Introduction to Video Classification Using Deep Learning

Waseem Gharbieh

Outline

- Introduction
- Datasets
- Approaches
- Transfer Learning

Introduction

Why Video?

- Most general way to perceive the world
- Temporal reasoning

Why Not Video?

- Computation (Depends on the number of frames)
- Data

Why Not Video?

- Computation (Depends on the number of frames)
- Data 

IMAGENET

Why Not Video?

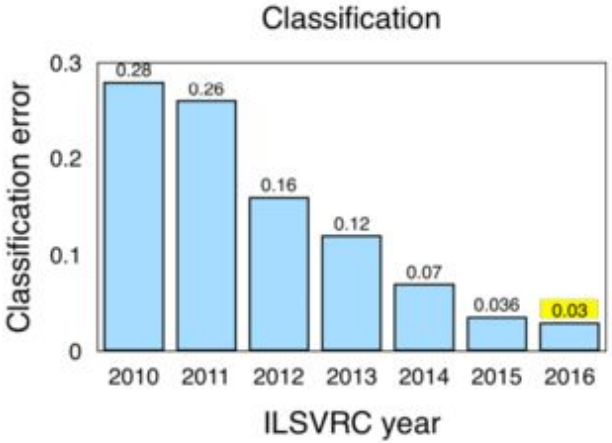
- Computation (Depends on the number of frames)
- Data 

IMAGENET
- Progress

Why Not Video?

- Computation (Depends on the number of frames)
- Data 

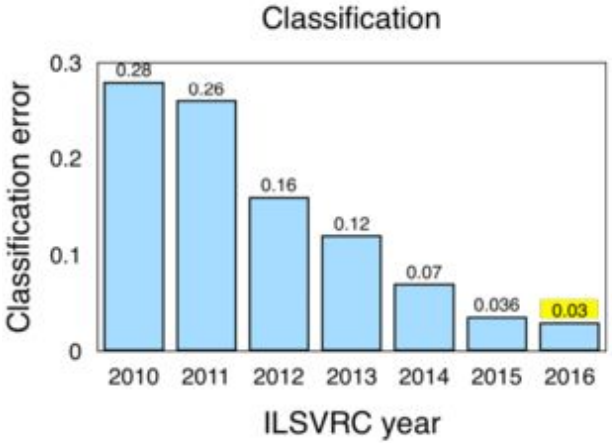
IMAGENET
- Progress



Why Not Video?

- Computation (Depends on the number of frames)
- Data 

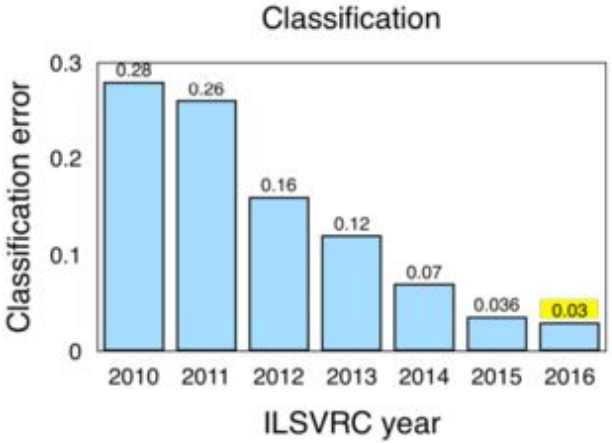
IMAGENET
- Progress
- A lot of applications do not require temporal reasoning



Why Not Video?

- Computation (Depends on the number of frames)
- Data 

IMAGENET
- Progress

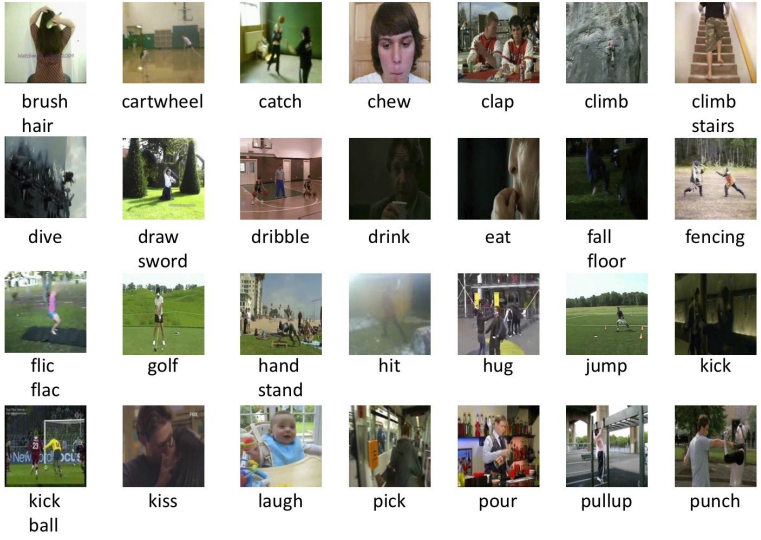


- A lot of applications do not require temporal reasoning
- Some applications that seem to require temporal reasoning can be solved using images

Datasets

HMDB-51

- A dataset of 6,766 videos collected from commercial movies as well as YouTube
- Contains 51 human motion classes
- Released November 2011 by Brown University



UCF-101

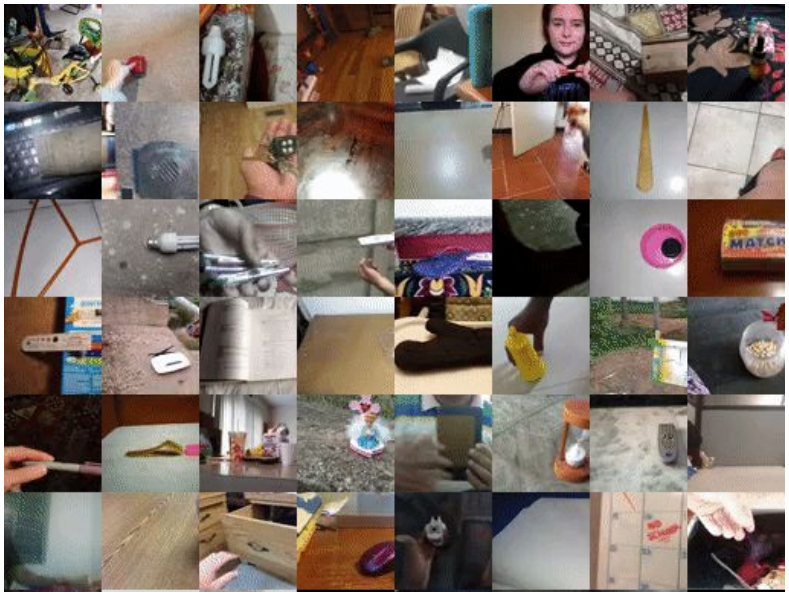
- A dataset of 13,320 videos collected from YouTube
- Contains 101 human actions classes
- Released November 2012 by University of Central Florida

Kinetics

- A dataset of 306,245 videos collected from YouTube
- Contains 400 human action classes
- Released May 2017 by Deepmind

Something-something

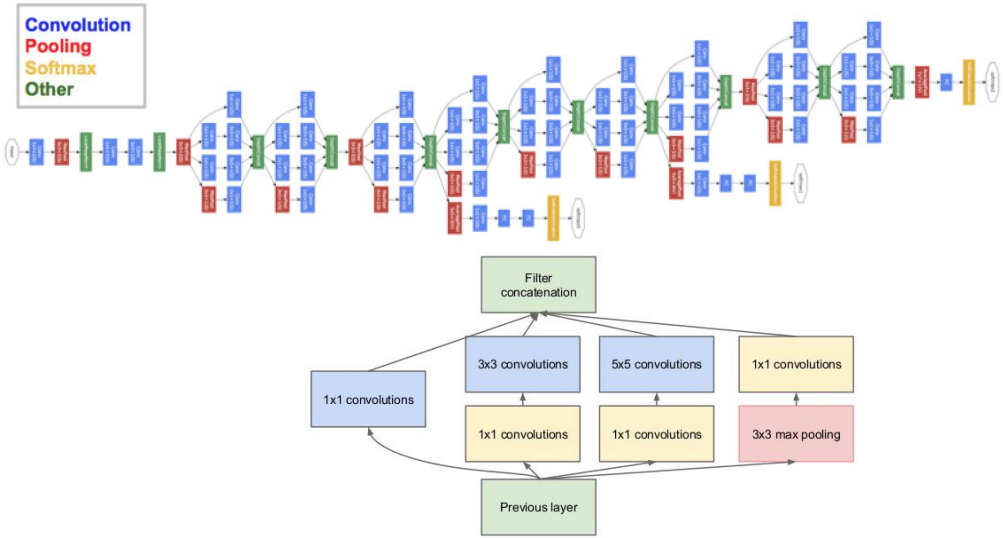
- A dataset of 108,499 videos collected using crowd workers
- Contains 174 common sense classes
- Released June 2017 by TwentyBN

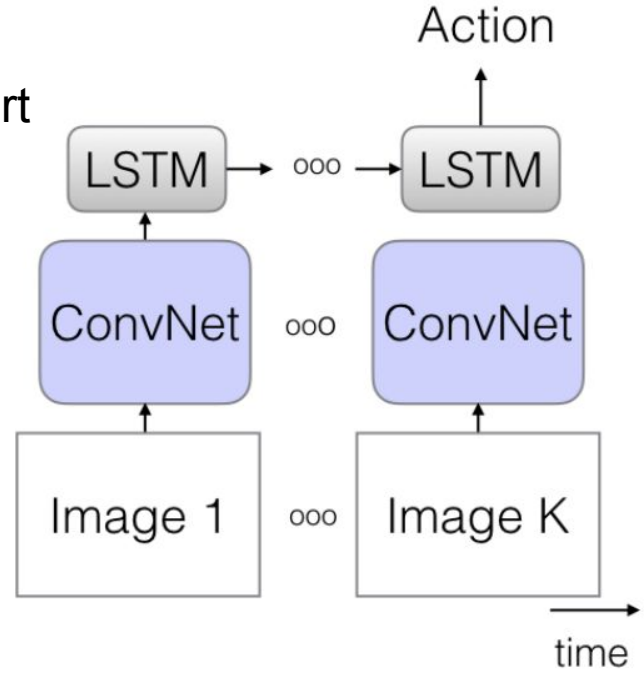


Approaches

LSTM

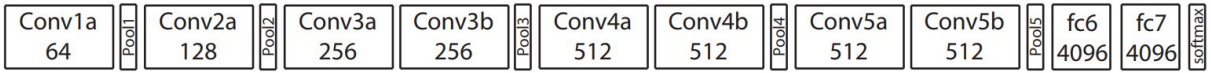
- ConvNet is Inception-V1
- LSTM aggregates video frames
- Only the output of the last frame is considered
- Input is 25 224x224 images sampled 5 frames apart

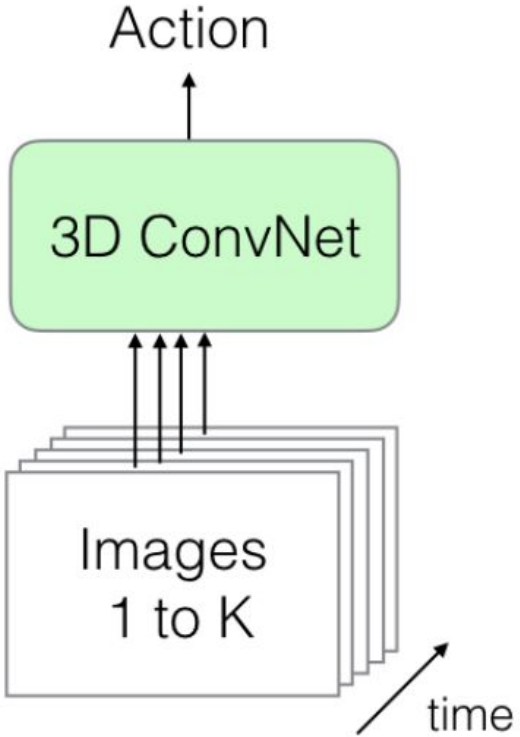




3D-ConvNet

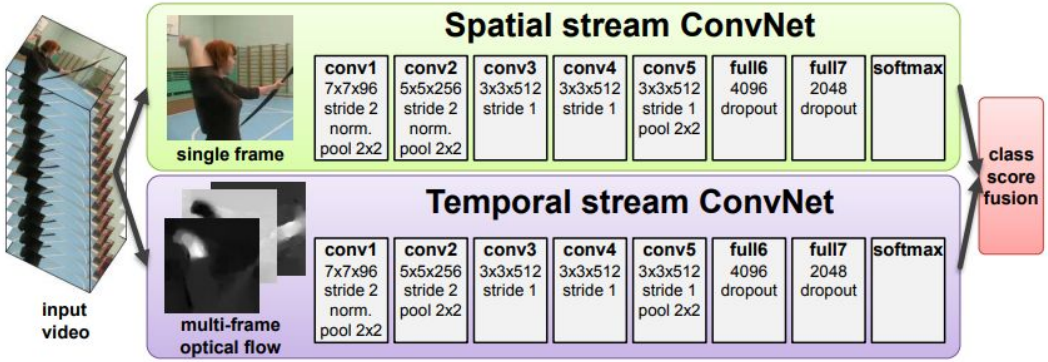
- 3D ConvNet based on work done by Du Tran, et al. 2015
- Input is 16 consecutive 112x112 images

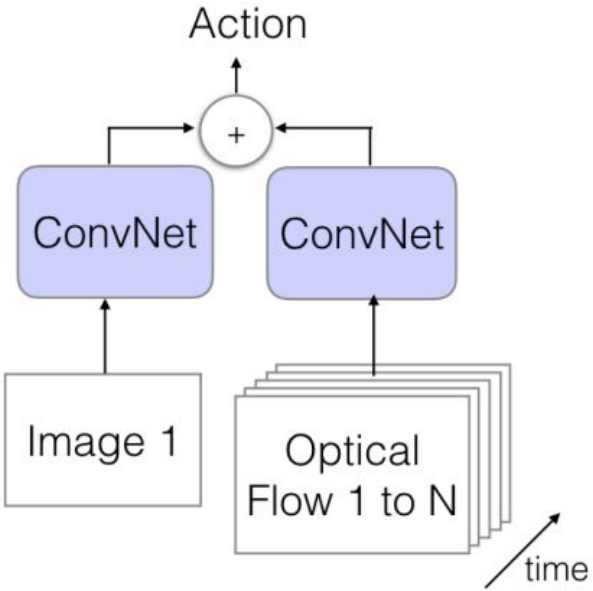




Two-Stream

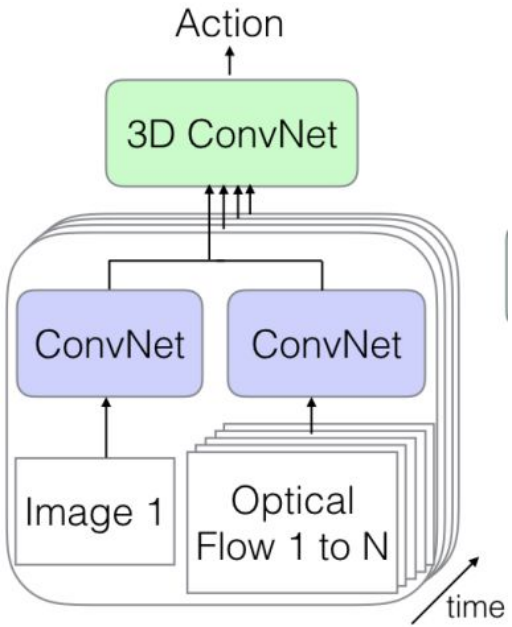
- ConvNet is based on work done by Simonyan and Zisserman, 2014 (ImageNet pretrained)
- Input is a 224x224 image + its 10 consecutive optical flow features





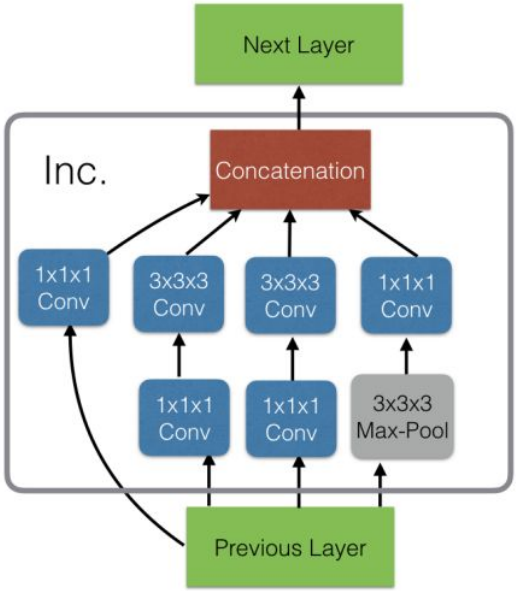
3D-Fused Two-Stream

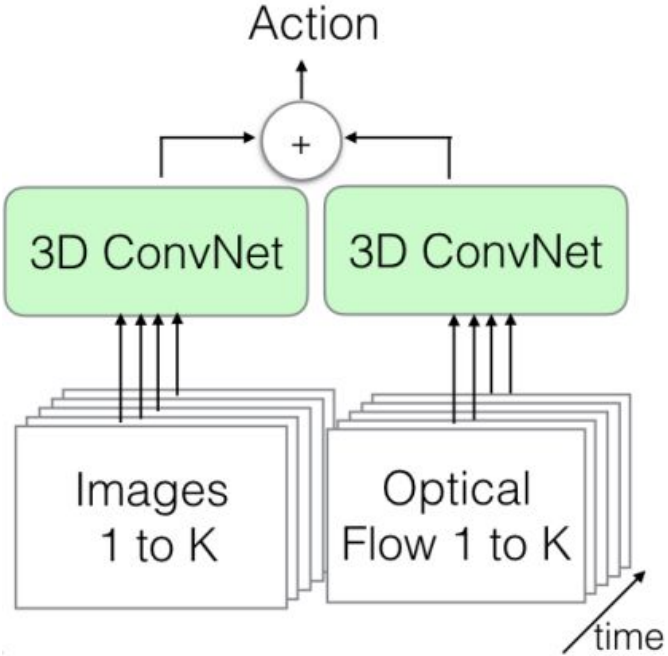
- ConvNet is Inception-V1
- Input is 5 224x224 images sampled 10 frames apart + 10 consecutive optical flow features for each image (so 50 in total)
- 3D ConvNet is 3x3x3 convolution with 512 channels followed by 3x3x3 maxpooling



Two-Stream 3D-ConvNet

- 3D ConvNet is inflated Inception-V1
- Input is 64 consecutive 224x224 images + their optical flow features





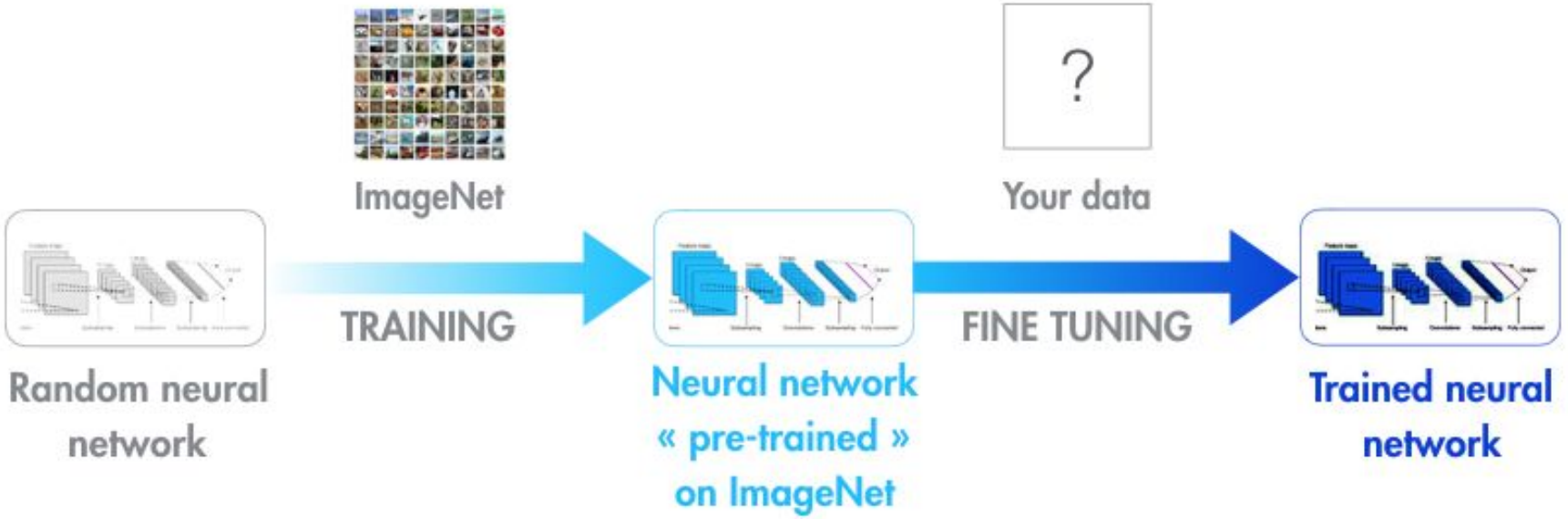
Results

- miniKinetics: Early version of Kinetics containing 213 classes with 120,000 videos

Architecture	UCF-101			HMDB-51			miniKinetics		
	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow	RGB	Flow	RGB + Flow
(a) LSTM	81.0	–	–	36.0	–	–	69.9	–	–
(b) 3D-ConvNet	51.6	–	–	24.3	–	–	60.0	–	–
(c) Two-Stream	83.6	85.6	91.2	43.2	56.3	58.3	70.1	58.4	72.9
(d) 3D-Fused	83.2	85.8	89.3	49.2	55.5	56.8	71.4	61.0	74.0
(e) Two-Stream I3D	**84.5**	**90.6**	**93.4**	**49.8**	**61.9**	**66.4**	**74.1**	**69.6**	**78.7**

Transfer Learning

Concept


ImageNet
?
Your data
TRAINING
FINE TUNING
Random neural network
Neural network « pre-trained » on ImageNet
Trained neural network

Transfer Learning From MiniKinetics

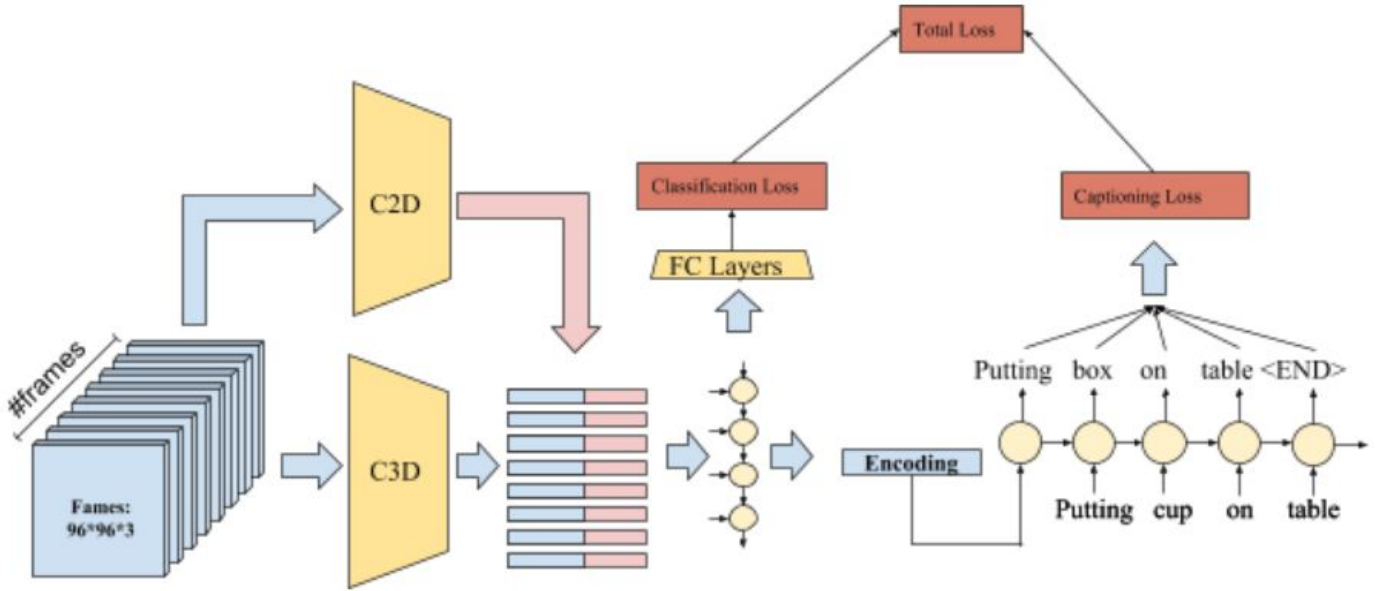
Architecture	UCF-101			HMDB-51		
	Original	Fixed	Full-FT	Original	Fixed	Full-FT
(a) LSTM	81.0 / 54.2	88.1 / 82.6	91.0 / 86.8	36.0 / 18.3	50.8 / 47.1	53.4 / 49.7
(b) 3D-ConvNet	– / 51.6	– / 76.0	– / 79.9	– / 24.3	– / 47.0	– / 49.4
(c) Two-Stream	91.2 / 83.6	93.9 / 93.3	94.2 / 93.8	58.3 / 47.1	66.6 / 65.9	66.6 / 64.3
(d) 3D-Fused	89.3 / 69.5	94.3 / 89.8	94.2 / 91.5	56.8 / 37.3	69.9 / 64.6	71.0 / 66.5
(e) Two-Stream I3D	93.4 / 88.8	97.7 / 97.4	98.0 / 97.6	66.4 / 62.2	79.7 / 78.6	81.2 / 81.3

Effect Of Dataset Size on Transfer Learning

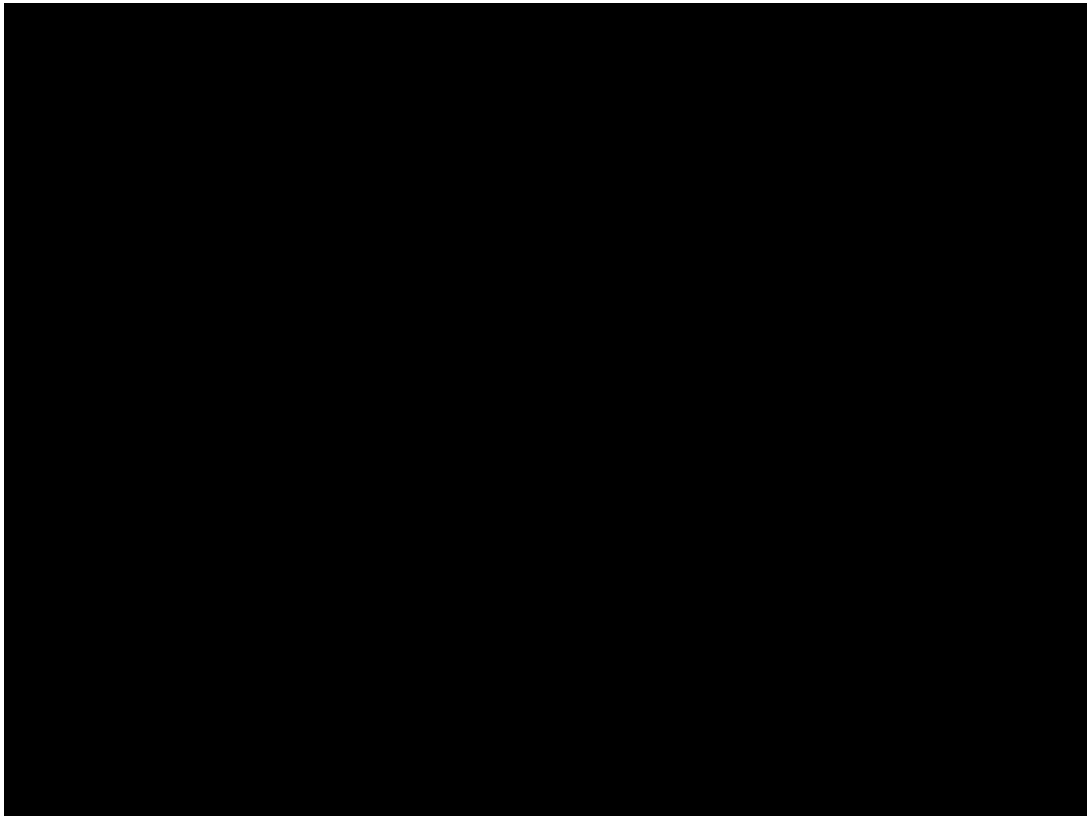
Model	UCF-101	HMDB-51
RGB-I3D, miniKinetics pre-training	91.8	66.4
RGB-I3D, Kinetics pre-training	95.4	74.5
Flow-I3D, miniKinetics pre-training	94.7	72.4
Flow-I3D, Kinetics pre-training	95.4	74.6
Two-Stream I3D, miniKinetics pre-training	96.9	76.3
Two-Stream I3D, Kinetics pre-training	**97.9**	**80.2**

Transfer Learning From One Task to Another

- Encoder Pretrained to perform classification
- Model was then tuned with 0.1 * Classification Loss + 0.9 * Captioning Loss



Demo



Thank You!

References

1. H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition, ICCV, 2011.
2. K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012
3. Kay, Will, et al. "The kinetics human action video dataset." arXiv preprint arXiv:1705.06950 (2017).
4. R. Goyal, et al. The” something something” video database for learning and evaluating visual common sense, ICCV, 2017.
5. Carreira, Joao, and Andrew Zisserman. "Quo vadis, action recognition? a new model and the kinetics dataset." CVPR, 2017.
6. Szegedy, Christian, et al. "Going deeper with convolutions." CVPR, 2015.
7. Tran, Du, et al. "Learning spatiotemporal features with 3d convolutional networks." ICCV, 2015.
8. Simonyan, Karen, and Andrew Zisserman. "Two-stream convolutional networks for action recognition in videos." NIPS. 2014.