CSE 234
Data Systems for Machine Learning

Arun Kumar

Exercise 3
Exercise

Q1) [2pts] Which of the following aspects of DL roughly resembles the pain of feature engineering in classical ML?

A. Training runtimes
B. Inference runtimes
C. Architecture selection
D. Hyperparameter tuning
E. Visualizing the weights
F. None of the above
Exercise

Q2) [2pts] What is the new programming paradigm ushered in by DL systems?

A. Imperative programming
B. Declarative programming
C. Functional programming
D. Differentiable programming
E. Integrable programming
F. None of the above
Q3) [2pts] Which of the following hyperparameter tuning approaches is the most popular in practice as per surveys?

A. Grid search
B. Random search
C. Hyperband
D. ASHA
E. Bayesian optimization
F. All of the above
Exercise

Q4) [2pts] Which of the model selection steps can potentially reduce the bias and raise the variance for a GLM?

A. Adding feature interactions
B. Dropping features
C. Subsampling the training dataset
D. Adding more training examples
E. Strengthening regularization
F. None of the above
Exercise

Q5) [3pts] When running Horovod on a 10-worker cluster, how many times does a worker talk to other workers to put together a full mini-batch gradient during SGD?

Q6) [4pts] Briefly explain 2 pros of Cerebro over task-parallel model selection systems.

Q7) [4pts] Briefly explain 2 cons of custom compiler stacks over TVM for DL inference.
Exercise

Q8) [6 x 3pts] Suppose you run Hyperband for hyperparameter tuning with the following knobs: \( R = 25; \ \eta = 5 \). It yields the following brackets in the same table format from the paper as explained in class.

<table>
<thead>
<tr>
<th>i</th>
<th>s = 2</th>
<th>s = 1</th>
<th>s = 0</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( n_i )</td>
<td>( r_i )</td>
<td>( n_i )</td>
</tr>
<tr>
<td>0</td>
<td>25</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>1</td>
<td>5</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>

A. What is the total number of unique hyperparameter configs tried?
B. How many configs were trained for 25 epochs?
C. How many configs were trained for at least 5 epochs?
D. How many configs were trained for at least 1 epoch?
E. How many configs got killed in total by early stopping?
Q9) [6 x 4pts] You are given the following properties of a model selection workload: model size 2 GB; 10 workers; 80 configs given together; 20 training epochs per config; 2 TB dataset with 10 billion data examples; SGD mini-batch size of 500. What is the absolute communication cost of the following systems approaches to executing this workload? Round to nearest TB.

A. Task parallelism with full replication
B. Task parallelism with full remote reads
C. BSP-style TF model averaging
D. Horovod
E. Asynchronous Parameter Server
F. Cerebro