

Auto-Parsing Network for Image Captioning and Visual Question Answering

Xu Yang^{1,*} Chongyang Gao^{2,*} Hanwang Zhang³ Jianfei Cai⁴

¹ School of Computer Science and Engineering, Southeast University

² Department of Computer Science, Dartmouth College

³ School of Computer Science and Engineering, Nanyang Technological University

⁴ Department of Data Science and AI, Monash University

xuyangaca@gmail.com chongyang.gao.gr@dartmouth.edu hanwangzhang@ntu.edu.sg
Jianfei.Cai@monash.edu

Abstract

We propose an Auto-Parsing Network (APN) to discover and exploit the input data’s hidden tree structures for improving the effectiveness of the Transformer-based vision-language systems. Specifically, we impose a Probabilistic Graphical Model (PGM) parameterized by the attention operations on each self-attention layer to incorporate sparse assumption. We use this PGM to softly segment an input sequence into a few clusters where each cluster can be treated as the parent of the inside entities. By stacking these PGM constrained self-attention layers, the clusters in a lower layer compose into a new sequence, and the PGM in a higher layer will further segment this sequence. Iteratively, a sparse tree can be implicitly parsed, and this tree’s hierarchical knowledge is incorporated into the transformed embeddings, which can be used for solving the target vision-language tasks. Specifically, we showcase that our APN can strengthen Transformer based networks in two major vision-language tasks: Captioning and Visual Question Answering. Also, a PGM probability-based parsing algorithm is developed by which we can discover what the hidden structure of input is during the inference.

1. Introduction

Nowadays, Transformer [57] based frameworks have been prevalently applied into vision-language tasks and impressive improvements have been observed in image captioning [16, 18, 30, 44], VQA [78], image grounding [38, 75], and visual reasoning [1, 50]. Researchers attribute the progress to the various advantages of Transformer, like the

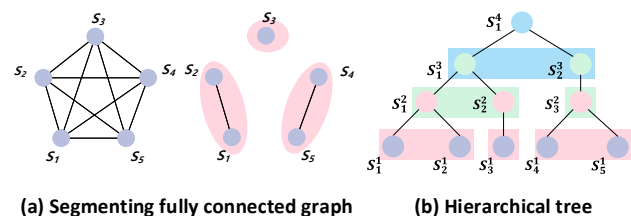


Figure 1. (a) The different graph priors entailed in the classic self-attention and our Probabilistic Graphical Model (PGM) constrained self-attention. Left: The classic self-attention pairs every two nodes in a graph and thus forms a fully connected graph. Right: Constrained by PGM, five nodes are segmented into three clusters. (b) By stacking our PGM constrained self-attention layers, a hierarchical tree can be automatically constructed. Hence we call our network as Auto-Parsing Network (APN).

efficient parallel computing [57], the ability to approximate any sequence-to-sequence function [79], and the exploitation of the fully connected graph prior [9] provided by self-attention as shown in the left part of Figure 1(a). Particularly, by the graph prior, although vision and language data have quite different superficial forms, their structural commonalities can be automatically abstracted, embedded, and transferred to narrow the domain gap.

However, the underlying structures of vision and language data are usually hierarchical and sparse, which are different from fully connected graphs, *e.g.*, a sentence or an image can be parsed as constituent trees of words or objects, respectively [9, 15, 56]. Without sparse and hierarchical constraints, this system may be overwhelmed by the trivial global dependencies and overlook the critical local context [33, 66, 68]. Taking VQA as an example, the question in the last row of Figure 4 asks, “What number is the hour hand on?” for an image containing both hour hand and minute hand. A system with the fully connected graph

*Both authors contributed equally to this research.

prior may make an incorrect prediction by directly exploiting the global dependency between “number” and “hand” due to their high co-occurrence frequency in the training set, and thus neglects the key local context “hour hand”. A similar problem is also observed in image captioning where noisy image scene graphs constructed by trivial dependencies may contribute less to the improvement [40].

To reduce trivial connections of fully connected graphs, researchers usually parse the input into some sparse and hierarchical structures, such as the filtered scene graphs [14, 23, 81] or sparse trees [52] and then exploit them into solving various vision and language tasks, *e.g.*, image captioning [70, 72], VQA [4, 32, 55], grounding [6, 37] and VCR [77]. However, these strategies require a large number of matched graph annotations [27, 39, 67] for training useful parsers [12, 51, 80, 82]; otherwise the domain shift can be induced, invalidating the parsed graphs.

To relieve the burden of incorporating hierarchical and sparse graph priors, inspired by Tree-Transformer [63], we propose a network that can learn to automatically parse inputs into trees during the end-to-end training without any additional graph annotations, hence named as **Auto-Parsing Network (APN)**. Specifically, we constrain the self-attention operation by a Probabilistic Graphical Model (PGM) [10, 29], which is parameterized by differentiable attention operations. As shown in the right part of Figure 1(a), the PGM helps to segment the input sequence into a few clusters. After each segmenting iteration, only the entities in the same cluster can attend to each other, and thus, the local context is embedded. Intuitively, each cluster can be considered as the parent of inside entities and these clusters together compose a new sequence. By stacking constrained self-attention layers, the new sequence at a lower layer will be further segmented by the PGM at a higher layer. For example, as in Figure 1(b), s_1^1 and s_2^1 in the first level are clustered into a new pseudo-parent node s_1^2 in the second level. Then, s_1^2 and s_2^2 are further clustered. Via this iterative way, a tree can be automatically parsed.

By APN, the local and global contexts can be accordingly embedded at lower and higher layers. Once we build an encoder-decoder based on APN, both source and target domains’ hierarchical structures can be automatically parsed, embedded, and transferred. We deploy the proposed APN in two fundamental vision-language tasks: image captioning [60, 65], and visual question answering [5]. Experiment results on both tasks show that our APN obtains consistent improvements compared to Transformer based models. Furthermore, we develop a parsing algorithm, which can generate constituent trees for vision and language inputs based on the calculated PGM probabilities. In this way, when the model infers, the hidden structure for each sample can be revealed.

In summary, we have the following contributions:

- Inspired by Tree-Transformer [63], we propose an Auto-Parsing Network (APN) which can unsupervisedly learn to parse trees for the inputs by imposing PGM probabilities on self-attention layers and exploiting hierarchical constraints into PGM probabilities.
- We design two different APNs for solving Image Captioning and visual Question Answering.
- We show that our APN achieves consistent improvements compared with the classic Transformer on both tasks.

2. Related Work

Structured Learning. Structured learning is a research hotspot [28, 43] since most data have hidden structures and the exploitation of these structures is beneficial to solving downstream tasks [11, 53]. The Probabilistic Graphical Model is one classic method which has been incorporated into deep networks to solve vision [13, 26] and NLP tasks [21, 61]. However, some of such models need to use the forward-backward algorithm [49] or the inside-outside algorithm [7] for inferring the probability during the training, which will unavoidably slow the training.

To accelerate the training, we follow **Tree-Transformer** [63] which applies the attention mechanism to calculate a constraint matrix to achieve soft segmentation. However, their constraint matrix is not a normalized probability and we reformulate their technique into a form of PGM. Based on this PGM probability, we design two different Auto-Parsing Networks to respectively solve image captioning and visual question answering. In this way, we can implicitly parse the hidden trees from the input data and the networks can be trained end-to-end without using the forward-backward or inside-outside algorithms.

Exploiting Graphs in Visual Reasoning. Image Captioning [60, 65] and Visual Question Answering [5] are two fundamental tasks in visual reasoning, that aim to generate a fluent sentence to describe a visual scene and provide answers to questions related to the visual contents, respectively. Thanks to the blooming of deep learning, encoder-decoder structures [70], attention mechanisms [3] including self-attention [20] and many other techniques, the performance of image captioning and VQA has been boosted significantly. Furthermore, considering that graph priors can transfer commonalities and mitigate the gap between visual and language domains, researchers explore how to use graphs [55, 76] properly in both tasks. On one hand, graphs can be transferred into latent variables by GCN [70, 72], which can be directly utilized by models. On the other hand, the input images and questions can be parsed into trees [63]. Because of structural correspondence, deep module networks [19, 78] can feed different types of parts of a parse tree into separate modules. However, the needs for additional data and parsers limit the use of graph priors. To

address this limitation, we parse inputs into trees based on PGMs and self-attention for different tasks automatically, without the need for any extra data and parsers.

3. Multi-Head Attention Revisit

We first revisit the multi-head attention operation, which is the elemental building block of our APN. Given the query, key, and value matrices, $\mathbf{Q} \in \mathbb{R}^{N_Q \times d}$, $\mathbf{K} \in \mathbb{R}^{N_K \times d}$, $\mathbf{V} \in \mathbb{R}^{N_K \times d}$,¹ Multi-Head attention [57] calculates an attended matrix \mathbf{O} as follows:

$$\begin{aligned}
 \text{Input: } & \mathbf{Q}, \mathbf{K}, \mathbf{V} \\
 \text{Att: } & \mathbf{A}_i = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{W}_i^{\mathbf{Q}}(\mathbf{K}\mathbf{W}_i^{\mathbf{K}})^T}{\sqrt{d}}\right) \\
 \text{Head: } & \mathbf{H}_i = \mathbf{A}_i\mathbf{V}\mathbf{W}_i^{\mathbf{V}}, \\
 \text{Multi-Head: } & \mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_8]\mathbf{W}^{\mathbf{H}}, \\
 \text{Output: } & \mathbf{O} = \text{FFN}(\mathbf{H}),
 \end{aligned} \tag{1}$$

where $\mathbf{W}_i^{\mathbf{Q}}, \mathbf{W}_i^{\mathbf{K}}, \mathbf{W}_i^{\mathbf{V}} \in \mathbb{R}^{d \times d_h}$, and $\mathbf{W}_i^{\mathbf{H}} \in \mathbb{R}^{d \times d}$ are all trainable matrices; the number of the heads is set to 8 and $d_h = d/8$; \mathbf{A}_i is the attention matrix for calculating the i -th head \mathbf{H}_i ; $[\cdot]$ is the concatenation operation; and FFN is the Position-wise Feed-Forward Network: FC-RELU-FC.

Specifically, when setting $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ to the same value, such mechanism is called self-attention or non-local convolution [62, 64]. From the Graph Network perspective, it is building a fully-connected graph [9]. For example, given a representation sequence $\mathbf{S} = \{s_1, s_2, \dots, s_T\} \in \mathbb{R}^{d \times T}$ and by setting $\mathbf{Q} = \mathbf{K} = \mathbf{V} = \mathbf{S}$, each two nodes s_i and s_j are connected by an edge weighted by the attention weight $\alpha_{i,j}$. In this way, dense and long-term dependencies between each two entities are embedded, which has been proven to be beneficial in various tasks.

However, since the underlying structures of data are usually hierarchical and sparse, a fully connected graph assumption may lead to trivial attention patterns [59], *e.g.*, only one entity is attended by the other entities, or the critical local context as discussed in Introduction is overlooked. To induce the sparse and hierarchical prior for capturing more meaningful relations, we impose the probabilistic graphical model (PGM) into the self-attention network and stack them to build an Auto-Parsing Network (APN).

4. Probabilistic Graphical Model

Generally, a PGM is defined as the product of the clique potential functions that are all conditioned on an input set [29, 41]:

$$P(\mathbf{x}|\mathbf{S}) = \frac{1}{Z(\mathbf{x}, \mathbf{S})} \prod_C \theta_C(\mathbf{x}_C|\mathbf{S}), \tag{2}$$

¹The number of the elements in \mathbf{K} and \mathbf{V} are the same and we use N_K to denote it.

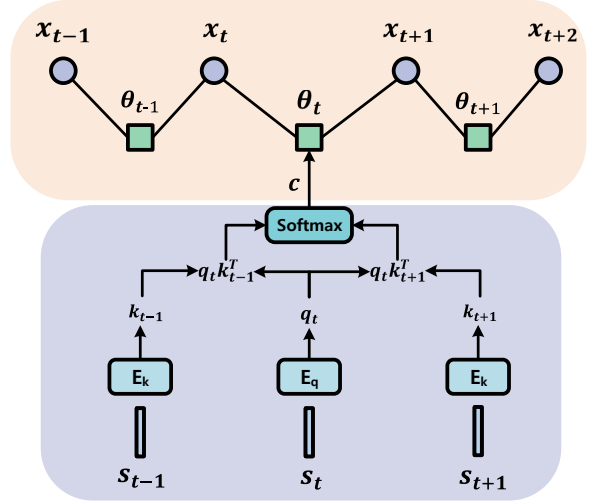


Figure 2. Illustration of our probabilistic graphical model (PGM), the top part is the PGM, and the bottom part is the implementation of the potential function. The E_k and E_q encode s into k and q , respectively.

where \mathbf{S} is the set of a sequential entities, $\theta_C(\cdot)$ is the potential function, and $Z(\cdot)$ is the partition function:

$$Z(\mathbf{x}, \mathbf{S}) = \sum_{\mathbf{x}} \prod_C \theta_C(\mathbf{x}_C|\mathbf{S}) \tag{3}$$

which makes the overall probabilities sum to 1.

In our case, since we focus more on the correlation between neighboring entities (*e.g.*, whether a few neighboring entities should be grouped together), we apply the pairwise PGM where each clique only includes the neighboring two entities and the PGM formula can be simplified as:

$$P(x_1, x_2, \dots, x_{T-1}|\mathbf{S}) = \frac{1}{Z(\mathbf{x}, \mathbf{S})} \prod_{t=1}^{T-1} \theta_t(x_t|\mathbf{S}), \tag{4}$$

where x_t is a binary hidden variable denoting whether s_t and s_{t+1} is connected to each other and $\theta_t(\cdot)$ measures the correlation between s_t and s_{t+1} , as visualized in the top part of Figure 2. For example, a large $\theta_3(x_3 = 1)$ indicates that s_3 and s_4 should be grouped together.

4.1. Network Parameterization of $\theta_t(\cdot)$

Specifically, we set $\theta_t(\cdot)$ as:

$$\begin{aligned}
 \theta_t(x_t = 1|\mathbf{S}) &= \text{Ber}(z_t^{t+1} = 1|c_t^{t+1})\text{Ber}(z_{t+1}^t = 1|c_t^t) \\
 \theta_t(x_t = 0|\mathbf{S}) &= 1 - \theta_t(x_t = 1|\mathbf{S}),
 \end{aligned} \tag{5}$$

where $\theta_t(x_t = 1|\mathbf{S})$ is the product of two Bernoulli distributions and $\theta_t(x_t = 0|\mathbf{S})$ is set to guarantee that the sum is 1. In this way, each $\theta_t(\cdot)$ is normalized and $Z(\mathbf{x}, \mathbf{S})$ in Eq. (4) is always 1. The Bernoulli distribution $\text{Ber}(z|c) = c^z(1-c)^{(1-z)}$ and z_t^{t+1} is a binary variable indicating whether s_t should be connected with s_{t+1} with the corre-

sponding probability c_t^{t+1} . Eq. (5) is designed based on the following intuition: if an entity s_t is close to its right neighbor s_{t+1} (i.e. $z_t^{t+1} = 1$) and that neighbor is also close to this entity (i.e. $z_{t+1}^t = 1$), then they should be tightly connected ($x_t = 1$).

To determine whether an entity s_t is close to its right neighbor s_{t+1} , we can compare their closeness with that between s_t and its left neighbor s_{t-1} , which naturally derives the following attention network parameterization of the probability c :

$$[c_t^{t-1}, c_t^{t+1}] = \text{ATT}(\mathbf{q}_t, [\mathbf{k}_{t-1}, \mathbf{k}_{t+1}]), \quad (6)$$

where $\mathbf{q}_t = s_t \mathbf{W}^q$, $\mathbf{k}_t = s_t \mathbf{W}^k$. Inputting the sequential set \mathcal{S} into this attention network, we can calculate c_t^{t-1} and c_t^{t+1} for each s_t . As in the bottom part of Figure 2, by bringing these probabilities back into Eq. (5) and (4), we can calculate the values of the potential functions and the $P(\mathbf{x}|\mathcal{S})$ for any binary hidden vector \mathbf{x} .

4.2. Segmenting a Sequence

An intuitive strategy of using the PGM with Eq. (4) to segment a sequence is to find the maximum of this probabilistic model:

$$\mathbf{x}^* = \underset{\mathbf{x}}{\text{argmax}}(P(\mathbf{x}|\mathcal{S})). \quad (7)$$

However, solving Eq. (7) over all possible \mathbf{x} using traditional methods like max-sum algorithm [10] is not fast and moreover it needs to be repeated in each forward pass during training, which is unbearable.

Instead, similar to the technique proposed in tree-transformer [63], we approximate this hard segmentation in a soft manner. Specifically, we calculate a $T \times T$ matrix M where if $i < j$:

$$\begin{aligned} M_{i,j} &= P(x_i = 1, x_{i+1} = 1, \dots, x_{j-1} = 1) \\ &= \prod_{t=i}^{j-1} \theta_t(x_t = 1|\mathcal{S}), \end{aligned} \quad (8)$$

if $i > j$, $M_{i,j} = M_{j,i}$, and $M_{i,i} = 1$. In this matrix, each $M_{i,j}$ is the marginal probability of clustering the entities from s_i to s_j , e.g., $M_{2,4}$ measures the probability of clustering the entities s_2, s_3, s_4 , and then if $M_{2,4}$ is large, s_2, s_3, s_4 are softly clustered.

After calculating M , we use it to revise the Head operation in Eq. (1) as follows:

$$\mathbf{H} = (\mathbf{M} \otimes \mathbf{A})\mathbf{V}\mathbf{W}^V, \quad (9)$$

where \otimes stands for the element-wise product. Intuitively, after revising, the original fully-connected graph becomes a sparser graph that contains a few clusters and the entities are only connected with each other in the same cluster. For example, as in the left of Figure 1 (a), s_i and s_j can be freely attended by each other in self-attention where the attended weights are $A_{i,j}$ and $A_{j,i}$, while in the right part, after constraining by M , they can only attend each other if $M_{i,j}$ has a large value, which denotes s_i and s_j are softly

clustered, e.g., s_1 and s_2 belong to the same cluster. In this way, the whole sequence is softly segmented into a few clusters. Note that we use the same M to constrain the self-attention weights of different heads since our PGM aims to segment the input sequence and all the heads should follow the segmented clusters to calculate the outputs.

4.3. Parsing a Tree

In this section, we first introduce how to automatically and implicitly parse a tree during the forward pass. Then we introduce an algorithm for explicitly parsing a tree after the forward pass, which can visualize how the hidden tree structure is incorporated into the embeddings.

Implicitly Parsing during Forward Pass. We have shown that by multiplying the self-attention weights with our PGM probability matrix M , the input sequence can be softly segmented into clusters. The segmented clusters can be considered as the parents of the entities in it, e.g., as shown in Figure 1 (b), $s_1^2 = \{s_1^1, s_2^1\}$ denotes that the parent s_1^2 has two children: s_1^1, s_2^1 .

However, simply stacking our PGM constrained layers cannot parse a tree since the entities in a lower-layer cluster may not still be in a higher-layer cluster, e.g., $M_{i,j}^l$ is large while $M_{i,j}^{l+1}$ is small. To amend this drawback, following Tree-Transformer [63], we modify the potential function (Eq. (5)) at the l -th level as:

$$\tilde{\theta}_t^l = \tilde{\theta}_t^{l-1} + (1 - \tilde{\theta}_t^{l-1})\theta_t^l, \quad (10)$$

where θ_t^l is the abbreviation of $\theta_t^l(x_t = 1|\mathcal{S}^l)$. Since $0 \leq \tilde{\theta}_t^{l-1} \leq 1$, $\tilde{\theta}_t^l$ is the convex combination of 1 and θ_t^l , $\tilde{\theta}_t^l$ will always be larger or equal than θ_t^l . Then we replace θ_t^l with $\tilde{\theta}_t^l$ in Eq. (8) to compute M^l and will have $M_{i,j}^{l+1} \geq M_{i,j}^l$. Then the entities which are clustered at a lower layer will still be clustered at a higher layer. By stacking these constrained layers, during the iterative segmenting, an implicitly tree can be automatically parsed during the forward pass. One toy example about this is shown in Figure 1 (b).

Explicitly Parsing after Forward Pass. With an inputting sequence, APN not only outputs the embeddings that incorporate the hidden hierarchical knowledge through the forward pass, but also calculates a series of potential probabilities (Eq. (10)) in each layer. We follow [63] to provide an algorithm to explicitly parse a sequence into a tree by potential probabilities, which is given in Algorithm 1. This algorithm parses a sequence into a tree from top to bottom recursively. **ParseTree**(l, i, j) denotes the process of using the potential function of the l -th level $\tilde{\theta}^l$ to segment a sequence that begins at s_i and ends at s_j . In line 1, it first determines whether s_i and s_j are neighbors: if they are, it segments s_i and s_j as the left and right leaves, respectively; otherwise, in Line 4, it finds the position p^* that has the minimum potential function value to segment the sequence into two parts: $\{s_{i:p^*}\}$ and $\{s_{p^*+1:j}\}$, as in Line 12 and 13.

Algorithm 1 Parsing Tree by Potential values

Input: $\tilde{\theta}_t^l$ for $t = 1 : T, l = 1 : L$
 $id \leftarrow$ minimum layer id
 $threshold \leftarrow$ Split point's threshold
 $l \leftarrow$ layer index; $i \leftarrow$ left position; $j \leftarrow$ right position
Output: ParseTree
ParseTree(l, i, j) :
 1: **if** $i - j \leq 1$ **then**
 2: **return** (i, j)
 3: **end if**
 4: $p^* \leftarrow \mathbf{argmin}_{p=i, \dots, j-1} (\tilde{\theta}_p^l)$
 5: $next \leftarrow \mathbf{max}(l - 1, id)$
 6: **if** $\tilde{\theta}_{p^*}^l > threshold$ **then**
 7: **if** $l == id$ **then**
 8: **return** (i, j)
 9: **end if**
 10: **return** **ParseTree**($next, i, j$)
 11: **end if**
 12: $LeftTree \leftarrow$ **ParseTree**($next, i, p^*$)
 13: $RightTree \leftarrow$ **ParseTree**($next, p^* + 1, j$)
 14: **return** ($LeftTree, RightTree$)

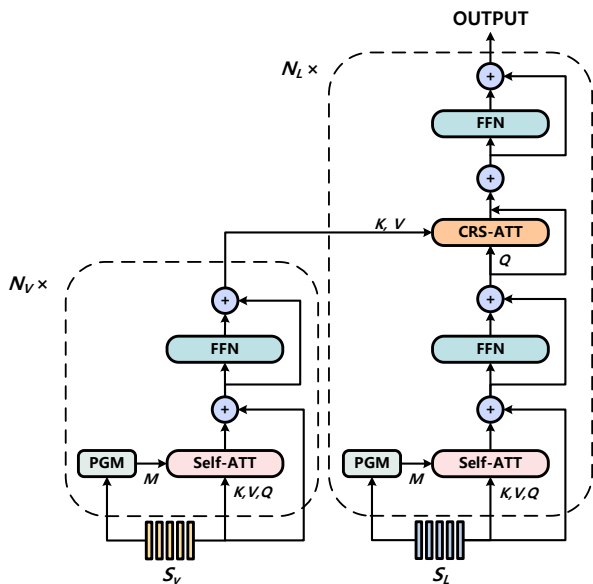


Figure 3. An overview of our APN for image captioning task. The FFN, CRS-ATT and Self-ATT denote the feed-forward network, cross-attention network and self-attention network.

Figure 4 shows the parsed results for captioning and VQA.

5. Image Captioning

5.1. Architecture and Objectives

The APN architecture for image captioning is shown in Figure 3. S_V and S_L are respectively the visual and language representation sequences. Note that the S_V is lin-

earized by sorting the RoIs from top-left to bottom-right. The left and right parts sketch the visual encoder and the language decoder, which are both stacked by 6 blocks, $N_V = N_L = 6$. We first train our APN by minimizing the cross-entropy loss:

$$L_{CE} = -\log P(Y^*), \quad (11)$$

where Y^* is the ground truth caption for the given image. Then, the model is further trained by maximizing a reinforcement learning (RL) based reward [47]:

$$R_{RL} = \mathbb{E}_{Y^s \sim P(Y)} [r(Y^s; Y^*)], \quad (12)$$

where Y^s is the sampled sentence and r is a sentence-level metric, e.g., the CIDEr-D [58] metric, for Y^s and Y^* .

5.2. Dataset, Settings and Metrics

Microsoft COCO Dataset [36] includes 123,287 images and each of them has 5 captions as labels. We conduct experiments on Karpathy split (113,287/5,000/5,000 train/val/test images) for offline testing and Official online test split (82,783/40,504/40,775 train/val/test images).

Settings. We preprocess the captions by the following steps. We change all words to lowercase and delete the words appearing less than 5 times. Then we trim the sentences to a maximum of 16 words. Finally, we get a vocabulary of 10,369 words. We use the visual features extracted by Up-Down [3]. The dimensionalities for both encoder and decoder are $d_h = 512$ (Eq. (1)); and the dimension of the inner-layer in feed-forward networks is 2048. We multiply all heads' attention matrix by the same M (Eq. (1)). We use Adam [25] optimizer following the settings in [57] with $warmup_steps = 20,000$. We first use cross-entropy loss (Eq. (11)) for 15 epochs and initialize the learning rate as $1e^{-5}$ and decay it by 0.8 every 5 epochs. Then the RL-based reward (Eq. (12)) is used for another 35 epochs where the learning rate is reset to $1e^{-5}$ and decay by 0.8 every 5 epochs. The batch size is 10.

5.3. Results

Ablation Studies. We conduct ablation studies to validate the effectiveness of our two key components: the probabilistic graphical model probability and the hierarchical constraint. Specifically, we design and compare the following ablative models. **BASE:** We use the architecture illustrated in Figure 3 without PGM. **PGM:** We incorporate all the PGM modules into the BASE architecture while we do not use the hierarchical constraints in Eq. (10). **APN:** We use the whole APN architecture where the potential probabilities are calculated by Eq. (10).

Results and Analysis. To evaluate the quality of the generated image captions, we measure the similarities between the generated caption with the ground-truth captions by five metrics: CIDEr-D [58], BLEU [45], METEOR [8], ROUGE [35] and SPICE [2].

Table 1. The CHAIR scores and captions’ recall scores in terms of Object, Attribute, Relation and Gender of ablative models on MS-COCO Karpathy split. \uparrow and \downarrow mean the higher the better and the lower the better, respective.

Models	Object \uparrow	Attribute \uparrow	Relation \uparrow	Gender \uparrow	CHAIRs \downarrow	CHAIRi \downarrow
BASE	28.2	9.2	22.3	62.1	12.4	9.5
PGM	29.4	10.3	23.4	63.3	11.6	8.7
APN	30.9	11.0	24.7	64.5	10.6	6.8

Table 2. The performances of various methods on MS-COCO Karpathy split. The metrics: B@N, M, R, C and S denote BLEU@N, METEOR, ROUGE-L, CIDEr-D and SPICE.

Models	B@4	M	R	C	S
GCN-LSTM [72]	38.2	28.5	58.3	127.6	22.0
SGAE [70]	38.4	28.4	58.6	127.8	22.1
CNM [71]	38.9	28.4	58.8	127.9	22.0
HIP [73]	39.1	28.9	59.2	130.6	22.3
ETA [31]	39.9	28.9	58.9	126.6	22.7
ORT [18]	38.6	28.7	58.4	128.3	22.6
AoANet [20]	38.9	29.2	58.8	129.8	22.4
\mathcal{M}^2 Transformer [16]	39.1	29.2	58.6	131.2	22.6
BASE	38.4	28.5	58.1	128.7	22.0
PGM	38.9	28.9	58.5	130.4	22.6
APN	39.6	29.2	59.1	131.8	23.0

The bottom section of Table 2 shows the performances of various baselines on MS-COCO Karpathy split. Compared with BASE, our PGM has a better performance on nearly all metrics. This suggests that the incorporated sparse assumption of our PGM can improve the qualities of the generated captions. APN achieves the best performance among all baselines, which demonstrates that the designed hierarchical property can constrain the attention to improve the model by enforcing a tree structure.

To validate whether more local contexts are exploited than the self-attention based Transformer, we also evaluate the recalls of different kinds of words: objects, attributes, relations, and genders. Specifically, we calculate the recall by counting whether a word in ground-truth captions appears in the generated captions. Note that since an image is assigned with five captions and different captions may use distinctive words, while we measure the recall of these words in only one generated caption. Thus the more distinctive a kind of words, *e.g.*, attribute, the recall is lower. We also exploit the bias measurements CHAIR [48] to validate whether our APN can avoid trivial or even negative global dependencies coming from the dataset bias.

All these scores are listed in Table 1 and we can find that our APN has the highest recalls of each kind of word and the lowest CHAIR scores. Both results demonstrate that our APN can exploit the local context to generate more descriptive words, and meantime avoid the negative global dependencies to generate less biased captions.

Qualitative Results. We visualize the language and image trees parsed by Algorithm 1 in the first two rows of Figure 4. To parse trees, we use the M of all N_L and N_V self-attention layers in Figure 3. The left and middle parts show the trees of the generated caption and the given image,

Table 3. Leaderboard of various methods using single model on the MS-COCO online test server.

Models	B@4		M		R		C	
	c5	c40	c5	c40	c5	c40	c5	c40
SCST [47]	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.0
LSTM-A [74]	35.6	65.2	27.0	35.4	56.4	70.5	116.0	118.0
Up-Down [3]	36.9	68.5	27.6	36.7	57.1	72.4	117.9	120.5
RFNet [22]	38.0	69.2	28.2	37.2	58.2	73.1	122.9	125.1
SGAE [70]	37.8	68.7	28.1	37.0	58.2	73.1	122.7	125.5
CNM [71]	38.4	69.3	28.2	37.2	58.4	73.4	123.8	126.0
ETA [31]	38.9	70.2	28.6	38.0	58.6	73.9	122.1	124.4
AoANet [20]	37.3	68.1	28.3	37.2	57.9	72.8	124.0	126.2
APN	38.9	70.2	28.8	38.0	58.7	73.7	126.3	127.6

respectively. Intuitively, our APN generates captions in a phrase-by-phrase manner instead of a word-by-word manner because our model considers previous words when generating new words at the phrase level as shown in Eq. (5). For example, in the first row of Figure 4, when generating the word “pizza”, our captioner first considers whether it is in the same cluster with “a pepperoni”. Furthermore, the PGM probability matrix M calculated by Eq. (8) contains the local context of clusters - phrases, and the phrase-level local information is incorporated when generating the captions. In this way, more details will be described, compared to BASE that only describes “a pizza”. Besides, the second case demonstrates that APN alleviates the bias. Since “person” and “board” have high co-occurrence in the training set, Base easily learns such bias to generate “person-play-board”, while APN learns the significant local context “dog-play-board” and then generates the right description. Also, by comparing language and visual trees, we can see that they have similar structures, *e.g.*, in the “pizza case”, the leftmost leaves of both trees focus on “a pepperoni pizza” and the rightmost leaves focus on “a table”. This observation suggests that the hidden structures are transferred from the language domain into the vision domain.

Comparisons to State-of-the-art methods. We compare our APN with various state-of-the-art models trained by the RL-based reward (Eq. (12)) which are GCN-LSTM [72], SGAE [70], CNM [71], HIP [73], AoA [20], ORT [18], ETA [31] and \mathcal{M}^2 Transformer [16]. Specifically, GCN-LSTM, SGAE, CNM, and HIP incorporate additional graph annotations into the model, *e.g.*, GCN-LSTM exploits the objects’ pairwise relationships acquired from a pre-trained relation classifier. For AoA, ORT, ETA and \mathcal{M}^2 Transformer, all of them exploit Transformer as the backbone. Note that we only compare with the captioners that have a similar scale as ours and do not compare with certain large-scale captioners like OSCAR [34].

Table 2 reports the results of various captioners on the Karpathy test split. We can see that our APN achieves a higher CIDEr-D score than the other state-of-the-art models. In particular, compared with the self-attention based models, *e.g.*, ORT and AOA, APN transfers the dense graph structure into a sparse one, which encourages the captioner to exploit more crucial local contexts for captioning. Also,

compared with the captioners that use graph data generated by other pre-trained parsers, *e.g.*, SGAE or HIP, our APN suffers less from the distribution shifts. We also report our single model’s performance on the MS COCO online test server in Table 3, which demonstrates that our model achieves better performances than the other ones in terms of most metrics. We do not compare with [44] because X-LAN needs larger batch size and more GPUs. Furthermore, we plan to extend our PGM framework to other state-of-the-art transformers, *e.g.*, AoANet [20], \mathcal{M}^2 Transformer [16] and X-LAN [44] in the future.

6. Visual Question Answering

6.1. Architecture

The APN architecture for VQA is illustrated in Figure 5, where S_V and S_L denote the embeddings of the images and the questions. Note that the S_V is linearized by sorting the RoIs from top-left to bottom-right. In experiments, we set $N_V = 1$, $N_L = 6$, and $N_C = 6$. In particular, N_V is set to 1 because the input visual features are extracted from the last few layers of Faster-RCNN [3, 46], which have already captured abundant high-level knowledge. The outputs of the cross-attention include two parts, which are added and then projected to a 3,129-dimensional vector, where 3,129 is the number of the most frequent answers. We follow [54] to train our model by the binary cross-entropy loss.

6.2. Dataset, Settings and Metrics

VQA-v2 Dataset [5] includes images from the MS-COCO dataset, with 3 questions for each image and 10 answers per question. It has 80k training images and 40k validation images available offline. The online evaluation provides test-dev and test-std splits, each of which has 80k images. For offline evaluation, the validation images are split into two sets, typically for validation and test, separately.

Settings. We set the dimensionality of the image feature, the question embedding and the multi-modal embedding to 2,048, 512 and 1,024, respectively. In the ablation studies, we set the hidden size of the multi-head attention to 512 and only train our model on the training set of the VQA2.0 dataset. We multiply all heads’ attention matrix by the same M (Eq. (1)). When compared with the other state-of-the-art models, we follow the conventions to set the hidden size of the multi-head attention to 1024 and also exploit the question-answer pairs in Visual Genome [27] to train our model. We apply the Adam [25] optimizer to train our model and follow MCAN [78] to set the learning rate to $\min(2.5te^{-5}, 1e^{-4})$, where t is the training epoch and after 10 epochs, the learning rate decays by 0.2 every 2 epochs. The batch size is 64 and the training epoch is 13.

Table 4. VQA-v2 Val accuracy scores of ablative baselines.

Models	Yes/No	Number	Other	Overall
BASE	83.30	47.95	57.30	65.84
PGM	83.90	48.63	57.92	66.51
APN	84.99	49.71	58.66	67.38

Table 5. VQA-v2 test-dev and test-std accuracy of various models.

Models	Test-dev				Test-std
	Yes/No	Number	Other	Overall	Overall
DCN [42]	83.51	46.61	57.26	66.87	66.97
VCTREE [52]	84.28	47.78	59.11	68.19	68.49
BAN [24]	85.42	54.04	60.52	70.04	70.35
DFAF [17]	86.09	53.32	60.49	70.22	70.34
MCAN [78]	85.82	53.26	60.72	70.63	70.90
TRRNet [69]	87.27	51.89	61.02	70.80	71.20
APN	87.44	52.68	61.18	71.14	71.33

6.3. Results

Ablation Studies. We design the similar baselines **BASE**, **PGM**, and **APN** as in Section 5 to test the effectiveness of our two key implementations: the PGM module and the hierarchical constraint. We report the accuracies of different question types on offline local validation split to compare the performances of these models, which are reported in Table 4. From this table, we can see that our APN achieves the highest accuracies on all question types. Hence, the extracted sparse and hierarchical structures are useful for VQA models to get more correct answers.

Qualitative Results. We use the M from all $N_c = 6$ self-attention layers of the decoder in Figure 5 to parse the trees. Two examples are illustrated in the last two rows of Figure 4 to demonstrate how the local contexts help the model correctly answer the questions. For example, we can find that the local context, *e.g.*, "old-fashioned sink", "hour hand" is gathered in the same cluster for both language and visual trees. In this way, our APN can accurately understand the questions and then attend to the right regions of an image to get the correct answers. Without recognizing the local context, BASE model may neglect the essential adjectives and thus fall into the wrong answers. For example, in the last row, the answer of BASE is "8" to the question, which is confused by the number under the minute hand.

Comparisons to State-of-the-art methods We compare our APN with certain state-of-the-art models that exploit graph priors or are designed based on the Transformer. Their performances are reported in Table 5. Note that we do not compare our method with the large-scale pre-training models like ERNIE-VIL [76]. From this table, we can see that our APN achieves higher accuracies than the others. For example, compared with VCTREE [52] which also incorporates sparse tree priors or TRRNet [69] which is built based on Transformer, our APN gets the highest accuracies. These comparisons suggest that APN can utilize the sparse tree more effectively than VCTREE, and using the sparse and hierarchical structure has advantages in solving VQA

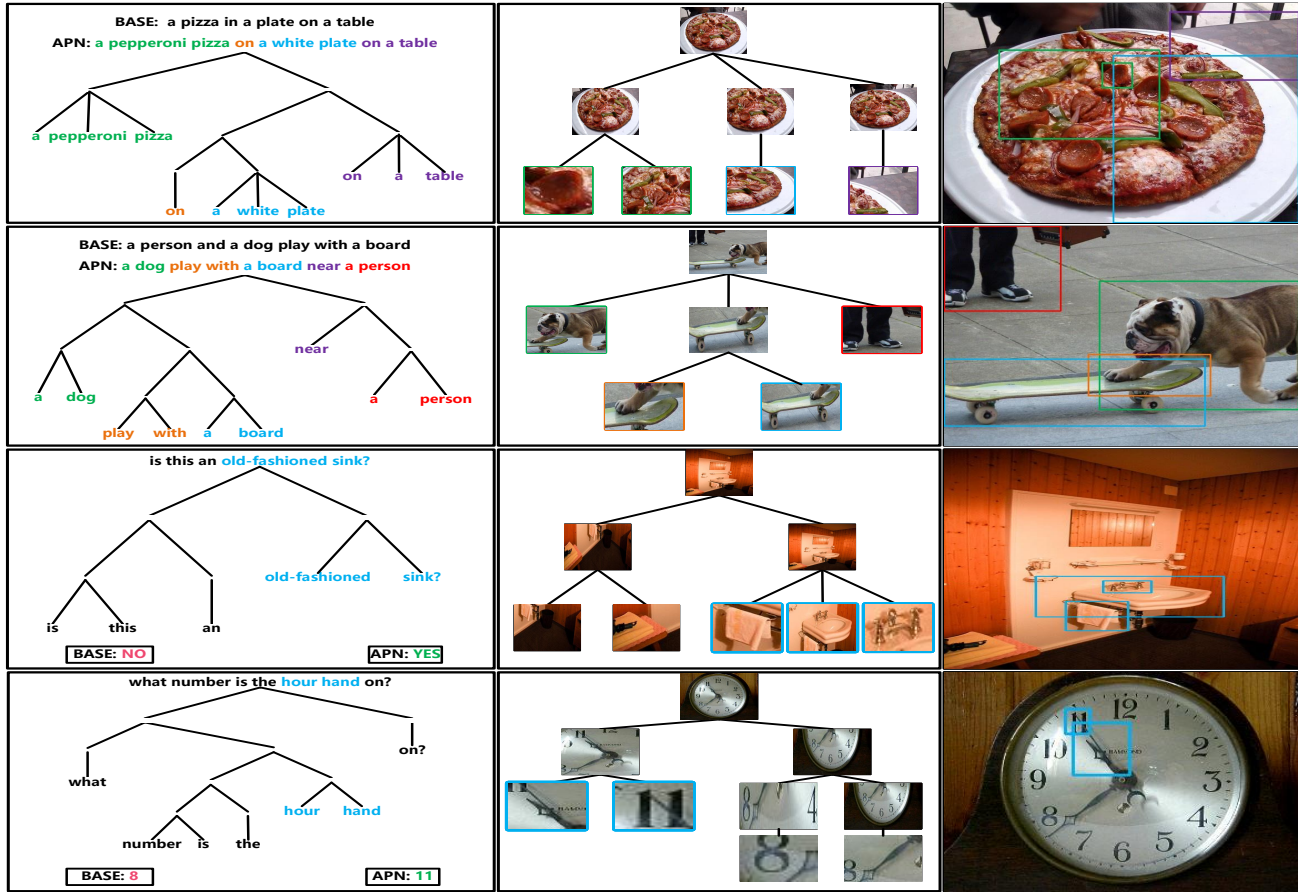


Figure 4. The parsed trees. The first and last two rows are examples for captioning and VQA, respectively. For captioning, the probabilities of both N_V and N_L layers in Figure 3 are used. For VQA, the probabilities of N_C layers in Figure 5 are used. The same colors between language and visual trees show the alignments. The red/green color in VQA cases indexes the correct/wrong answers.

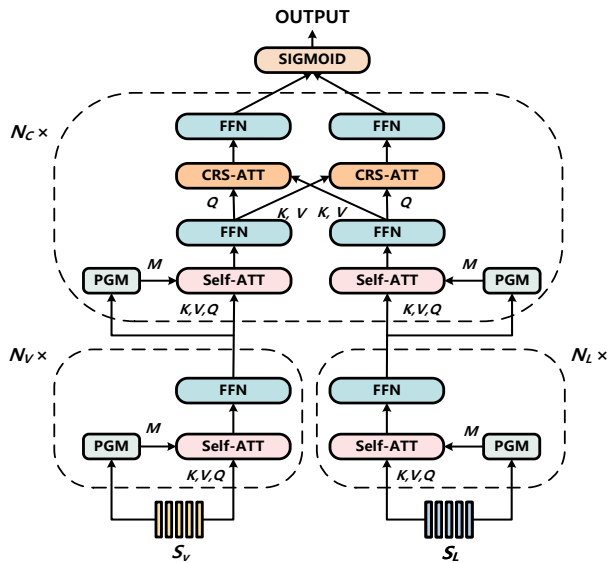


Figure 5. An overview of our APN for VQA. The FFN, CRS-ATT and Self-ATT denote the feed-forward network, cross-attention network and self-attention network, respectively.

task over the Transformer’s fully connected structure.

7. Conclusion

In this paper, we impose a Probabilistic Graphical Model (PGM) on the self-attention layers of a Transformer to incorporate the sparse assumption into the original fully connected one. Then, trivial global dependencies can be avoided, and critical local context can be discovered and exploited. Furthermore, we stacked the constrained self-attention layers and imposed hierarchical constraints on them by which a tree can be implicitly parsed. In this way, the model can unsupervisedly parse trees during the end-to-end training. A tree parsing algorithm was also provided, which exploits the calculated PGM probabilities to extract the hidden trees. Thus, we can figure out the hidden structure of each sample. We proposed two different APNs on Image Captioning and Visual Question Answering, and the results demonstrate that APN can improve both tasks compared with self-attention based Transformers.

Acknowledgements. This work is partially supported by NTU TIER2 and Monash FIT Start-up Grant.

References

- [1] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019. [1](#)
- [2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. [5](#)
- [3] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. [2](#), [5](#), [6](#), [7](#)
- [4] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48, 2016. [2](#)
- [5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. [2](#), [7](#)
- [6] Mohit Bajaj, Lanjun Wang, and Leonid Sigal. G3graphground: Graph-based language grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4281–4290, 2019. [2](#)
- [7] James K Baker. Trainable grammars for speech recognition. *The Journal of the Acoustical Society of America*, 65(S1):S132–S132, 1979. [2](#)
- [8] Satyanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005. [5](#)
- [9] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. [1](#), [3](#)
- [10] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006. [2](#), [4](#)
- [11] Matthew M Botvinick. Hierarchical models of behavior and prefrontal function. *Trends in cognitive sciences*, 12(5):201–208, 2008. [2](#)
- [12] Danqi Chen and Christopher D Manning. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750, 2014. [2](#)
- [13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014. [2](#)
- [14] Shizhe Chen, Qin Jin, Peng Wang, and Qi Wu. Say as you wish: Fine-grained control of image caption generation with abstract scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9962–9971, 2020. [2](#)
- [15] Noam Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT press, 2014. [1](#)
- [16] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020. [1](#), [6](#), [7](#)
- [17] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6639–6648, 2019. [7](#)
- [18] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, pages 11137–11147, 2019. [1](#), [6](#)
- [19] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 804–813, 2017. [2](#)
- [20] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4634–4643, 2019. [2](#), [6](#), [7](#)
- [21] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015. [2](#)
- [22] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 499–515, 2018. [6](#)
- [23] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. [2](#)
- [24] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, pages 1564–1574, 2018. [7](#)
- [25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [5](#), [7](#)
- [26] Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009. [2](#)
- [27] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017. [2](#), [7](#)
- [28] Alex Kulesza and Fernando Pereira. Structured learning with approximate inference. In *Advances in neural information processing systems*, pages 785–792, 2008. [2](#)

- [29] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001. [2](#), [3](#)
- [30] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. [1](#)
- [31] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8928–8937, 2019. [6](#)
- [32] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10313–10322, 2019. [2](#)
- [33] Shiyang Li, Xiaoyong Jin, Yao Xuan, Xiyong Zhou, Wenhu Chen, Yu-Xiang Wang, and Xifeng Yan. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems*, pages 5243–5253, 2019. [1](#)
- [34] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020. [6](#)
- [35] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004. [5](#)
- [36] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. [5](#)
- [37] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4673–4682, 2019. [2](#)
- [38] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1950–1959, 2019. [1](#)
- [39] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330, 1993. [2](#)
- [40] Victor Milewski, Marie-Francine Moens, and Iacer Calixto. Are scene graphs good enough to improve image captioning? *arXiv preprint arXiv:2009.12313*, 2020. [2](#)
- [41] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012. [3](#)
- [42] Duy-Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6087–6096, 2018. [7](#)
- [43] Sebastian Nowozin and Christoph H Lampert. *Structured learning and prediction in computer vision*. Now publishers Inc, 2011. [2](#)
- [44] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10971–10980, 2020. [1](#), [7](#)
- [45] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. [5](#)
- [46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. [7](#)
- [47] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, volume 1, page 3, 2017. [5](#), [6](#)
- [48] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2018. [6](#)
- [49] Ruslan Leont’evich Stratonovich. Conditional markov processes. In *Non-linear transformations of stochastic processes*, pages 427–453. Elsevier, 1965. [2](#)
- [50] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. [1](#)
- [51] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiabin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. [2](#)
- [52] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6619–6628, 2019. [2](#), [7](#)
- [53] Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011. [2](#)
- [54] Damien Teney, Peter Anderson, Xiaodong He, and Anton Van Den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4223–4232, 2018. [7](#)
- [55] Damien Teney, Lingqiao Liu, and Anton van Den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2017. [2](#)
- [56] Zhuowen Tu, Xiangrong Chen, Alan L Yuille, and Song-Chun Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of computer vision*, 63(2):113–140, 2005. [1](#)
- [57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017. [1](#), [3](#), [5](#)

- [58] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 5
- [59] Jesse Vig. Visualizing attention in transformer-based language representation models. *arXiv preprint arXiv:1904.02679*, 2019. 3
- [60] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2
- [61] Hao Wang and Dit-Yan Yeung. Towards bayesian deep learning: A framework and some existing methods. *IEEE Transactions on Knowledge and Data Engineering*, 28(12):3395–3408, 2016. 2
- [62] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3
- [63] Yau-Shian Wang, Hung-Yi Lee, and Yun-Nung Chen. Tree transformer: Integrating tree structures into self-attention. *arXiv preprint arXiv:1909.06639*, 2019. 2, 4
- [64] Sarah Wiegrefe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, 2019. 3
- [65] 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 *International conference on machine learning*, pages 2048–2057, 2015. 2
- [66] Mingzhou Xu, Derek F Wong, Baosong Yang, Yue Zhang, and Lidia S Chao. Leveraging local and global patterns for self-attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3069–3075, 2019. 1
- [67] Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. The penn chinese treebank: Phrase structure annotation of a large corpus. *Natural language engineering*, 11(2):207, 2005. 2
- [68] Baosong Yang, Jian Li, Derek F Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. Context-aware self-attention networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 387–394, 2019. 1
- [69] Xiaofeng Yang, Guosheng Lin, Fengmao Lv, and Fayao Liu. Trnnet: Tiered relation reasoning for compositional visual question answering. 2020. 7
- [70] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019. 2, 6
- [71] Xu Yang, Hanwang Zhang, and Jianfei Cai. Learning to collocate neural modules for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4250–4260, 2019. 6
- [72] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018. 2, 6
- [73] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2621–2629, 2019. 6
- [74] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *IEEE International Conference on Computer Vision, ICCV*, pages 22–29, 2017. 6
- [75] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10502–10511, 2019. 1
- [76] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020. 2, 7
- [77] Weijiang Yu, Jingwen Zhou, Weihao Yu, Xiaodan Liang, and Nong Xiao. Heterogeneous graph learning for visual commonsense reasoning. In *Advances in Neural Information Processing Systems*, pages 2769–2779, 2019. 2
- [78] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6281–6290, 2019. 1, 2, 7
- [79] Chulhee Yun, Srinadh Bhojanapalli, Ankit Singh Rawat, Sashank J Reddi, and Sanjiv Kumar. Are transformers universal approximators of sequence-to-sequence functions? *arXiv preprint arXiv:1912.10077*, 2019. 1
- [80] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018. 2
- [81] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5532–5540, 2017. 2
- [82] Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu. Fast and accurate shift-reduce constituent parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 434–443, 2013. 2