Day 9:

Speaker: Makarand Tapaswi, IIIT Hyderabad & Wadhwani AI

Title: Recent Advances in Video and Language Understanding.

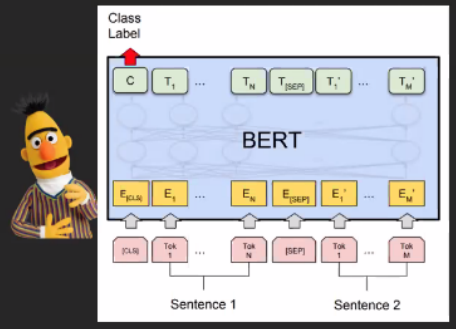
Evolution in CV

Large improvements in computer vision

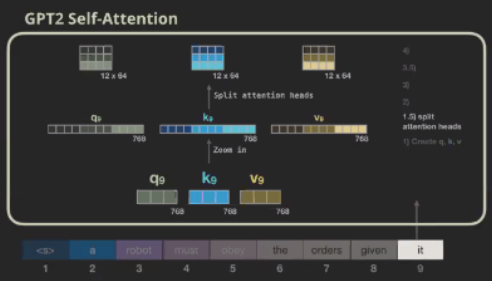
- Datasets with millions of annotated samples
- Convolutional Neural Networks with 100+ million learnable parameters
- Now, Transformers with billions of learnable parameters
- Coupled with smart augmentation strategies for self-supervised training

Recent Dev's in Sequence Models

Large improvements in language representations



OpenAI



- Massive text corpora are harnessed to build strong language models
- Examples: OpenAI's GPT and Google's BERT
- Such models show transfer learning capabilities to numerous NLP tasks

Multimodal Learning

Joint understanding of *Vision* and *Language*

Image-text Retrieval

A person eating a banana.

Image Captioning

A woman holding a banana up to her face.

Visual Question Answering

Q. Is the woman smiling?

A. Yes (?)

How we perceive ↓ → Yes for some & No for others

What does this picture show?

Objects and people:
- Mom
- Kid
- Book
- Teddy bear

What is happening?
- Mom is reading a book to her child

Raw image captioning

… but if it's a video?

The child starts crying after closing the book.

Rich information
- story in the book
- emotions of the parent and child
- why does the child cry?
- learn physics by observing that books on the shelf!

Slide Credit: Antonio Torralba
https://www.youtube.com/watch?v=AIEeakeXvMM

Example:

A multimodal audio-video-language stream

Real Life Problem

Joint understanding of *Video* and *Language*

Video Stories

Q. How did Jon Arryn die ?

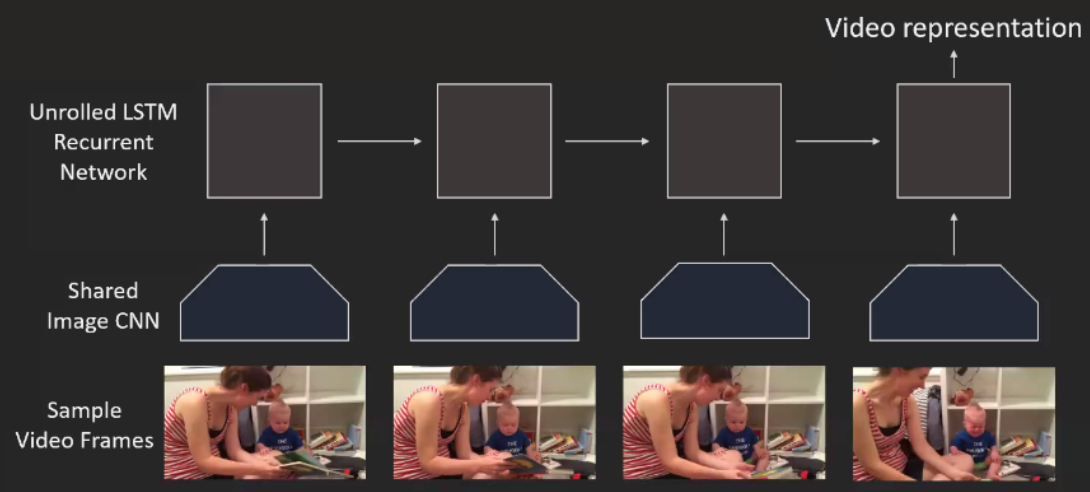
A. Fever *

Representation learning

How to bake a cake

Whirlwind tour of Video-Language Representations

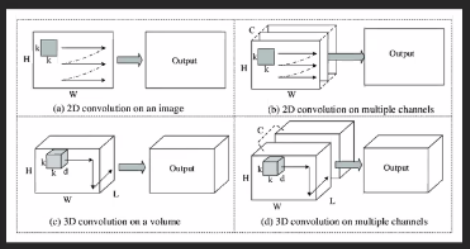
Video: CNN-LSTM Encoders

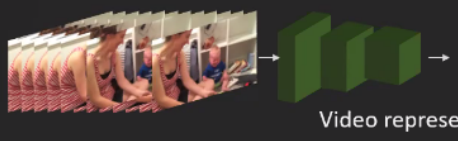


Reference: Long-term Recurrent Convolutional Networks for Visual Recognition and Description, PAMI 2014

Video: Time-aware Convolutional Networks

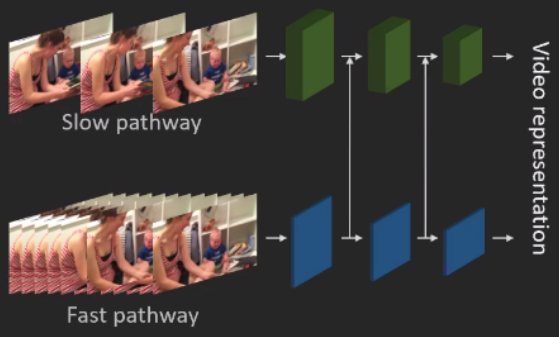
i3D
Inflating 2D convolutional networks to 3D





Reference: Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset, CVPR 2017

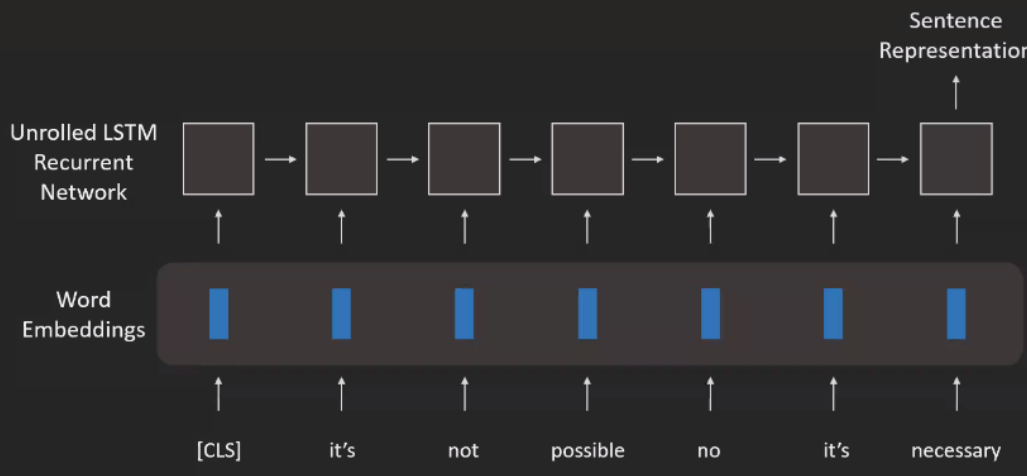
SlowFast
Convolutional Networks



Reference: SlowFast Networks for Video Recognition, ICCV 2019

The video frames → are passed in time, t

Sentence: Word2Vec + RNN/LSTMs

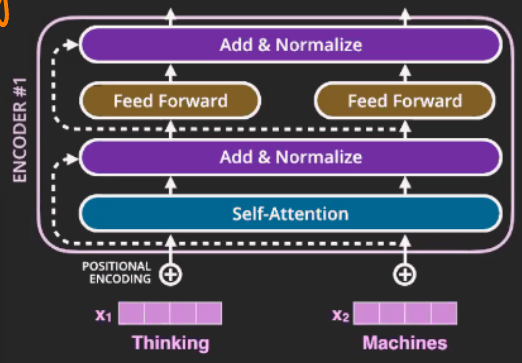


* Word Embedding included in LSTM

* Read → Attention is all you Need Paper !

* key, Query
* Self Attention

Transformers and the Self-Attention Layer



Reference: Attention is All you Need, NeurIPS 2017
Image Credits: The Illustrated Transformer, https://jalammar.github.io/illustrated-transformer/

BERT (very simplified diagram)!

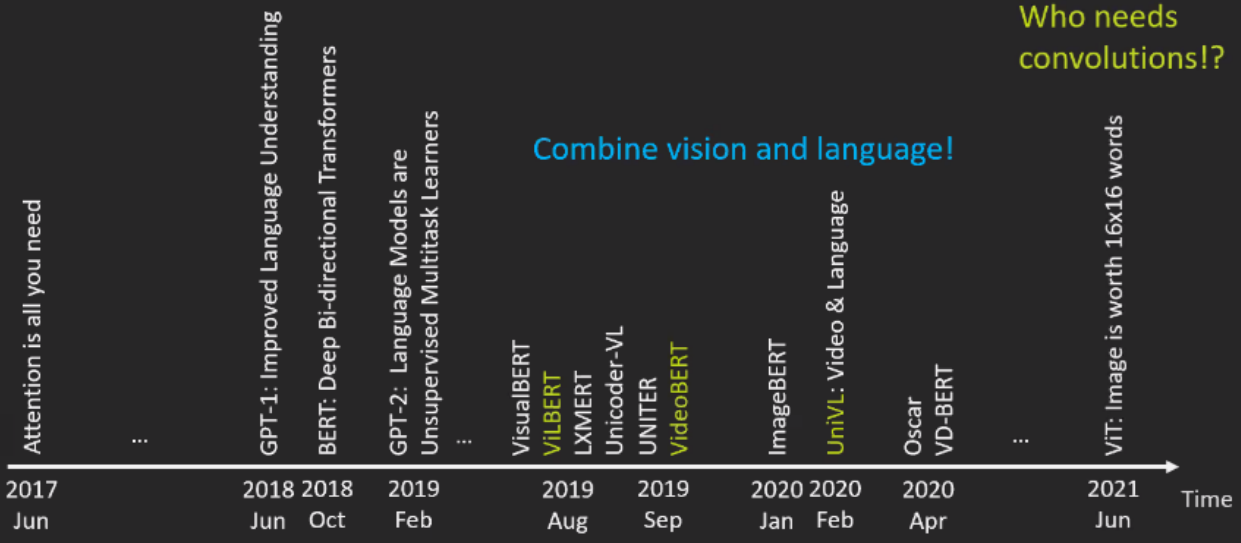
Reference: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv 2018.10

*Transform*ing vision and language!

Language is "solved"!?

Who needs convolutions!?

Combine vision and language!



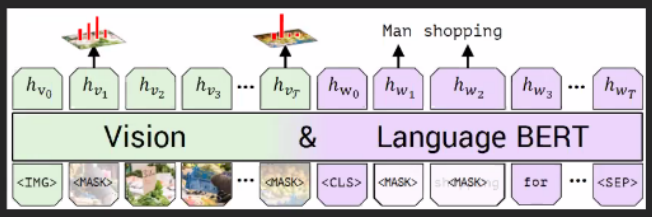
* to read →

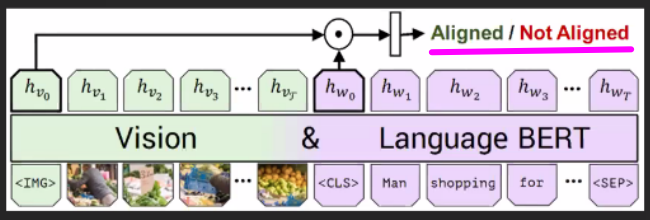
(Both vision & text learning multi model)

ViLBERT: Vision and Language BERT

Large-scale pretraining on image-caption pairs (Conceptual Captions)

a) Masked multi-modal learning



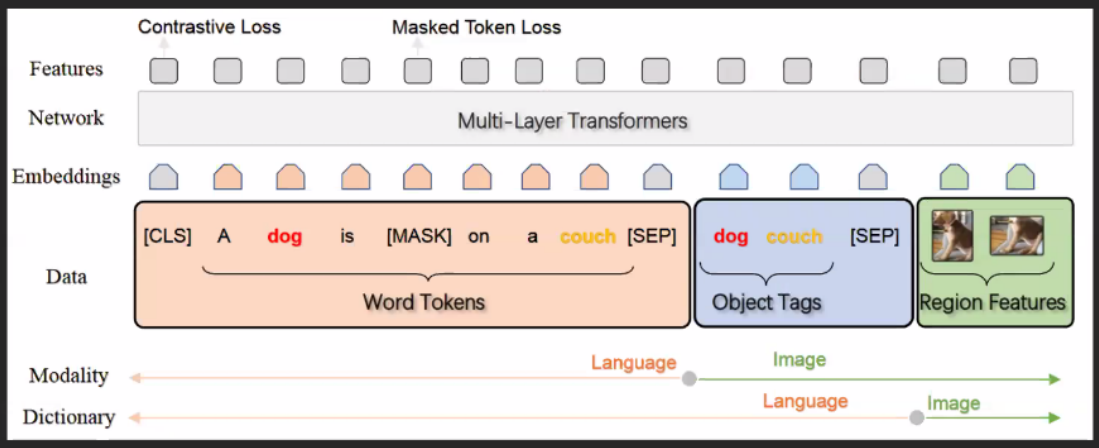


b) Multi-modal alignment prediction

Reference: ViLBert: Pretraining task-agnostic visio-linguistic representations for vision-and-language tasks, NeurIPS 2019

*

Oscar: Image regions, Object tags, Sentence



Reference: Oscar: Object-Semantics Aligned Pretraining for Vision-Language Tasks, ECCV 2020

Results:

Table 2: Detailed results on V+L tasks.

Method	Size	Text Retrieval R@1	Text Retrieval R@5	Text Retrieval R@10	Image Retrieval R@1	Image Retrieval R@5	Image Retrieval R@10	Text Retrieval R@1	Text Retrieval R@5	Text Retrieval R@10	Image Retrieval R@1	Image Retrieval R@5	Image Retrieval R@10
		1K Test Set						5K Test Set					
DVSA [14]	-	38.4	69.9	80.5	27.4	60.2	74.8	-	-	-	-	-	-
VSE++ [7]	-	64.7	-	95.9	52.0	-	92.0	41.3	-	81.2	30.3	-	72.4
DPC [46]	-	65.6	89.8	95.5	47.1	79.9	90.0	41.2	70.5	81.1	25.3	53.4	66.4
CAMP [42]	-	72.3	94.8	98.3	58.5	87.9	95.0	50.1	82.1	89.7	39.0	68.9	80.2
SCAN [18]	-	72.7	94.8	98.4	58.8	88.4	94.8	50.4	82.2	90.0	38.6	69.3	80.4
SCG [33]	-	76.6	96.3	99.2	61.4	88.9	95.1	56.6	84.5	92.0	39.2	68.0	81.3
PFAN [41]	-	76.5	96.3	99.0	61.6	89.6	95.2	-	-	-	-	-	-
Unicoder-VL [19]	B	84.3	97.3	99.3	69.7	93.5	97.2	62.3	87.1	92.8	46.7	76.0	85.3
12-in-1 [24]	B	-	-	-	65.2	91.0	96.2	-	-	-	-	-	-
UNITER [5]	B	-	-	-	-	-	-	63.3	87.0	93.1	48.4	76.7	85.9
UNITER [5]	L	-	-	-	-	-	-	66.6	89.4	94.3	51.7	78.4	86.9
OSCAR	B	88.4	99.1	99.8	75.7	95.2	98.3	70.0	91.1	95.5	54.0	80.8	88.5
OSCAR	L	**89.8**	**98.8**	**99.7**	**78.2**	**95.8**	**98.3**	**73.5**	**92.2**	**96.0**	**57.5**	**82.8**	**89.8**

(a) Image-text retrieval

Method	ViLBERT	VL-BERT	VisualBERT	LXMERT	12-in-1	UNITER$_B$	UNITER$_L$	OSCAR$_B$	OSCAR$_L$
Test-dev	70.63	70.50	70.80	72.42	73.15	72.27	**73.24**	73.16	**73.61**
Test-std	70.92	70.83	71.00	72.54	-	72.46	73.40	**73.44**	**73.82**

(b) VQA

Method	MAC	VisualBERT	LXMERT	12-in-1	UNITER$_B$	UNITER$_L$	OSCAR$_B$	OSCAR$_L$
Dev	50.8	67.40	74.90	-	77.14	**78.40**	78.07	**79.12**
Test-P	51.4	67.00	74.50	78.87	77.87	**79.50**	78.36	**80.37**

(c) NLVR2

Method	Test-dev	Test-std
LXMERT [39]	60.00	60.33
MMN [4]	-	60.83
12-in-1 [24]	-	60.65
NSM [12]	-	63.17
OSCAR$_B$	61.19	61.23
OSCAR$_B$*	**61.58**	**61.62**

(d) GQA

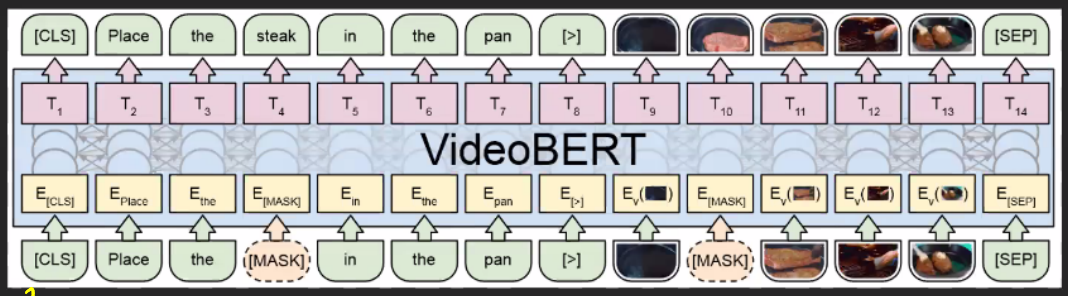
Method	cross-entropy optimization B@4	M	C	S	CIDEr optimization B@4	M	C	S
BUTD [2]	36.2	27.0	113.5	20.3	36.3	27.7	120.1	21.4
VLP [47]	36.5	28.4	117.7	21.3	39.5	29.3	129.3	23.2
AoANet [11]	37.2	28.4	119.8	21.3	38.9	29.2	129.8	22.4
OSCAR$_B$	36.5	30.3	123.7	23.1	40.5	29.7	137.6	22.8
OSCAR$_L$	**37.4**	**30.7**	**127.8**	**23.5**	**41.7**	**30.6**	**140.0**	**24.6**

(e) Image captioning on COCO

Method	in-domain CIDEr	in-domain SPICE	near-domain CIDEr	near-domain SPICE	out-of-domain CIDEr	out-of-domain SPICE	overall CIDEr	overall SPICE
UpDown [1]	78.1	11.6	57.7	10.3	31.3	8.3	55.3	10.1
UpDown + CBS [1]	80.0	12.0	73.6	11.3	66.4	9.7	73.1	11.1
UpDown + ELMo + CBS [1]	79.3	12.4	73.8	11.4	71.7	9.9	74.3	11.2
OSCAR$_B$	79.6	12.3	66.1	11.5	45.3	9.7	63.8	11.2
OSCAR$_B$ + CBS	80.0	12.1	80.4	**12.2**	75.3	**10.6**	79.3	**11.9**
OSCAR$_B$ + SCST + CBS	**83.4**	12.0	**81.6**	12.0	**77.6**	**10.6**	**81.1**	11.7
OSCAR$_L$	79.9	**12.4**	68.2	11.8	45.1	9.4	65.2	11.4
OSCAR$_L$ + CBS	78.8	12.2	78.9	**12.1**	77.4	10.5	78.6	**11.8**
OSCAR$_L$ + SCST + CBS	85.4	11.9	84.0	11.7	80.3	10.0	83.4	11.4

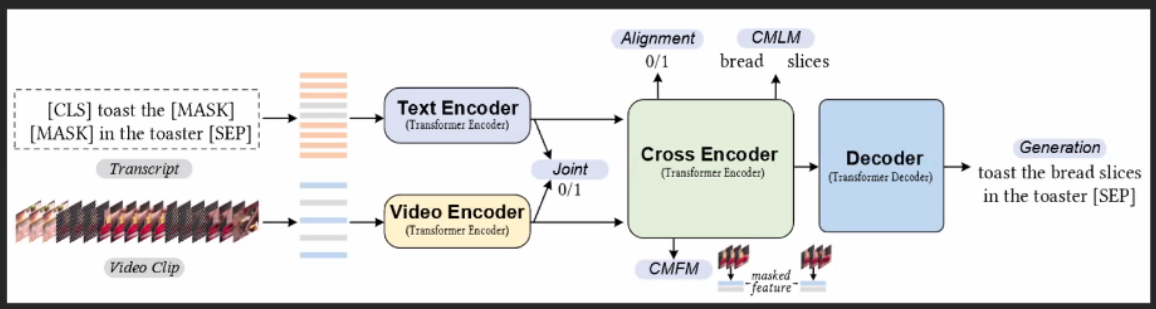
(f) Evaluation on NoCaps Val. Models are trained on COCO only without pre-training.

VideoBERT: Video and language, not different!



Reference: VideoBERT: A Joint Model for Video and Language Representation Learning, ICCV 2019

UniVL



Reference: UniVL: A Unified Video and Language Pre-Training Model for Multimodal Understanding and Generation, arXiv 2020.02

Two main streams in this talk

Video: Forest Gump chased by his friends with rocks.

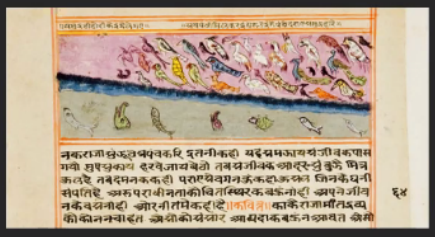
Human Interpretation of Stories →

Stories naturally engage vision and language

Drawing credit: Ananya Tapaswi, age 6

High in the mountains, lived a fire breathing dragon. Terrorizing the people of a nearby village had become its daily ritual. Fed up and frightened, the villagers finally go to their queen and ask for help.

Stories reflect human behavior



Panchatantra

Features stories about:

- Winning friends
- Losing friends
- Rash actions leading to losses
- How to face difficult situations
- Strategies of war and peace

Shakespeare

Features stories about:

- romance (Romeo & Juliet)
- manipulation and deceit (Macbeth)
- humor (A Midsummer Night's Dream)
- envy (The Merchant of Venice)

Story understanding

The ability of a machine to "watch" or "read" a story and perform human-like tasks

Search & Summarize

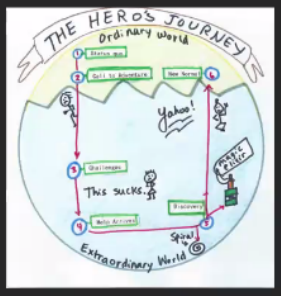
Story search

Movie summary

Answer Questions

Learn Human Behavior

Create new stories!



Learning from Movie texts

Transcripts and Subtitles with videos

Leonard: At least I didn't have to invent twenty-six dimensions just to make the math come out.

Leonard **stands up** and looks at Sheldon.

Sheldon: I didn't invent them, they're there.

Leonard: In what universe?.

00:07:00,733 --> 00:07:04,612
At least I didn't have to invent 26 dimensions to make the math come out.

00:07:04,773 --> 00:07:07,412
- I didn't invent them. They're there.
- In what universe?

Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video, BMVC 2006

Learning realistic human actions from movies, CVPR 2008

How to learn from multiple depictions

Plot Synopsis

Noting that the dire wolf is the sigil of the Stark family and there are as many pups as the Stark children, they take the pups in as companions.

Back at Winterfell, Catelyn informs her husband of a letter announcing the death of Lord Arryn, Eddard's old mentor and Catelyn's brother-in-law.

An additional message reports that the king himself is coming to Winterfell.

Film / TV episode

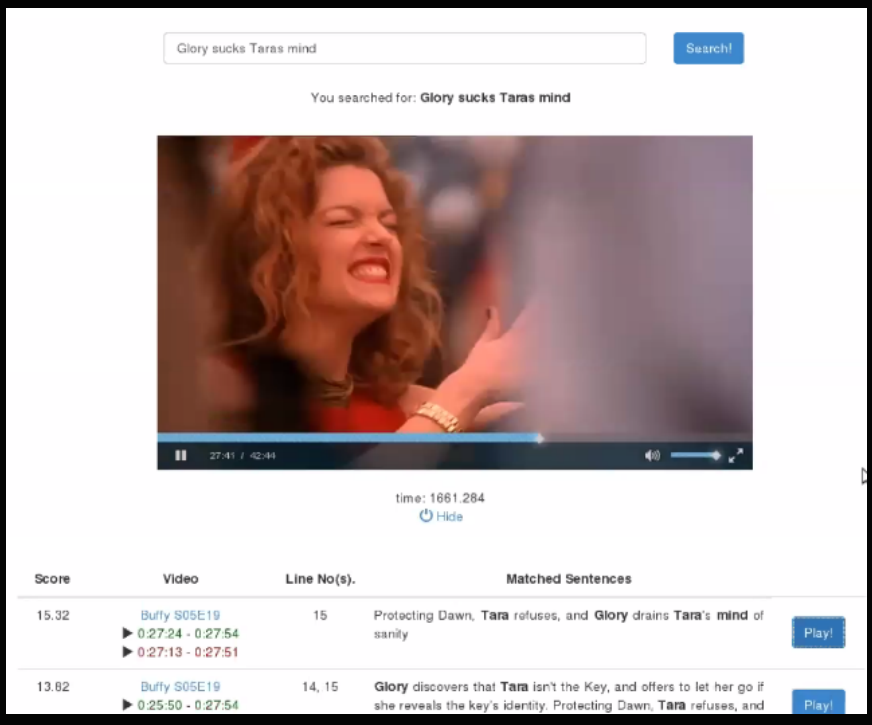
Book

"You did not come here to tell me crib tales. I know how little you like this place. What is it, my lady?"

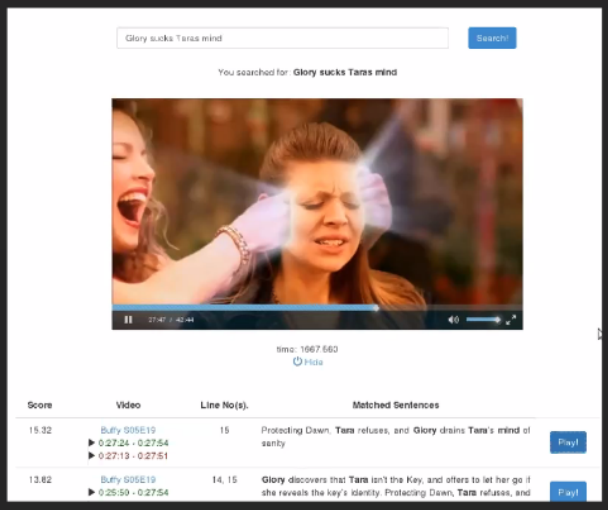
Catelyn took her husband's hand. "There was grievous news today, my lord. I did not wish to trouble you until you had cleansed yourself." There was no way to soften the blow, so she told him straight. "I am so sorry, my love. Jon Arryn is dead."

His eyes found hers, and she could see how hard it took him, as she had known it would. In his youth, ...

Search by Scene descriptions →



Search based on story plots



62 queries

85% found in plot

15% not localized in plot

53 queries

40 have overlap

13 don't overlap

[Tapaswi, IJMIR 2015]

Generate narrative captions

Plot synopses

1 sentence, video ~60+ s

Buffy performs the ritual in her bedroom then walks around her house looking for anything unusual.

[Tapaswi, IJMIR 2015]

Books

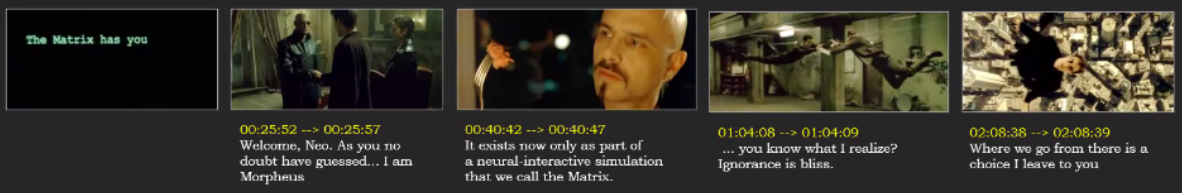
1 paragraph, video ~20 s

From an inside pocket of his black overcoat he pulled a slightly squashed box. Harry opened it with trembling fingers. Inside was a large, sticky chocolate cake with Happy Birthday Harry written on it in green icing.

[Tapaswi, CVPR 2015]

Characters are important!

MovieQA: Story question-answering



Movie:

- 200,000 frames
- 2,000 shots
- 1,000 dialogs

Questions:

Why does Cypher betray Morpheus?
How does Trinity save Neo?
What does Neo find at the Source?

- Long temporal dependencies
- Actions, interactions, emotions, intent

[Tapaswi, CVPR 2016]

* Character Identification Crucial in movies.

Identifying characters is crucial

Q. Who makes Indy return the crucifix after escaping from the grave robbers?

A1. The local sheriff
A2. Coronado
A3. No one, he keeps it
A4. The Boy Scout troop
A5. The grave robbers

How do we know characters' names?

- ✓ Dialog between characters
  "Hi, I'm Sheldon"
  "Raj, can you pass me the water?"
- ✓ Use weak annotations from dialog
- ✓ Multiple instance learning

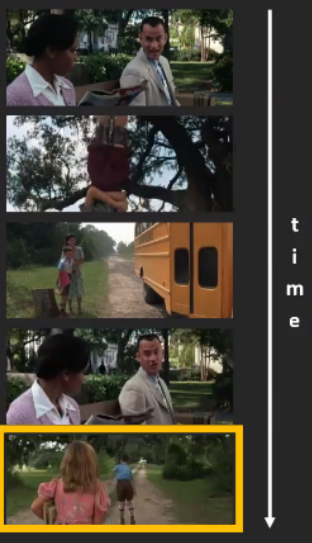
[Haurilet, WACV 2016]

More than facial appearance

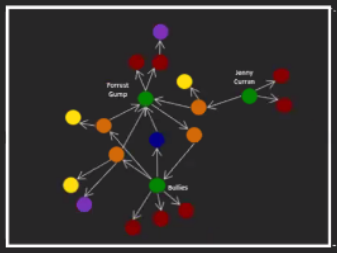
- Identity is more than just faces
- Learn clothing on-the-fly
- Speech patterns and speaking styles
- Energy based model to capture interactions across modalities

[Tapaswi, CVPR 2012]

Video clips of social situations



Situation Graph



To understand Characters behaviour

What are main aspects of human behaviour ?

Situations: argument, wedding, introduction, business meeting, work talk, fight, phone call, intimacy

Interactions: asks, orders, kisses, listens, leaves, hits, explains, yells, advises, suggests

Relationships: neighbor, friend, lover, parent, customer, stranger, colleague, spouse, child

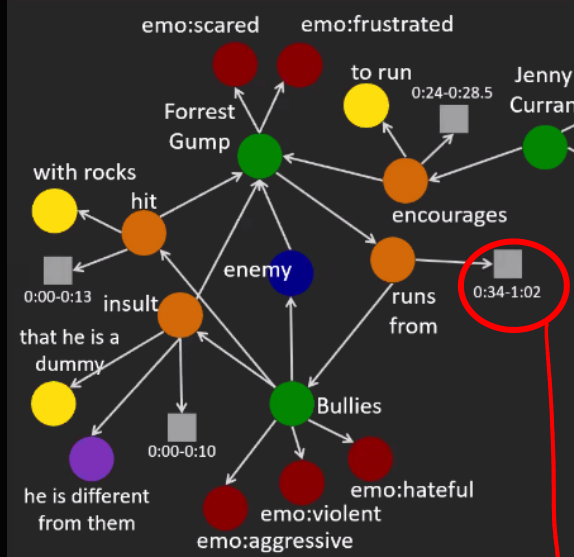
Attributes: female, adult, serious, nervous, upset, doctor, worried, happy, quiet, confused

tags for human behaviour

What are MovieGraphs?

Film Credit: Forrest Gump

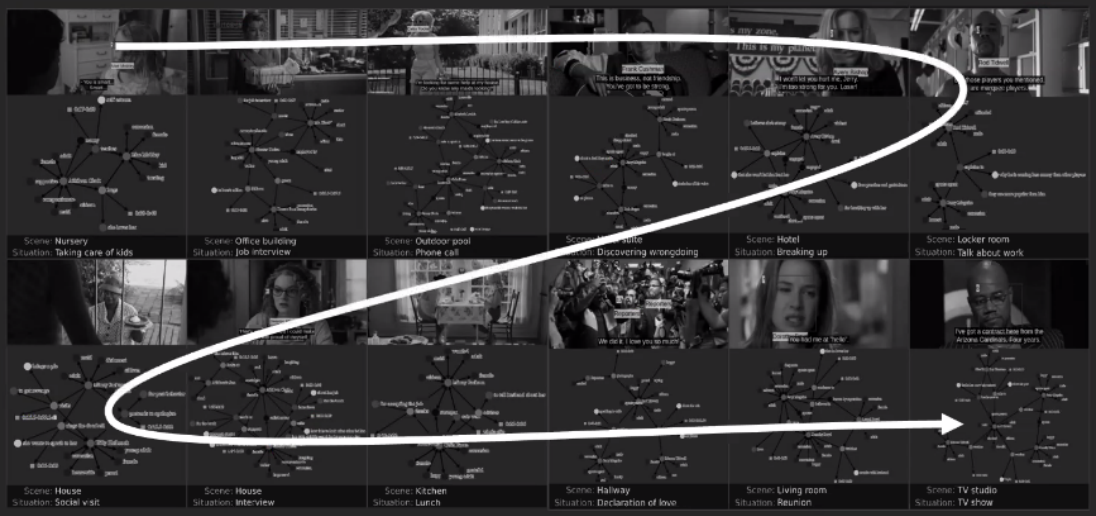
What are MovieGraphs?



Situation: Bullying
Scene: Field road
Description:
As Jenny and Forrest are walking down the road, three boys come along and start hitting Forrest. Jenny urges him to run from them. While Forrest runs, his leg braces begin to break apart.

→ Adding time stamps → sequence of events

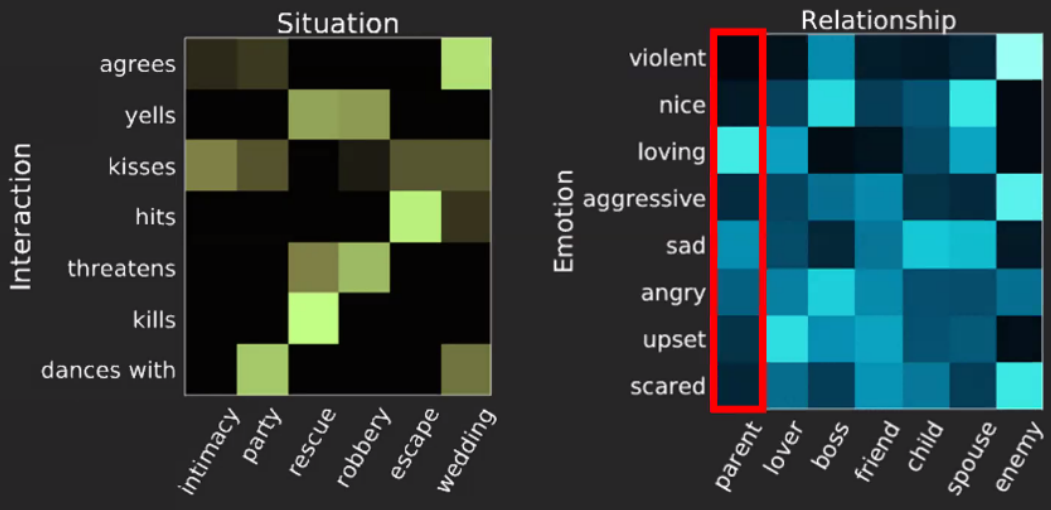
MovieGraphs Dataset



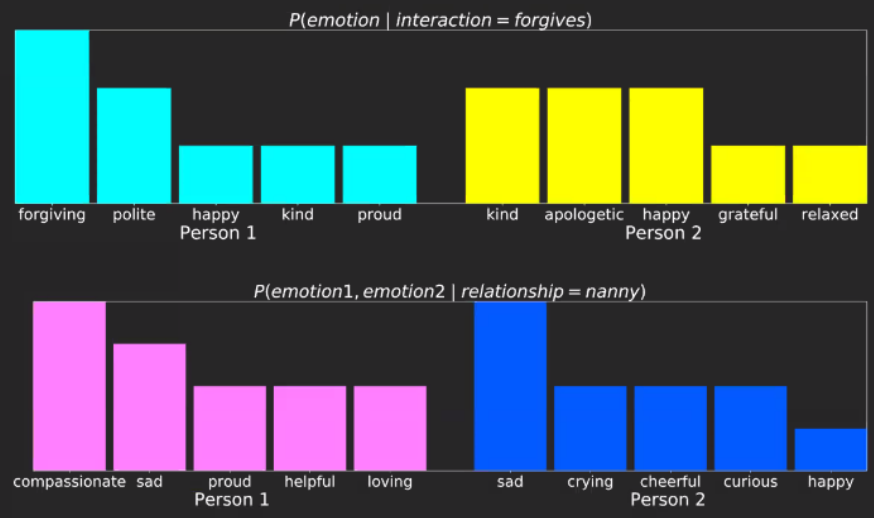
Exploring Social Common Sense

Social situations can be tricky!

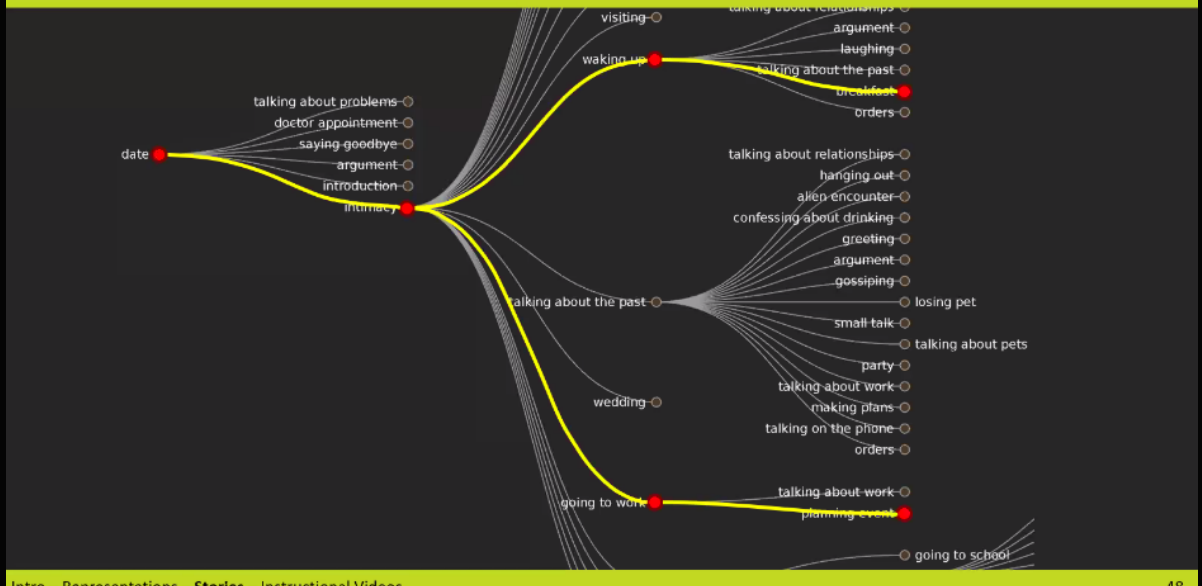
What happens when ...



Conditional Attributes



Situation Flow



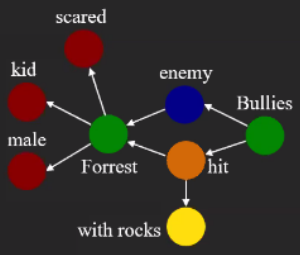
Graph based Video Retrieval

Graph-based Video Retrieval

Scoring similarity: graph query and video

graph movie latent id — component similarity

$$F(G, M, z) = \sum_c \phi(g_c, m_c, z)$$



Query Based Retrieval of Scenes.

Graph-based clip retrieval results - I

Query
Scene: Garden
Situation: Wedding

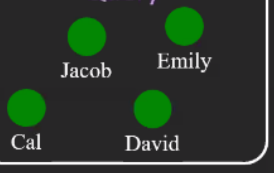
Rank 1

Jerry Maguire, 1996

Rank 2

Meet the Parents, 2000

Query



Crazy, Stupid, Love., 2011

Rank 1

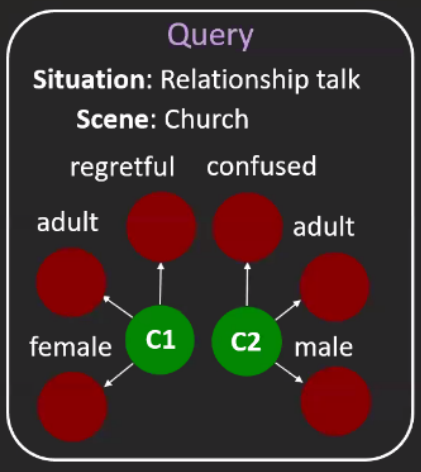
Jacob, Emily, Cal, David

Rank 2

Jacob, Cal

Graph-based clip retrieval results - II

Query
Situation: Relationship talk
Scene: Church



Rank 1

situation: argument
scene: banquet hall
C1: adult, female, heartbroken, angry
C2: adult, male, honest, remorseful

The Firm, 1993

Rank 2

situation: relationship talk
scene: church
C1: adult, female, regretful
C2: adult, male, confused

Four Weddings and a Funeral, 1994

Interactions and Relationships

A joint study of *interactions* and *relationships*

helps

laughs eats with

friends

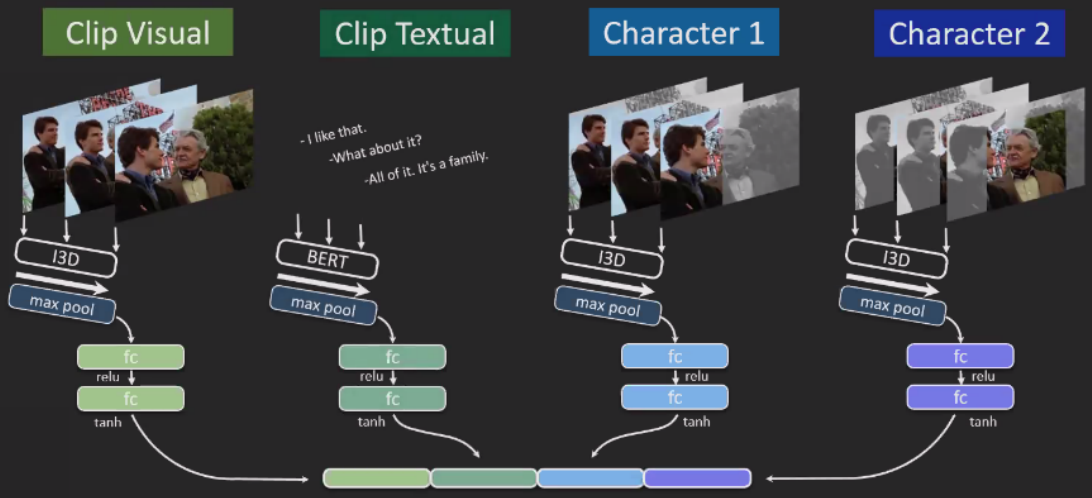
teacher

asks consoles

teaches

[Kukleva, CVPR 2020]

Multimodal embedding



Interaction prediction with multiple modalities

Clip Visual

walks 0.72
explains 0.68
watches 0.66

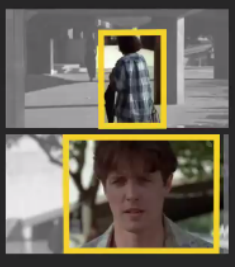
+

Clip Textual

watches 0.76
leaves 0.73
walks 0.72

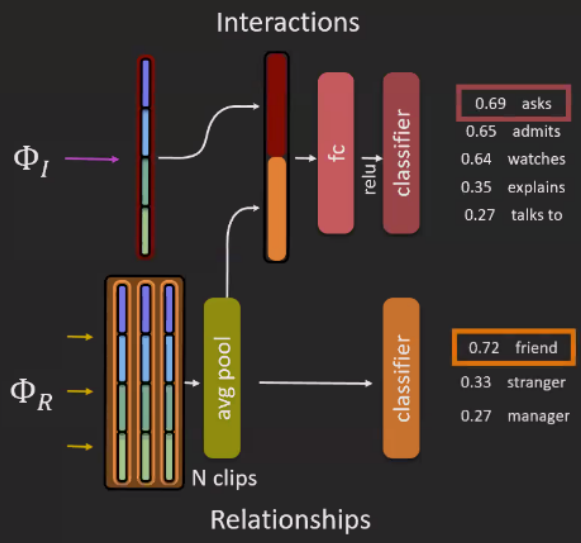
+

Tracks



leaves 0.75
walks 0.74
runs 0.7

Interactions, relationships, and multi-task learning



Joint prediction

Model	Interaction	Relationship
Random	1.0	6.7
Interaction	26.1	-
Relationship	-	26.8
Joint	26.3 (+0.2)	28.1 (+1.3)

Interactions

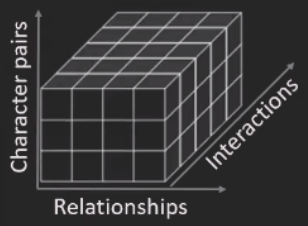
Hugs (siblings, child, lover)	+17%
Introduces (colleague, friend, stranger)	+14%
Runs (enemy, lover)	+12%
Talks to (all relationships)	-1%
Informs (many relationships)	-5%

Relationships

Sibling (hugs, informs, explains, watches)	+10%
Acquaintance (explains, greets)	+8%
Lover (suggests, kisses)	+7%
Parent (hugs, watches, suggests,explains)	-5%
Manager (talks to, suggests)	-8%

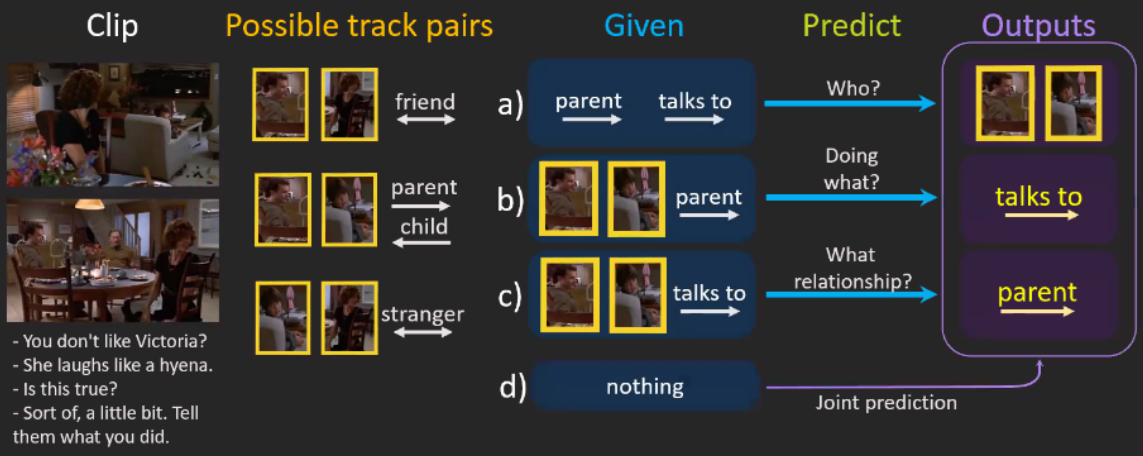
Which pair of characters are involved?

- Training with weak clip-level labels
- Unknown character pair is treated as a latent variable
- A max-margin loss with two tweaks is used to train models:
  - Later training epochs use hard negative sampling
  - Estimate $\hat{p}$ is sampled from a probability distribution rather than argmax
- At inference, we obtain a 3D tensor and compute argmax



Task	Given	Predict	Slice
Who?	interaction relationship	character	
What relationship?	character interaction	relationship	
Doing what?	character relationship	interaction	

Who, doing what, and what relationship?

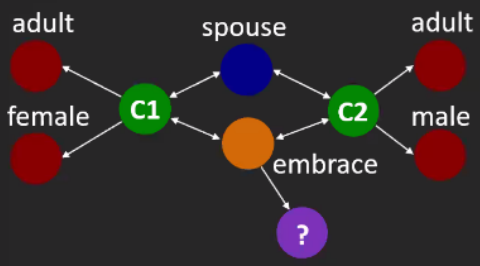


[Kukleva, CVPR 2020]

Reason prediction

Scene: House

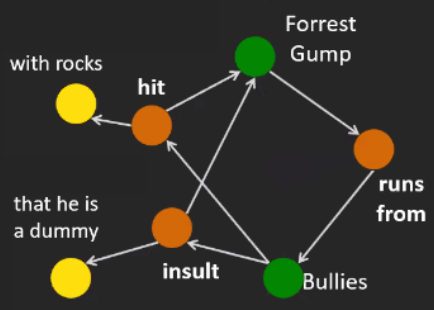
Situation: Talk about adoption



PRED: he is scared

GT: glad they'll have a baby

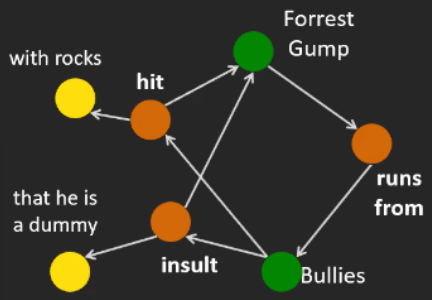
Interaction ordering

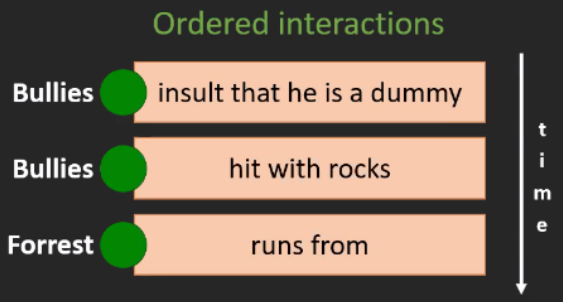


Unordered interactions

Forrest	runs from
Bullies	hit with rocks
Bullies	insult that he is a dummy

Interaction ordering





Revival of TV / Movie Understanding

Large Scale Movie Description Challenge (LSMDC)

His brow furrowed, SOMEONE looks down at the ground.

SOMEONE eyes him angrily, her jaw clenched.

SOMEONE heads off.

SOMEONE folds her arms.

SOMEONE approaches SOMEONE, who leans against the wall of the house.

LSMDC, 2015, https://sites.google.com/site/describingmovies/

Just identified → all human's as person

Large Scale Movie Description Challenge (LSMDC)

His brow furrowed, [PERSON1] looks down at the ground.

[PERSON2] eyes him angrily, her jaw clenched.

[PERSON1] heads off.

[PERSON2] folds her arms.

[PERSON1] approaches [PERSON3], who leans against the wall of the house.

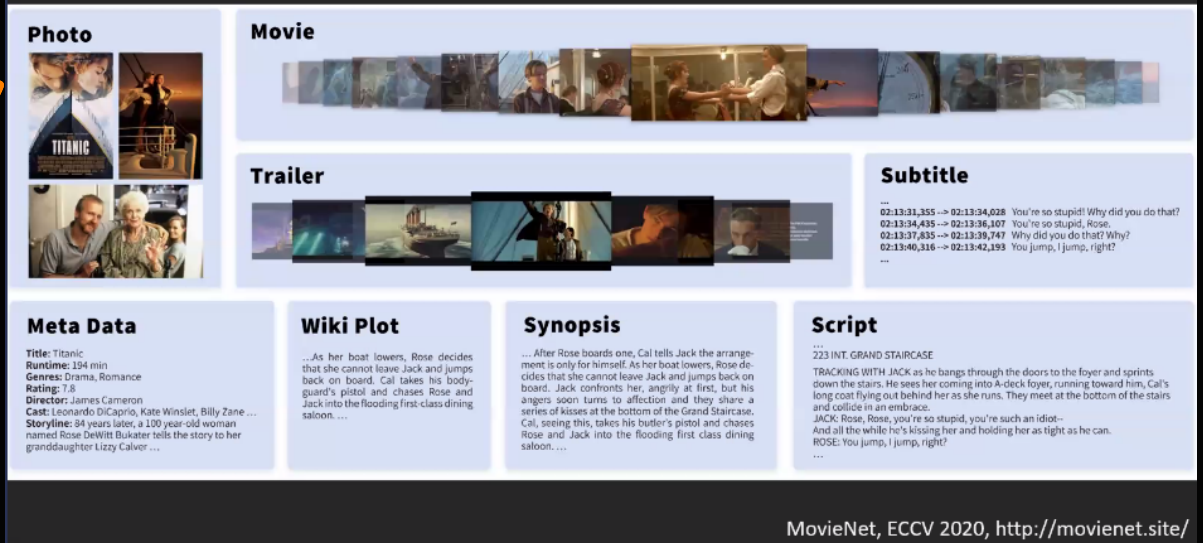
LSMDC, 2015, https://sites.google.com/site/describingmovies/

Condensed Movies

Condensed Movies, ACCV 2020
https://www.robots.ox.ac.uk/~vgg/research/condensed-movies/

* Full movie access & processing introduces copyright issues

MovieNet

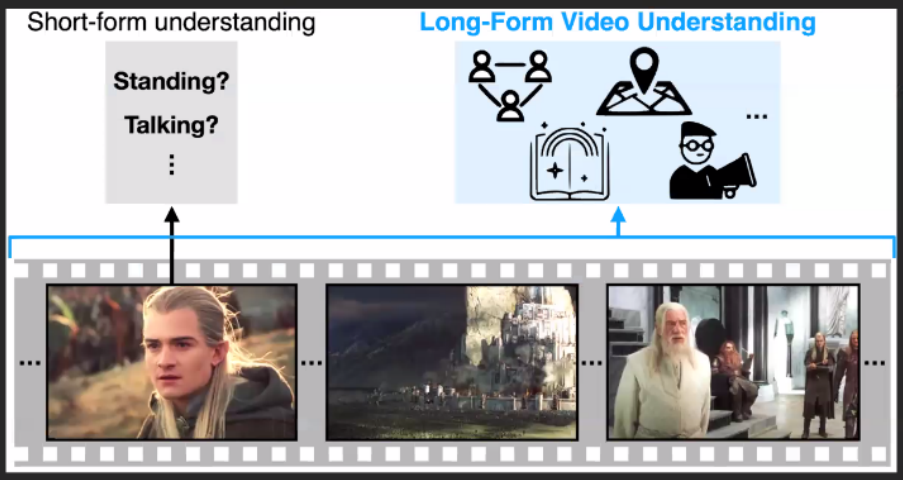


MovieNet, ECCV 2020, http://movienet.site/

* Japan → Researcher Adarsh.

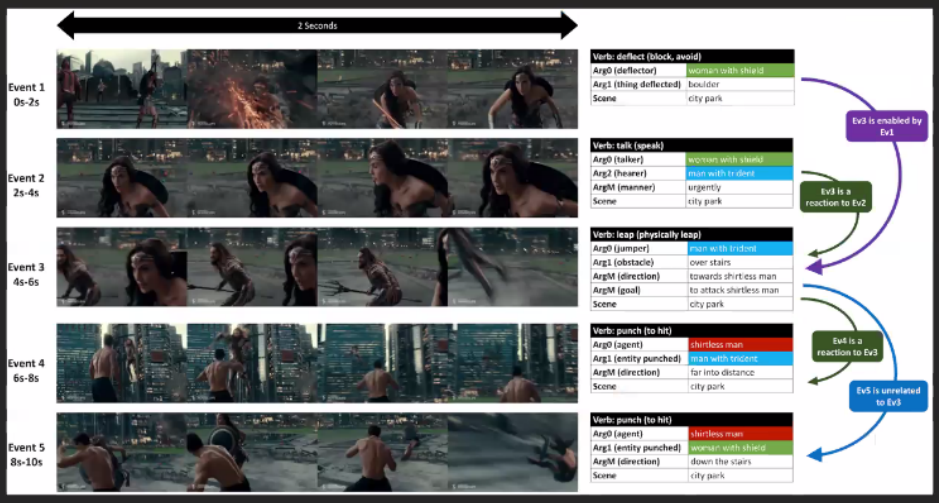
Extracting short freely available clips helps

Long-form Video Understanding



LVU, CVPR 2021, https://chaoyuan.org/lvu/

VidSitu: Video Situation Recognition



VidSitu, CVPR 2021, https://vidsitu.org/

VALUE challenge: Retrieval, QA, Captioning

Multi-channel Video

With both **Video Frames** and **Subtitle/ASR**

Diverse Video Domain

Diverse video content from **YouTube**, **TV Episodes** and **Movie Clips**

Various Datasets over Representative Tasks

11 datasets over 3 tasks: **Retrieval**, **Question Answering** and **Captioning**.

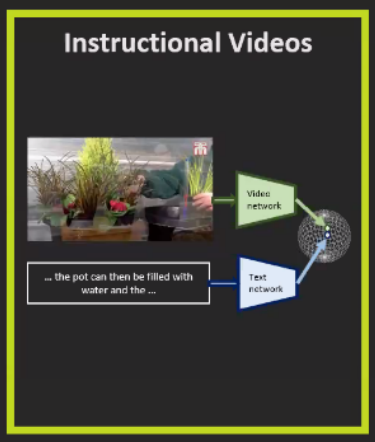
Leaderboard!

To track the advances in Video-and-Langauge research.

Attend Mohit Bansal's talk tomorrow!

VALUE Benchmark, https://value-benchmark.github.io/

Representation learning from Instructional videos



Can we learn without manual annotations?

Key idea 1

a **wide variety**
of instructional videos are available
at **large scale** on the Internet!

Key idea 2

Free supervision
using ASR output

WikiHow: filtering to visual instructional content

Scraped 130K tasks, kept 23K with potential visual cues

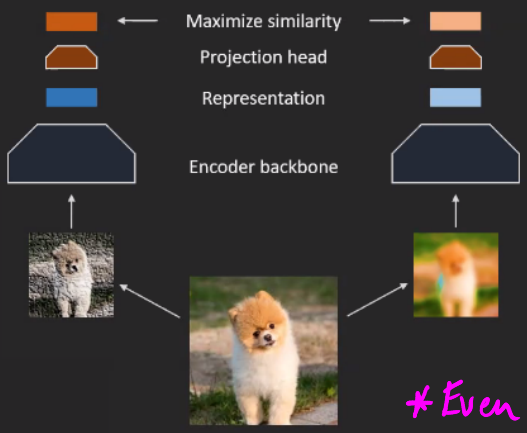
Dataset

HowTo100M

- Videos of people performing over 23k visual tasks
- 1.23 million Youtube videos (15 years!)
- Contains 130 millions of ~4s clips with narration
- Larger than any annotated video dataset

[Miech, ICCV 2019]

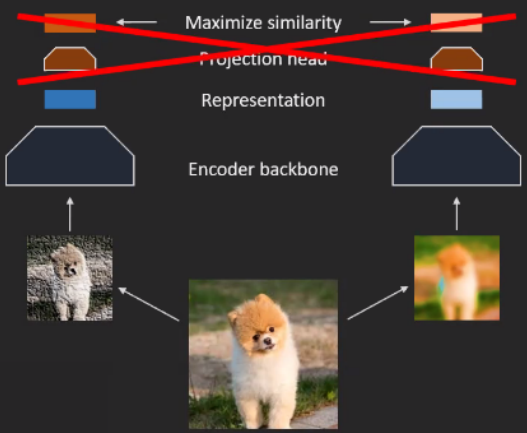
SimCLR: modern contrastive learning



* Even if labels are not known at the first place

Reference: A Simple Framework for Contrastive Learning of Visual Representations, ICML 2020

SimCLR: modern contrastive learning



Reference: A Simple Framework for Contrastive Learning of Visual Representations, ICML 2020

Intro – Representations – Stories – **Instructional Videos** 77

Video-text contrastive loss

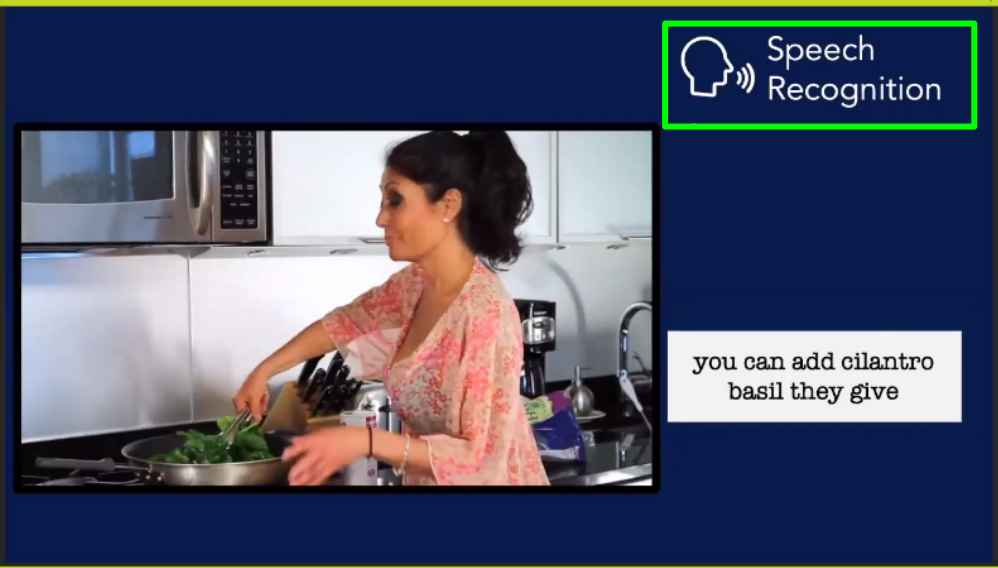
Video Credit: End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020

But the clips are not aligned …

Video Credit: End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020

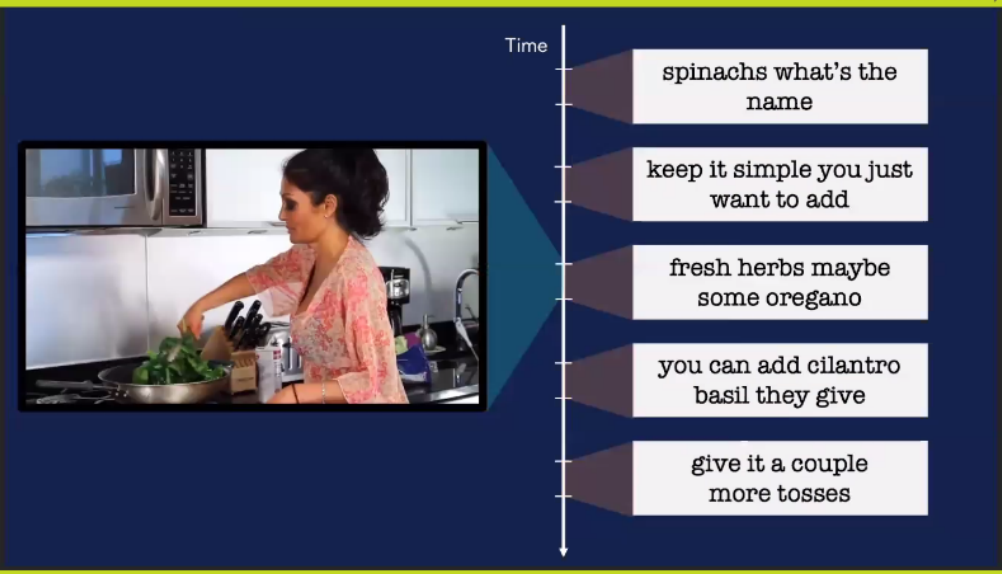
But the clips are not aligned …

Video Credit: End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020



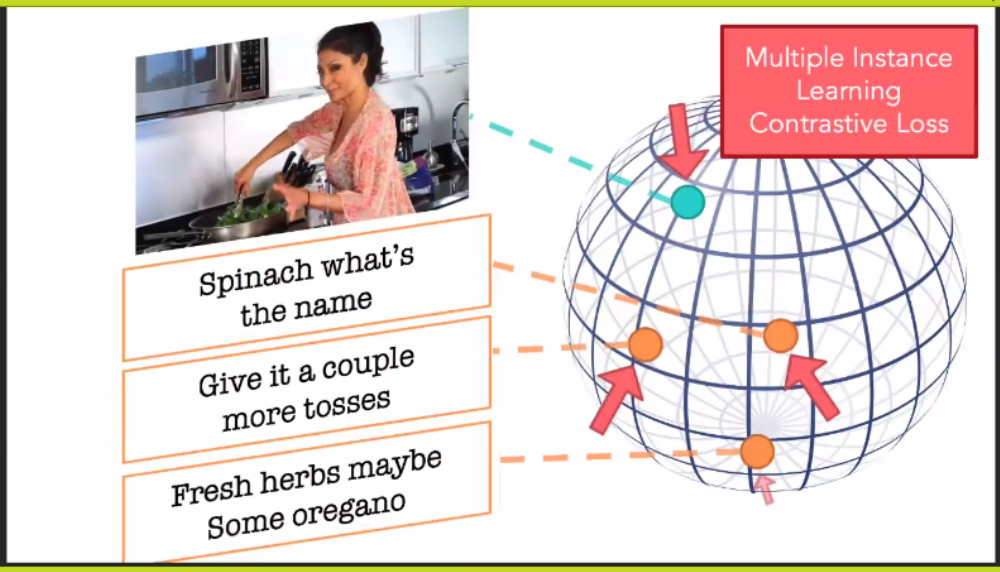
But the clips are not aligned …

Video Credit: End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020



But the clips are not aligned …

Video Credit: End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020



Multiple Instance Learning – Noise Contrastive Estimation

a) Standard Contrastive Loss

$$L_{i,j} = -\log \frac{\exp\left(\text{sim}(z_i, z_j)\right)}{\sum_k 1(k \neq i) \exp\left(\text{sim}(z_i, z_j)\right)}$$

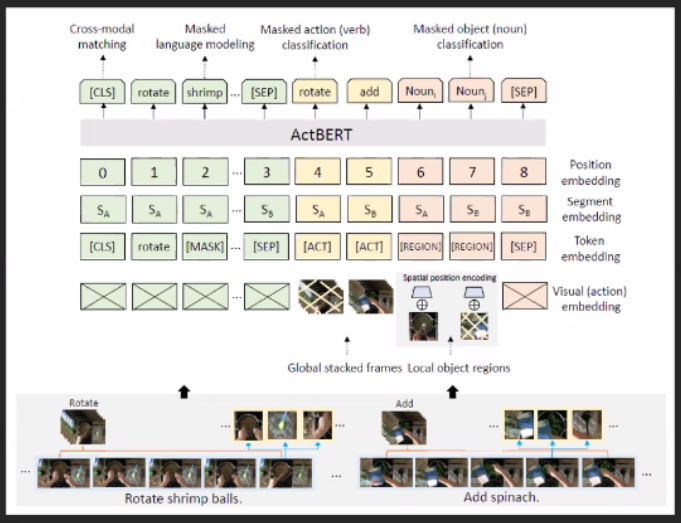
* Typo

b) MIL-NCE

$$L_{i,P_i} = -\log \frac{\sum_{j \in P_i} \exp\left(\text{sim}(z_i, z_j)\right)}{\sum_{j \in P_i} \exp\left(\text{sim}(z_i, z_j)\right) + \sum_{k \in N_i} \exp\left(\text{sim}(z_i, z_k)\right)}$$

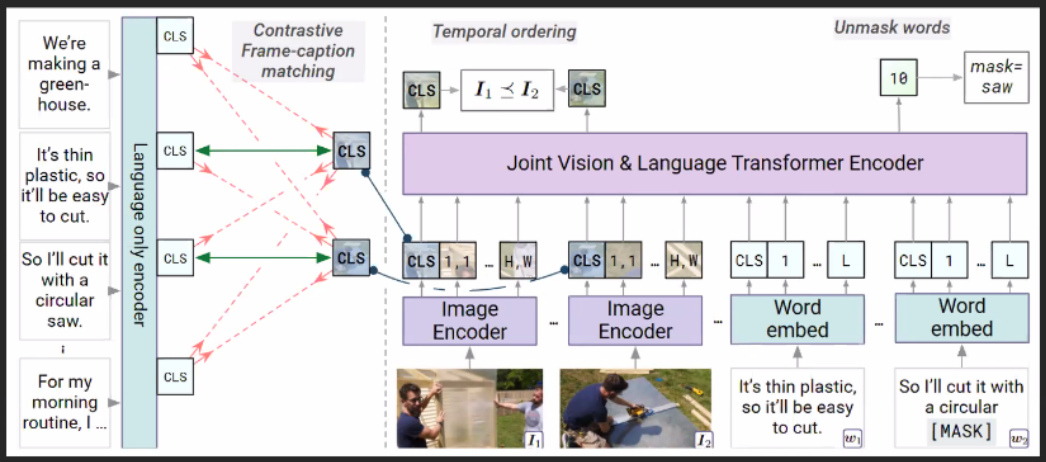
Reference: End-to-End Learning of Visual Representations from Uncurated Instructional Videos, CVPR 2020

ActBERT



Reference: ActBERT: Learning Global-Local Video-Text Representations, CVPR 2020

MERLOT: Multimodal knowledge models



Reference: MERLOT: Multimodal Neural Script Models, arXiv 2021.06

Thanks to my students, collaborators, adivsors!

If any of this is interesting for you …

Reach out!

I'm looking for motivated students to work on video-language problems

https://makarandtapaswi.github.io/students/