The Summary of ”X-Linear Attention Networks for Image Captioning” [1] Paper

Furkan Gül

I. Motivation

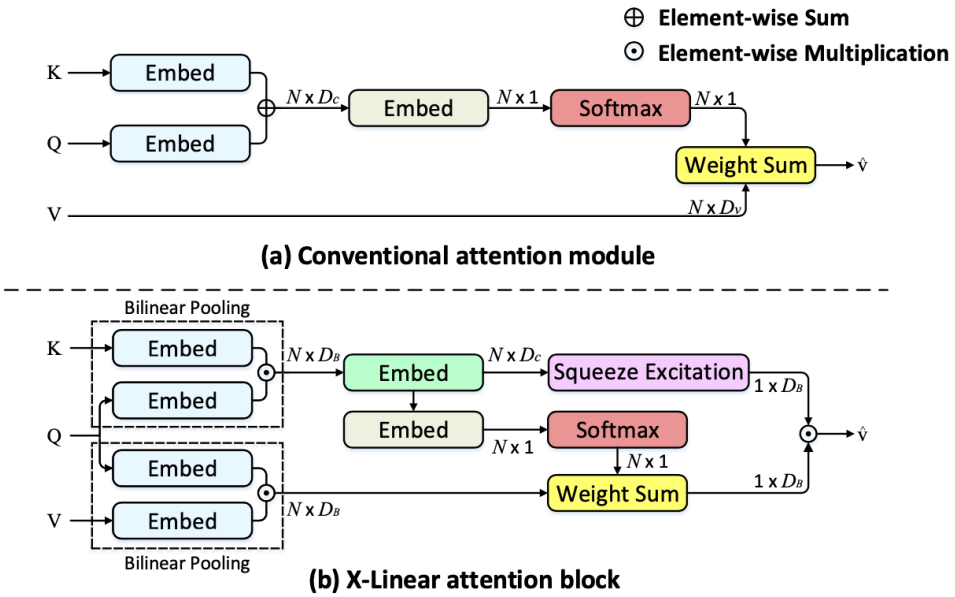

Fig. 1. Comparison between conventional attention mechanism and proposed X-Linear attention block for image captioning [1].

In Figure 1(a), the most conventional attention mechanism can be seen. It measures attention weights by linearly combining the query (state of decoder in the image captioning model) and key (image features outputted by encoder in the image captioning model). Then, the value (another image features outputted by encoder in the image captioning model) is implemented to the attention to calculate a weighted sum. The authors of the paper think that the conventional one inherently models just the 1st order feature relations, so it is not so beneficial in terms of exploiting all feature interactions. As, it is shown in Figure 1(b), a different attention mechanism, which is called X-Linear attention block, is proposed to capture higher-order relations. By stacking X-Linear attention blocks together in the image captioning model, X-Linear Attention Network (X-LAN) will be formed to exploit higher-order (2nd order feature interactions) intra-model in encoder and inter-model interactions in decoder.

II. Approach: X-Linear Attention Networks (X-LAN)

A. X-Linear Attention Block with an Extension

Several X-Linear attention blocks are stacked together to extract more higher (beyond infinity) order feature interactions. However, this implementation requires both high memory and high computational cost. To overcome this issue, a simpler and more effective algorithm is proposed to apply X-Linear attention block to represent infinity order interactions. This method additionally consists of encoding of query ($\boldsymbol{Q}$), each key ($\boldsymbol{k_i}$), and each value ($\boldsymbol{v_i}$) with Exponential Linear Unit (ELU) [2]. This methodology can be observed schematically in Figure 2.

B. X-LAN for Captioning

The proposed modeling framework together with X-Linear attention network (X-LAN) explained in the previous section for image captioning can be seen in Figure 3. To extract image regions, Faster R-CNN [3] is applied. Then, a stack of X-Linear attention blocks is added into the image encoder part to encode the region-level features with the higher-order intra-modal interaction. This will return a set of improved image-level & region-level features. In addition to the improved visual features, to extract multi-modal reasoning from the sentence decoder, X-Linear attention block is implemented there too. This boosted sentence generation process enables the improved representation of high order inter-modal interactions between image content and natural sentence.

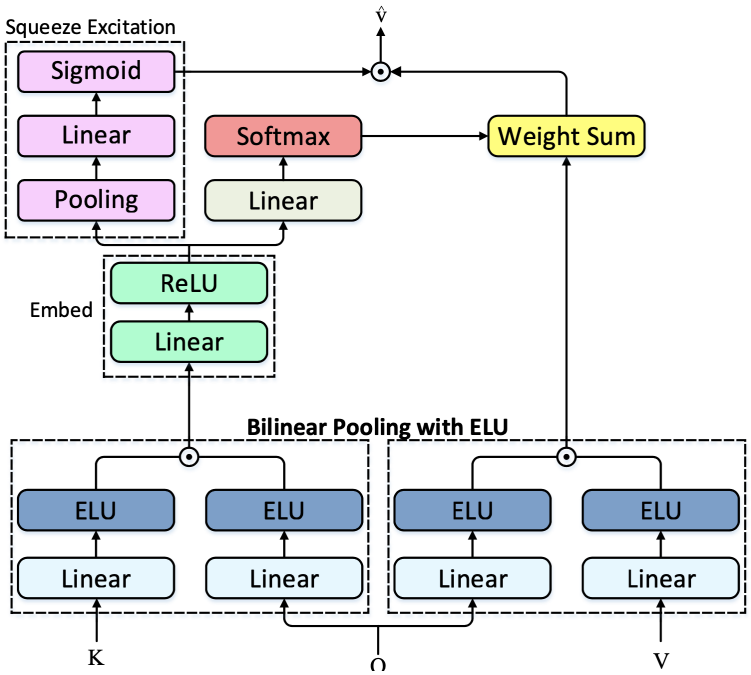

Fig. 2. Framework of X-Linear attention block with ELU to exploit infinity order feature interactions [1].

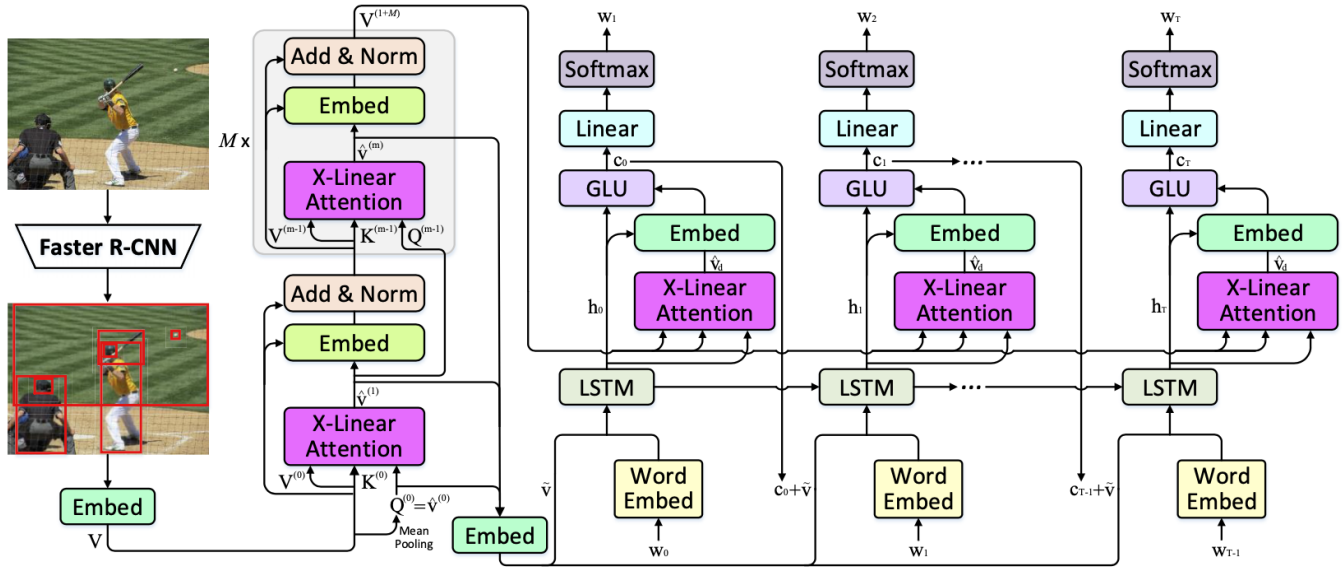

Fig. 3. Image captioning framework with X-Linear Attention Networks (X-LAN) [1].

III. Dataset & Results

All experiments in the paper are carried out on the most popular image captioning benchmark dataset, MSCOCO [4]. MSCOCO dataset which has more than 120000 images is used in their experiments to evaluate the suggested captioning model. There are at least 5 human-annotated sentence descriptions for every image in this dataset. They use Karparthy split [5] which is extensively preferred for reporting results in previous works. For both the validation and test phases with MSCOCO dataset, 5000 images are used. They convert all caption sentences to lower case, tokenize on white space, and discard words that have less than six occurrences. Thus, the vocabulary size for their experiments is fixed and 9488. For the evaluation of the caption model, following metrics are used: CIDEr [6], METEOR [7], SPICE [8], ROUGE [9], and BLEU [10].

A pre-trained Faster-RCNN [3] on ImageNet [11] and Visual Genome [12] datasets is applied to extract a 2048-dimensional bottom-up feature vectors [13] of input images. The original dimension of the feature vectors is 2048. But, they are linearly projected to a new space with the dimension of 1024 in the paper. One hot vector is used to represent each word. In X-Linear attention block with ELU, the dimensions of the bilinear query-key and the transformed bilinear feature are equal to 1024 and

	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]$^{\sum}$	-	-	-	32.8	26.7	55.1	106.5	-	-	-	-	35.4	27.1	56.6	117.5	-
RFNet [13]$^{\sum}$	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]$^{\sum}$	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]$^{\sum}$	-	-	-	-	-	-	-	-	81.0	-	-	39.0	28.4	58.9	129.1	22.2
HIP [39]$^{\sum}$	-	-	-	38.0	28.6	57.8	120.3	21.4	-	-	-	39.1	28.9	59.2	130.6	22.3
AoANet [12]$^{\sum}$	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$^{\sum}$	**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$^{\sum}$	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**

Fig. 4. Image captioning performance comparisons on the MSCOCO Karpathy test split. All values are reported as percentage (%). $\sum$ shows ensembled models.

512 respectively. The LTSM decoder's hidden size is 2024.

X-LAN is trained with a mini-batch size of 40, ADAM [14] optimizer, the beam size of 3. For the cross-entropy loss optimization, the learning rate schedule strategy in [15] is followed. They set the maximum number of iterations to 70 epochs due to the slow convergence rate issue in bilinear pooling. For the optimization of the CIDEr-D score, SCST is applied at the fixed learning rate of 5e-6, and the maximum number of iterations set to 35 epochs.

Model	B@1		B@2		B@3		B@4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
LSTM-A (ResNet-152) [40]	78.7	93.7	62.7	86.7	47.6	76.5	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
Up-Down (ResNet-101) [2]	80.2	95.2	64.1	88.8	49.1	79.4	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet (ResNet+DenseNet+Inception) [13]	80.4	95.0	64.9	89.3	50.1	80.1	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
SGAE (ResNet-101) [36]	81.0	95.3	65.6	89.5	50.7	80.4	38.5	69.7	28.2	37.2	58.6	73.6	123.8	126.5
GCN-LSTM (ResNet-101) [38]	80.8	95.2	65.5	89.3	50.8	80.3	38.7	69.7	28.5	[illegible]	[illegible]	73.4	125.3	126.5
AoANet (ResNet-101) [12]	81.0	95.0	65.8	89.6	51.4	81.3	39.4	71.2	29.1	38.5	58.9	74.5	126.9	129.6
HIP (SENet-154) [39]	81.6	**95.9**	66.2	90.4	51.5	81.6	39.3	71.0	28.8	38.1	59.0	74.1	127.9	130.2
X-LAN (ResNet-101)	81.1	95.3	66.0	89.8	51.5	81.5	39.5	71.4	29.4	38.9	59.2	74.7	128.0	130.3
X-LAN (SENet-154)	81.4	95.7	66.5	**90.5**	52.0	82.4	40.0	**72.4**	**29.7**	**39.3**	**59.5**	**75.2**	130.2	132.8
X-Transformer (ResNet-101)	81.3	95.4	66.3	90.0	51.9	81.7	39.9	71.8	29.5	39.0	59.3	74.9	129.3	131.4
X-Transformer (SENet-154)	**81.9**	95.7	**66.9**	**90.5**	**52.4**	**82.5**	**40.3**	**72.4**	29.6	39.2	**59.5**	75.0	**131.1**	**133.5**

Fig. 5. Leaderboard of several models on the online MSCOCO test server.

In Figure 4, they reports their results together with SCST [16], Up-Down [13], RFNet [17], GCN-LSTM [18], SGAE [19], and AoANet [20] on the Karpathy test split. As it can be seen in Figure 4, X-LAN model outperforms all SOTA models and becomes the new SOTA single-model for Karpathy split in terms of just cross-entropy loss related evaluation metrics.

In Figure 5, they report the performance of their ensembled models trained on the official MSCOCO evaluation server. At the time of submission, around the middle of 2020, they report that they take the lead in the leader board by outperforming all other submissions on the test server in terms of almost all metrics. However, when the leaderboard for MSCOCO captioning is examined detailly, there is no sign of proposed result.

REFERENCES

[1] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10968–10977, 2020.

[2] J. Barron. Continuously differentiable exponential linear units. *ArXiv*, abs/1704.07483, 2017.

[3] Shaoqing Ren, Kaiming He, Ross B. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015.

[4] Tsung-Yi Lin, M. Maire, Serge J. Belongie, J. Hays, P. Perona, D. Ramanan, Piotr Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014.

[5] A. Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:664–676, 2017.

[6] Ramakrishna Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.

[7] Michael J. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. In *WMT@ACL*, 2014.

[8] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.

[9] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*, 2004.

[10] Kishore Papineni, S. Roukos, T. Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[11] Jia Deng, W. Dong, R. Socher, L. Li, K. Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR 2009*, 2009.

[12] R. Krishna, Yuke Zhu, O. Groth, J. Johnson, Kenji Hata, J. Kravitz, Stephanie Chen, Yannis Kalantidis, L. Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016.

[13] Peter Anderson, X. He, C. Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[16] Steven J. Rennie, E. Marcheret, Youssef Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1195, 2017.

[17] W. Jiang, Lin Ma, Yu-Gang Jiang, W. Liu, and T. Zhang. Recurrent fusion network for image captioning. *ArXiv*, abs/1807.09986, 2018.

[18] Ting Yao, Yingwei Pan, Yehao Li, and T. Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

[19] X. Yang, Kaihua Tang, Hanwang Zhang, and J. Cai. Auto-encoding scene graphs for image captioning. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10677–10686, 2019.

[20] Lun Huang, Wenmin Wang, J. Chen, and Xiao-Yong Wei. Attention on attention for image captioning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4633–4642, 2019.