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

Lecture 1: Introduction
Course logistic

**Times:** Tues & Thurs 11-12:30

**Locations:** ICICS 246

**Instructor:** Leonid Sigal

**E-mail:** lsigal@cs.ubc.ca

**Office:** ICICS 119

**TA:** Shikib Mehri

**E-mail:** mehrishikib@gmail.com

**Course webpage:** [http://www.cs.ubc.ca/~lsigal/teaching.html](http://www.cs.ubc.ca/~lsigal/teaching.html)

**Discussion:** [piazza.com/ubc.ca/other/cpsc532](http://piazza.com/ubc.ca/other/cpsc532)
Course logistic

Times: Tues & Thurs 11-12:30
Locations: ICICS 246

If you have not registered for the course but want to take it, sign up on the sheet, come talk to me after class or schedule a meeting.

Course webpage: http://www.cs.ubc.ca/~lsigal/teaching.html
Discussion: piazza.com/ubc.ca/other/cpsc532l
About me ...

Associate Professor
2017 -

Senior Research Scientist
2009 - 2017

Postdoctoral Researcher
2007 - 2009

PhD, MSc
2001 - 2008
is however extremely hard to train a reasonable regressor with only 4 embedding vectors into auxiliary and testing dataset. In other words, we only need to further split the 6 emotion classes of Ekman dataset embedding vector is trained from video-level features to the corresponding in the semantic word vector space; a regressor function projected into a work, each class-level emotion textual name to the small number of emotion classes. Specifically, in our emotion prediction, our results are very promising.

We randomly select videos from each of the three training and testing classes with 5-round DAP). For comparison purposes, we created the following baselines: (a) using DAP (in the split of 4 auxiliary and 2 testing classes). It is also worth dictating, from which the test class labels are inferred. DAP assignment the highest attribution towards video emotion, that contains the highest attribution towards video emotion, are known, we computed the fraction of participants who discussed later. Since the ground-truth video emotion labels to guess the name of the emotion expressed in the clip. These clips are generated by different baseline techniques, as given all emotion keywords of the corresponding dataset unaware of project goals, were invited for the user study. As discussed earlier, another advantage of our encoding scheme is that we can identify the video clips that have high impact on the overall video emotion. A pilot study we performed indicated that emotions are sparsely expressed in the evaluation protocol of user study to evaluate the performance gain when the training classes bear some similarity between individual emotions is greater in YouTube-8. This result indicates zero-shot learning performs margin improvement over baselines than the two other datasets. This result indicates zero-shot learning performs even without any training examples on these categories, our examples of zero-shot emotion prediction. We highlight that qualitative results.

Fig. 7: Qualitative results of zero-shot emotion recognition. We show the keyframes of three successful cases: the frames of anger reaction of a football fan watching the game, the bottom row is for the grief reaction of fans who are sad about their favorite team losing the game, the top row is about a video of a boredom boy walking and lying on the couch; The bottom row is for the grief reaction of fans who are sad about their favorite team losing the game, the top row is about a video of a boredom boy walking and lying on the couch.
Anger
Boredom
Grief

have at most 4 embedding vectors into auxiliary and testing dataset. In other words, we only need to further split the 6 emotion classes of Ekman dataset embedding vector in the semantic word vector space; a regressor function projected into a work, each class-level emotion textual name emotion prediction, our results are very promising. We highlight that method can still classify these video successfully using the even without any training examples on these categories, our examples of zero-shot emotion prediction. We mention that the results of YouTube-24 have a largest margin improvement over baselines than the two other scenarios is that YouTube-8 contain less emotions than YouTube-8 dataset. An important difference between the two datasets is that contains the highest attribution towards video emotion, as mentioning that the results of YouTube-24 have a largest margin improvement over baselines than the two other scenarios is that YouTube-8 contain less emotions than YouTube-8 dataset. An important difference between the two datasets is that contains the highest attribution towards video emotion, as

Note that Ekman dataset is not used for this tasks due

Figure 4 shows the results. Our ITE+T1S approach produces the best accuracy, outperforming the second best technique on VideoStory-P14 and YouTube-24, but ITE+DAP is the second best technique on the second best technique on VideoStory-P14 and YouTube-24, so the semantic distance

between individual emotions is greater in YouTube-8. This suggests the T1S technique contributes the biggest performance gain when the training classes bear some similarity to the unseen test classes. However, when the training classes are very different from the testing classes, the ITE margin improvement over baselines than the two other scenarios is that YouTube-8 contain less emotions than YouTube-8 dataset. An important difference between the two datasets is that contains the highest attribution towards video emotion, as

As the first work on video emotion attribution, we define

We compare our T1S algorithm with Direct Attribution

vector smoothing by Eq (7). Four variants are compared: (a)

dicted, from which the test class labels are inferred. DAP

dimension of the word vectors of each test sample is pre-

We randomly select

On average, around

These clips are generated by different baseline techniques, as

to guess the name of the emotion expressed in the clip. Given all emotion keywords of the corresponding dataset unaware of project goals, were invited for the user study.

With only 4 embedding vectors.
What is Multi-modal Learning?
What is **Multi-modal Learning**?

- **Modality**: refers to a certain type of information and/or representation format in which information is stored.

- **Sensory modality**: one or more primary channels of communication.
What is Multi-modal Learning?

- **Modality:** refers to a certain type of information and/or representation format in which information is stored.

- **Sensory modality:** one or more primary channels of communication.
What is **Multi-modal Learning**?

- **Modality**: refers to a certain type of information and/or representation format in which information is stored.

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

**Visual (image)**

**Natural Language (text)**

**Auditory (voice / sound)**
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 and temporal synchrony in multimodal events

**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

The McGurk Effect

McGurk Effect (1976)

* video credit: OK Science

* Adopted from slides by Louis-Philippe Morency
Multimodal Research: Historical Perspective

The McGurk Effect

McGurk Effect (1976)

* Adopted from slides by Louis-Philippe Morency

* video credit: OK Science
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)

* 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]
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]
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
Deep Learning (a.k.a. representation learning)
- Better performance
- More interesting problems emerging

THIS IS OUR COURSE
Multimodal Research: Historical Perspective

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

Natural language description generation

[ Vinyals et al., 2015 ]
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 ]
Multimodal Research: Historical Perspective

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

Detecting objects based on linguistic descriptions

[ Ba et al., ICCV 2015 ]
Multimodal Research: Historical Perspective

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

Book-to-Movie alignment

[ Zhu et al, ICCV 2015 ]
Multimodal Research: Historical Perspective

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

Book-to-Movie alignment

[Zhu et al, ICCV 2015]
Multimodal Research: Historical Perspective

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

[ Kim & Sigal, CVPR 2015 ]
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

Visual question answering / dialog

[ Seo et al., NIPS 2017 ]
Multimodal Research: Historical Perspective

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

Narrative plot understanding

[ Iyyer et al., CVPR 2017 ]
**Multimodal Research: Historical Perspective**

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

- [Zhu et al., ICCV 2017]
- [Isola et al., CVPR 2017]

Image-to-image 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)

Video-to-Audio translation

[ Iyyer et al., NIPS 2016 ]
Multimodal Research: Historical Perspective

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

[ Iyyer et al., NIPS 2016 ]

Video-to-Audio translation
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 ]
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**
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 Anna Lee

Today, Google announced the acquisition of DNNresearch, a three-person Canadian research company that specializes in voice and image recognition. DNNresearch, which was founded last year within the University of Toronto's computer science department, specializes in object recognition and how objects belong to Google.

From left: Yisong Yue, Alex Krizhevsky and University Professor Geoffrey Hinton of the University of Toronto's Department of Computer Science. (Photo by John Gatto, University of Toronto)
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.

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.

Machine Learning Startup Acquired by ai-one

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.
Clever Hans

Clever Hans
(Orlov Trotter horse)

Wilhelm von Osten
Hans could get 89% of the math questions right.
Clever Hans

Clever Hans
(Orlov Trotter horse)

Wilhelm von Osten

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

Hans could get 89% of the math questions right
Is there zebra climbing the tree?

Yes
Pre-requisites

**Computer Science**

- **CPSC 340** (or equivalent)
- Python

**Mathematics**

- Calculus
- Linear Algebra
- Statistics

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

You will be given credits to use

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.

or use your own …

Nvidia GTX 1060 (with 6GB RAM) or above
Course structure

Approximately 50% of course will consist of lectures and optional readings.

Remaining 50% is reading and presentation of curated research papers on relevant topics.

4 programming assignments.

Final (individual or group) project.
Grading Criteria

- Assignments (programming) — 30% (total)
- Research papers — 20%
- Group project — 50%
Grading Criteria

• Assignments (programming) — 30% (total)

• Research papers — 20%

• Group project — 50%

NO LATE SUBMISSIONS — If you don’t complete the assignment, hand in what you have
Assignments (4 assignments and 30% of grade total)

- Assignment 1: **Neural Network Introduction** (5%)
- Assignment 2: **Convolutional Neural Networks** (5%)
- Assignment 3: **RNN Language Modeling** (10%)
- Assignment 4: **Neural Model for Image Captioning / Retrieval** (10%)

Assignments all use **Python Jupiter Notebooks**, use `handin` to hand everything in. Assignments always due at **5pm PST** on Fridays.
Assignments (4 assignments and 30% of grade total)

- Assignment 1: **Neural Network Introduction** (5%) — 🐍
- Assignment 2: **Convolutional Neural Networks** (5%) — 🍊
- Assignment 3: **RNN Language Modeling** (10%) — 🎨
- Assignment 4: **Neural Model for Image Captioning / Retrieval** (10%) — 📸

Assignments all use **Python Jupiter Notebooks**, use `handin` to hand everything in. Assignments always due at **5pm PST** on Fridays.
Assignments (4 assignments and 30% of grade total)

- Assignment 1: Neural Network Introduction (5%) — 🐍 python
- Assignment 2: Convolutional Neural Networks (5%) — PyTorch
- Assignment 3: RNN Language Modeling (10%) — PyTorch
- Assignment 4: Neural Model for Image Captioning / Retrieval (10%) — PyTorch

Assignments all use Python Jupiter Notebooks, use `handin` to hand everything in. Assignments always due at 5pm PST on Fridays.
Assignments (4 assignments and 30% of grade total)

- Assignment 1: **Neural Network Introduction** (5%) — Python
- Assignment 2: **Convolutional Neural Networks** (5%) — PyTorch
- Assignment 3: RNN **Language Modeling** (10%) — PyTorch
- Assignment 4: Neural Model for **Image Captioning / Retrieval** (10%) — PyTorch

Assignments all use **Python Jupiter Notebooks**, use `handin` to hand everything in. Assignments always due at **5pm PST** on Fridays.
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 (send me via e-mail your #1, #2, #3 choices)
- Will need to prepare slides and meet with me in person at least 2 days before your scheduled presentation for me to provide feedback.
- It is your responsibility to schedule these meetings.

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 Piazza
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)
• 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** (50% 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: 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**!