TVT: Two-View Transformer Network for Video Captioning

Ming Chen, Yingming Li, Zhongfei Zhang, Siyu Huang

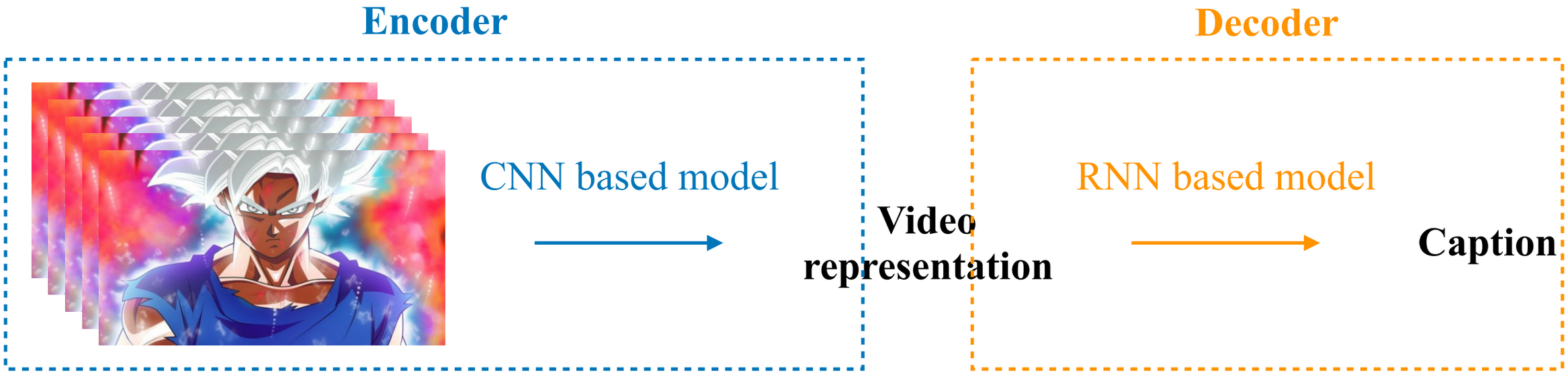
Published Date: Nov. 2018

WenWei

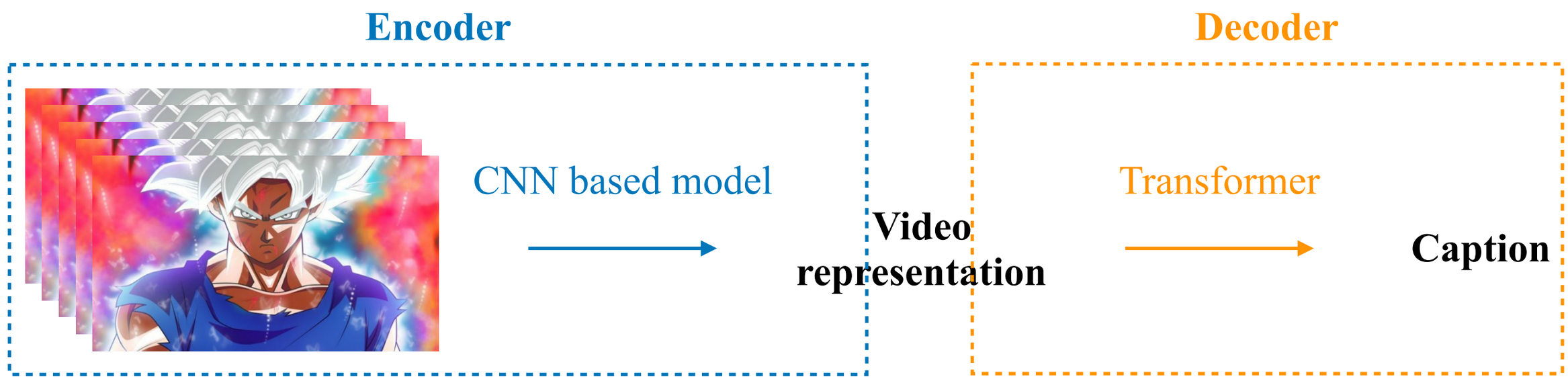
- Introduction
- Related Work
- Transformer
- Fusion Block
- Evaluation

Introduction

Previous video captioning:



Proposed video captioning:



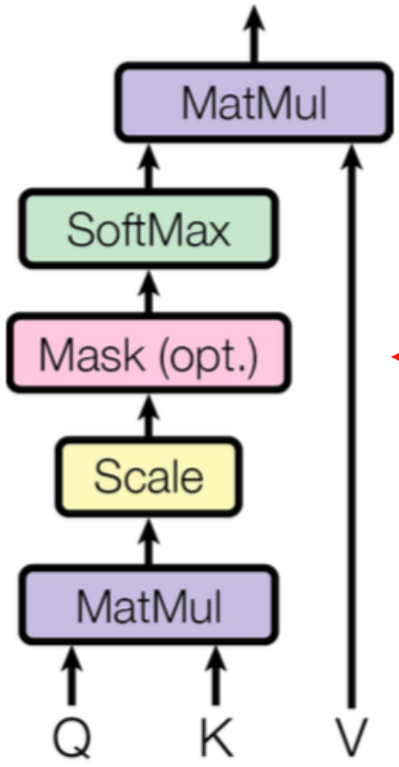
Related work

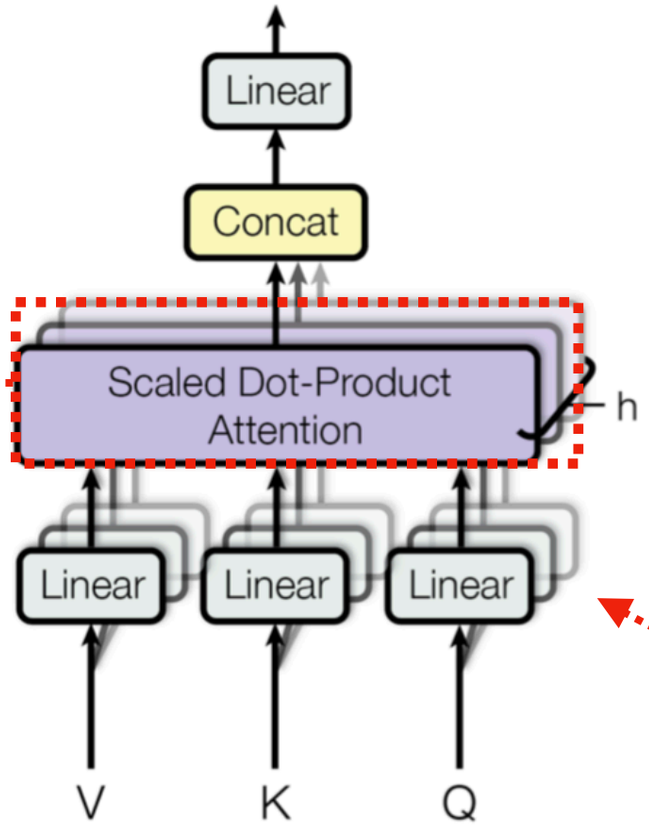
Advantage:

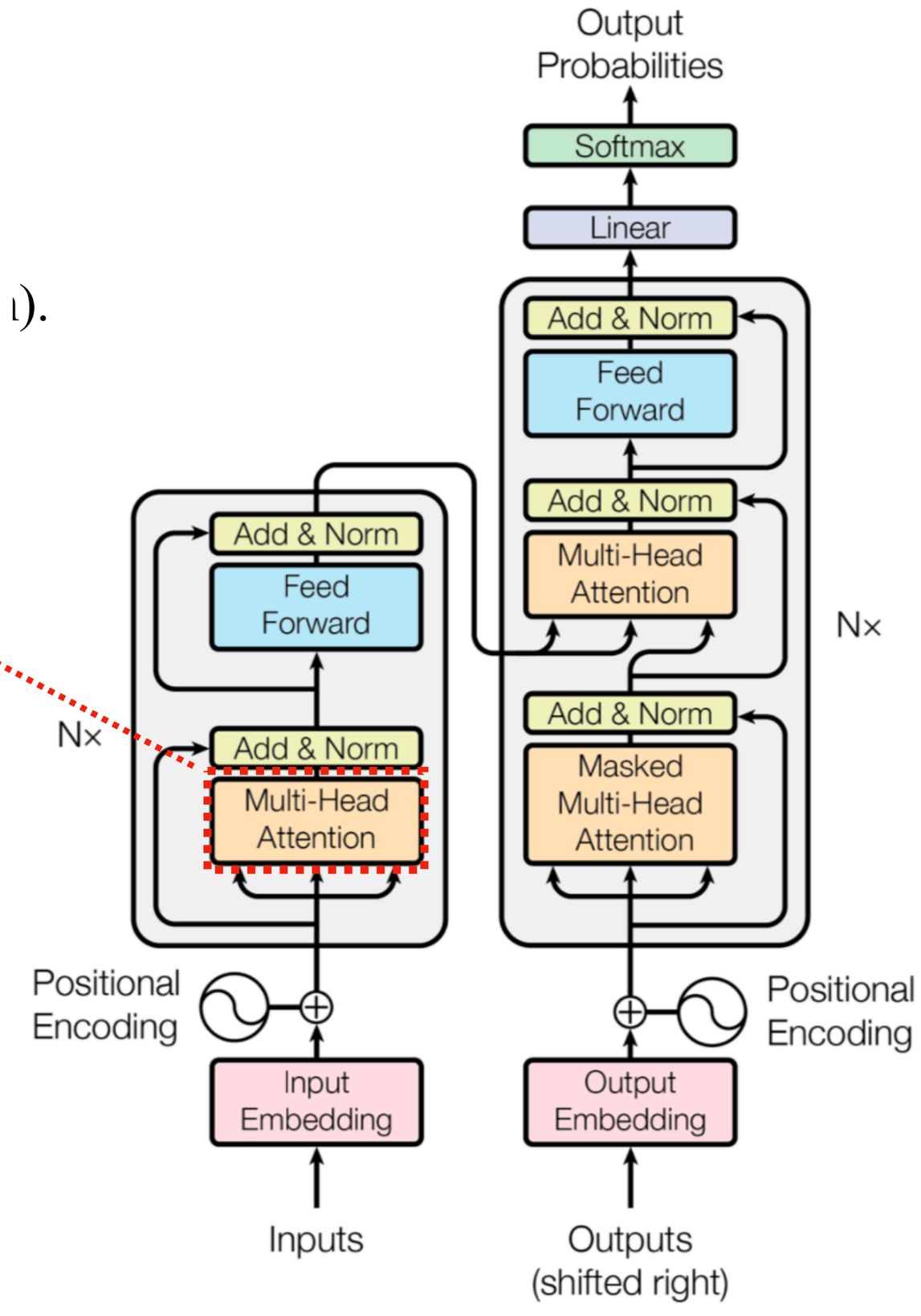
d Att

ing C ι).

nput



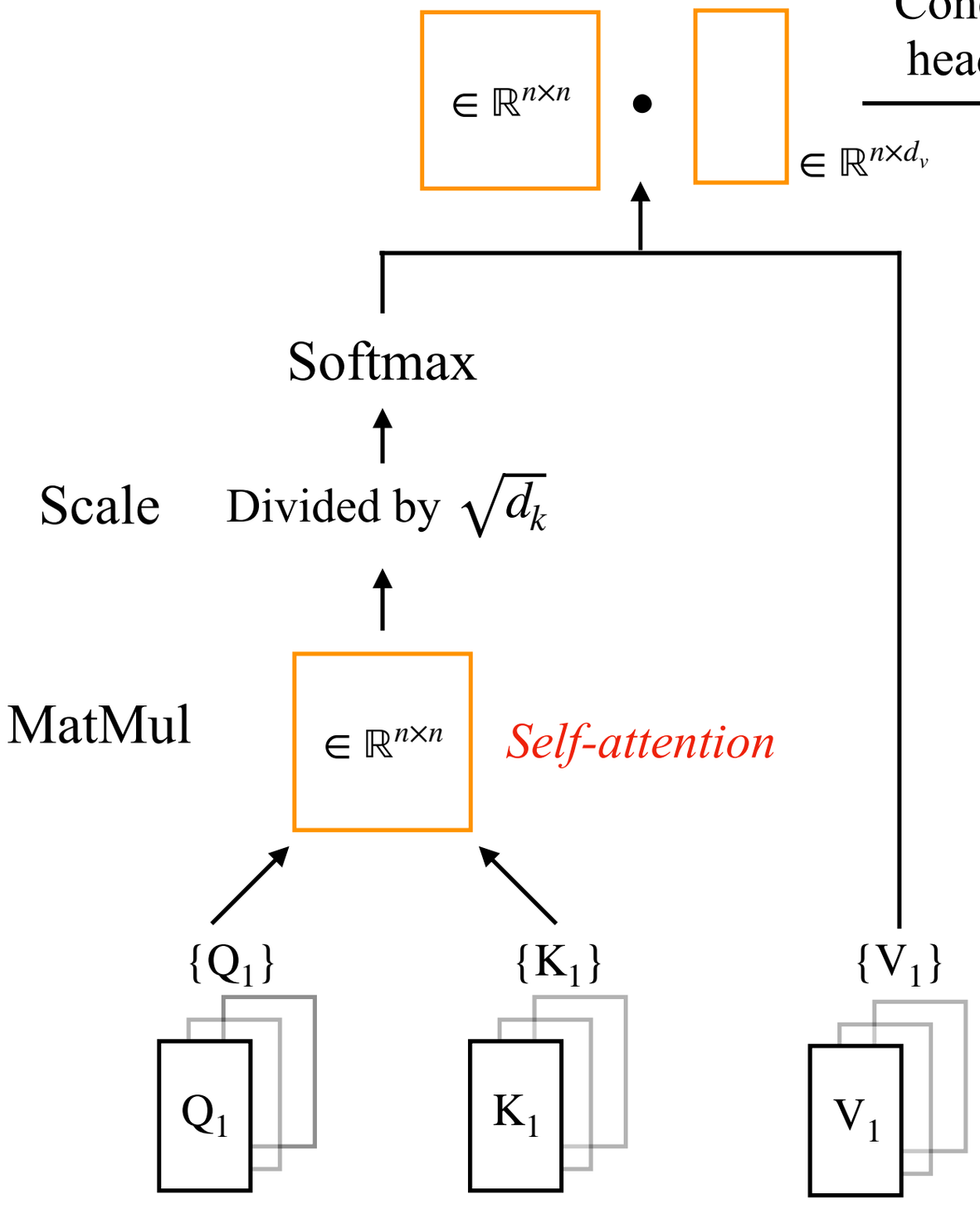




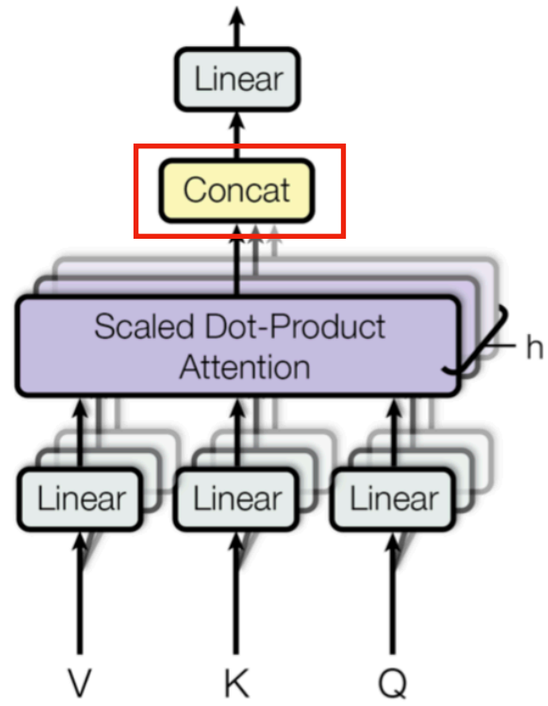
Resource: [Attention is all you need](#).

Related Work

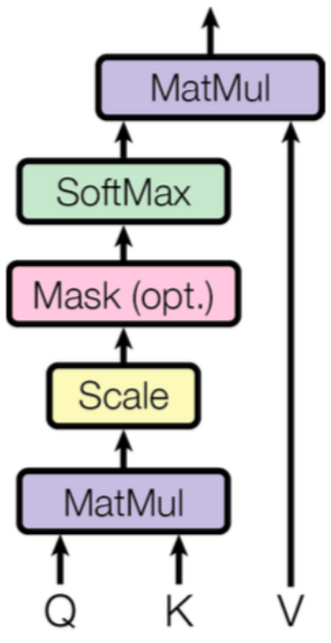
$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$



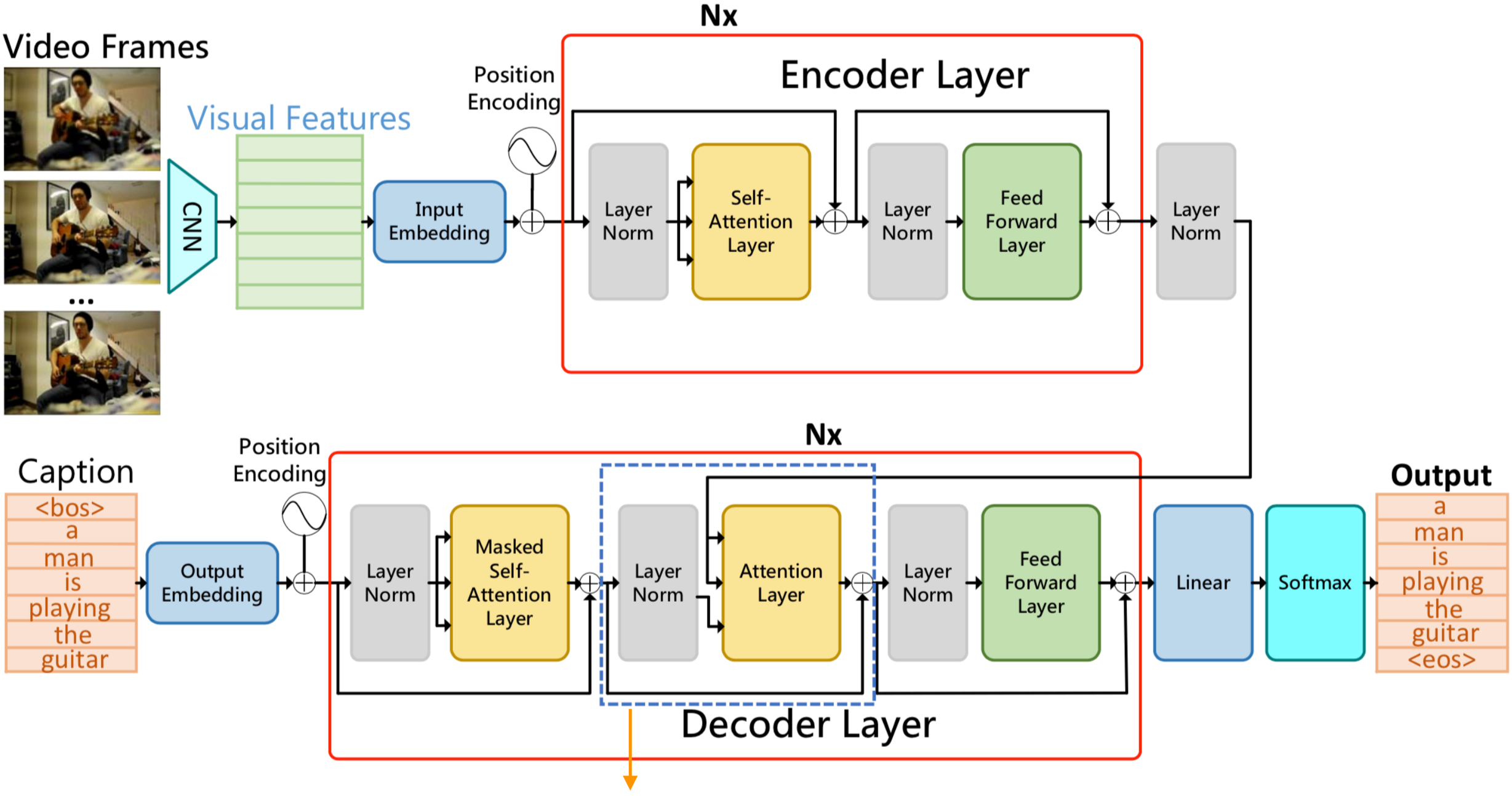
Multi-heads attention



Scaled Dot-Product Attention



Transformer

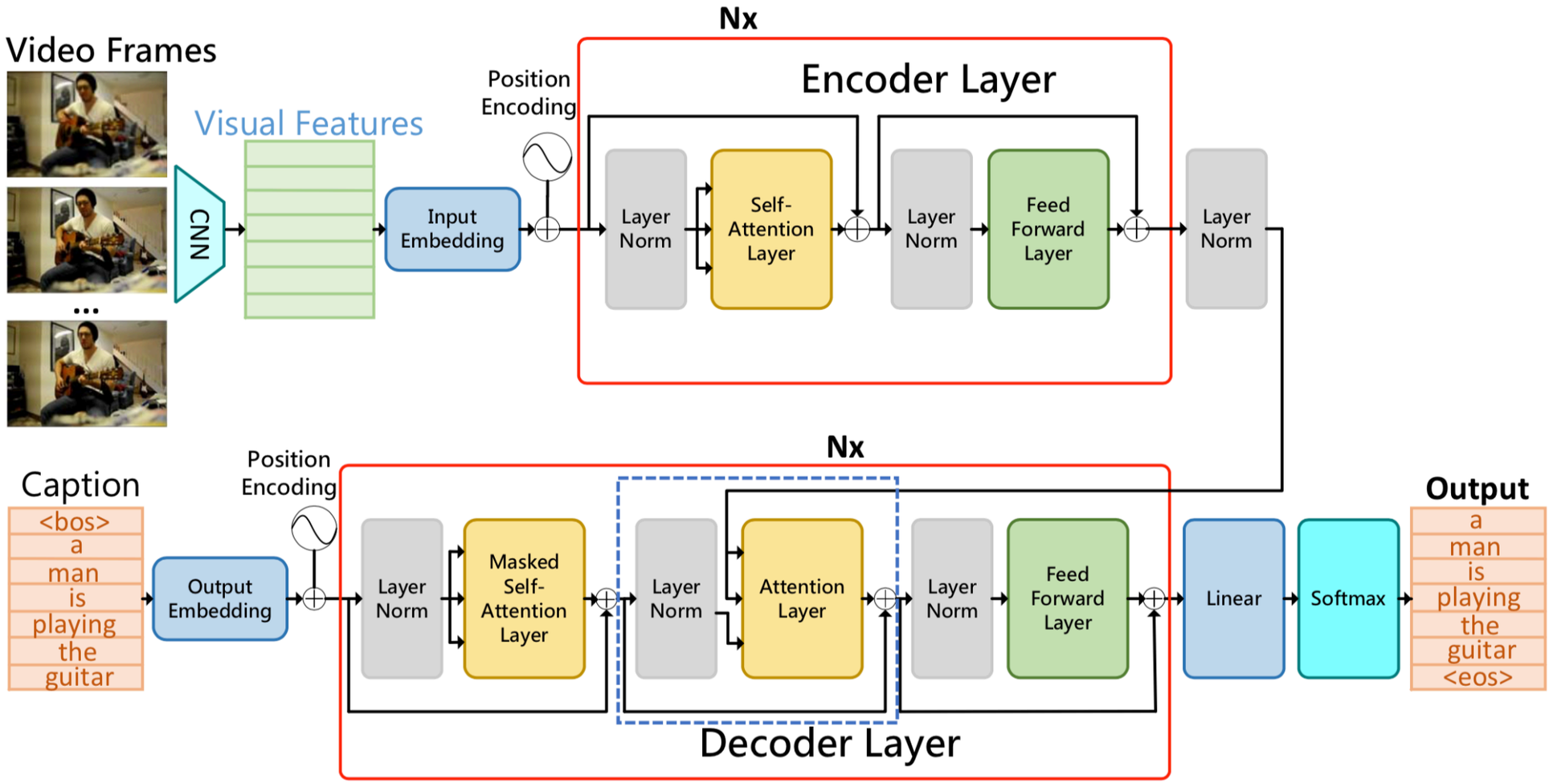


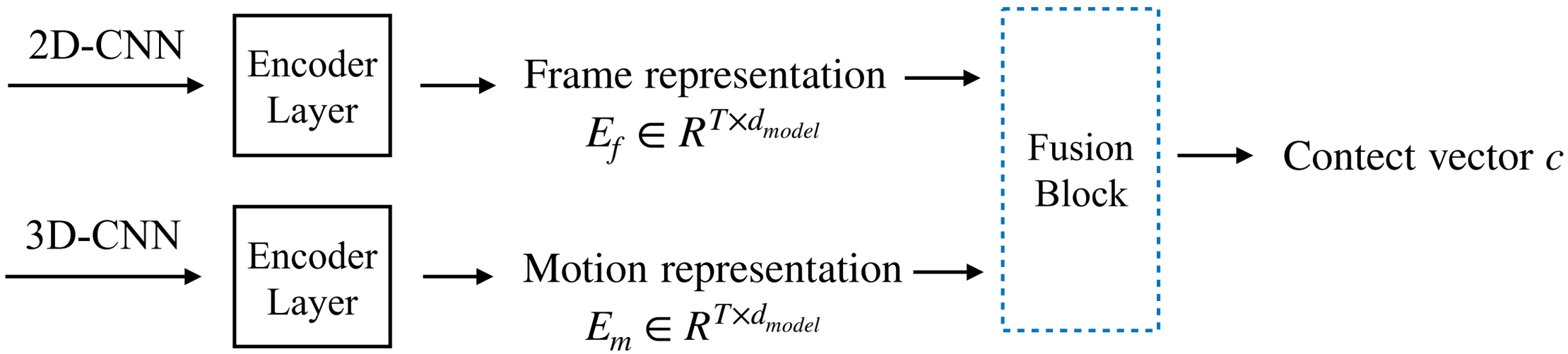
Fusion Block

How to combine **frame context** and **motion context**?

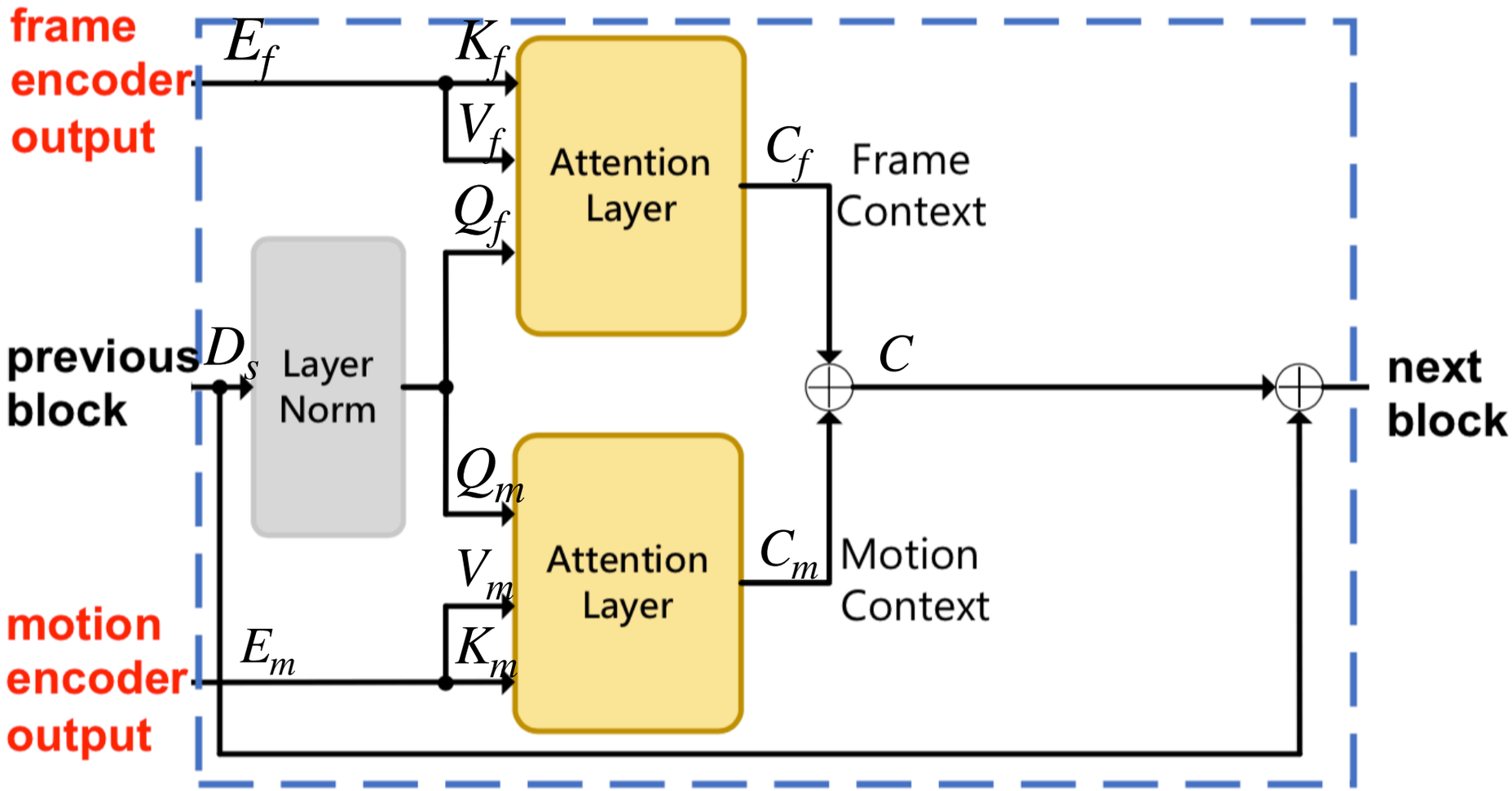
- Add-fusion block
- Attentive-fusion block

Transformer





Fusion Block



Add-fusion block

$$Q_f = \text{LayerNorm}(D_s)W_f^Q$$

$$K_f = E_f W_f^K$$

$$V_f = E_f W_f^V$$

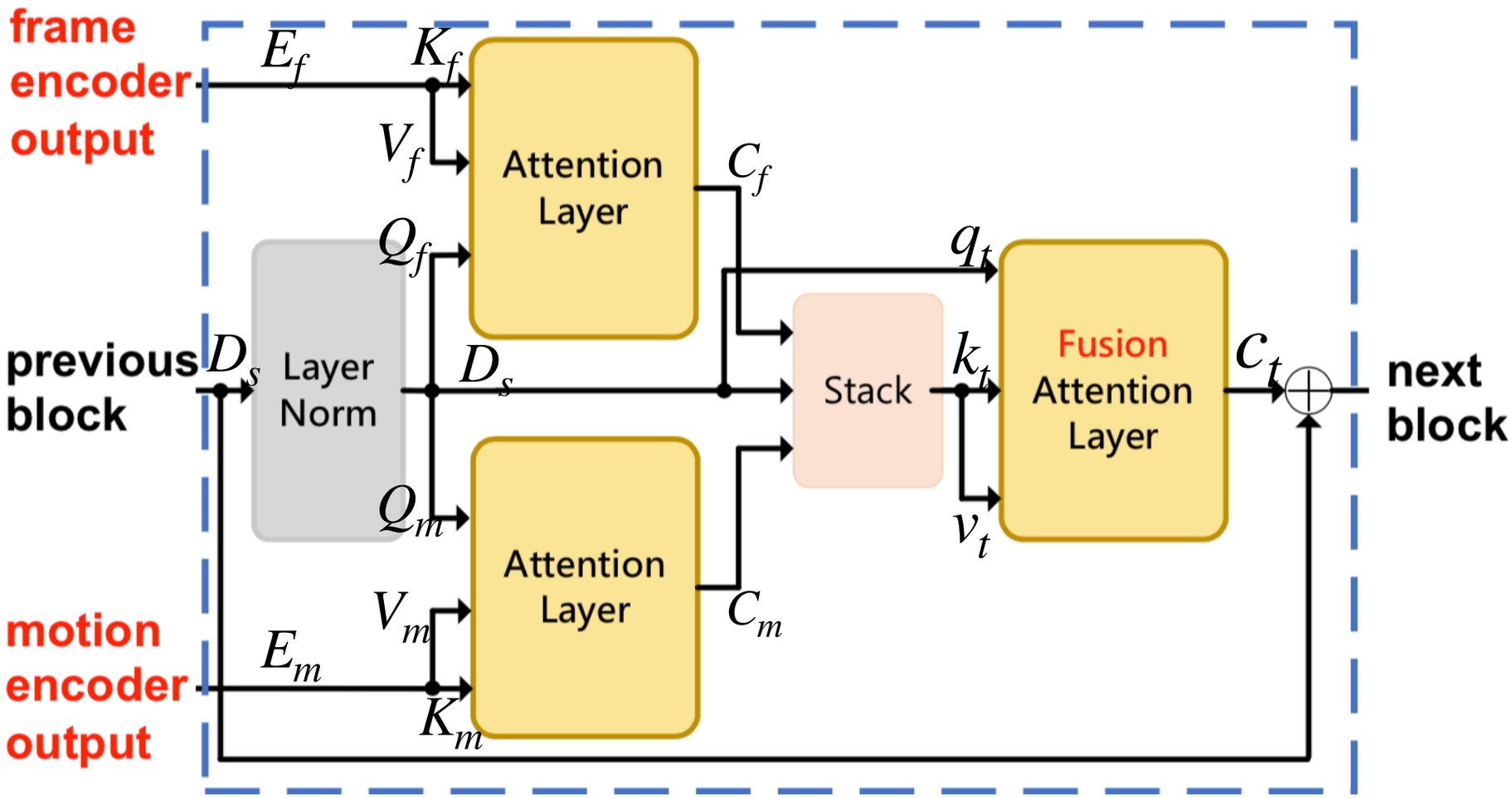
$$W_f^Q, W_f^K, W_f^V \in R^{d_{model} \times d_{model}}$$

$$C_f = \text{MultiHead}(Q_f, K_f, V_f)$$

$$C_m = \text{MultiHead}(Q_m, K_m, V_m)$$

★ $$C = \alpha C_f + (1 - \alpha) C_m$$

Fusion Block



Attentive-fusion block

$S_t = \text{Stack}(c_t^f, c_t^m, d_t)$, c_t^f, c_t^m, d_t are vectors of t – th column of C_f, C_m and D_s

$q_t = w^q d_t$

$K_t = s_t w^k$

$V_t = s_t w^v$

$c_t = \text{Attention}(q_t, K_t, V_t)$

Evaluation

MVSD(Microsoft Research Video Description):

- Image features: sample the videos at 5 *fps* and set the maximum number of frames as 50.
- Motion features: sample the videos at 25 *fps* and extract feature for every 64 frames.

MSR-VTT(Microsoft Research Video to Text):

- Image features: sample the videos at 3 *fps* and set the maximum number of frames as 60.
- Motion features: sample the videos at 15 *fps* and extract feature for every 64 frames.

2D-CNN model (Image features):

- RestNet-152 (2048-dimension)
- NasNet (4082-dimension)

3D-CNN model (Motion features):

- I3D (1024-dimension)

Evaluation

MVSD dataset

Models	BLEU@4	METEOR	CIDEr
LSTM-YT Venugopalan et al. (2015b)	33.29	29.07	-
S2VT Venugopalan et al. (2015a)	-	29.80	-
LSTM-I Dong et al. (2017)	44.60	29.70	-
SA Yao et al. (2015)	41.92	29.60	51.67
LSTM-E Pan et al. (2016b)	45.30	31.00	-
GRU-RCN Ballas et al. (2015)	43.26	31.60	68.01
h-RNN decoder Yu et al. (2016)	49.90	32.60	65.80
h-RNN encoder Pan et al. (2016a)	46.70	33.90	-
SCN-LSTM Gan et al. (2017)	51.10	33.50	77.70
TSA Pan et al. (2017)	52.80	33.50	74.00
M&M TGM Chen et al. (2017)	48.76	34.36	80.45
dualAFR Pu et al. (2018)	51.77	**36.41**	72.21
RecNet Wang et al. (2018a)	52.30	34.10	80.30
Att-TVT	**53.21**	35.23	**86.76**

Evaluation

Table 2: **Results on the MSR-VTT dataset.**

Models	BLEU	METEOR	ROUGE	CIDEr
VideoLAB Ramanishka et al. (2016)	39.10	27.70	60.60	44.10
Aalto Shetty and Laaksonen (2016)	39.80	26.90	59.80	45.70
v2t_navigator Jin et al. (2016)	40.80	28.20	60.90	44.80
CIDEnt-RL Pasunuru and Bansal (2017)	40.50	28.40	61.40	**51.70**
Dense-Cap Shen et al. (2017)	41.40	28.30	61.10	48.90
HRL Wang et al. (2018b)	41.30	**28.70**	**61.70**	48.00
Att-TVT	40.12	27.86	59.63	47.72
Att-TVT(+audio)	**42.46**	28.24	61.07	48.53

Evaluation

R: RestNet-152, N: NasNet, I: I3D

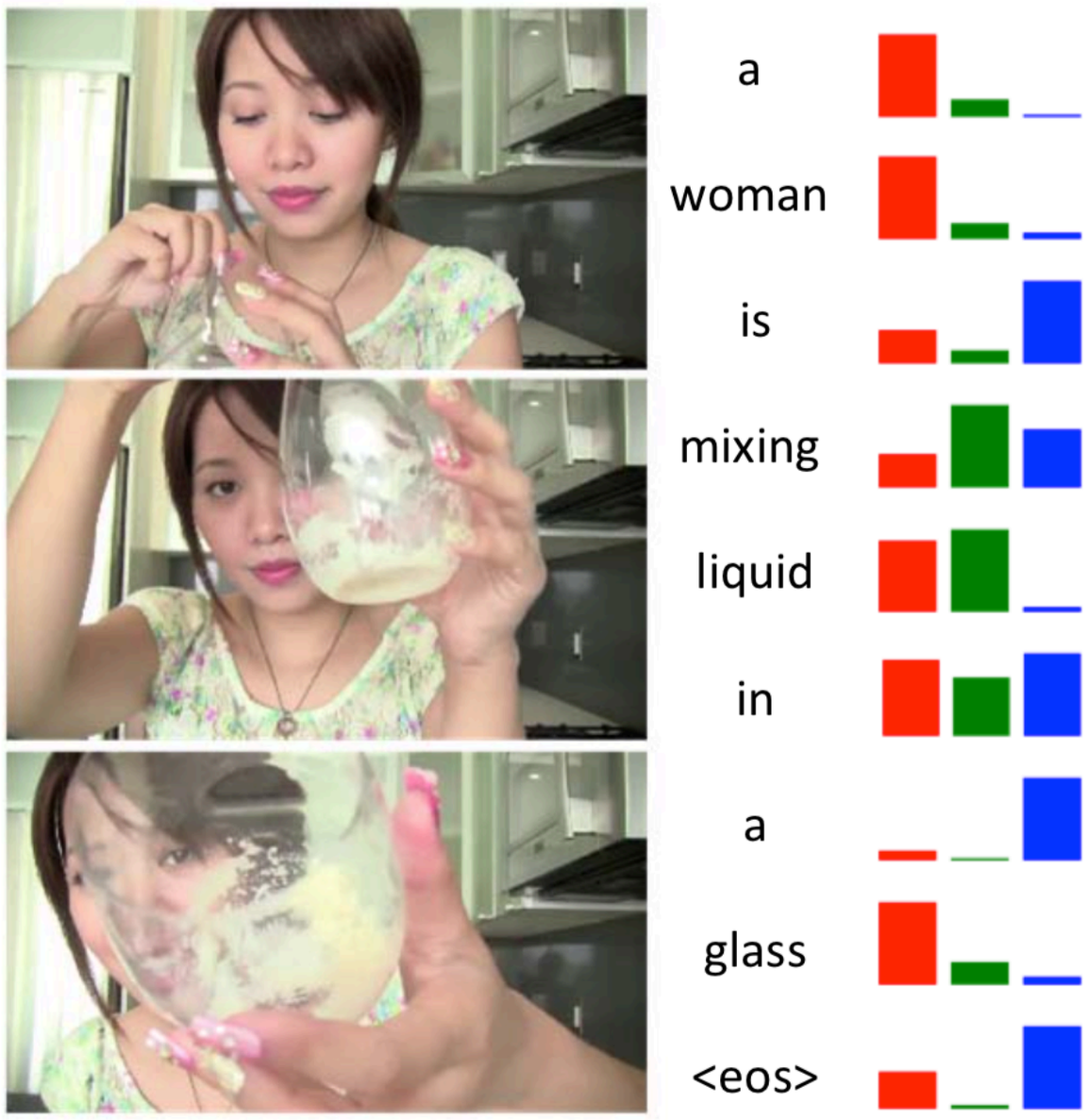
Base Model(R): zebras are eating.
Add-TVT(N+I): zebras are standing in a field.
Att-TVT(N+I): **zebras are playing with each other.**
GT: two zebras are playing with each other.

Base Model(R): a man and woman are singing.
Add-TVT(N+I): a man and woman are riding a bike.
Att-TVT(N+I): **a man and woman are riding a motorcycle.**
GT: a man and woman are riding a motorcycle.

Base Model(R): a man is playing a gun.
Add-TVT(N+I): the person is playing the music.
Att-TVT(N+I): **a group of people are playing the drums.**
GT: four men are playing musical instruments.

Base Model(R): a man is playing football.
Add-TVT(N+I): a group of men are fighting.
Att-TVT(N+I): **a man is doing martial arts.**
GT: a man is demonstrating martial arts.

Evaluation



- Red: frame representation.
- Green: motion representation.
- Blue: generated words.