Good morning!

CS 744: CLIPPER

Shivaram Venkataraman
Fall 2020
Course Project Proposals
- Due on Friday!
- See Piazza for template
- Submission instructions soon

Midterm details
- Open book, open notes
- Held in class time 9.30-10.45am Central Time
- Type / Upload photos (extra 15 mins)
MACHINE LEARNING: INFERENCE

Training
- ML model
- Learn
- Data parallel
- Pipeline parallel

Inference
- Query
- Prediction
- Netflix recommend movies

CLIPPER

Feedback

Training Data

Model

$X \rightarrow \hat{y}$
GOALS

- Interactive latencies (tail latency < 100ms)

- High throughput to handle load

- Improved prediction accuracy

- Generality (?)
MODEL CONTAINERS

interface Predictor<X,Y> {
    List<List<Y>> pred_batch(List<X> inputs);
}

Run using Docker containers

Can be replicated across machines

→ Model frameworks are black box for chipper!

→ Interface is implemented once per framework

→ Instantiate TF Shim with ResNet 18

→ Instantiate TF Shim with ResNet 50
MODEL ABSTRACTION LAYER

Caching

- Improve performance for frequent queries
- LRU eviction policy
- Important for feedback
Goals, Insight
- Increase latency (within SLO) for improved throughput
- Reduce RPC overheads
- GPU / BLAS acceleration

Approach
- Per container queues.
- Why?
ADAPTIVE BATCHING

AIMD: Additive Inc Multiplicative Dec (AIMD)

Why?
- Increase batch size carefully
- & latency quickly comes down if it's above SLO!

Delayed: Wait until batch exists

Why?
- Collect examples up to a certain time (5ms)
- Then dispatch

observe latency \( t < SLO \)

inc. batch

\( 0 \, 1 \, 2 \, 3 \, 4 \, 5 \)

Batch Size

\( 0 \, 1 \, 2 \, 3 \, 4 \)

Time

(inc) Should I wait?
(decr) Should I send?
Model Selection

- Improve Accuracy

Ensembles

Learn

Version 2

Version 3

“CAT”

“CAT”

“CAT”

“CAT”

Policy

“CAT”

Confident
SINGLE MODEL SELECTION

Multi-Arm Bandit formulation
- Explore vs Exploit
- Regret: Loss by not picking optimal action
- Goal: Minimize regret

Clipper
- Exp3 algorithm
- Single evaluation
- Scales to more models
  - Update weights based on feedback

Given you have N models, which one should I use?

Model Selection
- Model 1
- Model 2
**MULTI MODELS**

Ensemble
- Combine output from models (weighted average)
- How do we get the weights?

Robust Prediction
- React to model changes
- Output confidence score

\[ \alpha y_i + \beta z_i \]

*Binary classifier* → cat or log

Exp4 → update \( \alpha \) & \( \beta \)

Combine & threshold > 0.5 → cat log
STRAGGLER MITIGATION

Why do stragglers occur?

- If we wait for \( N \) model containers to reply, some of them might be slow?
- More replicas / scaling?

Approach

- Approx result based on whatever has finished
- Better approx now late! \( \rightarrow \) ML specific
SUMMARY

• Clipper: ML inference Workloads + Requirements
• Layered architecture provides generality
• Caching, Batching, Replication to improve latency, throughput
• Multi-Arm bandits to improve accuracy
DISCUSSION

https://forms.gle/FCVhPURqz7HSbDtg6
Consider a scenario where you run a model serving service that hosts a number of different applications. The traffic for some applications is sporadic (e.g. only a few hours where they are used). What are some advantages / disadvantages of using Clipper for such a service?

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>→ Adaptive batching, delayed</td>
<td>→ Each might be contented</td>
</tr>
<tr>
<td>→ multiple replicas, elasticity</td>
<td>→ how do this online + fashion?</td>
</tr>
<tr>
<td>→ Containerization, isolate apps</td>
<td></td>
</tr>
</tbody>
</table>

User provided code

```
10ms
```

```
EPC
```

```
Storage
```

```
App 1
```

```
App 2
```

```
Clipper
```
Larger ensembles use a lot of resources. 8-10 seems to be a good size. Balancing different things?

What's going on here?

Latency inflation is very low. Reasonable accuracy.