Pyraformer: Low-Complexity Pyramidal Attention for Long-Range Time Series Modeling and Forecasting

https://openreview.net/forum?id=0EXmFzUn5I

0. Introduction

- 시계열 예측에서 긴 과거 구간의 의존성을 정확히 모델링하는 것이 중요
- 기존 Transformer 기반 모델은 시퀀스 길이가 길어질수록 계산 복잡도가 급격히 증가
- Self-attention 구조는 장기 의존성 표현에는 효과적
- 시간·메모리 복잡도가 $O(N^2)$로 증가하여 장기 시계열 예측에 비효율적
- 다중 해상도 기반 Pyramidal Attention Module (PAM) 제안
- 신호 전달 경로 길이를 상수 수준으로 줄이면서 계산 복잡도를 선형 수준으로 유지

1. Overview

- 핵심 아이디어

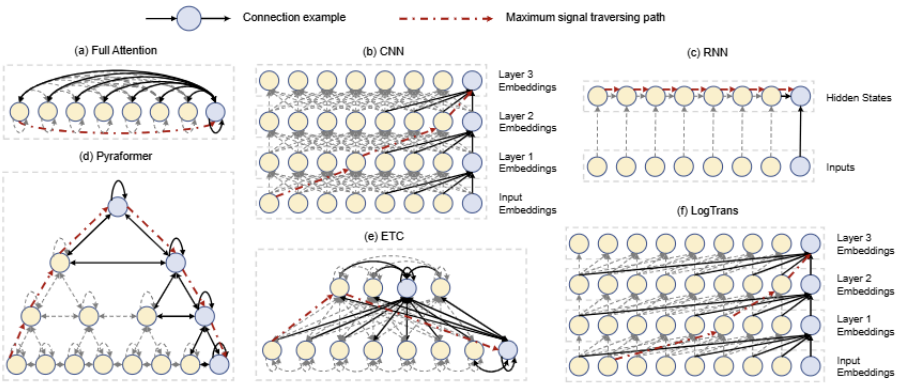
  시계열을 여러 해상도의 피라미드 구조로 표현하여, 서로 다른 시간 범위의 의존성을 효율적으로 학습

- 기본 구조
  - 시계열을 다중 스케일로 변환
  - 스케일 간 연결과 동일 스케일 내 이웃 연결을 동시에 사용
  - 전체 정보 전달 경로 길이를 입력 길이와 무관하게 상수로 유지
- 단기 예측과 장기 예측(long-range forecasting) 모두에서 성능 평가

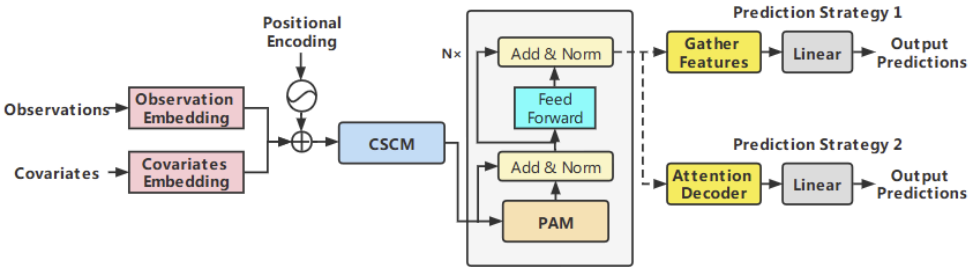
2. Challenges

- 장기 의존성 처리 문제
- 기존 self-attention은 시퀀스 길이가 길어질수록 계산량과 메모리 사용량이 급격히 증가
- 효율성과 정확도 균형
- 계산량을 줄이면 장기 패턴 학습 성능이 저하될 위험 존재
- 다중 스케일 표현 문제
- 서로 다른 시간 해상도의 정보를 효과적으로 통합하는 구조 설계가 어려움

3. Method



Method	Complexity per layer	Maximum path length
CNN (Munir et al., 2018)	$\mathcal{O}(L)$	$\mathcal{O}(L)$
RNN (Salinas et al., 2020)	$\mathcal{O}(L)$	$\mathcal{O}(L)$
Full-Attention (Vaswani et al., 2017)	$\mathcal{O}(L^2)$	$\mathcal{O}(1)$
ETC (Ainslie et al., 2020)	$\mathcal{O}(GL)$	$\mathcal{O}(1)$
Longformer (Beltagy et al., 2020)	$\mathcal{O}(L)$	$\mathcal{O}(L)$
LogTrans (Li et al., 2019)	$\mathcal{O}(L \log L)$	$\mathcal{O}(\log L)$
Pyraformer	$\mathcal{O}(L)$	$\mathcal{O}(1)$



- Pyramidal Attention Module (PAM)
  - 시계열을 피라미드형 계층 구조로 조직
  - 각 노드는 서로 다른 해상도의 시계열 요약 정보를 표현
  - Inter-scale connections
    - 서로 다른 스케일 간 정보 전달
    - 장기 의존성을 짧은 경로로 전달
  - Intra-scale neighbor connections
    - 동일 스케일 내 인접 시점 간 관계 학습
    - 국소 패턴 유지
- Coarser-Scale Construction Module (CSCM)
  - 원본 시계열을 점진적으로 더 거친 스케일로 요약
  - 피라미드 구조의 상위 노드 생성
- Prediction Module
  - PAM 출력 표현을 입력으로 받아 미래 시계열 값을 예측
  - 단일 스텝 및 다중 스텝 예측 모두 지원

---

4. Experiments

- 사용 데이터

  전력, 교통, 기후, 경제 등 공개 장기 시계열 벤치마크 데이터셋 사용
- 실험 설정
  - 비교 대상: 기존 Transformer 기반 시계열 모델 및 효율화 모델

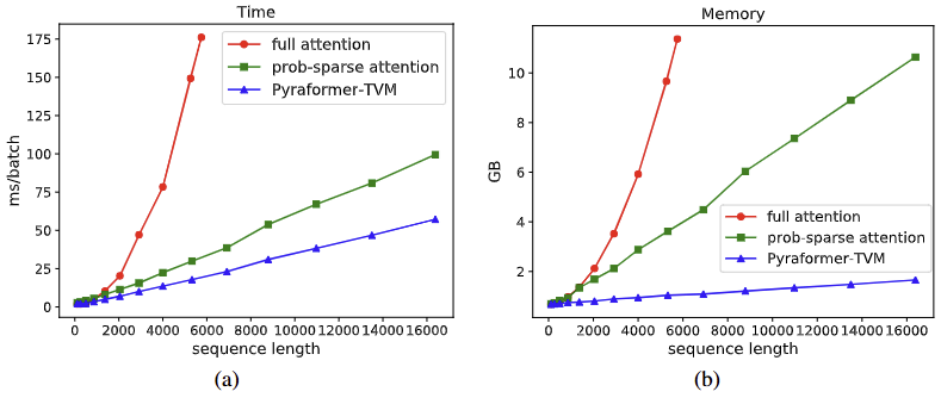
- 평가 지표: MAE, MSE, 장기 예측 정확도

- 추가 실험
  - PAM 구조 제거 실험으로 모듈 기여도 분석
  - 입력 길이 증가에 따른 시간·메모리 사용량 비교

5. Results

Methods	Parameters	Datasets	NRMSE	ND	Q-K pairs
Full-attention	$\mathcal{O}(N(HDD_K+DD_F))$	Electricity	0.328	0.041	456976
		Wind	0.175	0.082	589824
		App Flow	0.407	0.080	589824
LogTrans	$\mathcal{O}(N(HDD_K+DD_F))$	Electricity	0.333	0.041	50138
		Wind	0.173	0.081	58272
		App Flow	0.387	0.073	58272
Reformer	$\mathcal{O}(N(HDD_K+DD_F))$	Electricity	0.359	0.047	677376
		Wind	0.183	0.086	884736
		App Flow	0.463	0.095	884736
ETC	$\mathcal{O}(N(HDD_K+DD_F))$	Electricity	0.324	0.041	79536
		Wind	0.167	0.074	102144
		App Flow	0.397	0.069	102144
Longformer	$\mathcal{O}(N(HDD_K+DD_F))$	Electricity	0.330	0.041	41360
		Wind	0.166	0.075	52608
		App Flow	0.377	0.07	52608
Pyraformer	$\mathcal{O}(N(HDD_K+DD_F)$ $+(S-1)CD_K^2)$	Electricity	**0.324**	**0.041**	**17648**
		Wind	**0.161**	**0.072**	**20176**
		App Flow	**0.366**	**0.067**	**20176**

Methods	Metrics	ETTh1			ETTm1			Electricity		
		168	336	720	96	288	672	168	336	720
Informer	MSE	1.075	1.329	1.384	0.556	0.841	0.921	0.745	1.579	4.365
	MAE	0.801	0.911	0.950	0.537	0.705	0.753	0.266	0.323	0.371
	Q-K pairs	188040	188040	423360	276480	560640	560640	188040	188040	423360
LogTrans	MSE	0.983	1.100	1.411	0.554	0.786	1.169	0.791	1.584	4.362
	MAE	0.766	0.839	0.991	0.499	0.676	0.868	0.340	0.336	0.366
	Q-K pairs	74664	74664	216744	254760	648768	648768	74664	74664	216744
Longformer	MSE	0.860	0.975	1.091	0.526	0.767	1.021	0.766	1.591	4.361
	MAE	0.710	0.769	0.832	0.507	0.663	0.788	0.311	0.343	0.368
	Q-K pairs	63648	63648	249120	329760	1007136	1007136	63648	63648	249120
Reformer	MSE	0.958	1.044	1.458	0.543	0.924	0.981	0.783	1.584	4.374
	MAE	0.741	0.787	0.987	0.528	0.722	0.778	0.332	0.334	0.374
	Q-K pairs	1016064	1016064	2709504	5308416	14450688	14450688	1016064	1016064	2709504
ETC	MSE	1.025	1.084	1.137	0.762	1.227	1.272	0.777	1.586	4.361
	MAE	0.771	0.811	0.866	0.653	0.880	0.908	0.326	0.340	0.368
	Q-K pairs	125280	125280	288720	331344	836952	836952	125280	125280	288720
Pyraformer	MSE	**0.808**	**0.945**	**1.022**	**0.480**	**0.754**	**0.857**	**0.719**	**1.533**	**4.312**
	MAE	**0.683**	**0.766**	**0.806**	**0.486**	**0.659**	**0.707**	**0.256**	**0.291**	**0.346**
	Q-K pairs	**26472**	**26472**	**74280**	**57264**	**96384**	**96384**	**26472**	**26472**	**74280**



- 예측 성능
  - 장기 시계열 예측에서 기존 Transformer 및 효율화 모델 대비 더 높은 정확도 달성
  - 시퀀스 길이가 길어질수록 성능 우위가 뚜렷해짐
- 계산 효율성
  - 시간 및 메모리 복잡도가 입력 길이에 대해 **선형적으로 증가**
  - 기존 self-attention 대비 자원 사용량 크게 감소
- Ablation 결과
  - PAM 제거 시 장기 예측 성능 크게 하락
  - 다중 스케일 구조가 핵심 성능 요인으로 확인

6. Insight

- 연구 시사점
  - 다중 해상도 기반 hierarchical attention 구조는 장기 시계열 예측에서 매우 효과적
  - 계산 복잡도 문제를 구조적으로 해결한 점이 본 논문의 가장 큰 기여
- 피라미드 구조 설계와 하이퍼파라미터 선택이 성능에 민감할 가능성