In-context Fine-tuning for Time-series Foundation Models

Matthew Faw, Rajat Sen, Yichen Zhou, Abhimanyu Das
{matthewfaw, senrajat, yichenzhou, abhidas}@google.com

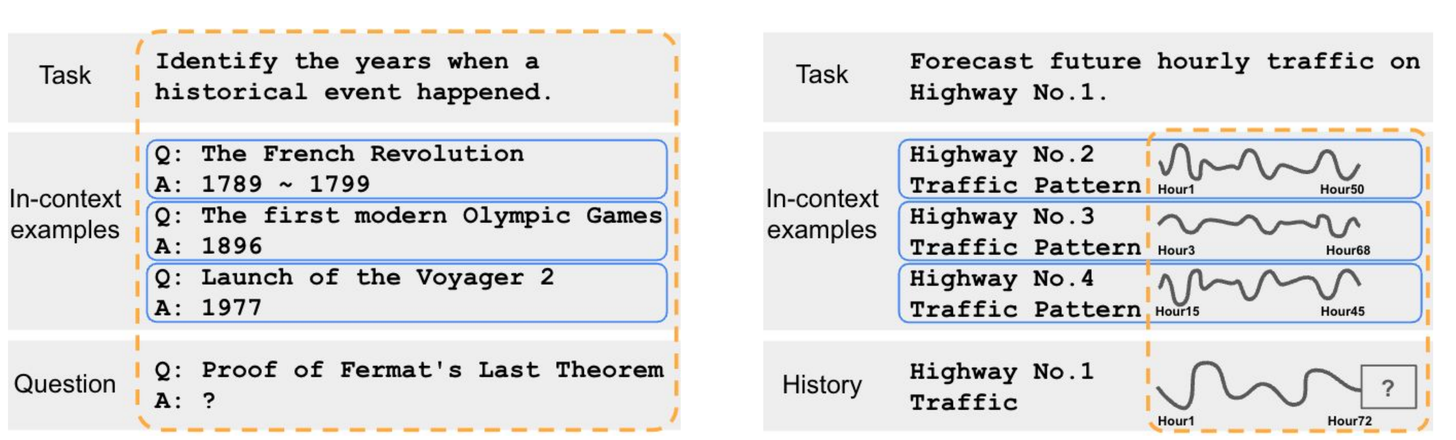
Motivation

- Time-series foundation models trained on hundred of billions of time-points consisting of time-series from various domains are gaining in popularity (see, e.g., [1-3])
- These models generalize to unseen datasets at inference time, i.e., do pretty well **zero-shot**.
- However, there are still areas of improvement:
  - Fine-tuning these foundation models on target domain datasets can boost performance
  - Fine-tuning *breaks* the zero-shot paradigm that precisely makes these timeseries foundation models so appealing to practitioners.
  - There is no clear way to prompt-tune these models

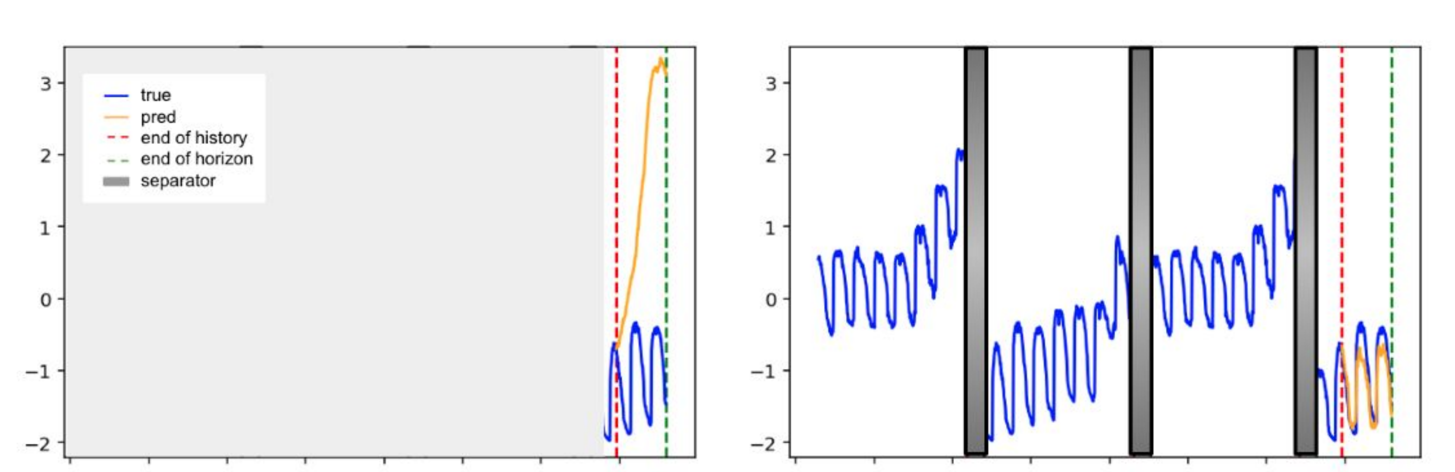
Goal: *Recover the benefits of fine-tuning a time-series foundation model by providing examples from a target dataset at inference time*

In-Context Examples

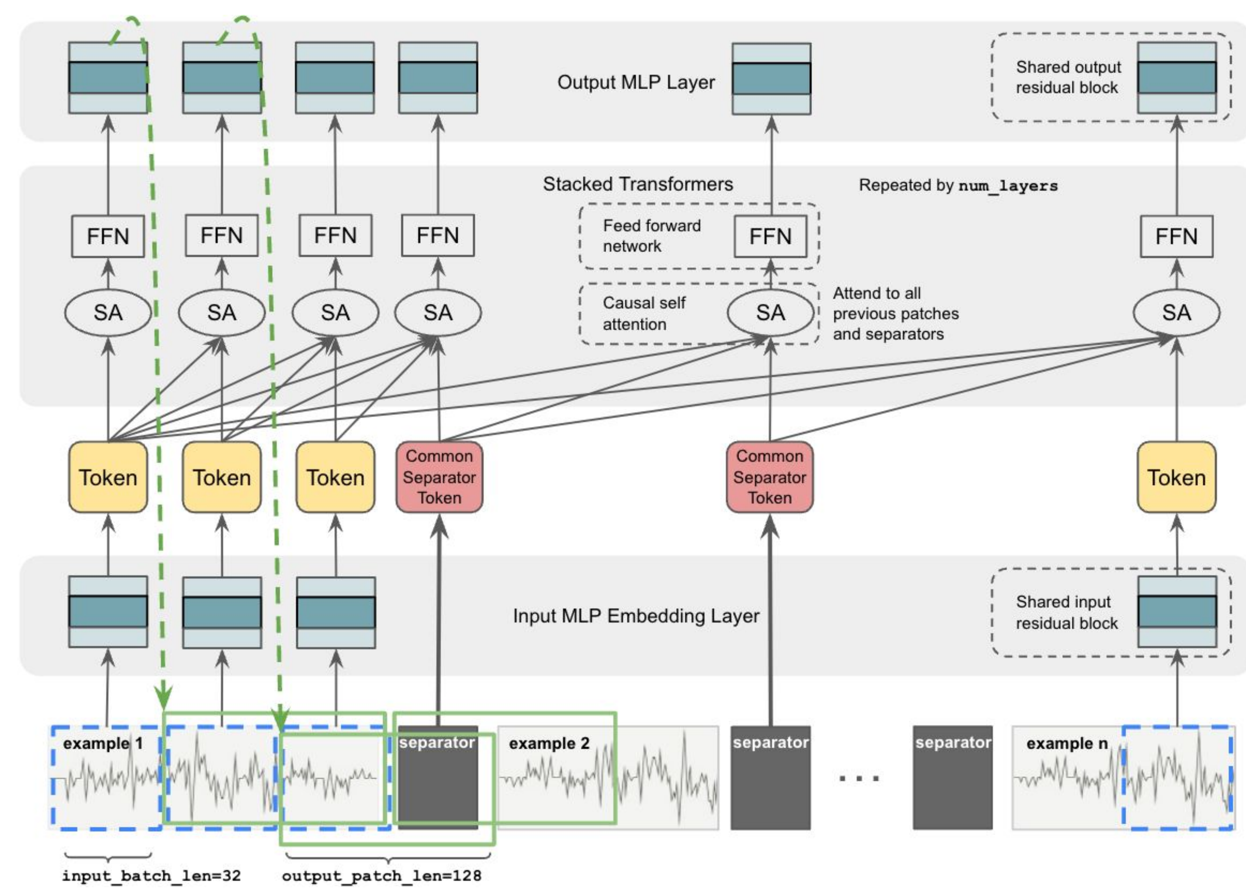
- Just like in NLP, we can potentially provide few shot time-series examples of other related forecasting tasks, provided the model is trained to handle them



- We augment each training example with additional time-series examples to improve model performance



Model Architecture



Training

- We start from the TimesFM-2.0, a 500M parameter model trained on > 400B time-points. The training corpus has variety of data sources like Wikipedia page visits, Google Trends, Synthetic time-series, and many smaller public time-series datasets from various domains including parts of LOTSA [1].

Table 2. Key statistics of LOTSA by domain.

	Energy	Transport	Climate	CloudOps	Web	Sales	Nature	Econ/Fin	Healthcare
# Datasets	30	23	6	3	3	6	5	23	6
# Obs.	16,358,600,896	4,900,453,419	4,188,011,890	1,518,268,292	428,082,373	197,984,339	28,547,647	24,919,596	1,594,281
%	59.17%	17.73%	15.15%	5.49%	1.55%	0.72%	0.09%	0.10%	0.01%

- A lot of this data can be grouped into related time-series mostly organized as smaller datasets – e.g., PEMSBAY has traffic data from related highways under Caltrans.

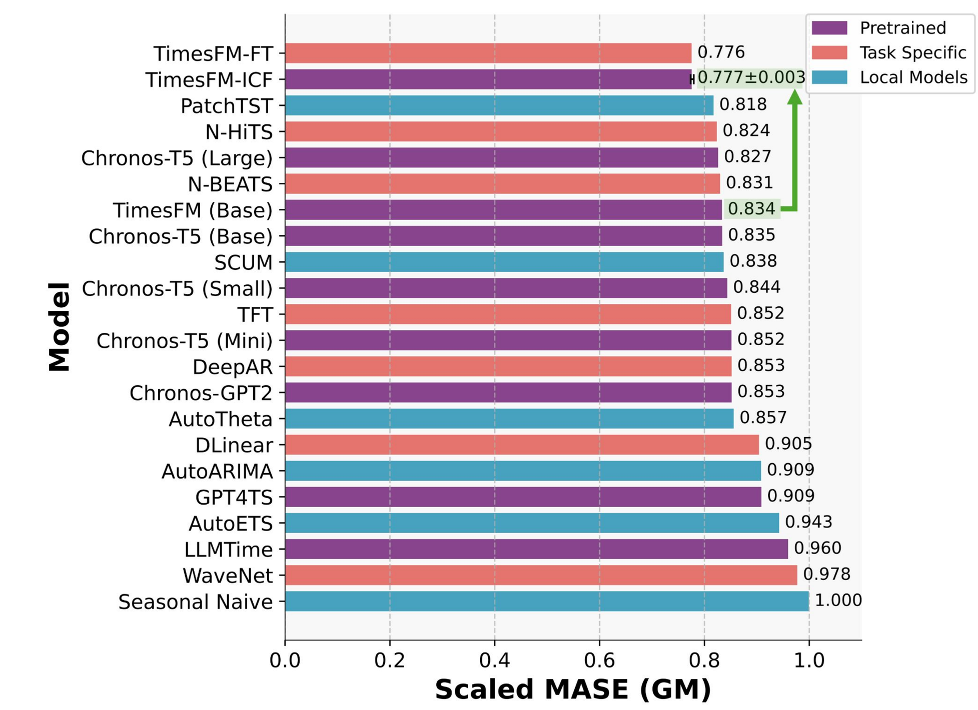
Dataset	Granularity	# Time series	# Time points
Synthetic		3,000,000	6,144,000,000
Electricity	Hourly	321	8,443,584
Traffic	Hourly	862	15,122,928
Weather (Zhou et al., 2021)	10 Min	42	2,213,232
Favorita Sales	Daily	111,840	139,179,538
LibCity (Wang et al., 2023)	15 Min	6,159	34,253,622
M4 hourly	Hourly	414	353,500
M4 daily	Daily	4,227	9,964,658
M4 monthly	Monthly	48,000	10,382,411
M4 quarterly	Quarterly	24,000	2,214,108
M4 yearly	Yearly	22,739	840,644
Wiki hourly	Hourly	5,608,693	239,110,787,496
Wiki daily	Daily	68,448,204	115,143,501,240
Wiki weekly	Weekly	66,579,850	16,414,251,948
Wiki monthly	Monthly	63,151,306	3,789,760,907
Trends hourly	Hourly	22,435	393,043,680
Trends daily	Daily	22,435	122,921,365
Trends weekly	Weekly	22,435	16,585,438
Trends monthly	Monthly	22,435	3,821,760

- We group windows of time-series that are either:
  - from the same dataset, or
  - from the same long time-series as related examples into one context i.e similar to packed examples.
- We maintain chronological order to avoid leakage due to our autoregressive decoding training strategy.
- We continue training the original TimesFM model with these packed example in decoder-only mode.
  - Crucially, while predicting the next patch, it can attend to the patches in previous time-windows (as well as the current one).
- We use separator tokens to distinguish windows.

Results

OOD Forecasting Benchmark

- We test on an OOD benchmark consisting of 23 datasets where out model, the original TimesFM and other foundation models like Chronos are zero-shot
- TimesFM-FT is a very strong bar because it is the base model fine-tuned on the training sets of each of these 23 datasets separately and then evaluated. TimesFM-ICF matches that without any extra training. Total time taken by TimesFM-ICF is significantly less.



Long-horizon forecasting on ETT

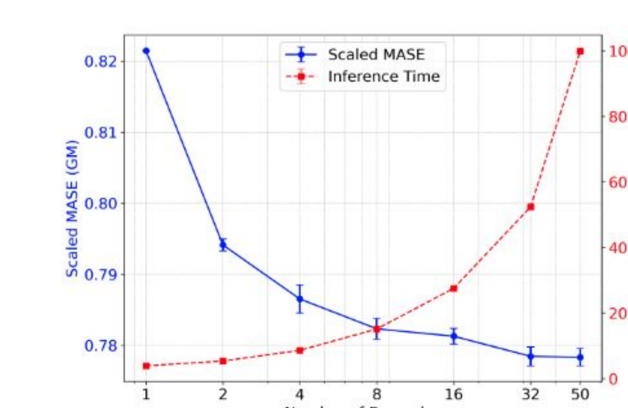
- We evaluate on the long-horizon forecasting benchmark of [4, 1] consisting of 4 Electricity Transformer Temperature (ETT) datasets with horizon lengths ranging from 96-720
- TimesFM-ICF rivals or outperforms all benchmarks, including TimesFM-FT, which was explicitly fine-tuned on the evaluation datasets

Table 1. MAE of TimesFM-ICF against other supervised and zero-shot methods on ETT Rolling Window, averaged over forecast horizons $\{96, 192, 336, 720\}$. See Table 9 for a detailed breakdown. We bold the numbers which are the best in every row, and including the ones that are within standard error of the best.

	Few-shot	Zero-shot				Task-specific							
Dataset	TimesFM-ICF	TimesFM (Base)	Moirai (Small)	Moirai (Base)	Moirai (Large)	TimesFM-FT	iTransformer	TimesNet	PatchTST	Crossformer	DLinear	SCINet	FEDformer
ETTh1	**0.405**	0.417	0.424	0.438	0.469	0.407	0.447	0.450	0.454	0.522	0.452	0.647	0.460
ETTh2	**0.378**	0.396	0.379	0.382	**0.377**	0.381	0.407	0.497	0.407	0.683	0.515	0.723	0.449
ETTm1	0.378	0.391	0.410	0.388	0.389	**0.371**	0.410	0.406	0.400	0.495	0.407	0.481	0.452
ETTm2	**0.307**	0.329	0.341	0.321	0.320	**0.306**	0.332	0.332	0.326	0.610	0.401	0.537	0.349

Ablation 1: Number of In-Context Examples

- Scaled MASE (GM) (+ inference time) vs number of in-context examples over the short context datasets in the OOD Benchmark
- Error decreases with number of in-context examples, but inference time increases



Ablation 2: Number of In-Context Examples

- Comparison against model trained with longer context length per window
- More shorter examples can be better than fewer long examples

Dataset	TimesFM-ICF	TimesFM (LH)	TimesFM (base)
OOD Benchmark	**0.777**	0.811	0.834

Google Research

References:
[1] Woo G, Liu C, Kumar A, Xiong C, Savarese S, Sahoo D. Unified training of universal time series forecasting transformers.
[2] Das A, Kong W, Sen R, Zhou Y. A decoder-only foundation model for time-series forecasting. In Forty-first International Conference on Machine Learning 2024 Jul 8.
[3] Ansari AF, Stella L, Turkmen C, Zhang X, Mercado P, Shen H, Shchur O, Rangapuram S, Arango SP, Kapoor S, Zschiegner J. Chronos: Learning the language of time series.
[4] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., and Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI conference on artificial intelligence, 2021.