MetaTPTrans: A Meta Learning Approach for Multilingual Code Representation Learning

Weiguo Pian, Hanyu Peng, Xunzhu Tang, Tiezhu Sun, Haoye Tian, Andrew Habib, Jacques Klein, Tegawendé F. Bissyandé

Language-agnostic vs. Language-specific Information
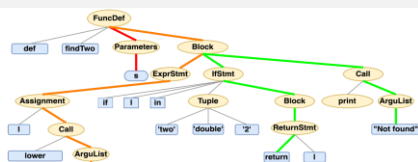
(a) Python code snippet and its AST
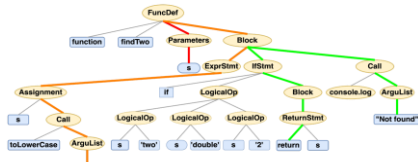
(b) JavaScript code snippet and its AST

- Language-agnostic information: The information can be directly extracted from the source code or AST by a unified model across different languages, without relying on any language-specific feature.

- Language-specific information: Language-specific underlying syntactic rules, the language-specific API names.

Motivation

- Multilingual training improves the performance of language models compared to training the models on single-language datasets separately.

- Existing multilingual training strategies overlooks the language-specific information

- We want to train a model that can extract both language-agnostic and language-specific information.

Approach

- Meta Learning From in Our Approach: Learning to generate parameters.
  - Mata Learner: Generate parameters for the Base Learner based on the language type (e.g. Python, Go).
  - Base Learner: Extract features from source code (TPTrans [1]).
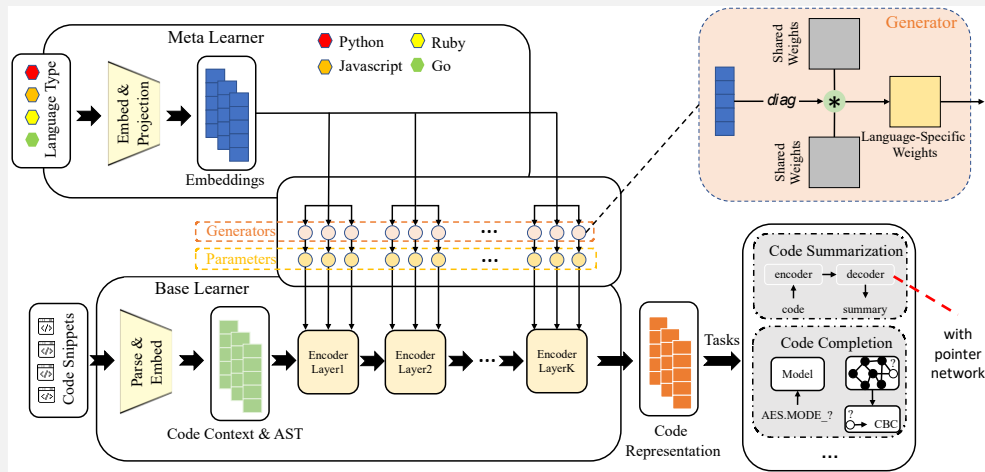
- Variants:
  - MetaTPTrans-α: Generate parameters for context token projection.
  - MetaTPTrans-β: Generate parameters for path encoding projection.
  - MetaTPTrans-γ: Generate parameters for both context token projection and path encoding projection.

Experimental Setup

- Dataset: CodeSearchNet [2].

- Tasks:
  - Code Summarization
  - Code Completion

Language	Samples per partition		
	Train	Valid	Test
Python	412,178	23,107	22,176
Ruby	48,791	2,209	2,279
JavaScript	123,889	8,253	6,483
Go	317,832	14,242	14,291
Total	902,690	47,811	45,229

[1] Integrating Tree Path in Transformer for Code Representation. In *NeurIPS 2021*.
[2] CodeSearchNet Challenge: Evaluating the State of Semantic Code Search. *arXiv:1909.09436, 2019*.

Experimental Results

Experimental Results of the Code Summarization task

Model	Python			Ruby			JavaScript			Go		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
code2seq (Single-language)	35.79	24.85	29.34	23.23	10.31	14.28	30.18	19.88	23.97	52.30	43.43	47.45
GREAT (Single-language)	35.07	31.59	33.24	24.64	22.23	23.38	31.20	26.84	28.86	50.01	46.51	48.20
CodeTransformer (Single.)	36.40	33.66	34.97	31.42	24.46	27.50	35.06	29.61	32.11	55.10	48.05	51.34
TPTrans (Single-language)	38.39	34.70	36.45	33.07	28.34	30.52	33.68	28.95	31.14	55.67	51.31	53.39
code2seq (Multilingual)	34.49	25.49	29.32	23.97	17.06	19.93	31.62	22.16	26.06	52.70	44.36	48.17
GREAT (Multilingual)	36.75	31.54	33.94	30.05	24.33	26.89	33.58	27.78	30.41	52.65	48.30	50.38
CodeTransformer (Multi.)	38.89	33.82	36.18	33.93	28.94	31.24	36.95	29.98	33.10	56.00	50.44	53.07
TPTrans (Multilingual)	39.71	34.66	37.01	33.07	28.34	35.55	34.92	30.01	32.33	56.48	52.02	54.16
MetaTPTrans-α	40.22	36.22	38.12	40.62	34.01	37.02	37.87	31.92	34.64	58.12	53.82	55.89
MetaTPTrans-β	39.97	36.12	37.94	40.44	33.69	36.76	38.87	32.66	35.50	58.86	54.24	56.45
MetaTPTrans-γ	40.47	35.19	37.65	40.58	32.04	35.81	37.90	30.11	33.56	58.20	53.38	55.68

Ablation Study of Code Summarization w/o Pointer Network

Model	Python			Ruby			JavaScript			Go		
	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
CodeTransformer (Multi.)	38.91	33.12	35.78	34.52	27.31	30.50	37.21	29.75	33.07	56.07	50.76	53.28
TPTrans (Multilingual)	38.78	34.72	36.64	38.05	32.35	34.97	36.35	30.06	32.90	56.49	51.99	54.15
MetaTPTrans-α	39.26	36.57	37.87	39.22	34.55	36.74	37.29	32.50	34.73	57.14	54.48	55.78
MetaTPTrans-β	38.34	37.32	37.82	38.87	36.07	37.42	37.35	34.06	35.63	56.56	55.14	55.84
MetaTPTrans-γ	38.50	36.96	37.71	38.38	33.99	36.05	37.72	32.62	34.98	56.49	54.30	55.38

Experimental Results of the Code Completion task

Model	Python		Ruby		JavaScript		Go	
	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
Transformer (Single-language)	47.57	69.86	44.39	62.24	37.57	53.15	40.21	59.65
CodeTransformer (Single.)	62.45	76.73	51.63	69.96	47.56	68.88	47.71	61.35
TPTrans (Single-language)	63.71	77.99	64.42	72.50	64.67	73.42	57.15	67.81
Transformer (Multilingual)	47.02	78.82	47.16	77.32	38.77	70.84	42.01	72.95
CodeTransformer (Multi.)	68.19	82.98	67.67	83.47	59.32	80.07	57.12	77.67
TPTrans (Multilingual)	69.81	84.10	72.14	82.27	67.45	81.17	60.45	79.03
MetaTPTrans-α	77.13	94.28	78.05	95.42	73.52	92.88	67.47	91.15
MetaTPTrans-β	71.75	86.26	73.82	86.85	72.71	86.90	66.74	85.21
MetaTPTrans-γ	67.12	90.72	71.89	93.61	69.55	90.97	61.60	88.99

Visualization of the Representation Learned by MetaTPTrans-α (left) vs. TPTrans (right)
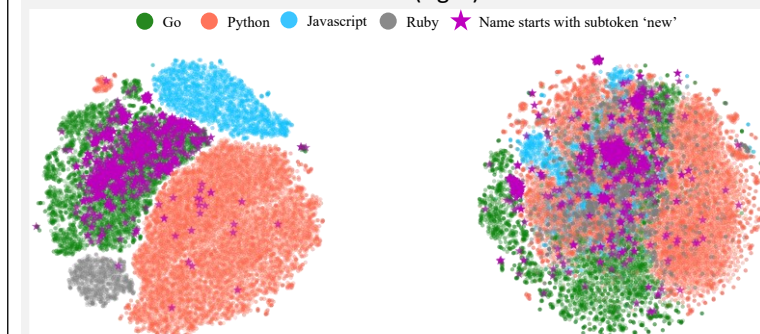
More language-specific information and more semantic information learned by our approach.