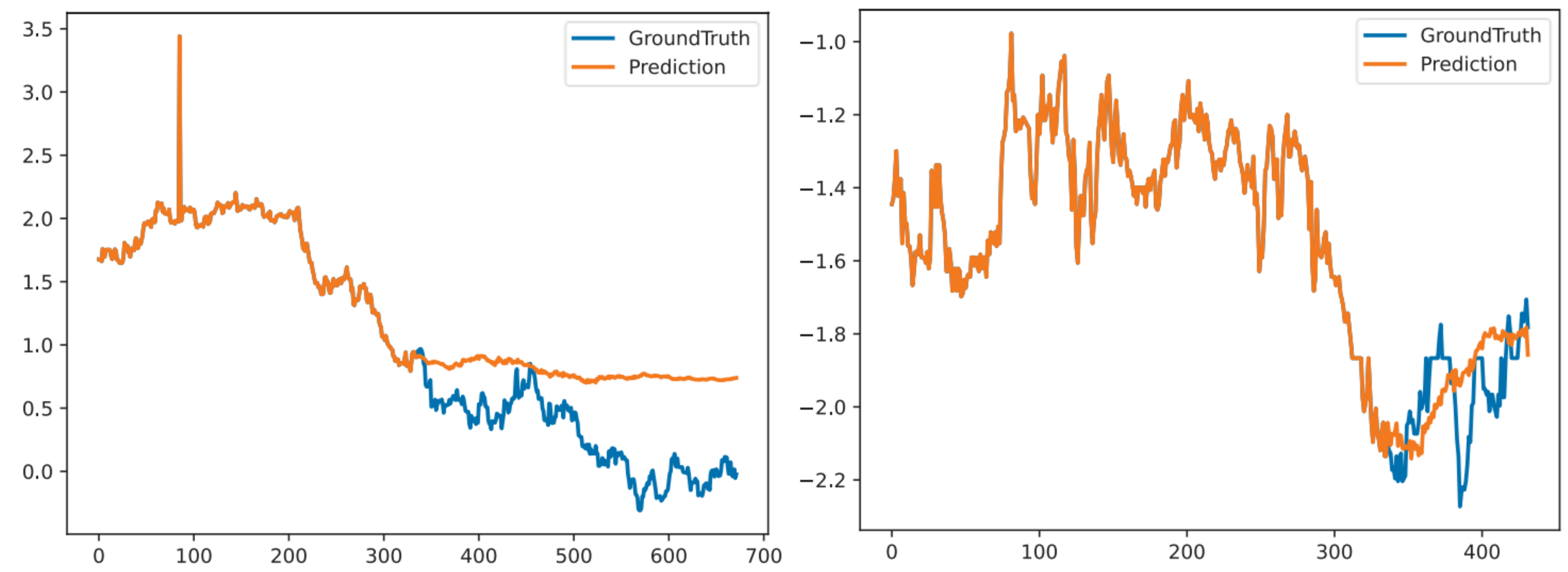
Generative Pretrained Hierarchical Transformer for Time Series Forecasting

Zhiding Liu[1], Jiqian Yang[1], Mingyue Cheng[1,*], Yucong Luo[1], Zhi Li[2]
[1]State Key Laboratory of Cognitive Intelligence,
University of Science and Technology of China, Hefei, China
[2]Shenzhen International Graduate School, Tsinghua University, Shenzhen, China

Task: Time series forecasting

Given the observation of past S time steps, predict the values of future T steps.



Motivation

- **Datasets**
  - (Pre-)Training on **a single dataset**, leading to suboptimal forecasting accuracy and transferability.
- **One-step generating schema**
  - A **customized head** is required for each forecasting task, hindering the generalizability of the pretrained models.
  - The **temporal dependencies within the predicted series** are inevitably overlooked, potentially leading to an inferior result.

Can we explore training **a single unified forecasting model** that generalizes well **across diverse data scenarios and forecasting settings**?

Method: Generative Pretrained Hierarchical Transformer

A. Pretraining Dataset

- Adopts the **channel-independent** assumption.
- Extend the methodology to the construction of the mixed pretraining dataset, **treating time series originating from various scopes as a whole** and **no extra information** is taken into account.
- The strategy can be therefore seamlessly applied to more diverse scenarios where the **covariate information may be missing and the data itself may be synthetic**.

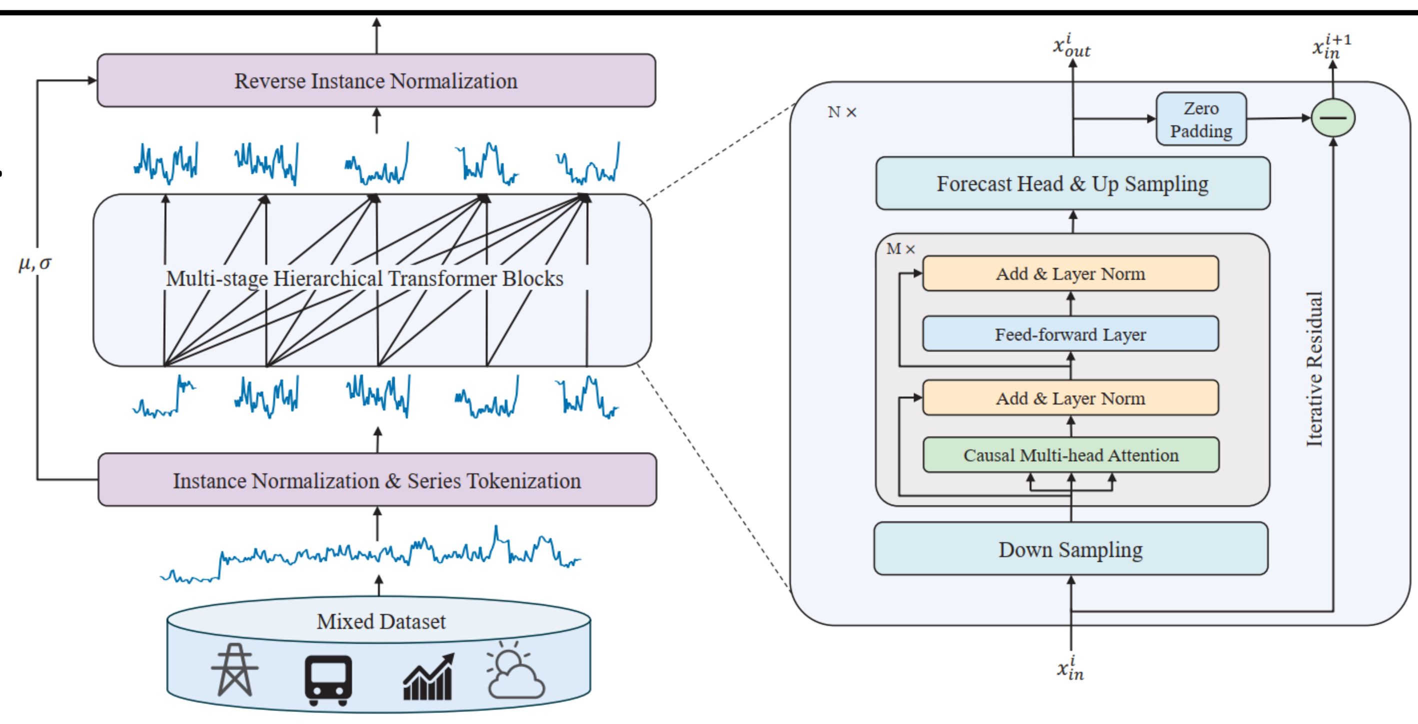


Illustration of the architecture of GPHT

B. The Hierarchical Structure

- We introduce a **token-level multi-stage** representation learning approach using hierarchical transformer blocks, where the sampling rate of each block varies.
- Can better capture the **multi-scale representation** of input series and better discover **commonalities hidden within mixed datasets** comprising various data scenarios.

C. The Optimization Target

- In pretraining, we formulate the pretraining task as a **standard language modeling** task, employing a **token-wise auto-regressive loss function** as the optimization target to fully leverage the mixed dataset and better capture temporal dependencies.
- In finetuning, we adopt a parameter-efficient tuning strategy where **only the forecasting heads are updated** to **strike a balance between maintaining generalizability and improving performance on a specific dataset**.

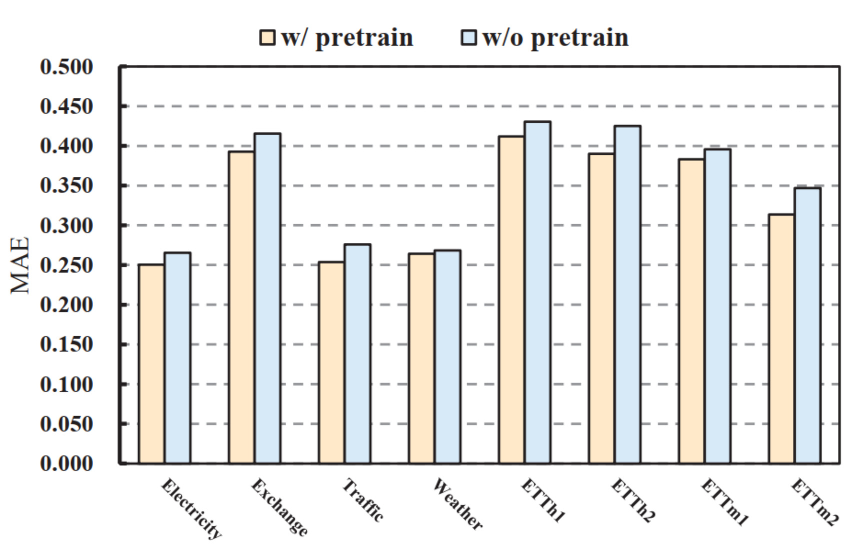
Experiments

Type		Ours				Self-supervised								Supervised							
Methods		GPHT*		GPHT		PatchTST		FPT		SimMTM		TimeMAE		PatchTST		iTransformer		TimesNet		DLinear	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	96	**0.128**	**0.219**	**0.128**	**0.219**	0.132	0.225	0.139	0.238	0.133	0.223	0.133	0.230	0.138	0.233	0.132	0.228	0.177	0.281	0.141	0.238
	192	0.147	**0.236**	**0.146**	**0.236**	0.148	0.241	0.155	0.252	0.147	0.237	0.150	0.246	0.153	0.247	0.154	0.249	0.193	0.295	0.154	0.251
	336	**0.165**	**0.255**	**0.165**	**0.255**	0.167	0.260	0.170	0.267	0.166	0.265	0.166	0.265	0.170	0.263	0.172	0.267	0.206	0.306	0.170	0.269
	720	0.206	**0.292**	0.207	**0.292**	0.205	**0.292**	0.208	0.299	0.203	0.297	**0.199**	0.296	0.206	0.295	0.204	0.296	0.223	0.320	0.205	0.302
Exchange	96	0.096	0.216	**0.087**	**0.207**	0.088	**0.207**	0.098	0.222	0.100	0.226	0.229	0.352	0.094	0.216	0.099	0.225	0.166	0.305	**0.087**	0.217
	192	0.183	0.304	0.172	**0.296**	0.186	0.308	0.209	0.327	0.210	0.332	0.653	0.581	0.191	0.311	0.206	0.329	0.303	0.413	**0.164**	0.298
	336	0.322	0.410	**0.309**	**0.400**	0.374	0.446	0.398	0.463	0.389	0.460	1.524	0.887	0.343	0.427	0.370	0.448	0.445	0.511	0.333	0.437
	720	0.833	0.685	**0.808**	**0.669**	0.857	0.692	1.010	0.747	1.104	0.800	2.525	1.193	0.888	0.706	0.963	0.746	1.389	0.899	0.988	0.749
Traffic	96	0.348	0.236	**0.346**	**0.234**	0.382	0.262	0.388	0.279	0.368	0.262	0.365	0.252	0.395	0.272	0.361	0.266	0.600	0.323	0.411	0.284
	192	0.374	0.248	**0.371**	**0.246**	0.385	0.261	0.411	0.287	0.373	0.251	0.383	0.260	0.411	0.278	0.378	0.271	0.612	0.327	0.423	0.289
	336	0.392	0.259	**0.388**	0.256	0.409	0.275	0.423	0.293	0.395	**0.254**	0.399	0.269	0.424	0.284	0.390	0.274	0.628	0.344	0.437	0.297
	720	0.428	0.284	**0.423**	**0.279**	0.438	0.291	0.449	0.307	0.432	0.290	0.438	0.291	0.453	0.300	0.424	0.291	0.657	0.349	0.467	0.316

Portion		5%										10%									
Methods		GPHT		FPT		SimMTM		PatchTST		iTransformer		GPHT		FPT		SimMTM		PatchTST		iTransformer	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	96	**0.143**	**0.237**	0.148	0.246	0.152	0.255	0.188	0.292	0.155	0.256	**0.140**	**0.233**	0.149	0.248	0.146	0.246	0.147	0.245	0.148	0.247
	192	**0.162**	**0.254**	0.163	0.259	0.167	0.268	0.202	0.304	0.172	0.272	**0.159**	**0.250**	0.164	0.261	0.163	0.262	0.162	0.258	0.167	0.266
	336	0.184	**0.275**	**0.181**	0.277	0.187	0.287	0.219	0.318	0.197	0.295	**0.180**	**0.271**	0.183	0.280	0.184	0.280	0.181	0.276	0.192	0.290
	720	0.238	0.321	**0.231**	**0.315**	0.240	0.326	0.264	0.351	0.261	0.344	0.231	**0.313**	0.234	0.318	0.242	0.325	**0.230**	0.315	0.244	0.329
ETTh1	96	**0.383**	**0.390**	0.478	0.474	0.537	0.502	0.505	0.481	0.580	0.520	**0.382**	**0.391**	0.453	0.454	0.482	0.467	0.450	0.448	0.557	0.514
	192	**0.426**	**0.416**	0.705	0.577	0.580	0.525	0.576	0.514	0.670	0.557	**0.424**	**0.418**	0.522	0.494	0.532	0.498	0.523	0.489	0.668	0.562
	336	**0.453**	**0.430**	0.736	0.571	0.603	0.543	0.672	0.554	0.726	0.577	**0.450**	**0.443**	0.571	0.522	0.561	0.523	0.523	0.494	0.684	0.559
	720	**0.433**	**0.440**	0.718	0.579	0.708	0.597	0.759	0.625	0.802	0.626	**0.427**	**0.442**	0.574	0.535	0.734	0.617	0.508	0.502	0.709	0.587

Main results & Few-shot evaluation

Methods		GPHT		FPT		PatchTST		DLinear	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Exchange	96	**0.098**	**0.219**	0.104	0.226	0.102	0.227	0.169	0.316
	192	**0.183**	**0.305**	0.218	0.333	0.205	0.325	0.230	0.374
	336	**0.321**	**0.411**	0.391	0.460	0.362	0.440	0.334	0.444
	720	0.824	0.682	0.978	0.734	0.991	0.745	**0.560**	**0.591**
Traffic	96	**0.411**	**0.291**	0.447	0.331	0.433	0.314	0.453	0.328
	192	**0.435**	**0.302**	0.461	0.335	0.447	0.319	0.464	0.330
	336	**0.460**	**0.316**	0.477	0.343	0.465	0.329	0.481	0.340
	720	0.521	0.353	**0.503**	0.356	0.504	0.354	0.506	**0.351**
Weather	96	**0.202**	**0.244**	0.216	0.264	0.207	0.259	0.239	0.297
	192	**0.248**	**0.283**	0.260	0.301	0.257	0.299	0.275	0.325
	336	**0.306**	**0.324**	0.328	0.351	0.340	0.350	0.323	0.360
	720	**0.389**	**0.377**	0.414	0.403	0.414	0.402	0.392	0.405



Zero-shot evaluation & Ablation on pretraining

Methods	Params	Training Time(per epoch)	Inference Speed(itr/s)
GPHT	37.98M(pretraining)/98.50K(finetuning)	20min(pretraining)/254.1s(finetuning)	0.34
FPT	105.20M(24.00M trainable)	3858.8s(finetuning)	0.69
SimMTM	62.14M(pretraining)/7.76M(finetuning)	73min(pretraining)/946.5s(finetuning)	5.98
PatchTST	4.27M	128.9s	9.02
iTransformer	5.28M	24.7s	26.39
TimesNet	150.64M	1179.6s	1.51

Computation cost comparison