

Decoding Strategies

Current Research in Decoding

Learning goals

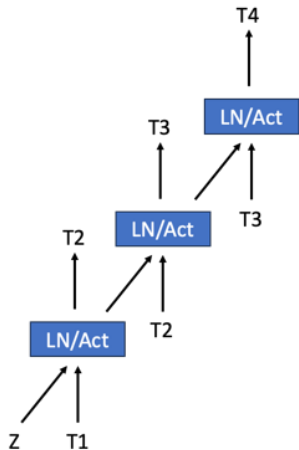
- Get to know speculative decoding
- Learn about Minimum Bayes Risk (MBR) Decoding

SPECULATIVE DECODING

► PyTorch

- Models generate the output sequence token by token
- A lot of forward passes are required to generate a long sequence of tokens
- Is there a way to generate the same output but also save time?
- **Idea:** attach multiple speculative heads to the model to predict the $N + 1^{st}$, $N + 2^{nd}$, $N + 3^{rd}$, etc. token


SPECULATIVE DECODING: ARCHITECTURE



► Speculator architecture

- They base the architecture on the Medusa paper ► Cai et al., 2024
- Make heads hierarchical, where each head stage predicts a token and then feeds it into the next head stage
- Z is the hidden state from the base model and $T1, \dots, T4$ are the generated tokens

SPECULATIVE DECODING: RESULTS

Figure: Speed comparison between vanilla decoding (left) and speculative decoding (right) 

SPECULATIVE DECODING: RESULTS

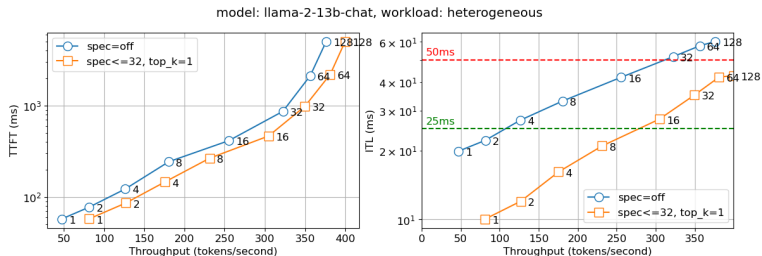


Figure: Time to first token (TTFT - left) and Inter-token latency (ITL - right) for Llama 13B with number of concurrent users indicated on the graph [PyTorch blog](#)

SPECULATIVE DECODING: TRAINING

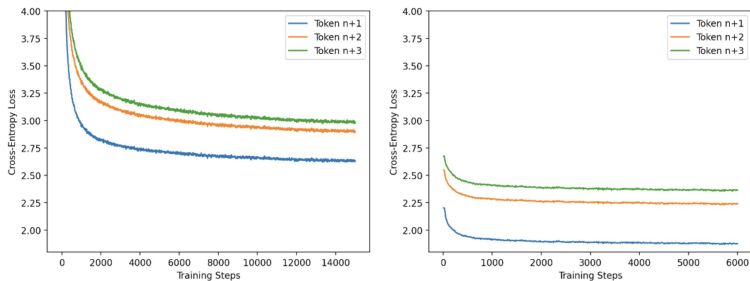


Figure: Per-head training loss curves for Llama2-13B speculator training, phase 1 and 2 [► PyTorch blog](#)

- They use a two phase approach to training a speculator to be more efficient
- **Phase 1:** Train on small batches with long sequences (4k tokens) with standard causal LM approach

SPECULATIVE DECODING: TRAINING

- **Phase 2:** Use large batches with short sequence lengths (256 tokens) generated from the base model
- Tune the heads to match the output of the base model

MINIMUM BAYES RISK (MBR) DECODING

► GitHub, suzyahyah

- MBR is based on bayesian decision theory, where one would pick an action based on minimizing the *Bayes Risk*:

$$\alpha^* = \operatorname{argmin}_{\alpha \in \mathcal{A}} \mathbb{E}_{\theta \sim p(\theta)} [\mathcal{L}(\theta, \alpha)]$$

- MBR Decoding involves choosing the bayes optimal action, where the action is a sequence
- Given a source input x (i.e. source language in machine translation), the space of possible hypothesis $h \in \mathcal{H}(x)$, a probability distribution over decoded sequences $p(y|x)$, and a loss function $\mathcal{L}(y, h)$, the MBR decode is given by:

$$h^* = \operatorname{argmin}_{h \in \mathcal{H}(x)} \mathbb{E}_{p(y|x)} [\mathcal{L}(y, h)]$$

MINIMUM BAYES RISK (MBR) DECODING

► GitHub, suzyahyah

- In theory we would like to have a distribution over reference sequences, which would be our $p(y|x)$
- But at inference time that is not available and we use $p_{\text{model}}(y|x)$ instead, as well as to construct $\mathcal{H}(x)$
- In practice MBR decoding has the following design choices (since its theoretical hypothesis space is infinite):
 - Construction of the hypothesis space $\mathcal{H}(x)$
 - Construction of the monte-carlo set of samples $y \in \mathcal{Y}$ to approximate $\mathbb{E}_{p(y|x)}$
 - The choice of loss function \mathcal{L} , like BLEU, precision, etc.
 - Choosing how to renormalise samples y from $p(y|x)$ with a small number of samples, the sequences are unlikely to be repeated and the monte-carlo estimate would give them all uniform probability