OptFormer: Towards Learning Universal Hyperparameter Optimizers with Transformers

Yutian Chen, Xingyou Song, Chansoo Lee, Zi Wang, Qiuyi Zhang, David Dohan, Kazuya Kawakami, Greg Kochanski, Arnaud Doucet, Marc'aurelio Ranzato, Sagi Perel, Nando de Freitas

Google AI Blog: ai.googleblog.com/2022/08/optformer-towards-universal.html
Paper: arxiv.org/abs/2205.13320
Code: github.com/google-research/optformer

NEURAL INFORMATION PROCESSING SYSTEMS

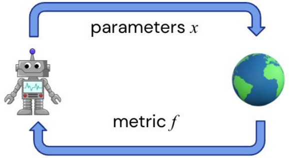
INTRODUCTION

Motivation

Data-driven hyperparameter optimization: Better priors + transfer learning

Problem: How to transfer-learn over *different search spaces and tasks*, containing unstructured text data?

Solution: Universal serialized data interface + Foundation Model for hyperparameter optimization



Google Vizier Database

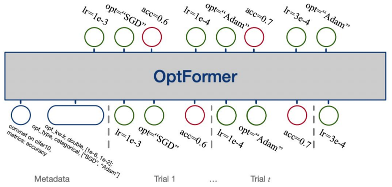
- 16M+ HPO studies from 5 years
- Covers data from ML, Ads, Search across Google
- Varying search spaces, metrics, horizons, algorithms, metadata

METHOD

OptFormer: One model, Two output types

Policy: $P(\mathbf{x}_{t+1} | m, \mathbf{x}_1, y_1, \dots, \mathbf{x}_t, y_t)$

Function regression: $P(y_t | m, \mathbf{x}_1, y_1, \dots, \mathbf{x}_t)$



Training

Standard supervised learning (behavioral cloning) w/ 250M param encoder-decoder Transformer

Inference as hyperparameter optimization

- Sample from prior

$$\mathbf{x}_{t+1} \sim P(\cdot | m, \mathbf{x}_1, y_1, \dots)$$

- **Augmented** with acquisition function

$$\mathbf{x}_{t+1} = \underset{\mathbf{x}^i \sim P(\mathbf{x}_{t+1}|\dots)}{\arg\max}\ u(P(y_{t+1}|\dots, \mathbf{x}^i))$$

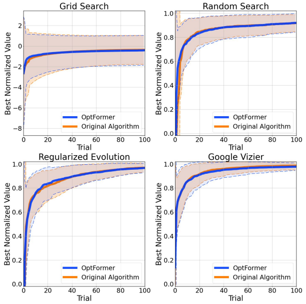
Data Serialization + Tokenization

- Normalized parameter + function values as integer token from [0, 1000)
- Metadata tokenized by standard English sentencepiece tokenizer

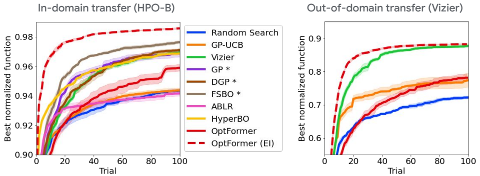
<table>
<tr><td>After preprocessing</td><td><name>:"convnet on cifar10",<metric>:"accuracy",<goal>:<MAXIMIZE>,
<algorithm>:"random_search"
&<name>:"opt_kw.lr",<type>:<DOUBLE>,<min_value>:1e-6,<max_value>:1e-2,
<scale_type>:<LOG>
&<name>:"opt_type",<type>:<CATEGORICAL>,<categories>:["SGD", "Adam"]
<831><0>⋆ <0> | <645><1>⋆ <999></td></tr>
<tr><td>Subwords after tokenization</td><td>name : " con v net on ci far 10 ", metric : " acc u racy ",
goal : MAXIMIZE , algorithm : " random _ search "
& name : " op t _ kw . lr ", type : DOUBLE , min_value : 1 e -6 ,
max_value : 1 e -2 , scale_type : LOG
& name : " op t _ type ", type : CATEGORICAL ,
categories : [" SG D ", " A dam "]
831 0 ⋆ 0 | 645 1 ⋆ 999</td></tr>
</table>

RESULTS

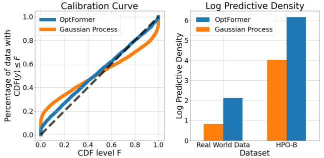
Simultaneous Algorithm Imitation



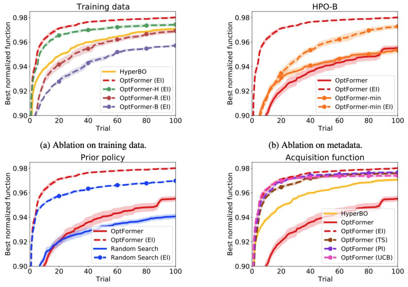
Bayesian-Augmented Policy Outperforms SOTA Google Vizier



Better Function Predictions than Gaussian Processes



Ablation Studies



(a) Ablation on training data.
(b) Ablation on metadata.
(c) Ablation on the prior policy.
(d) Ablation on the acquisition function.

CONCLUSION

Summary and Future:

- Text-based universal representations for hyperparameter optimizer
- Learn from huge HPO datasets
- Transformers = blackbox policies + function regressors
- Future: various extensions in search spaces, training algorithms, planning, multi-objective, ...