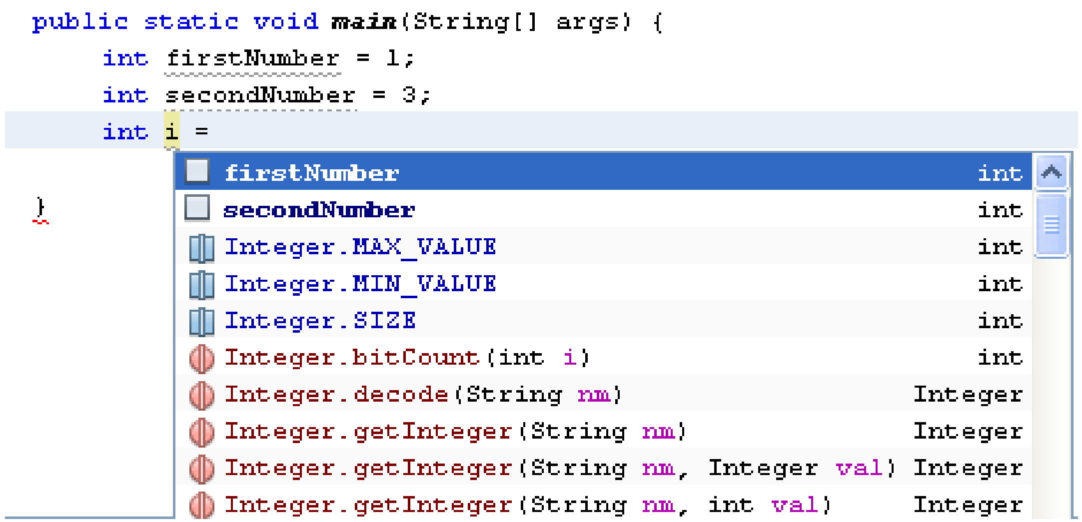
Code Completion with Neural Attention and Pointer Networks

Jian Li, Yue Wang, Michael R. Lyu, Irwin King
{jianli, yuewang, lyu, king}@cse.cuhk.edu.hk
IJCAI 2018, Stockholm, Sweden

香港中文大學
The Chinese University of Hong Kong

Code Completion

- An example:



- Static programming languages: compile-time type information
- Dynamic programming languages: learning-based language models

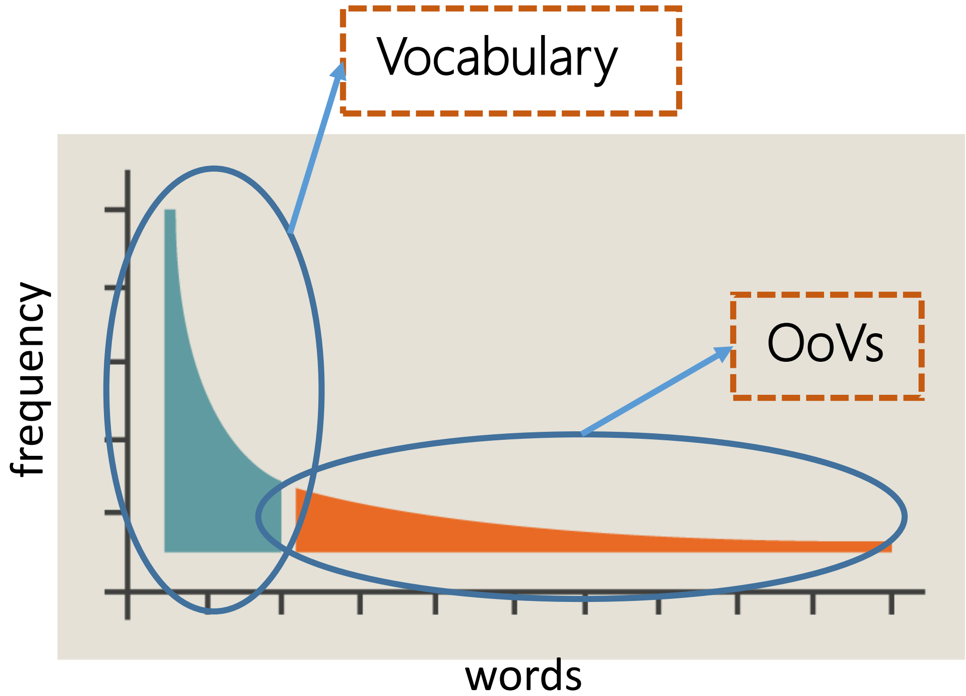
Code Completion with Language Models

- Simplified problem: given a sequence of code tokens, our task is to predict the next one token.
- Method: adapt neural language models for code completion.

Challenges

1. Long-range dependencies

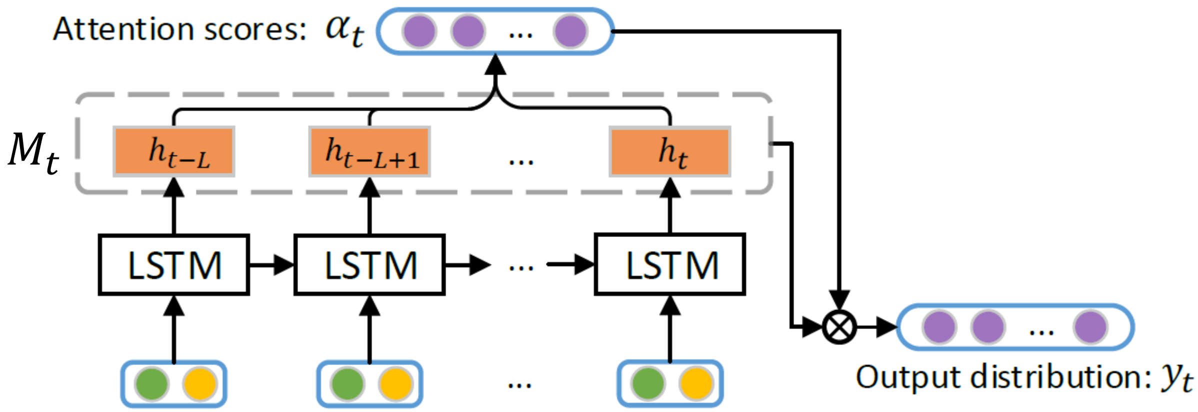
2. Out-of-Vocabulary (OoV) words



OoVs cannot be correctly predicted!

Attention Mechanisms

- Deal with long-range dependencies:
  - Context attention



$$
\begin{aligned}
A_t &= v^T \tanh(W^m M_t + (W^h h_t) 1_L^T) \\
\alpha_t &= softmax(A_t) \\
c_t &= M_t \alpha_t^T \\
&\boxed{G_t = \tanh(W^g [h_t; c_t])} \\
y_t &= softmax(W^v G_t + b^v)
\end{aligned}
$$

Attention Mechanisms

- Deal with long-range dependencies:
  - ➢ Abstract Syntax Tree (AST)

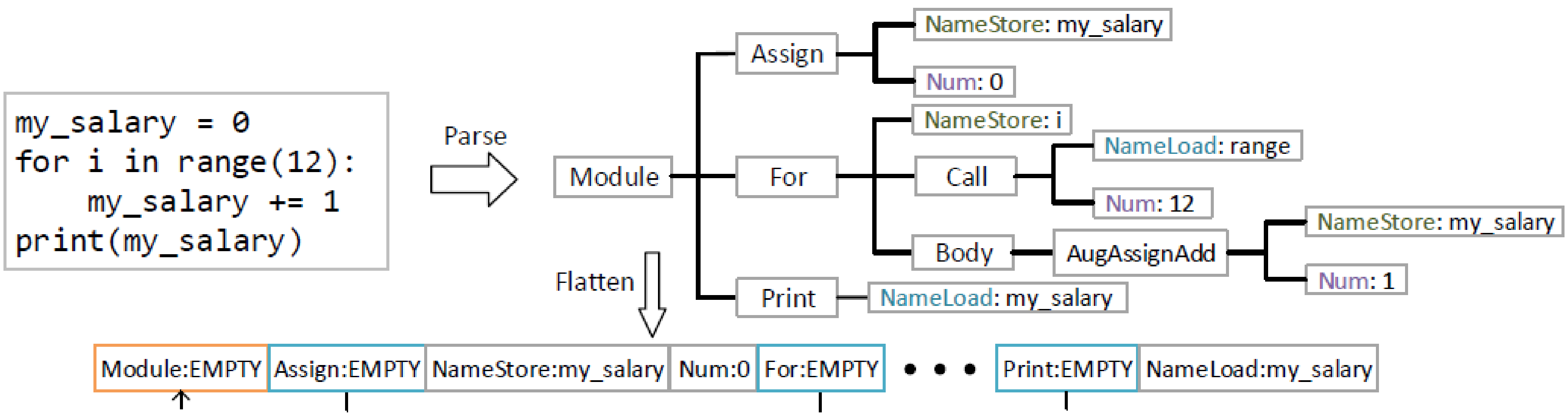
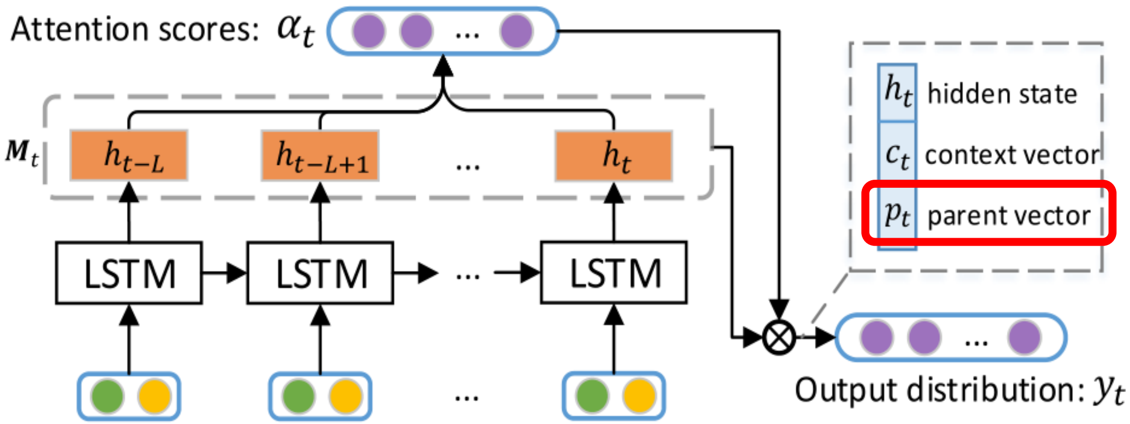


Figure 1: A Python program and its corresponding abstract syntax tree (AST). The dashed arrow points to a parent node.

Exploit the parent-children information on program's AST!

Attention Mechanisms

- Deal with long-range dependencies:
  - Parent attention



$$
\begin{aligned}
A_t &= v^T \tanh(W^m M_t + (W^h h_t) 1_L^T) \\
\alpha_t &= softmax(A_t) \\
c_t &= M_t \alpha_t^T \\
G_t &= \tanh(W^g [h_t; c_t; p_t]) \\
y_t &= softmax(W^v G_t + b^v)
\end{aligned}
$$

p_t is the parent vector storing the hidden state of the parent node.

Pointer Mixture Network

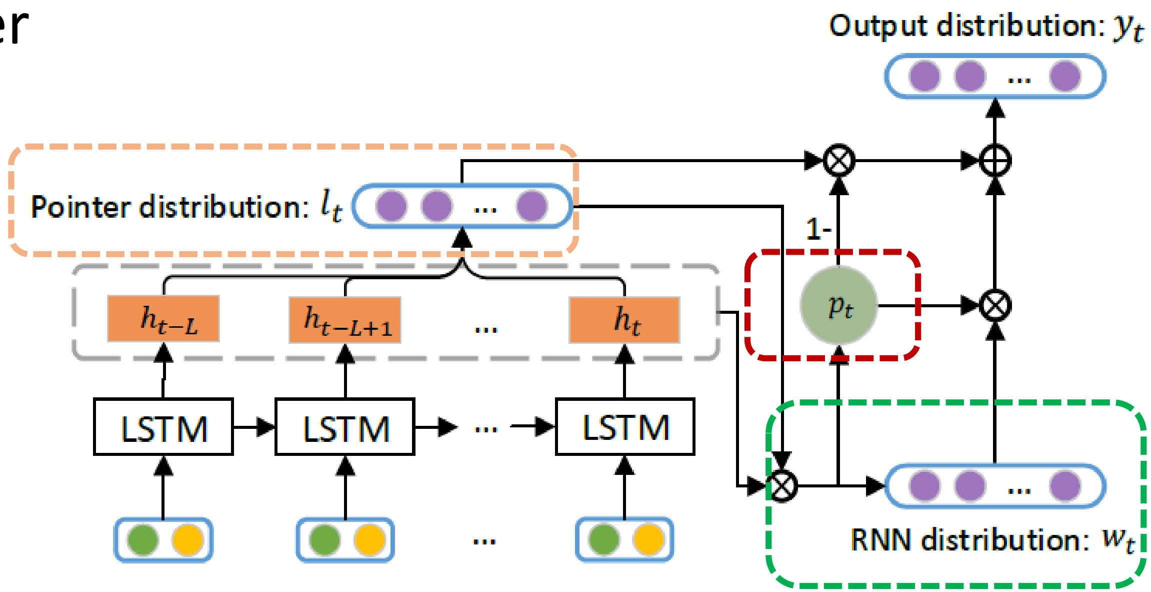
- Deal with OoV words:
  - Locally repeated terms are prevalent.

```
my_salary = 0
for i in range(12):
    my_salary += 1
print(my_salary)
```

  - Copy from local context to predict OoVs.
  - Learn *when and where* to copy.

Pointer Mixture Network

- Deal with OoV words:
  - Global RNN component
  - Local pointer component
    - Reuse the attention scores as the pointer distribution
  - Controller



$$p_t = \sigma(W^p[h_t; c_t] + b^p)$$
$$y_t = softmax([p_t w_t; (1 - p_t) l_t])$$

Experiments

- Dataset
  - JavaScript (JS) and Python (PY)
  - Query: remove the AST node (plus all the nodes to the right) from the node sequence and then attempt to predict the node.

Table 1: Dataset Statistics

	JS	PY
Training Queries	$10.7 * 10^7$	$6.2 * 10^7$
Test Queries	$5.3 * 10^7$	$3.0 * 10^7$
Type Vocabulary	95	329
Value Vocabulary	$2.6 * 10^6$	$3.4 * 10^6$

OoV problem!

Experiments

- Accuracies on next value prediction with different vocabulary sizes

Vocabulary Size OoV Rate / Localness	JS_1k 20% / 8%	JS_10k 11% / 3.7%	JS_50k 7% / 2%	PY_1k 24% / 9.3%	PY_10k 16% / 5.2%	PY_50k 11% / 3.2%
Vanilla LSTM	69.9%	75.8%	78.6%	63.6%	66.3%	67.3%
Attentional LSTM (ours)	71.7%	78.1%	80.6%	64.9%	68.4%	69.8%
Pointer Mixture Network (ours)	**73.2%**	**78.9%**	**81.0%**	**66.4%**	**68.9%**	**70.1%**

- OoV Rate denotes the percentage of AST nodes whose value is beyond the global vocabulary.
- Localness is the percentage of values who are OoV but occur in the context window, which is the upper-bound of the performance gain.

Vocabulary size ↑, OoV rate ↓, accuracy ↑, performance gain ↓

Experiments

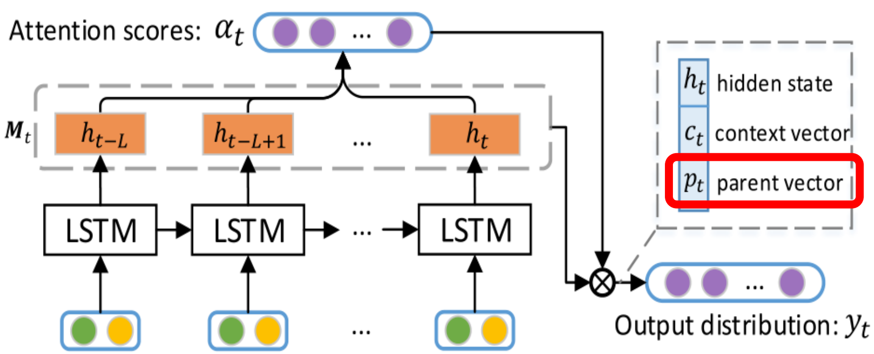
- Comparisons against the state-of-the-arts
  - Pointer Mixture Network only for predicting VALUE node

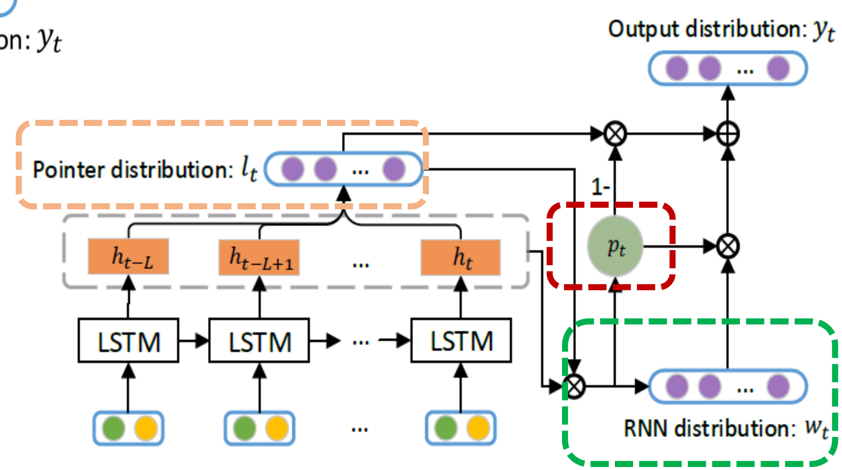
	JS		PY	
	TYPE	VALUE	TYPE	VALUE
Vanilla LSTM	87.1%	78.6%	79.3%	67.3%
Attentional LSTM (no parent attention)	88.1%	80.5%	80.2%	69.8%
Attentional LSTM (ours)	**88.6%**	80.6%	**80.6%**	69.8%
Pointer Mixture Network (ours)	-	81.0%	-	**70.1%**
LSTM [Liu *et al.*, 2016]	84.8%	76.6%	-	-
Probabilistic Model [Raychev *et al.*, 2016]	83.9%	**82.9%**	76.3%	69.2%

- Our model outperforms the state-of-the-art in almost all cases
- The proposed parent attention is also effective

Conclusion

1. Propose a parent attention mechanism for AST-based code completion.
2. Propose a pointer mixture network which learns to either generate a new value or copy an OoV value from local context.
3. Demonstrate the effectiveness of our model via experiments.






Q
A