

순열 대칭성을 활용한 비디오 프레임 순서 복원

Team SKKU_SME

GitHub Repository: <https://github.com/rebooot1/snu2026-skku-sme>

1. 과제 개요 및 접근 동기

이번 과제의 목표는 순서가 섞인 4장의 비디오 프레임과 하나의 캡션을 입력받아, 실제 시간 순서에 해당하는 순열을 찾는 것이다. 가능한 순열은 총 24개이며, 모델은 그중 하나를 정답으로 출력한다.

처음에는 캡션을 이해하고 사건의 전후 관계를 추론하는 능력이 가장 중요하다고 생각했다. 하지만 오답을 분석한 결과, 실제로는 시각적으로 비슷한 인접 프레임의 순서를 뒤바꾸는 경우가 많았다. 특히 사람이나 물체의 위치가 조금만 변하거나 움직임이 거의 없는 장면에서 예측이 불안정했다. 또한 동일한 네 장의 프레임이라도 입력 순서가 달라지면 모델의 답이 바뀌는 현상이 나타났다. 이는 모델이 시간 관계뿐 아니라 입력 위치나 프레임 번호에도 영향을 받고 있다는 뜻이다.

따라서 이 문제를 단순한 24개 클래스 분류가 아니라, **입력 순서가 달라져도 동일한 시간 관계를 복원해야 하는 순열 대칭성 문제**로 보았다. 이를 위해 학습 단계에서는 프레임 순서를 무작위로 바꾸는 순열 증강을 적용했고, 추론 단계에서는 여러 순열로 예측한 결과를 합쳐 최종 순서를 결정했다. 최종 시스템은 공개 평가에서 **0.94415**, 비공개 평가에서 **0.91869**의 EM 점수를 기록했다.

2. 전체 방법론 및 모델 구조

2.1 기본 모델과 LoRA 파인튜닝

기본 모델로는 **Qwen3.5-9B 비전-언어 모델**을 사용했다. 제한된 GPU 환경에서 학습하기 위해 전체 파라미터를 조정하는 대신 LoRA를 적용했다. LoRA는 attention의 q, k, v, o projection과 MLP의 gate, up, down projection에 적용했다. 주요 설정은 rank 16, alpha 32, dropout 0.05이다. 작은 움직임과 물체 위치 변화를 최대한 보존하기 위해 입력 해상도는 max_pixels=301056으로 설정했다.

처음에는 7B 모델로 학습과 추론 파이프라인을 확인한 뒤 9B 모델로 확장했다. 이후 PETO 기반 순열 증강 학습을 진행하고, 낮은 학습률로 한 번 더 추가 학습해 최종 백본 모델인 augA를 만들었다.

2.2 PETO 기반 순열 증강

PETO 학습에서는 프레임 4장의 입력 순서를 무작위로 변경하고, 변경된 입력 순서에 맞게

정답도 다시 계산했다. 예를 들어 원래 시간 순서가 [1, 2, 3, 4]인 샘플을 [3, 1, 4, 2]로 입력했다면, 모델이 출력해야 할 정답 역시 새롭게 부여된 프레임 번호를 기준으로 변환했다. 이 과정을 통해 모델이 특정 프레임 번호나 입력 위치를 외우는 것을 줄이고, 프레임 사이의 실제 시간 관계에 집중하도록 했다. 단순히 데이터 수를 늘리는 것이 아니라, 입력 순서가 바뀌어도 정답 관계는 유지된다는 문제의 구조를 학습에 반영한 것이 핵심이다.

2.3 K8 다중 뷰 합의 추론

추론 단계에서는 하나의 샘플을 서로 다른 8개의 프레임 순열로 입력했다. 각 순열에서 얻은 예측은 다시 원래 프레임 번호 기준으로 변환하고, 4×4 크기의 프레임-순위 투표 행렬에 누적했다. 행은 각 프레임을, 열은 해당 프레임이 배치될 시간 순위를 의미한다. 8번의 예측이 끝난 뒤에는 헝가리안 알고리즘을 사용해 전체 투표 점수가 가장 커지는 순열을 선택했다. 이 방식은 단순히 가장 많이 등장한 전체 순열을 고르는 것보다 프레임별 순위 정보를 함께 활용할 수 있다. 한 번의 입력에서 발생한 오류가 다른 입력 순서에서는 수정될 수 있기 때문에 단일 추론보다 결과가 안정적이었다.

2.4 검색 기반 후보 제약

최종 제출 시스템에는 학습 데이터를 이용한 검색 기반 후보 제약을 추가했다. 검증 데이터 분석 결과 다수의 데이터가 훈련 데이터와 동일하거나 유사한 데이터가 포함되어 있었다. 이에 테스트 추론 시점에도 학습 프레임을 비교해 동일한 이미지가 발견되면, 해당 학습 샘플에서 만들 수 있는 순열만 후보로 남겼다. 이후 제한된 후보 가운데 K8 투표 점수가 가장 높은 순열을 최종 결과로 선택했다. 검색 결과를 바로 정답으로 사용하는 것이 아니라, 후보 범위를 줄이는 보조 수단으로 활용했다.

3. 학습 및 추론 설정

공식 학습 데이터 9,535개 중 954개를 검증용으로 분리하고, 나머지 8,581개를 학습에 사용했다.

항목	설정
----	----

기본 모델	Qwen3.5-9B
-------	------------

파인튜닝	LoRA
------	------

LoRA 설정	r=16, alpha=32, dropout=0.05
---------	------------------------------

입력 해상도	max_pixels=301056
--------	-------------------

학습 순서	PETO v1 → augA 추가 학습
-------	----------------------

항목 설정

추론 방식 bf16, greedy decoding

결과 집계 K8 TTA + 헝가리안 알고리즘

평가 지표 Exact Match

학습은 7B 기본 모델에서 시작해 고해상도 입력과 TTA 효과를 확인한 뒤, 9B 모델로 확장하는 순서로 진행했다. PETO 학습 후에는 낮은 학습률로 추가 학습해 기존 성능을 유지하면서 순열 변화에 대한 안정성을 높였다. 최종 시스템은 RTX 3090 24GB 환경에서 재현했다. 총 819개 샘플을 추론하는 데 약 3.19시간이 걸렸으며, 최대 GPU 메모리 사용량은 약 19.18GB였다.

4. 실험 결과 및 분석

4.1 단계별 성능 변화

단계	주요 변경 사항	Public EM
7B SFT	기본 지도 파인튜닝	0.663
+ TTA	다중 순열 뷰 추론	0.785
+ 고해상도 입력 이미지 해상도 확대		0.808
9B 백본	모델 규모 확장	0.913
PETO	순열 증강 학습	0.92670
augA	낮은 학습률로 추가 학습	0.93542
최종 시스템	augA + 검색 기반 후보 제약	0.94415

비공개 평가 점수는 **0.91869**였다. 기본 7B SFT 모델에 TTA 추론을 적용하자 점수가 0.663에서 0.785로 상승했다. 모델을 다시 학습하지 않고 입력 순서만 다양하게 구성했음에도 0.122가 향상됐다. 이는 단일 입력 순서에 따라 모델의 예측이 크게 달라질 수 있다는 점을 보여준다.

입력 해상도를 높인 뒤에는 0.808을 기록했다. 이후 9B 모델로 확장했을 때 0.913까지 상승해 가장 큰 단일 향상을 보였다. PETO(Permutation-Equivariant Temporal Ordering) 순열 증강과 추가 학습을 적용하면서 각각 0.92670, 0.93542를 기록했고, 검색 기반 후보 제약을 추가한 최종 시스템은 0.94415를 달성했다. 결과적으로 모델 규모 확장뿐 아니라 순열 대칭성

을 학습과 추론에 함께 반영한 것이 성능 향상에 효과적이었다.

4.2 집계 방식 Ablation

K8 결과를 합치는 방법으로 hard vote, 소프트 확률 융합, Sinkhorn 기반 정렬, 신뢰도 가중 투표 등을 비교했다. 하지만 실험에서는 hard-vote 기반 합의를 안정적으로 넘어서는 방법을 찾지 못했다. 8개의 입력 뷰는 서로 다른 시각 정보를 제공하는 것이 아니라 같은 프레임을 다른 순서로 보여 준다. 따라서 모델이 작은 시각적 차이를 처음부터 인식하지 못하면 집계 방식만 복잡하게 바뀌도 오류를 해결하기 어렵다. 이 결과를 통해 후처리를 정교하게 만드는 것보다 변화 영역 확대, OCR, 프레임 차이 이미지처럼 새로운 시각 정보를 제공하는 방향이 더 중요하다고 판단했다.

4.3 Error Analysis

모델은 움직임이 작거나 장면 변화가 거의 없는 인접 프레임에서 자주 오류를 냈다. 특히 손의 위치, 물체의 이동, 화면 속 작은 글자처럼 일부 영역만 변하는 장면에서 순서를 뒤바꾸는 경우가 많았다. 또한 8개의 순열 뷰에서 같은 예측이 몇 번 나왔는지를 통해 모델의 불확실성을 확인할 수 있었다. 동일한 예측이 두 번만 나타난 샘플의 정확도는 약 0.19였지만, 8개의 뷰가 모두 같은 결과를 예측한 샘플은 약 0.94의 정확도를 보였다. 따라서 다중 뷰 일치 횟수는 최종 결과를 결정하는 것뿐 아니라, 모델이 불확실해하는 샘플을 선별하는 기준으로 사용할 수 있다.

4.4 공개·비공개 평가 차이

최종 시스템은 공개 평가에서 0.94415, 비공개 평가에서 0.91869를 기록해 약 0.025의 차이가 발생했다. 공개 평가 점수가 높은 모델이 새로운 데이터에서도 항상 안정적인 것은 아니었다. 따라서 모델을 선택할 때는 단일 리더보드 점수뿐 아니라 여러 검증 데이터에서의 성능과 다중 뷰 예측 일관성을 함께 확인할 필요가 있다.

4.5 시도했으나 채택하지 않은 방법

성능 향상을 위해 최종 방법 외에도 여러 아이디어를 실험했지만, 대부분은 검증 성능이 개선되지 않아 최종적으로 채택되지 않았다. 먼저 프레임을 두 장씩 비교하여 선후 관계를 예측한 뒤 전체 순서를 복원하는 pairwise 접근도 실험했으나, 전역적인 시간 흐름을 충분히 반영하지 못해 K8 추론보다 낮은 성능을 보였다. 이 밖에도 OCR 기반 텍스트 정보를 입력에 추가하거나, 프레임의 시각적 특징을 먼저 설명한 뒤 순서를 예측하는 2단계 프롬프트, Soft Voting이나 Sinkhorn 기반의 다양한 투표 집계 방식도 비교했지만, 기존 hard-vote 기반 K8 합의를 일관되게 넘어서는 결과는 얻지 못했다. 이러한 실험을 통해 복잡한 후처리나 추가 모듈을 도입하는 것보다, 순열 대칭성을 학습과 추론 과정에 일관되게 반영하는 현재의 접근이 가장 효과적이라는 결론을 얻었다.

5. 장점과 한계

제안한 방법의 가장 큰 장점은 입력 순서가 바뀌어도 실제 시간 관계는 유지된다는 문제의 구조를 학습과 추론에 직접 반영했다는 점이다. PETO 순열 증강은 모델이 입력 위치나 프레임 번호에 의존하는 것을 줄였고, K8 합의 추론은 한 번의 입력에서 발생하는 예측 변동을 완화했다. 별도의 추가 데이터 없이도 같은 샘플의 순서를 바꿔 학습과 추론에 활용할 수 있다는 점도 장점이다. 또한 단일 9B 모델을 기반으로 구성해 RTX 3090 24GB 환경에서도 실행할 수 있었다. 다중 뷰 일치 횟수를 통해 모델의 불확실성을 확인할 수 있다는 점도 실험 분석에 도움이 됐다.

반면 시각적으로 거의 동일한 인접 프레임을 구분하는 문제는 완전히 해결되지 않았다. 모델이 작은 변화를 인식하지 못하면 여러 순열로 추론하더라도 같은 오류가 반복될 수 있다. K8 추론은 하나의 샘플을 8번 처리하기 때문에 계산량도 증가한다. 향후에는 먼저 적은 수의 순열로 예측한 뒤, 결과가 불안정한 샘플에만 추가 추론을 수행하는 방식이 필요하다. 검색 기반 후보 제약 역시 동일 프레임이 존재할 때만 사용할 수 있으므로 일반적인 데이터에 그대로 적용하기 어렵다.

6. 결론 및 향후 과제

이번 과제에서는 비디오 프레임 순서 복원 문제를 순열 대칭성의 관점에서 접근했다. 학습 단계에서는 PETO 기반 순열 증강을 적용했고, 추론 단계에서는 8개의 순열 뷰와 헵가리안 알고리즘을 이용해 최종 순서를 결정했다. 그 결과 기본 7B SFT 모델의 공개 평가 점수 0.663을 최종 0.94415까지 향상시켰으며, 비공개 평가에서는 0.91869를 기록했다. 실험을 통해 단순히 더 큰 모델이나 복잡한 집계 기법을 사용하는 것보다, 문제에서 유지되는 성질을 학습과 추론 과정에 반영하는 것이 중요하다는 점을 확인했다.

다만 작은 시각적 변화를 구분하는 문제는 여전히 남아 있다. 향후에는 프레임 간 차이 영역을 자동으로 탐지해 확대하거나, OCR 정보를 추가해 시간 순서 판단에 필요한 단서를 더 제공하여 고도화할 수 있다. 또한 예측 일관성이 낮은 샘플에만 추가 추론을 적용해 정확도와 추론 비용 사이의 균형을 개선할 수 있을 것으로 생각한다.