

ITRANSFORMER: INVERTED TRANSFORMERS ARE EFFECTIVE FOR TIME SERIES FORECASTING

<https://arxiv.org/pdf/2310.06625>

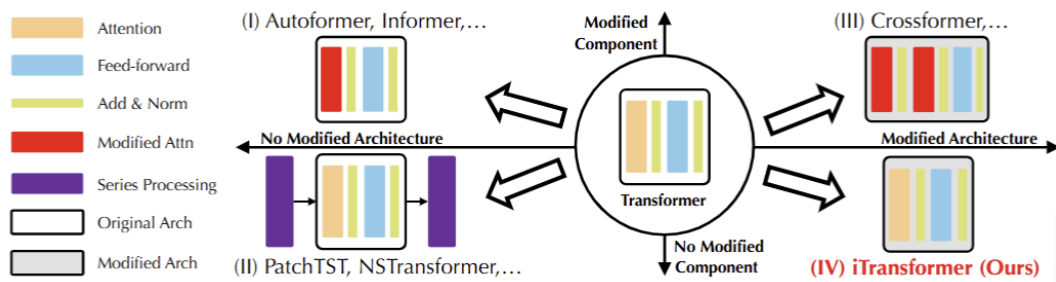
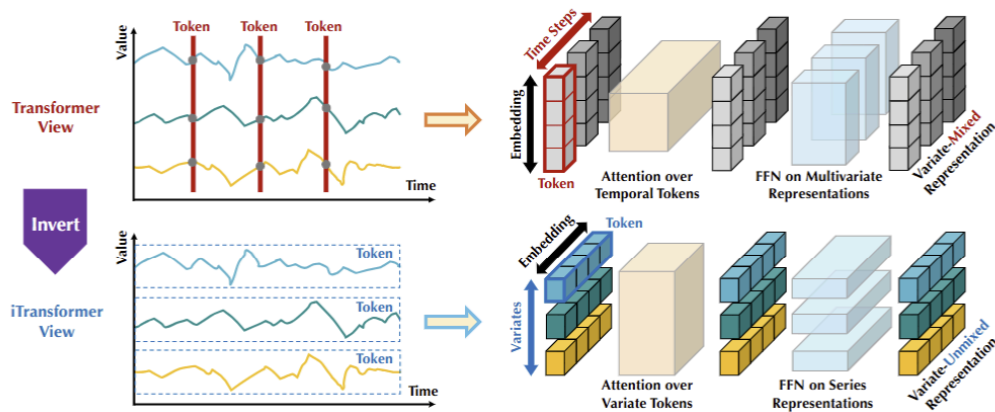
0. Introduction

- 시계열 예측은 금융, 에너지, 교통 등 다양한 분야의 핵심 문제
- Transformer 계열 모델은 장기 의존성 학습에 강점을 보였으나 다변량 시계열에서는 성능과 효율성에 한계 존재
- 기존 접근은 시간 축에 대한 attention에 집중
- 본 논문은 기존 Transformer 구조를 재해석하여 변수 차원에 attention을 적용하는 iTransformer를 제안
- 핵심 질문은 “시계열에서 정말 중요한 attention 축은 무엇인가”

1. Overview

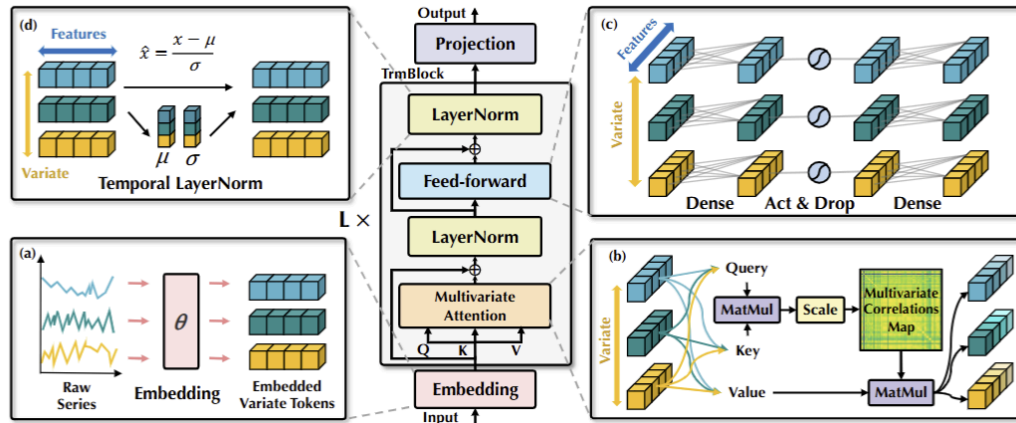
- iTransformer는 기존 Transformer의 입력 구조를 뒤집은 모델
- 시간 축이 아닌 변수 축을 토큰으로 사용
- 각 변수 간 상관관계를 attention으로 직접 학습
- 시간 정보는 embedding 형태로 간접 반영
- 장기 예측, 다변량 시계열에서 성능과 효율을 동시에 개선하는 것이 목표

2. Challenges



- 다변량 시계열은 변수 수가 많을수록 구조적 복잡성 증가
- 기존 Transformer는
 - 시간 길이에 비례한 계산 비용
 - 변수 간 관계를 간접적으로만 학습
- 장기 예측에서
 - 정보 희석
 - 불필요한 attention 계산 발생
- 모델 구조가 복잡해질수록 해석 가능성 저하

3. Method



- 입력 시계열을 (time × variable)에서 (variable × time) 관점으로 재구성
- 각 변수를 하나의 토큰으로 취급
- self-attention은 변수 간 관계를 학습
- 시간 정보는 positional embedding 또는 projection으로 처리
- 기존 Transformer 블록 구조는 유지
- 구조 변경만으로도 계산 복잡도 감소
- 별도의 복잡한 모듈 추가 없이 적용 가능

4. Experiments

- 다변량 시계열 예측 벤치마크 데이터셋 사용
 - ETT
 - Weather
 - Electricity
 - Traffic 등
- 기존 시계열 Transformer 계열 모델과 비교
 - Autoformer
 - Informer
 - FEDformer
 - PatchTST 등

- 평가 지표
 - MSE
 - MAE
- 다양한 예측 길이 설정으로 장기 예측 성능 평가

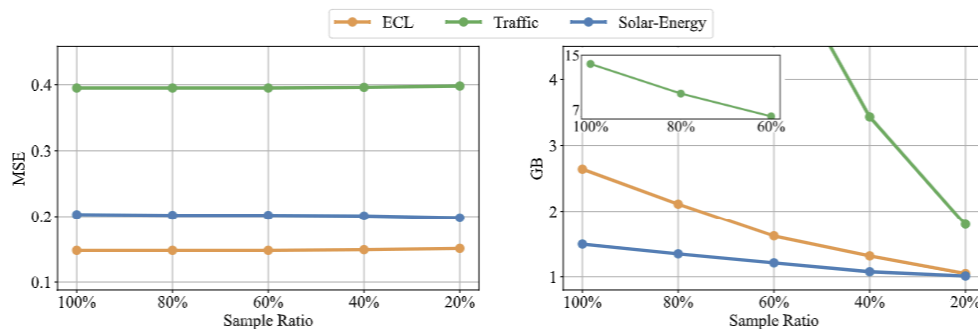
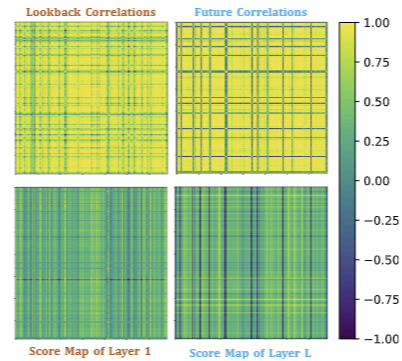
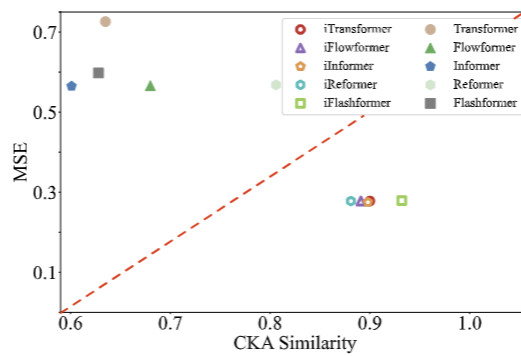
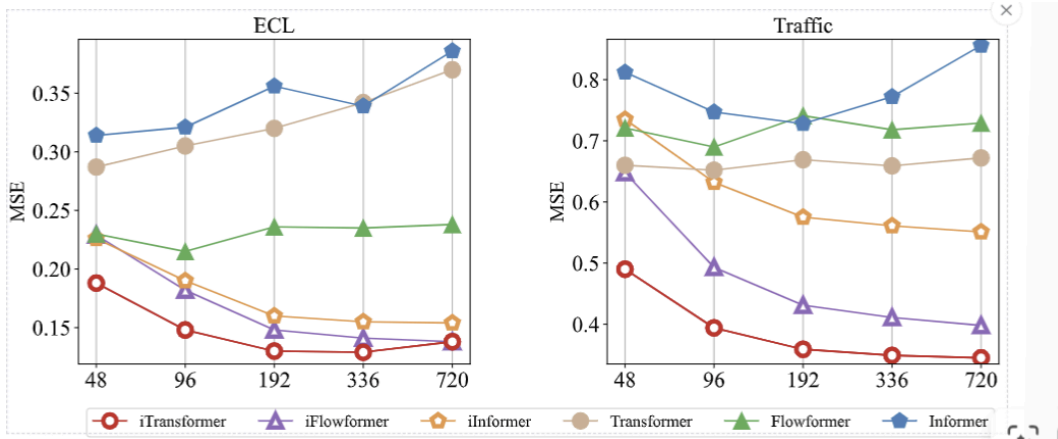
5. Results

Table 1: Multivariate forecasting results with prediction lengths $S \in \{12, 24, 36, 48\}$ for PEMS and $S \in \{96, 192, 336, 720\}$ for others and fixed lookback length $T = 96$. Results are averaged from all prediction lengths. Avg means further averaged by subsets. Full results are listed in Appendix F.4.

Models	iTransformer (Ours)		RLinear (2023)		PatchTST (2023)		Crossformer (2023)		TiDE (2023)		TimesNet (2023)		DLinear (2023)		SCINet (2022a)		FEDformer (2022)		Stationary (2022b)		Autoformer (2021)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ECL	0.178	0.270	0.219	0.298	0.205	0.290	0.244	0.334	0.251	0.344	0.192	0.295	0.212	0.300	0.268	0.365	0.214	0.327	0.193	0.296	0.227	0.338
ETT (Avg)	0.383	0.399	0.380	0.392	0.381	0.397	0.685	0.578	0.482	0.470	0.391	0.404	0.442	0.444	0.689	0.597	0.408	0.428	0.471	0.464	0.465	0.459
Exchange	0.360	0.403	0.378	0.417	0.367	0.404	0.940	0.707	0.370	0.413	0.416	0.443	0.354	0.414	0.750	0.626	0.519	0.429	0.461	0.454	0.613	0.539
Traffic	0.428	0.282	0.626	0.378	0.481	0.304	0.550	0.304	0.760	0.473	0.620	0.336	0.625	0.383	0.804	0.509	0.610	0.376	0.624	0.340	0.628	0.379
Weather	0.258	0.278	0.272	0.291	0.259	0.281	0.259	0.315	0.271	0.320	0.259	0.287	0.265	0.317	0.292	0.363	0.309	0.360	0.288	0.314	0.338	0.382
Solar-Energy	0.233	0.262	0.369	0.356	0.270	0.307	0.641	0.639	0.347	0.417	0.301	0.319	0.330	0.401	0.282	0.375	0.291	0.381	0.261	0.381	0.885	0.711
PEMS (Avg)	0.119	0.218	0.514	0.482	0.217	0.305	0.220	0.304	0.375	0.440	0.148	0.246	0.320	0.394	0.121	0.222	0.224	0.327	0.151	0.249	0.614	0.575

Table 2: Performance promotion obtained by our inverted framework. Flashformer means Transformer equipped with hardware-accelerated FlashAttention (Dao et al., 2022). We report the average performance and the relative MSE reduction (Promotion). Full results can be found in Appendix F.2.

Models		Transformer (2017)		Reformer (2020)		Informer (2021)		Flowformer (2022)		Flashformer (2022)	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ECL	Original	0.277	0.372	0.338	0.422	0.311	0.397	0.267	0.359	0.285	0.377
	+Inverted	0.178	0.270	0.208	0.301	0.216	0.311	0.210	0.293	0.206	0.291
	Promotion	35.6%	27.4%	38.4%	28.7%	30.5%	21.6%	21.3%	18.6%	27.8%	22.9%
Traffic	Original	0.665	0.363	0.741	0.422	0.764	0.416	0.750	0.421	0.658	0.356
	+Inverted	0.428	0.282	0.647	0.370	0.662	0.380	0.524	0.355	0.492	0.333
	Promotion	35.6%	22.3%	12.7%	12.3%	13.3%	8.6%	30.1%	15.6%	25.2%	6.4%
Weather	Original	0.657	0.572	0.803	0.656	0.634	0.548	0.286	0.308	0.659	0.574
	+Inverted	0.258	0.279	0.248	0.292	0.271	0.330	0.266	0.285	0.262	0.282
	Promotion	60.2%	50.8%	69.2%	55.5%	57.3%	39.8%	7.2%	7.7%	60.2%	50.8%



- iTransformer는 대부분의 데이터셋에서 SOTA 또는 준 SOTA 성능 달성
- 장기 예측 구간에서 성능 저하가 적음
- 변수 수가 많은 데이터에서 강점
- 모델 구조는 단순하지만 성능은 기존 복잡한 모델과 경쟁
- 계산 효율성과 메모리 사용 측면에서도 이점 확인
- 구조 변경만으로 성능이 개선된 점이 인상적

6. Insight

- 이 논문의 핵심은 새로운 모델 제안이 아니라 문제 재정의
- 시계열 예측에서 중요한 것은 시간 간 관계가 아니라
변수 간 구조적 상관관계일 수 있음을 명확히 보여줌
- attention을 어디에 쓰느냐가 모델 성능을 결정한다는 점을 실증
- 복잡한 모듈 추가보다 입력 표현 방식이 더 중요할 수 있음을 시사
- PatchTST 이후 시계열 Transformer 흐름과 자연스럽게 연결
- 이 논문은 시계열 모델 설계에서 "더 복잡하게"가 아니라 "더 정확하게 바라보는 것"의
중요성을 강조