Deep Learning for Vision and Language

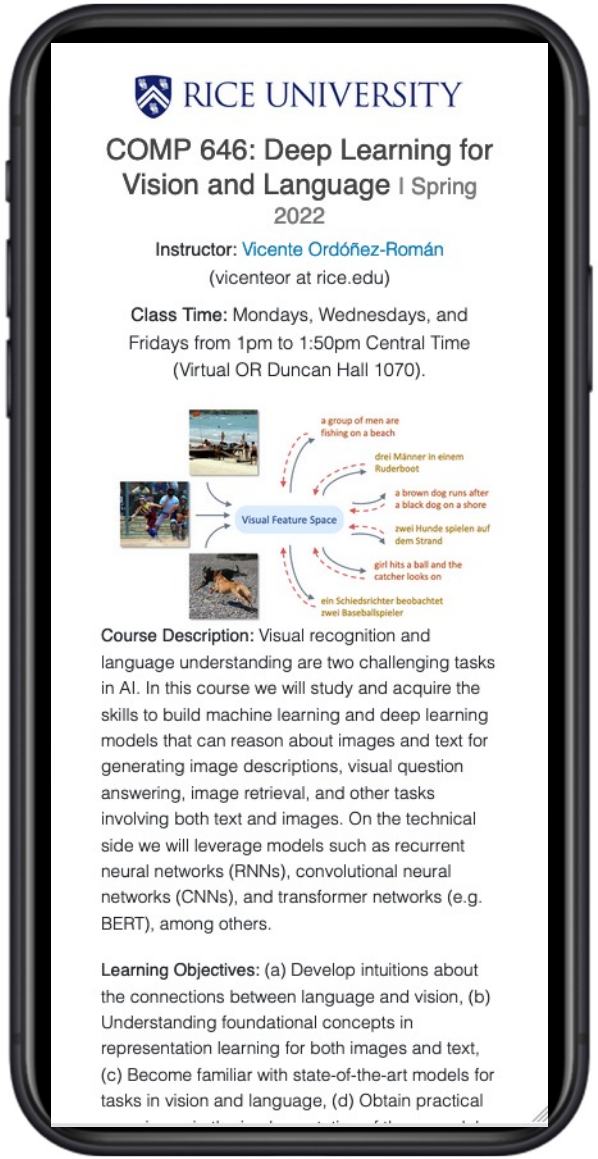
Welcome and Introduction

RICE UNIVERSITY

About the class

- COMP 646: Deep Learning for Vision and Language
- Instructor: **Vicente** Ordóñez (Vicente Ordóñez Román)
- Website: https://www.cs.rice.edu/~vo9/deep-vislang
- Location: Herzstein Hall 210
- Times: Tuesdays and Thursdays
  from 4pm to 5:15pm
- Office Hours: Tuesdays 10am to 11am (DH3098)
- Teaching Assistants: TBD
- Discussion Forum: Piazza (Sign-up Link on Rice Canvas and Class Website)

https://www.cs.rice.edu/~vo9/deep-vislang/

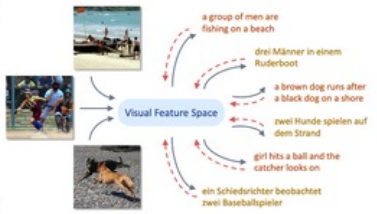


RICE UNIVERSITY

vislang

COMP 646: Deep Learning for Vision and Language | Spring 2023

Instructor: Vicente Ordóñez-Román (vicenteor at rice.edu), Office Hours: 10am to 11am on Thursdays at DH3098.
Class Time: Tuesdays and Thursdays from 4pm to 5:15pm Central Time (Herzstein Hall 210).
Piazza: link



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 involving both text and images. On the technical side we will leverage models such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer networks (e.g. BERT, GPT-3, ViTs), among others.

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

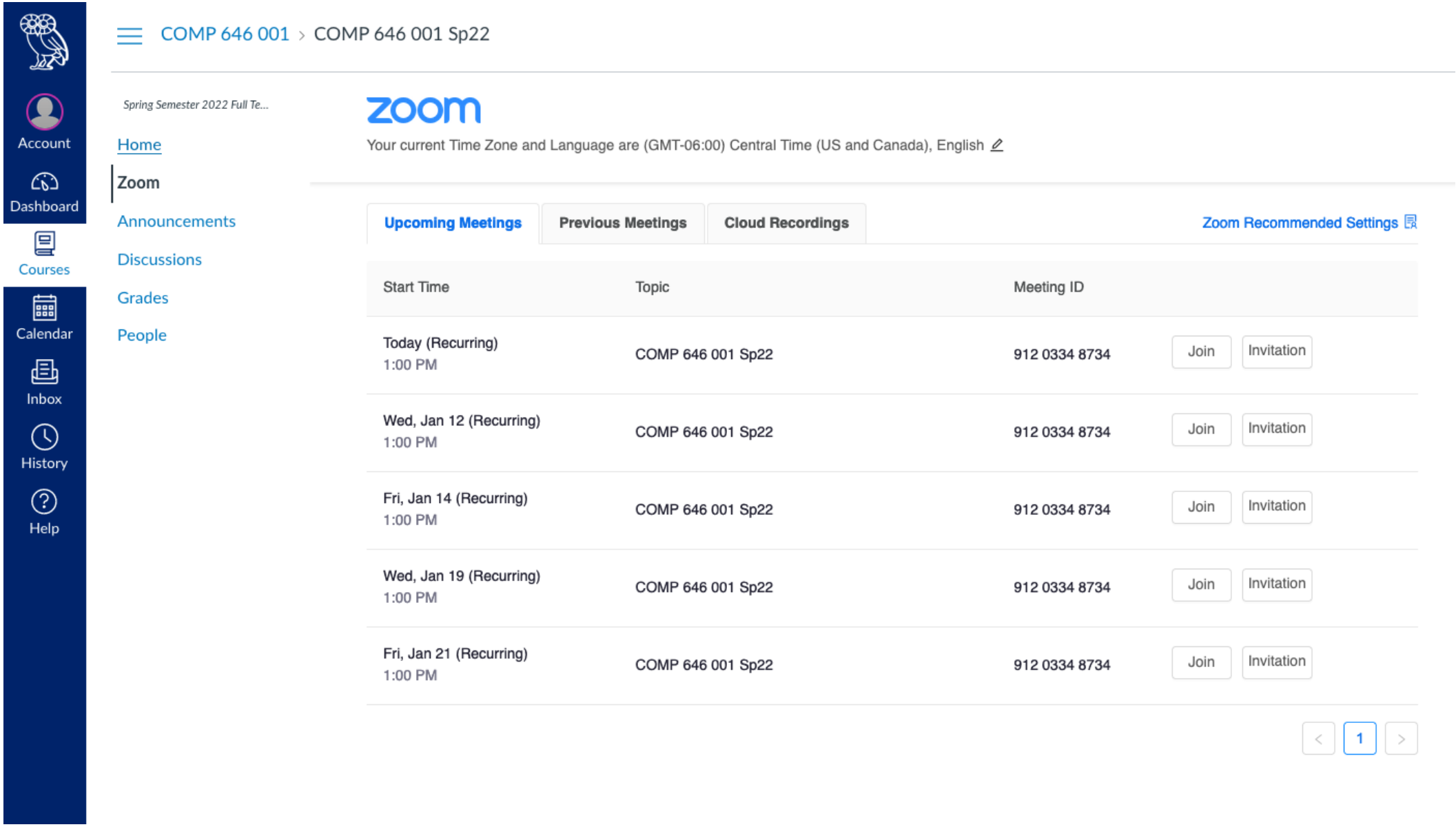
Prerrequisites: There are no formal pre-requisities for this class. However a basic command of machine learning, deep learning or computer vision will be useful when taking this class. Students should have knowledge of linear algebra, differential calculus, and basic statistics and probability. Moreover students are expected to have attained some level of proficiency in Python programming or be willing to learn Python programming. Students are encouraged to complete the following activity before the first lecture: [Primer on Image Processing].

Grading: Assignments: 30% (3 assignments), Class Project: 60%, Quiz: 10%

Schedule

Date	Topic
Tue, Jan 10	Introduction to Vision and Language
Thu, Jan 12	Machine Learning I: Supervised vs Unsupervised Learning, Linear Classifiers
Tue, Jan 17	Machine Learning II: Stochastic Gradient Descent / Regularization
Thu, Jan 19	Neural Networks: Multi-layer Perceptrons and Backpropagation
Tue, Jan 24	Computer Vision I: The Convolutional Operator and Image Filtering
Thu, Jan 26	Computer Vision II: Convolutional Neural Networks
Tue, Jan 31	Computer Vision III: Convolutional Neural Network Architectures: LeNet, AlexNet, GoogleNet, ResNets.
Thu, Feb 2	Computer Vision IV: Convolutional Neural Networks for Object Detection and Segmentation
Tue, Feb 7	Natural Language Processsing I: Introduction: Bag of Words, N-gram Language Models
Thu, Feb 9	Spring recess (No Scheduled Classes)

Zoom Links


Account
Dashboard
Courses
Calendar
Inbox
History
Help
COMP 646 001 › COMP 646 001 Sp22
Spring Semester 2022 Full Te...
Home
Zoom
Announcements
Discussions
Grades
People
zoom
Your current Time Zone and Language are (GMT-06:00) Central Time (US and Canada), English
Upcoming Meetings
Previous Meetings
Cloud Recordings
Zoom Recommended Settings
Start Time
Topic
Meeting ID
Today (Recurring) 1:00 PM
COMP 646 001 Sp22
912 0334 8734
Join
Invitation
Wed, Jan 12 (Recurring) 1:00 PM
COMP 646 001 Sp22
912 0334 8734
Join
Invitation
Fri, Jan 14 (Recurring) 1:00 PM
COMP 646 001 Sp22
912 0334 8734
Join
Invitation
Wed, Jan 19 (Recurring) 1:00 PM
COMP 646 001 Sp22
912 0334 8734
Join
Invitation
Fri, Jan 21 (Recurring) 1:00 PM
COMP 646 001 Sp22
912 0334 8734
Join
Invitation
1

About me -- Vicente

Associate Professor, 2021 - Present	RICE UNIVERSITY
Visiting Academic 2021 - Present	amazon alexa
Assistant Professor, 2016 - 2021	UNIVERSITY of VIRGINIA
Visiting Professor, 2019	Adobe Research
Visiting Researcher, 2015 - 2016	AI2 ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE
MS, PhD in CS, 2009-2015	THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL Stony Brook University

... 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.

vision, language and learning

vislang | RICE UNIVERSITY

home people demos publications

The vision, language and learning lab, *vislang*, at Rice University pursues fundamental research at the intersection of computer vision, natural language processing and machine learning. We aim to create intelligent systems that can learn from vast amounts of visual and textual information, that can integrate and enhance human experiences, and that can resolve complex tasks that typically require human intelligence.

Read about some of our work on bias in visual recognition in WIRED and Glamour. Some of our recent work on analyzing movies on TechXplore, and our work on generating images from text in the blogs of IBM and NVIDIA.

News and Announcements

- 07/2021. Two papers accepted to ICCV 2021, Reranking Transformers [arxiv] and MEDIRL [arxiv].
- 07/2021. After some wonderful five years at the University of Virginia, our group is in the process of moving to the Department of Computer Science at Rice University in Houston, Texas~!
- 06/2021. Our work on teaching machines compositional vision and language models is funded through a National Science Foundation CAREER Award [link]
- 06/2021. Tianlu Wang defends her PhD Dissertation *Measuring and Mitigating Biases in Vision and Language Models*, accepts position as

Visual Translator

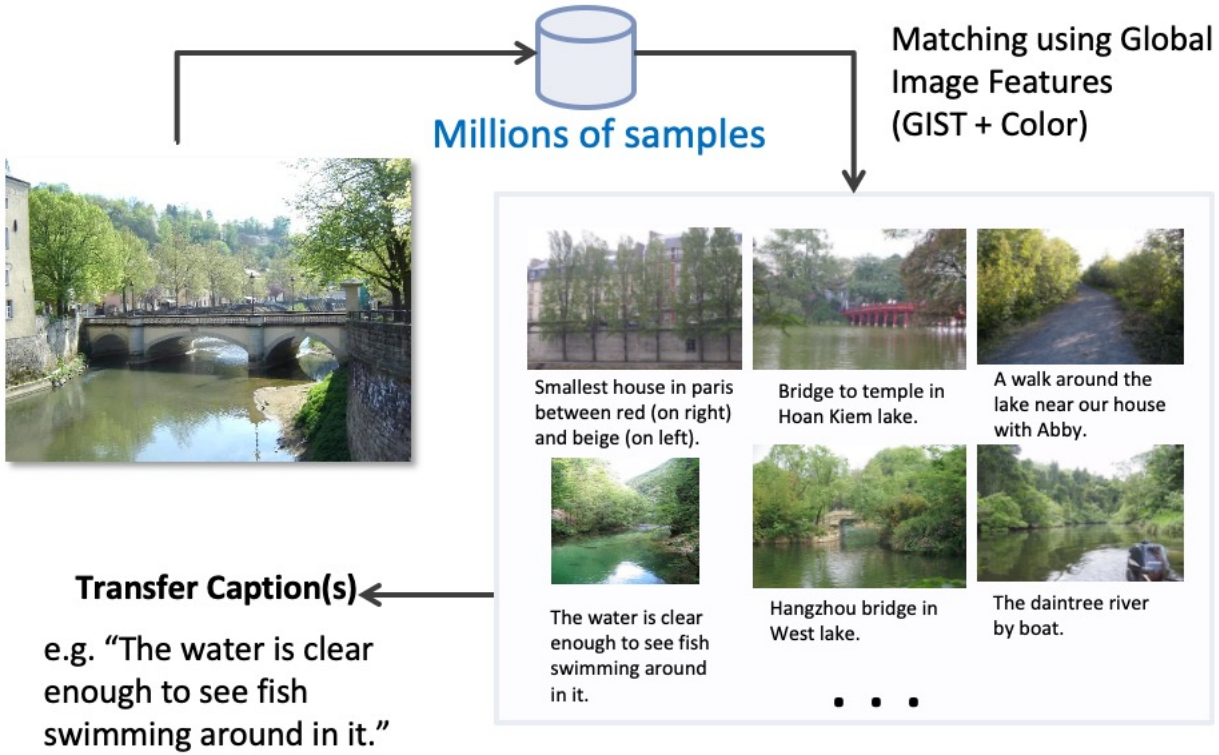
This demo attemps to translate a sentence in English into visual feature space and into a sentence in both German (Deutsch) and Japanese (日本語).

Facial Recognition Technologies in the Wild

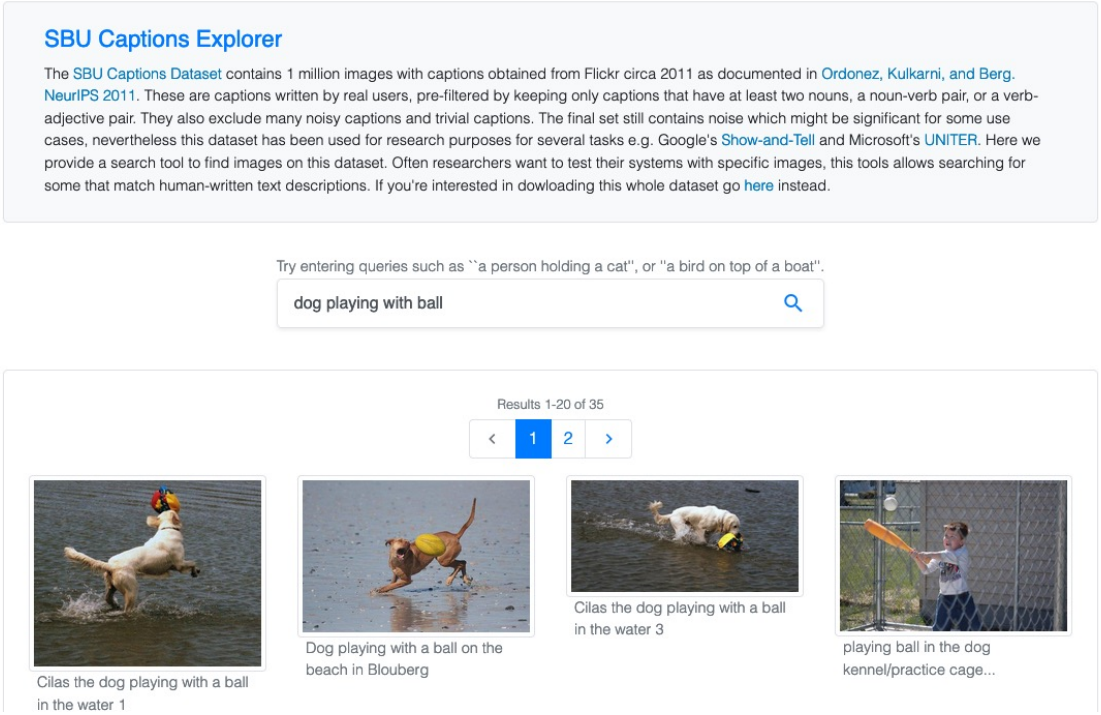
With colleagues Erik-Learned Miller, Jamie

Some of our work includes…

Describing images with language



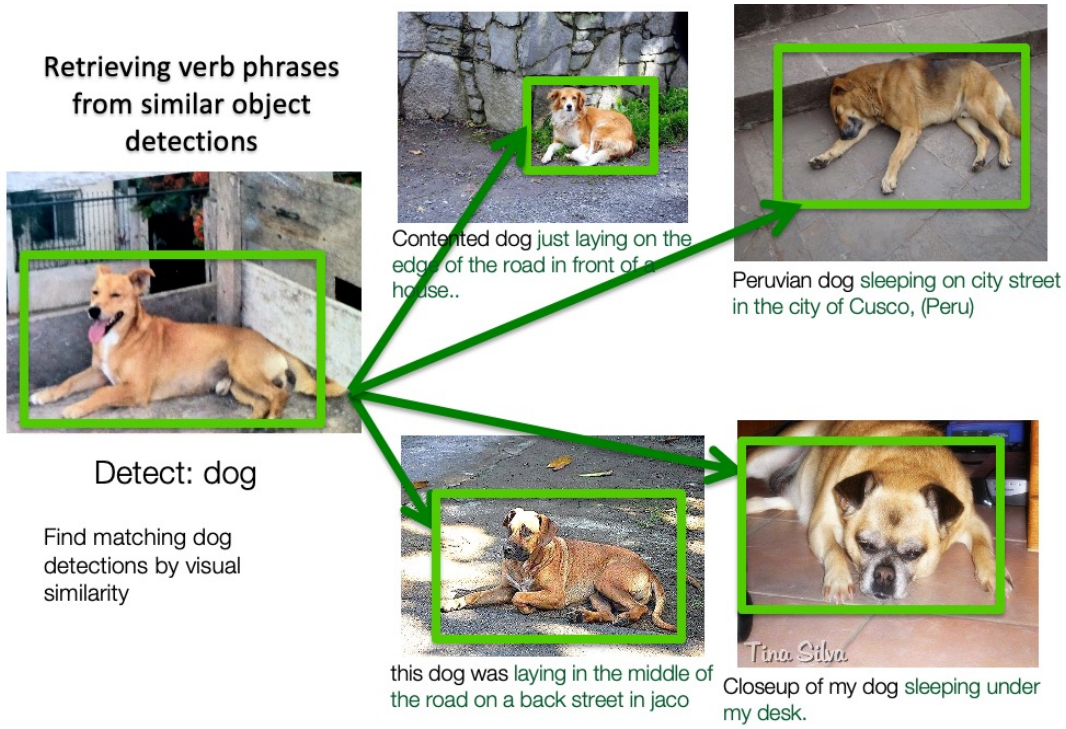
https://vislang.ai/sbu-explorer



Im2Text: Describing Images Using 1 Million Captioned Photographs
Vicente Ordonez, Girish Kulkarni, Tamara L. Berg.
Advances in Neural Information Processing Systems. **NIPS 2011**. Granada, Spain. December 2011.

Some of our work includes…

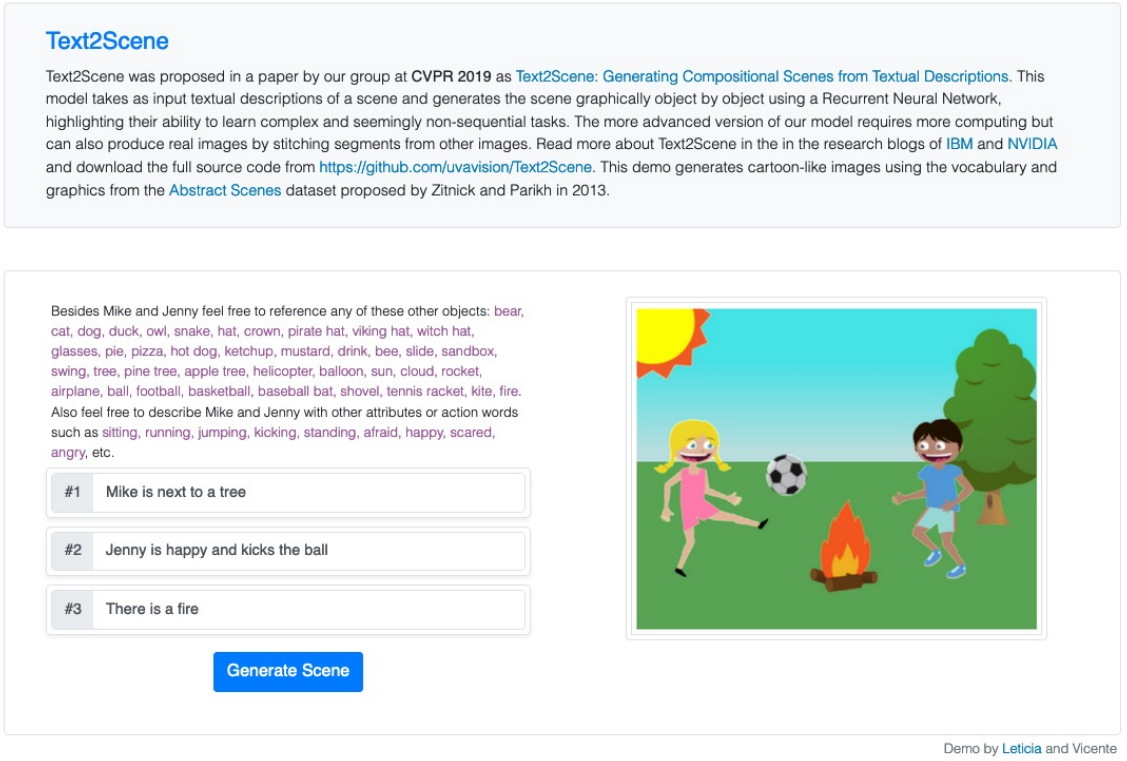
Describing images with language



Large Scale Retrieval and Generation of Image Descriptions
V. Ordonez, X. Han, P. Kuznetsova, G. Kulkarni, M. Mitchell, K. Yamaguchi, K. Stratos, A. Goyal, J. Dodge, A. Mensch, H. Daume III, A.C. Berg, Y. Choi, T.L. Berg.
International Journal of Computer Vision. **IJCV 2015**. [August 2016 Issue]. [pdf] [link] [bibtex]

Describing language with images

https://vislang.ai/text2scene

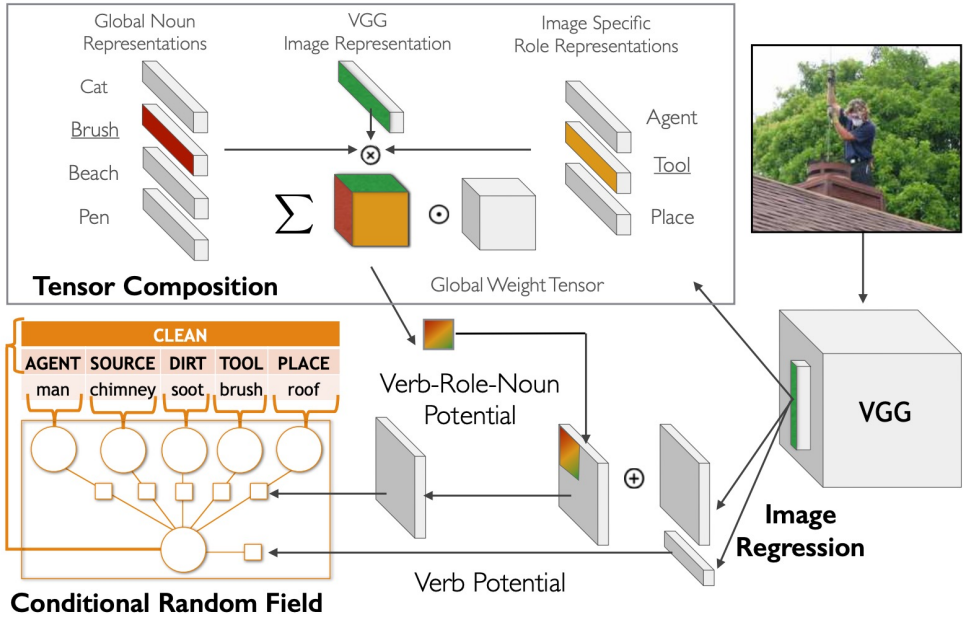


Some of our work includes…

Situation Recognition

CARRYING

ROLE	VALUE	ROLE	VALUE	ROLE	VALUE
AGENT	MAN	AGENT	WOMAN	AGENT	MAN
ITEM	BABY	ITEM	BUCKET	ITEM	TABLE
AGENTPART	CHEST	AGENTPART	HEAD	AGENTPART	BACK
PLACE	OUTSIDE	PLACE	PATH	PLACE	STREET

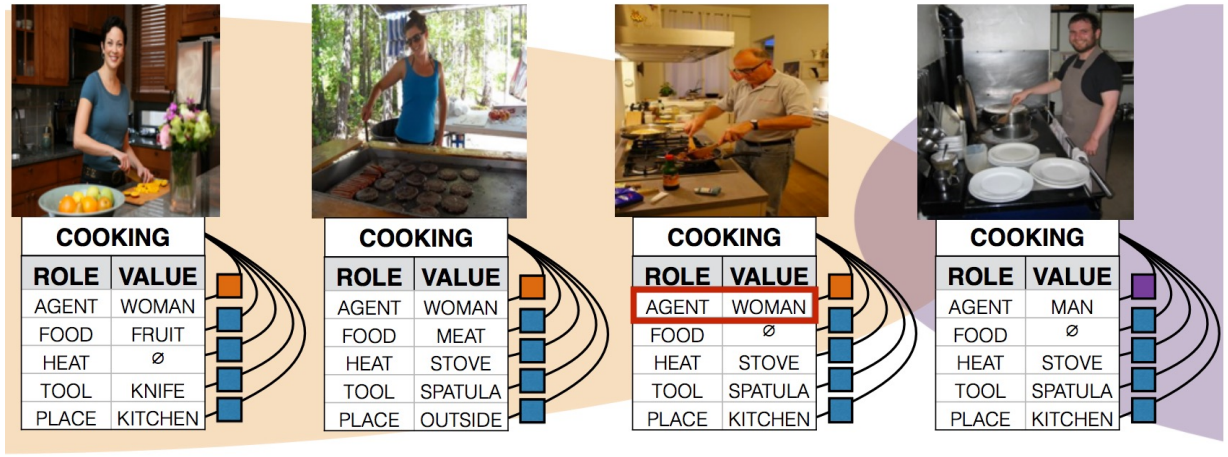


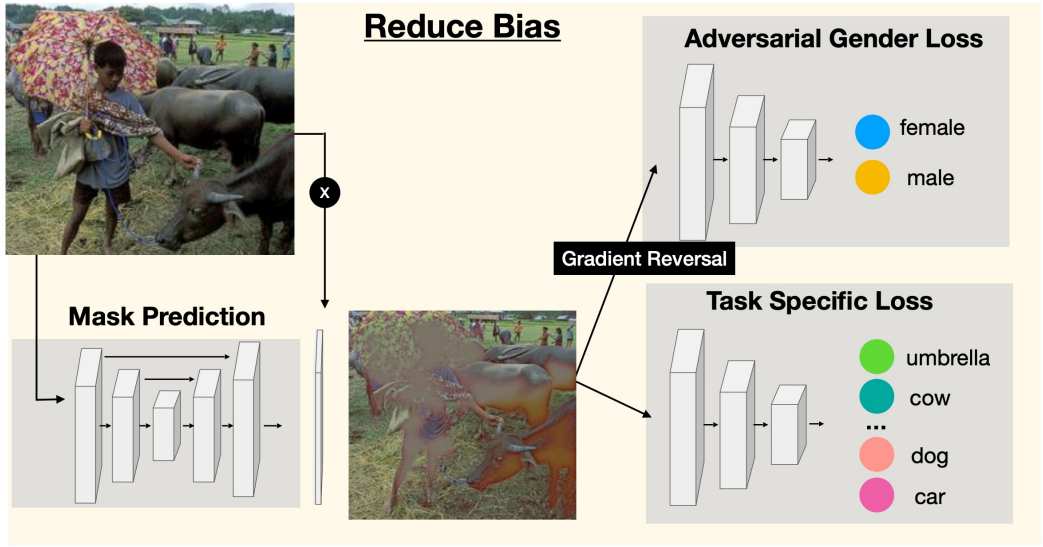
http://imsitu.org/

Commonly Uncommon: Semantic Sparsity in Situation Recognition
Mark Yatskar, Vicente Ordonez, Luke Zettlemoyer, Ali Farhadi.
Intl. Conference on Computer Vision and Pattern Recognition. **CVPR 2017**. Honolulu, Hawaii. July 2017. [pdf] [arXiv] [bibtex] [demo]

Some of our work includes…

Learning from Images with Textual Descriptions





https://www.vislang.ai/genderless

Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, Vicente Ordonez. International Conference on Computer Vision. **ICCV 2019**. Seoul, South Korea. October 2019. [arxiv] [code] [demo] [bibtex]

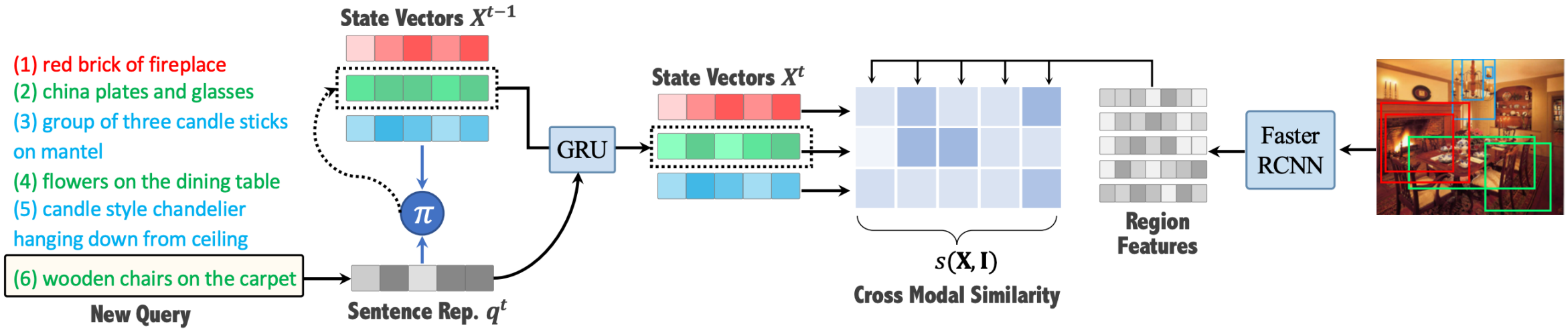
Some of our work includes…

Interactive Image Retrieval

Drill-down: Interactive Retrieval of Complex Scenes using Natural Language Queries
Fuwen Tan, Paola Cascante-Bonilla, Xiaoxiao Guo, Hui Wu, Song Feng, Vicente Ordonez. Conf. on Neural Information Processing Systems. **NeurIPS 2019**. Vancouver, Canada. December 2019. [arxiv] [code] [bibtex]

Some of our work includes…

Interactive Image Retrieval



Drill-down: Interactive Retrieval of Complex Scenes using Natural Language Queries
Fuwen Tan, Paola Cascante-Bonilla, Xiaoxiao Guo, Hui Wu, Song Feng, Vicente Ordonez. Conf. on Neural Information Processing Systems. **NeurIPS 2019**. Vancouver, Canada. December 2019. [arxiv] [code] [bibtex]

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

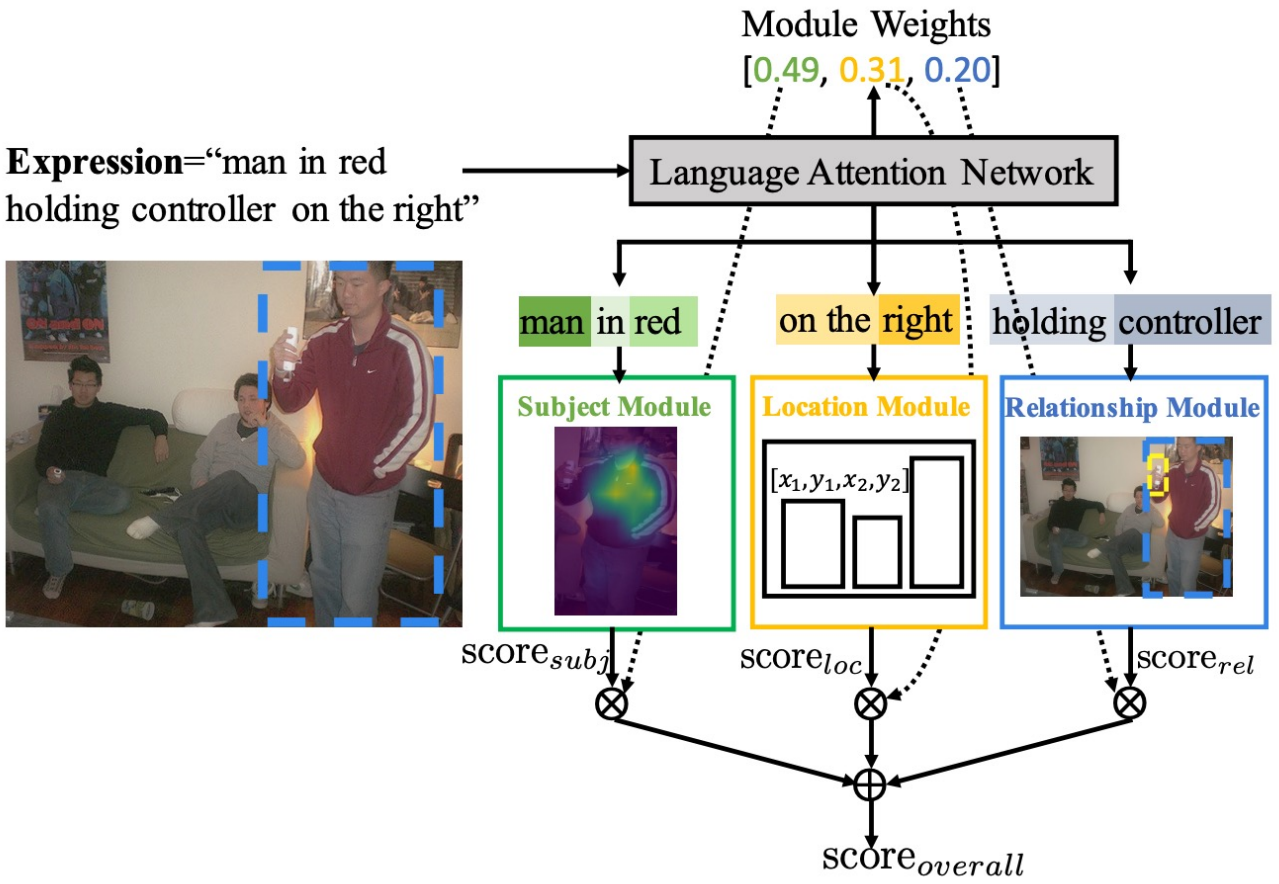
Drill-down: Image Retrieval System

The woman is wearing a pink coat

Drill-down: Image Retrieval System

they both have goggles

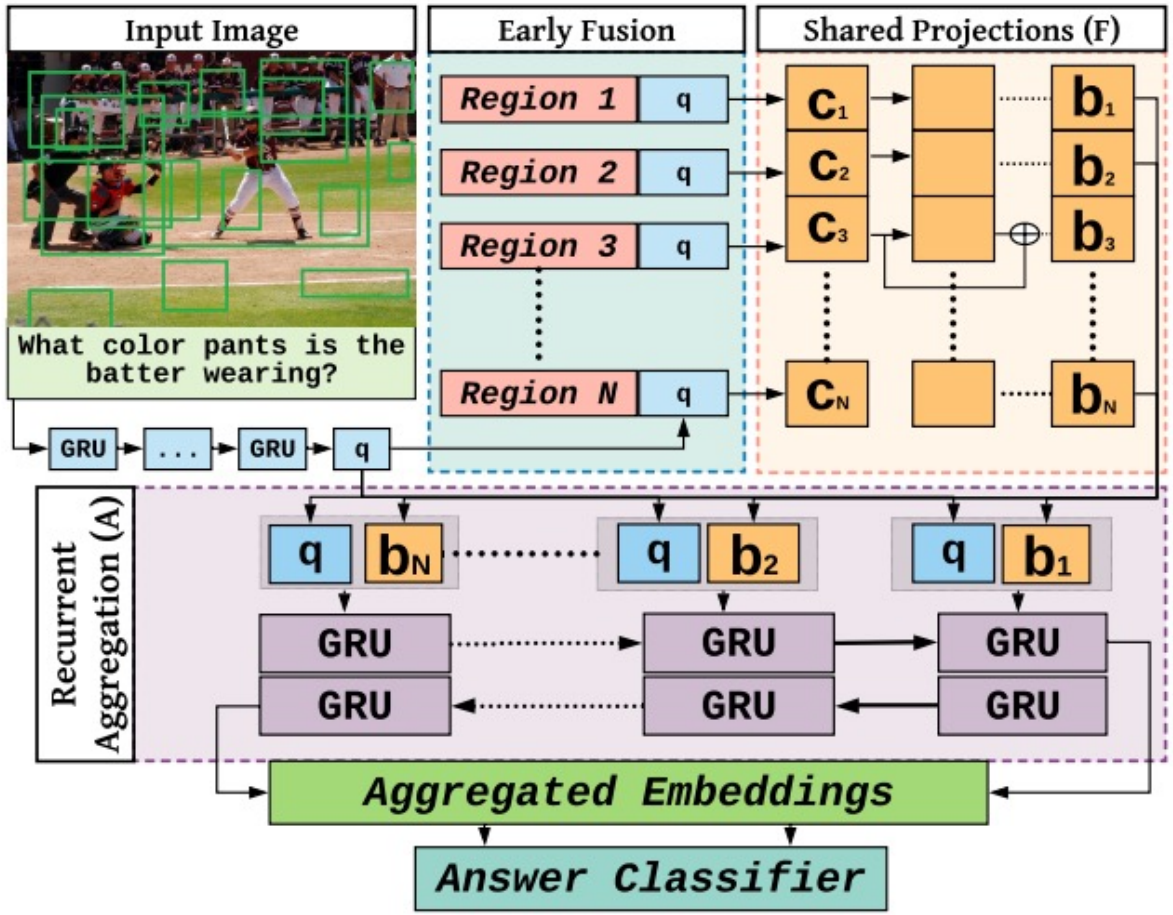
Referring Expression Comprehension



MAttNet: Modular Attention Network for Referring Expression Comprehension

Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, Tamara L.Berg

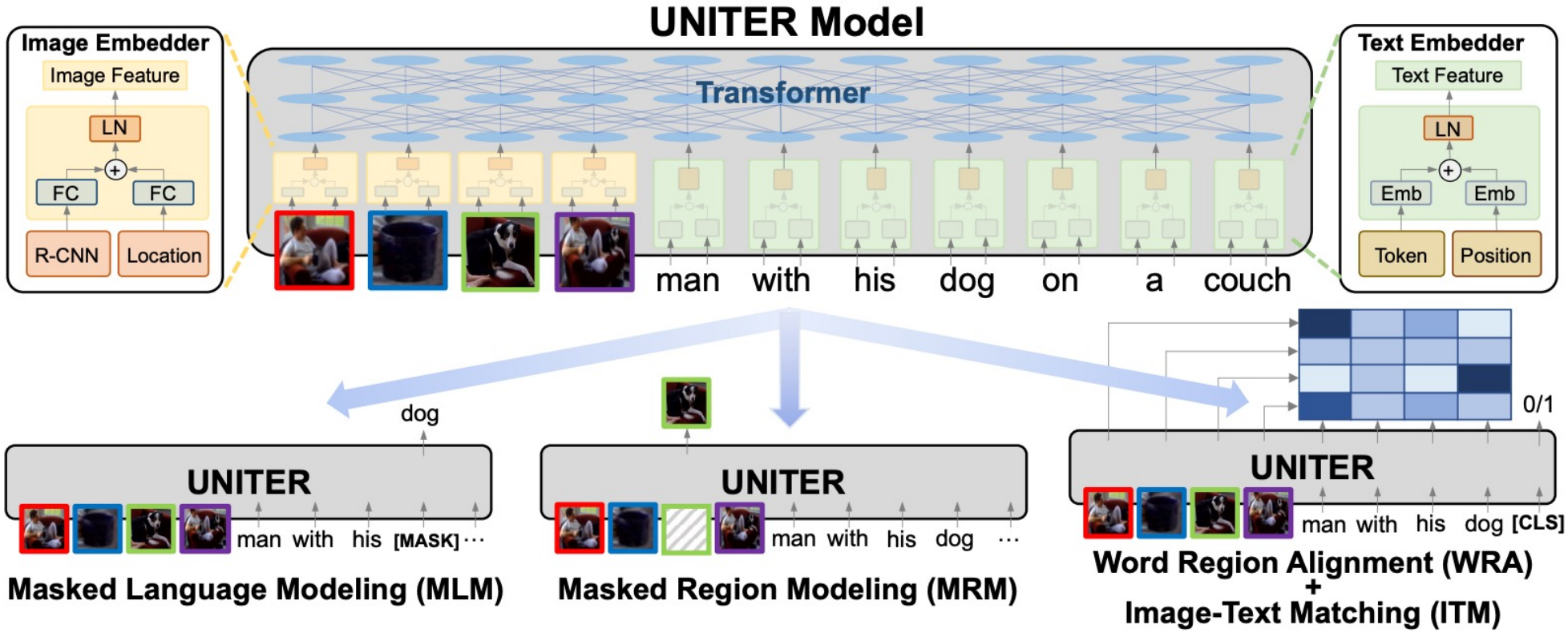
Visual Question Answering



Answer Them All! Toward Universal Visual Question Answering Models

Robik Shrestha, Kushal Kafle, Christopher Kanan

Vision-and-Language Transformers

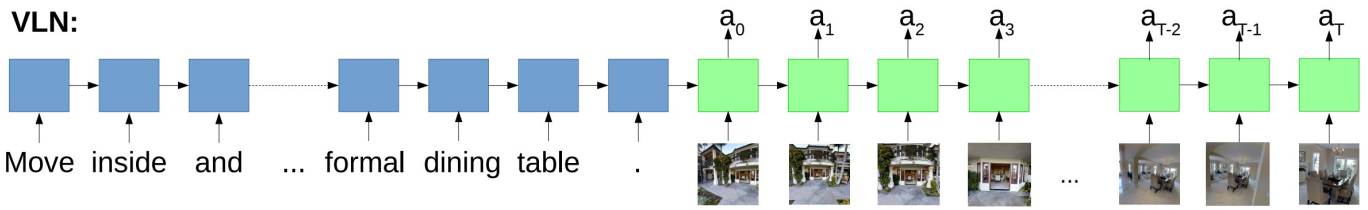


UNITER: UNiversal Image-TExt Representation Learning

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, Jingjing Liu

Vision-and-Language for Navigation

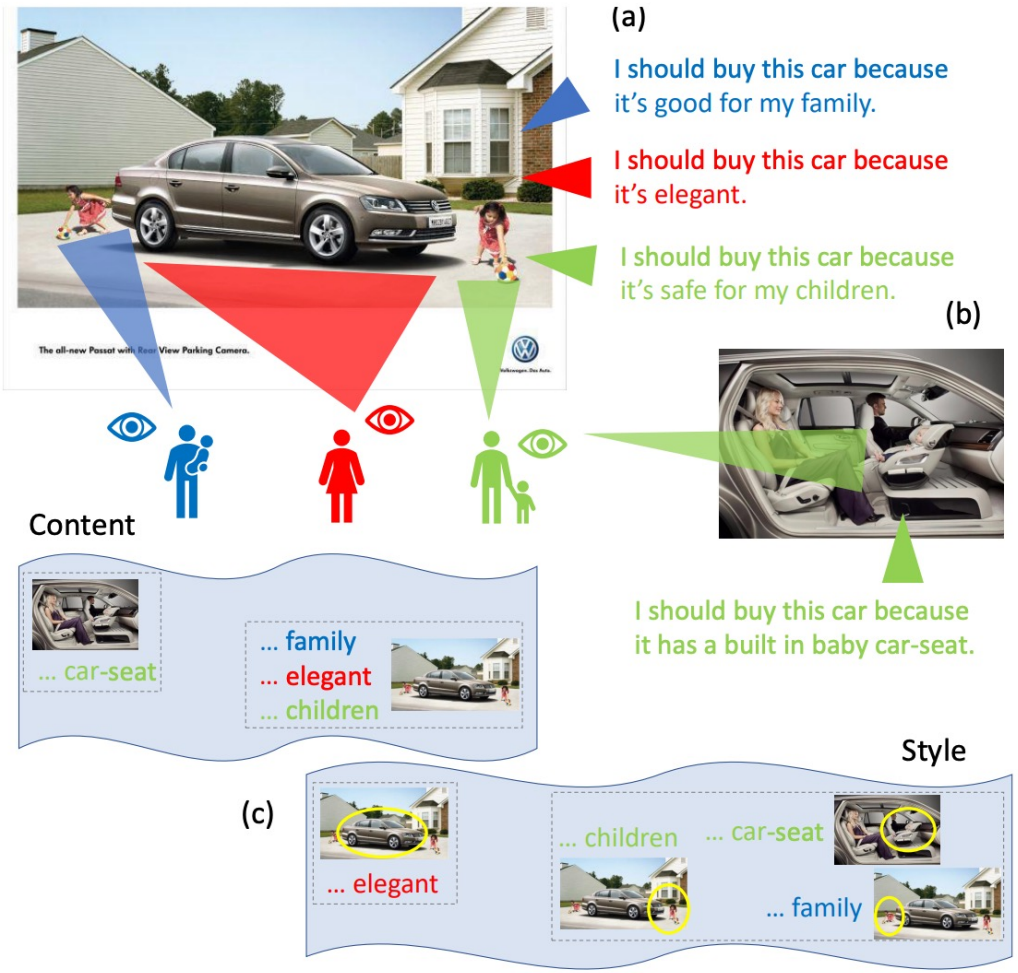
Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.



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

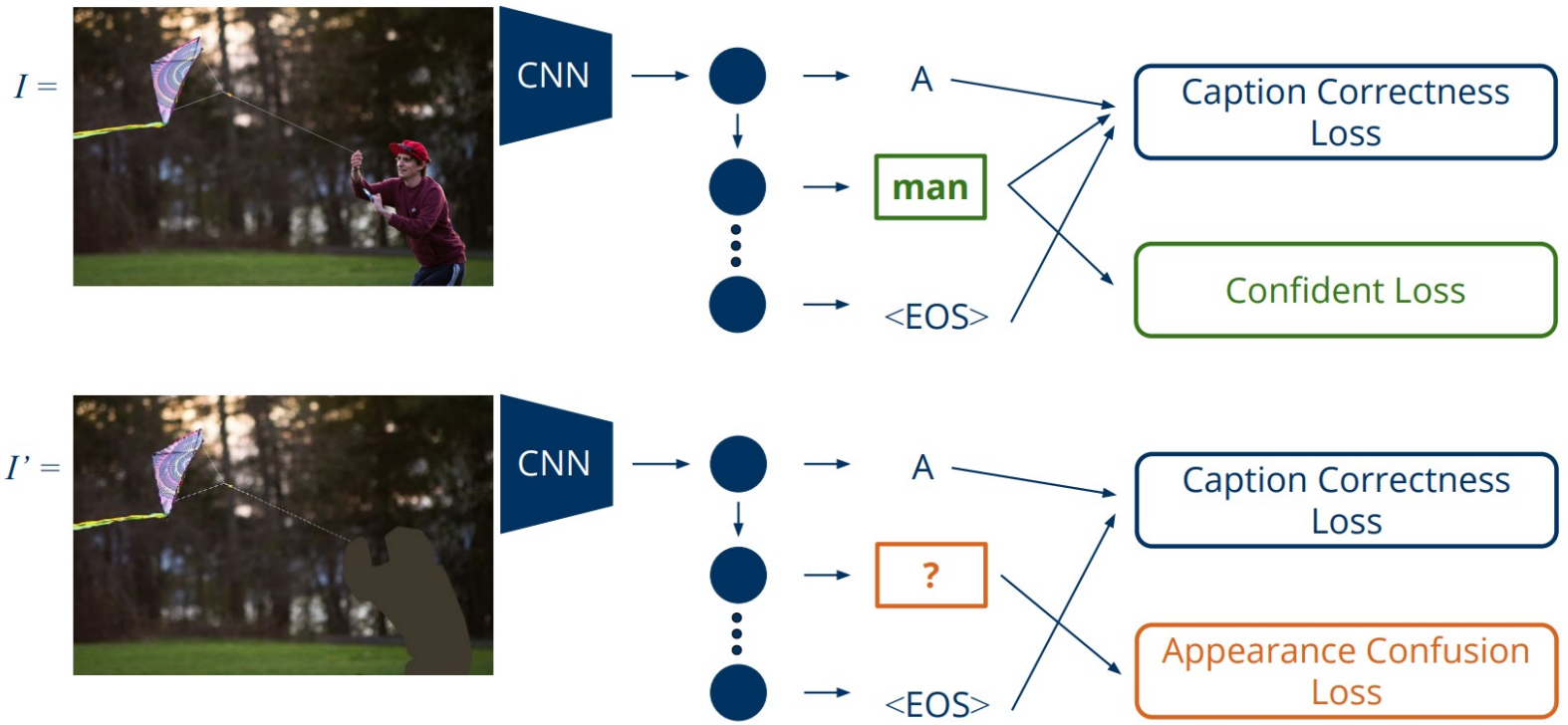
Personalized Image Retrieval



Cross-Modality Personalization for Retrieval

Nils Murrugarra-Llerena Adriana Kovashka
Department of Computer Science
University of Pittsburgh
{nineil, kovashka}@cs.pitt.edu

Fairness in Vision and Language Models



Women also Snowboard: Overcoming Bias in Captioning Models

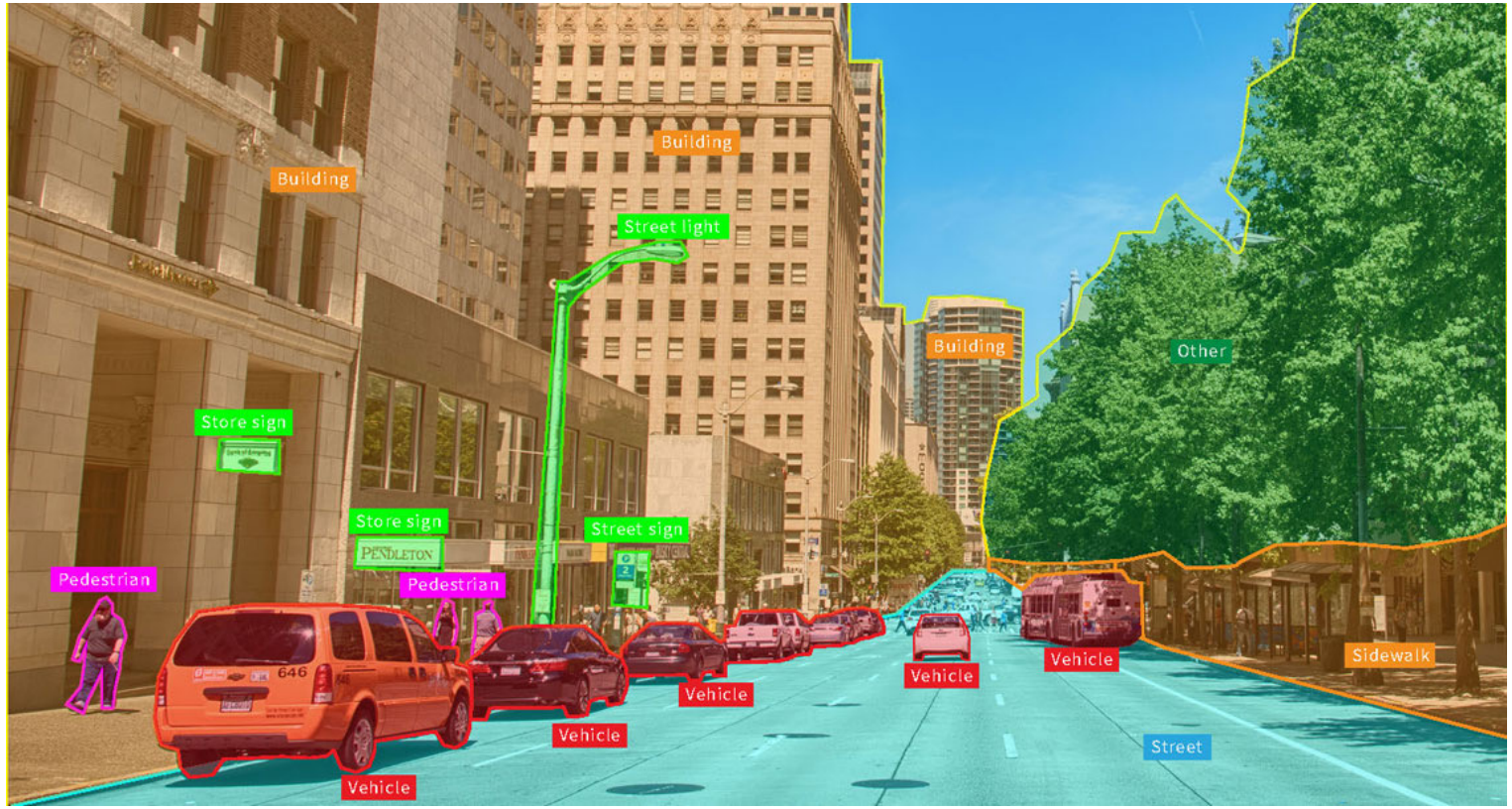
Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, Anna Rohrbach

Why Vision and Language Together?

- What makes us intelligent?

Why Vision and Language Together?

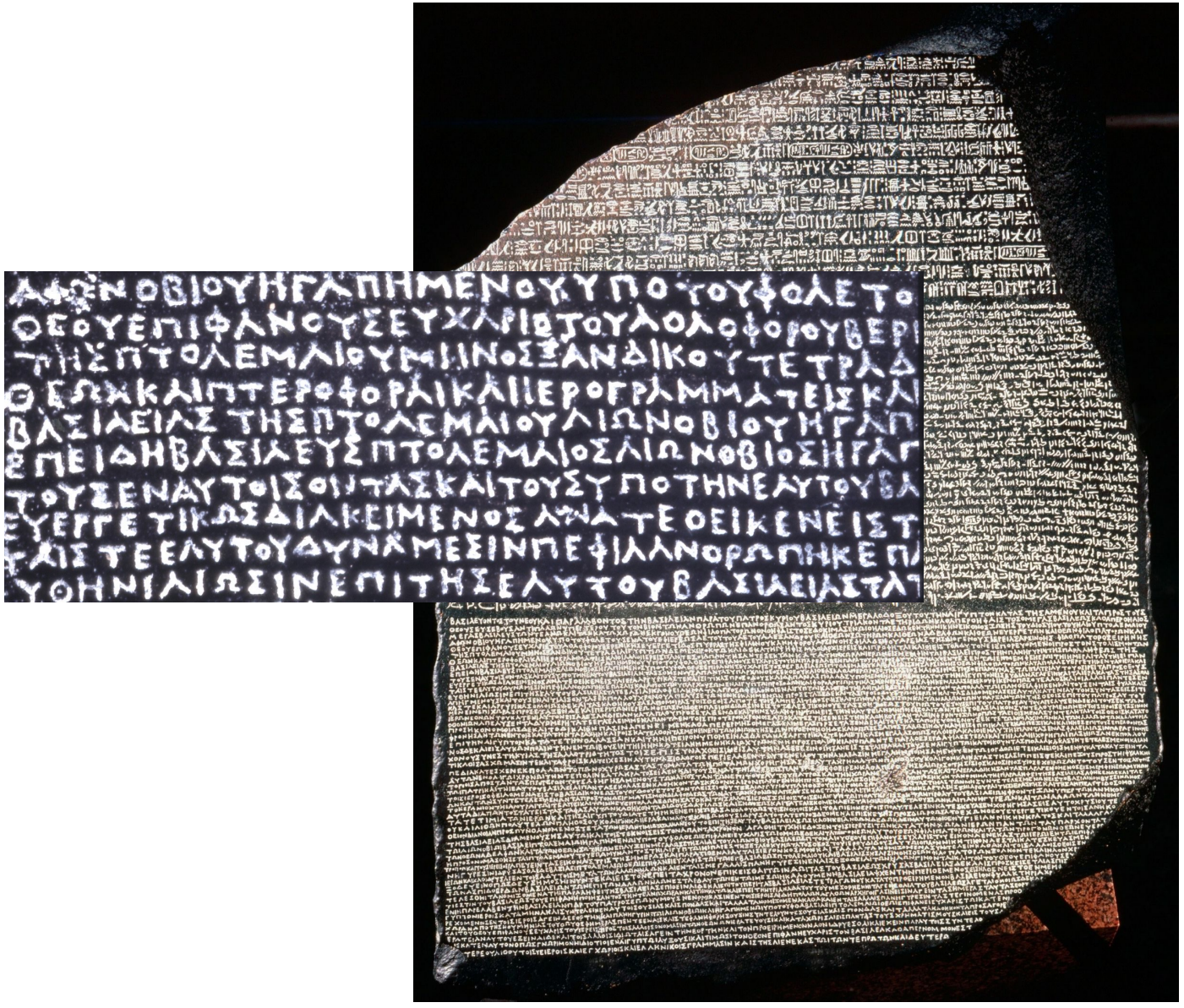
- What makes us intelligent?
- Vision is not just sensing – but interpreting what our eyes capture



https://appen.com/blog/computer-vision-vs-machine-vision/

Why Vision and Language Together?

- What makes us intelligent?
- Vision is not just sensing – but interpreting what our eyes capture
- Language is not just a sequence of symbols – but interpreting what do they mean – think of a foreign language to you



Can we learn language through pictures?

https://www.hameraypublishing.com/blogs/all/teaching-kids-about-the-structure-of-the-spanish-language

Vision and Language in Practice

- Searching products using language can be hard – e.g. I want to find a "rustic vintage curio with dark cherry finishes"

Rustic

Vintage

Hutch

Curio

Vision and Language in Practice

- Robotics: Instruction Following

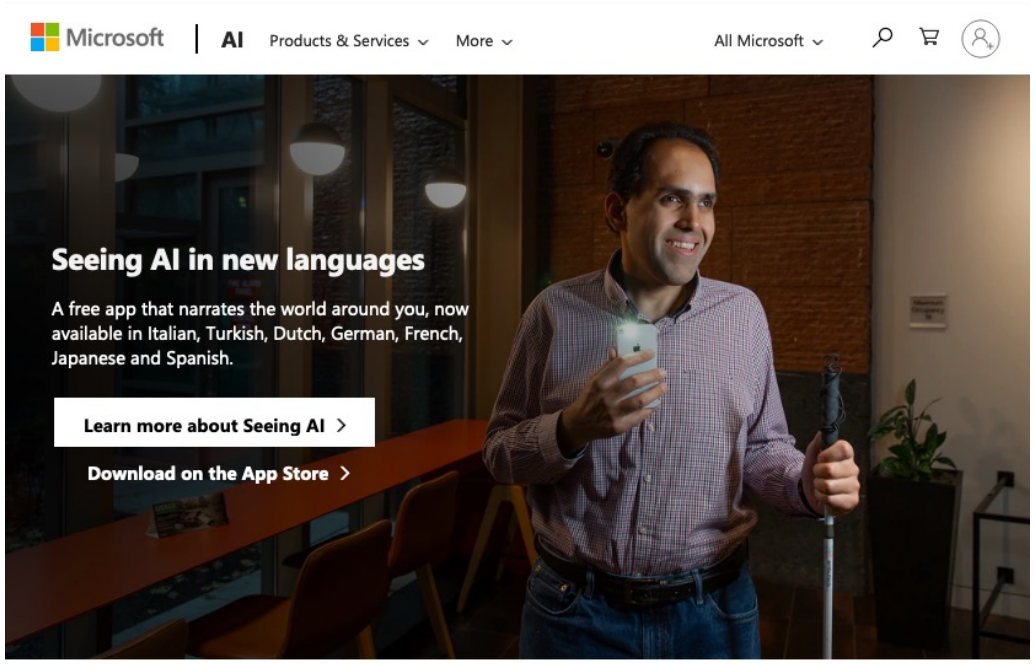
Amazon launches home robot Astro and giant Alexa display

Robot that can check on loved ones and pets is one of plethora of devices announced at big launch event

Astro is Amazon's first attempt at a home robot designed to be a roving smart platform for Alexa, video calling and many other services. Photograph: Amazon

Vision and Language in Practice

- Assistive Technologies



Complete multiple tasks with one app

Switch between channels to tune the description of what's in front of the camera.

What will we cover in this class?

In terms of tools

3 weeks	• Introduction to ML / Vision / NLP • Neural Networks (NNs) / Deep Learning.
2 weeks	• Convolutional Neural Networks (CNNs)
2 weeks	• Recurrent Neural Networks (RNNs, LSTMs, GRUs)
8 weeks	• Transformers (e.g. BERT, GPT, UNITER, etc) • State-of-the-art and Recent Developments

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…

https://www.cs.rice.edu/~vo9/deep-vislang/

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. Definitely know how to program at a college graduate level.
- You will benefit from knowing some 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 and concepts are needed. (matrices, vectors, vector spaces, chain rule of calculus, derivatives, gradients, bayes theorem, maximum likelihood estimation, least squares regression)

Grading for this class: COMP 646

- Assignments: 30pts (3 assignments: 10pts + 10pts + 10pts)
- **Class Project: 60pts**
- Quiz: 10pts

Total: 100pts

- Grade cutoffs: TBD

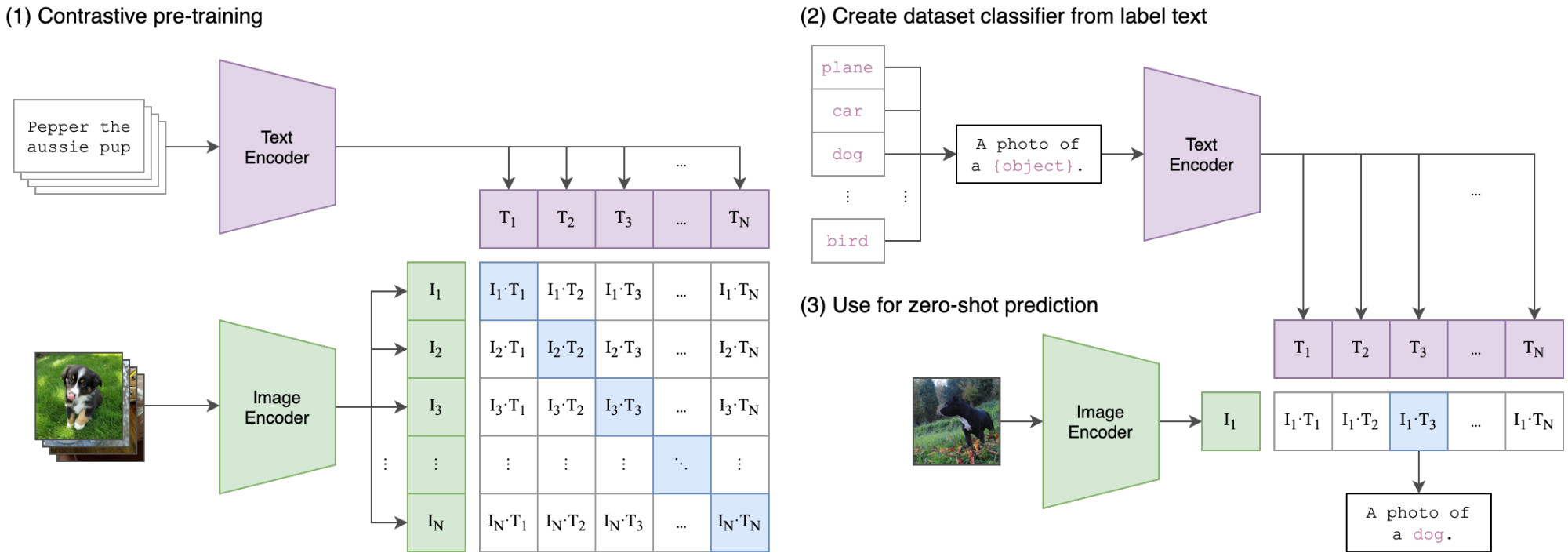
Class Project Timeline

- Class Project: 60pts
  - You can form a group: 3 students maximum per group
  - You can also work solo – 1 student groups.
- In ~3 weeks: Submit as a group a project proposal (1 page PDF)
- In ~5 weeks: Submit as a group a final project proposal (1 page PDF)
- In ~10 weeks: Submit a project progress report (2 page PDF)
- End of semester: Submit the following:
  - Project report PDF (4 pages)
  - Slides + Presentation (Video / Demo)
  - Source code + ideally an online demo (if appropriate)

New Project Requirement:
Take Advantage of One of the Following Recent Open Models for your Project

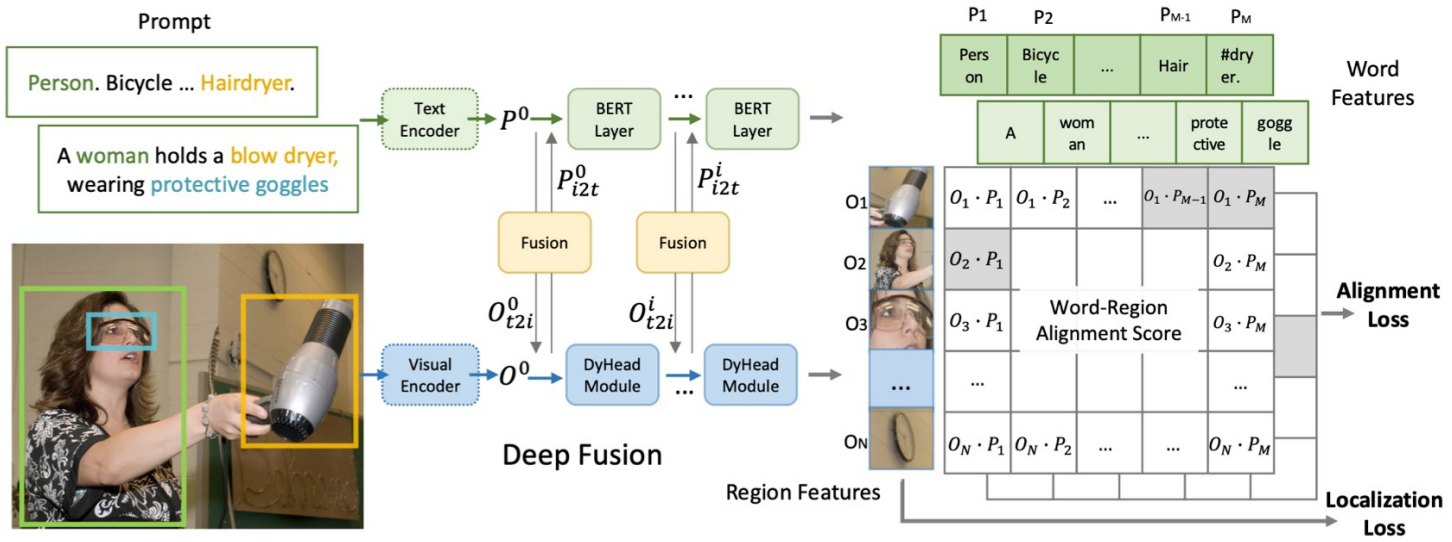
- CLIP by OpenAI (Images + Text)
- FLAN T5 by Google (Text)
- ChatGPT by OpenAI (Text)
- GLIP by Microsoft (Images + Text)
- Whisper by OpenAI (Speech to Text)
- Stable Diffusion v2 by Stability.AI (Text to Images)

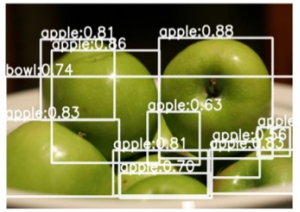
CLIP



https://github.com/openai/CLIP

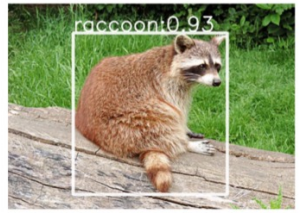
GLIP





Prompt : person. bicycle. car. motorcycle...

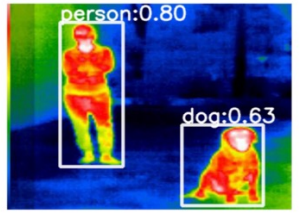
Prompt : aerosol can... lollipop... pendulum...



Prompt : raccoon

Prompt : pistol

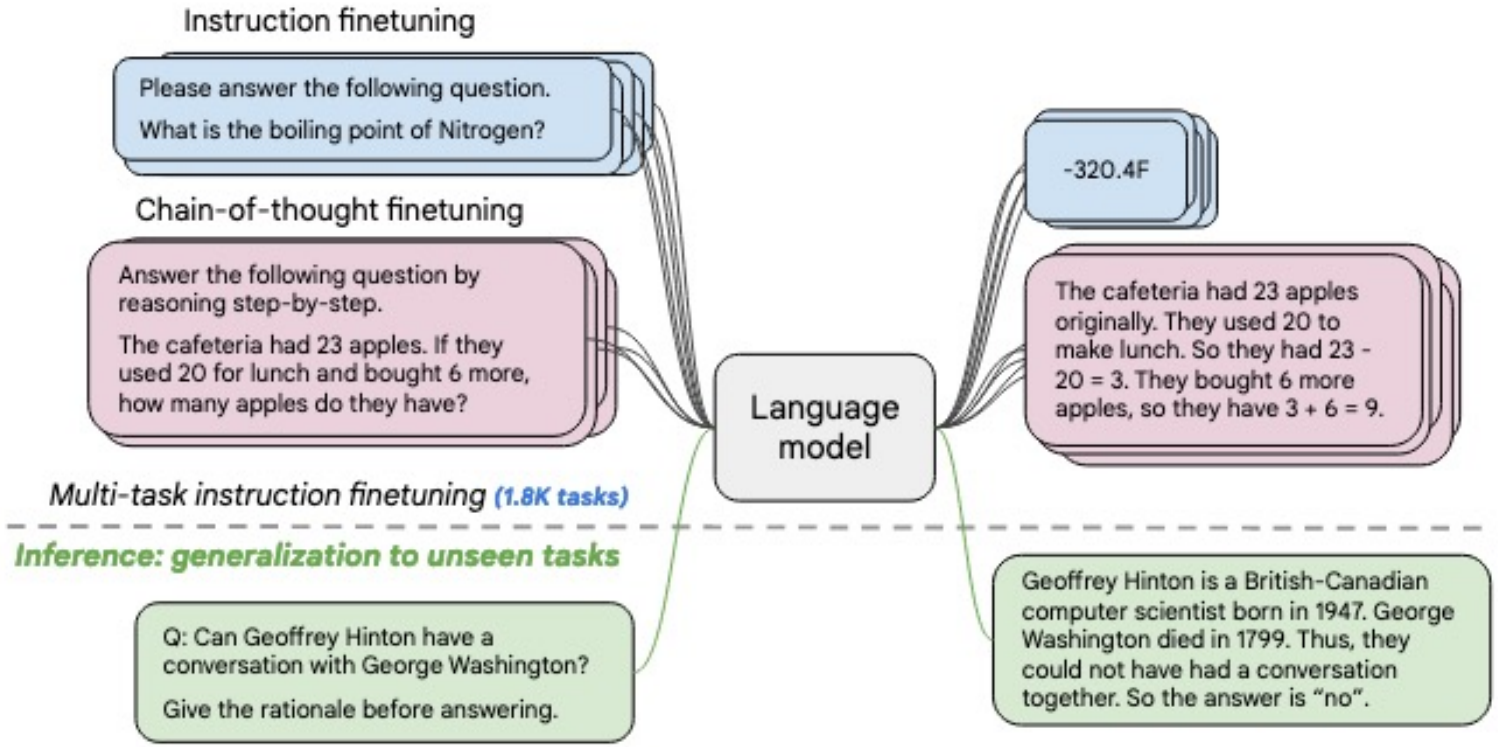
Prompt : there are some holes on the road



Prompt : person. dog.

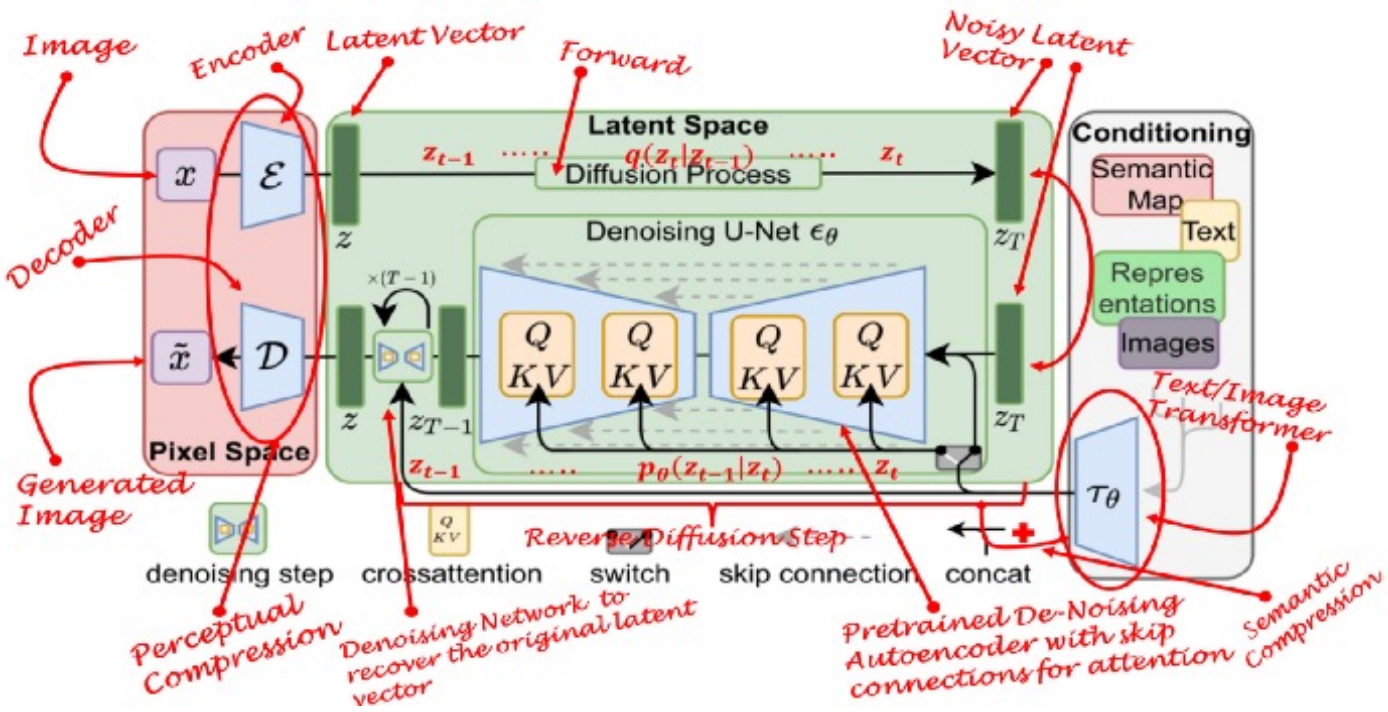
https://github.com/microsoft/GLIP

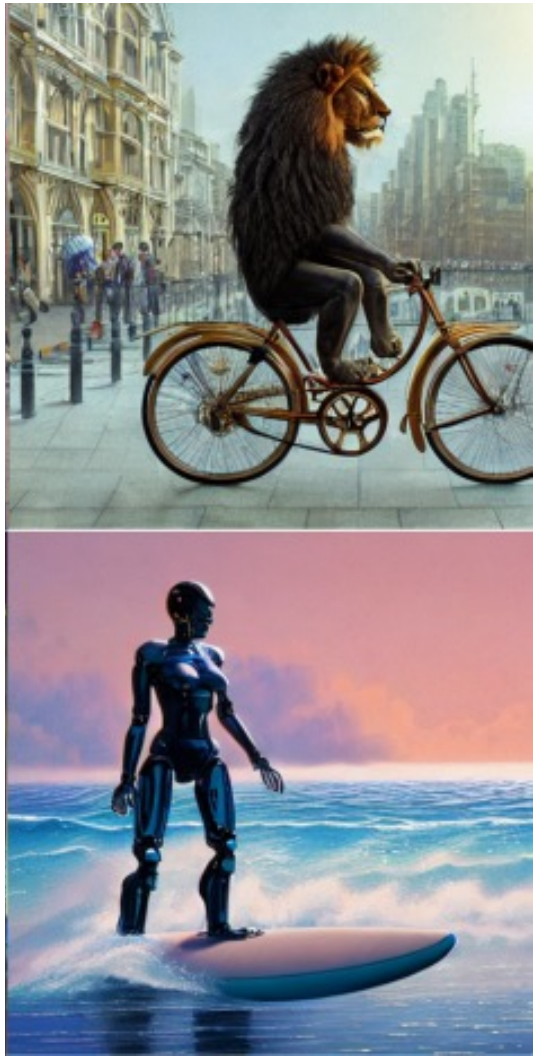
FLAN T5



https://huggingface.co/docs/transformers/model_doc/flan-t5

Stable Diffusion v2

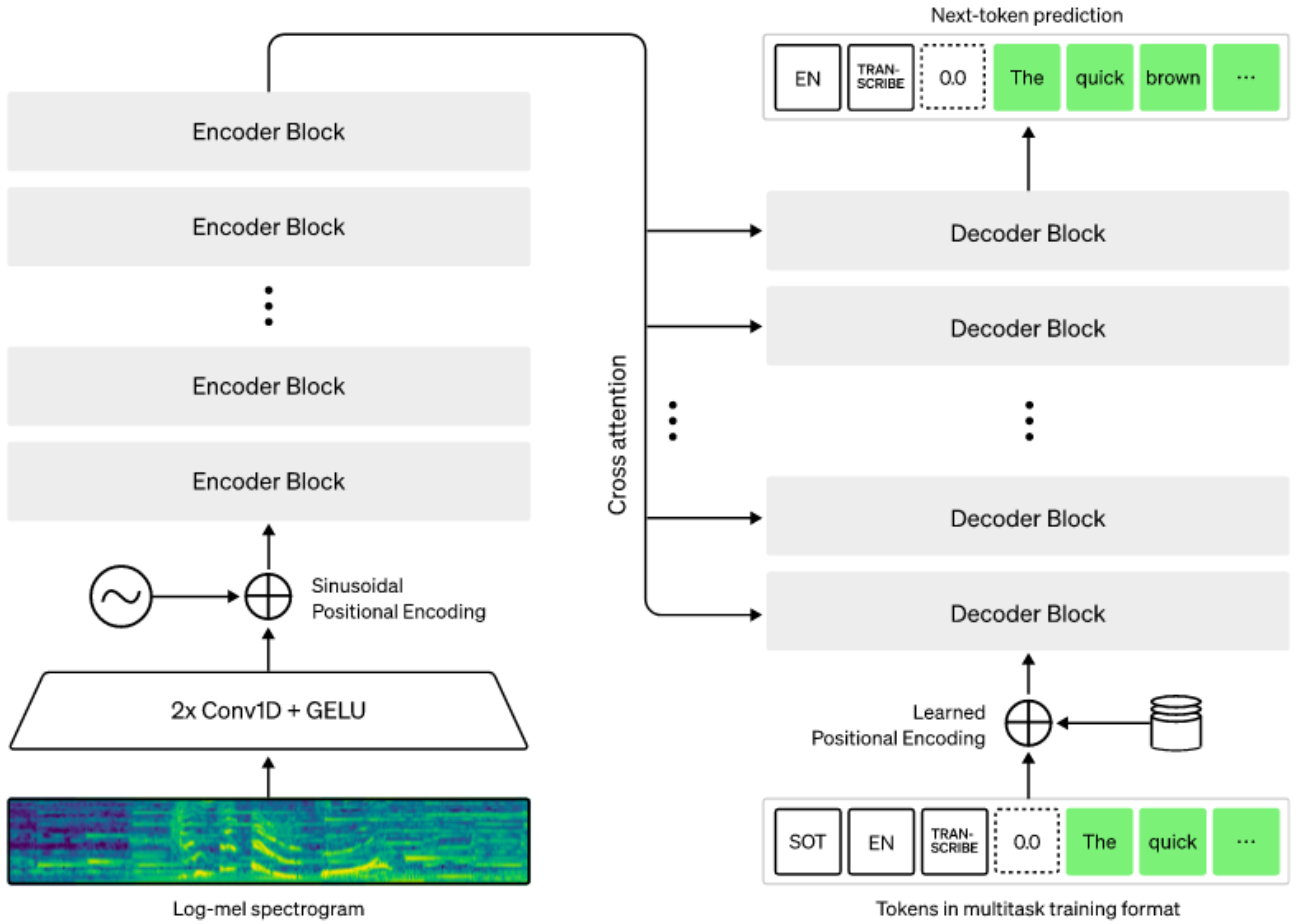




https://towardsdatascience.com/what-are-stable-diffusion-models-and-why-are-they-a-step-forward-for-image-generation-aa1182801d46

https://huggingface.co/stabilityai/stable-diffusion-2

Whisper



https://github.com/openai/whisper

You will benefit if you have / but not required

NVIDIA Ampere A100	$17,000
NVIDIA Tesla v100	$7,000
NVIDIA RTX 3090	$3,000
NVIDIA GTX 1080 Ti	$700

We will be using

https://pytorch.org/

HUGGING FACE

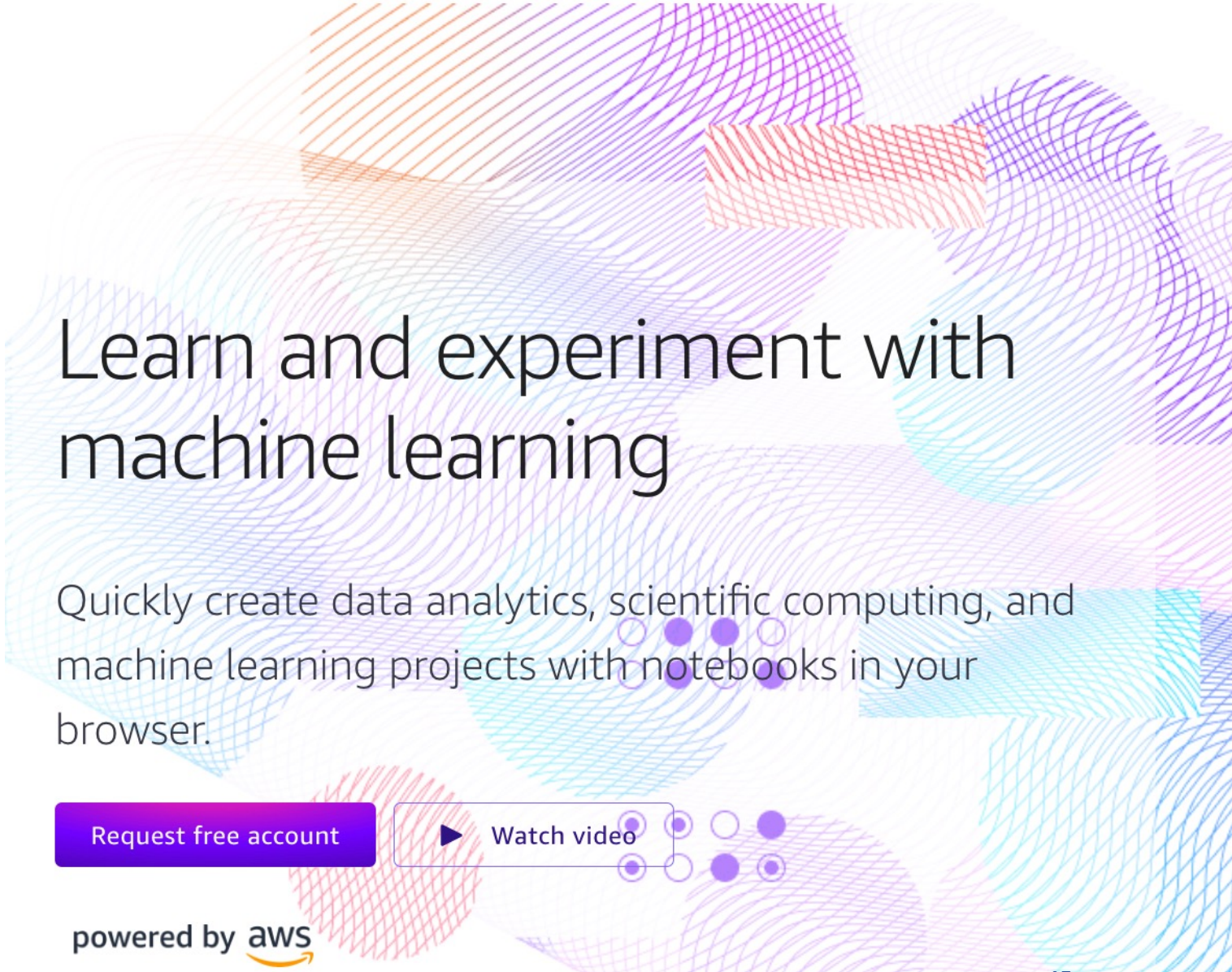
https://huggingface.co/

But you're free to use any other framework especially for your projects: e.g. Tensorflow, Apache MXNet, JAX

We will also be using…

https://colab.research.google.com/

Also try using:



https://aws.amazon.com/sagemaker/studio-lab/

Demos

https://vislang.ai/genderless

vislang | University of Virginia

home people demos publications

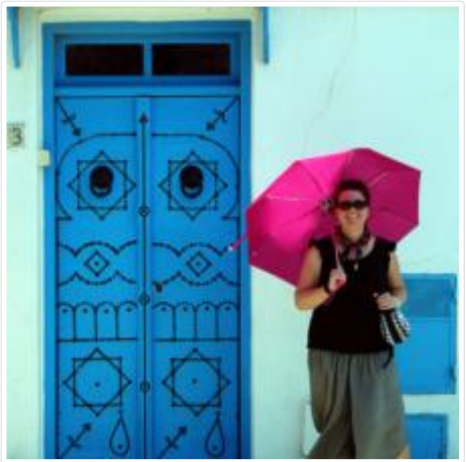
Genderless

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.

upload an image | paste image URL

Tap here to choose an image...

Original Image



Genderless Image

Demo by Lindsey and Vicente

Demos

https://vislang.ai/text2scene

vislang | University of Virginia

home people demos publications

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/uavision/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

For Next Class…

- Intro to Machine Learning

- You need to complete the following two activities:

Completing this [Primer on Image Processing], and optionally, the tutorial and assignment on [Image Classification] from my old Deep Learning for Visual Recognition class.

Questions?