Summarizing Source Code using a Neural Attention Model

University of Washington, Seattle, USA

Overview
Our task is to generate high-level summaries of the function of programming source code.

C#
```csharp
var x = "FundList[10].Amount";
int xIndex = Convert.ToInt32(Regex.Match(x, @"\d+"), Value);
```
Identify the number in given string

SQL
```sql
SELECT * FROM table
ORDER BY RANDOM() LIMIT 10;
```
Select random rows from my sql table

These summaries have many SE applications:
- Code Search
- Documentation
- Code Navigation

Neural Attention Model
We use an end-to-end model that jointly performs content selection using an attention mechanism, and surface realization using Long Short Term Memory networks.

We model the conditional next-word probability as:

\[
p(n_i | n_1, \ldots, n_{i-1}) \propto W \tanh(W_1 h_i + W_2 t_i)
\]

\(h_i\) is the hidden state of the LSTM cell at the time step \(i\).

The attention model computes a weighted sum \(t_i\) of the token embeddings of code \(C\) based on the LSTM hidden state. In this way, it selects the most useful tokens to generate the current word.

Future Work
1. Generation models using tree to sequence methods
2. Discovery and explanation of code idioms
3. Using language-code models for code synthesis

Code Summarization Dataset
We create a new dataset from programming QA websites containing 66K C# and 33K SQL examples.

Example Outputs
- C# How to convert string to int?
- C# How to get all child nodes in TreeView?
- C# How to call a URL from a web api post?
- SQL How to get random rows from a mysql database?
- SQL How to get the sum of a column in a single query?
- SQL How to call a URL from a web api post?