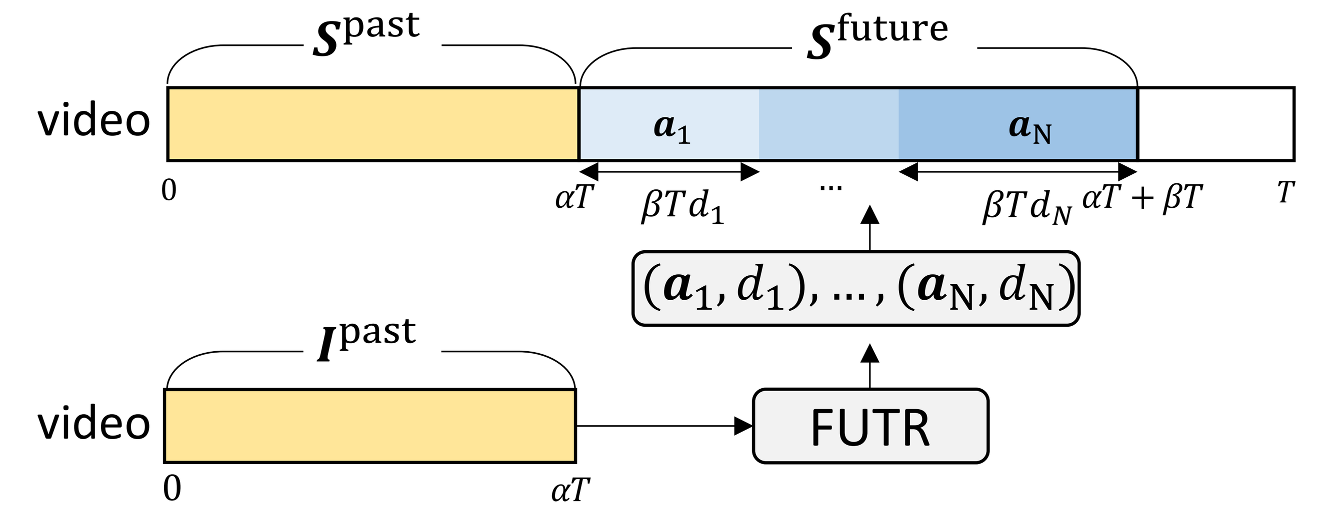
Future Transformer for Long-term Action Anticipation

paper website

Dayoung Gong[1] Joonseok Lee[1] Manjin Kim[1] Seong Jong Ha[2] Minsu Cho[1]

[1]Pohang University of Science and Technology (POSTECH) [2]NCSOFT

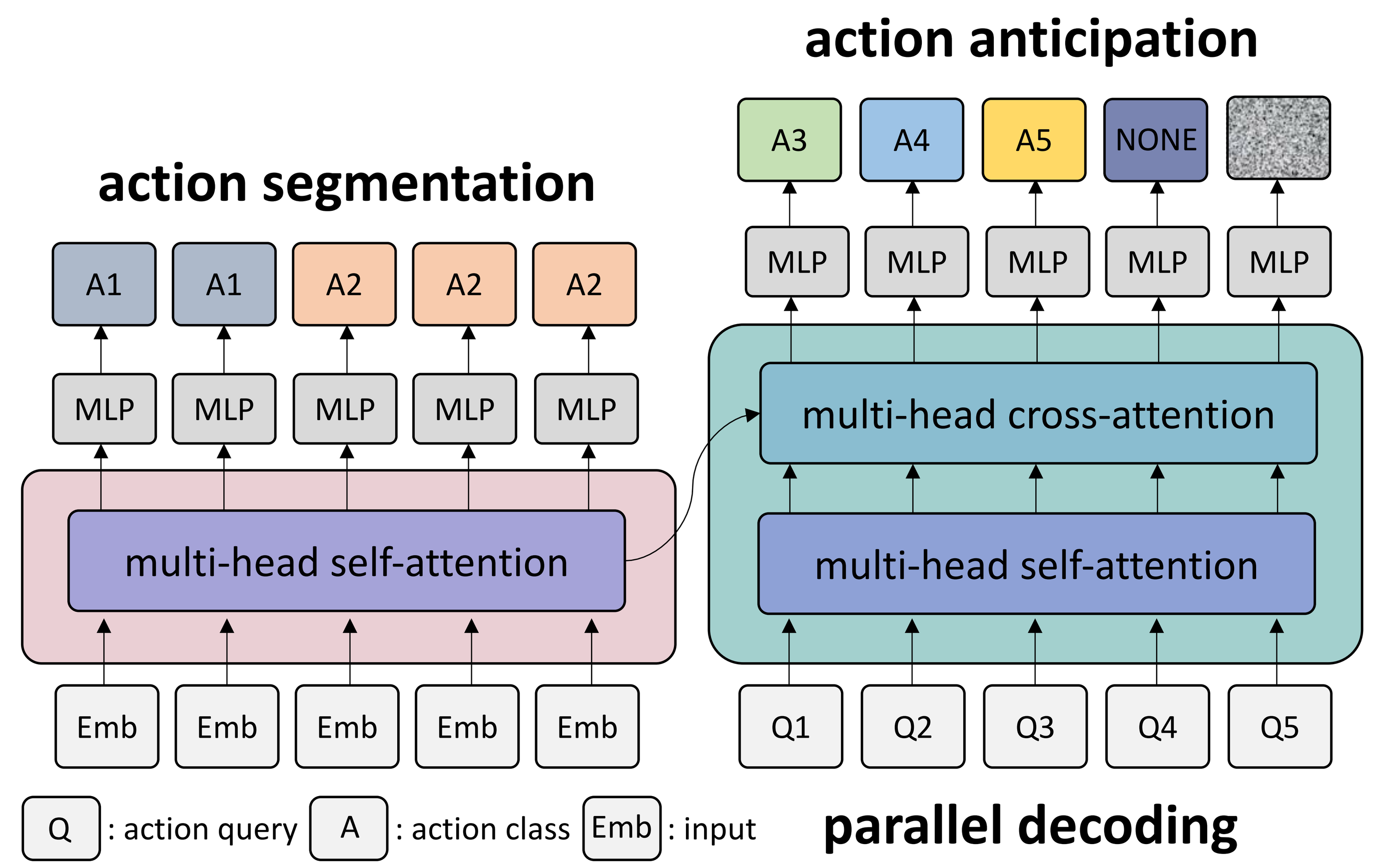
Long-term action anticipation



: Anticipating actions of future video frames with limited observation

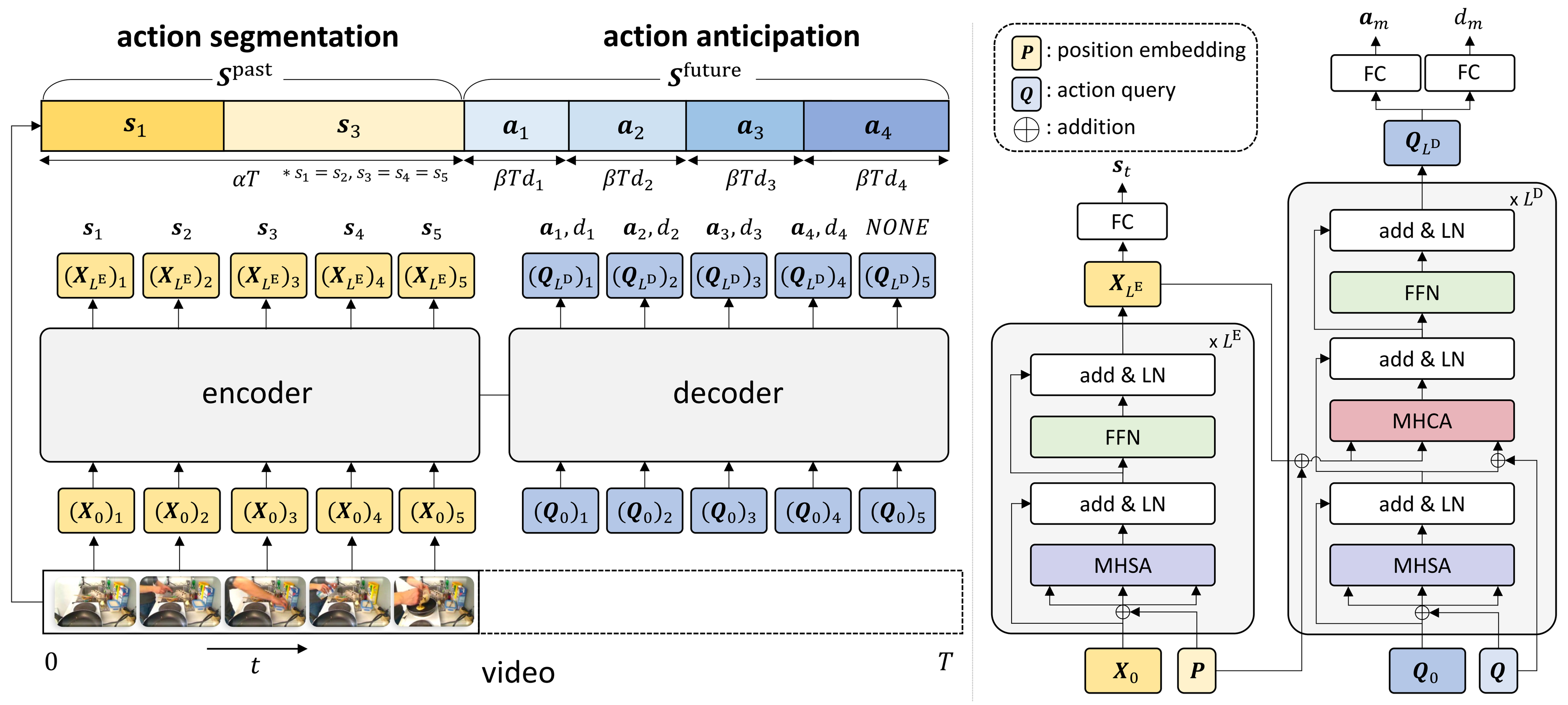
- Learning of long-term dependencies between past and future actions
- Inferring classes and durations of future actions

Contributions



- **An end-to-end attention neural network**, dubbed FUTR
- **Action segmentation (encoder) & action anticipation (decoder)**
- **Leveraging fine-grained visual features** of past actions
- **Long-term dependency modeling** between past and future actions
- **Parallel anticipation** for accurate and faster inference
- **State of the art** on long-term action anticipation benchmarks

Future Transformer (FUTR)



Action segmentation (Encoder)

- Input: observed frames (X_0)
- Output: logits of action segmentation
- Learning distinctive feature representations between past actions via self-attention
- L_{seg} : cross-entropy loss for action segmentation

Action anticipation (Decoder)

- Input: action query (Q_0) & output of the encoder (X_{L^E})
- Output: logits of action anticipation & predicted duration
- Learning long-term dependencies between past and future actions via self-attention and cross-attention
- $L_{anticipate}$: Cross-entropy loss for future action anticipation
- $L_{duration}$: L2 loss for duration prediction

Advantages of FUTR

- Faster inference time
- No error accumulation
- Long-term relations of past and future actions
- Utilizing fine-grained visual features of past actions

Experimental Results

Comparison with the state of the arts

dataset	methods	β ($\alpha=0.2$)				β ($\alpha=0.3$)			
		0.1	0.2	0.3	0.5	0.1	0.2	0.3	0.5
Breakfast	CNN (Farha et al., CVPR18')	12.78	11.62	11.21	10.27	17.72	16.87	15.48	14.09
	Temporal Agg. (Sener et al., ECCV20')	24.20	21.10	20.00	18.10	30.40	26.30	23.80	21.20
	Cycle Cons. (Farha et al., GCPR20')	25.88	23.42	22.42	21.54	29.66	27.37	25.58	25.20
	FUTR (ours)	**27.70**	**24.55**	**22.83**	**22.04**	**32.27**	**29.88**	**27.49**	**25.87**
50 Salads	Temporal Agg. (Sener et al., ECCV20')	25.50	19.90	18.20	15.10	30.60	22.50	19.10	11.20
	Cycle Cons. (Farha et al., GCPR20')	34.76	**28.41**	21.82	15.25	34.39	23.70	18.95	**15.89**
	FUTR (ours)	**39.55**	27.54	**23.31**	**17.77**	**35.15**	24.86	**24.22**	15.26

* α: observation rates, β: prediction rates

Analysis

Parallel vs. auto-regressive decoding

method	AR	causal mask	β ($\alpha=0.3$) 0.1	0.2	0.3	0.5	time (ms)
FUTR-A	✓	✓	27.10	25.41	23.28	20.51	14.68
FUTR-M	-	✓	31.82	28.55	26.57	24.17	5.70
FUTR	-	-	**32.27**	**29.88**	**27.49**	**25.87**	**3.91**

- Parallel decoding is efficient and effective.
- Bi-directional attention of action query is crucial.

Effect of action segmentation loss

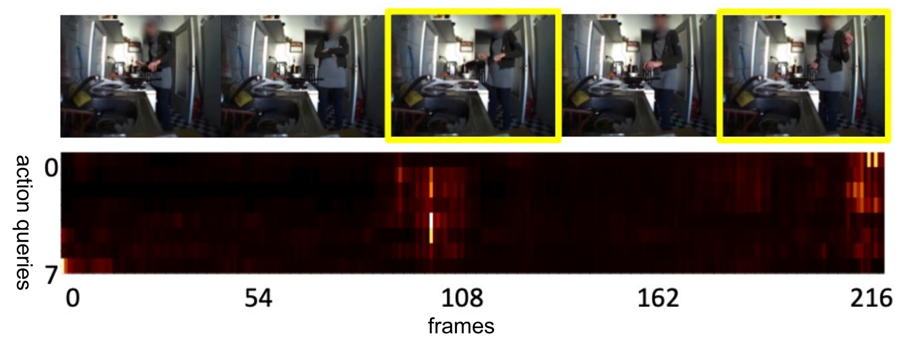
$\mathcal{L}^{seg}$	$\mathcal{L}^{action}$	$\mathcal{L}^{duration}$	β ($\alpha=0.3$) 0.1	0.2	0.3	0.5
-	✓	✓	28.31	25.85	24.91	22.50
✓	✓	✓	**32.27**	**29.88**	**27.49**	**25.87**

- Using action segmentation is effective.

Effect of long-term relations modeling

encoder	decoder	β ($\alpha=0.3$) 0.1	0.2	0.3	0.5
LSA	LSA	27.70	24.39	23.18	21.60
GSA	LSA	30.15	27.51	25.62	23.28
LSA	GSA	28.37	25.08	24.03	22.28
GSA	GSA	**32.27**	**29.88**	**27.49**	**25.87**

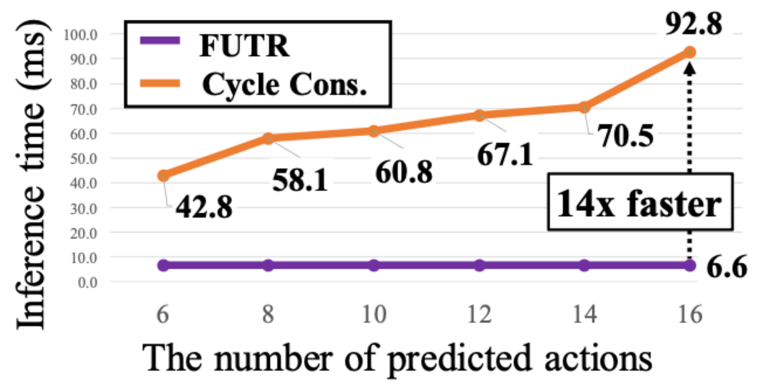
Cross-attention map visualization

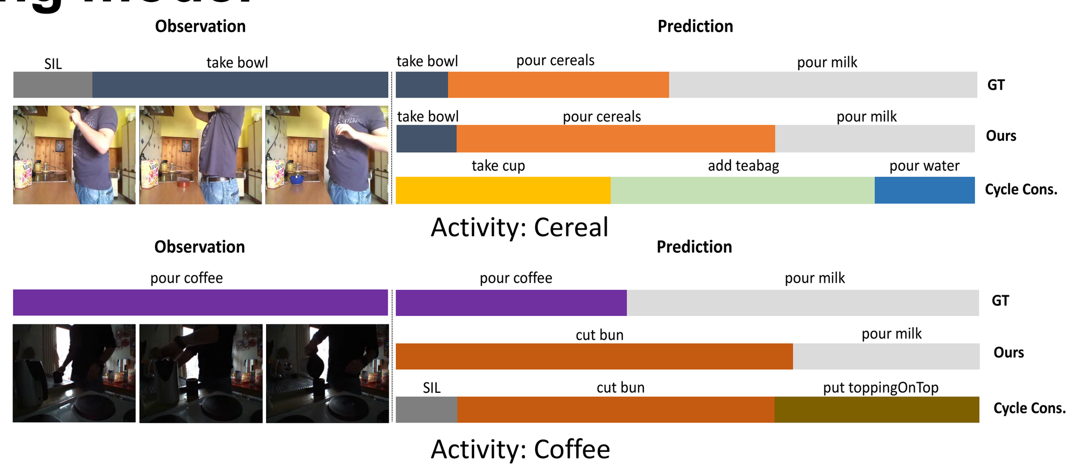


Activity: Fried egg

- Learning long-term relations between past and future action is crucial.
- FUTR detects important actions by using fine-grained visual features.

Comparison with the existing model





Activity: Cereal

Activity: Coffee

- FUTR is 14 x faster than Cycle Cons. (Farha et al. GCPR'20) when predicting 16 actions.
- FUTR utilizes fine-grained visual features without error accumulation.

Future Transformer for Long-term Action Anticipation

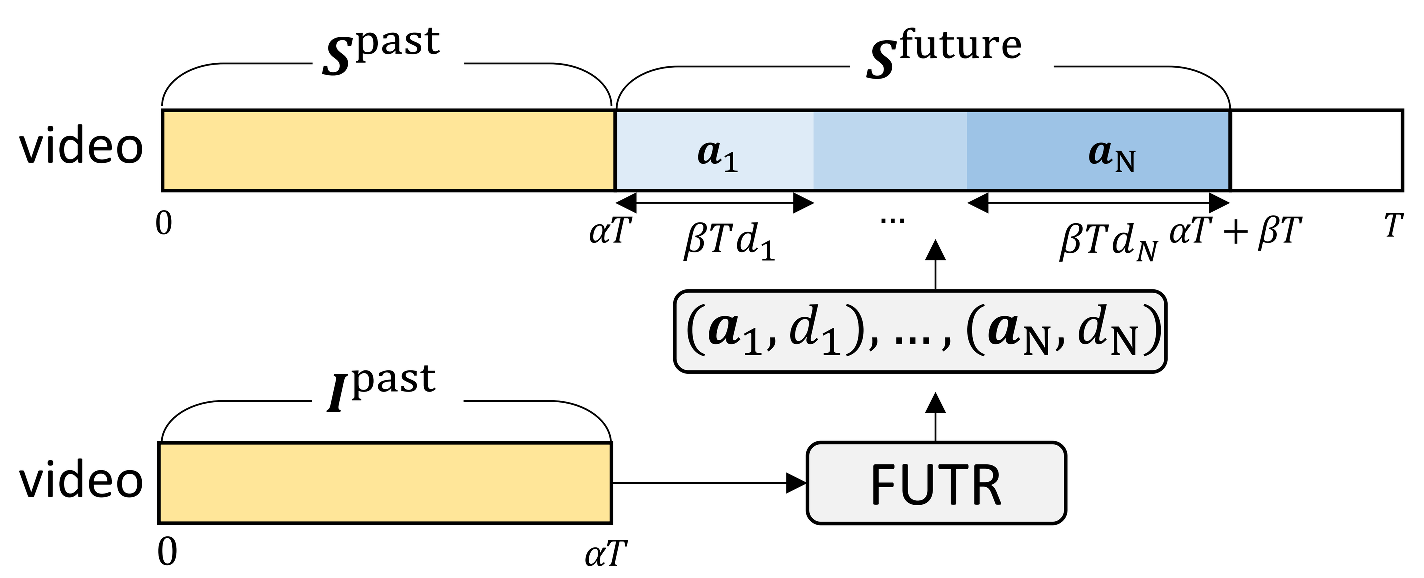
paper website

Dayoung Gong[1] Joonseok Lee[1] Manjin Kim[1] Seong Jong Ha[2] Minsu Cho[1]

[1]Pohang University of Science and Technology (POSTECH) [2]NCSOFT

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

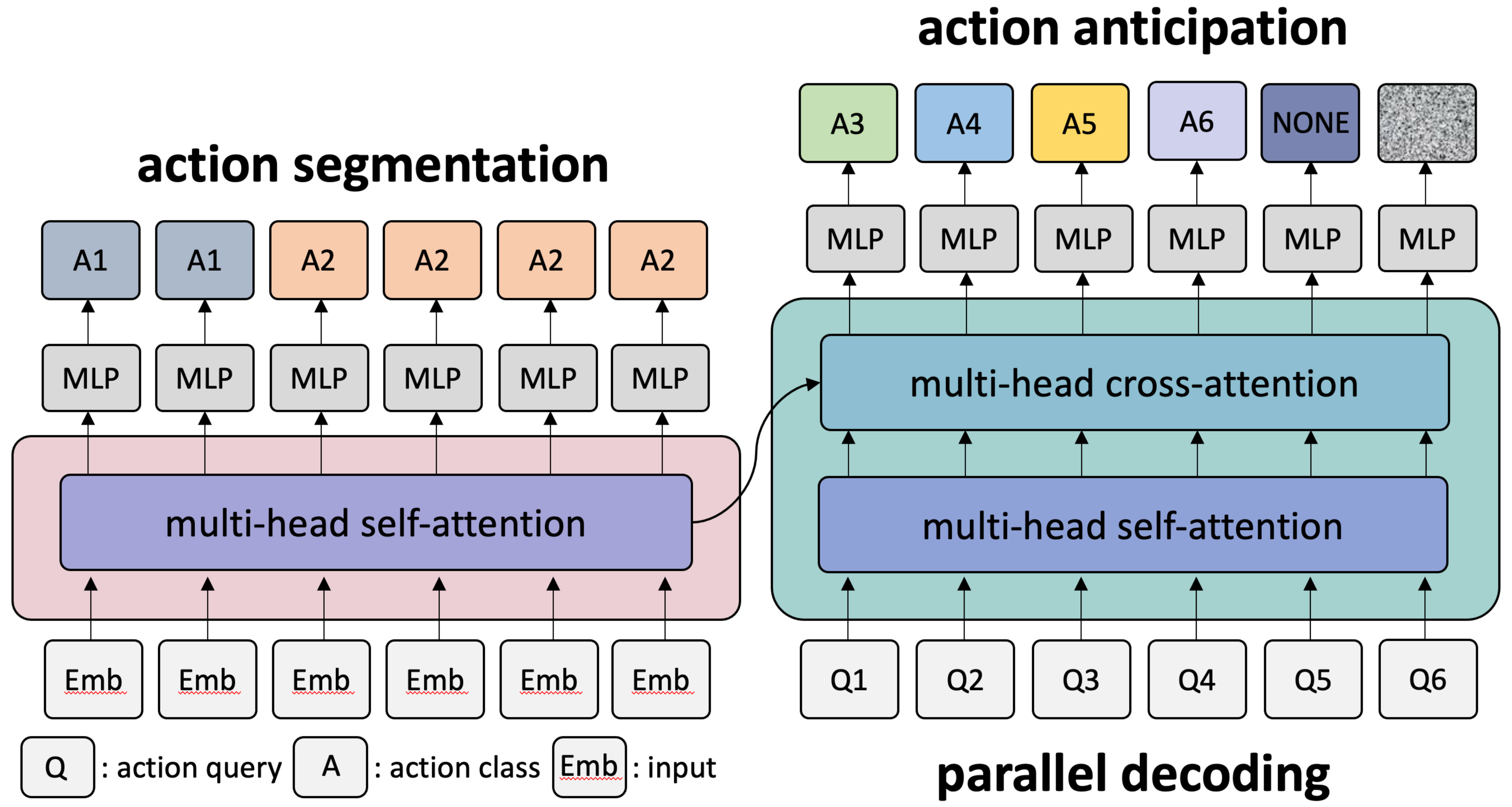
Long-term action anticipation



: Anticipating actions of future video frames with limited observation

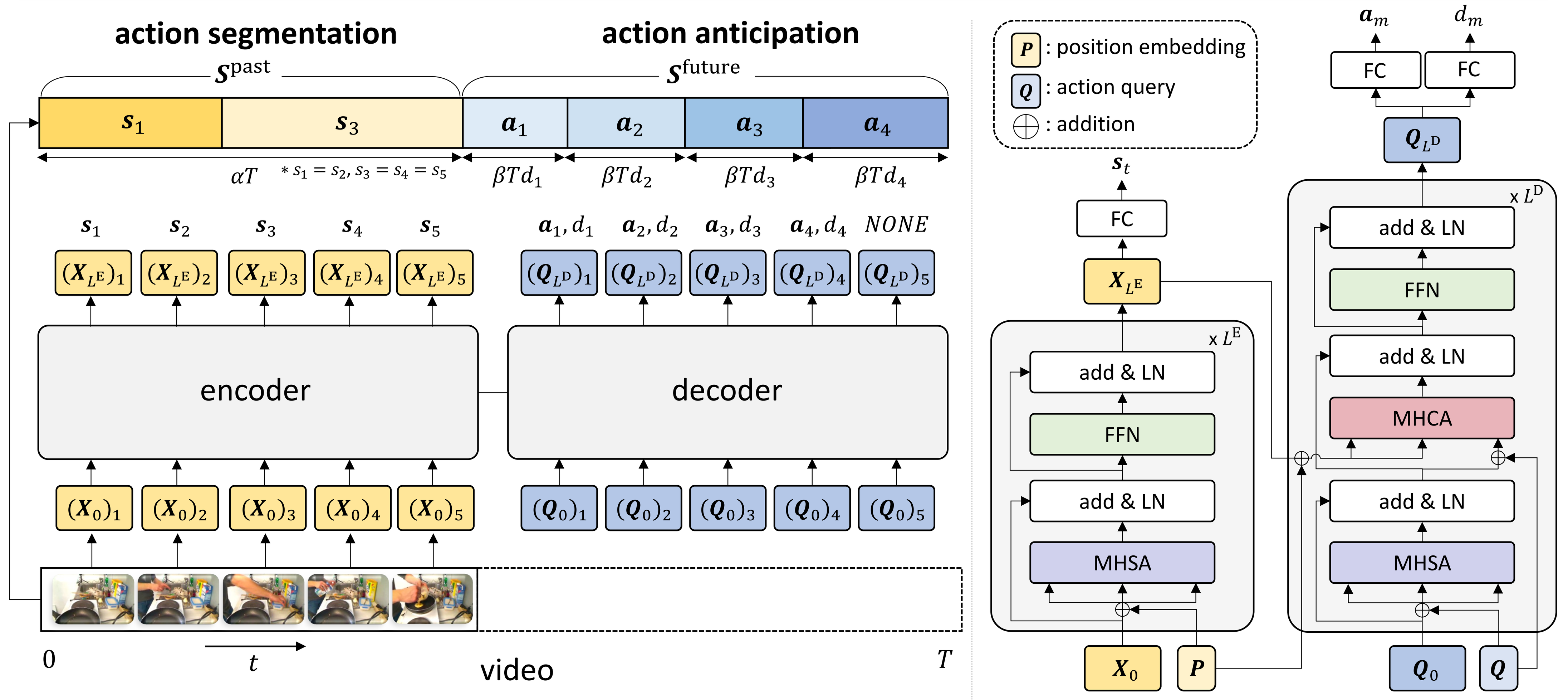
- Learning of long-term dependencies between past and future actions
- Inferring classes and durations of future actions

Contributions



- **An end-to-end attention neural network**, dubbed FUTR
- **Action segmentation (encoder)** & **action anticipation (decoder)**
- **Leveraging fine-grained visual features** of past actions
- **Long-term dependency modeling** between past and future actions
- **Parallel anticipation** for accurate and faster inference
- **State of the art** on long-term action anticipation benchmarks

Future Transformer (FUTR)



Action segmentation (Encoder)

- Input: observed frames (X_0)
- Output: logits of action segmentation
- Learning distinctive feature representations between past actions via self-attention
- L_{seg} : cross-entropy loss for action segmentation

Action anticipation (Decoder)

- Input: action query (Q_0) & output of the encoder (X_{L^E})
- Output: logits of action anticipation & predicted duration
- Learning long-term dependencies between actions via self-attention and cross-attention
- $L_{\text{anticipate}}$: Cross-entropy loss for future action anticipation
- L_{duration}: L2 loss for duration prediction

Advantages of FUTR

- Faster inference time
- No error accumulation
- Long-term relations of past and future actions
- Utilizing fine-grained visual features of past actions

Acknowledgement

This research was supported by NCSOFT, the IITP grant funded by MSIT (No.2019-0-01906, AI Graduate School Program - POSTECH), and the Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

Experimental Results

Comparison with the state of the arts

dataset	methods	β ($\alpha=0.2$) 0.1	0.2	0.3	0.5	β ($\alpha=0.3$) 0.1	0.2	0.3	0.5
Breakfast	CNN (Farha et al., CVPR18')	12.78	11.62	11.21	10.27	17.72	16.87	15.48	14.09
	Temporal Agg. (Sener et al., ECCV20')	24.20	21.10	20.00	18.10	30.40	26.30	23.80	21.20
	Cycle Cons. (Farha et al., GCPR20')	25.88	23.42	22.42	21.54	29.66	27.37	25.58	25.20
	FUTR (ours)	**27.70**	**24.55**	**22.83**	**22.04**	**32.27**	**29.88**	**27.49**	**25.87**
50 Salads	Temporal Agg. (Sener et al., ECCV20')	25.50	19.90	18.20	15.10	30.60	22.50	19.10	11.20
	Cycle Cons. (Farha et al., GCPR20')	34.76	**28.41**	21.82	15.25	34.39	23.70	18.95	**15.89**
	FUTR (ours)	**39.55**	27.54	**23.31**	**17.77**	**35.15**	24.86	**24.22**	15.26

* α: observation rates, β: prediction rates

Analysis

Parallel vs. auto-regressive decoding

method	AR	causal mask	β ($\alpha=0.3$) 0.1	0.2	0.3	0.5	time (ms)
FUTR-A	✓	✓	27.10	25.41	23.28	20.51	14.68
FUTR-M	-	✓	31.82	28.55	26.57	24.17	5.70
FUTR	-	-	**32.27**	**29.88**	**27.49**	**25.87**	**3.91**

- Parallel decoding is efficient and effective.
- Bi-directional attention of action query is crucial.

Effect of action segmentation loss

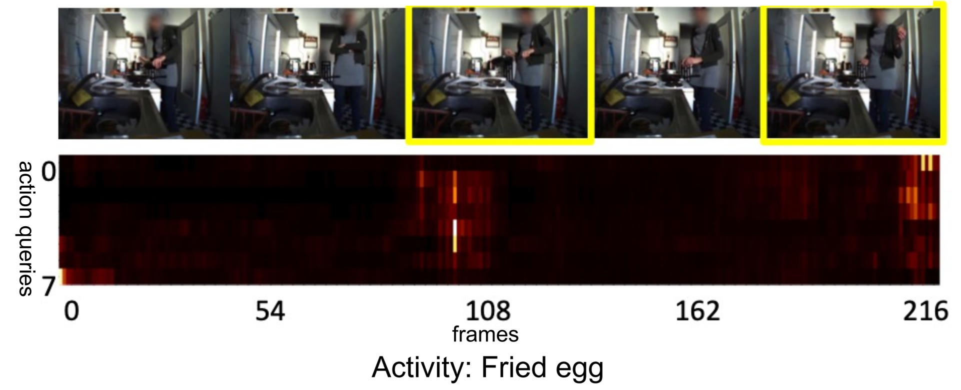
$\mathcal{L}^{\text{seg}}$	$\mathcal{L}^{\text{action}}$	$\mathcal{L}^{\text{duration}}$	β ($\alpha=0.3$) 0.1	0.2	0.3	0.5
-	✓	✓	28.31	25.85	24.91	22.50
✓	✓	✓	**32.27**	**29.88**	**27.49**	**25.87**

- Using action segmentation is effective.

Effect of long-term relations modeling

encoder	decoder	β ($\alpha=0.3$) 0.1	0.2	0.3	0.5
LSA	LSA	27.70	24.39	23.18	21.60
GSA	LSA	30.15	27.51	25.62	23.28
LSA	GSA	28.37	25.08	24.03	22.28
GSA	GSA	**32.27**	**29.88**	**27.49**	**25.87**

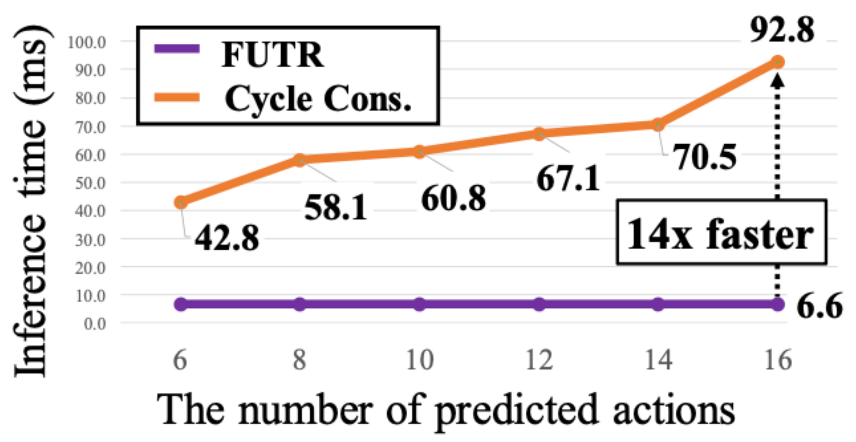
Cross-attention map visualization

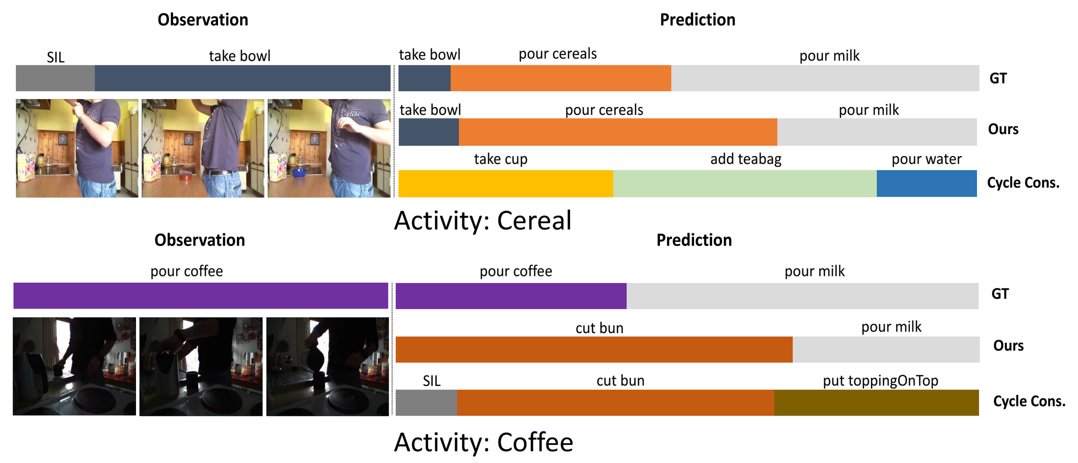


Activity: Fried egg

- Learning long-term relations between past and future action is crucial.
- FUTR detects important actions by using fine-grained visual features.

Comparison with the existing model





- FUTR is 14 x faster than Cycle Cons. (Farha et al. GCPR'20) when predicting 16 actions.
- FUTR utilizes fine-grained visual features without error accumulation.

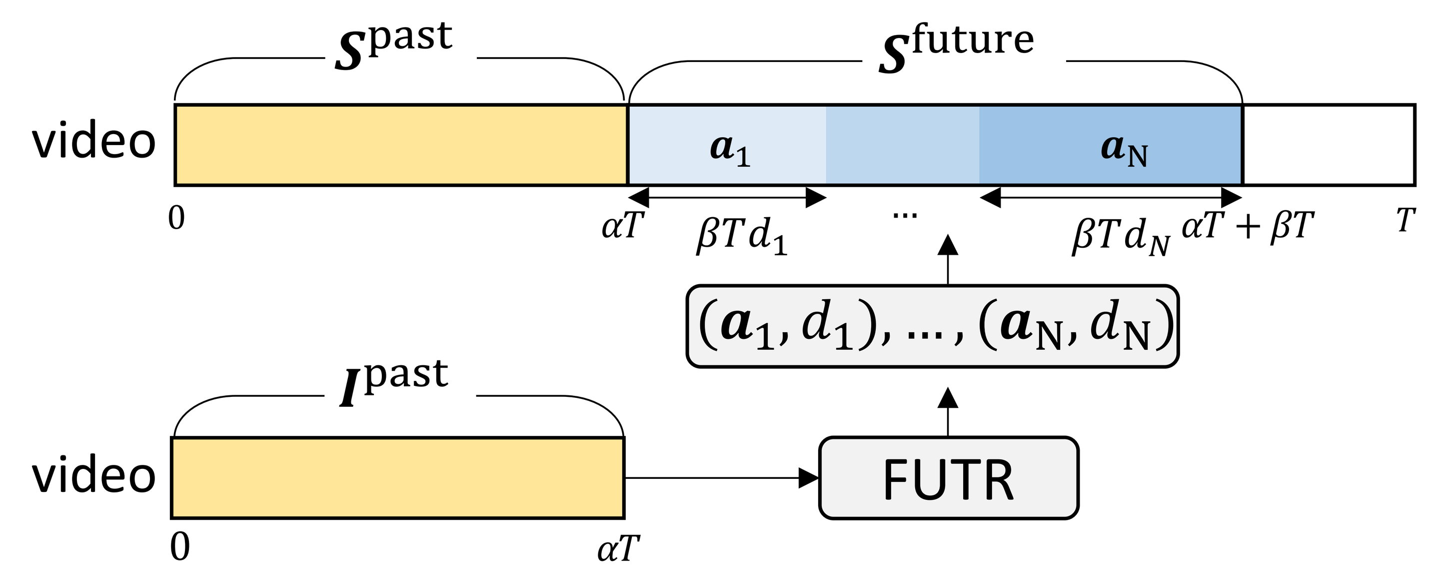
paper website

Future Transformer for Long-term Action Anticipation

Dayoung Gong[1] Joonseok Lee[1] Manjin Kim[1] Seong Jong Ha[2] Minsu Cho[1]
[1]Pohang University of Science and Technology (POSTECH) [2]NCSOFT

CVPR JUNE 19-24 2022 NEW ORLEANS • LOUISIANA

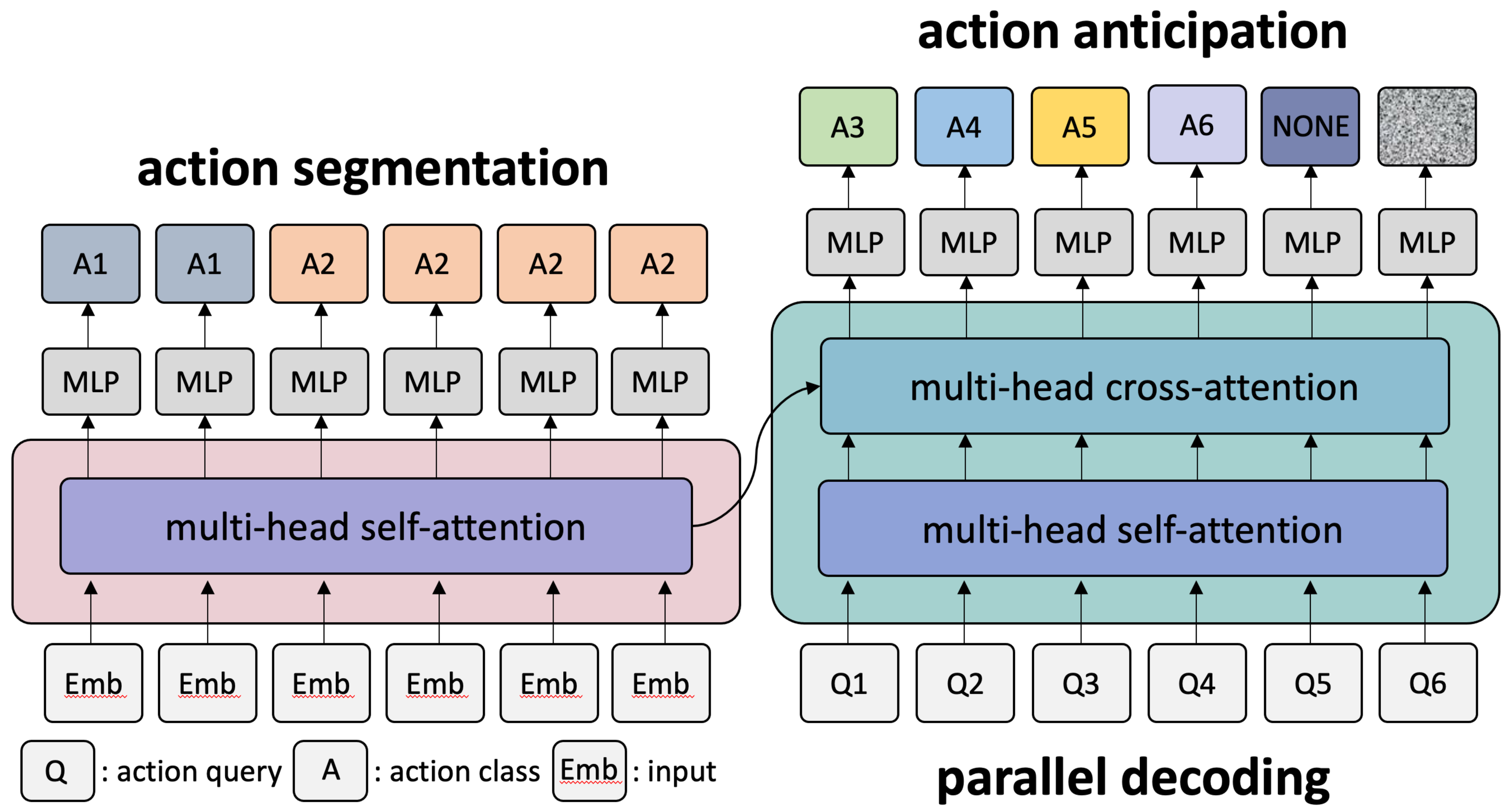
Long-term action anticipation



: Anticipating actions of future video frames with limited observation

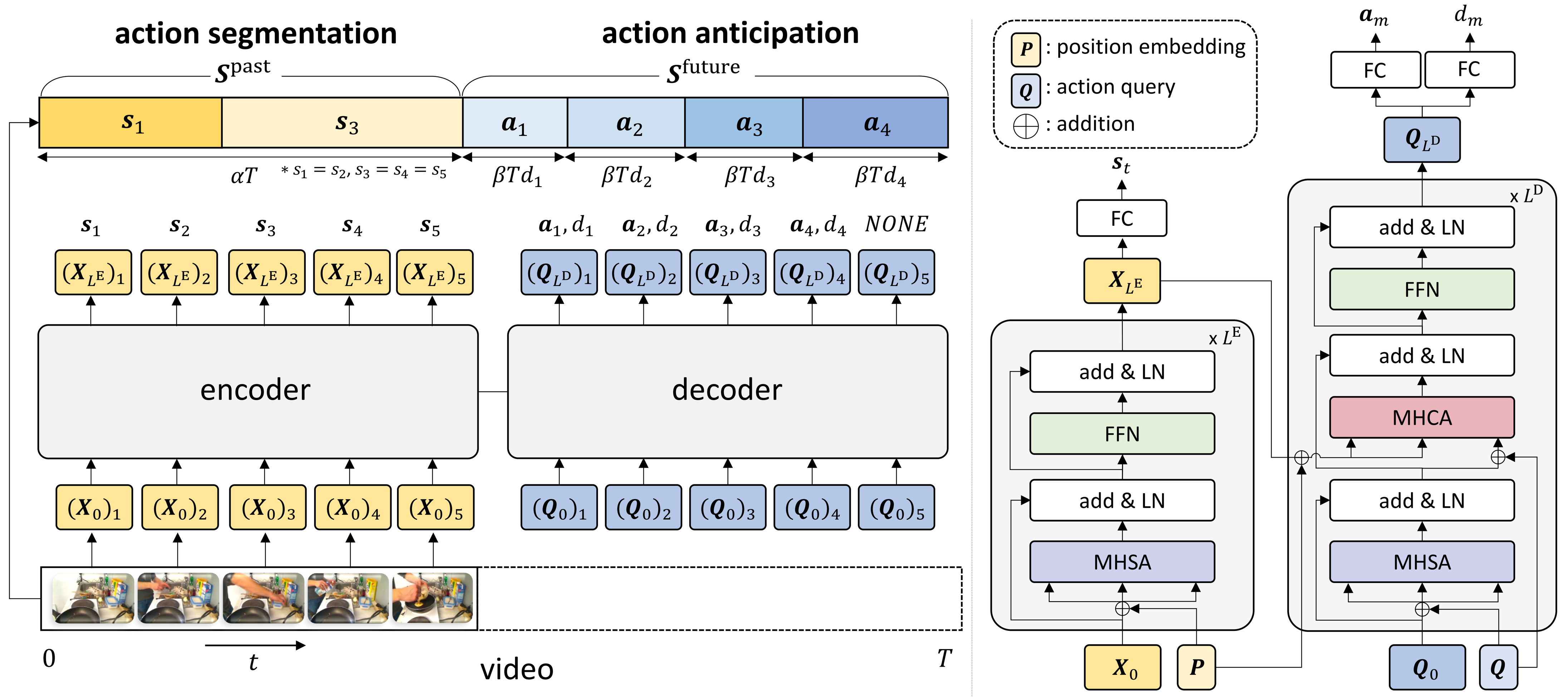
- Learning of long-term dependencies between past and future actions
- Inferring classes and durations of future actions

Contributions



- **An end-to-end attention neural network**, dubbed FUTR
- **Action segmentation (encoder) & action anticipation (decoder)**
- **Leveraging fine-grained visual features** of past actions
- **Long-term dependency modeling** between past and future actions
- **Parallel anticipation** for accurate and faster inference
- **State of the art** on long-term action anticipation benchmarks

Future Transformer (FUTR)



Action segmentation (Encoder)

- Input: observed frames (X_0)
- Output: logits of action segmentation
- Learning distinctive feature representations between past actions via self-attention
- L_{seg} : cross-entropy loss for action segmentation

Action anticipation (Decoder)

- Input: action query (Q_0) & output of the encoder (X_{L^E})
- Output: logits of action anticipation & predicted duration
- Learning long-term dependencies between actions via self-attention and cross-attention
- $L_{anticipate}$: Cross-entropy loss for future action anticipation
- $L_{duration}$: L2 loss for duration prediction

Advantages of FUTR

- Faster inference time
- No error accumulation
- Long-term relations of past and future actions
- Utilizing fine-grained visual features of past actions

Acknowledgements

- NCSOFT
- IITP grant funded by MSIT
- Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD

Experimental results

Comparison with the state of the arts

dataset	methods	β (α = 0.2) 0.1	0.2	0.3	0.5	β (α = 0.3) 0.1	0.2	0.3	0.5
Breakfast	CNN (Farha et al., CVPR18')	12.78	11.62	11.21	10.27	17.72	16.87	15.48	14.09
	Temporal Agg. (Sener et al., ECCV20')	24.20	21.10	20.00	18.10	30.40	26.30	23.80	21.20
	Cycle Cons. (Farha et al., GCPR20')	25.88	23.42	22.42	21.54	29.66	27.37	25.58	25.20
	FUTR (ours)	**27.70**	**24.55**	**22.83**	**22.04**	**32.27**	**29.88**	**27.49**	**25.87**
50 Salads	Temporal Agg. (Sener et al., ECCV20')	25.50	19.90	18.20	15.10	30.60	22.50	19.10	11.20
	Cycle Cons. (Farha et al., GCPR20')	34.76	**28.41**	21.82	15.25	34.39	23.70	18.95	**15.89**
	FUTR (ours)	**39.55**	27.54	**23.31**	**17.77**	**35.15**	24.86	**24.22**	15.26

* α: observation rates, β: prediction rates

Analysis

Parallel vs. auto-regressive decoding

method	AR	causal mask	β (α = 0.3) 0.1	0.2	0.3	0.5	time (ms)
FUTR-A	✓	✓	27.10	25.41	23.28	20.51	14.68
FUTR-M	-	✓	31.82	28.55	26.57	24.17	5.70
FUTR	-	-	**32.27**	**29.88**	**27.49**	**25.87**	**3.91**

- Parallel decoding is efficient and effective.
- Bi-directional attention of action query is crucial.

Effect of action segmentation loss

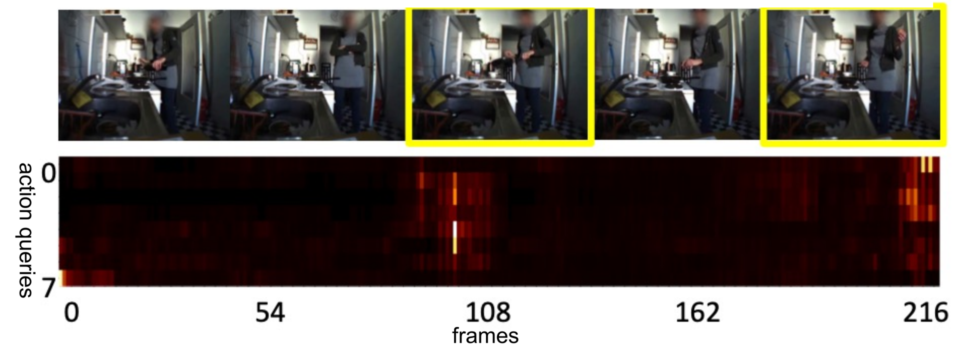
loss $\mathcal{L}^{seg}$	$\mathcal{L}^{action}$	$\mathcal{L}^{duration}$	β (α = 0.3) 0.1	0.2	0.3	0.5
-	✓	✓	28.31	25.85	24.91	22.50
✓	✓	✓	**32.27**	**29.88**	**27.49**	**25.87**

- Using action segmentation is effective.

Effect of long-term relations modeling

encoder	decoder	β (α = 0.3) 0.1	0.2	0.3	0.5
LSA	LSA	27.70	24.39	23.18	21.60
GSA	LSA	30.15	27.51	25.62	23.28
LSA	GSA	28.37	25.08	24.03	22.28
GSA	GSA	**32.27**	**29.88**	**27.49**	**25.87**

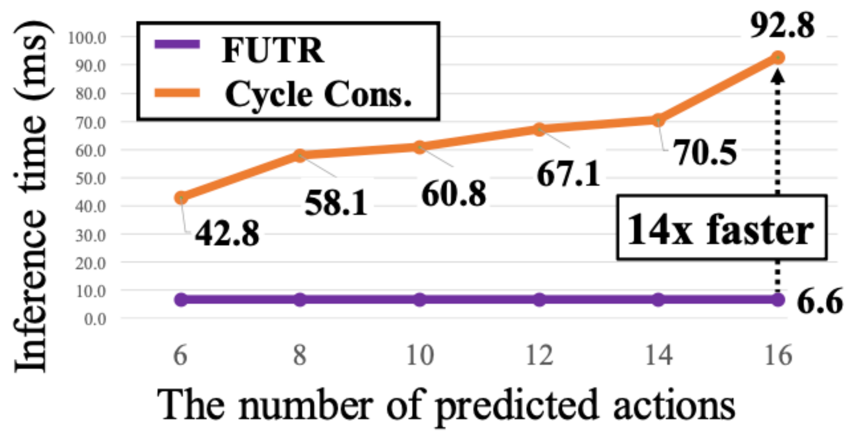
Cross-attention map visualization

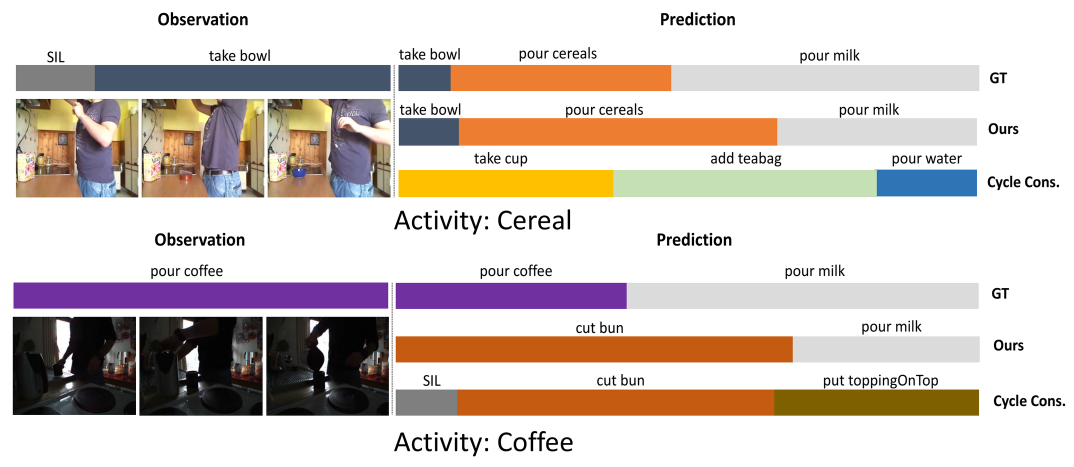


Activity: Fried egg

- Learning long-term relations between past and future action is crucial.
- FUTR detects important actions by using fine-grained visual features.

Comparison with the existing model





- FUTR is 14 x faster than Cycle Cons. (Farha et al. GCPR'20) when predicting 16 actions.
- FUTR utilizes fine-grained visual features without error accumulation.

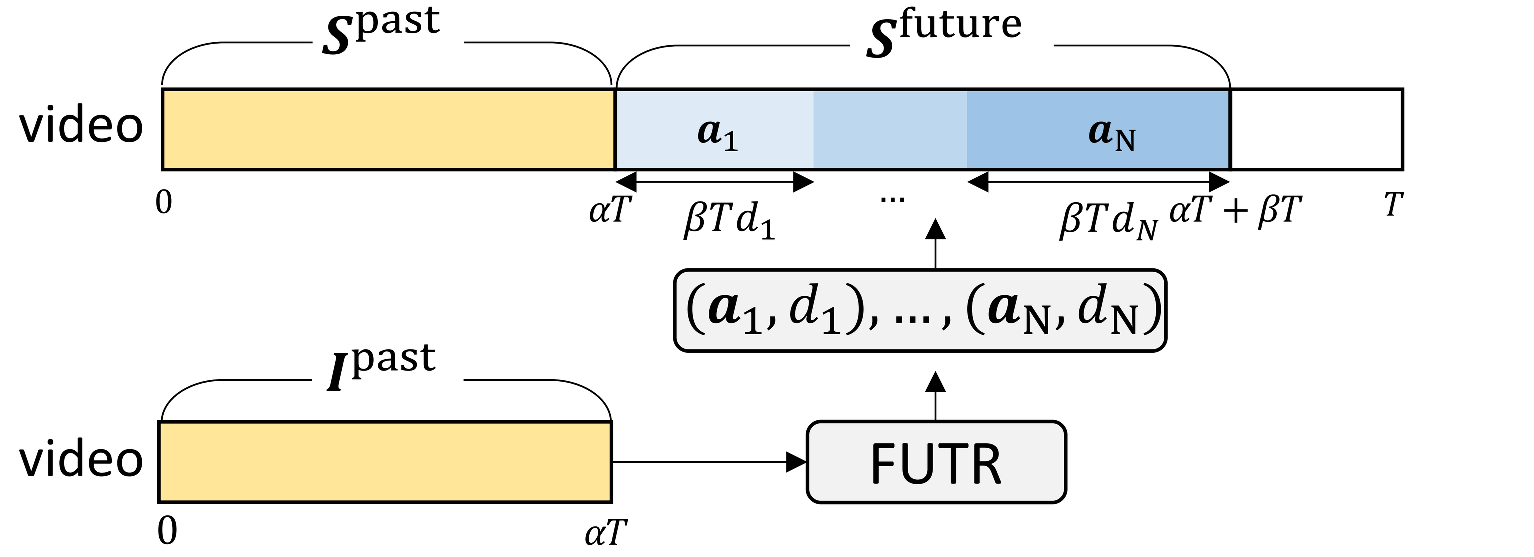
Future Transformer for Long-term Action Anticipation

paper website

Dayoung Gong[1] Joonseok Lee[1] Manjin Kim[1] Seong Jong Ha[2] Minsu Cho[1]

[1]Pohang University of Science and Technology (POSTECH) [2]NCSOFT

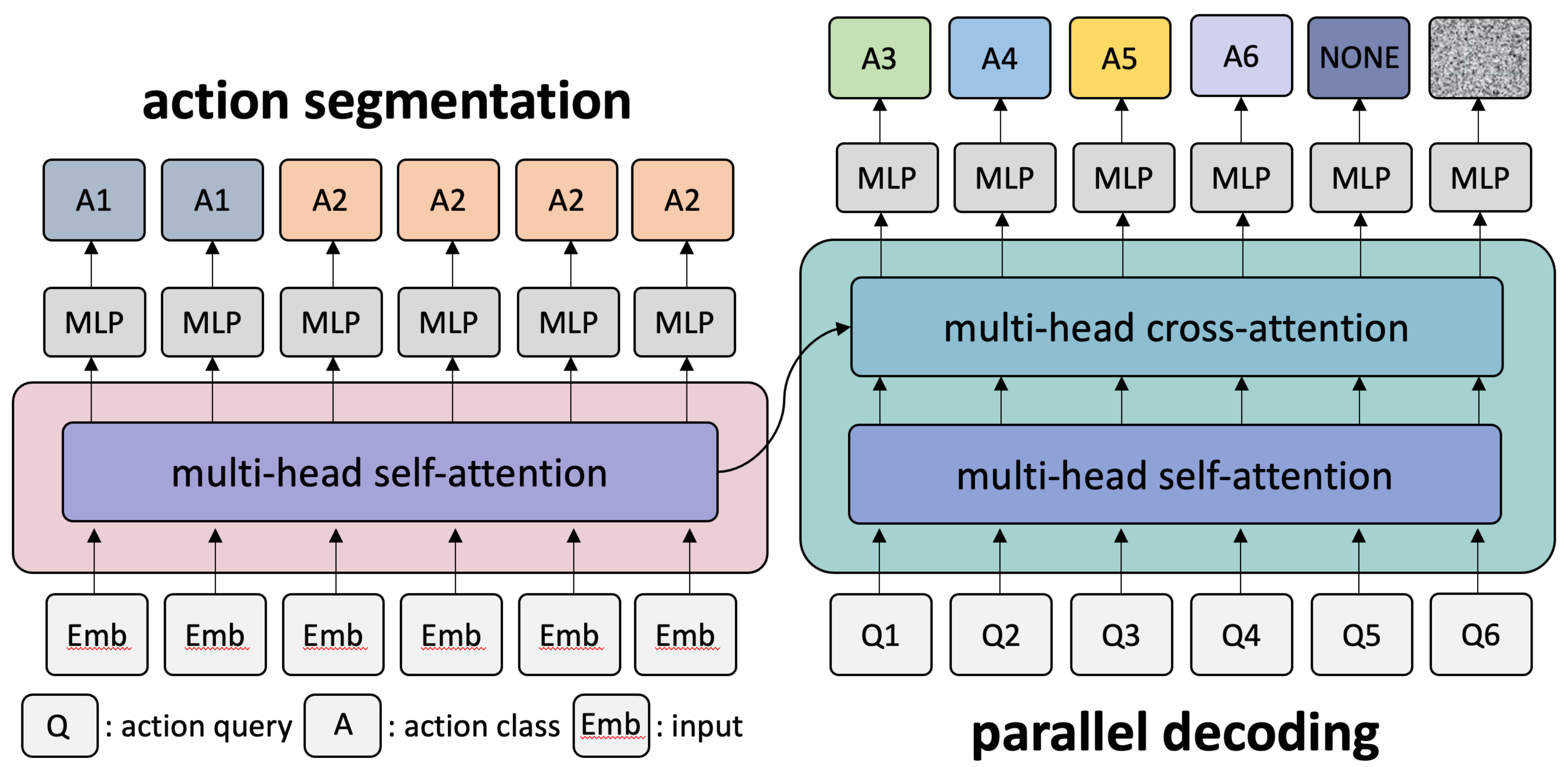
Long-term action anticipation



: Anticipating actions of future video frames with limited observation

- Learning of long-term dependencies between past and future actions
- Inferring classes and durations of future actions

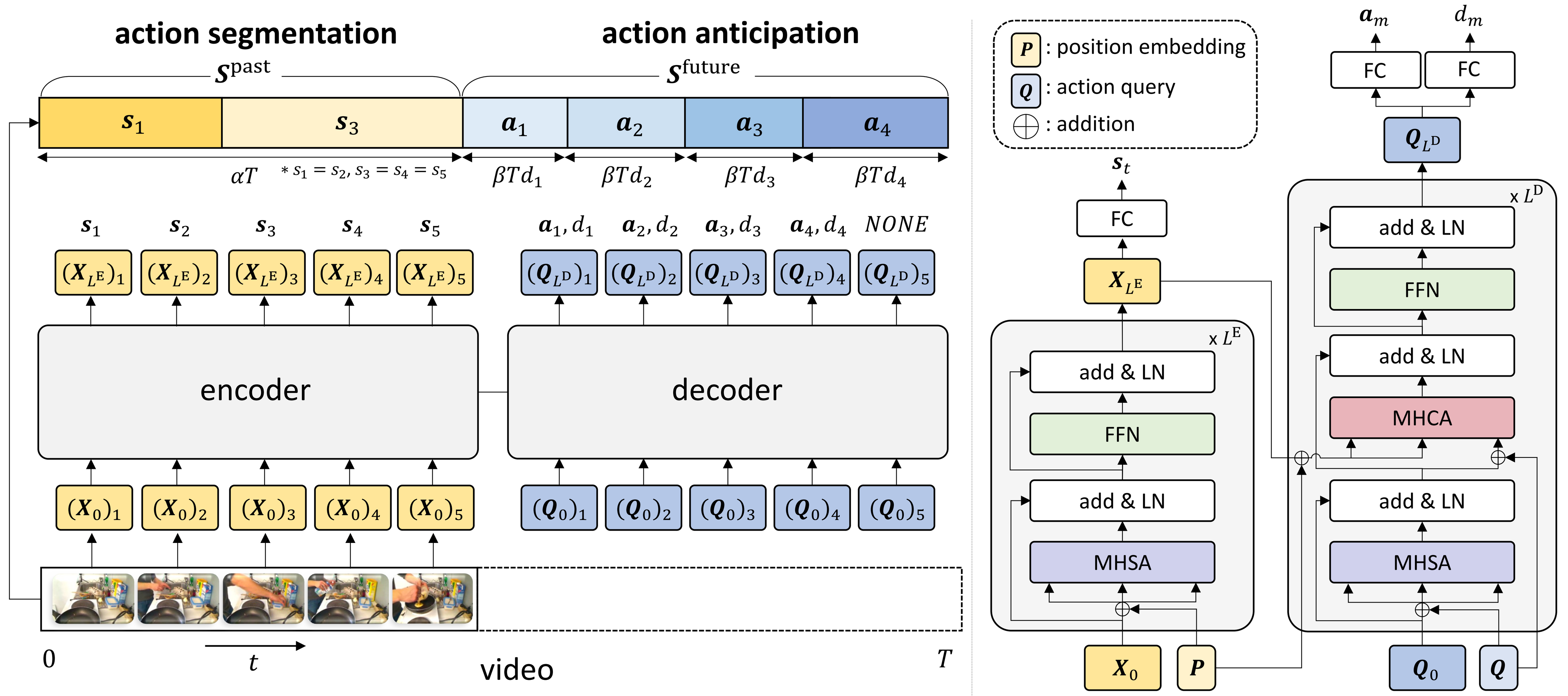
Contributions



parallel decoding

- **An end-to-end attention neural network**, dubbed FUTR
- **Action segmentation (encoder)** & **action anticipation (decoder)**
- **Leveraging fine-grained visual features** of past actions
- **Long-term dependency modeling** between past and future actions
- **Parallel anticipation** for accurate and faster inference
- **State of the art** on long-term action anticipation benchmarks

Future Transformer (FUTR)



Action segmentation (Encoder)

- Input: observed frames (X_0)
- Output: logits of action segmentation
- Learning distinctive feature representations between past actions via self-attention
- L_{seg} : cross-entropy loss for action segmentation

Action anticipation (Decoder)

- Input: action query (Q_0) & output of the encoder (X_{L^E})
- Output: logits of action anticipation & predicted duration
- Learning long-term dependencies between actions via self-attention and cross-attention
- $L_{anticipate}$: Cross-entropy loss for future action anticipation
- $L_{duration}$: L2 loss for duration prediction

Advantages of FUTR

- ✅ Faster inference time
- ✅ No error accumulation
- ✅ Long-term relations of past and future actions
- ✅ Utilizing fine-grained visual features of past actions

Acknowledgement

This research was supported by NCSOFT, the IITP grant funded by MSIT (No.2019-0-01906, AI Graduate School Program - POSTECH), and the Center for Applied Research in Artificial Intelligence (CARAI) grant funded by DAPA and ADD (UD190031RD).

Experimental Results

❑ Comparison with the state of the arts

dataset	methods	β ($\alpha=0.2$)				β ($\alpha=0.3$)			
		0.1	0.2	0.3	0.5	0.1	0.2	0.3	0.5
Breakfast	CNN (Farha et al., CVPR18')	12.78	11.62	11.21	10.27	17.72	16.87	15.48	14.09
	Temporal Agg. (Sener et al., ECCV20')	24.20	21.10	20.00	18.10	30.40	26.30	23.80	21.20
	Cycle Cons. (Farha et al., GCPR20')	25.88	23.42	22.42	21.54	29.66	27.37	25.58	25.20
	FUTR (ours)	**27.70**	**24.55**	**22.83**	**22.04**	**32.27**	**29.88**	**27.49**	**25.87**
50 Salads	Temporal Agg. (Sener et al., ECCV20')	25.50	19.90	18.20	15.10	30.60	22.50	19.10	11.20
	Cycle Cons. (Farha et al., GCPR20')	34.76	**28.41**	21.82	15.25	34.39	23.70	18.95	**15.89**
	FUTR (ours)	**39.55**	27.54	**23.31**	**17.77**	**35.15**	24.86	**24.22**	15.26

* α: observation rates, β: prediction rates

❑ Analysis

Parallel vs. auto-regressive decoding

method	AR	causal mask	β ($\alpha=0.3$)				time (ms)
			0.1	0.2	0.3	0.5	
FUTR-A	✓	✓	27.10	25.41	23.28	20.51	14.68
FUTR-M	-	✓	31.82	28.55	26.57	24.17	5.70
FUTR	-	-	**32.27**	**29.88**	**27.49**	**25.87**	**3.91**

- Parallel decoding is efficient and effective.
- Bi-directional attention of action query is crucial.

Effect of action segmentation loss

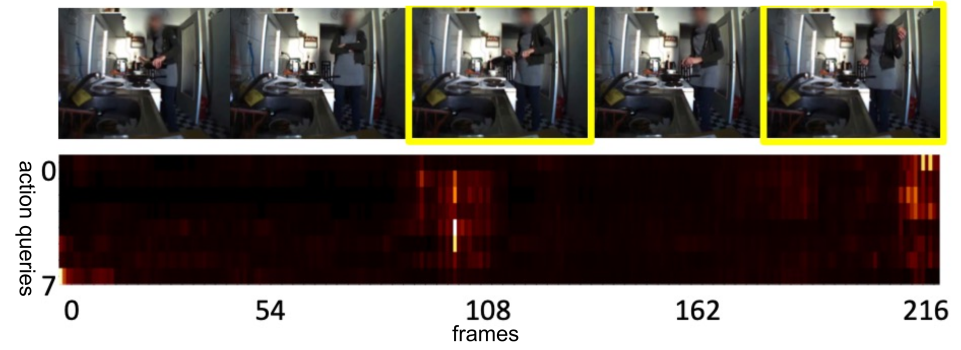
loss			β ($\alpha=0.3$)			
$\mathcal{L}^{seg}$	$\mathcal{L}^{action}$	$\mathcal{L}^{duration}$	0.1	0.2	0.3	0.5
-	✓	✓	28.31	25.85	24.91	22.50
✓	✓	✓	**32.27**	**29.88**	**27.49**	**25.87**

- Using action segmentation is effective.

Effect of long-term relations modeling

encoder	decoder	β ($\alpha=0.3$)			
		0.1	0.2	0.3	0.5
LSA	LSA	27.70	24.39	23.18	21.60
GSA	LSA	30.15	27.51	25.62	23.28
LSA	GSA	28.37	25.08	24.03	22.28
GSA	GSA	**32.27**	**29.88**	**27.49**	**25.87**

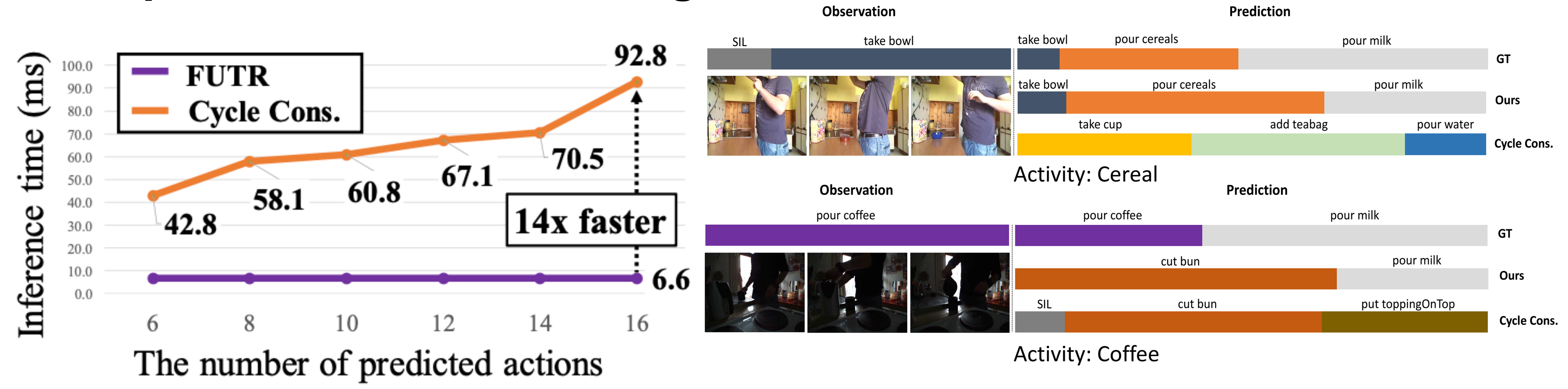
Cross-attention map visualization



Activity: Fried egg

- Learning long-term relations between past and future action is crucial.
- FUTR detects important actions by using fine-grained visual features.

❑ Comparison with the existing model



- FUTR is 14 x faster than Cycle Cons. (Farha et al. GCPR'20) when predicting 16 actions.
- FUTR utilizes fine-grained visual features without error accumulation.