Vision & Language

Introduction
About the class

- CS6501/CS4501: Vision and Language
- Instructor: **Vicente** Ordonez (Vicente Ordóñez Román)
- Website: [https://www.vicenteordonez.com/vislang/](https://www.vicenteordonez.com/vislang/)
- Location: Zoom – UVA Collab has the links for every class
- Times: Tuesdays & Thursdays 3:30pm to 4:45pm
- Office Hours: Fridays between 1pm and 3pm.
- Discussion Forum: Campuswire (you must have received an invitation)
Our TAs

Paola Cascante-Bonilla
Office Hours: Fridays between 3 and 4pm

Ziyan Yang
Office Hours: Wednesdays between 1 and 2pm

Anshuman Suri
Office Hours: Mondays between 1 and 2pm
Instructor: Vicente Ordóñez-Román (vicente at virginia.edu), Office Hours: Fridays between 1pm and 3pm (ET)
Teaching Assistant: Ziyao Yang (zy9cx at virginia.edu), Office Hours: Wednesdays between 1pm and 2pm (ET)
Teaching Assistant: Paola Casacante-Bonilla (pccb2 at virginia.edu), Office Hours: Fridays between 3pm and 4pm (ET)
Teaching Assistant: Anshuman Suri (astsw at virginia.edu), Office Hours: Mondays between 1pm and 2pm (ET)

Class Time: Tuesdays & Thursdays between 3:30PM and 4:45PM (ET).
Discussion Forum: https://campuswire.com/c/G8B171FE1

Course Description: Visual recognition and language understanding are two challenging tasks in AI. In this course we will study and acquire the skills to build machine learning and deep learning models that can reason about images and text for generating image descriptions, visual question answering, image retrieval, and other tasks. On the technical side we will leverage models such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer networks (e.g. BERT), among others.

Learning Objectives: (a) Develop intuitions about the connections between language and vision, (b) Understanding foundational concepts for representation learning for both images and text, (c) Become familiar with state-of-the-art models for tasks in vision and language, and (d) Obtain practical experience in the implementation of these models.

Prerequisites: It is recommended to have had a prior class in any of the following: Machine Learning, Computer Vision, Deep Learning for Visual Recognition, Natural Language Processing, Artificial Intelligence, or similar. Students are encouraged to complete the following activities before the first lecture: Completing this [Primer on Image Processing], and the tutorial and assignment on [Image Classification] from the Deep Learning for Visual Recognition class.

Schedule

<table>
<thead>
<tr>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tue, Aug 25th</td>
<td>Introduction to Vision and Language.</td>
</tr>
<tr>
<td>Thu, Aug 27th</td>
<td>Machine Learning Primer</td>
</tr>
<tr>
<td>Tue, Sep 1st</td>
<td>Computer Vision Primer</td>
</tr>
<tr>
<td>Thu, Sep 3rd</td>
<td>Natural Language Processing Primer</td>
</tr>
<tr>
<td>Tue, Sep 8th</td>
<td>Recurrent Neural Networks (RNNs, LSTMs, GRUs)</td>
</tr>
<tr>
<td>Thu, Sep 10th</td>
<td>Image Captioning Models and Datasets (e.g. Show-and-Tell, Show-Attend-and-Tell)</td>
</tr>
<tr>
<td>Tue, Sep 15th</td>
<td>Referring Expressions (Module Networks - MATNet)</td>
</tr>
<tr>
<td>Thu, Sep 17th</td>
<td>Transformer Models and Self-Attention (e.g. BERT, GPT, XLNet)</td>
</tr>
</tbody>
</table>
Welcome to Vision & Language #1

Hi Everyone welcome to the class on Vision and Language. This semester this class will focus on advanced topics but we will have the basics covered although probably quickly which may be not at the satisfaction of everyone, or maybe too slow if people in the class already have taken a machine learning class. Nevertheless by the third week of class we hope everyone has a more or less equal footing on basics and we can get started building systems and models that can reason about images and text and often jointly about images and text. The class emphasizes this semester on building something you can show off to your friends – that is I will require your project to have an end product that people can interact with in some reasonable capacity. I'm not interested in just a project report. Although a project report is also in order. I hope everyone enjoys the class and we all learn together. We are in the middle of a pandemic so excuse us if we are less than 100% and there are some issues on the way. Let's all have a good time here.

Comments

No one's commented here... yet

Be a maverick and get the conversation going.
<table>
<thead>
<tr>
<th>Start Time</th>
<th>Topic</th>
<th>Meeting ID</th>
<th>Join</th>
</tr>
</thead>
<tbody>
<tr>
<td>Today (Recurring)</td>
<td>Vision &amp; Language</td>
<td>984 9859 0421</td>
<td></td>
</tr>
<tr>
<td>3:30 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu, Aug 27 (Recurring)</td>
<td>Vision &amp; Language</td>
<td>984 9859 0421</td>
<td></td>
</tr>
<tr>
<td>3:30 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fri, Aug 28 (Recurring)</td>
<td>Office Hours -- Paola</td>
<td>962 4059 1595</td>
<td></td>
</tr>
<tr>
<td>3:00 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mon, Aug 31 (Recurring)</td>
<td>Office Hours -- Anshuman</td>
<td>984 8671 4505</td>
<td></td>
</tr>
<tr>
<td>1:00 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue, Sep 1 (Recurring)</td>
<td>Vision &amp; Language</td>
<td>984 9859 0421</td>
<td></td>
</tr>
<tr>
<td>3:30 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wed, Sep 2 (Recurring)</td>
<td>Office Hours -- Ziyang</td>
<td>960 4614 3694</td>
<td></td>
</tr>
<tr>
<td>1:00 PM</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## About me -- Vicente

<table>
<thead>
<tr>
<th>Position</th>
<th>Organization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assistant Professor, 2016 - Now</td>
<td>University of Virginia</td>
</tr>
<tr>
<td>Visiting Professor, 2019</td>
<td>Adobe Research</td>
</tr>
<tr>
<td>Visiting Researcher, 2015 - 2016</td>
<td>Allen Institute for AI</td>
</tr>
<tr>
<td>MS, PhD in CS, 2009-2015</td>
<td>The University of North Carolina at Chapel Hill</td>
</tr>
<tr>
<td></td>
<td>Stony Brook University</td>
</tr>
</tbody>
</table>

... also spent time at:

- Google
- Microsoft
- eBay
What is Vision and Language?

Anything at the intersection of Computer Vision and Natural Language Processing. Systems and models that depend a little bit on both.

- **Computer Vision**: How do we teach machines to process, represent and understand images? e.g. to recognize objects in images.

- **Natural Language Processing**: How do we teach machines to process, represent and understand text? e.g. to classify or generate text.
Some of our work includes...

Describing images with language

- Millions of samples
  - Matching using Global Image Features (GIST + Color)

- Transfer Caption(s)
  - e.g. “The water is clear enough to see fish swimming around in it.”

Im2Text: Describing Images Using 1 Million Captioned Photographs
Vicente Ordonez, Girish Kulkarni, Tamara L. Berg.
Some of our work includes...

Describing images with language

- Detect: dog
  - Find matching dog detections by visual similarity

- Retrieving verb phrases from similar object detections

- Peruvian dog sleeping on city street in the city of Cusco, Peru
- Closeup of my dog sleeping under my desk.

- Contexted dog just laying on the edge of the road in front of a house.

- This dog was laying in the middle of the road on a back street in Jaco.

Large Scale Retrieval and Generation of Image Descriptions
International Journal of Computer Vision. IJCV 2015. [August 2016 Issue]. [pdf] [link] [bibtex]
Some of our work includes...

Describing images with language

Obj2Text: Generating Visually Descriptive Language from Object Layouts
Xuwang Yin, Vicente Ordonez. Empirical Methods in Natural Language Processing. EMNLP 2017. Copenhagen, Denmark. September 2017. [pdf] [arxiv] [code] [bibtex]
(Oral presentation)
Some of our work includes...

Hierarchical Classification and Entry-level Categorization

From Large Scale Image Categorization to Entry-Level Categories
Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, Tamara L. Berg.
IEEE International Conference on Computer Vision. ICCV 2013. Sydney, Australia.
December 2013. [pdf] [supplemental material] [slides] [project page] [bibtex] (~Oral Presentation + Best Paper Award - Marr Prize!)
Some of our work includes...

Situation Recognition

http://imsitu.org/

Commonly Uncommon: Semantic Sparsity in Situation Recognition
Some of our work includes...

Generating Images from Text

https://www.vislang.ai/text2scene

Text2Scene: Generating Compositional Scenes from Textual Descriptions
Some of our work includes...

Learning from Images with Textual Descriptions

https://www.vislang.ai/genderless

Some of our work includes...

Interactive Image Retrieval

Drill-down: Interactive Retrieval of Complex Scenes using Natural Language Queries
Some of our work includes...

Interactive Image Retrieval

(1) red brick of fireplace
(2) china plates and glasses
(3) group of three candle sticks on mantel
(4) flowers on the dining table
(5) candle style chandelier hanging down from ceiling
(6) wooden chairs on the carpet

New Query

State Vectors $X^{t-1}$

GRU

State Vectors $X^{t}$

CROSS MODAL SIMILARITY $s(X, I)$

Region Features

Faster RCNN

Drill-down: Interactive Retrieval of Complex Scenes using Natural Language Queries
Drill-down: Image Retrieval System

Target
Drill-down: Image Retrieval System

Two people in a ski field
Drill-down: Image Retrieval System

The man is wearing a black hat
Drill-down: Image Retrieval System

The woman is wearing a pink coat
Drill-down: Image Retrieval System

they both have goggles
Some of our work includes...

**Referring Expressions**

**ReferItGame Dataset**

130k Referring expressions for 90k Objects in 19k images

**ReferItGame: Referring to Objects in Photographs of Natural Scenes**

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, Tamara L. Berg.

Empirical Methods on Natural Language Processing. EMNLP 2014. Doha, Qatar. October 2014. [pdf] [project page] [game] [bibtex] (~Oral presentation)
A lot of other work on Vision and Language!

**MAttNet: Modular Attention Network for Referring Expression Comprehension**

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, Tamara L.Berg
A lot of other work on Vision and Language!

**UNITER: UNiversal Image–TEXT Representation Learning**

A lot of other work on Vision and Language!

Guest Lecture: Biases in Vision and Language: Visual Question Answering

Kushal Kafle, Research Scientist at Adobe Research
Previously at RIT and extensive work in vision and language and VQA.

Answer Them All! Toward Universal Visual Question Answering Models

Robik Shrestha, Kushal Kafle, Christopher Kanan
A lot of other work on Vision and Language!

Guest Lecture: Visually Grounded Explanations for Physical Tasks

Nazneen Rajani, Research Scientist at Salesforce Research

Previously at UT Austin and extensive work on NLP, machine learning, explainable AI.

ESPRIT: Explaining Solutions to Physical Reasoning Tasks

Nazneen Fatema Rajani, Rui Zhang, Yi Chern Tan, Stephan Zheng, Jeremy Weiss, Aadit Vyas, Abhijit Gupta, Caiming Xiong, Richard Socher, Dragomir Radev
A lot of other work on Vision and Language!

**Instruction:** Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

**Guest Lecture:** Vision-and-Language Navigation (VLN)

**Peter Anderson**, Research Scientist at **Google Research**

Previously at Georgia Tech and extensive work in vision and language and VLN.

**Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments**

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, Anton van den Hengel
A lot of other work on Vision and Language!
A lot of other work on Vision and Language!

**Visual to Sound: Generating Natural Sound for Videos in the Wild**

Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, Tamara L. Berg

**Guest Lecture:** Grounding Vision to Sound and Audio

Yipin Zhou, Research Scientist at Facebook AI

Previously at UNC and extensive work with images, sound, and video.
A lot of other work on Vision and Language!

Guest Lecture: When is Grounding Helpful for Language and Vision Tasks?

Lisa Anne Hendricks, Research Scientist at DeepMind
Previously at UC Berkeley and extensive work with deep learning, computer vision, and vision and language.

Women also Snowboard: Overcoming Bias in Captioning Models
Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, Anna Rohrbach
What will we cover in this class?

In terms of tools

• Introduction to ML / Vision / NLP
• Neural Networks (NNs) / Deep Learning.
• Convolutional Neural Networks (CNNs)
• Recurrent Neural Networks (RNNs, LSTMs, GRUs)
• Transformers (e.g. BERT, GPT, UNITER, etc)
• Other topics as they arise: GANs, Reinforcement Learning, Probabilistic models.
What will we cover in this class?

In terms of topics

• Image Captioning
• Referring Expression Comprehension
• Visually-grounded Question Answering
• Learning from Text and Images
• Visually-grounded Dialog
• Retrieving Images from Natural Language Queries
• Generating Images from Text
• Multimodal Translation using both Images and Text
• Vision-Language Navigation
• Biases in Vision and Language Tasks
• Possibly more topics...
Pre-requisites

• No formal pre-requisites but...

• You need to know how to program with Python or be VERY motivated to learn as you go.

• You need to know some basics of Machine Learning or be VERY motivated to do some self-learning as you go.

• You need to be proficient on basic calculus, linear algebra, and statistics. Nothing advanced but the right basic terminology is needed.
Grading for this class: CS 4501

• Assignments: 300pts (2 assignments: 150pts + 150pts)
• Class Project: 600pts
• Peer Reviews/Participation: 100pts

• Grade cutoffs: A+ (1000pts), A (930pts), A- (900pts), B+ (870pts), B (830pts), B- (800pts), C+ (770pts), C (730pts), C- (700pts), D+ (670pts), D (630pts), D- (600pts).
Grading for this class: CS 6501

• Assignments: 300pts (3 assignments: 100pts + 100pts + 100pts)
• Class Project: 600pts
• Peer Reviews/Participation: 100pts

• Grade cutoffs: A+ (1000pts), A (930pts), A- (900pts), B+ (870pts), B (830pts), B- (800pts), C+ (770pts), C (730pts), C- (700pts), D+ (670pts), D (630pts), D- (600pts).
We will also be using...

https://colab.research.google.com/
However for your projects...

Genderless
Our group has produced several models and diagnostic methods for addressing gender bias in natural language processing and computer vision. Here we leverage our ICCV 2019 paper: Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In this paper we proposed a method to adversarially remove as much as possible from an image any features that could be predictive of whether a person will use a gendered word to describe it. We used a large dataset of images with captions and selected images that had references in the text such as "man" or "woman" and trained a model that can recognize the objects in the image but has as much difficulty as possible in predicting gender. When we applied this transformations to the image space, we can examine what the model is trying to do. Try your own images below and see what it does.
However for your projects...

Text2Scene

Text2Scene was proposed in a paper by our group at CVPR 2019 as Text2Scene: Generating Compositional Scenes from Textual Descriptions. This model takes as input textual descriptions of a scene and generates the scene graphically object by object using a Recurrent Neural Network, highlighting their ability to learn complex and seemingly non-sequential tasks. The more advanced version of our model requires more computing but can also produce real images by stitching segments from other images. Read more about Text2Scene in the in the research blogs of IBM and NVIDIA and download the full source code from https://github.com/uvadatasets/Text2Scene. This demo generates cartoon-like images using the vocabulary and graphics from the Abstract Scenes dataset proposed by Zitnick and Parikh in 2013.

Besides Mike and Jenny feel free to reference any of these other objects: bear, cat, dog, duck, owl, snake, hat, crown, pirate hat, Viking hat, witch hat, glasses, pie, pizza, hot dog, ketchup, mustard, drink, bee, slide, sandbox, swing, tree, pine tree, apple tree, helicopter, balloon, sun, cloud, rocket, airplane, ball, football, basketball, baseball bat, shovel, tennis racket, kite, fire. Also feel free to describe Mike and Jenny with other attributes or action words such as sitting, running, jumping, kicking, standing, afraid, happy, scared, angry, etc.

#1 Mike is next to a tree

#2 Jenny is happy and kicks the ball

#3 There is a fire

Generate Scene

Demo by Leticia and Vicente
Questions?