**Response to Editor**

This paper has been reviewed by three expert reviewers. All three reviewers question the motivation of the problem being studied in this paper. For example, Reviewer 1 writes "The current emoji prediction on smartphones is quite fast and efficient, so the motivation of using edge servers is not clear"; Reviewer 2 writes "It would be better for the authors to discuss potential drawbacks or limitations of the current inference-based approach"; Reviewer 3 writes "the need for distributed inference of mobile keyboard emoji prediction in the edge system is not well justified". I concur with the reviewers. This is certainly beyond the scope of a "Major Revision". I have been debating between "Reject" and "Revise and Resubmit as New", and decided to recommend the latter. Should the authors decide to revise and resubmit (as new).

**Comment E-1:** It is recommended that the author make a convincing case about the need for edge computing for this problem, by pointing out the limitations of the state of the art in this area.

**Response E-1:**

**Response to Reviewer 1**

Strengths:

1. Emoji prediction is an important topic. The idea of using edge servers for emoji prediction is interesting.

2. The theoretic analysis of the scheme is solid.

Weaknesses:

**Comment 1-1:** The motivation of the paper is not clear. The paper mentions a few related papers but does not introduce the current emoji prediction techniques used by Apps in Android/iOS. The current emoji prediction on smartphones is quite fast and efficient, so the motivation of using edge servers is not clear. It is better to introduce the current techniques and their performance first and then reconsider the motivation.

**Response 1-1:** Thanks for your suggestion.

通讯带宽换能量？

现在的emoji比较简单，未来的emoji比较复杂，不光着眼于现在，还有未来

**Comment 1-2:** The evaluations of the paper are not convincing. The paper does not show experimental results in real scenarios with real edge servers and smartphones. So the impact of using edge servers and model selection methods is unknown.

**Response 1-2:** Thanks for your comment. In fact, our work focuses on modeling and analysis of providing emoji prediction via edge servers. The deployment in real scenarios with edge servers and smartphones is not the main point of our work.

**Comment 1-3:** In the proposed method, the smartphone sends the user's message directly to edge servers in plaintext which may cause many security and privacy issues. Therefore, the proposed method is not practical.

**Response 1-3:** Thanks for your comment. The security and privacy issues of the smartphone sending the user’s message directly to edge servers in plaintext are not the main concern of our work. In fact, our proposed method can be well implemented through integrating the existing encryption or decryption technology on the smartphone for security and privacy. For example, the user's message can be encrypted as ciphertext and be sent to the selected edge server. In this way, the edge server first decrypts the ciphertext and then conducts inference. With the inference completed, the predicted emojis are sent back to the smartphone.

**Response to Reviewer 2**

Strengths:

+ They study the distributed inference and resource constraints, which differ from most existing works that focus on the statistical and neural network models.

+ They compare their G-LEAP approach with other baseline algorithms like Random, UCB, and Greedy.

+ They include many formula derivations in detail in the appendix for the algorithm design proof.

Weaknesses:

**Comment 2-1:** The simulation setting is not in detail, and it would be better if the authors could list the edge server model and what kind of smartphone is taken into the simulation test.

**Response 2-1:** Thanks for your comment. Firstly, the emoji prediction models along with the corresponding evaluated prediction accuracy values used in our simulations are listed in Table 3 of Section 7.1.1. Secondly, our simulations are conducted in a simulated environment running on our servers rather than a production-level environment with online smartphones. As the prediction model deployed on the smartphone is often chosen as a light-weight statistical model, *e.g.*, a Decision Tree model, most smartphones with either Android or iOS systems can adopt our G-LEAP scheme.

具体什么模型，size多大，神经网络的细节。具体什么手机（安卓/ios）

**Comment 2-2:** Not clear about the motivation or the problem definition of the emoji prediction, it would be better for the authors to give one or two examples or use cases for the emoji prediction

**Response 2-2:** Thanks for your comment.

Emoji prediction 最开始google的文献怎么介绍的 emoji prediction本身有没有意义

**Comment 2-3:** Lack the discussion or conclusion. It would be better for the authors to discuss potential drawbacks or limitations of the current inference-based approach.

**Response 2-3:** Thanks for your comment.

目前的inference没有考虑资源受限，是unconstrained

Related Work

**Comment 2-4:** What datasets are used in this paper for the mobile keyboard emoji prediction?

**Response 2-4:** Thanks for your comment. In our paper, we adopt a twitter emoji prediction dataset obtained from Kaggle [34] for our simulations.

论文里列出来 类似AAAI

**Comment 2-5:** Will this involve any privacy issue of inferring the user behavior?

**Response 2-5:** Thanks for your comment. The privacy issue is not the main point of our work. Nevertheless, for security and privacy, our proposed method can be well implemented through integrating the existing encryption or decryption technology on the smartphone. For example, the user's message can be encrypted as ciphertext and be sent to the selected edge server. In this way, the edge server first decrypts the ciphertext and then conducts inference. With the inference completed, the predicted emojis are sent back to the smartphone.

**Comment 2-6:** Shall the author give some justification for this dataset usage?

**Response 2-6:** Thanks for your comment. Here the adopted dataset [34] contains tweets gathered from Twitter. Twitter is a popular social platform where users often publish tweets containing emojis for expressing emotion. Thus, this dataset is suitable for training and testing emoji prediction models.

是代表性还是什么典型

**Comment 2-7:** The authors list four challenges when introducing the background of the emoji prediction, but the authors didn't mention or answer how they resolved these challenges.

**Response 2-7:** Thanks for your comment.

具体分析

**Comment 2-8:** When the author evaluates the performance of G-LEAP under different budget settings, not sure why the authors set fixed V = 10^4.

**Response 2-8:** Thanks for your comment. Recall that is a tunable positive constant which controls the relative importance of regret minimization to energy efficiency. In Section 7.5, we evaluate the effect of on the performance of G-LEAP. As shown in Fig. 5, we can observe that when the value of is larger than 500, the average regret converges to zero. Thus, we set for a good tradeoff between regret minimization (*i.e.*, online learning) and energy efficiency (*i.e.*, online control) during other simulations, including those evaluating the performance of G-LEAP under different budget settings.

具体分析设定值时，需要给个说明

**Comment 2-9:** The related work chapter primarily discusses the current state-of-art categories rather than introducing related work.

**Response 2-9:** Thanks for your comment.

Related work内容太少

**Response to Reviewer 3**

This paper studies the model selection problem with constrained energy consumption on smartphone for mobile emoji prediction. The target scenario is that a smartphone needs to select a subset of models available at an edge server for improving emoji prediction under the long-term energy constraint. The authors formulate the problem as a constrained combinatorial multi-armed bandit problem and introduce a Green Learning Aided emoji Prediction (G-LEAP) algorithm. The proposed algorithm is evaluated via a combination of theoretical analysis and simulation studies.

The paper has the following strengths in my opinion.

+ The proposed algorithm for model selection appears to be technically sound under the assumed system model.

+ The proposed solution has good technical depth.

+ The proposed algorithm is evaluated via a combination of theoretical analysis and simulation studies.

+ The paper is relatively well written and easy to follow.

+ The current version offers adequate new contributions over the conference version.

Meanwhile, I have the following concerns about the paper.

**Comment 3-1:** The problem being addressed is not well motivated. It is claimed in the Introduction that hosting multiple emoji prediction models on a smartphone locally incurs prohibitive energy consumption, but the paper does not offer any evidence to support the claim. In other words, the need for distributed inference of mobile keyboard emoji prediction in the edge system is not well justified. Since distributed inference introduces a set of new challenges, the problem tackled in this paper seems a bit contrived without strong motivation.

**Response 3-1:** Thanks for your comment.

需要系统的相关论文energy consumption

**Comment 3-2:** The Related Work section does not serve its purpose well. The paper mainly discusses prior works on the training of emoji prediction model, which leaves the impression that the problem tackled in this paper is entirely new. However, model selection via multi-armed bandit has been studied in the literature. A quick Google search found the following prior works.

1. A Multi-Armed Bandit Model Selection for Cold-Start User Recommendation, UMAP 2017

2. Model Selection for Contextual Bandit, NeurIPS 2019

3. Model Selection in Contextual Stochastic Bandit Problems, NeurIPS 2020

It is not clear whether any of the prior multi-armed bandit-based solution can be applied the solve the target problem, or if the problem being studied in this problem introduces any new challenges. The authors should cite and discuss prior works that use similar mechanisms and articulate why none of the existing solutions can be applied here.

**Response 3-2:** Thanks for your comment.

model selection via multi-armed bandit： 资源受限 constrained inference部分

以前的bandit算法可以用吗

**Comment 3-3:** The prediction accuracy for the same model across different rounds are assumed to be *i.i.d*. Why?

**Response 3-3:** Thanks for your comment. For a specific user, it is reasonable to assume that the user’s inputs follow a stationary distribution. In other words, the user’s inputs can be seen as samples drawn from the user’s corpus according to his/her language input habits. For a single user, the same model produces stationary prediction results across different rounds. Thus, the indicator random variables indicating the prediction correctness of the same model for a single user are *i.i.d.*, and the prediction accuracy is the expected value of the indicator random variables.

**Comment 3-4:** The end-to-end latency is model as the sum of the latencies of selected models, which needs be justified. A more reasonable model is that a user submits his typed text to an edge server, and the edge makes prediction using the selected models, possibly in parallel, and returns the prediction results. The model of end-to-end latency needs to be made more realistic.

**Response 3-4:** Thanks for your comment.

**Comment 3-5:** Similarly, since the smartphone only needs to send the input to the edge server once, the energy consumption cannot be modeled as the sum of the energy consumption of different models either.

**Response 3-5:** Thanks for your comment. In our work, each edge server deploys only one pretrained prediction model. Hence, during each round, the smartphone needs to send the input to the models on the selected edge servers or conduct local prediction. In this way, the total energy consumption is modeled as the sum of the energy consumption between the smartphone and the selected edge servers or on the smartphone for local prediction.

**Comment 3-6:** Simulation setting only considers 4 available models and the user can select 3, which somehow diminishes the practical value of the proposed solution. What is the cost of storing and running these four models locally? It would be better to show how the proposed mechanism perform over a wider range of these parameters.

**Response 3-6:** Thanks for your comment. We have conducted extensive simulations with more models over a wider range of the parameters. Specifically, the details are shown as follows.

* First, note that in our simulations we consider 5 available models in total, including one simple model on the smartphone and 4 models on edge servers. Considering the practical usage of G-LEAP, we have conducted extensive simulations with more available models and the results are shown in Section 7 of our revision.
* Second, the cost of deploying all the models locally on the smartphone consists of storage overhead, intensive computation and prohibitive energy consumption. This is why we leverage the power of edge computing in the design of G-LEAP.
* Third, Section 7.7 and Section 7.8 evaluate the performance of G-LEAP with different number of available models and different number of selected models in each round respectively.

需要更多的设定，比如10个里面选3个

Store cost以及running cost具体是什么

参数：10个里面选1个 或者选3个 变化一下

**Comment 3-7:** The end-to-end latency and energy consumption of these four models (Table 4) seem to be set arbitrarily with no justification. While this setting may help demonstrate the performance of the proposed mechanism, an unrealistic setting does not help reveal the true performance of the proposed solution in reality and again raises the question whether the problem studied in this paper is a real problem or not.

**Response 3-7:** Thanks for your comment. To reveal the performance of our proposed solution in real world, we conducted experiments using a smartphone and several servers. Particularly, the measurement details are shown as follows.

* First, for the end-to-end latency, we measure the model inference time by conducting model inference over a test dataset for times and take the average values.
* Second, we measure the energy consumption caused by selecting each emoji prediction model by adopting A/B tests on the smartphone. Specifically, we measure the energy consumption over 1 hour of the smartphone running a keyboard application with and without selecting each model for a number of times. The energy consumption of selecting each model can be estimated by calculating the difference of two energy consumption values and divide it by the number of repetition times.

需要实际的检测，justification，low median high组合一下

设置不合理，让人质疑在实际中部署的可能性