SlowFast Networks for Video Recognition

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, Kaiming He

2023.01.13 논문 리뷰

배성훈

SlowFast Networks for Video Recognition

- **Research Background**
  - 정지영상의 경우, 모든 orientation이 동일하면서 이동 불변성 (shifted invariant)의 특징을 가짐
  - Video의 경우, 모든 motion의 spatiotemporal orientation은 같지 않음

  - 즉, Video recognition에서 slow motion과 fast motion의 기여도가 다름
  - 이러한 이유로, 저자는 Spatial structures와 temporal events를 서로 분리해서 다룸
    - Spatial semantics of visual content는 느리게 변함
      예를 들어, 손을 흔드는 행위에서 '손'이란 객체의 identity는 고유함
    - 이러한 categorical semantics (colors, textures, lighting..)을 인식하는 것은 상대적으로 느리게 인식
    - 이와 반대로 motion은 상대적으로 빠르게 변화

  - 이런 직관을 통해 Two-pathway SlowFast 모델 제안

SlowFast Networks for Video Recognition

- **Motivation**
  - Retinal ganglion cells에 대한 연구에 motivate
  - 사람 시각 시스템에 대한 연구로, cell들이 80% Parveocellular(P-cells)와 15~20%의 Magnocellular(M-cells)로 이뤄짐
  - M-cells는 high temporal frequency에 대한 연산 및 fast-temporal change에 반응
    -> spatial detail이나 color에는 거의 반응을 하지 않음
  - P-cells는 spatial detail, color에 대해 반응, temporal information에는 천천히 반응
  - 이러한 연구 결과를 기반으로, SlowFast 모델 설계
  - 각 pathway가 P-cells와 M-cells의 비율 (80:20)인 것과 같이 각 pathway의 계산량이 80:20을 보임

  - 즉, 각 pathway가 M-cells와 P-cells의 기능을 하도록 설계
  - 인간 시각 시스템을 적용해 Video recognition 성능 향상

SlowFast Networks for Video Recognition

- **Method**:
  - **SlowFast Network**
  - 기존의 모델 (Two stream design)과 차이
    - 서로 다른 temporal speed를 탐지하지 못함
    - Optical flow 계산 X, End-to-End 학습

  - Slow pathway
    - ✓ Semantic information을 image or few sparse frames에서 추출 (lower temporal rate)
    - ✓ Low frame rates와 flow refreshing speed로 작동
  - Fast pathway
    - ✓ Motion 포착
    - ✓ High temporal resolution, Fast refreshing speed
    - ✓ 전체 계산량의 20%만 차지 -> 모델 경량화 기여
      Fewer channel을 가지고, spatial information 처리에 많은 자원을 쓰지 않도록 설계
    - ✓ Temporal pooling을 진행하지 않음 ->high frame rate로 동작해 temporal fidelity 유지

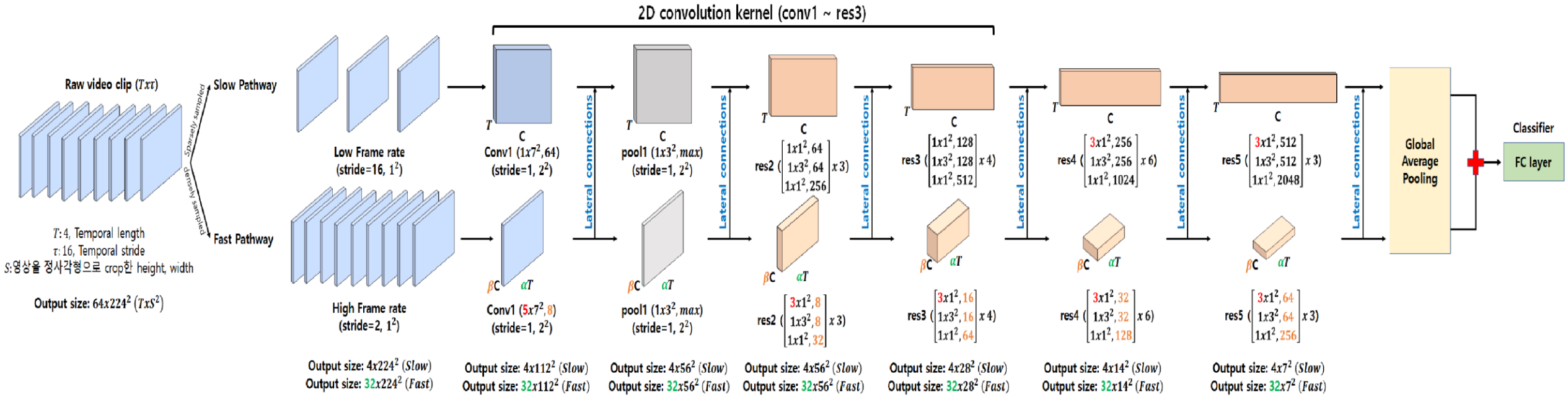
SlowFast Networks for Video Recognition

- **Method**:
  - **SlowFast Network**

    2개의 서로 다른 frame rates로 동작

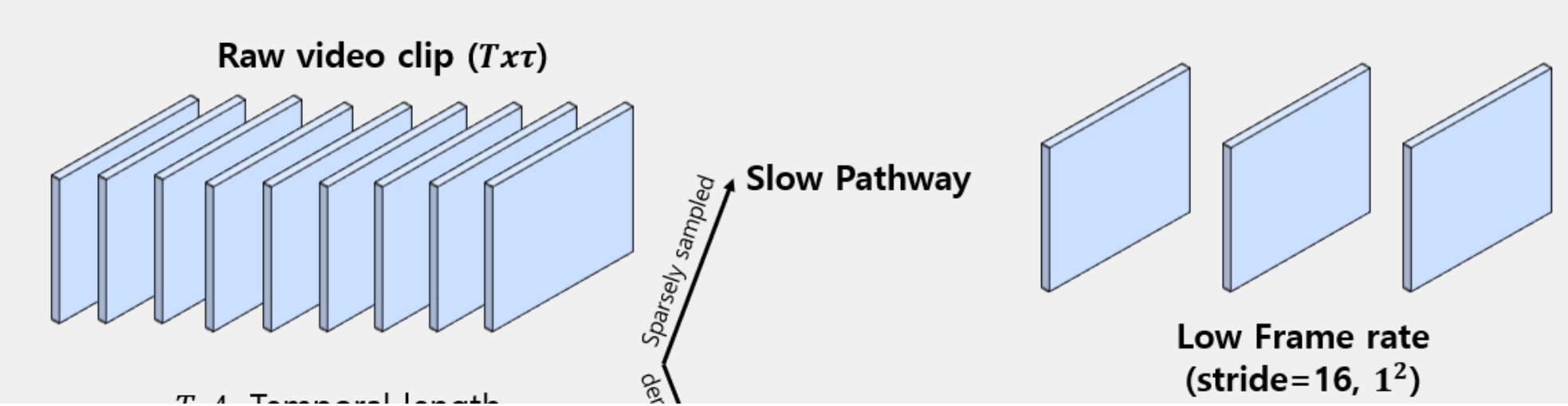
    Slow pathway와 Fast pathway로 분할해서 각각 spatial information, temporal information 학습

    각 과정에서 lateral connection을 사용해 Fast information을 Slow information과 합침



SlowFast Networks for Video Recognition

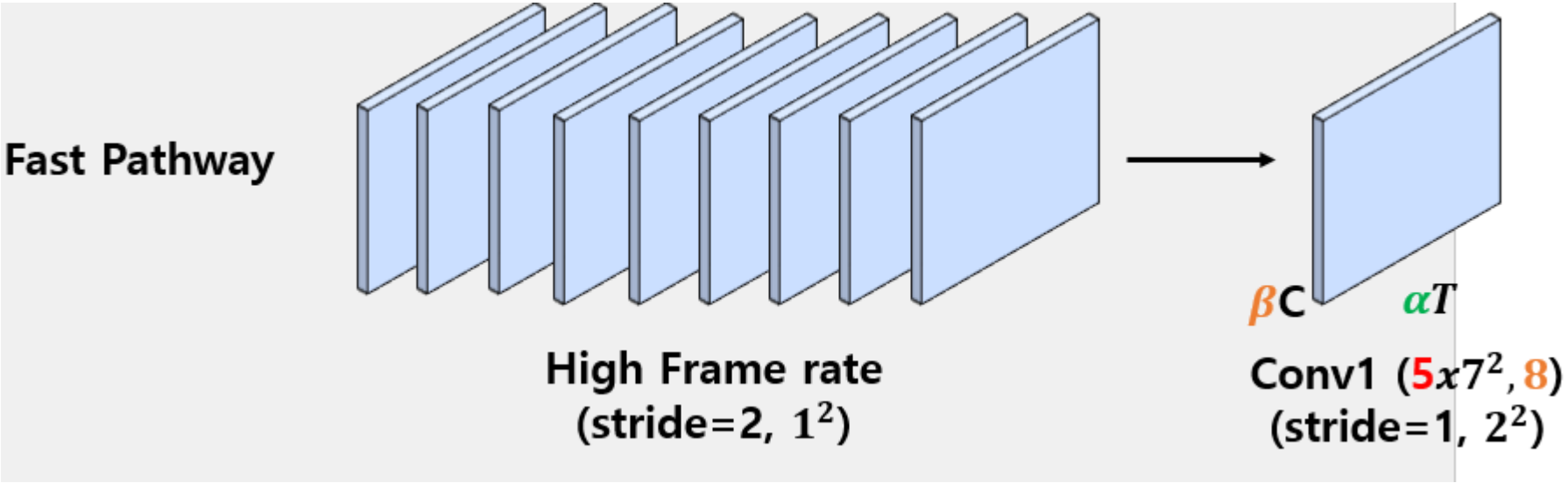
- **Method**:
  - **Slow pathway**
    - Input frames에 대해 temporal stride τ 적용 -> low frame rate
      τ: 16, **Raw video clip에서 16 frames 마다 샘플링**, 30fps 영상에서 대략 1초에 2frame만 추출
    - 샘플링된 frame 수가 T이면, Raw video의 frame 수는 $T\text{x}\tau$
    - 일반적인 C3D, I3D와 다르게, **non-degenerate temporal convolution** 사용 (**temporal kernel size > 1**)
      Slow pathway에서는 res4와 res5에만 적용
      -> 초기 layer에서 temporal convolution 사용이 정확도를 저하시키는 실험에 의거



SlowFast Networks for Video Recognition

- **Method**:
  - **Fast pathway:**
    - 더 큰 temporal resolution과 작은 channel capacity를 가짐

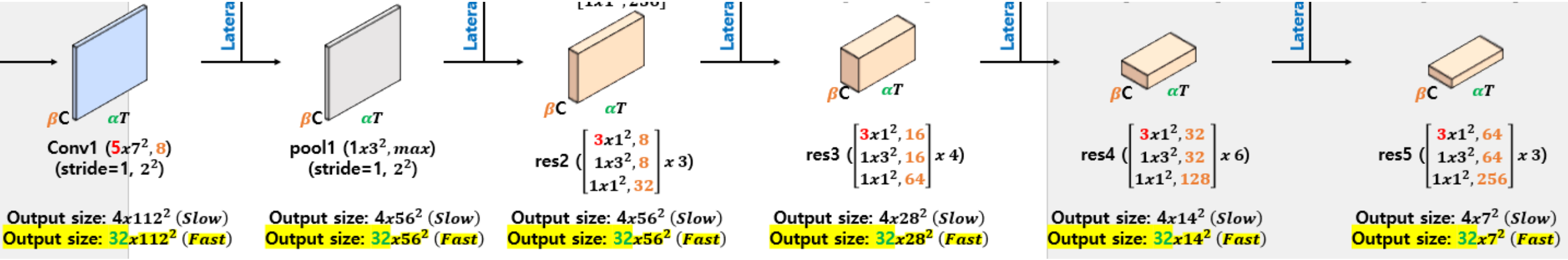
      α=8, $\beta=\frac{1}{8}$
    - 모든 block에서 **non-degenerate temporal convolution** 사용 (**temporal kernel size > 1**)
    - Detailed motion을 포착하기 위해 temporal convolution에 대해 fine temporal resolution 유지
      -> 이를 위해 **temporal downsampling layer X**



SlowFast Networks for Video Recognition

- **Method**:
  - **Fast pathway: High temporal resolution features**
    - High input resolution을 입력으로 받으면서 Network layer 전체에서 high resolution features 추구 (Temporal downsampling layer X, time-strided convolution X)
    - Downsampling은 global pooling layer 전까지 사용하지 않음
    - Feature tensor는 temporal dimension에 따라 항상 αT **frames**를 갖도록 **temporal fidelity** 유지

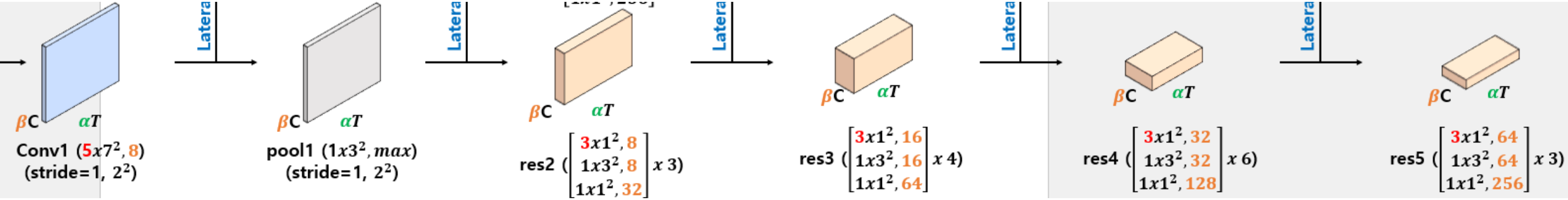
Fast pathway의 high temporal resolution 유지 ($\alpha T : 8x4$)



SlowFast Networks for Video Recognition

- **Method**:
  - **Fast pathway: Low channel capacity**
    - 좋은 accuracy 달성을 위해 lower channel capacity 사용
      -> 모델 경량화
    - Fast pathway는 slow pathway와 유사한 convolution layer로 구성되지만, Slow pathway의 β ($\beta < 1$) channels의 비율을 가짐 -> $\beta = \frac{1}{8}$
    - 공통된 layer의 FLOPs가 layer의 channel scaling 비율에 따라 Quadratic 할 수 있음
      -> Fast pathway가 Slow pathway보다 계산이 효율적 (모델 경량화 및 FLOPs={80:20})

Fast pathway의 Lower channel capacity ($\beta: \frac{1}{8}$)



SlowFast Networks for Video Recognition

- **Method**:
  - **Fast pathway: Low channel capacity**
    - Fast pathway는 spatial dimension에서 특별한 처리 과정이 없기 때문에 spatial modeling capacity가 낮음
    - Fast pathwa는 **spatial modeling ability가 약할 수록 temporal modeling ability가 강한 trade off** 가짐
    - 위의 개념을 motivate해, input spatial resolution의 감소, **color information 제거** -> accuracy 향상 기여
    - 모델 경량화 측면에서 less spatial capacity가 더 이득

Channel capacity ratio

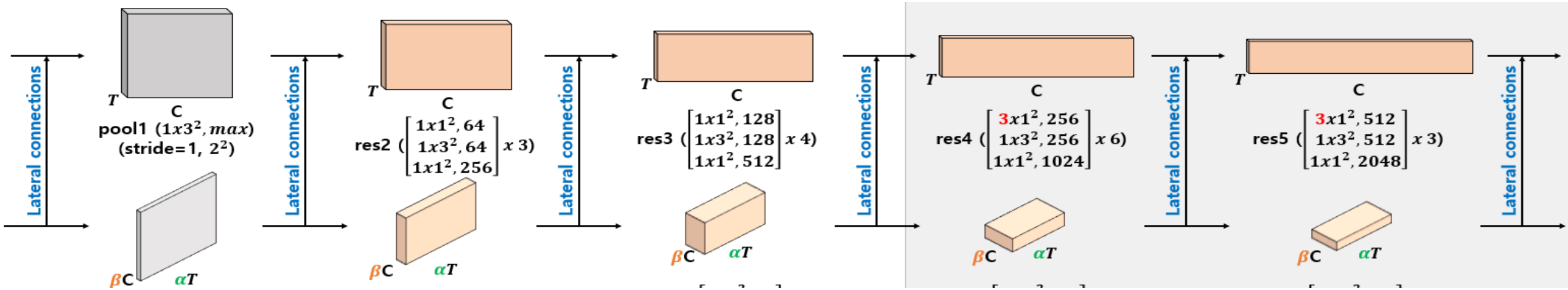
	top-1	top-5	GFLOPs
Slow-only	72.6	90.3	27.3
$\beta = 1/4$	75.6	91.7	54.5
1/6	**75.8**	92.0	41.8
1/8	75.6	**92.1**	36.1
1/12	75.2	91.8	32.8
1/16	75.1	91.7	30.6
1/32	74.2	91.3	**28.6**

Weaker spatial input to Fast pathway

Fast pathway	spatial	top-1	top-5	GFLOPs
RGB	-	**75.6**	**92.1**	36.1
RGB, β=1/4	*half*	74.7	91.8	34.4
gray-scale	-	**75.5**	**91.9**	**34.1**
time diff	-	74.5	91.6	34.2
optical flow	-	73.8	91.3	35.1

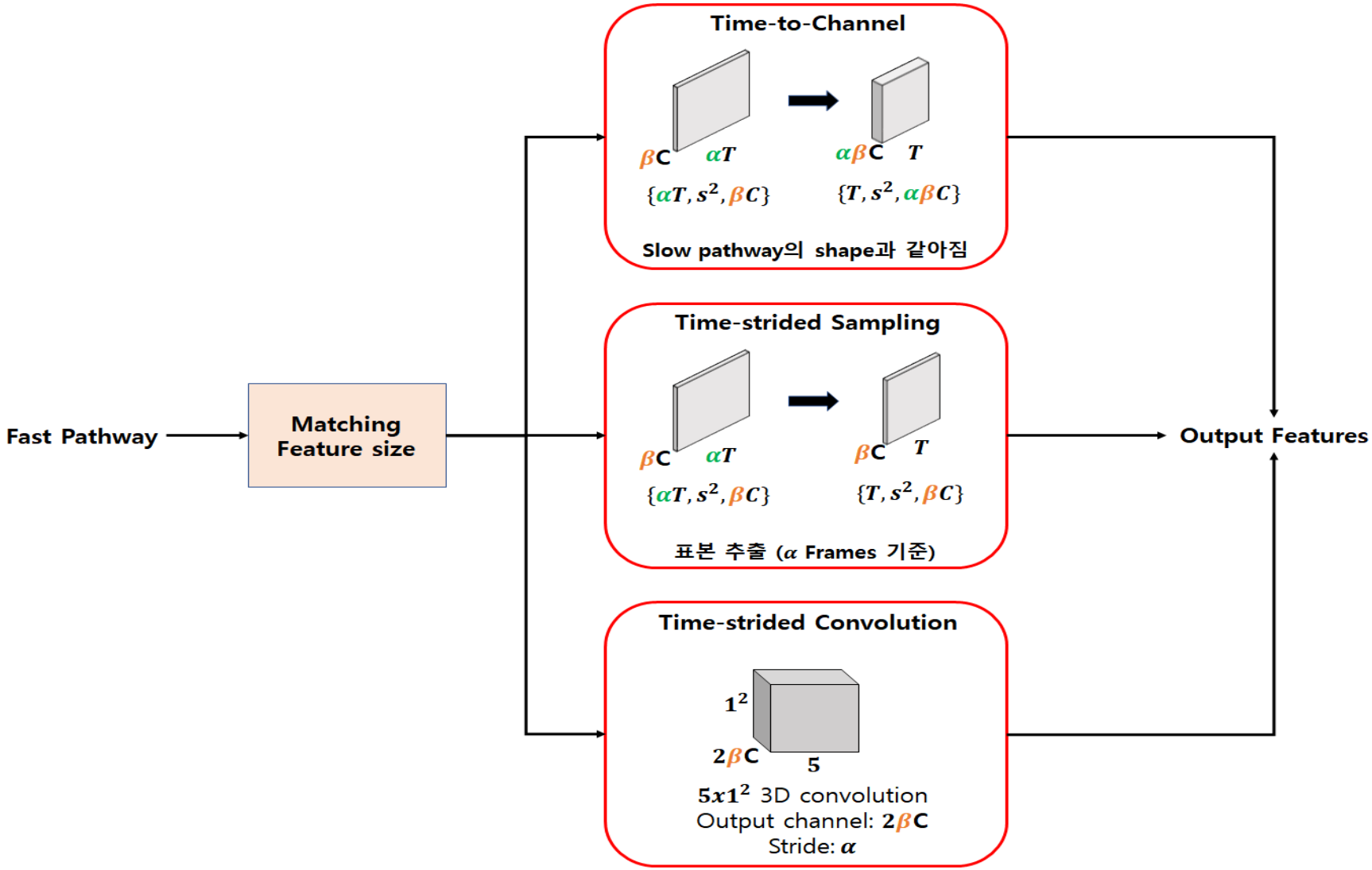
SlowFast Networks for Video Recognition

- **Method**:
  - **Lateral connections**
    - Object Detection에서 서로 다른 level의 spatial resolution과 semantics를 합치는 기법
    - 2 Pathway의 정보가 합쳐짐 (Fast pathway의 정보가 Slow pathway에 추가되는 **단방향 연결**)
    - 하나의 Lateral connection을 모든 stage마다 two pathways 사이에 추가
    - In ResNet. pool1, res2, res3, res4, res5 각각의 뒤에 추가
    - 2 Pathway의 서로 다른 temporal dimension을 맞추기 위한 **transformation** 수행
    - 이때 features를 pathway에 연결하는 다양한 전략을 사용해 성능 비교



SlowFast Networks for Video Recognition

- **Method**:
  - **Lateral connections: Transformation**



SlowFast Networks for Video Recognition

- **Experiments**: **Action classification**
  - **Kinetics-400, -600, Charades**
  - **Training**
    - Kinetics dataset: Non Pretrained random initialization으로 학습
    - Synchronized SGD
    - **Temporal domain**: Full-length video에서 clip random sampling ($\alpha T x \tau$ frames)
      Input temporal: Slow pathway (T), Fast pathway (αT)
    - **Spatial domain**: Video로부터 [256,320] pixels로 resize 후 224x224 pixels로 random crop or Horizontal flips
  - **Inference**
    - Temporal axis 따라 10개의 clip으로 sampling
    - 각 clip에 대해, 더 짧은 spatial side를 256pixels로 확장, 256x256의 3개의 crop (spatial dimension cover)
    - Prediction을 위해 Average softmax score
    - 사용된 View 수에서 spacetime "View"당 FLOP 계산 (spatial clip이 있는 temporal clip)
    - Spatial size: 256^2, 30개 Views (3 Spatial crop x 10 temporal clip)

SlowFast Networks for Video Recognition

- **Experiments**: **Action classification**
  - **Datasets**
    - Kinetics-400: 240K training videos, 20K validation, 400 human action categories
    - Kinetics-600: 392K training videos, 30K validation, 600 classes
      Single spatially centered flip -> FLOPs computational cost 계산
    - Charades: 9.8K training videos, 1.8K validation, 157 classes
      평균 최대 30초 간격으로 길게 수행된 action을 찍은 multi-label classification
  - **Metrics: mAP**

SlowFast Networks for Video Recognition

- **Experiments**: **Action classification**
  - Kinetics-400, Input sampling ($T\times\tau$), Backbones: ResNet-50/101, Non local(NL)
    - 기존 SOTA보다 2.1% 더 높은 Top-1 accuracy
    - ImageNet pretrained 없는 모델에서도 더 높은 Top-1 accuracy 달성 (73.9% -> 79.8% / **5.8% Up**)
    - Pretrained의 유무에 관계없이 성능 향상

Comparison with the state-of-the-art on Kinetics-400

model	flow	pretrain	top-1	top-5	GFLOPs×views
I3D [5]		ImageNet	72.1	90.3	108 × N/A
Two-Stream I3D [5]	✓	ImageNet	75.7	92.0	216 × N/A
S3D-G [61]	✓	ImageNet	77.2	93.0	143 × N/A
Nonlocal R50 [56]		ImageNet	76.5	92.6	282 × 30
Nonlocal R101 [56]		ImageNet	77.7	93.3	359 × 30
R(2+1)D Flow [50]	✓	-	67.5	87.2	152 × 115
STC [9]		-	68.7	88.5	N/A × N/A
ARTNet [54]		-	69.2	88.3	23.5 × 250
S3D [61]		-	69.4	89.1	66.4 × N/A
ECO [63]		-	70.0	89.4	N/A × N/A
I3D [5]	✓	-	71.6	90.0	216 × N/A
R(2+1)D [50]		-	72.0	90.0	152 × 115
R(2+1)D [50]	✓	-	73.9	90.9	304 × 115
SlowFast 4×16, R50		-	75.6	92.1	36.1 × 30
SlowFast 8×8, R50		-	77.0	92.6	65.7 × 30
SlowFast 8×8, R101		-	77.9	93.2	106 × 30
SlowFast 16×8, R101		-	78.9	93.5	213 × 30
SlowFast 16×8, R101+NL		-	**79.8**	**93.9**	234 × 30

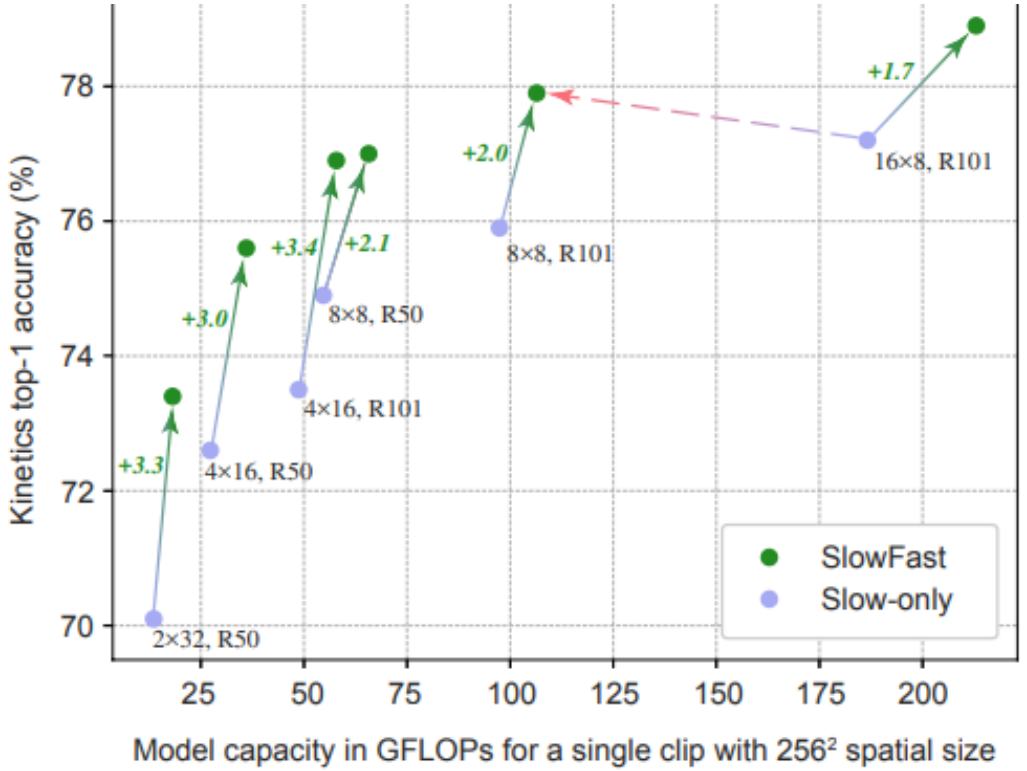
SlowFast Networks for Video Recognition

- **Experiments**: **Action classification**
  - Kinetics-400, Input sampling ($T\times\tau$), Backbones: ResNet-50/101, Non local(NL)
    - Low inference-time cost 달성
    - 기존의 연구는 temporal axis에 따라 clip이 상당히 **dense한 sampling** 사용 (Inference time: **> 100 views**)
    - In SlowFast, 많은 temporal clip이 필요하지 않음 (Spacetime view 당 cost 낮음, **36.1 GFLOPs**)
      -> **High temporal resolution**을 사용해 **모델 경량화** 기여

model	flow	pretrain	top-1	top-5	GFLOPs×views
I3D [5]		ImageNet	72.1	90.3	108 × N/A
Two-Stream I3D [5]	✓	ImageNet	75.7	92.0	216 × N/A
S3D-G [61]	✓	ImageNet	77.2	93.0	143 × N/A
Nonlocal R50 [56]		ImageNet	76.5	92.6	282 × 30
Nonlocal R101 [56]		ImageNet	77.7	93.3	359 × 30
R(2+1)D Flow [50]	✓	-	67.5	87.2	152 × 115
STC [9]		-	68.7	88.5	N/A × N/A
ARTNet [54]		-	69.2	88.3	23.5 × 250
S3D [61]		-	69.4	89.1	66.4 × N/A
ECO [63]		-	70.0	89.4	N/A × N/A
I3D [5]	✓	-	71.6	90.0	216 × N/A
R(2+1)D [50]		-	72.0	90.0	152 × 115
R(2+1)D [50]	✓	-	73.9	90.9	304 × 115
SlowFast 4×16, R50		-	75.6	92.1	36.1 × 30
SlowFast 8×8, R50		-	77.0	92.6	65.7 × 30
SlowFast 8×8, R101		-	77.9	93.2	106 × 30
SlowFast 16×8, R101		-	78.9	93.5	213 × 30
SlowFast 16×8, R101+NL		-	**79.8**	**93.9**	234 × 30

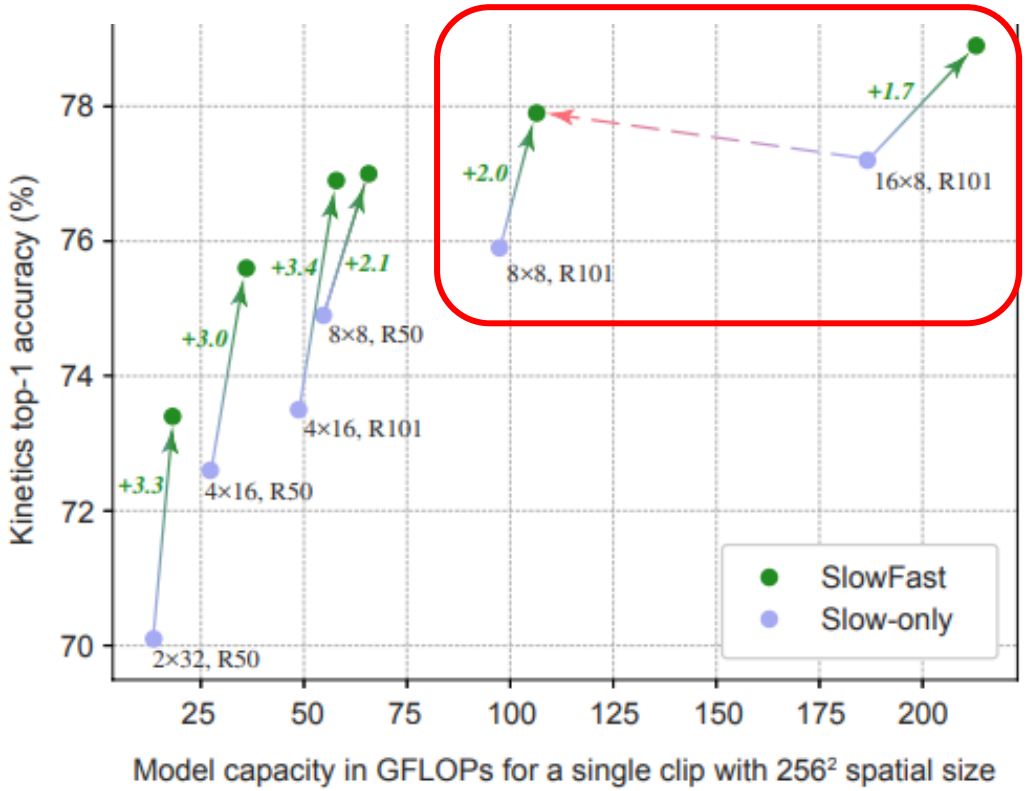
SlowFast Networks for Video Recognition

- **Experiments**: **Action classification**
  - **Trade off accuracy / complexity**
    - Slow-only와 SlowFast에 따른 성능 비교
    - **Horizontal axis**: 전체 Inference cost의 $\frac{1}{30}$에 비례하는 256^2 spatial size의 단일 input clip에 대한 model capacity 측정
    - Slow-only보다 SlowFast를 사용할 때 **더 낮은 cost로 더 좋은 성능**을 나타냄



SlowFast Networks for Video Recognition

- **Experiments**: Slow, Fast pathway 사용 전략에 따른 결과 비교
  - **Frame rate ratio α**
    - Slow pathway에서 frame 수를 2배로 늘리면, computation cost (horizontal axis)가 2배로 증가
    - 더 높은 frame rate에서 작동해도 computation cost의 작은 증가 -> 서로 다른 전략 모두 성능 향상 기여
    - 녹색 화살표: slow pathway 정보에 fast pathway 정보를 추가하는 이점
    - 빨간색 화살표: SlowFast가 더 높은 정확도와 적은 cost



SlowFast Networks for Video Recognition

- **Experiments**: **Action classification**
  - Kinetics-600, Input sampling ($T\times\tau$), Backbones: ResNet-50/101, Non local(NL)
    - Kinetics-600의 validation set은 Kinetics-400의 training set과 겹침
    - Kinetics-400을 pretrained X
    - SOTA인 ActivityNet보다 높은 성능 보임 (79.0% -> 81.8%)

Comparison with the state-of-the-art on Kinetics-600

model	pretrain	top-1	top-5	GFLOPs × views
I3D [3]	-	71.9	90.1	108 × N/A
StNet-IRv2 RGB [21]	ImgNet+Kin400	79.0	N/A	N/A
SlowFast 4×16, R50	-	78.8	94.0	36.1 × 30
SlowFast 8×8, R50	-	79.9	94.5	65.7 ×30
SlowFast 8×8, R101	-	80.4	94.8	106 × 30
SlowFast 16×8, R101	-	81.1	95.1	213 × 30
SlowFast 16×8, R101+NL	-	**81.8**	**95.1**	234 × 30

SlowFast Networks for Video Recognition

- **Experiments**: **Action classification**
  - Charades, Non local(NL)
    - Baseline (Slow-only) 사용시 (Pretrained Kinetics-400) mAP= 39.0
    - SlowFast에 NL과 Pretrained Kinetics-600 사용시 3.1 mAP 성능 향상 (42.1 -> 45.2)
    - 기존 SOTA보다 더 낮은 computation cost, 높은 mAP 달성

Comparison with the state-of-the-art on Charades

model	pretrain	mAP	GFLOPs × views
CoViAR, R-50 [59]	ImageNet	21.9	N/A
Asyn-TF, VGG16 [42]	ImageNet	22.4	N/A
MultiScale TRN [62]	ImageNet	25.2	N/A
Nonlocal, R101 [56]	ImageNet+Kinetics400	37.5	544 × 30
STRG, R101+NL [57]	ImageNet+Kinetics400	39.7	630 × 30
our baseline (Slow-only)	Kinetics-400	39.0	187 × 30
SlowFast	Kinetics-400	42.1	213 × 30
SlowFast, +NL	Kinetics-400	42.5	234 × 30
SlowFast, +NL	Kinetics-600	**45.2**	234 × 30

SlowFast Networks for Video Recognition

- **Experiments**: Slow, Fast pathway 사용 전략에 따른 결과 비교
  - **Fast pathway design ($Tx\tau = 4x16$)**
    - **Individual pathways**:
      - Slow, Fast pathway 각각 따로 사용했을 때 각각 모델 경량화 기여 (GFLOPs: 27.3, 6.4)
    - **SlowFast fusion**:
      - Naïve fusion (No lateral connection), two pathway의 Final output 연결 -> 73.5 mAP
      - Time-to-Channel, Time-strided sampling, Time-strided convolution: 모두 Slow-only보다 좋은 성능

SlowFast fusion

	lateral	top-1	top-5	GFLOPs
Slow-only	-	72.6	90.3	27.3
Fast-only	-	51.7	78.5	**6.4**
SlowFast	-	73.5	90.3	34.2
SlowFast	TtoC, sum	74.5	91.3	34.2
SlowFast	TtoC, concat	74.3	91.0	39.8
SlowFast	T-sample	75.4	91.8	34.9
SlowFast	T-conv	**75.6**	**92.1**	36.1

→ **Best mAP**

SlowFast Networks for Video Recognition

- **Experiments**: Slow, Fast pathway 사용 전략에 따른 결과 비교
  - **Fast pathway design (Channel capacity of Fast Pathway)**
    - 목적: detailed spatial representation 없이도 motion 포착에 대한 **lower channel capacity 유지**
      -> Channel ratio β로 제어
    - Best mAP: $\beta=\frac{1}{6},\frac{1}{8}$
    - $\beta=\frac{1}{32}\sim\frac{1}{4}$까지 모두 Slow-only 보다 더 좋은 성능 달성
    - $\beta=\frac{1}{32}$ 일 때, Fast pathway는 1.3 GFLOPs의 약간의 cost 증가가 있지만 1.6 mAP 성능 증가

Channel capacity ratio

	top-1	top-5	GFLOPs
Slow-only	72.6	90.3	27.3
$\beta = 1/4$	75.6	91.7	54.5
1/6	**75.8**	92.0	41.8
1/8	75.6	**92.1**	36.1
1/12	75.2	91.8	32.8
1/16	75.1	91.7	30.6
1/32	74.2	91.3	28.6

Best mAP ←

SlowFast Networks for Video Recognition

- **Experiments**: Slow, Fast pathway 사용 전략에 따른 결과 비교
  - **Fast pathway design (Weaker spatial input to fast pathway)**
    - Fast pathway에 서로 다른 weaker spatial input을 사용하는 연구
      1) Half spatial resolution (112x112) $\beta = \frac{1}{4}$ (FLOPs 유지 위해)
      2) Gray-scale input frames
      3) Time difference frames, 이전 frame에 현재 frame 빼서 계산
      4) Fast pathway input으로 optical flow 사용
    - 전체 방법이 Slow-only baseline 보다 좋은 성능
    - Grayscale은 RGB만큼의 성능, FLOPs는 5% 감소

Weaker spatial input to Fast pathway

Fast pathway	spatial	top-1	top-5	GFLOPs
RGB	-	**75.6**	**92.1**	36.1
RGB, β=1/4	*half*	74.7	91.8	34.4
gray-scale	-	**75.5**	**91.9**	**34.1**
time diff	-	74.5	91.6	34.2
optical flow	-	73.8	91.3	35.1

- 결론적으로, Fast pathway에 **weaker spatial input**과 **channel capacity** 적용은 **모델 경량화** 및 **temporally high resolution**에 영향을 줌

SlowFast Networks for Video Recognition

- **Experiments: Action Detection**
  - **AVA dataset**
    - Spatiotemporal localization of human action
    - 437 movie에서 추출
    - Spatiotemporal label은 초당 하나의 frame으로 제공
    - 모든 사람은 bounding box로 annotate
    - Action detection이 어려운 대신 Actor localization은 덜 어려움
    - 211K training, 57K validation video segments
    - 60 classes로 evaluate 60 classes에 대해 frame-level IoU threshold 0.5로 mAP

SlowFast Networks for Video Recognition

- **Experiments: Action Detection**
  - **Detection Architecture (약간의 수정)**
    - Res5의 Filters에 **Spatial stride=1, Dilation=2** 추가 -> Res5의 spatial resolution **2배** 증가
    - **Res5의 마지막 feature map에서 RoI features 추출**
      - 먼저, temporal axis에 따라 frame의 각 **2D RoI를** 복제해 **3D RoI**로 연장
      - **RoIAlign**을 통해 Spatially로 **RoI features**, Temporally로 **global average pooling** 계산
    - RoI features는 **max-pool** 되고 class별로 multi-label predictio을 위해 **sigmoid기반 classifier**로 전달
    - Action detection models과 공동으로 학습되지 않는 off-the-shelf person detector에 의해 Region proposals 계산
    - Person-detection model은 Detectron에 의해 학습됨: ResNeXt-101-FPN을 사용한 Faster R-CNN Backbone: Pretrained ImageNet & COCO human keypoint images
    - 학습된 모델의 detector를 AVA의 person detection에 맞춰 fine-tune
    - Action detection을 위한 Region proposals: confidence > 0.8로 detect된 Person boxes
      - Person class에 대해 Recall 91.1%, precision 90.7%

SlowFast Networks for Video Recognition

- **Experiments**: **Action Detection**
  - **Training**
    - Kinetics-400 classification models로부터 SlowFast weight 초기화
      - (**Pretrained Kinetics-400 model's weight** 사용 -> 빠른 수렴, Transfer learning)
    - **Step-wise learning rate** 사용
      - (특정 epoch마다 learning rate 감소), validation error가 포화될 때 learning rate를 10배 감소
    - AVA의 211k training dataset
      - 68 epoch, 각 epoch: 14000 iteration (첫 번째 1000 iteration에서 **linear warm-up**을 사용)
    - Weight decay: 10^{-7}
    - 다른 hyper parameter는 Kinetics Experiment와 같음
    - Input: 224x224 크기의 $\alpha T x \tau$ frames
    - Ground truth boxes는 training sample로 사용

SlowFast Networks for Video Recognition

- **Experiments: Action Detection**
  - **Inference**
    - 평가할 frame 주위의 αTxτ frames을 사용해 single clip에 대한 inference 수행
    - Height와 width 중 짧은 쪽이 256 pixels이 되도록 spatial dimension size 조정
    - Backbone feature extractor는 표준 Faster R-CNN에서 fully convolution을 사용해 계산

SlowFast Networks for Video Recognition

- **Experiments: Action Detection**
  - **Main results**
    - Pretrained Kinetics-400 사용한 경우 기존의 SOTA보다 **5.6 mAP** 높은 26.3 mAP 달성
    - Optical flow 없이 학습한 경우에도 **7.3 mAP** 높게 성능 달성
    - Kinetics-600 사용시 **0.5 mAP** 상승한 26.8 mAP 달성
    - Non local block 추가시 27.3 mAP로 상승
    - Test set에 대해 single crop set accuracy에 대해 27.1 mAP 달성
    - **Ground-truth boxes를 IoU > 0.9 겹치도록 예측한 Proposals**를 사용한 경우 28.2 mAP 달성 -> **SOTA**

model	flow	video pretrain	val mAP	test mAP
I3D [20]		Kinetics-400	14.5	-
I3D [20]	✓	Kinetics-400	15.6	-
ACRN, S3D [46]	✓	Kinetics-400	17.4	-
ATR, R50+NL [29]		Kinetics-400	20.0	-
ATR, R50+NL [29]	✓	Kinetics-400	21.7	-
9-model ensemble [29]	✓	Kinetics-400	25.6	21.1
I3D [16]		Kinetics-600	21.9	21.0
SlowFast		Kinetics-400	26.3	-
SlowFast		Kinetics-600	26.8	-
SlowFast, +NL		Kinetics-600	27.3	**27.1**
SlowFast*, +NL		Kinetics-600	**28.2**	-

SlowFast Networks for Video Recognition

- **Experiments: Action Detection**
  - **Main results**
    - AVA v2.2 에서는 29.0 mAP까지 증가
    - **16x8** SlowFast 모델 사용 시 29.8 mAP
    - Testing을 위해 **Multiple spatial scales**와 **Horizontal flip** 사용해 30.7 mAP 달성
    - 전체 **ensemble** 통해 Challenge: Test set에 대해 34.3 mAP 달성

SlowFast models on AVA v2.2

model	flow	video pretrain	val mAP	test mAP
SlowFast, 8×8		Kinetics-600	29.0	-
SlowFast, 16×8		Kinetics-600	29.8	-
SlowFast++, 16×8		Kinetics-600	**30.7**	-
SlowFast++, ensemble		Kinetics-600	-	**34.3**

SlowFast Networks for Video Recognition

- **Conclusion**
  - 본 논문에서는 사람의 인지 시스템을 모방해 새로운 Two-pathway Network를 제안
  - 각 Pathway의 효과에 대해 다양한 실험을 통해 입증했고, 다양한 전략을 사용해 성능 향상에 기여
  - 특히, Fast pathway design에 대한 저자의 고찰이 담겨있어 상당히 의미있는 논문이라고 생각함
  - 실험에 사용된 Dataset 모두에서 성능을 향상시키고 계산량을 감소시킨 만큼 추후 Action classification의 Backbone이 될 연구라고 생각함