X-Linear Attention Networks for Image Captioning

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

JD AI Research, Beijing, China

{panyw.ustc, tingyao.ustc, yehaoli.sysu}@gmail.com, tmei@jd.com

Abstract

Recent progress on fine-grained visual recognition and visual question answering has featured Bilinear Pooling, which effectively models the 2nd order interactions across multi-modal inputs. Nevertheless, there has not been evidence in support of building such interactions concurrently with attention mechanism for image captioning. In this paper, we introduce a unified attention block — X-Linear attention block, that fully employs bilinear pooling to selectively capitalize on visual information or perform multi-modal reasoning. Technically, X-Linear attention block simultaneously exploits both the spatial and channel-wise bilinear attention distributions to capture the 2nd order interactions between the input single-modal or multi-modal features. Higher and even infinity order feature interactions are readily modeled through stacking multiple X-Linear attention blocks and equipping the block with Exponential Linear Unit (ELU) in a parameter-free fashion, respectively. Furthermore, we present X-Linear Attention Networks (dubbed as X-LAN) that novelly integrates X-Linear attention block(s) into image encoder and sentence decoder of image captioning model to leverage higher order intra- and inter-modal interactions. The experiments on COCO benchmark demonstrate that our X-LAN obtains to-date the best published CIDEr performance of 132.0% on COCO Karpathy test split. When further endowing Transformer with X-Linear attention blocks, CIDEr is boosted up to 132.8%. Source code is available at https://github. com/Panda-Peter/image-captioning.

1. Introduction

Image captioning is the task of automatically producing a natural-language sentence to describe the visual content of an image. The essential practice of neural captioning models follows encoder-decoder paradigm [24, 33], which is derived from neural machine translation [30]. In between, Convolutional Neural Network (CNN) is utilized to encode an input image and Recurrent Neural Network (RNN) is adopted as sentence decoder to generate the output sentence, one word at each time step. Despite involving two
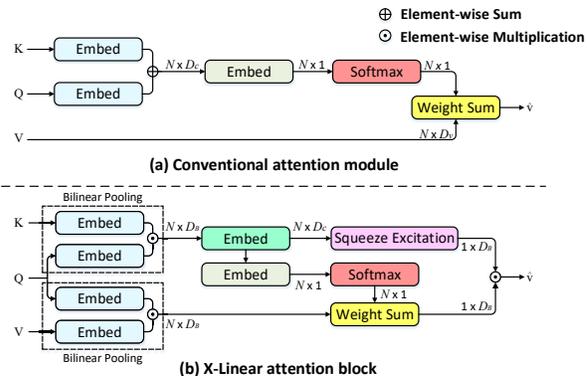

Figure 1. Comparison between conventional attention mechanism and our X-Linear attention block for image captioning. (a) Conventional attention mechanism linearly fuses query (Q) and key (K) via element-wise sum to compute spatial attention weight for each value (V), which characterizes the 1st order interaction between query and key. (b) X-Linear attention block fully capitalizes on bilinear pooling to capture the 2nd order feature interaction in between, and measures both spatial and channel-wise attention distributions. The two attention weights are adopted to accumulate the enhanced values of bilinear pooling on query and value.

different major modalities (visual content and textual sentence) in image captioning, such paradigm of approaches seldom explores the multi-modal interactions particularly at the early stage. In other words, vision and language are treated independently. That prompts the recent state-of-the-art methods [2, 35] to adopt visual attention mechanisms which trigger the interaction between visual content and natural sentence. Concretely, these visual attention mechanisms boost performance by learning to identify selective spatial regions in an image conditioning on current hidden state of language decoder, and in turn accumulating encoded region features with attention weights to guide decoding process. Figure 1(a) illustrates the most conventional attention measure which estimates attention weights via linearly fusing the given query (hidden state of sentence decoder) and key (encoded image features) from different modalities. The attention is then applied to the value (encoded image features) to derive a weighted sum. Nevertheless, we argue that the design of conventional attention inherently exploits only the 1st order feature interaction and is still lacking in efficacy. That severely limits the capacity of complex multi-modal reasoning in image captioning.

A natural way to mitigate the problem is to capture higher order interactions. We start our exploration from 2nd order interaction through the unique design of a unified attention block, namely X-Linear attention block, as shown in Figure 1(b). Technically, the outer product of key and query is computed through bilinear pooling to take all pairwise interactions between query and key into account. After bilinear pooling, two embedding layers are exploited to predict attention weights for each spatial region, followed by a softmax layer to normalize the spatial attention vector. In the meantime, the embedded outer product (feature map) is passed through a squeeze-excitation operation. The squeeze operation aggregates the feature map across spatial regions to produce a channel descriptor and the excitation operation performs the self-gating mechanism with a sigmoid on the channel descriptor to obtain the channel-wise attention vector. Finally, the outer product of query and value via bilinear pooling is weighted summated with the spatial attention vector, and we take the channel-wise multiplication of the sum and the channel-wise attention vector as the attended features. As such, our X-Linear attention block builds the 2nd order interactions and infers the joint representations for image features and hidden states. It is also appealing in view that a stack of the blocks is readily grouped to go beyond bilinear models and extract higher order interactions. In the extreme case, our model could create infinity order interactions by stacking numerous X-Linear attention blocks and we implement this via the kernel trick, e.g., Exponential Linear Unit (ELU), in practice.

By integrating X-Linear attention block(s) into image captioning structures, we present a new X-Linear Attention Networks (X-LAN) to leverage high order intra- and inter-modal interactions, respectively, in the encoder and decoder. Specifically, for image encoder, Faster R-CNN is firstly utilized to detect a set of image regions. After that, a stack of X-Linear attention blocks are adopted to encode the region-level features with the higher order intra-modal interaction in between, leading to a set of enhanced region-level and image-level features. Conditioned on the enhanced visual features induced by image encoder, we further employ X-Linear attention block in sentence decoder to perform multi-modal reasoning. This encourages the exploration of high order inter-modal interactions between visual content and natural sentence to boost sentence generation.

The main contribution of this work is the proposal of a unified X-Linear attention block that models the 2nd order interactions with both spatial and channel-wise bilinear attention. This also leads to the elegant view of how the block should be extended for mining higher or even infinity order interactions and how to integrate such block(s) into image captioning structure. Through an extensive set of experiments, we demonstrate that our new X-LAN model achieves new state-of-the-art performances on COCO dataset.

2. Related Work

Image Captioning. Image captioning is an active research area [2, 12, 19, 23, 24, 28, 33, 34, 35, 37, 39, 40, 41]. The early attempts [24, 33] exploit the encoder-decoder paradigm that firstly utilizes CNN to encoder image and then adopts RNN based decoder to generate the output word sequence, leading to promising results for this task. After that, a series of innovations have been proposed to boost image captioning by encouraging more interactions between the two different modalities via attention mechanism [5]. In particular, [35] integrates soft and hard attention mechanism into LSTM based decoder, aiming to select the most relevant image regions for word prediction at each decoding stage. [41] presents semantic attention that learns to selectively focus on the semantic attributes in image for sentence generation. Instead of fully performing visual attention as in [35], [23] proposes an adaptive attention model that dynamically decides whether to attend to image regions at each decoding stage. Furthermore, bottom-up and top-down attention mechanism [2] exploits visual attention at object level via bottom-up mechanism, and all salient image regions are associated with the output words through top-down mechanism for image captioning. [26] presents the look back method to integrate attention weights from previous time step into the measurement of attention at current time step, which suits visual coherence of human. Later on, the most recently proposed attention on attention module [12] enhances visual attention by further measuring the relevance between the attention result and the query.

Much of existing attention mechanisms in image captioning have concentrated on the exploration of only the 1st order feature interaction between image content and sentence, reflecting limited capacity of multi-modal reasoning. In contrast, we design a novel X-Linear attention block to capture higher and even infinity order interactions, which facilitate both single-modal feature enhancement and multi-modal reasoning for image captioning.

Bilinear Pooling. Bilinear pooling is an operation to calculate outer product between two feature vectors. Such technique can enable the 2nd order interaction across all elements in feature vectors and thus provide more discriminative representations than linear pooling. An early pioneering work [22] demonstrates the advantage of bilinear pooling for fine-grained visual recognition task. Local pairwise feature interactions are thus modeled by leveraging bilinear pooling over the outputs of two CNNs. Later on, [9] proposes compact bilinear pooling that efficiently compresses the high-dimensional bilinear pooling feature into compact one with a few thousand dimensions, but retains the same discriminative power in the meantime. [8] further extends compact bilinear pooling into multi-modal scenario where visual and textual representations are combined for visual question answering task. Instead of compact bilinear pool-

ing that needs complex computations, [16] proposes a flexible low-rank bilinear pooling structure with linear mapping and Hadamard product. Recently, [42] presents a hierarchical bilinear pooling model to aggregate multiple cross-layer bilinear pooling features for fine-grained visual recognition. [15] exploits low-rank bilinear pooling to construct bilinear attention network, aiming to learn bilinear attention distributions for visual question answering.

The aforementioned bilinear pooling techniques are mainly designed for fine-grained visual recognition or visual question answering. Instead, our X-Linear attention block is applicable to image encoder and sentence decoder to exploit higher order intra and inter-modal interactions for image captioning task.

3. X-linear Attention Networks (X-LAN)

In this section, we introduce a novel unified formulation of attention module, named X-Linear attention block, that fully capitalizes on bilinear pooling to capture the 2nd order feature interactions with both spatial and channel-wise bilinear attention. Moreover, we show a specific integration of X-Linear attention block into image encoder and sentence decoder to capture higher order intra- and inter-modal interactions, aiming to enhance visual information and perform complex multi-modal reasoning for image captioning.

3.1. Conventional Attention Module

We first provide a brief review of the most conventional attention module [35] applied in image captioning, which learns to selectively attend to salient image regions for sentence generation. Formally, at decoding time step t, conditioned on the query $\mathbf{Q}$ (current hidden state of sentence decoder $\mathbf{h}_t$), we can obtain the attention distribution $\boldsymbol{\alpha}^t$ over a set of keys $\mathbf{K} = \{\mathbf{k}_i\}_{i=1}^N$ (N local image features):

$$a_i^t = \mathbf{W}_a \left[\tanh\left(\mathbf{W}_k \mathbf{k}_i + \mathbf{W}_q \mathbf{Q}\right)\right], \boldsymbol{\alpha}^t = softmax\left(\mathbf{a}^t\right), \quad (1)$$

where $\mathbf{W}_a$, $\mathbf{W}_k$, and $\mathbf{W}_q$ are embedding matrices, and a_i^t denotes the i-th element in $\mathbf{a}^t$. In this sense, the normalized attention weight α_i^t for each local image feature (i-th element in $\boldsymbol{\alpha}^t$) is derived from the linear fusion of the given query and key via element-wise sum. Such way inherently exploits only the 1st order feature interaction between natural sentence and visual content for attention measurement. Next, attention module produces the attended image feature $\hat{\mathbf{v}}^t$ by accumulating all values $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^N$ (N local image features) with spatial attention weights: $\hat{\mathbf{v}}^t = \sum_{i=1}^N \alpha_i^t \mathbf{v}_i$.

3.2. X-Linear Attention Block

Though conventional attention module nicely triggers the interaction between different modalities, only the 1st order feature interaction is exploited, which reflects limited

capacity of complex multi-modal reasoning in image captioning. Inspired by the recent successes of bilinear pooling applied in fine-grained visual recognition [9, 42] or visual question answering [8, 15], we fully capitalize on bilinear pooling techniques to construct a unified attention module (X-Linear attention block) for image captioning, as depicted in Figure 1(b). Such design of X-Linear attention block strengthens the representative capacity of the output attended feature by exploiting higher order interactions between the input single-modal or multi-modal features.

In particular, suppose we have the query $\mathbf{Q} \in \mathbb{R}^{D_q}$, a set of keys $\mathbf{K} = \{\mathbf{k}_i\}_{i=1}^N$, and a set of values $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^N$, where $\mathbf{k}_i \in \mathbb{R}^{D_k}$ and $\mathbf{v}_i \in \mathbb{R}^{D_v}$ denote the i-th key/value pair. X-Linear attention block firstly performs low-rank bilinear pooling to achieve a joint bilinear query-key representation $\mathbf{B}_i^k \in \mathbb{R}^{D_B}$ between query $\mathbf{Q}$ and each key $\mathbf{k}_i$:

$$\mathbf{B}_i^k = \sigma\left(\mathbf{W}_k \mathbf{k}_i\right) \odot \sigma\left(\mathbf{W}_q^k \mathbf{Q}\right), \quad (2)$$

where $\mathbf{W}_k \in \mathbb{R}^{D_B \times D_k}$, and $\mathbf{W}_q^k \in \mathbb{R}^{D_B \times D_q}$ are embedding matrices, σ denotes ReLU unit, and $\odot$ represents element-wise multiplication. As such, the learnt bilinear query-key representation $\mathbf{B}_i^k$ conveys the 2nd order feature interactions between query and key.

Next, depending on all bilinear query-key representations $\{\mathbf{B}_i^k\}_{i=1}^N$, two kinds of bilinear attention distributions are obtained to aggregate both spatial and channel-wise information within all values. Most specifically, the spatial bilinear attention distribution is introduced by projecting each bilinear query-key representation into the corresponding attention weight via two embedding layers, followed with a softmax layer for normalization:

$$\mathbf{B}_i'^k = \sigma\left(\mathbf{W}_B^k \mathbf{B}_i^k\right), b_i^s = \mathbf{W}_b \mathbf{B}_i'^k, \boldsymbol{\beta}^s = softmax\left(\mathbf{b}^s\right), \quad (3)$$

where $\mathbf{W}_B^k \in \mathbb{R}^{D_c \times D_B}$ and $\mathbf{W}_b \in \mathbb{R}^{1 \times D_c}$ are embedding matrices, $\mathbf{B}_i'^k$ is the transformed bilinear query-key representation, and b_i^s is the i-th element in $\mathbf{b}^s$. Here each element β_i^s in $\boldsymbol{\beta}^s$ denotes the normalized spatial attention weight for each key/value pair. Meanwhile, we perform a squeeze-excitation operation [11] over all transformed bilinear query-key representations $\{\mathbf{B}_i'^k\}_{i=1}^N$ for channel-wise attention measurement. Concretely, the operation of squeeze aggregates all transformed bilinear query-key representations via average pooling, leading to a global channel descriptor $\bar{\mathbf{B}}$:

$$\bar{\mathbf{B}} = \frac{1}{N} \sum_{i=1}^N \mathbf{B}_i'^k. \quad (4)$$

After that, the followed excitation operation produces channel-wise attention distribution $\boldsymbol{\beta}^c$ by leveraging the self-gating mechanism with a sigmoid over the global channel descriptor $\bar{\mathbf{B}}$:

$$\mathbf{b}^c = \mathbf{W}_e \bar{\mathbf{B}}, \boldsymbol{\beta}^c = sigmoid\left(\mathbf{b}^c\right), \quad (5)$$
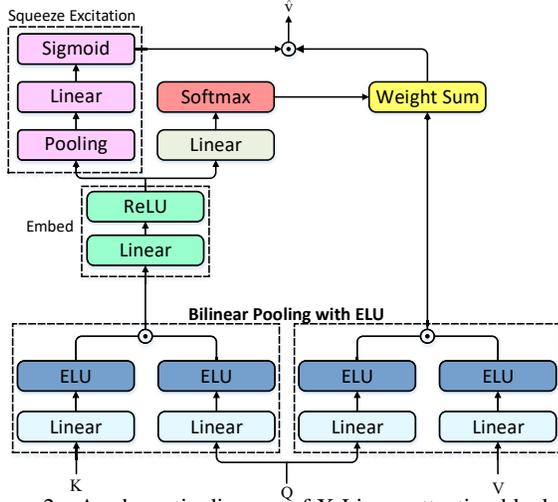
Figure 2. A schematic diagram of X-Linear attention block plus ELU to capture infinity order feature interactions.

where $\mathbf{W}_e \in \mathbb{R}^{D_B \times D_c}$ is embedding matrix.

Finally, our X-Linear attention block generates the attended value feature $\hat{\mathbf{v}}$ by accumulating the enhanced bilinear values with spatial and channel-wise bilinear attention:

$$\hat{\mathbf{v}} = F_{X-Linear}(\mathbf{K}, \mathbf{V}, \mathbf{Q}) = \boldsymbol{\beta}^c \odot \sum_{i=1}^{N} \boldsymbol{\beta}_i^s \mathbf{B}_i^v,$$
$$\mathbf{B}_i^v = \sigma(\mathbf{W}_v \mathbf{v}_i) \odot \sigma(\mathbf{W}_q^v \mathbf{Q}), \quad (6)$$

where $\mathbf{B}_i^v$ denotes the enhanced value of bilinear pooling on query $\mathbf{Q}$ and each value $\mathbf{v}_i$, $\mathbf{W}_v \in \mathbb{R}^{D_B \times D_v}$, and $\mathbf{W}_q^v \in \mathbb{R}^{D_B \times D_q}$ are embedding matrices. Accordingly, compared to conventional attention modules that simply explore 1st order interaction between query and key, X-Linear attention block produces the more representative attended feature since higher order feature interactions are exploited via bilinear pooling.

Extension with higher order interactions. In order to exploit higher order feature interactions, we further iterate the above process of bilinear attention measurement and feature aggregation using a stack of our X-Linear attention blocks. Formally, for the m-th X-Linear attention block, we firstly take the pervious output attended feature $\hat{\mathbf{v}}^{(m-1)}$ as input query, coupled with current input keys $\mathbf{K}^{(m-1)} = \{\mathbf{k}_i^{(m-1)}\}_{i=1}^N$, and values $\mathbf{V}^{(m-1)} = \{\mathbf{v}_i^{(m-1)}\}_{i=1}^N$:

$$\hat{\mathbf{v}}^{(m)} = F_{X-Linear}\left(\mathbf{K}^{(m-1)}, \mathbf{V}^{(m-1)}, \hat{\mathbf{v}}^{(m-1)}\right), \quad (7)$$

where $\hat{\mathbf{v}}^{(m)}$ is the output new attended feature. $\hat{\mathbf{v}}^{(0)}$, $\mathbf{K}^{(0)}$, and $\mathbf{V}^{(0)}$ denotes $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$, respectively. Then, all keys/values are further updated conditioned on the output new attended feature $\hat{\mathbf{v}}^{(m)}$:

$$\mathbf{k}_i^{(m)} = LayerNorm(\sigma(\mathbf{W}_m^k[\hat{\mathbf{v}}^{(m)}, \mathbf{k}_i^{(m-1)}]) + \mathbf{k}_i^{(m-1)}),$$
$$\mathbf{v}_i^{(m)} = LayerNorm(\sigma(\mathbf{W}_m^v[\hat{\mathbf{v}}^{(m)}, \mathbf{v}_i^{(m-1)}]) + \mathbf{v}_i^{(m-1)}),$$
$$(8)$$

where $\mathbf{W}_m^k$ and $\mathbf{W}_m^v$ are embedding matrices. Note that here each key/value is concatenated with the new attended

feature, followed with a residual connection and layer normalization as in [31]. We repeat the process (Eq.(7) and Eq.(8)) M times via stacking M X-Linear attention blocks, which captures higher ($2M^{th}$) order feature interactions.

Extension with infinity order interactions. One natural way to exploit more higher (even infinity) order feature interactions is to stack plenty of X-Linear attention blocks. Nevertheless, such way inevitably leads to a huge rise in memory demand and computational cost, not to mention the extreme case of stacking infinity blocks. Instead, we adopt a simple but effective method to enable our X-Linear attention block to model infinity order interactions by additionally encoding query $\mathbf{Q}$, each key $\mathbf{k}_i$, and each value $\mathbf{v}_i$ with Exponential Linear Unit (ELU) [4], as shown in Figure 2. That is, the infinity order feature interactions can be approximately modeled via performing bilinear pooling on two exponentially transformed features. Here we demonstrate that such approximation can be proved via Taylor expansion of each element in bilinear vector after exponential transformation. Specifically, given two feature vectors X and Y, the Taylor expansion of bilinear pooling over the exponentially transformed features can be expressed as:

$$\exp(W_X X) \odot \exp(W_Y Y)$$
$$= [\exp(W_X^1 X) \odot \exp(W_Y^1 Y), ..., \exp(W_X^D X) \odot \exp(W_Y^D Y)]$$
$$= [\exp(W_X^1 X + W_Y^1 Y), ..., \exp(W_X^D X + W_Y^D Y)] \quad (9)$$
$$= [\sum_{p=0}^{\infty} \gamma_p^1 (W_X^1 X + W_Y^1 Y)^p, ..., \sum_{p=0}^{\infty} \gamma_p^D (W_X^D X + W_Y^D Y)^p],$$

where W_X and W_Y are embedding matrices, D denotes the dimension of bilinear vector, W_X^i/W_Y^i is the i-th row in W_X/W_Y. Therefore, this expansion clearly shows that each element in bilinear vector after exponential transformation reflects infinity order interactions.

3.3. X-LAN for Image Captioning

Recall that our X-Linear attention is a unified attention block, it is feasible to plug X-Linear attention block(s) into image encoder and sentence decoder to capture higher order intra- and inter-modal interactions for image captioning. We next present how to integrate such block(s) into the encoder-decoder structure via our devised X-Linear Attention Networks (X-LAN), as illustrated in Figure 3.

3.3.1 Notation and Training Strategy

In the standard task of image captioning, we are given an image I to be described with a natural-language sentence $Y_{1:T}$. The sentence $Y_{1:T} = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_T\}$ is a sequence of T words, where $\mathbf{w}_t$ is the textual feature of the t-th word. The image I is represented as a set of spatial image region features $\mathbf{V} = \{\mathbf{v}_i\}_{i=1}^N$ by utilizing Faster R-CNN [27]. During training, given the ground-truth sentence $Y_{1:T}^\star$ for image I, we first train our X-LAN by minimizing the cross entropy loss $L_{CE}(\theta) = -\sum_{t=1}^{T} log(p_\theta(\mathbf{w}_t^\star | Y_{1:t-1}^\star))$,
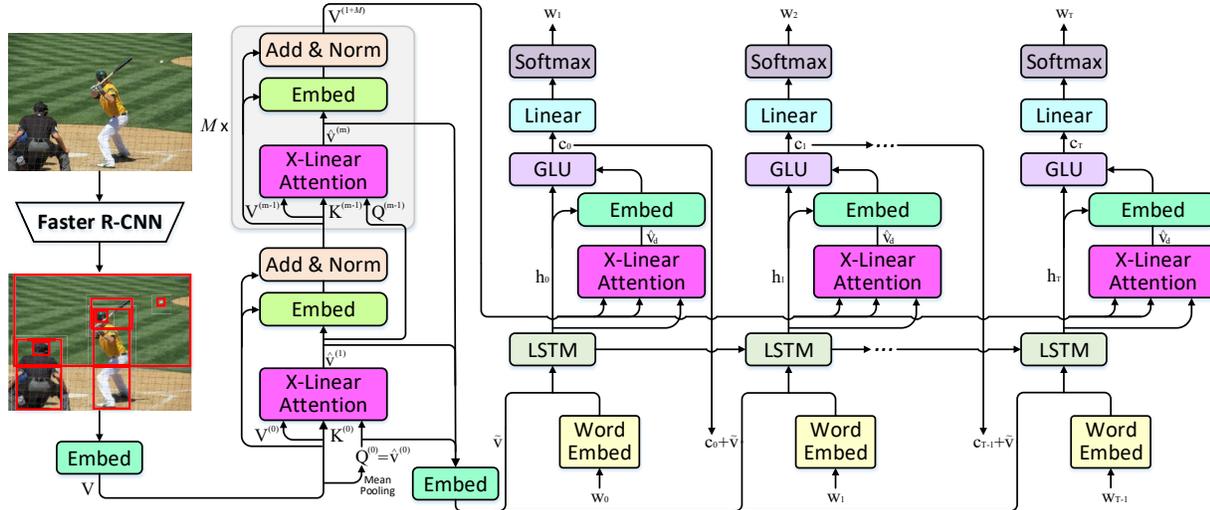
Figure 3. Overview of our X-Linear Attention Networks (X-LAN) for image captioning. Faster R-CNN is firstly utilized to detect a set of image regions. Next, a stack of X-Linear attention blocks are leveraged in image encoder to encode the region-level features with the higher order intra-modal interaction in between, leading to a set of enhanced region-level and image-level features. Depending on the enhanced visual features, X-Linear attention block is further adopted in sentence decoder to perform multi-modal reasoning. This encourages the exploration of high order inter-modal interactions between visual content and natural sentence to boost sentence generation.

where θ denotes the parameters of X-LAN. Next, our X-LAN can be further optimized with sentence-level reward via Self-Critical Sequence Training [28].

3.3.2 Encoder with X-Linear Attention

The image encoder is a module that transforms the input set of spatial image region features $\mathbf{V}$ into a series of intermediate states, which are enhanced with the contextual information in between. Here we fully employ our X-Linear attention block(s) to construct the image encoder. As such, the representative capacity of encoded image-level or region-level features are strengthened via capturing higher order intra-modal feature interactions.

Formally, the image encoder in X-LAN is composed of a stack of $(1 + M)$ identical layers ($M = 3$). Each layer includes two components: X-Linear attention block as in Eq.(7) and keys/values updating module as in Eq.(8). Specifically, for the first X-Linear attention block, we take the mean-pooled region feature $\hat{\mathbf{v}}^{(0)} = \bar{\mathbf{v}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{v}_i$ as the initial input query, coupled with the initial keys/values (*i.e.*, all region features $\mathbf{K}^{(0)} = \mathbf{V}^{(0)} = \mathbf{V}$). The output is thus the attended image-level feature $\hat{\mathbf{v}}^{(1)}$, which will be further fed into the next X-Linear attention block as input query. Meanwhile, the keys/values are updated conditioned on the attended image-level feature $\hat{\mathbf{v}}^{(1)}$. After that, we repeat the updating process of query and keys/values in M times via the subsequence M stacked layers. Accordingly, by performing feature enhancement via the image encoder with $(1 + M)$ X-Linear attention blocks, we can obtain $(1 + M)$ output attended image-level features $\{\hat{\mathbf{v}}^{(m)}\}_{m=1}^{1+M}$. Moreover, we treat the updated values $\mathbf{V}^{(1+M)}$ after the final X-Linear attention block as the enhanced region-level

features, which are endowed with the higher order intra-modal feature interactions in between.

3.3.3 Decoder with X-Linear Attention

The sentence decoder aims to generate the output sentence conditioned on the enhanced image-level and region-level visual features induced by the image encoder. To further encourage high order inter-modal interactions between visual content and natural sentence, we integrate our X-Linear attention block into attention-based LSTM decoder to perform multi-modal reasoning. In particular, at each decoding time step t, we firstly concatenate the mean-pooled region feature $\hat{\mathbf{v}}^{(0)}$ and all attended image-level features $\{\hat{\mathbf{v}}^{(m)}\}_{m=1}^{1+M}$, which is further transformed into the global image-level feature $\tilde{\mathbf{v}}$ through an embedding layer:

$$\tilde{\mathbf{v}} = \mathbf{W}_G[\hat{\mathbf{v}}^{(0)}, \hat{\mathbf{v}}^{(1)}, ..., \hat{\mathbf{v}}^{(1+M)}], \quad (10)$$

where $\mathbf{W}_G$ is embedding matrix. The input of LSTM is thus set as the concatenation of current input word $\mathbf{w}_t$, the global image-level feature $\tilde{\mathbf{v}}$, the previous LSTM hidden state $\mathbf{h}_{t-1}$, and the pervious context vector $\mathbf{c}_{t-1}$. After that, we take the output of LSTM $\mathbf{h}_t$ as input query of X-Linear attention block, whose keys/values are set as the enhanced region-level features $\mathbf{V}^{(1+M)}$ from image encoder. In this way, the output attended feature $\hat{\mathbf{v}}_d$ of X-Linear attention block is more representative by capturing the 2nd order interactions between image features and hidden state. Next, we measure current context vector $\mathbf{c}_t$ by concatenating the attended feature $\hat{\mathbf{v}}_d$ with current LSTM hidden state $\mathbf{h}_t$, followed with an embedding layer and a Gated Linear Unit (GLU) [6]. Such context vector $\mathbf{c}_t$ is finally leveraged for the prediction of next word $\mathbf{w}_{t+1}$ via a softmax layer.

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

4. Experiments

4.1. Dataset and Implementation Details

All the experiments are conducted on the most popular image captioning benchmark COCO [21]. The whole COCO dataset contains 123,287 images, which includes 82,783 training images, 40,504 validation images, and 40,775 testing images. Each image is equipped with five human-annotated sentences. Note that the annotations for official testing set are not provided and the evaluation over that testing set can only be conducted through online testing server. In addition, we adopt the widely adopted Karpathy split [14] for offline evaluation. There are 113,287 training images, 5,000 validation images, and 5,000 testing images in the Karpathy split. We pre-process all training sentences by converting them into lower case and dropping the words that occur less than 6 times, leading to the final vocabulary with 9,488 unique words.

We leverage the off-the-shelf Faster-RCNN pre-trained on ImageNet [7] and Visual Genome [18] to extract image region features [2]. Each original region feature is a 2,048-dimensional vector, which is transformed as the input region feature with the dimension $D_v = 1,024$. Each word is represented as "one-hot" vector. The dimensions of the bilinear query-key representation and the transformed bilinear feature (D_B and D_c) in X-Linear attention block is set as 1,024 and 512, respectively. We stack four X-Linear attention blocks (plus ELU) in the image encoder and the sentence decoder is equipped with one X-Linear attention block (plus ELU). The hidden layer size in LSTM decoder is set as 1,024. The whole image captioning architecture are mainly implemented with PyTorch, optimized with Adam [17]. For the training stage, we follow the training schedule in [31] to optimize the whole architecture with cross-entropy loss. The warmup steps are set as 10,000 and

the mini-batch size is 40. Since low-rank bilinear pooling may lead to slow convergence rate as indicated in [16], we set the maximum iteration as 70 epoches. For the training with self-critical training strategy, as in [28], we first select the initialization model which is trained with cross-entropy loss and achieves best CIDEr score on validation set. After that, the whole architecture is further optimized with CIDEr reward, when the learning rate is set as 0.00001 and the maximum iteration is 35 epoches. At the inference stage, we adopt the beam search strategy and set the beam size as 3. Five evaluation metrics, BLEU@N [25], METEOR [3], ROUGE-L [20], CIDEr [32], and SPICE [1], are simultaneously utilized to evaluate our model.

4.2. Performance Comparison

Offline Evaluation. Table 1 summaries the performance comparisons between the state-of-the-art models and our proposed X-LAN on the offline COCO Karpathy test split. Note that for fair comparison, we report the results for each run optimized with both cross entropy loss and CIDEr Score. Meanwhile, we separately show the performances for single models and ensemble/fused models. In general, our X-LAN consistently exhibits better performances than other single models, which include the non-attention baselines (LSTM, LSTM-A) and attention-based methods (SCST, RFNet, and others). The CIDEr score of our X-LAN can achieve 132.0% with CIDEr score optimization, which is to-date the best performance without any model ensemble and makes the absolute improvement over the best competitor AoANet by 2.2%. The performance improvements generally demonstrate the key advantage of exploiting higher and even infinity order interactions via our X-Linear attention block, that facilitate both single-modal feature enhancement and multi-modal reasoning for image captioning. In particular, LSTM-A improves LSTM by em-

Table 2. Leaderboard of the published state-of-the-art image captioning models on the COCO online testing server, where B@N, M, R, and C are short for BLEU@N, METEOR, ROUGE-L, and CIDEr scores. All values are reported as percentage (%).

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	37.6	58.5	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**

phasising semantic attributes at decoding stage. RFNet and Up-Down further boost the performances by involving attention mechanism that learns to identify selective spatial regions for sentence generation. Moreover, by exploiting rich semantic information in images (e.g., visual relations between objects or scene graph) for sentence generation, GCN-LSTM and SGAE exhibit better performance than Up-Down. Nevertheless, the performances of SGAE are lower than AoANet that enhances conventional visual attention by further measuring the relevance between the attention result and the query. This confirms that improving attention measurement is an effective way to enhance the interaction between visual content and natural sentence and thus boost image captioning. In addition, by integrating X-Linear attention block(s) into encoder and decoder, our X-LAN outperforms AoANet, which demonstrates the merit of mining higher and even infinity intra- and inter-modal interactions. Similar to the observations over single models, an ensemble version of our X-LAN by fusing four models with different initialized parameters obtains better performances than other ensemble models.

To fully verify the generalizability of our X-Linear attention block for image captioning, we include a variant of our X-LAN (named **X-Transformer**) by plugging X-Linear attention blocks into Transformer based encoder-decoder structure. Table 1 also shows the performance comparison between Transformer and our X-Transformer. Note that here Transformer denotes our implementation of Transformer-based encoder-decoder structure as in [29]. Similar to the observations in LSTM-based encoder-decoder structure, X-Transformer boosts up the performances by integrating X-Linear attention blocks into the Transformer-based encoder and decoder. The performance improvements again demonstrate the advantage of exploiting higher order interactions via our X-Linear attention block for image captioning.

Online Evaluation. In addition, we evaluate our X-LAN and X-Transformer on the official testing set by submitting the ensemble versions to online testing server. Table 2 details the performances over official testing images with 5 reference captions (c5) and 40 reference captions (c40). Note that here we adopt two common backbones (ResNet-

Figure 4. Examples of image captioning results by Up-Down and X-LAN, coupled with the corresponding ground truth sentences.

101 [10] and SENet-154 [11]) for online evaluation. The results clearly show that compared to all the other published state-of-the-art systems, our X-LAN and X-Transformer exhibit better performances across most metrics.

Qualitative Analysis. Figure 4 showcases several image captioning results of Up-Down and our X-LAN, coupled with human-annotated ground truth sentences (GT). Generally, compared with the captions of Up-Down which are somewhat relevant to image content and logically correct, our X-LAN produces more accurate and descriptive sentences by exploiting higher order intra- and inter-modal interactions. For example, Up-Down generates the phrase of "a truck is driving" that is inconsistent with the visual content for the last image, while "a group of cars stopped" in our X-LAN depicts the visual content more precise. This again confirms the advantage of capturing the high order interactions among image regions and meanwhile triggering high order interactions between different modalities for multi-modal reasoning via our X-Linear attention block.

Table 3. Ablation study on the use of X-Linear attention block(s) in image encoder and sentence decoder (optimized with cross-entropy loss), where B@N, M, R, and C are short for BLEU@N, METEOR, ROUGE-L, and CIDEr. All values are reported as percentage (%).

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**

(a) Up-Down: A bowl of apples and apples on a table
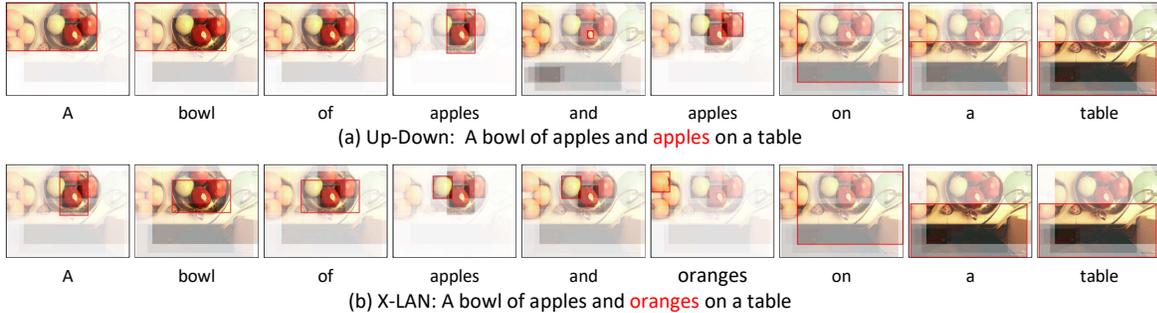

(b) X-LAN: A bowl of apples and oranges on a table

Figure 5. The visualization of attended image regions along the caption generation processes for (a) Up-Down and (b) X-LAN. At the decoding step for each word, we outline the image region with the maximum attention weight in red.

4.3. Experimental Analysis

Ablation Study. To fully examine the impact of X-Linear attention block(s) in both image encoder and sentence decoder for image captioning, we conduct ablation study by comparing different variants of our X-LAN in Table 3. We start from a base model which is a degraded version of our X-LAN by simply encoding images with Faster R-CNN and further leveraging LSTM with conventional attention module for sentence generation. This ablated base model obtains similar results to Up-Down. Next, we extend the base model by replacing the conventional attention module with our X-Linear attention block in sentence decoder, which obtains better performances. The results indicate that compared to conventional attention module that only exploits 1st order inter-modal interaction, our X-Linear attention block enhances the capacity of multi-modal reasoning via the modeling of higher order interactions. Furthermore, in order to show the relations between performance and the number of stacked X-Linear attention blocks in image encoder, we include a series of variants by integrating one more X-Linear attention blocks into encoder. In general, stacking more X-Linear attention blocks in image encoder can lead to performance improvements. That basically validates the effectiveness of modeling high order intra-modal interactions between image regions in encoder. However, no further significant performance improvement is observed when stacking four blocks. We speculate that the increased parameters by stacking more blocks might result in overfitting, which somewhat hinder the exploitation of more higher order interaction in this way. Recall that instead of stacking blocks to capture more higher order interactions, we present a simple but effective solution to enable even infinity order feature interactions by equip-

ping X-Linear attention block with ELU in a parameter-free fashion. As such, when upgrading our X-Linear attention block with ELU in both encoder and decoder, a larger performance gain is attained.

Attention Visualization. In order to better qualitatively evaluate the generated results with X-Linear attention block, we visualize the evolutions of attended image regions along the caption generation processes for Up-Down and X-LAN in Figure 5. We can observe that Up-Down sometimes focus on the irrelevant image region whose corresponding object should not be generated at that time step. For instance, at the 6th decoding step for Up-Down, the region containing irrelevant object "apples" is attended, which misleads decoder to produce "apples" again. Instead, by exploiting higher order feature interactions for multi-modal reasoning via X-Linear attention block, our X-LAN consistently focuses on the correct regions for image captioning.

5. Conclusions

We present a novel unified X-Linear attention block for image captioning, which models the 2nd order interactions with both spatial and channel-wise bilinear attention. The higher and even infinity order feature interactions can be readily modeled via staking multiple X-Linear attention blocks and equipping the block with Exponential Linear Unit (ELU). To verify our claim, we devise X-Linear Attention Networks (X-LAN) by integrating X-Linear attention block(s) into image encoder and sentence decoder to exploit higher order intra and inter-modal interactions for image captioning. Extensive experiments conducted on COCO dataset demonstrate the efficacy of X-Linear attention block and X-LAN. More remarkably, we obtain new state-of-the-art performances on this captioning dataset with X-LAN.

References

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

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*, 2005.

[4] Jonathan T Barron. Continuously differentiable exponential linear units. *arXiv preprint arXiv:1704.07483*, 2017.

[5] Kyunghyun Cho, Aaron Courville, and Yoshua Bengio. Describing multimedia content using attention-based encoder-decoder networks. *IEEE Trans. on Multimedia*, 2015.

[6] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In *ICML*, 2017.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[8] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016.

[9] Yang Gao, Oscar Beijbom, Ning Zhang, and Trevor Darrell. Compact bilinear pooling. In *CVPR*, 2016.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[11] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[12] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019.

[13] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *ECCV*, 2018.

[14] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[15] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *NIPS*, 2018.

[16] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *ICLR*, 2017.

[17] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[18] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[19] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Pointing novel objects in image captioning. In *CVPR*, 2019.

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

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[22] Tsung-Yu Lin, Aruni RoyChowdhury, and Subhransu Maji. Bilinear cnn models for fine-grained visual recognition. In *ICCV*, 2015.

[23] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017.

[24] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Explain images with multimodal recurrent neural networks. In *NIPS Workshop on Deep Learning*, 2014.

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

[26] Yu Qin, Jiajun Du, Yonghua Zhang, and Hongtao Lu. Look back and predict forward in image captioning. In *CVPR*, 2019.

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[28] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

[29] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, 2018.

[30] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

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

[32] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[34] Jing Wang, Yingwei Pan, Ting Yao, Jinhui Tang, and Tao Mei. Convolutional auto-encoding of sentence topics for image paragraph generation. In *IJCAI*, 2019.

[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[36] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019.

[37] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *CVPR*, 2017.

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

[39] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *ICCV*, 2019.

[40] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *ICCV*, 2017.

[41] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016.

[42] Chaojian Yu, Xinyi Zhao, Qi Zheng, Peng Zhang, and Xinge You. Hierarchical bilinear pooling for fine-grained visual recognition. In *ECCV*, 2018.