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

Inference

Training Data

Learn

Model

Adapt

Feedback

Query

Prediction

Application

CLIPPER

$x \xrightarrow{\mathcal{F}} \hat{y}$
GOALS

- Interactive latencies (tail latency < 100ms)

- High throughput to handle load

- Improved prediction accuracy

- Generality (?)
MODEL CONTAINERS

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

Run using Docker containers

Can be replicated across machines
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

Why?

Delayed: Wait until batch exists

Why?
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
MULTI MODELS

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

Robust Prediction
- React to model changes
- Output confidence score
Why do stragglers occur?

Approach
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?
(a) Latency

(b) Missing Predictions

(c) Accuracy