공학석사학위논문

주식 매매를 위한 강화학습에서의 대조적 표현학습 연구

Reinforcement Learning for Stock Trading based on Contrastive Representation Learning of Market States

2023 년 8 월

서울대학교 대학원

산업공학과

정 지 문

주식 매매를 위한 강화학습에서의 대조적 표현학습 연구

Reinforcement Learning for Stock Trading based on Contrastive Representation Learning of Market States

지도교수 박 종 헌

이 논문을 공학석사 학위논문으로 제출함

2023 년 6 월

서울대학교 대학원

산업공학과

정 지 문

정지문의 공학석사 학위논문을 인준함

2023 년 7 월

위 원 장 조 성 준 (인)

부위원장 이 경 식 (인)

위 원 박 종 헌 (인)

초록

주가 시계열 데이터는 불규칙성이 강해 각 상황의 특징을 정확하게 파악하기 어렵다. 이러한 어려움 때문에 효과적인 주식 매매 전략에 대한 연구가 오랫동안 진행되어 왔다. 최근에는 강화학습에 심층학습을 결합한 주식 매매 연구가 주목을 받고 있다. 강화학습은 관찰한 상태(state)를 통해 행동(action)을 선택하기 때문에 주가의 현재 정보를 잘 표현하는 상태 선정이 매우 중요하다. 기존 강화학습 주식 매매 연구들은 규칙기반으로 추출한 주가 시계열의 특징이나 차원축소 또는 군집화와 같은 방법을 적용하여 추출한 정보를 강화학습 모델의 상태로 사용하고 있다.

본 논문에서는 최근 시계열 특징 추출에서 우수한 성능을 보이고 있는 대조학습을 활용하여 주가 데이터의 표현(representation)을 추출하고, 이를 강화학습 주식 매매 모델의 상태로 활용하는 방법을 제안한다. 실제 주식 데이터를 통해 실험한 결과 대조학습으로 추출한 주가의 표현을 기술적 지표와 함께 강화학습 모델의 상태로 사용하였을 때 기존 연구들에 비해 더 높은 수익률을 낼 수 있음을 확인하였다. 또한 대조학습으로 추출한 표현을 상태로 사용 시 강화학습 모델이 더 소극적인 매매를 하는 특징을 발견하였다. 이러한 결과는 대조학습을 통해 추출한 주가 데이터의 표현이 강화학습 주식 매매 모델의 성능을 향상시키는 데 중요한 역할을 할 수 있음을 시사한다.

주요어: 강화학습, 주식 매매, 시계열, 표현학습, 대조학습, 심층학습
학번: 2021-23086

목차

표 목차

그림 목차

제 1 장 서론

1.1 연구 배경 및 동기

최근 강화학습 알고리즘을 활용한 주식 매매 결정 연구가 많은 관심을 받고 있다. 강화학습은 관찰된 상태(state)를 기반으로 행동(action)을 선택하게 하여 agent가 주식 시장에서 이익을 극대화하는 전략을 학습하는 방법론이다 [44]. 관측된 상태로 행동을 선택하기 때문에 주가 데이터의 정보를 효과적으로 표현하는 상태 선정이 매우 중요하다. 기존 연구들은 규칙기반 지표를 강화학습 모델의 상태로 사용하여 주가에 대한 정보를 설명하였다 [4, 15]. 그러나 이런 방법들은 규칙이 지정한 영향인자 외의 변화에 대해 설명력이 떨어져 주식 시장의 동적인 특성을 충분히 설명하기 어렵다.

최근 시계열에 대조학습(contrastive learning)을 적용하여 특징을 효과적으로 추출하는 연구들이 예측 및 이상치 탐지 분야에서 좋은 성능을 보이고 있다 [1, 2]. 시계열의 대조학습 연구는 데이터를 다양한 변형기법을 통해 증강하고 비슷한 패턴을 갖는 데이터 쌍을 구성하여 양성 데이터(positive pair) 쌍과 음성 데이터(negative pair) 쌍으로 구분한다 [7]. 양성 데이터 쌍과 음성 데이터 쌍의 유사점과 차이를 학습하며 시계열 데이터의 패턴을 잘 표현하는 표현을 생성한다.

본 논문에서는 이러한 대조학습을 통해 생성된 주식 데이터의 표현을 주식 매매 강화학습 모델의 상태로 활용하는 방법을 제시한다. 이를 통해 다양한 맥락(context)에서도 주가 데이터의 특징을 잘 추출할 수 있어, 이에 따라 주식 매매 수익률을 향상

시킬 수 있는 새로운 모델과 전략을 개발하게 되었다. 본 연구는 주식 매매 강화학습 모델이 더욱 효과적이고 의사결정을 지원할 수 있는 새로운 접근법을 제안한다.

1.2 연구 목적

본 연구에서는 주식 매매 강화학습 모델이 주가 데이터의 상태에 대한 정보를 더 정확하게 관측하여 더 나은 매매 의사결정을 할 수 있는 방법론을 제안한다. 대조학습을 통해 주가 시계열에 데이터의 특징을 추출하고 이를 강화학습 주식 매매 모델의 상태로 활용하여 주식 매매 의사결정을 돕는다. 실제 주식 데이터 셋에서 실험을 진행하며, 대조학습으로 추출한 표현의 효과를 평가하기 위해 기술적 지표만을 상태로 사용한 주식 매매 모델과 성능을 비교한다. 기존 연구들과 결과 비교를 통해 동일 기간에 대한 수익률이 향상됨을 보이고 대조학습을 통해 추출한 주가 시계열 표현의 주식 매매 강화학습 모델에서의 효과성을 보이는 것을 목적으로 한다. 연구 목적을 정리하면 다음과 같다.

(a) 대조학습을 이용한 표현학습으로 주가의 현재 상태에 대한 정보를 잘 반영하는 표현을 추출한다.

(b) 강화학습 주식 매매 모델의 상태로 주가 시계열 데이터의 표현을 활용하여 각 상황에서 최대 이익을 얻도록 의사결정을 돕는 딥러닝 기법을 연구한다.

1.3 문제 정의

본 논문의 문제 정의는 대조학습을 이용하여 추출한 주가 시계열 데이터의 표현과 주가에 대한 정보를 나타내는 OHLCV(Open, High, Low, Close, Volume) 및 기술적 지표(technical indicator)들을 강화학습의 상태로 활용하여 주식 매매 의사결정을 하는 것이다. 종목의 종가를 기준으로 일 단위로 주식을 매수, 매도 또는 아무것도 하지 않는 행동으로 주식 매매를 진행하게 된다.

주식 매매의 성과를 평가하기 위해, 여러 종목에 대해 대조학습을 통해 추출한 표현을 활용한 강화학습 모델의 수익률과 활용하지 않은 모델의 수익률을 비교한다. 이를 통해 대조학습을 통해 추출된 표현의 효과를 실험으로 확인하고자 한다. 본 논문의 문제를 정의한 그림이 1.1에 제시되어 있다.

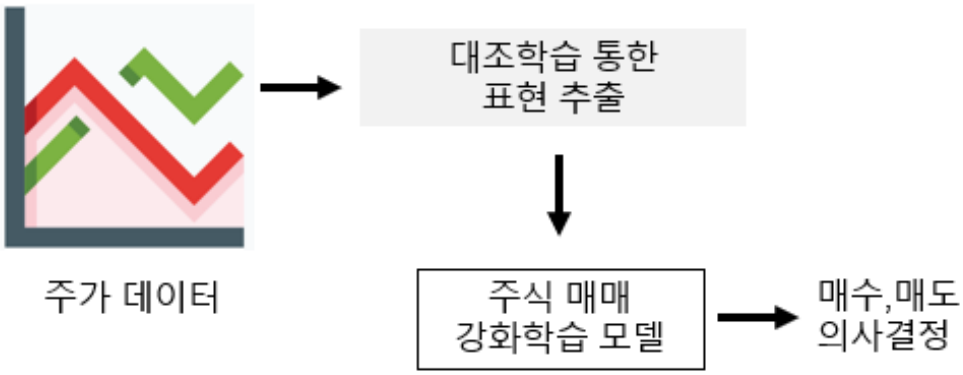

그림 1.1: 문제 정의

1.4 논문 구성

본 논문은 총 5장으로 구성된다. 2장에서는 본 연구에서 활용하는 대조학습과 강화학습에 대해 설명하고, 관련된 선행 연구를 소개한다. 3장에서는 제안 기법의 원리와 특징에 대해 자세히 설명한다. 4장에서는 실험에 사용된 데이터 셋과 실험 결과를 발표한다. 마지막 5장에서는 실험 결과 해석을 통해 연구의 결론에 대해 정리하고 향후 연구 방향을 제시한다.

제 2 장 배경 이론 및 관련 연구

2.1 배경 이론

2.1.1 표현학습(Representation Learning)

표현학습은 인공지능 모델이 데이터에서 주요한 패턴과 구조를 인식하고 추출하는 방식을 일컫는다 [10-12]. 표현학습을 통해 원시 데이터를 훨씬 더 사용 가능하고 효율적인 형태로 바꿔주어 모델은 예측, 분류, 의사결정 등의 작업을 훨씬 더 편리하게 수행할 수 있다.

표현학습은 주어진 데이터를 설명하는 핵심적인 요소나 개념을 찾아낼 수 있어 고차원 데이터를 다루는데 큰 이점이 있다. 심층학습 기반 표현학습은 복잡한 패턴과 다양한 수준의 데이터 추상화를 파악하기 위해 심층학습 알고리즘을 활용하므로 더욱 정확하고 의미 있는 데이터 표현을 제공한다 [22].

표현 학습은 다양한 분야에서 활용된다. 컴퓨터 비전에서 표현 학습은 이미지나 비디오 데이터에서 중요한 특징을 추출하는 데 사용된다 [28]. 이러한 특징은 사물 인식, 얼굴 인식, 감정 인식 등의 작업에서 기계가 이미지를 이해하는 데 도움을 준다. 또한 자율 주행 자동차와 같이 실시간으로 많은 양의 시각적 정보를 처리해야 하는 시스템에서도 표현 학습이 중요하게 사용된다. 자연어 처리에서 표현 학습은 텍스트 데이터의 의미 있는 특징을 추출하고 이해하는 데 사용된다 [30]. 이는 현재 번역, 감성 분석, 요약, 질문 응답 등의 작업을 수행하는 데 필수적인 기술이 되었다.

이처럼, 표현 학습은 복잡한 데이터에서 중요한 특징을 찾아내고 이를 인공지능 모델에 이용할 수 있도록 변환하는 데 중요한 역할을 한다. 이를 통해 모델은 더 효율적으로 작업을 수행하고, 더 정확한 예측을 하는 데 도움을 받을 수 있다.

2.1.2 대조학습(Contrastive Learning)

표현학습의 일부인 대조학습은 딥러닝 연구에서 사용되는 비지도 학습 방법으로 특징이 비슷한 샘플들은 같은 표현을 다른 샘플들은 다른 표현을 갖는 것을 목표로 학습을 진행한다 [28]. 데이터의 레이블이 제한적이거나 얻기 어려운 상황에서 유용하게 활용된다.

대조학습은 학습 데이터를 양성 데이터 쌍과 음성 데이터 쌍으로 구성한다. 양성 데이터 쌍은 비슷한 특징을 가진 두 샘플로 구성되며, 음성 데이터 쌍은 서로 다른 특징을 가진 두 샘플로 구성된다. 모델은 양성 데이터 쌍을 가깝게, 음성 데이터 쌍을 멀게 표현할 수 있도록 학습된다. 이를 통해 모델은 데이터의 중요한 특징을 추출하고 유사성 및 차이점을 인식하는 능력을 학습한다.

원본 데이터를 변형하거나 조정하여 새로운 학습 데이터를 생성하는 데이터 증강(augmentation) 방법을 적용하여 대조학습의 성능을 향상시키는 연구들이 활발히 진행되고 있다 [7, 30]. 원래 샘플을 변형한 샘플을 양성 데이터 쌍, 다른 샘플과 그 변형된 샘플을 음성 데이터 쌍으로 인식하여 학습을 진행하면 과적합을 방지하여 성능을 높이는데 도움이 된다.

대조학습은 데이터의 본질적인 특성과 패턴을 효과적으로 학습하여 다양한 분야에

서 활용될 수 있다. 컴퓨터 비전 분야에서 유사한 이미지들을 군집화 하거나 유사성을 평가하여 이미지 인식, 분류, 세분화 [29] 등의 문제를 해결하는데 효과적으로 사용되고 있으며, 자연어 처리 분야의 단어의 뜻이나 문맥의 의미를 이해하는 작업 등에서도 다양하게 적용되고 있다 [30].

2.1.3 강화학습(Reinforcement Learning)

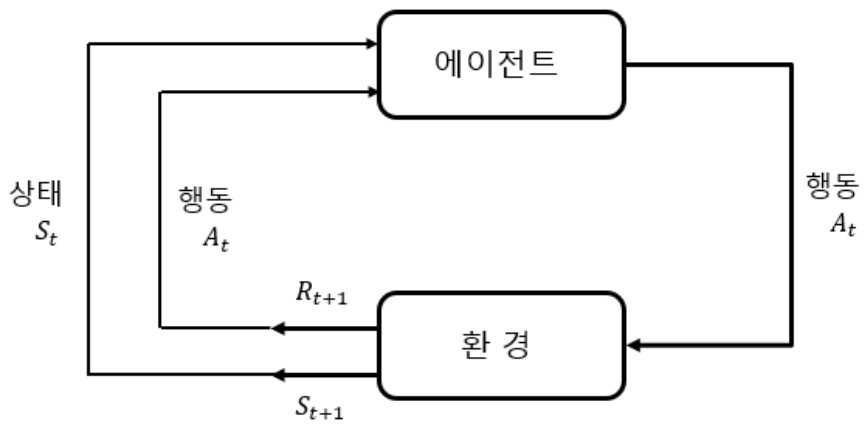

그림 2.1: 강화학습 개요

강화학습은 에이전트(agent)가 환경과 상호작용하며 행동을 결정하고, 그 결과에 따라 보상을 받아 학습하는 방법이다. 에이전트는 학습의 주체로 주어진 환경의 현재 상태에서 어떤 행동을 취할지 결정하고, 그 행동에 따른 보상을 받는다 [44]. 이러한 행동과 보상 반복 과정을 통해 강화학습은 최적의 행동 정책을 학습하게 된다. 그림

2.1이 강화학습의 원리를 간단히 설명한다.

강화학습은 상태, 행동, 보상, 상태 전이 확률 등을 포함하는 수학적 모델 마르코프 의사결정 과정(Markov Decision Process, MDP)을 기반으로 한다. 에이전트는 MDP의 상태, 보상과 특정 상태의 가치에 대한 값인 가치함수 그리고 특정 상태에서 특정 행동이 얼마나 좋은지에 대한 값인 Q함수(행동 가치 함수) 기반으로 최적의 행동 과정책을 찾는 과정을 수행한다.

최근에는 강화학습에 심층학습을 적용하여 가치함수나 Q함수, 정책 등을 근사하는 방식으로 높은 차원의 복잡한 문제에서도 훌륭한 성능을 보이고 있다 [43, 48]. 심층학습을 적용한 강화학습은 게임이나 로봇 제어, 자율 주행 등의 분야에 적용되어 복잡한 의사결정 문제를 해결하고, 최적의 전략을 발견하는데 사용되고 있다.

2.2. 관련 연구

2.2.1 시계열 특징 추출 연구

시계열 데이터의 특징 추출 연구는 규칙기반(rule-based) 방법 연구와 딥러닝 활용 방법 연구로 크게 구분된다. 규칙 기반 방법으로는 군집화 기법을 이용한 연구와 기술적 지표 생성 연구가 있다. 군집화를 적용한 연구에서는 길이가 다른 시계열 데이터의 유사도를 측정하기 위해 동적 시간 정합(Dynamic Time Warping)을 사용하고, 이를 기반으로 클러스터링을 진행한 후 클러스터 간의 특징을 추출한다 [12]. 또한, 시계열 데이터를 주파수 영역으로 분해하는 푸리에(Fourier) 변환, 웨이블렛(Wavelet) 변환 등의 방법을 활용하여 기존 트렌드를 벗어나는 전조 지표를 개발하는 연구들도 다수 존재한다 [13].

심층학습 기반 차원 축소 연구에서는 LSTM 방식이나 오토인코더(auto-encoder)를 활용하여 시계열 데이터를 차원 축소하는 방법이 연구되고 있다 [3, 5]. 에너지 또는 주가 데이터를 차원 축소하여 심층학습 기반 예측모델의 입력으로 사용하는 연구가 활발하게 이루어지고 있다 [22, 24].

시계열 데이터에 대조학습을 적용하여 특징을 추출하는 연구는 시계열 데이터를 특정 윈도우 크기로 분할한 후 유사한 패턴을 가진 데이터를 양성 데이터 쌍으로 선택하고 유사하지 않은 패턴을 가진 샘플을 음성 데이터 쌍으로 선택하여 학습을 진행한다 [20]. 데이터를 다양한 방식으로 증강하여 학습의 효과를 높이는 방향의 연구들이 진행되고 있다. 원본 시계열 데이터를 변형시켜 데이터를 증강한 후, 대조학습을 통해 특징을 학습하는 연구와[1] 시계열 데이터를 시간 영역과 주파수 영역으로 분해한 후

대조학습을 적용하여 시계열의 특징을 추출하는 연구 등이 존재한다 [2, 10].

2.2.2 주식 매매 강화학습 연구

강화학습을 활용한 주식 매매 연구는 다양한 알고리즘을 적용하여 진행되고 있다. DQN 알고리즘을 개선한 Double DQN, Dueling DQN 등의 알고리즘으로 [48, 49] 학습의 수렴을 더 안정화하여 성능을 개선한 연구가 존재한다. 정책 경사법(policy gradient) 계열의 알고리즘으로 매매 행동을 결정하는 정책을 최적화하는 시도들도 존재한다 [4, 46]. 정책 경사법 계열의 알고리즘은 행동의 수가 제한적인 DQN 계열 알고리즘에 비해 연속적인 행동을 선택할 수 있다는 장점이 있다.

주식 매매 강화학습의 상태 선정에 대한 연구는 규칙기반 상태와 심층학습 기반 상태 사용 연구들이 진행되고 있다. 규칙기반 연구로는 주가의 기술적 지표를 새롭게 개발하여 상태로 사용한 연구들이 존재하고 주식 막대(candle stick)의 통계량을 축소하여 활용하는 연구도 진행되었다 [14, 47]. 심층학습 기법을 활용한 연구로는 오토인코더를 사용하여 차원축소를 진행한 연구와 군집화를 통해 표현을 생성하여 강화학습 주식 매매 모델의 상태로 활용하는 연구가 존재한다 [3, 4]. 또 다른 연구로는 비지도 학습을 통해 시계열 데이터에 라벨을 부여하고, 라벨링을 기반으로 생성한 표현을 주식 매매에 활용하는 연구가 있다 [15]. 이를 통해 규칙기반 지표가 표현하지 못한 시장 흐름에 대한 정보를 강화학습 모델에 제공하여 좋은 성능을 보였다. 이 외에도 선물, 외환 거래, 암호화폐 등의 다양한 데이터에 적용해서 강화학습 매매 알고리즘의 성능을 평가해 본 연구가 존재한다 [25, 30].

제 3 장 대조학습을 이용한 강화학습 주식매매 기법

3.1 대조학습 기반 주가 시계열 표현 추출

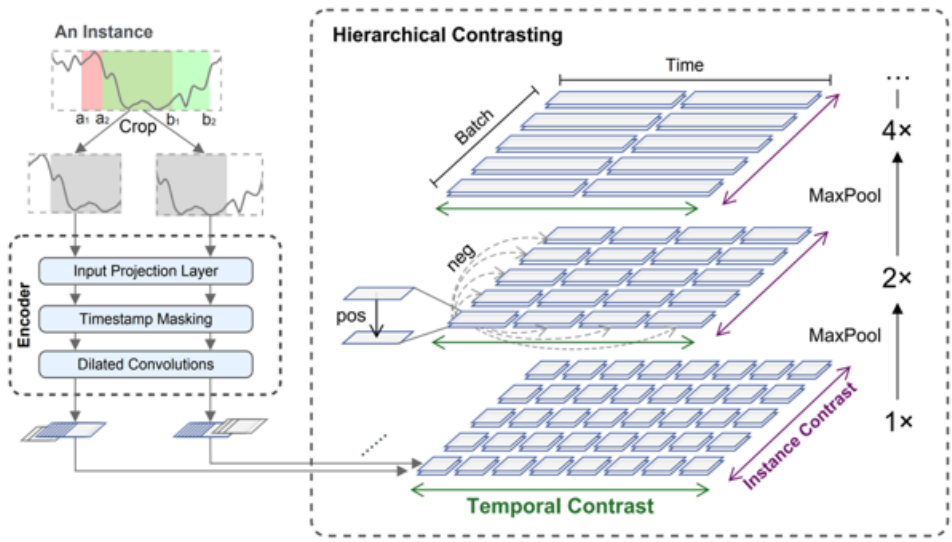

그림 3.1: Ts2Vec 방식의 대조학습[1]

본 논문은 2022 년에 발표된 시계열 대조학습 기법 논문 Ts2Vec 의[1] 대조학습 방식으로 주가 시계열 데이터의 표현을 추출하였다. Ts2vec 모델은 시계열의 중요한 특성인 크기를 바꾸지 않는 마스킹(masking)과 랜덤하게 자르는 방법으로 데이터를 증강한다. 시간이 겹치는 구간이 존재하도록 시계열을 2 개로 랜덤하게 잘라서 각 객체를 인코더를 통과시켜 표현을 추출한다. 이후 표현의 시간대가 겹치는

부분 내에서 서로 동일한 시간이면 양성 데이터 쌍, 다르면 음성 데이터 쌍으로 정의하여 동일 시간에 해당하는 부분의 유사도를 높이도록 학습해 표현을 생성한다.

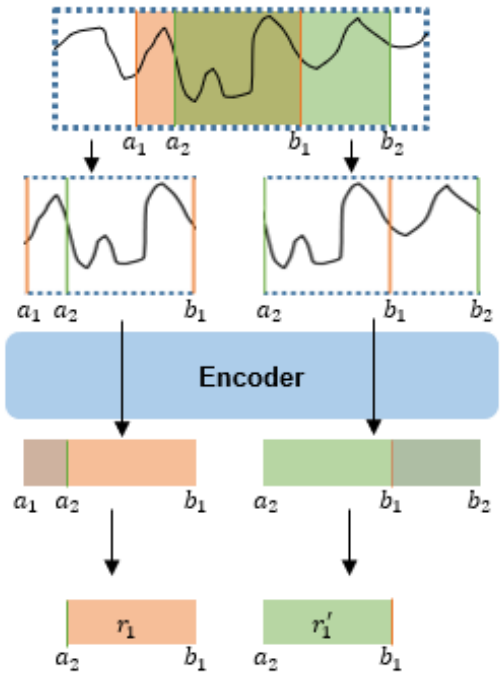

그림 3.2: Ts2Vec 방식의 표현 생성

Ts2Vec모델의 표현생성 방식은 다음과 같다. 하나의 시계열 데이터를 그림 3.2와 같이 겹치는 시간대가 생기도록 랜덤하게 자르는 부분을 선택해 $a_1 \sim b_1$ 의 객체와 $a_2 \sim b_2$ 의 객체를 생성한다. 두 개의 인스턴스를 인코더를 통과시켜 잘린 영역의 맥락을 반영한 표현을 생성한다. 이후 두개의 표현간 겹치는 시간대인 $a_2 \sim b_1$ 영역으로 표현을 각각 자른 후 r_1 과 r'_1 이라고 명명한다.

Ts2Vec모델의 대조학습은 시간적 손실 함수(temporal loss)와 객체 간 손실 함수

(instance loss)를 사용하여 진행된다. 그림 3.2에서 원본 시계열 데이터를 랜덤하게 자른 후 인코더를 통과시켜 생성한 표현 r_1과 r'_1은 같은 시점이더라도 각기 다른 주변 맥락을 반영하여 인코딩 되었기 때문에 서로 다른 값을 갖는다.

$$L_{temporal}^{(i,t)} = -log \frac{\exp\left(r_{i,t} \cdot r'_{i,t}\right)}{\sum_{t' \in \Omega} (\exp\left(r_{i,t} \cdot r'_{i,t'}\right) + \mathbb{I}_{\left[t \neq t\right]} \exp\left(r_{i,t} \cdot r_{i,t'}\right))} \qquad (3.1)$$

시간적 손실 함수는 r_1과 r'_1에서 동일 시점이면 양성 데이터 쌍, 다른 부분이면 음성 데이터 쌍으로 정의해 표현이 추출된 맥락이 다르더라도 동일 시간대이면 같은 값을 갖도록 학습을 진행한다. $L_{temporal}^{(i,t)}$은 수식 3.1과 같이 정의한다. t는 시계열 데이터의 타임스탬프(timestamp)이고, i는 타임스탬프의 인덱스이다. Ω는 두 객체 r_1과 r'_1의 겹치는 시간 구간을 나타낸다.

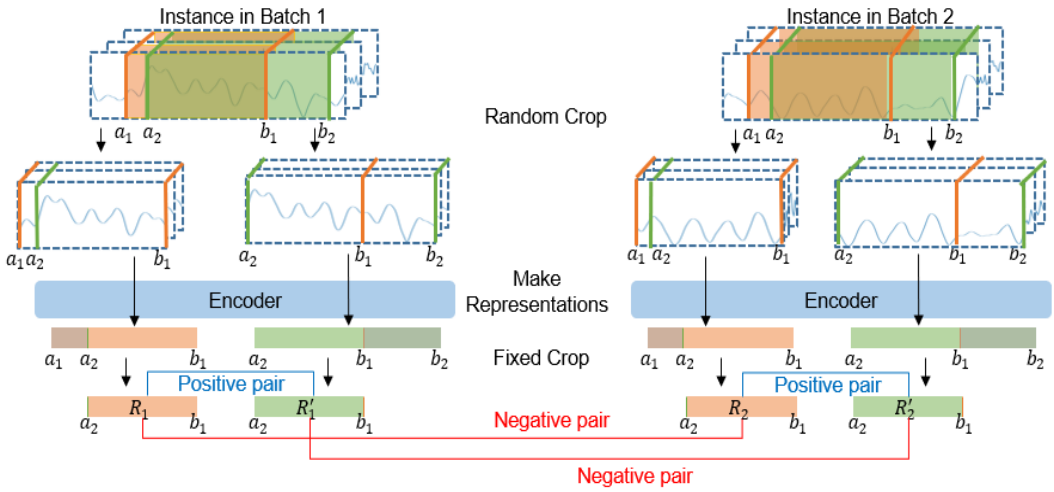

그림 3.3: Ts2Vec 방식의 배치 단위 대조학습

$$L_{instance}^{(i,t)} = -log \frac{\exp(r_{i,t} \cdot r'_{i,t})}{\sum_{j=1}^{B}(\exp(r_{i,t} \cdot r'_{j,t}) + \mathbb{1}_{[i \neq j]} \exp(r_{i,t} \cdot r_{j,t})} \qquad (3.2)$$

그림 3.3는 batch 내의 객체끼리 양성과 음성 데이터 쌍을 설정하는 과정을 묘사한다. 배치(batch)내의 서로 다른 시계열을 겹치는 시간대가 생기도록 랜덤하게 자르는 부분을 선택해 $a_1 \sim b_1$의 객체와 $a_2 \sim b_2$의 객체를 생성한다. 이후 배치별로 인코더를 통과시켜 각각 표현을 생성하고 겹치는 시간대인 $a_2 \sim b_1$영역으로 표현을 자른 후 R_1, R'_1과 R_2, R'_2를 생성한다. 동일 시간대의 같은 객체(R_1 경우 R'_1)는 양성 데이터 쌍, 다른 객체(R_1 경우 R_2, R'_2)는 음성 데이터 쌍으로 정의한 후 양성 데이터 쌍들의 유사도를 높이도록 학습한다. $L_{instance}^{(i,t)}$는 수식 3.2와 같이 정의한다. B는 학습 단위 배

치의 수, t는 시계열 데이터의 타임스탬프이고, i는 타임스탬프의 인덱스를 나타낸다.

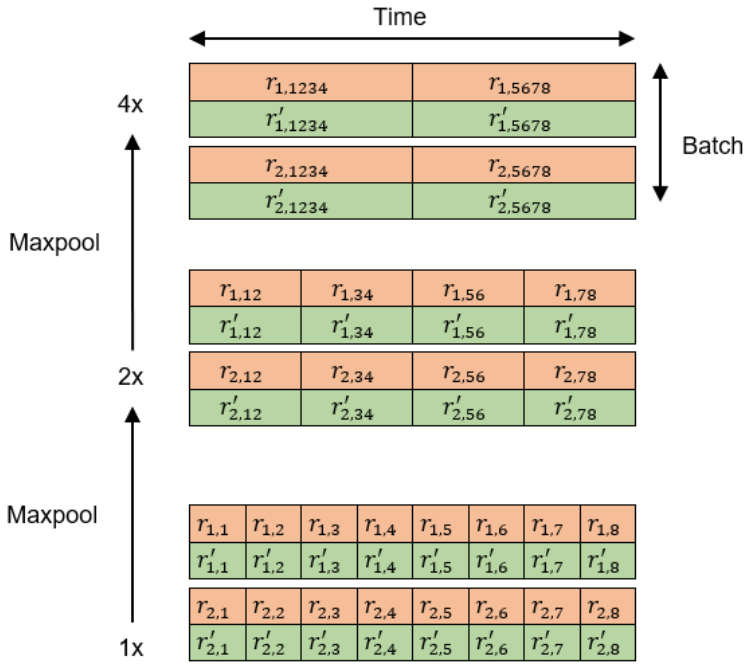

그림3.4: Ts2Vec방식의 계층적 손실함수

$$L_{total}^{(i,t)} = k * L_{temporal}^{(i,t)} + (1-k) * L_{instance}^{(i,t)} \tag{3.3}$$

최종 학습은 시간적 손실 함수와 객체간 손실 함수를 구하고 두 손실 함수를 특정 k비율로 결합하여 전체 손실 함수를 구한다. 이후 그림 3.4와 같이 표현에 최대 풀링(max pooling)을 적용하여 전체 손실 함수를 구하는 작업을 반복하고 이를 계층적(hierarchical)으로 반복하여 계층적 손실 함수(hierarchical loss) 값을 누적 업데이트

한다. 위 방법으로 구한 손실 함수 값을 최대 풀링을 진행한 총 횟수로 나눠준 값으로 최종 학습 손실 함수 값을 구한다.

3.2 강화학습 주식 매매 모델

3.2.1 배경 연구

본 논문에서 제안하는 강화학습 주식매매 모델은 2015년 Deep Mind에서 발표한 심층 Q 네트워크(DQN)모델 구조를 따른다 [27].

DQN은 Q 러닝의 Q함수를 인공신경망으로 대체하여 기존 연구에서는 불가능하던 입력 차원이 큰 복잡한 문제를 해결하였다. 이후 DQN을 활용한 연구들이 게임, 로봇 공학, 자율주행 등의 분야에서 뛰어난 성능을 보이며 다양한 연구가 진행되고 있다. 그림 3.5와 같이 Q 러닝은 에이전트가 환경과 상호작용하며, 이를 통해 수집한 경험을 바탕으로 Q함수를 업데이트하는 강화학습의 한 방식이다. 에이전트는 각 상태에서 Q함수 값이 가장 큰 행동을 선택함으로써 환경에서의 행동을 결정한다. 학습 초기에는 Q값이 잘 정의되지 않기 때문에, 에이전트는 일정한 확률로 무작위 행동을 선택하는 ε-탐욕(greedy) 전략을 사용하여 경험한 행동 중 Q함수값이 가장 큰 행동을 선택하거나 한 번도 경험해보지 않은 탐험적 행동을 선택하여 학습을 진행한다.

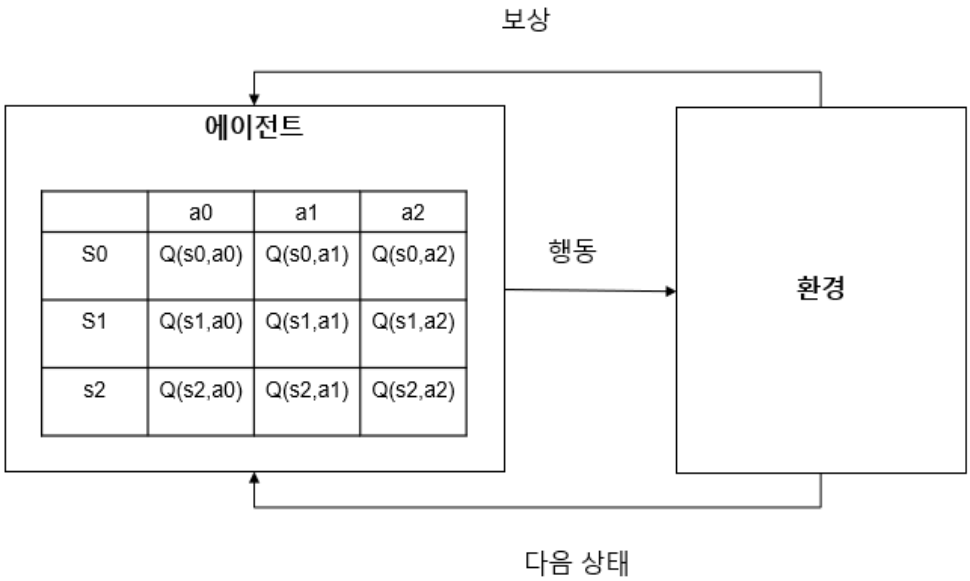

그림3.5: Q 러닝 개요

DQN 알고리즘은 Q 러닝의 Q함수를 인공 신경망을 사용하여 대체한다. 신경망이 이 Q함수를 학습함에 따라 에이전트가 어떤 상태에서 어떤 행동을 선택할지 결정할 수 있게 된다. DQN은 경험 반복(experience replay) 기법을 통해 에이전트의 경험을 저장하고 이를 무작위로 추출하여 학습에 사용해 학습의 안정성을 높였다. 또한 고정된 Q 타겟을 사용하여 학습 과정에서의 발산을 방지하였다.

3.2.2 마르코프 의사결정 과정(MDP) 정의

본 논문에서 제안하는 강화학습 주식매매 기법의 MDP는 다음과 같다.

3.2.2.1 상태(state)

강화학습 주식 매매 모델의 상태로 주가의 OHLCV(Open, High, Low, Close, Volume) 정보만을 사용했을 때 학습이 수렴하지 않아 OHLCV 정보 외에 주가에 대한 추가적인 정보를 줄 수 있는 기술적 지표와 직전 14일의 종가 가격 그리고 행동을 통한 포트폴리오 가치 개선율을 추가하여 49차원의 상태로 구성하였다. 상태의 모든 연속적인 값을 갖는 구성요소는 정규화를 진행하였다.

(a) OHLCV(Open, High, Low, Close, Volume): 당일 주식의 시가, 고가, 저가, 종가, 거래량 정보를 반영하여 알고리즘이 주식 시장의 현재 상황을 파악할 수 있다.

(b) 직전 14일의 일별 종가: 직전 2주 동안의 주식 종가 가격을 상태에 반영하여 알고리즘이 주식 시장의 과거 흐름을 파악할 수 있다.

(c) 직전 포트폴리오 가치 대비 현재 포트폴리오 가치: 직전 행동의 투자 성과를 반영하여 알고리즘이 다음 행동을 결정하는데 도움을 준다.

(d) 표현: 표현학습으로 추출한 주가 시계열 데이터의 특징을 상태에 반영한다. 본 논문에서는 대조학습 방식으로 표현을 추출한다.

(e) 기술적 지표(Technical Indicator, T.I): 주식의 가격과 거래량 정보(OHLCV)를 바탕으로 생성한 23개의 기술적 지표들을(RSI, CCI, SO, MFI, VO, EMA etc.) 강화학습의 상태에 반영해 알고리즘이 주식 시장의 흐름에 대한 추가적인 정보를 얻을 수 있다. 본 연구에서 사용한 기술적 지표의 의미와 수식을 아래에서 설명한다.

- RSI(Relative Strength Index)

  일정 기간 동안 주가의 전일 가격에 비해 상승한 변화량과 하락하 변화량의 평균값을 구한 후 상승한 변화량이 크면 과매수로, 하락한 변화량이 크면 과매도로 판단하여 주가 움직임의 속도와 변화 측정한다.

$$RSI = 100 - \{\frac{100}{1+\frac{Avg(Upward\ Price\ Change)}{Avg(Downward\ Price\ Change)}}\} \quad (3.4)$$

- CCI(Commodity Channel Index)

  현재 가격이 평균 주가와 얼마나 차이가 나는지 나타낸다.

$$CCI = \frac{Avg\left(\frac{High+Low+Close}{3}\right) - SMA\_of\_Avg\left(\frac{High+Low+Close}{3}\right)}{0.015 * Deviation} \quad (3.5)$$

- SO(Stochastic Oscillator)

  종가를 바탕으로 한 시장의 움직임을 측정하는 지표이다.

$$SO = \frac{\text{Avg}((\text{Recent Close}) - (\text{Lowest of Previous 14 Sessions}))}{((\text{Highest of Previous 14 Sessions}) - (\text{Lowest of Previous 14 Sessions})) * 100)} \quad (3.6)$$

- MFI(Money Flow Index)

  RSI 지표에 거래량 정보를 추가하여 거래량을 고려한 주가의 움직임 변화 측정한다.

$$MFI = 100 - \{100/(1 + \text{Sum}(\text{Positive Money Flow})/ \text{Sum}(\text{Negative Money Flow}))\} \quad (3.7)$$

$$Money\ Flow = (High + Low + Close)/3) * Volume$$

- VO(Volume Oscillator)

  일정 기간동안 단기 거래량 이동평균과 장기 거래량 이동평균을 내어 차이를 나타낸다. 거래량 변화를 통해 주가의 상승 또는 하락 추세를 판단한다.

$$VO = \{\frac{(\text{Shorter Period SMA of Volume}) - (\text{Longer Period SMA of Volume})}{\text{Longer Period SMA of Volume}}\} * 100 \quad (3.8)$$

- EMA(Exponential Moving Average)

  일정 기간 동안의 주가 평균인 이동평균의 한 유형으로 가장 최근 데이터에 더 큰 가중치를 부여하여 최근 가격 변동을 반영한 주가 추세 변화를 나타낸다.

$$EMA(t) = weight * current\_price + (1 - weight) * EMA(t-1) \quad (3.9)$$

3.2.2.2 행동(action)

행동은 기존 강화학습을 이용한 주식 매매 연구를[3, 4, 21] 참조하여 5개의 행동을 갖는 유형과 11개의 행동을 갖는 두 가지 유형으로 정의했다.

(a) 5개 행동 정의

- hold: 매수, 매도 모두 하지 않고 현재 상태를 유지한다.
- buy50: 현재 예산으로 살 수 있는 최대 주식 수의 50% 매수한다.
- buy100: 현재 예산으로 살 수 있는 최대 주식 수의 100%를 매수한다.
- sell50: 현재 보유한 주식의 50% 매도한다.
- sell100: 현재 보유한 주식의 100%를 매도한다.

(b) 11개 행동 정의

주식을 매수하고 매도하는 비율을 5개 20%단위로 세분화하여 hold를 포함하여 총 11개의 행동을 정의했다.

- hold: 매수, 매도 모두 하지 않고 현재 상태를 유지한다.
- buy20, buy40, buy 60, buy 80, buy 100: 현재 예산으로 살 수 있는 최대 주식 수의 20%, 40%, 60%, 80%, 100% 매수한다.
- sell20, sell40, sell60, sell80, sell100: 현재 보유한 주식의 20%, 40%, 60%, 80%, 100%를 매도한다.

3.2.2.3 보상(reward)

보상은 당일 주식 시장 개장 시점의 포트폴리오 가격 대비 행동 이후의 포트폴리오 가치 변동 정도를 수식 3.3과 같이 정의하여 강화학습 에이전트가 선택한 행동의 적절성을 평가한다.

$$Reward = \{ \frac{(\text{행동 이후의 포트폴리오 가치})}{(\text{당일 장 open시점의 포트폴리오 가치})} - 1 \} * 100 \qquad (3.10)$$

3.2.3 제안 기법

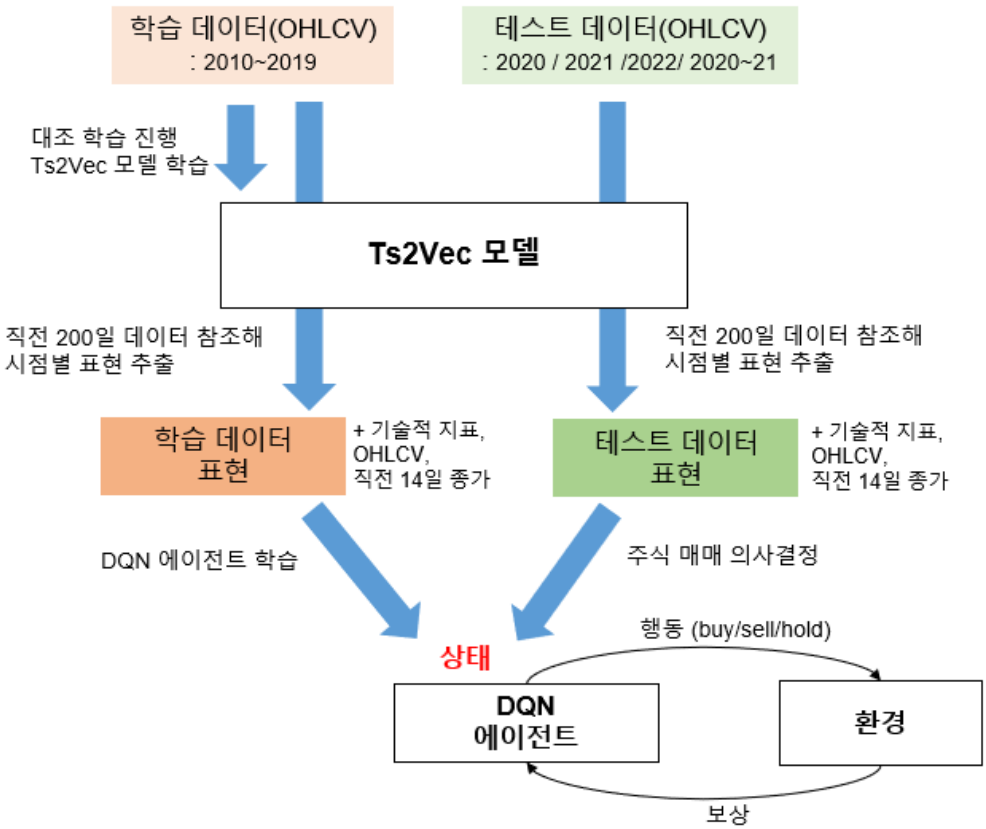

그림 3.6: 대조학습을 이용한 강화학습 주식 매매 모델 구조

그림 3.6은 본 논문이 제안하는 대조학습을 이용한 주식 매매 강화학습 모델의 구조를 설명한다. 제안 기법은 학습 데이터 기간의 주가 OHLCV 데이터를 입력 받아 3.1에서 설명한 대조학습 방법을 통해 Ts2Vec모델을 학습시킨다. 이후 학습된 Ts2Vec 모델에 학습 데이터와 테스트 데이터를 입력하여 당일과 직전 200일의

OHLCV 값을 참조한 각 시점의 표현을 생성한다.

학습된 Ts2Vec 모델이 생성한 학습 데이터의 표현을 주가의 기술적 지표와 OHLCV 정보, 포트폴리오 가치 개선 정도 정보와 함께 DQN 기반 주식 매매 모델의 상태로 사용하여 에이전트가 각 상태에서 최적의 행동을 선택하도록 DQN 모델을 학습한다. 이후 테스트 데이터의 표현을 주가의 기술적 지표와 OHLCV 정보, 포트폴리오 가치 개선 정도 그리고 직전 14일 종가와 함께 학습된 DQN 기반 주식 매매 모델의 상태로 사용하여 테스트 기간에 대해 주식 매매 의사결정을 내린다.

제 4 장 실험 결과

4.1 실험 데이터

실험에 사용한 주가 데이터는 NASDAQ의 시가총액 상위100위 안의 종목 중 상장된 지 20년 이상 된 각기 다른 산업분야의 종목을 10개를 선정했다. 선정된 종목은 아래 표 4.1과 같다.

2010.01.01~2021.12.31기간의 데이터 중 2010.01.01~ 2019.12.31까지의 데이터를 시계열 특징 추출 대조학습 모델과 주식 매매 강화학습 학습 데이터로 사용하고, 테스트 데이터는 2020.01.01~2020.12.31, 2021.01.01~2021.12.31, 2022.01.01~2022.12.31, 2020.01.01~2021.12.31 기간의 4가지 기간으로 구분하여 실험을 진행했다. 아래 그림 4.1과 4.2가 2010~2022년까지의 각 종목의 종가 그래프를 보여준다.

표 4.1: 평가 주식 종목

	기업명	상장 코드
1	Adobe	ADBE
2	Amazon	AXP
3	Advanced Micro Devices	AMD
4	American Express	AXP
5	Costco	COST
6	eBay	EBAY
7	Intel	INTC
8	McDonald's	MCD
9	Microsoft	MSFT
10	Pepsi	PEP

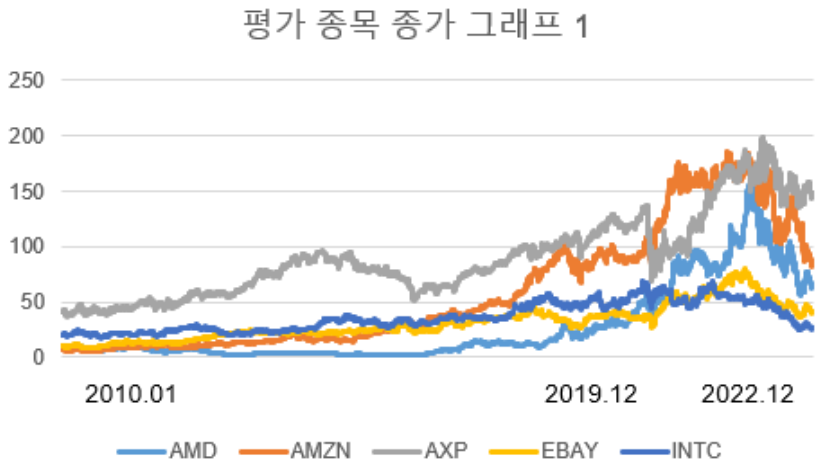

그림 4.1: 평가 종목 종가 그래프 1

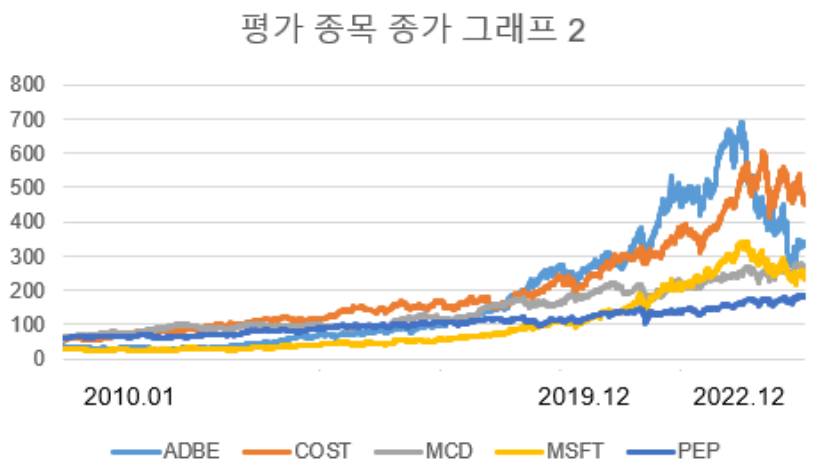

그림 4.2: 평가 종목 종가 그래프 2

4.2 사용 모델

대조학습으로 추출된 주가 표현이 주식 매매 강화학습 모델에서 어떤 효과를 갖는지 검증하기 위해 모델의 상태와 행동 종류를 다양하게 설정하고 모델의 특징을 파악하기 위해 학습 시 표현의 크기와 거래 수수료 유무를 다르게 하여 각 조건에서의 모델의 특징을 확인하는 실험을 진행하였다.

4.2.1 상태별 분류

(a) 기술적 지표 모델(T.I 모델)

표현을 사용하지 않고 23개의 기술적 지표와 포트폴리오 개선 정도, 직전 14일의 종가만을 강화학습의 상태로 사용한 모델을 'T.I모델'이라고 표기한다. 심층 학습을 통해 추출한 표현을 상태로 사용한 모델과 성능을 비교하여 표현의 효과를 파악할 수 있는 베이스 모델로 사용된다.

대조학습으로 추출된 주가 표현이 주식 매매 강화학습 모델에서 어떤 효과를 갖는지 검증하기 위해 모델의 상태와 행동 종류를 다양하게 설정하고 모델의 특징을 파악하기 위해 학습 시 표현의 크기와 거래 수수료 유무를 다르게 하여 각 조건에서의 모델의 특징을 확인하는 실험을 진행하였다.

(b) 대조학습 추출 표현+ 기술적 지표 모델(Ts2Vec+ T.I 모델)

Ts2Vec의[1] 대조학습 기법을 통해 추출한 주가 데이터의 표현을 T.I 모델(a)의 상태와 함께 사용한 모델을 'Ts2Vec+ T.I 모델'이라고 표기한다. 각 시점의 표현은 직전 200일의 주가 OHLCV 정보를 과거 데이터로 학습된 Ts2Vec 모델로 인코딩

시켜 시점 별 크기 160의 표현을 생성하였다.

4.2.2 행동별 분류

(a) 5개 행동 모델

매수, 매도 행동을 하지 않는 hold, 현재 예산으로 살 수 있는 최대 주식 수의 50%를 매수하는 buy50, 현재 예산으로 살 수 있는 최대 주식 수의 100%를 매수하는 buy100, 현재 보유한 주식수의 최대 50%를 매도하는 sell50, 현재 보유한 주식수의 최대 100%를 매도하는 sell100의 총 5가지 행동으로 주식 매매 의사결정을 진행하는 모델이다. 매매 의사결정 문제를 단순화하여 대조학습으로 추출한 표현의 효과를 파악한다.

(b) 11개 행동 모델

5개 행동 모델의 매수, 매도 비율을 20% 단위로 세분화하여 hold와 buy 20, 40, 60, 80, 100% 행동, 그리고 sell 20, 40, 60, 80, 100%의 총 11개 행동으로 확장했다. 행동의 수가 다양한 상황에서 대조학습으로 추출한 표현의 효과를 비교 검증할 수 있다.

4.2.3 거래 유형별 분류

강화학습 주식 매매 모델이 주식 매매를 진행할 때마다 적용되는 수수료를 다르게 설정하여 모델의 특징을 실험하였다. 주식을 매수 시에는 수수료가 발생하지 않고 매도 시에만 매도 금액의 0.1% 수수료를 적용하는 모델과 매매 수수료가 없는 모델로 각각 학습을 진행하여 수익률과 거래 특징을 분석하였다.

4.3 실험 조건

4.3.1 대조학습 실험 조건

주가 시계열 데이터의 표현을 추출하기위한 Ts2Vec 모델의 하이퍼파라미터(hyperparameter)를 다양한 조합으로 설정하여 실험을 진행하였다. 실험 결과 주식 매매 강화학습 모델의 최고 평균 수익률을 내는 최적 조합은 표 4.2와 같다. 이를 최적 조건으로 선정하여 이후 실험을 진행하였다.

(a) 학습 에폭(epoch): 대조학습 모델 훈련 반복 수 하이퍼파라미터로 {500, 1000, 3000}의 세가지 값으로 실행 진행하였다.

(b) 배치(batch) 데이터 수: 모델 훈련 시 하나의 배치에 들어가는 시계열의 크기로 {200, 300}의 두가지 값으로 실험 진행하였다.

(c) 인코더(encoder) 입력: Ts2Vec 모델의 인코더로 시점 t의 표현을 생성하기 위해 참조하는 과거 데이터의 수 파라미터로 {100, 200, 300} 값을 통해 다양한 길이의 과거 데이터를 참조하여 실험 진행하였다.

(d) 표현 크기: 인코더를 통과한 후 생성되는 최종 표현을 크기 160의 임베딩으로 생성하였다.

표 4.2: 대조학습 하이퍼파라미터 최적 조합

하이퍼파라미터	최적 값
에폭(epoch)	3000
배치 데이터 수	300
인코더 입력	200
표현 크기	80, 160

4.3.2 강화학습 실험 조건

강화학습 주식 매매 모델 하이퍼파라미터를 다음과 같이 세팅하여 실험을 진행하였다. 실험 결과 주식 매매 강화학습 모델의 최고 평균 수익률을 내는 최적 조합인 아래 표 4.3 조건으로 이후 실험을 진행하였다.

(a) Q 네트워크 크기: 다양한 은닉층으로 구성된 신경망 구조를 실험하였다. 사용된 구조는 뉴런 256개의 단일 네트워크와 3개의 층을 가진 [128, 256, 128] 신경망 그리고 [512, 1024, 512]의 더 복잡한 네트워크 구조를 통해 성능의 차이를 비교하였다.

(b) γ(gamma): 강화학습에서 미래 보상의 할인율을 결정하는 파라미터로 {0.3, 0.8, 0.9}의 세 가지 값을 실험하였다.

(c) 타겟 네트워크 업데이트: 타겟 네트워크의 업데이트 주기를 결정하는 파라미터로, {1, 10, 20, 100}의 네 가지 값으로 실험을 진행했다.

(d) 메모리: 경험 반복(experience replay)을 위한 메모리의 크기를 결정하는 파라미터로, {2000, 4000, 10000}의 세 가지 값을 실험하였다.

(e) 배치 크기: Q 네트워크 업데이트에 사용되는 샘플의 수를 결정하는 파라미터로, {256, 512}의 두 가지 값을 사용하였다.

(f) 에피소드: 강화학습 훈련의 반복 수를 의미하는 파라미터로, {1000, 2000, 3000}의 세 가지 값을 실험하였다.

(g) ε (epsilon): ε-탐욕 정책에서 탐색의 정도를 결정하는 파라미터로, {0.3, 0.5, 0.7, 0.9}의 네 가지 값을 사용하고, 에피소드의 진행 정도에 비례하여 감소시킨 값을 학습에 적용하였다.

다양한 하이퍼파라미터 조합을 통해 실험을 진행하였으며, 최적 파라미터 조합은 아래 표 4.3과 같다. 이를 최적 조건으로 선정하여 이후 실험을 진행하였다.

표 4.3: 강화학습 하이퍼파라미터 최적 조합

하이퍼파라미터	**최적 값**
Q 네트워크	[128, 256, 128]
ɣ(gamma)	0.9
타겟 네트워크 업데이트	10
메모리 크기	2000
배치 크기	512
에피소드	2000
ε (epsilon)	0.3

4.4 실험 결과

4.4.1 수익률 평가

2010.01.01~2019.12.31 기간의 데이터로 학습시킨 강화학습 주식 모델로 테스트 데이터에 대해 수익률 평가를 진행하였다. 테스트 데이터로 수익률 평가를 진행하기 전 학습 데이터로 훈련시킨 강화학습 모델의 보상이 에피소드가 진행됨에 따라 상승하며 일정한 값으로 수렴하는지를 점검했다.

강화학습 주식 매매 모델의 상태로 주가의 OHLCV 정보만을 사용했을 때 학습이 수렴하지 않아 OHLCV 정보와 함께 기술적 지표 23개와 포트폴리오 가치 개선 정도 그리고 직전 14일의 가격을 강화학습의 상태로 활용하여 학습의 수렴을 에피소드별 보상 그래프를 통해 확인하였다. 이후 대조학습을 통해 생성한 표현을 기술적 지표와 함께 상태로 사용한 모델에서도 학습이 수렴함을 확인하였다.

그림 4.3-4.12를 통해 평가를 진행한 모든 종목에서 대조학습으로 추출한 크기 160의 표현을 기술적 지표와 함께 상태로 사용한 모델(repr_160)이 미사용 모델(no_repr) 대비 더 높은 보상으로 학습이 수렴함을 확인할 수 있다. 그림의 가로축은 학습 에피소드 0~2000이고, 세로축은 에피소드별 누적 보상의 값을 나타낸다. 대조학습으로 추출한 표현이 강화학습 모델 학습 시 주가 흐름에 대한 추가적인 정보를 더 제공하여 효과적인 의사결정을 하도록 도움을 주었음을 알 수 있다.

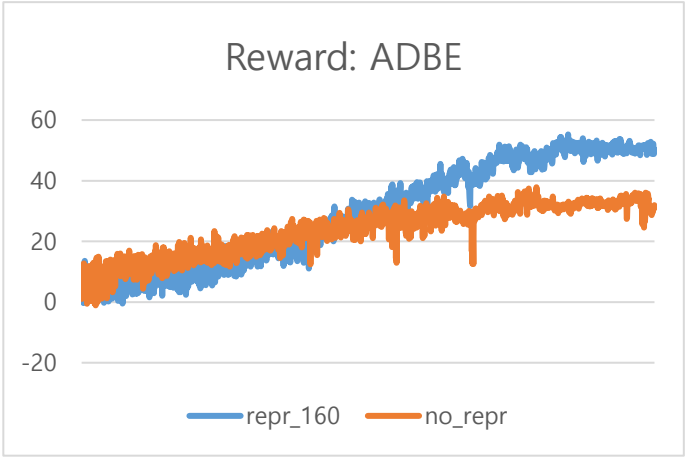

그림 4.3: 강화학습 모델 학습 그래프 ADBE

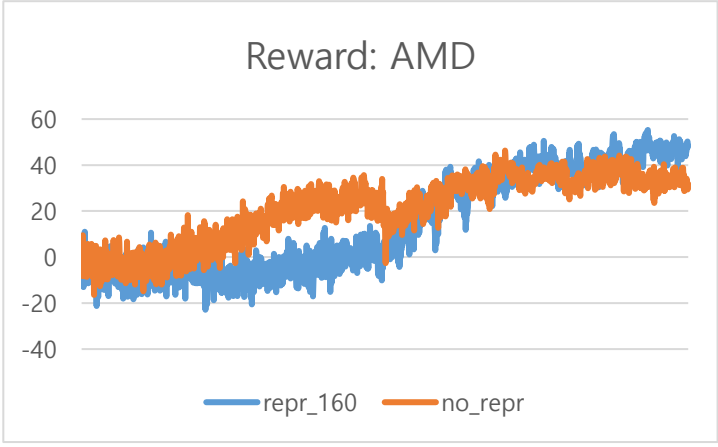

그림 4.4: 강화학습 모델 학습 그래프 AMD

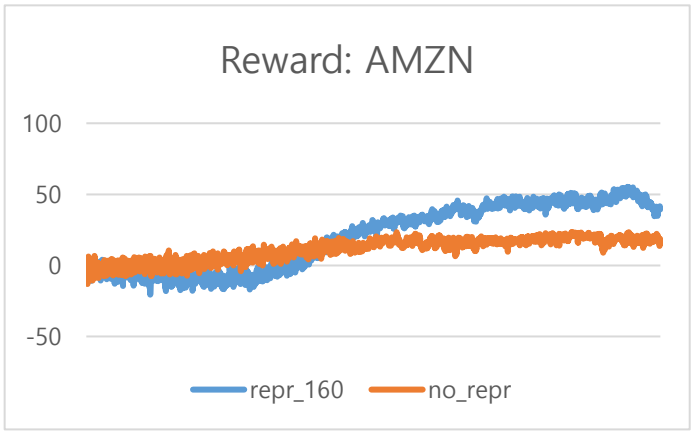

그림 4.5: 강화학습 모델 학습 그래프 AMZN

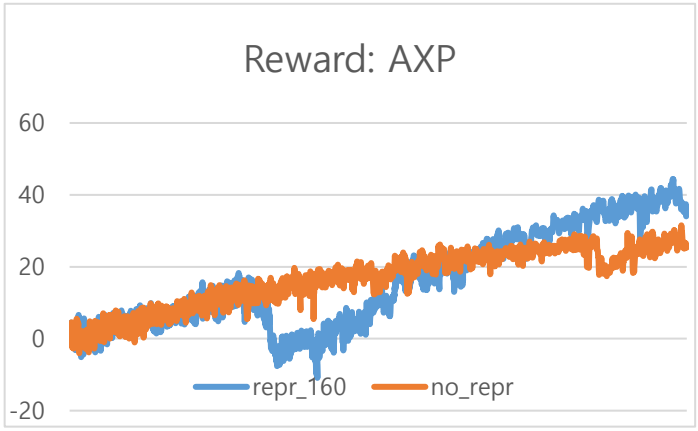

그림 4.6: 강화학습 모델 학습 그래프 AXP

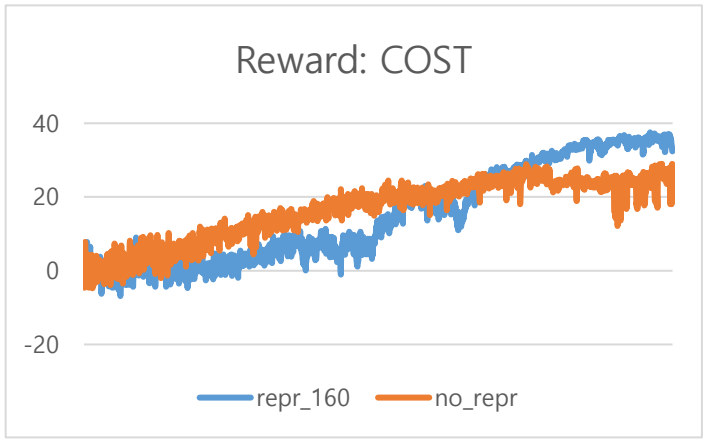

그림 4.7: 강화학습 모델 학습 그래프 COST

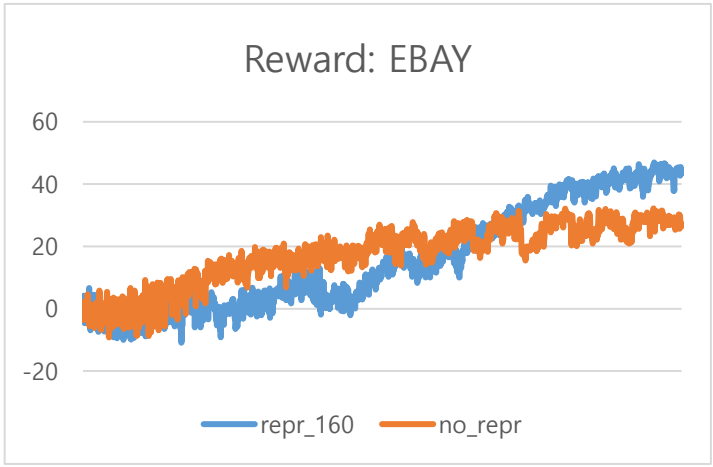

그림 4.8: 강화학습 모델 학습 그래프 EBAY

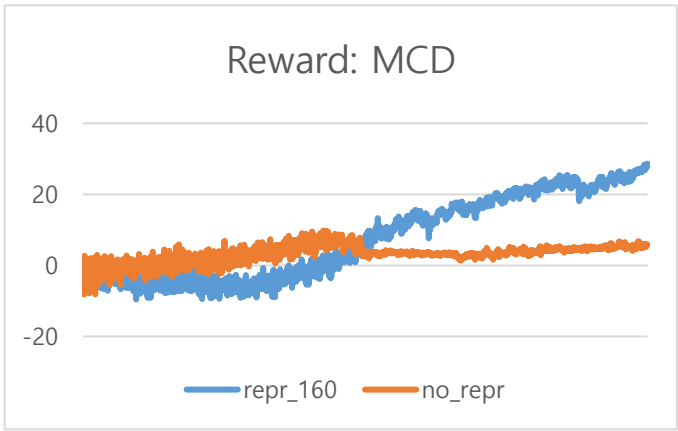

그림 4.9: 강화학습 모델 학습 그래프 MCD

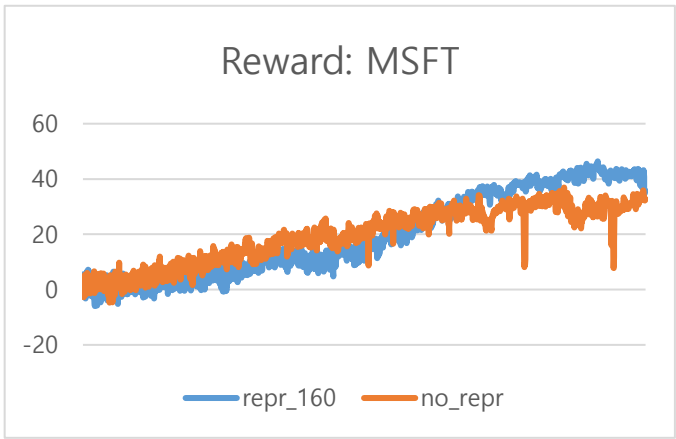

그림 4.10: 강화학습 모델 학습 그래프 MSFT

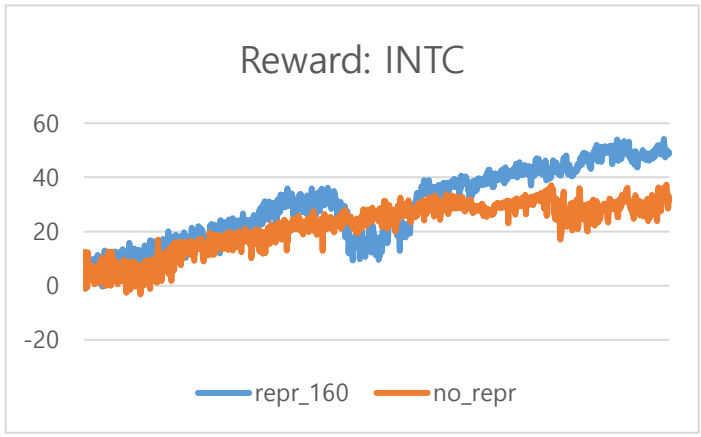

그림 4.11: 강화학습 모델 학습 그래프 INTC

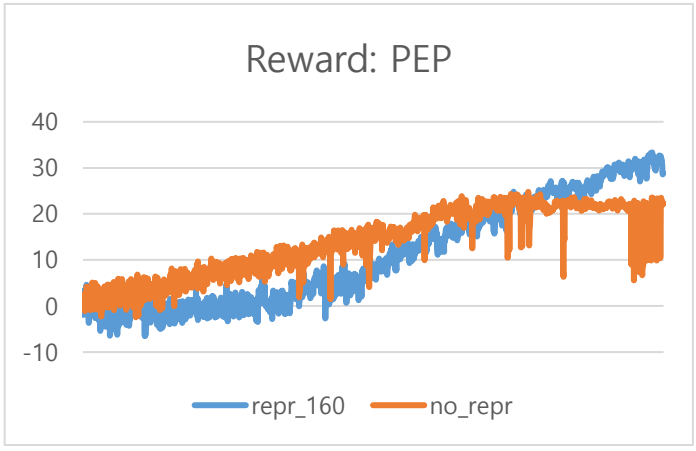

그림 4.12: 강화학습 모델 학습 그래프 PEP

2020, 2021, 2022 그리고 2020~2021년의 테스트 기간에 대해 총 10개의 NASDAQ 상장 주식을 학습 데이터(2010~2019)로 훈련한 강화학습 모델이 매매 의사결정을 진행하여 초기 투자금 대비 수익률을 구하였다. 이 때 주식 매도 거래 수수료를 0.1%로 설정하였다. 개별 종목의 수익률을 전체 종목수로 나눈 평균 수익률로 각기 다른 상태를 활용하는 강화학습 주식 매매 모델의 성능을 비교하였다.

T.I 모델은 강화학습 모델의 상태로 기술적 지표와 OHLCV 정보, 포트폴리오 개선 정도 그리고 직전 14일의 종가 데이터만을 사용한 모델로 대조학습 기법을 사용하여 추출한 표현을 상태로 사용한 모델과 성능을 비교하여 표현의 효과를 파악할 수 있는 베이스 성능으로 사용된다. Ts2Vec+ T.I 모델은 학습된 Ts2Vec 모델을 통해 추출한 표현을 T.I 모델의 상태에 추가하여 강화학습 모델의 상태로 활용한 모델이다.

5개 행동을 갖는 DQN 모델에서 제안기법을 실험한 결과는 아래 표 4.4 과 같다. 실험 결과 전체 대조학습을 통해 추출한 표현을 사용한 강화학습 모델이 기술적 지표만을 사용한 모델 대비 모든 테스트 기간에서 더 높은 수익률을 기록했다. 위 결과를 통해 5개 행동으로 정의한 강화학습 주식 매매 모델에서 대조학습을 통해 추출한 표현이 수익률을 향상시키는데 도움이 됨을 확인할 수 있다.

표 4.4: 5개 행동 모델 실험 결과

	Model	ADBE	AMD	AMZN	AXP	COST	EBAY	INTC	MCD	MSFT	PEP	Average(%)
2020	**T.I**	19.7	51.2	44.5	22.3	14.0	8.9	8.8	-9.6	-16.9	-21.2	12.2
	T.I+Ts2Vec	21.6	51.2	25.4	-7.1	14.0	22.1	80.5	-17.1	-16.9	-21.2	**15.2**
2021	**T.I**	4.3	74.1	0.6	12.4	22.3	8.9	14.1	-8.1	-1.4	-14.1	11.3
	T.I+Ts2Vec	14.1	74.1	0.6	26.9	22.3	10.9	-3.8	-8.1	-1.4	-14.1	**12.2**
2022	**T.I**	-18.7	-36.4	-41.5	24.3	-6.2	3.5	-49.5	-5.4	-10.1	1.5	-13.8
	T.I+Ts2Vec	-16.0	-28.6	-41.9	-3.2	21.6	-17.7	-49.5	7.8	-13.3	3.3	**-13.8**
2020 ~2021	**T.I**	29.0	157.8	38.1	78.2	36.8	27.9	18.5	2.8	15.2	-8.4	39.6
	T.I+Ts2Vec	31.7	157.8	54.4	29.7	18.9	28.7	66.9	6.9	75.9	30.8	**50.2**

11개의 행동을 갖는 DQN 모델에서도 5개 행동을 갖는 모델과 같이 대조학습을 통한 표현의 영향을 확인할 수 있다. 11개 행동 모델의 수익률을 평가한 실험결과는 아래 표 4.5와 같다. 2021년과 2022년에서 대조학습으로 추출한 표현을 상태로 사용한 강화학습 모델이 기술적 지표만을 상태로 사용한 모델 대비 높은 수익률을 기록하였다. 5개 행동 모델과 마찬가지로 대부분의 경우에서 대조학습으로 추출한 표현을 상태로 사용한 모델이 더 적은 거래 수를 갖는 경향이 존재했다. 위 결과를 통해 행동의 수가 더 적은 경우가 대조학습을 통해 추출한 표현을 사용한 강화학습 주식 매매 모델의 수익률을 향상시키는데 도움이 될 수 있음을 알 수 있다.

표 4.5: 11개 행동 모델 실험 결과

	Model	ADBE	AMD	AMZN	AXP	COST	EBAY	INTC	MCD	MSFT	PEP	Average(%)
2020	**T.I**	47.7	0	73.7	-5	0	41.3	-19.4	4.1	0	0	14.2
	T.I+Ts2Vec	30.8	0	7.5	-16.6	0	5.3	28.8	28.9	0	0	8.5
2021	**T.I**	17.4	0	0	36.5	0	19.6	1.4	0	0	0	7.5
	T.I+Ts2Vec	6.2	0	0	25.1	0	41.3	12.2	0	0	0	8.5
2022	**T.I**	-39.3	-55.3	-49.6	-13.5	-17.7	-36.9	0	-0.5	-26.9	5	-23.5
	T.I+Ts2Vec	17.9	-11.9	-23.2	22.1	9.5	-27.8	0	2.1	-13.5	-12	-3.7
2020 ~2021	**T.I**	68.4	0	76	30.6	87.1	85.9	-14.8	31.9	109.2	27.6	50.2
	T.I+Ts2Vec	39.7	0	24.4	34.3	44.3	19	31.6	33.4	40.3	6.4	27.3

4.4.2 거래 분석

강화학습 주식 매매 모델의 거래 특징을 알아보기 위해 테스트 기간 동안 각 행동별 출현 횟수를 분석해보았다. 거래 수수료를 0.1%로 학습한 모델의 80% 이상 경우에서 Ts2Vec+ T.I 모델의 보유(hold) 행동의 비율이 T.I 모델 대비 더 높음을 확인했다. 이를 통해 대조학습으로 추출한 표현이 강화학습 주식 매매 모델에게 추가적인 정보를 제공하여 각 상황에서 더 신중한 의사결정을 하도록 유도했음을 알 수 있다. Adobe 종목의 각 테스트 기간별 행동의 분포를 아래 표 4.6과 4.7에 나타내었다. 5개 행동 모델과 11개 행동 모델 모두에서 대조학습 표현 사용 모델이 미사용 모델 대비 보유 행동의 수가 더 많거나 같았다.

표 4.6 : 5개 행동 모델 ADBE 거래 분석

	Model	Hold	Buy50	Buy100	Sell50	Sell100
2020	T.I	60	92	5	95	0
	T.I+Ts2Vec	170	29	14	38	1
2021	T.I	142	54	1	55	0
	T.I+Ts2Vec	186	14	21	31	0
2022	T.I	71	89	2	90	0
	T.I+Ts2Vec	177	10	30	35	0
2020 ~2021	T.I	240	128	5	132	0
	T.I+Ts2Vec	368	23	47	67	0

평가 종목 중 대표로 Adobe 종목의 전체 테스트 기간에 대한 거래를 아래에 그림 4.13-4.20에서 소개한다. 2020년 테스트 기간동안 5개 행동 모델 경우 T.I 모델과 Ts2Vec+ T.I 모델의 거래를 주가 그래프 위에 표현하였다. 그래프의 가로축은 테스트 기간의 날짜이고 세로축은 주가를 나타낸다. 그래프의 색깔 점은 순서대로 빨간색은buy100, 오렌지색은 buy50, 파란색은 sell100, 초록색은 sell50 행동을 나타내며 색깔 점이 없는 모든 부분은 hold 행동을 취한날을 의미한다. 표현을 활용한 Ts2Vec+ T.I 모델에서 보유 비율이 더 높아서 주식 매매 거래가 더 적게 일어났음을 거래 그래프를 통해 시각적으로 확인할 수 있다. 이와 반대로 거래 수수료를 0%로 설정하여 학습한 모델에서는 보유 행동의 수가 T.I+ Ts2Vec 모델에서 더 적게 나타나는 경향을 보였다. 거래 수수료의 금액에 따라 모델의 거래 특징이 달라짐을 파악하였다.

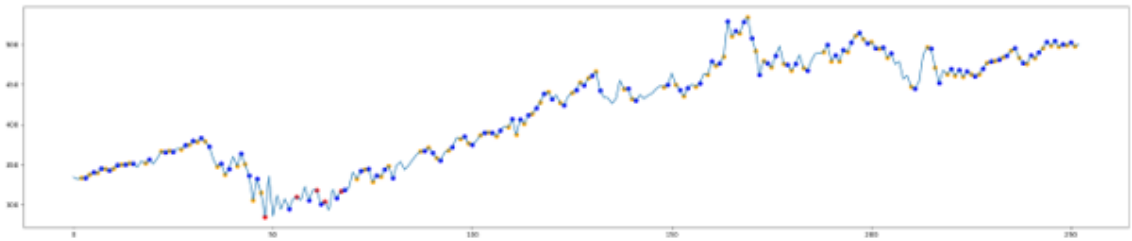

그림 4.13: T.I 모델 ADBE 2020년 거래

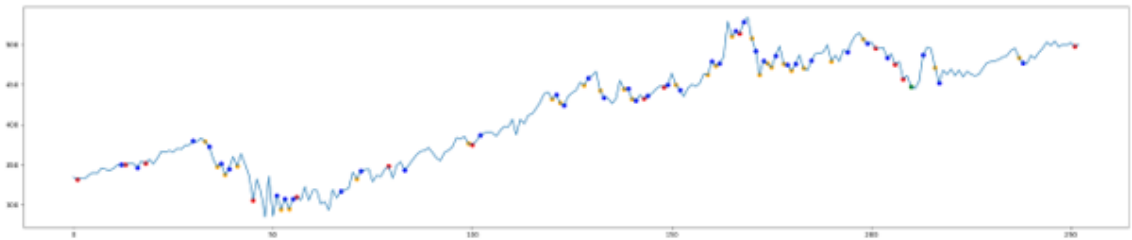

그림 4.14: T.I+Ts2Vec 모델 ADBE 2020년 거래

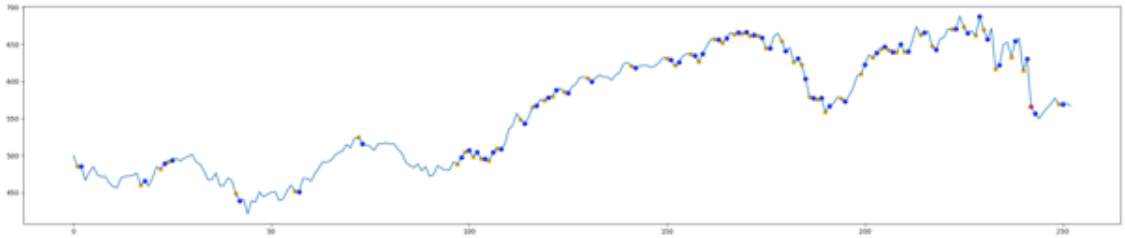

그림 4.15: T.I 모델 ADBE 2021년 거래

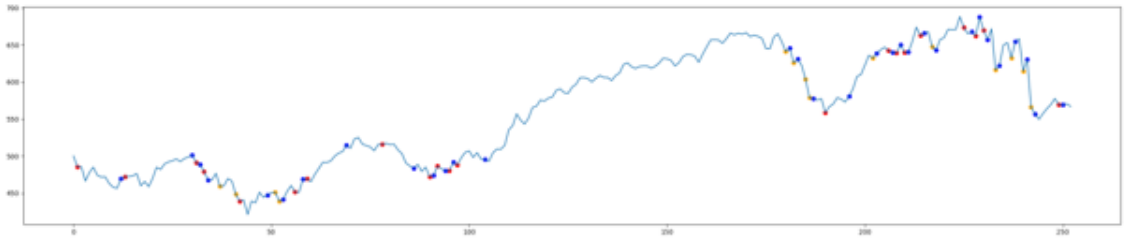

그림 4.16: T.I+Ts2Vec 모델 ADBE 2021년 거래

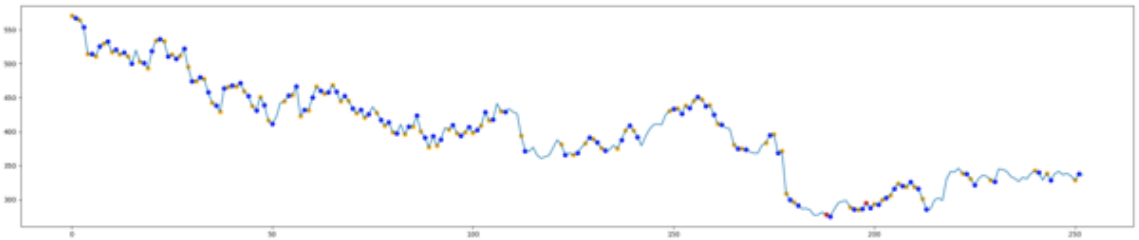

그림 4.17: T.I 모델 ADBE 2022년 거래

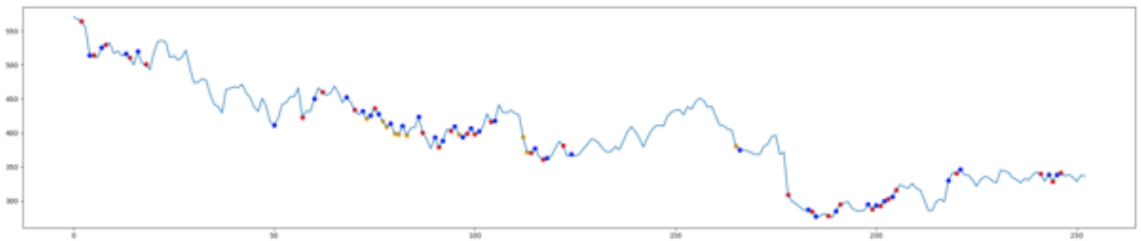

그림 4.18: T.I+Ts2Vec 모델 ADBE 2022년 거래

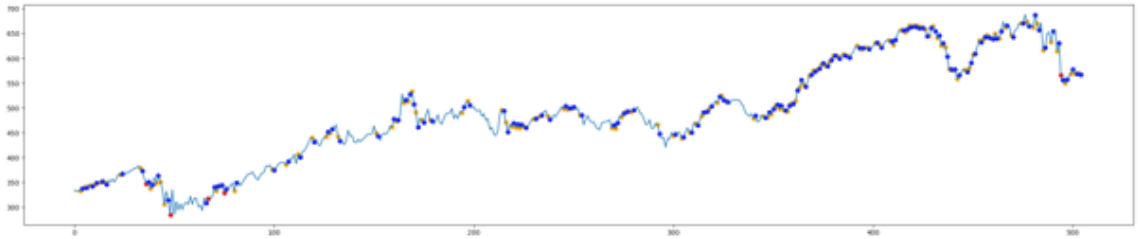

그림 4.19: T.I 모델 ADBE 2020-2021년 거래

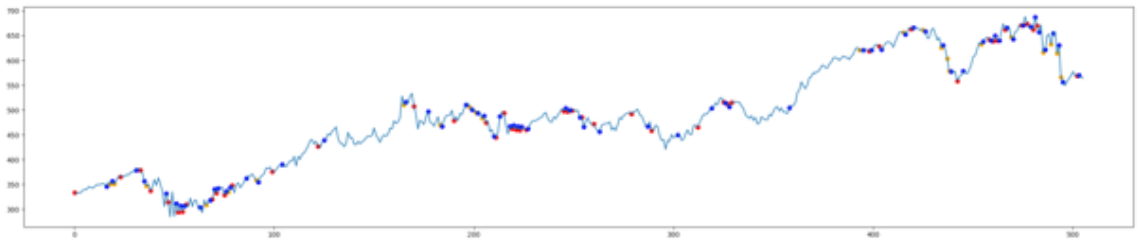

그림 4.20: T.I+Ts2Vec 모델 ADBE 2020-2021년 거래

4.4.3 타 매매법 비교

전체 평가 종목에 대해 연초에 주식을 매수 후 연말에 전량을 판매하는 매수 후 보유 전략(Buy&Hold)을 적용한 수익률을 평가하였다. 각 테스트 기간별 강화학습 주식 매매 모델의 수익률과 매수 후 보유 전략 수익률은 아래 표 4.7 과 같다. 2020년과 2021년은 대부분의 종목에서 주가가 상승한 주식 시장이었다. 이 기간의 매수 후 보유 전략의 전체 종목 평균 수익률은 2020년 30.8%, 2021년30.1%, 2020~2021년 70.2%로 동일 기간의 T.I+ Ts2Vec모델의 수익률인 15.2%, 12.1%, 50.2보다 높았다. 반면 하락장인 2022년에는 매수 후 보유 전략의 수익률은 -29.4%를 기록했고, T.I+ Ts2Vec 모델의 수익률은 -13.8%를 기록했다. 주가가 연평균 30% 이상 상승이 가능한 시장에서는 다수의 거래를 하는 강화학습 주식 매매 모델보다 매수 후 보유 전략이 더 큰 수익률을 낼 수 있지만, 하락장에서는 강화학습 주식 매매 모델이 상황에 따른 매매 의사결정을 진행해 매수 후 보유 전략보다 더 높은 수익률을 얻을 수 있다. 위 실험 결과를 바탕으로 주식 시장 상황에 따라 다른 전략을 선택하여 수익률을 높이는 주식 투자 기법을 시도해 볼 수 있다.

표 4.7: 매수 후 보유 전략 수익률 비교

	Model	ADBE	AMD	AMZN	AXP	COST	EBAY	INTC	MCD	MSFT	PEP	Average(%)
2020	T.I+Ts2Vec	21.5	51.2	25.3	-7.1	14.0	22.1	80.5	-17.1	-16.9	-21.2	15.2
	Buy&Hold	49.5	86.8	71.6	-3.9	29.3	38.4	-18.1	6.9	38.5	9.2	30.8
2021	T.I+Ts2Vec	14.1	74.1	0.6	26.9	22.2	10.9	-3.8	-8.1	-1.4	-14.1	12.1
	Buy&Hold	16.8	55.9	4.6	38.6	49.3	29.1	3.7	27.5	54.5	20.4	30.1
2022	T.I+Ts2Vec	-16.0	-28.6	-41.9	-3.2	21.6	-17.7	-49.5	7.8	-13.3	3.3	-13.8
	Buy&Hold	-40.4	-56.9	-50.7	-12.2	-19.4	-37.9	-50.3	-1.9	-28.4	4.4	-29.4
2020 ~2021	T.I+Ts2Vec	31.6	157.8	54.4	29.7	18.9	28.7	66.9	6.9	75.9	30.7	50.2
	Buy&Hold	69.6	193.1	75.7	30.0	94.8	83.2	-15.4	33.5	109.4	27.9	70.2

제 5 장 결론

5.1 결론

본 논문에서는 대조학습을 통해 추출한 주가 시계열 데이터의 표현을 강화학습 주식 매매 모델의 상태로 활용하여 수익률을 향상시킬 수 있도록 하는 방법을 제시하였다.

강화학습 모델의 상태를 기술적 지표, OHLCV, 포트폴리오 가치 증가율 그리고 직전 14일 종가로 사용한 모델과 대조학습을 통해 추출한 표현을 추가한 모델을 각각 NASDAQ 종목 10개의 2010~2019년 데이터로 학습하여 테스트 기간별 수익률을 평가하였다. 행동의 종류 수에 따라 표현이 효과가 달라지는지를 확인하기 위해 5개 행동 과 11개 행동 경우를 각각 평가하였고, 거래 수수료의 차이에 따른 모델의 수익률 차이 및 특성을 확인하였다.

대조학습으로 추출한 표현을 주식 매매 강화학습 모델의 상태로 학습한 결과 표현을 사용하지 않은 경우 대비 더 높은 보상으로 학습이 수렴하여 모델 성능을 개선할 수 있음을 확인하였다. 각 테스트 기간별 수익률 평가 결과 여러 상태를 사용한 강화학습 모델 중 대조학습으로 추출한 표현을 상태로 사용한 경우가 5개 행동 모델의 모든 테스트 기간에서 가장 높은 수익률을 기록했다. 실험을 통해 직전 200일의 주가 정보를 참조한 대조학습의 경우 크기 160의 표현이 가장 높은 수익률을 갖음을 확인하였다. 행동의 수를 11개로 확장한 모델 평가를 통해 행동의 수와 무관하게 제안 기

법이 동일한 효과를 갖음을 보였다. 테스트 기간내 행동의 발생 분포를 분석해본 결과 대조학습 표현을 사용한 모델에서 미사용 모델 대비 아무것도 하지 않는 hold 행동의 비율이 많아 매매 의사결정을 더 적게 내리는 경향성을 확인했다.

결론적으로 대조학습을 통해 추출한 주가 데이터의 표현을 통해 강화학습 주식 매매 모델이 시장 흐름에 대해 더 높은 이해가 가능하도록 학습을 하고, 더 신중한 거래를 하도록 유도하여 주식 매매 수익률을 향상시킬 수 있음을 실제 주가 데이터를 통해 확인하였다. 평가 결과 Ts2Vec의 대조학습 방식으로 추출한 크기 160의 표현 활용한 5개 행동의 모델이 가장 우수한 성능을 보였다. 연초에 주식을 매수 후 연말에 전량을 매도하는 매수 후 보유 전략과의 수익률 비교를 통해 주가의 상승과 하락 추세 예측에 따라 본 논문의 제안 기법과 매수 후 보유 전략 중 어떤 전략을 선택하는 게 유리한지에 대한 통찰을 얻을 수 있었다.

이러한 결과들은 강화학습을 활용한 주식 투자 전략 개발에 있어, 새로운 방향성을 제시하는 연구 결과라고 할 수 있다. 강화학습 상태를 개선하는 연구에서 본 논문에서 사용한 Ts2Vec기법 이외의 다양한 시계열의 대조학습 기법을 적용하여 주가 데이터의 특징을 추출해 수익률을 높이는 연구들을 추가로 진행해 볼 수 있다.

5.2 향후 연구

실제 주식 데이터 평가를 통해 대조학습으로 추출한 표현이 주식 매매 강화학습 모델의 수익률을 높이고 거래 빈도를 더 줄이는 효과를 확인하였다. 그러나 표현이 주가 데이터로부터 어떤 특징을 추출해서 이런 결과를 도출하는지에 대한 설명력은 부족한 상태이다.

차원 축소와 군집화 기법을 활용해 표현과 주가 데이터와의 유사도를 비교한 연구들을[16, 24] 참조하여 대조학습 및 타 표현학습 기법을 통해 생성한 표현과 주가 데이터와의 시점 별 유사도를 비교해보는 연구를 진행해볼 수 있다. 유사도 비교 분석을 통해 대조학습을 통해 추출한 표현이 기술적 지표 및 타 표현학습 기법 기반 표현보다 원본 주가 데이터와 유사하다면 대조학습의 표현이 강화학습 주식 매매 모델에 주가의 현재 상태에 대한 추가적인 유용한 정보를 제공하여 각 상황에서 최선의 의사결정을 내리도록 돕는다는 주장을 뒷받침할 수 있다.

다른 측면으로는 강화학습 알고리즘을 변경하여 본 논문의 제안기법을 실험해 제안기법의 성능과 실용성을 파악해 볼 수 있다. 또한 본 연구에서 사용한 DQN 알고리즘은 행동의 종류를 이산적으로만 설정 가능하여 행동의 수가 제한적이다. 정책 경사법 계열의 알고리즘을 사용하면 행동을 연속적으로 정의하여 강화학습 모델이 매 상황별로 구매하는 주식의 수를 연속적으로 설정할 수 있다. 행동이 연속적인 상황에서의 수익률과 거래 유형을 분석해보는 연구를 진행해 볼 수 있다.

마지막으로 본 연구에서는 데이터 셋을 NASDAQ 시가총액 100위 안의 종목 중 서로 다른 산업분야의 종목 10개로 선정했다. 향후 연구에서는 데이터를 주가의 움직

임에 따라 상승, 하강 또는 횡보 유형으로 분리해 학습 데이터와 테스트 데이터의 트렌드 유형에 따른 수익률 분석을 해볼 수 있다. 대조학습을 통해 추출한 표현을 활용한 강화학습 모델이 학습 데이터와 테스트 데이터의 트렌드 유형에 따라 수익률 경향성과 어떤 거래 특성을 갖는지 파악한다면 본 연구에서 제안한 기법을 더 효과적으로 활용할 수 있을 것으로 기대된다.

참고 문헌

[1] Y. Zhihan, Y. Wang, J. Duan, T. Yang, C. Huang, Y. Tong. “Ts2vec: Towards universal Representation of time series”, Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 36. No. 8. (2022)

[2] G. Woo, C. Liu, D. Sahoo, A. kumar, S. Hoi..“CoST: Contrastive Learning of Disentangled Seasonal-Trend Representations for Time Series Forecasting”, arXiv preprint arXiv:2202.01575 (2022)

[3] X. Liu, Z. Xiong, S. Zhong, H. Yang, A. Walid. “Practical Deep Reinforcement Learning Approach for Stock Trading”, arXiv (2018)

[4] DY Park, KH Lee. “Practical Algorithmic Trading Using State Representation Learning and Imitative Reinforcement Learning”, IEEE Access. Vol.9 (2021)

[5] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, L. Sun. "Transformers in time series: A survey." arXiv 2202.07125 (2022)

[6] D. Cheng, F. Yang, S. Xiang, J. Liu. "Financial time series forecasting with multi-modality graph neural network", Pattern Recognition 121 (2022)

[7] S. Uchida, B.K Iwana. “An empirical survey of data augmentation for time series classification with neural networks”, PLOS ONE (2021)

[8] K. Lei, B. Zhang, Y. Li, M. Yang, Y. Shen..“Time-driven feature-aware jointly deep reinforcement learning for financial signal Representation and algorithmic trading”, Elsevier Vol. 140 (2020)

[9] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, M. Long.“TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis”, arXiv:2210.02186 (2022)

[10] E. Eldele, M. Ragab, Z. Chen, M. Wu, C. Kwoh, X. Li, C. Guan. “Time-Series Representation Learning via Temporal and Contextual Contrasting”, arXiv:2106.14112 (2021)

[11] J.Y Franceschi, A. Dieuleveut, M. Jaggi. “Unsupervised Scalable Representation Learning for Multivariate Time Series”, NeurIPS (2019)

[12] Q. Lei, J. Yi, R. Vaculin, L. Wu, I. Dhilon. “Similarity Preserving Representation Learning for Time Series Clustering”, arXiv:1702.03584 (2021)

[13] Z. Dai, H. Zhu, J. Kang. “New technical indicators and stock returns predictability” Elsevier Vol. 140 (2021) pp.127-142

[14] R. Kusuma, T. Ho, WC. Kao, Y, Ou, K. Hua. “Using Deep Learning Neural Networks and Candlestick Chart Representation to Predict Stock Market ”, arXiv:19 03.12258(2019)

[15] N. Botteghi, M. Poel, C. Brune. “Unsupervised Representation Learning in Deep Reinforcement Learning”, arXiv:2208.14226 (2022)

[16] X. Yang, H. Yang, Q. Chen, R. Zhang, L. Yang, B. Xiao, C. Wang. “FinRL: A Deep Reinforcement Learning Library for Automated Stock Trading in Quantitative Finance”, arXiv:2011.09607 (2020)

[17] X. Wu, H. Chen, J. Wang, L. Troiano, V. Loia, H. Fujita. “Adaptive stock trading strategies with deep reinforcement learning methods”, Elsevier Vol. 538 (2020), pp.142-158

[18] C. Ma, J. Zhang, J. Liu, L. Ji, F. Gao.“A parallel multi-module deep reinforcement learning algorithm for stock trading”, Elsevier Vol. 449 (2021), pp.290-302

[19] F. Soleymani, E. Paquet. “Financial Portfolio optimization with online deep reinforcement learning and restricted stacked autoencoder—DeepBreath”,Elsevier Vol. 156 (2020)

[20] X. Yang, Z. Zhang, R. Cui. “TimeCLR: A self-supervised contrastive learning framework for univariate time series Representation”, Elsevier Vol. 245 (2022)

[21] J. Poppelbaum, G. Chadha, A. Schwung. “Contrastive learning based self-supervised time-series analysis”, Elsevier Vol. 117 (2022)

[22] S. Albahli, T. Nazir, A. Mehnmood, A. Irtaza, A. Alkhalifah, W. Albattah. “AEI-DNET: A Novel DenseNet Model with an Autoencoder for the Stock Market Predictions Using Stock Technical Indicators”, Electronics Vol.11 (2022)

[23] Z. Dai, X. Dong, J. Kang, L. Hong. “Forecasting stock market returns: New technical indicators and two-step economic constraint method”, Elsevier Vol.53 (2020)

[24] B.S.Bini, T. Mathew. “Clustering and Regression Techniques for Stock Prediction”, Elsevier Vol.24 (2016), pp.1248-1255

[25] W. Bao, X. Liu.“Multi-agent deep reinforcement learning for liquidation strategy analysis”, ICML (2019)

[26] H. Yang, X. Liu, S. Zhong, A. Walid. “Deep reinforcement learning for automated stock trading: an ensemble strategy”, ICAIF (2020)

[27] V. Minh, K. Kavukcuoglu, D. silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller. “Playing Atari with Deep Reinforcement Learning”, NIPS (2013)

[28] T. Chen, S. Kornblith, M. Norouzi, G. Hinton. “A Simple Framework for Contrastive Learning of Visual Representations”, ICML(2020)

[29] E. Kharitonov, M. Riviere, G. Synnaeve, L. Wolf, P. Mazare, M. Douze, E. Dupoux. “Data Augmenting Contrastive Learning of Speech Representations in the Time Domain”, IEEE (2021)

[30] Z. Jiang, J. Liang, “Cryptocurrency Portfolio management with deep reinforcement learning”, Intelligent Systems Conference (IntelliSys), IEEE (2017), pp. 905–913..

[31] S. Hochreiter and J. Schmidhuber, “Long short-term memory”, Neural computation, 9 (1997)

[32] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling”, arXiv preprint arXiv:1412.3555

[33] Lu Wang, Wei Zhang, Xiaofeng He, Hongyuan Zha, “Supervised reinforcement learning with recurrent neural Network for dynamic treatment recommendation,”International Conference on Knowledge Discovery & Data Mining (2018) pp2447-2456

[34] J. Y. Hong, S. H. Park, and J.-G. Baek “Ssdtw: Shape segment dynamic time warping” Expert Systems with Applications, 150 (2020)

[35] E. Jang, S. Gu, and B. Poole, Categorical reparameterization with gumbel softmax, arXiv preprint arXiv:1611.01144 (2016)

[36] B. Lim and S. Zohren, “Time-series forecasting with deep learning: a survey”, Philosophical Transactions of the Royal Society A, 379 (2021), pp. 20200209

[37] M. Liu, A. Zeng, Z. Xu, Q. Lai, and Q. Xu, "Time series is a special sequence: Forecasting with sample convolution and interaction", arXiv preprint arXiv:2106.09305 (2021)

[38] Z. Yue, Y. Wang, J. Duan, T. Yang, C. Huang, and B. Xu, "Learning timestamp-level Representations for time series with hierarchical contrastive loss", arXiv e-prints (2021), pp. arXiv–2106

[39] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, Fedformer: "Frequency enhanced decomposed transformer for long-term series forecasting" arXiv preprint arXiv:2201.12740 (2022)

[40] Y. Kim, W. Ahn, KJ. Oh, D. Enke, "An intelligent hybrid trading system for discovering trading rules for the futures market using rough sets and genetic algorithms", Applied Soft Computing, vol. 55 (2017), pp. 127-140

[41] F. Bertoluzzoa, M. Corazza. "Testing different reinforcement learning configurations for financial trading: introduction and applications", Procedia Economics and Finance, vol. 3 (2012), pp. 68-77

[42] R. Neuneier, "Enhancing Q-Learning for optimal asset allocation", Advances in Neural Information Processing Systems (1997)

[43] D. Silver, G. Lever, N. Hess, T. Degris, D. Wierstra, M. Riedmiller, "Deterministic policy gradient algorithms", International Conference on Machine Learning, vol. 32 (2014)

[44] Richard S. Sutton and Andrew G. Barto, "Reinforcement learning: an introduction", MIT Press. (1998)

[45] Y. Deng, F. Bao, Y. Kong, Z. Ren, Q. Dai. "Deep direct reinforcement learning for financial signal Representation and trading", IEEE Trans. Neural Netw. Learn. Syst., vol. 28, no. 3, Mar. 2017, pp. 653–664

[46] Y. Li, W. Zheng, and Z. Zheng, "Deep robust reinforcement learning for practical algorithmic trading", IEEE Access, vol. 7 (2019), pp. 108014–108022,.

[47] D. Fengqian, L. Chao, "An adaptive financial trading system using deep reinforcement learning with candlestick decomposing features", IEEE Access, vol. 8 (2020), pp. 63666–63678

[48] H. Van Hasselt, A. Guez, D. Silver, "Deep reinforcement learning with double Q-learning" ,in Proc. AAAI, vol. 30, no. 1 (2016), pp. 2094–2100.

[49] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver, "Rainbow: Combining improvements in deep reinforcement learning" ,in Proc. AAAI, vol. 32, no. 1 (2018), pp. 3215–3222.

[50] H. v. Hasselt, A. Guez, D. Silver, "Deep reinforcement learning with Double Q Learning, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence", AAAI Press (2016), pp. 2094–2100.

[51] Z. Li, S. Xue, W. Lin, M. Tong. "Training a robust reinforcement learning controller for the uncertain system based on policy gradient method", Neurocomputing 316 (2018), pp.313–321.

[52] E. Chong, C. Han, F. C. Park, "Deep learning networks for stock market analysis and prediction: Methodology, data Representations,and case studies" ,Elsevier Vol.83 (2017), pp.187-205

[53] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, et al., "Human-level control through deep reinforcement learning", Nature 518 (2015), pp.529–533

Abstract

Reinforcement Learning for Stock Trading based on Contrastive Representation Learning of Market States

Ji Moon, Jeong
Department of Industrial Engineering
The Graduate School
Seoul National University

Stock data are highly irregular, making it difficult to accurately characterize each situation. This challenge has long spurred research on effective stock trading strategies. Recently, due to the advancements in deep learning, research on stock trading using reinforcement learning has been actively carried out. Since reinforcement learning selects actions based on observed states, choosing a state that accurately represents current stock price information is crucial.

This paper proposes a method to extract the representation of stock price data using contrastive learning, which has recently shown excellent performance in time-series feature extraction, and to use it as the state in reinforcement learning stock trading models. Experimental results using the representation of stock prices extracted through contrastive learning demonstrated higher profitability than previous studies using rule-based indicators. Furthermore, results found that the reinforcement learning model conducts less trading when using the representation extracted by contrastive learning

as the state. These results suggest that the representation of stock price data extracted through contrastive learning could play a crucial role in enhancing the performance of reinforcement learning-based stock trading models.

Keywords: Reinforcement learning, Stock trading, Time series, Representation learning, Contrastive learning, Deep learning

Student Number: 2021-23086