I am interested in developing Machine Learning (ML) algorithms that optimize for data, model, and system efficiency; and how these methods solve real-world problems in Natural Language Processing (NLP) and Computer Vision (CV). As Artificial Intelligence becomes increasingly data and computationally driven, the importance of learning data representations with efficiently optimized methods cannot be overemphasized. Also, the rather steep increase in the size and costs of ML models in recent years means that participation in the development of the most prominent models is limited to a small number of teams within resource-rich institutions. The limited resources that have been available to me as an independent researcher have made me very conscious and calculative when designing experiments. During my PhD at CMU, I would like to focus on designing methods that enable us to build efficient models and systems with less computation and data resources, thereby lowering the barrier to shaping the ML technologies we create. Below, I discuss my research experiences and intended future directions.

**Data Efficiency:** The most significant constraints to developing and deploying real-world ML systems in the Global South are extremely low-resourced data and compute constraints. In my first research project, I developed a sign-to-text&speech system to reduce the communication barrier between the hearing-impaired community and the mainstream populace in sub-Saharan Africa. Prior to this work, there was no available dataset for sign languages in Africa. To this end, I created the first ever large-scale and publicly available sign language dataset in Africa (with the help of a TV sign language broadcaster and two special education schools in Nigeria). After developing several baseline models using image classification and object detection (with You-Only-Look-Once and single-shot-detector) architectures, I deployed the best-performing model (accuracy up to 98%) for real-time conversion of sign words/phrases to text and speech. The work was published at IJCAI-22 after a spotlight presentation at the 2021 NeurIPS ML4D workshop. However, while the work garnered awards and media fanfare for solving a critical-yet-overlooked societal problem, the pains of creating a pioneer dataset with my meager resources got me thinking about how to build data-efficient algorithms for low-resourced scenarios even beyond NLP. And in cases where some form of data is available, how do we approach the problems that arise when evaluating in out-of-domain settings or class imbalance which is prevalent in many CV tasks? I am interested in exploring optimal transport and geometric data manipulations to solve this family of problems.

**Efficiency in algorithms and methods:** Fuelled by the desire to enable the collaborative development of updatable ML models, I started looking for answers through Federated Learning (FL). I realized that the most significant bottlenecks in FL are communication and computation efficiency, including how "merging" disparate updates from different workers through federated averaging heavily degrades model performance due to the heterogeneous nature of the data on which the workers are trained. These bottlenecks spurred my research in exploring ways to utilize efficient evolutionary algorithms as second-order methods for FL. Leveraging particle swarm optimization, I successfully decreased local computational costs in federated training. I have also explored several other ways to cut costs with FL, which has resulted in numerous failures. For example, I tried implementing decentralized mutable torrents for faster (and cheaper) communication among clients in decentralized FL and neural architectural search in a federated setting. The latter led to another set of problems; I had to find novel ways to combine local architectures globally without using stochastic weight averaging. This inspired me to research and design more efficient optimization methods beyond the traditional gradient descent approach, and I became interested in empirically understanding the behavior of neural networks; specifically, how they are learned and what they learn. To this effect, I plan on investigating methods to speed up local and global convergence in a computationally cheap way while training neural networks.

**Efficiency in systems and applications:** In one of my recent works, I investigated ways to modify the architecture of Vision Transformers (ViTs) to make them deployable on mobile and edge devices. These devices represent platforms on which large models would be primarily deployed in real-world settings. They typically have much less computing power and memory bandwidth - requiring models with fewer parameters, smaller sizes, and low inference latency. After studying several transformer-based architectures designed for edge devices, I identified several common patterns in design choices. In particular, most of these architectures were MLP-based or sparse models; some even combined transformer and convolutional neural networks. Quantization, pruning, and other sparsification techniques are often applied to the trained models. Previous work has shown that there always exists a tradeoff between compression and model performance. I am very interested in designing novel methods that enable us to train, evaluate and deploy neural networks in

resource-constrained environments without sacrificing model performance.

Compression techniques like the Lottery Ticket Hypothesis have also been used to train sparser neural networks with similar test performance as their dense counterparts. However, recent work has also shown that these techniques are less robust to class imbalance; and sometimes result in less interpretable models that output less confident predictions. This is another exciting challenge I want to tackle; I want to develop sparser models that are also very robust to similar problems like the ones highlighted above.

**Growing professionally:** My bachelor's degree timeline was extended due to interruptions by Covid19 lockdowns and academic union strike actions. Fortunately, this gave me more time than a typical undergraduate to refine my research interests beyond what my immediate environment offers. I have always spent more time pursuing my interests than trying to score perfect grades in classes. I built the relevant research fundamentals by completing online coursework extensively in maths and ML. Then I started leveraging ML Collective - an open-access research lab where self-motivated researchers from diverse backgrounds utilize its resources to support their research - and successfully collaborated with researchers from top-tier institutions worldwide. Through various ML engineering internships, I have acquired relevant experience in how research translates into scalable real-world solutions. And recently, I wrote a grant proposal and won a $115k grant from the Algorand Foundation that enabled me to assemble a team of five ML/software engineers and data scientists to build an open-source platform for real-time social media opinion mining for digital assets. Asides from my work experiences, I have led several communities focused on peer-to-peer collaborative development in my campus and city, including Google Developer Students Club, Artificial Intelligence Plus Club, and DSN's City Chapter. I have personally taught and mentored hundreds of ML enthusiasts in my city and Western Africa at large. I aim to become a research professor exploring efficiency problems in ML while mentoring students to foster growth and collaboration in efficient ML.

**Moving forward:** The most significant path to making reasonable progress in my career goals is obtaining a doctoral degree. I am interested in much of CMU's current research on efficient ML. I relish the opportunity to work with Professors Emma Strubell, Beidi Chen, Albert Gu, Ameet Talwalkar, Virginia Smith. I have taken time to go through their work and, having perused their research websites, developed a profound sense of what working in their labs looks like. My devotion to growth and impactful work, coupled with my unique experiences, allow me to get the most out of the graduate program at CMU. The resources and breadth of ML research performed at CMU will provide me with the experience I need to pursue my passion in computer science.