2026 01 14
발표 자료

광운대학교 로봇학과
FAIR Lab

김한서

KWANGWOON UNIVERSITY

이번 주 진행사항

- Why Attention Fails
  - 논문 리뷰

Why Attention Fails

Why Attention Fails: The Degeneration of Transformers into MLPs in Time Series Forecasting

Liang Zida
Shanghai Jiaotong University
greek-guardian@sjtu.edu.cn

Jiayi Zhu
Shanghai Jiaotong University
18161778290@sjtu.edu.cn

Weiqiang Sun*
Shanghai Jiaotong University
sunwq@sjtu.edu.cn

- arXiv 등록일: 2025-09-25
- 인용 수: 2회(Google Scholar, 2026-01-09)
- Published at ICLR 2026
- 링크: https://arxiv.org/pdf/2509.20942
- 시계열 예측에서 Transformer의 Attention Mechanism이 제대로 동작하는지 실험한 논문

Why Attention Fails

Background

- 문제점
  - 기존 Multi-Head Attention이 데이터 간 문맥적 관계를 거의 학습하지 못함
  - 모델 성능의 대부분이 Feed-Forward Network에서 나와 단순 MLP처럼 작동함
- 원인 분석
  - Attention Replacement, Perturbation 등 4가지 실험을 통한 검증

Why Attention Fails

실험 방법

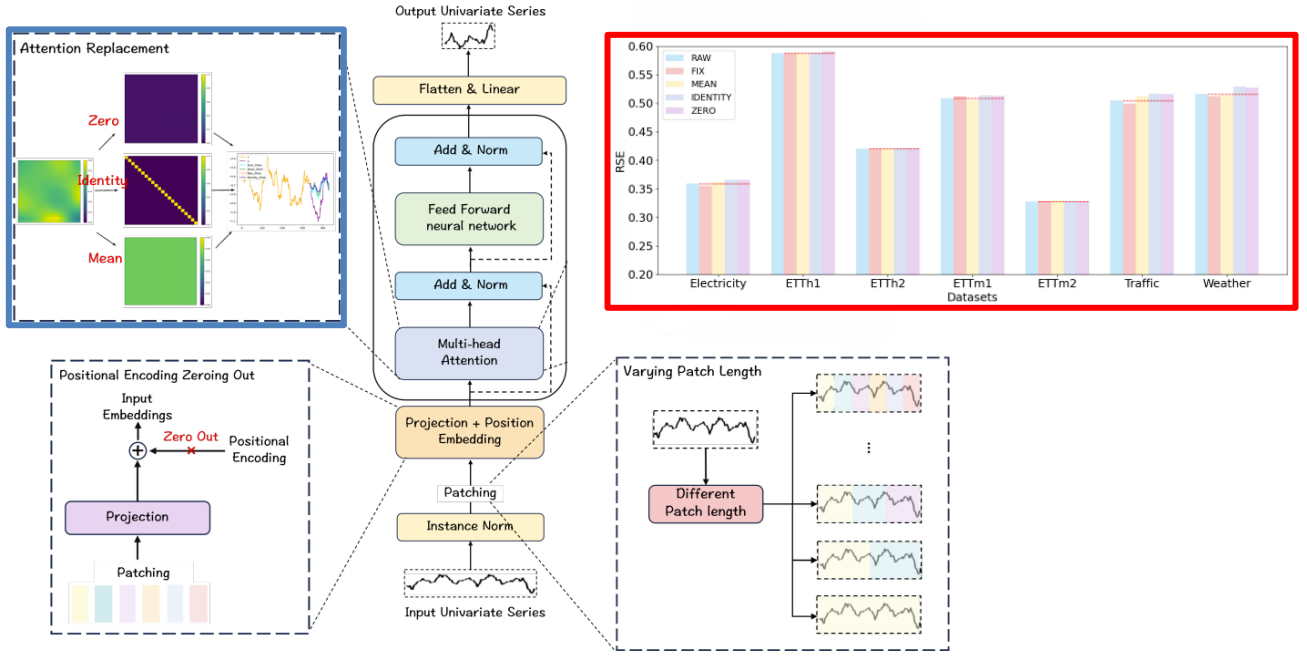


Figure 2: Experimental setup. This figure illustrates the four experiments conducted in this section.

- Attention Replacement
  - Attention 행렬을 영행렬(Zero), 단위행렬(Identity), 평균행렬(Mean), 고정 학습행렬(Fixed) 등으로 대체
  - 모델 성능에 유의미한 변화가 없거나 오히려 성능이 향상되는 모습을 보임

$$A = softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)\ Attention(Q,K,V) = AV,\ A \in \{\mathbf{0}, I, A_{mean}, A_{fixed}\}$$

*A : Attention 행렬
*QK^T: 토큰 간의 유사도
*d_k: Key 벡터의 차원
*$\sqrt{d_k}$: 안정적인 연산을 위한 스케일링 계수

Why Attention Fails

실험 방법

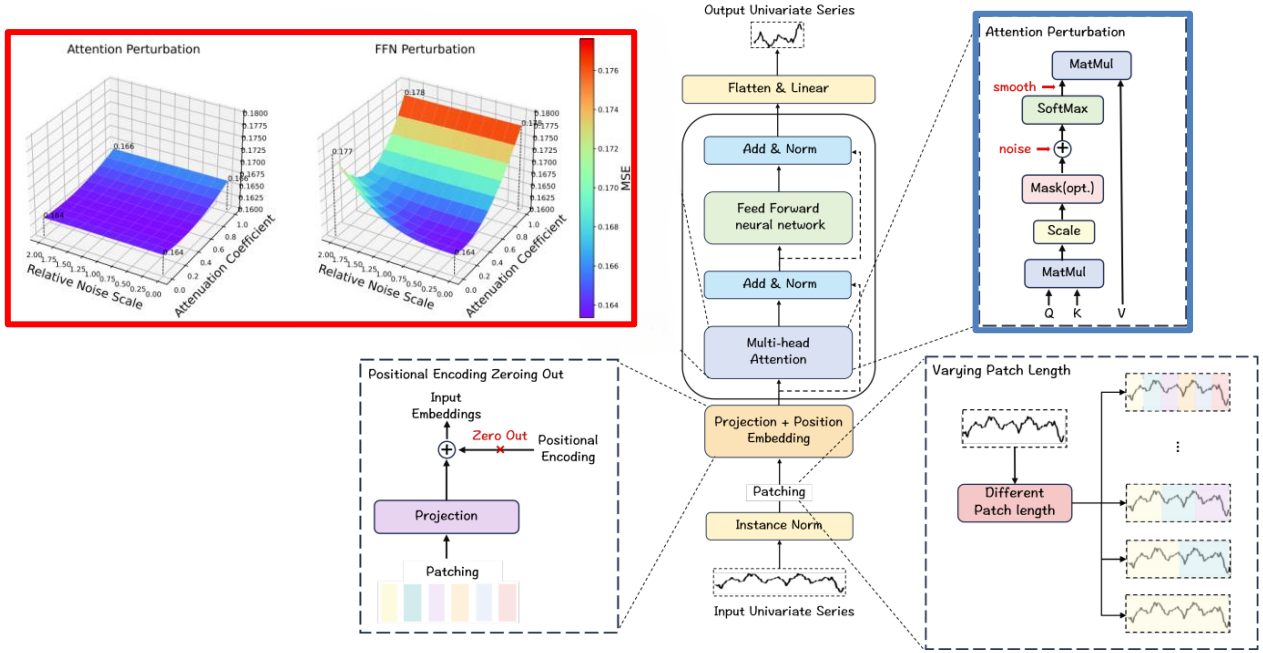


Figure 2: Experimental setup. This figure illustrates the four experiments conducted in this section.

- Attention Perturbation
  - 학습이 완료된 Multi-Head Attention과 Feed-Forward Network에 동일한 Perturbation을 적용하여 각 모듈이 예측에 얼마나 기여하는지 확인함

$$A = (1-\alpha)\cdot softmax\left(\frac{QK^{\top}}{\sqrt{d_k}} + \epsilon_{T\times T}\right) + \frac{\alpha}{T}\cdot 1_{T\times T}$$

*α: 섭동(노이즈)의 강도 조절 파라미터
*$\epsilon_{T\times T}$: 가우시안 노이즈 행렬
*$1_{T\times T}$: 모든 원소 1인 행렬, $Smoothing$ 역할

Why Attention Fails

실험 방법

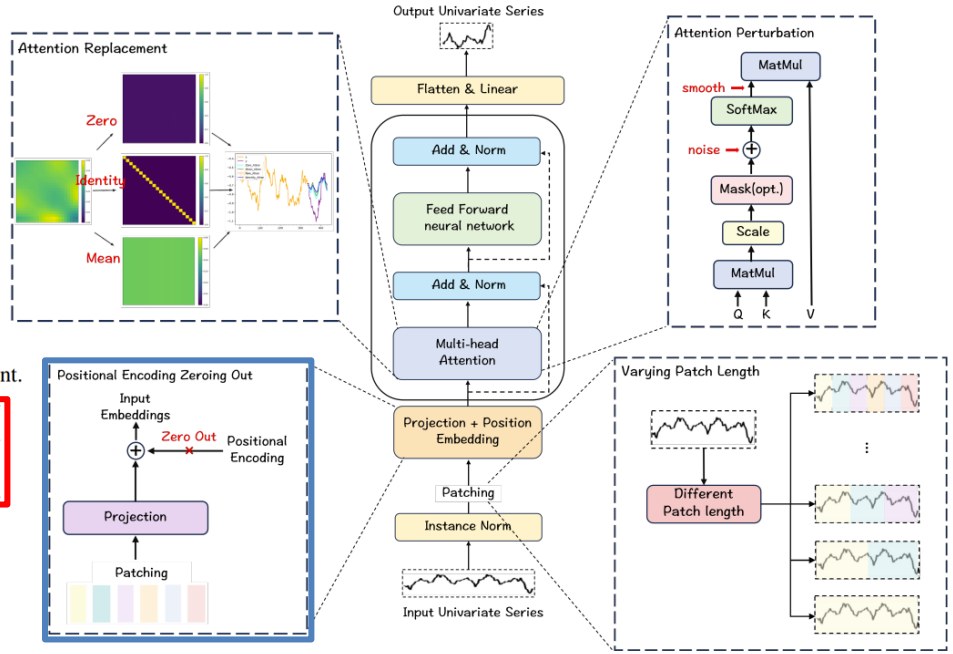


Figure 2: Experimental setup. This figure illustrates the four experiments conducted in this section.

Table 2: Positional encoding zeroing out experiment.

Model	Metric	PosEnc	Zero PosEnc
ViT	Accuracy	89.8%	32.4%
PatchTST	MSE	0.142	0.142

- Positional Encoding Zeroing Out
  - 학습이 완료된 모델에서 토큰 간의 시간적 위치 정보를 제공하는 Positional Encoding을 0으로 설정
  - Attention Mechanism은 순열 불변성(Permutation-invariant)을 가지기 때문에, 위치 정보가 없으면 시간적 의존성을 처리할 수 없어 성능이 급감해야 함.

Why Attention Fails

실험 방법

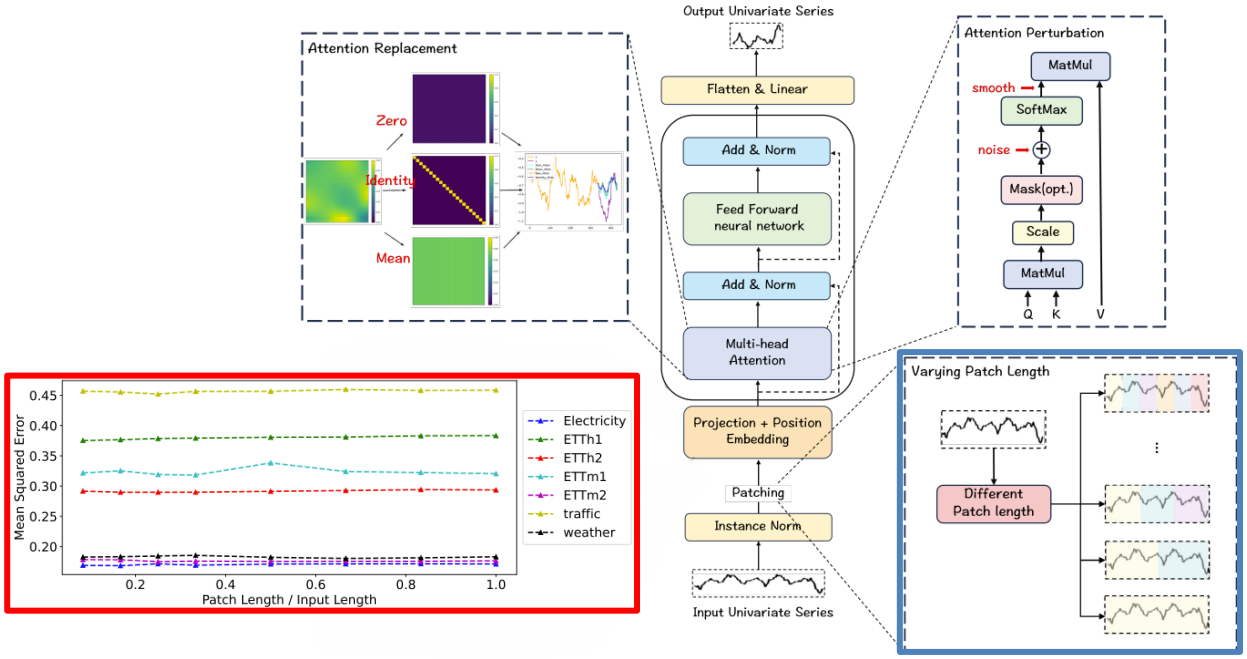


Figure 2: Experimental setup. This figure illustrates the four experiments conducted in this section.

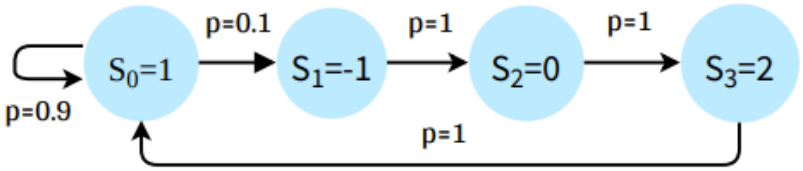
- Varying Patch Length
  - 패치 길이가 증가함에 따라 토큰의 수는 감소하며, 결과적으로 토큰 간 Attention의 영향력은 약화됨
  - 패치 길이를 입력 시퀀스 전체 길이와 동일하게 설정하여 모델을 단일 토큰 기반의 단순한 MLP 구조로 축소

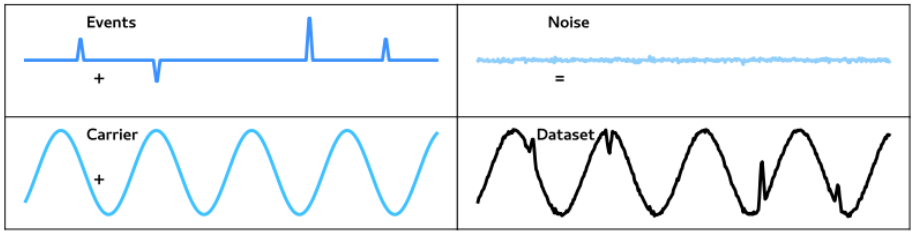
Why Attention Fails

FAIR Lab

KWANGWOON UNIVERSITY

Toy Dataset 설계





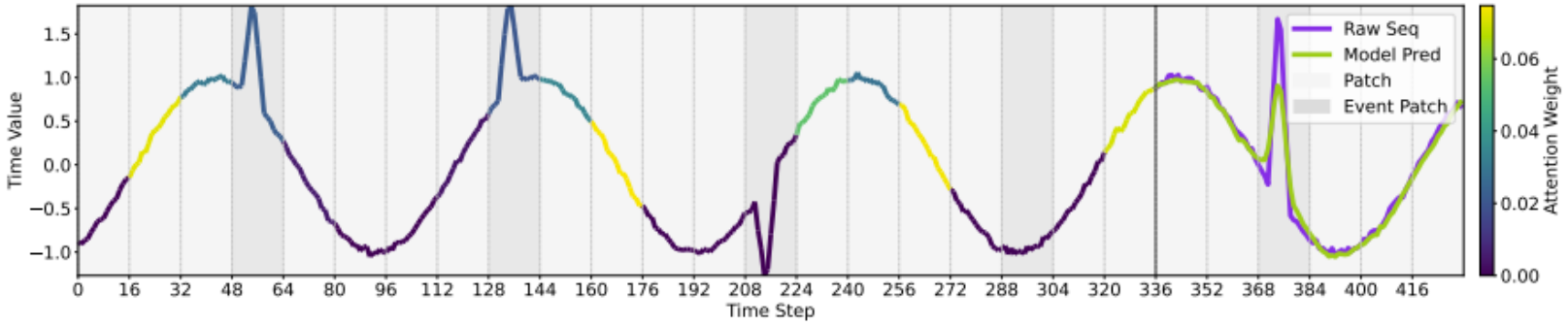
- 설계 목적
  - 실제 시계열 데이터는 복잡하여 패치 간의 관계를 정량화 하기 어려움, 따라서
    해석 가능성이 높은 State Machine 기반의 합성 데이터 생성 $x(t) = x_{carrier}(t) + x_{event}(t) + x_{noise}(t)$

Why Attention Fails

FAIR Lab

KWANGWOON UNIVERSITY

Toy Dataset 실험 결과



- 패치 간 의존성 포착 실패 분석
  - 예측에 가장 중요한 정보가 담긴 이벤트 패치 구간에서 Attention Weight가 0에 가깝게 수렴함
  - 선형 임베딩의 한계로 인해 토큰들이 의미 있는 상관관계를 가질 수 있는 유의미한 Latent Space 형성 실패

Why Attention Fails

결론 및 향후 방향

- 결론
  - Attention Mechanism 자체의 수정보다 Representation Learning의 개선이 우선됨
- 향후 방향
  - VQ-VAE와 같이 시계열을 Discrete Codebook으로 매핑하여 Attention이 더 효과적으로 작동할 수 있는 구조를 제안하였음

Why Attention Fails

실험 세팅

- 사용한 모델: PatchTST
- 재현 실험 데이터셋: ETTh1

Experiment	ETTh1
Learning rate	10^{-3}
Epoch	100
Batch size	128
Loss function	MSE
Seq_len	336
Pred_len	96
d_model	128
d_ff	256
Patch length	16

Why Attention Fails

FAIR Lab

KWANGWOON UNIVERSITY

재현 실험 (ETTh1)

- Attention Replacement를 적용한 실험 결과, 논문 수치와 유사한 수치가 나옴

	ETTh1 Paper		ETTh1 Reproduction	
Pred len(Mode)	MSE	MAE	MSE	MAE
96 (Raw)	0.404	0.418	0.398	0.425
96 (Identity)	0.381	0.403	0.384	0.402
96 (Zero)	0.384	0.405	0.381	0.399
96 (Mean)	0.392	0.412	0.394	0.415