Language Technologies Institute

Carnegie Mellon University

# Advanced Multimodal Machine Learning

## Lecture 1.2: Challenges and applications

Louis-Philippe Morency

* Original version co-developed with Tadas Baltrusaitis

# Objectives

- Quick review of the 5 technical challenges in multimodal machine learning
- Course syllabus and weekly schedule
- Identify tasks/applications of multimodal machine learning
- Knowledge of available datasets to tackle the challenges
- Appreciation of current state-of-the-art

# Core Technical Challenges

# Core Challenges in “Deep” Multimodal ML

Representation

Alignment

Fusion

Translation

Co-Learning

**Multimodal Machine Learning: A Survey and Taxonomy**

By Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency

https://arxiv.org/abs/1705.09406

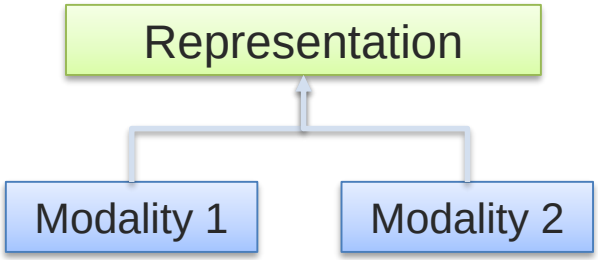
- ☑ **5 core challenges**
- ☑ **37 taxonomic classes**
- ☑ **253 referenced citations**

These challenges are non-exclusive.

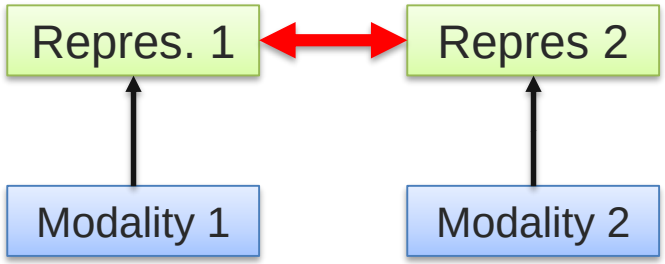
# Core Challenge 1: Representation

**Definition:** Learning how to represent and summarize multimodal data in away that exploits the complementarity and redundancy.

## A Joint representations:

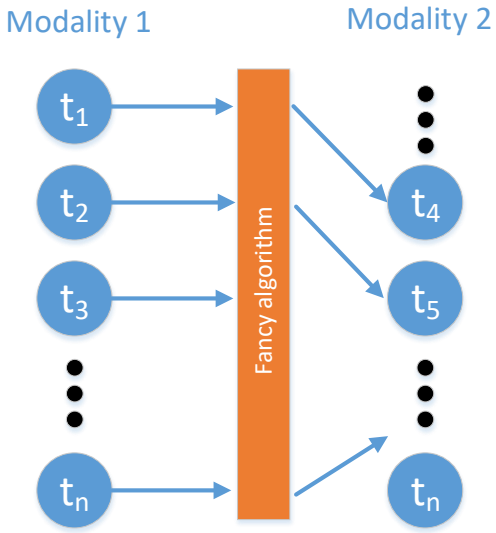



## B Coordinated representations:



# Core Challenge 2: Alignment

**Definition:** Identify the direct relations between (sub)elements from two or more different modalities.




**A Explicit Alignment**

The goal is to directly find correspondences between elements of different modalities

**B Implicit Alignment**

Uses internally latent alignment of modalities in order to better solve a different problem

# Core Challenge 3: Fusion

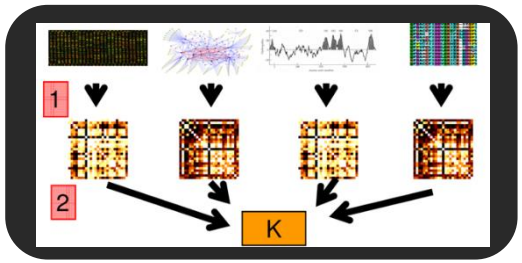
**Definition:** To join information from two or more modalities to perform a prediction task.

B **Model-Based (Intermediate) Approaches**

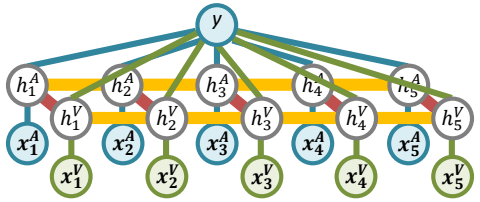
1) **Deep neural networks**

2) **Kernel-based methods**

3) **Graphical models**



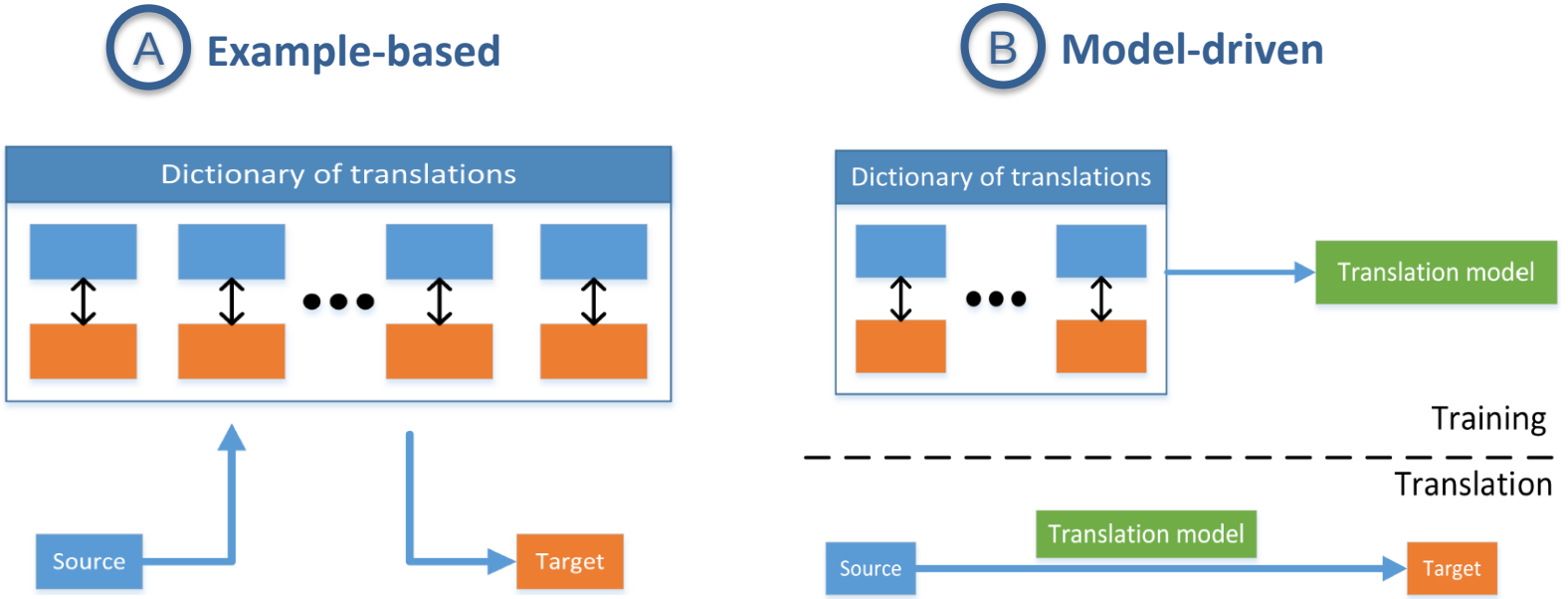

Multiple kernel learning




Multi-View Hidden CRF

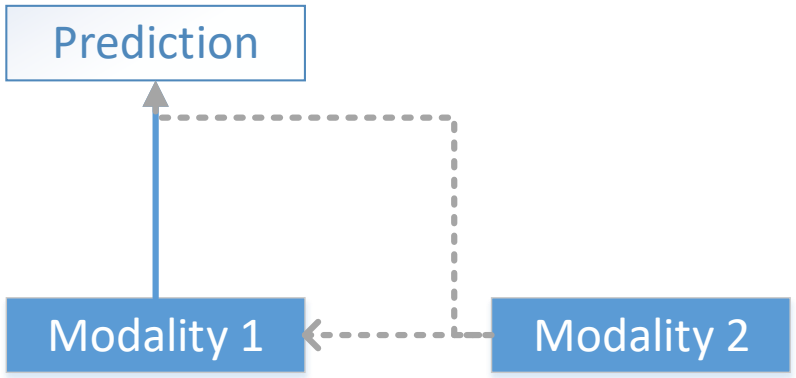
# Core Challenge 4: Translation

**Definition:** Process of changing data from one modality to another, where the translation relationship can often be open-ended or subjective.

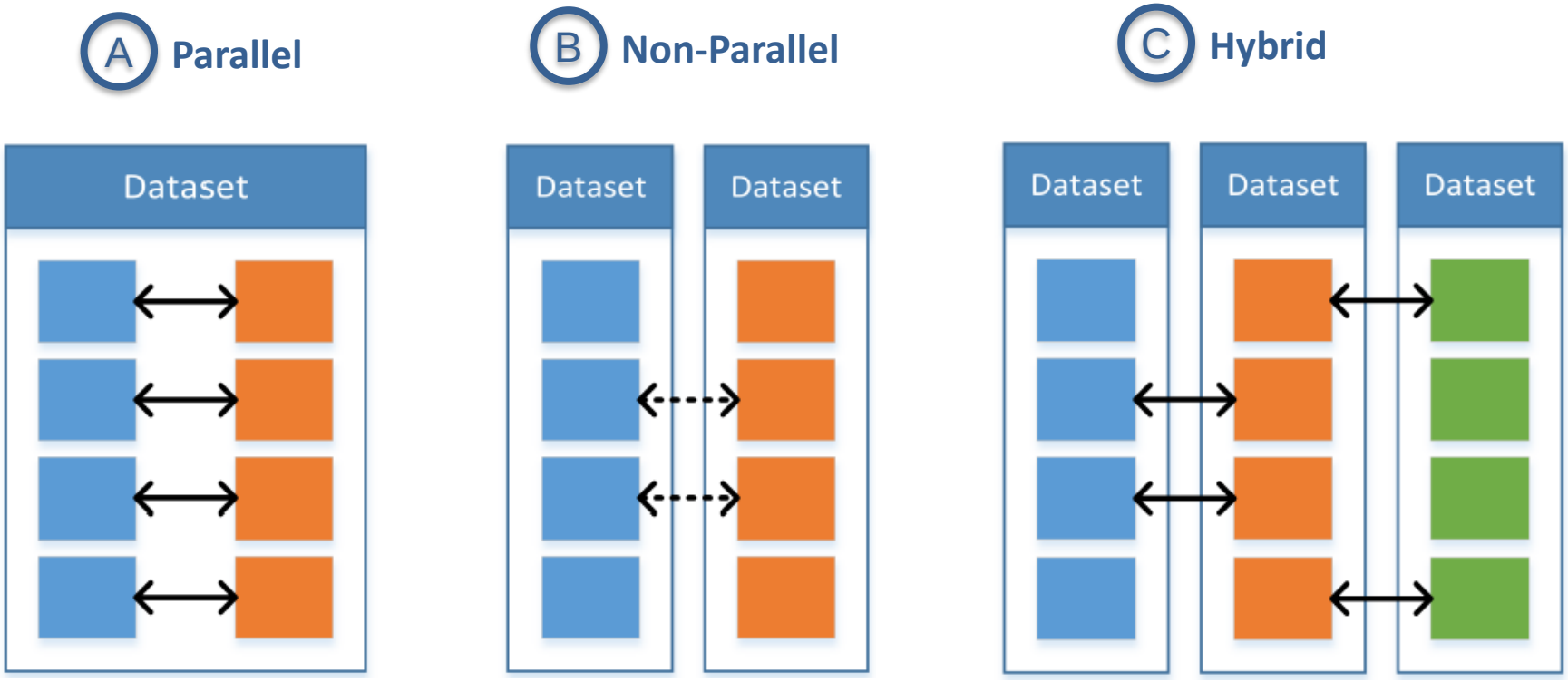


# Core Challenge 5: Co-Learning

**Definition:** Transfer knowledge between modalities, including their representations and predictive models.



# Core Challenge 5: Co-Learning



# Taxonomy of Multimodal Research [ https://arxiv.org/abs/1705.09406 ]

## Representation

- Joint
  - Neural networks
  - Graphical models
  - Sequential
- Coordinated
  - Similarity
  - Structured

## Translation

- Example-based
  - Retrieval
  - Combination
- Model-based
  - Grammar-based
  - Encoder-decoder
  - Online prediction

## Alignment

- Explicit
  - Unsupervised
  - Supervised
- Implicit
  - Graphical models
  - Neural networks

## Fusion

- Model agnostic
  - Early fusion
  - Late fusion
  - Hybrid fusion
- Model-based
  - Kernel-based
  - Graphical models
  - Neural networks

## Co-learning

- Parallel data
  - Co-training
  - Transfer learning
- Non-parallel data
  - Zero-shot learning
  - Concept grounding
  - Transfer learning
- Hybrid data
  - Bridging

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, Multimodal Machine Learning: A Survey and Taxonomy

# Multimodal Applications

[ https://arxiv.org/abs/1705.09406 ]

| | Challenges | | | | |
|---|---|---|---|---|---|
| Applications | Representation | Translation | Fusion | Alignment | Co-learning |
| **Speech Recognition and Synthesis** | | | | | |
| Audio-visual Speech Recognition | ✓ | | ✓ | ✓ | ✓ |
| (Visual) Speech Synthesis | ✓ | ✓ | | | |
| **Event Detection** | | | | | |
| Action Classification | ✓ | | ✓ | | ✓ |
| Multimedia Event Detection | ✓ | | ✓ | | ✓ |
| **Emotion and Affect** | | | | | |
| Recognition | ✓ | | ✓ | ✓ | ✓ |
| Synthesis | ✓ | ✓ | | | |
| **Media Description** | | | | | |
| Image Description | ✓ | ✓ | | ✓ | ✓ |
| Video Description | ✓ | ✓ | ✓ | ✓ | ✓ |
| Visual Question-Answering | ✓ | | ✓ | ✓ | ✓ |
| Media Summarization | ✓ | ✓ | ✓ | | |
| **Multimedia Retrieval** | | | | | |
| Cross Modal retrieval | ✓ | ✓ | | ✓ | ✓ |
| Cross Modal hashing | ✓ | | | | ✓ |

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency, Multimodal Machine Learning: A Survey and Taxonomy

# Course Syllabus

# Lecture Schedule

| Classes | Lectures |
|---|---|
| **Week 1**<br>8/29 & 8/31 | **Course introduction**<br>• Research and technical challenges<br>• Multimodal applications and datasets |
| **Week 2**<br>9/5 & 9/7 | **Basic mathematical concepts**<br>• Language, image and audio representat<br>• Loss functions and basic neural networ |
| **Week 3**<br>9/12 & 9/14 | **Convolutional neural networks and optimi**<br>• Neural network optimization<br>• Convolutional neural networks |
| **Week 4**<br>9/19 & 9/21 | **Recurrent neural networks**<br>• Backpropagation Through Time<br>• Gated networks and LSTM |

Thursday 8/31 in NSF 3305

Project preferences due on Monday night

Pre-proposal due on Sunday 9/17

# Lecture Schedule

| Classes | Lectures |
|---|---|
| **Week 5**<br>9/26 & 9/28 | **Multimodal representation learning**<br>• Multimodal auto-encoders<br>• Multimodal joint representations |
| **Week 6**<br>10/3 & 10/5 | ***First project assignment - Presentat*** |
| **Week 7**<br>10/10 & 10/12 | **Multivariate statistics and coordinated representations**<br>• Deep canonical correlation analysis<br>• Non-negative matrix factorization |
| **Week 8**<br>10/17 & 10/19 | **Multimodal alignment and attention models**<br>• Explicit alignment and dynamic time warping<br>• Implicit alignment and attention models |

Thursday in NSH 1305, 5pm-6:20pm.
Proposal due: 10/8.

# Lecture Schedule

| Classes | Lectures |
|---|---|
| **Week 9**<br>10/24 – 10/26 | **Multimodal optimization**<br>• Practical deep model optimization<br>• Variational approaches |
| **Week 10**<br>10/31 & 11/2 | **Probabilistic graphical models**<br>• Boltzmann distribution and CRFs<br>• Continuous and fully-connected CRFs |
| **Week 11**<br>11/7 & 11/9 | ***Mid-term project assignment - Pre*** |
| **Week 12**<br>11/14 & 11/16 | **Multimodal fusion and new directions**<br>• Multi-kernel learning and fusion<br>• New directions in multimodal machine learning |

Thursday in GHC-6115.
Midterm due on 11/12.

# Lecture Schedule

| Classes | Lectures |
|---|---|
| **Week 13**<br>11/21 & 11/23 | ***Thanksgiving week (+ Project preparation)*** |
| **Week 14**<br>11/28 & 11/30 | **Advanced multimodal representations**<br>• Image and video description<br>• Guest lecture |
| **Week 15**<br>**12/4** & 12/5<br>** Final ** | ***Final project assignment - Present*** |

**Monday** in GHC-6115.
Final project due: 12/10.

# Process for Selecting your Course Project

- Thursday 8/31: Lecture describing available multimodal datasets and research topics
- Monday 9/4: Submit a short paragraph listing your top 3 choices
- Tuesday 9/5: in later part of the lecture, we will do a “speed dating” session to meet teammates
- Sunday 9/17: pre-proposals are due. You should have selected your teammates and dataset.

# Course Project

$$
\begin{aligned}
i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \\
f_t &= \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \\
c_t &= f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\
o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \\
h_t &= o_t \tanh(c_t)
\end{aligned}
$$

- Pre-proposal (in 2 weeks)
  - Define your dataset and research task
- First project assignment (in 5 weeks)
  - Experiment with unimodal representations
  - Explore/discuss simple baseline model(s)
- Midterm project assignment (in 10 weeks)
  - Implement and evaluate state-of-the-art model(s)
  - Discuss new multimodal model(s)
- Final project assignment (in 14 weeks)
  - Implement and evaluate new multimodal model(s)
  - Discuss future directions

# 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




# Affect recognition

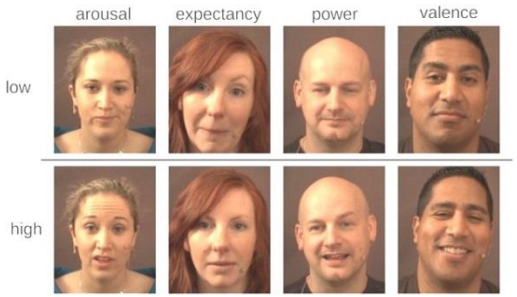
- Emotion recognition
  - Categorical emotions – happiness, sadness, etc.
  - Dimensional labels – arousal, valence
- Personality/trait recognition
  - Not strictly affect but human behavior
  - Big 5 personality
- Sentiment analysis
  - Opinions

# 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



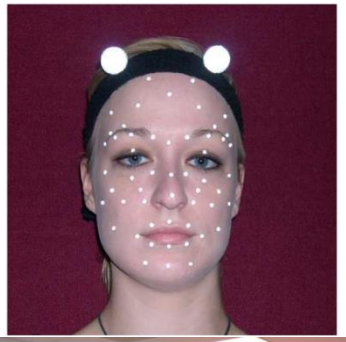

AVEC 2011/2012

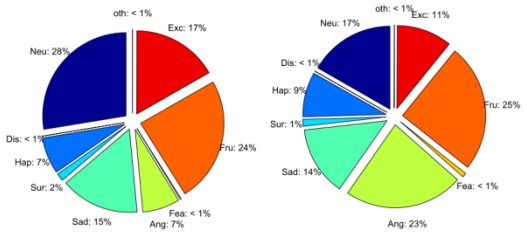
AVEC 2013/2014

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 5

- 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

# Other Affect Datasets

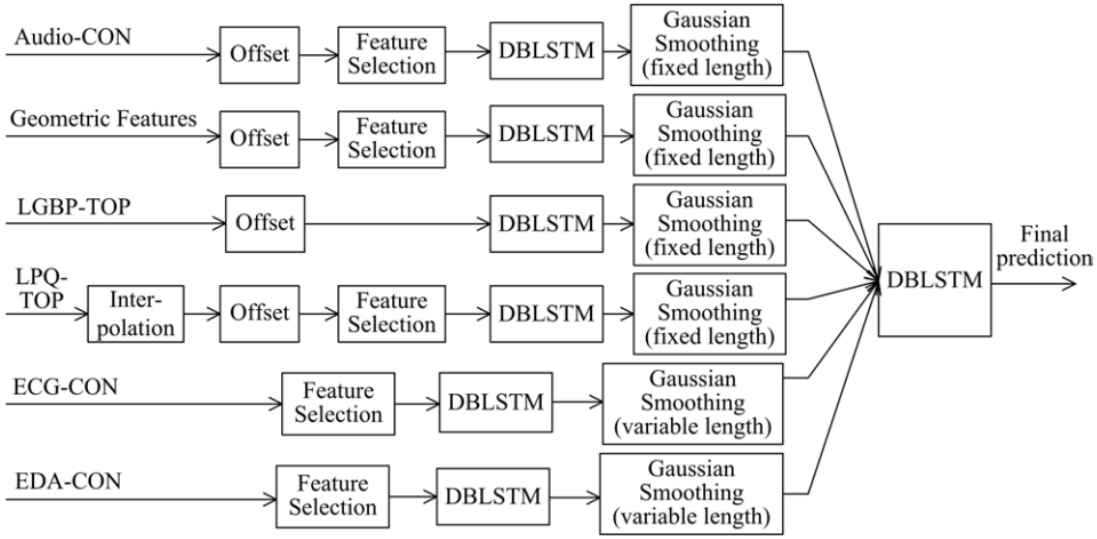
- MMDB: Multimodal Dyadic Behaviour Database

# Affect recognition technical challenges

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

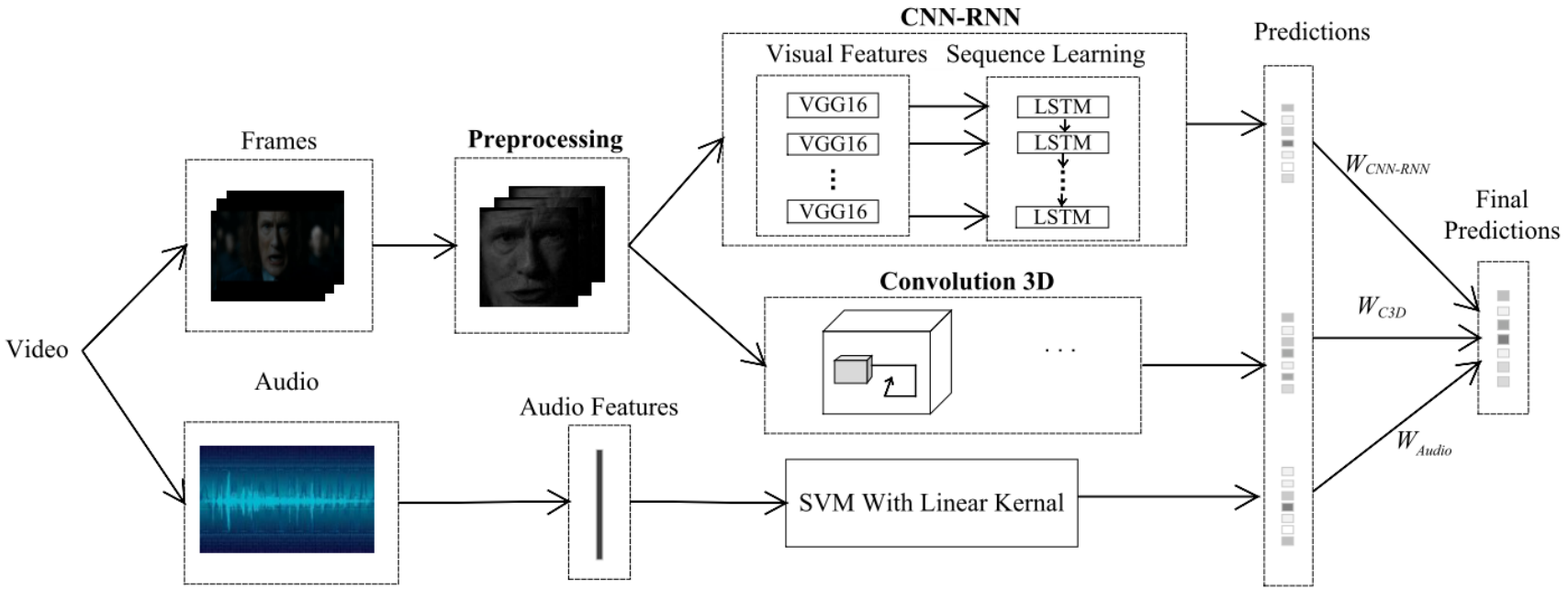
# Affect recognition Challenges

- 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 Challenges

- 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


"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 (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

# Evaluating Image Captioning Results - 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. |

# Evaluating Image Captioning Results - 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 |

# Evaluating Image Captioning Results - 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 |

# 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

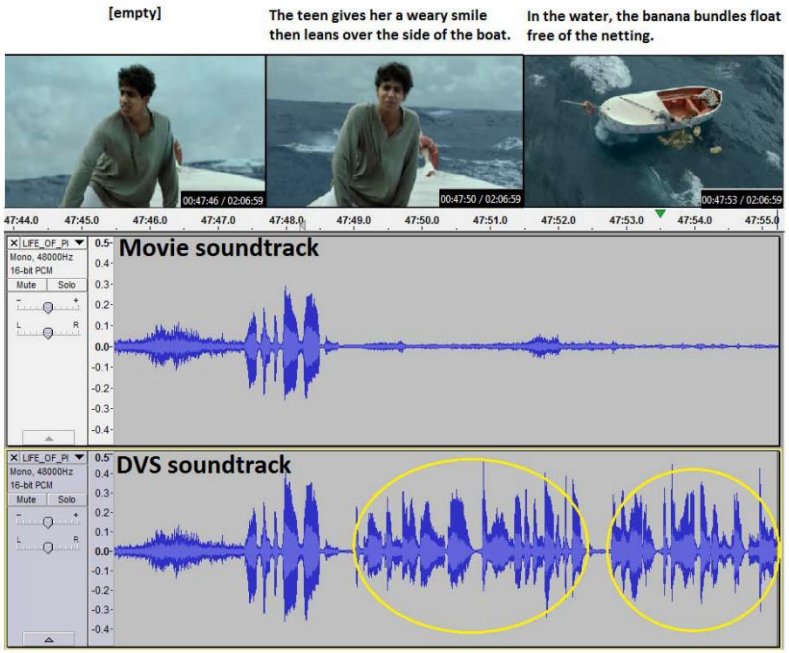
**AD**: Abby gets in the basket.

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

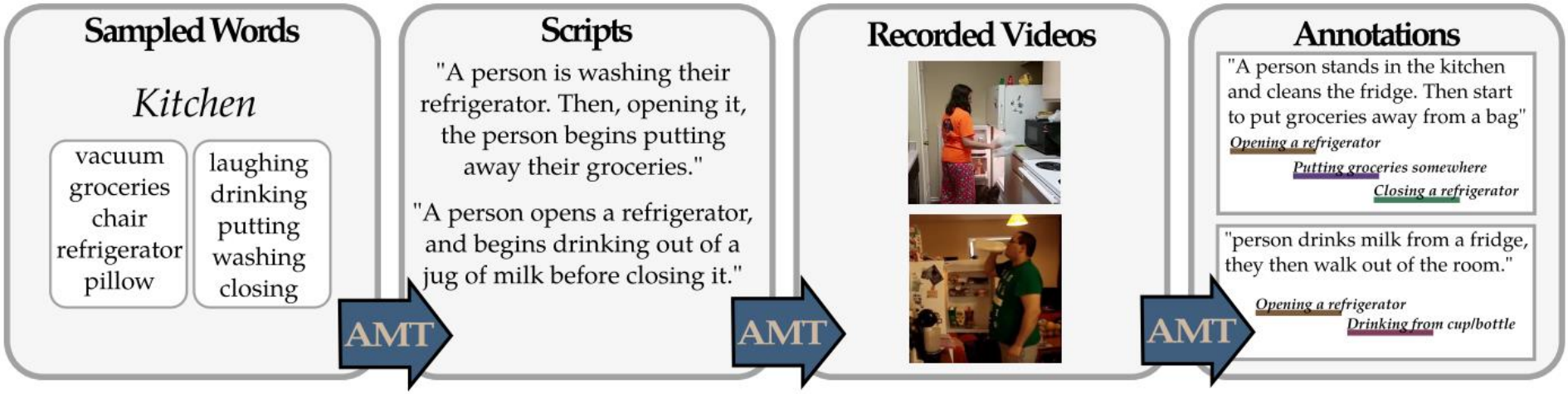


# Media description dataset 2 - Video captioning

- Large Scale Movie Description and Understanding Challenge (LSMDC) hosted at ECCV 2016 and ICCV 2015
- 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

# 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

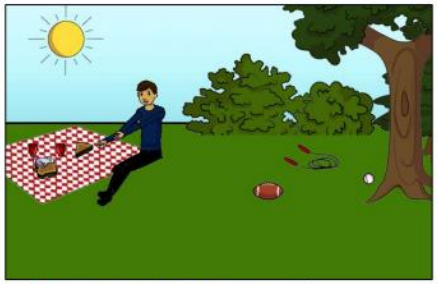


# Media Description dataset 3 - VQA

- 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?

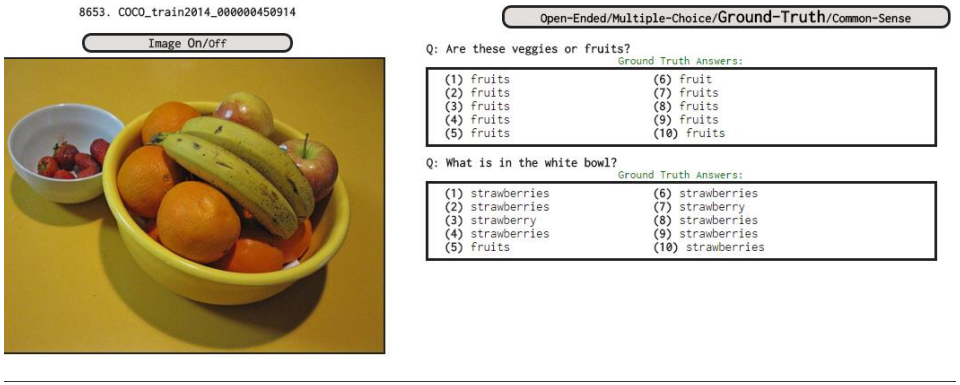


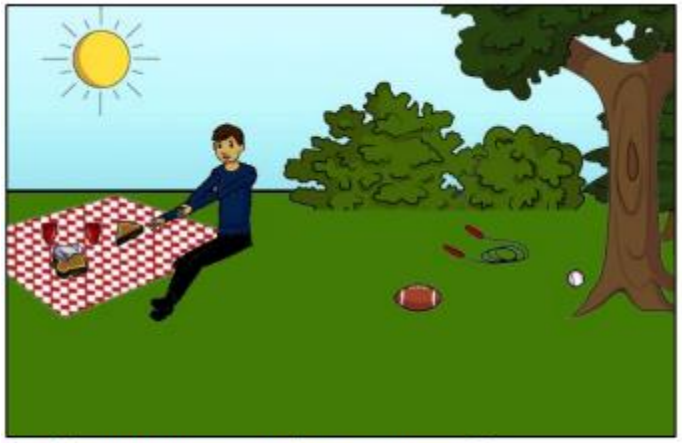
Is this person expecting company?
What is just under the tree?

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



# VQA Challenge 2016 and 2017

- Two challenges organized these past two years (link)
- 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

# Media Description – other VQA datasets

COCOQA
**Q: What is the color of the desk?**
**A: white**
**Q: What are on the white desk?**
**A: computers**

COCOQA
**Q: What is the color of the dresses?**
**A: purple**
**Q: What are three women dressed up and on?**
**A: phones**

DAQUAR
**Q: What is the object close to the wall?**
**A: whiteboard**
**Q: What is the object in front of the sofa?**
**A: table**

DAQUAR
**Q: What is the largest object?**
**A: sofa**
**Q: How many windows are there?**
**A: 2**

VQA
**Q: How many bikes are there?**
**A: 2**
**Q: What number is the bus?**
**A: 48**

VQA
**Q: How many pickles are on the plate?**
**A: 1**
**Q: What is the shape of the plate?**
**A: round**


VQA
**Q: What does the sign say?**
**A: stop**
**Q: What shape is this sign?**
**A: octagon**

VQA
**Q: What type of trees are here?**
**A: palm**
**Q: Is the skateboard airborne?**
**A: yes**

# Media Description – Other VQA datasets

- DAQUAR
  - Synthetic QA pairs based on templates
  - 12468 human question-answer pairs

- COCO-QA
  - Object, Number, Color, Location
  - Training: 78736
  - Test: 38948

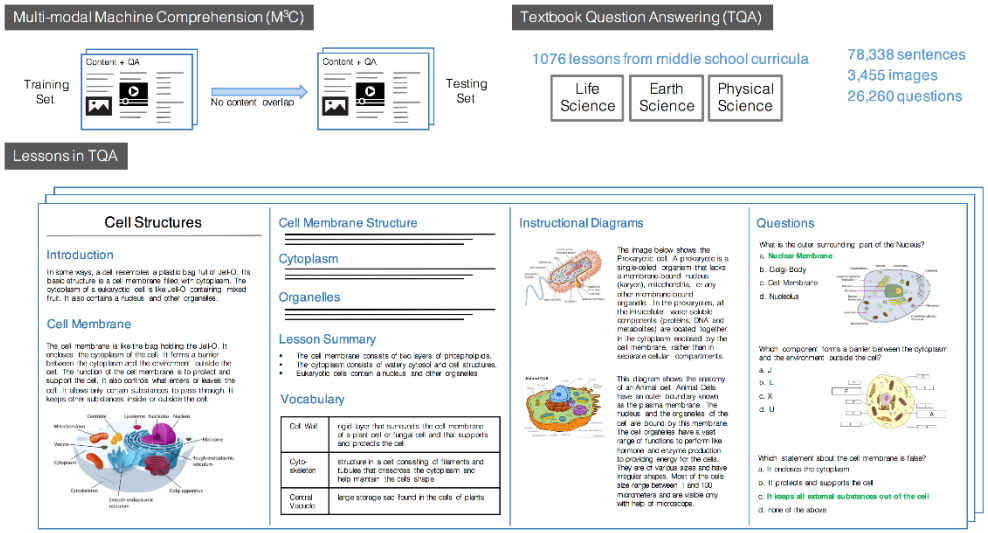
# Media Description – other VQA datasets

- Visual Madlibs
  - Fill in the blank Image Generation and Question Answering
  - 360,001 focused natural language descriptions for 10,738 images
  - collected using automatically produced fill-in-the-blank templates designed to gather targeted descriptions about: people and objects, their appearances, activities, and interactions, as well as inferences about the general scene or its broader context


1. This place is a park.
2. When I look at this picture, I feel competitive.
3. The most interesting aspect of this picture is the guys playing shirtless.
4. One or two seconds before this picture was taken, the person caught the frisbee.
5. One or two seconds after this picture was taken, the guy will throw the frisbee.
6. Person A is wearing blue shorts.
7. Person A is in front of person B.
8. Person A is blocking person B.
9. Person B is a young man wearing an orange hat.
10. Person B is on a grassy field.
11. Person B is holding a frisbee.
12. The frisbee is white and round.
13. The frisbee is in the hand of the man with the orange cap.
14. People could throw the frisbee.
15. The people are playing with the frisbee.

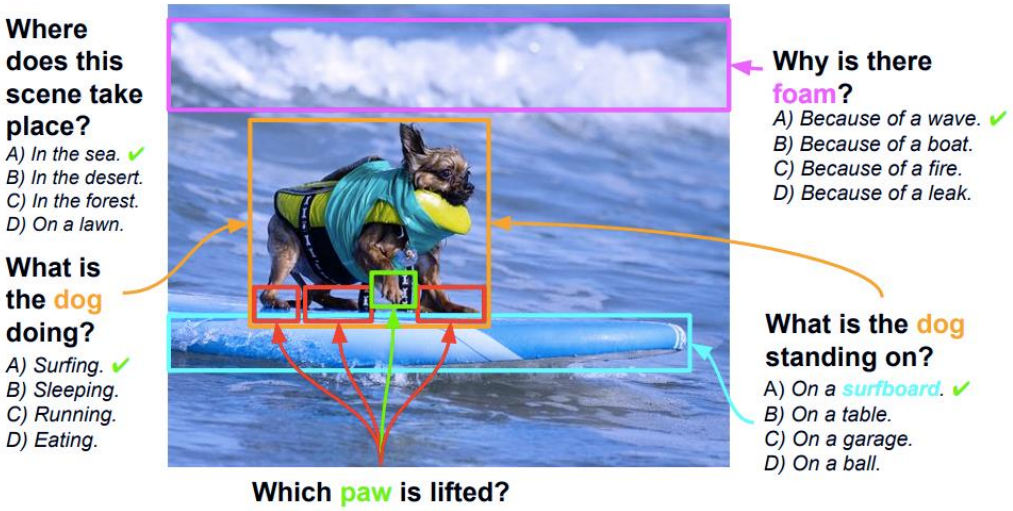
# Media Description – other VQA datasets

- Textbook Question Answering
  - Multi-Modal Machine Comprehension
  - Context needed to answer questions provided and composed of both text and images
  - 78338 sentences, 3455 images
  - 26260 questions



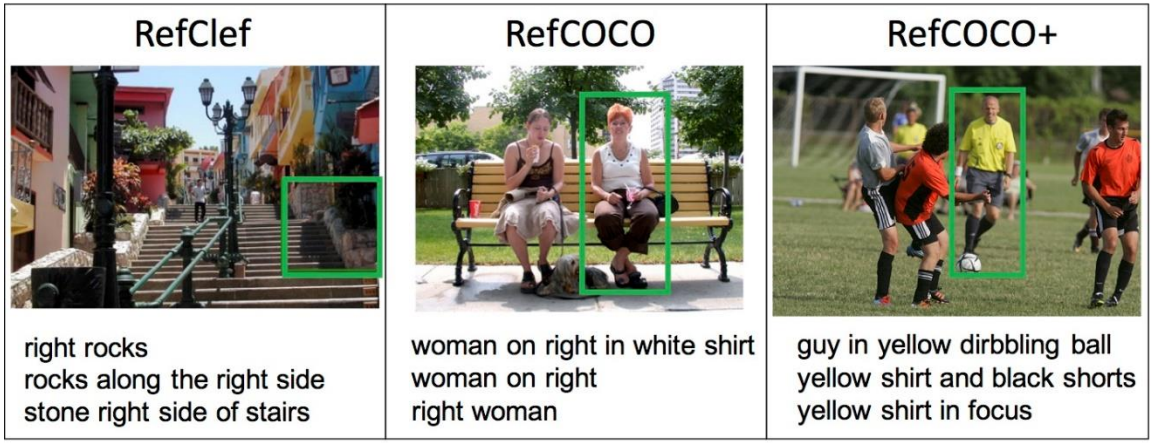
# Media Description – other VQA datasets

- Visual7W
  - Grounded Question Answering in Images
  - 327,939 QA pairs on 47,300 COCO images
  - 1,311,756 multiple-choices, 561,459 object groundings, 36,579 categories
  - what, where, when, who, why, how and which



# Media Description – Referring Expression datasets

- Referring Expressions:
  - Generation (Bounding Box to Text) and Comprehension (Text to Bounding Box)
  - Generate / Comprehend a noun phrase which identifies a particular object in an image
  - Many datasets!
    - RefClef
    - RefCOCO (+, g)
    - GRef



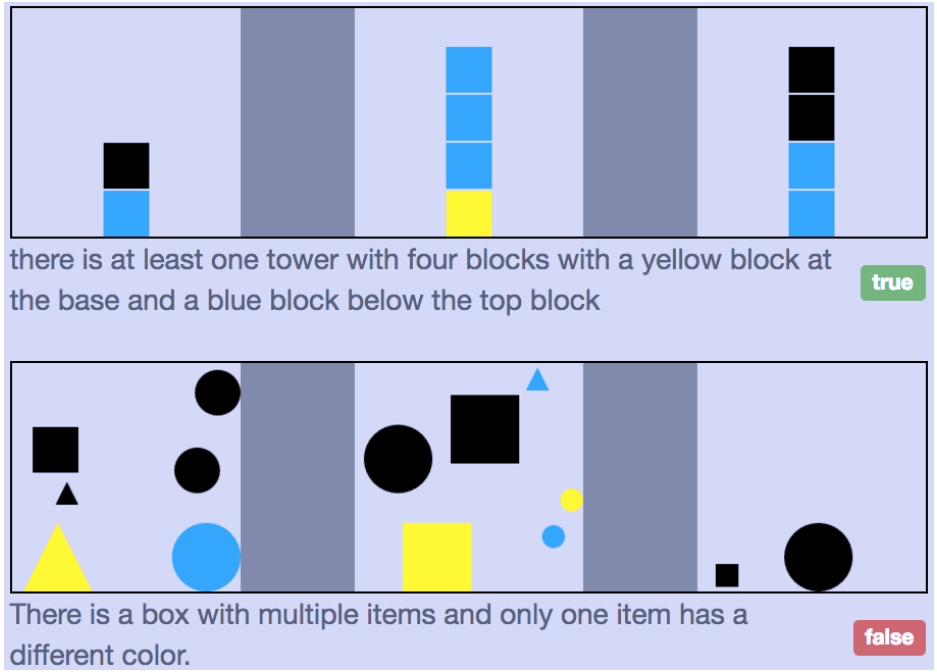
# Media Description - Referring Expression datasets

- GuessWhat?!
  - Cooperative two-player guessing game for language grounding
  - Locate an unknown object in a rich image scene by asking a sequence of questions
  - 821,889 questions+answers
  - 66,537 images and 134,073 objects

| Questioner | Oracle |
|---|---|
| Is it a vase? | Yes |
| Is it partially visible? | No |
| Is it in the left corner? | No |
| Is it the turquoise and purple one? | Yes |

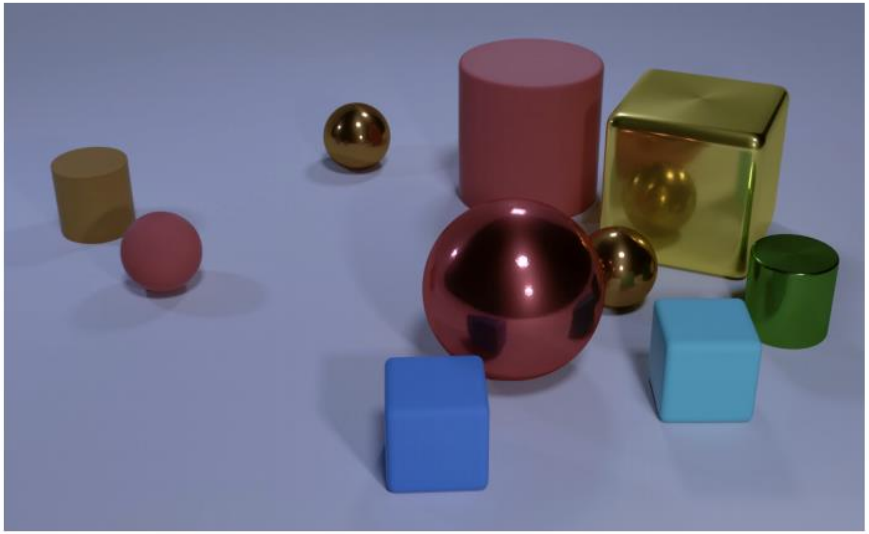
# Media Description – Visual Reasoning

- Cornell NLVR
  - 92,244 pairs of natural language statements grounded in synthetic images
  - Determine whether a sentence is true or false about an image



# Media Description - Visual Reasoning

- CLEVR
  - A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning
  - Tests a range of different specific visual reasoning abilities
  - Training set: 70,000 images and 699,989 questions
  - Validation set: 15,000 images and 149,991 questions
  - Test set: 15,000 images and 14,988 questions



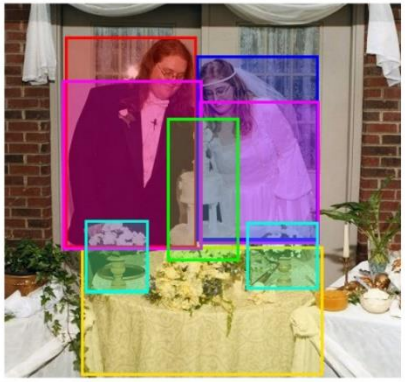
**Q:** Are there an equal number of large things and metal spheres?
**Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
**Q:** How many objects are either small cylinders or metal things?

# Media Description - other datasets

- Flickr30k Entities
  - Region-to-Phrase Correspondences for Richer Image-to-Sentence Models
  - 158k captions
  - 244k coreference chains
  - 276k manually annotated bounding boxes

A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

During a gay pride parade in an Asian city, some people hold up rainbow flags to show their support.
A group of youths march down a street waving flags showing a color spectrum.
Oriental people with rainbow flags walking down a city street.
A group of people walk down a street waving rainbow flags.
People are outside waving flags .



A couple in their wedding attire stand behind a table with a wedding cake and flowers.
A bride and groom are standing in front of their wedding cake at their reception.
A bride and groom smile as they view their wedding cake at a reception.
A couple stands behind their wedding cake.
Man and woman cutting wedding cake.

# 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.

(a) Translations

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

(b) Independent descriptions

# Other Media Description Datasets

- MVSO: Multilingual Visual Sentiment Ontology. There are multiple derivatives of this as well
- Dataset from the AAAI'16 paper 'Listen, Attend, and Walk: Neural Mapping of Navigational Instructions to Action Sequences'. The data is provided with the code here.
- Visual Relation dataset: learning relations between objects based on language priors.
- Visual genome is another dataset in this area.
- Pinterest : Contains 300 million sentences describing over 40 million 'pins'

# 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.

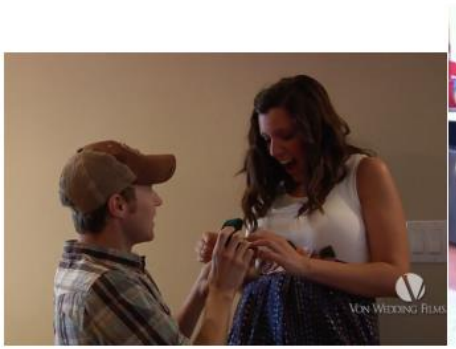
Abby clasps her hands around his face and kisses him passionately.

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

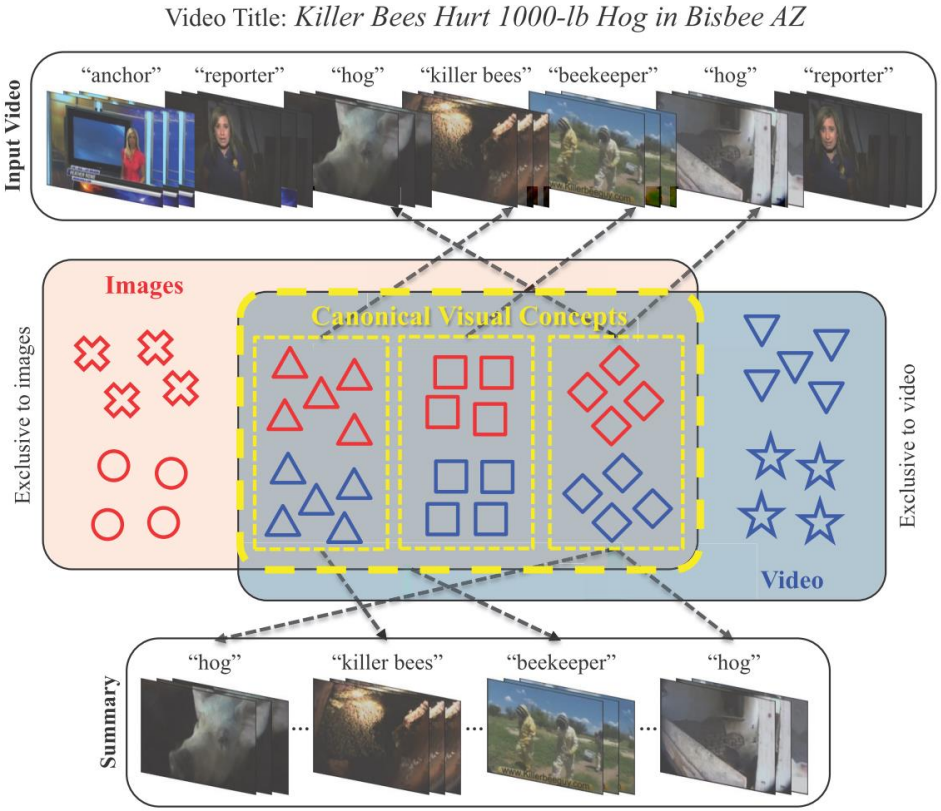
- What's Cooking - cooking action dataset
  - **melt butter**, **brush oil**, etc.
  - **taste**, **bake** etc.
- Audio-visual, ASR captions
  - 365k clips
  - Quite noisy
- Surprisingly many cooking datasets:
  - TACoS, TACoS Multi-Level, YouCook

# 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

# 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

# Other Multimodal Datasets

- 1) Youtube 8M
  - https://research.google.com/youtube8m/
- 2) Youtube Bounding Boxes
  - https://research.google.com/youtube-bb/
- 3) Youtube Open Images
  - https://research.googleblog.com/2016/09/introducing-open-images-dataset.html
- 4) YFCC 100M
  - https://webscope.sandbox.yahoo.com/catalog.php?datatype=i&did=67
- 5) VIST
  - http://visionandlanguage.net/VIST/

# 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™

# Course Project Guidelines

- Dataset should have at least two modalities:
  - Natural language and visual/images
- Teams of 3 or 4 students are preferred
  - No individual projects
- The project should explore algorithmic novelty
- Possible venues for your final report:
  - NAACL 2018, ACL 2018, IJCAI 2017, ICML 2018
  - We can help you these paper submissions
  - Language & vision papers have high acceptance rates at NLP conferences

# 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 15 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/