<SNU AI 대회 보고서>

삼육대학교 채은채
삼육대학교 채수민
덕성여자대학교 강민주

1. 전체적인 방법론 및 모델 구조- 채수민

: 문장 하나와 순서가 섞인 4장의 이미지가 주어졌을 때 올바른 배열(24개 순열 중 하나)을 예측하는 문제로 , 단순히 24-클래스 분류가 아니라 “문장-이미지 그라운딩”, “프레임 간 관계 추론이 결합된 구조 문제” 로 정의 하였다.

1) 파이프 라인

(1) DINOv2로 프레임의 전역, 공간, feature 추출한다.

(2) MiniLM으로 문장을 임베딩이 아닌 단어 토큰 feature로 변환한다.

(3) 각 프레임의 patch feature가 문장, 단어 토큰을 쿼리로 cross-attention하여 이미지-텍스트
그라운딩 수행한다. 수행한 결과를 contact/곱/절대차로 프레임 표현 융합한다.

(4) 학습 가능한 슬롯 임베딩+ order-CLS 토큰과 함께 2-layer 관계 Transformer에 입력해
프레임 간 상대적 선후관계를 학습한다.

(5) permutation/ position/ pairwise/ no-ordering 4개 헤드를 동시에 학습하는 멀티태스크를
출력한다.

네 헤드는 서로 다른 표현에서 같은 정답을 다른 방식으로 예측하도록 하여 서로의 정규화 역할을 하도록 설계하였다.

2. 제안한 아이디어의 동기와 핵심 기여

1) 동기

: 4프레임을 연결한 후 24-way softmax로 직접 분류하면, 실제 데이터에서 비율이 높은 “이미 정순(1,2,3,4)” 클래스로 쉽게 붕괴한다. 학습 로그에서 already-ordered-exact-accuracy는 1에 가깝고 needs-order-exact 정확도는 0에 가까워지는 것이 대표적이다. 24-way one-hot 하나만으로는 “두 프레임 중 무엇이 먼저인가” 같은 고밀한 관계 신호가 명시적으로 전달되지 않는다. 이를 해결하기 위해서는 5가지가 있다.

2) 핵심기여

(1) 멀티태스크 보조 손실

: pairwise position no-ordering 손실을 추가하고 inverse sqrt 클래스 가중치를 손실에 적용해 정순 편향 붕괴를 다층으로 방지한다.

(2) 텍스트 정렬 순위 손실

: “문장은 대체로 시간순으로 서술된다.”는 약한 가정하에 앞선 프레임일수록 문장 앞쪽
단어에 attention이 집중되도록 유도한다.

(3) 추론시 Fusion Decoding

: 24개 순열 후보 마다 permutaion/ position/pairwise 세 헤드의 로그확률을 가중합해 점수를 매기고 최댓값을 선택한다. 단일 argmax가 아닌 3-헤드 합의 기반 앙상블적 디코딩이며 결합 가중치는 그리드 서치로 튜닝한다.

학습 및 추론 과정

2. 모델 학습 과정

2.1 기본 아키텍처 및 학습 파이프라인 검증학습은 DINOv2와 MiniLM을 완전히 고정한 상태로 네 프레임의 전역/공간 특징과 문장의 단어 토큰 특징을 사전에 1회 추출하여 캐시로 저장하고, 이후 캐시된 특징만을 이용해 관계 Transformer 및 예측 head를 학습하는 방식으로 진행되었다. 각 이미지의 patch 특징은 문장 토큰에 대해 cross-attention을 수행하여 이미지별로 어떤 단어와 의미적으로 연결되는지를 학습하며(grounding),(24-way),(4-way),(이진), 순서 존재 여부 분류(이진)를 동시에 수행하는 멀티태스크 손실로 학습되었다.

3. 추론 과정

추론 단계에서는 학습된 세 가지 예측 head(순열 예측, 프레임별 위치 예측, 프레임 쌍별 선후 관계 예측)의 출력을 결합하여 최종 순서를 결정하는 융합 디코딩 방식을 적용하였다. 초기 구현에서는 순열 예측 head의 출력만으로 최종 답을 결정하였으나, 이는 학습 과정에서 함께 최적화된 위치·쌍별 예측 정보를 추론 단계에서 활용하지 못하는 한계가 있었다. 이에 24개의 유효한 순열 각각에 대해 순열 예측 확률, 위치 예측 확률, 쌍별 선후 관계 확률을 가중합한 점수를 계산하고, 점수가 최대인 순열을 최종 예측으로 채택하는 방식으로 개선하였다.

4. 결론 및 향후 과제

최빈값 기준선 대비 약 2배 수준의 검증 정확도(약 32~33%)를 확보하였으며, 이 과정에서 코드 결함 검증 절차의 중요성, 조기 종료 및 정규화 설정이 학습 결과에 미치는 영향, 그리고 멀티태스크 head 정보를 추론 단계에서 결합하는 것의 효과를 확인하였다. 다만 현재 수준은 고정된 사전학습 특징의 표현력에 의해 제한되는 것으로 판단되며, 향후에는 GPU 자원을 확보하여 대형 비전 인코더로의 전환 및 종단간 미세조정을 시도함으로써 추가적인 성능 개선을 도모할 필요가 있다.

성능 향상을 위해 적용한 주요 기법

2. 멀티 헤드 융합 디코딩

모델은 순열 예측, 프레임별 위치 예측, 프레임 쌍별 선후 관계 예측, 순서 존재 여부 예측을 동시에 학습하는 멀티태스크 구조를 취하고 있으나, 초기 구현에서는 추론 시 순열 예측 head의 출력만을 최종 답으로 사용하고 있었다. 이에 따라 학습 과정에서 함께 최적화된 위치·쌍별 예측 정보가 추론 단계에서 전혀 활용되지 못하는 구조적 비효율이 존재하였다. 이를 개선하기 위하여, 24개의 유효한 순열 각각에 대해 순열 예측 로그/위치 예측 로그/ 쌍별 선후 관계 로그확률을 가중합하여 점수를 산출하고, 점수가 최대인 순열을 최종 예측으로 채택하는 융합 디코딩 기법을 도입하였다.

3. 클래스 불균형 완화

학습 데이터의 정답 분포를 분석한 결과, 전체 24개 순열 클래스 중 순서가 이미 올바른 경우가 전체의 약 15.5%를 차지하여 나머지 23개 클래스에 비해 상대적으로 빈도가 높은 불균형이

확인되었다. 초기에는 역제곱근 기반의 클래스 가중치를 손실 함수에 적용하였으나, 가중치가 과도하게 작용하고 있음을 확인하였다. 이에 클래스 가중치의 적용 범위를 특정 구간(0.5배~2.0배)으로 제한하여, 다수 클래스가 학습 과정에서 사실상 배제되지 않도록 조정하였다.

4. 정규화 강화를 통한 과적합 완화

dropout 비율을 0.10에서 0.25로, weight decay를 1e-4에서 3e-4로 상향 조정하여 모델의 일반화 성능을 높이고자 하였다. 아울러 학습률을 3e-4에서 1.5e-4로 낮추고, 조기 종료 인내를 4 epoch에서 8 epoch로 완화하여, 검증 지표의 자연스러운 변동으로 인해 학습이 지나치게 이른 시점에 종료되는 것을 방지하였다.

5. 순열 제약 학습을 위한 Sinkhorn 정규화

프레임별 위치 예측 head는 네 개의 슬롯을 각각 독립적인 4-way 분류로 학습하는 구조로 인해, "각 위치는 정확히 하나의 프레임에만 배정되어야 한다"는 순열 고유의 제약을 학습 과정에서 명시적으로 반영하지 못하는 한계가 있었다. 이에 위치 예측 결과에 대해 행과 열을 반복적으로 정규화하는 Sinkhorn 알고리즘을 적용하여 이중확률행렬에 근사시키고, 이를 기반으로 한 보조 손실을 전체 손실 함수에 추가하였다.

8. 모델 및 특징 표현력 확장

정체 구간을 극복하기 위한 추가적인 시도로, DINOv2가 추출하는 공간적 지역 토큰의 해상도를 4×4에서 8×8로 확장하여 프레임 내 세부적인 위치 변화를 더 정밀하게 포착하도록 하였으며, 이에 맞추어 관계 Transformer의 은닉 차원을 256에서 320으로, 레이어 수를 2에서 3으로 확대하였다. 아울러 GPU 환경에서는 DINOv2 자체를 소형에서 중형 모델로 교체하거나, 사전학습된 비전 인코더의 마지막 일부 계층을 unfreeze하여 원본 이미지 기반으로 종단간 미세조정을 수행하는 방안을 마련하였다.

9. 결론

본 절에서 서술한 기법들은 크게 추론 방식 개선, 학습 안정화, 구조적 제약 반영 시도, 모델 다양성 확보, 표현력 확장의 다섯 범주로 구분할 수 있다. 이 중 실질적으로 검증 정확도 향상에 기여한 것으로 확인된 기법은 융합 디코딩, 클래스 가중치 조정, 정규화 강화, 디코딩 가중치 탐색이었으며, Sinkhorn 정규화는 가설과 달리 효과가 없거나 오히려 특정 head의 학습을 저해하는 것으로 확인되어 철회하였다. 이러한 과정은 성능 개선을 위한 기법 도입 시 각 기법의 효과를 개별적으로 검증하는 절차가 병행되어야 함을 시사한다.

##1. 전체 성능 비교

구분	모델 및 주요 설정	Valid raw exact	Valid fused exact	결과 해석

무작위 기준선	24개 순열 중 무작위 선택	4.17%	-	24-class 문제의 이론적 무작위 성능
최빈값 기준선	모든 표본을 [1,2,3,4]로 예측	15.50%	15.50	순서 변경 표본을 전혀 해결하지 못함
X-CLIP zero-shot	24개 순열 후보를 X-CLIP으로 직접 비교	약 4.00%	-	무작위 수준으로 시간 순서 일반화에 실패
X-CLIP 학습 pilot	2 epoch candidate-ranking 학습	약 12.00%	-	최빈값 기준선보다 낮음
초기 관계 모델	초기 DINOv2·MiniLM 관계 Transformer 설정	약 4.35%	-	순열 분류가 무작위 수준에 머묾
Test2 3-epoch pilot	100% shuffle, class weight 미사용, AMP 사용	3.99%	5.17	프레임 순서를 실제로 학습하지 못함
Test2 FP32 pilot	100% shuffle, class weight 미사용, AMP 해제	3.57%	10.42	전체 순열 성능은 낮음
Shuffle 제거 실험	shuffle 0%, class weight 미사용, FP32	15.52%	15.73	정확도 상승처럼 보이지만 사실상 정순 클래스 편향
Shuffle 제거 실험 + 가중치	shuffle 0%, class weight 사용, FP32	15.17%	15.59	정순 클래스 붕괴를 해결하지 못함
복수 설정 동시 변경	shuffle 50%, class weight 제거, 정규화·loss 등을 동시에 변경	-	30.89	기존 최고 모델보다 낮아 과소적합 가능성 확인

DINOv2+MiniLM 기본 모델	Frozen feature, 관계 Transformer, 멀티태스크 학습 및 fusion	15.57%	31.41	같은 epoch에서 fusion이 raw head보다 크게 우수
최고 체크인트	DINOv2+MiniLM 기본 구조의 최고 validation checkpoint	-	32.98	현재까지 확인된 가장 높은 검증 성능
재현 실험 범위	다른 실행 환경 또는 반복 실행	-	32.14%~33.30%	약 1.16%포인트 범위에서 유사한 성능으로 수렴

##6. 방법론의 장점과 한계

1. 방법론의 장점

1) 이미지와 문장을 직접 연결하는 grounding 구조
: 단순히 이미지 특징과 문장 전체 임베딩을 결합하지 않고, 각 이미지 patch가 문장의 단어 토큰을 cross-attention으로 참조하도록 설계하였다. 이를 통해 각 프레임이 문장 내 어떤 동작 또는 상태 표현과 관련되는지를 프레임별로 학습할 수 있다.

2) 순열 문제의 구조를 반영한 멀티태스크 학습
: 24개 순열 분류만 수행하지 않고, 프레임별 위치, 두 프레임 간 선후관계, 정순 여부를 함께 학습하였다. 동일한 정답을 서로 다른 수준에서 감독함으로써 단일 분류 손실보다 풍부한 시간 관계 정보를 제공한다.

2. 방법론의 한계

1) 고정된 이미지 특징의 표현력 한계
: Frozen DINOv2는 객체와 장면의 의미적 특징에는 강하지만, 프레임 순서 판단에 필요한 작은 손동작이나 상태 변화에 직접 최적화돼 있지 않다. 따라서 외형이 거의 동일한 인접 프레임에서는 선후관계를 구분하지 못하는 사례가 발생한다.

2) 시간 정보가 없는 단일 이미지 인코더
: 네 프레임을 각각 독립적으로 인코딩한 뒤 관계 Transformer에서 결합하므로, 비디오 모델처럼 움직임의 연속성이나 optical flow를 직접 학습하지 못한다. 프레임 사이의 변화 방향은 고수준 특징의 차이만으로 간접 추론해야 한다.

3. 장점과 한계를 연결한 향후 개선 방향

확인된 한계	향후 개선 방향
미세한 프레임 변화 구분 부족	고해상도 patch, 차분 feature, optical flow 추가
Frozen DINOv2 한계	마지막 block 부분 fine-tuning