

CMSC828G: Special Topics in Systems for Machine Learning

Speculative Decoding

Instructor: Yafan Huang



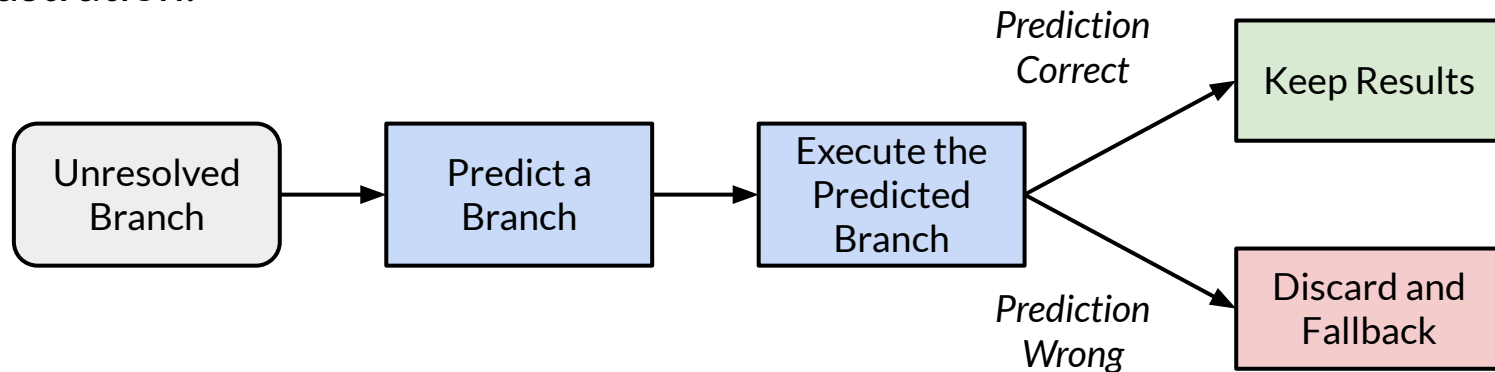
DEPARTMENT OF
COMPUTER SCIENCE

Motivation

- Autoregressive decoding generates one token at a time.
 - Each iteration requires an expensive target-model execution, often bottlenecked by memory bandwidth.
 - Sequential token dependencies limit opportunities for parallel execution.
-
- Can we reduce the number of expensive model executions by performing some cheaper work in advance?

Speculative Execution in CPUs

- Illustration:



- Toy example:

```
if (x>0) y=a+b;  
else y=a-b;
```

1. Predict: assuming "x>0";
2. Execute "a+b" ahead;
3. Verify real x value;
4. Keep or discard.

From Speculative Execution to LLM Decoding

CPU Speculative Execution

Branch prediction

Speculative instructions

Branch resolution

Commit correct execution

Squash incorrect execution

LLM Speculative Decoding

Draft token prediction

Speculative tokens

Target model verification

Accepted verified tokens

Reject drafts and correct output

Speculative Decoding

- A **smaller draft model** proposes multiple tokens; the **larger target model** verifies them in fewer expensive steps.
- Helps most when decode is the bottleneck and draft predictions are accurate.



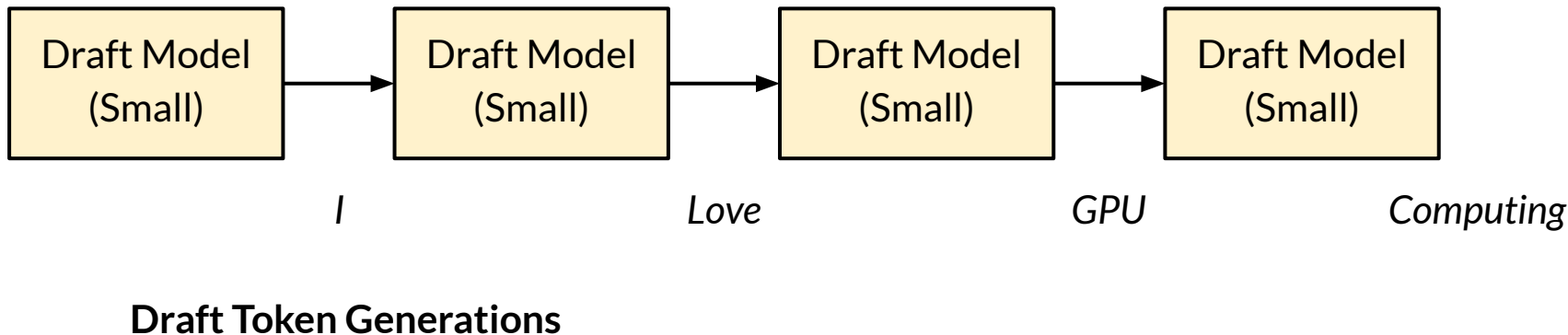
Draft Model
(Small)

Target Model
(Large)

Speculative Decoding

- A **smaller draft model** proposes multiple tokens; the **larger target model** verifies them in fewer expensive steps.
- Helps most when decode is the bottleneck and draft predictions are accurate.

Draft Inference (multiple pass in a decoding loop)

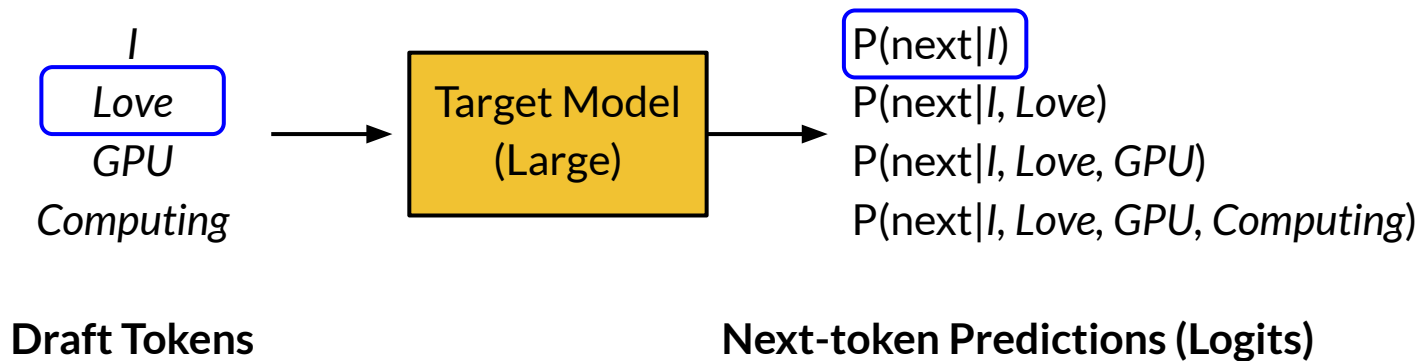


Speculative Decoding

- A **smaller draft model** proposes multiple tokens; the **larger target model** verifies them in fewer expensive steps.
- Helps most when decode is the bottleneck and draft predictions are accurate.

Target Inference (one pass with multiple tokens)

Target Model Verification



Why Does Speculation Help (or Hurt)?

- Speculation is beneficial only when the saved execution time exceeds the speculation overhead.
 - What if the draft model is too expensive?
 - What if most speculative tokens are rejected?
 - What if parallel verification itself becomes expensive?
 - What if we do serving on edge devices?

Research Challenges and Opportunities

Challenges

Drafting Overhead

Prediction Quality

Verification Efficiency

System Integration

Research Questions

Can we generate speculative tokens more cheaply?

Can we improve the chance of finding useful future tokens?

Can we efficiently verify multiple candidate sequences?

Batching, KV cache, and other resource management.

Student Presentations

- [ICML'24] Medusa: Simple LLM Inference Acceleration Framework with Multiple Decoding Heads
- [ASPLOS'24] SpecInfer: Accelerating Large Language Model Serving with Tree-based Speculative Inference and Verification