1D CNN-GRU Cross Learning for Multivariate TSF

(Time Series Forecasting)

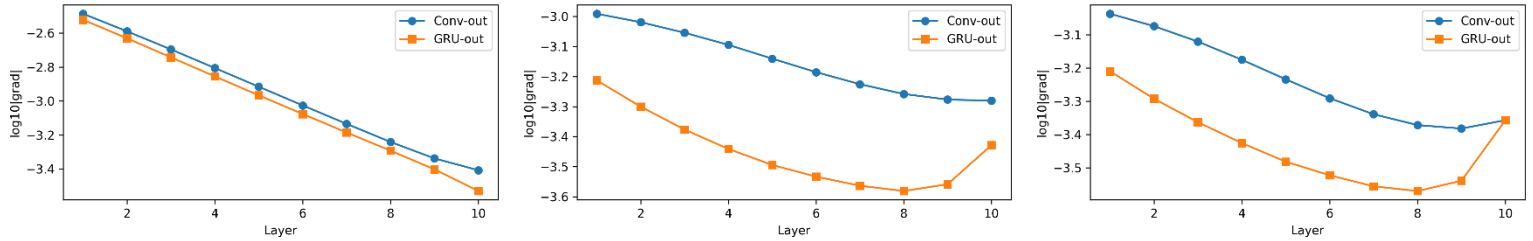
연세대학교 지능형데이터·최적화학과
System Intelligence Lab
박힘찬

Table of Contents

Q1. 핵심 설계 철학

제안 모델의 설계 철학이 무엇인가?

- 시계열 데이터는 **짧은 구간의 모양** (국소 패턴) 과 **긴 흐름** (문맥) 이 동시에 중요함
- **CNN** 는 얕은 층에서는 국소 패턴을, 층이 깊어질수록 큰 구간에서의 흐름도 함께 보게 됨
- **GRU** 는 얕은 층에서는 단기 변동을, 층이 깊어질수록 시계열 데이터의 문맥을 안정적으로 반영하게 됨
- 이 둘을 **최하단에서만 합치면 "국소 패턴" 과 "문맥" 을 따로 학습하게 되어** CNN 은 문맥을 알 수 없어 어느 모양을 더 강조해야 하는지 알지 못하고, GRU 는 국소 패턴을 알 수 없어 급변 직후 무엇이 바뀌었는지 변화에 대한 단서를 제때 받지 못하게 됨
- 실제로 층마다 연결을 끊고 끝에서만 합치거나 두 경로를 평균할 경우 RMSE 가 일관되게 나빠지고, Gradient 흐름도 불균형해지는 것을 확인할 수 있었음



ETTm2 데이터셋에 대한 Gradient Flow 그림, 왼쪽에서부터 제안 모델, ① Late-Fusion (끝에서만 합침), ② No-Fusion (두 경로 산술평균) 모델의 출력

Q1. 핵심 설계 철학

제안 모델의 설계 철학이 무엇인가?

- 제안 모델은 **반대 경로가 보고 있는 '지금의 상태' 를 이용**해 변수 별로 필터를 매번 새로 만들어 **각 정보를 가볍게 보정**하고, 그 **소량을 서로 가중 합**하여 **국소 무늬 (CNN) 를 문맥 (GRU) 에 맞춰 다듬고, 문맥 (GRU) 을 국소 무늬 (CNN) 에 맞추어 보강**하는 것을 **층마다 반복**하도록 만듦
- CNN 신호를 GRU 에 조금씩 섞으면 GRU 의 업데이트/망각 판단이 주입된 피크, 경사, 꺾임 등의 모양 (CNN) 신호를 토대로 더 분명하게 이루어짐
- GRU 신호를 CNN 에 조금씩 섞으면 CNN 가 문맥 정보 (GRU) 를 이용하여 지금 시간대 / 상황이 중요한지 강조하거나 줄일 수 있음
- 서로 간의 정보 참조를 약화시키거나 제거할 경우 성능 악화가 발생한다는 것이 간접 증거임

Datasets	Phase-Drift (↓= 정합↑)	SPAI (↑= 정합↑)	Δ RMSE % @ H = 720 (EMA)	Δ RMSE % @ H = 720 (Static-DW)
ETTh2	**0.0417**	**0.8193**	+ 18.5 %	+ 17.5 %
ETTm2	**0.0208**	**0.8221**	+ 8.8 %	+ 8.2 %
ETTh1	0.0833	0.6530	–	+ 3.3 %
ETTm1	0.0885	0.6496	–	+ 1.8 %
Weather	0.0486	0.7675	+ 4.3 %	+ 1.4 %

Q1. 핵심 설계 철학

제안 모델의 설계 철학이 무엇인가?

- 실험을 통해 확인해 본 결과,
- GRU 신호를 CNN 에 섞으면 오차가 안정적으로 감소하고, 필터/출력이 시간대에 맞춰 다르게 동작하는 것을 확인할 수 있었음

Datasets	Δ RMSE % (GRU → CNN)	Δ RMSE % (Both)	dCor(Conv 커널 계수, TOD) (GRU → CNN)	dCor(Conv 출력 변화, TOD) (GRU → CNN)
ETTh2	+ 1.8 %	+ 8.3 %	0.11	0.19
ETTm2	+ 1.7 %	+ 5.9 %	0.04	0.10
ETTh1	+ 0.1 %	+ 1.9 %	0.18	0.34
ETTm1	+ 1.3 %	+ 2.2 %	0.07	0.16
Weather	+ 2.0 %	+ 2.4 %	0.19	0.18

Datasets	Δ RMSE % (GRU → CNN)	Δ RMSE % (Both)	dCor(Conv 커널 계수, TOD) (Both)	dCor(Conv 출력 변화, TOD) (Both)
ETTh2	+ 1.8 %	+ 8.3 %	0.44	0.61
ETTm2	+ 1.7 %	+ 5.9 %	0.37	0.32
ETTh1	+ 0.1 %	+ 1.9 %	0.71	0.83
ETTm1	+ 1.3 %	+ 2.2 %	0.79	0.58
Weather	+ 2.0 %	+ 2.4 %	0.61	0.48

- Δ RMSE : (none − mode) / none * 100 으로, 클 수록 좋음
- dCor (distance correlation, TOD (하루 주기), 범위 0~1) : 값이 클수록 필터의 모양/세기나 Conv 출력의 크기/형태가 더 뚜렷하게 변화하여, 문맥에 맞춘 필터/출력 조정이 강하게 일어났음을 의미함

Q1. 핵심 설계 철학

제안 모델의 설계 철학이 무엇인가?

- 실험을 통해 확인해 본 결과, CNN 신호를 GRU 에 섞으면 단독 성능은 떨어지는 것으로 확인됨
- 그러나 CNN ↔ GRU 상호 섞음의 경우, 피크/꺾임과 같은 급변 상황에서 GRU 의 반응이 더 크고, 더 빨리 반응하는 것으로 확인됨, 단독 보다 양방향에서 효과가 안정적임

Datasets	Δ RMSE % (CNN→GRU)	Δ RMSE % (Both)	Spearman rank corr. 증가량 (CNN→GRU)	피크 반응 증폭률 % (CNN→GRU)	반응 선행/지연 (스텝) (CNN→GRU)
ETTh2	+ 0.9 %	+ 8.3 %	– 0.03	– 0.9 %	+ 0.96
ETTm2	– 0.6 %	+ 5.9 %	– 0.03	– 3.2 %	+ 0.91
ETTh1	– 1.0 %	+ 1.9 %	+ 0.05	+ 0.2 %	+ 0.91
ETTm1	– 1.2 %	+ 2.2 %	+ 0.03	+ 3.4 %	+ 0.36
Weather	– 0.6 %	+ 2.4 %	– 0.07	– 13.2 %	+ 0.21

Datasets	Δ RMSE % (CNN→GRU)	Δ RMSE % (Both)	Spearman rank corr. 증가량 (CNN→GRU)	피크 반응 증폭률 % (CNN→GRU)	반응 선행/지연 (스텝) (CNN→GRU)
ETTh2	+ 0.9 %	+ 8.3 %	+ 0.05	+ 9.8 %	+ 1.52
ETTm2	– 0.6 %	+ 5.9 %	+ 0.16	+ 13.0 %	+ 1.05
ETTh1	– 1.0 %	+ 1.9 %	+ 0.13	+ 11.4 %	+ 1.32
ETTm1	– 1.2 %	+ 2.2 %	+ 0.24	+ 15.6 %	+ 1.25
Weather	– 0.6 %	+ 2.4 %	+ 0.12	+ 28.6 %	+ 0.82

- Spearman rank corr. (-1 ~ 1) : Conv 변화와 GRU 변화가 같은 방향으로 움직이는 정도가 얼마나 늘었는지를 의미
- 피크 반응 증폭률 % : GRU 반응이 none 보다 얼마나 더 커졌는지의 증가율
- 반응 선행/지연 : 양수면 반응이 더 빨라짐, (ETTh는 1 당 1시간(1스텝), ETTm은 15분(1스텝), Weather는 10분(1스텝))

Q1. 핵심 설계 철학

SIL System Intelligence Lab

왜 단독 사용 보다 상호 결합이 더 좋은가?

- 실험을 통해 확인해 본 결과, CNN ↔ GRU 의 상호 결합이 단방향 결합 이상의 성능을 보임

1) 제안 모델의 설계 의도인 잘 포착하는 정보의 유형이 다른 두 모델을 계층별로 특징을 결합하는 구조가 단독 구조 에서와 달리 **상호 보완적으로 정보를 활용**하여 풍부한 표현을 얻음

2) 단방향 에서와 달리 CNN ↔ GRU 의 **상호 결합이 각 경로의 특성을 보강하면서 불필요한 간섭 (noise) 을 줄였음**을 시사함

3) CNN ↔ GRU 상호 결합을 통해 한 모델의 출력이 다른 모델의 입력으로 거듭 활용되어 **특성 재사용이 증가**하고 **서로 보완하는 새로운 패턴을 학습**하여 **전체 모델의 표현력이 확장**된 것으로 사료됨

1) Bi-branch bidirectional coupled interaction fusion network for multi-retinal diseases diagnosis
2) Two-Stream Network for Sign Language Recognition and Translation
3) Dual Path Networks

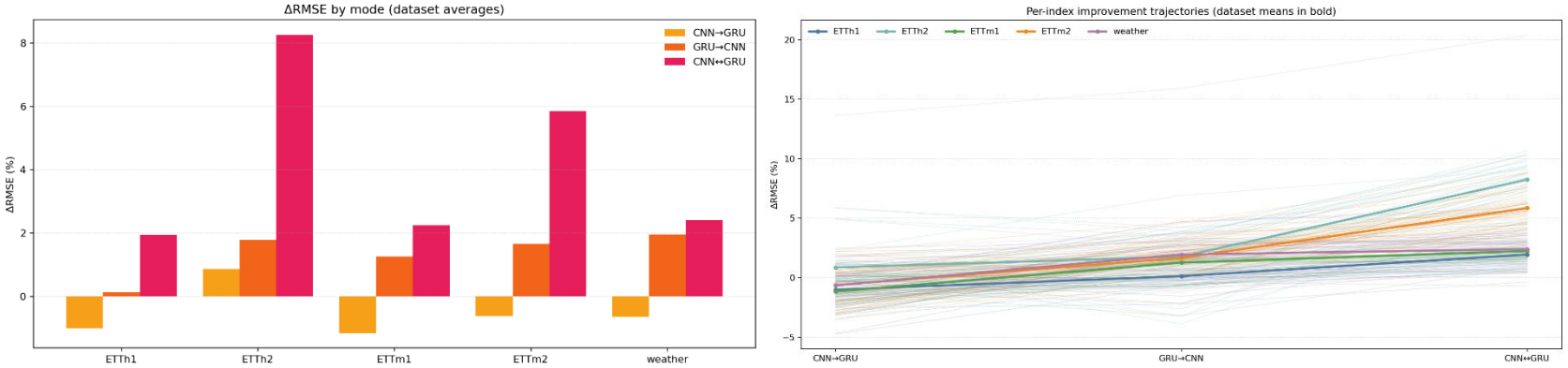
Q1. 핵심 설계 철학

왜 단독 사용 보다 상호 결합이 더 좋은가?

- 실험을 통해 확인해 본 결과, CNN ↔ GRU 의 상호 결합이 단방향 결합 이상의 성능을 보임

1) 제안 모델의 설계 의도인 잘 포착하는 정보의 유형이 다른 두 모델을 계층별로 특징을 결합하는 구조가 단독 구조 에서와 달리 **상호 보완적으로 정보를 활용**하여 풍부한 표현을 얻음

   - ➢ CNN ↔ GRU 상호 결합이 단방향 결합보다 평균적으로 더 좋은 성능을 보임
   - ➢ CNN ↔ GRU 의 상호 보완이 보편적으로 일어나며 보완 효과가 실제 오차 감소로 이어짐



각 데이터셋 · 모델에 대한 Δ RMSE 분포

- Δ RMSE : (none − mode) / none * 100 으로, 클 수록 좋음

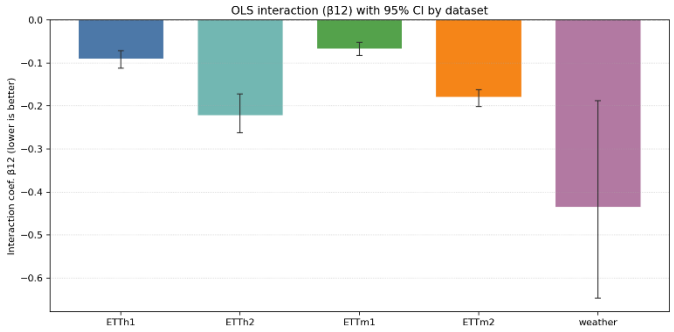
Q1. 핵심 설계 철학

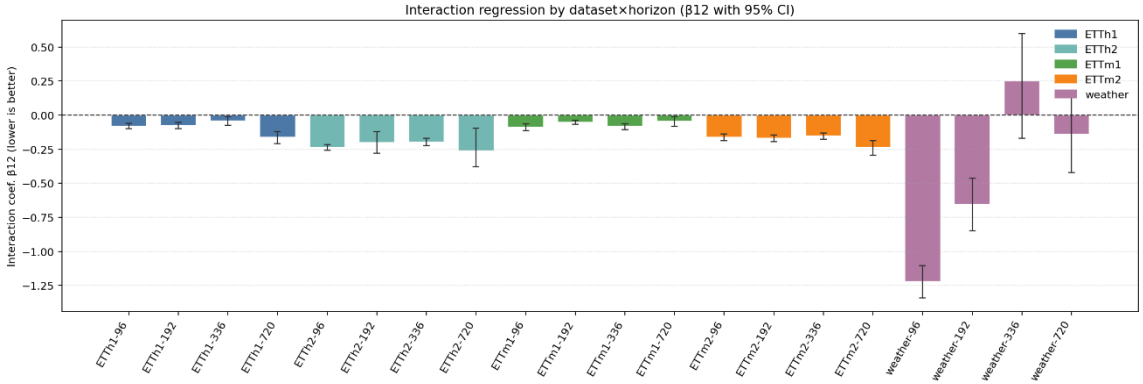
왜 단독 사용 보다 상호 결합이 더 좋은가?

- 실험을 통해 확인해 본 결과, CNN ↔ GRU 의 상호 결합이 단방향 결합 이상의 성능을 보임

2) 단방향 에서와 달리 CNN ↔ GRU 의 **상호 결합이 각 경로의 특성을 보강하면서 불필요한 간섭 (noise) 을 줄였음**을 시사함

➢ 회귀식을 작성하여 검증한 결과, 평균적으로 모든 데이터셋에서 상호 결합이 단방향 결합들의 합 이상의 초과 이득이 있었으며 Horizon 별 결과에서도 Weather (Horizon 336, 720) 을 제외한 모든 결과에서 초과 이득이 있음이 확인됨





각 데이터셋 · Horizon 에 대한 β_{12} 막대와 95% CI (신뢰구간) 막대
β_{12} 는 낮을 수록 상호 결합의 효과가 단방향 결합 둘을 더한 것보다 초과 이득이 있다는 의미이며,
95% CI 막대가 0 미만이면 그 초과 이득이 통계적으로 유의미하다는 것을 의미함

Q1. 핵심 설계 철학

왜 단독 사용 보다 상호 결합이 더 좋은가?

2) 단방향 에서와 달리 CNN ↔ GRU 의 **상호 결합이 각 경로의 특성을 보강하면서 불필요한 간섭 (noise) 을 줄였음**을 시사함

- 회귀식은 다음과 같음

$$\underbrace{\text{RMSE}}_{\text{관측된 오차}} = \beta_0 + \beta_1 \underbrace{[\text{CNN} \rightarrow \text{GRU}]}_{\text{켜짐(1)/꺼짐(0)}} + \beta_2 \underbrace{[\text{GRU} \rightarrow \text{CNN}]}_{\text{켜짐(1)/꺼짐(0)}} + \beta_{12} \underbrace{[\text{CNN} \leftrightarrow \text{GRU}]}_{\text{둘 다 켜졌을 때 1}}$$

- 이때 각 경우에서 β 는 다음처럼 해석됨
  - ✓ (결합 없음) : $\text{RMSE}_{none} = \beta_0$
  - ✓ CNN → GRU : $\text{RMSE}_{cg\_only} = \beta_0 + \beta_1$
  - ✓ CNN → GRU : $\text{RMSE}_{gc\_only} = \beta_0 + \beta_2$
  - ✓ CNN ↔ GRU : $\text{RMSE}_{both} = \beta_0 + \beta_1 + \beta_2 + \beta_{12}$
- 따라서 그림에서 설명하는 상호작용항 β_{12} 는 직관적으로 아래와 같으며, $\beta_{12} < 0$ 이면 단방향 두 개를 더한 수준 이상의 초과 이득이 있음을 의미함

$$\beta_{12} \approx \text{RMSE}_{both} - \text{RMSE}_{cg\_only} - \text{RMSE}_{gc\_only} + \text{RMSE}_{none}$$

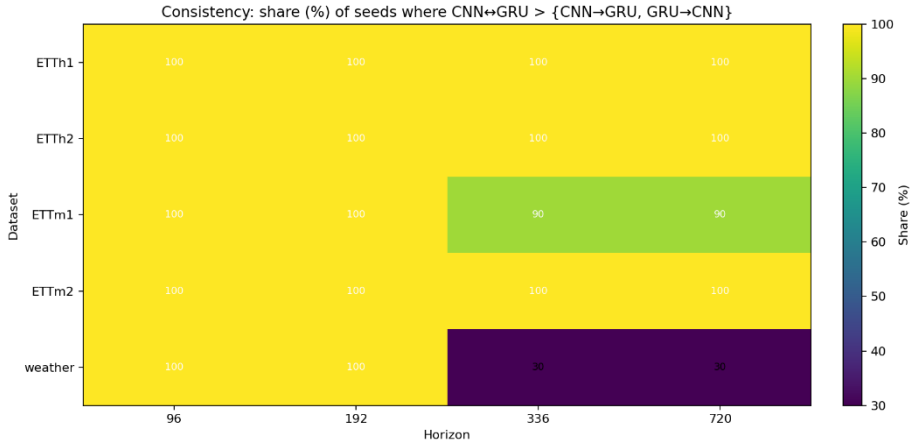
Q1. 핵심 설계 철학

왜 단독 사용 보다 상호 결합이 더 좋은가?

- 실험을 통해 확인해 본 결과, CNN ↔ GRU 의 상호 결합이 단방향 결합 이상의 성능을 보임

2) 단방향 에서와 달리 CNN ↔ GRU 의 **상호 결합이 각 경로의 특성을 보강하면서 불필요한 간섭 (noise) 을 줄였음**을 시사함

  - ➢ 또한 CNN ↔ GRU 상호 결합의 효과가 대부분의 경우에서 두 단방향 결합 이상의 효과가 있는 것으로 확인되어 간섭 억제 · 보강이 안정적인 현상임을 의미



CNN ↔ GRU 상호 결합이 두 단방향 결합 모두보다 더 좋은 Seed 의 비율 (%)
각 데이터셋 · Horizon 별로 Seed 10회 반복 실험함

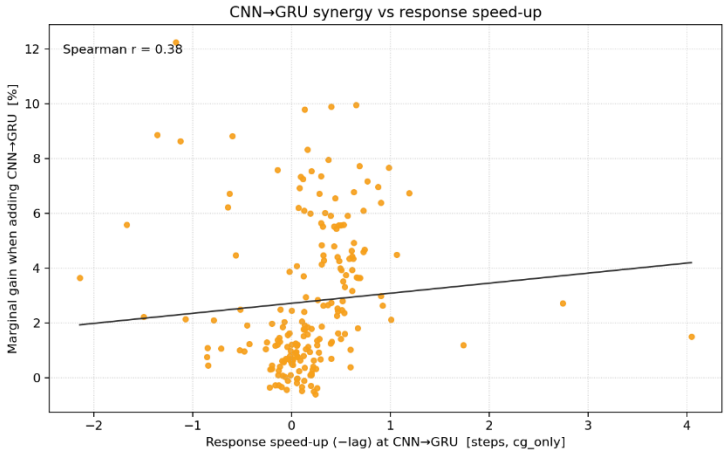
Q1. 핵심 설계 철학

왜 단독 사용 보다 상호 결합이 더 좋은가?

- 실험을 통해 확인해 본 결과, CNN ↔ GRU 의 상호 결합이 단방향 결합 이상의 성능을 보임

3) CNN ↔ GRU 상호 결합을 통해 한 모델의 출력이 다른 모델의 입력으로 거듭 활용되어 **특성 재사용이 증가**하고 **서로 보완하는 새로운 패턴을 학습**하여 **전체 모델의 표현력이 확장**된 것으로 사료됨

➢ CNN 의 로컬 단서 (급변) 가 GRU 의 업데이트 / 망각을 열어 이미 학습한 상태 / 특징을 다시 쓰고 문맥 · 모양의 새 조합을 쉽게 학습하도록 만듦 → 표현력 확장



CNN → GRU 가 급변에 더 빨리 반응할수록 (x) 상호 결합에서 얻는 이득이 더 커짐 (y)

- x(좌) : 반응 선행/지연 (스텝), 값이 클수록 더 빠르게 반응
- y(좌) : ΔRMSE(Both) - ΔRMSE(GRU → CNN)
- 급변 : CNN 변화량이 큰 순간 (CNN 블록이 크게 반응하는 피크, 급경사, 꺾임 등)

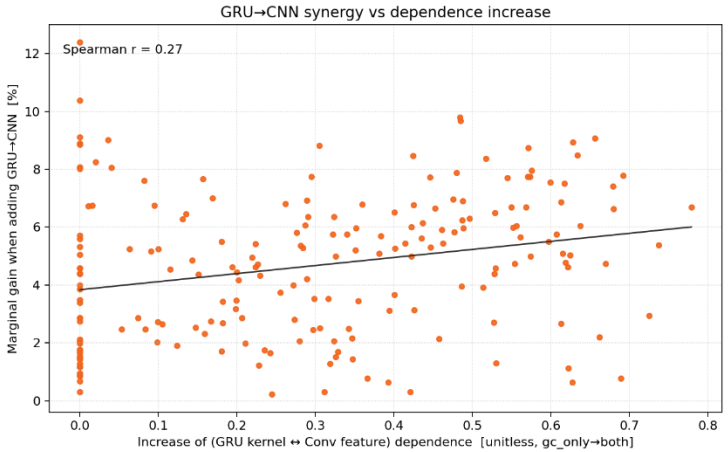
Q1. 핵심 설계 철학

왜 단독 사용 보다 상호 결합이 더 좋은가?

- 실험을 통해 확인해 본 결과, CNN ↔ GRU 의 상호 결합이 단방향 결합 이상의 성능을 보임

3) CNN ↔ GRU 상호 결합을 통해 한 모델의 출력이 다른 모델의 입력으로 거듭 활용되어 **특성 재사용이 증가**하고 **서로 보완하는 새로운 패턴을 학습**하여 **전체 모델의 표현력이 확장**된 것으로 사료됨

  - ➢ 급변을 빨리 포착할수록 상호 결합에서 더 이득을 보임 (특성 재사용과 새 패턴 학습이 쉬워짐 의미)
  - ➢ 문맥에 맞춰 필터 / 출력이 더 결합될수록 양방향에서의 추가 이득이 커짐 (문맥 재주입으로 필터 / 출력 재배치가 일어나 필요한 특징을 더 강하게 재사용함을 의미)



문맥에 맞춘 정렬 (x) 이 더 커질수록, 상호 결합에서 얻는 추가 이득이 더 커짐 (y)

- x(우) : distance correlation(Both) – distance correlation(GRU → CNN)
- y(우) : ΔRMSE(Both) - ΔRMSE(CNN → GRU)
- distance correlation (TOD (하루 주기), 범위 0~1) : 값이 클수록 필터의 모양/세기나 Conv 출력의 크기/형태가 더 뚜렷하게 변화하여, 문맥에 맞춘 필터/출력 조정이 강하게 일어났음을 의미함

Q2. 모델 설계 목적

SIL System Intelligence Lab

이 모델이 어떤 특징을 가진 데이터에 대해 예측을 잘 하도록 디자인되었음을 보여야 함

① **주기적 골격은 뚜렷하지만 위상 · 진폭이 조금씩 흔들리는 데이터**에 대해, 제안 모델은 CNN ↔ GRU 간 정보 참조와 매 배치 · 채널 · 시점별로 커널을 다르게 만들어 위상 · 진폭차를 보정하도록 제안함

② **변수들이 묶음으로 함께 움직이는 경향 (공변) 이 강한 데이터**에 대해, 제안 모델은 우선 변수들을 강하게 혼합하여 공통 변화를 포착한 후, 층마다 CNN ↔ GRU 서로의 상태를 기준으로 갱신하도록 하여 변수 간 상호작용 정보가 각 층에서 재주입 · 재정렬되도록 만듦

③ **저주파 골격 (추세 · 계절성) 이 강하고 고주파 (잔차) 는 상대적으로 낮은 데이터**에 대해, 제안 모델은 GRU 경로가 장기 의존과 상태 요약을 담당하고, CNN 경로가 피크 · 경사 · 꺾임 같은 형태 단서를 잡으며 Cross-Modal Attention 이 저주파 골격 정보 위의 요동 중 맥락적으로 의미 있는 부분만 강화하도록 만듦

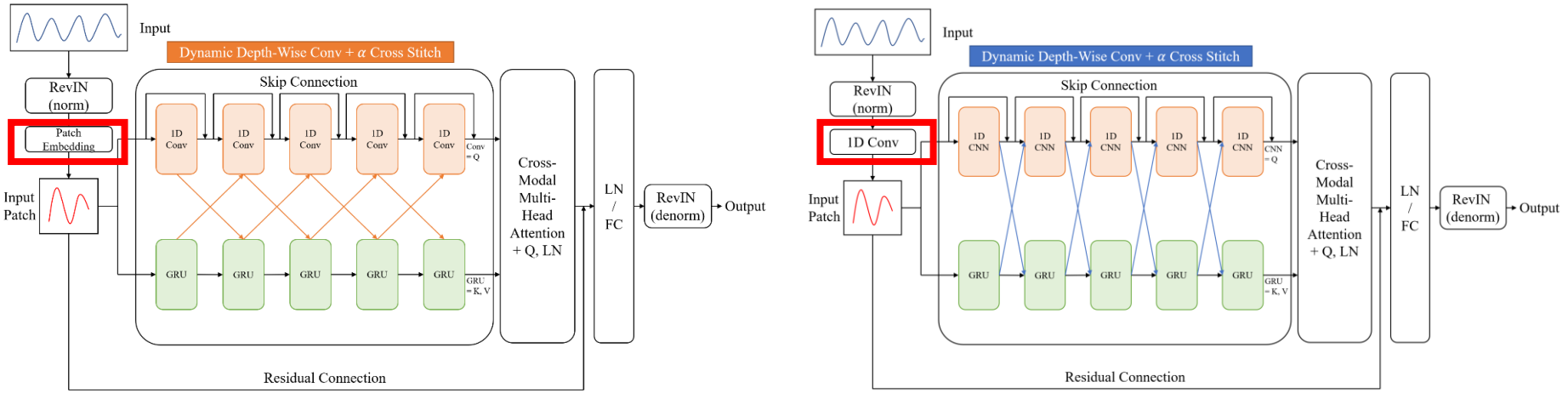
④ **검증 / 테스트 구간의 평균 · 분산 분포 변화가 큰 데이터**에 대해, RevIN (Reversible Instance Normalization) 을 도입하여 평균 / 분산 불일치를 줄임

➢ ETTm2, ETTh2 는 이 네 가지 데이터적 조건에 부합하는 데이터셋으로 이에 다른 데이터셋보다 유의한 성능을 얻을 수 있었음

Q3. 패치 임베딩과 상관관계

“패치 임베딩”이 변수 간 상관관계를 보았다고 할 수 있는가?

- 제안 모델의 “패치 임베딩” 층은 1D Conv 하나로 되어 있는 층이기 때문에, 이 층은 지도 학습을 수행하는 선형 혼합층이므로 이전에 주장했던 “패치 임베딩 층이 변수 간 상관관계를 학습했다” 고 주장하는 것은 옳지 못함
- 따라서 제안 모델의 해당 층 구성을 고려했을 때 “**1D Conv**” 층으로 명명하고, 이 층이 하는 역할은 “**입력 배치의 모든 변수를 받아 그 변수들의 공통 변화 신호를 만든다**” 고 설명하는 것이 더 정확함

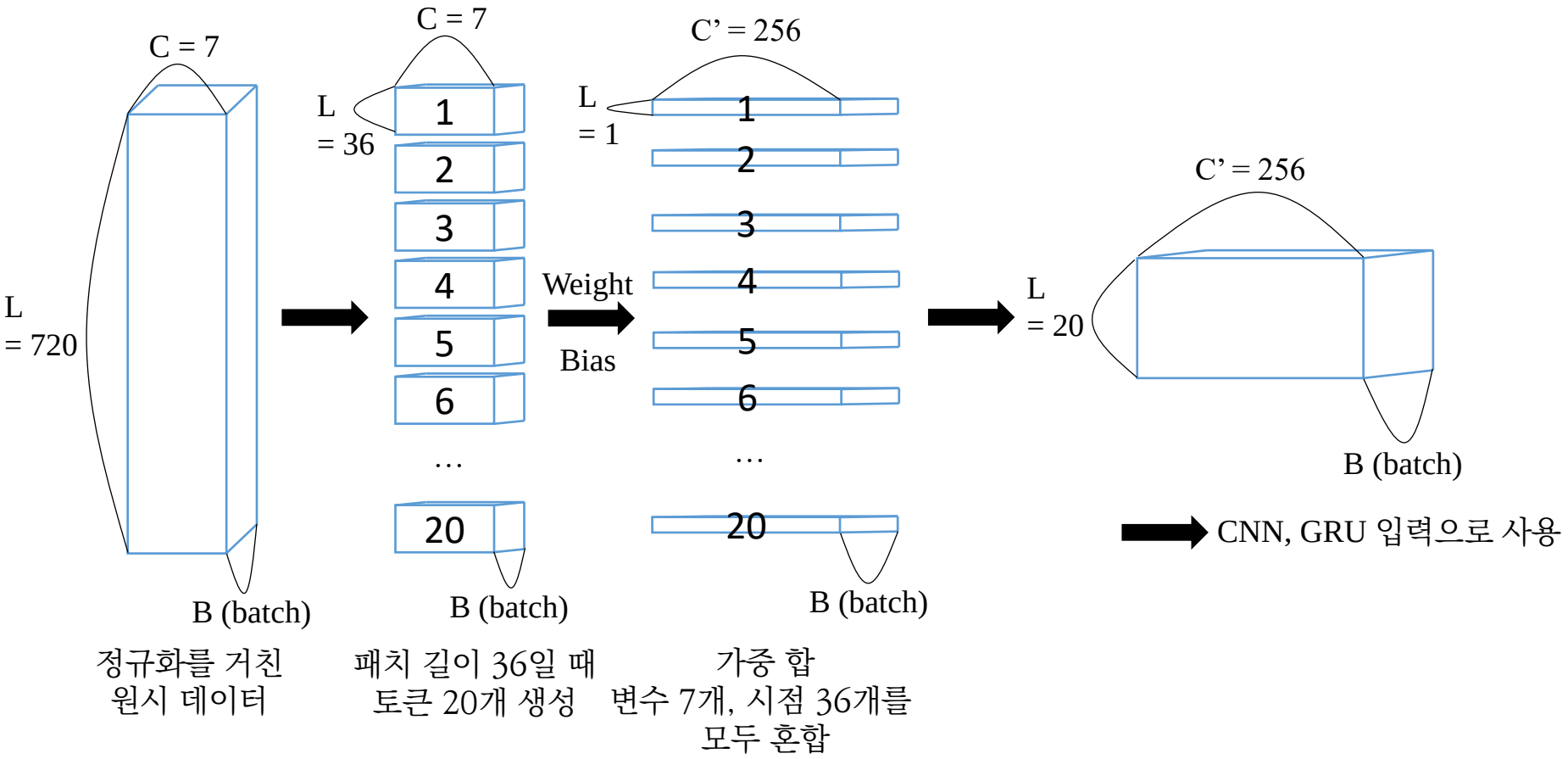


이전 발표에서의 모델 요약 (좌) 갱신된 모델 요약 (우)

Q3. 패치 임베딩과 상관관계

“패치 임베딩”이 변수 간 상관관계를 보았다고 할 수 있는가?

- 해당 층에서 이루어지는 연산을 그림으로 표현하면 다음과 같음 (기존 1D Conv 연산과 동일)



(패치 길이 36), (L), (W, b 연산 후 C’)는 모두 하이퍼파라미터

Q3. 패치 임베딩과 상관관계

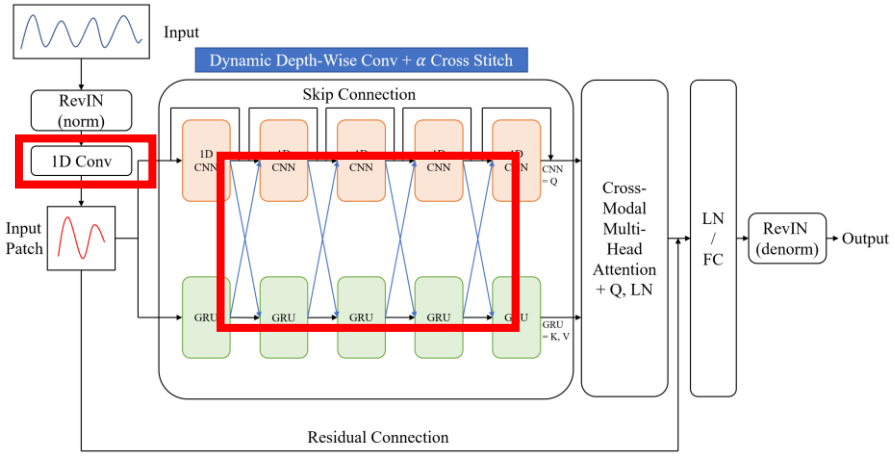
“1D Conv”층이 왜 필요한가?

- **시계열 데이터에서 국소 패턴과 긴 문맥을 동시에 보는 것**은 다음의 문제가 있음
  - (1) **긴 입력 길이** (L=720)로 인해 초반 학습이 불안정해지고 특히 GRU 는 정보를 효율적으로 유지하기 어렵다는 문제가 있음
  - (2) **입력 변수가 여러 개**이므로 변수를 따로 처리하게 되면 제안 모델의 CNN 은 다른 변수의 상황과 무관하게 학습하고 GRU 는 국소 단서 없이 장기 기억할 성분을 결정하게 됨
- 제안 모델은 1D Conv 층을 통해서 (1) **긴 입력 길이를 짧게 압축**하고 (2) **모든 변수와 패치당 시간 정보를 가중 합**하면서 각 토큰에 여러 변수가 같은 때에 어떻게 움직였는지 정보가 담김
- 이를 통해 모델은 원시 시계열을 입력 받는 것이 아닌 적절히 섞여 정돈된 토큰을 입력 받아 상대적으로 더 좋은 성능을 보일 수 있음

Q4. Depth-Wise 사용 이유

왜 표준 Conv1D 가 아니라 Depth-Wise 를 이용하는가

- 제안 모델의 "**동적 Depth-Wise Conv**" 는 앞 "설계 철학"에서 언급했던 "**반대 경로가 보고 있는 '지금의 상태' 를 이용**해 변수 별로 필터를 매번 새로 만들어 **각 정보를 가변게 보정**하고, 그 **소량을 서로 가중 합**하여 **국소 무늬 (CNN) 를 문맥 (GRU) 에 맞춰 다듬고, 문맥 (GRU) 을 국소 무늬 (CNN) 에 맞추어 보강**하는 것을 **층마다 반복**하도록" 하는 장치를 말함
- 따라서, 제안 모델은 '**변수를 혼합**' 하는 '**1D Conv**' 층 + '**서로를 보정**' 하는 '**동적 Depth-Wise Conv**' 층으로 역할이 나누어져 있음



Q4. Depth-Wise 사용 이유

왜 표준 Conv1D 가 아니라 Depth-Wise 를 이용하는가

- 제안 모델의 '변수를 혼합' 하는 '1D Conv' 층 + '서로를 보정' 하는 '동적 Depth-Wise Conv' 층을 제거하고 **표준 Conv1D 로 교체하면 다음과 같은 우려가 있음**
- (1) 표준 Conv1D 의 기능 (특징 추출) 은 이미 CNN 블록, 1D Conv 층에서 수행되어 교차 보정을 수행해야 하는 해당 위치에서 반복할 필요가 없고 **변수 축 정보가 변형되어 상호 보정 자체의 결합 안정성이 떨어짐**
- (2) 표준 Conv1D 는 한 번 학습된 커널을 일관되게 적용하여 시계열 데이터에서 위상 · 진폭의 흔들림을 **동적 Depth-Wise Conv 에서처럼 즉시 · 가볍게 조정해줄 수 없음**
- (3) **계산 효율 측면**에서 표준 Conv1D 의 사용이 제안 모델보다 불리함

	1D Conv 층 + 동적 Depth-Wise Conv (제안 모델)	1D Conv 층 + 표준 Conv1D	(변수 7 → 256 변환 블록) + 표준 Conv1D
앞 단 연산량	1,290,240	1,290,240	3,870,720
뒷 단 연산량	2,967,552	55,050,240	1,981,808,640
연산량 총합	4,257,792	56,340,480	1,985,679,360

제안 모델에서의 해당 구조 연산량과
구조를 표준 Conv1D 로 교체하였을 때 연산량 비교

감사합니다

연세대학교 지능형데이터·최적화학과
System Intelligence Lab
박힘찬

Q4. Depth-Wise 사용 이유

왜 표준 Conv1D 가 아니라 Depth-Wise 를 이용하는가

- 연산량 계산에 대해 상세 수식을 제시하면 다음과 같음
- 입력 시계열 길이 L=720, 특성 차원 d=256, 합성곱 커널 k=3, 입력 변수 수 C=7 (ETT 데이터셋 기준), 블록 깊이=7 (블록당 교차 경로 2개이므로 블록당 합성곱 2회) 일 때,
  - 앞 단 "1D Conv 층"
    - 출력 토큰 수 T=20, 출력 채널 d=256, 입력 채널 C=7, 커널 P(k)=36
    - MACs (연산량) $= \mathrm{T} \times d \times C \times P = 20 \times 256 \times 7 \times 36 = 1{,}290{,}240$ MACs
  - 블록 내 "동적 Depth-Wise Conv"
    - 커널 생성: $d \times d \times k = 256 \times 256 \times 3 = 196{,}608$ MACs
    - Depth-Wise Conv 1회: $T \times d \times k = 20 \times 256 \times 3 = 15{,}360$ MACs
    - 따라서 "동적 Depth-Wise Conv" 1회 반복 시 211,968 MACs
    - HP2 기준 블록 7개, 양방향이므로 $7 \times 12 \times 211{,}968 = 2{,}967{,}552$ MACs
  - 따라서 구조 총합 $1{,}290{,}240 + 2{,}967{,}552 = 4{,}257{,}792$ MACs

Q4. Depth-Wise 사용 이유

왜 표준 Conv1D 가 아니라 Depth-Wise 를 이용하는가

- 연산량 계산에 대해 상세 수식을 제시하면 다음과 같음
- 입력 시계열 길이 L=720, 특성 차원 d=256, 합성곱 커널 k=3, 입력 변수 수 C=7 (ETT 데이터셋 기준), 블록 깊이=7 (블록당 교차 경로 2개이므로 블록당 합성곱 2회) 일 때,
  - 앞 단 "1D Conv 층"
    - 출력 토큰 수 T=20, 출력 채널 d=256, 입력 채널 C=7, 커널 P(k)=36
    - MACs (연산량) = $\mathrm{T} \times d \times C \times P = 20 \times 256 \times 7 \times 36 = 1{,}290{,}240$ MACs
  - 표준 Conv1D 적용
    - 앞 단 1d Conv 를 적용했으므로 입력/출력 d=256, 길이 T=20, 커널 k=3
    - MACs (연산량) = $\mathrm{T} \times d \times d \times k = 20 \times 256 \times 256 \times 3 = 3{,}932{,}160$ MACs
    - 블록 7개, 양방향이므로 $7 \times 2 \times 3{,}932{,}160 = 55{,}050{,}240$ MACs
  - 따라서 구조 총합 $1{,}290{,}240 + \ 55{,}050{,}240 = 56{,}340{,}480$ MACs

Q4. Depth-Wise 사용 이유

왜 표준 Conv1D 가 아니라 Depth-Wise 를 이용하는가

- 연산량 계산에 대해 상세 수식을 제시하면 다음과 같음
- 입력 시계열 길이 L=720, 특성 차원 d=256, 합성곱 커널 k=3, 입력 변수 수 C=7 (ETT 데이터셋 기준), 블록 깊이=7 (블록당 교차 경로 2개이므로 블록당 합성곱 2회) 일 때,
  - 앞 단 "1D Conv 층" 제거, 입력 채널을 표준 Conv1D 로 교체
    - 길이 L=720, 출력 채널 d=256, 입력 채널 C=7, 커널 k=3
    - MACs (연산량) $= L \times d \times C \times k = 720 \times 256 \times 7 \times 3 = 3{,}870{,}720$ MACs
  - 표준 Conv1D 적용
    - MACs (연산량) $= L \times d \times d \times k = 720 \times 256 \times 256 \times 3 = 141{,}557{,}760$ MACs
    - 블록 7개, 양방향이므로 $7 \times 2 \times 141{,}557{,}760 = 1{,}981{,}808{,}640$ MACs
  - 따라서 구조 총합 $3{,}870{,}720 + 1{,}981{,}808{,}640 = 1{,}985{,}679{,}360$ MACs