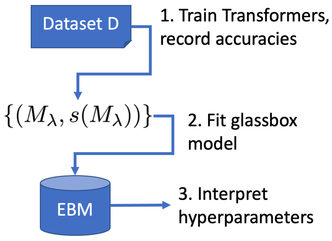
Post-Hoc Interpretation of Transformer Hyperparameters with Explainable Boosting Machines

Kiron Deb,* Xuan Zhang,* Kevin Duh
Johns Hopkins University

* These authors contributed equally to this work.

Post-Hoc Interpretation of Transformer Hyperparameters

Goal: To improve our understanding of hyperparameters in practice.



Type	Goal	Example Result
Prescriptive	Model Building	Given past experience, we recommend setting embedding size to 256 and attention head to 8 on Dataset D.
Descriptive (this work)	Post-Hoc Understanding	Given N models that are trained on dataset D, we find that embedding size influences BLEU more than attention heads.

Hyperparameter Search Datasets

A dataset on hyperparameter search for Transformer-based machine translation:
Reproducible and Efficient Benchmarks for Hyperparameter Optimization of Neural Machine Translation Systems, Zhang and Duh, TACL, 2020

Language Pairs	BPE (1k)	#layers	#embed	#hidden	#att_heads	init_lr (10^{-4})
zh-en; ru-en; ja-en; en-ja	10, 30, 50	2, 4	256, 512, 1024	1024, 2048	8, 16	3, 6, 10
sw-en	1, 2, 4, 8, 16, 32	1, 2, 4, 6	256, 512, 1024	1024, 2048	8, 16	3, 6, 10
so-en	1, 2, 4, 8, 16, 32	1, 2, 4	256, 512, 1024	1024, 2048	8, 16	3, 6, 10

* 2245 (hyperparameters, BLEU) pairs in total

Explainable Boosting Machines

Explainable Boosting Machine (EBM) is a generalized additive model with the form:

$$g(y) = \beta_0 + \sum_j f_j(x_j) + \sum_{ij} f_{ij}(x_i, x_j)$$

x: hyperparameters y: BLEU

f_j: feature function for feature x_j that is learnt through bagging and gradient boosting.

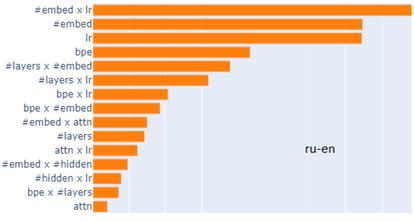
f_{ij}: models pairwise integrations between features.

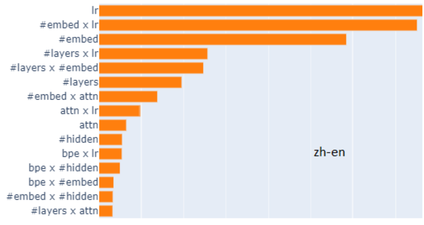
f_j f_{ij} Can be arbitrary shape functions based on 1 or 2 variables (hyperparamters) -> easy to interpret

Hyperparameter Analysis with EBM

I. Hyperparameter Importance

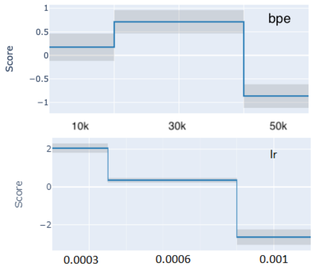
Hyperparameter importance score: $|f_j(x_j)|$





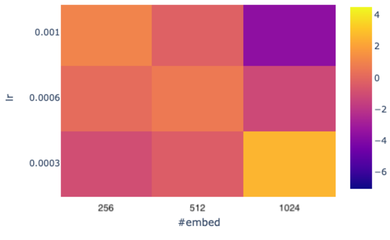
II. Single Hyperparameter Analysis

Score: higher score indicates a higher chance to get a higher BLEU score. $f_j(x_j)$



III. Pairwise Interaction Analysis

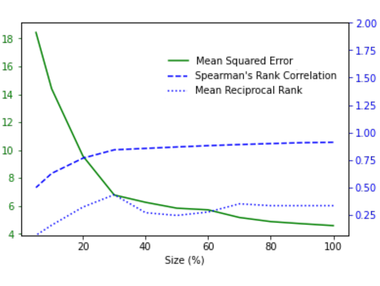
Score: higher score indicates a higher chance to get a higher BLEU score. $f_{ij}(x_i, x_j)$



Robustness Analysis of EBM

When can EBM be applied for this problem?

I. Varying Data Sizes



II. Transferability