X-Linear Attention Networks for Image Captioning

"a group of zebras grazing in a filed with a rainbow in the sky"

"two little girls eating donuts in a room"

"a group of skiers flying through the air while riding skis"

Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei

Vision and Multimedia Lab, JD AI Research
panyw.ustc@gmail.com

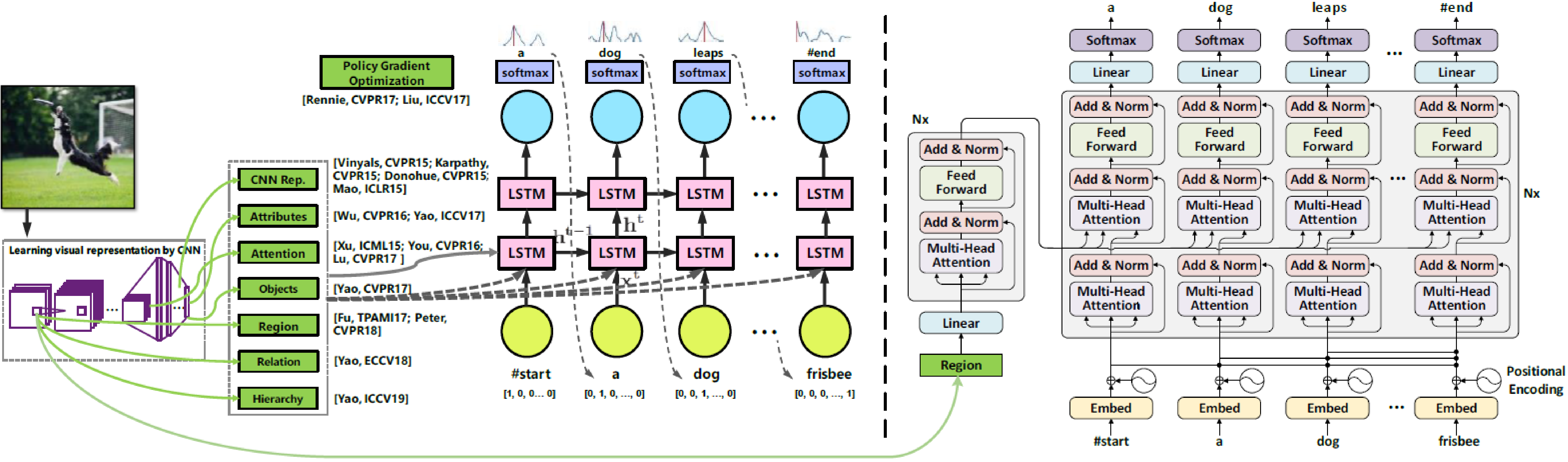
Code:

Mainstream: CNN Encoder + LSTM Decoder

[Google15, Stanford15, Berkeley15, Baidu/UCLA15, UdeM15, Rochester16, UAdelaide16, Virginia Tech17, THU17, MSR17&18, IBM17, U of Oxford & Google17, JD AI18&19]

Transformer-based encoder-decoder

[Sharma, ACL18]



(a) CNN encoder plus LSTM decoder

(b) Transformer-based encoder-decoder

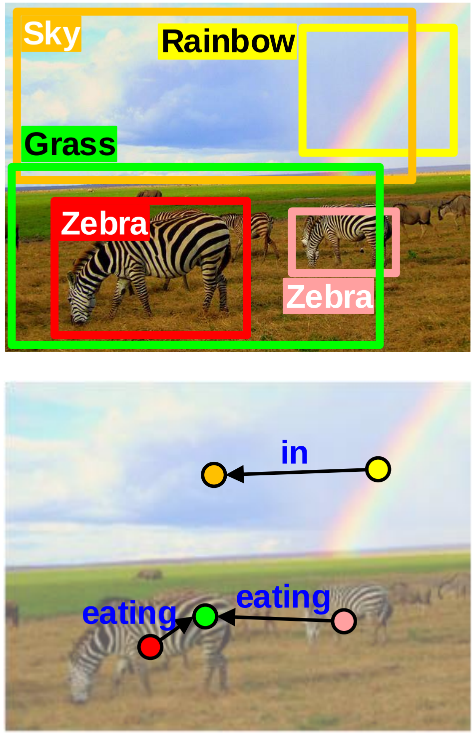
Phase I (past 5 years) – V/L independent

Enhance visual features with X

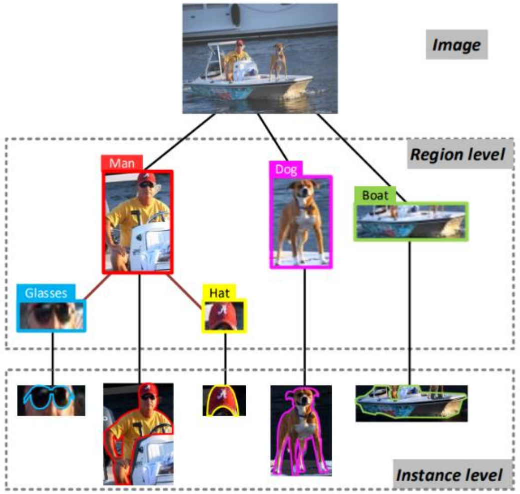
X = visual attributes
[You, CVPR16; Wu, CVPR16; Yao, ICCV17]

X = object / entity recognition
[Yao, CVPR17; Li, CVPR19]

X = region / relation
[Peter, CVPR18; Yao, ECCV18]



X = instance / hierarchy
[Yao, ICCV19]

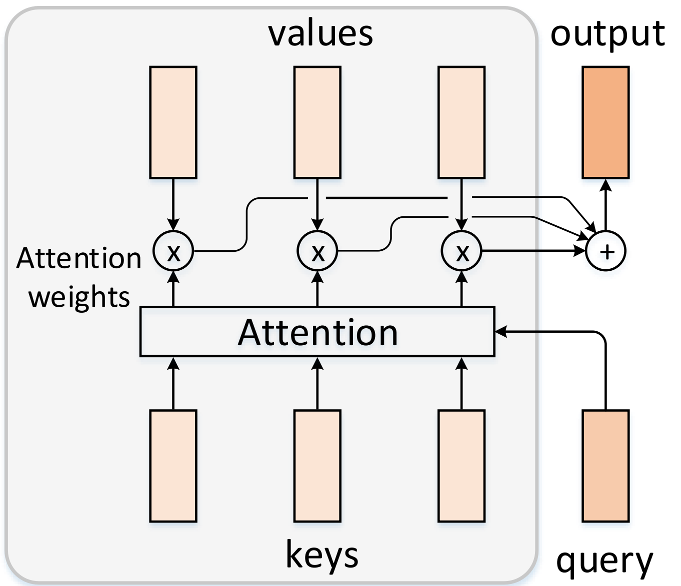


Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei, "Boosting Image Captioning with Attributes." In ICCV, 2017.
Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei, "Incorporating Copying Mechanism in Image Captioning for Learning Novel Objects." In CVPR, 2017.
Ting Yao, Yingwei Pan, Yehao Li and Tao Mei. "Exploring Visual Relationship for Image Captioning." In ECCV, 2018.
Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. "Pointing Novel Objects in Image Captioning." In CVPR, 2019.
Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei, "Hierarchy Parsing for Image Captioning." In ICCV, 2019.
Quanzeng You, et al. "Image captioning with semantic attention." In CVPR, 2016.
Anderson Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." In CVPR, 2018.

Phase II (present) – V/L interacted

Integrate encoder/decoder via X attention mechanism

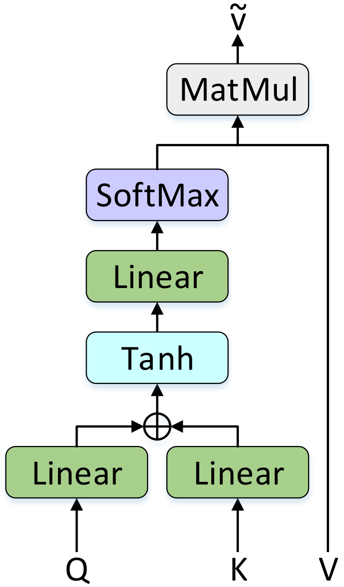
Memory (key-value pairs)



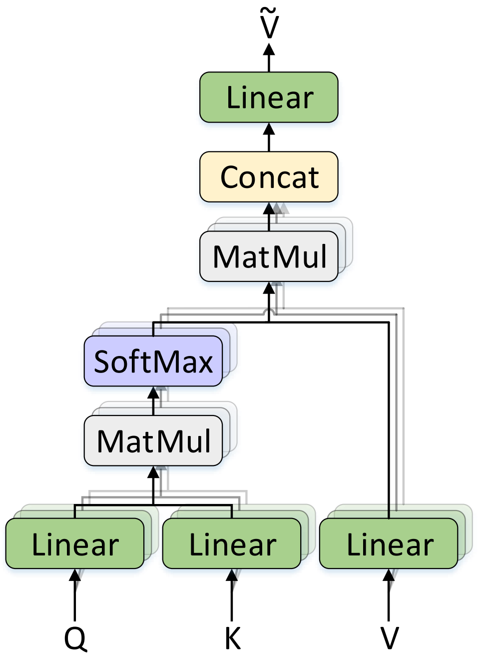
Query (Q): Hidden state from language decoder

Keys (K) =Values (V): Region-level representations from image encoder

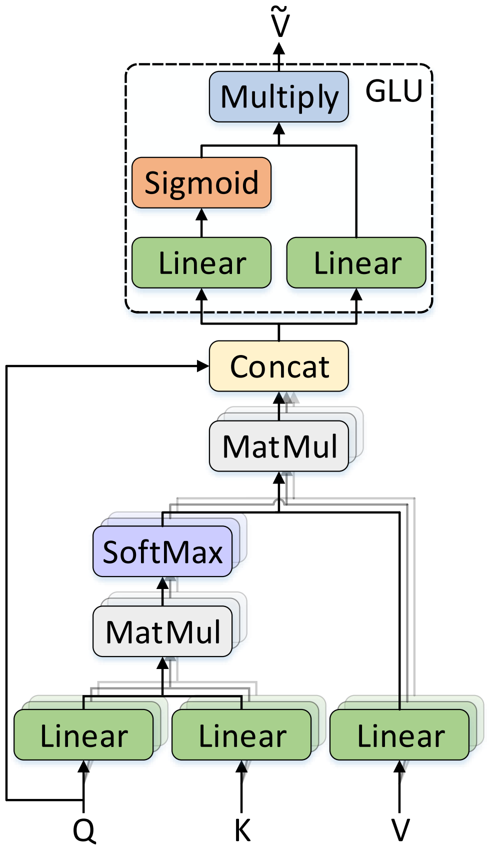
X = visual/top-down attention
[Xu, ICML15; Peter, CVPR18]



X = multi-head attention
[Sharma, ACL18]



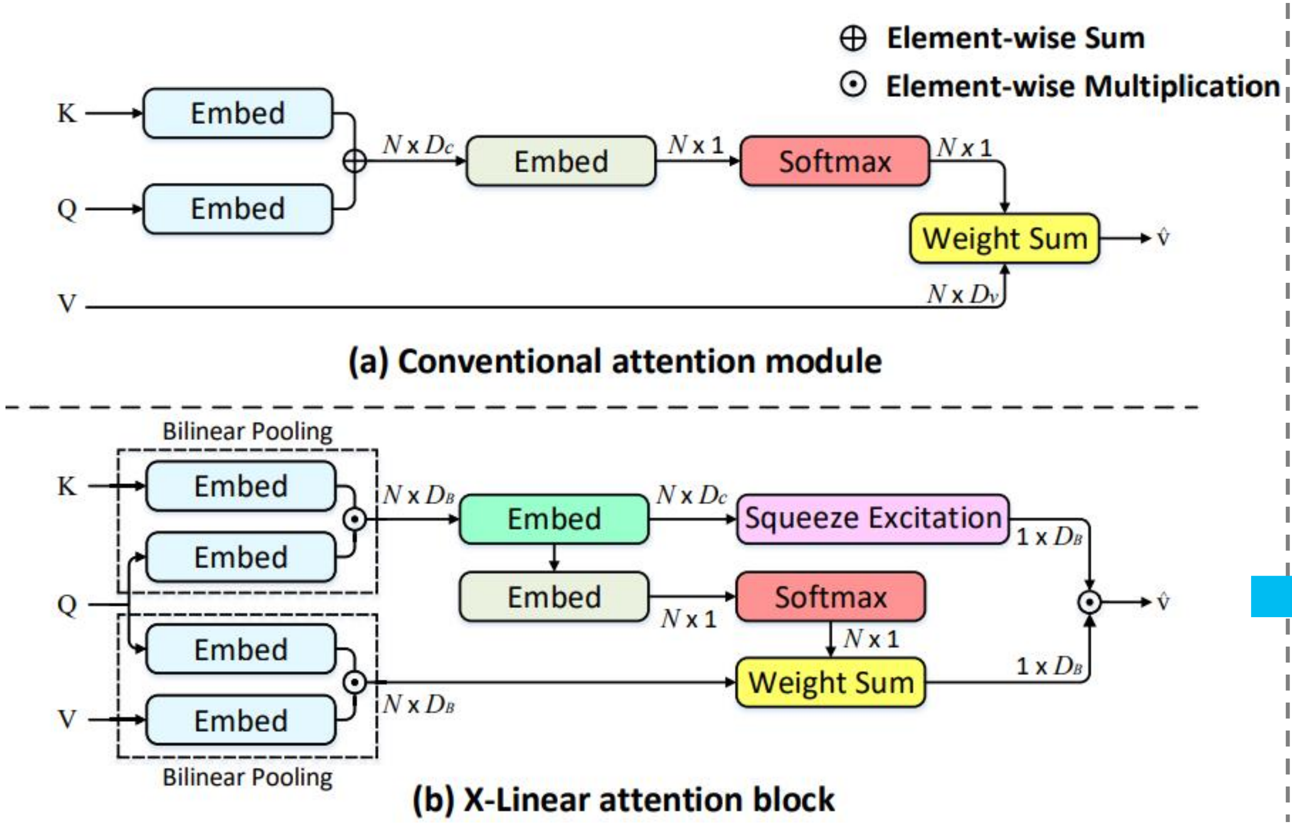
X = attention on attention
[Huang, ICCV19]



Kelvin Xu, et al. “Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.” In ICML, 2015.
Anderson Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." In CVPR, 2018.
Piyush Sharma, et al. “Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning.” In ACL, 2018.
Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. “Attention on Attention for Image Captioning.” In ICCV, 2019.

X-Linear Attention Block

- Motivation
  - Conventional attention: linear fusion of query and key -> 1st order interaction
  - X-Linear attention: bilinear pooling over query and key -> 2nd order interaction



Given query $\mathbf{Q} \in \mathbb{R}^{D_q}$ and a set of keys/values $\mathbf{K} = \{\mathbf{k}_i\}_{i=1}^N$ $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^N$

Bilinear query-key representation between query and each key:

$$\mathbf{B}_i^k = \sigma\left(\mathbf{W}_k \mathbf{k}_i\right) \odot \sigma\left(\mathbf{W}_q^k \mathbf{Q}\right)$$

Based on $\{\mathbf{B}_i^k\}_{i=1}^N$, we measure two kinds of bilinear attention distributions:

Spatial bilinear attention weights:

$$\mathbf{B}'^k_i = \sigma\left(\mathbf{W}_B^k \mathbf{B}_i^k\right), b_i^s = \mathbf{W}_b \mathbf{B}'^k_i, \boldsymbol{\beta}^s = softmax\left(\mathbf{b}^s\right)$$

Channel-wise bilinear attention weights:

$$\bar{\mathbf{B}} = \frac{1}{N}\sum_{i=1}^N \mathbf{B}'^k_i \ , \mathbf{b}^c = \mathbf{W}_e \bar{\mathbf{B}}, \boldsymbol{\beta}^c = sigmoid\left(\mathbf{b}^c\right)$$

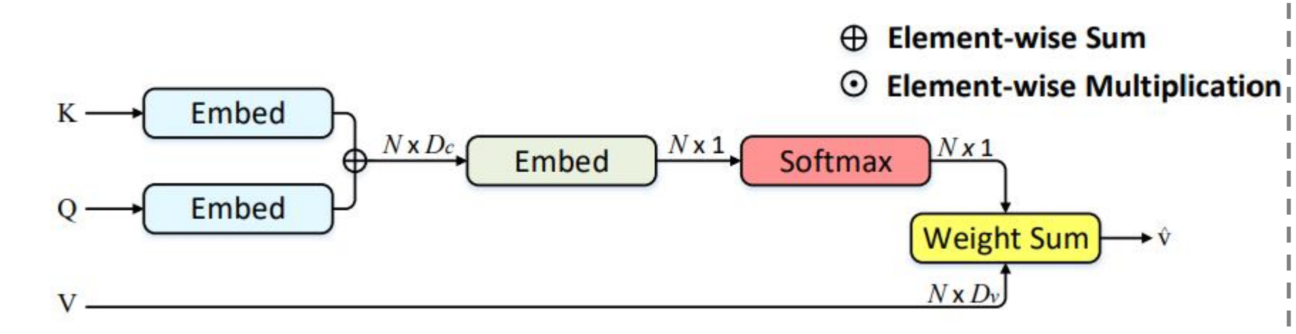
The final output attended value feature in X-Linear attention block:

$$\hat{\mathbf{v}} = F_{X-Linear}\left(\mathbf{K}, \mathbf{V}, \mathbf{Q}\right) = \boldsymbol{\beta}^c \odot \sum_{i=1}^N \boldsymbol{\beta}_i^s \mathbf{B}_i^v,$$
$$\mathbf{B}_i^v = \sigma\left(\mathbf{W}_v \mathbf{v}_i\right) \odot \sigma\left(\mathbf{W}_q^v \mathbf{Q}\right),$$

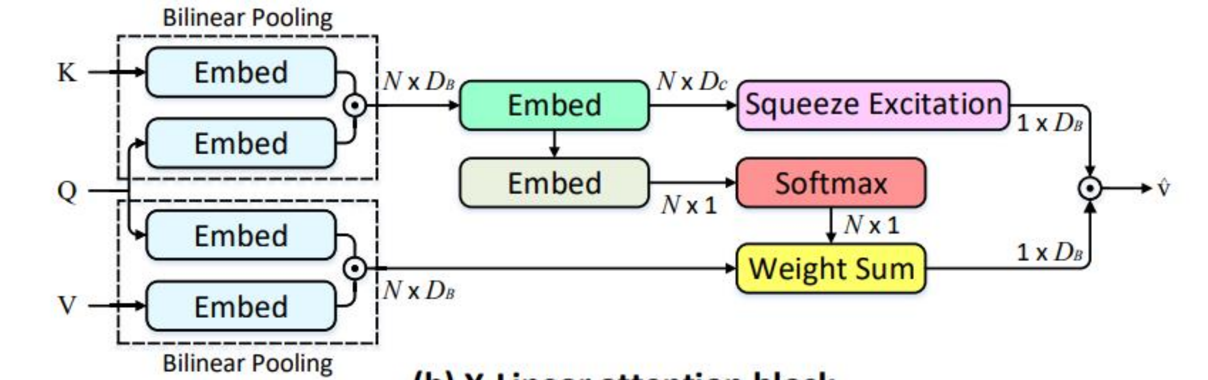
X-Linear Attention Block (+Exponential Linear Unit)

- Motivation
  - Conventional attention: linear fusion of query and key -> 1st order interaction
  - X-Linear attention: bilinear pooling over query and key -> 2nd order interaction **-> infinity order interaction**

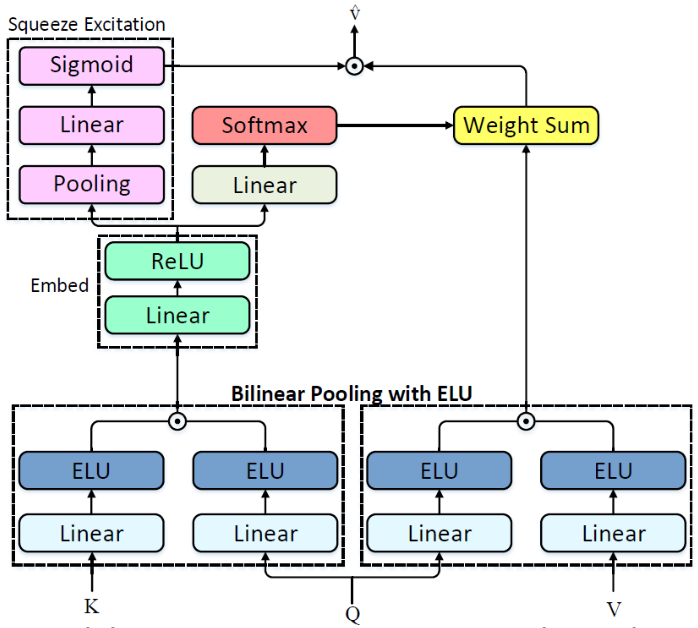
$$
\begin{aligned}
&\exp(\mathrm{W}_X \mathrm{X}) \odot \exp(\mathrm{W}_Y \mathrm{Y}) \\
&= [\exp(\mathrm{W}_X^1 \mathrm{X}) \odot \exp(\mathrm{W}_Y^1 \mathrm{Y}), \ldots, \exp(\mathrm{W}_X^D \mathrm{X}) \odot \exp(\mathrm{W}_Y^D \mathrm{Y})] \\
&= [\exp(\mathrm{W}_X^1 \mathrm{X} + \mathrm{W}_Y^1 \mathrm{Y}), \ldots, \exp(\mathrm{W}_X^D \mathrm{X} + \mathrm{W}_Y^D \mathrm{Y})] \\
&= [\sum_{p=0}^{\infty} \gamma_p^1 (\mathrm{W}_X^1 \mathrm{X} + \mathrm{W}_Y^1 \mathrm{Y})^p, \ldots, \sum_{p=0}^{\infty} \gamma_p^D (\mathrm{W}_X^D \mathrm{X} + \mathrm{W}_Y^D \mathrm{Y})^p],
\end{aligned}
$$



(a) Conventional attention module



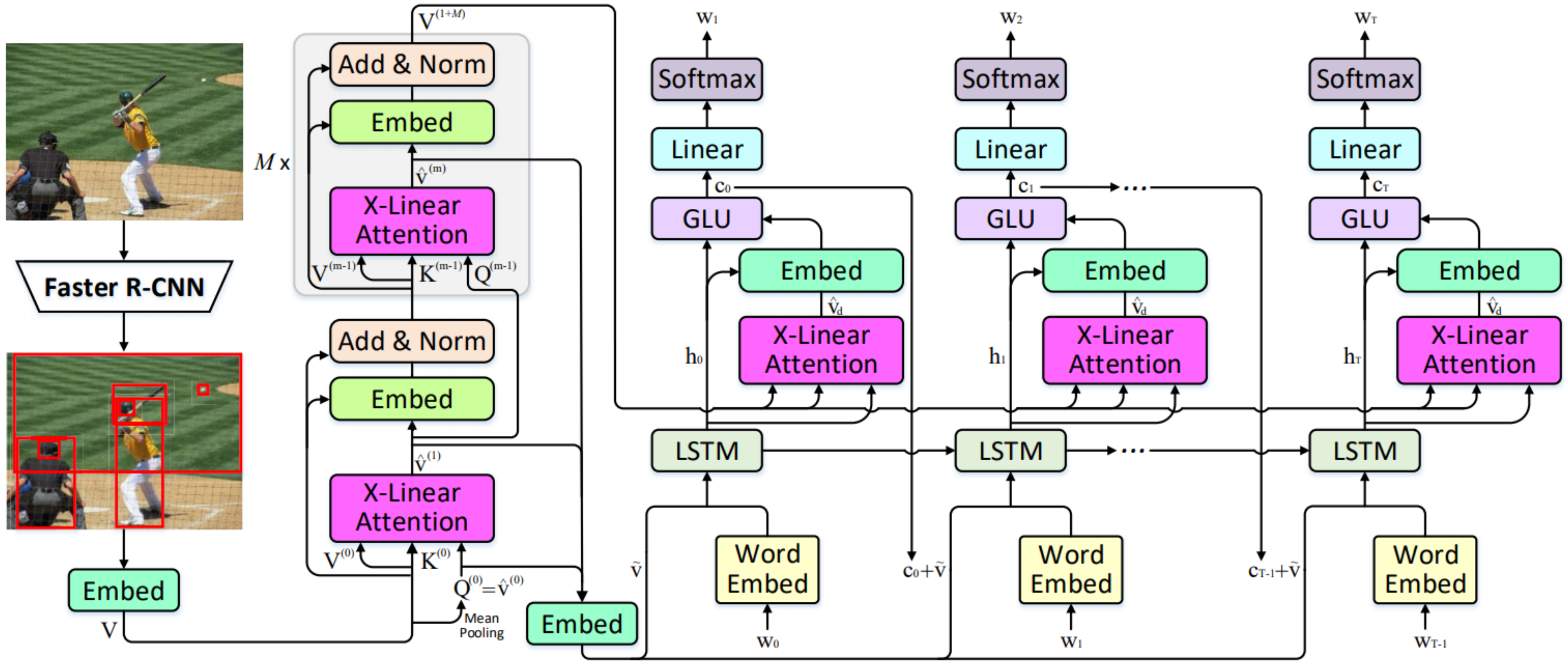
(b) X-Linear attention block



(c) X-Linear attention block (+ELU)

Jonathan T Barron. Continuously differentiable exponential linear units. arXiv preprint arXiv:1704.07483, 2017.

Image Captioning with X-Linear Attention Networks



- X-Linear attention in encoder: encode the region-level features with high order intra-modal interaction
- X-Linear attention in decoder: perform multi-modal reasoning depending on high order inter-modal interaction

Experiments on COCO Karpathy test split

Table 1. Performance comparisons on COCO Karpathy test split, where B@N, M, R, C and S are short for BLEU@N, METEOR, ROUGE-L, CIDEr and SPICE scores. All values are reported as percentage (%). Σ indicates model ensemble/fusion.

	Cross-Entropy Loss								CIDEr Score Optimization							
	B@1	B@2	B@3	B@4	M	R	C	S	B@1	B@2	B@3	B@4	M	R	C	S
LSTM [33]	-	-	-	29.6	25.2	52.6	94.0	-	-	-	-	31.9	25.5	54.3	106.3	-
SCST [28]	-	-	-	30.0	25.9	53.4	99.4	-	-	-	-	34.2	26.7	55.7	114.0	-
LSTM-A [40]	75.4	-	-	35.2	26.9	55.8	108.8	20.0	78.6	-	-	35.5	27.3	56.8	118.3	20.8
RFNet [13]	76.4	60.4	46.6	35.8	27.4	56.5	112.5	20.5	79.1	63.1	48.4	36.5	27.7	57.3	121.9	21.2
Up-Down [2]	77.2	-	-	36.2	27.0	56.4	113.5	20.3	79.8	-	-	36.3	27.7	56.9	120.1	21.4
GCN-LSTM [38]	77.3	-	-	36.8	27.9	57.0	116.3	20.9	80.5	-	-	38.2	28.5	58.3	127.6	22.0
LBPF [26]	77.8	-	-	37.4	28.1	57.5	116.4	21.2	80.5	-	-	38.3	28.5	58.4	127.6	22.0
SGAE [36]	77.6	-	-	36.9	27.7	57.2	116.7	20.9	80.8	-	-	38.4	28.4	58.6	127.8	22.1
AoANet [12]	77.4	-	-	37.2	28.4	57.5	119.8	21.3	80.2	-	-	38.9	29.2	58.8	129.8	22.4
X-LAN	**78.0**	**62.3**	**48.9**	**38.2**	**28.8**	**58.0**	**122.0**	**21.9**	80.8	65.6	51.4	39.5	**29.5**	**59.2**	132.0	**23.4**
Transformer [29]	76.1	59.9	45.2	34.0	27.6	56.2	113.3	21.0	80.2	64.8	50.5	38.6	28.8	58.5	128.3	22.6
X-Transformer	77.3	61.5	47.8	37.0	28.7	57.5	120.0	21.8	**80.9**	**65.8**	**51.5**	**39.7**	**29.5**	59.1	**132.8**	**23.4**
	Ensemble/Fusion															
SCST [28]$^{\Sigma}$	-	-	-	32.8	26.7	55.1	106.5	-	-	-	-	35.4	27.1	56.6	117.5	-
RFNet [13]$^{\Sigma}$	77.4	61.6	47.9	37.0	27.9	57.3	116.3	20.8	80.4	64.7	50.0	37.9	28.3	58.3	125.7	21.7
GCN-LSTM [38]$^{\Sigma}$	77.4	-	-	37.1	28.1	57.2	117.1	21.1	80.9	-	-	38.3	28.6	58.5	128.7	22.1
SGAE [36]$^{\Sigma}$	-	-	-	-	-	-	-	-	81.0	-	-	39.0	28.4	58.9	129.1	22.2
HIP [39]$^{\Sigma}$	-	-	-	38.0	28.6	57.8	120.3	21.4	-	-	-	39.1	28.9	59.2	130.6	22.3
AoANet [12]$^{\Sigma}$	78.7	-	-	38.1	28.5	58.2	122.7	21.7	81.6	-	-	40.2	29.3	59.4	132.0	22.8
X-LAN$^{\Sigma}$	**78.8**	**63.4**	**49.9**	**39.1**	**29.1**	**58.5**	**124.5**	**22.2**	81.6	66.6	52.3	40.3	29.8	59.6	133.7	23.6
X-Transformer$^{\Sigma}$	77.8	62.1	48.6	37.7	29.0	58.0	122.1	21.9	**81.7**	**66.8**	**52.6**	**40.7**	**29.9**	**59.7**	**135.3**	**23.8**

- X-Transformer: Replace the attention module in Transformer with our X-Linear attention block
- Model ensemble: Fuse four models with different initialized parameters

Evaluations on COCO test server and ablation study

Model	Group	B@4		METEOR		ROUGE-L		CIDEr-D	
		c5	c40	c5	c40	c5	c40	c5	c40
X-LAN	**Pan, et al., CVPR'20**	**40.3**	**72.4**	**29.6**	**39.2**	**59.5**	**75.0**	**131.1**	**133.5**
AoANet	Huang, et al., ICCV'19	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
HIP	Yao, et al., ICCV'19	39.3	71.0	28.8	38.1	59.0	74.1	127.9	130.2
GCN-LSTM	Yao, et al., ECCV'18	38.7	69.7	28.5	37.6	58.5	73.4	125.3	126.5
RFNet	Jiang, et al., ECCV'18	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
Up-Down	Anderson, et al., CVPR'18	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
LSTM-A	Yao, et al., , ICCV'17	35.6	65.2	27	35.4	56.4	70.5	116	118
Watson Multimodal	Rennie, et al., CVPR'17	34.4	63.6	26.8	35.3	55.9	70.4	112.3	114.6
G-RMI	Liu, et al., ICCV'17	33.1	62.4	25.5	33.9	55.1	69.4	104.2	107.1
MetaMind/VT_GT	Lu, et al., CVPR'17	33.6	63.7	26.4	35.9	55	70.5	104.2	105.9
DLTC@MSR	Gan, et al., CVPR'17	33.1	63.1	25.7	34.8	54.3	69.6	100.3	101.3
reviewnet	Yang, et al., NIPS'16	31.3	59.7	25.6	34.7	53.3	68.6	96.5	96.9

Image Encoder	Sentence Decoder	B@1	B@2	B@3	B@4	M	R	C	S
Faster R-CNN	LSTM + Conventional attention	76.4	60.3	46.7	36.1	27.9	56.7	114.1	20.9
Faster R-CNN	LSTM + X-Linear attention	76.9	60.9	47.3	36.6	28.2	57.0	117.0	21.2
Faster R-CNN + 1×X-Linear attention	LSTM + X-Linear attention	77.3	61.5	47.9	37.1	28.5	57.3	118.2	21.6
Faster R-CNN + 2×X-Linear attention	LSTM + X-Linear attention	77.5	61.9	48.4	37.7	28.6	57.7	119.4	21.6
Faster R-CNN + 3×X-Linear attention	LSTM + X-Linear attention	77.7	62.2	48.6	37.8	28.6	57.7	120.0	21.6
Faster R-CNN + 4×X-Linear attention	LSTM + X-Linear attention	77.8	62.3	48.7	37.8	28.6	57.8	120.4	21.6
Faster R-CNN + 4×X-Linear attention (+ELU)	LSTM + X-Linear attention (+ELU)	**78.0**	**62.3**	**48.9**	**38.2**	**28.8**	**58.0**	**122.0**	**21.9**

- X-Linear attention block in sentence decoder enhances the capacity of multi-modal reasoning
- Stacking more X-Linear attention blocks in image encoder can lead to performance improvements
- A larger performance gain is attained when upgrading X-Linear attention block with ELU

New video-language pre-training dataset: **Auto-captions on GIF**

http://www.auto-video-captions.top/2020/

- A large-scale video-language pre-training dataset
- 163,183 GIF videos and 164,378 sentences
- Automatically harvested and filtered from a hundred million web pages
- Offer a fertile ground for designing vision-language pre-training techniques

INTRODUCTION

The goal of this challenge is to offer a fertile ground for designing vision-language pre-training techniques that facilitate the vision-language downstream tasks (e.g., video captioning this year). Meanwhile, to further motivate and challenge the multimedia community, we provide a large-scale video-language pre-training dataset (namely "Auto-captions on GIF") for contestants to solve such challenging but emerging task.

The contestants are asked to develop video captioning system based on Auto-captions on GIF dataset (as pre-training data) and the public MSR-VTT benchmark (as training data for downstream task). For the evaluation purpose, a contesting system is asked to produce at least one sentence for each test video. The accuracy will be evaluated against human pre-generated sentence(s).

This monkey on the back of horse

Disney made the best cake of all time using projection

The dry driver returns to his car and presents his mate with kebab

Tiny squid flopping around on the rocky bottom of fish tank