Project Information

Project ID: 210310187

Project Name: Auto tuner for vector indexing parameters

Time Planning: Basic BOHB optimization in mid term.

Introduction

For solving the optimization of milvus hyperparameters, we use the Bayesian Optimization and Hyperband(BOHB)1 as our parameter search method.

Implementation

There are several level hyperparameters in milvus, including index_type, index_params and search_params. To get an end-to-end solution for index, we use BOHB in different level. For Index Type, to eminent the randomness of BO for index_type(which means BO may not fully explore some specific type due to init poor performance), we set two index type optimization mode(Loop and BO). ## Loss Function We use laplace method to conver the constraint BO to unconstraint version. Our loss function is as below:

$$Loss = sign(recall, threshold) - query\_per\_sec$$

$$Sign(recall, threshold) = \begin{cases} recall - threshold & recall > threshold \\ 100000 * (threshold - x) & recall <= threshold, \end{cases}$$

100000 is just a large number for Lagrange method, **threshold is set to 95**.

Method

Hardwareware Information

CPU: Intel Core i7-8700 CPU @ 4.6GHz
RAM: 2182MiB / 32083MiB

Index Type Optimization

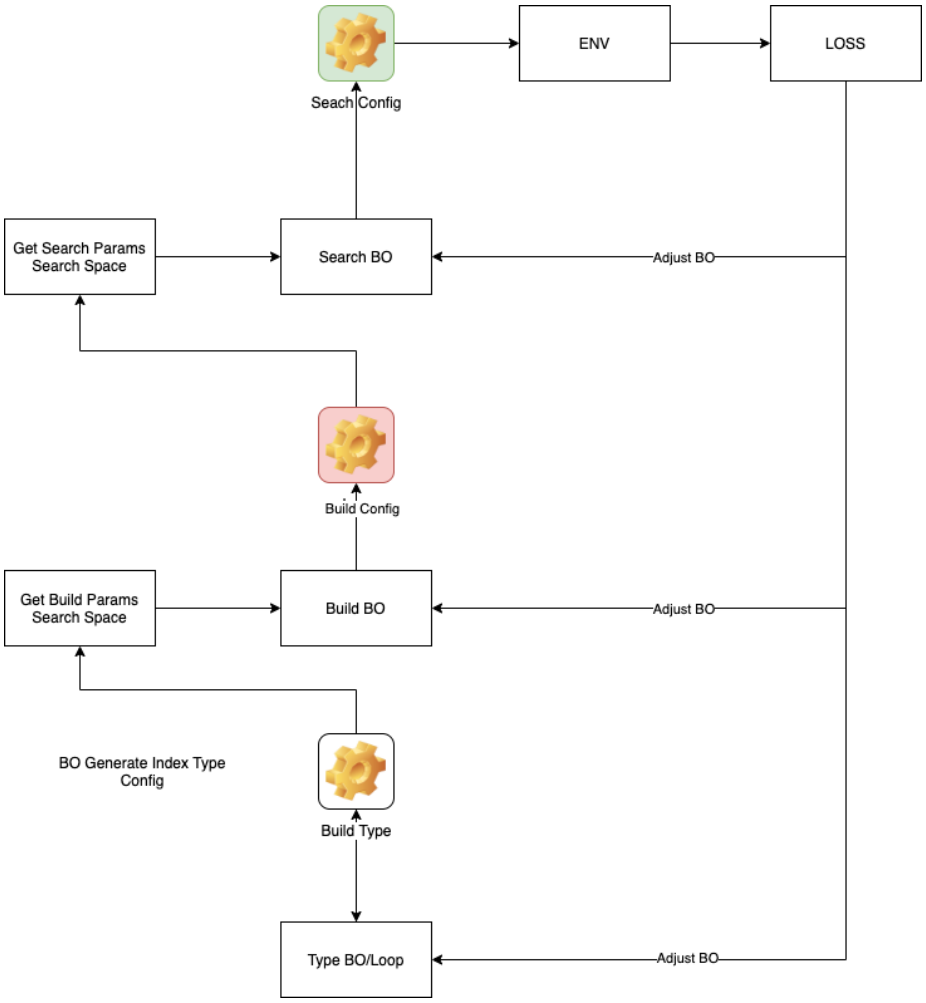

Figure 1: Model Architecture

Method	index_type	M	efConstruction	ef	recall	query_per_sec	loss
BOHB(Index Type Loop)	'HNSW'	17	445	114	99.68	16782.6	-16777.9
BOHB(Index Type BO)	'HNSW'	22	274	106	99.75	18522	-18517.2

	index_type	nlist	M	nprobe	recall	query_per_sec	loss
Grid Search	'HNSW'	4	158	200	97.11	18331.748252	-18329.638252

Index Parameters Optimization

IVF_FLAT

	index_type	nlist	nprobe	recall	query_per_sec	loss
BOHB	2883	54	99.68	14911	-14906.3	
Grid Search	'IVF_FLAT'	14601	101	100.0	14402.032758	-14397.032758

IVF_SQ8

	index_type	nlist	nprobe	recall	query_per_sec	loss
BOHB	'IVF_SQ8'	8405	46	98.86	13827.5	-13823.7
Grid Search	'IVF_SQ8'	5401	101	99.49	13080.62997	-13076.13997

IVF_PQ

	index_type	nlist	nprobe	recall	query_per_sec	loss
BOHB	'IVF_PQ'	3800	205	98.1	1289.0043055892756	1285.9043055892757
Grid Search (note: Loss is not correct in the wandb table)	'IVF_IVF_PQ'	1	1	95.08	1733.677784	7629.256438

HNSW

Method	index_type	M	efConstruction	ef	recall	query_per_sec	loss
BOHB	'HNSW'	18	92	157	99.85	17868.6	-17863.8
Grid Search	'HNSW'	4	158	200	97.11	18331.748252	-18329.638252

TODO:

- Add time Measure to current BO method and progress bar.
- Try to solve the cold-start problem using the feature and best index choice prior.

References

- BOHB