Language Technologies Institute

Carnegie Mellon University

Advanced Multimodal Machine Learning

Lecture 1.2: Challenges and applications

Louis-Philippe Morency
Tadas Baltrušaitis

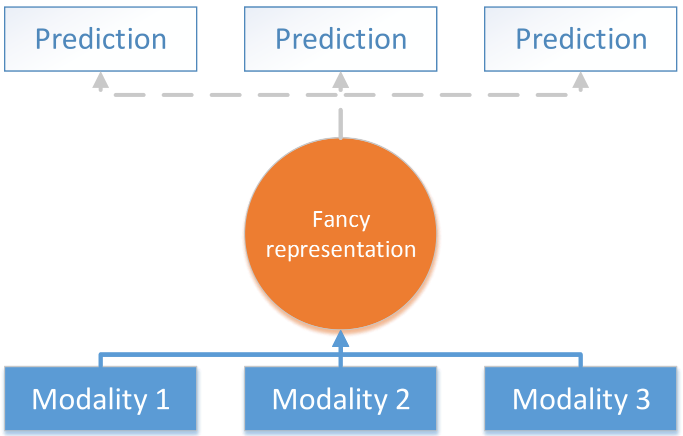
Objectives

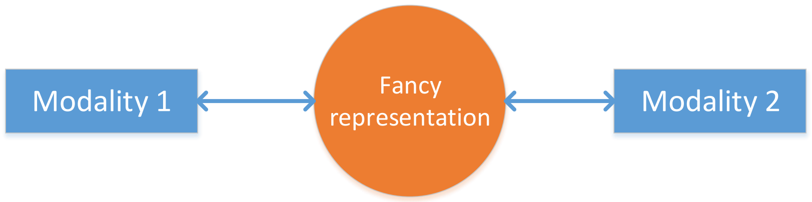
- Identify the 5 technical challenges in multimodal machine learning
- Identify tasks/applications of multimodal machine learning
- Knowledge of available datasets to tackle the challenges
- Appreciation of current state-of-the-art

Research and technical challenges

Challenge 1 - Multimodal representation

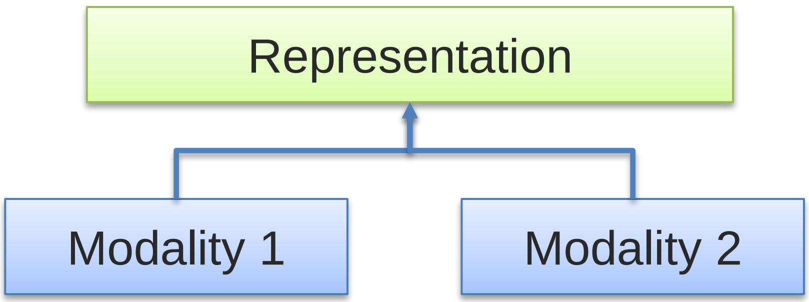
- “Computer interpretable description of the multimodal data (e.g., vector, tensor)”
  - Missing modalities
  - Heterogeneous data (symbols vs signals)
  - Static vs. sequential data
  - Different levels of noise
- Focus throughout the course, particularly weeks 3-7





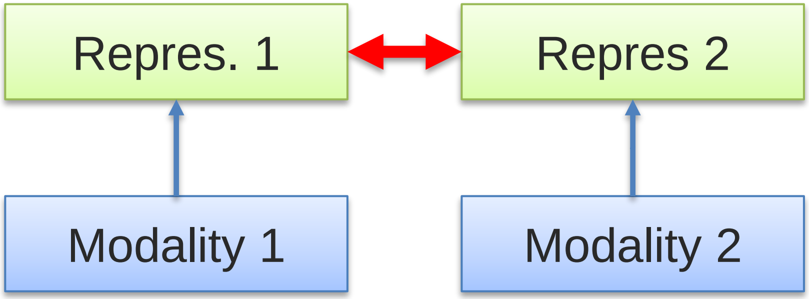
Challenge 1 - Multimodal representation types

A **Joint representations:**



- Simplest version: modality concatenation (early fusion)
- Can be learned supervised or unsupervised
- Multimodal factor analysis

B **Coordinated representations:**



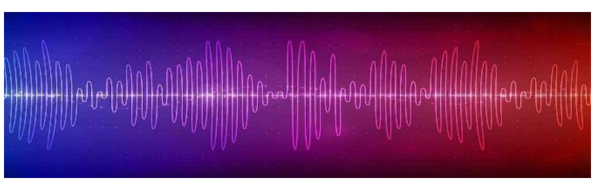
- Similarity-based methods (e.g., cosine distance)
- Structure constraints (e.g., orthogonality, sparseness)

Challenge 2 - Multimodal Translation / Mapping

- **Visual animations**

- **Image captioning**

- **Speech synthesis**



Translation / mapping:

"Process of changing data from one modality to another"

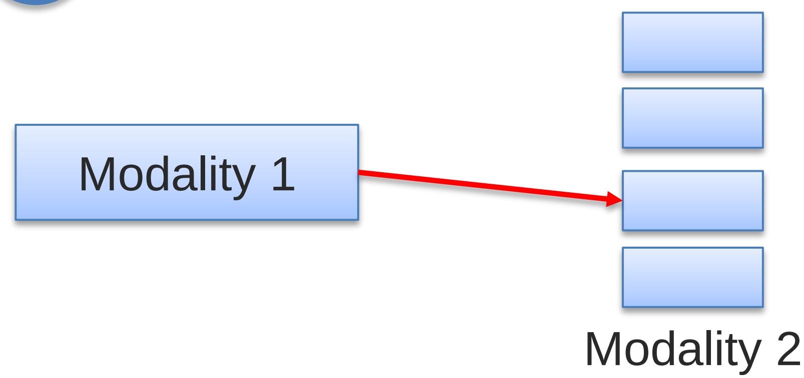
Challenges:

I. Different representations
II. Multiple source modalities
III. Open ended translations
IV. Subjective evaluation
V. Repetitive processes

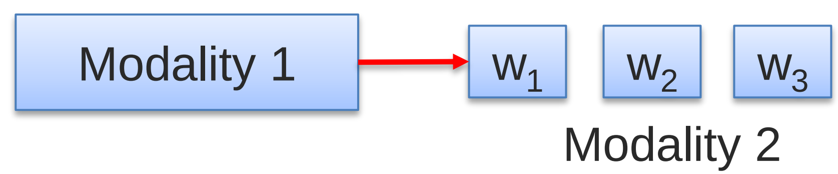
Challenge 2 - Multimodal Translation / Mapping

- Two major types
  - Example based
  - Generative
- Focus of some of the invited talks and Week 4, 5, 9

A Bounded (example based) translations:

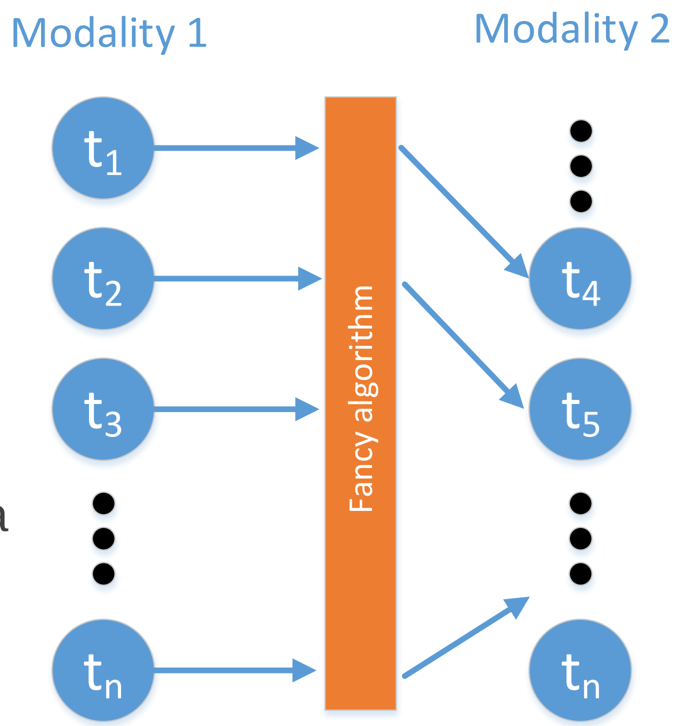


B Open-ended (generative) translations:



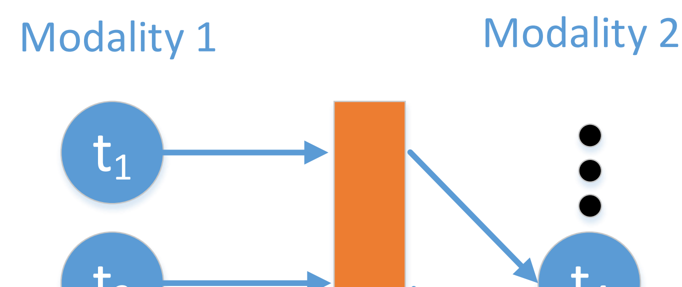
Challenge 3 - Alignment

- Alignment of (sub)components of multimodal signals
- Examples
  - Images with captions
  - Recipe steps with a how-to video
  - Phrases/words of translated sentences
- Two types
  - Explicit – alignment is the task in itself
  - Latent – alignment helps when solving a different task (for example “Attention” models)
- Focus of week 9

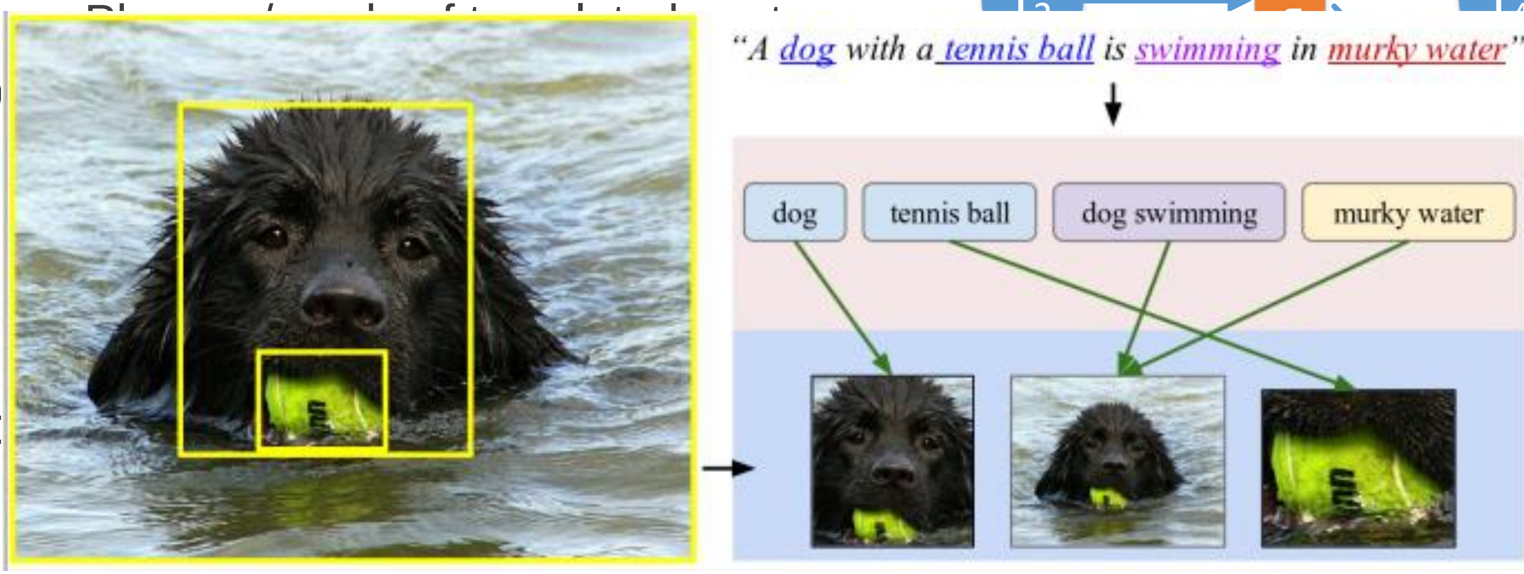


Challenge 3 - Alignment

- Alignment of (sub)components of multimodal signals
- Examples
  - Images with captions
  - Recipe steps with a how-to video

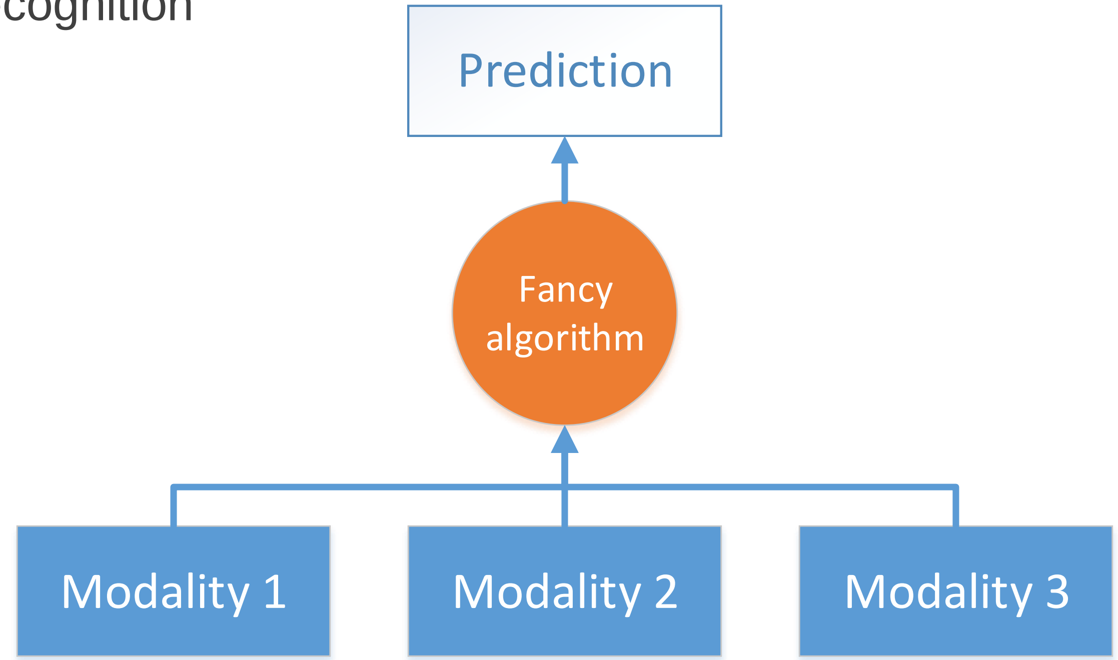


- Tw
- Fo



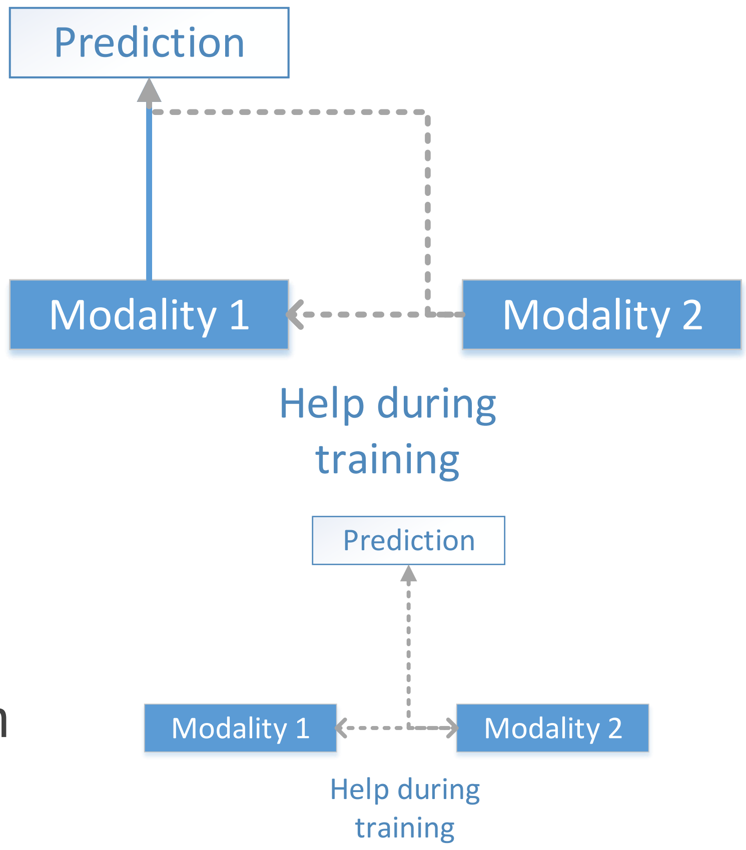
Challenge 4 - Multimodal Fusion

- Process of joining information from two or more modalities to perform a prediction
  - One of the earlier and more established problems
  - e.g. audio-visual speech recognition, multimedia event detection, multimodal emotion recognition
- Two major types
- Model Free
  - Early, late, hybrid
- Model Based
  - Kernel Methods
  - Graphical models
  - Neural networks
- Focus of Week 12



Challenge 5 – Co-learning

- How can one modality help learning in another modality?
  - One modality may have more resources
  - Bootstrapping or domain adaptation
  - Zero-shot learning
- How to alternate between modalities during learning?
  - Co-training (term introduced by Avrim Blum and Tom Mitchell from CMU)
- Transfer learning



Challenge 5 – Co-learning

- How can c
  learning in
  - One m
    resour
  - Bootst
    adapta
- How to alte
  modalities
  - Co-trai
    Avrim I
    CMU)
- Transfer le

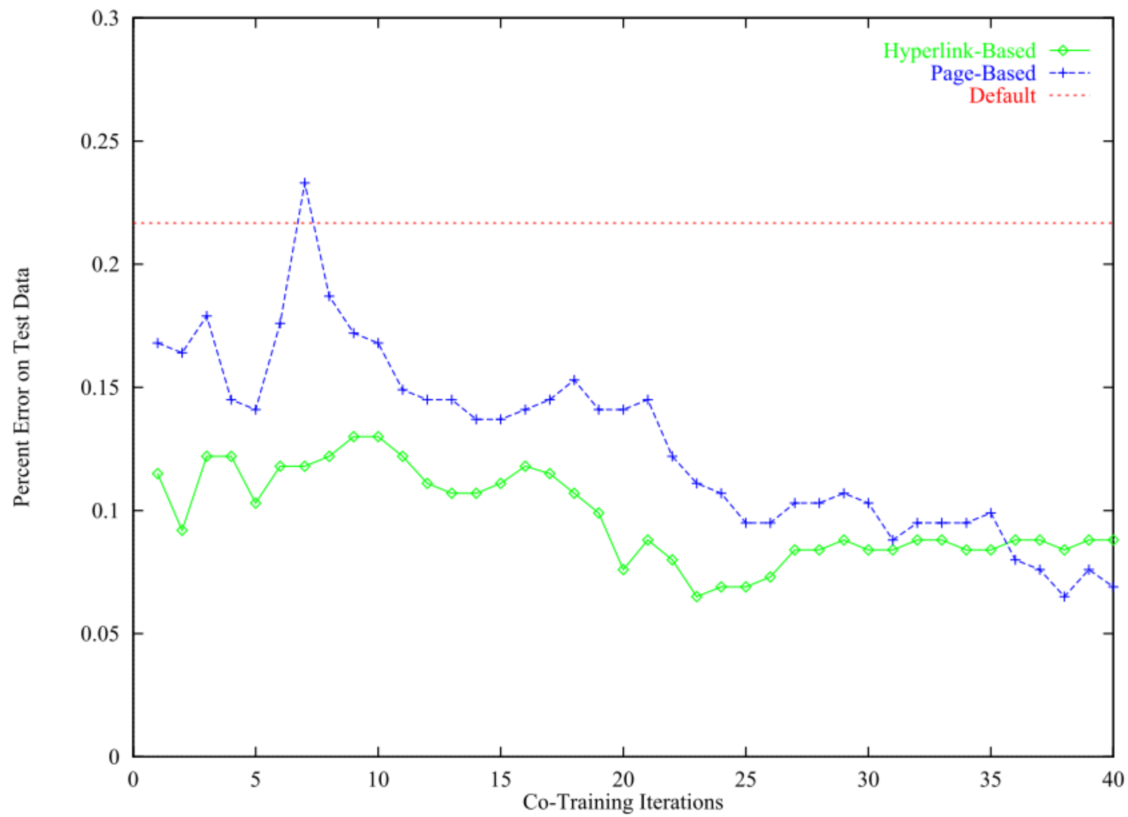


Figure 2: Error versus number of iterations for one run of co-training experiment.

lodality 2

dality 2

training

Core research problems recap

- Representations
  - Unimodal vs Multimodal
  - Joint vs Coordinated
  - Complementary (redundancy) vs Supplementary (adds extra)
- Alignment
  - Latent vs Explicit
- Translation
  - Example based vs Generative
- Fusion
  - Model free vs Model-full
- Co-learning

Actual tasks and datasets

Real world tasks tackled by MMML

- Affect recognition
  - Emotion
  - Personality traits
  - Sentiment
- Media description
  - Image captioning
  - Video captioning
  - Visual Question Answering
- Event recognition
  - Action recognition
  - Segmentation
- Multimedia information retrieval
  - Content based/Cross-media

"man in black shirt is playing guitar." "construction worker in orange safety vest is working on road." "two young girls are playing with lego toy." "boy is doing backflip on wakeboard."

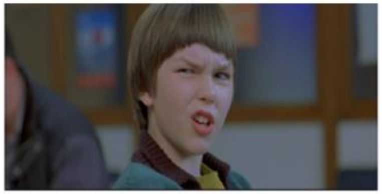
(a) answer-phone (a) get-out-car (a) fight-person (b) push-up (b) cartwheel

Affect recognition

- Emotion recognition
  - Categorical emotions – happiness, sadness, etc.
  - Dimensional labels – arousal, valence
- Personality/trait recognition
  - Not strictly affect but human behavior
  - Big 5 personality
- Sentiment analysis
  - Opinions

Affect recognition dataset 1

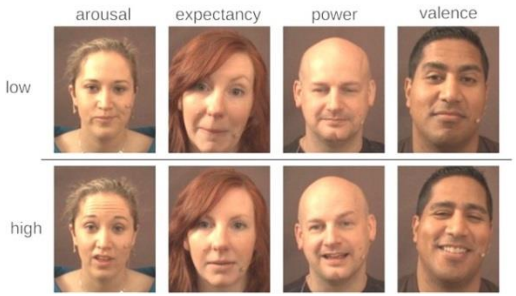
- AFEW – Acted Facial Expressions in the Wild (part of EmotiW Challenge)
- Audio-Visual emotion labels – acted emotion clips from movies
  - 1400 video sequences of about 330 subjects
- Labelled for six basic emotions + neutral
- Movies are known, can extract the subtitles/script of the scenes
- Part of EmotiW challenge





Affect recognition dataset 2

- Three AVEC challenge datasets 2011/2012, 2013/2014, 2015, 2016
- Audio-Visual emotion recognition
- Labeled for dimensional emotion (per frame)
- 2011/2012 has transcripts
- 2013/2014/2016 also includes depression labels per subject
- 2013/2014 reading specific text in a subset of videos
- 2015/2016 includes physiological data



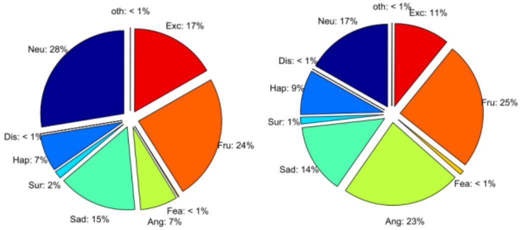
[AVEC 2011/2012]

[AVEC 2013/2014]

[AVEC 2015/2016]

Affect recognition dataset 3

- The Interactive Emotional Dyadic Motion Capture (IEMOCAP)
- 12 hours of data
- Video, speech, motion capture of face, text transcriptions
- Dyadic sessions where actors perform improvisations or scripted scenarios
- Categorical labels (6 basic emotions plus excitement, frustration) as well as dimensional labels (valence, activation and dominance)
- Focus is on speech



Affect recognition dataset 4

- Persuasive Opinion Multimedia (POM)
- 1,000 online movie review videos
- A number of speaker traits/attributes labeled – confidence, credibility, passion, persuasion, big 5…
- Video, audio and text
- Good quality audio and video recordings

Positive opinions
(5-star ratings)

Negative opinions
(1- or 2-star ratings)

Affect recognition dataset 4

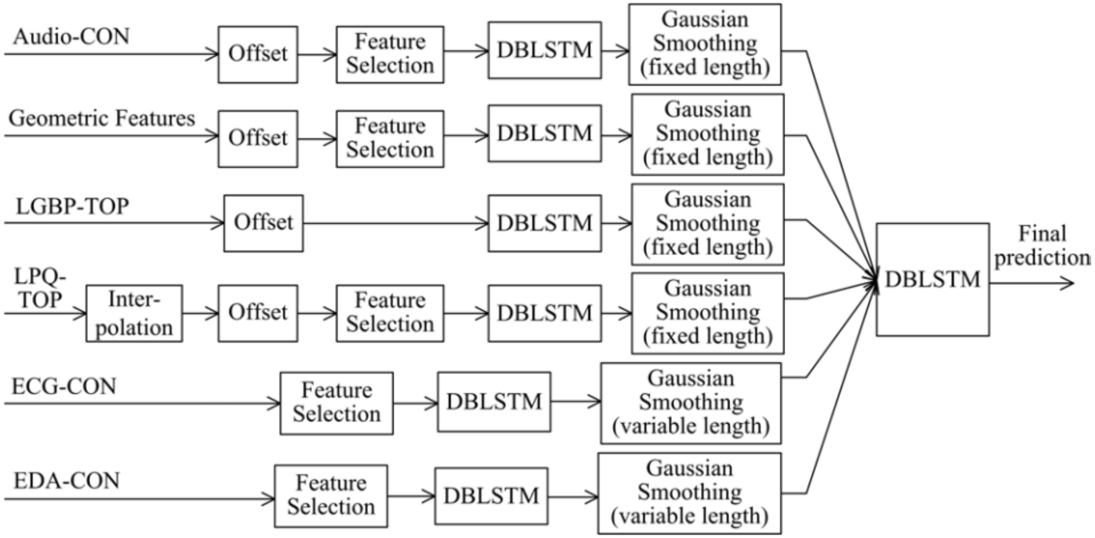
- Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos (MOSI)
- 89 speakers with 2199 opinion segments
- Audio-visual data with transcriptions
- Labels for sentiment/opinion
  - Subjective vs objective
  - Positive vs negative

Affect recognition technical challenges

- What technical problems could be addressed?
  - Fusion
  - Representation
  - Translation
  - Co-training/transfer learning
  - Alignment (after misaligning)

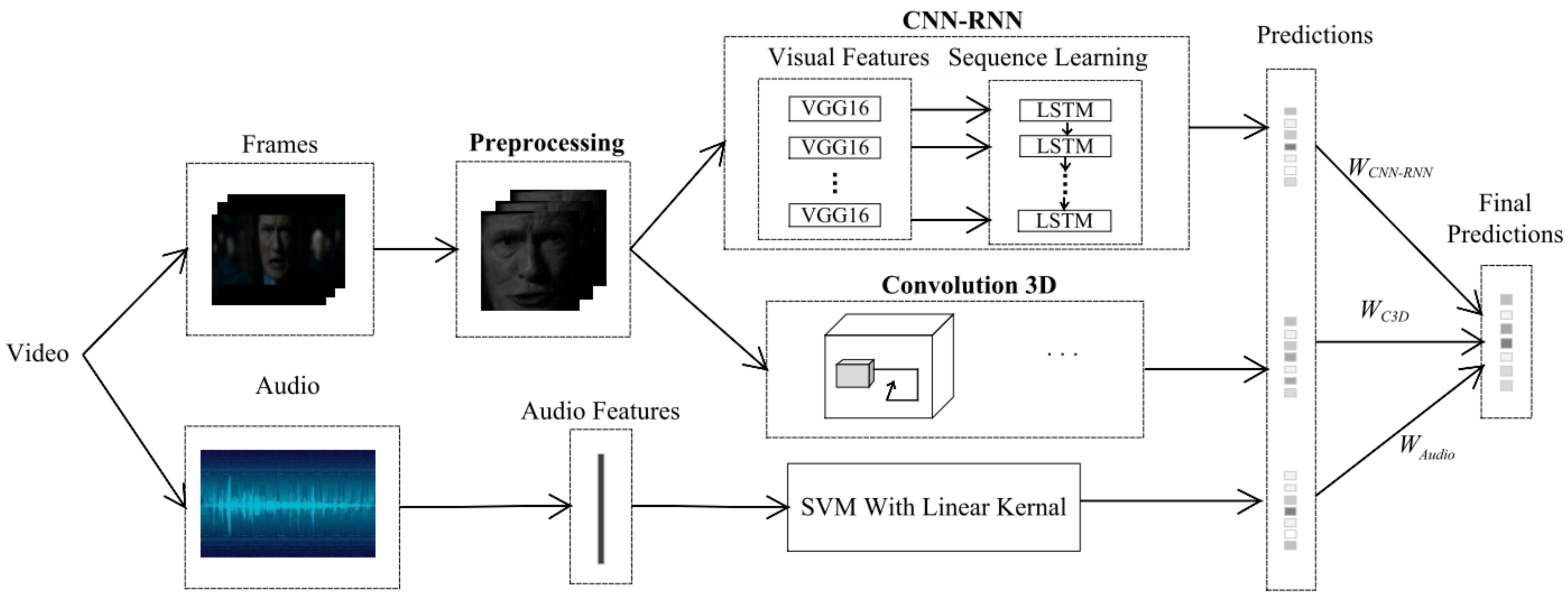
Affect recognition state of the art

- AVEC 2015 challenge winner:
  - Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks
- Will learn more about such models in week 4



Affect recognition state of the art 2

- EmotiW 2016 winner
- Video-based Emotion Recognition Using CNN-RNN and C3D Hybrid Networks



Media description

- Given a piece of media (image, video, audio-visual clips) provide a free form text description
- Earlier work looked at classes/tags/etc.


"man in black shirt is playing guitar."


"construction worker in orange safety vest is working on road."


"two young girls are playing with lego toy."


"boy is doing backflip on wakeboard."


"girl in pink dress is jumping in air."


"black and white dog jumps over bar."


"young girl in pink shirt is swinging on swing."


"man in blue wetsuit is surfing on wave."

Media description dataset 1 – MS COCO

- Microsoft Common Objects in COntext (MS COCO)
- 120000 images
- Each image is accompanied with five free form sentences describing it (at least 8 words)
- Sentences collected using crowdsourcing (Amazon Mechanical Turk)
- Also contains object detections, boundaries and keypoints

The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

Media description dataset 1 – MS COCO

- Has an evaluation server
  - Training and validation - 80K images (400K captions)
  - Testing – 40K images (380K captions), a subset contains more captions for better evaluation, these are kept privately (to avoid over-fitting and cheating)
- Evaluation is difficult as there is no one "correct" answer for describing an image in a sentence
- Given a candidate sentence it is evaluated against a set of "ground truth" sentences

State-of-the-art on MS COCO

- A challenge was done with actual human evaluations of the captions (CVPR 2015)

M1	Percentage of captions that are evaluated as better or equal to human caption.
M2	Percentage of captions that pass the Turing Test.
M3	Average correctness of the captions on a scale 1-5 (incorrect - correct).
M4	Average amount of detail of the captions on a scale 1-5 (lack of details - very detailed).
M5	Percentage of captions that are similar to human description.

State-of-the-art on MS COCO

- A challenge was done with actual human evaluations of the captions (CVPR 2015)

	M1	M2	M3	M4	M5
Human[5]	0.638	0.675	4.836	3.428	0.352
Google[4]	0.273	0.317	4.107	2.742	0.233
MSR[8]	0.268	0.322	4.137	2.662	0.234
Montreal/Toronto[10]	0.262	0.272	3.932	2.832	0.197
MSR Captivator[9]	0.250	0.301	4.149	2.565	0.233
Berkeley LRCN[2]	0.246	0.268	3.924	2.786	0.204
m-RNN[15]	0.223	0.252	3.897	2.595	0.202
Nearest Neighbor[11]	0.216	0.255	3.801	2.716	0.196

State-of-the-art on MS COCO

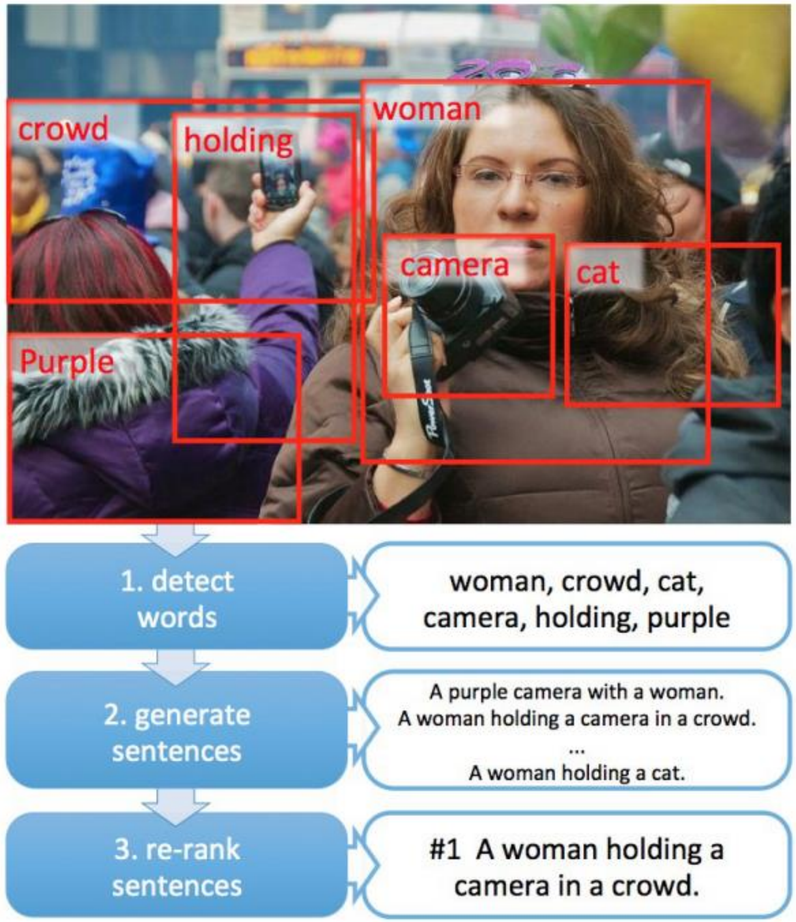
	CIDEr-D	Meteor	ROUGE-L	BLEU-1	BLEU-2
Google[4]	0.943	0.254	0.53	0.713	0.542
MSR Captivator[9]	0.931	0.248	0.526	0.715	0.543
m-RNN[15]	0.917	0.242	0.521	0.716	0.545
MSR[8]	0.912	0.247	0.519	0.695	0.526
Nearest Neighbor[11]	0.886	0.237	0.507	0.697	0.521
m-RNN (Baidu/ UCLA)[16]	0.886	0.238	0.524	0.72	0.553
Berkeley LRCN[2]	0.869	0.242	0.517	0.702	0.528
Human[5]	0.854	0.252	0.484	0.663	0.469

State-of-the-art on MS COCO

- Currently best results are:
  - Google - **Show and Tell: A Neural Image Caption Generator**
  - MSR – **From Captions to Visual Concepts and Back**
- Nearest neighbor does surprisingly well on the automatic evaluation benchmarks
- Humans perform badly on automatic evaluation metrics (shows something about the metrics)

State-of-the-art on MS COCO

- Currently b



oup of people
oping at an
loor market.

re are many
etables at the
stand.

metrics (sh … metrics)

Media description dataset 2 - Video captioning

- MPII Movie Description dataset
  - [A Dataset for Movie Description]
- Montréal Video Annotation dataset
  - [Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research]

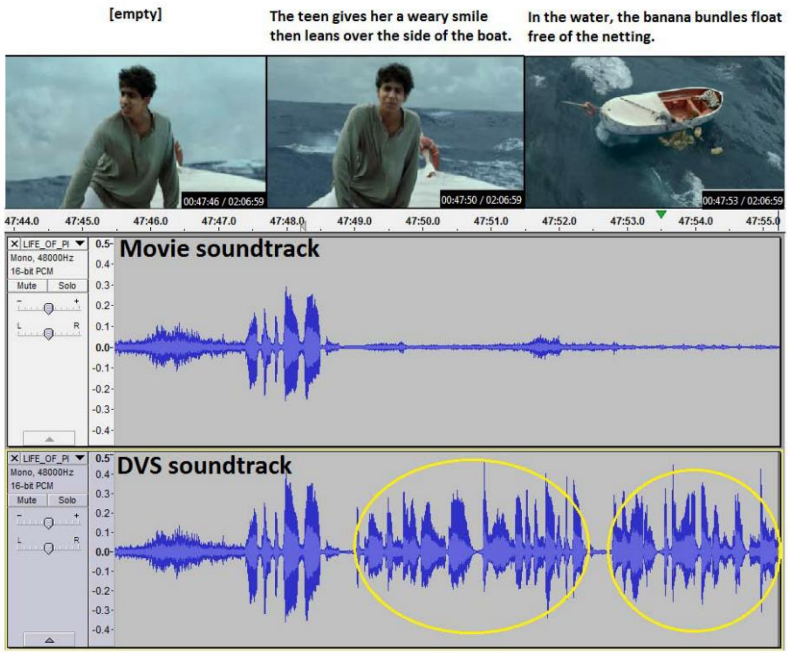
AD: Abby gets in the basket.

Mike leans over and sees how high they are.

Abby clasps her hands around his face and kisses him passionately.

Media description dataset 2 - Video captioning

- Both based on audio descriptions for the blind (Descriptive Video Service - DVS tracks)
- MPII – 70k clips (~4s) with corresponding sentences from 94 movies
- Montréal – 50k clips (~6s) with corresponding sentences from 92 movies
- Not always well aligned
- Quite noisy labels
- Single caption per clip

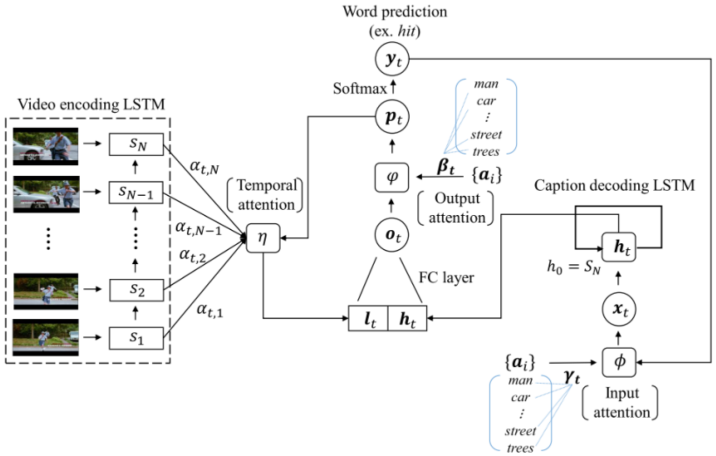


Media description dataset 2 - Video captioning

- Large Scale Movie Description and Understanding Challenge (LSMDC)
- Combines both of the datasets and provides three challenges
  - Movie description
  - Movie annotation and Retrieval
  - Movie Fill-in-the-blank
- Nice challenge, but beware
  - Need a lot of computational power
  - Processing will take space and time

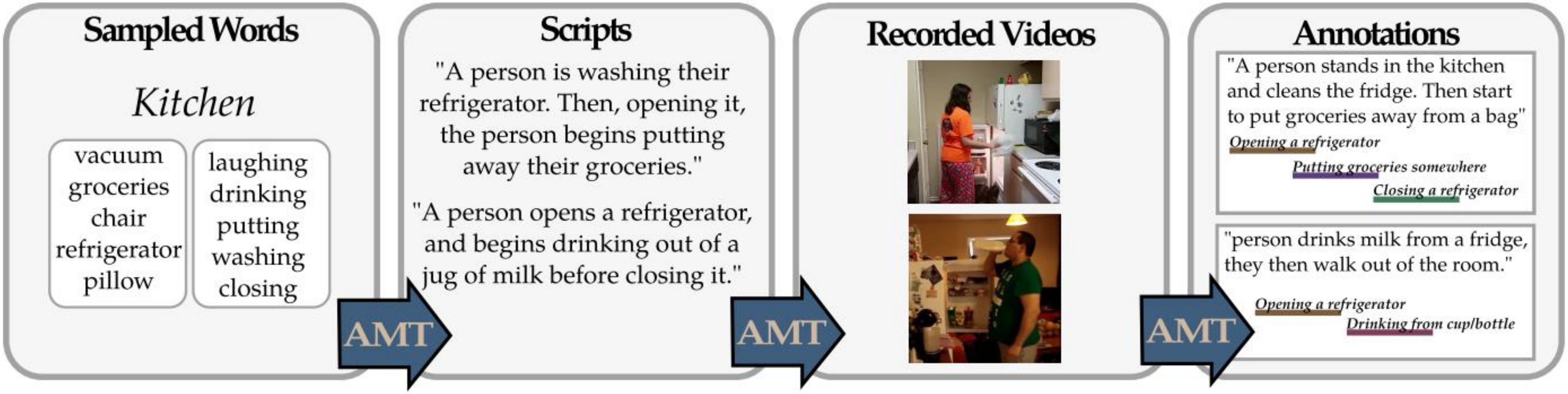
Video description state-of-the-art

- Good source for results - Describing and Understanding Video & The Large Scale Movie Description Challenge (LSMDC), hosted at ECCV 2016 and ICCV 2015
- Video Captioning and Retrieval Models with Semantic Attention
- Video Description by Combining Strong Representation and a Simple Nearest Neighbor Approach



Charades Dataset –video description dataset

- http://allenai.org/plato/charades/
- 9848 videos of daily indoors activities
- 267 different users
- Recording videos at home
- Home quality videos

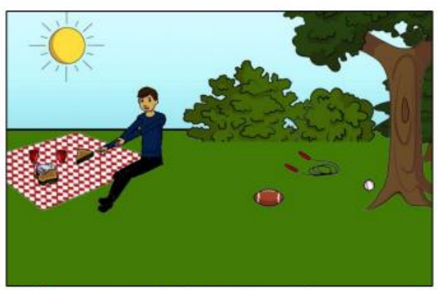


Media Description dataset 3 - VQA

- Task - Given an image and a question answer the question (http://www.visualqa.org/)

What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

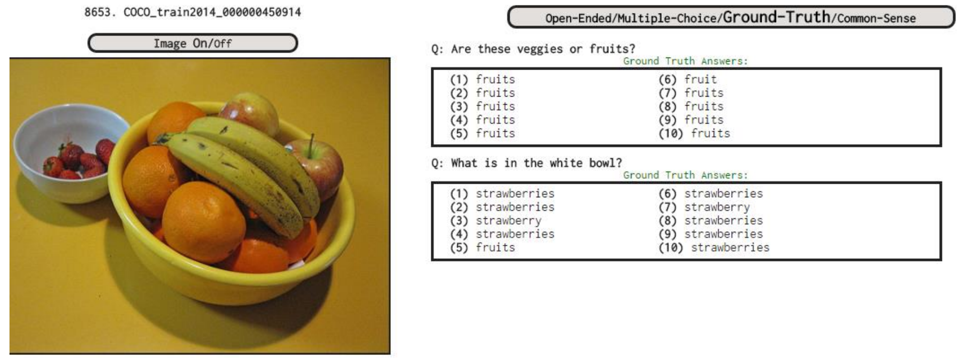


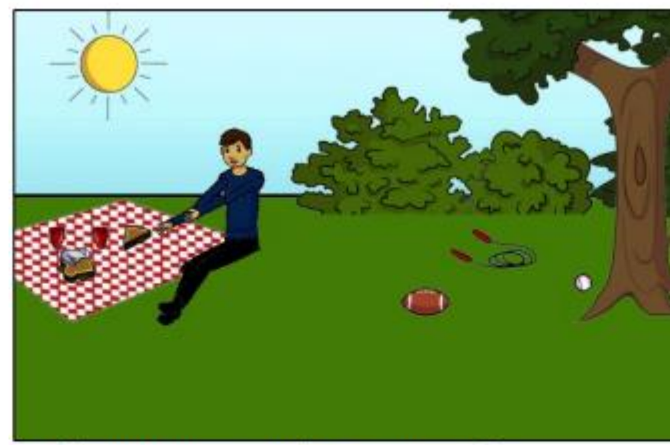
Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Media Description dataset 3 - VQA

- Real images
  - 200k MS COCO images
  - 600k questions
  - 6M answers
  - 1.8M plausible answers
- Abstract images
  - 50k scenes
  - 150k questions
  - 1.5M answers
  - 450k plausible answers





Is this person expecting company?
What is just under the tree?

VQA state-of-the-art

- Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding
- Winner is a representation/deep learning based model
- Currently good at yes/no question, not so much free form and counting

	By Answer Type			Overall
	Yes/No	Number	Other	
UC Berkeley & Sony[14]	83.79	38.9	58.64	66.9
Naver Labs[10]	83.78	37.67	54.74	64.89
DLAIT[5]	83.65	39.18	52.62	63.97
snubi-naverlabs[25]	83.64	38.43	51.61	63.4
POSTECH[11]	81.85	38.02	53.12	63.35
Brandeis[3]	82.53	36.54	51.71	62.8
VTComputerVison[19]	80.31	37.87	52.16	62.23
MIL-UT[7]	82.39	36.7	49.76	61.82

VQA 2.0

- Just guessing without an image lead to ~51% accuracy
  - So the V in VQA “only” adds 14% increase in accuracy
- [VQA v2.0](#) is attempting to address this
- Does not seem to be released just yet ☹

Multimodal Machine Translation

- Generate an image description in a target language, given an image an one or more descriptions in a source language
- http://www.statmt.org/wmt16/multimodal-task.html
- 30K multilingual captioned images (German and English)

1. Brick layers constructing a wall.
2. Maurer bauen eine Wand.

1. The two men on the scaffolding are helping to build a red brick wall.
2. Zwei Mauerer mauern ein Haus zusammen.

1. Trendy girl talking on her cellphone while gliding slowly down the street
2. Ein schickes Mädchen spricht mit dem Handy während sie langsam die Straße entlangschwebt.

1. There is a young girl on her cell-phone while skating.
2. Eine Frau im blauen Shirt telefoniert beim Rollschuhfahren.

(a) Translations

(b) Independent descriptions

Media description technical challenges

- What technical problems could be addressed?
  - Translation
  - Representation
  - Alignment
  - Co-training/transfer learning
  - Fusion


The man at bat readies to swing at the pitch while the umpire looks on.


A large bus sitting next to a very tall building.


AD: Abby gets in the basket.


Mike leans over and sees how high they are.


Abby clasps her hands around his face and kisses him passionately.


What color are her eyes?
What is the mustache made of?

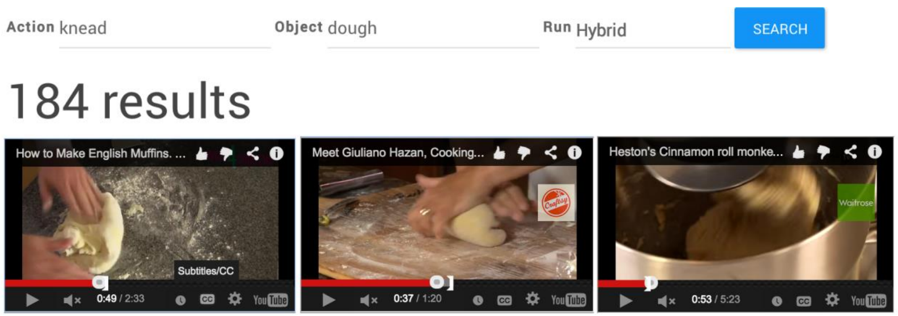

How many slices of pizza are there?
Is this a vegetarian pizza?

Event detection

- Given video/audio/ text detect predefined events or scenes
- Segment events in a stream
- Summarize videos

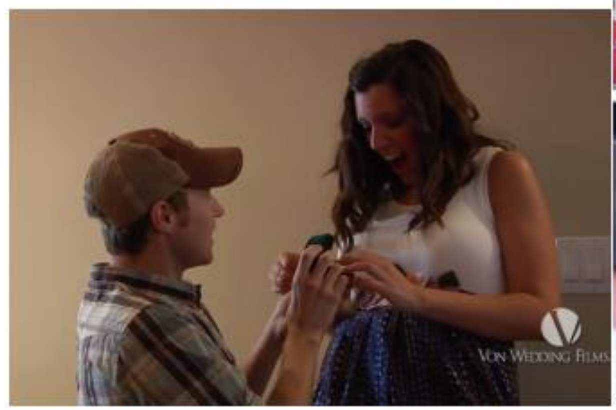
Event detection dataset 1

- What's Cooking - cooking action dataset
  - **melt butter**, **brush oil**, etc.
  - **taste**, **bake** etc.
- Audio-visual, ASR captions
  - 365k clips
  - Quite noisy
- Surprisingly many cooking datasets:
  - TACoS, TACoS Multi-Level, YouCook



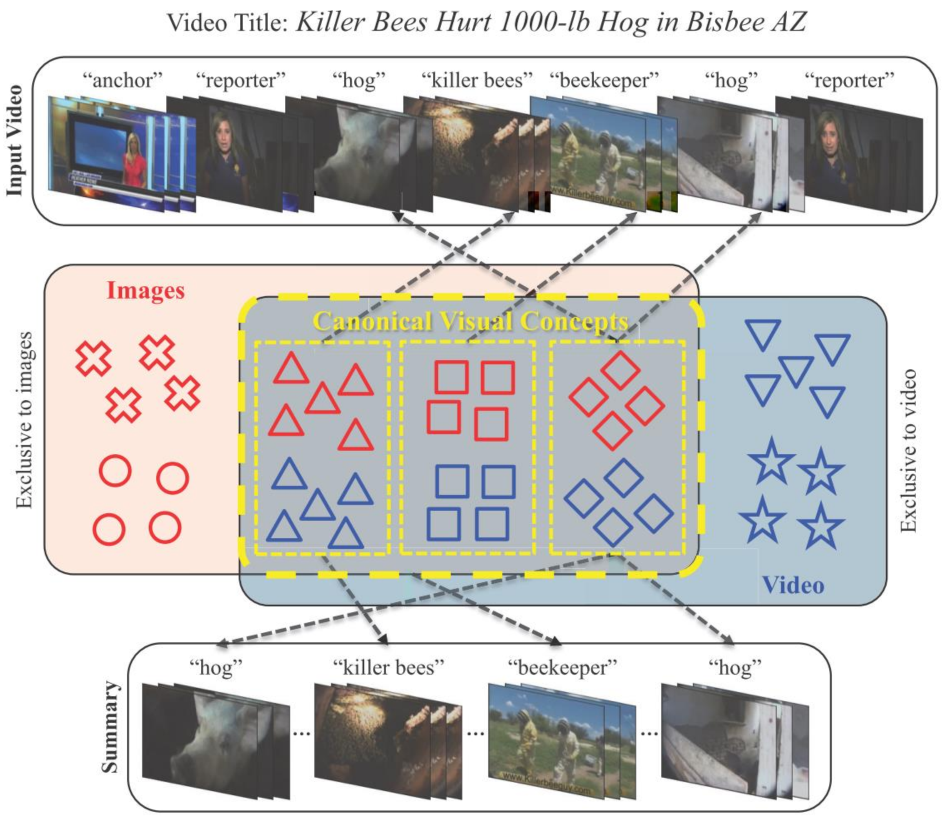
Event detection dataset 2

- Multimedia event detection
  - TrecVid Multimedia Event Detection (MED) 2010-2015
  - One of the six TrecVid tasks
  - Audio-visual data
  - Event detection



Event detection dataset 3

- [Title-based Video Summarization dataset]()
- 50 videos labeled for scene importance, can be used for summarization based on the title



Event detection dataset 4

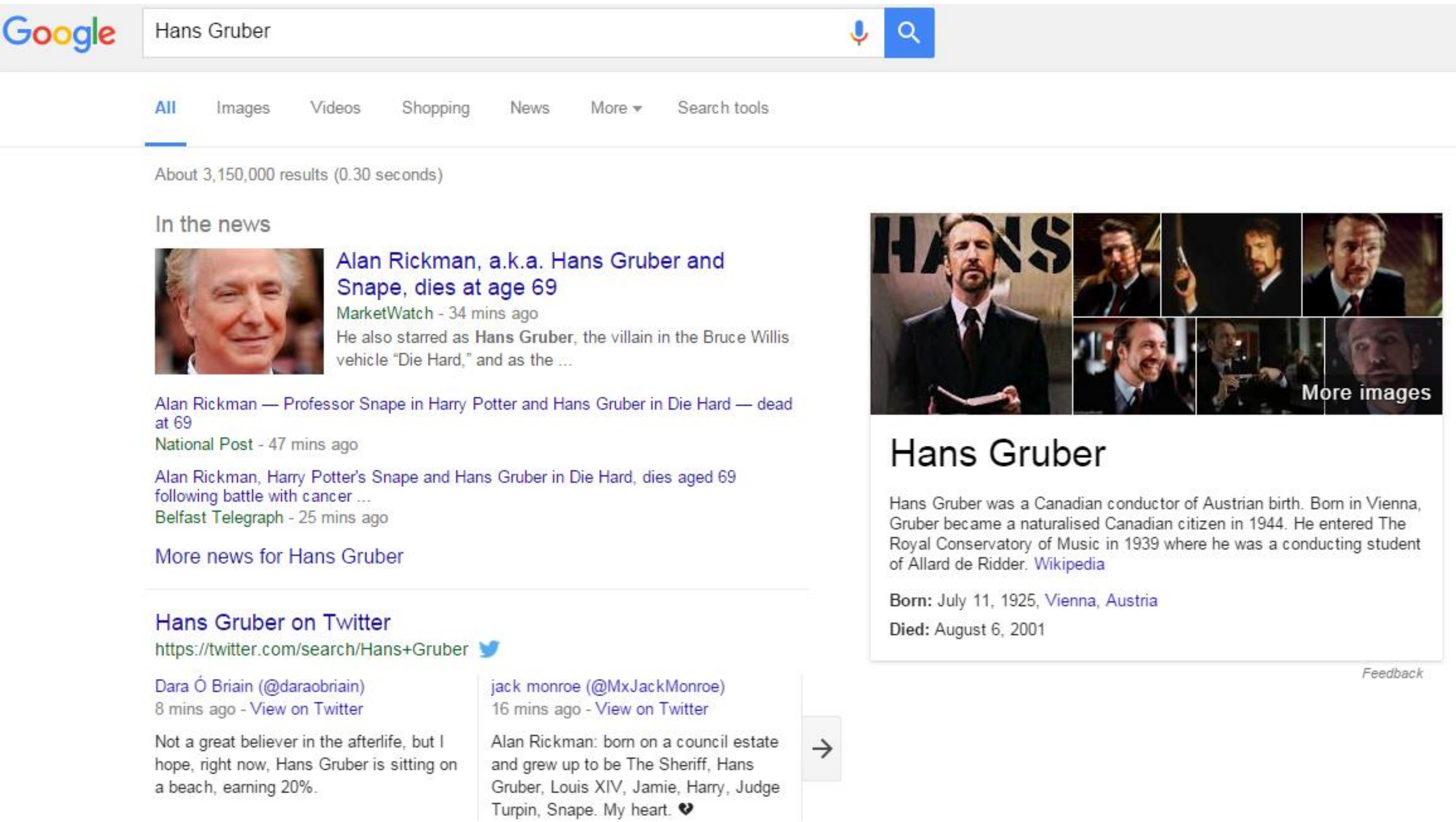
- MediaEval challenge datasets
  - Affective Impact of Movies (including Violent Scenes Detection)
  - Synchronization of Multi-User Event Media
  - Multimodal Person Discovery in Broadcast TV

Event detection technical challenges

- What technical problems could be addressed?
  - Fusion
  - Representation
  - Co-learning
  - Mapping
  - Alignment (after misaligning)

Cross-media retrieval

- Given one form of media retrieve related forms of media, given text retrieve images, given image retrieve relevant documents
- Examples:
  - Image search
  - Similar image search
- Additional challenges
  - Space and speed considerations


Google
Hans Gruber
All
Images
Videos
Shopping
News
More
Search tools
About 3,150,000 results (0.30 seconds)
In the news
Alan Rickman, a.k.a. Hans Gruber and Snape, dies at age 69
MarketWatch - 34 mins ago
He also starred as Hans Gruber, the villain in the Bruce Willis vehicle "Die Hard," and as the ...
Alan Rickman — Professor Snape in Harry Potter and Hans Gruber in Die Hard — dead at 69
National Post - 47 mins ago
Alan Rickman, Harry Potter's Snape and Hans Gruber in Die Hard, dies aged 69 following battle with cancer ...
Belfast Telegraph - 25 mins ago
More news for Hans Gruber
Hans Gruber on Twitter
https://twitter.com/search/Hans+Gruber
Dara Ó Briain (@daraobriain)
8 mins ago - View on Twitter
Not a great believer in the afterlife, but I hope, right now, Hans Gruber is sitting on a beach, earning 20%.
jack monroe (@MxJackMonroe)
16 mins ago - View on Twitter
Alan Rickman: born on a council estate and grew up to be The Sheriff, Hans Gruber, Louis XIV, Jamie, Harry, Judge Turpin, Snape. My heart.
HANS
More images
Hans Gruber
Hans Gruber was a Canadian conductor of Austrian birth. Born in Vienna, Gruber became a naturalised Canadian citizen in 1944. He entered The Royal Conservatory of Music in 1939 where he was a conducting student of Allard de Ridder. Wikipedia
Born: July 11, 1925, Vienna, Austria
Died: August 6, 2001
Feedback

Cross-media retrieval datasets

- MIRFLICKR-1M
  - 1M images with associated tags and captions
  - Labels of general and specific categories
- NUS-WIDE dataset
  - 269,648 images and the associated tags from Flickr, with a total number of 5,018 unique tags;
- Yahoo Flickr Creative Commons 100M
  - Videos and images
- Wikipedia featured articles dataset
  - 2866 multimedia documents (image + text)
- Can also use image and video captioning datasets
  - Just pose it as a retrieval task

Cross-media retrieval challenges

- What technical problems could be addressed?
  - Representation
  - Translation
  - Alignment
  - Co-learning
  - Fusion

Technical issues and support

Challenges

- To those used to only dealing with text or speech
  - Space will become an issue working with image and video data
  - Some datasets are in 100s of GB (compressed)
- Memory for processing it will become an issue as well
  - Won’t be able to store it all in memory
- Time to extract features and train algorithms will also become an issue
- Plan accordingly!
  - Sometimes tricky to experiment on a laptop (might need to do it on a subset of data)

Available tools

- Use available tools in your research groups
    - Or pair up with someone that has access to them
- Find a GPU!
- We will be getting AWS credit for some extra computational power
    - Will allow for training in the cloud
- Google Cloud Platform credit as well

Google Cloud Platform

amazon web services™

Before next class

- Let us know
  - what challenge and dataset interests you
  - What computational power you have available
  - Instructions how to do this will be sent today/tomorrow - Piazza
- Will reserve 10 minutes next week for “speed dating” and finding partners for projects
- Reading group assignment
  - Representation Learning: A Review and New Perspectives
  - Questions announced on Friday
  - Answer using Gradescope (instructions will follow soon)

Reference book for the course

- Good reference source
- Is not focused on multimodal but good primer for deep learning
- http://www.deeplearningbook.org/