Is Space-Time Attention All You Need for Video Understanding?

https://arxiv.org/abs/2102.05095

0. Introduction

- 비디오 이해(video understanding)에서 기존 2D/3D convolution 기반 모델의 한계를 극복하고자 함.
- NLP에서 성공한 self-attention 모델을 비디오 도메인에 적용 가능 여부 탐색.
- 기존 CNN 기반 모델은 지역적 정보에만 집중하며, 장거리 의존성 학습에 한계 존재.
- 연구 핵심 기여: 완전 self-attention 기반 비디오 모델(TimeSformer) 제안, 효율적 학습 및 장기 비디오 처리 가능.

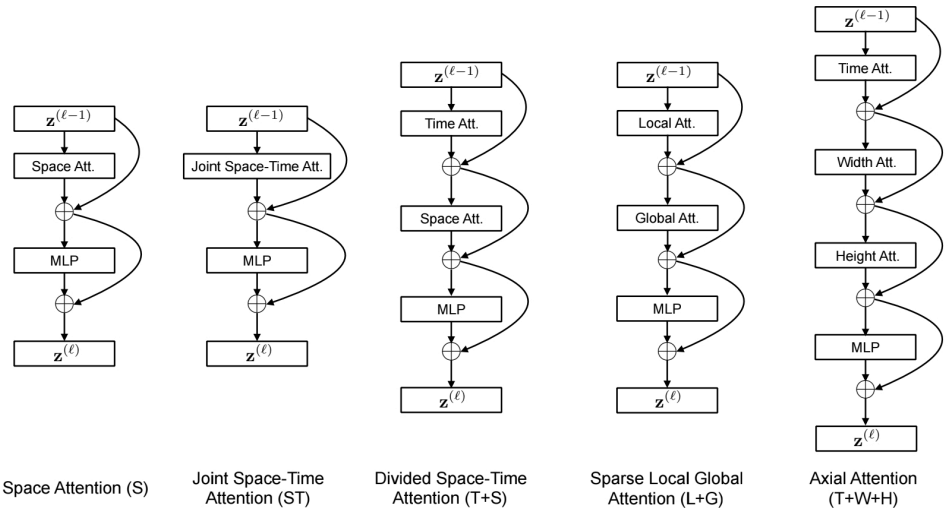
1. Overview

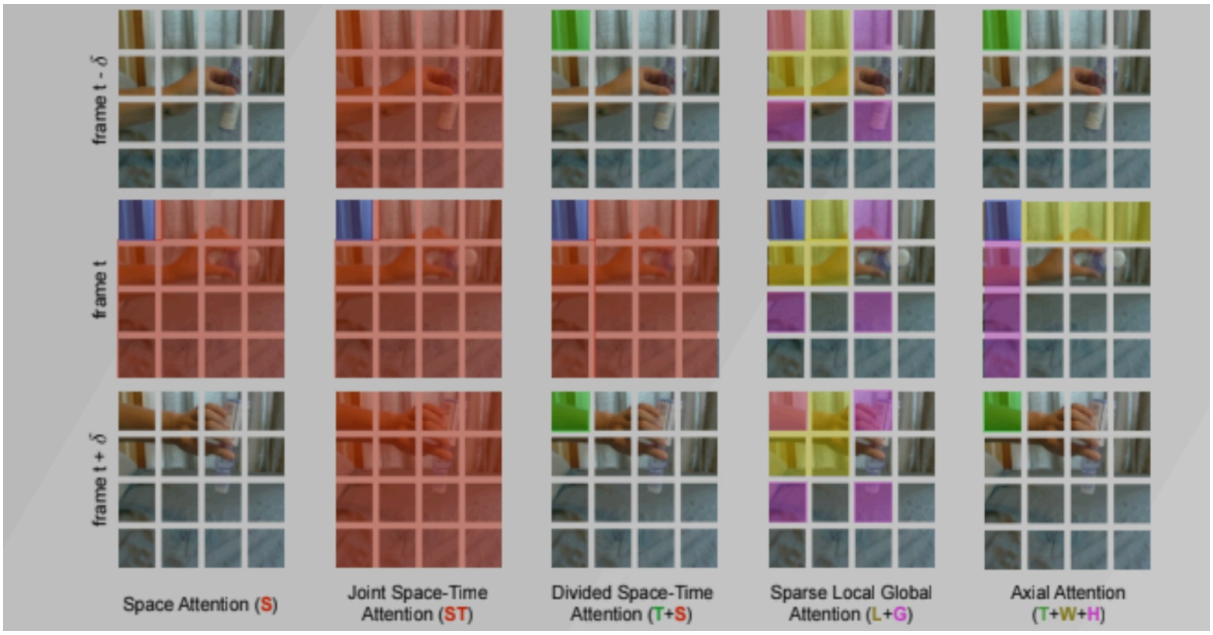
- TimeSformer는 Vision Transformer(ViT)를 기반으로, 프레임 패치를 순차적으로 처리하여 시공간적 feature 학습.
- divided attention 구조 사용, temporal과 spatial attention을 블록 내에서 분리 적용.
- convolution-free 모델로 기존 state-of-the-art 성능 달성 및 효율성 개선.

2. Challenges

- 비디오의 시공간적 길이와 고해상도 처리 시 self-attention 계산 비용 급증.
- 장거리 의존성을 학습하면서도 메모리와 시간 효율 유지 필요.
- 기존 CNN 기반 모델 대비 inductive bias가 적어 충분한 데이터와 최적화가 요구됨.
- 다양한 self-attention 설계 선택과 효율성 trade-off 존재.

3. Method





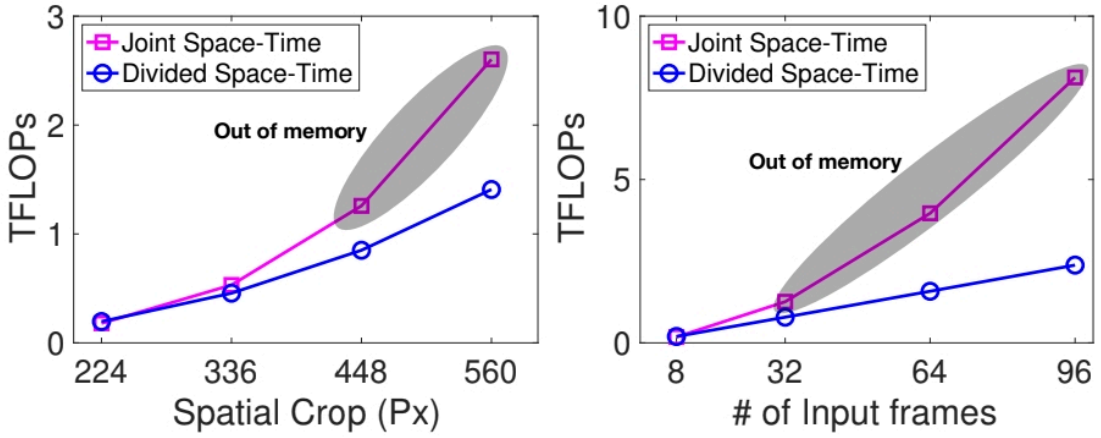
- 입력: 각 프레임을 패치 단위로 분할 후 선형 임베딩, positional encoding 추가.
- 모델 구조: Transformer encoder 기반, 블록 내에서 spatial과 temporal attention 분리 적용.
- 학습: 일반 classification 손실 사용, 하이퍼파라미터는 벤치마크 데이터셋에 최적화.
- 추가 설계: 효율적 self-attention을 위해 패치 단위 sequence 처리, long-range 비디오 처리 가능.

4. Experiments

Attention	Params	K400	SSv2
Space	85.9M	76.9	36.6
Joint Space-Time	85.9M	77.4	58.5
Divided Space-Time	121.4M	**78.0**	**59.5**
Sparse Local Global	121.4M	75.9	56.3
Axial	156.8M	73.5	56.2

- 데이터셋: Kinetics-400/600, 총 수십만 비디오 클립, train/validation 분할 사용.
- 비교 대상: 3D CNN 기반 SOTA 모델들.
- 평가 지표: Top-1, Top-5 accuracy.
- 실험 변수: attention 설계(divided, joint 등), 모델 크기, clip 길이.

5. Results

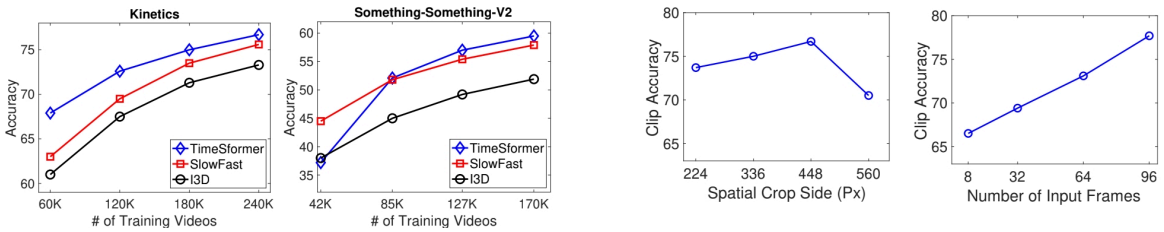


Model	Pretrain	K400 Training Time (hours)	K400 Acc.	Inference TFLOPs	Params
I3D 8x8 R50	ImageNet-1K	444	71.0	1.11	28.0M
I3D 8x8 R50	ImageNet-1K	1440	73.4	1.11	28.0M
SlowFast R50	ImageNet-1K	448	70.0	1.97	34.6M
SlowFast R50	ImageNet-1K	3840	75.6	1.97	34.6M
SlowFast R50	N/A	6336	76.4	1.97	34.6M
TimeSformer	ImageNet-1K	**416**	75.8	**0.59**	121.4M
TimeSformer	ImageNet-21K	**416**	**78.0**	**0.59**	121.4M

Table 2. Comparing TimeSformer to SlowFast and I3D. We observe that TimeSformer has lower inference cost despite having a larger number of parameters. Furthermore, the cost of training TimeSformer on video data is much lower compared to SlowFast and I3D, even when all models are pretrained on ImageNet-1K.

Method	Pretraining	K400	SSv2
TimeSformer	ImageNet-1K	75.8	**59.5**
TimeSformer	ImageNet-21K	**78.0**	**59.5**
TimeSformer-HR	ImageNet-1K	77.8	62.2
TimeSformer-HR	ImageNet-21K	**79.7**	**62.5**
TimeSformer-L	ImageNet-1K	78.1	**62.4**
TimeSformer-L	ImageNet-21K	**80.7**	62.3

Table 3. Comparing the effectiveness of ImageNet-1K and ImageNet-21K pretraining on Kinetics-400 (K400) and Something-Something-V2 (SSv2). On K400, ImageNet-21K pretraining leads consistently to a better performance compared to ImageNet-1K pretraining. On SSv2, ImageNet-1K and ImageNet-21K pretrainings lead to similar accuracy.



Method	SSv2	Diving-48**
SlowFast (Feichtenhofer et al., 2019b)	61.7	77.6
TSM (Lin et al., 2019)	63.4	N/A
STM (Jiang et al., 2019)	64.2	N/A
MSNet (Kwon et al., 2020)	64.7	N/A
TEA (Li et al., 2020b)	65.1	N/A
bLVNet (Fan et al., 2019)	**65.2**	N/A
TimeSformer	59.5	74.9
TimeSformer-HR	62.2	78.0
TimeSformer-L	62.4	**81.0**

- Kinetics-400/600에서 기존 SOTA 대비 동등하거나 우수한 정확도 달성.
- Divided attention 구조가 가장 높은 성능 기록.
- 학습 속도 및 inference 효율 3D CNN 대비 우수.
- 긴 비디오 클립 처리 가능, accuracy 소폭 저하에도 효율적 처리.
- Ablation study: spatial/temporal attention 분리 시 성능 향상 확인.

6. Insight

- self-attention만으로 convolution 없이도 비디오 이해 가능.
- 모델은 장기 의존성 학습에 강점, 데이터 풍부 환경에서 효과적.
- 실무 적용 시 효율성과 확장성 측면에서 장점 크지만, 데이터 부족 환경에서는 inductive bias 부족 단점 존재.
- 후속 연구: attention 구조 개선, 효율적 연산, 다양한 비디오 태스크 확장 가능성.