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

Lecture 1: Introduction
**Course logistic**

**Times:** Tues & Thurs 9:30-11:00am

**Locations:** DMP 101

**Instructor:** Leonid Sigal

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

**Office:** ICICS 119

**TA:** Mohit Bajaj

**E-mail:** mbajaj01@cs.ubc.ca

**Discussion:** piazza.com/ubc.ca/winterterm22018/cpsc532s
Course logistic

Times: Tues & Thurs 9:30-11:00am
Locations: DMP 101

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: [https://www.cs.ubc.ca/~lsigal/teaching18.html](https://www.cs.ubc.ca/~lsigal/teaching18.html)
Discussion: [piazza.com/ubc.ca/winterterm22018/cpsc532s](piazza.com/ubc.ca/winterterm22018/cpsc532s)
About me …

Associate Professor
2017 -

Senior Research Scientist
2009 - 2017

Postdoctoral Researcher
2007 - 2009

PhD, MSc
2001 - 2008
Anger
Boredom
Grief

is however extremely hard to train a reasonable regressor into auxiliary and testing dataset. In other words, we only embedding vector is trained from video-level features to the corresponding 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.

encoded feature. Thus considering the difficulty of zero-shot method can still classify these video successfully using the even without any training examples on these categories, our

inherent difficulties of the zero-shot learning task, we consider the results to be very promising. Given the

of ITE+T1S is effective under different zero-shot learning conditions. Given the

on VideoStory-P14 and YouTube-24, so the semantic distance two scenarios is that YouTube-8 contain less emotions than 24, but ITE+DAP is the second best technique on the absolute percentage points. We observe that AvgP+T1S is baseline by 3.6, 4.8, and 1.2 absolute percentage points produces the best accuracy, outperforming the second best (in the split of 4 auxiliary and 2 testing classes). It

duplicate experiments. In the zero-shot setting, no instances are known, we computed the fraction of participants who are invited for the user study. We compare our T1S algorithm with Direct Attribution

w2
R(K)=

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

as

that contains the highest attribution towards video emotion, these clips are generated by different baseline techniques, as and clip computed from the video, participants are asked to guess the name of the emotion expressed in the clip. We show the keyframes of three successful cases: the frames of top row shows a video clip of an anger parade; the middle 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. We randomly select videos from each of the three

4.4 Video Emotion Attribution

Random sampling

Chance

VideoStory-P

0

36.5%

50.4%

68.9%

7.14%

10%

16.7%

34.6%

37.6%

52%

52%

76.5%

72.4%

60%

36.5%

52%

48.3%

72.4%

76.5%

72.4%

76.5%
is however extremely hard to train a reasonable regressor.

Having at most 4 embedding vectors, each class-level emotion textual name is trained from video-level features to the corresponding dimension of the word vectors of each test sample is pre-determined. This allows us to infer the test class labels. For DAP, at test time each keyframe is passed through the encoding scheme, and the attribution towards video emotion is computed. The keyframe with the highest attribution is then used to guess the name of the emotion expressed in the clip.

Given all emotion keywords of the corresponding dataset to emotion in our three datasets. We randomly select 0.66 videos from each of the three datasets. For each video, we extract a 2-second video clip and clip computed from the video, participants are asked to guess the name of the emotion expressed in the clip.

As discussed earlier, another advantage of our encoding scheme is that we can identify the video clips that have the highest impact on the overall video emotion. A pilot study we performed indicated that emotions are sparsely expressed in videos. On average, around 100% of the participants correctly guessed the emotion expressed in the clip.

Figure 4 shows the results. Our ITE+T1S approach produces the best accuracy, outperforming the second best technique on the YouTube-8 dataset. An important difference between the two scenarios is that YouTube-8 contains less emotions than the other datasets. For YouTube-24, we randomly split the training and testing classes with 5-round cross-validation. The experiments show the combination of ITE+T1S is effective under different zero-shot learning tasks. We consider the results to be very promising.

Table 1 summarizes the average accuracy of different techniques on 16 emotion categories. We observe that AvgP+T1S, which is the average of predictions using different video-level feature representation (AvgP or T1S), produces the best accuracy, outperforming the second best technique on VideoStory-P14 and YouTube-24, but ITE+DAP is the second best technique on YouTube-8. The results of YouTube-8 have the largest margin improvement over baselines than the two other datasets. For each technique, we present the absolute percentage points. We observe that AvgP+T1S is the best technique on YouTube-8, producing the best accuracy, outperforming the second best technique by 3.6, 4.8, and 1.2 absolute percentage points, respectively and the random baseline by 8.1, 6.3, and 15.9 absolute percentage points.
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.
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

* 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*
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
The McGurk Effect

McGurk Effect (1976)

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

* Adopted from slides by Louis-Philippe Morency

* video credit: OK Science
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

Dongwook Yoon

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

* 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!”

Story generation
Multimodal Research: Historical Perspective

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

Detecting objects based on linguistic descriptions

Corn Poppy

Papaver rhoes (common names include corn poppy, corn rose, field poppy, Flanders poppy, red poppy, red weed, coquelicot, and, due to its colour, which is said to cause them, as headache and hiccough) is a species of flowering plant in the poppy family, Papaveraceae. This poppy, native to Europe, is notable as an agricultural weed (hence the “corn” and “field”) and as a symbol of fallen soldiers. P. rhoes is sometimes 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.

[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)

... 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 ]
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)

Image-to-image translation

[ Zhu et al., ICCV 2017 ]

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

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

Video-to-Audio translation

[ Iyyer et al., NIPS 2016 ]
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

*course is heavy on *practical* deep learning*
Google snaps up object recognition startup

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

Google has acquired 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 now belongs to Google.

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

Hans could get 89% of the math questions right
The horse was smart, just not in the way van Osten thought!

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

AI agent

Yes
# Pre-requisites

## Computer Science

<table>
<thead>
<tr>
<th>CPSC 340 (or equivalent)</th>
<th>Needed for Assignments</th>
</tr>
</thead>
</table>

## Mathematics

<table>
<thead>
<tr>
<th>Calculus</th>
<th>Linear Algebra</th>
<th>Statistics</th>
</tr>
</thead>
</table>

**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%

**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 Canvas to hand everything in. Assignments always due at **5pm PST** on due date.
Assignments (4 assignments and 30% of grade total)

• Assignment 1: Neural Network Introduction (5%) — Python

• Assignment 2: Convolutional Neural Networks (5%) — PyTorch

Assignments all use Python Jupiter Notebooks, use Canvas to hand everything in. Assignments always due at 5pm PST on due date.
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

Assignments all use Python Jupiter Notebooks, use Canvas to hand everything in. Assignments always due at 5pm PST on due date.
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 Canvas to hand everything in. Assignments always due at **5pm PST** on due date.
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) [7.5%]

• 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 in person at least 2 days before your scheduled presentation for me to provide feedback.

• It is your responsibility to schedule these meetings.

• You will also need to argue against one of the papers [2.5%]
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) [7.5%]

• 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 in person at least 2 days before your scheduled presentation for me to provide feedback.

• It is your responsibility to schedule these meetings.

• You will also need to argue against one of the papers [2.5%]

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)
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** (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 (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 Akhtar

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

Explored image-to-image translation techniques to translate musical styles.
Project Example: Robust Adversarial Detection

Bayesian Neural Network and variational inference for detecting and analyzing adversarial attacks
Classification with few samples using transfer learning techniques

Project Example: Classification with Tree Priors

by Saeid Naderiparizi and Setareh Cohan
Project Example: Semi-supervised Image Captioning

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

by Bicheng Xu, Weirui Kong, Jiaxuan Chen
**Project Example:** Visual Question Answering

Improve interaction between two agents

- End-to-end differentiability
- Discriminator for human-like questions

---

**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

Automatic synthesis of realistic images from text