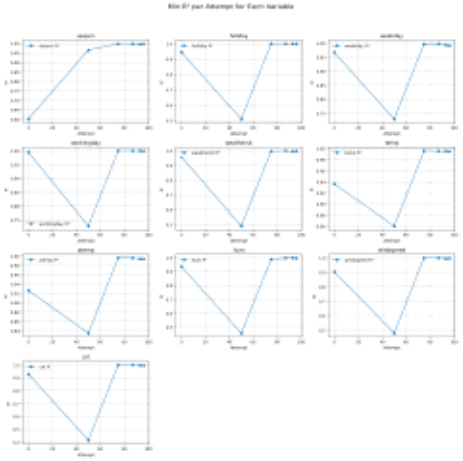
| 2025 학년도 1학기 유레카 탐구대회 | ※ 번호<br>(미기재) |
|---|---|
| 과제연구 과목 | 연구 주제 |
| 정보 | AI의 재귀 학습의 최대 성능 및 붕괴 임계점 분석 |

## 1. 탐구 동기와 목적

기계학습과 대형 언어 모델 모델의 급격한 발전을 계기로, AI가 생성한 데이터가 웹상에 개시되고 이를 AI가 다시 학습하는 재귀적 관계가 구축되고 있다. 이는 AI가 학습할 수 있는 정보의 다양성을 제한하고, 궁극적으로는 AI 모델의 성능이 하락하는 AI 붕괴의 위협이 있다. 본 연구자는 AI 모델 붕괴 현상을 해결하기 위해, 실험적으로 AI 붕괴를 관측하고자 한다. 또한, 재귀적 학습을 고려한 최적의 인간-AI 데이터 학습 비율의 존재성을 구하며, 이를 변수의 특성에 따른 손실함수를 사용하여 확인해보고자 한다.



## 2. 탐구 내용

### 가. 선행 연구 고찰 및 탐구의 독창성

Ilia Shumailov 등(2023)의 연구는 AI가 스스로 생성한 데이터를 반복적으로 학습할 때 정보의 다양성이 줄어들고, 결과적으로 원본 데이터 분포에서 멀어지는 현상을 실험적으로 확인했다. 이들은 위키 기반 이미지 데이터를 기반으로 CNN과 Diffusion 모델을 활용해, 재귀적 학습을 거듭할수록 출력물이 점차 단조롭고 왜곡되는 경향을 확인했다. 해당 연구는 AI 붕괴가 실제로 발생할 수 있는 위험이라는 점을 처음으로 실험적으로 보여주었지만, 어느 시점부터 모델이 성능을 잃기 시작하는지에 대한 구체적인 임곗값이나 인간-생성 데이터의 비율과 같은 조건은 다루지 않았다.

본 연구는 다양한 방식의 경향성을 갖는 데이터, 특히 시계열 데이터에 특화된 구조인 LSTM 기반 모델을 활용한다. 이를 통해 AI 붕괴 현상이 일반 데이터 환경에서 어떠한 특성이 나타나는지를 확인하고자 한다. 또한, 인간이 생성한 데이터와 AI가 생성한 데이터를 조합했을 때 학습 성능 변화를 실험적으로 분석함으로써, 장기적인 관점에서 AI 모델을 안정적으로 유지하기 위한 최적의 데이터 비율을 확인하고자 한다. 본 연구는 AI 붕괴에 대한 실증적 확인 및 정량적인 데이터 평가를 진행한다는 점에서 기존 연구와 차별화된다.

### 나. 탐구 절차 및 방법

실험 1에서는 LSTM 모델을 활용해 AI 붕괴 현상을 정량적으로 분석하고 성능 붕괴의 임계점을 파악한다. 1세대 LSTM 모델을 학습시켜 대조군을 확보한 뒤, 같은 구조를 출력하는 모델을 기반으로 이전 세대의 예측 데이터를 다음 세대의 학습 데이터로 사용하는 방식으로 작동하도록 하였다. 총 45세대까지 반복 학습을 진행한 후, 각 세대의 예측값과 실제값, 손실함수를 계산하여 저장하고, 오차 추이를 시각화하여 성능 붕괴가 시작되는 시점을 분석하였다. 실험 2에서는 인간 데이터와 AI 데이터의 구성 비율에 따른 모델 성능 적합성을 분석하였다. 이를 위해 1세대 LSTM이 생성한 데이터를 100% AI 데이터로, 원본 데이터를 100% 인간 데이터로 간주하였으며, 이진 탐색 알고리즘을 활용해 다양한 비율의 혼합 데이터를 구성하였다. 각 비율에 대해 같은 구조의 모델을 학습시킨 후, 10세대에 걸쳐 예측 오차를 비교, 가장 낮은 오차를 기록한 비율을 최적의 인간-AI 데이터 비율로 설정하였다.

**다. 탐구 결과**

1번 실험에서는 같은 구조의 LSTM 모델을 45세대까지 반복 학습한 결과, 성능 저하 현상이 특정 세대 이후 반복적으로 나타났다. 3세대까지는 손실함수 값이 안정적으로 유지되었으나, 6세대에서 손실함수가 급격히 증가하는 양상을 확인했다. 이후 약 7세대 주기로 오차가 상승하는 경향이 반복되었다. 특히 20세대 이후부터는 손실함수의 최댓값이 주기적으로 증가하는 현상이 나타났으며, 모든 변수에서 평균 약 5% 수준의 손실 증가가 관찰되었다. season, weekday, hum 등의 변수는 지표가 0.92 이하로 떨어지는 구간이 반복적으로 발생하였고, hum, atemp, temp 변수에서는 MAE 및 RMSE 기준 손실이 점진적으로 증가하는 양상이 뚜렷했다. 또한, 30세대 이후에는 손실함수 값이 급격히 진동하고 예측 성능이 저하되는 현상이 집중적으로 발생하였다.

2번 실험에서는 인간 데이터와 AI 생성 데이터의 비율을 조정하며 예측 성능을 측정한 결과, 손실함수에서 인간 데이터 비율이 87.5%일 때 가장 낮은 오차값을 기록하였다. 인간 데이터와 AI 데이터가 50%씩 혼합되었을 때 거의 모든 변수에서 성능이 급격히 저하되었으며, 이후 인간 데이터 비율이 75% 이상으로 증가함에 따라 손실함수 값이 다시 감소하였다. 연속형 변수인 temp, cnt는 RMSE 손실함수 기준으로 75% 구간에서 급격한 오차 감소가 나타났고, season, holiday 등의 변수는 MAE 손실함수 기준으로 96-100% 구간에서 손실 값이 다시 증가하였다. hum과 holiday 변수는 이상값에 민감한 HUBER 손실함수에서, weekday, workingday는 주기성을 고려한 Log-Cosh 손실함수에서 각각 50% 지점에서 가장 큰 손실을 보였으며, 이후 87.5% 구간에서 성능이 회복되었다. 전체적으로 인간 데이터 비율이 87.5%일 때 가장 안정적이고 낮은 손실 값을 보이는 경향이 공통으로 확인되었다.

**3. 결론 및 발전 가능성**

본 연구는 입출력 동일 구조의 LSTM 모델을 기반으로, 반복 학습 세대에 따른 성능 변화와 인간-AI 데이터 비율에 따른 손실함수의 민감도를 분석하였다. 그 결과, 반복 학습이 일정 세대 이상 진행되면 성능이 저하되는 경향이 뚜렷하게 나타났으며, 이는 과적합이나 데이터 적합의 비선형적 구조적 한계에서 비롯됨을 의미한다. 또한, AI 생성 데이터가 일정 비율 이상 포함될 때 모든 손실함수에서 성능이 급격히 저하되었고, 인간 데이터 비율이 87.5%일 때 예측 정확도가 가장 높게 나타났다. 이는 해당 지점에서 AI의 데이터 제작과 AI 붕괴 사이의 적합 지점임을 나타내며, 최적의 인간-AI 데이터의 비율이 존재함을 실증적으로 확인하였다.

본 연구 결과를 토대로 향후 연구에서는 변수별 손실함수의 세부적인 성능 차이를 더욱 정밀하게 분석하고 시각화하는 평가 체계를 구축할 필요가 있다. 또한 자연어 모델 및 Transformer 등 고급 알고리즘 등을 통한 데이터 구조 변경, 노이즈의 특성, 인간 데이터의 오류 검정 등 실제 환경에서 복합적으로 작용하는 요소들을 반영한 다양한 조건에서 손실함수의 유연성과 안정성을 검증하는 후속 실험이 요구된다. 아울러 인간 데이터 확보에 따른 비용과 시간의 제약을 고려할 때의 비용을 고려하여, AI 모델의 효율성을 극대화하는 방법론적 개발 또한 중요한 과제 중 하나이다.