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.

<table>
<thead>
<tr>
<th>Type</th>
<th>Goal</th>
<th>Example Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prescriptive</td>
<td>Model Building</td>
<td>Given past experience, we recommend setting embedding size to 256 and attention head to 8 on Dataset D.</td>
</tr>
<tr>
<td>Descriptive</td>
<td>Post-Hoc Understanding</td>
<td>Given N models that are trained on dataset D, we find that embedding size influences BLEU more than attention heads.</td>
</tr>
</tbody>
</table>

Hyperparameter Search Datasets


Language Pairs | BPE (#k) | #layers | #embed | #hidden | #att_heads | init_lr (10⁻⁴) |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>zh-en; ru-en; ja-en; en-ja</td>
<td>10, 30, 50</td>
<td>2, 4</td>
<td>256, 512, 1024</td>
<td>1024, 2048</td>
<td>8, 16</td>
<td>3, 6, 10</td>
</tr>
<tr>
<td>sw-en</td>
<td>1, 2, 4, 8, 16, 32</td>
<td>1, 2, 4, 6</td>
<td>256, 512, 1024</td>
<td>1024, 2048</td>
<td>8, 16</td>
<td>3, 6, 10</td>
</tr>
<tr>
<td>so-en</td>
<td>1, 2, 4, 8, 16, 32</td>
<td>1, 2, 4, 6</td>
<td>256, 512, 1024</td>
<td>1024, 2048</td>
<td>8, 16</td>
<td>3, 6, 10</td>
</tr>
</tbody>
</table>

* 2245 (hyperparameters, BLEU) pairs in total

Explainable Boosting Machines

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

\[ y(\mathbf{x}) = \beta_0 + \sum_j f_j(x_j) + \sum_{i,j} f_{ij}(x_i, x_j) \]

- \( \mathbf{x} \): 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 (hyperparameters) \( \rightarrow \) 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