International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Compression Based Multi-Modal Routing for Network Optimization

Navin Kumar Sehgal

Department of IT, IITM College of Engineering, Bahadurgarh, India

Abstract
The internet technology is expanding at a fast rate. With evolution of 6G network it is possible to transmit data faster enabling real time applications like Autonomous vehicles, remote healthcare systems, digital education and immersive learning systems. These applications are QoS sensitive and need low latency, low error rates and higher efficiency. Traditional algorithms rely on static routing methods making them incapable of handling dynamic, heterogenous and high-volume traffic environments. This paper proposes compression based multi modal framework integrating Mico LLMs and data compression methods. The framework is based on multimodal metrics like traffic pattern, congestion information, connection type, topology, jitter and round-trip time. These parameters help calculate network load and routing conditions for better speeds autoencoders are used to compress packet data while preserving traffic characteristics. A micro LLM is used to analyses traffic patterns in multi modal ways including data, topology and congestion and therefore helps in generating routing decisions for faster data transmission. Micro LLMs also learn the network parameters using feedback mechanism based on network performance. Based on a network metrics data set it was found that compressions reduce packet size by about 30% and also maintain network dynamics. These methods improve bandwidth, suggesting routing paths that avoid congestion for real time applications

Keywords: Multi-modal routing, compression-aware networking, micro-LLMs, autoencoder compression, QoS optimization, 6G networks.

Introduction

There is rapid evolution in networking technologies including 6G, cloud computing, edge computing
Th result is development of newer applications that work with high-speed data. As the data traffic speeds increase traffic management becomes even more essential as there is need for reliable, high throughput and low latency communication [2] The newer wireless networks running on 6G are focused on connecting trillions of devices, keeping delay rates to a minimum [1]. The capabilities are augmented with AI support applications in autonomous and immersive systems
The problem with above applications is that that there is sudden increase in speed requirements in these applications making their network operating environment very dynamic. This requires adaptable routing mechanisms that can adapt for different scenarios. Traditional algorithms like shortest path routing though efficient are static in nature and have a fixed set of network metrics for route calculation [15].
These algorithms fare poorly in dynamic application requirements. As an example, educations applications that need live streaming lectures and virtual labs need better routing for jitter free experience.

With advent of AI the Machine learning techniques are being used for traffic management. These analyses traffic patterns and adapt to changing network needs [3]. Reinforcement Learning is a good approach to learn from continuous interaction with network environment and improve with time. [12]. More recently graph-based network structures have been introduced to represent network topology and traffic relationships. They use neural models to predict routing and have shown to improve network performance [13]. Another methodology considers data traffic not as static but tries to learn intent which can be used in forming network policies [16]. LLMs have been explored to interpret network context information for routing decisions. [9].

But all above methods focus primarily on path selection without reducing traffic load. They sometimes fail to address network congestion issues caused by redundant traffic. The paper suggests reducing packet size before transmission can improve utilization of bandwidth. Further many of the above systems rely on single modality like latency or only topology without considering the heterogenous nature of modern networks. Today's network generates lot of multi modal information related to networks like text logs, topology structures and traffic patterns which when combined can provide a more comprehensive view of network state.

The paper also suggests leveraging the power of micro LLMS on routing devices for better routing decisions. The multimodal information can be used to fine tune these micro LLMs in a feedback loop that can help to predict better routes The above framework integrates deep auto encoders for compressions and micro LLMs for routing which use network information to construct a network context. Micro LLMs get finetuned in a feedback system that can improve prediction accuracy over time.

Literature Review

Networking research focus has shifted to improving Quality of Service (QoS) that can be achieved by optimizing network performance metrics. Many routing algorithms use metrics like round trip time for calculating shortest paths. Metrics like bandwidth, latency and packet loss were added for calculation. But these algorithms need much more computing when used in large networks. Also, these algorithms were static and ignored sudden change in network requirements. Cloud and wireless networks are dynamic in nature and need real time routing systems and require more adaptive and data driven routing methods.

Traffic flows can be engineered to optimize networks and use distributed network paths. Traditional traffic flow control relies on mathematical models or heuristic search. The resource allocation is such that network congestion is kept at a minimum level. These methods however require knowledge of current and future demand pattens for heterogeneous and large networks. Multi path routing and loading balancing techniques were incorporated to handle the problem to some extent but these have issues with packet ordering and an increase in network complexity.

According to Boutaba [3] Machine learning algorithms can be used for analyzing traffic patterns , predicting traffic behavior and also performing network anomaly detection, route optimization with congestion control . Supervised machine learning models are preferable for predicting network state based on the previous traffic data. The algorithms can change routing policies based on prediction of traffic flow and possible congestion. Unsupervised machine learning algorithms are more suitable for abnormal traffic, security threats and any network anomalies. Clustering algorithms can group traffic flows into different traffic classes and apply similar routing strategies in a group, Deep learning algorithms can capture long time traffic patterns. They can learn to organize traffic patterns into hierarchical representations based on priority and bandwidth requirements but this approach requires bigger data sets and higher computational needs.

In their paper Netravali and Alizadeh [17] state that reinforcement learning is another method for network optimization. Rewards and penalties can be applied based on feedback in the form of network performance which helps learn optimal actions. Exploration of different routing strategies is done to gradually learn policies which maximize rewards which is useful in dynamic network environment. Reinforcement learning for adaptive video streaming has been demonstrated for dynamic traffic flow adjustment Reinforcement learning algorithms can automatically discover complex policies not present in static routing. However, this method requires long training and exploration time which is not possible in real time network environment as it can degrade performance in short term.

Yang [13] states that physical structure of the network plays an important role in selecting transmission paths. For networks with topology information graph neural networks can be applied that can also add traffic related information into GNN. Traffic flow relationships can be learned by GNN models that can later predict routing paths more accurately. Feng and Shen [5] have created Graph Router which is a framework that uses this method for network reasoning based on GNN . GNNs allow combining local node information with global network architecture for graph-based reasoning . However, GNNs are computationally expensive in case of larger networks.

According to Manias and Chouman [16] data being transmitted can be considered not as static but suitable for interpretation of intent The intent can be specified in natural language and can be used in routing policies This makes it feasible to focus on overall network management instead of low-level details This method has been demonstrated to optimize routing decisions . However, this required sophisticated understanding of networks and routing policies

LLMs can convert network information to natural language for analysis and decision making as stated in paper by Hu [9]. Router Bench is a framework created by Kan [8] that evaluates LLM based routing .Such systems can analyse performance metrics, network logs and other textual data related to networks. Micro LLMs can be used for network analysis .

Zhang [14] stated that traditional network routing works at network layer using network layer protocols like IP addressing Cross layer information can helps gain better understanding of the traffic flow. This approach improves IOT networks which are more network dependent . Cross layer information has been demonstrated to be effective with edge devices , for example physical layer data speed along with application layer latency can help make better routing decisions.

In the paper by Baltrusaitis and Ahuja [2] muti modal learning techniques integrate heterogeneous data. Multi modal learning can combine different features for better decision accuracy for example, routing systems can combine congestion metrics, topology information, and service descriptions to make more informed routing choices. However, integrating heterogeneous data can if not done without correct context information can misalign network data.

Most algorithms studies above are single modal or capture only a part of the network context information. Compression although done at presentation layer level in traditional algorithms has not been practiced at packet level. Many routing algorithms suggested above cannot work in dynamic and real time environments

Neural network-based architecture mentioned in literature review lack continuous feedback learning mechanism to improve their decision making instead of one-time training. Overcoming these challenges requires an integrated approach to packet compression, multi modal analytics and intelligent routing. The proposed system integrates autoencoder for packet compression, multi modal analysis which allows relevant visible network congestion state diagram to be added to network information and micro LLM to enhance speed and reduce resource consumption.

Methodology

The proposed system suggests a compression based multi modal routing that is based on micro LLM for routing decisions. The system performs processing as a sequential pipeline along with feedback learning loop. It is suitable for fast and dynamic network environment like 6G, Cloud and edge systems. The system performs functions like packet compression with autoencoders, finding context from network metrics, Micro LLM for routing, continuous network monitoring and feedback loop for improvement of model. It performs network optimization using reduction in traffic load and improvement in routing efficiency.

The first layer of the system captures network traffic data which can be multimedia, IOT or interactive communication data. It also identifies latency, jitter, congestion and resource utilization metrics. The data can become larger in volume and dynamic. The next steps involve compression the data packets using autoencoder that is trained to output latent representation of packet data while preserving essential information. It also performs dimensionality reduction, removing redundant information thereby reducing bandwidth consumption, lesser congestion and improvement of network efficiency and optimizing network load. A decoder is used to reconstruct the data as per requirement.

The next step in the system is derivation of multi modal network context based on Round Trip Time (RTT), Network Congestion which may be represented as a diagram, Jitter information, traffic patterns, topology diagram and types of applications running can be part of cross layer information. The topology and congestion diagram along with the other network metrics are converted to natural language and added to network context in natural language. This is done using algorithms that have conversion rules.

The natural language context information derived is input to Micro LLM that is already finetuned for route calculation.

Therefore, selects the optimised path and also prioritises traffic. It proves better in handling dynamic conditions and handling multiple factors. The routing path selected by Micro LLM is used by the router for packet forwarding and manage traffic. Adjustment of network path happens dynamically based on real time traffic conditions leading to reduced packet delays, efficient use of packet resources. The system performs continuous monitoring of network metrics like latency, throughput, jitter.

This prevents performance degradation by identifying congestion points and assess network efficiency. Improvement of routing decisions happens by fine tuning model parameters, improving long term performance and evolution of routing policies. The system therefore combines the best of autoencoders, multi modal context interpretation and model-based reasoning as shown in Figure 1- System Architecture.

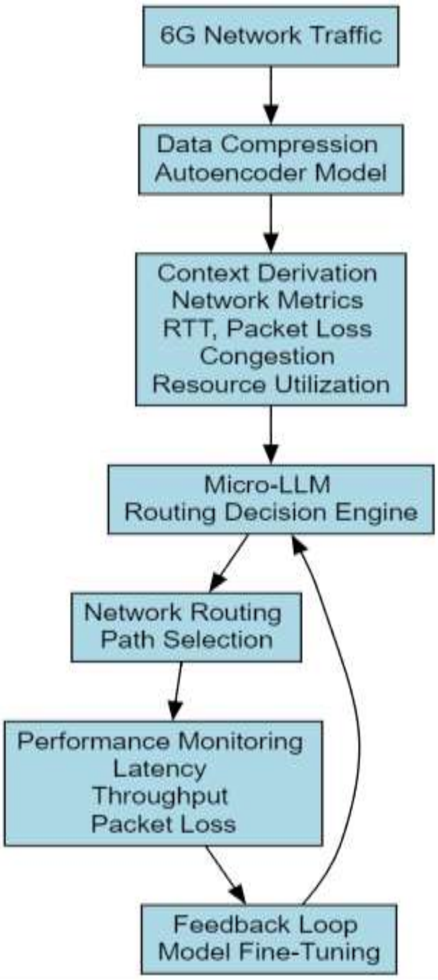

Figure 1 – System Architecture

Data Analysis for the above study was done using 6G network data downloaded from Kaggle. The dataset consists of network metrics like RTT, Latency, Jitter, Congestion indicators. Initial data exploratory analysis was done to analyse number of records, columns and any missing values. Data was cleaned for accuracy. missing values were filled with mean/median.

Duplicate and inconsistent records were removed and data was normalised and scaled. One hot encoding was used to convert traffic and connection type data to numerical values. Feature engineering on existing dataset was carried out by adding columns RTT, Jitter, Congestion level, Traffic movement, traffic rate, packet size as the original dataset consists of only columns – timestamp, bytes sent, bytes received and network anomaly.

Research Gap found based on study of previous research found use of single modality most, no compression at packet level, absence of network context and limited use of lightweight models.

System architecture was devised which could compress data, derive context from network metrics, use micro LLMs and use performance monitoring in dynamic environments and use feedback to improve network decisions Graphs were created on the dataset to find the role of the proposed system on the network optimisation.

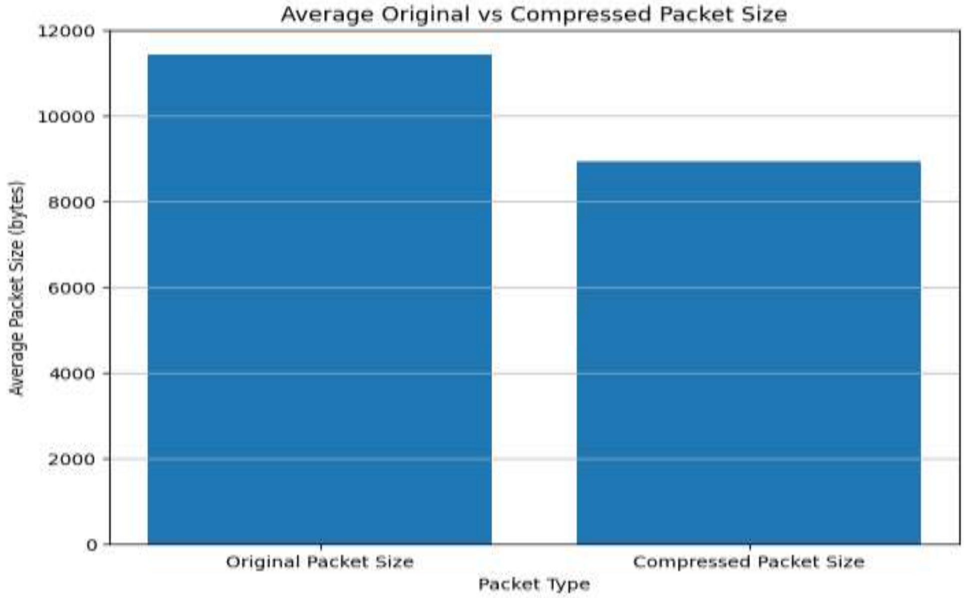

Figure 2 – Original vs Compressed packet size

Figure 2 – Original vs Compressed packet size compares the average packet size and compressed packet size. Figure 2 show the effectiveness of packet compression. It was observed that Average original packet size was 11,400 bytes and Average compressed packet size was 8,900 bytes. This also proved the usefulness of autoencoder in creating compact representations of data packets while maintaining data integrity. Reduced packet size effects network optimization due to less bandwidth usage, reduced delay, lower congestion.

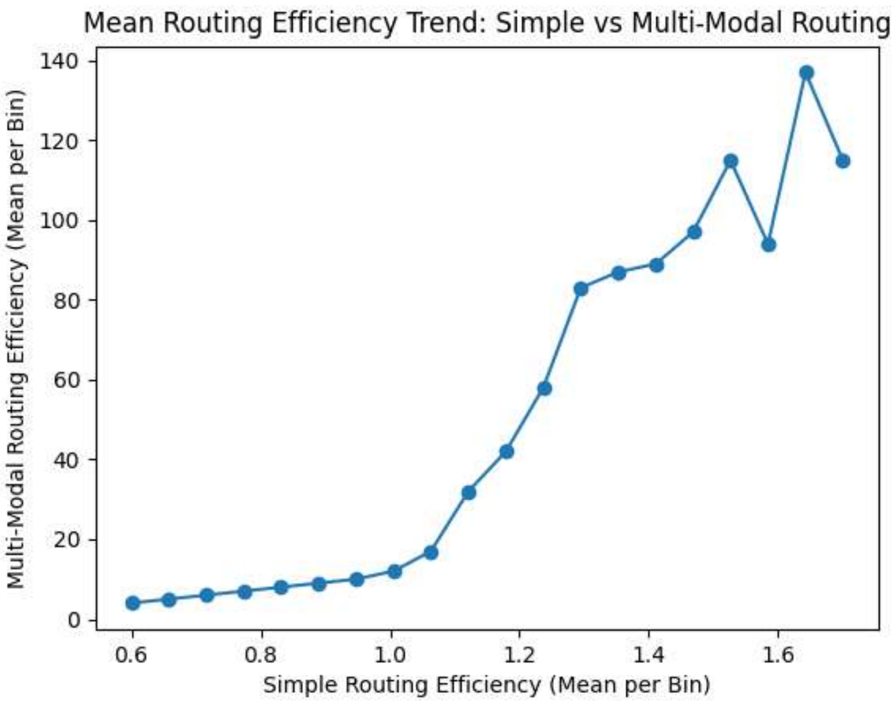

Figure 3 – Simple vs Multi-Modal Routing Efficiency

Figure 3- (Simple vs Multi-Modal Routing Efficiency) found routing efficiency calculated from RTT to increase faster for multi modal routing as compared to single mode routing. Routing efficiency is calculated from RTT, packet loss, congestion, network topology, traffic movement and only RTT in case of single modal form. For lower values of single mode frequency multi modal frequency grows slowly. But for higher values of single mode frequency multi modal frequency increases sharply. The figure points to nonlinear relationship between two types of routing efficiencies

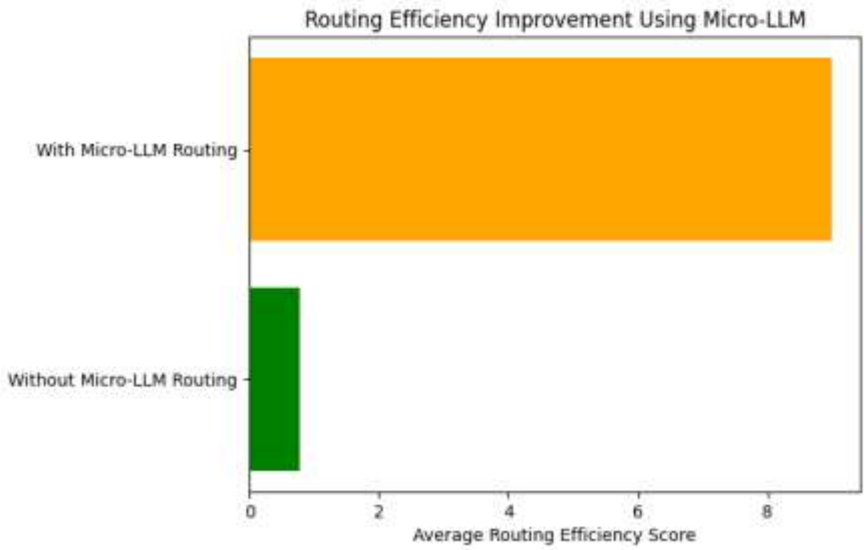

Figure 4 – Simple vs Micro-LLM Routing

Figure 4 (Simple vs Micro-LLM Routing) compares routing efficiency with Micro-LLM and without micro LLM usage. It shows improvement in network optimization as micro LLMs are context aware and suggest better decision paths. Routing efficiency values are computed from topology, congestion, traffic movement and resource utilisation and only RTT in case of simple routing efficiency. The efficiencies are averaged across all network samples. The formula used for simple routing efficiency is

$$\text{Efficiency Simple} = \left(1/ (\text{RTT} + 1)\right) + (1 + (\text{Packet Loss \_1})) \tag{1}$$

Were

RTT is round trip time

Packet loss is number of packets not reaching destination

$$\text{Efficiency\_MicroLLM}=(1 / (\text{RTT} + 1) + (1/ (\text{Packet Loss} + 1)) + \alpha \times \text{Topology} + \beta\times \text{Connection} + \gamma \times \text{Congestion} + \delta \times \text{Traffic Movement} + \in \times \text{Resource Utilization} \tag{2}$$

where RTT is round trip time

Topology is type of topology used converted to numeric code

Connection is connection type wired or wireless

Congestion is measured using delay in packet transmission

Traffic movement is based on number of bytes transmitted

Resource Utilization is amount of buffering used

The results demonstrate difference in routing efficiency which is very low without micro LLM (0.7 -1) and better with micro LLM (8-9) Adding multi modal context to micro LLM improves routing decisions better .it can interpret complex inputs and handle dynamic network environments.

Results and Discussion

Experiments carried out based on the dataset showed effectiveness of compression based multi model routing framework. It evaluated three key aspects including packet size reduction, routing efficiency with network context and routing using micro-LLM. Autoencoder compression reduced the packet size from 11.4 kb to 8.9 kb preserving all characteristics. The overall efficiency in whole transmission was observed to be 20-22% for current network. It shall lead to lower consumption of bandwidth, lower delays, better network scalability. The second component pf the system also proved to reduce latency and jitter. Routing based on network topology, connection type, traffic patterns and congestion levels formed a multi modal effect for improving network performance. The third graph of the experiment showed that routing using Micro-LLMs improved from 0.77 to 8.99. This showed that Micro-LLM can interpret complex network context. As the proposed architecture adds a feedback loop for continuous monitoring and improvement of the network. The feedback loop fine-tuned Micro-LLM to improve latency, throughput and jitter. The integration of all three systems provided several benefits like improved routing, adaptive traffic management and less congestion.

Conclusion

The study introduces a new approach in high traffic networks like cloud and 6G. It also introduces using autoencoders for packet compression. It focuses on multi modal approach rather than single modal approach for understanding networks. Further it proposes development of Micro-LLM on routing engines. The framework provides scalable system for next generation communication. It can benefit a lot of domains including Internet of Things (IOT), Edge computing systems, Cloud Systems Future directions include incorporating reinforcement learning in micro deep networks, deploying micro agents at edge nodes, incorporating security aware routing in real world 6G environments.

References

1. Akyildiz, I. F., Kak, A., & Nie, S. “6G and beyond: The future of wireless communications systems”. IEEE Access, 2020.
2. Baltrušaitis, T., Ahuja, C., & Morency, L. P. “Multimodal machine learning: A survey”. IEEE TPAMI, 2019.
3. Boutaba, R., et al. “Machine learning for networking: A survey”. IEEE Communications Surveys & Tutorials, 2018
4. Chiang, M., & Zhang, T. “Fog and IoT research opportunities”. IEEE IoT Journal, 2016

5. Feng, T., Shen, Y., & You, J. . “Graph Router: A graph-based router for LLM selections”. NeurIPS, 2024
6. Goodfellow, I., Bengio, Y., & Courville, A.. “Deep Learning”. MIT Press, 2016.
7. Hinton, G., & Salakhutdinov,. “Reducing dimensionality of data with neural networks”. Science, 2006
8. Hu, Q. J., et al. .” RouterBench: Benchmark for multi-LLM routing systems”, 2024.
9. Kan, K. B., et al. . “Mobile-LLaMA for network analysis in 5G networks”. IEEE Network, 2024
10. Kreutz, D., et al. . “Software-defined networking: A comprehensive survey”. Proceedings of the IEEE.
11. Letaief, K. B., et al. (2019). The roadmap to 6G. IEEE Communications Magazine.
12. Mao, H., et al. (2017). Neural adaptive video streaming with reinforcement learning. SIGCOMM.
13. S utton, R., & Barto, A. (2018). Reinforcement Learning: An Introduction.
14. Yang, J., et al. (2025). QoS-aware routing for edge computing networks.
15. [Zhang, K., et al. (2017). Delay-constrained offloading for mobile edge computing.
16. A. S. Tanenbaum and D. J. Wetherall,”Computer Networks”, Prentice Hall, 2011.
17. D. M. Manias, A. Chouman, and A. Shami, “Semantic routing for enhanced performance of LLM-assisted intent-based 5G core network management and orchestration,” IEEE GLOBECOM, 2024.
18. H. Mao, R. Netravali, and M. Alizadeh, “Neural adaptive video streaming with reinforcement learning,” ACM SIGCOMM, 2017.