Google DeepMind

OmniPred: Language Models as Universal Regressors

Xingyou (Richard) Song, Oscar Li, Chansoo Lee, Bangding (Jeffrey) Yang, Daiyi Peng, Sagi Perel, Yutian Chen

Introduction

Background: Performance prediction is a powerful tool to predict outcome metrics of a system or model from a set of parameters, but traditionally restricted to methods which are only applicable to a specific task.

Our **OmniPred** Proposal: Train language models as universal end-to-end performance predictors over lots of (x,y) evaluation data from diverse experiments across Google.

Result: Through only textual representations of (x,y), language models are capable of very precise numerical regression and if given the opportunity to train over multiple tasks, can significantly outperform traditional regressors.

Training and Evaluation

Regular generative pretraining over token sequences. Inference uses regular temperature decoding over y-token logits.

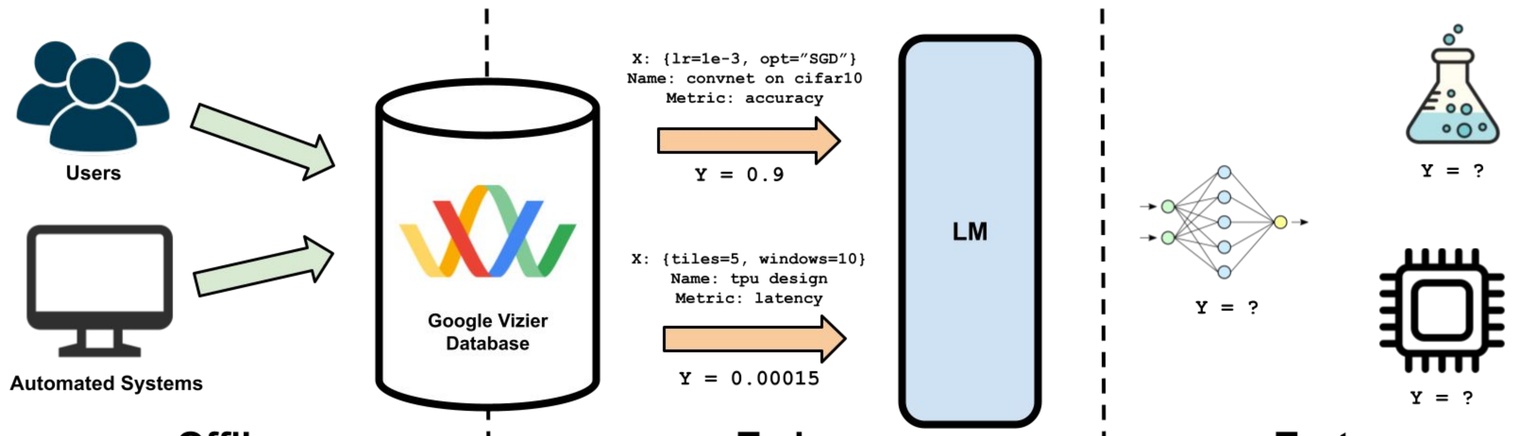
Study error defined as:

$$\frac{1}{y_{\max} - y_{\min}} \frac{1}{|\mathcal{D}^{\text{test}}|} \sum_{(x,y) \in \mathcal{D}^{\text{test}}} |f(x) - y|$$

Property	Statistic
# Studies	$\mathcal{O}$(70M+)
# Trials	$\mathcal{O}$(120B+)
# Distinct Users	$\mathcal{O}$(14K)

Vizier Database

Overview



Unified Textual + Token Representations

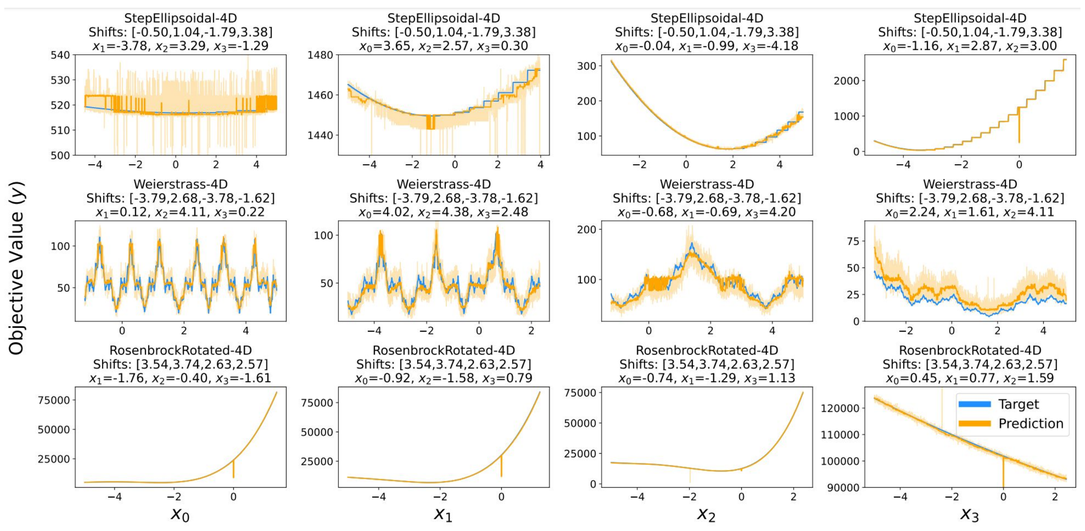
Represent suggestions x using JSON, and objectives y using custom digit-by-digit tokenization. Metadata (title, username, description, etc.) also can be used.

Independence from normalizations or constraints allows unification across any search space or objective scales.

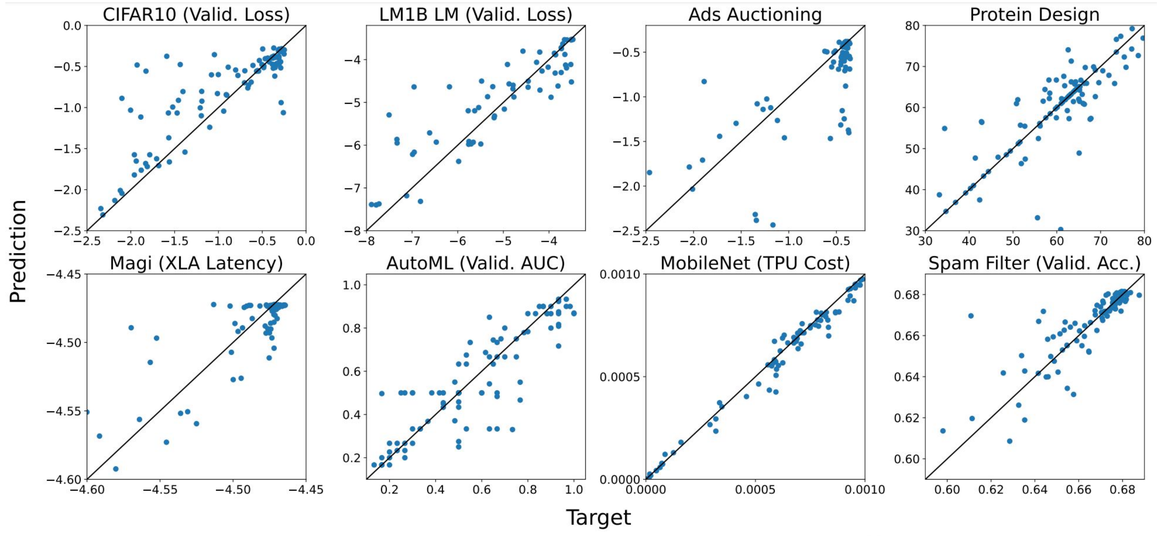
	Language Model Textual Representation
x	`batch_size:128,kernel:'rbf',learning_rate:0.5,model:'svm',optimizer:'sgd'`
m	`title:'classification',user:'some-person',description:'spam detection', objective:'accuracy'`
y	`<+><1><2><3><E-2>`

Simultaneous Regression:

Model is able to simultaneously regress over multiple different objectives and tasks.



Simultaneous regression over synthetic BBOB functions of varying scales.



Simultaneous regression over real-world objectives with different search spaces.

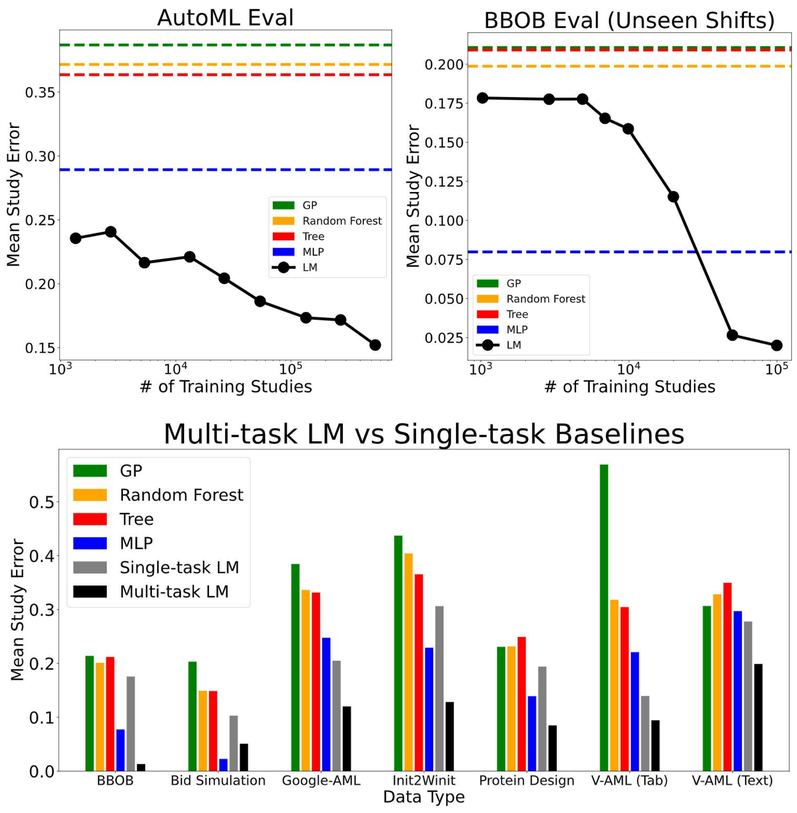
Multi-task Transferrability:

Model can transfer, i.e. improve performance on a single task using knowledge gained from similar but non-equivalent tasks.

Right: ***More training tasks*** lead to better performance prediction. Multi-task LM eventually outperforms traditional baselines as training data increases, over multiple domains. Interestingly, a single-task model remains a competitive choice!

Bottom: Multi-task training diminished if data is anonymized (replacing ***textual metadata*** with task id).

Datasets (# Training Studies)	Mean Study Error (↓) Original	Anonymized
BBOB (50K)	**0.03**	0.46
BBOB (Full 1M)	**0.01**	FAIL
AutoML (26.3K)	**0.19**	0.44
AutoML (Full 540K)	**0.15**	0.43



Adapting to Unseen Data

To deal with unseen tasks from random users, we can ***finetune*** from pretrained checkpoint. Gains over single-task can be large, depending on task similarity.

Method	Mean Study Error (↓)
Single-task (LM)	0.28
Pretrain (LM)	0.68
Pretrain + Finetune (LM)	**0.21**
MLP Baseline	0.25
Tree Baseline	0.32
Random Forest	0.32
Gaussian Process	0.42

Pretraining Dataset	Mean Study Error (↓) on AutoML Before Finetuning	After Finetuning
None (Single-Task)	0.98	0.20
BBOB	0.98	0.45
AutoML	**0.15**	**0.15**
Entire Vizier	0.31	**0.15**