시계열 예측 연구 서베이

스마트데이터연구실
최장호
10.10.24

시계열 데이터 예측 최신 연구 조사

- ICLR, ICML, NeurIPS 위주로 시계열 데이터 예측 최신 연구 조사
  - 최근 3개년도까지 확장 고려
  - 교통 예측 연구도 일부 포함 고려

2023	NeurIPS	BasisFormer: Attention-based Time Series Forecasting with Learnable and Interpretable Basis	어텐션 메커니즘, 해석 가능한 기저
2023	NeurIPS	Large Language Models Are Zero-Shot Time Series Forecasters	대규모 언어 모델, 제로샷 예측,
2023	NeurIPS	Predict, Refine, Synthesize: Self-Guiding Diffusion Models for Probabilistic Time Series Forecasting	확산 모델, 시계열 예측, 자기 유도 메커니즘
2024	ICLR	CARD: Channel Aligned Robust Blend Transformer for Time Series Forecasting	시계열, 채널 정렬
2024	ICLR	ModernTCN: A modern pure convolution structure for general time series analysis	합성곱, 시계열 분석
2024	ICLR	TimeMixer: Decomposable Multiscale Mixing for Time Series Forecasting	다중 스케일, 시계열 예측
2024	ICLR	TEMPO: Prompt-based Generative Pre-trained Transformer for Time Series Forecasting	프롬프트 기반, 생성형 모델
2024	ICLR	Generative Learning for Financial Time Series with Irregular and Scale-Invariant Patterns	금융 시계열, 생성적 학습
2024	ICLR	Soft Contrastive Learning for Time Series	대조 학습, 시계열
2024	ICLR	DAM: A foundation model for forecasting	기초 모델, 예측
2024	ICLR	iTransformer: Inverted Transformers Are Effective for Time Series Forecasting	역전된 트랜스포머, 시계열 예측
2024	ICML	Revitalizing Multivariate Time Series Forecasting: Learnable Decomposition with Inter-Series Dependencies and Intra-Series Variations Modeling	학습 가능한 분해, 다변량 시계열,
2024	ICML	RobustTSF: Towards Theory and Design of Robust Time Series Forecasting with Anomalies	강건한 시계열 예측, 이상치 처리

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions

History of attention in machine learning

- Learning to combine foveal gilpses with a third-order Boltzmann machine, Hugh Larochelle and Geoffrey Hinton, NeurIPS
- Learning where to Attend with Deep Architectures for Image Tracking, Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas,

- Sequence to Sequence Learning with Neural Networks, Ilya Sutskever, Oriol Vinyals, Quoc V. Le
- On the Properties of Neural Machine Translation: Encoder-Decoder Approaches, Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, Yoshua Bengio, 2014, SSST-8
- Neural Machine Translation by Jointly Learning to Align and Translate, Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio, ICLR 2015

- **Attention is all you need (Transformer)**
- BERT, GPT2, Transformer-XL…

Self-Attention (Intra-Attention)

- 주의 메커니즘은 단일 시퀀스의 서로 다른 위치들을 관련지어 동일 시퀀스의 표현을 계산하는 방식
- 이 메커니즘은 기계 독해, 추상적 요약, 이미지 설명 생성 등에서 매우 유용한 것으로 입증 됨

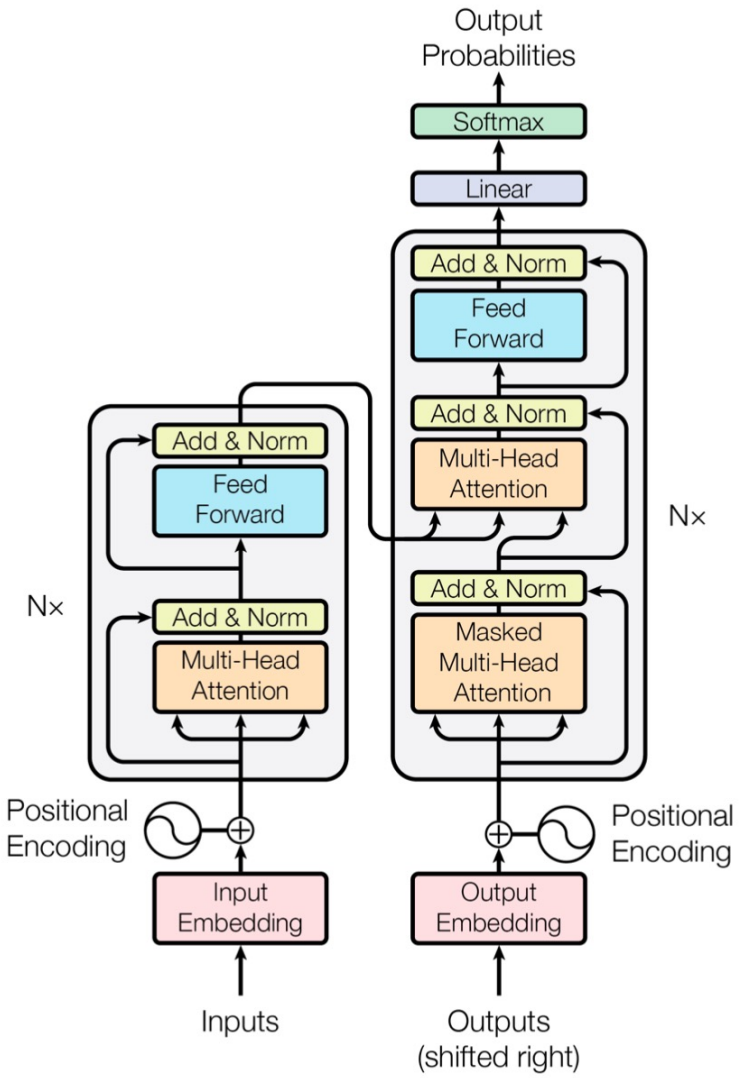
- **왜 Self-Attention인가?**
  - 각 층당 전체 계산 복잡도를 최소화
  - 병렬로 처리 가능한 계산의 양을 최대화하고, 필요한 순차 연산의 최소 수를 기준으로 측정
  - 서로 다른 층 유형으로 구성된 네트워크에서 두 입력 및 출력 위치 사이의 최대 경로 길이를 최소화

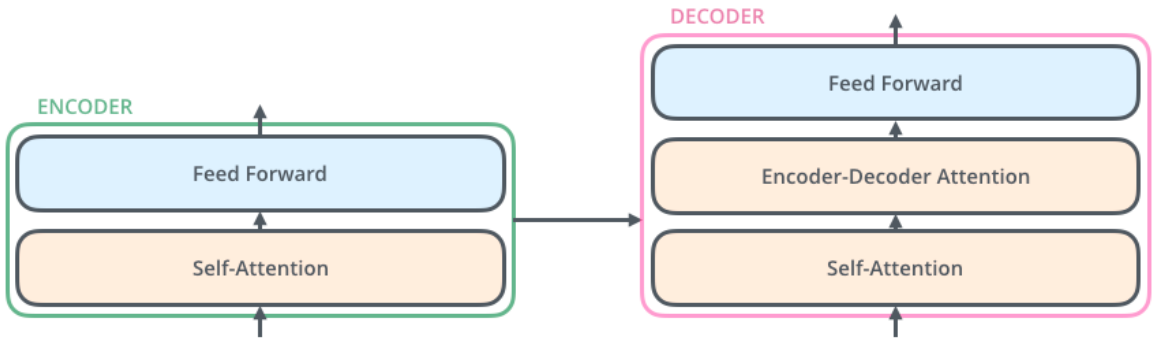
Transformer

- Core Ideas
  - Depth with non-linearities and gradient descent
  - Model dependencies over the whole range of the input sequence
  - Considering words positions
  - No recurrent connections
  - Whole model can be computed in an efficient feed-forward fashion

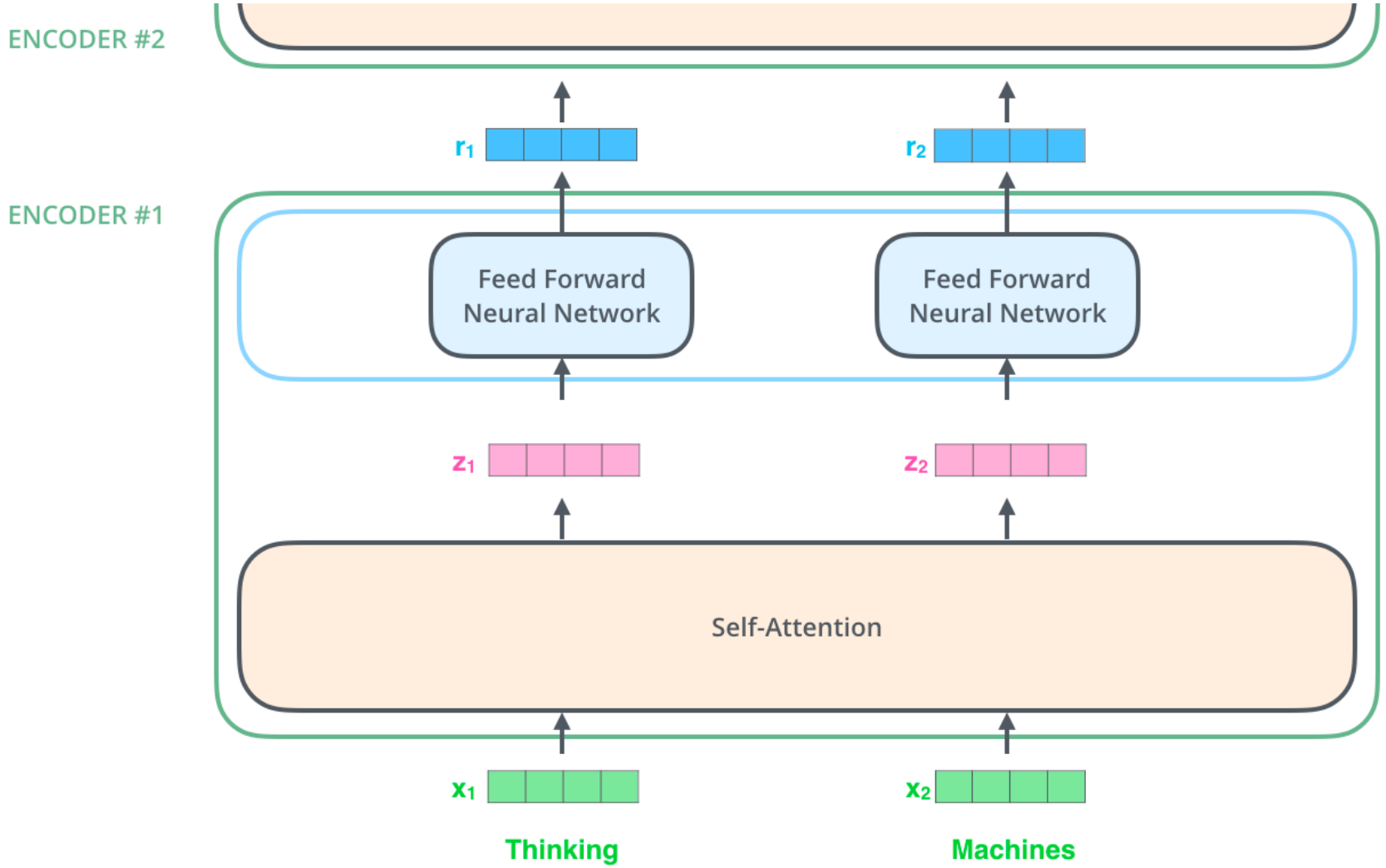
- Solutions
  - Attention is a strong enough mechanism to learn
  - key, query, and value are all the same vectors (with minor linear transformation)
  - attend to themselves and stacking such self-attention provides sufficient nonlinearity and representational power to learn very complicated functions

Transformer – 구조

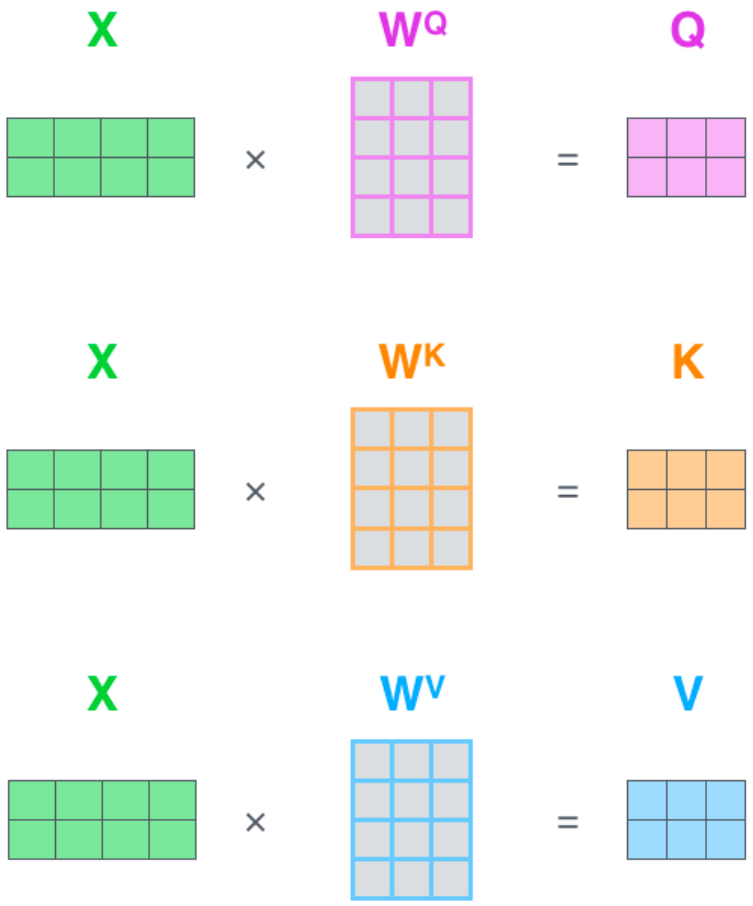




Transformer – Encoder



Transformer – Self-attention



$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) V = Z$$

Transformer– Multi-head Attention

1) This is our input sentence*

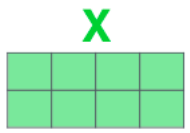
2) We embed each word*

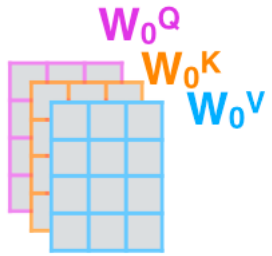
3) Split into 8 heads. We multiply X or R with weight matrices

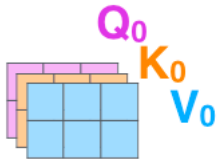
4) Calculate attention using the resulting Q/K/V matrices

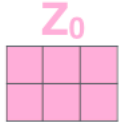
5) Concatenate the resulting Z matrices, then multiply with weight matrix W^O to produce the output of the layer

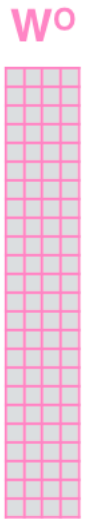
Thinking Machines

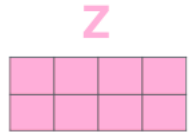


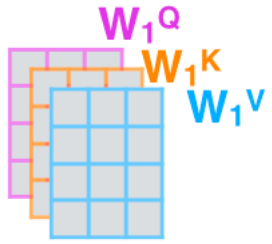


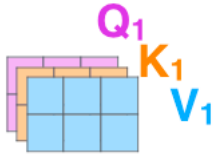


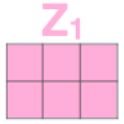




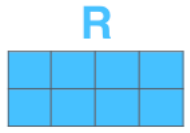


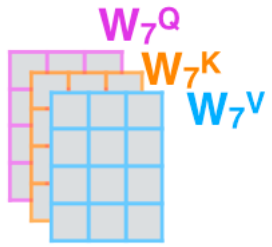


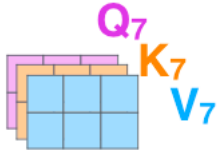


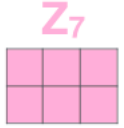


* In all encoders other than #0, we don't need embedding. We start directly with the output of the encoder right below this one

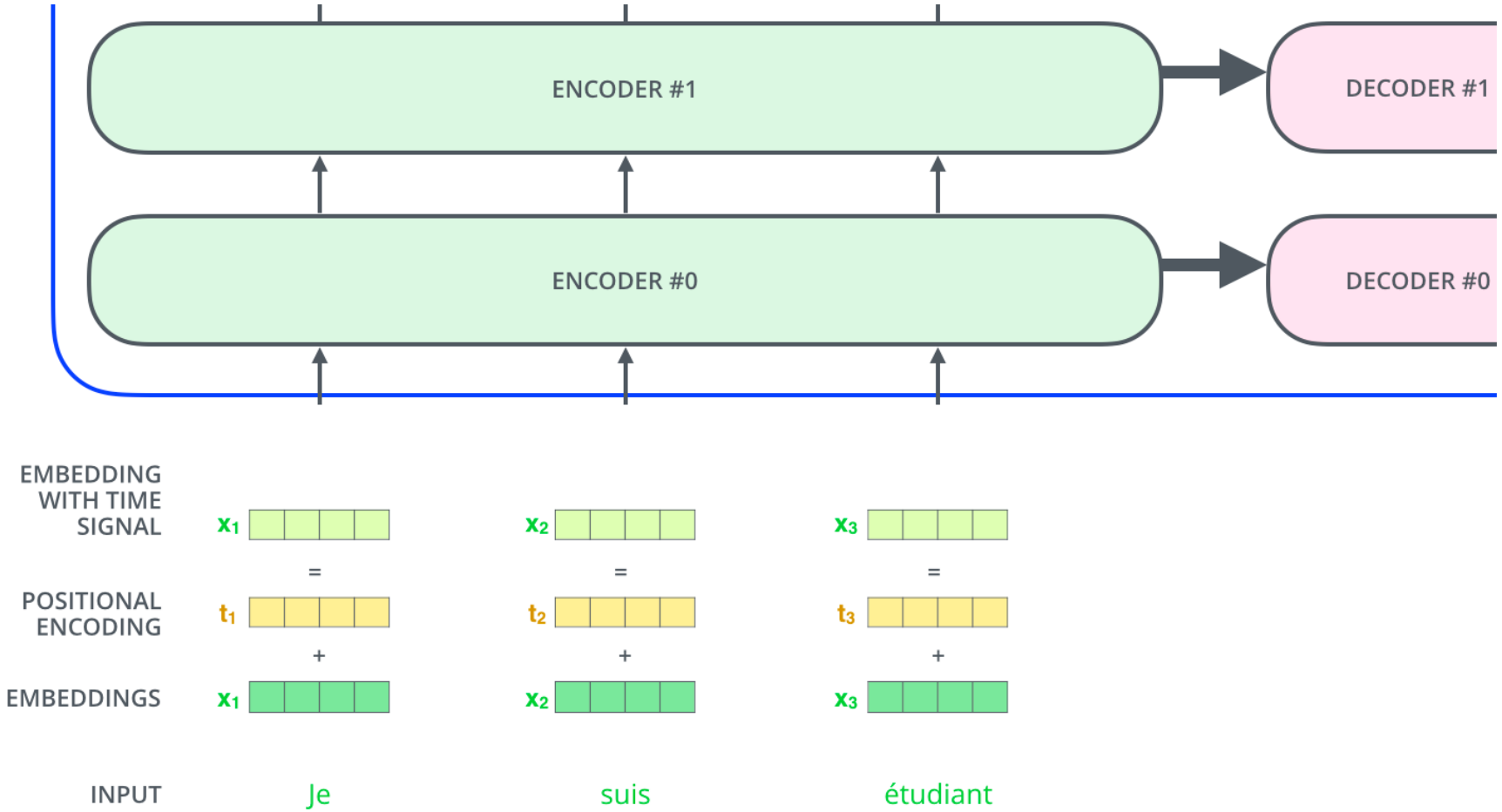




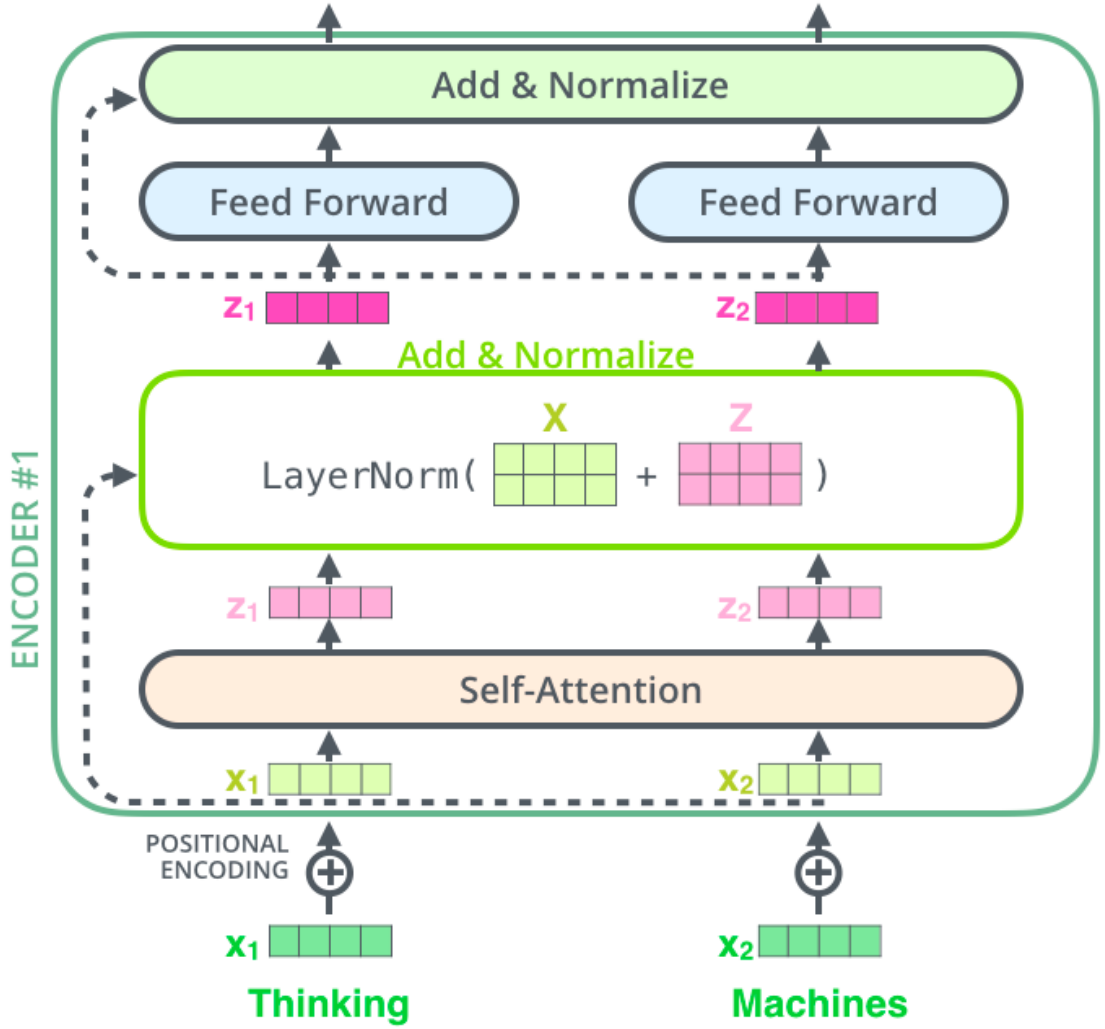




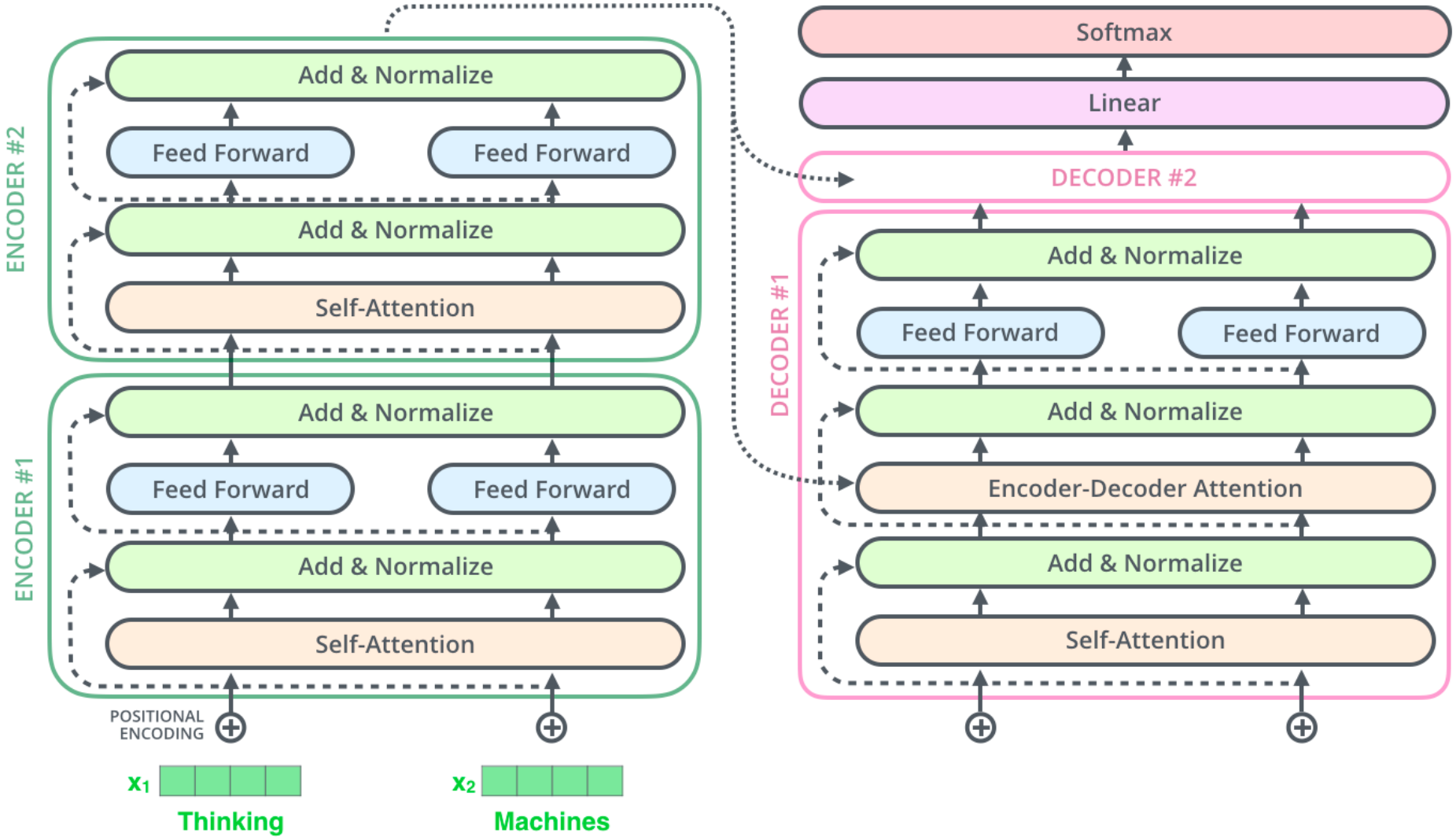
Positional Encoding



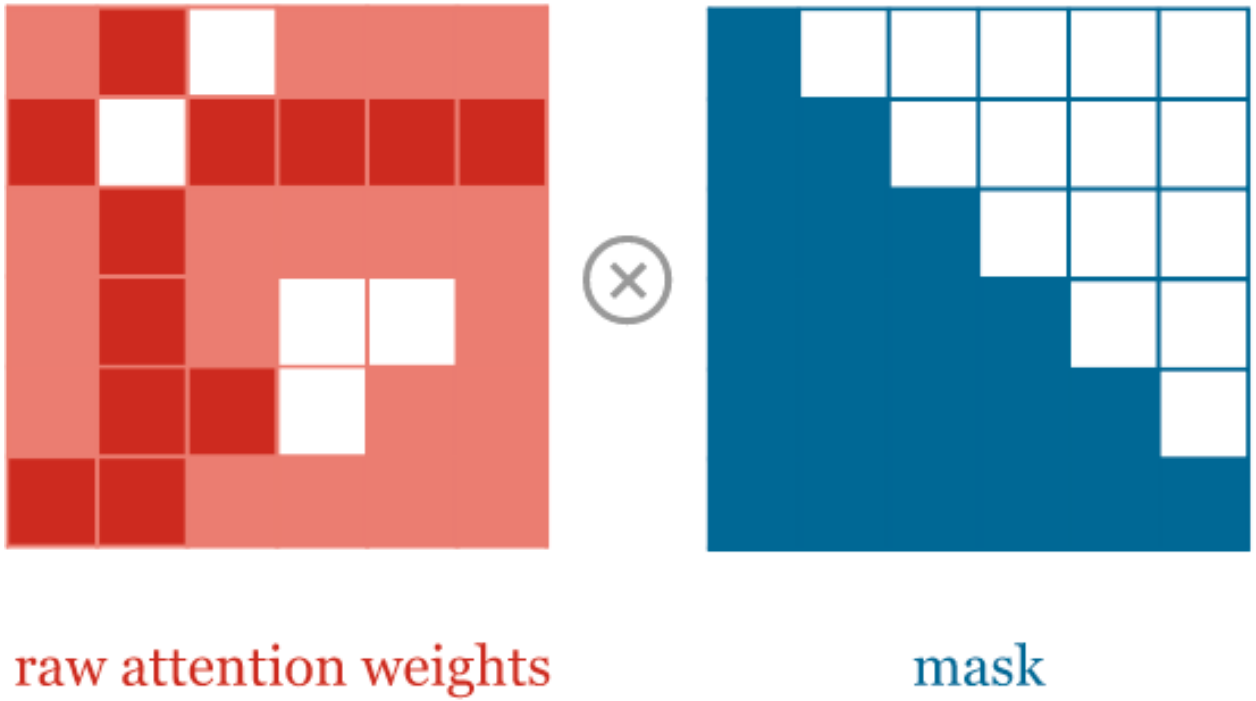
Transformer - Residuals

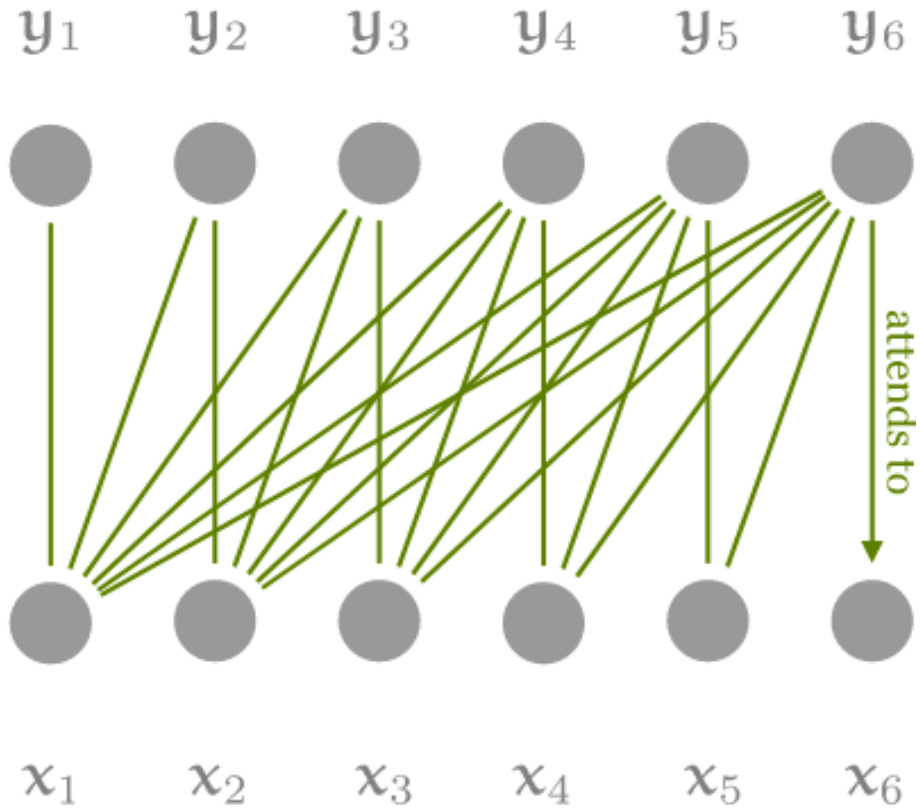


Transformer – Decoder



Transformer – Masking Self-Attention





Informer

- Informer는 장기 시계열 예측을 위해 특별히 설계된 고급 딥러닝 모델로, Zhou 등(2021)의 논문 Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting에서 소개되었습니다.
- Informer의 주요 동기는 긴 시계열 데이터를 처리할 때, 기존의 Transformer 아키텍처의 효율성과 효과를 개선
- Informer의 주요 특징:
  - **긴 시퀀스 처리**: Informer는 전통적인 Transformer 모델이 계산 과부하와 비효율성으로 어려움을 겪는 긴 시계열 예측 작업에서 특히 효과적
  - **효율적인 어텐션 메커니즘**: Informer는 Transformer의 자기 어텐션 메커니즘에서 발생하는 이차 복잡성 문제를 해결하기 위해 ProbSparse Attention이라는 효율적인 어텐션 메커니즘을 제안하여 계산 부하를 크게 줄임
    - ProbSparse Attention: 긴 시퀀스에서 각 토큰에 대한 어텐션 점수를 계산하는 대신, Informer는 가장 정보가 많은 일부 쿼리에 대해서만 어텐션을 선택적으로 계산
  - Self-Attention Distillation: 계층 간 유사한 어텐션 점수를 집계하여 중복 정보를 줄임
  - **생성적 디코더**: Informer는 생성적 스타일의 디코더를 사용하여 시계열 데이터를 효율적으로 예측하며, 관련된 입력 데이터에 집중함으로써 예측 성능을 더욱 향상시킴
  - **O(log L) 시간 복잡도**: ProbSparse 어텐션 메커니즘 및 기타 최적화 덕분에 Informer는 시퀀스 길이에 대해 로그 시간 복잡도를 달성하여, 매우 긴 시퀀스를 처리해야 하는 작업에서도 높은 확장성을 제공

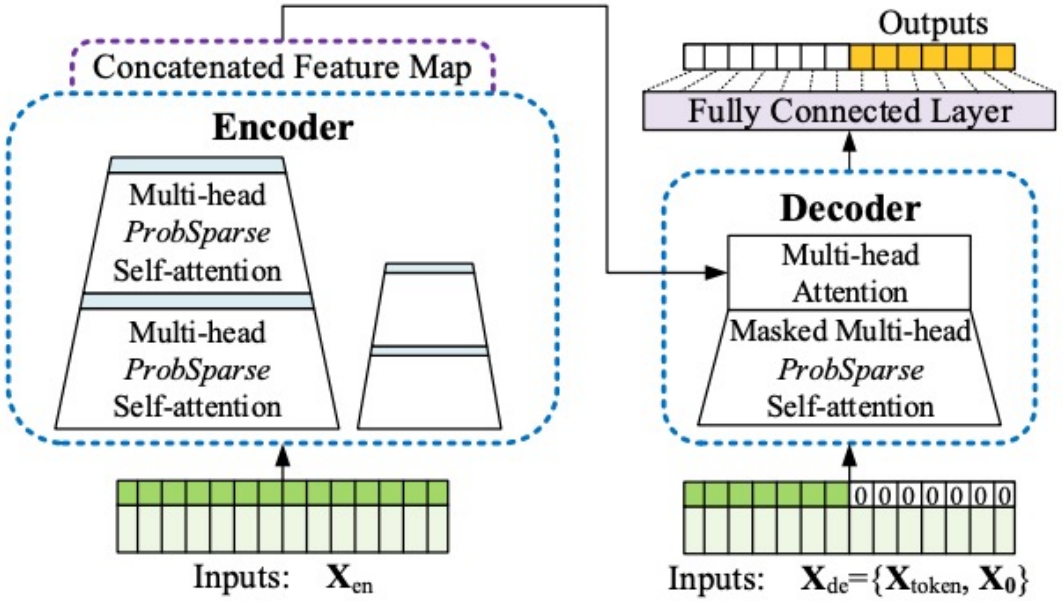
Informer – 구조

인코더:

- Informer의 인코더는 입력된 시계열 데이터를 처리하고 의미 있는 패턴과 트렌드를 추출합
- **ProbSparse 어텐션 메커니즘**을 활용하여 가장 "정보성 있는" 쿼리를 선택함으로써 계산량을 줄임
- 인코더는 입력 시퀀스의 표현을 학습하며, 이 표현은 단기 및 장기 의존성을 모두 포착

디코더:

- 디코더는 **생성적 스타일의 디코더**로, 입력 데이터의 인코딩된 표현을 바탕으로 미래 시계열 포인트를 예측하도록 설계
- 디코딩 과정은 일반적으로 생성적 프로세스의 자기회귀적 특성을 사용하여 미래 값을 예측하는 방식이지만, 긴 예측을 위해 최적화 됨:
  - 예측된 시퀀스에서 이전 단계에 주목하는 **멀티헤드 자기 어텐션 층**
  - 인코더에서 나온 관련된 출력을 참조하는 **크로스 어텐션 층**
  - 주목된 특징을 처리하고 예측된 시계열 값을 출력하는 **피드포워드 층**



Informer – ProbSparse Attention

- **Kullback-Leibler (KL) Divergence as a Proxy:**
  - 어떤 쿼리가 가장 정보성 있는지를 결정하기 위해, ProbSparse 어텐션은 각 쿼리의 어텐션 가중치 확률 분포와 균등 분포 간의 Kullback-Leibler(KL) 발산을 사용
  - 발산이 높은 쿼리(즉, 그들의 어텐션 분포가 균등하지 않은 쿼리)는 특정 키-값 쌍에 강한 어텐션을 보이기 때문에 정보성 있는 것으로 간주
  - KL 발산 값이 가장 높은 상위 K개의 쿼리만 선택함으로써, ProbSparse 어텐션은 어텐션 계산의 수를 줄입니다.

Informer – 부족한 부분?

- ProbSparse
  - 선택적 어텐션 메커니즘은 중요한 정보를 간과할 가능성
  - 매우 복잡하거나 노이즈가 많은 데이터셋에서는 작은 세부 사항도 정확한 예측을 위해 중요할 수 있음
- 짧은 시퀀스에 대한 제한된 적용성?
  - ProbSparse 어텐션 메커니즘을 사용하는 오버헤드

BasisFormer: Attention-based Time Series Forecasting with Learnable and Interpretable Basis

저자: Zelin Ni, Hang Yu, Shizhan Liu, Jianguo Li, Weiyao Lin
발표: NeurIPS 2023 메인 컨퍼런스 트랙

- 기존의 시계열 예측 모델들은 특정 데이터셋에 맞춤화된 기저를 생성하거나 각 시계열과 뚜렷한 상관관계를 가지는 기저를 동시에 만족시키는 데 한계가 있었음
- 시계열 예측을 위한 효과적이고 해석 가능한 기저(basis)를 학습하는 end-to-end 아키텍처를 제안
- 제안 모델 (BasisFormer):
  - Basis 모듈: 시계열의 과거와 미래 부분을 두 개의 다른 뷰로 취급하고 대조 학습을 사용하여 기저를 획득
  - Coef 모듈: 양방향 교차 주의 메커니즘을 통해 시계열과 과거 뷰의 기저 간 유사도 계수를 계산
  - Forecast 모듈: 유사도 계수를 기반으로 미래 뷰의 기저를 선택하고 통합하여 정확한 미래 예측을 수행
- 학습 가능하고 해석 가능한 기저 사용, end-to-end 아키텍처로 설계, 단변량 및 다변량 시계열 예측 모두에 효과적
- **성능**: 6개 데이터셋에 대한 광범위한 실험을 통해 검증
  - 단변량 예측 작업에서 기존 최고 성능 대비 11.04% 향상
  - 다변량 예측 작업에서 기존 최고 성능 대비 15.78% 향상
- 코드 공개: https://github.com/nzl5116190/Basisformer

BasisFormer – 구조

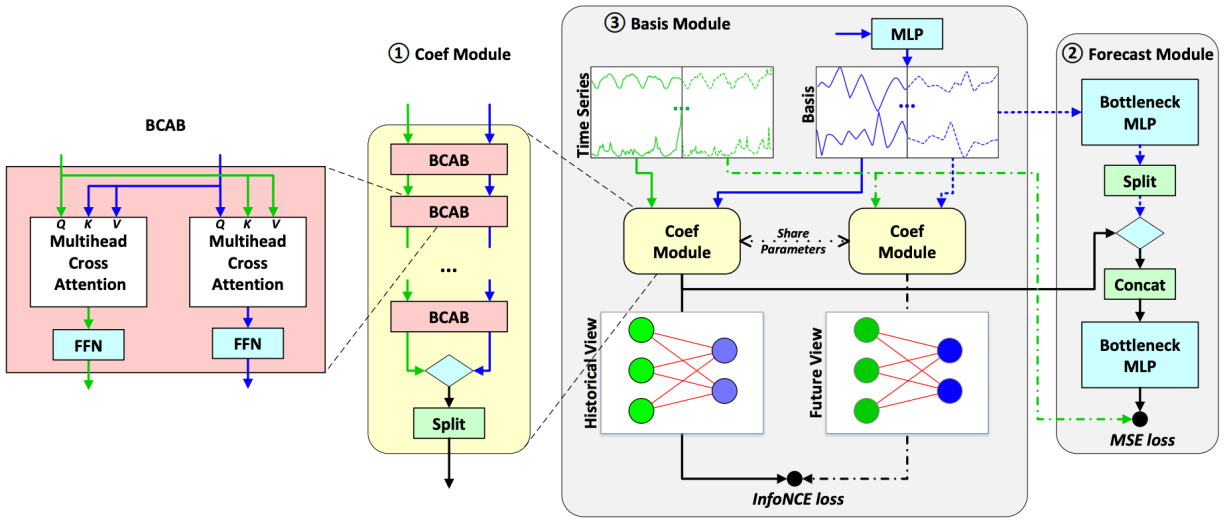


Figure 1: The architecture of BasisFormer, consisting of ① the Coef module, ② the Forcast module, and ③ the Basis module. The green and blue lines denote the data flow of the set of time series and basis vector repespetively. The cyan diamond denotes tensor dot product. Note that the dot-dash line, which denotes the data flow of the future part of the time series, is only included during training but removed during inference.

- 이 구조를 통해 BasisFormer는 시계열 데이터의 특성을 효과적으로 포착하고, 과거 패턴을 기반으로 미래를 정확하게 예측
- 특히, 학습 가능한 기저를 사용함으로써 모델의 적응성과 해석 가능성을 동시에 향상

Basis 모듈:적응형 자기 지도 학습을 통해 기저(basis)를 획득
- 시계열의 과거와 미래 부분을 두 개의 다른 뷰로 취급합니다.
- 대조 학습(contrastive learning)을 사용하여 기저를 학습합니다.
- 두 가지 손실 함수를 사용합니다: InfoNCE 손실과 smoothness 손실

Coef 모듈:시계열과 기저 벡터 간의 유사도를 측정
- 양방향 교차 주의(bidirectional cross-attention) 메커니즘을 사용합니다.
- 시계열과 기저를 각각 그래프의 노드로 표현하는 이분 그래프(bipartite graph)를 활용합니다.
- 과거 뷰의 시계열과 기저 간 유사도 계수를 계산합니다.

Forecast 모듈: 선택된 기저를 바탕으로 정확한 미래 예측을 수행
- Coef 모듈에서 계산된 유사도 계수를 기반으로 작동
- 미래 뷰의 기저를 선택하고 통합

End-to-end 아키텍처: 모든 모듈이 통합되어 한 번에 학습
학습 가능한 기저: 데이터셋의 특성을 정확히 반영하는 기저를
학습해석 가능한 구조: 각 모듈의 역할이 명확하게 정의되어 있어 모델의 동작을 이해하기 쉬움
유연한 적용: 단변량 및 다변량 시계열 예측 모두에 효과적으로 적용 가능

BasisFormer – Cont’d

해결하려는 문제점:

- 기존의 시계열 예측 모델들은 특정 데이터셋에 맞춤화된 기저(basis)를 생성하거나 각 시계열과 뚜렷한 상관관계를 가지는 기저를 동시에 만족시키는 데 한계가 있음
- 효과적인 시계열 예측을 위해서는 데이터셋의 특성을 정확히 반영하면서도 각 시계열과 명확한 상관관계를 가지는 기저가 필요함

주요 기여점:

- 학습 가능하고 해석 가능한 기저를 활용하는 end-to-end 시계열 예측 아키텍처인 BasisFormer를 제안
- 적응형 자기 지도 학습을 통해 시계열의 과거와 미래 부분을 두 개의 다른 뷰로 취급하고 대조 학습을 사용하여 기저를 획득하는 방법을 개발
- 양방향 교차 주의 메커니즘을 통해 시계열과 과거 뷰의 기저 간 유사도 계수를 계산하는 Coef 모듈을 설계
- 유사도 계수를 기반으로 미래 뷰의 기저를 선택하고 통합하여 정확한 미래 예측을 수행하는 Forecast 모듈을 개발
- 6개 데이터셋에 대한 광범위한 실험을 통해 단변량 예측 작업에서 기존 최고 성능 대비 11.04%, 다변량 예측 작업에서 15.78%의 성능 향상을 달성

BasisFormer – 부족한 부분?

부족한 부분:

- 논문에서 직접적으로 언급된 한계점은 없지만, 모델의 복잡성이 증가함에 따라 계산 비용과 학습 시간이 증가
- 제안된 모델의 해석 가능성에 대한 더 깊이 있는 분석이 필요
- 다양한 도메인과 더 큰 규모의 데이터셋에 대한 추가적인 검증이 필요

의견:

- “Interpretable " 이라고 할 수 있을까?