Towards a Unified Logic Perspective of Stochastic And-Or Grammars

Kewei Tu TUKW@SHANGHAITECH.EDU.CN
ShanghaiTech University, No. 8 Building, 319 Yueyang Road, Shanghai 200031, China

Abstract

Stochastic And-Or grammars extend traditional stochastic grammars of language to model other types of data such as images and events. In this short paper, we discuss our work in progress that aims to provide a unified framework and a probabilistic logic interpretation of stochastic And-Or grammars. We also present a general inference algorithm of stochastic And-Or grammars and discuss its tractability.

1. Introduction

Stochastic grammars, traditionally used to model artificial and natural languages, have been extended to model other types of data such as images (Zhu & Mumford, 2006; Jin & Geman, 2006; Zhao & Zhu, 2011) and events (Ivanov & Bobick, 2000; Ryoo & Aggarwal, 2006; Zhang et al., 2011; Pei et al., 2011). One prominent example of such extensions is stochastic And-Or grammars (AOG) (Zhu & Mumford, 2006). A stochastic AOG simultaneously represents hierarchical compositions (i.e., a pattern can be decomposed into a certain configuration of sub-patterns) and reconfigurations (i.e., a pattern may have multiple alternative configurations), and in this way it can compactly model 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, in a unified way.

Although stochastic AOGs have been applied or adapted to solve a variety of problems, such as image scene parsing (Zhao & Zhu, 2011), video event parsing (Pei et al., 2011), and event causal parsing (Fire & Zhu, 2013), there has been little work to provide a theoretic framework that unifies the applications of stochastic AOGs to different types of data. Besides, there are several probabilistic models proposed in the areas of stochastic relational learning and probabilistic logic that share certain features with stochastic AOGs (e.g., stochastic logic programs (Muggleton, 1996), tractable Markov logic (Domingos & Webb, 2012)), but their relations with stochastic AOGs have not been thoroughly studied. In this short paper, we report our work in progress that aims to provide a unified framework and a probabilistic logic interpretation of stochastic AOGs that is agnostic to the type of data being modeled. We focus on the context-free subclass of stochastic AOGs, which is amenable to efficient inference and learning and serves as the skeleton of more complex stochastic AOGs. The potential impacts of our work include the following.

- A unified framework of stochastic AOGs can help us generalize and improve existing ad hoc approaches for modeling, inference and learning. It also facilitates applications of stochastic AOGs to novel data and problems and enables the research of general-purpose inference and learning algorithms of stochastic AOGs.
- A logic interpretation of stochastic AOGs can be used to clarify their relationship with existing stochastic relational models and probabilistic logics. It may facilitate the incorporation of ideas from the area of stochastic relational learning into stochastic AOGs, thus improving or enhancing existing AOG-based approaches. It may also contribute to the research of novel (tractable) stochastic relational models.

The rest of the paper is organized as follows. Section 2 introduces stochastic AOGs and their applications. Section 3 presents a preliminary logic interpretation of stochastic context-free AOGs. Section 4 studies the inference algorithm of AOGs and its tractability. Section 5 summarizes our ongoing and planned research.

2. Stochastic And-Or Grammars

An AOG is an extension of a constituency grammar used in natural language parsing (Manning & 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 specifies additional relations between the entities. A stochastic AOG models the uncertainty in the composition

Proceedings of the 31^{st} International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32. Copyright 2014 by the author(s).

by defining a probabilistic distribution over the set of valid compositions.

We first define stochastic context-free AOGs, which is the simplest form of stochastic AOGs. We follow the work of Tu et al. (2013) and give a definition that is agnostic to the type of the data being modeled. A stochastic context-free AOG is defined as a 5-tuple $\langle \Sigma, N, S, R, P \rangle$. Σ is a set of terminal nodes representing atomic patterns that are not decomposable; N is a set of nonterminal nodes representing decomposable patterns, which is divided into two disjoint sets: And-nodes N^{AND} and Or-nodes N^{OR}; $S \in N$ is a start symbol that represents a complete entity; R is a set of grammar rules, each of which represents the generation from a nonterminal node to a set of nonterminal or terminal nodes; P is the set of probabilities assigned to the grammar rules. The set of grammar rules R is divided into two disjoint sets: And-rules and Or-rules.

- An And-rule represents the decomposition of a pattern into a configuration of non-overlapping sub-patterns. It takes the form of $A \rightarrow a_1 a_2 \ldots a_n$, where $A \in N^{\mathrm{AND}}$ is a nonterminal And-node and $a_1 a_2 \ldots a_n$ is a set of terminal or nonterminal nodes representing the sub-patterns. A set of relations are specified between the sub-patterns and between the nonterminal node A and the sub-patterns, which configure how these sub-patterns form the composite pattern represented by A. The probability of an And-rule is specified by the energy terms defined on the relations.

- An Or-rule represents an alternative configuration of a composite pattern. It takes the form of $O \rightarrow a$, where $O \in N^{\mathrm{OR}}$ is a nonterminal Or-node, and a is either a terminal or a nonterminal node representing a possible configuration. The set of Or-rules with the same left-hand side can be written as $O \rightarrow a_1 | a_2 | \ldots | a_n$. The probability of an Or-rule specifies how likely the alternative configuration represented by the Or-rule is selected.

With a stochastic context-free AOG, one can generate a valid 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 until the data sample contains only terminal nodes (atomic patterns). Given a data sample consisting of atomic patterns, one can also infer its compositional structure by parsing the data sample with the stochastic context-free AOG.

In a stochastic context-free AOG, a relation can only be specified within an And-rule, i.e., between a pattern and its immediate sub-patterns or between two sub-patterns of the same parent. In more complex stochastic AOGs, however, relations are allowed to be specified between any two nodes. 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 & Zhu, 2011). However, such relations make parsing much more difficult and hence approximate inference such as MCMC is typically employed.

Stochastic AOGs are first proposed to model images (Zhu & Mumford, 2006; Zhao & 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 are later extended to model events, in particular the temporal and causal composition of events from atomic actions (Pei et al., 2011) and fluents (Fire & Zhu, 2013). More recently, these two types of AOGs are used jointly to model objects, scenes and events from the simultaneous input of video and text (Tu et al., 2014).

3. A Logic Perspective of Stochastic Context-free AOG

In this section we describe our preliminary interpretation of a stochastic context-free AOG as a probabilistic logical system. There are two types of entities in the universe of the logic: normal entities and parameters. There is a bijection between normal entities and parameters. For each normal entity x, we use $x.\theta$ to denote its parameter (so "$.\theta$" can be seen as a unary function). For example, in a text AOG, a normal entity represents a word or phrase in a sentence and its parameter is the start/end positions of the word or phrase; in an image AOG, a normal entity represents an image patch and its parameter contains the position and coverage of the patch.

There are two types of formulas in the logic: And-rules and Or-rules. Each And-rule takes the following form.

$$\begin{aligned}
&\forall x\, \exists y_1, y_2, \ldots, y_n, \\
&\qquad A(x) \rightarrow \bigwedge_{i=1}^{n} \left(B_i(y_i) \wedge R_i(x, y_i)\right) \\
&\qquad\qquad \wedge R_\theta(x.\theta, y_1.\theta, y_2.\theta, \ldots, y_n.\theta)
\end{aligned}$$

The unary relations A and B_i typically denote entity types but could also denote attributes. Each binary relation R_i typically denotes the `HasPart` relation but 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. R_θ denotes the relation between the entity parameters, e.g., in an image AOG R_θ could constrain the relative positions between all the entities. R_θ is typically factorized to the conjunction of a set of binary relations. There can be uncertainty associated with R_θ (e.g., a bivariate normal distribution over the relative position between two entities) and otherwise the And-rule is deterministic.

Each Or-rule takes the following form.

$$\forall x, A(x) \to B(x)$$

Each Or-rule is associated with the conditional probability of $A(x) \to B(x)$ being true when the grounded left-hand side $A(x)$ is true. We require that for each 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 sets of rules. First, Or-rules with the same left-hand side are mutually exclusive, i.e., for any two Or-rules $\forall x, A(x) \to B_i(x)$ and $\forall x, A(x) \to B_j(x)$, we have $\forall x, A(x) \to B_i(x) \uparrow B_j(x)$ where $\uparrow$ is the Sheffer stroke. Second, given a grounding of $A(x)$ the Or-rules with $A(x)$ as the left-hand side cannot be false at the same time, i.e., $\forall x, A(x) \to \bigvee_i B_i(x)$ where i ranges over all such rules.

We can divide the set of unary relations into two categories: those that appear in the left-hand side of rules (which we call nonterminal relations) and those that do not (which we call terminal relations). The set of nonterminal relations is further divided into two sub-categories: And-relations (those appearing in the left-hand side of And-rules) and Or-relations (those appearing in the left-hand side of Or-rules). We require that these two sub-categories are disjoint. There also exists a special nonterminal relation S that does not appear in the right-hand side of any rule, which corresponds to the start symbol of the AOG. We require that in the universe of the logic, there exists exactly one normal entity s such that $S(s)$ is true.

The above interpretation is still preliminary and there is a few important issues that need to be addressed. In particular, we need to enforce that every possible world contains exactly a tree or DAG of entities (in which each edge represents a binary relation) that is consistent with the logic.

Now we give the logic interpretation of two example AOGs. The first is an image AOG of simple line drawing. Each normal entity represents an image patch and its parameter is a 2D vector representing the position of the patch. Each nonterminal relation denotes a class of line drawing patterns while each terminal relation denotes a line segment of a specific orientation. The start relation S denotes a class of line drawing images. All the parent-child binary relations in And-rules are the `HasPart` relation. The parameter relations in And-rules specify relative positions between entities.

The second example is an event AOG grounded to video frames (Pei et al., 2011). There are two types of normal entities: event entities and object entities. An event entity represents an event, whose parameter is the start/end time of the event. An object entity represents an object or human at a specific time interval, whose parameter contains the time interval information and the spatial information of the object. A nonterminal relation denotes either an event type or an object type, while a terminal relation denotes an atomic object. The start relation S denotes an event type. There are three types of And-rules:

- In an event-subevent rule, each unary relation denotes an event type and all the parent-child binary relations are the `HasPart` relation; the parameter relation specifies the temporal relations between the events in terms of the start/end time.
- In an object-subobject rule, each unary relation denotes an object type and all the parent-child binary relations are the `HasPart` relation; the parameter relation specifies the spatial relations between the objects and enforces the alignment of the time intervals of the objects.
- In an event-object rule, the unary relation in the left-hand side denotes an event type and the unary relations in the right-hand side denote object types; the parent-child binary relations denote the `Agent`, `Patient` and `Location` relations between the event and its initiator(s), target(s) and location respectively; the parameter relation specifies the spatial relations between the objects and enforces the alignment of the time intervals of the event and objects.

4. Inference Tractability

The main inference problem associated with stochastic AOGs is parsing, i.e., given a data sample consisting of only terminal nodes, the task is to infer the most likely parse of the data sample. In our logic perspective, this is equivalent to identifying the most likely possible world in which terminal relations of the leaf entities of the parse tree or DAG match the terminal nodes in the data sample. A related inference problem is to compute the marginal probability of a data sample. In the logic perspective, this is equivalent to compute the probability summation of the possible worlds that match the data sample.

Both of the two inference problems can be solved with a bottom-up dynamic programming algorithm. Algorithm 1 shows the inference algorithm that returns the probability of the most likely parse under the assumption that there is no recursive rules in the AOG. 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 need to replace the max operation with sum in line 21 and 29 of Algorithm 1. We are working to extend the algorithm to handle recursive rules.

Now we analyze the time complexity of Algorithm 1. We

Algorithm 1 Parsing with non-recursive AOG

```
input  a data sample $X$ consisting of a set of terminal nodes
       $\{x_i\}$, the AOG $G$
output  the probability $p^*$ of the most likely parse
 1: find the topological ordering $Q$ of terminals and non-
    terminals of $G$ such that the nodes in the right-hand
    side of a rule always proceed the node in the left-hand
    side.
 2: create an empty map $M$
 3: for each node $a$ in $Q$ do
 4:   $M[a] \leftarrow \{\}$ /* Initialize the set of partial parses
      rooted at $a$. Each partial parse is to be represented
      by its root parameter and probability. */
 5:   if $a$ is a terminal then
 6:     for all $x \in X$ that matches $a$ do
 7:       $M[a] \leftarrow M[a] \cup \{\langle x.\theta, 1.0\rangle\}$
 8:     end for
 9:   else if $a$ is an And-node then
10:     $\langle b_1, \ldots, b_n\rangle \leftarrow$ the child nodes of $a$
11:     for all $\langle\langle\theta_1, p_1\rangle, \ldots, \langle\theta_n, p_n\rangle\rangle \in M[b_1] \times \ldots \times$
        $M[b_n]$ do
12:       if $\langle\theta_1, \ldots, \theta_n\rangle$ matches the relations specified
          between $\langle b_1, \ldots, b_n\rangle$ then
13:         $\phi \leftarrow$ the parent parameter computed from
            $\langle\theta_1, \ldots, \theta_n\rangle$ and the parent-child relations
14:         $M[a] \leftarrow M[a] \cup \{\langle\phi, \prod_{i=1}^{n} p_i\rangle\}$
15:       end if
16:     end for
17:   else /* $a$ is an Or-node */
18:     for all child node $b$ of $a$ and its probability $q$ do
19:       for all $\langle\theta, p\rangle \in M[b]$ do
20:         if $\exists\langle\theta', p'\rangle \in M[a], \theta' = \theta$ then
21:           $M[a] \leftarrow M[a] \cup \{\langle\theta, \max\{p', pq\}\rangle\}$
22:         else
23:           $M[a] \leftarrow M[a] \cup \{\langle\theta, pq\rangle\}$
24:         end if
25:       end for
26:     end for
27:   end if
28: end for
29: return $\max\{p : \langle\theta, p\rangle \in M[S]\}$ /* $S$: start symbol */
```

assume that the AOG is in Chomsky normal form, i.e., in each And-rule the composite pattern is decomposed into two sub-patterns[1]. In a complete run of Algorithm 1, line 1 takes $O(|G|)$ time; line 6–8 take $O(|\Sigma| \times |X|)$ time; line 10–16 take $O(|N^{\text{AND}}| \times |C|^2)$ time where C denotes the set of partial parses (compositions of a subset of terminals) of the data sample that is valid according to the AOG; line 18–26 take $O(|N^{\text{OR}}| \times |C|)$ time; line 29 takes no more than $O(|C|)$ time.

In the worst case when all possible partial compositions of terminals from the data sample are valid, $|C|$ becomes exponential in $|X|$, the number of terminals in the data sample, and therefore Algorithm 1 is intractable. However, we show that the algorithm is tractable under the assumption that only a very small fraction of all possible compositions are valid.

Composition Sparsity Assumption. *For any data sample and any integer α, the number of valid compositions of size α is polynomial in the size of the data sample $|X|$, where the size of a composition is defined as the number of terminals it contains.*

Since the maximal size of a composition cannot exceed the size of the data sample, the total number of valid compositions $|C|$ is now also polynomial. Therefore, the time complexity of Algorithm 1 becomes polynomial in the data sample size $|X|$ and the AOG size $|G|$.

Our assumption is reasonable in many scenarios. For example, in a stochastic context-free grammar of language (a special case of stochastic AOG), for a sentence of length m, a valid composition is a substring of the sentence and the number of substrings of size α is $m - \alpha + 1$. As another example, in the Hierarchical Space Tiling image AOG (Wang et al., 2013), for an image of size $m = n \times n$, a valid composition is a rectangle region of the image and the number of rectangle regions of size α is:

$$\sum_d (n - d + 1)(n - \frac{\alpha}{d} + 1) \leq n^3$$

where d ranges over all the factors of α such that $d \leq \sqrt{\alpha}$.

5. Discussion

We have discussed our work in progress that aims to provide a unified framework of stochastic AOGs and its probabilistic logic interpretation. We have also discussed the inference algorithm of AOGs and its tractability under the composition sparsity assumption. Currently we are working to refine the probabilistic logic interpretation of AOGs and study its relation with existing work such as stochastic logic programs (Muggleton, 1996) and tractable Markov logic (Domingos & Webb, 2012). We are also trying to enhance the inference algorithm so it could handle recursive rules. In the future we plan to accommodate existing AOG learning algorithms into the unified framework and develop novel learning algorithms.

[1] Note that with proper transformation, the Chomsky normal form of a context-free grammar has a size polynomial in the size of the original grammar (Lange & Leiß, 2009).

References

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

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

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

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

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

Manning, Christopher D. and Schütze, Hinrich. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-13360-1.

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

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

Rothrock, Brandon, Park, Seyoung, and Zhu, Song-Chun. Integrating grammar and segmentation for human pose estimation. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 3214–3221. IEEE, 2013.

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

Tu, Kewei, Pavlovskaia, Maria, and Zhu, Song-Chun. Unsupervised structure learning of stochastic and-or grammars. In *Advances in Neural Information Processing Systems*, pp. 1322–1330, 2013.

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

Wang, Shuo, Wang, Yizhou, and Zhu, Song-Chun. Hierarchical space tiling for scene modeling. In *Computer Vision–ACCV 2012*, pp. 796–810. Springer, 2013.

Zhang, Zhang, Tan, Tieniu, and Huang, Kaiqi. An extended grammar system for learning and recognizing complex visual events. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(2):240–255, February 2011. ISSN 0162-8828.

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

Zhu, Song-Chun and Mumford, David. A stochastic grammar of images. *Found. Trends. Comput. Graph. Vis.*, 2 (4):259–362, 2006. ISSN 1572-2740.