A Time Series Is Worth 64 Words Long Term Forecasting With Transformers

https://arxiv.org/abs/2211.14730

0. Introduction

- 시계열 예측은 다양한 실세계 애플리케이션에서 중요하지만, 긴 범위의 예측은 여전히 어려움.
- 기존 Transformer 기반 모델들은 복잡하거나 비효율적.
- 저자는 간단하고 효율적인 구조인 TimesNet을 제안함.

1. Overview

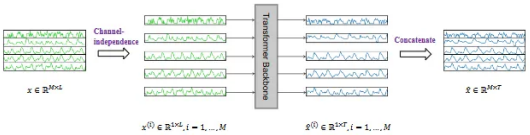
- 핵심 아이디어: 시계열을 패치 단위의 토큰으로 나누어 처리 → 자연어처럼 취급
- 각 시계열을 64개 패치 토큰으로 줄여 Transformer 입력에 맞게 변환.
- 이 구조를 통해 연산량 감소 및 예측 정확도 향상을 도모.

2. Challenges

- 시계열 데이터는 길고, 연속적이며, 동적 주기성과 잡음이 포함됨.
- 기존 Transformer는 입력이 길어질수록 계산량이 급증 → 스케일 문제.
- 시계열은 자연어보다 구조적으로 시간 종속성이 강함.

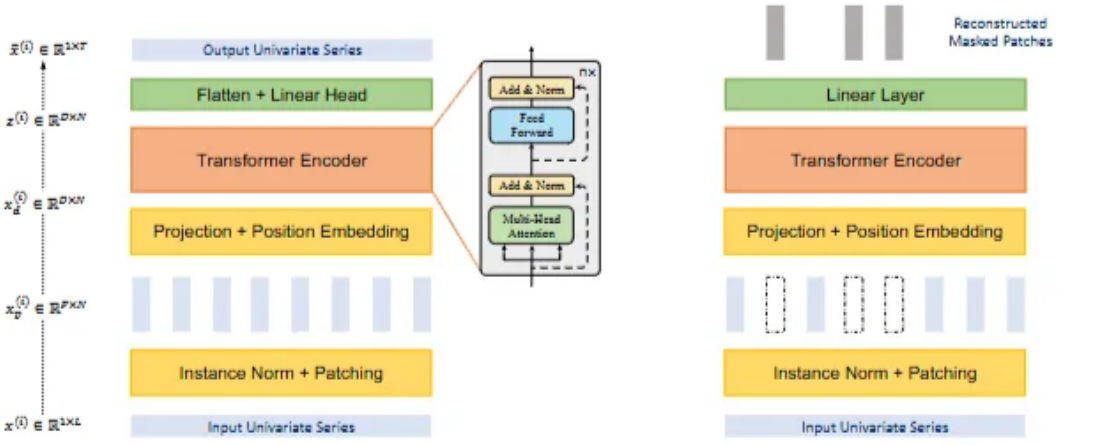
3. Method

- 시계열을 윈도우로 나누고, 각 윈도우를 축약하여 64개의 단어로 표현.
- 이 과정을 통해 Transformer 구조에 자연스럽게 적합하게 조정.
- Multi-layer Transformer로 시계열 패턴을 학습하고 예측 수행.



(a) PatchTST Model Overview

- Channel-independence : 다채널 시계열 데이터를 각 채널별로 분리해 독립적으로 분석
- Channel-mixing : 여러 채널을 혼합해 채널 간 상호작용을 학습 (예: attention 메커니즘)
- Channel-independent 모델
  - 채널 간 상호작용을 포착하기 어려울 수 있으나, 오히려 더 나은 성능을 보임
  - 유연성: 각 채널이 독립적으로 transformer layer를 통과 → 채널별 attention map 학습 → 다양한 패턴 학습 가능
  - 데이터 의존성: channel-mixing은 많은 데이터 필요 / channel-independence는 적은 데이터에도 효율적
  - 오버피팅: channel-mixing은 빠르게 오버피팅 / channel-independence는 안정적인 최적화와 높은 예측 성능



(b) Transformer Backbone (Supervised)

(c) Transformer Backbone (Self-supervised)

(b) ~ Patching

- **정의**: 인접한 데이터를 일정 단위(patch length)로 묶고 stride 단위로 슬라이딩
- **장점**: token 수 감소 → 시간·공간 복잡도 낮춤 → 긴 시계열 학습 가능 → 예측 성능 향상
- **프로세스**:

    - 채널 분리 + patching
    - vanilla transformer encoder 통과
    - linear layer로 예측
- **확장**: self-supervised learning을 추가 적용해 성능 향상 시도

(c) ~ Self-Supervised Representation Learning

- **정의**: 라벨 없이 의미 있는 표현 학습
- **방식**: 일부 patch를 마스킹 → 모델이 마스킹된 부분을 복원하도록 학습
- **설계**: 마스킹된 patch와 인접 patch가 겹치지 않도록 구성 (patch length = stride)
- **효과**:
    - 채널 내 패턴 이해도 증가
    - 일반화된 표현 학습
    - transfer learning에 활용 가능

4. Experiments

Datasets	Weather	Traffic	Electricity	ILI	ETTh1	ETTh2	ETTm1	ETTm2
Features	21	862	321	7	7	7	7	7
Timesteps	52696	17544	26304	966	17420	17420	69680	69680

Table 2: Statistics of popular datasets for benchmark.

- 총 6개 시계열 데이터셋(ETT, Electricity, Exchange Rate 등)을 사용해 비교 실험.
- 기존 SOTA 모델들과 비교 (Informer, Autoformer, FEDformer 등).
- 평가 지표로는 Mean Squared Error와 Mean Absolute Error 사용
- 다양한 horizon(예측 범위)에 대한 성능 측정.

5. Results

Models		PatchTST/64		PatchTST/42		DLinear		FEDformer		Autoformer		Informer		Pyraformer		LogTrans	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	**0.149**	**0.198**	0.152	0.199	0.176	0.237	0.238	0.314	0.249	0.329	0.354	0.405	0.896	0.556	0.458	0.490
	192	**0.194**	**0.241**	0.197	0.243	0.220	0.282	0.275	0.329	0.325	0.370	0.419	0.434	0.622	0.624	0.658	0.589
	336	**0.245**	**0.282**	0.249	0.283	0.265	0.319	0.339	0.377	0.351	0.391	0.583	0.543	0.739	0.753	0.797	0.652
	720	**0.314**	**0.334**	0.320	0.335	0.323	0.362	0.389	0.409	0.415	0.426	0.916	0.705	1.004	0.934	0.869	0.675
Traffic	96	**0.360**	**0.249**	0.367	0.251	0.410	0.282	0.576	0.359	0.597	0.371	0.733	0.410	2.085	0.468	0.684	0.384
	192	**0.379**	**0.256**	0.385	0.259	0.423	0.287	0.610	0.380	0.607	0.382	0.777	0.435	0.867	0.467	0.685	0.390
	336	**0.392**	**0.264**	0.398	0.265	0.436	0.296	0.608	0.375	0.623	0.387	0.776	0.434	0.869	0.469	0.734	0.408
	720	**0.432**	**0.286**	0.434	0.287	0.466	0.315	0.621	0.375	0.639	0.395	0.827	0.466	0.881	0.473	0.717	0.396
Electricity	96	**0.129**	**0.222**	0.130	**0.222**	0.140	0.237	0.186	0.302	0.196	0.313	0.304	0.393	0.386	0.449	0.258	0.357
	192	**0.147**	**0.240**	0.148	**0.240**	0.153	0.249	0.197	0.311	0.211	0.324	0.327	0.417	0.386	0.443	0.266	0.368
	336	**0.163**	**0.259**	0.167	0.261	0.169	0.267	0.213	0.328	0.214	0.327	0.333	0.422	0.378	0.443	0.280	0.380
	720	**0.197**	**0.290**	0.202	0.291	0.203	0.301	0.233	0.344	0.236	0.342	0.351	0.427	0.376	0.445	0.283	0.376
ILI	24	**1.319**	**0.754**	1.522	0.814	2.215	1.081	2.624	1.095	2.906	1.182	4.657	1.449	1.420	2.012	4.480	1.444
	36	1.579	0.870	**1.430**	**0.834**	1.963	0.963	2.516	1.021	2.585	1.038	4.650	1.463	7.394	2.031	4.799	1.467
	48	**1.553**	**0.815**	1.673	0.854	2.130	1.024	2.505	1.041	3.024	1.145	5.004	1.542	7.551	2.057	4.800	1.468
	60	**1.470**	**0.788**	1.529	0.862	2.368	1.096	2.742	1.122	2.761	1.114	5.071	1.543	7.662	2.100	5.278	1.560
ETTh1	96	**0.370**	0.400	0.375	**0.399**	0.375	**0.399**	0.376	0.415	0.435	0.446	0.941	0.769	0.664	0.612	0.878	0.740
	192	0.413	0.429	0.414	0.421	**0.405**	**0.416**	0.423	0.446	0.456	0.457	1.007	0.786	0.790	0.681	1.037	0.824
	336	**0.422**	0.440	0.431	**0.436**	0.439	0.443	0.444	0.462	0.486	0.487	1.038	0.784	0.891	0.738	1.238	0.932
	720	**0.447**	0.468	0.449	**0.466**	0.472	0.490	0.469	0.492	0.515	0.517	1.144	0.857	0.963	0.782	1.135	0.852
ETTh2	96	**0.274**	0.337	**0.274**	**0.336**	0.289	0.353	0.332	0.374	0.332	0.368	1.549	0.952	0.645	0.597	2.116	1.197
	192	0.341	0.382	**0.339**	**0.379**	0.383	0.418	0.407	0.446	0.426	0.434	3.792	1.542	0.788	0.683	4.315	1.635
	336	**0.329**	0.384	0.331	**0.380**	0.448	0.465	0.400	0.447	0.477	0.479	4.215	1.642	0.907	0.747	1.124	1.604
	720	**0.379**	**0.422**	**0.379**	**0.422**	0.605	0.551	0.412	0.469	0.453	0.490	3.656	1.619	0.963	0.783	3.188	1.540
ETTm1	96	0.293	0.346	**0.290**	**0.342**	0.299	0.343	0.326	0.390	0.510	0.492	0.626	0.560	0.543	0.510	0.600	0.546
	192	0.333	0.370	**0.332**	0.369	0.335	**0.365**	0.365	0.415	0.514	0.495	0.725	0.619	0.557	0.537	0.837	0.700
	336	0.369	0.392	**0.366**	0.392	0.369	**0.386**	0.392	0.425	0.510	0.492	1.005	0.741	0.754	0.655	1.124	0.832
	720	**0.416**	**0.420**	0.420	0.424	0.425	0.421	0.446	0.458	0.527	0.493	1.133	0.845	0.908	0.724	1.153	0.820
ETTm2	96	0.166	0.256	**0.165**	**0.255**	0.167	0.260	0.180	0.271	0.205	0.293	0.355	0.462	0.435	0.507	0.768	0.642
	192	0.223	0.296	**0.220**	**0.292**	0.224	0.303	0.252	0.318	0.278	0.336	0.595	0.586	0.730	0.673	0.989	0.757
	336	**0.274**	**0.329**	0.278	**0.329**	0.281	0.342	0.324	0.364	0.343	0.379	1.270	0.871	1.201	0.845	1.334	0.872
	720	**0.362**	**0.385**	0.367	**0.385**	0.397	0.421	0.410	0.420	0.414	0.419	3.001	1.267	3.625	1.451	3.048	1.328

Table 3: Multivariate long-term forecasting results with supervised PatchTST. We use prediction lengths $T \in \{24, 36, 48, 60\}$ for ILI dataset and $T \in \{96, 192, 336, 720\}$ for the others. The best results are in **bold** and the second best are underlined.

- PatchTST/42: 모든 벤치마크에서 기존 Transformer 모델 및 DLinear보다 우수
- PatchTST/64: 더 긴 입력으로 예측 성능 추가 향상
- Transformer 모델 대비 MSE 21.0%, MAE 16.7% 감소
- DLinear 대비도 대부분의 데이터셋에서 성능 우위

Models		PatchTST						DLinear		FEDformer		Autoformer		Informer	
		Fine-tuning		Lin. Prob.		Sup.									
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	**0.144**	**0.193**	0.158	0.209	0.152	0.199	0.176	0.237	0.238	0.314	0.249	0.329	0.354	0.405
	192	**0.190**	**0.236**	0.203	0.249	0.197	0.243	0.220	0.282	0.275	0.329	0.325	0.370	0.419	0.434
	336	**0.244**	**0.280**	0.251	0.285	0.249	0.283	0.265	0.319	0.339	0.377	0.351	0.391	0.583	0.543
	720	**0.320**	**0.335**	0.321	0.336	**0.320**	**0.335**	0.323	0.362	0.389	0.409	0.415	0.426	0.916	0.705
Traffic	96	**0.352**	**0.244**	0.399	0.294	0.367	0.251	0.410	0.282	0.576	0.359	0.597	0.371	0.733	0.410
	192	**0.371**	**0.253**	0.412	0.298	0.385	0.259	0.423	0.287	0.610	0.380	0.607	0.382	0.777	0.435
	336	**0.381**	**0.257**	0.425	0.306	0.398	0.265	0.436	0.296	0.608	0.375	0.623	0.387	0.776	0.434
	720	**0.425**	**0.282**	0.460	0.323	0.434	0.287	0.466	0.315	0.621	0.375	0.639	0.395	0.827	0.466
Electricity	96	**0.126**	**0.221**	0.138	0.237	0.130	**0.222**	0.140	0.237	0.186	0.302	0.196	0.313	0.304	0.393
	192	**0.145**	**0.238**	0.156	0.252	0.148	0.240	0.153	0.249	0.197	0.311	0.211	0.324	0.327	0.417
	336	**0.164**	**0.256**	0.170	0.265	0.167	0.261	0.169	0.267	0.213	0.328	0.214	0.327	0.333	0.422
	720	**0.193**	**0.291**	0.208	0.297	0.202	**0.291**	0.203	0.301	0.233	0.344	0.236	0.342	0.351	0.427

Table 4: Multivariate long-term forecasting results with self-supervised PatchTST. We use prediction lengths $T \in \{96, 192, 336, 720\}$. The best results are in **bold** and the second best are underlined.

- Fine-tuning이 최상 성능, 그러나 linear probing만으로도 supervised보다 좋음
- 대형 데이터셋(Weather, Traffic, Electricity)에서 특히 뛰어난 일반화 성능을 보임

Models		PatchTST						DLinear		FEDformer		Autoformer		Informer	
		Fine-tuning		Lin. Prob.		Sup.									
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	**0.145**	**0.195**	0.163	0.216	0.152	0.199	0.176	0.237	0.238	0.314	0.249	0.329	0.354	0.405
	192	**0.193**	**0.243**	0.205	0.252	0.197	**0.243**	0.220	0.282	0.275	0.329	0.325	0.370	0.419	0.434
	336	**0.244**	**0.280**	0.253	0.289	0.249	0.283	0.265	0.319	0.339	0.377	0.351	0.391	0.583	0.543
	720	0.321	0.337	**0.320**	0.336	**0.320**	**0.335**	0.323	0.362	0.389	0.409	0.415	0.426	0.916	0.705
Traffic	96	0.388	0.273	0.400	0.288	**0.367**	**0.251**	0.410	0.282	0.576	0.359	0.597	0.371	0.733	0.410
	192	0.400	0.277	0.412	0.293	**0.385**	**0.259**	0.423	0.287	0.610	0.380	0.607	0.382	0.777	0.435
	336	0.408	0.280	0.425	0.307	**0.398**	**0.265**	0.436	0.296	0.608	0.375	0.623	0.387	0.776	0.434
	720	0.447	0.310	0.457	0.317	**0.434**	**0.287**	0.466	0.315	0.621	0.375	0.639	0.395	0.827	0.466

Table 5: Transfer learning task: PatchTST is pre-trained on Electricity dataset and the model is transferred to other datasets. The best results are in **bold** and the second best are underlined.

- 일부 데이터셋에서는 fine-tune시 supervised보다 소폭 낮지만, 대부분의 기존 Transformer 모델보다 뛰어남
- 학습 횟수(epochs)가 적기 때문에 시간 효율도 우수함

Models		PatchTST								FEDformer	
		P+CI		CI		P		Original			
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Weather	96	**0.152**	**0.199**	0.164	0.213	0.168	0.223	0.177	0.236	0.238	0.314
	192	**0.197**	**0.243**	0.205	0.250	0.213	0.262	0.221	0.270	0.275	0.329
	336	**0.249**	**0.283**	0.255	0.289	0.266	0.300	0.271	0.306	0.339	0.377
	720	**0.320**	**0.335**	0.327	0.343	0.351	0.359	0.340	0.353	0.389	0.409
Traffic	96	**0.367**	**0.251**	0.397	0.271	0.595	0.376	-	-	0.576	0.359
	192	**0.385**	**0.259**	0.411	0.276	0.612	0.387	-	-	0.610	0.380
	336	**0.398**	**0.265**	0.423	0.282	0.651	0.391	-	-	0.608	0.375
	720	**0.434**	**0.287**	0.457	0.309	-	-	-	-	0.621	0.375
Electricity	96	**0.130**	**0.222**	0.136	0.231	0.196	0.307	0.205	0.318	0.186	0.302
	192	**0.148**	**0.240**	0.164	0.263	0.215	0.323	-	-	0.197	0.311
	336	**0.167**	**0.261**	0.168	0.262	0.228	0.338	-	-	0.213	0.328
	720	**0.202**	**0.291**	0.219	0.312	0.244	0.345	-	-	0.233	0.344

- PatchTST는 L이 증가할수록 예측 성능이 꾸준히 향상
- 기존 모델들은 L 증가 시 성능 개선이 거의 없음 -> 시계열 정보 흡수에 비효율

- 대부분의 데이터셋과 예측 길이에서 SOTA 성능 달성.
- 특히 긴 시계열 예측에서 성능 우수.
- 연산량 및 파라미터 수에서도 경량화된 모델 구조로 효율 확보.

6. Insight

- 시계열을 언어처럼 다룰 수 있다는 새로운 접근법 제시.

- 입력 압축(tokenization)과 간단한 구조만으로도 복잡한 모델보다 우수한 결과 가능.
- 미래의 시계열 연구에서 언어 모델 기반 접근의 가능성을 시사.
- 시계열 데이터 표현 학습 (Representation Learning for Time Series)을 활용하는 TS2Vec, TNC (Temporal Neighborhood Coding), CPC (Contrastive Predictive Coding) 논문 읽어볼 예정