Backpropagation of Syntax Errors in Context-Sensitive Languages

Breandan Considine, Jin Guo, Xujie Si

Main Idea

- Matrices over $\mathbb{Z}_2^n$ are useful structures for studying finite state machines
- The operators $\{\text{XOR}, \land, \top\}$ are *functionally complete* logical primitives
- We use them to implement probabilistic context-sensitive program repair

Algebraic Parsing

Given a CFG, $\mathcal{G}' : \langle \Sigma, V, P, S\rangle$ in Chomsky Normal Form (CNF), we can define a *recognizer*, $R : \mathcal{G}' \to \Sigma^n \to \mathbb{B}$ for bounded strings $\sigma : \Sigma^n$ using the following construction. Let 2^V be our domain, 0 be $\varnothing$, $\oplus$ be $\cup$, and $\otimes$:

$$X \otimes Y := \{\, w \mid \langle x, y\rangle \in X \times Y, (w \to xy) \in P \,\}$$

Valiant (1975) shows that if we let $\sigma_r^{\uparrow} := \{V \mid (V \to \sigma_r^{\uparrow}) \in P\}$, initialize the matrix $M^0_{r+1=c}(\mathcal{G}', e) := \sigma_r^{\uparrow}$ and solve for its fixpoint $M^* = M + M^2$,

$$M^0 := \begin{pmatrix} \varnothing & \sigma_1^{\uparrow} & \varnothing & \cdots & \varnothing \\ & & \ddots & \ddots & \vdots \\ & & & & \varnothing \\ & & & & \sigma_n^{\uparrow} \\ \varnothing & \cdots & & & \varnothing \end{pmatrix} \Rightarrow M^* = \begin{pmatrix} \varnothing & \sigma_1^{\uparrow} & \cdots & \mathcal{T} \\ & \ddots & \ddots & \vdots \\ & & & \sigma_n^{\uparrow} \\ \varnothing & \cdots & & \varnothing \end{pmatrix}$$

the recognizer is then defined as: $R(\mathcal{G}', \sigma) := S \in \mathcal{T}? \iff \sigma \in \mathcal{L}(\mathcal{G})?$

Galois Connection

- CYK parsers can be lowered onto $\mathbb{Z}_2^{|V|\times n\times n}$ or $\mathcal{M} : (\mathbb{Z}_2^p \to \mathbb{Z}_2)^{|V|\times n\times n}$
- $\mathcal{M}^*$ can be solved for directly using Gaussian elimination or XOR-SAT
- Enables sketch-based synthesis in σ or $\mathcal{G}$: just use variables for holes!
- We can encode using the characteristic function, i.e., $\mathbb{1}_{\subseteq V} : V \to \mathbb{Z}_2^{|V|}$
- $\oplus, \otimes$ are defined as $\boxplus, \boxtimes$, so that the following diagram commutes:

$$\begin{tikzcd}
\langle \mathcal{G}', \Sigma^{n-1}\rangle & \textbf{Set} & \textbf{Bit} & \langle \mathcal{G}', \Sigma^{n-1}\rangle \arrow[dr] & \textbf{SAT} \\
\textbf{Rubix} & M \times M \arrow[ul, leftrightarrow] \arrow[r, leftrightarrow, "\mathbb{1}^{2^{n\times n}}", "\mathbb{1}^{-2^{n\times n}}"'] \arrow[d, "+"', "*"] & \mathbb{Z}_2^{|V|^{n\times n}} \times \mathbb{Z}_2^{|V|^{n\times n}} \arrow[rr, leftarrowtail, "\varphi^{2^{n\times n}}", "\varphi^{-2^{n\times n}}"'] \arrow[d, "+"', "*"] & & \mathcal{M} \times \mathcal{M} \arrow[d, leftrightarrow, "+"', "*"] \\
\textbf{Matrix} & 2^V \times 2^V \arrow[r, leftrightarrow, "\mathbb{1}^2", "\mathbb{1}^{-2}"'] \arrow[d, "\oplus"', "\otimes"] & \mathbb{Z}_2^{|V|} \times \mathbb{Z}_2^{|V|} \arrow[rr, leftarrowtail, "\varphi^2", "\varphi^{-2}"'] \arrow[d, "\boxplus"', "\boxtimes"] & & \mathcal{V} \times \mathcal{V} \arrow[d, leftrightarrow, "\boxplus"', "\boxtimes"] \\
\textbf{Vector} & 2^V \arrow[r, leftrightarrow, "\mathbb{1}", "\mathbb{1}^{-1}"'] & \mathbb{Z}_2^{|V|} \arrow[rr, leftarrowtail, "\varphi", "\varphi^{-1}"'] & & \mathcal{V}
\end{tikzcd}$$

Brozozowski's Derivative

Valiant's $\otimes$ operator, which unifies known factors in a binary CFG, implies a left- and right-quotient, which yield the set of nonterminal forests that may appear to either side of a known factor and its corresponding root.

Valiant's $\otimes$	Left Quotient	Right Quotient
$x \otimes y = \{\, w \mid (w \to xy) \,\}$	$\frac{\partial f}{\partial \overleftarrow{x}} = \{\, y \mid (w \to xy) \,\}$	$\frac{\partial f}{\partial \overrightarrow{y}} = \{\, x \mid (w \to xy) \,\}$
x w	x w	x w
y	y	y

The left quotient coincides with Brzozowski's derivative (1964) over regular languages, here lifted into the context-sensitive setting (our work).

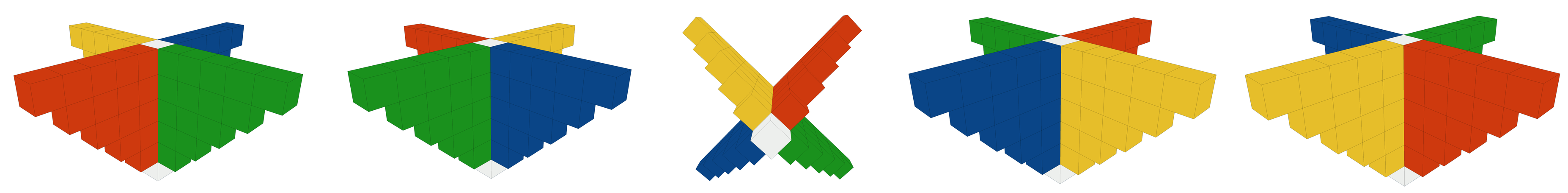
Context Sensitivity

It is well-known that the family of CFLs is not closed under intersection. For example, consider $\mathcal{L}_\cap := \mathcal{L}(\mathcal{G}_1) \cap \mathcal{L}(\mathcal{G}_1)$ defined in the following way:

$$P_1 := \{\, S \to LR,\ L \to ab \mid aLb,\ R \to c \mid cR \,\}$$
$$P_2 := \{\, S \to LR,\ R \to bc \mid bRc,\ L \to a \mid aL \,\}$$

$\mathcal{L}_\cap$ is equivalent to the language $\{\, a^d b^d c^d \mid d > 0 \,\}$, which is not a CFL. We can encode $\bigcap_{i=1}^{c} \mathcal{L}(\mathcal{G}_i)$ as a polygonal prism with upper-triangular matrices adjoined to each rectangular face. Specifically, we intersect all terminals $\Sigma_\cap := \bigcap_{i=1}^{c} \Sigma_i$, then for each $t \in \Sigma_\cap$, construct an equivalence class $E(t, \mathcal{G}_i) = \{w_i \mid (w_i \to t) \in P_i\}$ and glue them together at each σ_i:

$$\bigwedge_{t \in \Sigma_\cap} \bigwedge_{j=1}^{c-1} \bigwedge_{i=1}^{|\sigma|} E(t, \mathcal{G}_j) \equiv_{\sigma_i} E(t, \mathcal{G}_{j+1})$$



Orientations of a $\bigcap_{i=1}^{4} \mathcal{L}(\mathcal{G}_i) \cap \Sigma^6$ configuration, reprojected into 2-space.

As $c \to \infty$, this shape approximates a circular cone whose symmetric axis intersects orthogonal CNF unit productions $w_i \to t$, with $S_i \in \mathcal{T}_i$ represented by bitvectors on the base perimeter. Equations of this form are equiexpressive with the family of CSLs realizable by finite CFL intersection.

Fpt Software

McGill UNIVERSITY

Mila