Stochastic And-Or Grammars: A Unified Framework and Logic Perspective*

Kewei Tu
School of Information Science and Technology
ShanghaiTech University, Shanghai, China
tukw@shanghaitech.edu.cn

Abstract

Stochastic And-Or grammars (AOG) extend traditional stochastic grammars of language to model other types of data such as images and events. In this paper we propose a representation framework of stochastic AOGs that is agnostic to the type of the data being modeled and thus unifies various domain-specific AOGs. Many existing grammar formalisms and probabilistic models in natural language processing, computer vision, and machine learning can be seen as special cases of this framework. We also propose a domain-independent inference algorithm of stochastic context-free AOGs and show its tractability under a reasonable assumption. Furthermore, we provide two interpretations of stochastic context-free AOGs as a subset of probabilistic logic, which connects stochastic AOGs to the field of statistical relational learning and clarifies their relation with a few existing statistical relational models.

1 Introduction

Formal grammars are a popular class of knowledge representation that is traditionally confined to the modeling of natural and computer languages. However, several extensions of grammars have been proposed over time to model other types of data such as images [Fu, 1982; Zhu and Mumford, 2006; Jin and Geman, 2006] and events [Ivanov and Bobick, 2000; Ryoo and Aggarwal, 2006; Pei *et al.*, 2011]. One prominent type of extension is stochastic And-Or grammars (AOG) [Zhu and Mumford, 2006]. A stochastic AOG simultaneously models compositions (i.e., a large pattern is the composition of several small patterns arranged according to a certain configuration) and reconfigurations (i.e., a pattern may have several alternative configurations), and in this way it can compactly represent a probabilistic distribution over a large number of patterns. Stochastic AOGs can be used to parse data samples into their compositional structures, which help solve multiple tasks (such as classification, annotation, and segmentation of the data samples) in a unified manner. In this paper we will focus on the context-free subclass of stochastic AOGs, which serves as the skeleton in building more advanced stochastic AOGs.

Several variants of stochastic AOGs and their inference algorithms have been proposed in the literature to model different types of data and solve different problems, such as image scene parsing [Zhao and Zhu, 2011] and video event parsing [Pei *et al.*, 2011]. Our first contribution in this paper is that we provide *a unified representation framework* of stochastic AOGs that is agnostic to the type of the data being modeled; in addition, based on this framework we propose *a domain-independent inference algorithm* that is tractable under a reasonable assumption. The benefits of a unified framework of stochastic AOGs include the following. First, such a framework can help us generalize and improve existing ad hoc approaches for modeling, inference and learning with stochastic AOGs. Second, it also facilitates applications of stochastic AOGs to novel data types and problems and enables the research of general-purpose inference and learning algorithms of stochastic AOGs. Further, a formal definition of stochastic AOGs as abstract probabilistic models makes it easier to theoretically examine their relation with other models such as constraint-based grammar formalism [Shieber, 1992] and sum-product networks [Poon and Domingos, 2011]. In fact, we will show that many of these related models can be seen as special cases of stochastic AOGs.

Stochastic AOGs model compositional structures based on the relations between sub-patterns. Such probabilistic modeling of relational structures is traditionally studied in the field of statistical relational learning [Getoor and Taskar, 2007]. Our second contribution is that we provide *probabilistic logic interpretations* of the unified representation framework of stochastic AOGs and thus show that stochastic AOGs can be seen as a novel type of statistical relational models. The logic interpretations help clarify the relation between stochastic AOGs and a few existing statistical relational models and probabilistic logics that share certain features with stochastic AOGs (e.g., tractable Markov logic [Domingos and Webb, 2012] and stochastic logic programs [Muggleton, 1996]). It may also facilitate the incorporation of ideas from statistical relational learning into the study of stochastic AOGs and at the same time contribute to the research of novel (tractable) statistical relational models.

*This work was supported by the National Natural Science Foundation of China (61503248).

2 Stochastic And-Or Grammars

An AOG is an extension of a constituency grammar used in natural language parsing [Manning and Schütze, 1999]. Similar to a constituency grammar, an AOG defines a set of valid hierarchical compositions of atomic entities. However, an AOG differs from a constituency grammar in that it allows atomic entities other than words and compositional relations other than string concatenation. A stochastic AOG models the uncertainty in the composition by defining a probabilistic distribution over the set of valid compositions.

Stochastic AOGs were first proposed to model images [Zhu and Mumford, 2006; Zhao and Zhu, 2011; Wang *et al.*, 2013; Rothrock *et al.*, 2013], in particular the spatial composition of objects and scenes from atomic visual words (e.g., Garbor bases). They were later extended to model events, in particular the temporal and causal composition of events from atomic actions [Pei *et al.*, 2011] and fluents [Fire and Zhu, 2013]. More recently, these two types of AOGs were used jointly to model objects, scenes and events from the simultaneous input of video and text [Tu *et al.*, 2014].

In each of the previous work using stochastic AOGs, a different type of data is modeled with domain-specific and problem-specific definitions of atomic entities and compositions. Tu et al. [Tu *et al.*, 2013] provided a first attempt towards a more unified definition of stochastic AOGs that is agnostic to the type of the data being modeled. We refine and extend their work by introducing parameterized patterns and relations in the unified definition, which allows us to reduce a wide range of related models to AOGs (as will be discussed in section 2.1). Based on the unified framework of stochastic AOGs, we also propose a domain-independent inference algorithm and study its tractability (section 2.2). Below we start with the definition of stochastic context-free AOGs, which are the most basic form of stochastic AOGs and are used as the skeleton in building more advanced stochastic AOGs.

A *stochastic context-free AOG* is defined as a 5-tuple $\langle \Sigma, N, S, \theta, R \rangle$:

- Σ is a set of terminal nodes representing atomic patterns that are not decomposable;
- N is a set of nonterminal nodes representing high-level patterns, which is divided into two disjoint sets: And-nodes and Or-nodes;
- $S \in N$ is a start symbol that represents a complete pattern;
- θ is a function that maps an instance of a terminal or nonterminal node x to a parameter θ_x (the parameter can take any form such as a vector or a complex data structure; denote the maximal parameter size by m_θ);
- R is a set of grammar rules, each of which takes the form of $x \to C$ representing the generation from a nonterminal node x to a set C of nonterminal or terminal nodes (we say that the rule is "headed" by node x and the nodes in C are the "child nodes" of x).

The set of rules R is further divided into two disjoint sets: And-rules and Or-rules.

- An And-rule, parameterized by a triple $\langle r, t, f \rangle$, represents the decomposition of a pattern into a configuration of non-overlapping sub-patterns. The And-rule specifies a production $r : A \to \{x_1, x_2, \ldots, x_n\}$ for some $n \geq 2$, where A is an And-node and $x_1, x_2, \ldots, x_n$ are a set of terminal or nonterminal nodes representing the sub-patterns. A relation between the parameters of the child nodes, $t(\theta_{x_1}, \theta_{x_2}, \ldots, \theta_{x_n})$, specifies valid configurations of the sub-patterns. This so-called *parameter relation* is typically factorized to the conjunction of a set of binary relations. A *parameter function* f is also associated with the And-rule specifying how the parameter of the And-node A is related to the parameters of the child nodes: $\theta_A = f(\theta_{x_1}, \theta_{x_2}, \ldots, \theta_{x_n})$. We require that both the parameter relation and the parameter function take time polynomial in n and m_θ to compute. There is exactly one And-rule that is headed by each And-node.
- An Or-rule, parameterized by an ordered pair $\langle r, p \rangle$, represents an alternative configuration of a pattern. The Or-rule specifies a production $r : O \to x$, where O is an Or-node and x is either a terminal or a nonterminal node representing a possible configuration. A conditional probability p is associated with the Or-rule specifying how likely the configuration represented by x is selected given the Or-node O. The only constraint in the Or-rule is that the parameters of O and x must be the same: $\theta_O = \theta_x$. There typically exist multiple Or-rules headed by the same Or-node, and together they can be written as $O \to x_1 | x_2 | \ldots | x_n$.

Note that unlike in some previous work, in the definition above we assume deterministic And-rules for simplicity. In principle, any uncertainty in an And-rule can be equivalently represented by a set of Or-rules each invoking a different copy of the And-rule.

Fig. 1(a) shows an example stochastic context-free AOG of line drawings. Each terminal or nonterminal node represents an image patch and its parameter is a 2D vector representing the position of the patch in the image. Each terminal node denotes a line segment of a specific orientation while each nonterminal node denotes a class of line drawing patterns. The start symbol S denotes a class of line drawing images (e.g., images of animal faces). In each And-rule, the parameter relation specifies the relative positions between the sub-patterns and the parameter function specifies the relative positions between the composite pattern and the sub-patterns.

With a stochastic context-free AOG, one can generate a compositional structure by starting from a data sample containing only the start symbol S and recursively applying the grammar rules in R to convert nonterminal nodes in the data sample until the data sample contains only terminal nodes. The resulting compositional structure is a tree in which the root node is S, each non-leaf node is a nonterminal node, and each leaf node is a terminal node; in addition, for each appearance of And-node A in the tree, its set of child nodes in the tree conforms to the And-rule headed by A, and for each appearance of Or-node O in the tree, it has exactly one child node in the tree which conforms to one of the Or-rules headed by O. The probability of the compositional structure is the product of the probabilities of all the Or-rules used in the generation process. Fig. 1(b) shows an image and its com-

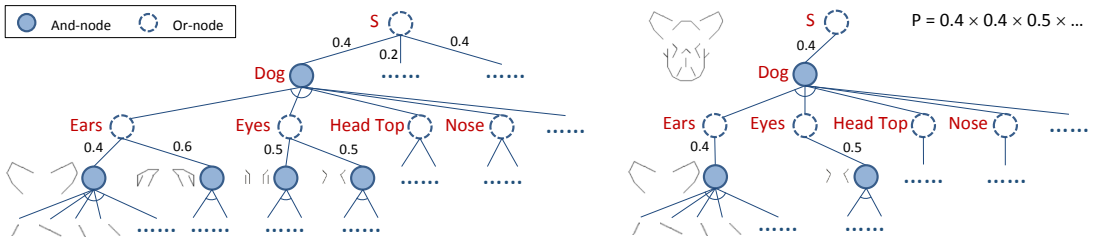

Figure 1: (a) A graphical representation of an example stochastic AOG of line drawings of animal faces. Each And-rule is represented by an And-node and all of its child nodes in the graph. The spatial relations within each And-rule are not shown for clarity. Each Or-rule is represented by an Or-node and one of its child nodes, with its probability shown on the corresponding edge. (b) A line drawing image and its compositional structure generated from the example AOG. Again, the spatial relations between nodes are not shown for clarity. The probability of the compositional structure is partially computed at the top right.

positional structure generated from the example AOG in Fig. 1(a). Given a data sample consisting of only atomic patterns, one can also infer its compositional structure by parsing the data sample with the stochastic context-free AOG. We will discuss the parsing algorithm later.

Our framework is flexible in that it allows different types of patterns and relations within the same grammar. Consider for example a stochastic AOG modeling visually grounded events (e.g., videos of people using vending-machines). We would have two types of terminal or nonterminal nodes that model events and objects respectively. An event node represents a class of events or sub-events, whose parameter is the start/end time of an instance event. An object node represents a class of objects or sub-objects (possibly in a specific state or posture), whose parameter contains both the spatial information and the time interval information of an instance object. We specify temporal relations between event nodes to model the composition of an event from sub-events; we specify spatial relations between object nodes to model the composition of an object from its component sub-objects as well as the composition of an atomic event from its participant objects; we also specify temporal relations between related object nodes to enforce the alignment of their time intervals.

Note that different nonterminal nodes in an AOG may share child nodes. For example, in Fig.1 each terminal node representing a line segment may actually be shared by multiple parent nonterminal nodes representing different line drawing patterns. Furthermore, there could be recursive rules in an AOG, which means the direct or indirect production of a grammar rule may contain its left-hand side nonterminal. Recursive rules are useful in modeling languages and repetitive patterns.

In some previous work, stochastic AOGs more expressive than stochastic context-free AOGs are employed. A typical augmentation over context-free AOGs is that, while in a context-free AOG a parameter relation can only be specified within an And-rule, in more advanced AOGs parameter relations can be specified between any two nodes in the grammar. This can be very useful in certain scenarios. For example, in an image AOG of indoor scenes, relations can be added between all pairs of 2D faces to discourage overlap [Zhao and Zhu, 2011]. However, such relations make inference much more difficult. Another constraint in context-free AOGs that is sometimes removed in more advanced AOGs is the non-overlapping requirement between sub-patterns in an And-rule. For example, in an image AOG it may be more convenient to decompose a 3D cube into 2D faces that share edges [Zhao and Zhu, 2011]. We will leave the formal definition and analysis of stochastic AOGs beyond context-freeness to future work.

2.1 Related Models and Special Cases

Stochastic context-free AOGs subsume many existing models as special cases. Because of space limitation, here we informally describe these related models and their reduction to AOGs and leave the formal definitions and proofs in the supplementary material [Tu, 2016].

Stochastic context-free grammars (SCFG) are clearly a special case of stochastic context-free AOGs. Any SCFG can be converted into an And-Or normal form that matches the structure of a stochastic AOG [Tu and Honavar, 2008]. In a stochastic AOG representing a SCFG, each node represents a string and the parameter of a node is the start/end positions of the string in the complete sentence; the parameter relation and parameter function in an And-rule specify string concatenation, i.e., the substrings must be adjacent and the concatenation of all the substrings forms the composite string represented by the parent And-node.

There have been a variety of grammar formalisms developed in the natural language processing community that go beyond the concatenation relation of strings. For examples, in some formalisms the substrings are interwoven to form the composite string [Pollard, 1984; Johnson, 1985]. More generally, in a grammar rule a linear regular string function can be used to combine lists of substrings into a list of composite strings, as in a linear context-free rewriting system (LCFRS) [Weir, 1988]. All these grammar formalisms can be represented by context-free AOGs with each node representing a list of strings, the node parameter being a list of start/end positions, and in each And-rule the parameter relation and parameter function defining a linear regular string function. Since LCFRSs are known to generate the larger class of mildly context-sensitive languages, context-free AOGs when

instantiated to model languages can be at least as expressive as mildly context-sensitive grammars.

Constraint-based grammar formalisms [Shieber, 1992] are another class of natural language grammars, which associate so-called feature structures to nonterminals and use them to specify constraints in the grammar rules. Such constraints can help model natural language phenomena such as English subject-verb agreement and underlie grammatical theories such as head-driven phrase structure grammars [Pollard and Sag, 1988]. It is straightforward to show that constraint-based grammar formalisms are also special cases of context-free AOGs (with a slight generalization to allow unary And-rules), by establishing equivalence between feature structures and node parameters and between constraints and parameter relations/functions.

In computer vision and pattern recognition, stochastic AOGs have been applied to a variety of tasks as discussed in the previous section. In addition, several other popular models, such as the deformable part model [Felzenszwalb *et al.*, 2008] and the flexible mixture-of-parts model [Yang and Ramanan, 2011], can essentially be seen as special cases of stochastic context-free AOGs in which the node parameters encode spatial information of image patches and the parameter relations/functions encode spatial relations between the patches.

Sum-product networks (SPN) [Poon and Domingos, 2011] are a new type of deep probabilistic models that extend the ideas of arithmetic circuits [Darwiche, 2003] and AND/OR search spaces [Dechter and Mateescu, 2007] and can compactly represent many probabilistic distributions that traditional graphical models cannot tractably handle. It can be shown that any decomposable SPN has an equivalent stochastic context-free AOG: Or-nodes and And-nodes of the AOG can be used to represent sum nodes and product nodes in the SPN respectively, all the node parameters are set to null, parameter relations always return true, and parameter functions always return null. Because of this reduction, all the models that can reduce to decomposable SPNs can also be seen as special cases of stochastic context-free AOGs, such as thin junction trees [Bach and Jordan, 2001], mixtures of trees [Meila and Jordan, 2001] and latent tree models [Choi *et al.*, 2011].

2.2 Inference

The main inference problem associated with stochastic AOGs is parsing, i.e., given a data sample consisting of only terminal nodes, infer its most likely compositional structure (parse). A related inference problem is to compute the marginal probability of a data sample. It can be shown that both problems are NP-hard (see the supplementary material [Tu, 2016] for the proofs). Nevertheless, here we propose an exact inference algorithm for stochastic context-free AOGs that is tractable under a reasonable assumption on the number of valid compositions in a data sample. Our algorithm is based on bottom-up dynamic programming and can be seen as a generalization of several previous exact inference algorithms designed for special cases of stochastic AOGs (such as the CYK algorithm for text parsing).

Algorithm 1 shows the inference algorithm that returns the probability of the most likely parse. After the algorithm terminates, the most likely parse can be constructed by recursively backtracking the selected Or-rules from the start symbol to the terminals. To compute the marginal probability of a data sample, we simply replace the max operation with sum in line 20 of Algorithm 1.

Algorithm 1: Parsing with a stochastic context-free AOG

Input: a data sample X consisting of a set of non-duplicate instances of terminal nodes, a stochastic context-free AOG G in Chomsky normal form

Output: the probability p^* of the most likely parse of X

1: Create an empty map M /* $M[i, O, \theta, T]$ stores the probability of a valid composition of size i with root Or-node O, parameter θ, and set T of terminal instances. */
2: **for all** $x \in X$ **do**
3: $a \leftarrow$ the terminal node that x is an instance of
4: $\theta \leftarrow$ the parameter of x
5: **for all** Or-rule $\langle O \rightarrow a,\ p \rangle$ in G **do**
6: $M[1, O, \theta, \{x\}] \leftarrow p$
7: **for** $i = 2$ **to** $|X|$ **do**
8: **for** $j = 1$ **to** $i - 1$ **do**
9: **for all** $\langle O_1, \theta_1, p_1 \rangle : M[j, O_1, \theta_1, T_1] = p_1$ **do**
10: **for all** $\langle O_2, \theta_2, p_2 \rangle : M[i - j, O_2, \theta_2, T_2] = p_2$ **do**
11: **for all** And-rule $\langle A \rightarrow O_1 O_2,\ t,\ f \rangle$ in G **do**
12: **if** $t(\theta_1, \theta_2) = True$ and $T_1 \bigcap T_2 = \emptyset$ **then**
13: $\phi \leftarrow f(\theta_1, \theta_2)$
14: $T \leftarrow T_1 \bigcup T_2$
15: **for all** Or-rule $\langle O \rightarrow A,\ p_O \rangle$ in G **do**
16: $p \leftarrow p_O p_1 p_2$
17: **if** $M[i, O, \phi, T]$ is null **then**
18: $M[i, O, \phi, T] \leftarrow p$
19: **else**
20: $M[i, O, \phi, T] \leftarrow \max\{p, M[i, O, \phi, T]\}$
21: **return** $\max_\theta M[|X|, S, \theta, X]$ /* S is the start symbol */

In Algorithm 1 we assume the input AOG is in a generalized version of Chomsky normal form, i.e., (1) each And-node has exactly two child nodes which must be Or-nodes, (2) the child nodes of Or-nodes must not be Or-nodes, and (3) the start symbol S is an Or-node. By extending previous studies [Lange and Leiß, 2009], it can be shown that any context-free AOG can be converted into this form and both the time complexity of the conversion and the size of the new AOG is polynomial in the size of the original AOG. We give more details in the supplementary material [Tu, 2016].

The basic idea of Algorithm 1 is to discover valid compositions of terminal instances of increasing sizes, where the size of a composition is defined as the number of terminal instances it contains. Size 1 compositions are simply the terminal instances (line 2–6). To discover compositions of size $i > 1$, the combination of any two compositions of sizes j and $i - j$ ($j < i$) are considered (line 7–20). A complete parse of the data sample is a composition of size $|X|$ with its root being the start symbol S (line 21).

The time complexity of Algorithm 1 is $O(|X|^2 c^2 |G|(|X| + |G|))$ where $c = \max_i |C_i|$ and C_i is the set of valid compositions of size i in the data sample X. In the worst case when all possible compositions of terminal instances from the data sample are valid, we have $c = \binom{|X|}{\lfloor |X|/2 \rfloor}$ which is exponential

in $|X|$. To make the algorithm tractable, we restrict the value of c with the following assumption on the input data sample.

Composition Sparsity Assumption. *For any data sample X and any positive integer $i \leq |X|$, the number of valid compositions of size i in X is polynomial in $|X|$.*

This assumption is reasonable in many scenarios. For text data, for a sentence of length m, a valid composition is a substring of the sentence and the number of substrings of size i is $m - i + 1$. For image data, if we restrict the compositions to be rectangular image patches (as in the hierarchical space tiling model [Wang *et al.*, 2013]), then for an image of size $m = n \times n$ it is easy to show that the number of valid compositions of any specific size is no more than n^3.

3 Logic Perspective of Stochastic AOGs

In a stochastic AOG, And-rules model the relations between terminal and nonterminal instances and Or-rules model the uncertainty in the compositional structure. By combining these two types of rules, stochastic AOGs can be seen as probabilistic models of relational structures and are hence related to the field of statistical relational learning [Getoor and Taskar, 2007]. In this section, we manifest this connection by providing probabilistic logic interpretations of stochastic AOGs. By establishing this connection, we hope to facilitate the exchange of ideas and results between the two previously separated research areas.

3.1 Interpretation as Probabilistic Logic

We first discuss an interpretation of stochastic context-free AOGs as a subset of first-order probabilistic logic with a possible-world semantics. The intuition is that we interpret terminal and nonterminal nodes of an AOG as unary relations, use binary relations to connect the instances of terminal and nonterminal nodes to form the parse tree, and use material implication to represent grammar rules.

We first describe the syntax of our logic interpretation of stochastic context-free AOGs. There are two types of formulas in the logic: And-rules and Or-rules. Each And-rule takes the following form (for some $n \geq 2$).

$$\forall x \, \exists y_1, y_2, \ldots, y_n, A(x) \rightarrow \bigwedge_{i=1}^{n} (B_i(y_i) \wedge R_i(x, y_i)) \\ \wedge R_\theta(\theta(x), \theta(y_1), \theta(y_2), \ldots, \theta(y_n))$$

The unary relation A corresponds to the left-hand side And-node of an And-rule in the AOG; each unary relation B_i corresponds to a child node of the And-rule. We require that for each unary relation A, there is at most one And-rule with $A(x)$ as the left-hand side. The binary relation R_i is typically the `HasPart` relation between an object and one of its parts, but R_i could also denote any other binary relation such as the `Agent` relation between an action and its initiator, or the `HasColor` relation between an object and its color. Note that these binary relations make explicit the nature of the composition represented by each And-rule of the AOG. θ is a function that maps an object to its parameter. R_θ is a relation that combines the parameter relation and parameter function in the And-rule of the AOG and is typically factorized to the conjunction of a set of binary relations.

Each Or-rule takes the following form.

$$\forall x, A(x) \rightarrow B(x) \ : p$$

The unary relation A corresponds to the left-hand side Or-node and B to the child node of an Or-rule in the AOG; p is the conditional probability of $A(x) \rightarrow B(x)$ being true when the grounded left-hand side $A(x)$ is true. We require that for each true grounding of $A(x)$, among all the grounded Or-rules with $A(x)$ as the left-hand side, exactly one is true. This requirement can be represented by two additional sets of constraint rules. First, Or-rules with the same left-hand side are mutually exclusive, i.e., for any two Or-rules $\forall x, A(x) \rightarrow B_i(x)$ and $\forall x, A(x) \rightarrow B_j(x)$, we have $\forall x, A(x) \rightarrow B_i(x) \uparrow B_j(x)$ where $\uparrow$ is the Sheffer stroke. Second, given a true grounding of $A(x)$, the Or-rules with $A(x)$ as the left-hand side cannot be all false, i.e., $\forall x, A(x) \rightarrow \bigvee_i B_i(x)$ where i ranges over all such Or-rules. Further, to simplify inference and avoid potential inconsistency in the logic, we require that the right-hand side unary relation B of an Or-rule cannot appear in the left-hand side of any Or-rule (i.e., the second requirement in the generalized Chomsky normal form of AOG described earlier).

We can divide the set of unary relations into two categories: those that appear in the left-hand side of rules (corresponding to the nonterminal nodes of the AOG) and those that do not (corresponding to the terminal nodes). The first category is further divided into two sub-categories depending on whether the unary relation appears in the left-hand side of And-rules or Or-rules (corresponding to the And-nodes and Or-nodes of the AOG respectively). We require these two sub-categories to be disjoint. There is also a unique unary relation S that does not appear in the right-hand side of any rule, which corresponds to the start symbol of the AOG.

Now we describe the semantics of the logic. The interpretation of all the logical and non-logical symbols follows that of first-order logic. There are two types of objects in the universe of the logic: normal objects and parameters. There is a bijection between normal objects and parameters, and function θ maps a normal object to its corresponding parameter. A possible world is represented by a pair $\langle X, L \rangle$ where X is a set of objects and L is a set of literals that are true. We require that there exists exactly one normal object $s \in X$ such that $S(s) \in L$. In order for all the deterministic formulas (i.e., all the And-rules and the two sets of constraint rules of all the Or-rules) to be satisfied, the possible world must contain a tree structure in which:

1. each node denotes an object in X with the root node being s;
2. each edge denotes a binary relation defined in some And-rule;
3. for each leaf node x, there is exactly one terminal unary relation T such that $T(x) \in L$;
4. for each non-leaf node x, there is exactly one And-node unary relation A such that $A(x) \in L$, and for the child nodes $\{y_1, y_2, \ldots, y_n\}$ of

x in the tree, $\{B_i(y_i)\}_{i=1}^n \cup \{R_i(x, y_i)\}_{i=1}^n \cup \{R_\theta(\theta(x), \theta(y_1), \theta(y_2), \ldots, \theta(y_n))\} \subseteq L$ according to the And-rule associated with relation A;

5. for each node x, if for some Or-node unary relation A we have $A(x) \in L$, then among all the Or-rules with A as the left-hand side, there is exactly one Or-rule such that $B(x) \in L$ where B is the right-hand side unary relation of the Or-rule, and for the rest of the Or-rules we have $\neg B(x) \in L$.

We enforce the following additional requirements to ensure that the possible world contains no more and no less than the tree structure:

1. No two nodes in the tree denote the same object.
2. X and L contain only the objects and relations specified above.

The probability of a possible world $\langle X, L \rangle$ is defined as follows. Denote by R^{Or} the set of Or-rules. For each Or-rule $r : \forall x, A(x) \rightarrow B(x)$, denote by p_r the conditional probability associated with r and define $g_r := \{x \in X | A(x) \in L \wedge B(x) \in L\}$. Then we have:

$$P(\langle X, L \rangle) = \prod_{r \in R^{Or}} p_r^{|g_r|}$$

In this logic interpretation, parsing corresponds to the inference problem of identifying the most likely possible world in which the terminal relations and parameters of the leaf nodes of the tree structure match the atomic patterns in the input data sample. Computing the marginal probability of a data sample corresponds to computing the probability summation of the possible worlds that match the data sample.

Our logic interpretation of stochastic context-free AOGs resembles tractable Markov logic (TML) [Domingos and Webb, 2012; Webb and Domingos, 2013] in many aspects, even though the two have very different motivations. Such similarity implies a deep connection between stochastic AOGs and TML and points to a potential research direction of investigating novel tractable statistical relational models by borrowing ideas from the stochastic grammar literature. There are a few minor differences between stochastic AOGs and TML, e.g., TML does not distinguish between And-nodes and Or-nodes, does not allow recursive rules, enforces that the right-hand side unary relation in each Or-rule is a sub-type of the left-hand side unary relation, and disallows a unary relation to appear in the right-hand side of more than one Or-rule.

3.2 Interpretation as a Stochastic Logic Program

Stochastic logic programs (SLP) [Muggleton, 1996] are a type of statistical relational models that, like stochastic context-free AOGs, are a generalization of stochastic context-free grammars. They are essentially equivalent to two other representations, independent choice logic [Poole, 1993] and PRISM [Sato and Kameya, 2001]. Here we show how a stochastic context-free AOG can be represented by a pure normalized SLP [Cussens, 2001]. Since several inference and learning algorithms have been developed for SLPs and PRISM, our reduction enables the application of these algorithms to stochastic AOGs.

In our SLP program, we have one SLP clause for each And-rule and each Or-rule in the AOG. The overall structure is similar to the probabilistic logic interpretation discussed in section 3.1. For each And-rule, the corresponding SLP clause takes the following form:

$$\begin{aligned} 1.0 : a(X, P) \text{ :- } & b_1(X_1, P_1), b_2(X_2, P_2), \cdots, b_n(X_n, P_n), \\ & append([X_1, \ldots, X_n], X), r_1(X, X_1), r_2(X, X_2), \\ & \cdots, r_n(X, X_n), r_\theta(P, P_1, \ldots, P_n). \end{aligned}$$

The head $a(X, P)$ represents the left-hand side And-node of the And-rule, where X represents the set of terminal instances generated from the And-node and P is the parameters of the And-node. In the body of the clause, b_i represents the i-th child node of the And-rule, r_i represents the relation between the And-node and its i-th child node, $append(\ldots)$ states that the terminal instance set X of the And-node is the union of the instance sets from all the child nodes, and r_θ represents a relation that combines the parameter relation and parameter function of the And-rule. For relations r_i and r_θ, we need to have additional clauses to define them according to the type of data being modeled.

For each Or-rule in the AOG, if the right-hand side is a nonterminal, then we have:

$$p : \ a(X, P) \text{ :- } b(X, P).$$

where p is the conditional probability associated with the Or-rule, a and b represent the left-hand and right-hand sides of the Or-rule respectively, whose arguments X and P have the same meaning as explained above. If the right-hand side of the Or-rule is a terminal, then we have:

$$p : \ a([t], [\ldots]).$$

where t is the right-hand side terminal node and the second argument represents the parameters of the terminal node.

Finally, the goal of the program is

$$\text{:- } s(X, P).$$

which represents the start symbol of the AOG, whose arguments have the same meaning as explained above.

4 Conclusion

Stochastic And-Or grammars extend traditional stochastic grammars of language to model other types of data such as images and events. We have provided a unified representation framework of stochastic AOGs that can be instantiated for different data types. We have shown that many existing grammar formalisms and probabilistic models in natural language processing, computer vision, and machine learning can all be seen as special cases of stochastic context-free AOGs. We have also proposed an inference algorithm for parsing data samples using stochastic context-free AOGs and shown that the algorithm is tractable under the composition sparsity assumption. In the second part of the paper, we have provided interpretations of stochastic context-free AOGs as a subset of first-order probabilistic logic and stochastic logic programs. Our interpretations connect stochastic AOGs to the field of statistical relational learning and clarify their relation with a few existing statistical relational models.

References

[Bach and Jordan, 2001] Francis R Bach and Michael I Jordan. Thin junction trees. In *NIPS*, 2001.

[Choi *et al.*, 2011] Myung Jin Choi, Vincent YF Tan, Animashree Anandkumar, and Alan S Willsky. Learning latent tree graphical models. *The Journal of Machine Learning Research*, 12:1771–1812, 2011.

[Cussens, 2001] James Cussens. Parameter estimation in stochastic logic programs. *Machine Learning*, 44(3):245–271, 2001.

[Darwiche, 2003] Adnan Darwiche. A differential approach to inference in bayesian networks. *Journal of the ACM (JACM)*, 50(3):280–305, 2003.

[Dechter and Mateescu, 2007] Rina Dechter and Robert Mateescu. And/or search spaces for graphical models. *Artificial intelligence*, 171(2):73–106, 2007.

[Domingos and Webb, 2012] Pedro Domingos and William Austin Webb. A tractable first-order probabilistic logic. In *AAAI*, 2012.

[Felzenszwalb *et al.*, 2008] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[Fire and Zhu, 2013] A. Fire and S.C. Zhu. Using causal induction in humans to learn and infer causality from video. In *35th Annual Cognitive Science Conference (CogSci)*, 2013.

[Fu, 1982] King Sun Fu. *Syntactic pattern recognition and applications*, volume 4. Prentice-Hall Englewood Cliffs, 1982.

[Getoor and Taskar, 2007] Lise Getoor and Ben Taskar. *Introduction to statistical relational learning*. MIT press, 2007.

[Ivanov and Bobick, 2000] Yuri A. Ivanov and Aaron F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):852–872, 2000.

[Jin and Geman, 2006] Ya Jin and Stuart Geman. Context and hierarchy in a probabilistic image model. In *CVPR*, 2006.

[Johnson, 1985] Mark Johnson. Parsing with discontinuous constituents. In *ACL*, 1985.

[Lange and Leiß, 2009] Martin Lange and Hans Leiß. To CNF or not to CNF? an efficient yet presentable version of the CYK algorithm. *Informatica Didactica*, 8:2008–2010, 2009.

[Manning and Schütze, 1999] Christopher D. Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999.

[Meila and Jordan, 2001] Marina Meila and Michael I Jordan. Learning with mixtures of trees. *The Journal of Machine Learning Research*, 1:1–48, 2001.

[Muggleton, 1996] Stephen Muggleton. Stochastic logic programs. *Advances in inductive logic programming*, 32:254–264, 1996.

[Pei *et al.*, 2011] Mingtao Pei, Yunde Jia, and Song-Chun Zhu. Parsing video events with goal inference and intent prediction. In *ICCV*, 2011.

[Pollard and Sag, 1988] Carl Pollard and Ivan A. Sag. *Information-based Syntax and Semantics: Vol. 1: Fundamentals*. Center for the Study of Language and Information, Stanford, CA, USA, 1988.

[Pollard, 1984] Carl Pollard. Generalized context-free grammars, head grammars and natural language. *Ph.D. diss., Stanford University*, 1984.

[Poole, 1993] David Poole. Probabilistic horn abduction and bayesian networks. *Artificial intelligence*, 64(1):81–129, 1993.

[Poon and Domingos, 2011] Hoifung Poon and Pedro Domingos. Sum-product networks : A new deep architecture. In *UAI*, 2011.

[Rothrock *et al.*, 2013] Brandon Rothrock, Seyoung Park, and Song-Chun Zhu. Integrating grammar and segmentation for human pose estimation. In *CVPR*, 2013.

[Ryoo and Aggarwal, 2006] M. S. Ryoo and J. K. Aggarwal. Recognition of composite human activities through context-free grammar based representation. In *CVPR*, 2006.

[Sato and Kameya, 2001] Taisuke Sato and Yoshitaka Kameya. Parameter learning of logic programs for symbolic-statistical modeling. *Journal of Artificial Intelligence Research*, pages 391–454, 2001.

[Shieber, 1992] Stuart M Shieber. *Constraint-based grammar formalisms: parsing and type inference for natural and computer languages*. MIT Press, 1992.

[Tu and Honavar, 2008] Kewei Tu and Vasant Honavar. Unsupervised learning of probabilistic context-free grammar using iterative biclustering. In *ICGI*, 2008.

[Tu *et al.*, 2013] Kewei Tu, Maria Pavlovskaia, and Song-Chun Zhu. Unsupervised structure learning of stochastic and-or grammars. In *NIPS*, 2013.

[Tu *et al.*, 2014] Kewei Tu, Meng Meng, Mun Wai Lee, Tae Eun Choe, and Song-Chun Zhu. Joint video and text parsing for understanding events and answering queries. *IEEE MultiMedia*, 2014.

[Tu, 2016] Kewei Tu. Stochastic And-Or grammars: A unified framework and logic perspective (supplementary material). `http://sist.shanghaitech.edu.cn/faculty/tukw/ijcai16-sup.pdf`, 2016.

[Wang *et al.*, 2013] Shuo Wang, Yizhou Wang, and Song-Chun Zhu. Hierarchical space tiling for scene modeling. In *ACCV*. 2013.

[Webb and Domingos, 2013] William Austin Webb and Pedro Domingos. Tractable probabilistic knowledge bases with existence uncertainty. In *AAAI Workshop: Statistical Relational Artificial Intelligence*, 2013.

[Weir, 1988] David Jeremy Weir. Characterizing mildly context-sensitive grammar formalisms. *Ph.D. diss., University of Pennsylvania*, 1988.

[Yang and Ramanan, 2011] Yi Yang and Deva Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011.

[Zhao and Zhu, 2011] Yibiao Zhao and Song Chun Zhu. Image parsing with stochastic scene grammar. In *NIPS*, 2011.

[Zhu and Mumford, 2006] Song-Chun Zhu and David Mumford. A stochastic grammar of images. *Found. Trends. Comput. Graph. Vis.*, 2(4):259–362, 2006.