From Cloud to Edge: Rethinking Generative AI for Low-Resource Design Challenges

Sai Krishna Revanth Vuruma, Ashley Margetts, Jianhai Su, Faez Ahmed, Biplav Srivastava

UNIVERSITY OF South Carolina


01.

Motivation


INTRODUCTION

- Generative AI has been shown to revolutionize design processes and optimize functional performance across diverse domains ranging from fashion to medicine.
- However, due to its heavy demand on resources, it is usually trained on large computing infrastructure and often made available as a cloud-based service.

SITUATIONS WE SEEK TO ADDRESS

- Designing for users in rural, disadvantaged regions, or special use situations
- Repairing equipments in the field, especially
  - remote areas with limited resources or
  - emergency situations

THE NEED FOR TINYML

- Unique challenges in applications for remote areas and limited resources could greatly benefit from compact, offline AI models.
- Why?
  - These regions frequently face challenges such as limited internet connectivity, limited memory and computational constraints.
  - Engineering design problems in remote areas often possess unique contextual nuances - from local material constraints to specific environmental considerations and product usage patterns.

CURRENT MODELS

LLMs

- Remarkably capable in tasks like natural language generation & understanding.
- Limited however to few modalities (text being the most common).

Diffusion Models

- Can generate high-quality images from text prompts.
- Require significant prompt engineering for optimal results.

CURRENT MODELS

VLMs

- Responses can be modified through human-like conversation.
- Performance is limited by prompt engineering and supporting models.

Multi-modal LLMs

- Can encode information across multiple modalities (text, image, audio, etc).
- Displayed great potential in conceptualizing design but lack precision (Picard et al. 2023)

02.

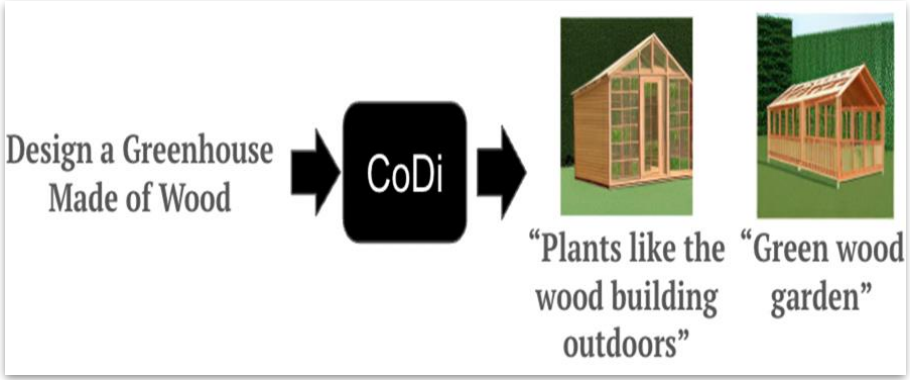
GenAI for Local Design

PREDICTING POTENTIAL IMPACTS

- Analysis of failure reports identified the most common pitfall in designing for the developing world as ”Lacking the contextual knowledge needed for significant impact” (Wood and Mattson 2016).
- This highlights gaining understanding of a community’s needs before embarking on design efforts.
- By predicting these social impacts, models can make better-informed decisions about an ideal design.

PRODUCT DESIGN & REPAIR

- Offline models provide an opportunity to customize a generated design to give a set of constraints.
- The current state of the model does not provide particularly helpful design outputs but demonstrates a potentially effective




03.

Discussion


POTENTIAL SOLUTIONS

- **Model Compression & Optimization:** Approaches such as Model Pruning, Quantization and Knowledge Distillation can help in making these large models smaller for inference and hence more accessible.
- **Edge Computing:** By processing data closer to the problem location, the need for high-speed internet connectivity can be reduced (Singh and Gill 2023).
- **Evaluation Metrics:** When applied to offline models, we must consider metrics that have low-resource evaluation capabilities and prioritize low cost performance in designs.

CONCLUSION

- By leveraging TinyML concepts such as Edge Computing and optimized model architectures we can make ML models smaller and hence more accessible to a wider population.
- Offline ML models, trained on relevant local data, can better capture and respond to the nuances of local design problems, offering more tailored and effective design solutions.
- In addition, these features also enable TinyML models to provide increased user privacy due to data being processed locally, reduced power consumption and internet usage.

From Cloud to Edge: Rethinking Generative AI for Low-Resource Design Challenges

Sai Krishna Revanth Vuruma[1], **Ashley Margetts**[2], **Jianhai Su**[3], **Faez Ahmed**[2], **Biplav Srivastava**[3]

[1] Department of Computer Science and Engineering, University of South Carolina, Columbia, South Carolina, USA
[2] Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[3] AI Institute, University of South Carolina, Columbia, South Carolina, USA
svuruma@email.sc.edu, amargett@mit.edu, suj@email.sc.edu, faez@mit.edu, biplav.s@sc.edu

Abstract

Generative Artificial Intelligence (AI) has shown tremendous prospects in all aspects of technology, including design. However, due to its heavy demand on resources, it is usually trained on large computing infrastructure and often made available as a cloud-based service. In this position paper, we consider the potential, challenges, and promising approaches for generative AI for design on the edge, i.e., in resource-constrained settings where memory, compute, energy (battery) and network connectivity may be limited.

Adapting generative AI for such settings involves overcoming significant hurdles, primarily in how to streamline complex models to function efficiently in low-resource environments. This necessitates innovative approaches in model compression, efficient algorithmic design, and perhaps even leveraging edge computing. The objective is to harness the power of generative AI in creating bespoke solutions for design problems, such as medical interventions, farm equipment maintenance, and educational material design, tailored to the unique constraints and needs of remote areas. These efforts could democratize access to advanced technology and foster sustainable development, ensuring universal accessibility and environmental consideration of AI-driven design benefits.

Introduction

In the rapidly evolving landscape of engineering design, the integration of Artificial Intelligence (AI) has catalyzed a transformative shift. Generative AI has been shown to revolutionize design processes (Regenwetter, Nobari, and Ahmed 2022) and optimize functional performance across diverse domains ranging from fashion (Sbai et al. 2018) to medicine (Walters and Murcko 2020). However, a crucial aspect often overlooked in this digital renaissance is the need for AI accessibility in remote or resource-constrained environments — a realm where tiny, offline machine learning (ML) models may be needed.

The need for tiny AI models in remote areas exists for two main reasons. Firstly, these regions frequently face challenges such as limited internet connectivity and inadequate computational resources. In such scenarios, cloud-dependent large AI models are impractical. Tiny ML models tailored for offline use can overcome these barriers, bringing the power of advanced design tools to isolated communities. This democratization not only fuels local innovation but also ensures that cutting-edge AI-assisted design solutions are not the exclusive domain of well-resourced urban centers.

Unique challenges in applications such as medicine, agriculture, and education in remote areas could greatly benefit from compact, offline AI models. In medical intervention, AI-driven design tools could assist in creating medical devices, such as a ventilator, tailored for limited-resource settings, focusing on affordability and ease of use while ensuring high efficacy. For the repair of farm equipment such as irrigation pumps, AI can play a pivotal role in predicting equipment failures or guiding repairs, adapted to the specific machinery and agricultural practices of different regions. In the context of school supplies, AI models could help design educational materials like toys (Jain 2023) that are not only engaging and culturally relevant but also accessible to students with limited resources.

Secondly, engineering design problems in remote areas often possess unique contextual nuances - from local material constraints to specific environmental considerations and product usage patterns. Offline ML models, trained on relevant local data, can better capture and respond to these nuances, offering more tailored and effective design solutions in spaces such as reliable sewage repair or low-cost greenhouse design.

Moreover, these small AI models align with the principles of sustainable technology deployment. By reducing reliance on large data centers and continuous internet connectivity, they contribute to lower energy consumption and a smaller carbon footprint, which is particularly crucial in ecologically sensitive remote areas (Prakash et al. 2023).

The challenge lies in distilling the complexity of generative models into lightweight versions without significant loss of functionality and performance. This requires innovative approaches in model compression, efficient algorithm design, and perhaps the use of edge computing architectures.

Review of Current Models

Generative AI techniques revolutionize the way machines can learn and replicate human behavior. They use Machine Learning (ML) techniques to create new data that is realistic

READ OUR PAPER!

AI4Society

THANK YOU ALL

Contact Information

Sai Vuruma
svuruma@email.sc.edu

LEARN MORE HERE!

AI4Society