Learning Spatiotemporal Features with 3D Convolutional Networks
(IEEE 2015)

2기 박소영

00. INTRODUCTION

Generic video descriptor의 필요성

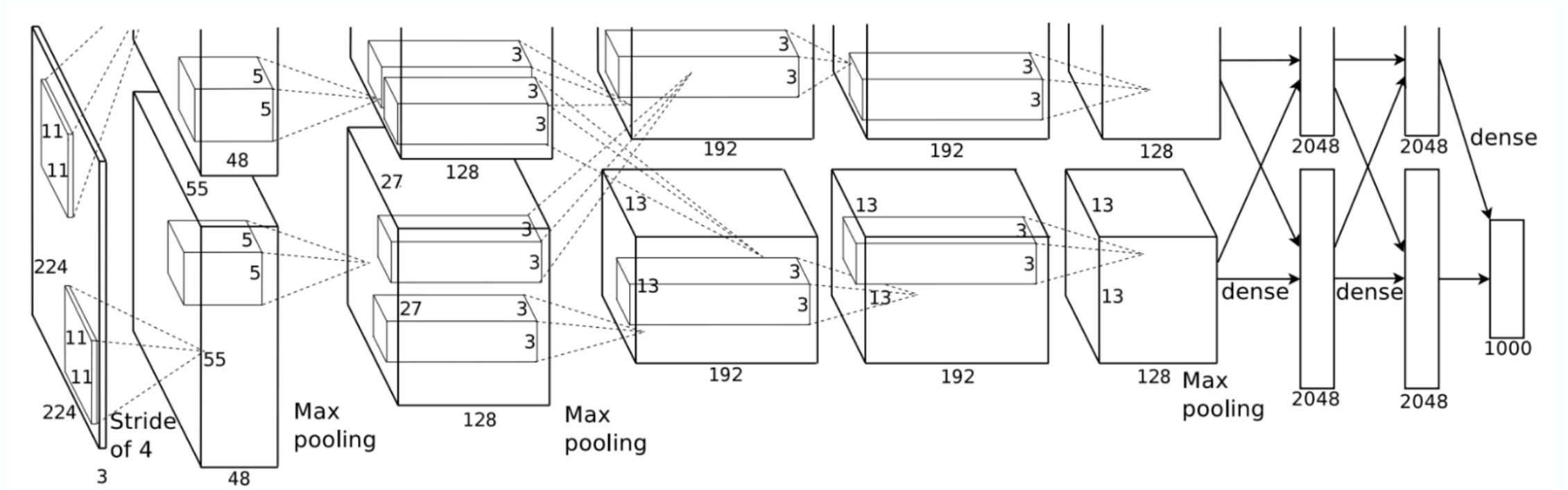
Video를 이해하기 위해 지금까지 다양한 접근 방법을 가져옴
=> Generic video descriptor가 필요하다

좋은 video descriptor란?

1) Generic : **다양한 종류**를 모두 이해할 수 있도록 일반적이어야 한다
2) Compact : 처리, 저장 등을 **더 큰 범위**에서 수행 가능해야 한다
3) Efficient : 수천 개의 video가 **real system**에서 동작 가능해야 한다
4) Simple : **더 간단한 모델**에서도 잘 동작해야 한다

00. INTRODUCTION

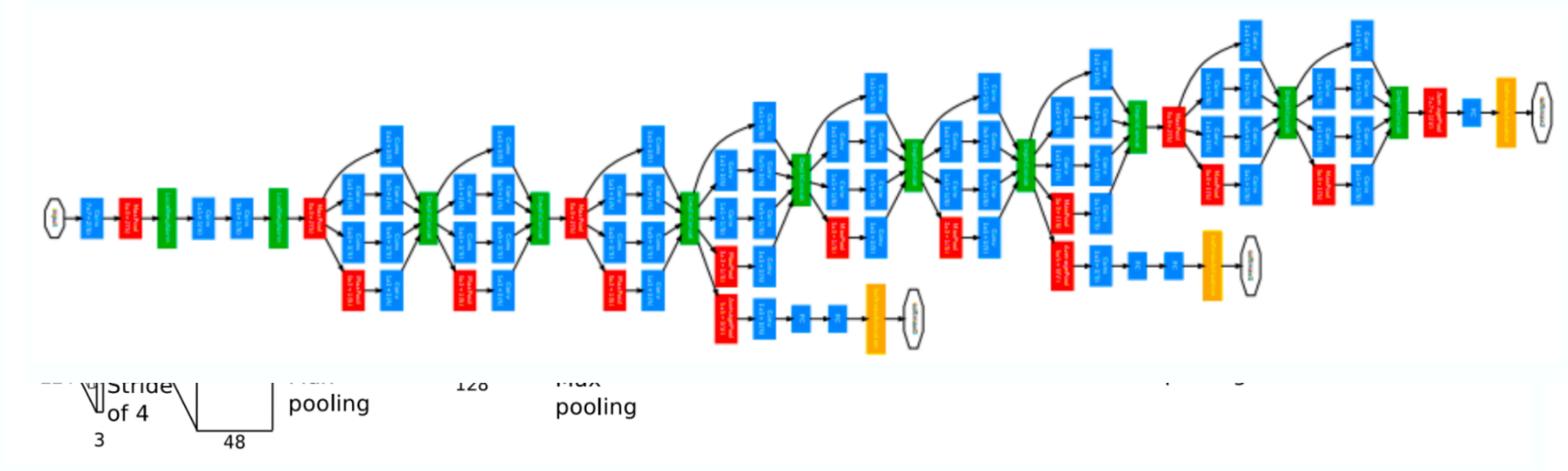
그동안의 다양한 image task 모델들



지난 수 년 동안 image 영역에서 feature 추출 위한 다양한 model을 사용할 수 있게 되었다

00. INTRODUCTION

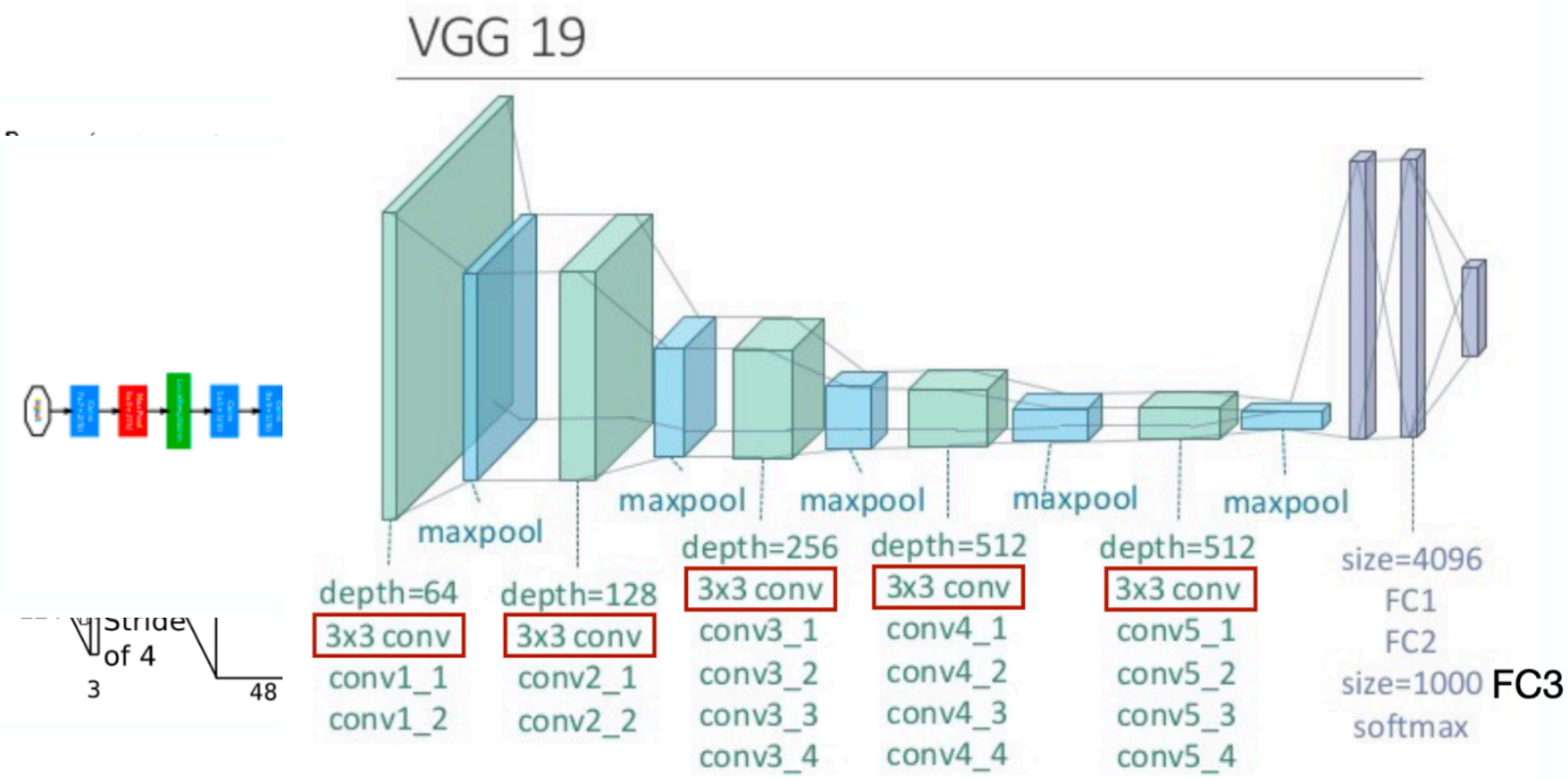
그동안의 다양한 image task 모델들



지난 수 년 동안 image 영역에서 feature 추출 위한 다양한 model을 사용할 수 있게 되었다

00. INTRODUCTION

그동안의 다양한 image task 모델들



지난 수 년 동안 image 영역에서 feature 추출 위한 다양한 model을 사용할 수 있게 되었다

00. INTRODUCTION

C3D

그러나 이 모델들은 image base deep feature -> video 영역에 적합하지 않다
Motion modeling에 적합하지 않기 때문

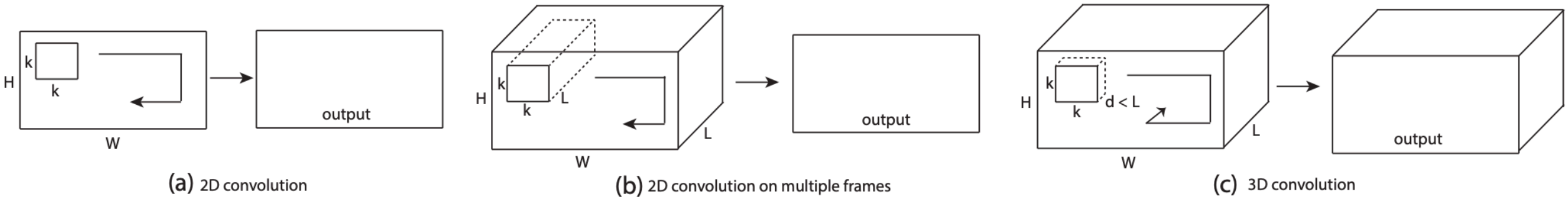
이 논문에서, 저자들은 **3D ConvNet**을 이용해 **spatiotemporal feature**를 학습

- **간단한 선형 분류로 학습한 feature**만으로도 다양한 video 분석 task에 좋은 결과
- **최초로 3D ConvNet을 제시한 것은 아님.** 그러나 다양한 task에서 좋은 결과

시공간적인 특징을 학습하기에 2D conv에 비해 3D conv가 더 적합
3D conv에서 3x3x3 conv kernel이 가장 좋은 성능을 보였다
C3D가 4개의 다른 벤치마크 중에서 가장 뛰어난 성능을 보였다

01. METHOD

2D conv vs 3D conv



(a) 2D convolution (b) 2D convolution on multiple frames (c) 3D convolution

(a) Image에 2D conv 적용, Output: image
(b) 복수 개의 image에 2D conv 적용, Output: image
(c) 3D conv 적용, Output: volume

2D conv는 매번 conv 작동될 때마다 '시각적인 정보' 를 잃는다

01. METHOD

시각적인 정보를 잃어?

Two-stream convolutional networks for action recognition in videos, *NIPS*, 2014

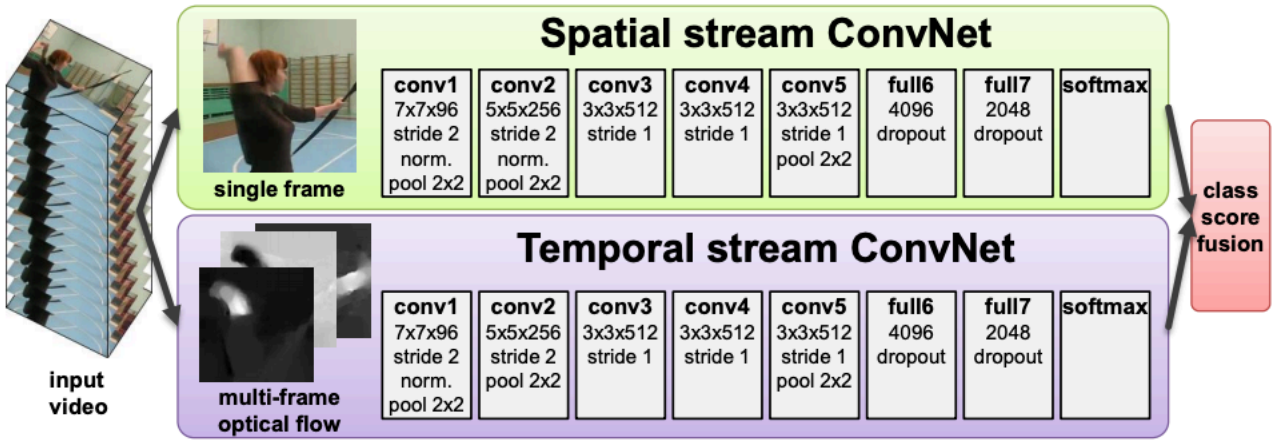


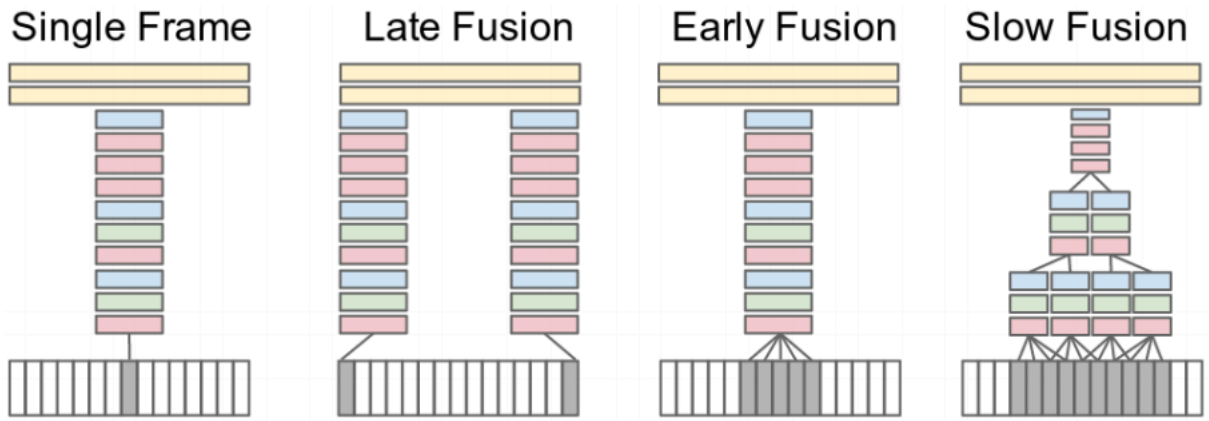
Figure 1: **Two-stream architecture for video classification.**

Temporal stream ConvNet은 복수의 frame을 input으로 가져도
처음 Conv layer 후에 시간적 정보가 완벽하게 소실된다

01. METHOD

시각적인 정보를 잃어?

Large-scale video classification with convolutional neural networks, *CVPR*, 2014



대부분의 network에서 input의 시간적 신호를 첫 Conv layer 이후 잃어버림
그러나 3D Conv를 사용한 Slow Fusion에서는 좋은 결과를 보임
=> 3D Conv가 그 이유일 것

01. METHOD

UCF101(중간 사이즈)을 학습하는 C3D

모든 Conv kernel은 size d : kernel temporal depth

Conv stride : 1

Pooling: max pooling / 2x2x2 kernel**(첫번째 pooling제외)** / stride 1 output 1/8

01. METHOD

UCF101(중간 사이즈)을 학습하는 C3D

논문의 주요 관심사 : Temporal information!

(1) Homogeneous temporal depth

d : 1,3,5,7 (depth-n 으로 명명)

(2) Varying temporal depth

Increase: 3-3-5-5-7

Decrease: 7-5-5-3-3

01. METHOD

UCF101(중간 사이즈)을 학습하는 C3D

논문의 주요 관심사 : Temporal information!

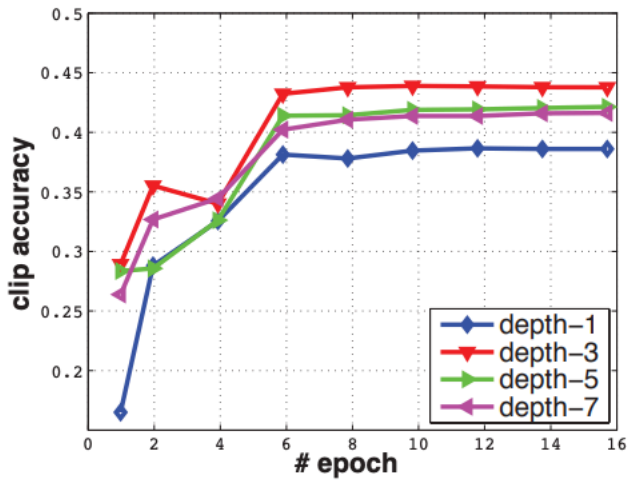
(1) Homogeneous temporal depth

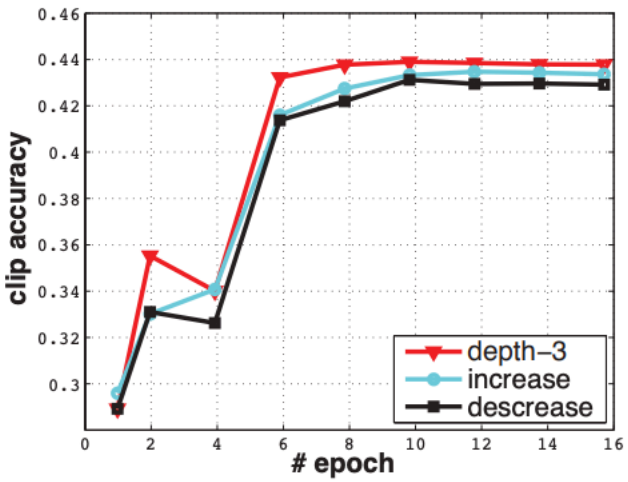
d : 1,3,5,7 (depth-n 으로 명명)

(2) Varying temporal depth

Increase: 3-3-5-5-7

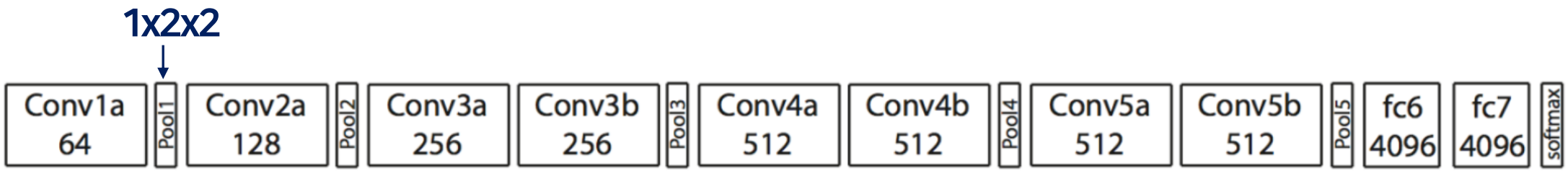
Decrease: 7-5-5-3-3





01. METHOD

Large size data를 학습하는 C3D



모든 Conv kernel은 size d : 3x3x3

Conv stride : 1x1x1

Pooling: max pooling / 2x2x2 kernel**(첫번째 pooling제외)** / stride 2x2x2 output 1/8

01. METHOD

Large size data를 학습하는 C3D

Sports-1M

Method	Number of Nets	Clip hit@1	Video hit@1	Video hit@5
DeepVideo's Single-Frame + Multires [18]	3 nets	42.4	60.0	78.5
DeepVideo's Slow Fusion [18]	1 net	41.9	60.9	80.2
Convolution pooling on 120-frame clips [29]	3 net	**70.8***	**72.4**	**90.8**
C3D (trained from scratch)	1 net	44.9	60.0	84.4
C3D (fine-tuned from I380K pre-trained model)	1 net	46.1	61.1	85.2

DeepVideo: 이전에 3D 적용했던 모델

01. METHOD

What does C3D learn?

C3D는 처음 몇 개의 frame에서 외관에 초점을 맞추다가 두드러지는 움직임을 추적한다

02. EXPERIMENT

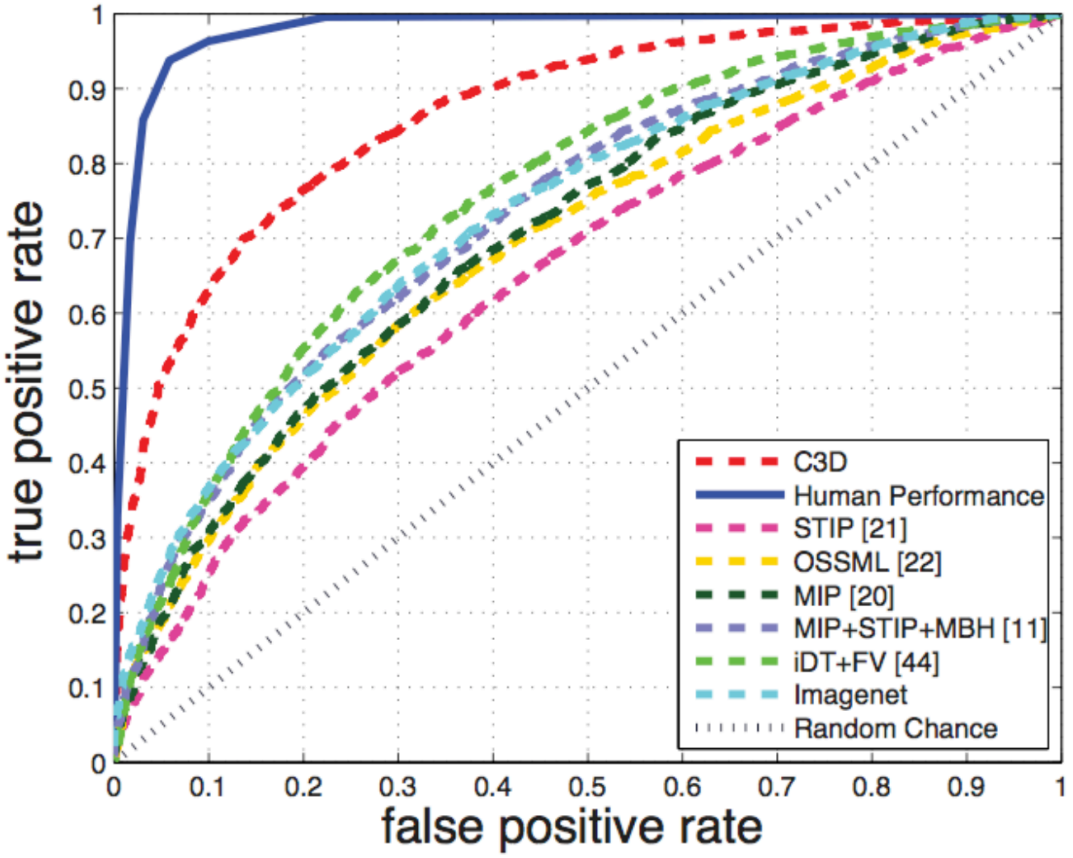
(1) Action Recognition

Method	Accuracy (%)
Imagenet + linear SVM	68.8
iDT w/ BoW + linear SVM	76.2
Deep networks [18]	65.4
Spatial stream network [36]	72.6
LRCN [6]	71.1
LSTM composite model [39]	75.8
C3D (1 net) + linear SVM	82.3
C3D (3 nets) + linear SVM	**85.2**
iDT w/ Fisher vector [31]	87.9
Temporal stream network [36]	83.7
Two-stream networks [36]	88.0
LRCN [6]	82.9
LSTM composite model [39]	84.3
Conv. pooling on long clips [29]	88.2
LSTM on long clips [29]	88.6
Multi-skip feature stacking [25]	89.1
C3D (3 nets) + iDT + linear SVM	**90.4**

Table 3. **Action recognition results on UCF101**. C3D compared with baselines and current state-of-the-art methods. Top: simple features with linear SVM; Middle: methods taking only RGB frames as inputs; Bottom: methods using multiple feature combinations.

02. EXPERIMENT

(2) Action Similarity Labeling



Method	Features	Model	Acc.	AUC
[21]	STIP	linear	60.9	65.3
[22]	STIP	metric	64.3	69.1
[20]	MIP	metric	65.5	71.9
[11]	MIP+STIP+MBH	metric	66.1	73.2
[45]	iDT+FV	metric	68.7	75.4
Baseline	Imagenet	linear	67.5	73.8
Ours	**C3D**	linear	**78.3**	**86.5**

02. EXPERIMENT

(3) Scene and object recognition

Dataset	[4]	[41]	[8]	[9]	Imagenet	C3D
Maryland	43.1	74.6	67.7	77.7	**87.7**	**87.7**
YUPENN	80.7	85.0	86.0	96.2	96.7	**98.1**

03. CONCLUSION

~

* 비디오 분석에서 3D conv가 시공간적 feature를 학습하는 것이 가능하다
* 3D conv에 가장 최적인 temporal kernel length를 찾으려 했다
* C3D가 외관과 motion info를 동시에 modeling할 수 있으며 2D conv를 능가
* 선형 classifier를 도입함으로써 최신 방법에 아주 가까운 성능을 낼 수 있다

04. FIN

QNA

Q & a