Towards Training Stronger Video Vision Transformers for EPIC-KITCHENS-100 Action Recognition

Abstract

최근 비전 트랜스포머 연구가 급증하면서 영상 인식, 포인트 클라우드 분류 및 비디오 이해와 같은 다양한 도전적인 컴퓨터 비전 애플리케이션에 대한 놀라운 잠재력을 입증했습니다. 이 문서에서는 EPIC-KITCHENS-100 Action Recognition 데이터 세트에 대한 보다 강력한 비디오 비전 트랜스포머 학습에 대한 경험적 결과를 제시합니다. 구체적으로, 우리는 비디오 비전 트랜스포머에 대한 학습 기법(예: 증강, 해상도, 초기화 등)을 탐색합니다. 학습 방법을 통해 단일 ViViT 모델은 EPIC-KITCHENS-100 데이터 세트의 검증에서 47.4%의 성능을 달성하여 원본 문서 [1]에 보고된 성능을 3.4% 능가했습니다. 비디오 트랜스포머는 동사-명사 액션 예측 모델에서 특히 좋습니다. 이로 인해 비디오 트랜스포머의 전반적인 액션 예측 정확도가 컨볼루션 변압기보다 눈에 띄게 높아집니다. 놀랍게도, 최고의 비디오 트랜스포머조차도 동사 예측에서 컨볼루션 네트워크를 저하시킵니다. 따라서 우리는 비디오 비전 트랜스포머와 일부 컨볼루션 비디오 네트워크를 결합하고 EPIC-KITCHENS-100 액션 인식 대회에 솔루션을 제시합니다.

1. Introduction

컴퓨터 비전 분야의 최근 발전은 이미지 인식[6, 25], 포인트 클라우드 분류[23] 및 비디오 이해[1,2]와 같은 다양한 컴퓨터 비전 애플리케이션에서 놀라운 잠재력을 입증한 트랜스포머 기반 모델 제품군의 급속한 확장을 목격했습니다. 그들은 적절한 증강 전략의 조합이 주어졌을 때, 컨볼루션 네트워크의 성능을 대체하는 것으로 나타났습니다 [16].

이 문서에서는 비디오 비전 트랜스포머의 학습 기술에 대한 최근 탐색을 보고합니다. 특히, ViViT [1]를 기본 모델로 채택하고, 네트워크 초기화뿐만 아니라 데이터 소스의 품질, 증강, 입력 해상도의 영향을 조사했습니다. 그 결과 ViViT는 Epic-Kitchen-100 데이터 세트의 액션 인식 정확도에서 47.4%를 달성할 수 있습니다. 또한, ViViT는 액션 분류에서는 컨볼루션 네트워크보다 성능이 뛰어나지만 동사 분류에서는 컨볼루션 네트워크보다 성능이 떨어집니다. 이것은 그들의 앙상블이 최종 정확도를 높이는 데 도움이 될 수 있다는 것을 의미합니다. 비디오 트랜스포머와 컨볼루션 트랜스포머를 결합함으로써 이 백서는 마침내 Epic-Kitchen-100 Action Recognition 과제에 대한 솔루션을 제시합니다.

2. Training video vision transformers

우리는 인수 분해된 인코더가 포함된 ViViT-B/16x2를 기본 모델로 사용합니다. 두 개의 분류 헤드가 동일한 클래스 토큰에 연결되어 입력 비디오 클립의 동사와 명사를 각각 예측합니다. 먼저 공개적으로 사용 가능한 대규모 비디오 데이터셋에서 네트워크를 사전 학습한 다음 서사 키친 데이터셋에서 ViViT를 미세 조정합니다.

Model	Dataset	Resolution	Top 1	Views
ViViT-B/16x2 Fact. encoder	K400	224	78.6	4×3
		320	80.6	4×3
	K700	224	69.7	4×3
		320	71.5	4×3
	SSV2	224	63.6	1×1
		320	-	1×1
	K400-Raft	224	60.5	4×3
	K400-Tvl1	224	65.4	4×3

1번 테이블. 키네틱스 400, 700 및 SSV2에 대한 사전 학습 ViViT 입력 해상도 Y의 각 데이터 세트 X를 사용한 사전 학습은 X-Y로 표시된다. 예를 들어, 입력 해상도 224로 K400에서 학습된 초기화 가중치에 대해서는 K400-224로 표시합니다.

2.1. Initialization Preparation

[14, 13, 18]에서 사용되는 감독된 사전 학습[17, 7, 1, 4]과 감독되지 않은 모델[10, 9]과 같이 사전 학습된 모델을 준비하는 방법은 여러 가지가 있습니다. 여기서는 감독된 사전 학습을 채택하여 더 나은 다운스트림 성능을 제공합니다. 모델은 먼저 Kinetics 400[11], Kinetics 700[3] 및 Something-Something-V2[8]에 대해 학습됩니다. 저마다. 이 파트에서는 성능을 높이기 위해 대부분 DeepViT[25]의 학습 레시피를 따릅니다. 특히 AdamW를 최적화 도구[12]로 사용하고 기본 학습률을 0.0001로 설정합니다. 가중치 감소는 0.1로 설정되어 있습니다. [1]의 초기화 접근법에 따라 ImageNet21k에서 사전 학습된 ViT 가중치로 ViViT 모델을 초기화하고 코사인 학습 속도 스케줄로 30 epoch로 모델을 학습합니다. 학습은 시작 학습률이 1e-6인 2.5 epoch로 준비됩니다. 우리는 색상 확대, 혼합 및 레이블 스무딩을 활성화합니다. 이 모형은 droppath 비율이 0.2로 정규화됩니다. Kinetics 및 SSV2에 대한 결과는 1번 테이블과 같습니다. 또한 Raft[15]와 TVL1[21]을 사용하여 추출한 Kinetics 400의 optical flow에 대해 ViViT를 학습했습니다.

2.2. Training video transformers on Epic-Kitchen

Epic-Kitchen의 비디오 트랜스포머 학습을 위해 초기화, 데이터 소스의 품질, 증강, 입력 해상도, 액션 계산 전략 및 임시 샘플링 보폭의 측면에서 학습 레시피를 요약합니다. 최적화 도구, 기본 학습률을 포함한 학습 매개변수입니다. 학습 일정은 총 50 epoch로 설정되었으며 10 epoch로 워밍업됩니다. 그 결과는 2번 테이블에서 관찰할 수 있습니다. 달리 지정되지 않은 경우 간격에 1을 사용하여 프레임을 샘플링합니다.

초기화: ImageNet-21K, Kinetics400, Kinetics700 및 SSV2의 사전 학습을 통해 초기화를 방지합니다. SSV2 초기화를 중단한 이유는 SSV2도 복잡한 시공간 상호작용을 가진 자기 중심 액션 인식 데이터 세트이기 때문입니다. 강력한 초기화(ImageNet21K에서 Kinetics400까지, 나아가 Kinetics700까지)를 사용하면 액션 인식 정확도가 현저하게 향상되는 것을 관찰할 수 있습니다. 만약 우리가 동사와 명사 한정의 개선을 분해한다면, 우리는 더 강한 초기화 모델이 명사 예측에 가장 많은 개선을 가져다준다는 것을 알 수 있습니다. 그러나 K700에서 SSV2(0.1%)로 초기화를 대체하여 더 높은 동사 예측 정확도를 관찰할 수 있지만 명사 예측(1.4%)은 현저하게 감소하였습니다. 따라서 최종 제출 시 SSV2로 초기화된 모델은 포함하지 않았습니다.

ID	Init.	Qual.	Res.	Aug.	Top1 A	Top1 V	Top1 N	Top1 A*
A	IN21K	256	224	CJ	36.1	62.4	48.2	-
B	**K400-224**	256	224	CJ	37.2	61.7	50.9	-
C	K400-224	**512**	224	CJ	38.4	62.7	52.2	-
D	**K700-224**	512	224	CJ	39.6	63.5	53.3	-
E	K700-224	512	224	**CJ+**	42.8	65.2	56.2	-
F	K700-224	512	**320**	CJ+	45.2 46.3†	67.4	58.9	42.4 43.4
G	**SSV2-224**	512	320	CJ+	44.5 45.7†	**67.5**	57.5	- -
I	K700-224	512	**384**	CJ+	45.8 **47.0**†	67.2	**59.0**	42.5 -
[1]	-	-	224	CJ*	44.0	66.4	56.8	-

2번 테이블. EPIC-KITCHENS-100에서 ViViT 미세 조정. Init은 학습 전 데이터 집합을 나타냅니다. Qual은 변환이 수행되기 전의 입력 비디오의 짧은 쪽 길이를 나타냅니다. 원본 비디오의 크기를 조정합니다. Res는 모델에 대한 입력 비디오의 해상도를 나타냅니다. Aug는 랜덤 자르기 및 랜덤 플립 외에 확대 전략을 나타냅니다. A, V, N은 각각 액션, 동사, 명사 예측 정확도를 나타냅니다. A*는 테스트 세트의 액션 예측 정확도를 나타냅니다. CJ+와 CJ는 각각 랜덤 컬러 지터를 나타내며, 섞임 없이 랜덤 지우기로 믹스를 절단합니다. CJ*는 [1]에서 사용된 다른 증강 전략을 나타냅니다. †행동 예측은 각 관점을 통합하기 전에 먼저 계산됩니다. 파란색 글꼴은 각 실험의 변화를 강조 표시합니다. 성능 열의 굵게 표시된 글꼴은 최상의 성능을 나타내는 모델입니다.

데이터 출처 품질: 하드 드라이브 I/O에 대한 부담을 완화하여 학습 속도를 높이기 위해 비디오의 짧은 부분을 각각 256과 512로 조정합니다. 입력 데이터 소스의 품질을 높이면 액션, 동사, 명사 예측이 각각 1.2%, 1.0%, 1.3% 향상될 수 있음을 관찰할 수 있습니다.

증강: 더 강한 증강(혼합 [22], 절단 혼합 [20] 및 랜덤 지우기 [24])을 활용하는 이점을 관찰합니다. 랜덤 컬러 지터링만 사용하는 것에 비해 증강 효과가 강하면 액션 예측이 3.2%, 동사 예측이 1.7%, 명사 예측이 2.9% 향상됩니다.

입력 해상도: 입력 해상도를 추가로 변경합니다. 입력 해상도를 224에서 320으로 높이면 액션 예측이 약 2.4%, 동사 예측이 약 2.2%, 명사 예측이 약 2.7% 향상됩니다. 예측 정확도의 포화도는 입력 해상도를 320에서 384로 추가로 높일 때 관찰되며, 액션 예측의 개선은 0.6%만 관찰됩니다.

ID	Temp Sampling Rate	Top1 A	Top1 V	Top1 N
F	2	45.2 46.3†	67.4	58.9
I	3	46.4 47.4†	68.4	59.6
[1]	2	44.0	66.4	56.8

3번 테이블. ViViT. A, V, N의 시간 샘플링 속도를 변경하는 것은 각각 작용, 동사 및 명사 예측 정확도를 나타냅니다. † 액션 예측은 각 관점을 통합하기 전에 먼저 계산됩니다.

Model	Optical Flow	Top1 A	Top1 V	Top1 N
ViViT-B/16x2-Flow-A	Raft	34.6 35.4†	66.8	43.5
ViViT-B/16x2-Flow-B	TVL1	34.5 35.1†	66.4	43.3

4번 테이블. optical flow로 ViViT를 학습합니다. A, V, N은 각각 액션, 동사, 명사 예측 정확도를 나타냅니다. † 행동 예측은 각 관점을 통합하기 전에 먼저 계산됩니다.

예측 정확도는 입력 해상도를 320에서 384로 추가로 높일 때 관찰되며 작용 예측의 개선은 0.6%만 관찰됩니다. 액션 점수 계산: 표에 †가 있는 숫자로 표시된 것처럼 액션 점수를 다르게 계산하면 액션 예측 결과가 달라질 수 있습니다. 각 비디오 클립에 대해 두 가지 예측을 수행하므로 다중 뷰에 대한 액션 예측을 집계하는 두 가지 방법이 있습니다. 동사 $P_v^i \in \mathbb{R}^{1\times N_v}$ 와 명사 $P_n^i \in \mathbb{R}^{1\times N_n}$에 대해 각각 예측이 있다고 가정하면, 여기서 N_v와 N_n은 동사와 명사의 클래스 수를 나타내고 i는 뷰에 대한 색인을 나타낸다. 예측을 집계하는 첫 번째 방법은 다음과 같습니다.

$$P_a = (\sum_i P_v^\top)(\sum_i P_n) \,, \qquad (1)$$

여기서 $P_a \in \mathbb{R}^{N_v\times N_n}$은 액션에 대한 예측을 나타냅니다. 이 접근법은 액션 예측을 직접 계산하기 전에 먼저 다중 뷰에 대한 동사와 명사 예측을 집계합니다. 두 번째 접근법은 각 관점을 통합하기 전에 각 관점에 대한 액션 예측을 각각 계산합니다.

$$P_a = \sum_i (P_v^\top P_n) \,. \qquad (2)$$

2번 테이블에서 볼 수 있듯이, 각 뷰에 대한 조치 점수를 합산하면 다른 변종보다 약 1% 더 높은 성능을 얻을 수 있습니다. 더 중요한 것은 이러한 개선이 테스트 세트에도 반영될 수 있다는

것입니다.

시간 샘플링 stride. Epic-Kitchen 데이터 집합은 상대적으로 높은 FPS를 가지므로, 하나의 프레임을 간격(즉, 시간 샘플링 속도가 2임을 의미)으로 하는 샘플링 프레임이 시간 적용 범위에 불충분할 수 있습니다. 32프레임 샘플링 시 1초만 적용됩니다. 따라서 임시 샘플링 속도도 줄였고 그 결과는 테이블 5번에 나와 있습니다. 보시다시피, 임시 샘플링 속도를 약간 수정하면 성능이 현저하게 향상될 수 있습니다. 그 이유 중 하나는 시간적 여유가 더 길기 때문일 것입니다. 또 다른 가능한 이유는 샘플링 속도 3에 의해 생성된 FPS가 사전 학습 FPS에 가깝기 때문입니다. ViVi-B/16x2-I의 최종 단일 모델 성능은 [1]에서 보고된 성능을 3.4% 능가합니다.

Model	Training	Top1		
		A	V	N
TimeSformer 8×32	original	34.4	57.1	51.3
	ours-224	39.4	63.9	51.7
	ours-320	**42.5**	**65.2**	**55.0**

5번 테이블. EPIC-KITCHENS-100에 대한 TimeSporm의 결과는 각각 액션, 동사 및 명사 예측 정확도를 나타냅니다. 모든 액션 정확도는 액션 예측을 집계하기 전에 계산하여 얻습니다.

2.3. Training video transformers with optical flow

더 나은 motion feature를 포착하기 위해 optical flow는 또 다른 데이터 소스로 활용합니다. optical flow를 데이터 소스로 사용하는 비디오 트랜스포머는 앞서 언급한 것과 동일한 학습 방법을 사용하여 학습됩니다. Raft[15]와 TVL1[21]을 사용하여 각각 추출된 optical flow인 두 가지 optical flow 모델을 학습했습니다. 결과는 제시되어 있습니다.

2.4. Other transformer based models

우리가 사용하는 또 다른 트랜스포머 기반 비디오 분류 모델은 TimeSformer [2]입니다. TimeSformer의 경우 K600에서 사전 학습된 오픈 소스 모델을 직접 사용하고 먼저 기본 설정을 유지하고 15개 Epoch에 대해 학습했습니다. 그런 다음 우리는 비교해서 우리의 학습 레시피를 사용했습니다. 이것은 우리의 학습 레시피가 액션 예측 정확도에서 오리지널을 5% 향상시킨다는 것을 보여줍니다. 입력 해상도를 더 높이면 액션 예측 정확도가 3% 향상됩니다.

3. Training convolutional video networks

비디오 비전 트랜스포머는 성능이 우수할 수 있지만, 컨볼루션 네트워크에서도 보완적인 예측이 필요합니다. 다음 부분에서 볼 수 있듯이 CSN[17]과 SlowFast[7]와 같은 컨볼루션 네트워크는 동사 예측에 상대적으로 강합니다.

우리는 ir-CSN-152와 SlowFast-16 x 8-101을 우리의 기본 모델로 사용합니다. ViViT 모델의 학습

과정과 유사하게, 우리는 Kinetics 700에서 이 두 모델을 학습함으로써 사전 학습된 가중치를 구합니다 [3]. EPICKITCHENS-100 데이터셋에 대한 학습을 위해 최적화 프로그램 및 학습 속도 일정 등을 포함하여 ViViT와 동일한 학습 매개 변수를 사용합니다. 우리는 [5]를 따르고 학습 중에 배치 표준 평균과 분산을 동결합니다. 그 결과는 6번 테이블에서 볼 수 있습니다. 보시다시피 배치 표준 평균과 분산을 동결하면 액션 인식 정확도가 약 1.3% 향상됩니다. mixup, cutmix 및 랜덤 소거 기능을 적용하면 검증 및 테스트 세트 모두에서 더 향상된 결과를 얻을 수 있습니다. 그러나 ViViT에서의 실험 결과와 달리 입력 해상도를 높이면 검증 세트의 성능이 실제로 향상되지만 테스트 세트의 정확도는 향상되지 않습니다. 따라서, 우리는 SlowFast 16x8-101의 학습 해상도를 224로 유지하고 있습니다. 입력 해상도가 224x224에 불과한 경우에도 동사 예측 측면에서 대부분의 ViViT를 능가할 수 있다는 점이 흥미롭습니다.

ID	Model	F. BN	Res.	Aug.	Top1 A	Top1 V	Top1 N	Top1 A*
A	ir-CSN-152	×	224	CJ	41.0 42.4†	66.4	52.4	37.8 -
B	ir-CSN-152	✓	224	CJ	42.7 43.9†	67.6	55.1	- 40.9
C	ir-CSN-152	✓	224	CJ+	43.5 44.5†	68.4	55.9	- **42.5**
D	ir-CSN-152	✓	320	CJ+	45.1 **46.2**†	**69.0**	**57.2**	- 42.4
-	SlowFast-16×8-101	✓	224	CJ+	43.0 43.9†	68.2	55.1	- -

6번 테이블. EPIC-KITCHENS-100의 ir-CSN-152 및 SlowFast-16x8-101 미세 조정. F.BN은 동결된 배치 평균 및 분산을 나타냅니다. Res는 모델에 대한 입력 비디오의 해상도를 나타냅니다. Aug는 랜덤 자르기 및 랜덤 플립 외에 확대 전략을 나타냅니다. A, V, N은 각각 액션, 동사, 명사 예측 정확도를 나타냅니다. A*는 테스트 세트의 액션 예측 정확도를 나타냅니다. † 액션 예측은 각 관점을 통합하기 전에 먼저 계산됩니다. 파란색 글꼴은 각 실험의 변화를 강조 표시합니다. 성능 열의 굵게 표시된 글꼴은 최상의 성능을 나타내는 모델입니다.

하나의 비디오 클립에 대해 더 긴 기간을 다루기 위해 CSN 모델에는 추가로 Long-term Feature Banks(LFB)[19]를 사용합니다. 이러한 실험을 위해, 우리는 Epic-Kitchen 학습을 받은 ir-CSN-152s로 모델을 초기화하고, 2개의 예열 epoch로 이전과 동일한 기본 학습률로 10개 epoch로 모델을 추가 학습합니다. 그 결과는 7번 테이블과 같습니다. 학습을 초기화하는 데 사용된 원래 모델에서 추출한 기능을 사용할 때 명사 예측에 대한 ir-CSN-152-C의 개선이 관찰됩니다. ViViT 기능을 feature bank로 사용하면 명사 예측이 더욱 향상되어 최종 액션 예측 정확도가 크게 향상됩니다. 그에 비해 동사의 정확도는 거의 영향을 받지 않습니다.

ID	Model	LFB Feature	Top1		
			A	V	N
-	ir-CSN-152-B	-	43.9†	67.6	55.1
E	ir-CSN-152-B	ir-CSN-152-B	42.9†	66.9	54.7
F	ir-CSN-152-B	ViViT-B/16x2-F	47.3†	67.6	60.1
-	ir-CSN-152-C	-	44.5†	68.4	55.9
G	ir-CSN-152-C	ir-CSN-152-C	44.8†	68.1	56.8
H	ir-CSN-152-C	ViViT-B/16x2-F	**47.3†**	**68.1**	**60.3**

7번 테이블. ir-CSN-152에 LFB 적용. † 액션 예측은 각 관점을 통합하기 전에 먼저 계산됩니다. 성능 열의 굵게 표시된 글꼴은 최상의 성능을 나타내는 모델입니다.

Model Name	Top1		
	A	V	N
ir-CSN-152-B	43.9	67.6	55.1
ir-CSN-152-C	44.5	68.4	55.9
ir-CSN-152-F	47.2	67.6	60.1
ir-CSN-152-G	44.8	68.1	56.8
ir-CSN-152-H	47.3	68.1	60.3
SlowFast-16×8-101	43.9	68.2	55.1
ViViT-B/16x2-Flow-A	35.4	66.8	43.5
ViViT-B/16x2-Flow-B	35.1	66.4	43.3
ViViT-B/16x2-F	46.3	67.4	58.9
ViViT-B/16x2-H	47.0	67.2	59.0
ViViT-B/16x2-I	47.4	68.4	59.6
TimeSformer-320	42.5	65.2	55.0
Overall (Val)	51.7	72.4	62.6
Overall (Test)	48.5	69.2	60.3

8번 테이블. 앙상블 모델 목록입니다. 표에 나열된 모든 성능은 집계하기 전에 각 보기에 대한 액션을 계산하는 것입니다.

4. Ensembling models

서로 다른 모델의 상호 보완적인 예측을 활용하기 위해 제시된 모델의 선택된 부분 집합을 결합했습니다. 선택한 모델 세트가 8번 테이블에 나와 있습니다. 모델들의 앙상블은 액션 예측에서 최고 성능의 성능을 4.3% 향상시킵니다. 우리가 얻은 최종 시험 정확도는 액션 예측 48.5%, 동사 예측 69.2%, 명사 예측 60.3%입니다.

5. Conclusion

본 문서에서는 EPICKITCHENS-100 액션 인식 과제에 대한 NAT 솔루션을 소개합니다. 우리는 보다 강력한 비디오 비전 트랜스포머를 학습하고, 여러 비디오 비전 트랜스포머와 컨볼루션 비디오 인식 모델을 조합하여 성능을 강화하기 시작했습니다.