Stack Attention: Improving the Ability of Transformers to Model Hierarchical Patterns

Brian DuSell and David Chiang

UNIVERSITY OF NOTRE DAME

ETH zürich

An Attention Mechanism for Recursive Syntax

Standard attention doesn't have a good way of dealing with recursion. Two examples:

- **Theoretical:** Can't model balanced brackets (under certain assumptions) (Hahn, 2020)
  [ ] (( ) (([(((( ))))])))
- **Empirical:** Brittle on center embedding (Lakretz et al., 2022)
  The **keys** that the **man** near the cabinet **holds are** ...

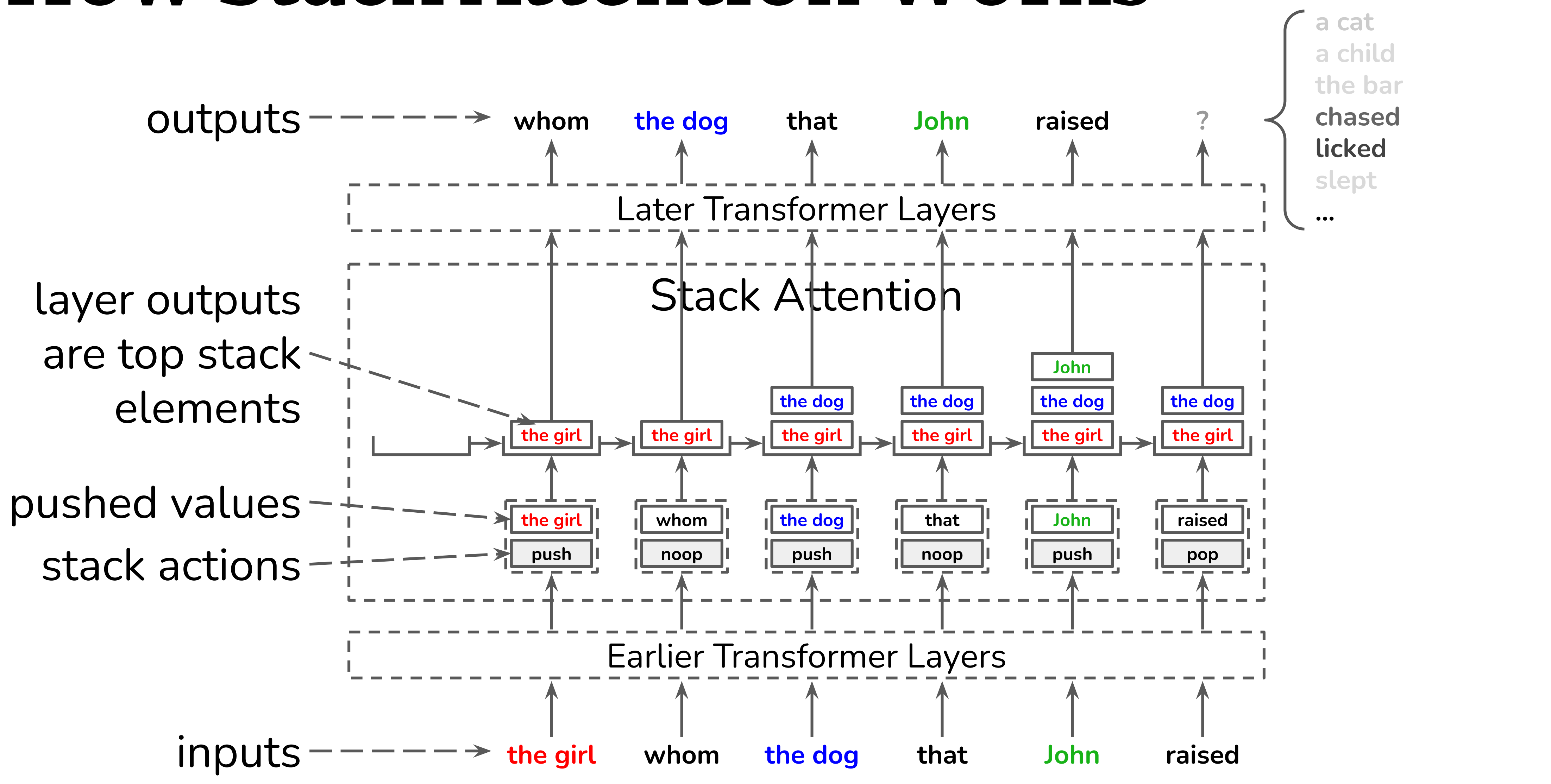
Our solution: Syntax is deeply connected to **stacks**, so we propose a new self-attention mechanism based on differentiable stacks called **stack attention**.

Features

1. **Differentiable end-to-end** with standard backprop; no changes to training algorithm required
2. Syntactically **unsupervised**; no parse trees required in training data
3. **Generative**; no future context required, works with standard decoding algorithms

No prior work satisfies 2 and 3 at the same time. Stack attention can be used as a **drop-in replacement** for standard attention.

How Stack Attention Works



Stack Attention = **Differentiable Stack** = attention over **partial syntax trees**

Two Flavors of Stack Attention

Superposition (Sup)

- *Superposition* of three stack actions (push, noop, pop)
- **Faster**
- Less expressive
- **Special case** of nondeterministic

Nondeterministic (Nd)

- Based on *nondeterministic* pushdown automata (PDAs)
- Recognizes **all context-free languages**
- Slower

Serial Time Complexity

Attention	Serial Time
SDPA	$O(n^2)$
Superposition	$O(n^2)$
Nondeterministic	$O(n^3)$

Parallel Time Complexity

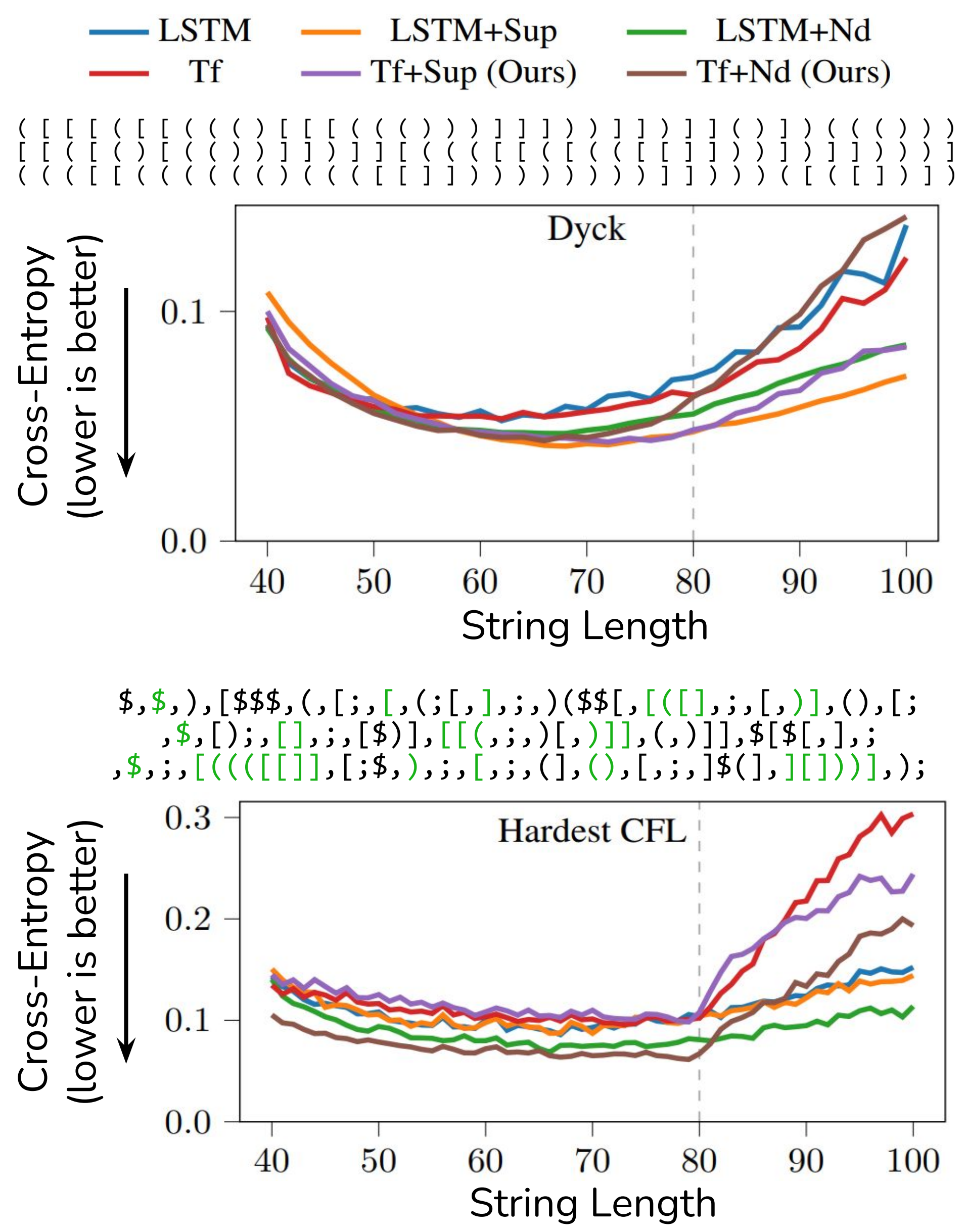
Attention	Implemented	Parallel CKY	Theoretical
SDPA	$O(\log n)$	–	–
Superposition	$O(n)$	–	$O((\log n)^2)$
Nondeterministic	$O(n^2)$	$O(n \log n)$	$O((\log n)^2)$

Wall-Clock Runtime on Natural Language Modeling

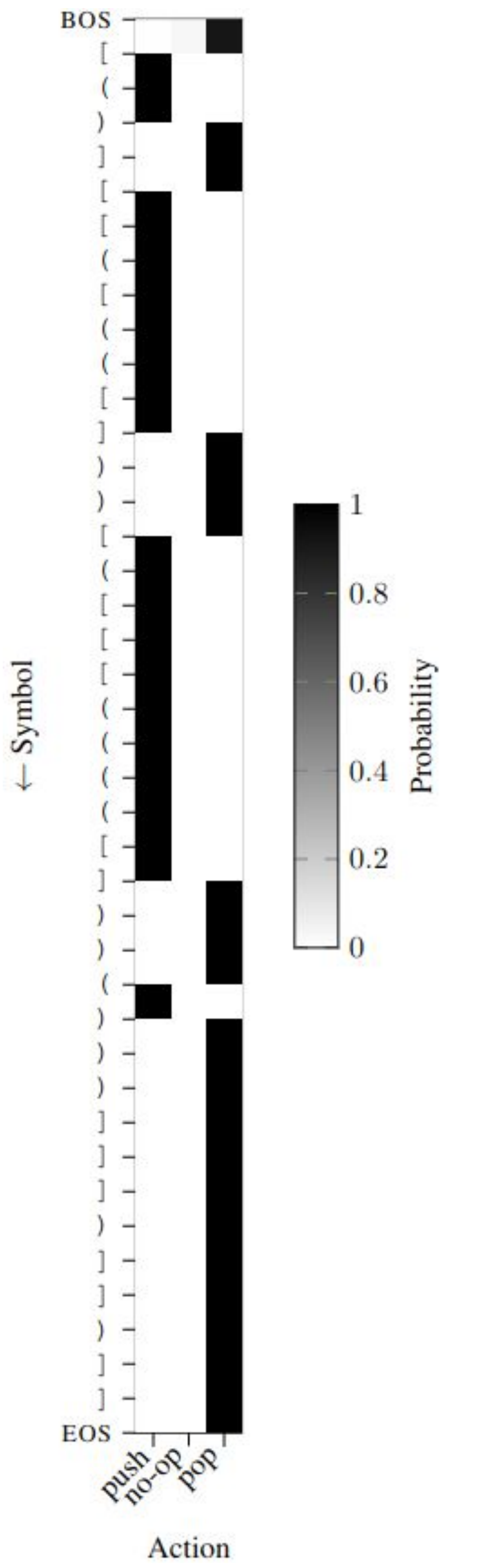
Model	Examples/s	Minutes/Epoch	GPU Memory
Tf	859	0.8	394 MB
Tf+Sup	345	1.9	397 MB
Tf+Nd	27	24.3	1.91 GB

Results

Context-Free Language Modeling



Learned Stack Actions for Balancing Brackets



Natural Language Modeling on Penn Treebank (Perplexity)

Model	Params.	Val. ↓	Test ↓
Tf	10,051,072	115.11	92.84
Tf+Sup (Ours)	10,050,304	122.94	98.67
Tf+Nd (Ours)	9,861,898	**110.59**	**88.54**

Future Work

- Runtime improvements, parallelization across timestep dimension
- Interpretability of learned syntactic structure
- Benchmarking for **data efficiency** (e.g., BabyLM) and **hierarchical inductive bias** (e.g., McCoy et al., 2020)