災害に強い IoT システムの研究開発

代表研究者　　李　鶴　　室蘭工業大学大学院工学研究科准教授

1 はじめに

Society 5.0 で目指す IoT 社会は、さまざまなモノが繋がることで情報を集約し、リアルタイムに解析することで効率や生産性を向上させる豊かな世界である。近年、多くの IoT デバイスが普及しており、エリクソンの予測によれば、2022 年までに IoT デバイスは 180 億台近くに達する見込みである。実際に、登別市では官民連携により灯油タンクに IoT デバイスを取り付けて残量を把握できるスマートメーターの設置が推進されている（2018 年 4 月室蘭民報朝刊)。このような進展により、日常生活や産業における IoT デバイスの利便性は飛躍的に向上している。

一方で、災害発生後にこれらの IoT デバイスを適用することは大きなチャレンジである。災害時には、ネットワークインフラが損傷を受ける可能性が高く、従来の集中型クラウドコンピューティングでは対応が困難となる。これに対して、エッジコンピューティングは分散型の処理能力を提供することで、災害時における迅速な対応と高い信頼性を実現する技術として注目されている。エッジデバイスが独立して動作し、高効率で情報処理を行う能力は、災害対応の迅速性と効果を大きく向上させる。

インテリジェントエッジコンピューティングは、IoT デバイスに近い場所にコンピューティングリソースを配置することで、リアルタイム処理を実現し、クラウドコンピューティングの遅延を低減する災害に強い技術パラダイムである。この技術は、セマンティックセグメンテーションや大型言語モデル（LLM）の推論タスクなど、さまざまな高性能コンピューティングタスクにおいて大きな可能性がある。特に、災害後にビデオ監視、医療画像解析、自然言語処理などの要求を対応には、エッジコンピューティングが効率的かつ迅速な処理を提供することで、リアルタイムの意思決定を支援することが期待されている。

しかし､災害後のエッジ環境へのデプロイメントには､計算リソースの消費と消費電力効率の課題が伴う。通常､電力不足やバッテリー寿命が限られているため､消費電力を最小限に抑えることが求められる｡特に､セマンティックセグメンテーションや LLM 推論タスクは､高精度な結果を得るために大量の計算を必要とし､エッジデバイスの消費電力を増加させる要因となる。したがって、消費電力の効率の向上は、エッジコンピューティングシステムの設計において重要な課題である。

本研究では、インテリジェントエッジコンピューティング環境における消費電力効率の高いセマンティックセグメンテーションと LLM 推論の実現を目指す。具体的には、タスクのオフロードに関するパラメータの最適化を通じて、消費電力と性能のバランスを取る方法を探求する。セマンティックセグメンテーションでは、ビデオ解像度、ニューラルネットワークの深さ、および圧縮率などのパラメータを最適化し、LLM 推論では、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、効率的なタスク処理を実現する。

2 研究成果

2-1 インテリジェントエッジコンピューティングにおける消費電力効率の高いセマンティックセグメンテーションの構築

（1）問題の提出

セマンティックセグメンテーションは、コンピュータビジョンの重要な分野であり、画像内の異なる意味に基づいてピクセルをグループ化する。ビデオ監視、医療画像解析、植物病害検出などの多くの重要な分野では、大量の異なるデータストリームを処理するためにセマンティックセグメンテーションモデルが必要とされる。通常、ほとんどのセグメンテーションアプリケーションは、高精度モデルの処理に膨大な計算が必要なため、ローカルデバイスではなく高性能なクラウドサーバーにデプロイされる。

インテリジェントエッジコンピューティングは、データソースに近いネットワークのエッジにコンピューティングサーバーを配置し、ローカルデバイスでリアルタイムのセマンティックセグメンテーションをサポートし、クラウドベースの処理に伴う遅延を軽減する。この高度なエッジコンピューティングは、人工知能および機械学習アルゴリズムを使用してエッジ層でデータを分析し、よりスマートな意思決定と効率的なリ

公益財団法人電気通信普及財団
研究調査助成報告書　第39号　2024年度

ソース配分を可能にする。しかし、他の AI タスクとは異なり、エッジ環境にデプロイされるセマンティックセグメンテーションタスクは、より多くの計算リソースと複雑なモデルのデプロイメントを必要とする。これにより、そのようなタスクを処理するエッジデバイスの消費電力が増加する。さらに、エッジデバイスの電力供給は常に限られており、消費電力効率が重要な課題となる。したがって、特定の環境やアプリケーションに関係なく、エッジデバイスの長期間かつ安定した運用を保証するためには、消費電力を削減することが重要である。

エッジコンピューティングとセマンティックセグメンテーションに関連する研究は、エッジコンピューティングにおける消費電力効率の課題に対処するために、タスクのオフロード場所を最適化することを強調している。セマンティックセグメンテーションでは、オフロードパラメータの最適化が性能効率を向上させる大きな機会となる。セマンティックセグメンテーションの性能は、ユーザーの視点からの画像解像度やサーバーの視点からのニューラルネットワークモデルの複雑さなど、さまざまなパラメータによって大きく影響を受ける。具体的には、ユーザーの視点からの精度要件とサーバー側からの電力消費制約が密接に関連している。例えば、低解像度または圧縮されたフレームを処理することで電力消費を削減できるが、同時にタスクのセマンティック精度が損なわれる。したがって、タスクをオフロードする前に適切なパラメータを割り当てることは、消費電力効率を大幅に向上させることができる。

エッジコンピューティングのテストベッドで予備実験を行い、タスクの精度と電力消費に影響を与えるオフロードパラメータを特定する。これらの重要なパラメータを特定した後、パラメータを変更して電力消費とタスク精度に与える影響を観察する。ビデオ形式、ニューラルネットワークのサイズ、期待される精度などのパラメータを適切に割り当てることで、消費電力効率を大幅に向上させることができることが実験結果から示されている。したがって、エッジコンピューティングシステムがタスクに適切なパラメータを割り当てる方法を決定した後、システムはより低い電力消費で動作できる。しかし、タスクの複雑さのため、エッジコンピューティングでセマンティックセグメンテーションタスクをオフロードするために適切なパラメータを自動的に割り当てることは困難である。一般的なコンピューティングタスクをエッジサーバーに割り当てる場合とは異なり、セマンティックセグメンテーションタスクには、未知の効果がある複数の異なる設定可能なパラメータがある。

その結果、本研究では、エッジコンピューティングにおけるセマンティックセグメンテーションタスクのオフロードに関する消費電力効率を改善するための学習戦略を提案する。まず、インテリジェントエッジコンピューティングでセマンティックセグメンテーションタスクをオフロードする際の消費電力と異なる効果パラメータを見つける学習問題をマルチアームバンディット（MAB）問題として定式化する。次に、定式化された MAB 問題を解決するために、UCB-CoR アルゴリズムを提案する。ここで、CoR は「Clustering of Regions」を意味し、UCB アルゴリズムの探索における反復回数を減らすためのパラメータの事前パーティショニングを指す。UCB アルゴリズムを採用する動機は、その広く知られた理論的保証と、さまざまなアプリケーションでの実績のある性能である。UCB アルゴリズムは、漸近的な設定でほぼ最適な後悔境界を達成することが示されており、多数のアームが存在し、時間の限られたシナリオで魅力的な選択肢となる。。

（２）システムの設計

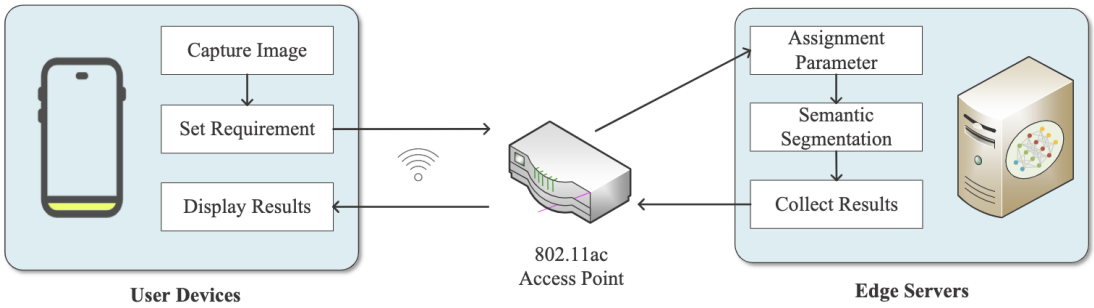


図１−１インテリジェントエッジコンピューティングフレームワーク

システムの概要として、ユーザーのデバイスがビデオフレーム（タスク）をキャプチャし、それをエッジサーバーに送信してセマンティックセグメンテーションを実行する。タスクの数は N で表し、ユーザーは提

供された参照値のセットから最小許容 mIoU 値を選ぶ。エッジサーバーがタスクを受け取ると、mIoU の要件と以前の決定に基づいて低電力のパラメータを選択する。このシステムはラウンドごとに動作し、各ラウンドで適切なパラメータを割り当てる。

エッジコンピューティング環境においてセマンティックセグメンテーションのタスクをオフロードする際、消費電力と mIoU に影響を与える 3 つの主要なパラメータがある。これらのパラメータは、ビデオ解像度、ニューラルネットワークの深さ、および圧縮率である。しかし、これらのパラメータ間の関係は未知であり、非常に複雑である。目標は、ユーザーが定義した最小 mIoU 制約を満たしながら消費電力を最小限に抑えることである。エッジサーバーの消費電力は主にタスク処理中に発生する。タスクのパラメータが割り当てられると、それに応じてタスクが処理され始める。タスクが完了すると、処理されたデータはワイヤレスネットワークを介してユーザーに送信される。計算に伴う消費電力はエッジサーバーによって直接提供され、各ラウンドで選択されたパラメータに依存する。

データ伝送の消費電力は伝送時間と伝送電力によって決まる。タスクの伝送時間はタスクのサイズと伝送速度によって決まる。伝送速度は、チャンネルの帯域幅、平均受信信号電力、およびノイズ電力から導出される。チャンネルの帯域幅が広いほど、伝送速度は速くなり、伝送にかかる時間が短くなる。伝送にかかる時間が短くなると、伝送に必要な消費電力も減少する。このシステムの目標は、エッジサーバーがタスクを効率的に処理し、ユーザーの mIoU 要件を満たしながら消費電力を最小限に抑えることである。このために、適切なビデオ解像度、ニューラルネットワークの深さ、および圧縮率を選択する必要がある。

（３）問題の定式化

エッジサーバーの消費電力は主にタスク処理中に発生する。タスクパラメータが割り当てられると、そのパラメータに基づいてタスクが適応され、処理が開始される。タスクが完了すると、処理されたデータがワイヤレスネットワークを介してユーザーに送信される。計算に伴う消費電力は、エッジサーバーによって直接測定され、選択されたパラメータに応じて変動する。データ伝送に関しては、消費電力は伝送時間と伝送電力によって決まる。タスクの伝送時間はタスクのサイズと伝送速度によって決定される。伝送速度は Shannon-Hartley の定理を用いて計算され、チャンネル帯域幅、平均受信信号電力、およびノイズ電力に依存する。システム全体の消費電力は、タスク処理中のエッジサーバーによる消費電力とデータ伝送中の消費電力の合計として表される。

セマンティックセグメンテーションの評価において、mIoU（平均交差結合）は非常に重要な指標である。ユーザーがタスクをアップロードする前に、最小許容 mIoU 値を指定する。私たちの目標は、ユーザーの mIoU 要件を満たしながら、システム全体の消費電力を最小限に抑えることである。この問題を解決するために、私たちは上限信頼境界（UCB）アルゴリズムを使用する。このアルゴリズムは、最適なパラメータを選択することで、時間の経過とともに総消費電力を効率的に最小化することを目指す。UCB アルゴリズムは、探索と活用のバランスを調整し、環境に関する知識が蓄積されるにつれて、最適なパラメータの探索に集中する。このアルゴリズムを使用することで、セマンティックセグメンテーションタスクの消費電力効率を大幅に改善することが可能になる。

（４）実験結果

実験では、エッジサーバーに事前に学習された 3 つのセマンティックセグメンテーションモデルを配置する。これらのモデルは VOC 2012 データセットを使用する。モデルの推論プロセスを加速するために、TensorRT を基盤の最適化エンジンとして使用し、モデルを ONNX 形式に変換してシームレスな互換性を確保する。動画の圧縮率は 100 と 20 の 2 種類、動画の解像度セットは 128、256、320、480、512 の 5 種類、ニューラルネットワークの深さセットは 18、34、50 の 3 種類を使用し、これらの組み合わせをセクション III で得られた測定値に基づいて設定する。H.264 動画圧縮標準を使用し、すべてのタスクを 30 フレーム/秒で処理することを仮定し、各タスクの処理時間を 1 秒と設定する。

これにより、3 つのパラメータセットの組み合わせから 30 種類の異なるパラメータが生じる。エッジサーバーが類似した環境で動作することを仮定し、各パラメータを 2 つのエッジサーバーに配置し、合計で 60 のパラメータ（アーム）を作成する。図 1-2 には、アルゴリズムのイテレーション結果と異なる閾値における UCB アルゴリズムの正解率が示されている。ユーザーごとの平均 SNR を 30 dB とし、チャネル帯域幅を 20 MHz に設定する。ユーザーの mIoU 要件を以下のように分類する：mIoU 要件なし、$36 < \text{mIoU} \leq 46$、$46 < \text{mIoU} \leq 52$、$52 < \text{mIoU} \leq 57$。これに応じてパラメータも 3 つのカテゴリに分け、各カテゴリに 20 のパラメータを設定する。IEEE 802.11ac 無線チャネルモデルを使用してシステムを構築する。

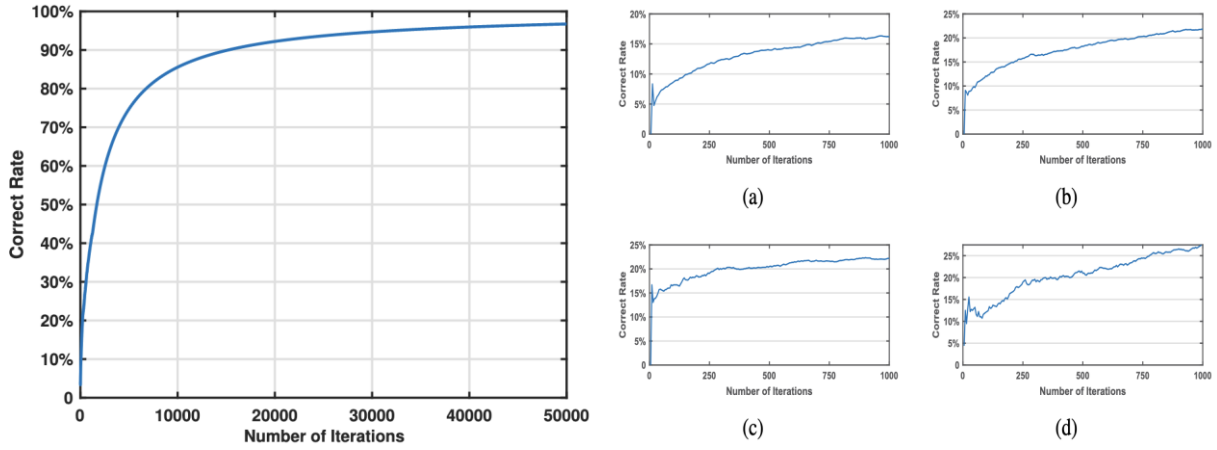


図１−２パラメーターに対するアルゴリズムのイテレーション結果および異なる閾値における UCB アルゴリズムの正解率

アルゴリズムの収束を評価するために、60 のパラメータで 50,000 回のイテレーションを行った結果、最適な構成パラメータの選択率が上昇し、40,000 回のイテレーション後に収束する。イテレーション数が無限に近づくにつれ、アルゴリズムが正しいパラメータを選択する確率は 100%に限りなく近づくことがわかる。

次に、ユーザーの mIoU 要件を満たすために、3 つの mIoU 閾値を設定し、60 の構成パラメータを分類する。イテレーション数を 1,000 に設定し、結果を図 1-2 右に示す。図 1-2 右(a)では、mIoU 要件なしで 60 のパラメータ全てに対して 1,000 回のイテレーションを行う。図 1-2 右(b)-(d)では、特定の mIoU 要件を満たすパラメータをイテレーションする（36 < mIoU ≤ 46、46 < mIoU ≤ 52、および 52 < mIoU ≤ 57）。アルゴリズムのイテレーション開始時には大きな変動が見られ、その後ゆっくりと上昇する。これは、探索段階で正しいパラメータ（最も低い電力消費量）を選択するためである。イテレーション数が総消費電力に与える影響を分析するために、構成パラメータ数を固定し、イテレーション数を 1,000 に設定して 100 イテレーションごとに総消費電力を記録する。アルゴリズムを 10 回実行し、消費電力データを 10 セット取得して、平均、最大、最小値を計算する（図 1-3）。

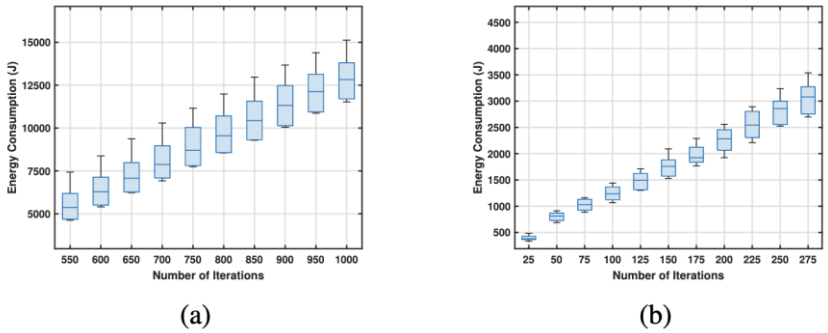
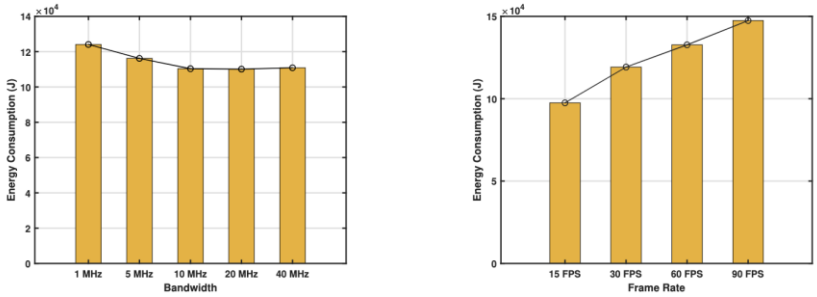


図１−３異なるイテレーションにおける消費電力値の変動

次に、UCB-CoR アルゴリズムを使用して、システムの消費電力に対するチャネル帯域幅とビデオフレームレートの影響を評価する。5 つのチャネル帯域幅（1MHz、5MHz、10MHz、20MHz、40MHz）で 10,000 回のイテレーションを行った結果、チャネル帯域幅が増加するとシステムの消費電力が減少するものの、一定の限界があることが示される（図 1-4a)。次に、4 つのビデオフレームレート（15FPS、30FPS、60FPS、90FPS）で 10,000 回のイテレーションを行った結果、ビデオフレームレートが増加するにつれてシステムの消費電力も増加することが示される（図 1-4b)。

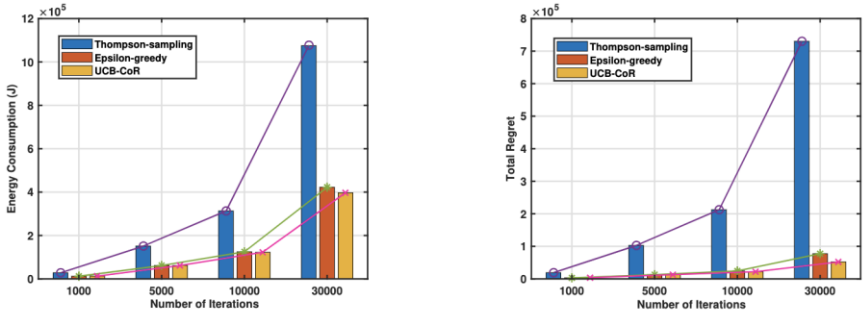


(a) (b)

図１－４異なるイテレーションにおける消費電力値の変動

UCB-CoR アルゴリズムの性能を他の一般的なアルゴリズム（トンプソンサンプリング、イプシロングリーディアルゴリズム）と比較する。イテレーション数が同じ場合の総消費電力と総後悔値を評価した。初期段階では、イプシロングリーディアルゴリズムと UCB-CoR アルゴリズムの性能はほぼ同じであったが、イテレーション数が増えるにつれて UCB-CoR アルゴリズムの優位性が明らかになった（図１－５a）。総後悔値の比較結果も同様の傾向を示している（図１－５b）。最後に、UCB アルゴリズムと提案した UCB-CoR アルゴリズムを比較した。イテレーション数が増えるとともに、消費電力の差が縮小するが、UCB-CoR アルゴリズムの方が依然として優れていることが示された。

本研究では、エッジコンピューティングにおけるセマンティックセグメンテーションの消費電力効率を改善するためのパラメータ選択問題を、UCB-CoR アルゴリズムを使用して解決した。シミュレーション結果は、UCB-CoR アルゴリズムが他のアルゴリズムに比べて消費電力を効果的に最小化することを示した。将来的には、追加のパラメータを考慮し、エッジ学習を研究に組み込むことを計画している。



(a) (b)

図１－５総消費電力と後悔値の比較

2-2 エッジコンピューティングにおける大型言語モデルの生成推論：消費電力効率のアプローチ

（１）問題の提出

OpenAI の GPT シリーズなどの大型言語モデル（LLM）は、複数の言語にまたがるさまざまな自然言語処理（NLP）タスクにおいて強力な能力を示している。これらのモデルは、トレーニングおよび推論の両方において、非常に多くの計算資源とストレージリソースを消費する。そのため、デバイスや環境リソースの制約を考えると、ほとんどの LLM アプリケーションは通常、高性能なクラウドサーバーにホスティングされる。

インテリジェントエッジコンピューティングは、データソースに近い場所にコンピューティングサーバーを配置することで、ローカルデバイスでのリアルタイム推論を可能にする。この戦略により、クラウドへのデータ伝送の遅延を大幅に削減し、プライバシーも向上する。しかし、一般的なエッジ AI タスクとは異なり、LLM 推論タスクはより多くの計算資源とストレージのサポートを必要とする。エッジデバイスのバッテリー寿命が通常限られているため、タスク推論の消費電力を最小化することが、エッジデバイスの安定かつ信頼性の高い運用を確保するために重要である。

エッジコンピューティングに関する研究は、タスクのオフロード先を最適化してタスク処理の消費電力効

率を向上させることに焦点を当てている。通常、これらの方法は、異なるオフロードデバイスなど、1 つの構成のみを選択することに焦点を当てている。LLM 推論タスクでは、推論の消費電力は LLM のパラメータ数、入力テキストの長さなど、さまざまな要因によって影響を受ける。そのため、複数のオフロード構成の選択を最適化する方が、消費電力効率を向上させるためには良い選択肢となる。一方で、LLM 推論タスクのタスク要件を満たすことも重要である。タスク要件は、推論精度や推論時間、またはその組み合わせである場合がある。特に、ユーザーの視点からのタスク要件は、サーバーの視点からの消費電力要件と密接に関連している。例えば、低消費電力デバイスやパラメータ数の少ない LLM を使用することで消費電力を削減できるが、それによりタスクの推論時間が長くなったり、低精度になる可能性がある。そのため、複数のオフロード構成の選択を最適化しながら、各ユーザーのタスク要件を満たすことは難しい課題である。

この問題に対処するために、エッジコンピューティングのテストベッドで予備実験を行い、タスク要件と消費電力に対して重要なオフロード構成を特定する。これらの構成を変更することで、消費電力とタスク要件に与える影響を観察する。セクション III で詳述する実験結果は、適切な構成を選択することで消費電力効率を大幅に向上させ、タスク要件を確保できることを示している。しかし、エッジ環境での LLM 推論タスクのオフロードに適した構成を自動的に割り当てることは複雑な課題である。一般的なエッジコンピューティングタスクとは異なり、LLM 推論タスクには複数の構成があり、それらの組み合わせが消費電力とタスク要件に予測不可能な影響を与えるため、構成の選択はより複雑である。

本研究では、LLM 推論タスクをエッジ環境にオフロードする際の消費電力効率を改善するための新しいアプローチを提案する。この問題をマルチアームバンディット（MAB）問題としてモデル化することで、さまざまな構成がエッジコンピューティングで LLM タスクを処理する際の消費電力にどのように影響するかを理解する。この MAB 問題を解決するために、上限信頼境界（UCB）アルゴリズムを使用する。漸近的な設定では、UCB は最適な後悔境界に近い値を達成することが示されており、マルチアームおよび大規模ホライズンシナリオにおいて魅力的な選択肢となる。

（２）問題の定式化

私たちは、生成推論タスクをより消費電力効率的にするためにインテリジェントエッジコンピューティングを使用するシステムを提案する。このシステムでは、ユーザーが自身のデバイスからタスクをエッジサーバーにアップロードし、最低許容推論速度と最低許容 AGI 値を選ぶことができる。タスクの数は N で表し、各タスクには特定の最低許容推論速度と最低許容 AGI 値が設定される。これらのパラメータは、実験によって得られた参照値のセットから提供される。

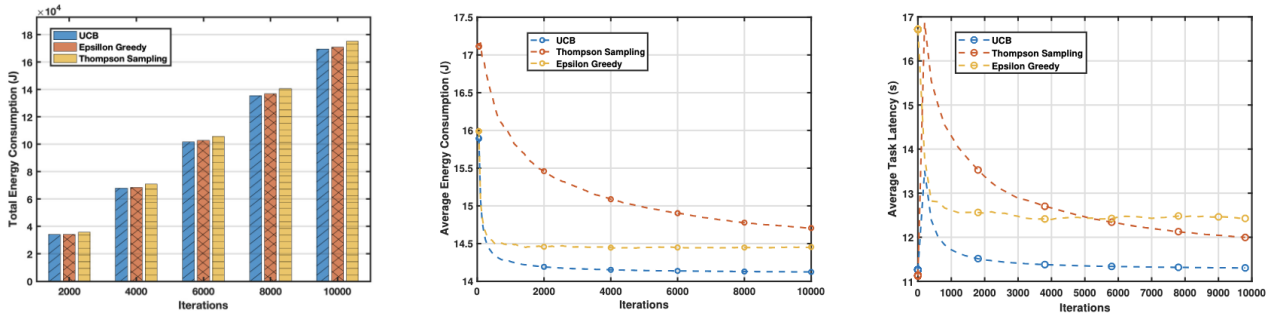
システムは離散的なラウンドで動作し、各ラウンドごとに特定の構成値を割り当てる。エッジデバイスの種類とモデルパラメータの数という 2 つの主要な構成が、タスクの消費電力、推論速度、精度に大きな影響を与える。しかし、これらの構成間の相互作用は非常に複雑であり、予測が難しいものだ。私たちの目標は、タスクの推論速度と精度の制約を満たしながら、タスクの消費電力を最小限に抑えることだ。そのため、エッジデバイスの種類を有限のセットから選び、モデルパラメータの数を有限のセットから選ぶ必要がある。

エッジサーバーにおける主な消費電力源は、タスクの生成推論だ。タスクの構成値が選択されると、エッジサーバーはこの構成に基づいて調整し、生成推論を開始する。タスクが完了すると、生成されたテキストはワイヤレスネットワークを介してユーザーに送信される。エッジサーバーは、生成推論プロセス中の消費電力を記録する。

データ伝送の消費電力は、伝送時間と伝送電力によって決まる。タスクの伝送時間は、タスクのサイズと伝送速度によって決定される。伝送速度は、チャンネルの帯域幅、受信信号の平均電力、ノイズ電力に依存する。チャンネルの帯域幅が広いほど、伝送速度は速くなり、伝送にかかる時間が短くなる。伝送にかかる時間が短くなると、伝送に必要な消費電力も減少する。このシステムの目標は、エッジサーバーがタスクを効率的に処理し、ユーザーの最低許容推論速度と AGI 値を満たしながら消費電力を最小限に抑えることだ。そのためには、適切なビデオ解像度、ニューラルネットワークの深さ、および圧縮率を選ぶ必要がある。これにより、エッジサーバーは最適な消費電力効率で動作し、タスクの要件を満たすことができる。

（３）実験結果

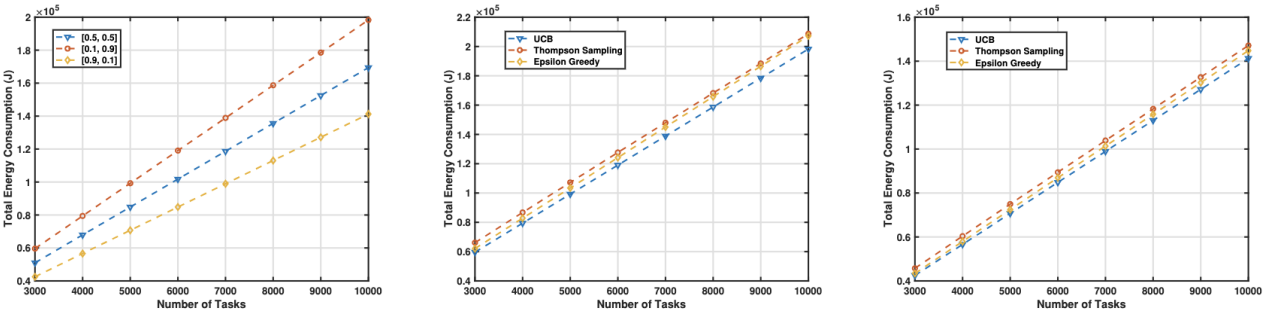
シミュレーションフレームワークのために、各ユーザーの平均信号対雑音比（SNR）を 50dB に設定し、チャネル帯域幅を 30MHz に設定する。さらに、ワイヤレスチャネルモデルには IEEE 802.11ac 標準を採用する。比較アルゴリズムとして、トンプソンサンプリングアルゴリズムとイプシロングリーディアルゴリズムの 2 つの一般的に使用されるマルチアームバンディット（MAB）問題のアルゴリズムを選ぶ。



(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