일반논문 (Regular Paper)
방송공학회논문지 제30권 제3호, 2025년 5월 (JBE Vol.30, No.3, May 2025)
https://doi.org/10.5909/JBE.2025.30.3.402
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

시계열 전력 데이터 예측을 위한 디퓨전 모델 기반 마스킹 데이터 증강 방법

이 재 호[a)], 임 종 경[a)], 배 성 호[a)‡]

A Diffusion Model-based Masking Data Augmentation Method for Improving Time Series Power Forecasting Performance

Jaeho Lee[a)], Jongkyung Im[a)], and Sung-Ho Bae[a)‡]

요 약

본 연구는 태양광 발전량 예측을 위해 소규모 시계열 데이터에서 사용할 수 있는 새로운 데이터 증강 기법을 제안한다. 최근 디퓨전(diffusion) 모델이 괄목할만한 데이터 생성 능력을 보임에 따라, 본 연구에서는 증강 데이터의 품질을 높이기 위해 디퓨전(diffusion) 모델을 활용한 데이터 증대 기법을 제안한다. 이때, 디퓨전(diffusion) 모델을 이용해 시계열 데이터의 모든 특징을 한번에 생성할 경우, 특징 간 상관 관계를 제대로 고려하지 못하는 문제를 발견했다. 이를 해결하기 위해, 원본 샘플의 특징 중 하나를 마스킹 한 다음, 마스킹 한 부분을 디퓨전(diffusion) 모델로 생성하여 원본과 유사한 특성을 유지하면서 증강 효과를 가지는 마스킹 데이터 증강 방법을 개발하였다. 또한 생성 데이터와 원본 데이터 수에 따른 과도한 외부 데이터 의존성을 방지하기 위해 생성 데이터 중 원본 데이터와 유사도가 높은 상위 20개의 데이터만 추출하여 학습에 사용하였다. 실험 결과 원본 데이터를 학습한 모델과 비교했을 때 R^2-$score$는 약 217% 개선이 있었고, RMSE의 경우 약 35% 정도의 성능의 개선이 있었고, SMAPE 지표에서도 약 12.9%의 성능 개선이 관찰되었다.

Abstract

This study proposes a novel data augmentation technique for small-scale time series data to improve solar power generation forecasting. With recent advancements in diffusion models demonstrating remarkable data generation capabilities, we introduce a data augmentation method utilizing diffusion models to enhance the quality of augmented data. However, we observed that generating all features of a time series data with diffusion models fails to properly account for inter-feature dependencies. To address this, we developed a masked data augmentation approach, where one feature of the original sample is masked and regenerated using the diffusion model. This method preserves the characteristics of the original sample while achieving an augmentation effect. Additionally, to mitigate excessive reliance on a large number of externally generated data, we filtered the generated data by selecting the top 20 samples most similar to the original data and used these for training. As a result of the experiment, there was about 217% improvement in R^2-$score$, and about 35% improvement in RMSE, and about 12.9% improvement in SMAPE compared to the model trained with the original data.

Keyword : Tabular and time series data, Diffusion model, Masking, Data augmentation

Ⅰ. 서 론

태양광 발전량 예측은 일사량, 전운량, 기온과 같은 기상 요소를 통해 태양광 발전 시스템이 생산할 전력량을 사전에 추정하는 것을 의미한다. 태양광 발전량 예측을 하는 이유는 태양광 발전은 출력 변동성이 높아 전력 계통의 안정성에 영향을 줄 수 있다. 따라서 정확한 발전량 예측을 통해 전력 수급 계획을 세우면 계통의 안정성을 유지하는 데 도움이 된다[1]. 그리고 발전량 예측을 통해 에너지 저장 시스템의 최적 운영과 부하 관리가 가능해져 에너지 효율성을 높일 수 있다.

기존의 태양광 발전량 예측을 하는 방법은 간단하게 물리 기반 모델과 통계적 모델이 있다. 물리 기반 모델은 태양의 위치, 지리적 정보 등을 활용하여 일사량과 발전량을 계산하는 방식이다[2]. 이는 시스템의 물리적 특성을 고려하지만, 환경 요인을 정확히 반영하기 어렵다. 통계적 모델은 과거의 발전량 데이터와 기상 데이터를 기반을 통해 미래의 발전량을 예측한다[3]. 이는 데이터의 패턴을 활용하지만, 비선형적이고 복잡한 관계를 충분히 모델링하지 못한다.

최근에는 기계학습 모델이 발전되어 이를 사용해서 태양광 발전량 예측을 하는데, 기계학습 모델을 사용하기 위해선 충분한 고품질 데이터의 확보가 중요하다[4]. 기상 데이터의 경우 기본적으로 자연 현상을 기반으로 수집되므로 관측 가능한 시간과 공간이 제한되어 있어 충분한 데이터를 모으는 데 한계가 있다. 게다가 시계열 데이터의 특성으로 인해 단순한 데이터 증강 기법을 적용하기도 어렵다.

시계열 데이터는 시간적인 연속성과 패턴을 가진다. 따라서 만약 기존의 이미지나 텍스트 데이터에서 흔히 사용하는 데이터 증강 기법을 그대로 적용한다면 데이터의 시간적 의존성과 연속성을 훼손하게 된다. 즉, 일반적으로 사용하는 기법을 시계열 데이터에 적용하기에는 한계가 있다[5].

따라서, 본 연구의 주요 목표는 시계열 데이터의 특성을 유지하면서도 데이터를 증강할 수 있는 새로운 접근 방식을 제안하여 예측 성능을 높이는 것이다. 이를 위해, 본 연구에서는 디퓨전(diffusion) 모델 기반의 마스킹 데이터 증강 기법을 최초로 제안한다. 이 방법은 기존 데이터의 각 특성 간 관계를 유지하면서도 새로운 데이터를 생성할 수 있도록 설계되었다. 구체적으로, 원본 데이터의 일부 특성을 마스킹하고 이를 디퓨전(diffusion) 모델로 재생성하여, 원본 데이터의 분포를 최대한 보존하면서도 증강 효과를 달성한다. 추가로 생성된 데이터 중 원본 데이터와의 유사도가 높은 데이터를 선별하여 학습에 활용함으로써, 외부 생성 데이터에 대한 과도한 의존성을 줄이고 모델의 안정성을 높였다.

이러한 기법을 통해 본 연구에서는 소규모 시계열 데이터 환경에서도 높은 예측 성능을 확보할 수 있었다. 실험 결과, 디퓨전(diffusion) 모델을 활용한 데이터 증강은 약 221%의 성능 향상을 이루었으며, 예측 지표인 MAE와 RMSE에서도 각각 약 35.5%와 36%의 개선을 달성하였다. 이를 통해 제안된 접근법이 데이터 증강 및 예측 성능 향상에서 기존 방법 대비 우수한 성능을 보임을 확인하였다.

Ⅱ. 본 론

1. 관련 연구

1.1 딥러닝 기반 태양광 발전량 예측 모델

딥러닝은 전력 수급 관리와 태양광 발전량의 예측 정확도 향상을 위해 널리 활용되고 있다[4]. 딥러닝 기반 모델은 주로 재생에너지 발전량 예측 시장에서 요구되는 안정적이고 신뢰성 있는 예측 결과를 도출하기 위해 개발되었다.

특히, Convolutional Neural Network (CNN)와 Gated Recurrent Unit (GRU)을 결합한 CNN-GRU[6] 모델은 시계열 데이터의 공간적 및 시간적 특징을 효과적으로 학습할

a) 경희대학교(Kyung Hee University)
‡ Corresponding Author : 배성호(Sung-Ho Bae)
E-mail: shbae@khu.ac.kr
Tel: +82-31-201-2593
ORCID: https://orcid.org/0000-0002-3389-1159
※이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.RS-2025-00155911, 인공지능융합혁신인재양성(경희대학교))와, 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터사업의 연구결과로 수행되었음 (IITP-2025-RS-2023-00259004).
· Manuscript February 4, 2025; Revised April 14, 2025; Accepted April 14, 2025.

수 있다. CNN은 합성곱 연산을 통해 입력 데이터의 지역적 패턴을 추출하는 데 뛰어난 성능을 보이고, GRU는 순환 신경망의 한 종류로 장기 의존성을 학습하는 데 적합하다. 따라서 CNN-GRU 모델은 기상 데이터의 공간적 특징과 시간적 변화를 동시에 고려하여 태양광 발전량을 예측하는 데 유리하다. 따라서 우리는 제안하는 데이터 증강 기법을 통해 생성된 데이터를 CNN-GRU 모델을 사용하여 평가한다.

1.2 LSGAN

Generative Adversarial Networks (GAN)은 생성자(Generator)와 판별자(Discriminator)라는 두 개의 신경망을 서로 경쟁시키며 학습하는 방식으로 데이터를 생성하는 모델이다[7]. 이 두 신경망은 서로를 속이기 위해 경쟁하는 과정에서 생성자는 점점 더 진짜와 유사한 데이터를 생성하게 된다. GAN은 특히 이미지 생성 및 보완에 큰 성과를 보여 왔지만, 기울기 소실(Gradient Vanishing)과 모드 붕괴(Mode Collapse)와 같은 여러 한계가 존재한다. GAN의 이러한 한계를 극복하기 위해 제안된 것이 Least Squares Generative Adversarial Networks (LSGAN)이다[8].

LSGAN은 전통적인 GAN에서 사용되는 교차 엔트로피 손실 함수 대신 최소 제곱 손실 함수를 사용하여, 기울기 소실 문제를 완화하고 모델 훈련을 보다 안정적으로 만들어 데이터 증강이 필요한 환경에서 중요한 대안으로 떠오르고 있다.

1.3 디퓨전(diffusion) 모델

디퓨전(diffusion) 모델은 데이터에 노이즈를 제거해가면서 고품질 데이터 샘플을 생성하는 모델이다[9]. 디퓨전(diffusion) 모델은 데이터 분포를 모드 붕괴 없이 다양한 샘플을 생성할 수 있으며, 특히 Fréchet Inception Distance (FID)와 인셉션 점수 측면에서 경쟁력 있는 성능을 보인다. 확산 모델은 주로 디노이즈 과정을 학습하는 데 집중한다. 각 단계에서 노이즈가 추가된 샘플을 입력받아 해당 노이즈 수준을 제거하도록 학습함으로써, 노이즈 점수 매칭 기법을 활용한 데이터 분포 기울기를 간접적으로 추정한다. 이러한 점진적인 샘플링 접근 방식은 상태 간 보간이나 손상된 입력 복구에도 유리하며, 생성 과정 전반의 안정성과 높은 샘플 품질을 보장한다.

이 모델은 특히 GAN이 직면하는 불안정한 훈련 문제를 완화한다. 변분 훈련 목표를 통해 수렴을 안정적으로 유도하므로 다양한 유형의 데이터에서도 견고하게 작동한다. 이러한 점 때문에 확산 모델은 고품질의 다양한 샘플을 요구하는 이미지 생성, 데이터 증강, 프라이버시 보존 합성 데이터 등에서 핵심적인 역할을 수행하며, 전통적인 GAN의 한계를 극복하는 강력한 대안으로 주목받고 있다.

최근에는 디퓨전 모델을 시계열 데이터 증강에 활용하는 연구도 활발히 이루어지고 있다. 예를 들어, 센서 기반 시계열 데이터의 주파수 보간을 통해 행동 예측 정확도를 높인 연구[10]나, 시계열 데이터의 예측 및 결측치 대체를 위해 디퓨전 모델을 적용한 연구[11] 등이 있다. 이러한 연구들은 디퓨전 모델이 시계열 데이터 증강 및 분석에 효과적임을 보여주며, 본 연구와 같은 소규모 시계열 데이터 환경에서도 유용하게 적용될 수 있음을 시사한다.

디퓨전 모델에는 다양한 종류가 존재하며, 대표적으로 Denoising Diffusion Probabilistic Models (DDPM)[12], Denoising Diffusion Implicit Models (DDIM)[13], Latent Diffusion Model[14], 그리고 최근 주목받고 있는 Stable Diffusion[14] 등이 있다. 본 연구에서는 이들 중 가장 기본이 되는 DDPM을 사용하였다. DDPM을 선택한 이유는 현재 데이터셋의 크기가 매우 적기 때문이다. 상대적으로 더 복잡한 구조를 가진 최신 디퓨전(diffusion) 모델들은 큰 데이터셋에서 강력한 성능을 발휘하지만, 데이터의 크기가 작을 경우 모델이 과적합 될 가능성이 높아진다. 반면, DDPM은 상대적으로 간단한 구조를 가지면서도 안정적으로 데이터를 생성할 수 있는 장점이 있다. 이러한 특성은 데이터의 규모와 모델의 복잡성 간의 균형을 맞추는 데 적합하여, 본 연구의 목적에 부합한다고 판단하였다. 따라서, 본 연구에서는 DDPM을 기반으로 데이터 증강을 수행하여 모델의 안정성과 학습의 효율성을 높이는 데 중점을 두었다.

2. 데이터 정의

본 연구에서 데이터 증강 기법을 적용할 데이터는 세종

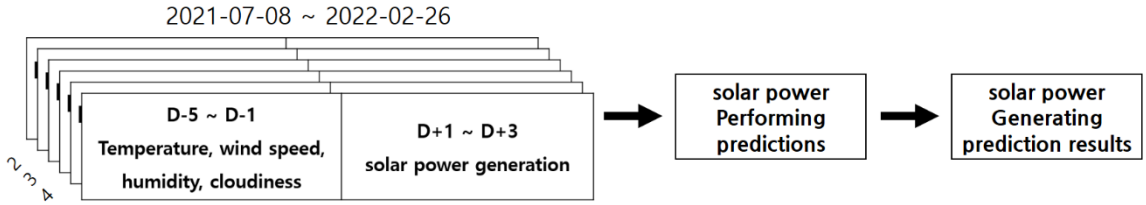


그림 1. 데이터 정의
Fig. 1. Data definition

특별자치시의 각 지역에 있는 태양광 발전기의 발전량과 해당 시점에서 세종시의 기상 관측된 데이터이다. 그림 1을 살펴보면 2021년 7월 8일부터 2022년 2월 26일까지의 데이터로, 총 234개의 묶음으로 구성되어 있다. 각 묶음은 과거 5일의 데이터와 미래 3일의 데이터를 포함하고, 각각의 날짜에는 오전 8시부터 오후 8시까지 매 시간별로 기온, 풍속, 습도, 전운량에 대한 데이터와 해당 조건에서 생성된 태양광 발전량 데이터가 포함되어 있다.

본 연구에서는 데이터의 시간적 순서를 유지하면서 학습셋과 테스트셋을 분리하였다. 학습셋은 2021년 7월 8일부터 2021년 12월 31일까지의 데이터를 사용하였고, 테스트셋은 2022년 1월 1일부터 2022년 2월 26일까지의 데이터를 사용하였다. 테스트셋에는 약 57개의 묶음이 포함되어 있고 이는 전체 데이터의 약 23%를 차지한다.

그러나 데이터의 수가 적기 때문에, 기존의 학습 방법으로는 일반화된 성능을 보장하기 어려운 상황이다. 특히 시계열 데이터 특성상 임의로 데이터를 증강하기 어려우며, 기존의 증강 기법을 그대로 적용할 수 없다는 점에서 추가적인 문제가 발생한다.

따라서 본 연구에서는 D시점 기준으로 D-5일부터 D-1일까지의 데이터와 D+1일부터 D+3일의 예측값을 하나의 묶음으로 보고, 이러한 묶음을 새로운 방식으로 증강하는 아이디어를 제안한다. 우리는 시계열 데이터임에도 불구하고 얻고자 하는 태양광 생성량을 예측할 때 사용할 데이터를 기준일로부터 과거 5일로 고정했다[15]. 이때 데이터 증강은 D-5일부터 D-1일까지의 기상 데이터와 D+1일부터 D+3일까지의 생성된 태양광 발전량 데이터 중 특정 요소를 마스킹하여 해당 부분을 정답 데이터로 설정하고, 마스킹되지 않은 나머지 데이터를 입력 데이터로 활용하여 마스킹된 부분을 복원하는 방식으로 진행된다. 따라서 기존의 시계열 데이터 학습 과정과 다르게 각 데이터 묶음을 섞어서 학습하는 방식이 가능하다는 가정하에, 새로운 데이터를 생성해 기존 데이터에 추가하는 증강 전략을 고안한다.

3. 제안 방법 : 디퓨전(diffusion) 모델 기반 데이터 증강

제안 방법으로, 최근 많은 주목을 받는 디퓨전(diffusion) 모델을 활용하는 전략을 제안한다. 디퓨전(diffusion) 모델은 데이터의 구조적 특성을 보존하면서도 새로운 데이터를 생성할 수 있는 장점을 가지고 있다. 이러한 방식은 GAN과 같은 기존의 생성 모델들보다 더 안정적인 학습을 보장하며, 모드 붕괴와 같은 문제를 피할 수 있다[16].

따라서 본 연구에서는 디퓨전(diffusion) 모델을 테이블형 데이터, 특히 시계열 데이터에 맞게 변형하여 적용하는 방법을 제안한다. 이를 통해 기존의 GAN 기반 데이터 증강 방식보다 더 안정적이고 현실적인 데이터를 생성할 수 있었다. 구체적으로는, 테이블 데이터의 구조적 특성을 고려한 디퓨전 프로세스를 설계하여, 학습 시 데이터의 패턴을 최대한 보존하면서 새로운 데이터를 증강하는 방향으로 연구를 진행하였다.

추가적으로 모든 데이터를 한번에 학습시키는 것이 아니라 특성 간 유사성을 기반으로 데이터를 그룹화하여 학습하는 마스킹 기반 접근법을 설계하였다. 그림 2는 디퓨전(diffusion) 모델 기반 마스킹 데이터 증강 기법의 전체적인 과정을 보여준다. 먼저, 원본 테이블형 데이터에 대해 마스킹 과정을 수행하여 일부 특성을 선택적으로 가린다. 이 마스킹된 데이터를 입력으로 하여 Forward Diffusion Process

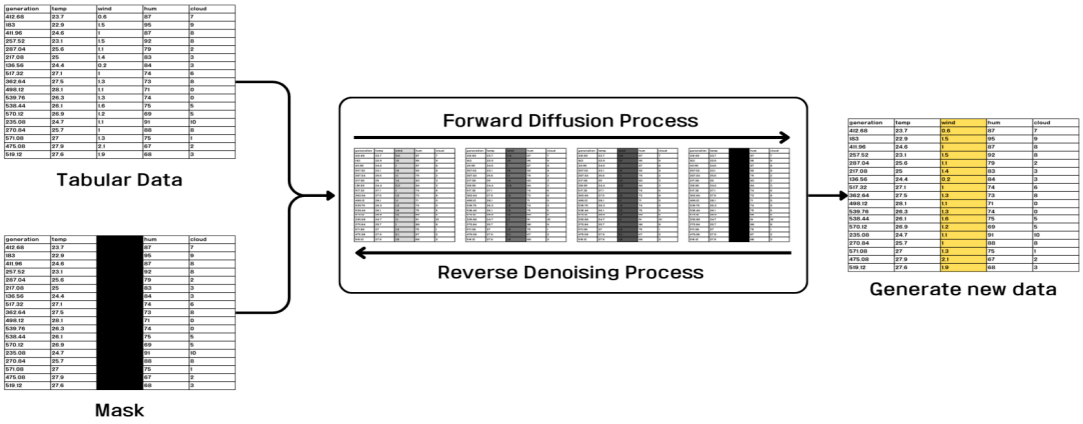


그림 2. 마스킹 기반 Diffusion model
Fig. 2. Masking-based Diffusion Model

알고리즘 1

Algorithm generates tabular data using diffusion model selective augmentation for enhanced machine learning performance.

Initialize the diffusion model $M_{\text{diffusion}}$ with parameters input_dim, hidden_dim, num_layers
Use a masking mechanism to Identify F_{g}, the complement of F_{c}
for each feature group F_{g}:
Split D into D_{c}(condition data) and D_{g}(label data).
while not converged:
$X_t \leftarrow add\_noize(X_{t-1}, t, \beta)$
$X_{t-1} \leftarrow M_{diffusion}(X_t)$
$loss \leftarrow MSE(X_{t-1}, X_{true})$
Generate N_g samples for each data point in D_c using $M_{diffusion}$
Combine all generated samples into G_{total}
Fit $KNN(D)$ model on the original dataset D
for each sample in G_{total}:
Compute similarity scores with D
Select top $k$samples G_k with the highest similarity scores.

와 Reverse Denoising Process 단계를 거치며 학습이 진행된다.

이 과정에서 마스킹되지 않은 특성의 정보를 기반으로 마스킹된 부분을 생성한다. 이를 통해 디퓨전(diffusion) 모델은 데이터 내의 특성 간 상관 관계를 학습하고, 새로운 데이터를 생성할 때 이러한 관계를 유지하도록 한다. 제안된 모델에 대한 학습 과정을 알고리즘 1에 기술한다.

4. 실험 및 결과

4.1 생성 데이터 분석

실제 데이터와 생성된 데이터 간의 유사도를 판별하기 위해 각 데이터를 t-SNE 기법[17]을 활용하여 생성된 데이터와 기존 데이터를 2차원 공간에 시각화한다. 그림 3을 확인하면 그래프에서 파란색 점은 실제 데이터를 나타내고, 주

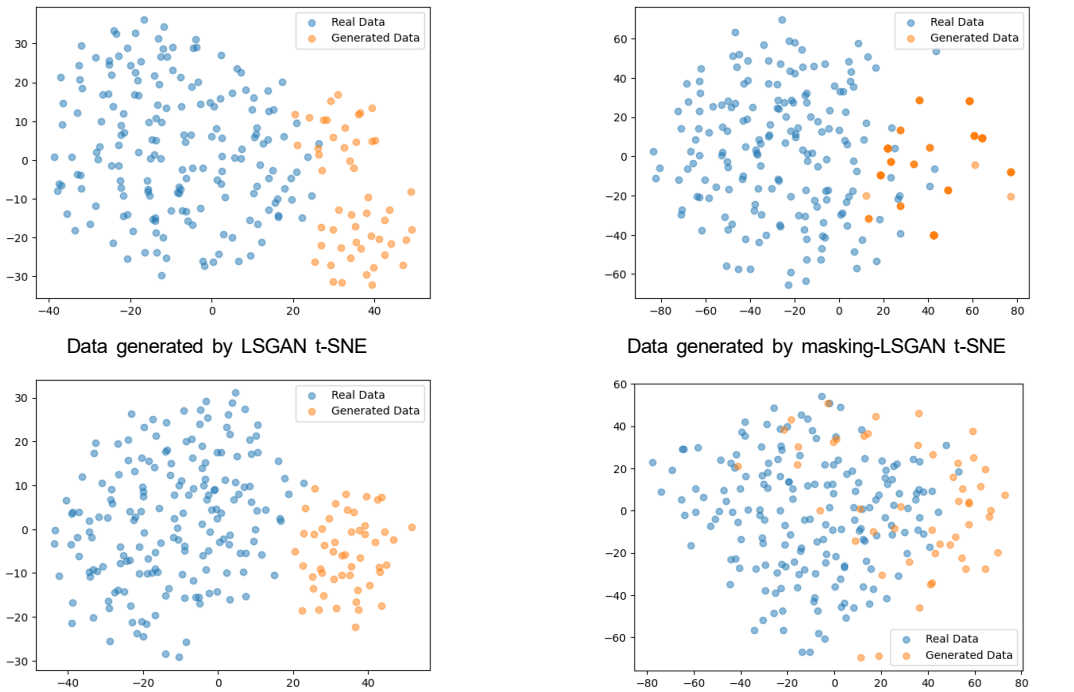


그림 3. 서로 다른 생성 방법에 따라 생성된 데이터와 실제 데이터의 t-SNE 시각화 결과
Fig. 3. t-SNE visualization of real and generated data using different generative methods

황색 점은 생성된 데이터를 나타낸다. 기존의 LSGAN 및 디퓨전(diffusion) 모델로 생성된 데이터의 경우 실제 데이터 분포와 일부 겹치는 모습을 보이지만, 생성된 데이터가 원본 데이터의 분포를 완전히 재현하지는 못한다. 이에 따라 마스킹 기법을 각각의 모델에 적용한 결과 생성된 데이터의 분포 재현력이 전반적으로 향상되었다. 특히 마스킹 기법을 적용한 디퓨전(diffusion) 모델은 생성된 데이터가 실제 데이터의 클러스터 내부에 밀접하게 분포하여 보다 정밀하게 원본 특성을 반영하는 모습을 보인다. 이는 기존의 LSGAN이나 마스킹이 적용되지 않은 디퓨전(diffusion) 모델과 비교했을 때 실제 데이터의 구조적 특성을 더욱 정교하게 학습했음을 의미한다. 즉, LSGAN 기반 모델도 실제 데이터 공간에 일부 근접한 데이터를 생성하지만, 디퓨전(diffusion) 모델에 비해서는 원본 특성 반영 수준이 상대적으로 낮다. 이는 디퓨전(diffusion) 모델이 특성 간 관계를 보다 정밀하게 파악하고, 데이터 생성 시 우수한 성능을 보인다. 결론적으로, 마스킹이 적용된 디퓨전(diffusion) 모델이 가장 뛰어난 성능을 보이고 본 연구의 제안 방식이 효과적임을 확인할 수 있다.

4.2 학습 성능 평가

생성된 데이터를 통해 학습을 진행하면 성능 향상이 있는지를 확인하기 위해 CNN-GRU 모델을 사용하여 기존 데이터와 동일한 방식으로 학습을 진행하였다. 구체적으로는 3개의 1차원 합성곱 계층과 2개의 GRU 계층 및 완전연결 계층으로 구성되었으며, 각 계층의 필터 수, 스트라이드, GRU 유닛 수 등의 하이퍼파라미터는 Keras Tuner의 Hyperband 탐색 알고리즘을 통해 최적화하였다. 모든 계층에는 ELU 활성화 함수를 사용하였으며, Adam 옵티마이저와 평균 절대 오차를 손실 함수로 사용하여 최대 100회 반복 학습하였고, 조기 종료를 적용하여 과적합을 방지하였다. 전체 학습 데이터의 10%는 검증용으로 사용하였으며, 배치 크기는 64로 설정하였다.

본 연구에서는 모델 예측 성능을 종합적으로 평가하기

위해 Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), R^2-$score$, Symmetric Mean Absolute Percentage Error (SMAPE)를 사용한다. MAE는 예측값과 실제값 간의 절대적 편차의 평균을 나타내며, 예측 오차의 평균 크기를 직관적으로 파악할 수 있어 모델이 전반적으로 어느 정도 정확하게 예측하는지를 확인하는 수단으로 사용한다. RMSE (Root Mean Squared Error)는 오차를 제곱한 뒤 평균을 내고 다시 제곱근을 취하는 방식으로, 상대적으로 큰 오차에 더 큰 가중치를 부여한다. 이를 통해 MAE로는 드러나지 않는 예측 극단치에 대한 민감도를 확보하여 모델의 안정성과 신뢰성 여부를 평가하는 데 사용을 한다. 마지막으로 R^2-$score$는 모델이 실제 데이터의 변동성을 얼마나 설명하고 있는지를 비율로 나타내는 척도로, 선형 회귀 모델에서 흔히 사용되며 예측치가 실제값의 분산을 재현하는 정도를 가시화한다. SMAPE는 예측값과 실제값 간의 비율 기반 차이를 고려한 지표로 예측 오차를 정규화해 비율로 나타내어 양방향 오차를 대칭적으로 반영하여 시계열 예측에서 자주 사용된다. 이로써 단순 오차 크기를 넘어 데이터 분포 특성까지 반영하여 모델이 전체 데이터 패턴을 적절히 학습하였는지 종합적으로 판단하는 기준을 제공한다. 위에서 언급한 MAE, RMSE, R^2-$score$, SMAPE는 각각 아래와 같이 계산된다:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad (1)$$

$$RMSE = \sqrt{\sum_{i=1}^{n}\frac{(\hat{y}_i - y_i)^2}{n}} \quad (2)$$

$$R^2 - score = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (3)$$

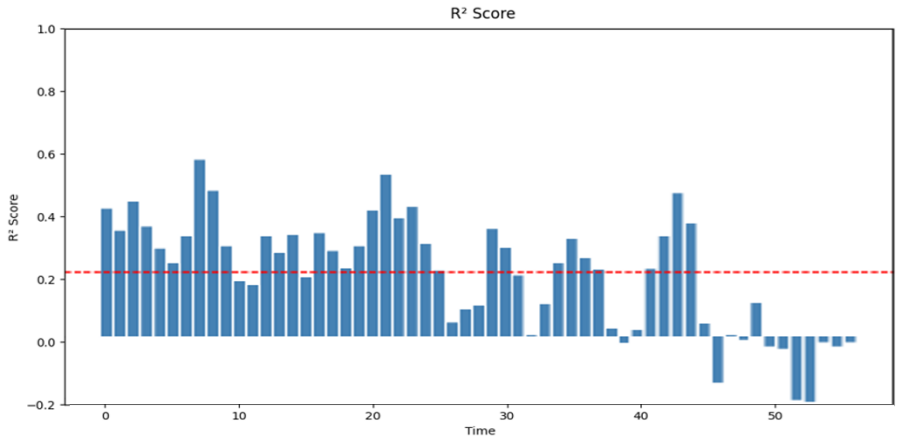
$$SMAPE = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y}_i - y_i|}{(|\hat{y}_i| + |y_i|)/2} * 100\% \quad (4)$$

MAE, RMSE, R^2-$score$, SMAPE에서 y_i는 실제 데이터 값을 의미하고 $\hat{y}_i$는 모델을 통해 예측한 값을 의미한다. n은 전체 데이터의 개수로 결과적으로 관측값과 예측값 간의 평균적인 절대 차이를 나타낸다. 그리고 R^2-$score$에서 $\bar{y}$는 실제 데이터의 평균을 의미하는 것으로 모델이 데이터를 얼마나 잘 설명하는지 나타내는 척도이다. 이러한 지표를 병행하여 활용함으로써, 본 연구에서는 예측 모델의 정밀성, 안정성, 설명력을 균형 있게 평가하고, 증강 데이터 활용에 따른 성능 향상을 다각도로 검증한다.

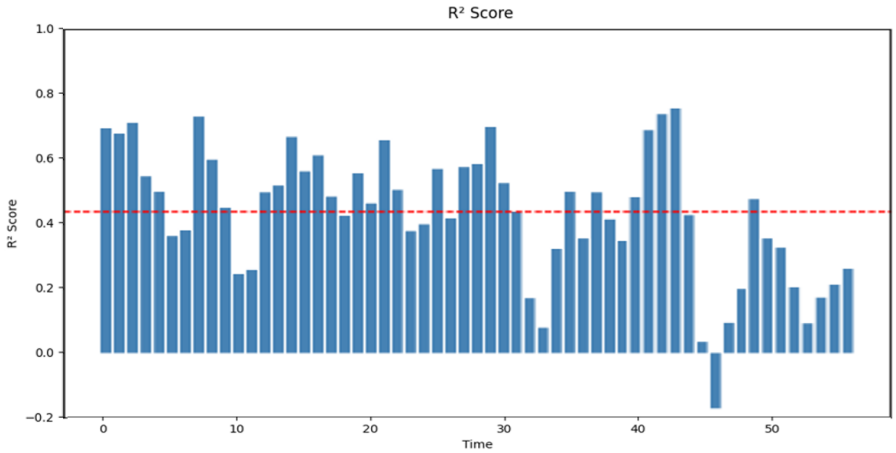
그림 4에 따르면 기존의 데이터만을 사용하여 학습한 경우 R^2-$score$는 0.2105, MAE는 261.84, RMSE는 374.62, SMAPE는 79.39%로 측정되었다. 디퓨전(diffusion) 모델로 생성된 데이터를 추가로 활용한 경우 R^2-$score$는 0.1835, MAE는 265.86, RMSE는 380.99, SMAPE는 78.03%로 오히려 성능이 저하되었다. 이는 단순한 디퓨전 기반 생성 데이터가 예측 모델의 성능 개선에 직접적인 도움을 주지 못할 수 있음을 의미한다. 반면, 본 연구에서 제안한 마스킹 기반 디퓨전(diffusion) 모델을 통해 생성한 데이터를 활용한 경우 R^2-$score$는 0.6673, MAE는 172.96, RMSE는 243.19, SMAPE는 69.10%로 대폭 개선되었다. 기존 데이터만을 사용한 경우 대비 R^2-$score$는 약 217% 증가했고, MAE는 33.9%, RMSE는 35.1%, SMAPE는 12.9% 감소한 수치이다. 이는 마스킹 기법이 디퓨전(diffusion) 모델의 학습 성능을 극적으로 향상시켰으며 예측 정밀도와 안정성 측면 모두에서 가장 우수한 결과를 달성했음을 보여준다.

추가적으로, 선행 연구에서 사용된 LSGAN 기반 생성 모델의 경우 R^2-$score$는 0.4363, MAE는 218.97, RMSE는 316.55, SMAPE는 78.60%로 나타났다. 이는 기존 데이터만 사용했을 때보다 성능이 개선되었지만 마스킹 기반 디퓨전(diffusion) 모델에 비해 성능이 떨어진다. 또한, 마스킹 기법을 적용한 Masking-LSGAN의 경우 R^2-$score$는 0.4259, MAE는 221.67, RMSE는 319.45, SMAPE는 76.72%로 측정되어 LSGAN과 성능 차이는 크지 않았다. 두 모델 모두 원본 데이터만을 활용한 경우보다 예측 성능이 향상되긴 했지만, 마스킹 디퓨전(diffusion) 모델만큼의 뚜렷한 개선은 나타나지 않았다.

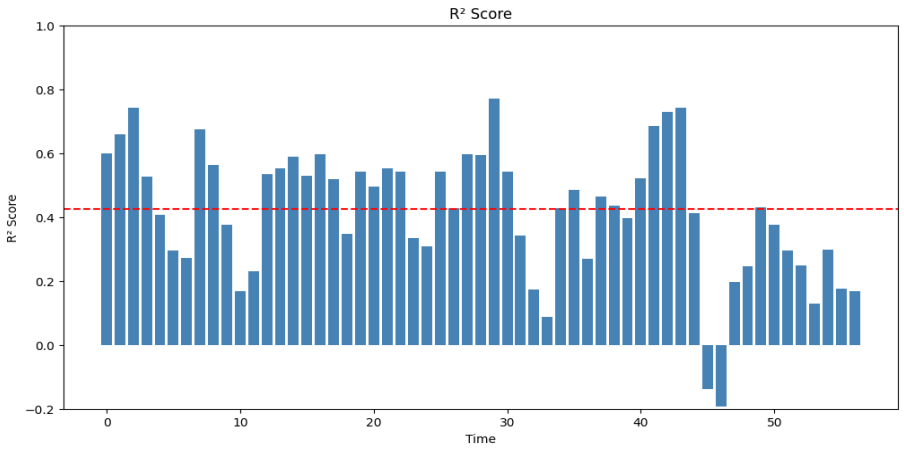
결과적으로, 본 연구에서 제안한 마스킹 기반 디퓨전(diffusion) 모델이 모든 비교군 중에서 가장 높은 예측 정확도와 가장 낮은 오차율을 달성하였으며, 증강 데이터 생성 방식으로서 가장 효과적인 접근임을 실험을 통해 입증



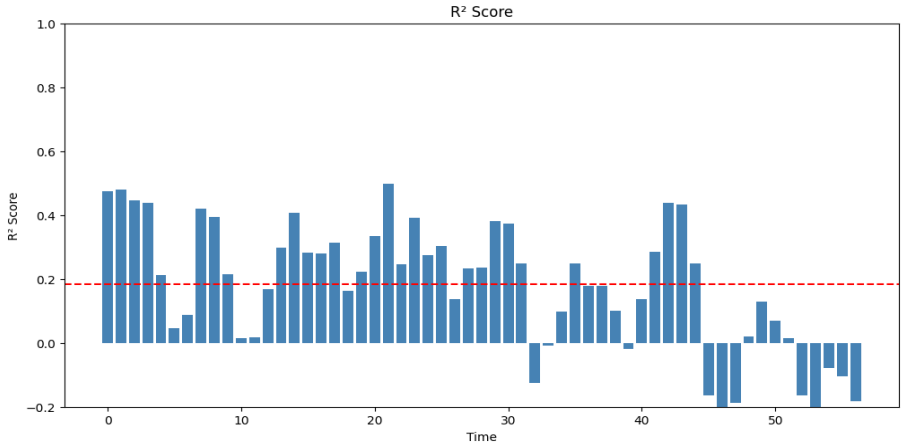
R^2-$score$ of the original data



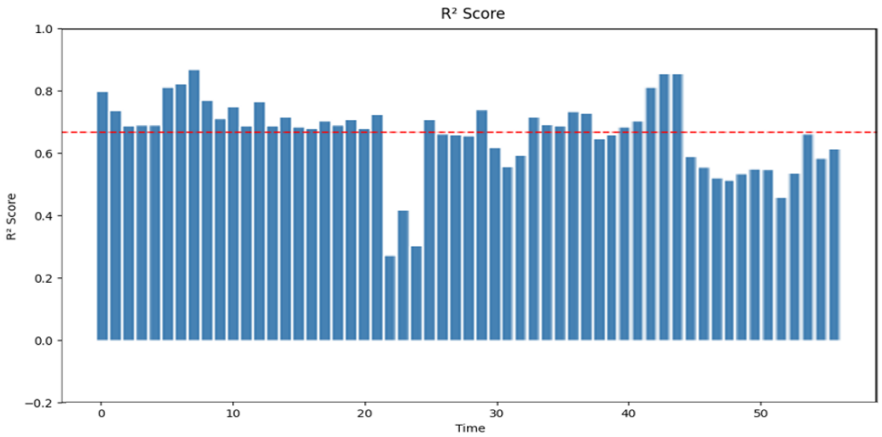
R^2-$score$ of data generated by LSGAN



R^2-$score$ of data generated by masking-LSGAN



R^2-$score$ of data generated by Diffusion Model



R^2-$score$ of data generated by Masking-based Diffusion Model

그림 4. 데이터의 R^2-$score$

Fig. 4. R^2-$score$ of data

표 1. 디퓨전(diffusion) 모델을 통해 데이터 증대를 한 결과 평가

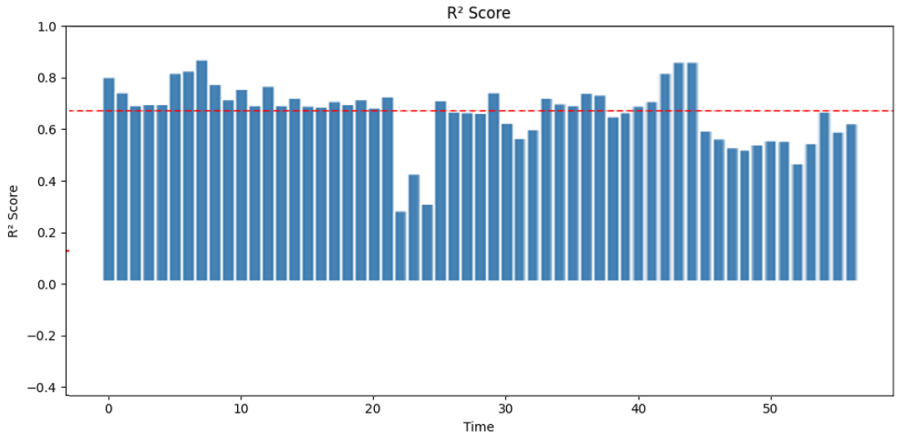
Table 1. Evaluation of data augmentation with diffusion model

Method	MAE ↓	RMSE ↓	R^2-$score$ ↑	SMAPE ↓
Original data	261.84	374.62	0.2105	79.39%
LSGAN[9]	218.97	316.55	0.4363	78.60%
Masking-LSGAN	221.67	319.45	0.4259	76.72%
Diffusion Model	265.86	380.98	0.1835	78.03%
Masking-based Diffusion Model	172.96	243.19	0.6673	69.10%

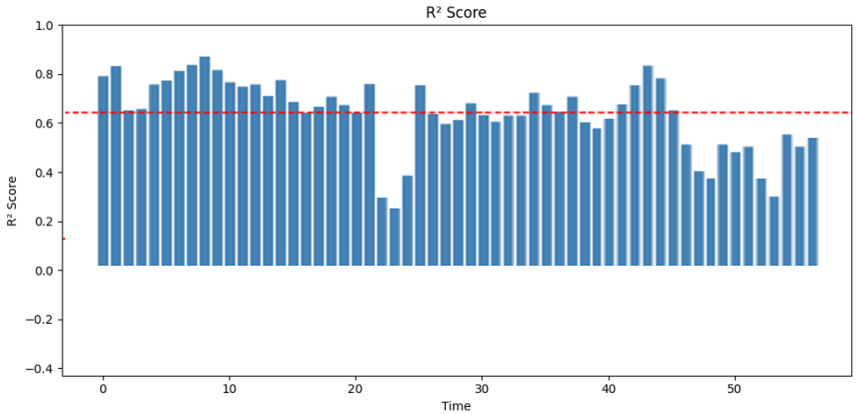
하였다. 실험의 전체적인 결과는 다음 표 1을 통해서 확인할 수 있다.

4.3 마스킹 수에 따른 성능 비교

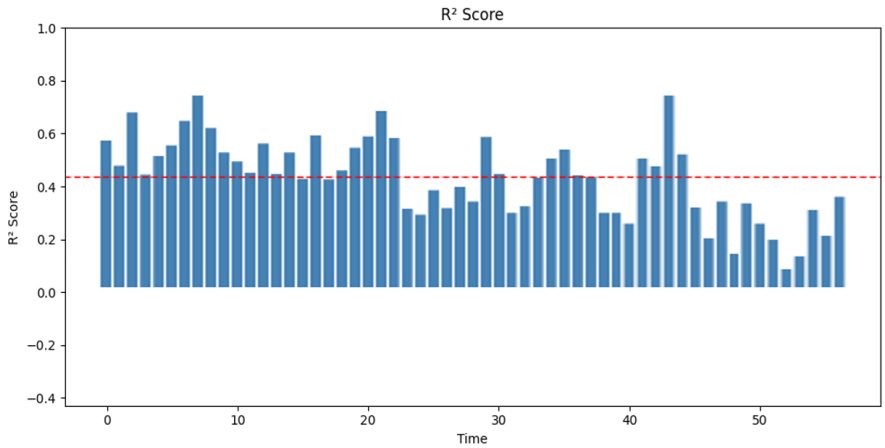
앞선 실험을 통해 마스킹을 사용한 것이 더 높은 성능을



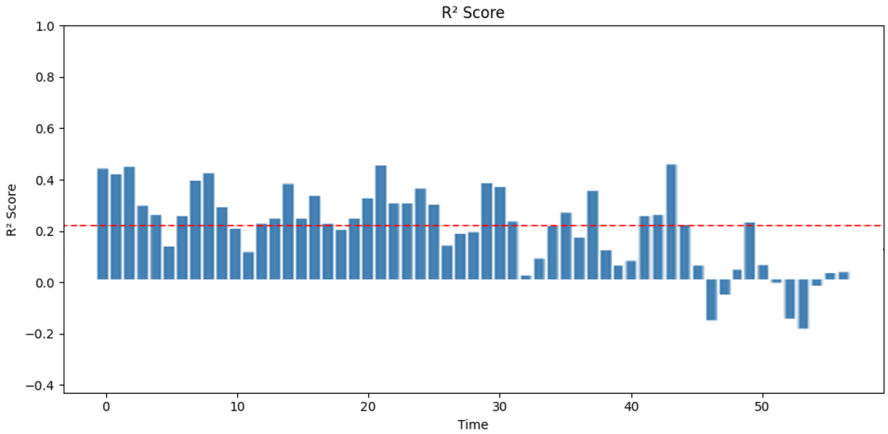
1 masking



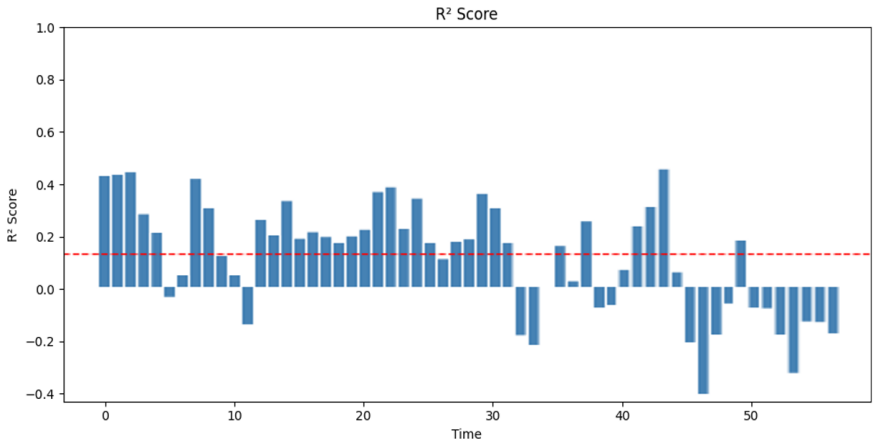
2 masking



3 masking



4 masking



5 masking

그림 5. 마스킹 수에 따른 성능 비교

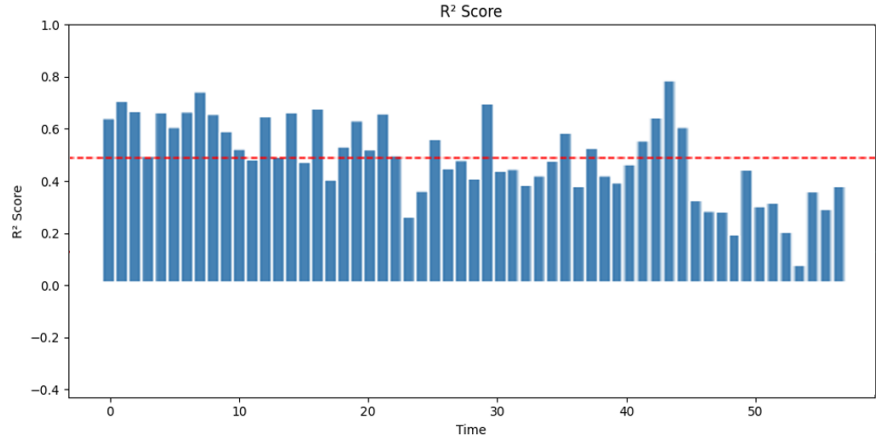
Fig. 5. Performance comparison based on the number of masking

보임을 알 수 있었다. 마스킹의 개수가 모델 성능에 어떤 영향을 미치는지 구체적으로 분석하기 위해 최적의 마스킹 개수를 찾는 추가 실험을 진행하였다. 그리고 이 과정에서 다양한 마스킹 개수를 적용한 모델의 성능을 평가하였다.

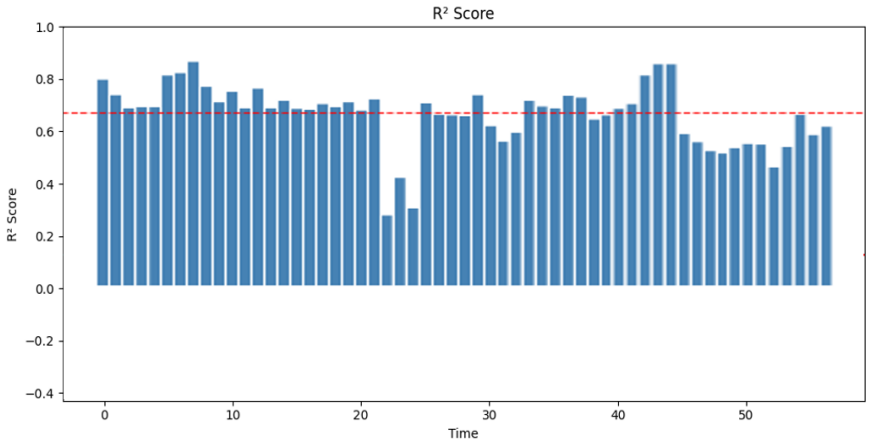
그림 5를 통해 마스킹 수에 따른 모델 성능 평가 결과에서 마스킹을 1개만 적용한 경우가 가장 우수한 성능을 보였다. R^2-*score*, MAE, RMSE, SMAPE의 평가 지표를 종합적으로 비교해볼 때, 마스킹 1개 적용 시 R^2-*score*는 0.6673로 가장 높은 값을 기록하였으며, 이는 모델이 데이터 변동성을 가장 잘 설명하고 있음을 시사한다. 또한, MAE, RMSE, SMAPE에서도 각각 179.92, 243.19, 69.10%를 기록하며, 평균 오차와 변동성을 가장 효과적으로 줄였음을 확인할 수 있다.

반면, 마스킹 수가 증가할수록 성능은 감소하는 경향을 보였다. 마스킹 2개를 적용했을 때는 R^2-*score*가 0.6365로 감소하였고, MAE와 RMSE는 각각 179.92과 254.22, SMAPE는 76.39%으로 증가하였다. 마스킹 3개 적용 시에는 R^2-*score*가 0.4262로 더 낮아졌으며, MAE는 224.04, RMSE는 319.37, SMAPE는 81.02%로 성능이 더 떨어졌다.

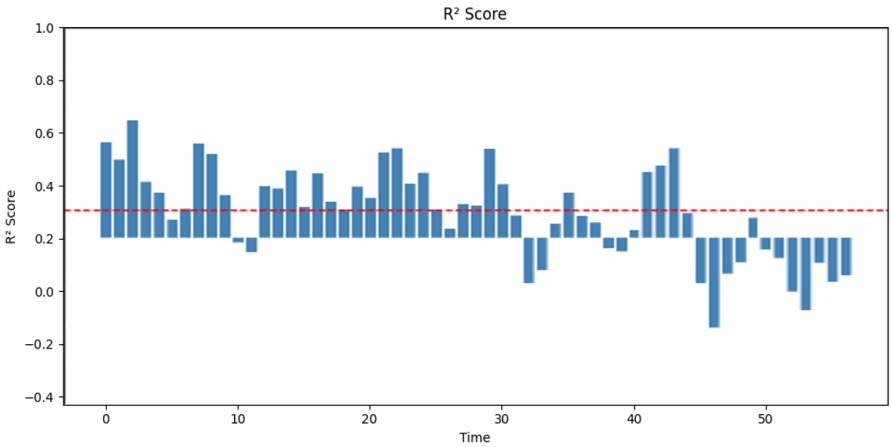
마스킹 수가 4개와 5개로 늘어날수록 성능 저하가 발생하였는데, 마스킹 4개의 경우 R^2-*score*는 0.2128, MAE는 258.95, RMSE는 374.08, SMAPE는 86.84%로 악화되었으며, 마스킹 5개에서는 R^2-*score*가 0.1255로 가장 낮은 값을 기록하였고, MAE와 RMSE는 각각 271.73과 394.28,



Augmented Samples: 10



Augmented Samples: 20



Augmented Samples: 30

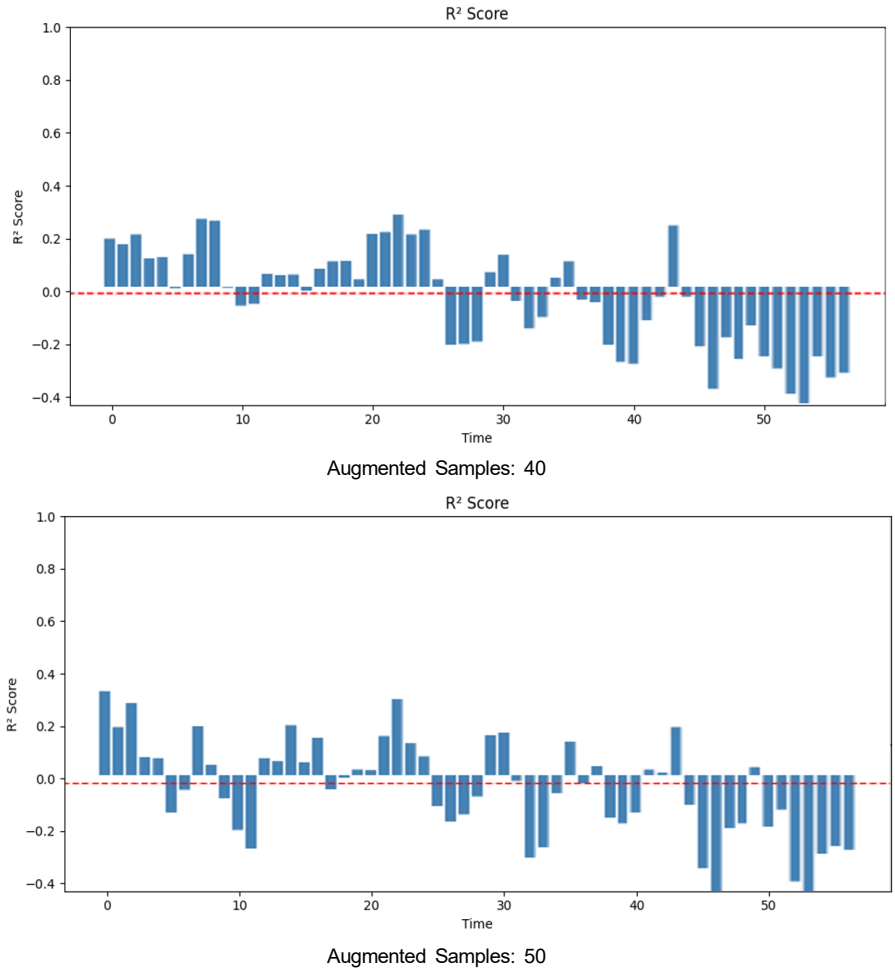


그림 6. 증강 데이터 개수에 따른 모델 성능 변화
Fig. 6. Performance variation according to the amount of augmented data used

SMAPE는 84.82%로 큰 오차와 변동성을 보였다.

이러한 결과는 마스킹 수가 증가할수록 데이터 특성 간의 상관 관계를 충분히 고려하지 않고 데이터가 생성되기 때문에 발생한 문제로 해석할 수 있다. 마스킹을 과도하게 적용할 경우, 데이터 생성 과정에서 기존 데이터의 분포와 다른 값들이 생성되어 모델이 학습 과정에서 왜곡된 패턴을 학습하게 되고, 이는 예측 성능의 저하로 이어진다. 실제로, 마스킹을 늘려 데이터 생성 실험을 진행한 결과, 생성된 데이터의 분포가 원본 데이터의 분포와 큰 차이를 보이는 경향이 확인되었다. 이는 모델이 중요한 정보를 효과적으로 학습하지 못하게 하는 주요 원인으로 작용했음을 시사한다. 반면, 마스킹을 적절히 적용한 1개 경우에는 모델이 학습 가능한 정보의 균형을 효과적으로 유지하여 예측 성능을 극대화할 수 있었음을 의미한다. 따라서 본 연구에서는 마스킹 1개를 적용하는 것이 최적의 모델 성능을 보장하는 설정임을 확인하였다.

4.4 증강 데이터 사용량에 따른 성능 변화 분석

앞선 실험에서는 마스킹 기반 디퓨전 모델을 통해 생성된 데이터를 전체적으로 활용한 결과 성능 향상이 이루어졌음을 확인하였다. 그러나 실제 환경에서는 어느 정도의 증강 데이터를 사용하는 것이 가장 효율적이고, 어느 시점부터 성능 개선이 한계에 도달하는지에 대한 분석이 필요하다. 이를 위해 기본 데이터 177개를 기반으로 추가적으로

10개, 20개, 30개, 40개, 50개의 증강 데이터를 점진적으로 추가하여 학습 성능의 변화를 관찰하였다.

그림 6은 증강 데이터 양에 따른 모델 성능 변화를 시각화한 결과이다. 실험 결과 증강 데이터가 20개까지 추가되는 경우에는 성능이 크게 향상되었다. R^2-$score$는 0.6673, MAE는 172.96, RMSE는 243.19, SMAPE는 69.10%로 기본 데이터만 사용할 때보다 모든 지표에서 뚜렷한 개선이 있었다. 이는 일정 수준의 고품질 증강 데이터가 모델의 일반화 성능을 높이는 데 유효함을 보여준다.

하지만 30개 증강 데이터를 사용했을 경우 R^2-$score$는 0.1285로 급격히 하락하였고, MAE는 273.21, RMSE는 393.60, SMAPE는 80.25%로 오히려 성능이 악화되었다. 40개, 50개로 증강 데이터 수가 더 늘어날수록 이러한 성능 저하 현상은 심화되었다. 특히 50개 증강 데이터 사용 시에는 R^2-$score$가 -0.0300, SMAPE가 103.92%까지 치솟아 모델이 데이터의 실제 분포를 제대로 학습하지 못하고 오히려 혼란을 초래했음을 알 수 있다.

이러한 결과는 디퓨전 모델이 생성한 데이터라 하더라도 지나치게 많은 데이터를 사용하면 원본 분포에서 벗어난 잡음이 모델 학습을 방해할 수 있다는 점을 시사한다. 생성 데이터의 품질뿐만 아니라 사용량의 균형이 중요하며 본 실험에서는 약 10~20개 즉, 5~10%의 증강 데이터가 최적 범위로 판단된다.

결론적으로, 마스킹 기반 디퓨전 모델로 생성한 증강 데이터를 활용할 때에는 적절한 개수의 고품질 데이터를 선별적으로 사용하는 전략이 모델의 예측 성능을 극대화하는 데 효과적이라는 것을 확인할 수 있었다.

Ⅲ. 결 론

본 논문에서는 소규모 시계열 데이터 기반 태양광 발전량 예측 문제에 대해 마스킹 기반의 디퓨전(diffusion) 모델을 활용한 테이블형 데이터 증강 기법을 제안한다. 제안된 마스킹 전략은 기존의 디퓨전 모델 기반 증강 기법과 차별화되는 접근으로 특성 간의 관계를 유지하면서도 데이터 특성값 범위 차이로 인한 왜곡을 최소화하였다. 그 결과, 생성된 데이터는 기존 데이터 분포에 근접한 형태를 보였다.

이러한 데이터 증강 기법을 적용한 결과, 원본 데이터만을 사용했을 때와 비교해 예측 성능이 크게 향상되었음을 확인한다. 이는 디퓨전(diffusion) 모델을 통한 증강 방식이 LSGAN 대비 우수한 데이터 품질과 예측 성능 향상을 제공한다는 점을 보인다.

결론적으로, 본 연구가 제안한 마스킹 기반의 디퓨전(diffusion) 모델 기반 테이블형 데이터 증강 기법은 기존 데이터 대비 정보량과 다양성을 효과적으로 확장하여 예측 성능을 높였다.

본 연구는 디퓨전(diffusion) 모델의 데이터 증강 기법이 시계열 데이터에서 효과적임을 보였지만, 여전히 몇 가지 발전 가능성이 남아 있다. 특히, 데이터셋마다 최적의 마스킹 방법이 달라질 수 있기 때문에 최적의 마스킹 수를 자동으로 결정하는 알고리즘 개발이 요구된다.

참 고 문 헌 (References)

[1] J. Hong, “Analysis of Deep Learning Model for Prediction of Solar Power Generation Based on Big Data,” Ph.D. dissertation, Mokwon University, Department of IT Engineering, Republic of Korea, pp. 1–123, 2023. http://www.riss.kr/link?id=T16623906&outLink=K

[2] J. Song, Y. Jeong, and S. Lee, “Analysis of prediction model for solar power generation,” Journal of Digital Convergence, Vol. 12, No. 3, pp. 243–248, Mar. 2014.
doi: https://doi.org/10.14400/JDC.2014.12.3.243

[3] Y. Lee, D. Kim, W. Shin, C. Kim, and H. Kim, “A Comparison of Machine Learning Models in Photovoltaic Power Generation Forecasting,” Journal of the Korean Institute of Industrial Engineers, Vol. 47, No. 5, pp. 444–458, Oct. 2021.
doi: https://doi.org/10.7232/JKIIE.2021.47.5.444

[4] D. Shin, and C. Kim, “Short Term Forecast Model for Solar Power Generation using RNN-LSTM,” Journal of Advanced Navigation Technology, Vol. 22, No. 3, pp. 233–239, Jun. 2018.
doi: https://doi.org/10.12673/JANT.2018.22.3.233

[5] W. Qingsong, S. Liang, Y. Fan, S. Xiaomin, G. Jingkun, W. Xue, and X. Huan, “Time Series Data Augmentation for Deep Learning: A Survey,” arXiv preprint, arXiv:2002.12478, 2020.
doi: https://doi.org/10.48550/arXiv.2002.12478

[6] J. Kim, and N. Moon, “CNN-GRU-Based Feature Extraction Model of Multivariate Time-Series Data for Regional Clustering,” Advances in Computer Science and Ubiquitous Computing, Lecture Notes in Electrical Engineering, vol. 715, pp. 401–405, January 2021.
doi: https://doi.org/10.1007/978-981-15-9343-7_55

[7] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley,

S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Networks,” arXiv preprint, arXiv:1406.2661, Jun. 2014.
doi: https://doi.org/10.48550/arXiv.1406.2661

[8] Y. Li, M. Zhang, and C. Chen, “A Deep-Learning Intelligent System Incorporating Data Augmentation for Short-Term Voltage Stability Assessment of Power Systems,” Applied Energy, Vol. 308, No.118347, 2022.
doi: https://doi.org/10.1016/j.apenergy.2021.118347

[9] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, “Deep Unsupervised Learning Using Nonequilibrium Thermodynamics,” Proceedings of the International Conference on Machine Learning (ICML), Lille, France, pp. 2256 – 2265, Jul. 2015.

[10] L. Lin, Z. Li, R. Li, X. Li, and J. Gao, “Diffusion Models for Time Series Applications: A Survey,” arXiv preprint, arXiv:2305.00624, 2023.
doi: https://doi.org/10.48550/arXiv.2305.00624

[11] C. Meijer and L. Y. Chen, “The Rise of Diffusion Models in Time-Series Forecasting,” arXiv preprint, arXiv:2401.03006v2, 2024.
doi: https://doi.org/10.48550/arXiv.2401.03006

[12] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” arXiv preprint, arXiv:2006.11239, 2020.
doi: https://doi.org/10.48550/arXiv.2006.11239

[13] J. Song, C. Meng, and S. Ermon, “Denoising Diffusion Implicit Models,” Proceedings of the International Conference on Learning Representations (ICLR), 2021. https://openreview.net/forum?id=St1giarCHLP

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10684 – 10695, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.01042

[15] W. Lee and Y. Kim, “Predicting Photovoltaic Power Generation with Random Forests,” Proceedings of the Korean Institute of Information Processing Society 2016 Fall Conference, pp. 397 – 400, Oct. 2016.

[16] P. Dhariwal and A. Nichol, “Diffusion Models Beat GANs on Image Synthesis” Advances in Neural Information Processing Systems (NeurIPS), Vol. 34, pp. 8780 – 8794, 2021.
doi: https://doi.org/10.48550/arXiv.2105.05233

[17] L. van der Maaten and G. Hinton, “Visualizing Data using t-SNE,” Journal of Machine Learning Research, Vol. 9, pp. 2579 – 2605, 2008.

[18] M. Gulati and P. Roysdon, “TabMT: Generating Tabular Data with Masked Transformers,” Advances in Neural Information Processing Systems, 2023.

[19] N. Dua, S. N. Singh, V. B. Semwal, et al., “Inception inspired CNN-GRU hybrid network for human activity recognition” Multimedia Tools and Applications, Vol. 82, pp. 5369 – 5403, 2023.
doi: https://doi.org/10.1007/s11042-021-11885-x

[20] S. Lee, D. Cho, and M. Lee, “Spatializing the Pearson’s Correlation Coefficient: An Experimental Comparison of Three Relevant Techniques” Journal of the Korean Geographical Society, Vol. 53, No. 5, pp. 761 – 776, 2018.

저 자 소 개

이 재 호

- 2019년 3월 ~ 2025년 2월 : 경희대학교 응용과학대학 응용수학과 학사
- ORCID : https://orcid.org/0009-0000-9987-1671
- 주관심분야 : 시계열 데이터 분석, 금융

임 종 경

- 2010년 3월 ~ 2017년 2월 : 경희대학교 전자전파공학과 학사
- 2017년 6월 ~ 2020년 1월 : 세메스 주식회사 S/W 엔지니어
- 2021년 2월 ~ 2021년 10월 : LG이노텍 S/W 엔지니어
- 2022년 3월 ~ 현재 : 경희대학교 인공지능학과 석박·통합과정
- ORCID : https://orcid.org/0009-0005-7695-0752
- 주관심분야 : 객체 탐지, 모델 경량화

저 자 소 개

배 성 호

- 2004년 3월 ~ 2011년 2월 : 경희대학교 전자정보대학 컴퓨터공학 및 전자공학 공학사(복수전공)
- 2011년 2월 ~ 2016년 8월 : KAIST 전기 및 전자공학과 공학박사
- 2016년 7월 ~ 2017년 8월 : MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) 박사 후 연구원
- 2017년 9월 ~ 현재 : 경희대학교 전자정보대학 컴퓨터공학과 부교수
- ORCID : https://orcid.org/0000-0002-3389-1159
- 주관심분야 : 영상 처리, 비디오 압축, 모델 경량화