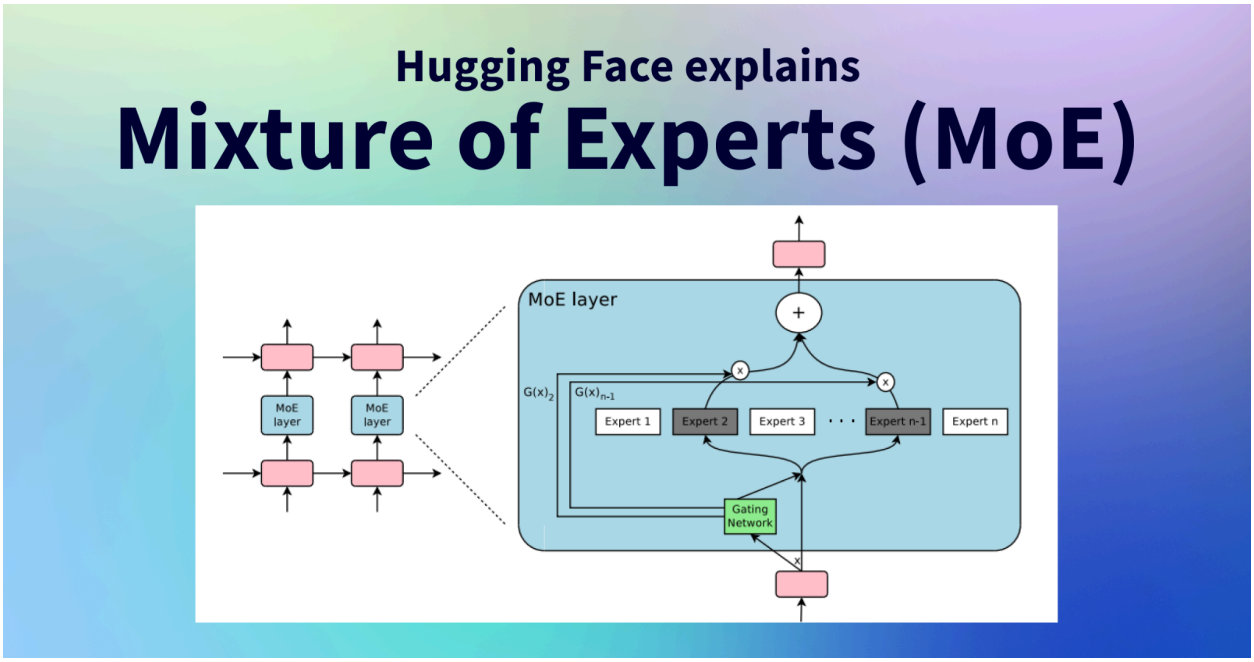
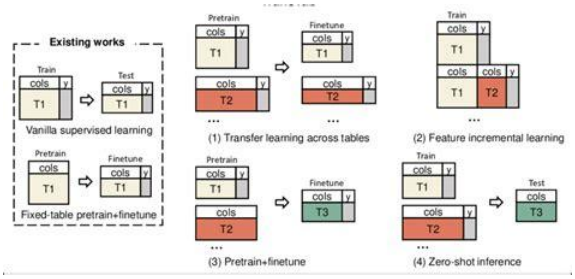
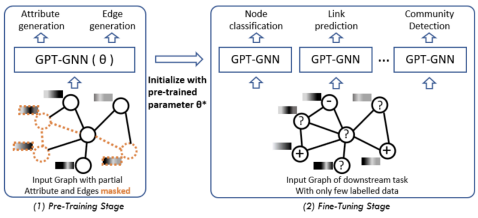


Introduction

Foundation models—large neural networks pre-trained on massive datasets—have revolutionized NLP and vision by enabling zero-shot transfer and light fine-tuning for new tasks. Early breakthroughs like BERT and GPT showed the power of pre-training on text; today, researchers are extending the same recipe to non-text data such as time-series, graphs, and tables. In this survey, we focus on TIME-MoE, a time-series foundation model, and explain why building a single, versatile network for sequence data can unlock faster, more accurate forecasting across domains.





Architectural Foundations

At the heart of TIME-MoE lie three key components:

1. **Mixture-of-Experts (MoE)**
   A collection of specialized subnetworks, or “experts,” each trained on different patterns. A lightweight router selects only a few experts per time-step, letting the model scale to billions of parameters without a corresponding explosion in compute.

2. **Unified Prediction Head**
   Instead of separate output layers for each forecasting horizon or downstream task, a single multi-task head simultaneously supports multiple objectives—e.g., predicting 1, 8, 32, and 64 steps ahead in one pass.

3. **Transformer Backbone (Dense vs. Sparse)**

    - **Dense Transformers** use full self-attention across all tokens, yielding strong performance but quadratic compute cost as sequence length grows.

    - **Sparse Transformers** (here, MoE) reduce that cost by routing tokens through only selected experts, enabling efficient scaling to very large model sizes.

Key Results

- **Zero-Shot Performance:** Out of the box, TIME-MoE reduces forecasting error by 20–30% compared to previous transformer-based time-series models—no extra fine-tuning needed.

- **Fast Fine-Tuning:** A single epoch of additional training brings performance on par with or exceeding dense transformers that require much longer training schedules.

- **Efficiency Gains:** Sparse routing makes training about 80% faster and inference about 40% faster than an equivalently sized dense model, dramatically cutting compute overhead.

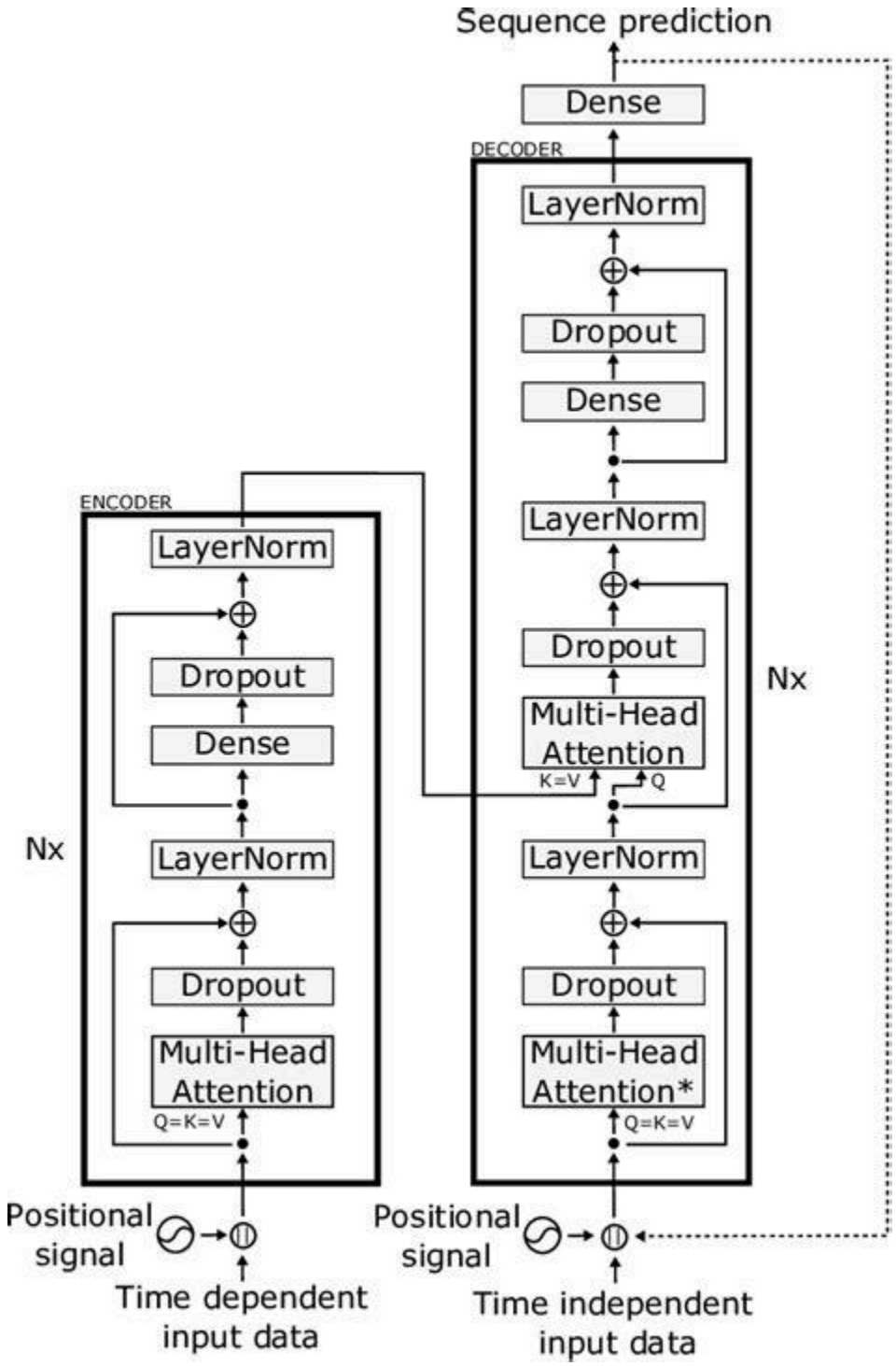
Placing TIME-MoE in Context

Foundation models are also emerging for other structured data:

- **Tabular Data:** Transformer-based models can impute missing cells, classify columns, and transfer knowledge between datasets with similar schemas.

- **Graphs:** Large graph neural networks pre-train on tasks like edge prediction and node masking; hybrid methods convert graph fragments into text so language models can help reason over structure.

TIME-MoE applies the same foundational recipe—pre-training on diverse data, sparse scaling, and unified heads—to time-series, demonstrating that one network can handle many forecasting tasks across domains.



Conclusion & Future Directions

By consolidating multiple forecasting tasks into a single, scalable model, TIME-MoE cuts development and compute costs while improving accuracy. Looking ahead, key challenges include:

1. **Expert Interpretability:** How can we understand and label the roles each expert subnetwork plays?

2. **Irregular Sampling:** Many real-world series have gaps or variable time steps—how best to incorporate them?

3. **Multimodal Fusion:** Can one foundation model seamlessly combine time-series with text annotations, images, or tabular metadata?

Simplifying architecture and data requirements, TIME-MoE paves the way for high-quality, adaptable forecasting in industries from energy and finance to healthcare and IoT.

References:

Forecast error comparison. Adapted from Ansari *et al.*, *Billion-Scale Time Series Foundation Models with Mixture of Experts*, ICML 2024.