Qeios

Peer Review

Review of: "Draft Model Knows When to Stop: A Self-Verification Length Policy for Speculative Decoding"

Muhua Zhu[1]

1. Independent researcher

This paper proposes a new approach, named SVIP to speed up speculative decoding by adaptively adjusting draft length. Extensive experiments across multiple models and generation lengths demonstrate the superior performance of the proposed approach.

Strengths:

1. Although the idea of draft length control is not new, the idea of deciding draft length by relying on information solely from the draft model is relatively novel.
2. SVIP is a flexible and training-free approach that can be applied together with a variety of autoregressive LLMs. Experimental results also demonstrate the effectiveness of the approach.

Weaknesses:

1. Some content can be improved to make the paper more understandable. See the details in the following Suggestions.

Suggestions:

1. In order to derive Eq. 4 from Eq. 1 in Section 2.1, the authors relate β to KL divergence by using TVD as a bridge. But, to make the paper more self-contained and understandable, I think TVD should be defined and explained in the section. The same thing should be done for ``Pinsker's inequality'', because, with no definition of the inequality, we cannot tell how TVD is related to KL divergence.
2. Regarding the experiments, the authors use wall-time speedup as performance evaluation. Maybe we should also evaluate the change in generation quality caused by draft length control.

Grammatical errors:

1. ``autoregressive, which'' --> ``autoregressively, which''

Declarations

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/MFZO51