More on Time-Series Foundation Models

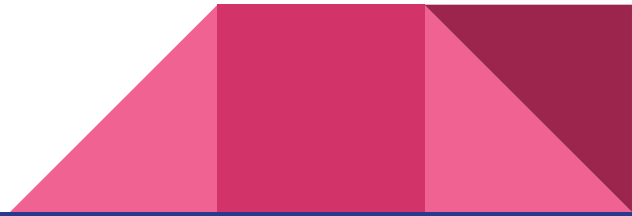
by Dustin Nguyen

Motivation

- Foundation models revolutionized NLP & vision
- Time-series data under-served
- Need one model for many forecasting tasks

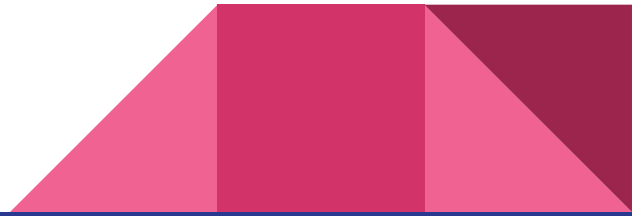
What Is TIME-MoE?

- Decoder-only transformer + MoE layers
- Billions of parameters, sparse routing
- Single head for multi-horizon forecasts



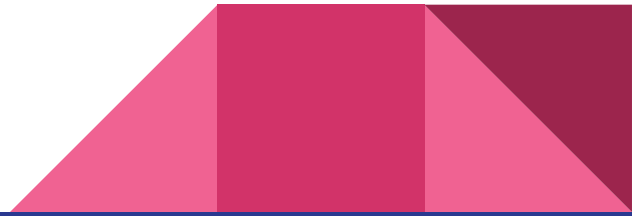
Mixture-of-Experts Explained

- Many small “expert” subnetworks
- Router selects top-k experts per token
- Scales parameter count without quadratic cost



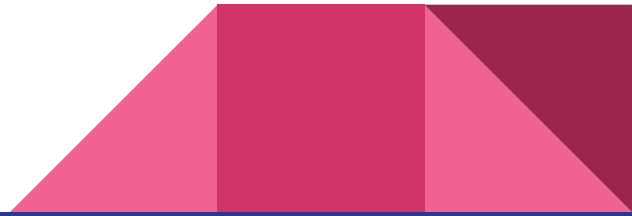
Pre-training

- 300 billion time points
- Nine domains: energy, finance, weather, healthcare…
- Mixed sampling resolutions



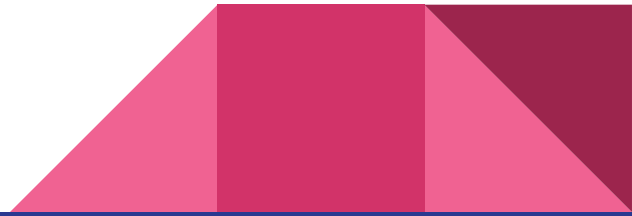
Multi-Horizon Head

- Predicts 1, 8, 32, 64 steps simultaneously
- Composite loss: Huber + expert-balance
- Shares knowledge across horizons



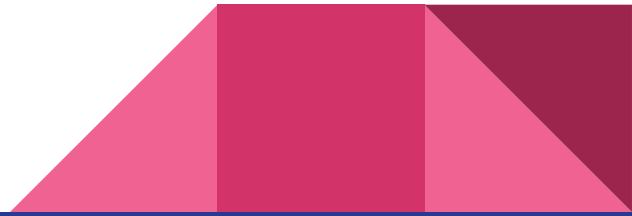
Zero-Shot & Fine-Tuning Results

- 20–30% error reduction zero-shot
- Matches/exceeds dense models in one epoch fine-tune
- Benchmarks: 6 unseen datasets



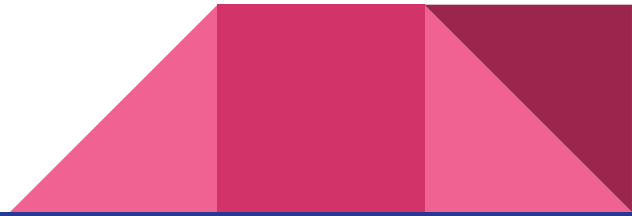
Efficiency Gains

- 78% faster training
- 39% faster inference
- Comparable cost to 100M-param dense model



Other Modalities at a Glance

- **Tabular:** cell imputation, schema transfer
- **Graph:** edge prediction, node masking
- **Vision/Text:** patch/image & wordpiece pre-training

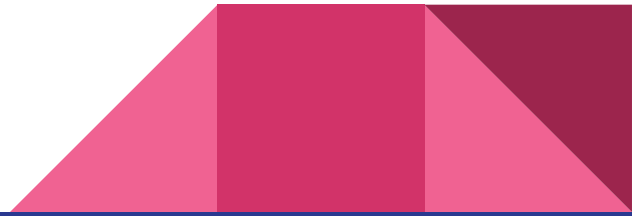


Challenges

- Expert interpretability
- Irregular or sparse sampling
- Multimodal fusion (time + text + image)

Future Directions

- Dynamic expert allocation & pruning
- Self-supervised objectives (contrastive, causal)
- Cross-modal pre-training frameworks



Conclusion & References

- TIME-MoE: one model, many forecasts
- 20–30% better zero-shot, huge speedups
- Ansari et al., “Billion-Scale Time Series Foundation Models with Mixture of Experts,” ICML 2024