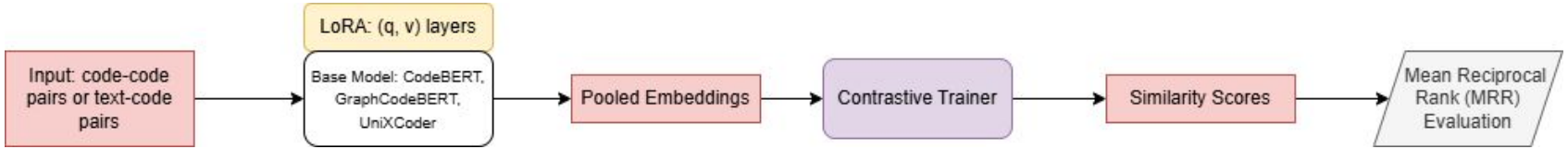
LoRACode: LoRA Adapters for Code Embeddings

Saumya Chaturvedi, Aman Chadha, Laurent Bindschaedler

ICLR
International Conference On Learning Representations

LoRACode introduces:

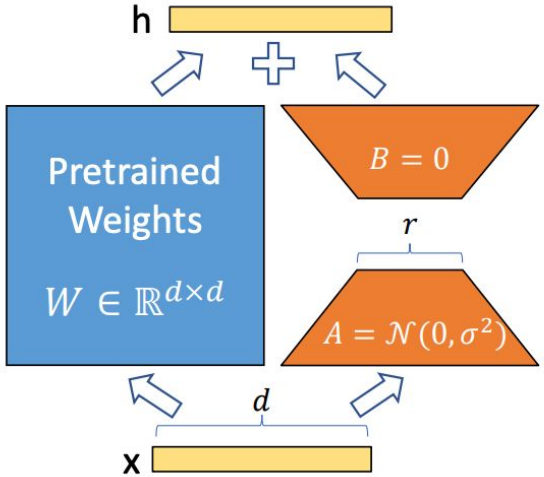
- PEFT approach to code search using **low rank adapters**
- Parameter Efficient Finetuning using only **1.83%-1.85%** of parameters
- Up to **86.69%** increase in **Mean Reciprocal Rank** for Text2Code and **9.1%** increase in MRR for Code2Code search tasks.



LoRACode Architecture

Related Works

- Existing Models (CodeBERT, GraphCodeBERT, UniXcoder)
- Limitations:
  - Limited scalability and ability to generalize over tasks
  - Lower Mean Reciprocal Ranks for code search
  - Struggle to separate language semantic from syntactic components
- LoRA adapters and the myriads of applications:
  - Jina AI's jina-embeddings-v3 utilizes low rank adapters for multilingual text search.
- Contrastive fine tuning



LoRA Attention

Design of LoRACode

- Source-included multilingual code search for text-to-code and code-to-code queries.
- LoRA-based adapters on CodeBERT, UniXcoder
- Contrastive Training & Similarity Scores
- MRR-based evaluation

Hyperparameters

Table 1: Training hyperparameters

Hyperparameter	**Value**
Batch Size	16 per device
Epochs	1 (rapid evaluation)
Learning Rate Scheduler	1000 warmup steps
Logging	Every 200 steps
Save Strategy	End of each epoch
Evaluation Strategy	No intermediate eval

Table 2: LoRA hyperparameters

Hyperparameter	**Value**
Ranks	16, 32, 64
LoRA Alpha	32, 64, 128
Target Modules	Query and Value
Dropout	10%

Experiments

- Training Setup: 2 H100 GPUs, 25 min fine-tuning
- CodeSearchNet dataset: 2 million code samples

Text2Code

- Initial taskwise adapters using CSN dataset
- Training different adapters for different programming languages

Code2Code

- Taskwise distinction across all languages

Results: MRR for Text2code

	Ruby	Go	PHP	Python	Java	Javascript
LoRACode - Combined (rank 64)	42.83	48.34	20.88	28.60	33.08	30.55
LoRACode - Combined (rank 32)	43.96	48.98	21.86	29.85	34.15	31.38
UnixCoder	44.06	49.59	22.31	29.76	34.47	32.05
GraphCodeBERT	20.80	12.48	8.08	10.38	8.60	7.30
CodeBERT	0.37	0.15	0.03	0.06	0.04	0.06
Starencoder	4.41	1.85	0.57	2.14	1.89	1.55

Initial LoRA configuration: task-wise adapters
minimal improvement

Solution: Language-Wise Adapters!

- Using CosQA dataset (Python only) showed **14.8%** increase in MRR!
  - dataset size
  - human annotation
  - different programming languages captured different context
- Training adapters separately per language
- Factors contributing to performance difference:
  - data duplicacy on merging datasets
  - difference in dataset size across languages
  - syntactical nuances across languages

Increase in MRR and NDCG for search on CosQA

	UniXCoder	**LoRACode - Combined (r=64)**
MRR	31.36	**36.02**
NDCG@10	35.64	**40.44**

Results: Language-specific adapter

Languages	MRR		NDCG		# Samples	Train Time	% Increase
	Base	LoRA	Base	LoRA			
Ruby	44.06	**45.78**	47.95	**49.77**	1558	7:01	3.90
Go	49.59	**82.88**	53.66	**85.35**	10456	47:36	67.13
PHP	35.22	**52.46**	24.73	**56.54**	15078	1:08:39	48.94
Python	29.76	**55.56**	32.83	**59.49**	15739	2:10:39	86.69
Java	34.47	**53.47**	37.94	**57.45**	10308	1:25:19	31.91
Javascript	32.05	**38.75**	35.04	**42.35**	3627	16:41	20.9

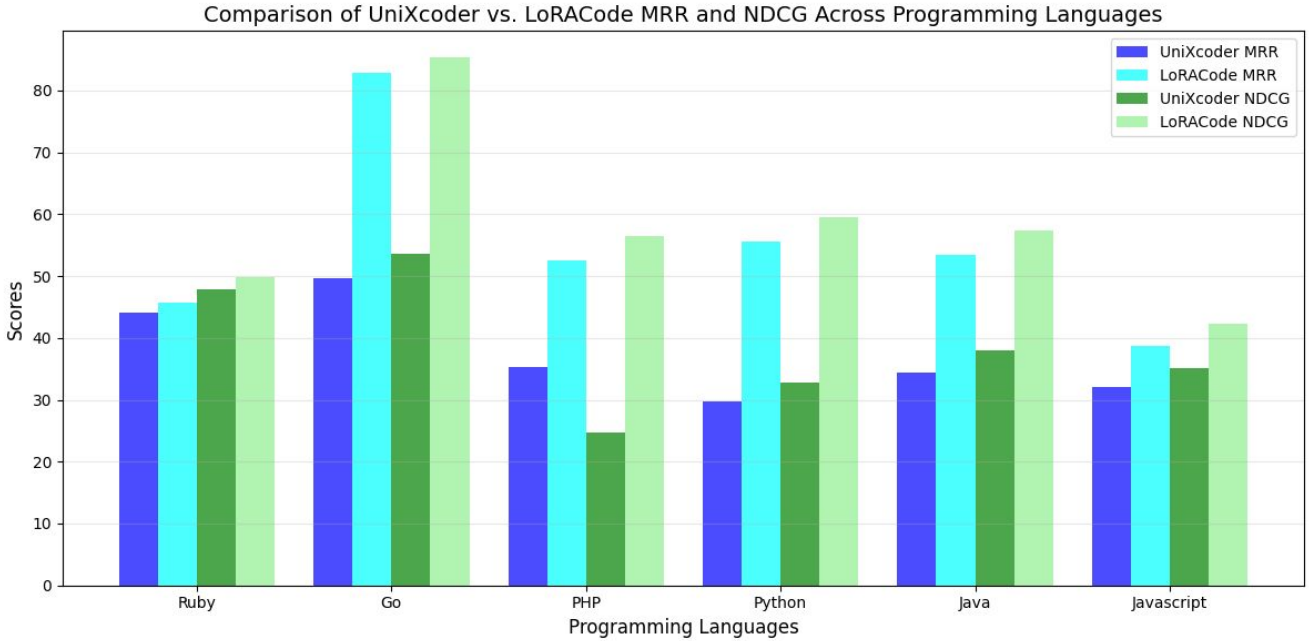
Language-wise adapters: massive improvement!

Results: MRR for Code2code

	C	PHP	Java	C++	C#	Javascript	Python
LoRACode - Combined (rank 64)	**41.07**	**44.18**	**48.84**	**48.91**	**48.69**	**48.56**	**48.27**
LoRACode - Combined (rank 32)	40.72	43.53	47.75	47.95	47.81	47.70	47.50
UnixCoder	37.64	42.56	45.84	45.51	46.01	46.50	46.68
GraphCodeBERT	32.81	37.93	30.86	34.04	31.74	39.53	20.30
CodeBERT	27.45	30.47	24.80	10.54	25.53	25.56	5.48
Starencoder	17.48	39.78	35.59	39.50	35.31	40.41	25.93

MRR scores across languages on combined code2code fine-tuning

Conclusion



LoRACode enhances code embeddings for both Code2code and Text2code tasks.

Parallels across languages and the impact on different datasets is yet to be explored and left as future work.

Thank you!