THE UNIVERSITY OF BRITISH COLUMBIA

Topics in AI (CPSC 532S):
Multimodal Learning with Vision, Language and Sound

Lecture 1: Introduction

Course **logistic**

Times: Tues & Thurs 11-12:00pm

Locations: Zoom (lectures will be recorded)

Instructor: Leonid Sigal

E-mail: lsigal@cs.ubc.ca

Office: Zoom

Course webpage: https://www.cs.ubc.ca/~lsigal/teaching20_Term2.html

Discussion: piazza.com/ubc.ca/winterterm22021/cpsc532s2012020w/home

Course **logistic**

Times: Tues & Thurs 11-12:00pm

Locations: Zoom (lectures will be recorded)

Instructor: Leonid Sigal

E-mail: lsigal@cs.ubc.ca

Office: Zoom

TA: Tanzila Rahman

E-mail: tanzila.himu@gmail.com

Suhail Mohammed

suhail33@cs.ubc.ca

Course webpage: https://www.cs.ubc.ca/~lsigal/teaching20_Term2.html

Discussion: piazza.com/ubc.ca/winterterm22021/cpsc532s2012020w/home

Course **logistic**

Times: Tues & Thurs 11-12:00pm

Locations: Zoom (lectures will be recorded)

Instructor: Leonid Sigal

E-mail: lsigal@cs.ubc.ca

Office: Zoom

TA: Tanzila Rahman

E-mail: tanzila.himu@gmail.com

Suhail Mohammed

2020 GTA Award

suhail33@cs.ubc.ca

Course webpage: https://www.cs.ubc.ca/~lsigal/teaching20_Term2.html

Discussion: piazza.com/ubc.ca/winterterm22021/cpsc532s2012020w/home

Course **logistic**

Times: Tues & Thurs 11-12:00pm

Locations: Zoom (lectures will be recorded)

While lectures **will be recorder**, I would strongly encourage you to come and listen and **participate live**. This is a grad course, discussion in class is an important component of the course.

Course webpage: https://www.cs.ubc.ca/~lsigal/teaching20_Term2.html

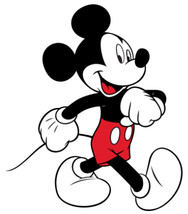
Discussion: piazza.com/ubc.ca/winterterm22021/cpsc532s2012020w/home

About **me …**

Associate Professor
2017 -

THE UNIVERSITY OF BRITISH COLUMBIA

Senior Research Scientist
2009 - 2017



Disney Research

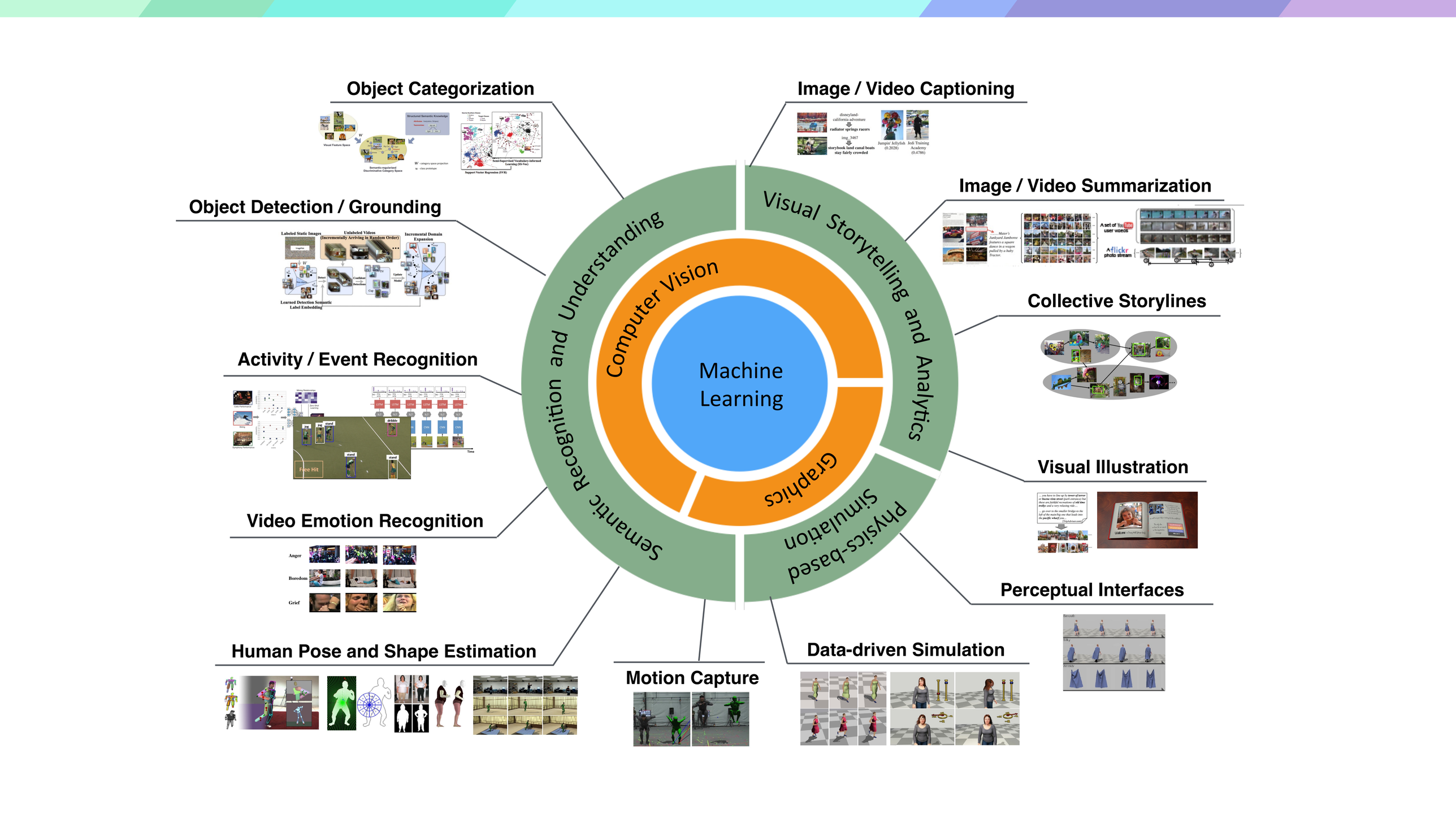
Postdoctoral Researcher
2007 - 2009

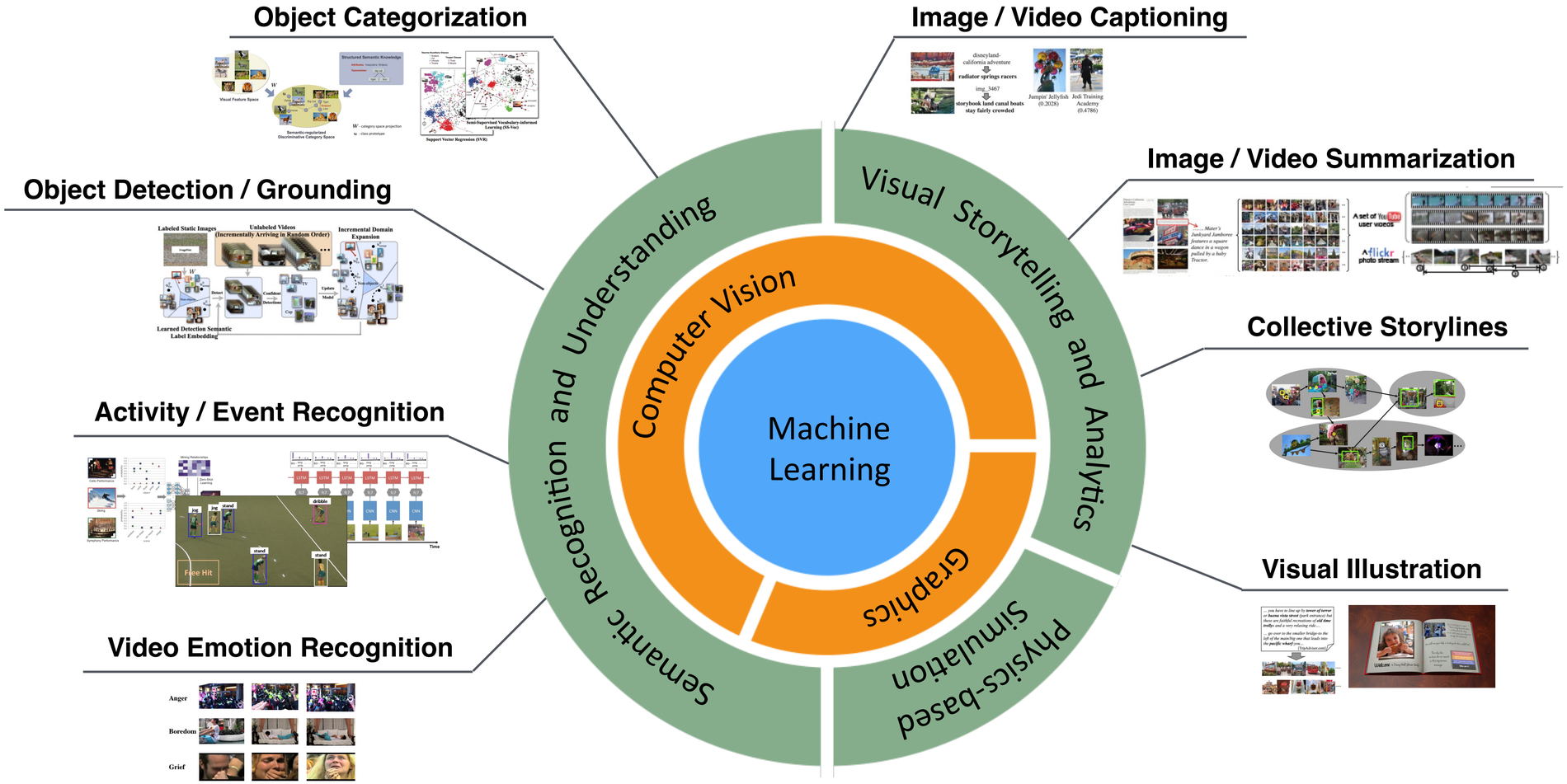
UNIVERSITY OF TORONTO

PhD, MSc
2001 - 2008

BROWN

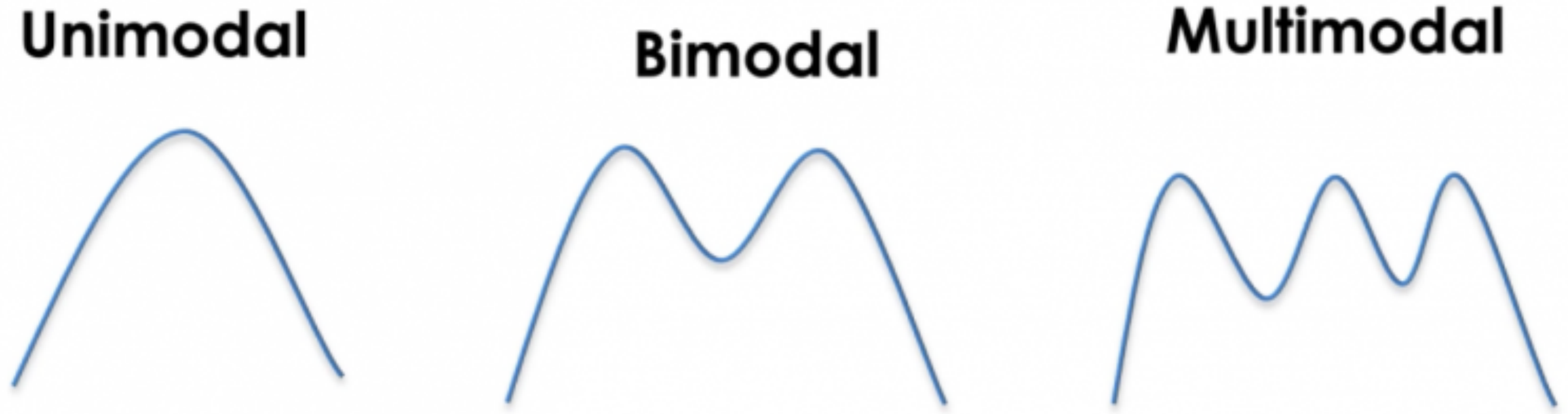
BOSTON UNIVERSITY


Machine Learning
Computer Vision
Graphics
Visual Storytelling and Analytics
Physics-based Simulation
Semantic Recognition and Understanding
Object Categorization
Image / Video Captioning
Image / Video Summarization
Object Detection / Grounding
Collective Storylines
Activity / Event Recognition
Visual Illustration
Video Emotion Recognition
Perceptual Interfaces
Human Pose and Shape Estimation
Motion Capture
Data-driven Simulation
Anger
Boredom
Grief
Free Hit


Object Categorization
Visual Feature Space
Structured Semantic Knowledge
Semantic-regularized Discriminative Category Space
W - category space projection
w - class prototype
Semi-Supervised Vocabulary-informed Learning (SS-Voc)
Support Vector Regression (SVR)
Image / Video Captioning
disneyland-california adventure
radiator springs racers
img_3467
storybook land canal boats
stay fairly crowded
Jumpin' Jellyfish (0.2028)
Jedi Training Academy (0.4786)
Image / Video Summarization
Master's Junkyard Jamboree features a square dance in a wagon pulled by a baby Tractor.
A set of YouTube user videos
A flickr photo stream
Object Detection / Grounding
Labeled Static Images
Unlabeled Videos (Incrementally Arriving in Random Order)
Incremental Domain Expansion
Learned Detection Semantic Label Embedding
Collective Storylines
Activity / Event Recognition
Free Hit
Time
Visual Illustration
Video Emotion Recognition
Anger
Boredom
Grief
Computer Vision
Machine Learning
Graphics
Visual Storytelling and Analytics
Physics-based Simulation
Semantic Recognition and Understanding

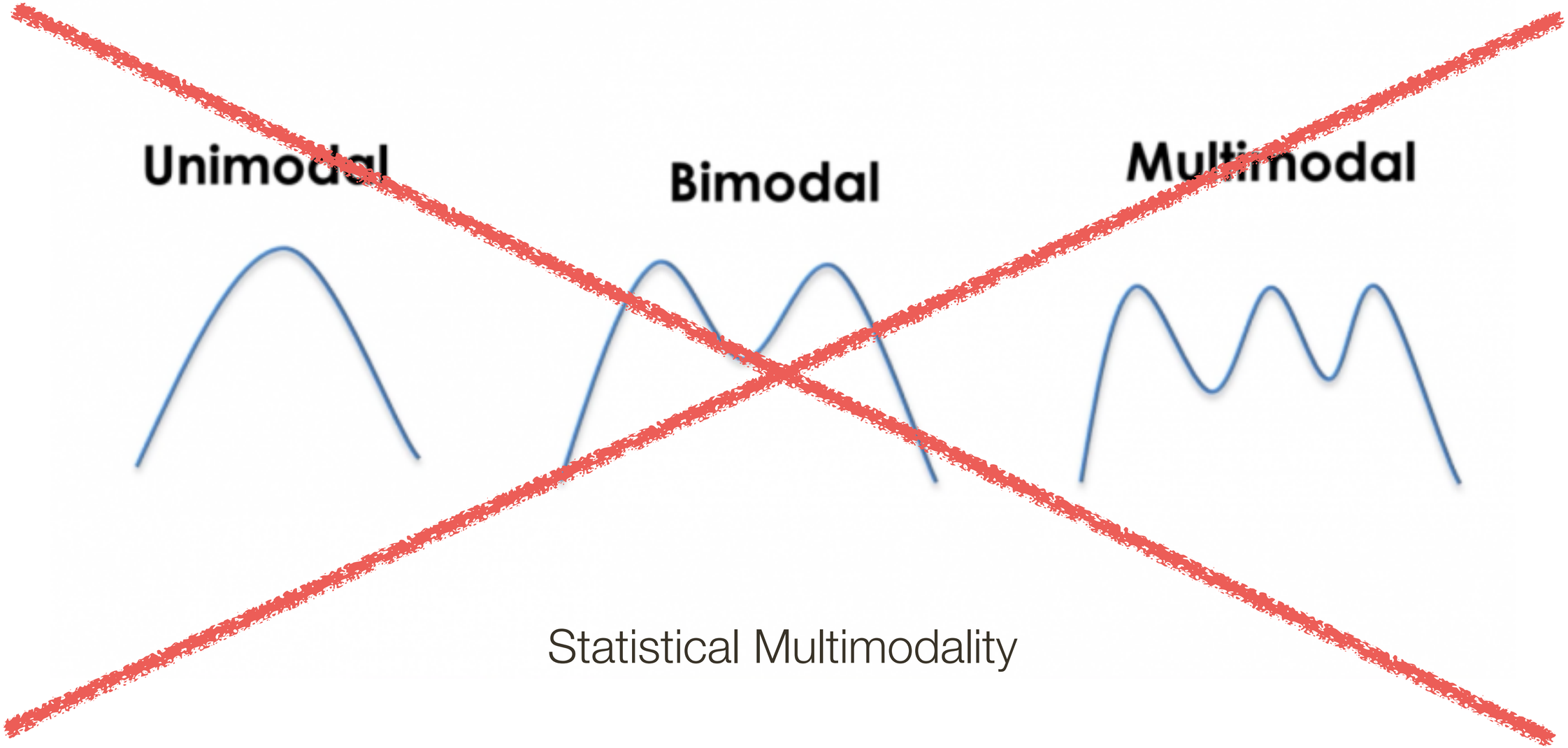
What is **Multi-modal Learning**?

What is **Multi-modal Learning**?



Statistical Multimodality

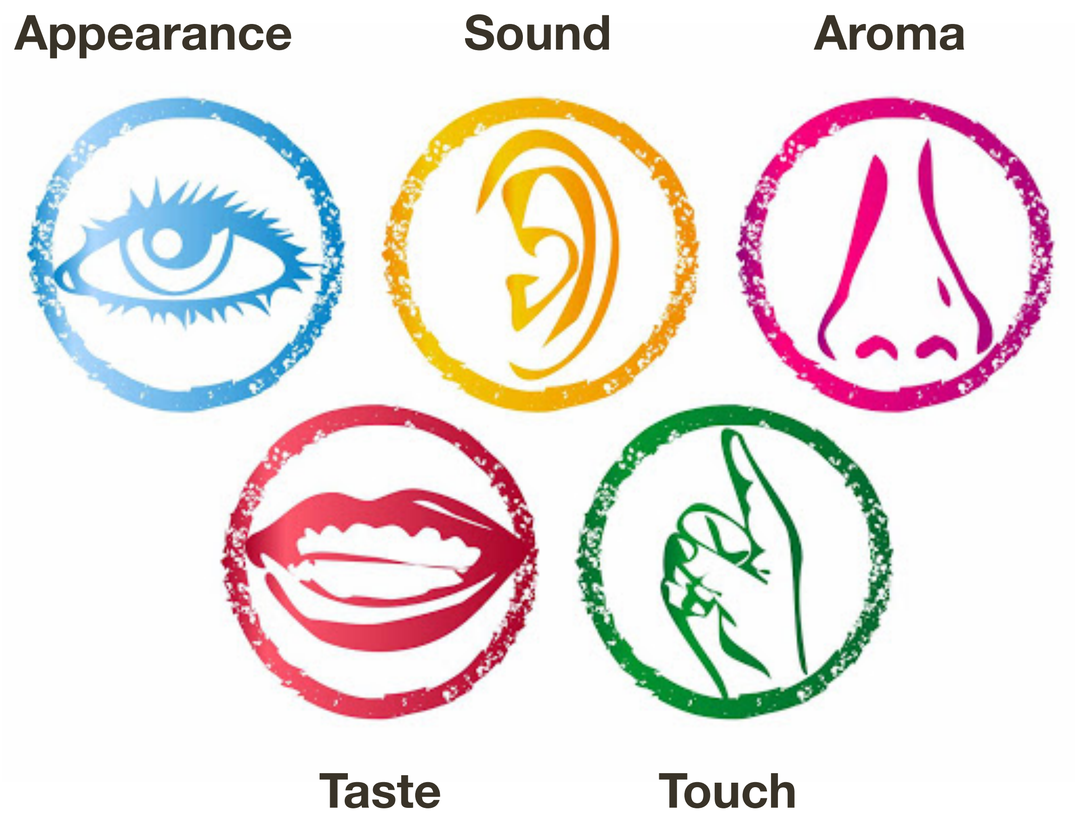
What is **Multi-modal Learning**?



Statistical Multimodality

What is **Multi-modal Learning**?

- **Sensory modality:** one or more primary channels of perception.



What is **Multi-modal Learning**?

- **Sensory modality:** one or more primary channels of perception.

  We will focus on modalities that are central to **communication**:

Appearance **Sound**

What is **Multi-modal Learning**?

- **Sensory modality:** one or more primary channels of perception.

We will focus on modalities that are central to **communication**:

Visual
— Perception
— Gestures
— Eye gaze

Appearance **Sound**

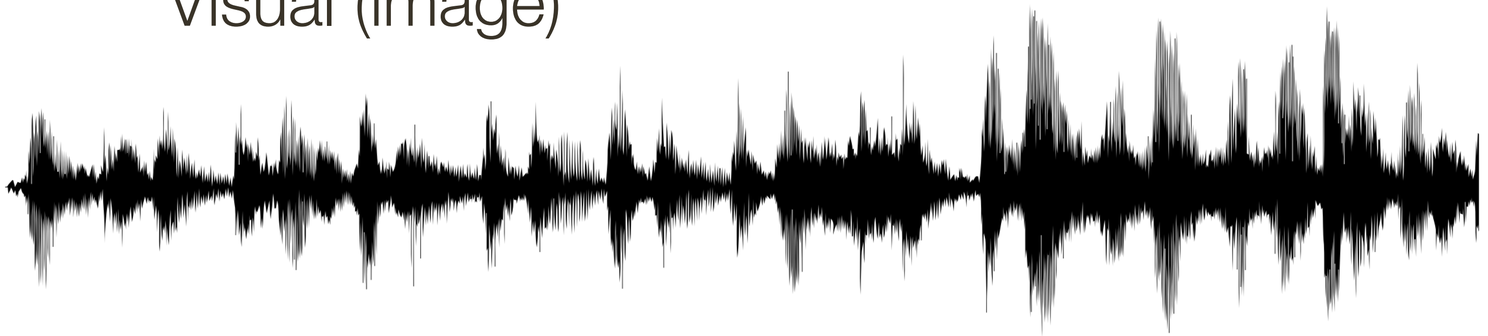
Verbal
lexicon — words
syntax — parts of speech
pragmatics — discourse

Vocal
prosody — intonation
music — tone, pitch

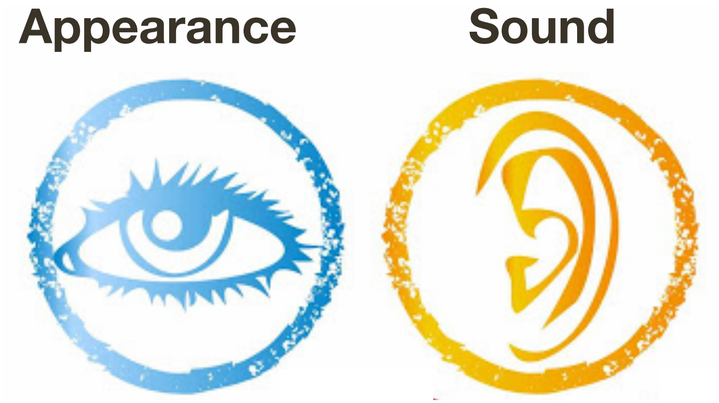
What is **Multi-modal Learning**?

- **Sensory modality:** one or more primary channels of perception/communication.
- **Modality:** refers to a certain type of information and/or representation format in which information is stored.

Visual (image)



Auditory (voice / sound)



Owl Wisdom, Omens, Vision of the night No bird has as much myth and mystery surrounding it than the owl. Part of this mystical aura is due to the fact that the bird is nocturnal and the night time has always seemed mysterious to humans. The owl is a symbol of the feminine, the moon, and the night. Because of it association with the moon it has ties to fertility and seduction. The owl is bird of magic and darkness of prophecy and wisdom.

Natural Language (text)

Visual (drawings)

What is **Multi-modal Learning**?

Totalitarian view: nearly everything can be cast as multi-modal learning problem

e.g., image classification is a **translation** from **image** modality to **label** modality

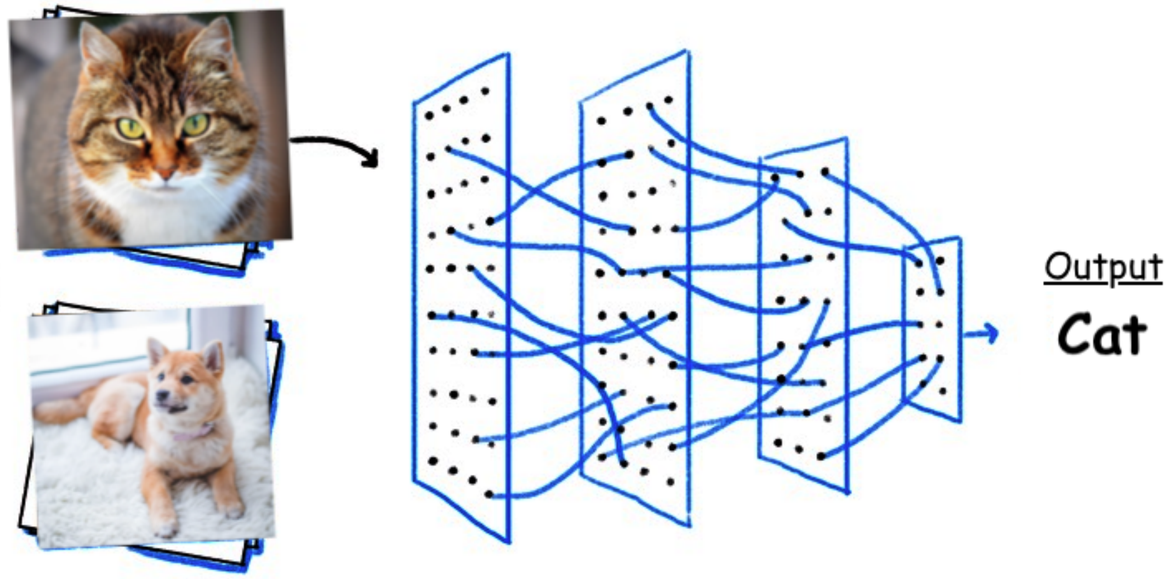


Image Source: https://towardsdatascience.com/10-papers-you-should-read-to-understand-image-classification-in-the-deep-learning-era-4b9d792f45a7

Multimodal Research: Historical Perspective

Four **eras** of multimodal research (according to Louis-Philippe (LP) Morency, CMU)

— The “**behavioral**” era (1970s until late 1980s)

— The “**computational**” era (late 1980s until 2000)

— The “**interaction**” era (2000 - 2010)

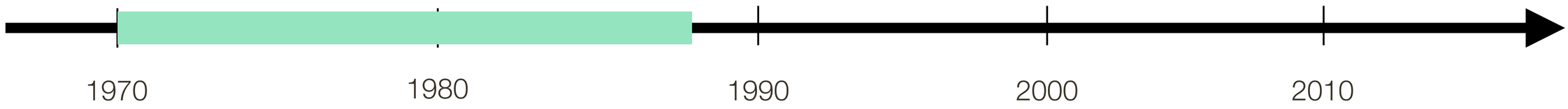
— The “**deep learning**” era (2010 until now)

* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Studies of multi-sensory integration in **Psychology**

e.g., infant's perception of substance and temporal synchrony in multimodal events



* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Studies of multi-sensory integration in **Psychology**

e.g., infant's perception of substance and temporal synchrony in multimodal events

Degeneracy — observed effect of sensory systems being able to co-educate each other without external teacher (self-supervision)

(e.g., babies randomly moving arms, until the experience a change in haptic state, grabbing something)

* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Studies of multi-sensory integration in **Psychology**

e.g., infant's perception of substance and temporal synchrony in multimodal events

Re-entry — ability of our brain to inter-relate multiple simultaneous representations across modalities

(e.g., when a person experiences a visual percept of an apple, other modalities are invoked, such that a person can experience its smell, taste, etc.)

* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Studies of multi-sensory integration in **Psychology**

e.g., infant's perception of substance and temporal synchrony in multimodal events

Transparency

* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Studies of multi-sensory integration in **Psychology**

e.g., infant's perception of substance and temporal synchrony in multimodal events

TRIVIA

Geoffrey Hinton ("father of deep learning")
received B.A. in Experimental Psychology
from King's College in Cambridge

* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

McGurk Effect (1976)

* video credit: **OK Science**

* Adopted from slides by Louis-Philippe Morency

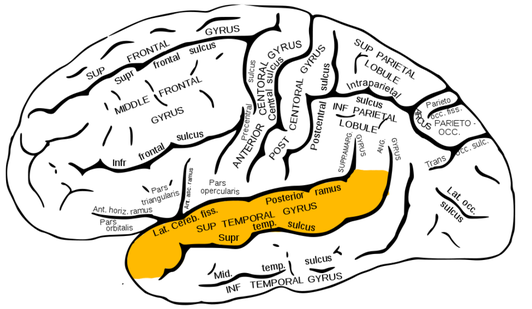
Multimodal Research: Historical Perspective

McGurk Effect (1976)

* video credit: **OK Science**

* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective



The McGurk Effect

Superior Temporal Sulcus is responsible for merging visual and auditory signals in the brain [Beauchamp et al. 2010].

McGurk Effect (1976)

* video credit: **OK Science**

* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Audio-visual speech recognition (motivated by McGurk effect)

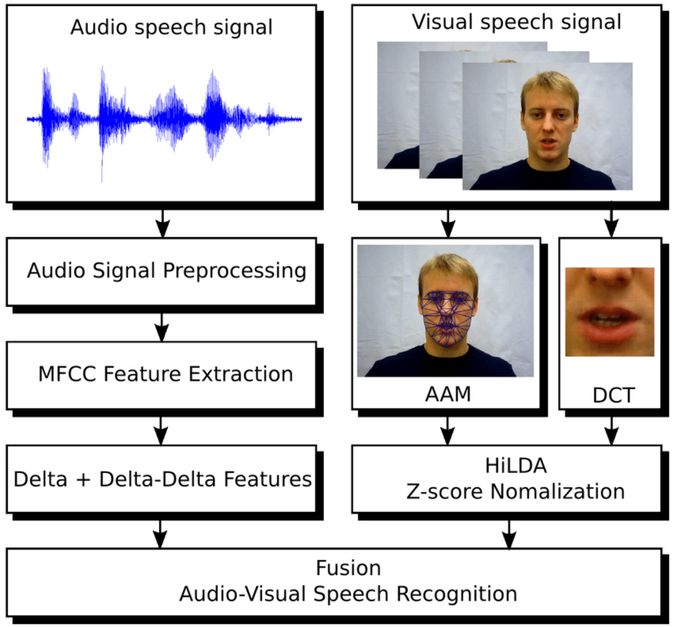


Image from: Palecek and Chaloupka

* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Audio-visual speech recognition (motivated by McGurk effect)

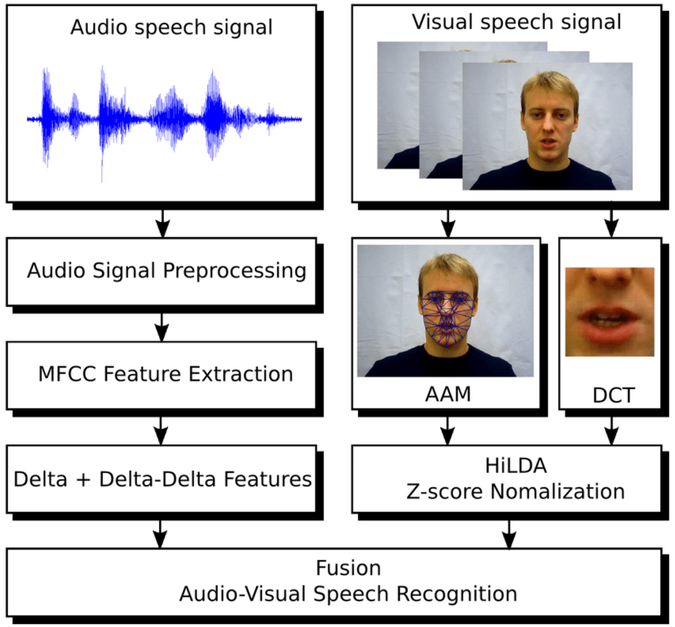


Image from: Palecek and Chaloupka

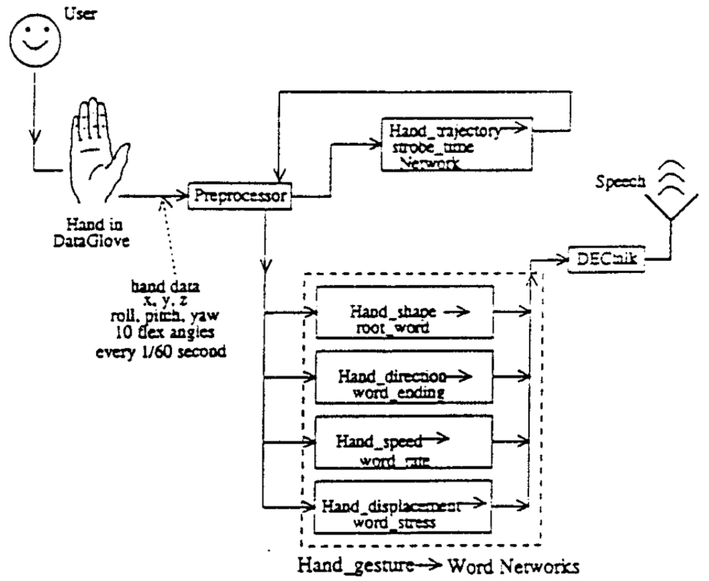
Turns out most of the information is **redundant** rather than **complementary** (i.e., improved robustness but not overall quality of speech recognition)

* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Audio-visual speech recognition (motivated by McGurk effect)

Multi-modal and multi-sensory interfaces



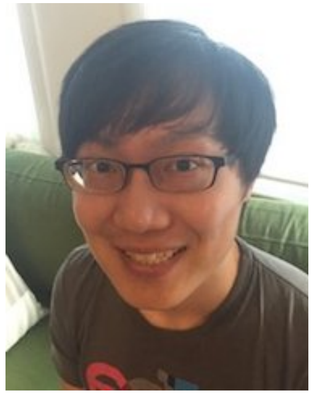
GloveTalk by S. Fels and G. Hinton [CHI'95]

* Adopted from slides by Louis-Philippe Morency

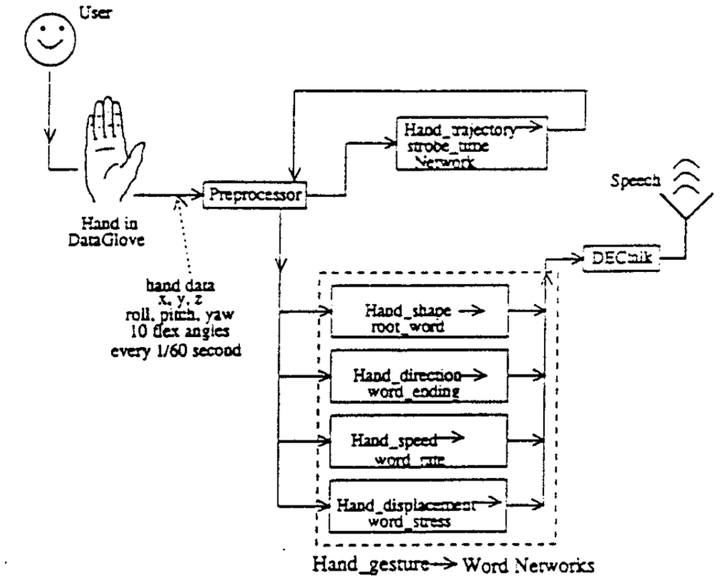
Multimodal Research: Historical Perspective

Audio-visual speech recognition (motivated by McGurk effect)

Multi-modal and multi-sensory interfaces



Dongwook Yoon



GloveTalk by S. Fels and G. Hinton [CHI'95]

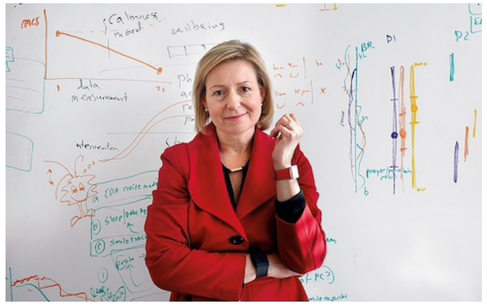
* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Audio-visual speech recognition (motivated by McGurk effect)

Multi-modal and multi-sensory interfaces

Affective Computing



Rosalind Picard

affectiva

* Adopted from slides by Louis-Philippe Morency

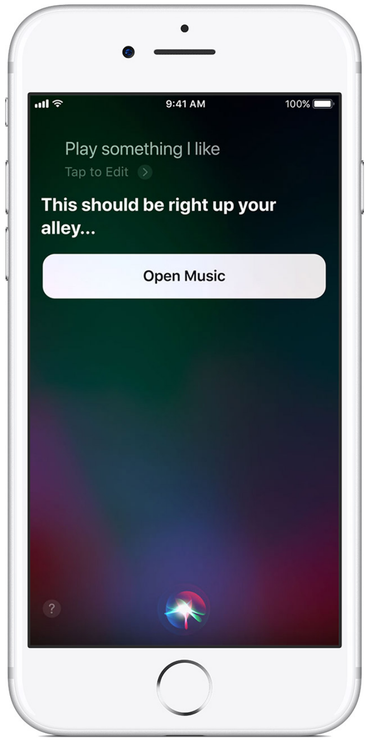
Multimodal Research: Historical Perspective

Modeling human multi-modal interactions

- Huge multi-laboratory efforts

AMI Project [2001-2006, IDIAP]
- 100+ hours of meeting recordings
- Synchronized video and audio
- Transcribed and annotated

CALO Project [2003-2008, SRI]
- Cognitive assistant that learns and organizes
- Personalized assistant that learns



Siri was spun as an output of multi-modal interaction projects

* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Modeling human multi-modal interactions

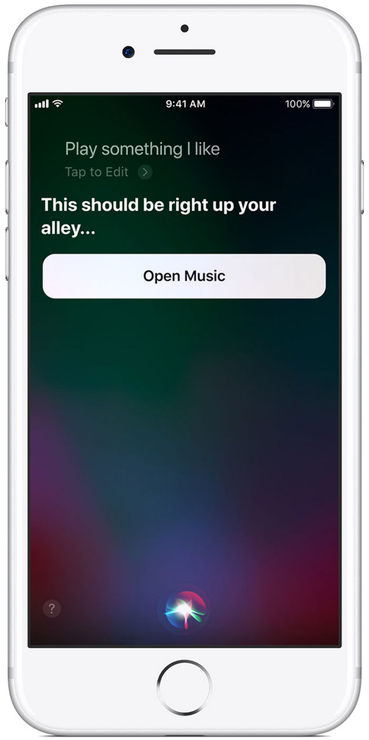
- Huge multi-laboratory efforts

Multimedia information retrieval

- Lots of challenges and progress

Research Tasks and Challenges:

- Shot boundary detection, story segmentation, search
- Semantic event, character and object detection



Siri was spun as an output of multi-modal interaction projects

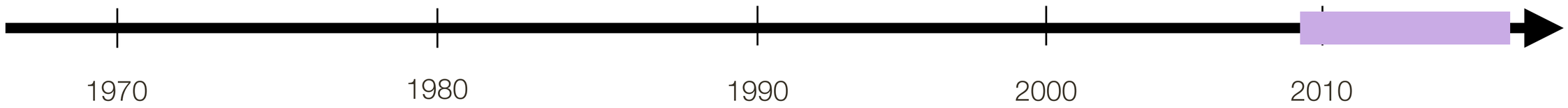
* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)

- Better performance
- More interesting problems emerging

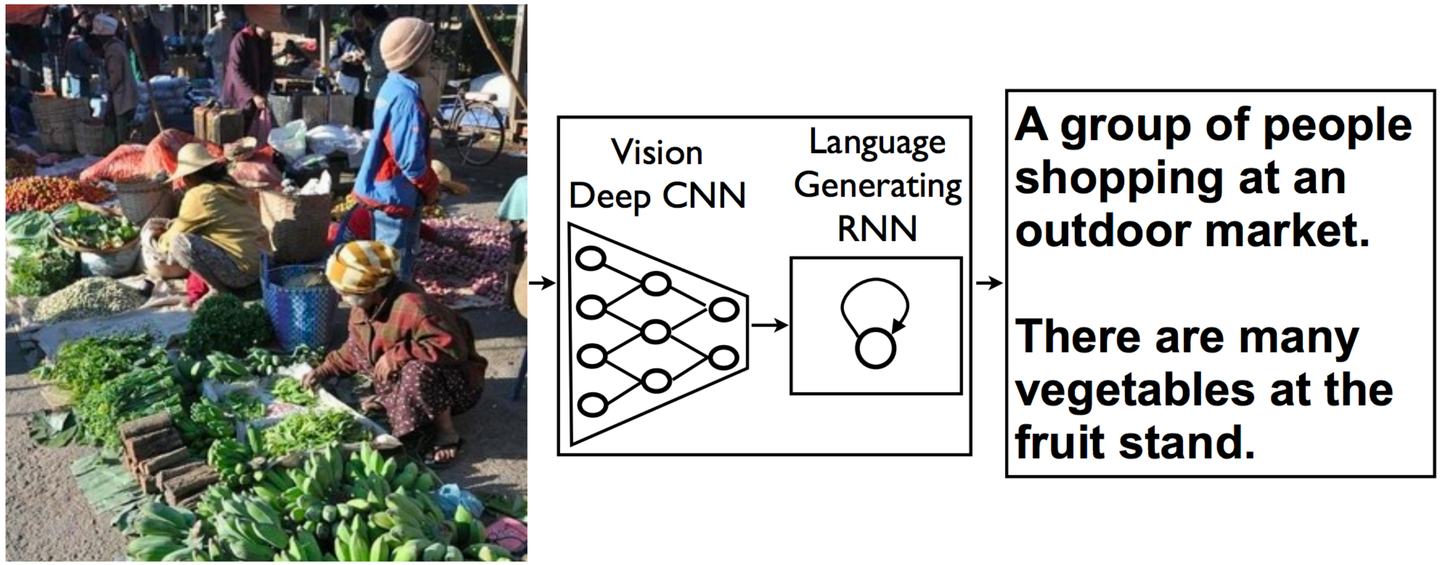
THIS IS OUR COURSE



* Adopted from slides by Louis-Philippe Morency

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)



[Vinyals *et al.*, 2015]

Natural language description generation

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)

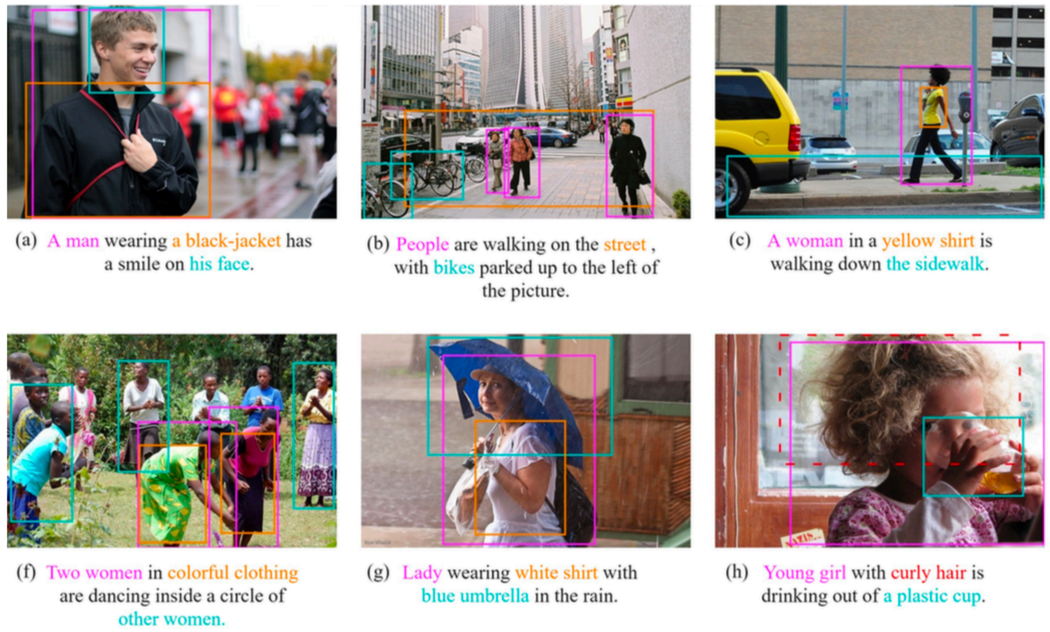
A few miles before tioga road reached highway 395 and the town of lee vining, smith turned onto a narrow blacktop road. On either side were parched, grassy open slopes with barbed-wire fences marking property lines. Cattle and horses grazed under trees whose black silhouettes stood stark against the gold-velvet mountains. Marty burst into song: “ home , home on the range, where the deer and the antelope play! Where seldom is heard a discouraging word and the skies are not cloudy all day!”

[Zhu et al, ICCV 2015]

Story generation

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)



[Bajaj *et al.*, ICCV 2019]

Language grounding

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)

[Ba *et al.*, ICCV 2015]

Corn Poppy

Papaver rhoeas (common names include corn poppy, corn rose, field poppy, Flanders poppy, red poppy, red weed, coquelicot, and, due to its odour, which is said to cause them, as headache and headwark) is a species of flowering plant in the poppy family, Papaveraceae. This poppy, a native of Europe, is notable as an agricultural weed (hence the "corn" and "field") and as a symbol of fallen soldiers.
P. rhoeas sometimes is so abundant in agricultural fields that it may be mistaken for a crop. The only species of Papaveraceae grown as a field crop on a large scale is Papaver somniferum, the opium poppy.

The plant is a variable annual, forming a long-lived soil seed bank that can germinate when the soil is disturbed. In the northern hemisphere it generally flowers in late spring, but if the weather is warm enough other flowers frequently appear at the beginning of autumn. The flower is large and showy, with four petals that are vivid red, most commonly with a black spot at their base. Like many other species of Papaver, it exudes a white latex when the tissues are broken.

.....

Detecting objects based on linguistic descriptions

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)

[Zhu et al, ICCV 2015]

Book-to-Movie alignment

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)

[Zhu et al, ICCV 2015]

Book-to-Movie alignment

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)

*... you have to line up by **tower of terror** or **buena vista street** (park entrance) but these are faithful recreations of **old time trollys** and a very relaxing ride....*

*... go over to the smaller bridge-to the left of the main/big one that leads into the **pacific wharf**-you...*

[TripAdvisor.com]

[Kim & Sigal, CVPR 2015]

Auto illustration

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)

Q1: *What color is the bowl ?*

GT answer: *White*
Predicted answer: *White*
Rank of GT: *1*

Q2: *Do you see any people?*

GT answer: *No*
Predicted answer: *No, just the cat*
Rank of GT: *2*

Q3: *What color is the cat ?*

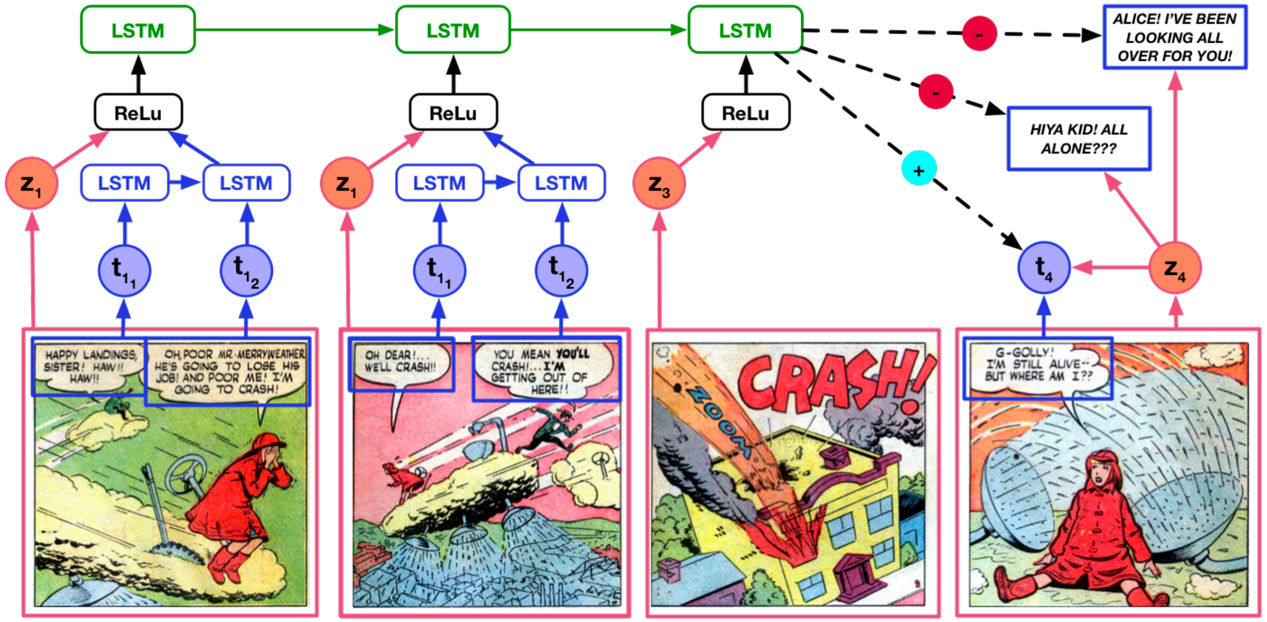
GT answer: *Grey, white, and black*
Predicted answer: *Grey, black and white*
Rank of GT: *6*

[Seo *et al.*, NIPS 2017]

Visual question answering / dialog

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)



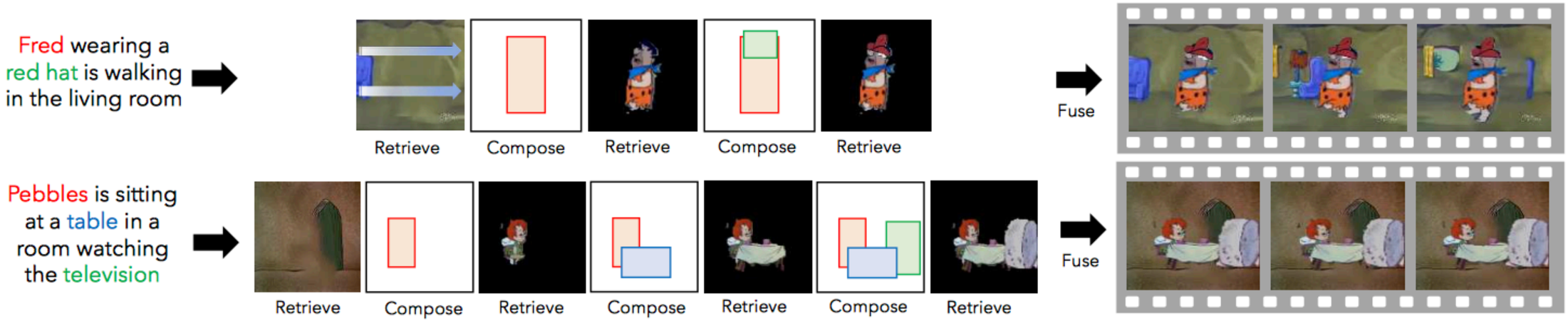
[Iyyer *et al.*, CVPR 2017]

Narrative plot understanding

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)

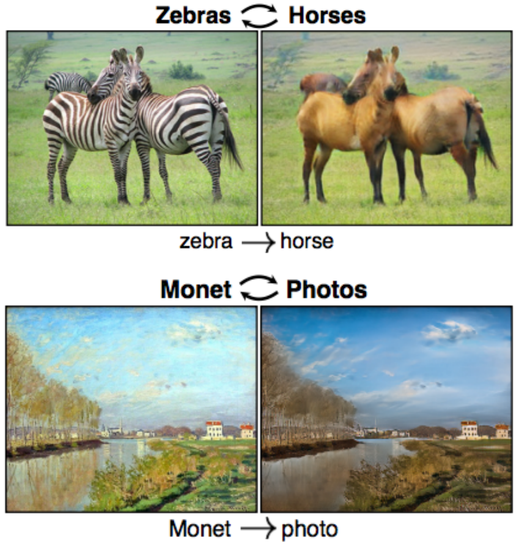
[Gupta *et al.*, ECCV 2018]



Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)

[Zhu *et al.*, ICCV 2017]



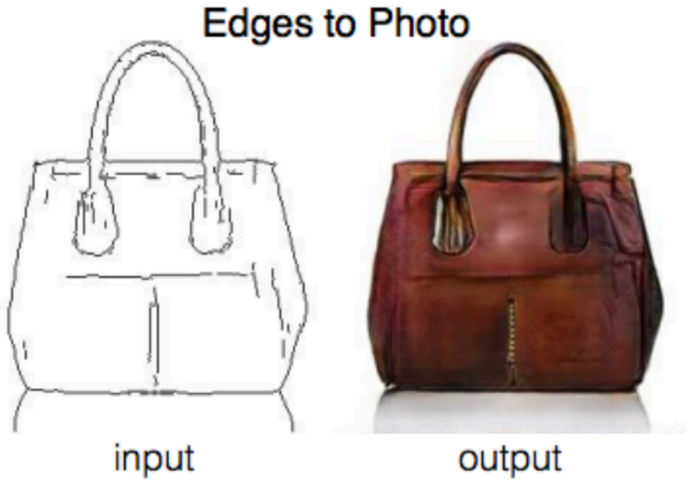
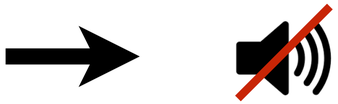


Image-to-image translation

[Isola *et al.*, CVPR 2017]

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)



[Iyyer *et al.*, NIPS 2016]

Video-to-Audio translation

Multimodal Research: Historical Perspective

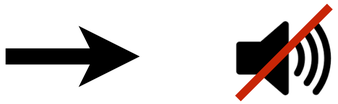
Deep Learning (a.k.a. representation learning)

[Iyyer *et al.*, NIPS 2016]

Video-to-Audio translation

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)



[Iyyer *et al.*, NIPS 2016]

Video-to-Audio translation

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)

LECTURING	
ROLE	VALUE
AGENT	TEACHER
AUDIENCE	STUDENT
PLACE	CLASSROOM

TIPPING	
ROLE	VALUE
AGENT	PEOPLE
ITEM	BOAT
AGENT PART	WEIGHT
PLACE	LAKE

JUMPING	
ROLE	VALUE
AGENT	WOMEN
SOURCE	LAND
OBSTACLE	-
DESTINATION	LAND
PLACE	OUTDOORS

CLIMBING	
ROLE	VALUE
AGENT	MAN
OBSTACLE	ROCK
TOOL	ROPE
PLACE	JUNGLE

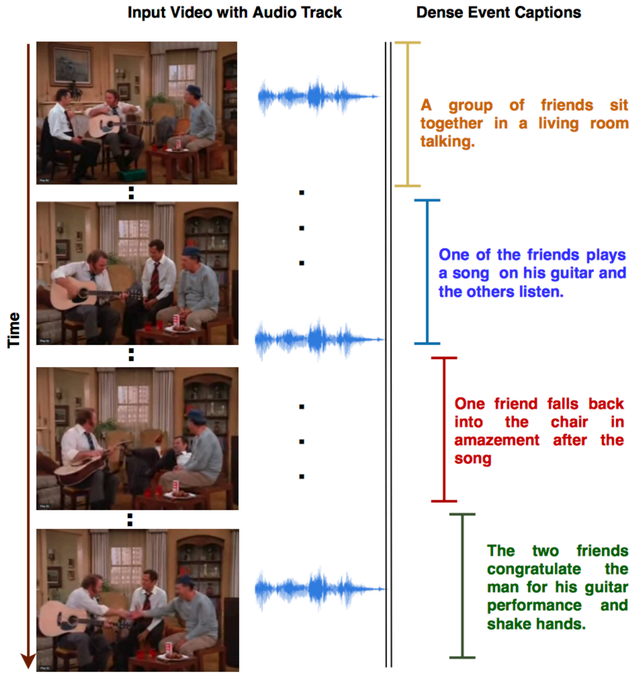
ENCOURAGING	
ROLE	VALUE
AGENT	DAD
RECEIVER	SON
PLACE	LAWN

[Suhail *et al.*, ICCV 2019]

Situational Recognition

Multimodal Research: Historical Perspective

Deep Learning (a.k.a. representation learning)



[Rahman *et al.*, ICCV 2019]

Dense Video Captioning

Key Challenges of Multimodal Learning

- Representation learning in each and across modalities
- Alignment between representations in different modalities
- Translation between modalities

Key Challenges of Multimodal Learning

- Representation learning in each and across modalities
- Alignment between representations in different modalities
- Translation between modalities

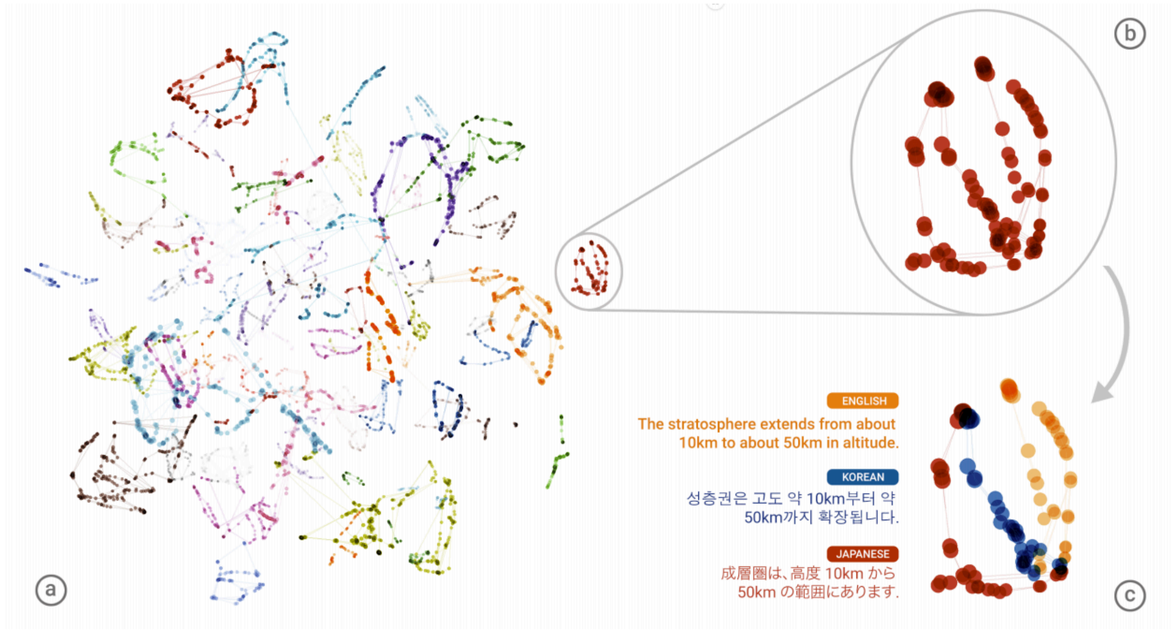
What's another phrase for "**representation learning**"?

Key Challenges of Multimodal Learning

- Representation learning in each and across modalities
- Alignment between representations in different modalities
- Translation between modalities

One translation model learned across many languages, actually improves the performance in translation over direct training on:

English -> German
German -> English
French -> English



[Johnson *et al.*, ArXiv 2017 from Google]

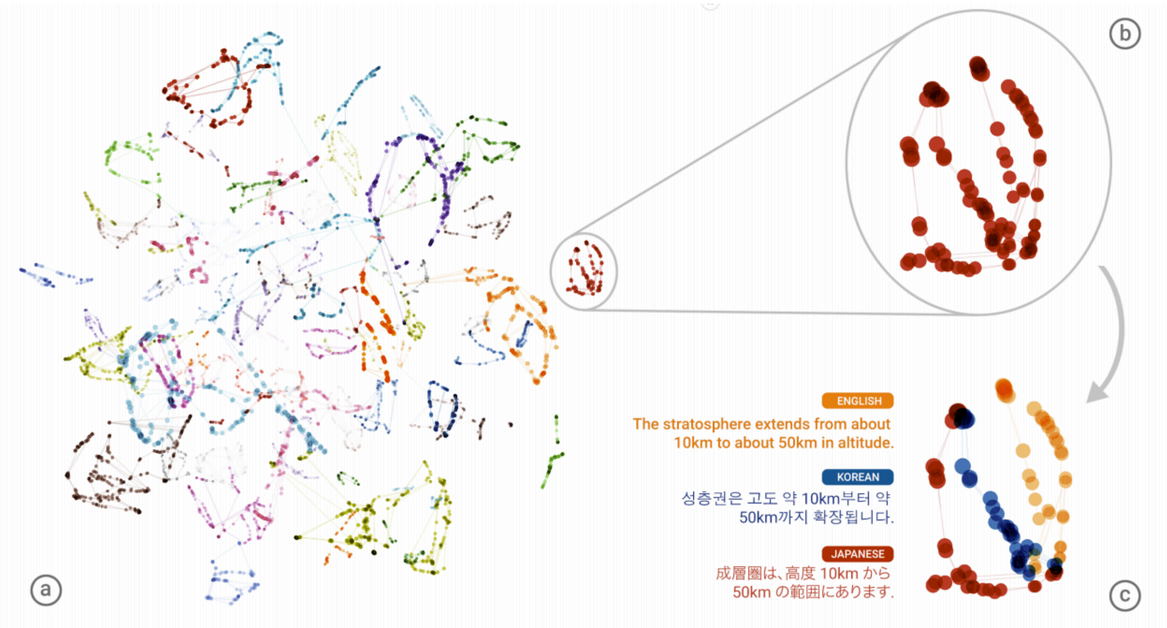
Key Challenges of Multimodal Learning

- Representation learning in each and across modalities
- Alignment between representations in different modalities
- Translation between modalities

One translation model learned across many languages, actually improves the performance in translation over direct training on:

English -> German
German -> English
French -> English

Allows translation between languages pairs never trained on before

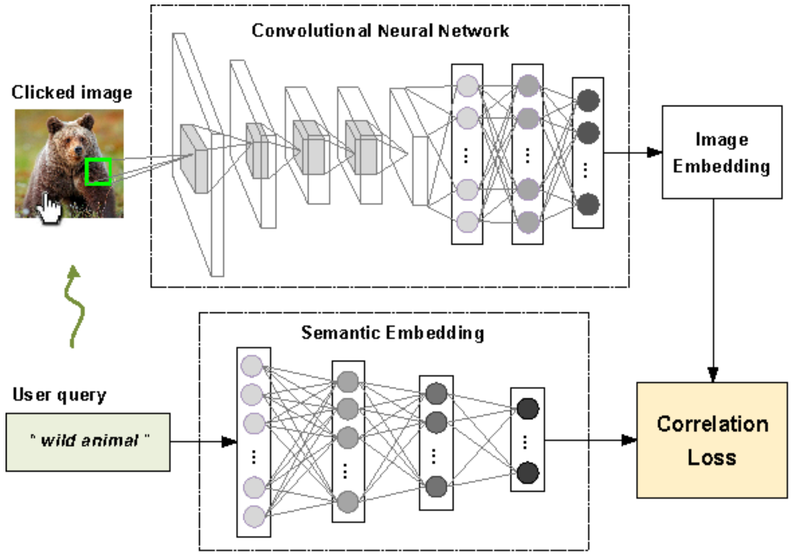


[Johnson *et al.*, ArXiv 2017 from Google]

Key Challenges of Multimodal Learning

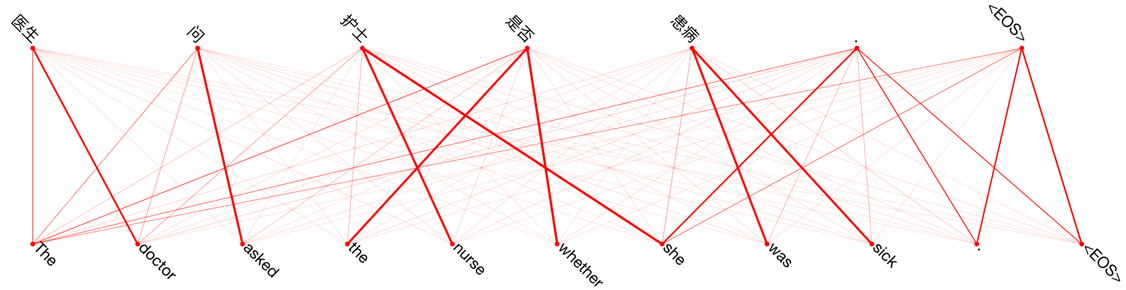
- Representation learning in each and across modalities
- Alignment between representations in different modalities
- Translation between modalities

Explicit



Implicit

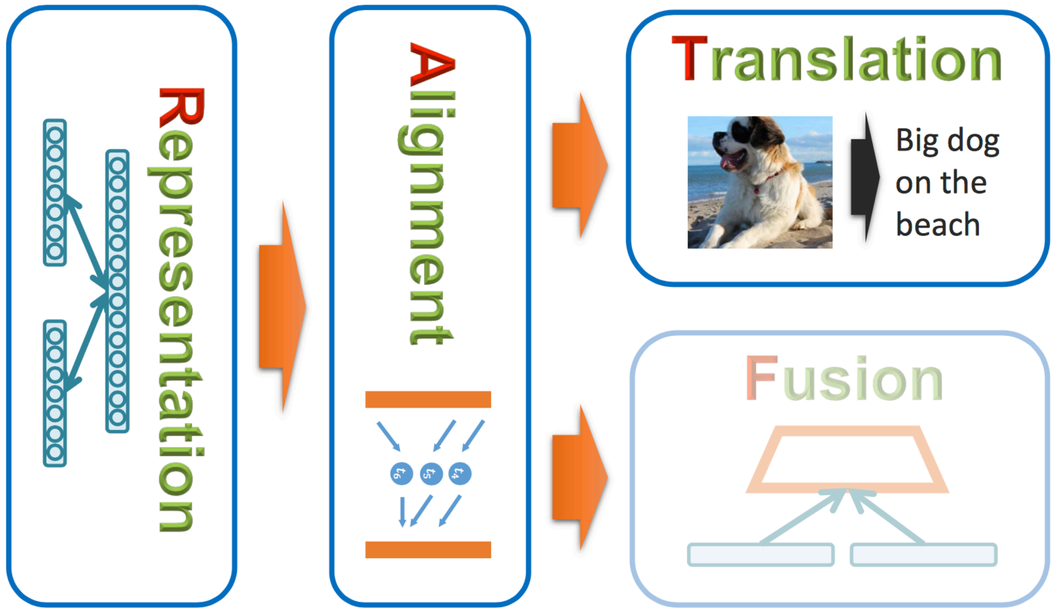
(e.g., using neural attention)



http://player-eric.com/2020/02/20/nmtvis-attention/

Key Challenges of Multimodal Learning

- Representation learning in each and across modalities
- Alignment between representations in different modalities
- Translation between modalities



* Adopted from slides by Louis-Philippe Morency

Objectives of the course

- Acquire **fundamentals and background** that would allow one to follow research in Computer Vision and on intersection of Vision + Language
- Ability to **design, build and apply deep learning architectures** for multi-modal problems (Vision + Language in particular)
- Obtain **overview of research trends** in Computer Vision and ML related to topics of the course
- Ability to define research problems, read and present research papers

Objectives of the course

- Acquire **fundamentals and background** that would allow one to follow research in Computer Vision and on intersection of Vision + Language
- Ability to **design, build and apply deep learning architectures** for multi-modal problems (Vision + Language in particular)
- Obtain **overview of research trends** in Computer Vision and ML related to topics of the course
- Ability to define research problems, read and present research papers

course is heavy on *practical* deep learning

Deep Learning

Google snaps up object recognition startup DNNresearch

Google has acquired a research startup founded within the University of Toronto, whose work includes object recognition.

by Josh Lowensohn @Josh / 13 March 2013, 9:22 am AEDT

2 / 0 / 0 / 0 / / more +

Google has acquired a three-person Canadian research company that specializes in voice and image recognition.

DNNresearch, which was founded last year within the the University of Toronto's computer science department, specializes in object recognition and now belongs to Google.

From left: Ilya Sutskever, Alex Krizhevsky and University Professor Geoffrey Hinton of the University of Toronto's Department of Computer Science. *(photo by John Guatto, University of Toronto)*

* slide from Dhruv Batra

Deep Learning

Google snaps up object recognition startup DNNresearch

Google has acquired a research startup founded within the University of Toronto, whose work includes object recognition.

by Josh Lowensohn @Josh / 13 March 2013, 9:22 am AEDT

2 / 0 / 0 / 0 / more +

Google has acquired a three-person Canadian research company that specializes in voice and image recognition.

DNNresearch, which was founded last year within the the University of Toronto's computer science department, specializes in object recognition and now belongs to Google.

From left: Ilya Sutskever, Alex Krizhevsky and University Geoffrey Hinton of the University of Toronto's Departmen Science. *(photo by John Guatto, University of Toronto)*

8TH ANNUAL CRUNCHIES AWARDS Celebrate the Best of Tech in 2014 Get Your Tickets Now

Google Acquires Artificial Intelligence Startu For More Than $500M

Posted Jan 26, 2014 by Catherine Shu (@catherineshu)

11.3k SHARES

Next story

CrunchBase

DeepMind

FOUNDED 2011

OVERVIEW

DeepMind is a cutting edge artificial intelligence company. We combine the best techniques from machine learning and systems neuroscience to build powerful general-purpose learning algorithms. Founded by Demis Hassabis, Shane Legg and Mustafa Suleyman, the company is based in London and supported by some of the ... entrepreneurs and investo... first commercial ...

« Search needs a shake-up Songbirds use grammar rules »

Machine Learning Startup Acquired by ai-one

Press Release

For Immediate Release: August 4, 2011

San Diego artificial intelligence startup acquired by leading provider of machine learning SDKs as market for advanced applications gets hot.

San Diego CA – ai-one announced today that it acquired Auto-Semantics, a local start-up providing artificial intelligence services to corporate IT departments. The acquisition is the latest in a series of joint-ventures and acquisitions by ai-one that consolidates its leadership position within the emerging market for machine learning technologies.

Yann LeCun

December 9, 2013 ·

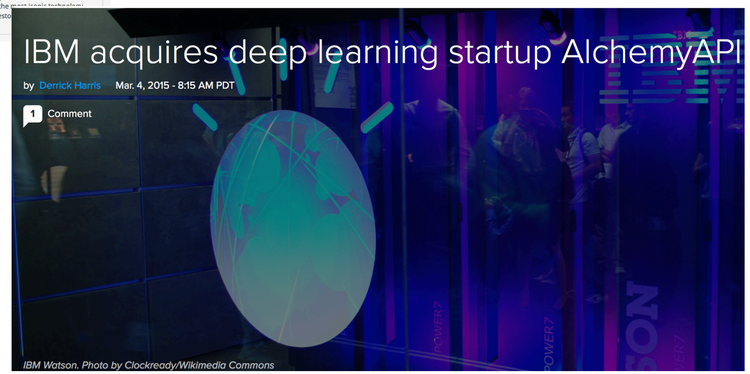
Big news today!

Facebook has created a new research laboratory with the ambitious, long-term goal of bringing about major advances in Artificial Intelligence.

IBM acquires deep learning startup AlchemyAPI

by Derrick Harris Mar. 4, 2015 - 8:15 AM PDT

1 Comment



IBM Watson. Photo by Clockready/Wikimedia Commons

* slide from Dhruv Batra

Clever Hans

Clever Hans

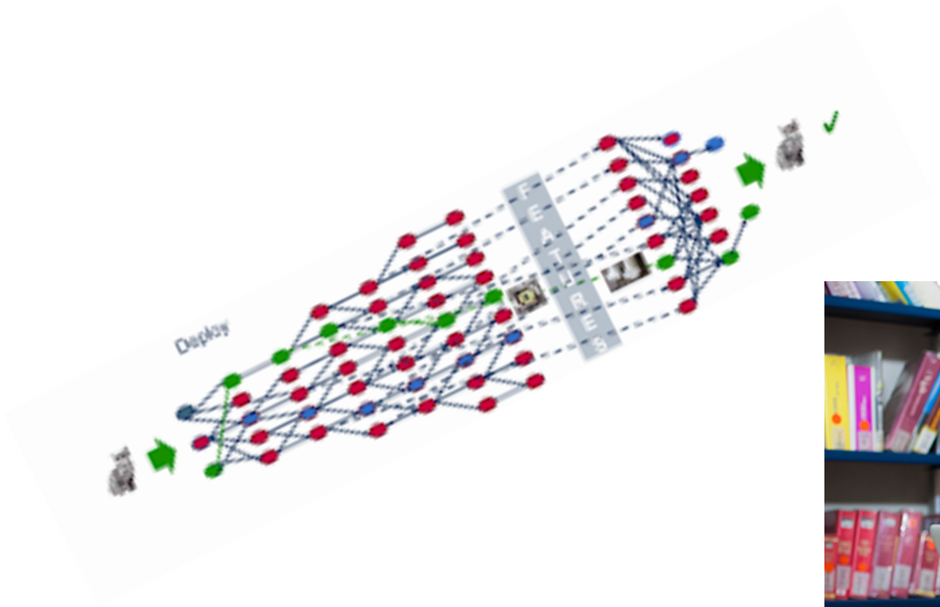
Hans could get 89% of the math questions right

Clever Hans

The horse was **smart**, just not in the way van Osten thought!

Hans could get 89% of the math questions right

Clever DNN



Visual **Question Answering**

Is there zebra climbing the tree?

AI agent

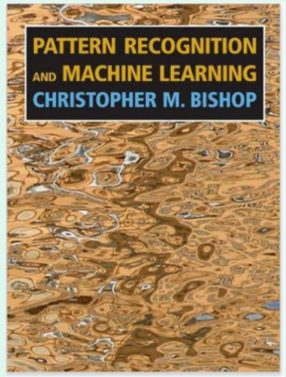
Yes

What will be covered **specifically**?

Course Webpage

Pre-requisites

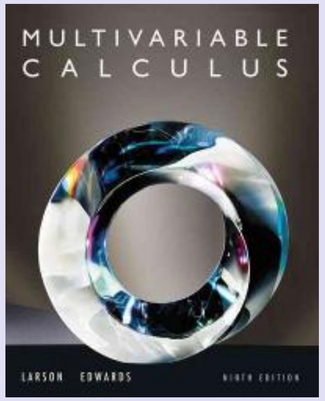
Computer Science



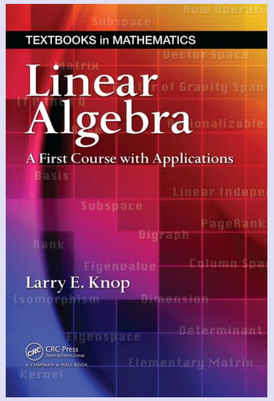
CPSC 340
(or equivalent)

Needed for
Assignments

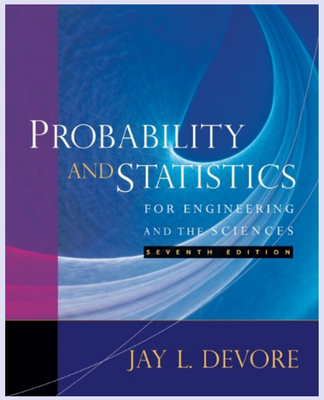
Mathematics



Calculus



Linear Algebra



Statistics

Helpful (but not necessary): some background in Computer Vision or NLP

Additional **Requirement**

Option 1: Google Colab

Additional **Requirement**

Option 1: Google Colab

Option 2: Own hardware

Nvidia GTX 1060 (with 6GB RAM) or above

Additional **Requirement**

Option 1: Google Colab

Option 2: Own hardware

Nvidia GTX 1060 (with 6GB RAM) or above

Option 3: Cloud services

As student you get $25 in credits

As student you get $100 in credits

You will need to provision the VM and ensure you keep track of spendings. As long as VM is running you are being charged, even if you are not running the code.

Course **structure**

Approximately 60% of course will consists of lectures and optional readings

Remaining 40% is reading  and presentation of  curated research papers on relevant topics

5 programming assignments

Final (individual or **group**) project

Course **structure**

Approximately 60% of course will consists of lectures and optional readings

Remaining 40% is reading  and presentation of  curated research papers on relevant topics

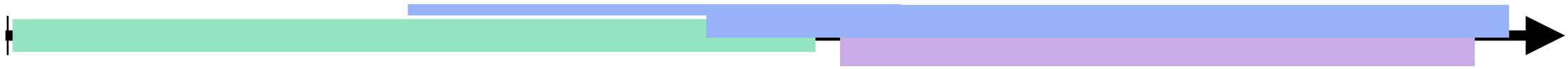
+ **guest speakers**

5 programming assignments

Final (individual or **group**) project

Grading Criteria

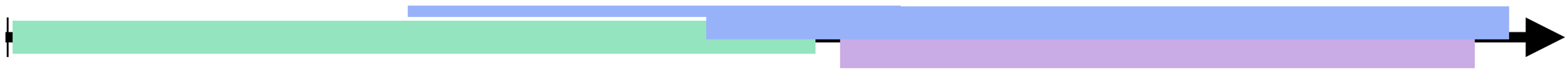
- **Assignments** (programming) — 40% (total)
- **Research papers** — 20%
- **Project** — 40%



Grading Criteria

- **Assignments** (programming) — 40% (total)
- **Research papers** — 20%
- **Project** — 40%

NO LATE SUBMISSIONS — If you don't complete the assignment, hand in what you have



Assignments (5 assignments and 40% of grade total)

- Assignment 0: **Introduction to PyTorch** (0%)
- Assignment 1: **Neural Network Introduction** (5%) — 

Assignments all use **Python Jupiter Notebooks**, use `Canvas` to hand everything in. Assignments always due at **11:59pm PST** on due date.

Assignments (5 assignments and 40% of grade total)

- Assignment 0: **Introduction to PyTorch** (0%)
- Assignment 1: **Neural Network Introduction** (5%) — 
- Assignment 2: **Convolutional Neural Networks** (5%) — PYTORCH

Assignments all use **Python Jupiter Notebooks**, use `Canvas` to hand everything in. Assignments always due at **11:59pm PST** on due date.

Assignments (5 assignments and 40% of grade total)

- Assignment 0: **Introduction to PyTorch** (0%)
- Assignment 1: **Neural Network Introduction** (5%) — 
- Assignment 2: **Convolutional Neural Networks** (5%) — PYTORCH
- Assignment 3: RNN **Language Modeling and Translation** (10%) — PYTORCH

Assignments all use **Python Jupiter Notebooks**, use `Canvas` to hand everything in. Assignments always due at **11:59pm PST** on due date.

Assignments (5 assignments and 40% of grade total)

- Assignment 0: **Introduction to PyTorch** (0%)
- Assignment 1: **Neural Network Introduction** (5%) — python™
- Assignment 2: **Convolutional Neural Networks** (5%) — PYTORCH
- Assignment 3: RNN **Language Modeling and Translation** (10%) — PYTORCH
- Assignment 4: Neural Model for **Image Captioning / Retrieval** (10%) — PYTORCH

Assignments all use **Python Jupiter Notebooks**, use `Canvas` to hand everything in. Assignments always due at **11:59pm PST** on due date.

Assignments (5 assignments and 40% of grade total)

- Assignment 0: **Introduction to PyTorch** (0%)
- Assignment 1: **Neural Network Introduction** (5%) — python™
- Assignment 2: **Convolutional Neural Networks** (5%) — PYTORCH
- Assignment 3: RNN **Language Modeling and Translation** (10%) — PYTORCH
- Assignment 4: Neural Model for **Image Captioning / Retrieval** (10%) — PYTORCH
- Assignment 5: Advanced Architectures **Graph NN** and **GANs** (10%) — PYTORCH

Assignments all use **Python Jupiter Notebooks**, use `Canvas` to hand everything in. Assignments always due at **11:59pm PST** on due date.

Assignments (5 assignments and 40% of grade total)

I reserve the right to **change** release and due dates for the assignments to accommodate constraints of the course, do not take the dates on web-page as “set in stone”.

Research Papers (reviews and presentation, 20% of grade total)

Presentation - 10%

- You will need to **present 1 paper** individually or as a group (group size will be determined by # of people in class)
- Pick a paper from the syllabus individually (we will have process to pick #1, #2, #3 choices)
- Will need to prepare slides and **meet with me or TA** for feedback
- It is your responsibility to schedule these meetings
- Very likely is that I will ask you to **record** these presentation and we will make these avialble

Research Papers (reviews and presentation, 20% of grade total)

Presentation - 10%

- You will need to **present 1 paper** individually or as a group (group size will be determined by # of people in class)
- Pick a paper from the syllabus individually (we will have process to pick #1, #2, #3 choices)
- Will need to prepare slides and **meet with me or TA** for feedback
- It is your responsibility to schedule these meetings
- Very likely is that I will ask you to **record** these presentation and we will make these avialble

Reading **Reviews** - 10%

- Individually, one for every class after the first half of semester
- Due 11:59pm a day before class where reading assigned, submitted via Canvas

Good **Presentation**

- You are effectively taking on responsibility for being an instructor for part of the class (**take it seriously**)
- What makes a **good presentation**?
  - High-level overview of the problem and motivation
  - Clear statement of the problem
  - Overview of the technical details of the method, including necessary background
  - Relationship of the approach and method to others discussed in class
  - Discussion of strengths and weaknesses of the approach
  - Discussion of strengths and weaknesses of the evaluation
  - Discussion of potential extensions (published or potential)

Reading **Reviews**

- Designed to make sure you read the material and have thought about it prior to class (to stimulate discussion)

  - Short summary of the paper (3-4 sentences)
  - Main contributions (2-3 bullet points)
  - Positive / negative points (2-3 bullet points each)
  - What did you not understand (was unclear) about the paper (2-3 bullet points)

Final **Project** (40% of grade total)

- Group project (groups of 3 are encouraged, but fewer maybe possible)
- Groups are self-formed, you will not be assigned to a group
- You need to come up with a project proposal and then work on the project as a group (each person in the group gets the same grade for the project)
- Project needs to be **research** oriented (not simply implementing an existing paper); you can use code of existing paper as a starting point though

Project proposal + class presentation: 15%
Project + final presentation (during finals week): 35%

Final **Project** (40% of grade total)

- Group project (groups of 3 are encouraged, but fewer maybe possible)
- Groups are self-formed, you will not be assigned to a group
- You need to come up with a project proposal and then work on the project as a group (each person in the group gets the same grade for the project)
- Project needs to be **research** oriented (not simply implementing an existing paper); you can use code of existing paper as a starting point though

Project proposal + class presentation: 15%
Project + final presentation (during finals week): 35%

Sample **Project Ideas**

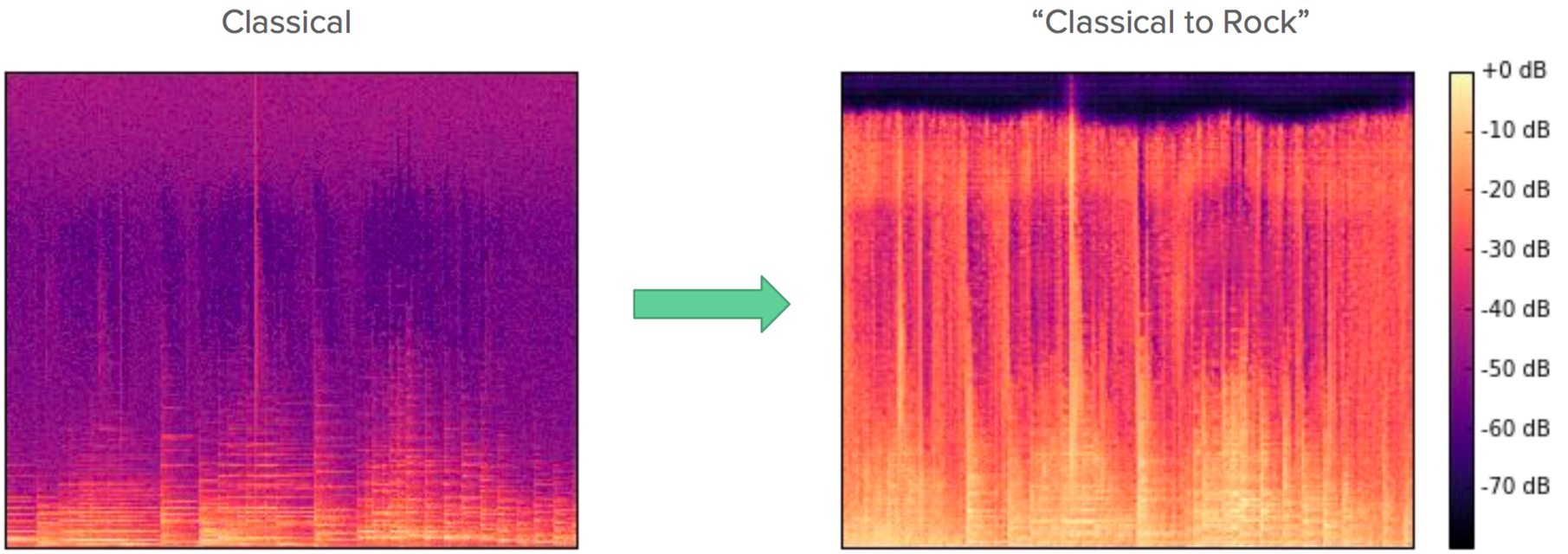
- Translate an image into a cartoon or Picasso drawing better than existing approaches (e.g., experiment with loss functions, architectures)
- Generating video clips by retrieving images relevant to lyrics of songs
- Generating an image based on the sounds or linguistic description
- Compare different feature representation and role of visual attention in visual question answering
- Storyboarding movie scripts
- Grounding a language/sound in an image

… there are **endless possibilities** … think **creatively** and **have fun**!

Project Example: Dreaming of Music

by Sijia (Candice) Tian, Alexandra Kim, Itrat Akhtrt

Evaluate the effectiveness of using visual music representation (spectrograms) to do classification and modify music using deep learning



Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *arXiv preprint arXiv:1703.10593* (2017).

Explored image-to-image translation techniques to translate musical styles

Project Example: Robust Adversarial Detection

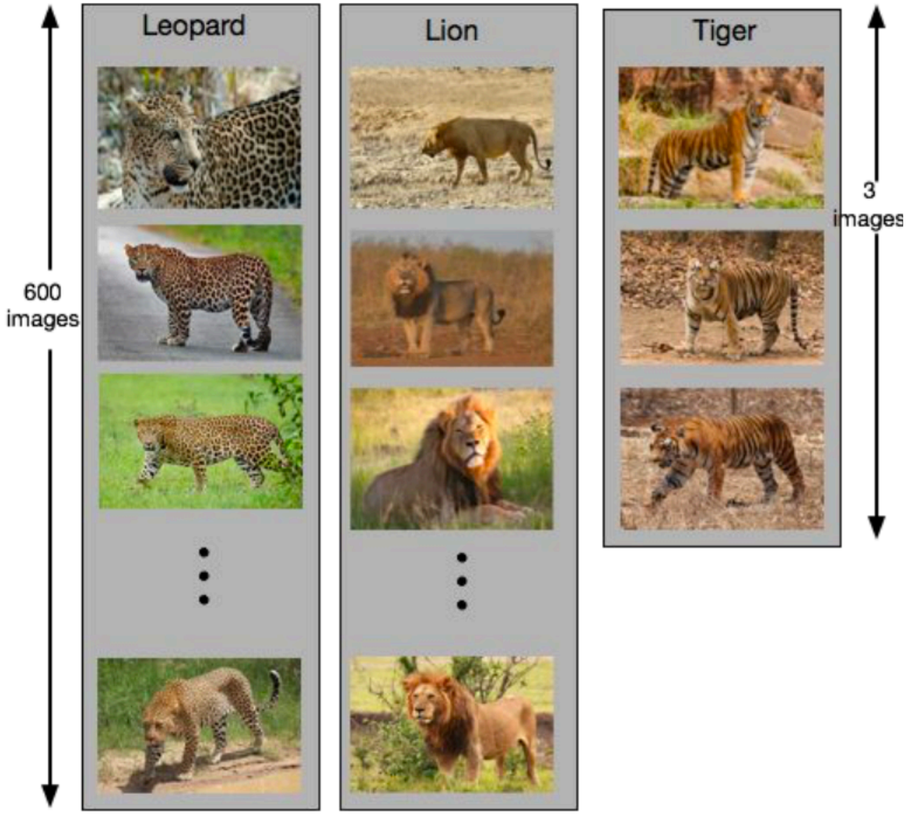
by Michael and Marjan

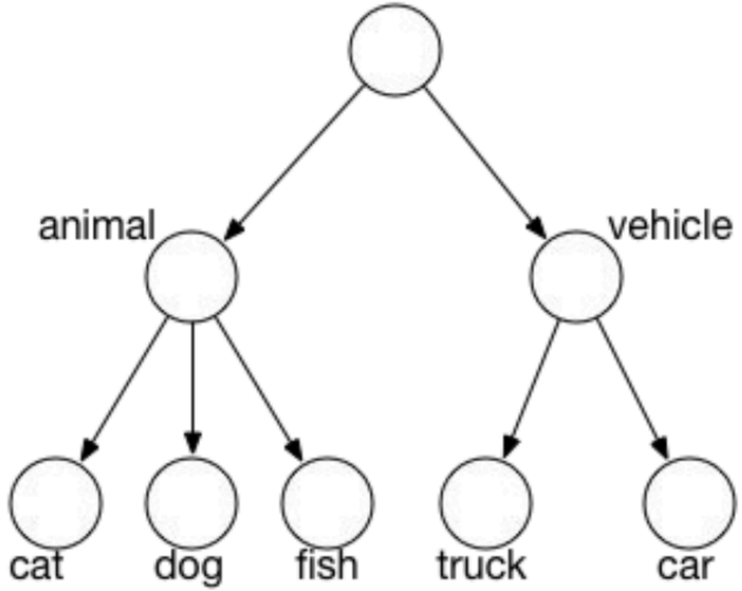
Bayesian Neural Network and variational inference for detecting and analyzing adversarial attacks

Project Example: Classification with Tree Priors

by Saeid Naderiparizi and Setareh Cohan

Classification with few samples using transfer learning techniques

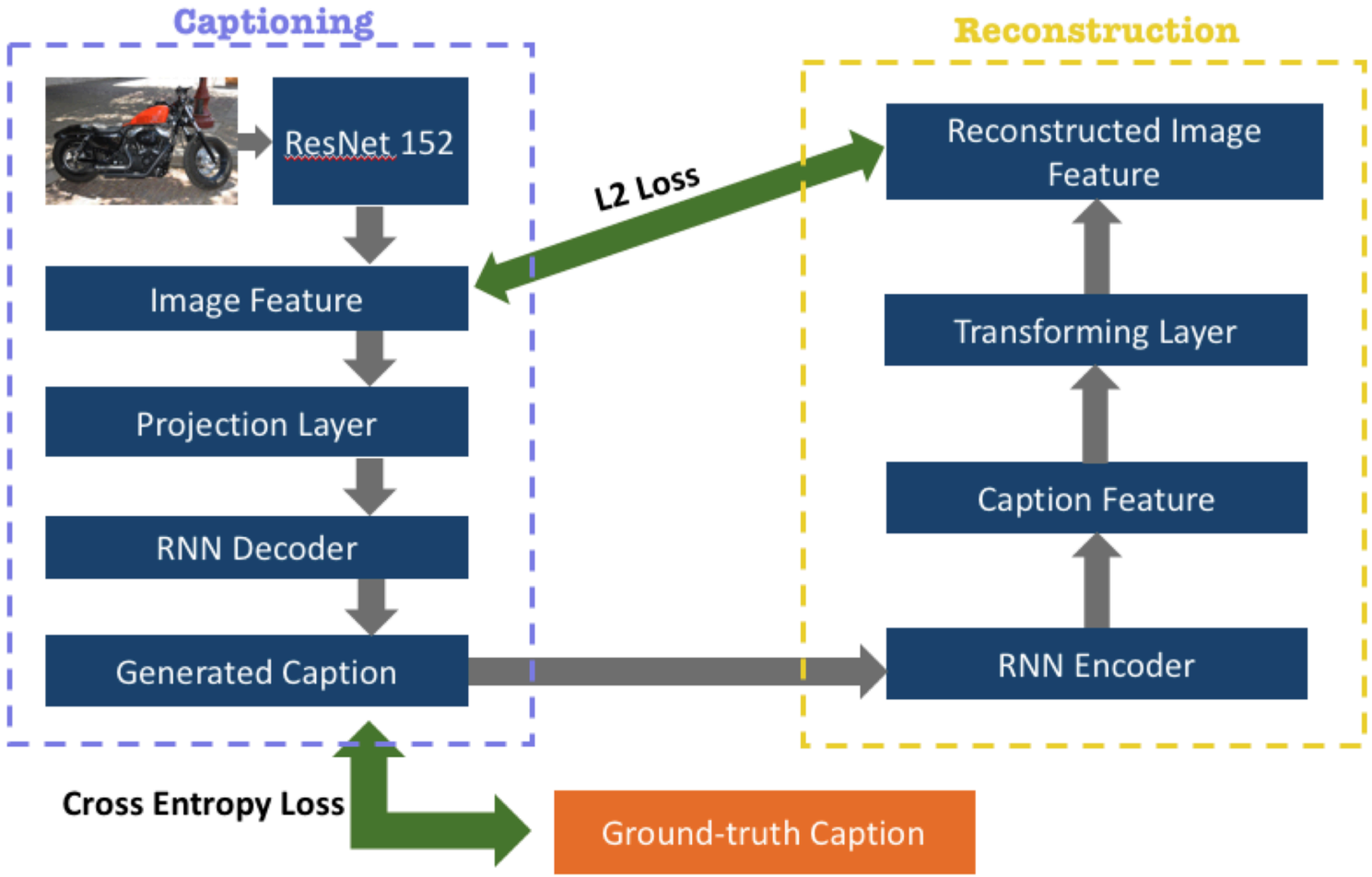




Project Example: Semi-supervised Image Captioning

by Bicheng Xu, Weirui Kong, Jiaxuan Chen

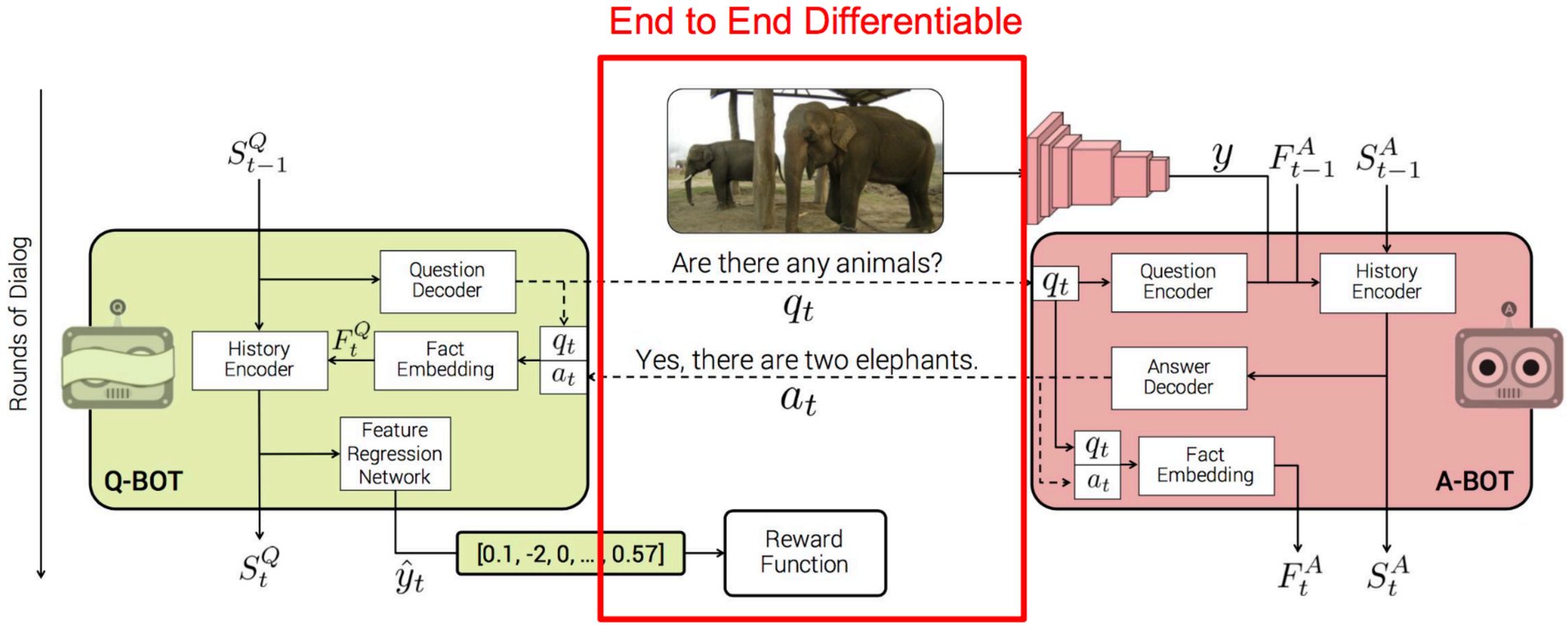
Effective use of unlabeled data during training of an image captioning network



Project Example: Visual Question Answering

by Siddhesh Khandelwal, Mohit Bajaj, Gursimran Singh

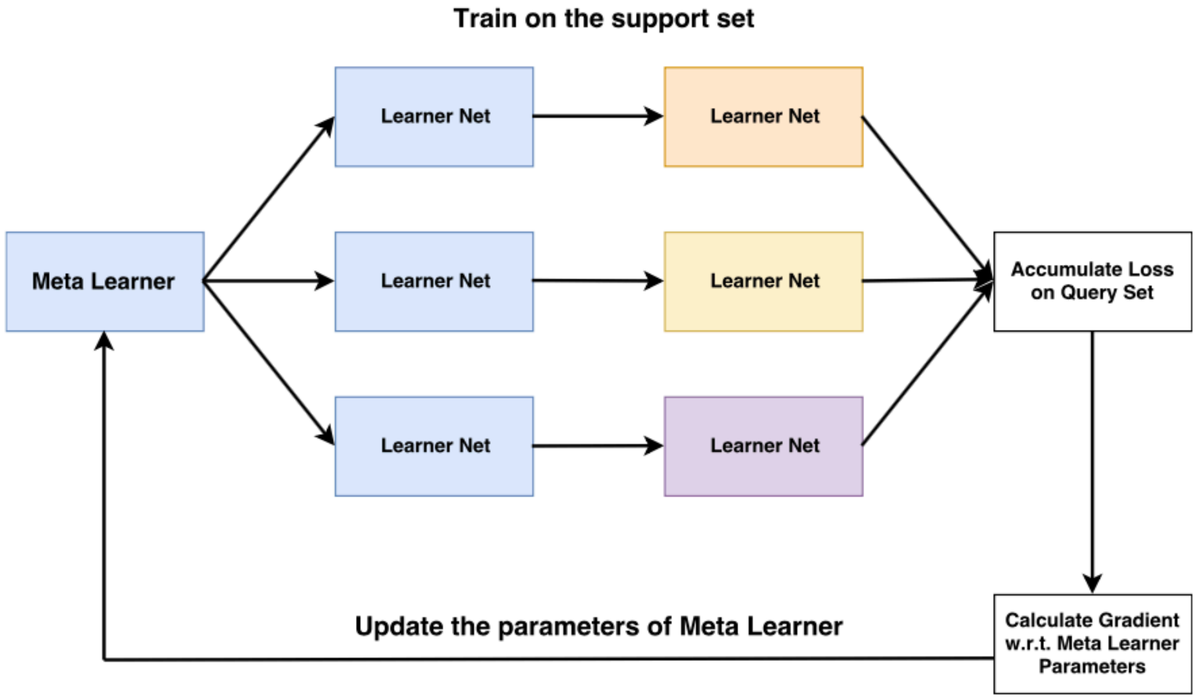
Improve interaction between two agents
- End-to-end differentiability
- Discriminator for human-like questions

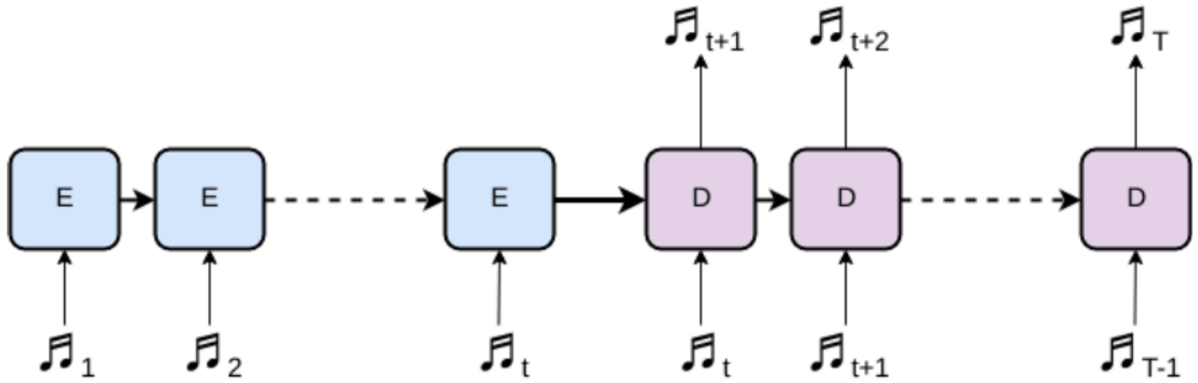


Das, Abhishek, et al. "Learning cooperative visual dialog agents with deep reinforcement learning." *arXiv preprint arXiv:1703.06585* (2017).

Project Example: Few Shot MIDI Music Generation

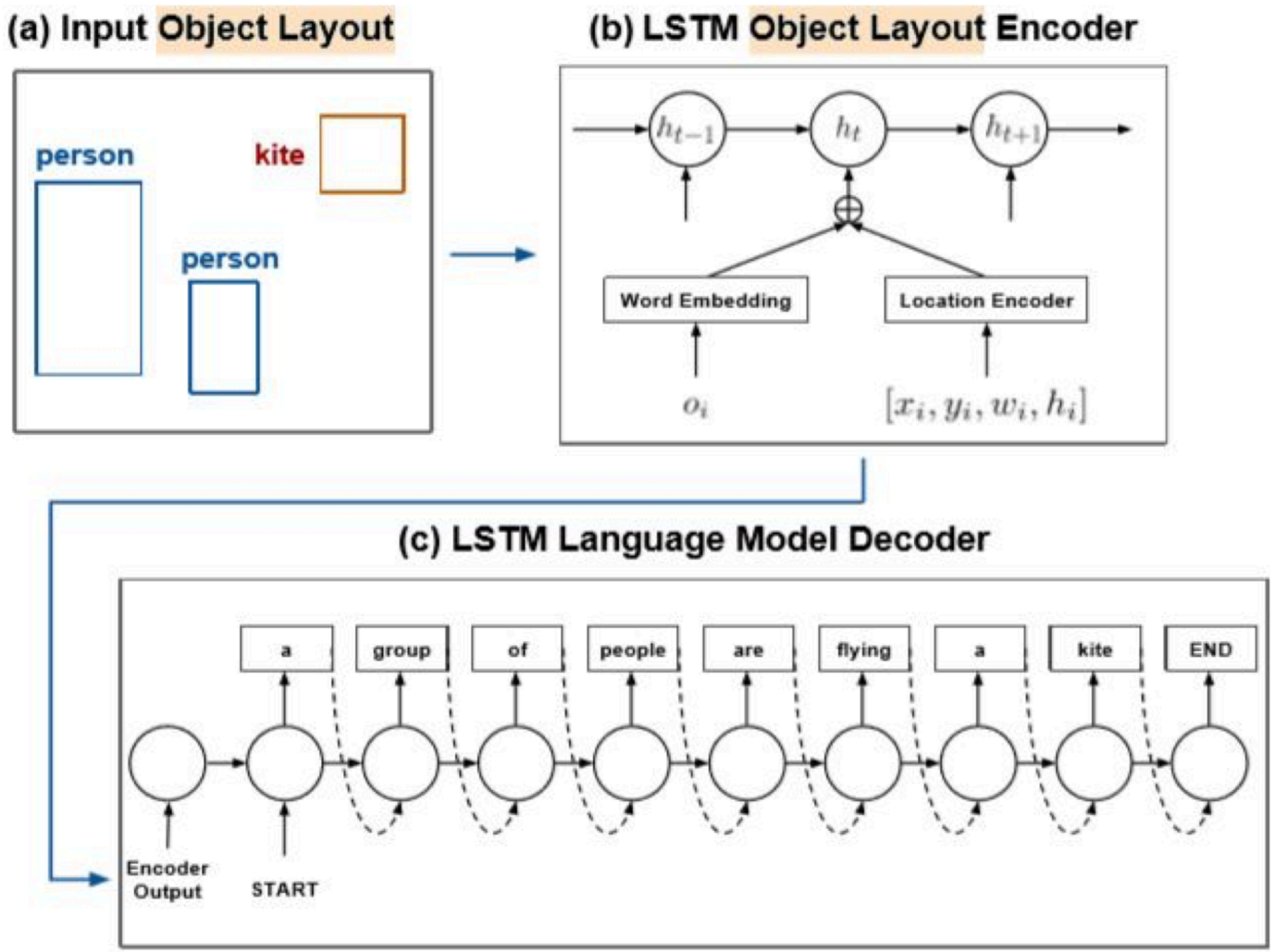
by Ben, Suhail, Anand





Project Example: Visually Descriptive Language from Layout

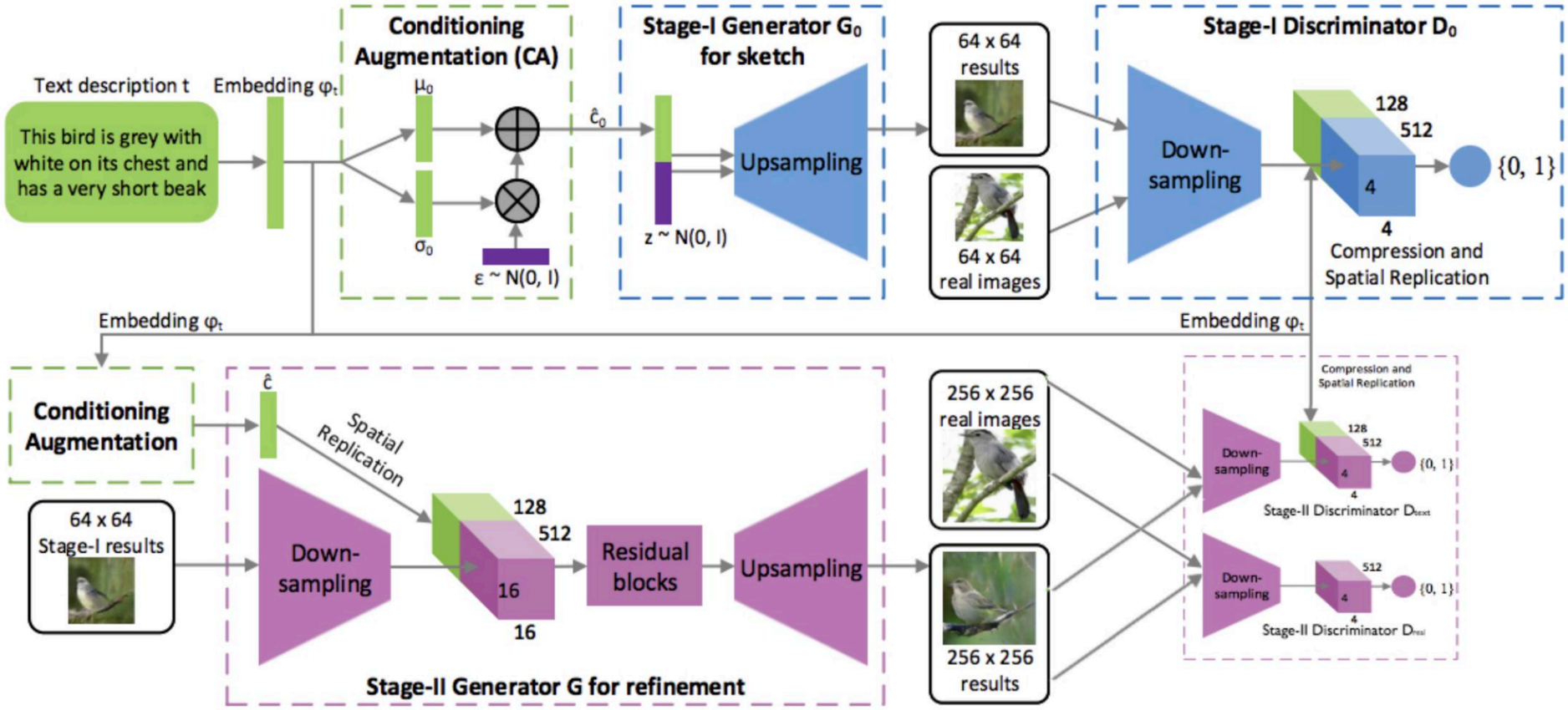
by Ke Ma, Wen Xiao, Sing Zeng



Project Example: StackGAN with Different Losses

by Polina Zablotskaia

Automatic synthesis of realistic images from text



H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In ICCV, 2017.