	495K	88.41
ViS4mer* (Islam & Bertasius, 2022)	Kinetics-600	495K	87.11
S5 (Wang et al., 2023)	Kinetics-600	495K	90.42
S5+LSMCL (Wang et al., 2023)	Kinetics-600	495K	**90.81**
Learnable VTM	Kinetics-600	495K	88.55

COIN

Algorithm	PT Dataset	#PT Samples	Accuracy
VideoGraph (Hussein et al., 2019b)	Kinetics-400	306K	65.50
Timeception (Hussein et al., 2019a)	Kinetics-400	136M	71.30
GHRM (Zhou et al., 2021)	Kinetics-400	495K	75.50
D-sprv (Lin et al., 2022)	HowTo100M	136M	89.90
ViS4mer (Islam & Bertasius, 2022)	Kinetics-600	495K	88.17
S5 (Wang et al., 2023)	Kinetics-600	495K	90.14
S5+LSMCL (Wang et al., 2023)	Kinetics-600	495K	90.70
Learnable VTM	Kinetics-600	495K	**91.26**

BreakFast

Thank you!

Poster: 12 Thu, 11AM-2PM
https://neurips.cc/virtual/2024/poster/93137

seonholee@mcl.korea.ac.kr