BBT-RGB

.::. LREC-COLING 2024

Qiushi Sun
qiushisun@u.nus.edu

April 29, 2024

Introduction
- LREC-COLING 2024

Paper: Make Prompt-based Black-Box Tuning Colorful: Boosting Model Generalization from Three Orthogonal Perspectives (BBT-RGB)

Authors: Qiushi Sun, Chengcheng Han, Nuo Chen, Renyu Zhu, Jingyang Gong, Xiang Li, Ming Gao

Overview

Black-Box Tuning
- Background Knowledge

- Language models are becoming larger

Black-Box Tuning

- Background Knowledge

- Language models are becoming larger
- It is prohibitively expensive to fine-tune the entire model for each task

Black-Box Tuning
- Background Knowledge

- Language models are becoming larger
- It is prohibitively expensive to fine-tune the entire model for each task
- Prompt tuning is good, but we still need BP through the entire model

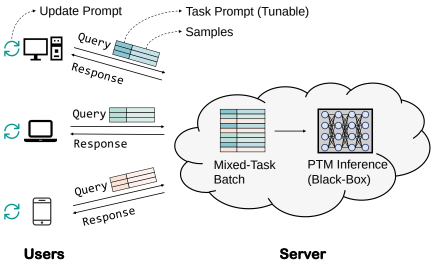
Black-Box Tuning
- Background Knowledge

- Language models are becoming larger
- It is prohibitively expensive to fine-tune the entire model for each task
- Prompt tuning is good, but we still need BP through the entire model
- Optimize prompts without BP?

Black-Box Tuning
- Background Knowledge

- Users usually do not have enough computing resources to run LLMs
- Providers often do not open-source model weights due to commercial reasons

Fig: Query the LMs deployed on the server.



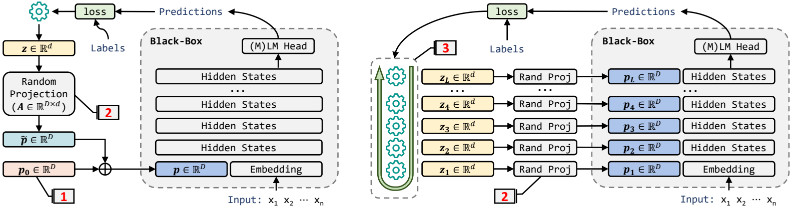
Is it possible to **optimize the prompt without BP**? → Black-Box Tuning![1]

[1] Black-box tuning for language-model-as-a-service. ICML 2022

Black-Box Tuning

- Previous Works: BBT & BBTv2[2]

Using an LLM as the backbone, completing downstream classification tasks under black-box settings by prompt-learning. The process will not use model gradient or parameters for backpropagation, only utilizing the model output.



(a) BBT (b) BBTv2

Fig: The architecture of previous works: BBT and BBTv2

[2] BBTv2: Towards a Gradient-Free Future with Large Language Models. EMNLP 2022

Black-Box Tuning

- Areas for improvement

- Lack of flexibility in using a single DFO algorithm.

Black-Box Tuning

- Areas for improvement

- Lack of flexibility in using a single DFO algorithm.
- Simple label words does not fully utilize the information returned by the black box.

Black-Box Tuning

- Areas for improvement

- Lack of flexibility in using a single DFO algorithm.
- Simple label words does not fully utilize the information returned by the black box.
- Unstable initialization in few-shot scenarios.

Black-Box Tuning

- Make Prompt-based BBT Colorful: Boosting Model Generalization from Three Orth. Perspectives

We optimize Black-Box tuning from three aspects.

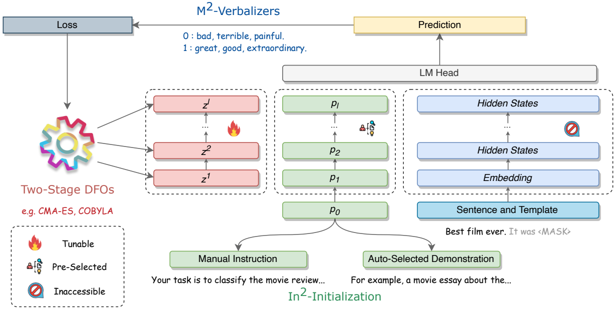


Fig: The proposed method: BBT-RGB

Plug-and-play, Server/User-friendly and Effective.

BBT-RGB
- Two-Stage DFOs

The previously used evolutionary algorithm has a much higher convergence rate than the search-based algorithm, which might cause fast overfitting.

- Combining different DFOs for continuous prompt optimization.
- Leveraging the advantages of two different kinds of DFOs.
- For the mitigation of the overfitting problem.

Algorithm 1 Two-Stage DFOs

Input: popsize:λ, intrinsic dimension:d
Input: budget1:$b1$, budget2:$b2$, backbone:f_{model}
Output: hidden variable: z

```
function Two-Stage DFO
    repeat
        for each hidden layer do
            Update z by Evolutionary DFO
        end for
    until b1 times f_model call
    for each hidden layer do
        repeat
            Update z by Search-based DFO
        until b2//d times f_model call
    end for
end function
```

BBT-RGB

- Two-Stage DFOs (Case Study)

- CMA-ES: Evolutionary strategy, uses population of solutions to explore parameter space.
- COBYLA: Local search, uses linear approximations for constrained optimization.

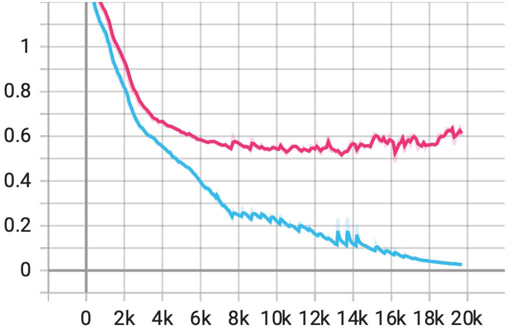


Fig: Comparison on SST-2

BBT-RGB

- M^2 Verbalizers

How to fully utilize limited signals ?

BBT-RGB

- M^2 Verbalizers

How to fully utilize limited signals ?

Multi-**M**ixed verbalizers → Fully exploiting the information from the Black-box.

Diversified Verbalizers Construction

1. Manual verbalizer selection.
2. Search-based verbalizer construction based on word importance estimation by TF-IDF.
3. Auto verbalizer generation based on neural nets (similar to LM-BFF).

Confidence of each category is represented by the avg prediction probability

BBT-RGB

- In^2 Initialization

Instruction and **In**-context learning → Better init. for prompt-based tuning!

- Appropriate init. has proven to play an essential role in prompt-based tuning.
- Part1: Task-specific manual Instruction
- Part2: Iterate through the entire training set and take each sample as a demonstration.
- Assessed together on a small dev set.

Performance of BBT-RGB
- Comparing with Multiple Baselines

Method	SST-2 acc	Yelp P. acc	AG's News acc	DBPedia acc	MRPC F1	SNLI acc	RTE acc	Avg.
Gradient-Based Methods								
Model Fine-Tuning	85.39 ±2.84	91.82 ±0.79	86.36 ±1.85	97.98 ±0.14	**77.35** ±5.70	54.64 ±5.29	**58.60** ±6.21	78.88
Prompt Tuning	68.23 ±3.78	61.02 ±6.65	84.81 ±0.66	87.75 ±1.48	51.61 ±8.67	36.13 ±1.51	54.69 ±3.79	63.46
P-Tuning v2	64.33 ±3.05	**92.63** ±1.39	83.46 ±1.01	97.05 ±0.41	68.14 ±3.89	36.89 ±0.79	50.78 ±2.28	70.47
Adapter	83.91 ±2.90	90.99 ±2.86	86.01 ±2.18	**97.99** ±0.07	69.20 ±3.58	57.46 ±6.63	48.62 ±4.74	76.31
LoRA	**88.49** ±2.90	90.21 ±4.00	**87.09** ±0.85	97.86 ±0.17	72.14 ±2.23	**61.03** ±8.55	49.22 ±5.12	78.01
BitFit	81.19 ±6.08	88.63 ±6.69	86.83 ±0.62	94.42 ±0.94	66.26 ±6.81	53.42 ±10.63	52.59 ±5.31	74.76
Gradient-Free Methods								
Manual Prompt	79.82	89.65	76.96	41.33	67.40	31.11	51.62	62.56
In-Context Learning	79.79 ±3.06	85.38 ±3.92	62.21 ±13.46	34.83 ±7.59	45.81 ±6.67	47.11 ±0.63	60.36 ±1.56	59.36
BBT	89.56 ±0.25	91.50 ±0.16	81.51 ±0.79	79.99 ±2.95	61.56 ±4.34	46.58 ±1.33	52.59 ±2.21	71.90
BBTv2	90.33 ±1.73	92.86 ±0.62	85.28 ±0.49	93.64 ±0.68	77.01 ±4.73	57.27 ±2.27	56.68 ±3.32	79.01
BBT-RGB (ours)	**92.89** ±0.26	**94.20** ±0.48	**85.60** ±0.41	**95.32** ±0.73	**80.49** ±1.84	**63.79** ±0.66	**62.82** ±1.20	**82.15**

Table: BBT-RGB vs Gradient-based/Gradient-free Baselines

Analysis

- Across different PTMs

Comparison of BBT-RGB and baselines on the large versions of GPT-2, BART, and T5.

- Directly applicable across different model architectures.

LM	Method	SST-2	AG's News	DBPedia
Decoder-Only Models				
GPT-2	BBT	75.53 ±1.98	77.63 ±1.89	77.46 ±0.69
	BBTv2	83.72 ±3.05	79.96 ±0.75	91.36 ±0.73
	BBT-RGB	**86.32** ±0.97	**82.01** ±0.81	**93.52** ±1.13
Encoder-Decoder Models				
T5	BBT	89.15 ±2.01	83.98 ±1.87	92.76 ±0.83
	BBTv2	91.40 ±1.17	85.11 ±1.11	93.36 ±0.80
	BBT-RGB	**92.91** ±0.97	**85.50** ±1.32	**93.74** ±0.56
BART	BBT	77.87 ±2.57	77.70 ±2.46	79.64 ±1.55
	BBTv2	89.53 ±2.02	81.30 ±2.58	87.10 ±2.01
	BBT-RGB	**92.63** ±1.43	**82.76** ±1.74	**88.26** ±1.06

Table: Experiments on different backbones.

Analysis

- Across different PTMs

Comparison of BBT-RGB and baselines on the large versions of GPT-2, BART, and T5.

- Directly applicable across different model architectures.
- Notable performance improvement.

LM	Method	SST-2	AG's News	DBPedia
Decoder-Only Models				
GPT-2	BBT	75.53 ±1.98	77.63 ±1.89	77.46 ±0.69
	BBTv2	83.72 ±3.05	79.96 ±0.75	91.36 ±0.73
	BBT-RGB	**86.32** ±0.97	**82.01** ±0.81	**93.52** ±1.13
Encoder-Decoder Models				
T5	BBT	89.15 ±2.01	83.98 ±1.87	92.76 ±0.83
	BBTv2	91.40 ±1.17	85.11 ±1.11	93.36 ±0.80
	BBT-RGB	**92.91** ±0.97	**85.50** ±1.32	**93.74** ±0.56
BART	BBT	77.87 ±2.57	77.70 ±2.46	79.64 ±1.55
	BBTv2	89.53 ±2.02	81.30 ±2.58	87.10 ±2.01
	BBT-RGB	**92.63** ±1.43	**82.76** ±1.74	**88.26** ±1.06

Table: Experiments on different backbones.

Analysis
- Cost-Effectiveness

- Better Performance
- Moderate Parameter Modifications
- More Stability

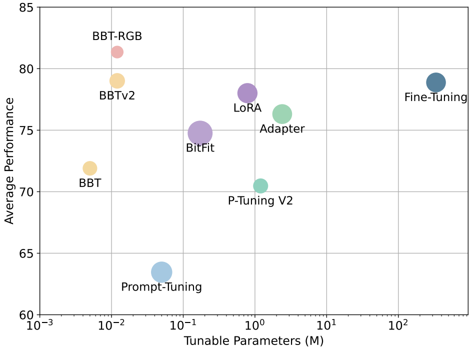


Fig: Cost-effective analysis

Acknowledgement

BBT-RGB is also derived from a prize-winning solution of the *First International Algorithm Case Competition: PLM Tuning Track, Guangdong-Hong Kong-Macao Greater Bay Area*.

The End

Thank You!