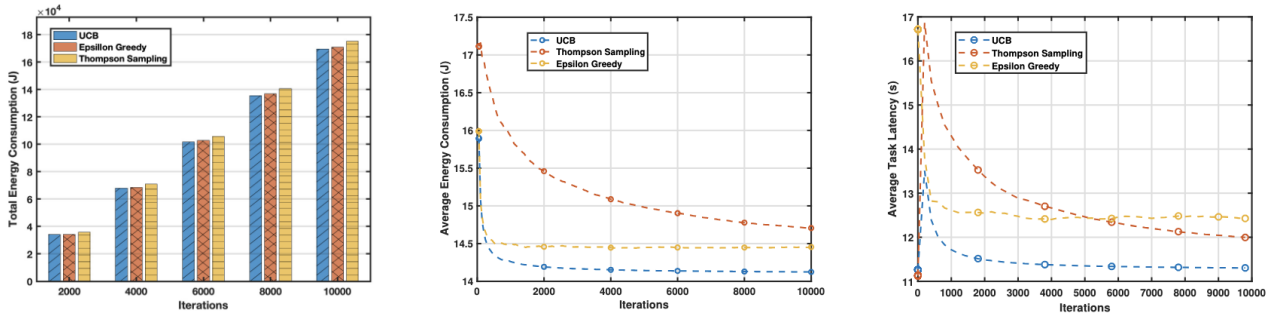
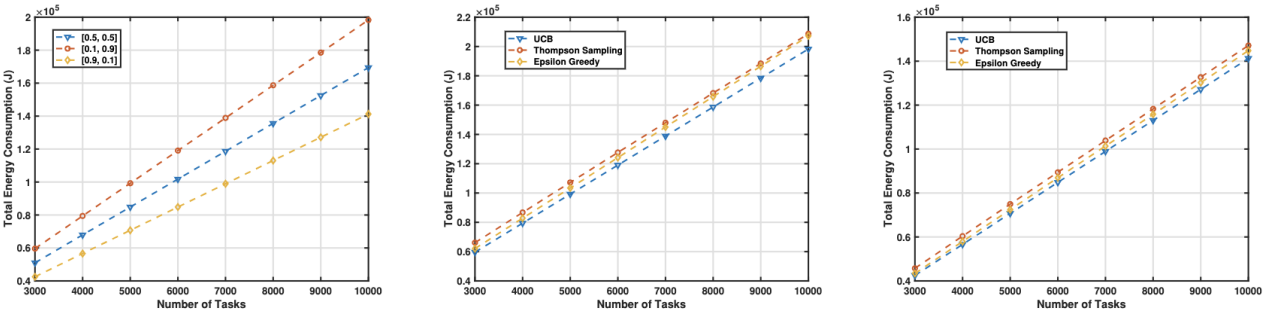


(a) (b) (c)

図２－１消費電力と遅延の比較

まず、正確性の要件を $\lambda_n > 20$、推論速度の要件を $\delta_n > 2$ と設定し、7,000 回のイテレーションで 3 つのアルゴリズムの総消費電力を評価する。総消費電力に関する比較結果は図２－１a に示される。トンプソンサンプリングは他のアルゴリズムに比べて高い消費電力を示す。0 から 1,000 回のイテレーションでは、イプシロングリーディと UCB アルゴリズムの性能はほぼ同じで、平均約 16,000 ジュールである。しかし、イテレーション数が増えるにつれて（1,000 から 7,000 回）、UCB アルゴリズムの優位性が明らかになる。7,000 回のイテレーション時点で、イプシロングリーディアルゴリズムの総消費電力は約 110,400 ジュールに達し、一方で UCB アルゴリズムは約 99,200 ジュールの低いレベルを維持する。この傾向は、MAB 問題における消費電力効率の最適化において UCB アルゴリズムがイプシロングリーディアルゴリズムよりも効果的であることを示している。



(a) (b) (c)

図２－２アルゴリズム性能分析

次に、10,000 回のイテレーションで 3 つのアルゴリズムの平均消費電力を評価する。結果は図２－１b に示される。すべてのアルゴリズムの平均消費電力はイテレーション数が増えるにつれて減少するが、UCB アルゴリズムが最初に収束する。UCB アルゴリズムは 1,000 回のイテレーション後に収束する。10,000 回のイテレーション後、UCB アルゴリズムの平均消費電力は 14.2 ジュールであり、イプシロングリーディアルゴリズムは 14.5 ジュール、トンプソンサンプリングアルゴリズムは 14.7 ジュールである。3 つのアルゴリズムの平均タスク遅延は図２－１c に示される。UCB アルゴリズムは常に最も低い平均遅延を維持することができる。

タスクには異なる要件があるため、比較のために異なる要件を持つ 3 つのセットを追加する。元の要件を $[0.5, 0.5]$ と定義し、タスクの半分は正確性に敏感（$\lambda_n > 40$）、他の半分は推論速度に敏感（$\delta_n > 9$）とする。さらに、$[0.1, 0.9]$ と $[0.9, 0.1]$ という 2 つのユニークな組み合わせを選ぶ。前者はタスクの 90%が正確性に敏感で、10%が推論速度に敏感であることを意味し、後者はタスクの 90%が推論速度に敏感で、10%が正確性に敏感であることを意味する。シミュレーション結果は図２－２a に示される。

正確性に敏感なシナリオでは、タスク数が 10,000 に達するとシステムの総消費電力は約 202,000 ジュールになる。この観察は、厳しい正確性基準を満たすために、モデルパラメータ数の多いモデル（例：LLaMA-33B）を選ぶ傾向があることを示している。一方、推論速度を重視するシナリオでは、タスク数が 10,000 に達するとシステムの総消費電力は約 142,000 ジュールになる。この場合、アルゴリズムは通常、推論速度が速いモデルを選び、これらのモデルは通常、パラメータ数が少なく、消費電力が低い。LLaMA-7B は、推論速度に敏

感なタスクの要件に適しているモデルの代表例である。3 つのアルゴリズムの性能を 2 つの異なるシナリオで比較する。図２−２b は正確性重視の制約下でのアルゴリズムの動作を示し、図２−２c は推論速度重視の制約下でのアルゴリズムの動作を示す。

実験は、LLM 推論タスクのオフロードにおいて最適な構成値を選択する際の UCB アルゴリズムの優位性を明確に示している。特に、UCB アルゴリズムは他のアルゴリズムよりも一貫して迅速に最適な構成値を特定し、正確な決定を下す能力を示している。さらに、このアルゴリズムは、3 つの異なるタスク要件においても迅速に収束することで、その優れた堅牢性を証明している。

３ まとめ

本研究では、災害に強い IoT システムの構築に向けたインテリジェントエッジコンピューティング技術を用いたエネルギー効率の高いセマンティックセグメンテーションおよび大型言語モデル（LLM）の推論手法を提案した。災害時においても迅速かつ正確な情報処理が求められる IoT システムにおいて、エッジコンピューティングの利用はクラウドへの依存を減らし、リアルタイム処理を実現するための有効な手段である。セマンティックセグメンテーションにおいては、ビデオ解像度、ニューラルネットワークの深さ、および圧縮率といったパラメータを最適化することで、エネルギー消費を削減しながら高精度な結果を得る方法を示した。具体的には、マルチアームバンディット問題として定式化し、上限信頼境界（UCB）アルゴリズムを用いることで、効率的なリソース配分とエネルギー消費の最小化を実現した。また、LLM の推論においても、エッジデバイスの種類とモデルパラメータの数を調整することで、エネルギー効率の高いタスク処理を達成した。異なるタスク要件に応じた最適な構成を選択することで、エネルギー消費と推論性能のバランスを取る手法を実証した。UCB アルゴリズムの適用により、エッジ環境での LLM 推論タスクにおいても優れたエネルギー効率を示した。これらの成果は、災害時における IoT システムの持続可能な運用と高信頼性を支える重要な技術基盤となる。

【参考文献】

[1] Yuan, X., Li, H., Ota, K., & Dong, M. (2023). Building energy efficient semantic segmentation in intelligent edge computing. IEEE Transactions on Green Communications and Networking.

[2] Yuan, X., Li, H., Ota, K., & Dong, M. (2024). Generative Inference of Large Language Models in Edge Computing: An Energy Efficient Approach. The 20th International Wireless Communications & Mobile Computing Conference (IWCMC).

〈発　表　資　料〉

題　　名	掲載誌・学会名等	発表年月
AI in SAGIN: Building deep learning service-oriented space-air-ground integrated networks	*IEEE Network*	2022 年 8 月
Learning IoV in 6G: Intelligent edge computing for Internet of Vehicles in 6G wireless communications	*IEEE Wireless Communications*	2023 年 3 月
HyScaler: a dynamic, hybrid VNF scaling system for building elastic service function chains across multiple servers	*IEEE Transactions on Network and Service Management*	2023 年 5 月
Building energy efficient semantic segmentation in intelligent edge computing	*IEEE Transactions on Green Communications and Networking*	2023 年 10 月
A multi-scale self-supervised hypergraph contrastive learning framework for video question answering	*Neural Networks*	2023 年 11 月
Generative Inference of Large Language Models in Edge Computing: An Energy Efficient Approach	*2024 International Wireless Communications and Mobile Computing*	2024 年 5 月