2025년도 한국통신학회 추계종합학술발표회

Decision Transformer-based Hierarchical AI Model Selection for Energy-Efficient Future Networks

Abate Selamawit Chane and Seong Ho Jeong*
Hankuk University of Foreign studies
selamchane@hufs.ac.kr, *shjeong@hufs.ac.kr

에너지 효율적 미래 네트워크를 위한 의사결정 트랜스포머 기반 계층적 인공지능 모델 선택

셀라마윗, 정성호
한국외국어대학교

Abstract

As next-generation networks such as 6G evolve toward AI-native architectures, both network operations and end-user applications are increasingly dependent on computationally intensive AI models. This paradigm shift introduces a significant energy footprint, driven by two major contributors: computation energy and transmission energy. Despite growing awareness, there remains a lack of systematic studies addressing the combined impact of AI computation and transmission on overall energy consumption. To quantify this impact, we conducted an experiment using six optical character recognition (OCR) models, two deployed on-device and four at the network edge. The results reveal a steep rise in energy consumption with model size and deployment distance. Building on these insights, we propose an energy-aware adaptive model selection framework that dynamically balances performance and energy efficiency. The framework leverages a decision transformer-based reinforcement learning mechanism to intelligently select the most suitable model from a hierarchical model pool spanning device, edge, and cloud deployments.

Ⅰ. Introduction

The rapid expansion of artificial intelligence (AI) technologies across diverse domains has led to an unprecedented reliance on large-scale AI models. This reliance, however, comes with a significant rise in energy consumption, primarily from two sources: computation energy, driven by resource-intensive training and inference of large models, and transmission energy, incurred during the transfer of data and inference tasks across devices, edge servers, and cloud infrastructures.

Ⅱ. Method

During the inference phase, the conventional practice of transmitting every task to large models residing in centralized servers can lead to significant and unnecessary energy expenditure. Not all tasks require the precision or complexity of such heavy models. The optimal model choice depends on the nature of the input data and the criticality of the task; simple or non-critical inferences may achieve sufficient accuracy with much lower computational cost. An adaptive and context-aware inference strategy can therefore yield substantial energy savings by reducing both transmissions overhead and the computation burden of large, centralized models.

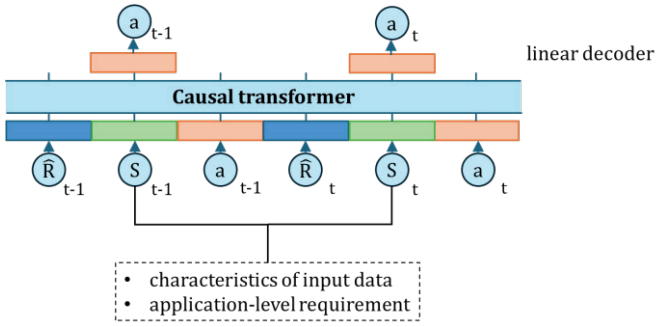

Fig. 1. Decision transformer Architecture

Fig.1 illustrates that the state for the decision transformer integrates both the intrinsic properties of the input data: capturing its complexity and the application-level requirements such as accuracy tolerance and latency targets. The table below presents an analysis of the energy consumption (EC) in KWh of various OCR models evaluated over 1,000 inference requests.

Table.1 Comparison of various OCR models

Models	size	Location	Accuracy	Latency	EC
Tesseract	~100K	On-device	0.334	83.38	0.001
Paddle OCR	~4M	On-device	0.71	14.91	0.00041
Easy OCR	~22M	Edge	0.79	40.58	0.00044
TrOCR-Base	~333M	Edge	0.881	82.99	0.00326
TrOCR-Large	~608M	Edge	0.91	118.84	0.005825
Moon dream	~2B	Edge	0.885	248.4	0.01304

Ⅲ. Conclusion

This proposal demonstrates the potential of integrated advanced decision transformer-based reinforcement learning techniques for adaptive, energy-efficient AI model selection.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (RS-2024-00436887) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

REFERENCES

[1] Zhang, P., Xiao, Y., Li, Y., Ge, X., Shi, G., & Yang, Y. (2023). Toward net-zero carbon emissions in network AI for 6G and beyond. *IEEE Communications Magazine*.