Day 16

Speaker: Prof. Andrew Zisserman, University of Oxford, UK

Title: Recognizing Human Actions in videos

DeepMind

UNIVERSITY OF OXFORD

Recognizing Human Actions in Video

Andrew Zisserman

Video Understanding

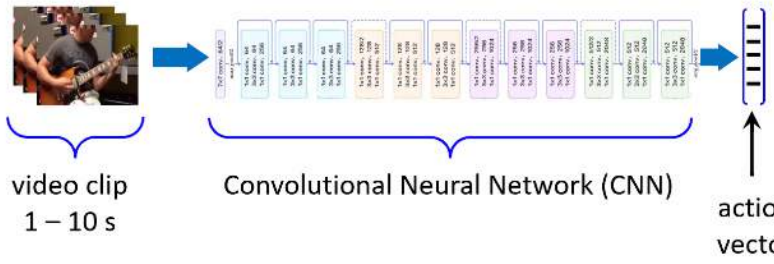
- What is in the video?
  - objects, animals, people …
- Where is it?
  - 3D scene spatial layout
  - object shape
  - human pose …
- What is happening?
  - **actions**
  - activities …

Objective: how to learn to recognize human actions without explicit supervision?

Representing human actions

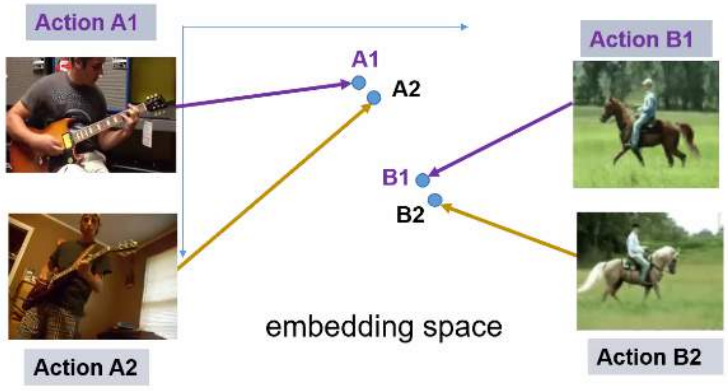
Learn a video clip embedding for action recognition

Map from video clip to a vector



Use action vector for classification, localization, retrieval …

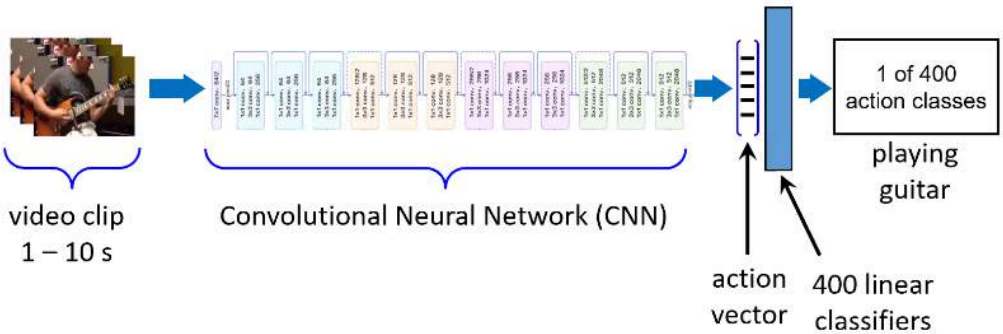
Objective of the action embedding



Action vectors of the same (semantic) action should be

- close, and
- distinct from those of different actions

Example: Using the action vector for linear classification



Outline

Part I: Learning to represent human actions using strong (explicit) supervision

- The Kinetics human action dataset

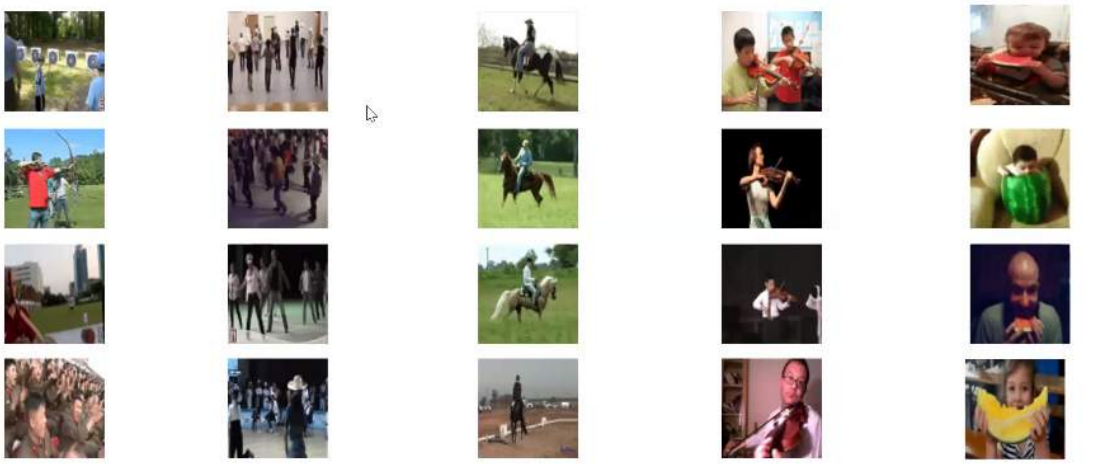
How can we learn to represent human actions without explicit supervision?

Part II: Using multi-modal (audio-visual) self-supervised learning

Part III: Using weak supervision from videos with text

The Kinetics Human Action Dataset

The Kinetics Human Action Video Dataset



Motivation

Objective: A large scale human action classification video dataset

- An ImageNet for human action recognition
  - Trimmed videos
  - Actions performed by humans
  - Action classification
- Large enough to use for architecture design and comparison
- Large enough to pre-train networks for other tasks, e.g.
  - Temporal action localization in untrimmed videos

Kinetics datasets overview

- Stats:

	Year	Action classes	Clips per class	Total
Kinetics-400	2017	400	400-1000	300k
Kinetics-600	2018	600	600-1000	500k
Kinetics-700	2019	700	600-1000	650k

- 10s clips
- Every clip from a different YouTube video:
  - huge variety in people, viewpoint, scenes, execution ...

- *The Kinetics Human Action Video Dataset*. Kay, Carreira, Simonyan, Zhang, Hillier, Vijayanarasimhan, Viola, Green, Back, Natsev, Suleyman, Zisserman, arXiv 2017
- *A Short Note about Kinetics-600*. Carreira, Noland, Banki-Horvath, Hillier, Zisserman, arXiv 2018
- *A Short Note on the Kinetics-700 Human Action Dataset*. Carreira, Noland, Hillier, Zisserman, arXiv 2019

Datasets available for download from: cvdfoundation

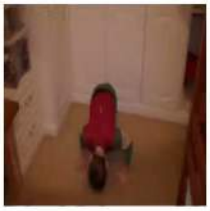
Action Classes

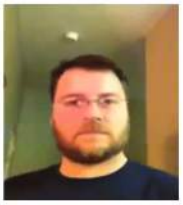
Person Actions (Singular)
e.g. waving, blinking, running, jumping

Person-Person Actions
e.g. hugging, kissing, shaking hands

Person-Object Actions
e.g. opening door, mowing lawn, washing dishes

Person Actions (Singular)

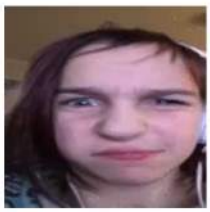


Head stand

Shaking Head

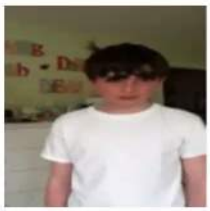


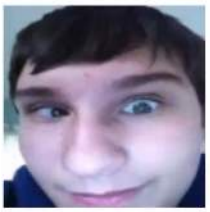

More Person Actions (Singular) - faces




Raising eyebrows

Crossing eyes






Person-Person Actions






Shaking Hands


Massaging Back

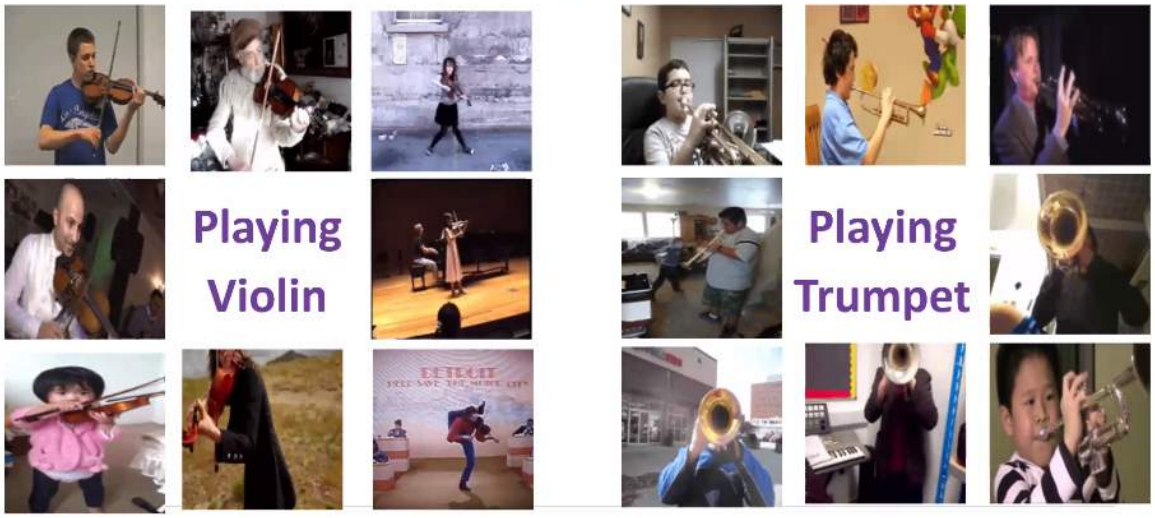






Google DeepMind

Person-Object Actions



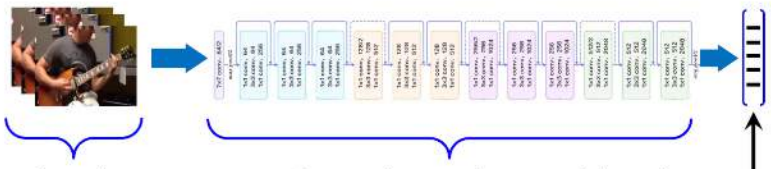
Google DeepMind

More random stuff many people do

Objective – human action classification

Learn a video clip embedding for action recognition

Map from video clip to a vector



Strong supervision: train network to classify actions on the Kinetics dataset

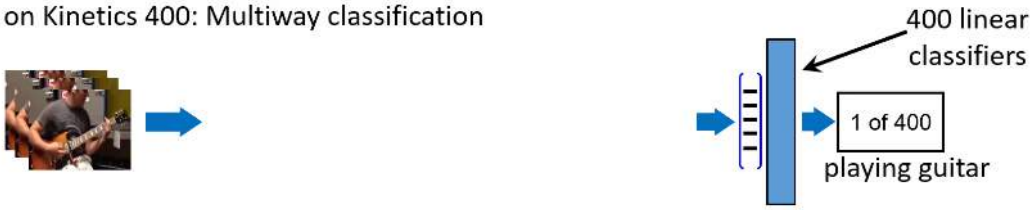
Use trained network for `downstream' tasks

1. Train on Kinetics 400: Multiway classification

2. Use network for new task, e.g. classification on a new video dataset: UCF-101

Use trained network for `downstream' tasks

1. Train on Kinetics 400: Multiway classification



2. Use network for new task, e.g. classification on a new video dataset: UCF-101

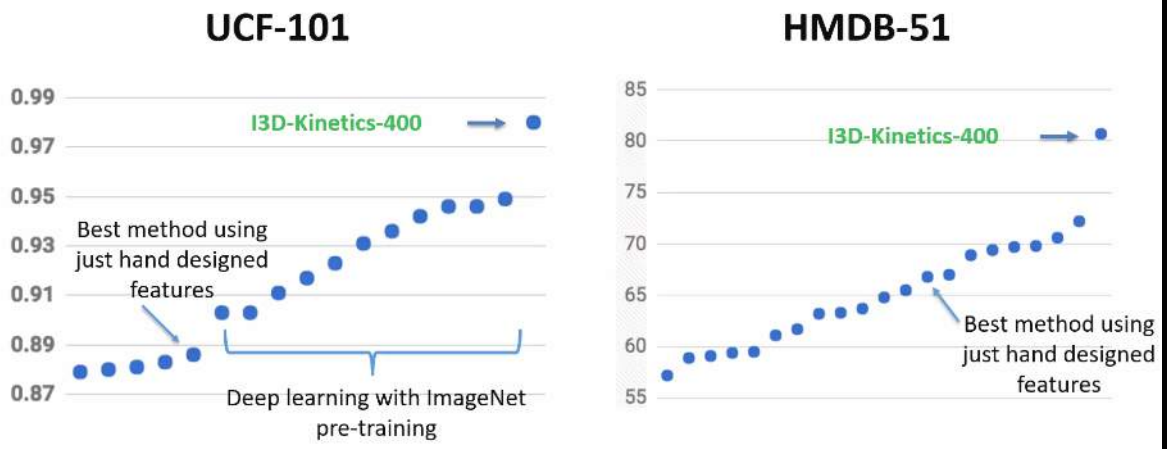
UCF-101 and HMDB-51 video human action datasets

UCF-101

HMDB-51

Dataset	Year	Actions	Clips	Total	Videos
HMDB-51 [15]	2011	51	min 102	6,766	3,312
UCF-101 [20]	2012	101	min 101	13,320	2,500

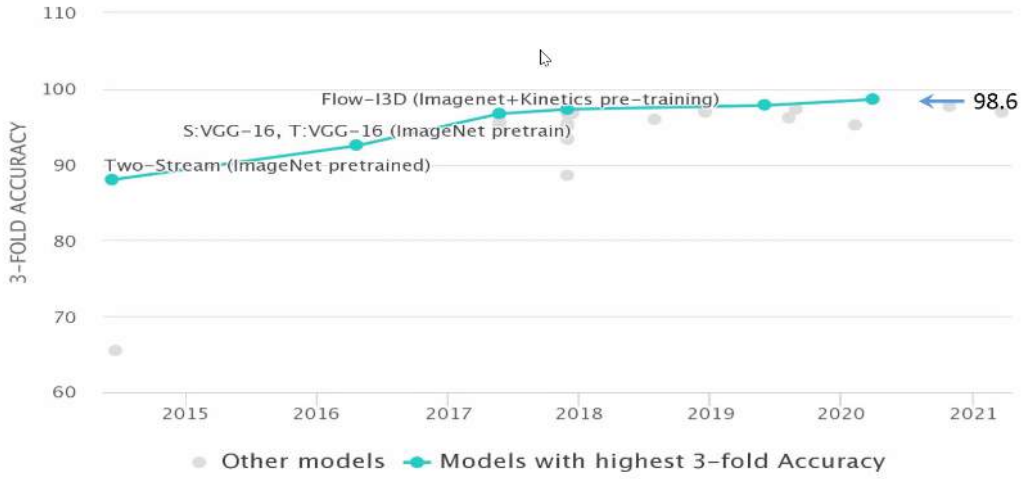
I3D-Kinetics-400 transfer performance (two stream, flow+RGB)



Compilation of results from actionrecognition.net

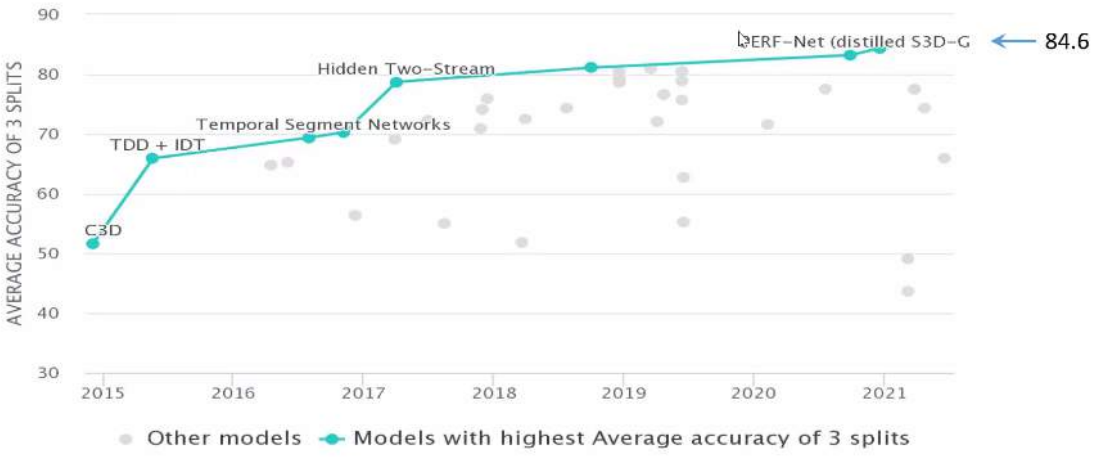
2019: HMDB-51 at 82.48

Performance on UCF-101



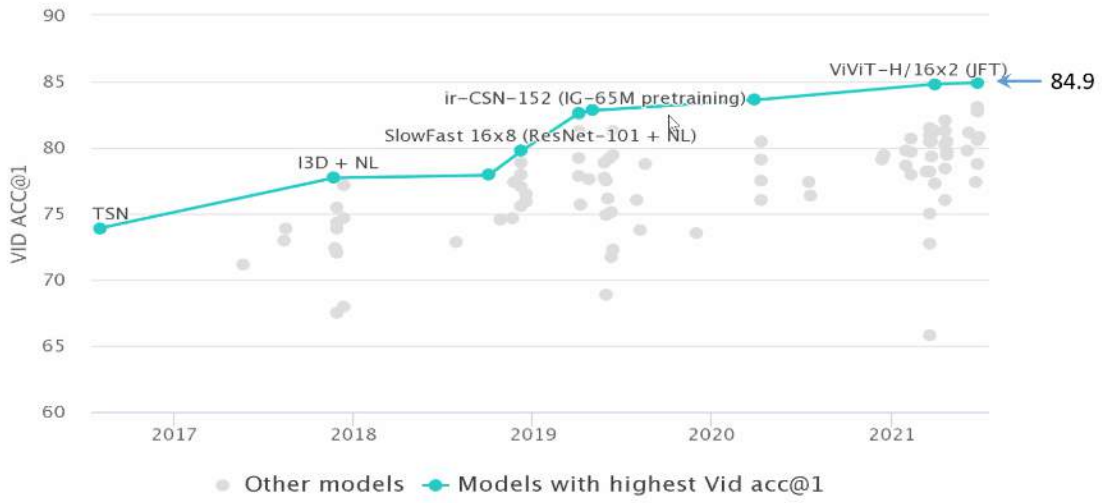
Compilation of results from `papers with code' website paperswithcode.com

Performance on HMDB-51



Compilation of results from `papers with code' website paperswithcode.com

Performance on Kinetics-400 val



Compilation of results from `papers with code' website paperswithcode.com

Lessons learnt over four years of strong supervision ....

Positives:

- Many new video architectures designed:
  - Spatio-temporal 3D CNNs: I3D, S3D, R(2+1)D, SlowFast, TSM, ...
  - Video transformers: Timesformer, ViVit, V-Swin, MViT, ...
- Pre-training on Kinetics improves transfer performance on downstream tasks

Negatives:

- Other datasets more suitable for ego-centric (1st person) video, e.g. EPIC-Kitchens
- Many classes in Kinetics do not require temporal information to be classified

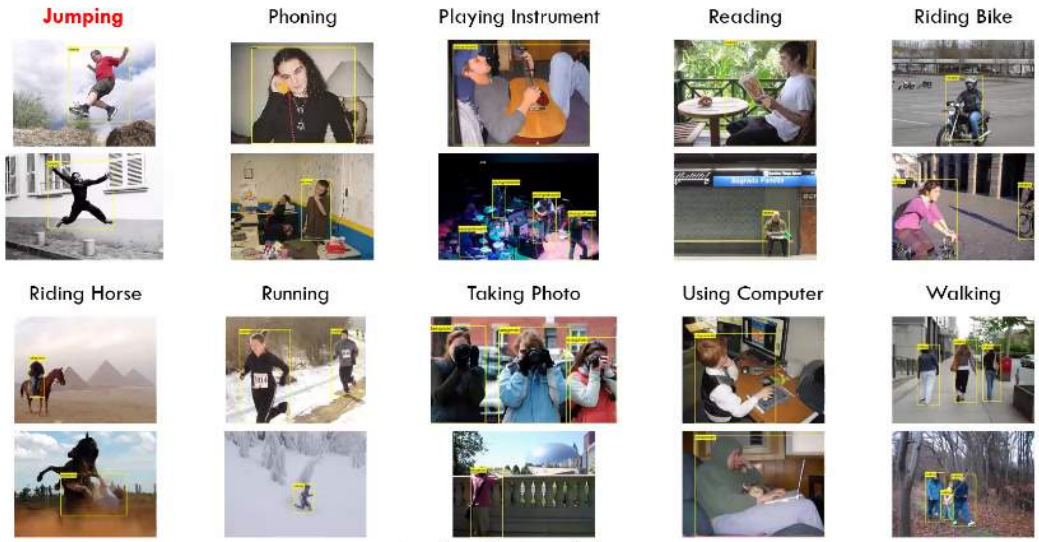
Video

A temporal sequence of frames

What is required to recognize the action?

- a single frame?
- a bag of frames (unordered)?
- an ordered sequence of frames?
- ...

Action Classification on Static Frames



PASCAL VOC Action Classification Challenge

Some actions require motion for classification

- Sitting down/standing up; closing/opening something
- Different dance styles ….

Dancing Macarena | Dancing Charleston | Zumba

Lessons learnt over four years of strong supervision ….

- Many classes in Kinetics do not require temporal information to be classified
  - Motion adds a few percent over frame level (2D image) classification
  - Other datasets such as Something-Something-V2 & Charades, are more suitable for evaluating temporal development
  - For discussion on this see: "Only Time Can Tell: Discovering Temporal Data for Temporal Modeling" Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, Lorenzo Torresani, 2019

Part II

Learning Human Action Representations using Audio-Visual Self-Supervision

Why Self-Supervision?

Rather than "strong supervision", e.g. on Kinetics:

1. Expense of producing a new dataset for each new task
2. Some areas are supervision-starved, e.g. medical data, where it is hard to obtain annotation
3. Untapped/availability of vast numbers of unlabelled images/videos
   - Facebook: one billion images uploaded per day
   - 300 hours of video are uploaded to YouTube every minute

Self-Supervised Learning

The Scientist in the Crib: What Early Learning Tells Us About the Mind
by Alison Gopnik, Andrew N. Meltzoff and Patricia K. Kuhl

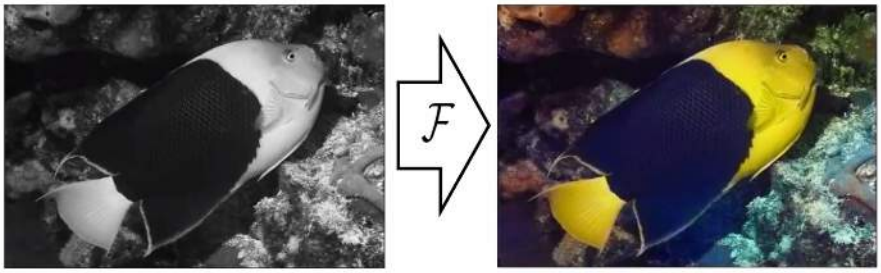
The Development of Embodied Cognition: Six Lessons from Babies
by Linda Smith and Michael Gasser

What is Self-Supervision?

- A form of unsupervised learning where the data provides the **supervision**
- In general, withhold some part of the data, and task the network with predicting it
- The task defines a proxy loss, and the network is forced to learn what we really care about, e.g. a semantic representation, in order to solve it
- Inspiration from NLP learning methods such as word2vec

Image example: colourization

Train network to predict pixel colour from a monochrome input



Grayscale image: *L* channel

$\mathbf{X} \in \mathbb{R}^{H \times W \times 1}$

Concatenate (L,ab)

$(\mathbf{X}, \hat{\mathbf{Y}})$

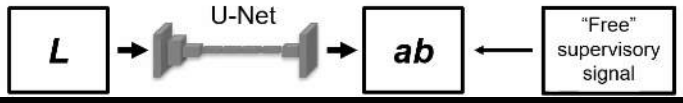


Image example: colourization

Train network to predict pixel colour from a monochrome input

"Colorful Image Colorization", Richard Zhang, Phillip Isola, Alexei A. Efros, ECCV 2016

Self-supervised learning for images

1. Pre-training: self-supervised network training on ImageNet using a proxy task

2. Supervised training of network for downstream task either by linear probe or initialization for fine tuning

- Recent proxy tasks:
  - Instance discrimination using a contrastive loss: SimCLR, MoCo, NNCLR
  - Regression (teacher-student) loss: BYOL, SimSiam
  - Clustering: DeepCluster, SwAV, SeLA

- Surpass performance of strong supervision (training with class labels) on ImageNet classification and on a number of downstream tasks, e.g.
  - PASCAL VOC segmentation & object detection,
  - NYU depth, …

Self-supervised learning for video

Video, beyond images, naturally …

- extends and develops sequentially in time,
- has motion (optical flow stream),
- has multiple modalities (audio stream)

Video inherits all the image proxy tasks (e.g. instance discrimination, clustering) at the frame level, but also has proxy tasks particular to the domain

Audio-Visual Co-supervision

Sound and frames are:

- Semantically consistent
- Synchronized

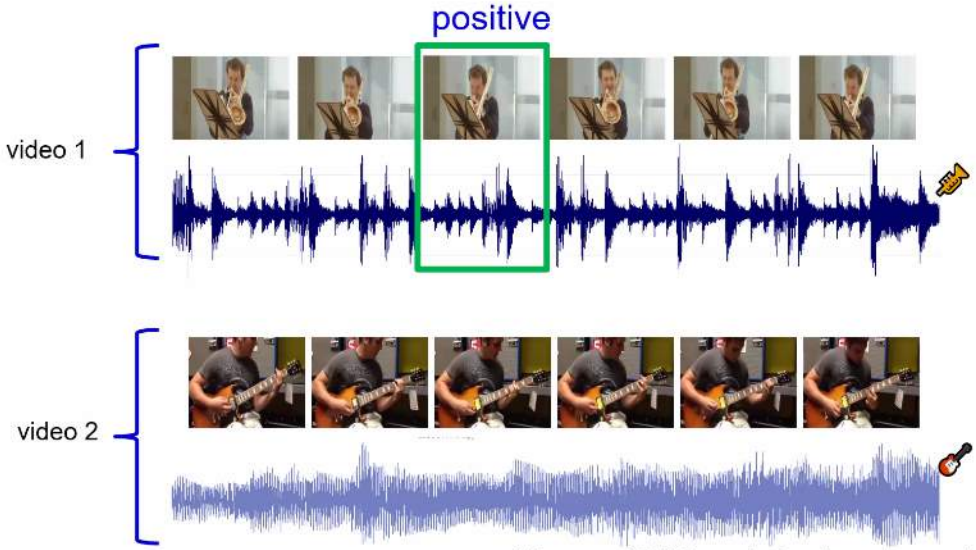
Audio-Visual Co-supervision

Objective: use vision and sound to learn from each other

- Sound and frames are (i) semantically consistent, and (ii) synchronized

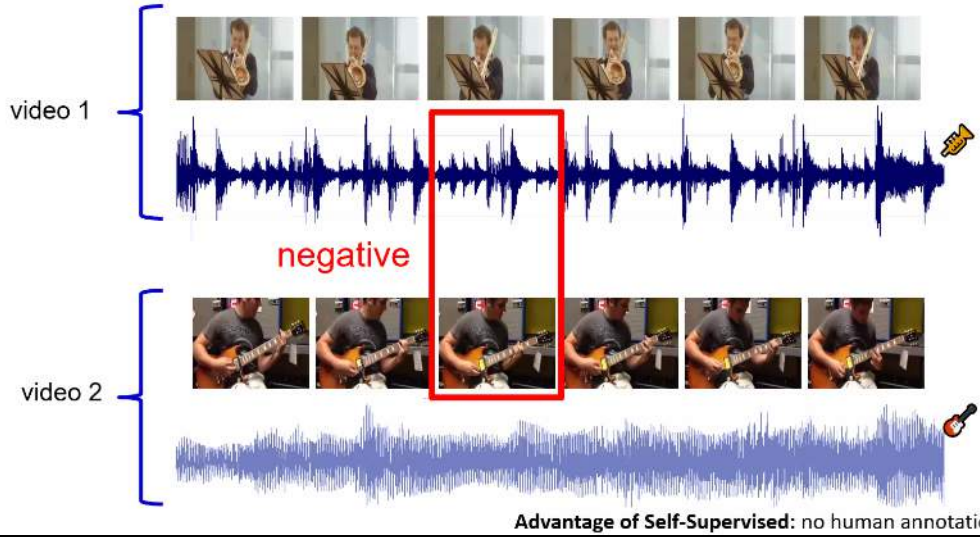
- Two types of proxy task:
  1. Predict audio-visual **correspondence**
  2. Predict audio-visual **synchronization**

Obtaining positives and negatives



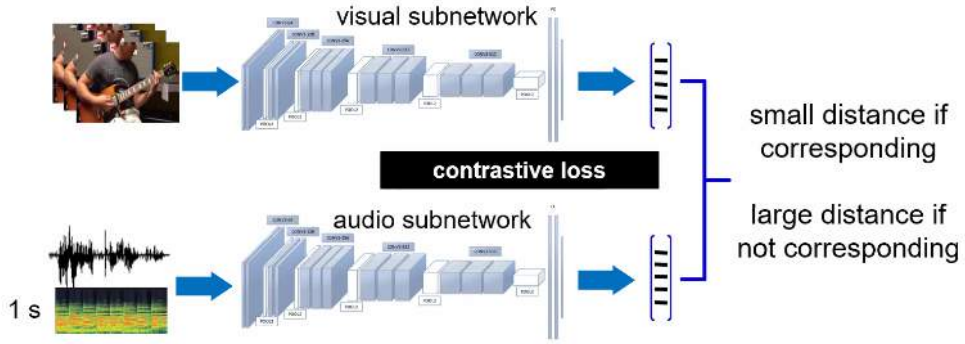
Advantage of Self-Supervised: no human annotation required!

Obtaining positives and negatives



Advantage of Self-Supervised: no human annotation required!

Learning from Correspondence



The network is trained from scratch with contrastive loss to:

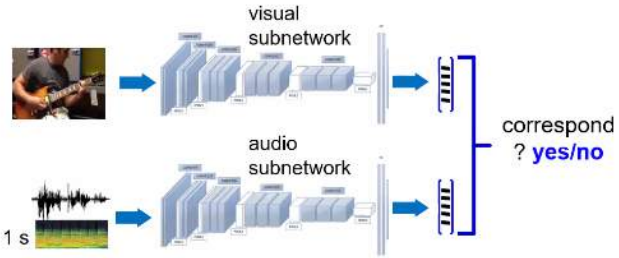
- Minimise distance between positive pairs (same video clip)
- Maximise distance between negative pairs (different video clips)

on hundreds of hours of video

Use audio and visual features

What can be learnt by watching and listening to videos?

- Good representations
  - Visual features
  - Audio features
- Intra- and cross-modal retrieval
  - Aligned audio and visual embeddings
- "What is making the sound?"
  - Learn to localize objects that sound



"Objects that Sound", Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Multi-modal Self-Supervision from Generalized Data Transformations

Mandela Patrick, Yuki M. Asano, Polina Kuznetsova, Ruth Fong, João F. Henriques, Geoffrey Zweig, Andrea Vedaldi, ICCV 2021

(a) **Models pretrained on Kinetics-400.**

Method	Architecture	Top-1 Acc% HMDB	UCF
Supervised	R(2+1)D-18	70.4	95.0
CPD [62]† *	3D-Resnet50	57.7	88.7
RotNet3D [49]	3D-ResNet18	33.7	62.9
DPC [38]	3D-ResNet34	35.7	75.7
Multisensory [77]	3D-ResNet18	-	82.1
XDC [5]	R(2+1)D-18	47.1	84.2
AVSlow-Fast [107]	AVSlowFast	54.6	87.0
AVTS [55]	MC3-18	56.9	85.8
AVID [71]	custom R(2+1)D	**60.8**	87.5
GDT (ours)	R(2+1)D-18	60.0	**89.3**

(b) **Models pretrained on other datasets.**

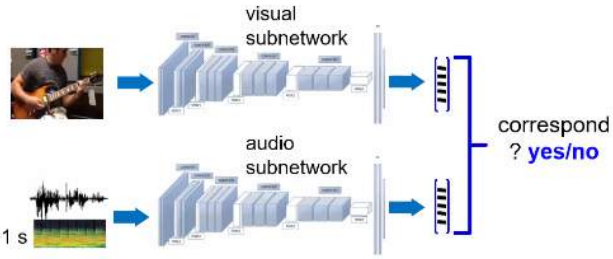
Method	Dataset	Top-1 Acc% HMDB	UCF
Supervised	Kinetics-400	70.4	95.0
GDT (ours)	VGGSound (170K)	62.1	89.4
XDC [5]	AudioSet (1.8M)	61.0	91.2
AVTS [55]	AudioSet (1.8M)	61.6	89.0
AVID [71]	Audioset (1.8M)	64.7	91.5
GDT (ours)	AudioSet (1.8M)	**66.1**	**92.5**
MIL-NCE [68]*	HowTo100M	61.0	91.3
ELo [83]	Youtube-2M	67.4	93.8
XDC [5]	IG65M	67.4	94.2
GDT (ours)	IG65M	**72.8**	**95.2**

Audio-visual self-supervision using a Noise Contrastive Estimation (NCE) loss

Use audio and visual features

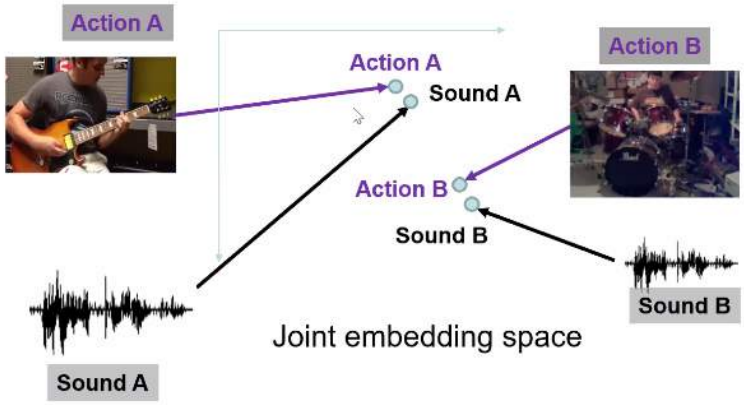
What can be learnt by watching and listening to videos?

- Good representations
  - Visual features
  - Audio features
- Intra- and cross-modal retrieval
  - Aligned audio and visual embeddings
- "What is making the sound?"
  - Learn to localize objects that sound



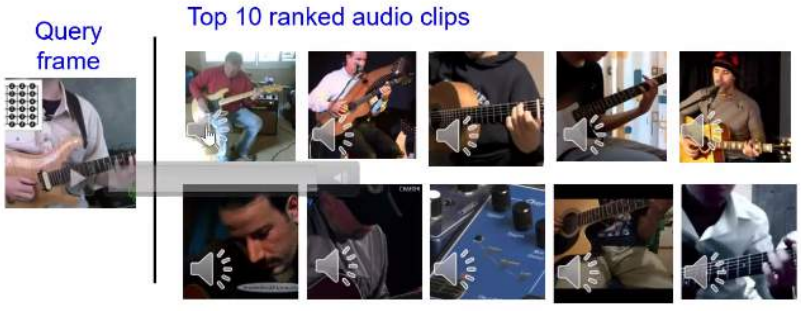
"Objects that Sound", Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Joint audio-video embedding



Query on image, retrieve audio

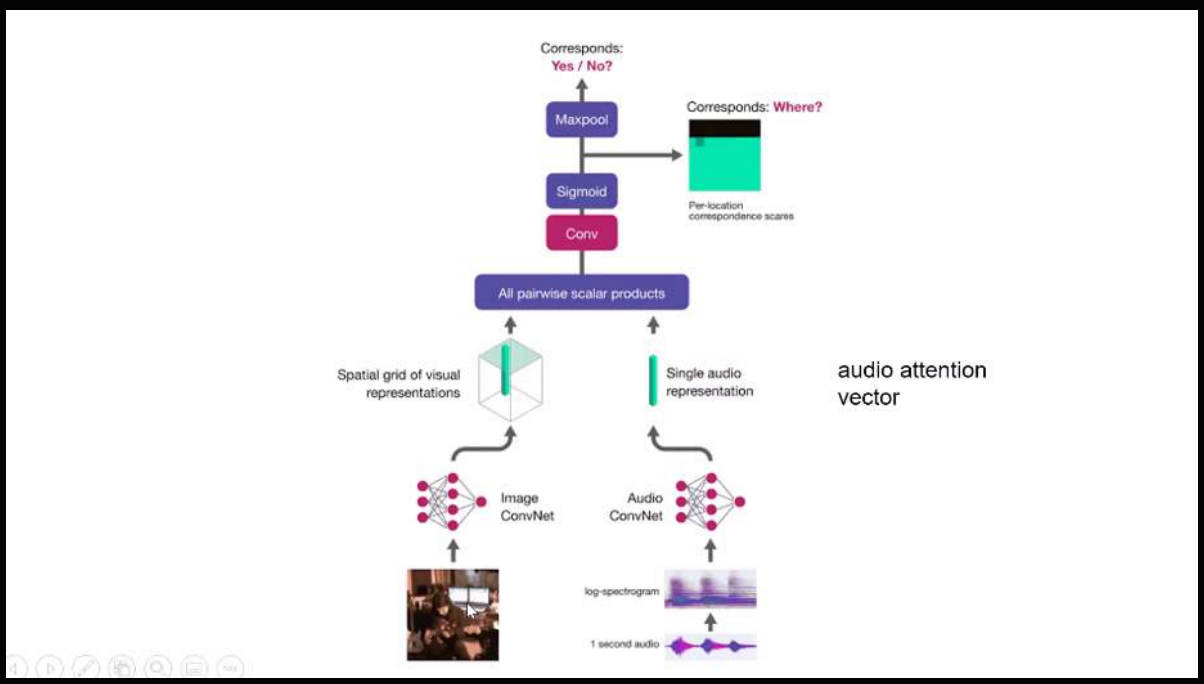
Search in 200k video clips of AudioSet



"Objects that Sound", Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Query on audio, retrieve video

Search in 200k video clips of AudioSet



Objects that Sound: object localization

Input: audio and video frame

- Frame by frame
- No motion information
- No memory
- No smoothing

frame | frame+heatmap | heatmap

"Objects that Sound", Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Objects that Sound: object localization

Input: audio and video frame

frame | frame+heatmap | heatmap

- Frame by frame
- No motion information
- No memory
- No smoothing

"Objects that Sound", Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Objects that Sound: object localization

Input: audio and video frame

frame | frame+heatmap | heatmap

- Frame by frame
- No motion information
- No memory
- No smoothing

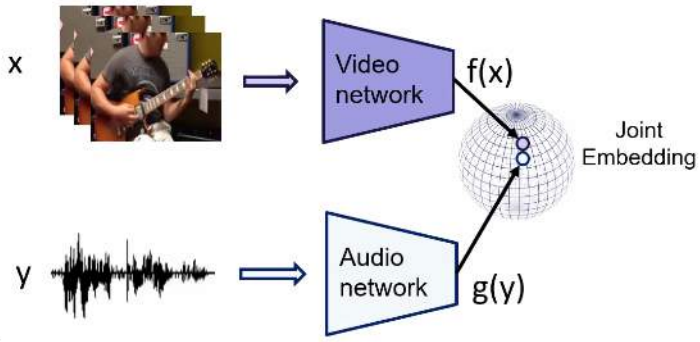
"Objects that Sound", Arandjelović and Zisserman, ICCV 2017 & ECCV 2018

Summary Point: learn a joint audio-video embedding

Architecture: Dual Encoder

Separate networks for video and audio encoding

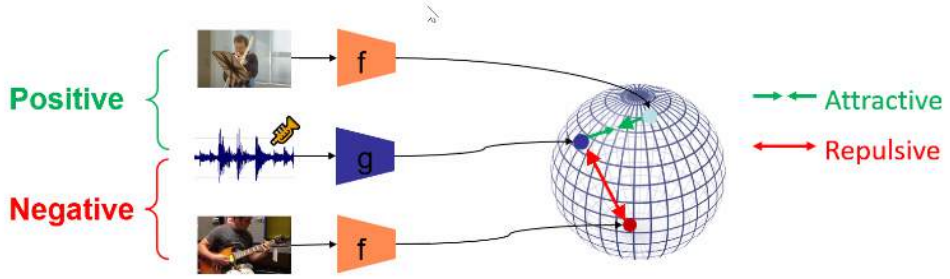
Score similarity, e.g. by $f(x)^\top g(y)$



Multi-Modal Contrastive Learning

Goal: Learn a joint multimodal space where embeddings of modalities that are semantically similar are close, and far otherwise

How? Using a ***contrastive*** approach.



This is an "old" idea: DeVise, *Frome et al.*, ***NeurIPS2013*** and WSABIE, *Weston et al.* ***IJCAI 2011.***

Multi-Modal Contrastive Objective

Noise Contrastive Estimation (NCE) loss

$$\max_{f,g} \log \left(\frac{\exp(f(x)^\top g(y))}{\exp(f(x)^\top g(y)) + \sum_{(x',y')\in\mathcal{N}} \exp(f(x')^\top g(y'))} \right)$$

Video

Audio

Video network

Audio network

Positive Video-Audio pair

Negative Video-Audio pairs

Softmax over positive and multiple negatives

Contrastive learning: recap and pros/cons

Summary: use a contrastive loss to match embeddings across modalities!

Pros

- **Easy to understand**
- **As a by-product:** learn a joint embedding space where cross modal retrieval is possible

Cons

- **Negative requirements:** computationally intensive
- **Negative choice matters:** risk too easy negatives as well as too hard negatives

Audio-Visual Co-supervision

Objective: use vision and sound to learn from each other

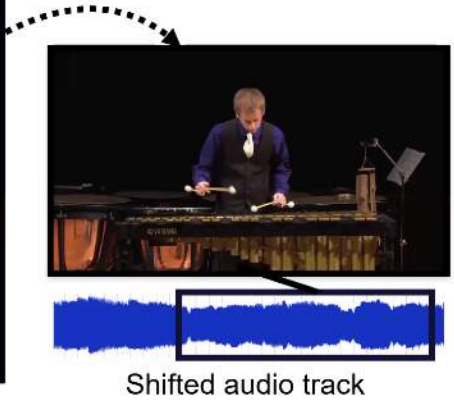
- Sound and frames are (i) semantically consistent, and (ii) synchronized

- Two types of proxy task:
  1. Predict audio-visual **correspondence**
  2. Predict audio-visual **synchronization**

Self-supervised Training

Audio-Visual Scene Analysis with Self-Supervised Multisensory Features, Andrew Owens, Alyosha Efros, ECCV 2018

Misaligned Audio



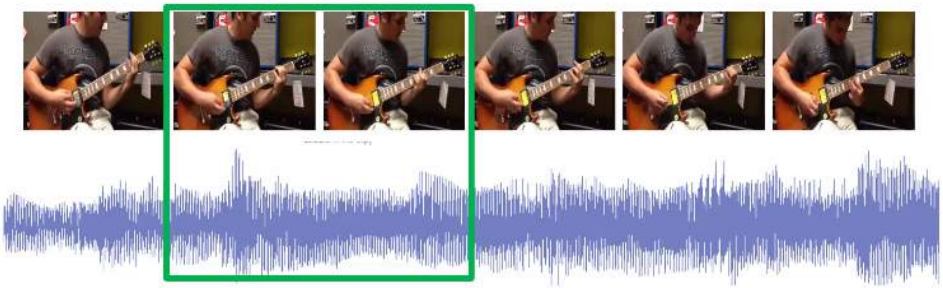
Shifted audio track

Audio-Visual Scene Analysis with Self-Supervised Multisensory Features, Andrew Owens, Alyosha Efros, ECCV 2018

Audio-Visual Synchronization

- **Positive samples:** in sync
- **Negative samples:** out of sync (introduce temporal offset)

positive



"Out of time: automated lip sync in the wild", Chung & Zisserman, ACCV Workshop, 2016

"Audio-Visual Scene Analysis with Self-Supervised Multisensory Features", Owens & Efros, 2018

Audio-Visual Synchronization

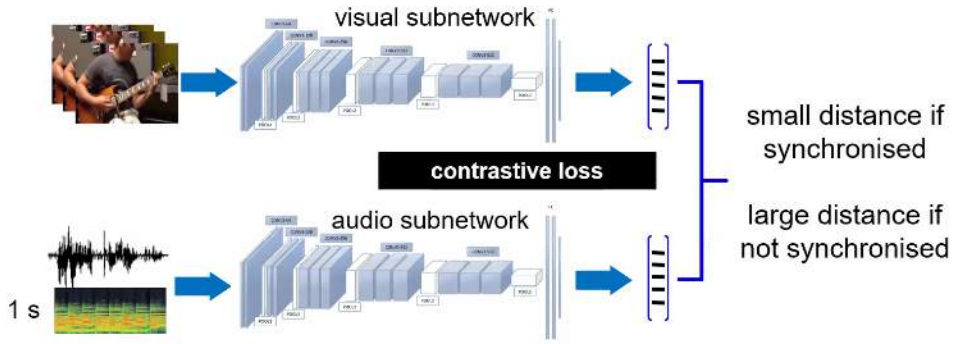
- **Positive samples:** in sync
- **Negative samples:** out of sync (introduce temporal offset)

negative

"Out of time: automated lip sync in the wild", Chung & Zisserman, ACCV Workshop, 2016

"Audio-Visual Scene Analysis with Self-Supervised Multisensory Features", Owens & Efros, 2018

Learning from Synchronization



The network is trained from scratch with contrastive loss to:

- Minimise distance between positive pairs (same temporal position)
- Maximise distance between negative pairs (different temporal positions)

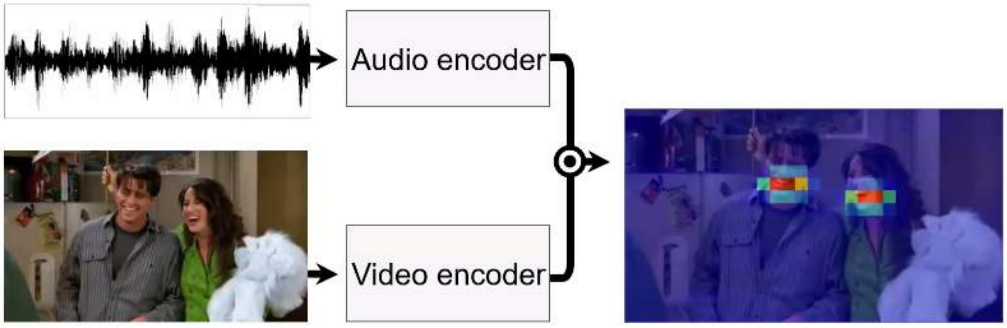
on hundreds of hours of video

Audio-Visual Scene Analysis with Self-Supervised Multisensory Features, Andrew Owens, Alyosha Efros, ECCV 2018

Audio-Visual Scene Analysis with Self-Supervised Multisensory Features, Andrew Owens, Alyosha Efros, ECCV 2018

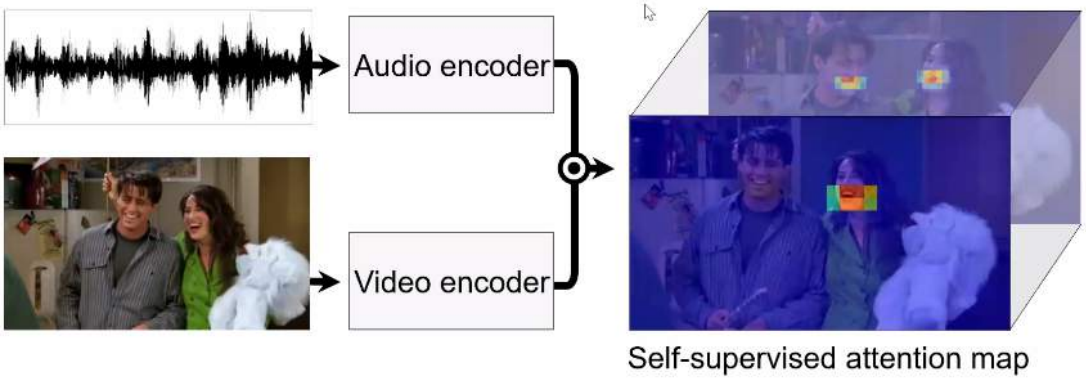
Audio-Visual Scene Analysis with Self-Supervised Multisensory Features, Andrew Owens, Alyosha Efros, ECCV 2018

Self-supervised synchronization on talking heads



Self-Supervised Learning of Audio-Visual Objects from Video
T. Afouras, A. Owens, J. S. Chung, A. Zisserman, ECCV 2020

Self-supervised synchronization on talking heads



Self-Supervised Learning of Audio-Visual Objects from Video
T. Afouras, A. Owens, J. S. Chung, A. Zisserman, ECCV 2020

Active Speaker Detection

Examples from the *Friends* series

Blue = active speaker
Red = inactive speaker

Active Speaker Detection

Examples from the *Friends* series

Blue = active speaker
Red = inactive speaker

Active Speaker Detection

Examples from the *Friends* series

Blue = active speaker
Red = inactive speaker

Active Speaker Detection

Examples from *Sesame Street*

Blue = active speaker
Red = inactive speaker

Active Speaker Detection

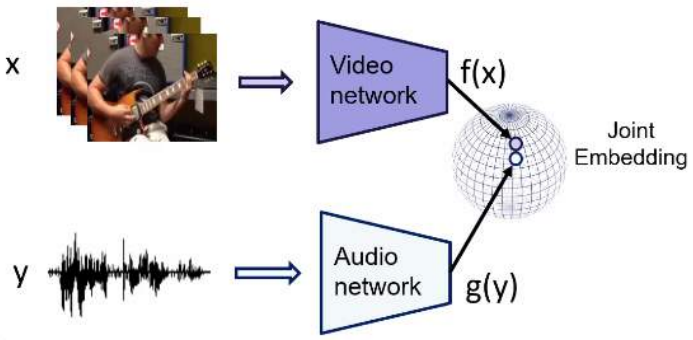
Examples from *The Simpsons*

Blue = active speaker
Red = inactive speaker

Summary Point: learn a joint audio-video embedding

Architecture: Dual Encoder

Separate networks for video and audio encoding



Score similarity, e.g. by $f(x)^\top g(y)$

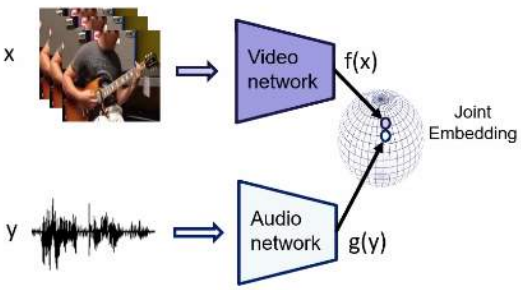
Self-supervised contrastive learning using: correspondence or synchronization

Question

What is the advantage of a dual encoder over a joint network?

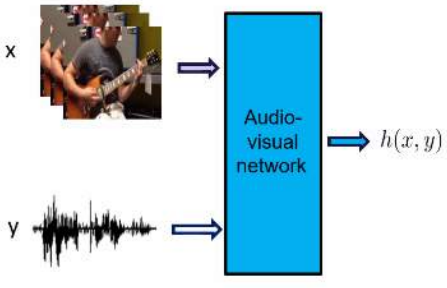
What is the disadvantage?

Architecture: Dual Encoder



Score similarity by $f(x)^\top g(y)$

Architecture: Joint Audio-Visual Network

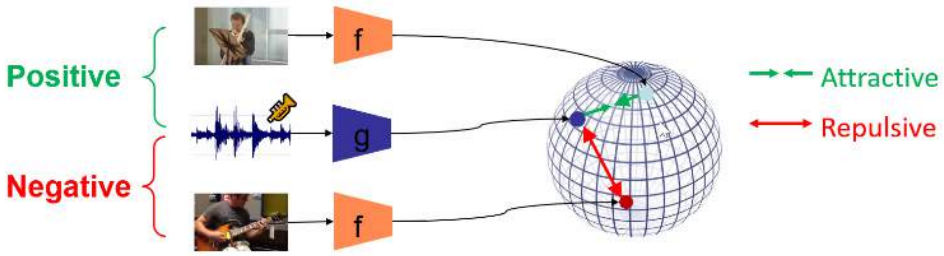


Score similarity by $h(x, y)$

Multi-Modal Contrastive Learning

Goal: Learn a joint multimodal space where embeddings of modalities that are semantically similar are close, and far otherwise

How? Using a ***contrastive*** approach.

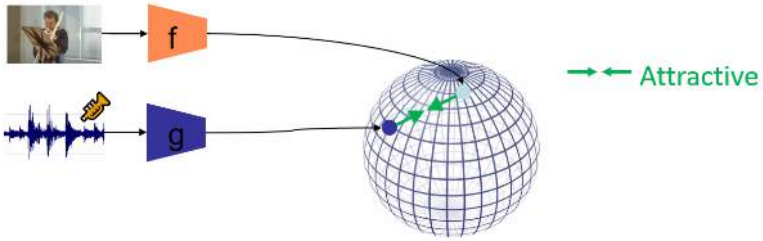


This is an "old" idea: DeVise, *Frome et al.*, ***NeurIPS2013*** and WSABIE, *Weston et al.* ***IJCAI 2011***.

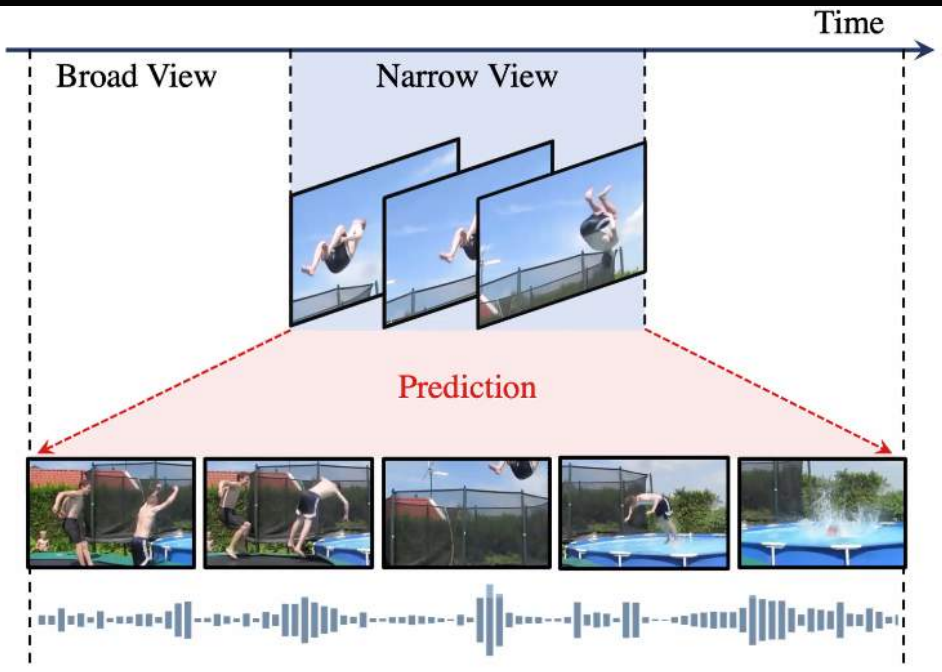
Multi-Modal Regression Learning

Goal: Learn a joint multimodal space where embeddings of modalities that are semantically similar are close, and far otherwise

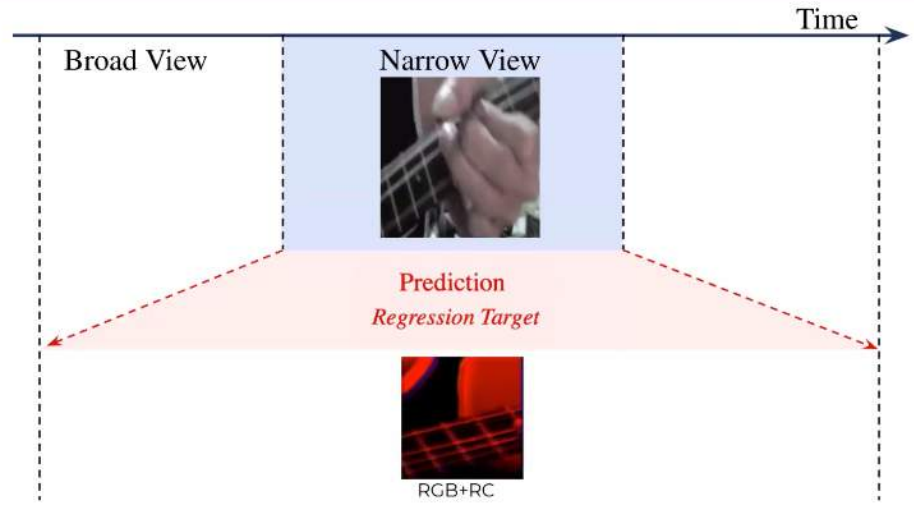
How? Predict the embedding vector of the other modality



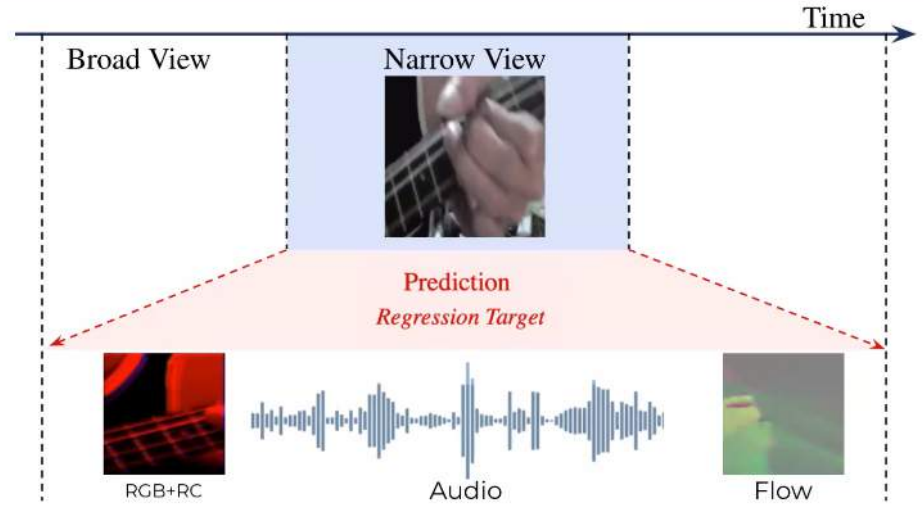
No negatives required!



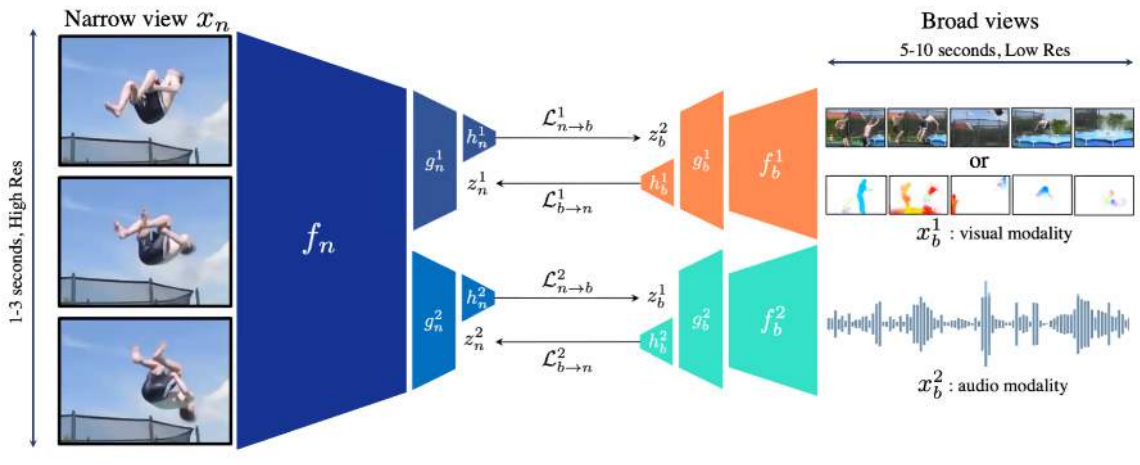
Example: Narrow and broad views



Example: Narrow and broad views



BraVe architecture



Training loss

Global loss

$$\mathcal{L}(x) = \underbrace{\mathcal{L}_{n\to b}(x)}_{\text{Narrow}\to\text{Broad}} + \underbrace{\mathcal{L}_{b\to n}(x)}_{\text{Broad}\to\text{Narrow}}$$

Narrow to Broad Loss

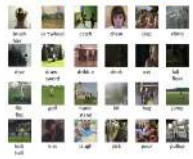
$$\mathcal{L}_{n\to b}(x) = \left\| \frac{h_n(z_n)}{\|h_n(z_n)\|_2} - \text{sg}\left[\frac{z_b}{\|z_b\|_2}\right] \right\|_2^2$$

Broad to Narrow Loss

$$\mathcal{L}_{b\to n}(x) = \left\| \frac{h_b(z_b)}{\|h_b(z_b)\|_2} - \text{sg}\left[\frac{z_n}{\|z_n\|_2}\right] \right\|_2^2$$

Standard evaluation datasets

HMDB51



Video Classification (51 classes)

Kinetics

Video Classification (600 classes)

UCF101

Video Classification (101 classes)

AudioSet

Audio Classification (527 classes)

ESC-50

Audio Classification (50 classes)

Comparison to SoTA: audio-visual learning

					UCF101		HMDB51		K600	ESC-50	AS
Method	Backbone (#params)	Dataset	Years	$\mathcal{M}$	Linear	FT	Linear	FT	Linear	Linear	MLP
ELo [66]	R(2+1)D-50 (46.9M)	YT8M	13	VFA		93.8	64.5	67.4			
AVID [57]	R(2+1)D-50 (46.9M)	AS	1	VA		91.5		64.7		89.2	
GDT [63]	R(2+1)D-18 (33.3M)	AS	1	VA		92.5		66.1		88.5	
MMV [4]	R(2+1)D-18 (33.3M)	AS	1	VA	83.9	91.5	60.0	70.1	55.5	85.6	29.7
XDC [5]	R(2+1)D-18 (33.3M)	AS	1	VA		93.0		63.7		84.8	
XDC [5]	R(2+1)D-18 (33.3M)	IG65M	21	VA		95.5		68.9		85.4	
BraVe:V↔A (ours)	TSM-50 (23.5M)	AS	1	VA	**93.4**	95.6	69.1	75.3	**71.1**	92.1	**36.4**
BraVe:V↔FA (ours)	TSM-50 (23.5M)	AS	1	VFA	93.2	95.8	70.2	76.9	70.3	92.6	36.3
BraVe:V↔FA (ours)	TSM-50x2 (93.9M)	AS	1	VFA	92.8	**96.5**	**70.6**	**79.3**	70.5	**92.9**	**36.4**
Supervised [12, 44, 66, 85]						96.8	71.5	75.9	82.4	94.7	43.9

Regression learning: recap and pros/cons

Summary: regress representations across views and modalities!

Pros	*Cons*
• **No negatives needed:** less importance of batch size	• **Harder to understand:** No notion of global/local optima
• **Works very well**	• **Sensitive to hyperparameters:** wrong hyperparameters can make the training collapse

Part III

Learning Human Action Representations using weak supervision from text

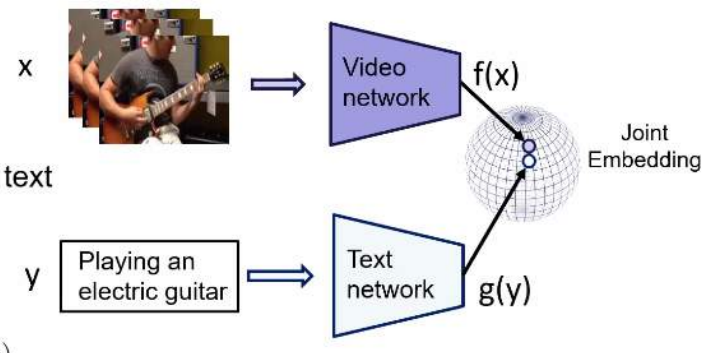
Objective: learn a joint text-video embedding

Architecture: Dual Encoder

Separate networks for video and text encoding

Score similarity, e.g. by $f(x)^\top g(y)$

How to obtain aligned text descriptions?



How to obtain video clips with aligned text descriptions?

Solutions:

1. Narrated instructional videos
   - HowTo100M dataset
2. Stock footage websites
   - WebVid-2M

What are narrated instructional videos?

Slide credit: A. Miech and J-B. Alayrac

The HowTo100M dataset

- 23K human **tasks** scraped from WikiHow
- 1.2M unique YouTube **videos** (duration 15 years)
- 136M **clips** with **narration** transcribed into text (mostly from ASR)
- Larger than any existing manually annotated captioning dataset

Dataset	Clips	Captions	Videos	Duration	Source	Year
Charades [48]	10k	16k	10,000	82h	Home	2016
MSR-VTT [58]	10k	200k	7,180	40h	Youtube	2016
YouCook2 [67]	14k	14k	2,000	176h	Youtube	2018
EPIC-KITCHENS [7]	40k	40k	432	55h	Home	2018
DiDeMo [15]	27k	41k	10,464	87h	Flickr	2017
M-VAD [52]	49k	56k	92	84h	Movies	2015
MPII-MD [43]	69k	68k	94	41h	Movies	2015
ANet Captions [26]	100k	100k	20,000	849h	Youtube	2017
TGIF [27]	102k	126k	102,068	103h	Tumblr	2016
LSMDC [44]	128k	128k	200	150h	Movies	2017
How2 [45]	185k	185k	13,168	298h	Youtube	2018
HowTo100M	**136M**	**136M**	**1.221M**	**134,472h**	Youtube	2019

HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips, Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, Josef Sivic, *ICCV2019*

DeepMind

Inria informatiques mathématiques

ENS

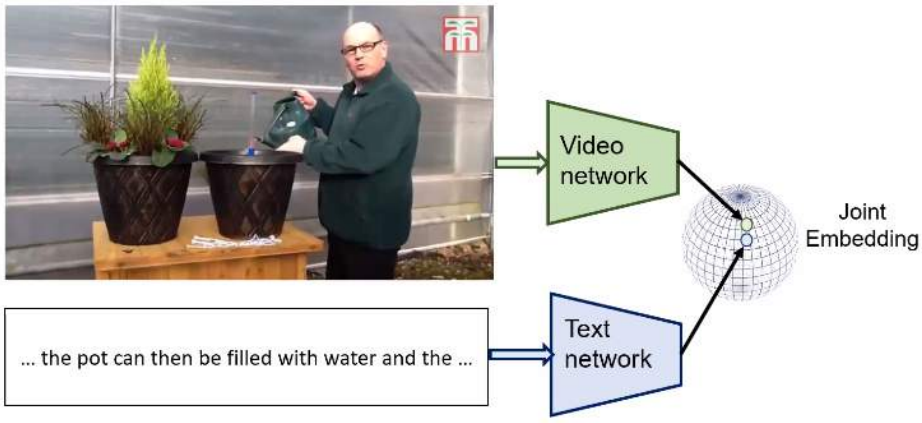
End-to-End Learning of Visual Representations from Uncurated Instructional Videos

A. Miech, J-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, A. Zisserman

CVPR 2020

Trained networks available

Objective: learn a joint text-video embedding



"Devise: A deep visual-semantic embedding model", A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. A. Ranzato, and T. Mikolov, NIPS, 2013.

Automatic Speech Recognition

you can add cilantro basil

add some cream to it

cream bacon spinach

keep it simple you just want to add

some fresh herbs maybe some oregano

you can add cilantro basil

they give it a couple more copies

Gotta start plating is to food yummy

fabulous tasty medierranean fish

Multiple Instance Learning approach

In MIL, we instead consider multiple positive candidate pairs

spinachs what's the name

keep it simple you just want to add

fresh herbs maybe some oregano

you can add cilantro basil they give

it a couple more copies when you

Multiple candidate positives

$\mathcal{P}$ Positive candidates

y^3 07:29	sander as you're going over this entire
y^1 07:32	area otherwise the end all product won't
y 07:34	be as flat as you would like it so just
y^2 07:37	be aware now once you have them enjoy
y^4 07:40	your sanding down your one on round

Learning embeddings for video and narrations

Combine Noise Contrastive Estimation (NCE) and Multiple Instance Learning (MIL)

Softmax over postive candidates and negatives

$$\max_{f,g} \sum_{i=1}^{n} \log \left(\frac{\sum_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)}}{\sum_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)} + \sum_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}} \right)$$

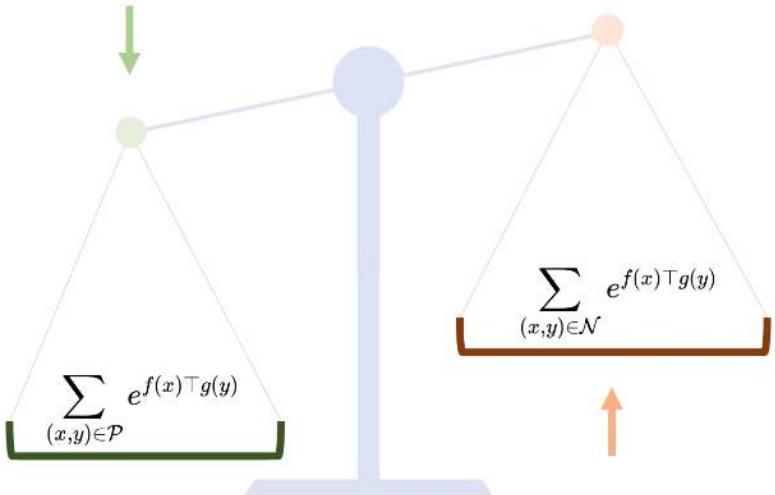
Negative sampling

Video network f(x)

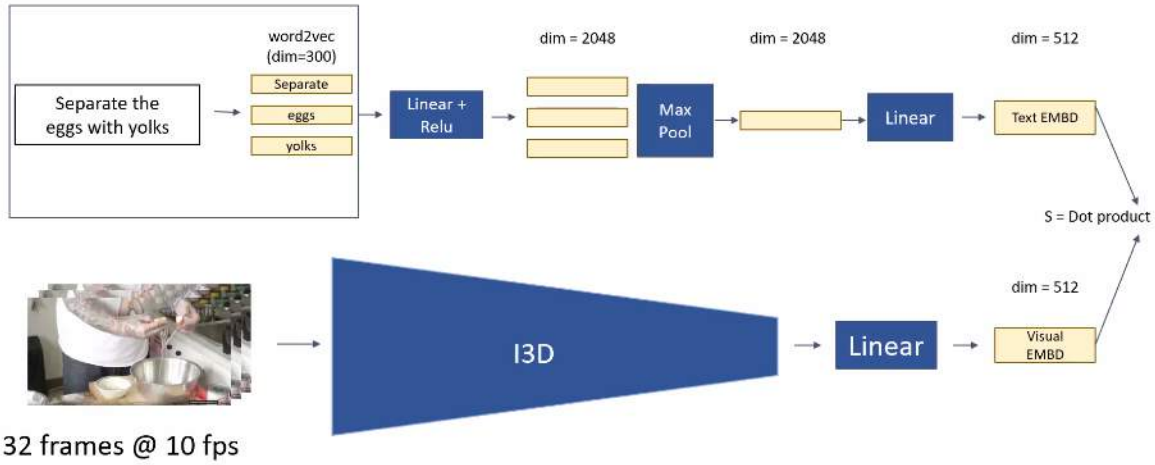
Text network g(y)

"Bag" of positive candidate

Noise-contrastive estimation: A new estimation principle for unnormalized statistical models, Gutmann and Hyvarinen, *AISTAT 2010*



Network Architecture



- Train with MIL-NCE objective on entire Howto100M training set

Trained model scores

$\mathcal{P}$ Positive candidates	
.50	main body of the laptop cover the
.63	duct tape with aluminum cover all
.61	remaining gaps edges with aluminum
.56	tape use the leftover poster board to
.50	create the keyboard keys I made my

$\mathcal{P}$ Positive candidates	
.67	spinach what's the name
.57	keep it simple you just want to add
.58	fresh herbs maybe some oregano
.59	you can add cilantro basil they give
.50	it a couple more copies and when you

UCF-101 and HMDB-51

Dataset	Year	Actions	Clips	Total	Videos
HMDB-51 [15]	2011	51	min 102	6,766	3,312
UCF-101 [20]	2012	101	min 101	13,320	2,500

HMDB & UCF-101 action recognition

Method	Backbone	Frozen	HMDB-51	UCF-101
OPN [43]	VGG-M-2048	✗	23.8	59.6
Shuffle and Learn [51]*	S3D	✗	35.8	68.7
Wang *et al.* [73]	C3D	✗	33.4	61.2
Geometry [24]	FlowNet	✗	23.3	55.1
Fernando *et al.* [23]	AlexNet	✗	32.5	60.3
ClipOrder [81]	R(2+1)D	✗	30.9	72.4
3DRotNet [36]*	S3D	✗	40.0	75.3
DPC [29]	3D-ResNet34	✗	35.7	75.7
CBT [67]	S3D	✓	29.5	54.0
CBT [67]	S3D	✗	44.6	79.5
AVTS [40]	I3D	✗	53.0	83.7
Ours	I3D	✓	**56.6**	**83.3**
		✗	**58.8**	**89.6**

How to obtain video clips with aligned text descriptions?

Solutions:

1. Narrated instructional videos
   - HowTo100M dataset
2. Stock footage websites
   - WebVid-2M

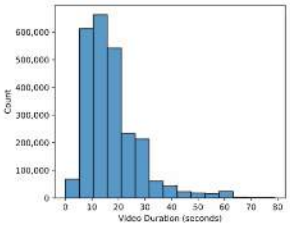
WebVid-2M Video-Caption Dataset

2.5M video-text pairs from stock footage websites

"Runners feet in a sneakers close up. realistic three dimensional animation."

"Female cop talking on walkie talkie, responding emergency call, crime prevention"

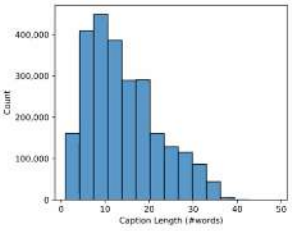
"Billiards, concentrated young woman playing in club"



"Lonely beautiful woman sitting on the tent looking outside. wind on the hair and camping on the beach near the colors of water and shore. freedom and alternative tiny house for traveler lady drinking"

"Kherson, ukraine - 20 may 2016: open, free, rock music festival crowd partying at a rock concert. hands up, people, fans cheering clapping applauding in kherson, ukraine - 20 may 2016. band performing"

"Cabeza de toro, punta cana/ dominican republic - feb 20, 2020: 4k drone flight over coral reef with manta"



Captions written manually to encourage use of video clips

- less noisy than narrated instructional videos

Frozen in Time:

A Joint Video and Image Encoder for End-to-End Retrieval

Max Bain, Arsha Nagrani, Gül Varol, Andrew Zisserman

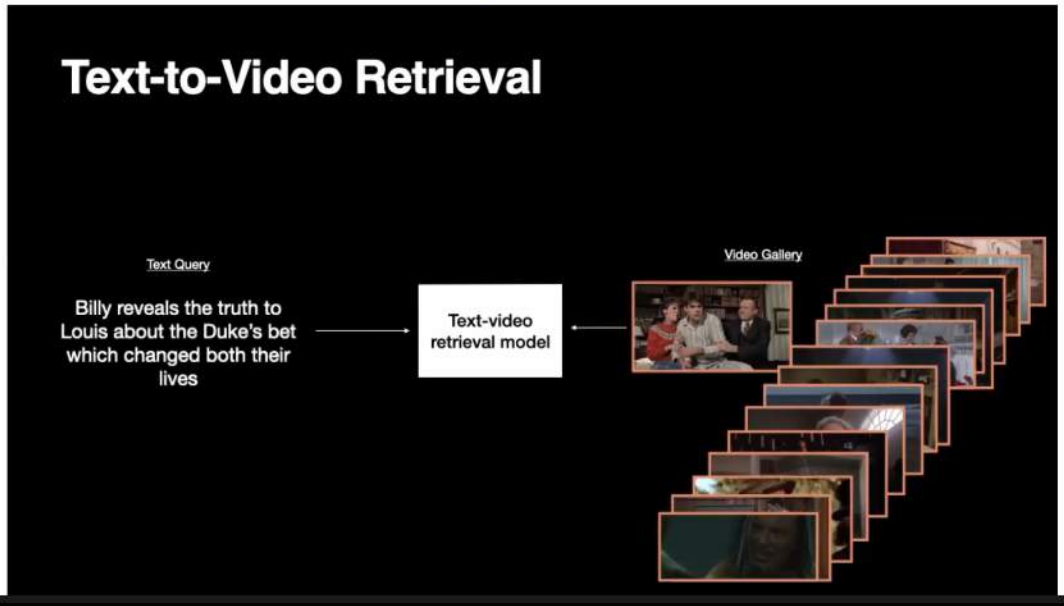
ICCV 2021

Objective 1

Text-to-video retrieval:

- Given a (large scale) library of video clips
- And a text description of a clip
- Retrieve the clip corresponding to the description

Objective 1

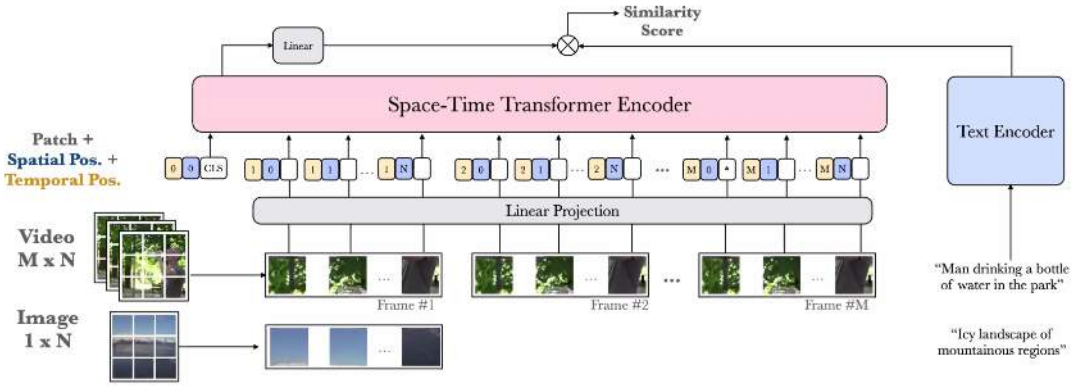


Objective 2: learn from images and videos with captions

Frozen in Time: A Joint Video and Image Encoder

- Transformer based architecture for text-to-video retrieval that can naturally ingest images and videos
  - Visual encoder that accepts a variable-length sequence
  - Treating images as 1-frame videos, ***frozen in time***
  - End-to-end-training (for images and video) from pixels
- Train on
  - Conceptual Captions 3M (images with captions)
  - WebVid-2M (video clips with captions)

End-to-end retrieval

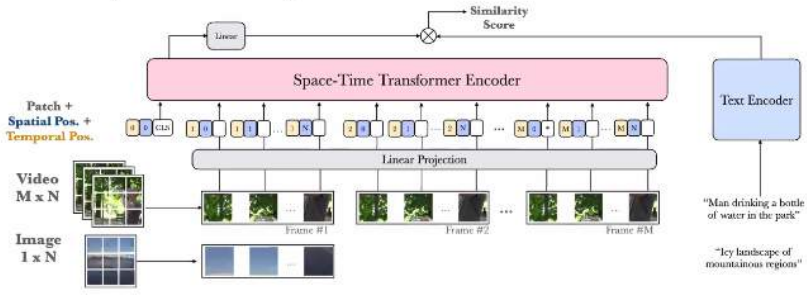


- Dual encoder for efficient retrieval

End-to-end retrieval

Video encoder:

- inspired from Timesformer [1]
- initialized from ViT [2] weights pretrained on ImageNet
  - Temporal embeddings zero-initialized

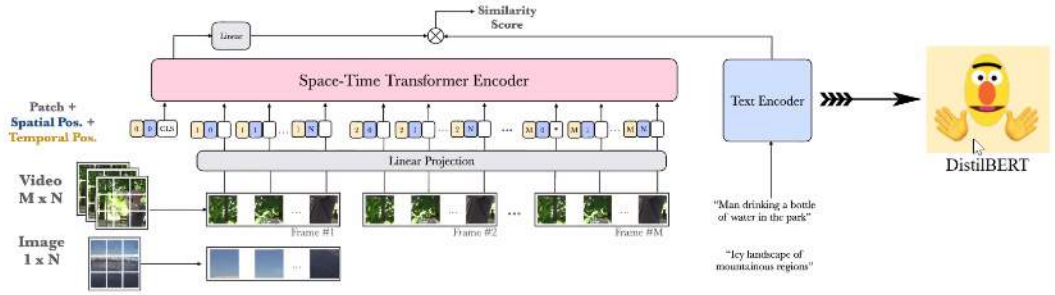


[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? ICML, 2021.
[2] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. ICLR, 2021.

End-to-end retrieval

Text encoder:

- initialized from DistilBERT [1]



Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf.
DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv, 2019.

Comparison to the state of the art

MSRVTT benchmark

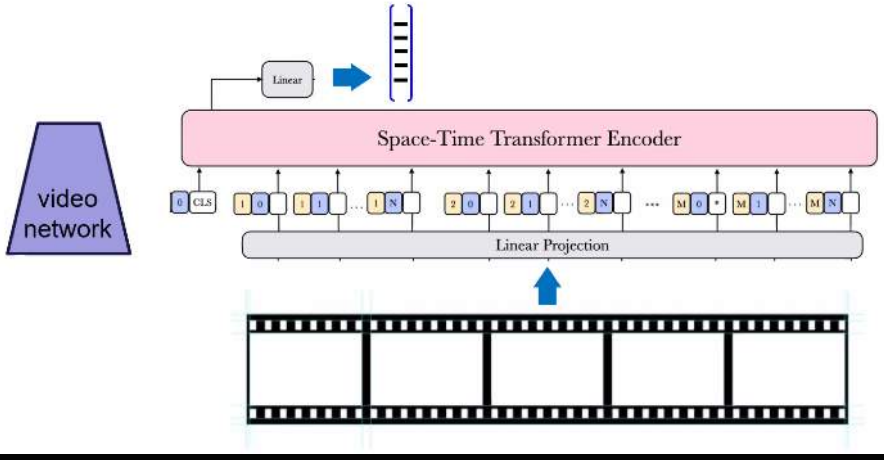
Method	E2E†	Vis Enc. Init.	Visual-Text PT	#pairs PT	R@1	R@5	R@10	MedR
JSFusion [62]	✓	-	-	-	10.2	31.2	43.2	13.0
HT MIL-NCE [35]	✓	-	HowTo100M	100M	14.9	40.2	52.8	9.0
ActBERT [67]	✓	VisGenome	HowTo100M	100M	16.3	42.8	56.9	10.0
HERO [27]	✓	ImageNet, Kinetics	HowTo100M	100M	16.8	43.4	57.7	-
VidTranslate [22]	✓	IG65M	HowTo100M	100M	14.7	-	52.8	
NoiseEstimation [2]	✗	ImageNet, Kinetics	HowTo100M	100M	17.4	41.6	53.6	8.0
CE [29]	✗	Numerous experts†	-		20.9	48.8	62.4	6.0
UniVL [31]	✗	-	HowTo100M	100M	21.2	49.6	63.1	6.0
ClipBERT [25]	✓	-	COCO, VisGenome	5.6M	22.0	46.8	59.9	6.0
AVLnet [44]	✗	ImageNet, Kinetics	HowTo100M	100M	27.1	55.6	66.6	4.0
MMT [15]	✗	Numerous experts†	HowTo100M	100M	26.6	57.1	69.6	4.0
Support Set [39]	✗	IG65M, ImageNet	-	-	27.4	56.3	67.7	3.0
Support Set [39]	✗	IG65M, ImageNet	HowTo100M	100M	30.1	58.5	69.3	**3.0**
Ours	✓	ImageNet	CC3M	3M	25.5	54.5	66.1	4.0
Ours	✓	ImageNet	CC3M, WebVid-2M	5.5M	**31.0**	**59.5**	**70.5**	**3.0**
Zero-shot								
HT MIL-NCE [35]	✓	-	HowTo100M	100M	7.5	21.2	29.6	38.0
Ours	✓	ImageNet	CC3M, WebVid-2M	5.5M	**18.7**	**39.5**	**51.6**	**10.0**

Using the joint text-video embedding for retrieval

Using the joint text-video embedding for retrieval

Offline: encode video dataset

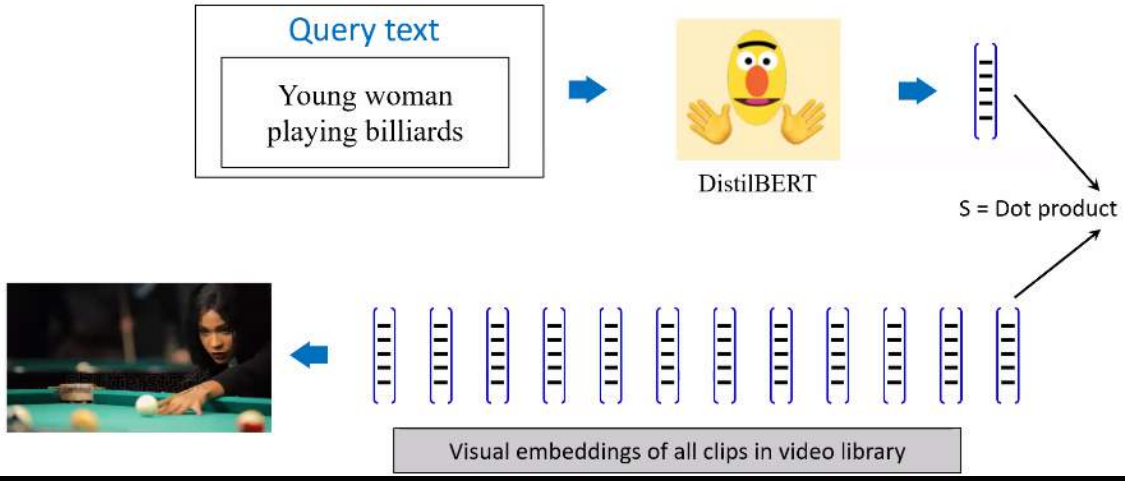
- use video network to generate representations of all video clips



Using the joint text-video embedding for retrieval

Online: encode query text and search for match

- use approximate nearest neighbours (ANN) for fast search



Real-Time Video Search Demo

Enter your search term... max display: 8

Paper

M. Bain, A. Nagrani, G. Varol, A. Zisserman.
Frozen in Time: A Joint Video and Image Encoder for End to End Paper.
ICCV, 2021.
(hosted on ArXiv)

[Bibtex]

Real-Time Video Search Demo

busy street in india max display: 8

Similarity: 1.58 Similarity: 1.57 Similarity: 1.57 Similarity: 1.55

Similarity: 1.55 Similarity: 1.54 Similarity: 1.54 Similarity: 1.54

Real-Time Video Search Demo

family camping in a field max display: 8

Similarity: 1.47 Similarity: 1.45 Similarity: 1.45 Similarity: 1.44

Similarity: 1.44 Similarity: 1.43 Similarity: 1.42 Similarity: 1.41

Real-Time Video Search Demo

women dancing in a street max display: 8

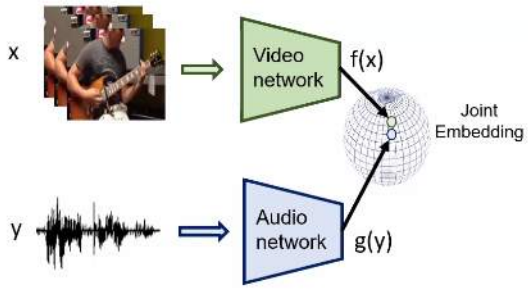
Similarity: 1.35 Similarity: 1.33 Similarity: 1.32 Similarity: 1.29

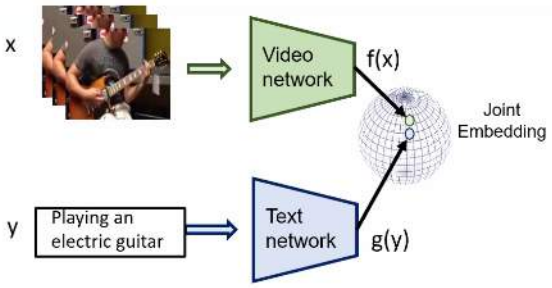
Similarity: 1.26 Similarity: 1.25 Similarity: 1.25 Similarity: 1.25

Summary

Summary: approaches to learn without strong (explicit) supervision:

- Audio-visual multi-modal self-supervised learning
- Weak supervision from text





Challenges for video understanding

1. Extended temporal sequences (beyond 10s)

- Need new datasets to explore this (so far have instructional & cooking videos)
- Will also drive new architectures, e.g. with memory

2. Multi-modality and multiple information streams

- Learn from both audio and aligned text
- And also: Speech/ASR, scene text, multi-lingual …
- Required to fully understand what is happening in a video