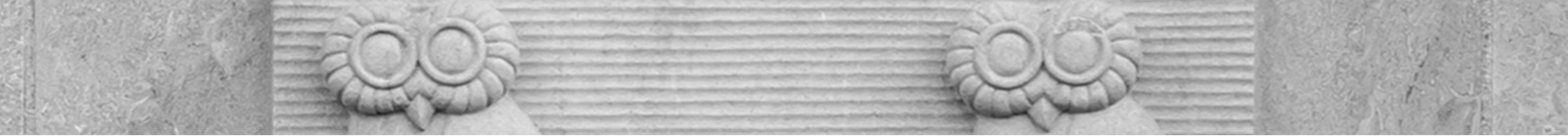
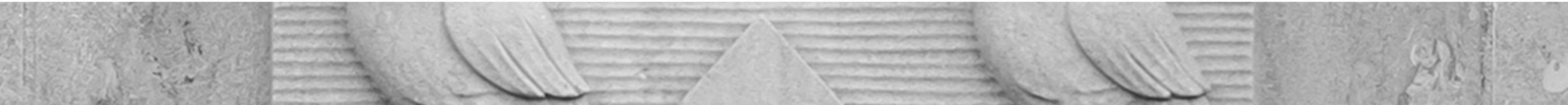


Deep Learning for Vision and Language

Welcome and Introduction

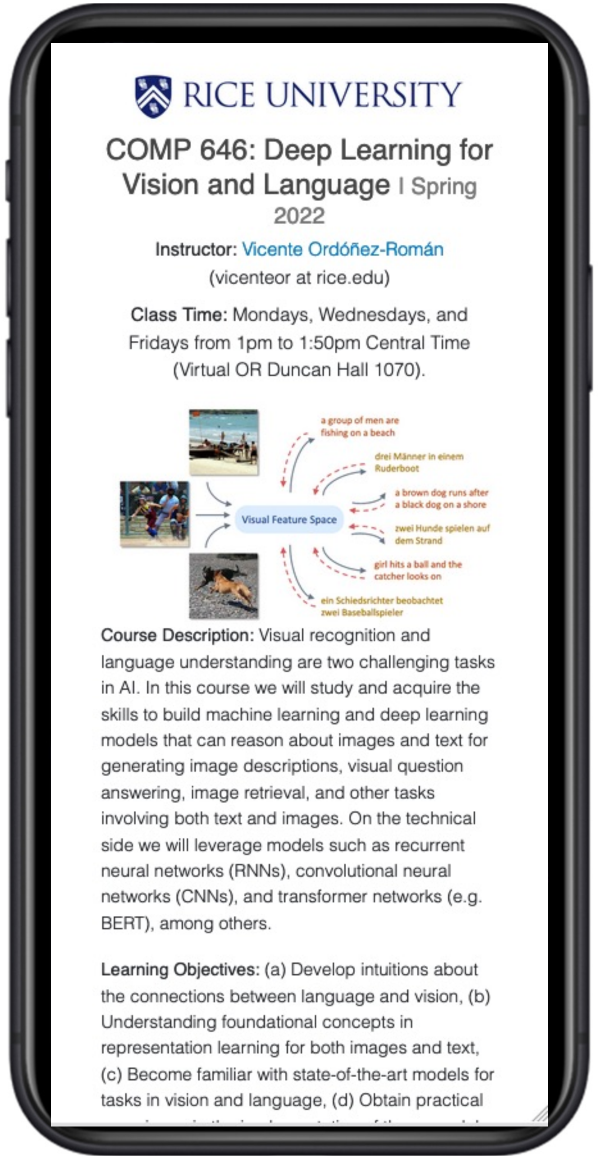
RICE UNIVERSITY



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: Keck Hall 100
- Times: Tuesdays and Thursdays
  from 4pm to 5:15pm
- Office Hours: TBD (Duncan Hall 2080)
- Teaching Assistants: TBD
- Discussion Forum: Piazza (Sign-up Link on Rice Canvas)

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



RICE UNIVERSITY

COMP 646: Deep Learning for Vision and Language | Spring 2022

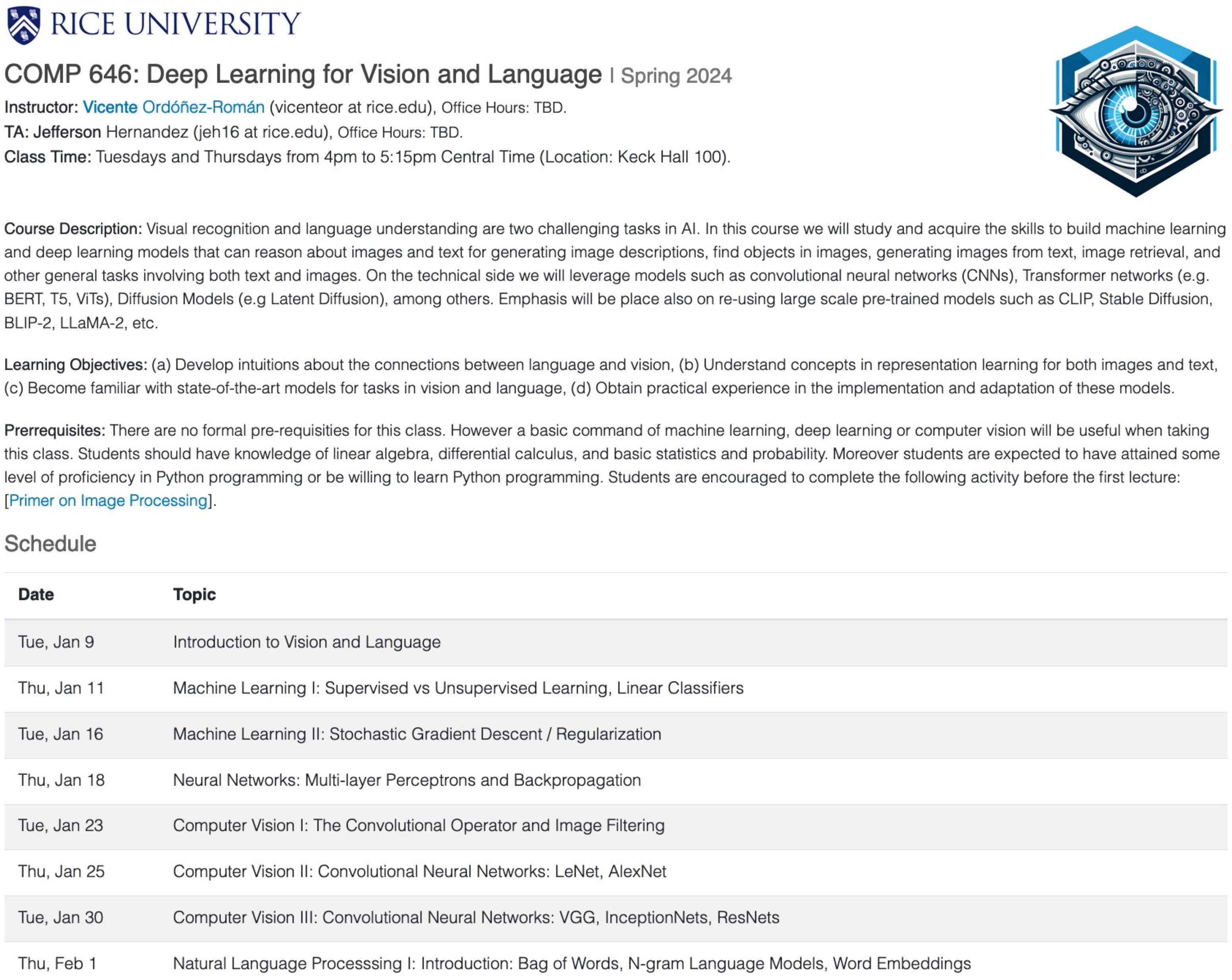
Instructor: Vicente Ordóñez-Román (vicenteor at rice.edu)

Class Time: Mondays, Wednesdays, and Fridays from 1pm to 1:50pm Central Time (Virtual OR Duncan Hall 1070).

a group of men are fishing on a beach
drei Männer in einem Ruderboot
a brown dog runs after a black dog on a shore
Visual Feature Space
zwei Hunde spielen auf dem Strand
girl hits a ball and the catcher looks on
ein Schiedsrichter beobachtet zwei Baseballspieler

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), 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



RICE UNIVERSITY

COMP 646: Deep Learning for Vision and Language | Spring 2024

Instructor: Vicente Ordóñez-Román (vicenteor at rice.edu), Office Hours: TBD.
TA: Jefferson Hernandez (jeh16 at rice.edu), Office Hours: TBD.
Class Time: Tuesdays and Thursdays from 4pm to 5:15pm Central Time (Location: Keck Hall 100).

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, find objects in images, generating images from text, image retrieval, and other general tasks involving both text and images. On the technical side we will leverage models such as convolutional neural networks (CNNs), Transformer networks (e.g. BERT, T5, ViTs), Diffusion Models (e.g Latent Diffusion), among others. Emphasis will be place also on re-using large scale pre-trained models such as CLIP, Stable Diffusion, BLIP-2, LLaMA-2, etc.

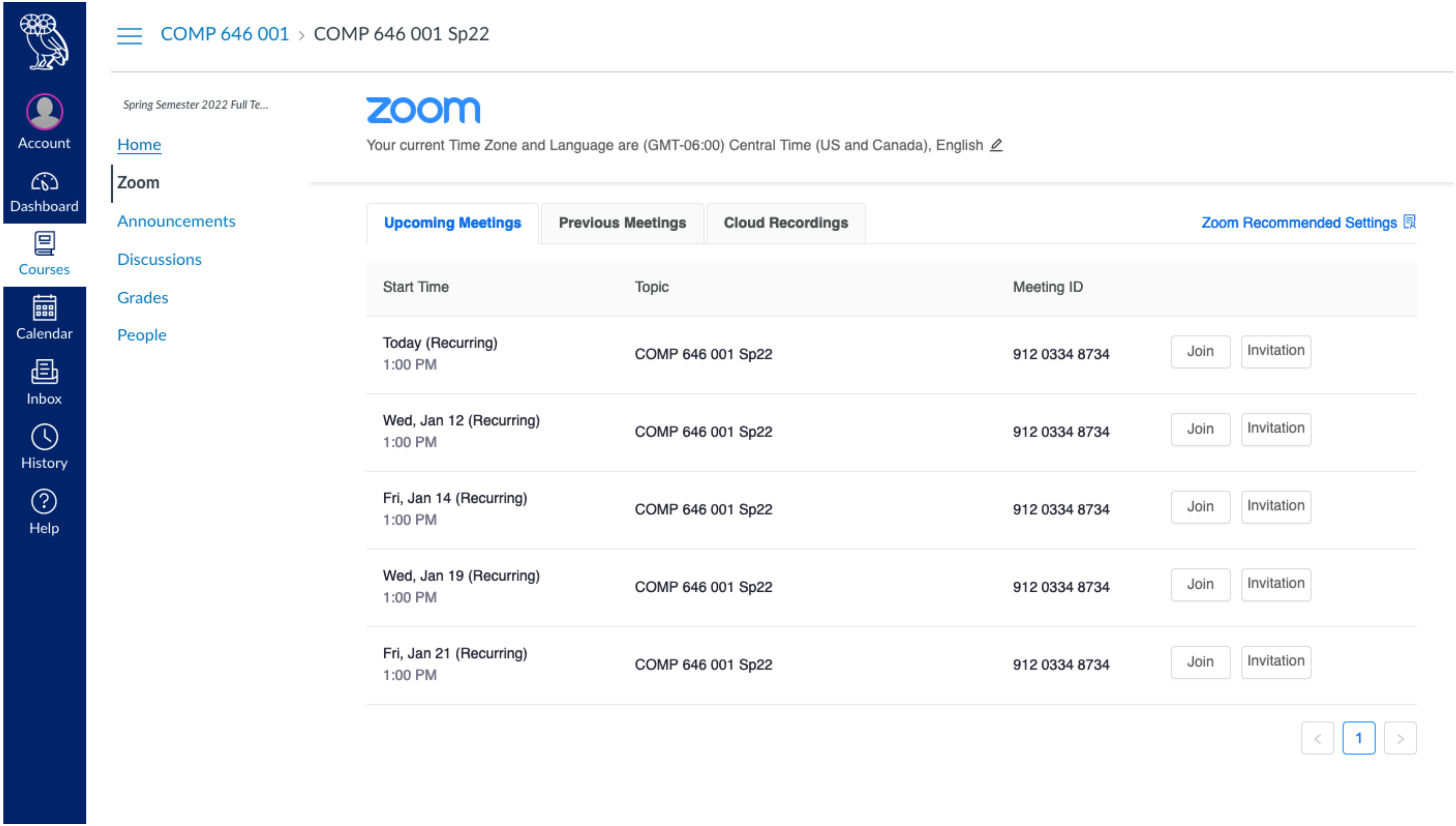
Learning Objectives: (a) Develop intuitions about the connections between language and vision, (b) Understand concepts in representation learning for both images and text, (c) Become familiar with state-of-the-art models for tasks in vision and language, (d) Obtain practical experience in the implementation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

Schedule

Date	Topic
Tue, Jan 9	Introduction to Vision and Language
Thu, Jan 11	Machine Learning I: Supervised vs Unsupervised Learning, Linear Classifiers
Tue, Jan 16	Machine Learning II: Stochastic Gradient Descent / Regularization
Thu, Jan 18	Neural Networks: Multi-layer Perceptrons and Backpropagation
Tue, Jan 23	Computer Vision I: The Convolutional Operator and Image Filtering
Thu, Jan 25	Computer Vision II: Convolutional Neural Networks: LeNet, AlexNet
Tue, Jan 30	Computer Vision III: Convolutional Neural Networks: VGG, InceptionNets, ResNets
Thu, Feb 1	Natural Language Processsing I: Introduction: Bag of Words, N-gram Language Models, Word Embeddings

Zoom Links



Your current Time Zone and Language are (GMT-06:00) Central Time (US and Canada), English

Start Time	Topic	Meeting ID
Today (Recurring) 1:00 PM	COMP 646 001 Sp22	912 0334 8734
Wed, Jan 12 (Recurring) 1:00 PM	COMP 646 001 Sp22	912 0334 8734
Fri, Jan 14 (Recurring) 1:00 PM	COMP 646 001 Sp22	912 0334 8734
Wed, Jan 19 (Recurring) 1:00 PM	COMP 646 001 Sp22	912 0334 8734
Fri, Jan 21 (Recurring) 1:00 PM	COMP 646 001 Sp22	912 0334 8734

About me -- Vicente

Associate Professor, 2021 - Present	RICE UNIVERSITY
Visiting Academic 2021 - Present	amazon
Assistant Professor, 2016 - 2021	UNIVERSITY of VIRGINIA
Visiting Professor, 2019	Adobe 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:

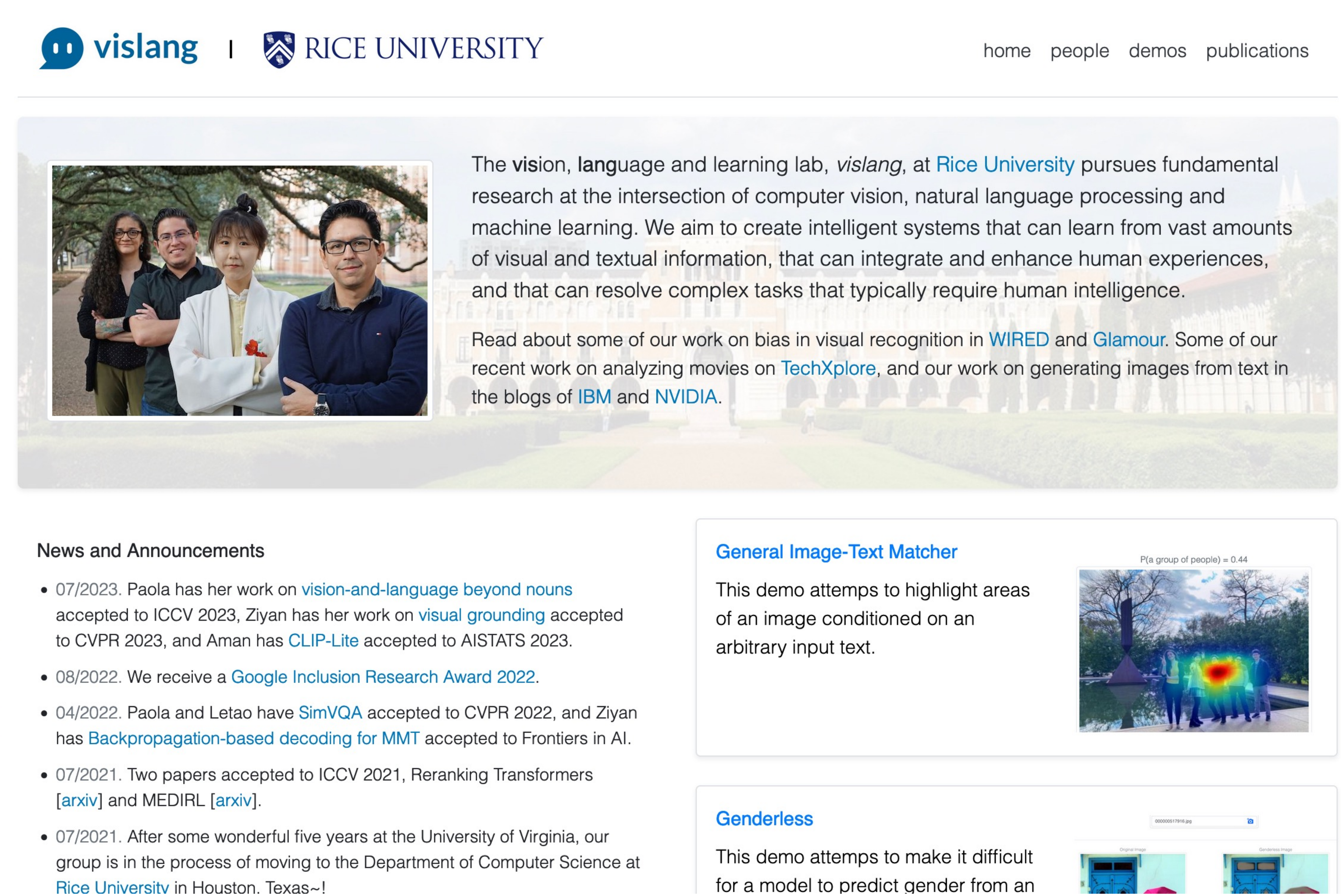
Google Microsoft ebay

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, **lang**uage and learning



vislang | RICE UNIVERSITY

home people demos publications

The **vis**ion, **lang**uage 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

- 07/2023. Paola has her work on vision-and-language beyond nouns accepted to ICCV 2023, Ziyan has her work on visual grounding accepted to CVPR 2023, and Aman has CLIP-Lite accepted to AISTATS 2023.
- 08/2022. We receive a Google Inclusion Research Award 2022.
- 04/2022. Paola and Letao have SimVQA accepted to CVPR 2022, and Ziyan has Backpropagation-based decoding for MMT accepted to Frontiers in AI.
- 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~!

General Image-Text Matcher

This demo attemps to highlight areas of an image conditioned on an arbitrary input text.

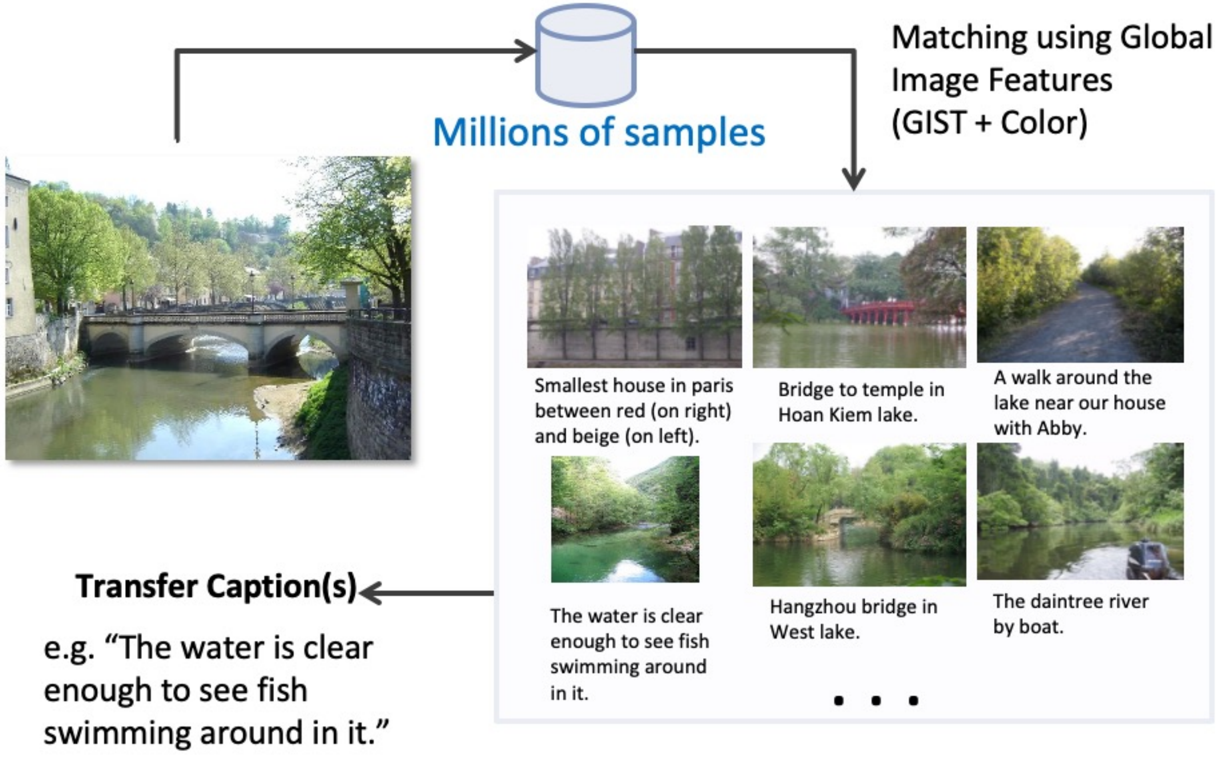
Genderless

This demo attemps to make it difficult for a model to predict gender from an

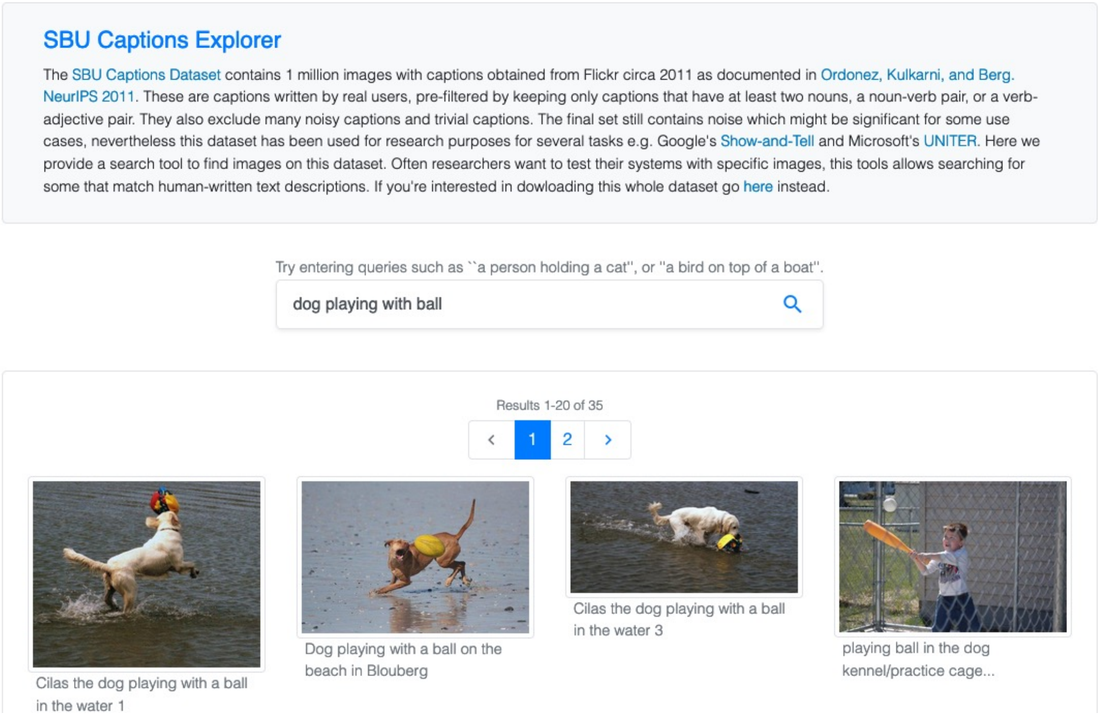
Some of our work includes…

Describing images with language

https://vislang.ai/sbu-explorer


Millions of samples
Matching using Global Image Features (GIST + Color)
Smallest house in paris between red (on right) and beige (on left).
Bridge to temple in Hoan Kiem lake.
A walk around the lake near our house with Abby.
The water is clear enough to see fish swimming around in it.
Hangzhou bridge in West lake.
The daintree river by boat.
• • •
Transfer Caption(s)

e.g. "The water is clear enough to see fish swimming around in it."



SBU Captions Explorer

The SBU Captions Dataset contains 1 million images with captions obtained from Flickr circa 2011 as documented in Ordonez, Kulkarni, and Berg. NeurIPS 2011. These are captions written by real users, pre-filtered by keeping only captions that have at least two nouns, a noun-verb pair, or a verb-adjective pair. They also exclude many noisy captions and trivial captions. The final set still contains noise which might be significant for some use cases, nevertheless this dataset has been used for research purposes for several tasks e.g. Google's Show-and-Tell and Microsoft's UNITER. Here we provide a search tool to find images on this dataset. Often researchers want to test their systems with specific images, this tools allows searching for some that match human-written text descriptions. If you're interested in dowloading this whole dataset go here instead.

Try entering queries such as ``a person holding a cat'', or ''a bird on top of a boat''.

dog playing with ball

Results 1-20 of 35

Cilas the dog playing with a ball in the water 1

Dog playing with a ball on the beach in Blouberg

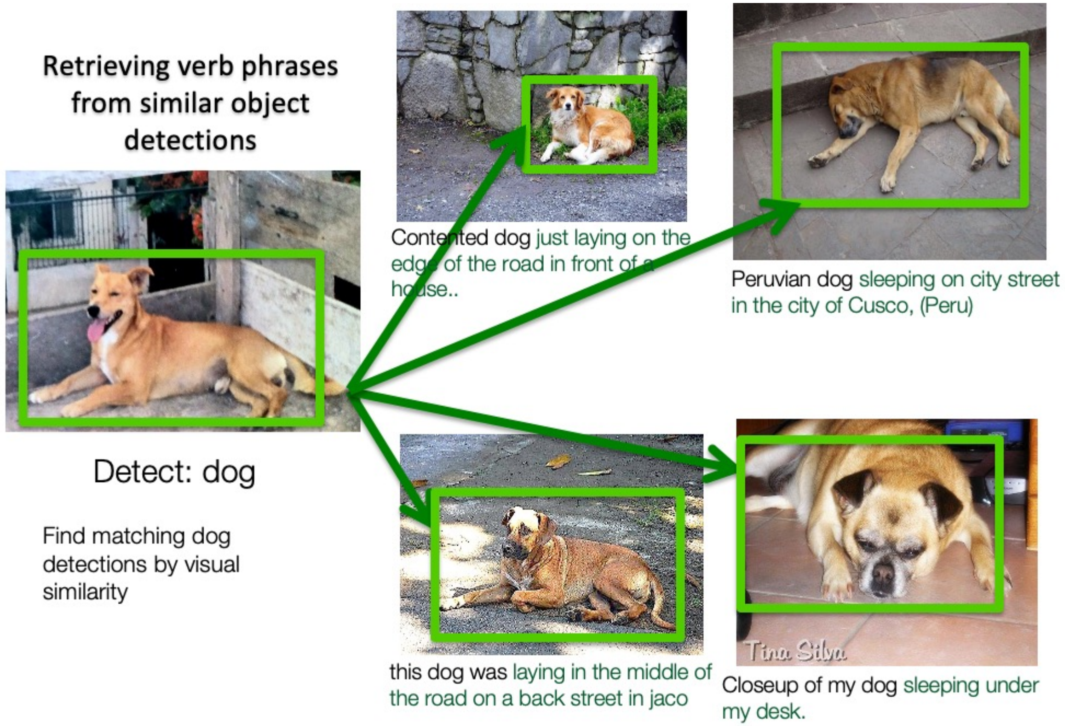
Cilas the dog playing with a ball in the water 3

playing ball in the dog kennel/practice cage...

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



Retrieving verb phrases from similar object detections

Detect: dog

Find matching dog detections by visual similarity

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

Peruvian dog sleeping on city street in the city of Cusco, (Peru)

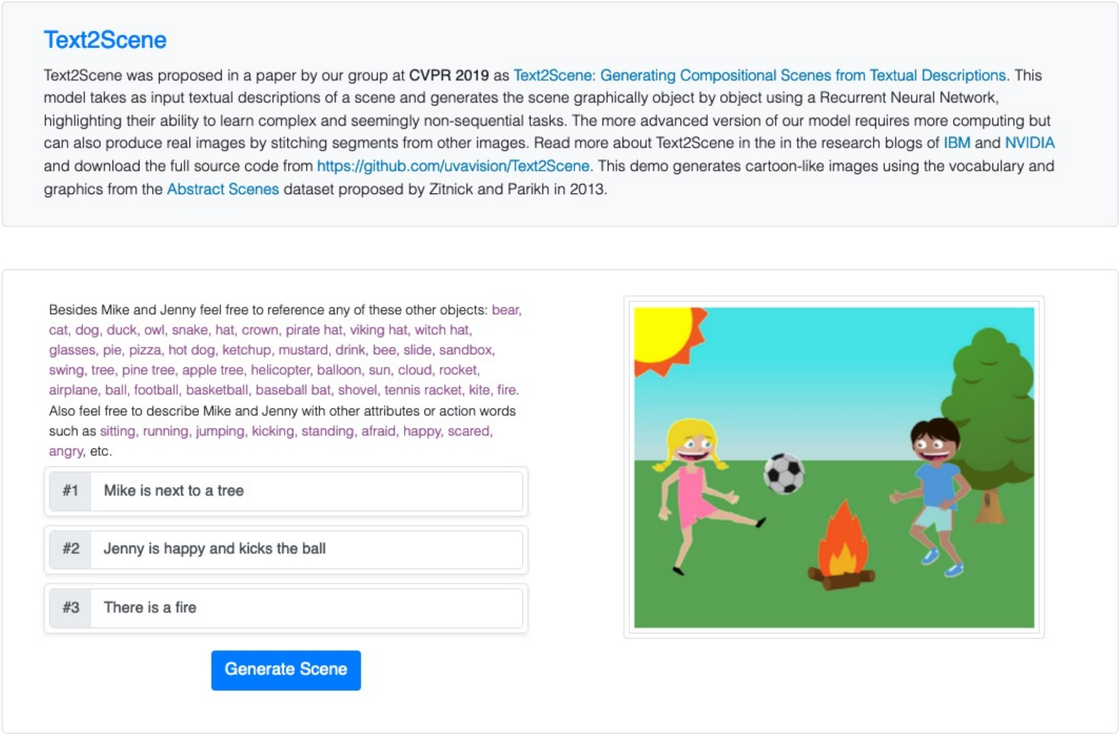
this dog was laying in the middle of the road on a back street in jaco

Closeup of my dog sleeping under my desk.

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



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/uvavision/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

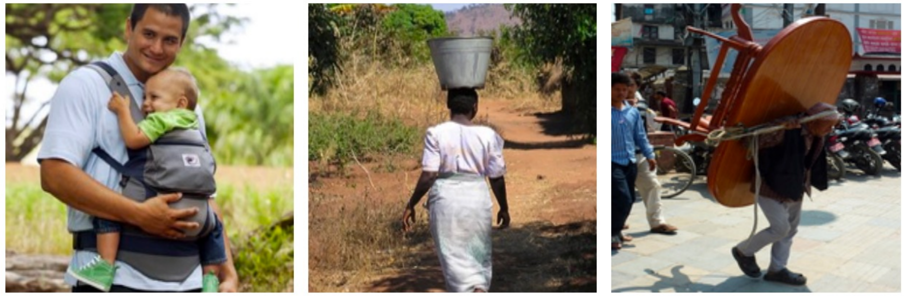
#2 Jenny is happy and kicks the ball

#3 There is a fire

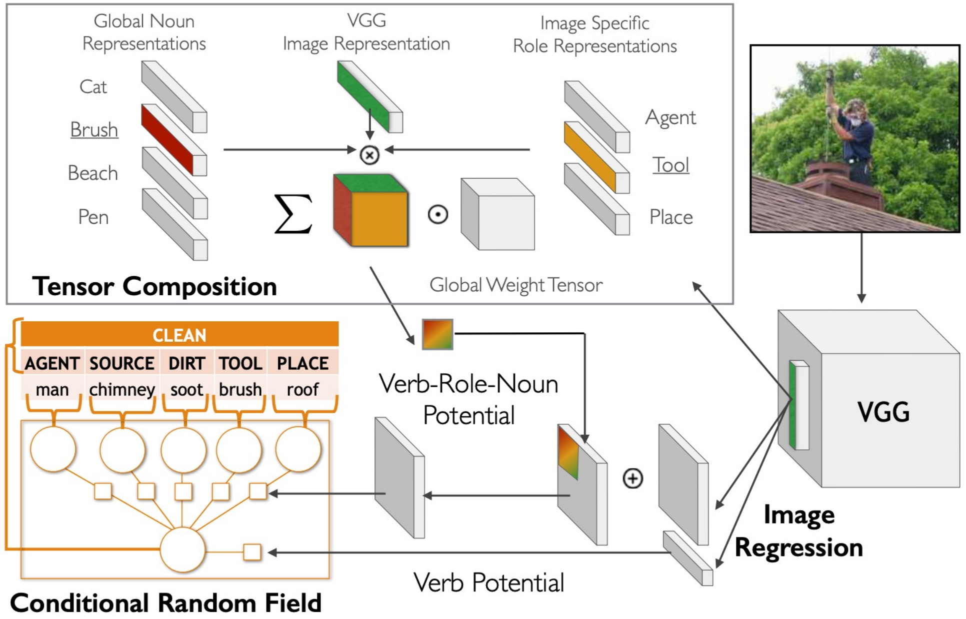
Demo by Leticia and Vicente

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

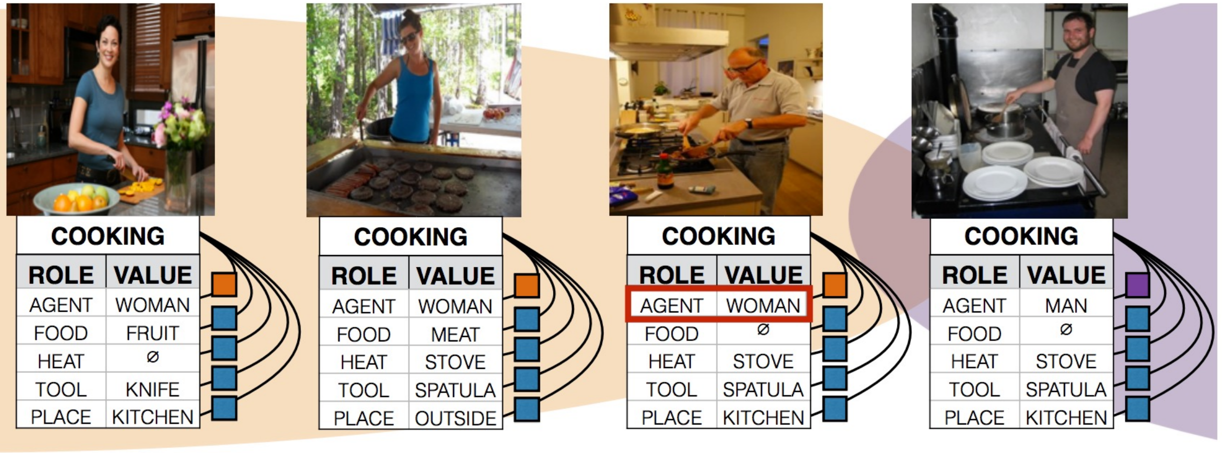


http://imsitu.org/

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



COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	FRUIT
HEAT	∅
TOOL	KNIFE
PLACE	KITCHEN

COOKING

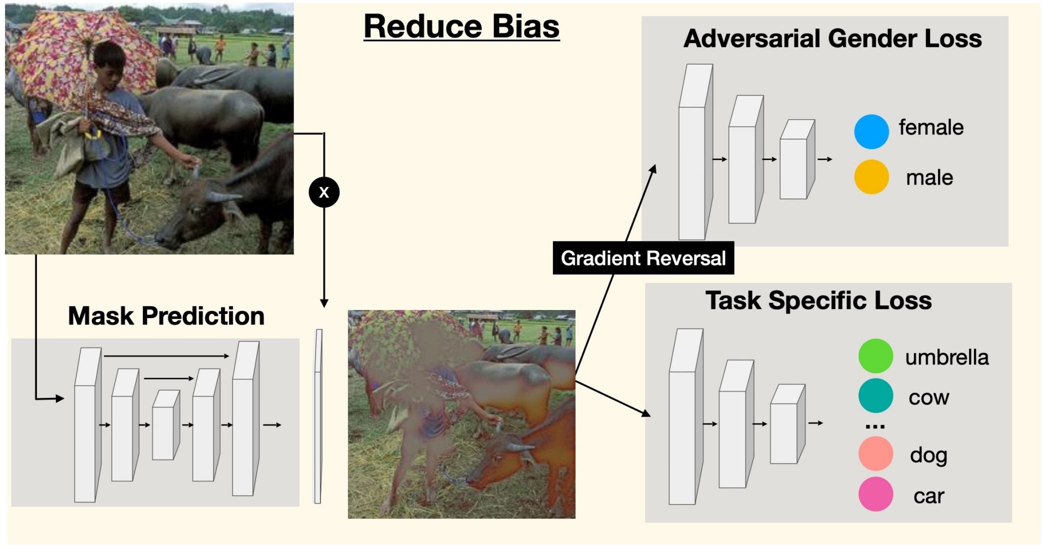
ROLE	VALUE
AGENT	WOMAN
FOOD	MEAT
HEAT	STOVE
TOOL	SPATULA
PLACE	OUTSIDE

COOKING

ROLE	VALUE
AGENT	WOMAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

COOKING

ROLE	VALUE
AGENT	MAN
FOOD	∅
HEAT	STOVE
TOOL	SPATULA
PLACE	KITCHEN

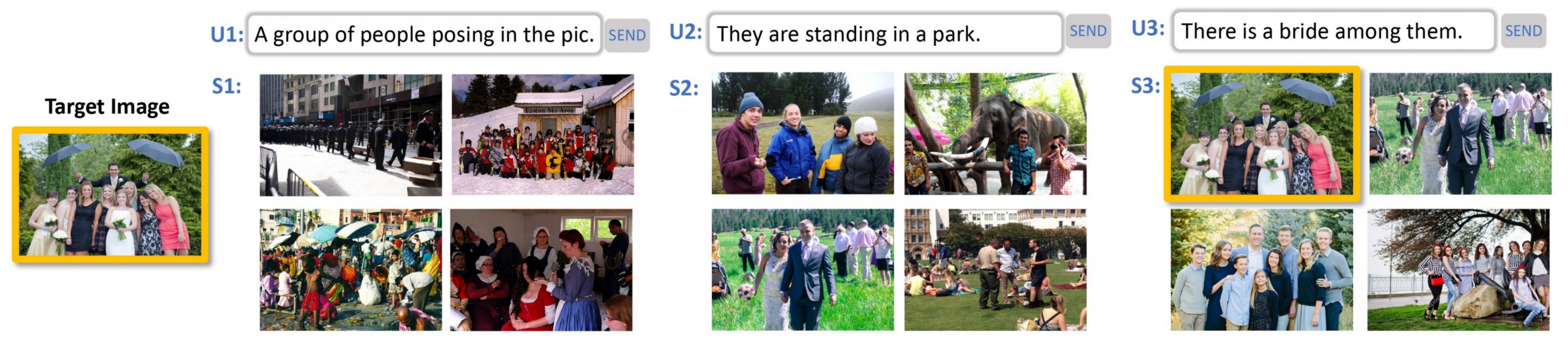


https://www.vislang.ai/genderless

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



U1: A group of people posing in the pic.

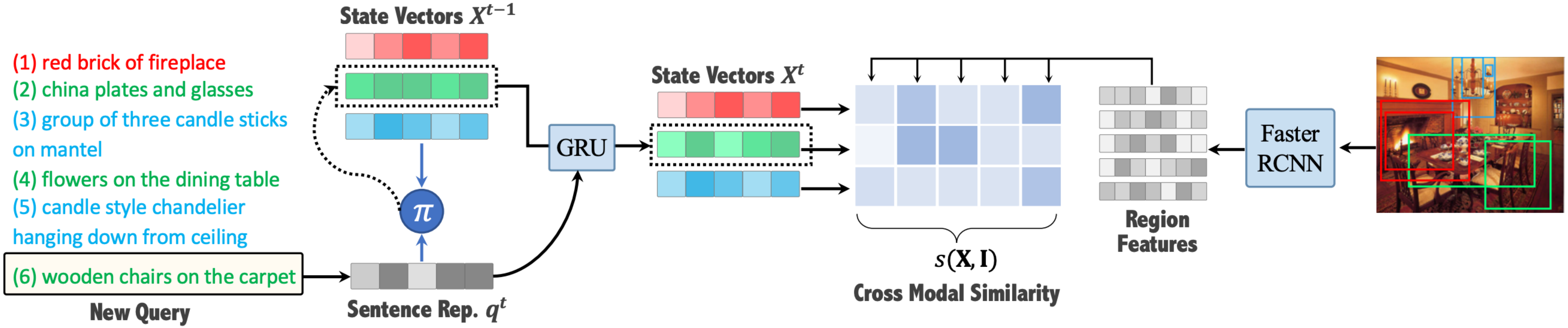
U2: They are standing in a park.

U3: There is a bride among them.

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



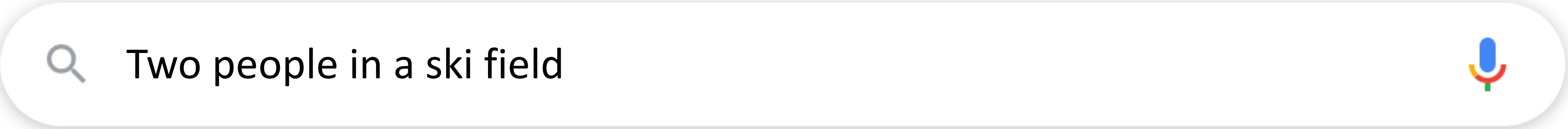
State Vectors X^{t-1}
(1) red brick of fireplace
(2) china plates and glasses
(3) group of three candle sticks on mantel
(4) flowers on the dining table
(5) candle style chandelier hanging down from ceiling
(6) wooden chairs on the carpet
New Query
π
Sentence Rep. q^t
GRU
State Vectors X^t
$s(\mathbf{X}, \mathbf{I})$
Cross Modal Similarity
Region Features
Faster RCNN

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]

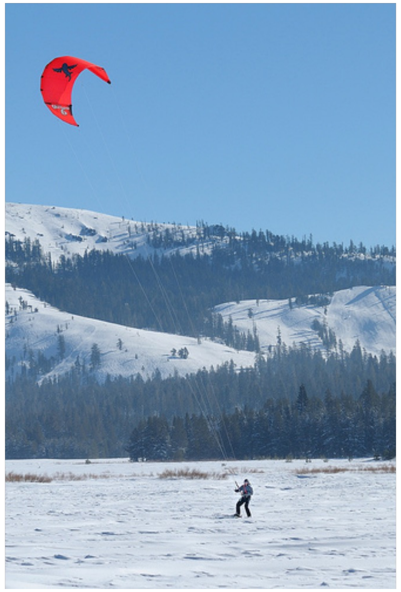
Drill-down: Image Retrieval System

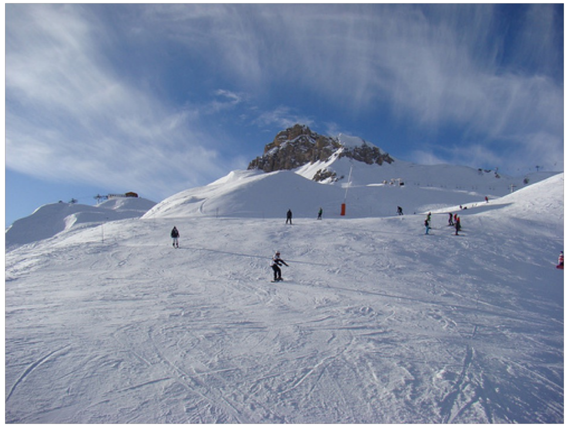
Target

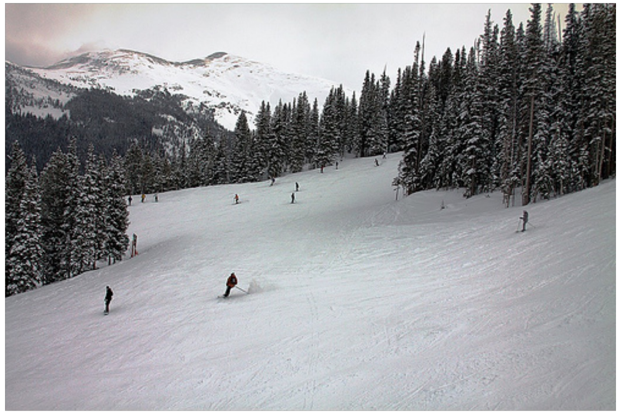
Drill-down: Image Retrieval System



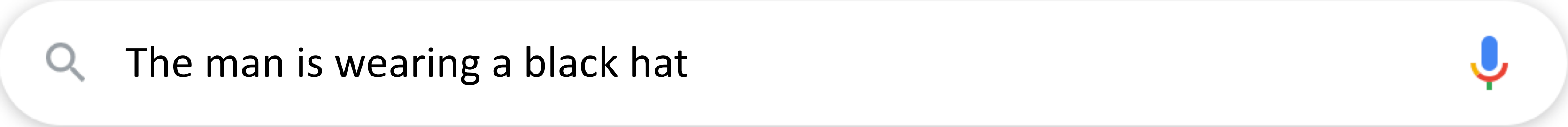
Two people in a ski field





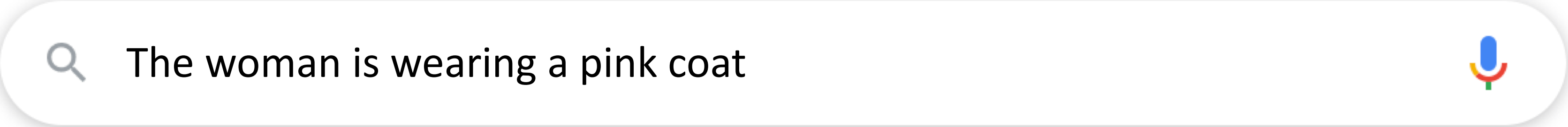


Drill-down: Image Retrieval System

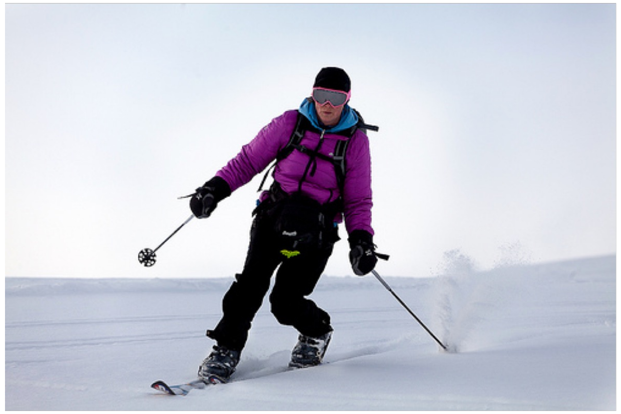


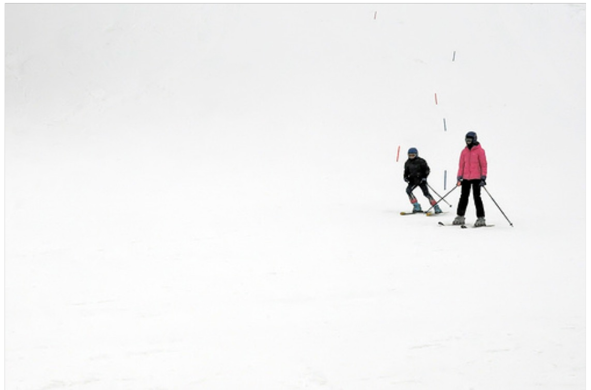
The man is wearing a black hat

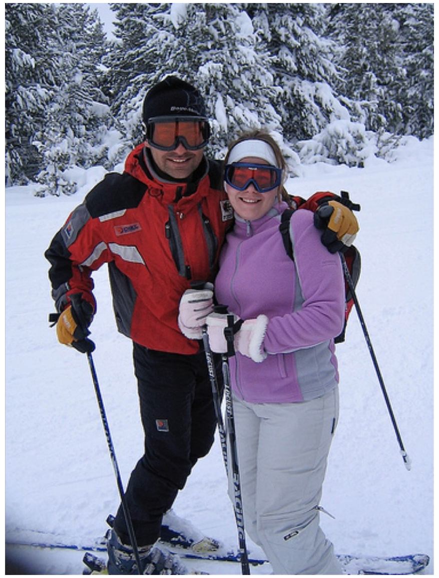
Drill-down: Image Retrieval System



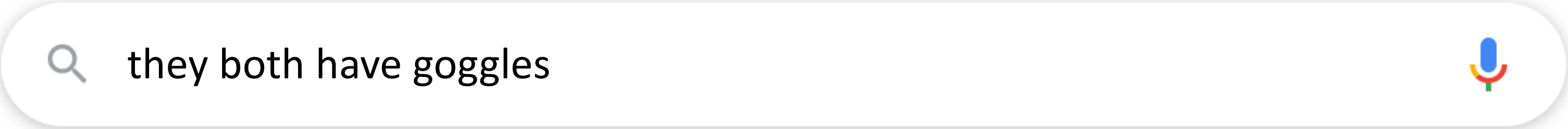
The woman is wearing a pink coat



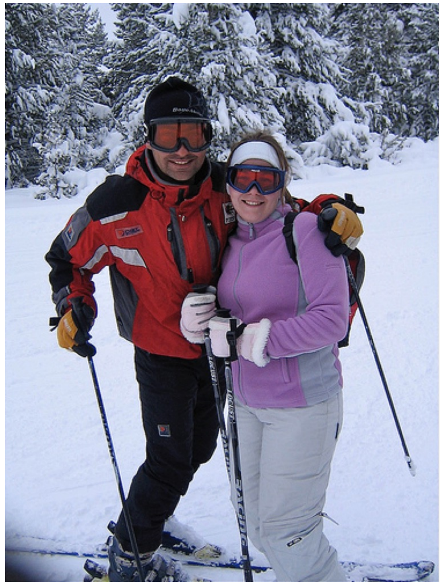


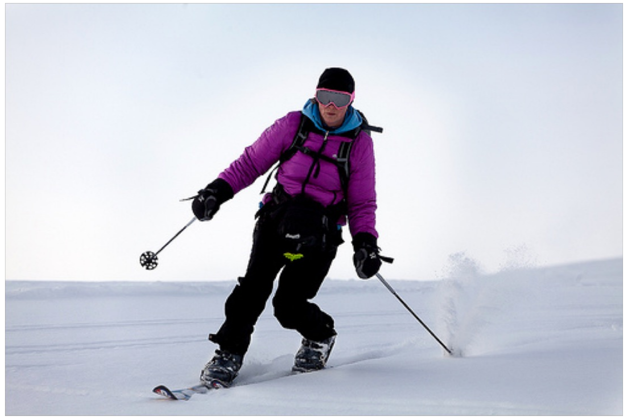


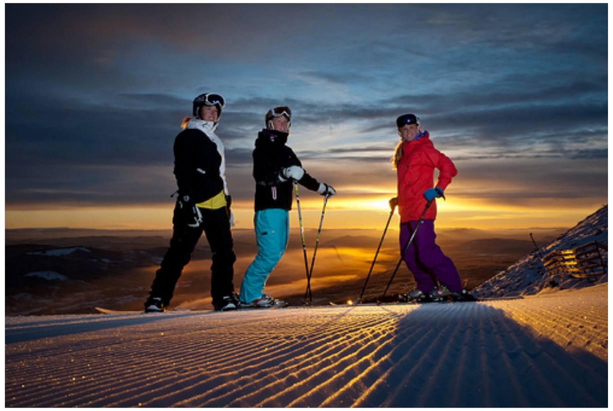
Drill-down: Image Retrieval System



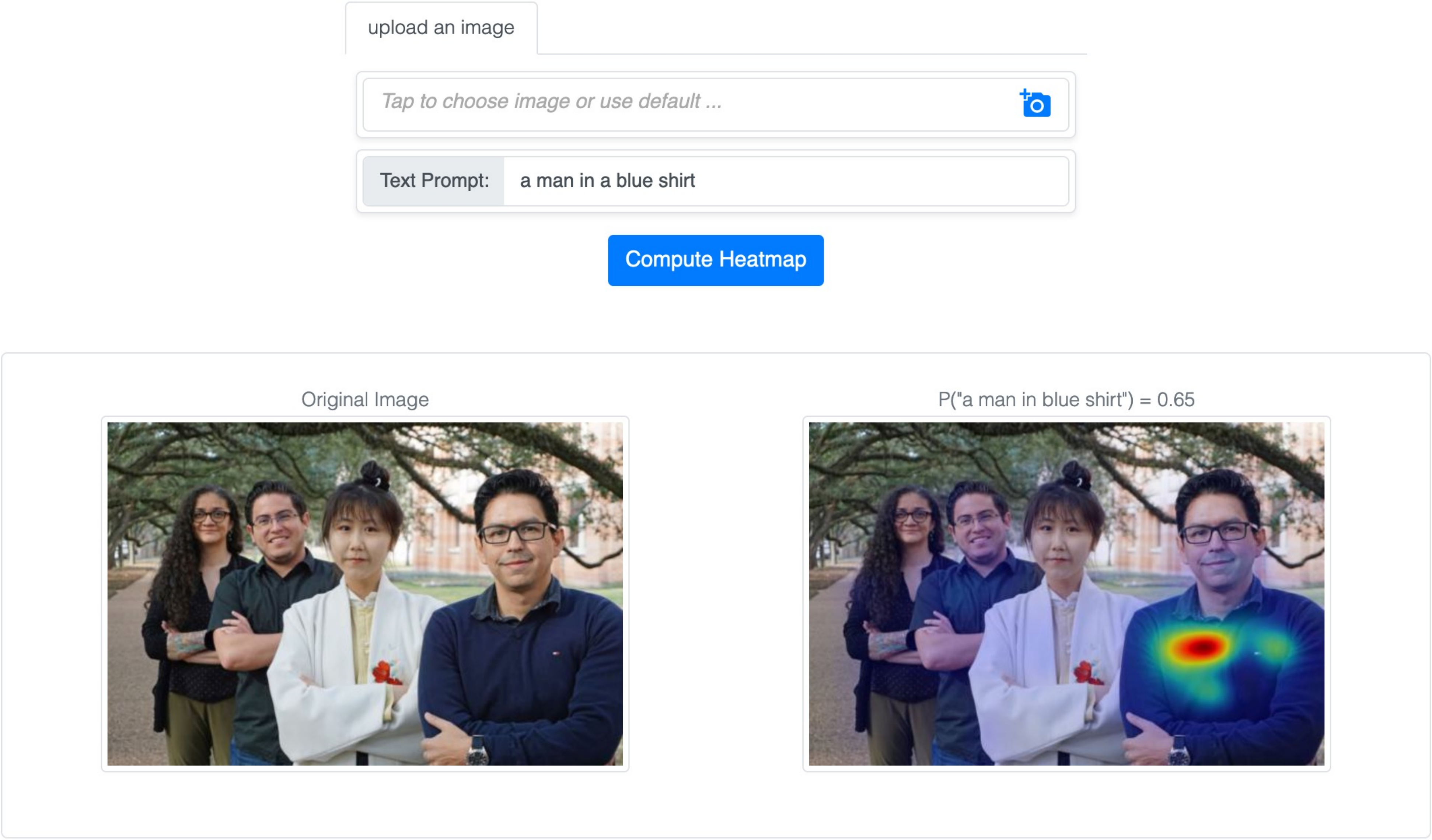
they both have goggles







Visual Grounding

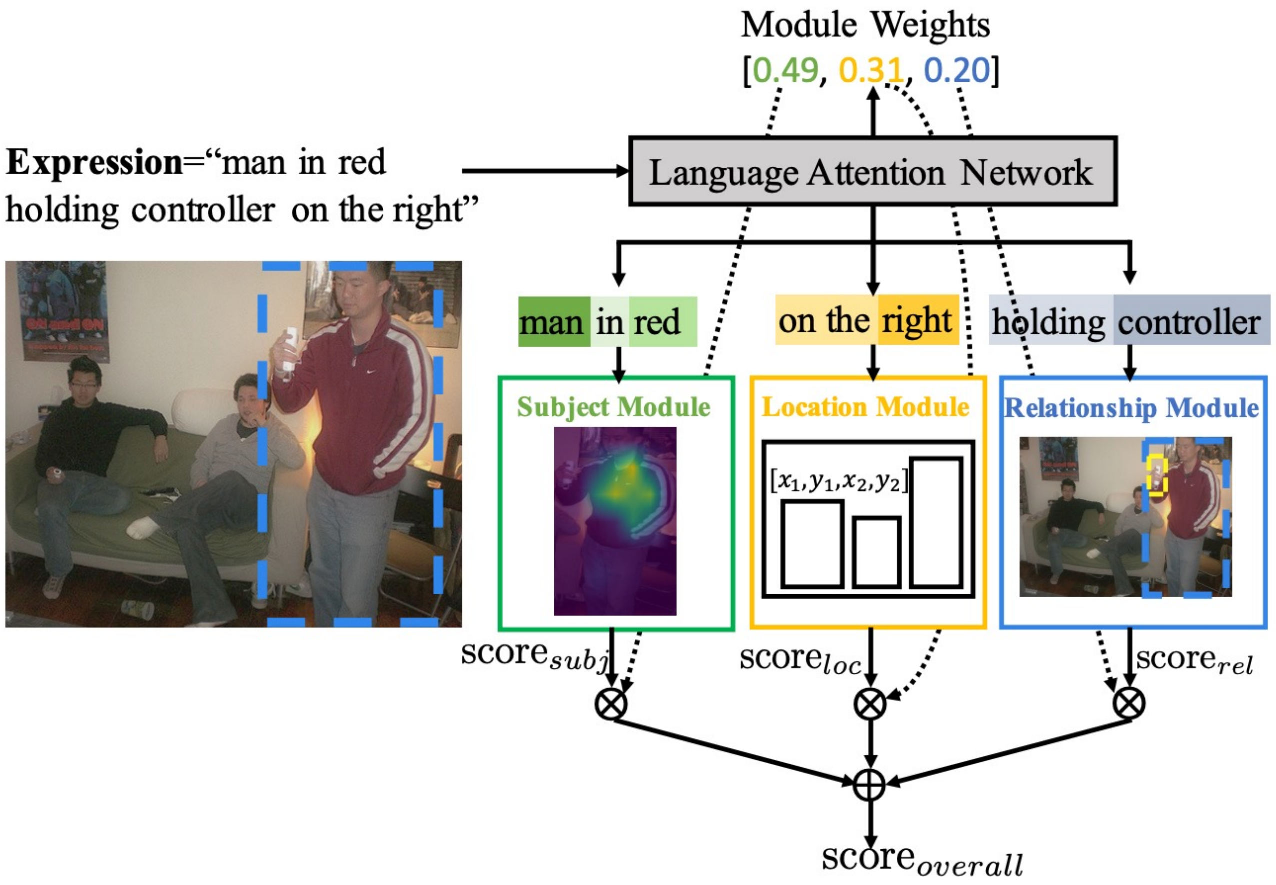


Text Prompt: a man in a blue shirt

P("a man in blue shirt") = 0.65

Demo by Ziyan and Vicente

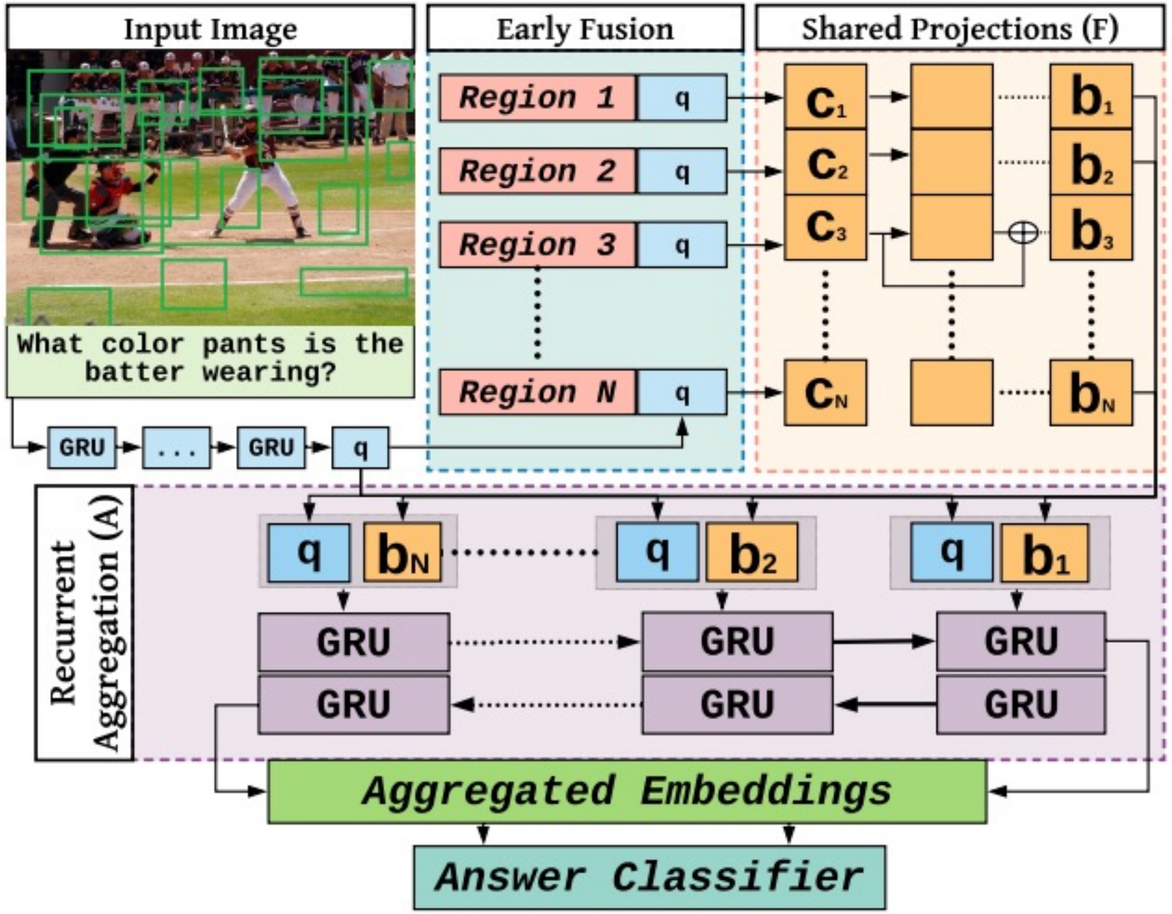
Referring Expression Comprehension



MAttNet: Modular Attention Network for Referring Expression Comprehension

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

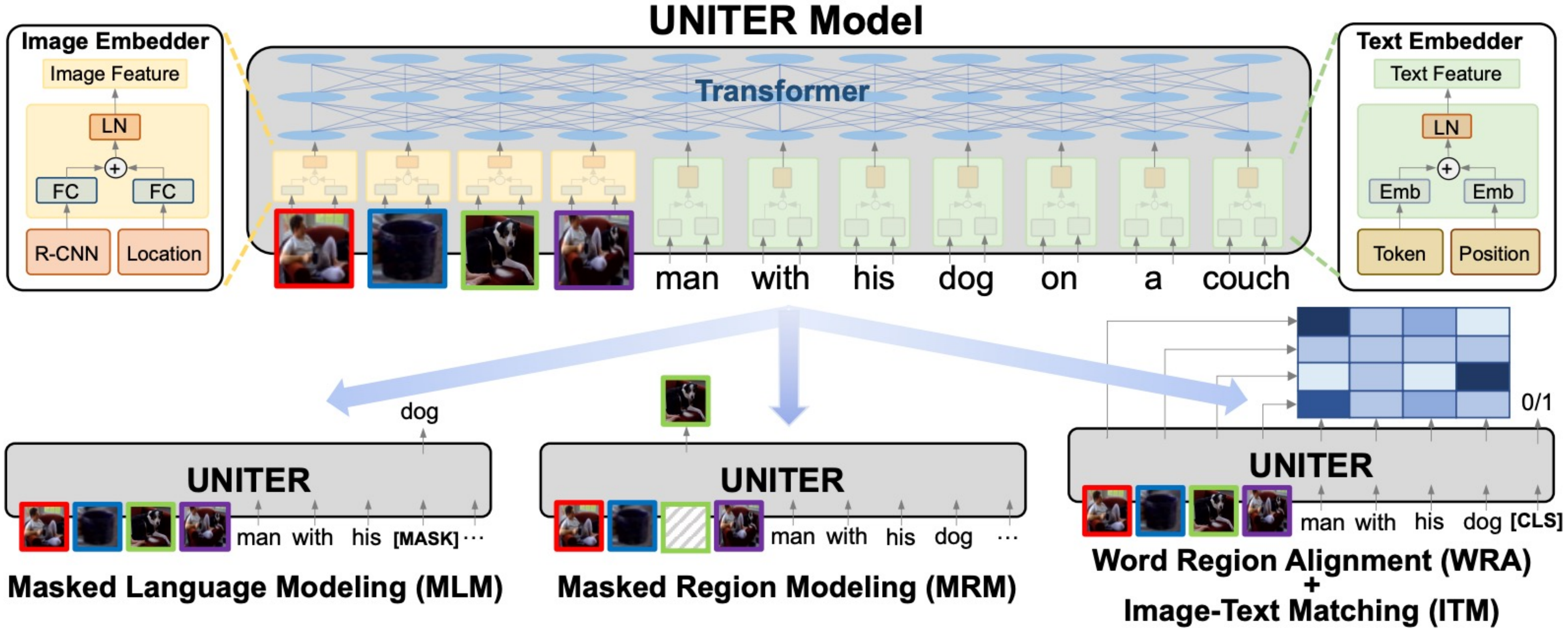
Visual Question Answering



Answer Them All! Toward Universal Visual Question Answering Models

Robik Shrestha, Kushal Kafle, Christopher Kanan

Vision-and-Language Transformers

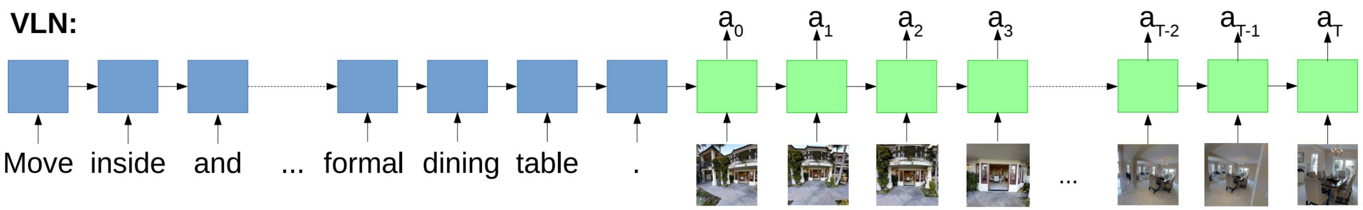


UNITER: UNiversal Image-TExt Representation Learning

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

Vision-and-Language for Navigation

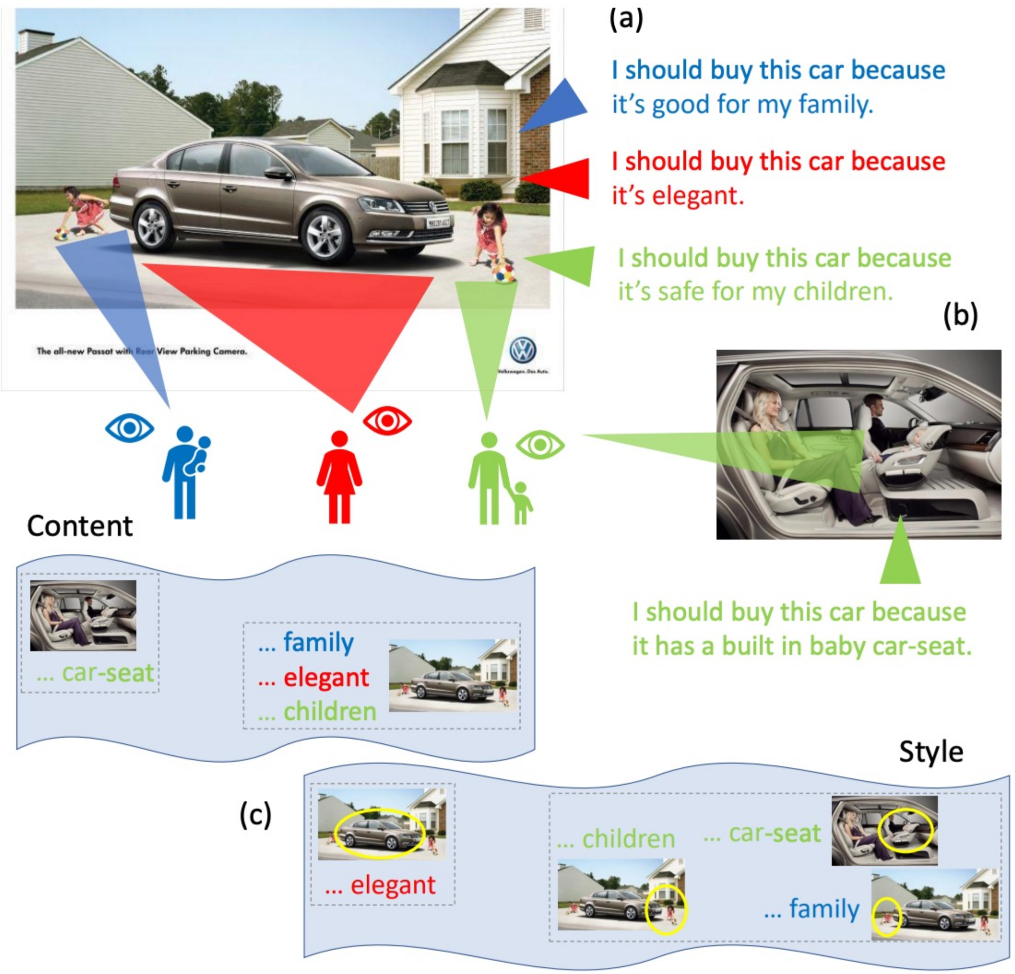
Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.



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

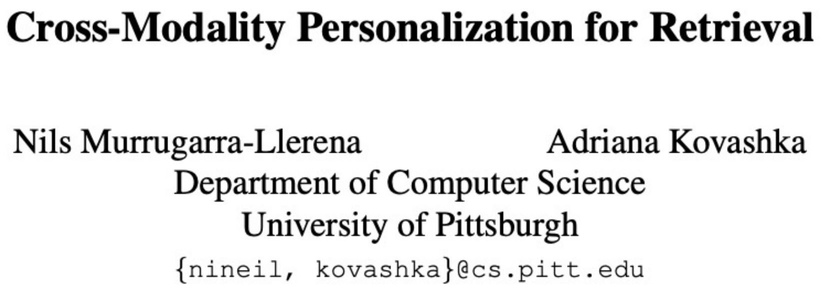


I should buy this car because it's good for my family.

I should buy this car because it's elegant.

I should buy this car because it's safe for my children.

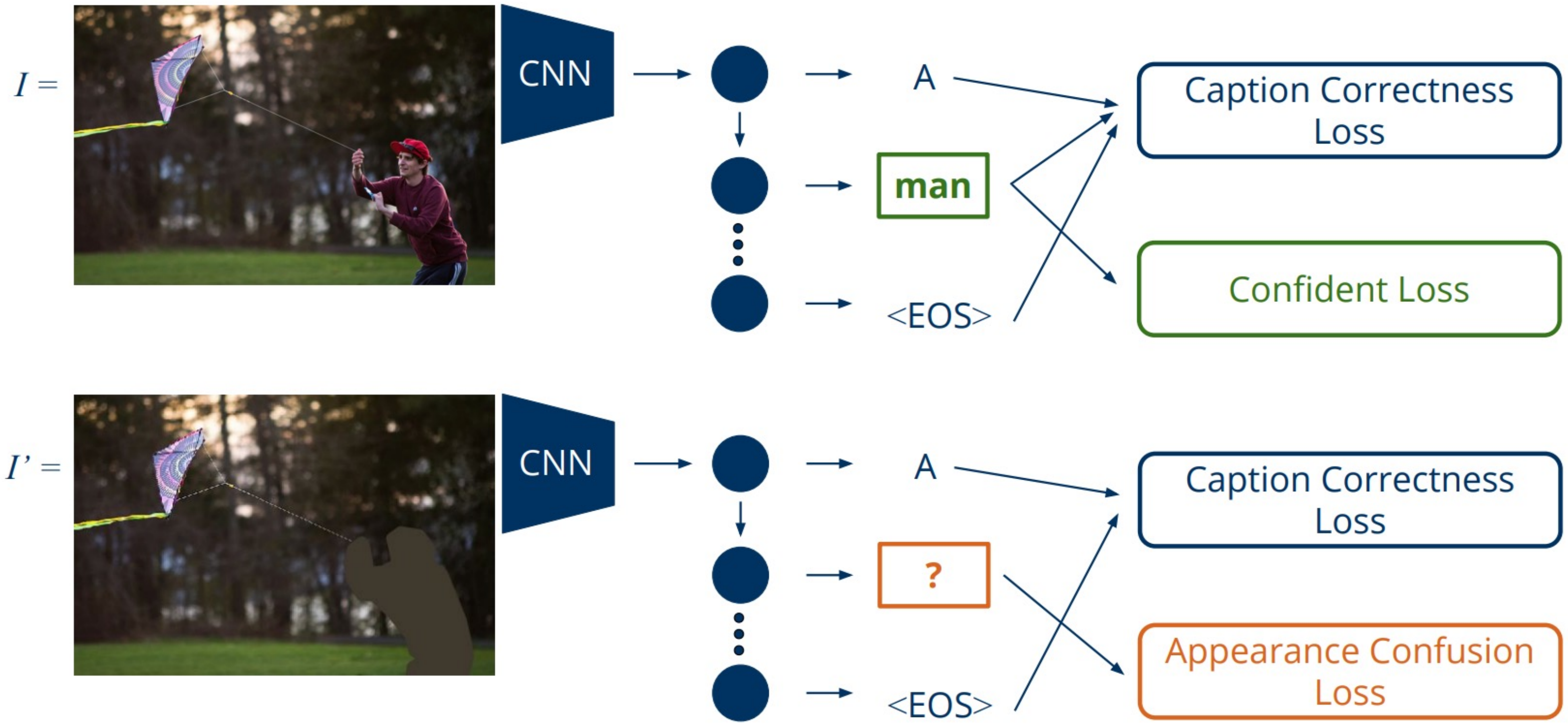
I should buy this car because it has a built in baby car-seat.



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

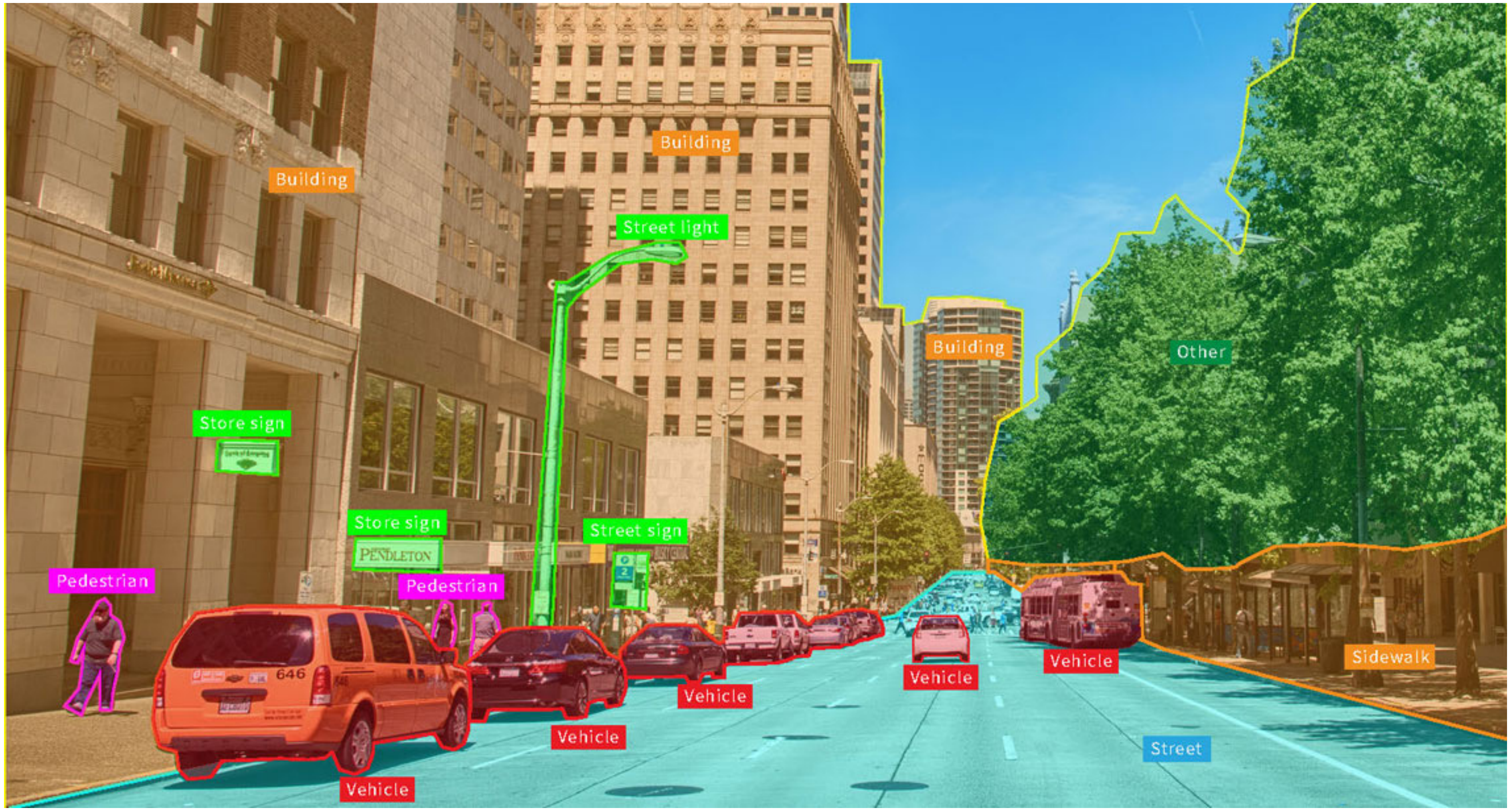
Kaylee Burns, Lisa Anne Hendricks, Kate Saenko, Trevor Darrell, Anna Rohrbach

Why Vision and Language Together?

- What makes us intelligent?

Why Vision and Language Together?

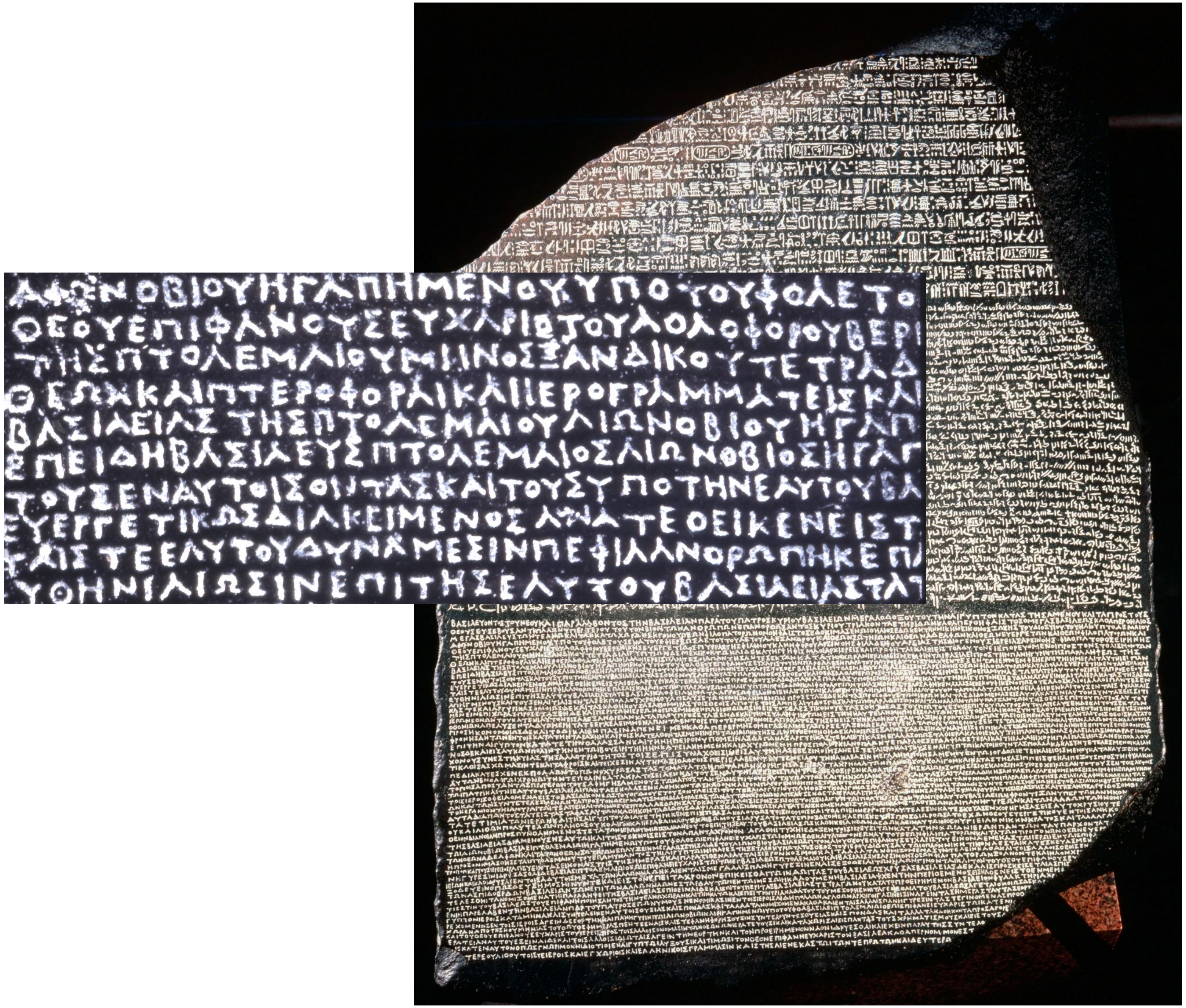
- What makes us intelligent?
- Vision is not just sensing – but interpreting what our eyes capture



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

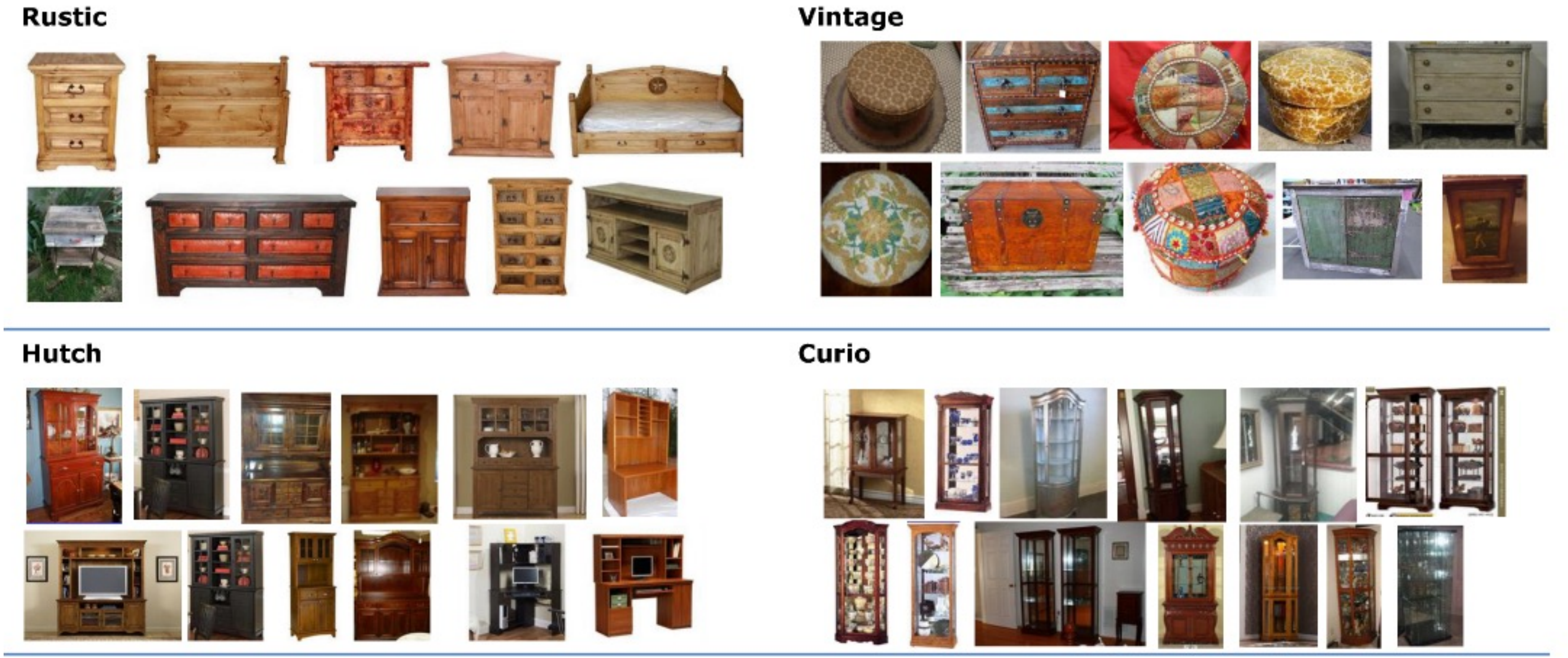


Can we learn language through pictures?

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

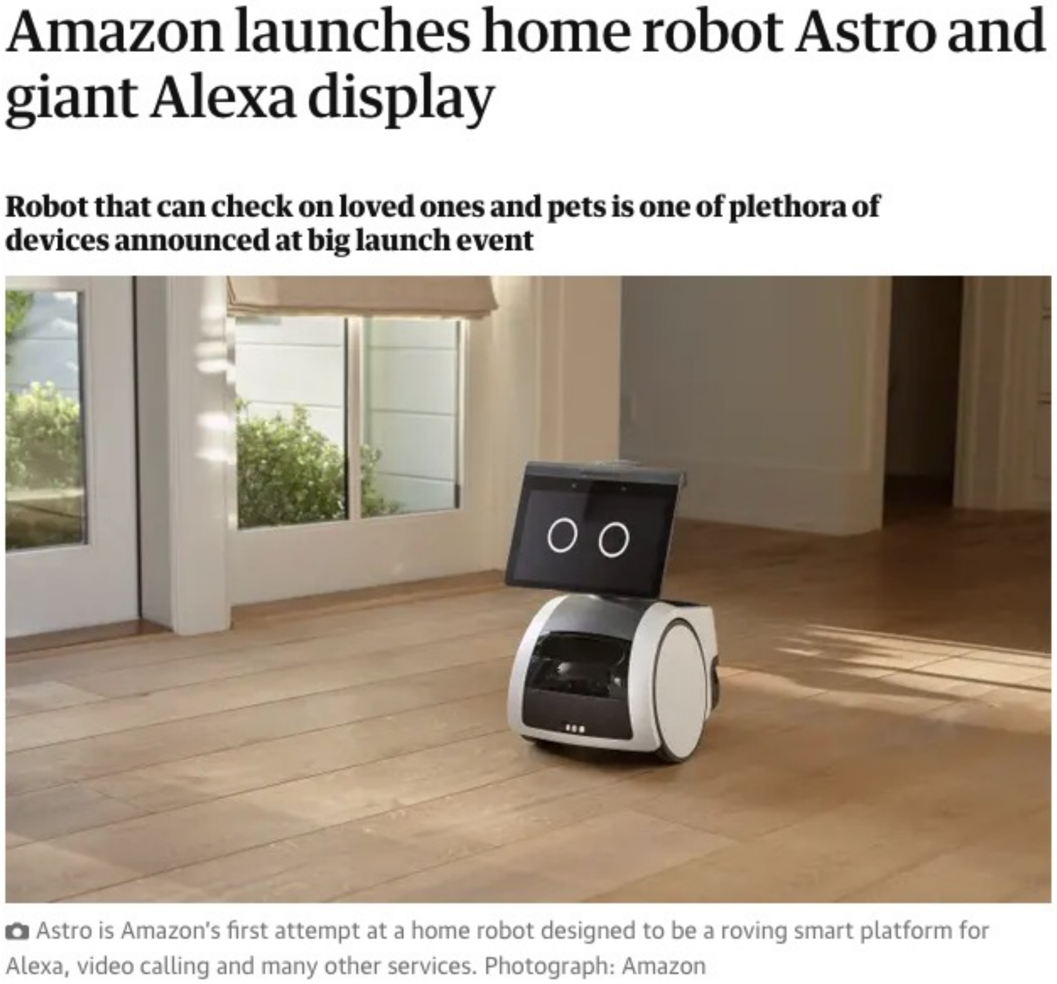
Vision and Language in Practice

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



Vision and Language in Practice

- Robotics: Instruction Following



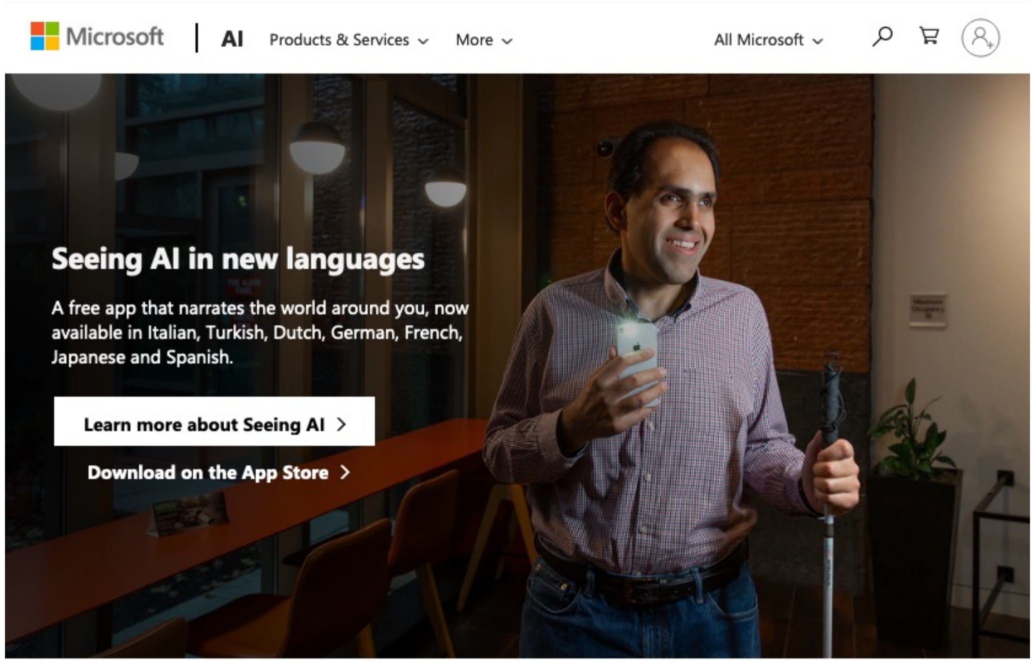
Amazon launches home robot Astro and giant Alexa display

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



Seeing AI in new languages

A free app that narrates the world around you, now available in Italian, Turkish, Dutch, German, French, Japanese and Spanish.

Complete multiple tasks with one app

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

Scene

An experimental feature to describe the scene around you

Color

Describes the perceived color

What will we cover in this class?

In terms of tools

3 weeks

2 weeks

2 weeks

8 weeks

- 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, FLAN-T5, etc)
- Diffusion Models (e.g. Stable Diffusion, ControlNet)
- 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 but no harsher than those indicated in our syllabus.

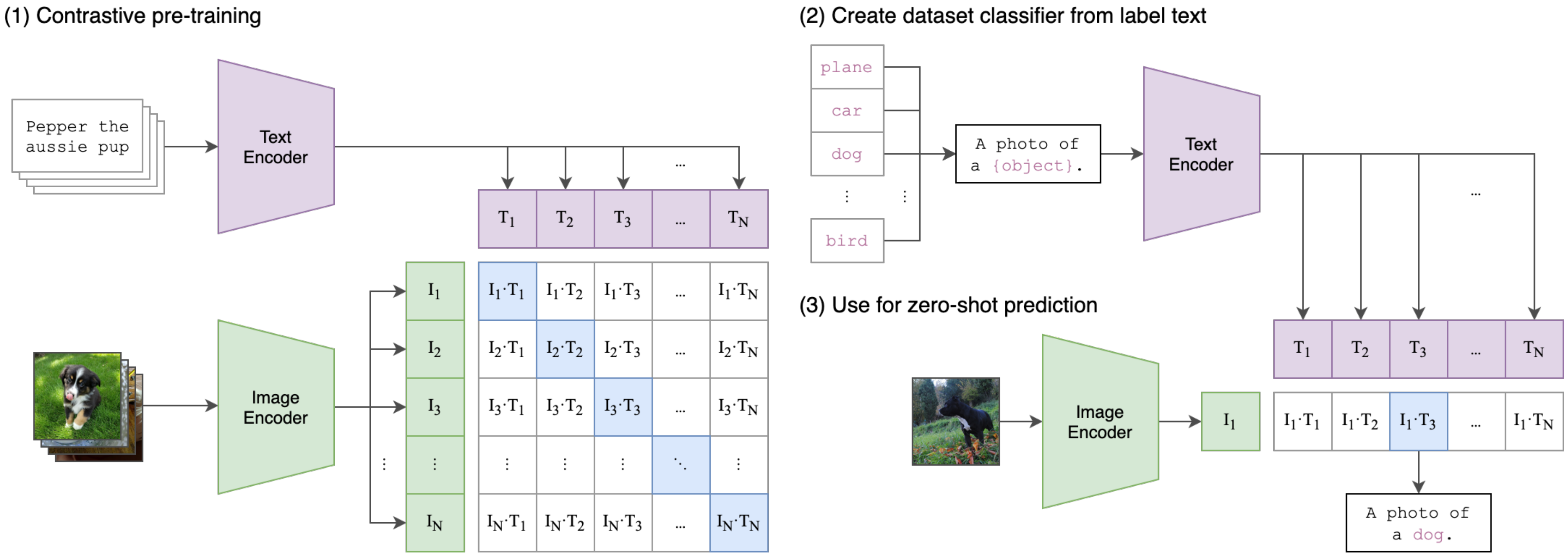
Class Project Timeline

- 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. Do not start from zero.

- CLIP by OpenAI or SigLIP or OpenCLIP or MetaCLIP (Images + Text)
- FLAN-T5 by Google (Text)
- Llama-2 by Meta or Vicuna-7B or Mistral-7B (Text)
- GLIP by Microsoft, or OwL-ViT (Images + Text)
- Whisper by OpenAI (Speech to Text)
- StableDiffusion v1.5 (or SDXL) models (Text to Images)
- LLaVa v1.5 (Multimodal LLM)

CLIP



(1) Contrastive pre-training

Pepper the aussie pup

Text Encoder

Image Encoder

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

(2) Create dataset classifier from label text

plane

car

dog

⋮

bird

A photo of a {object}.

Text Encoder

T_1 T_2 T_3 ... T_N

(3) Use for zero-shot prediction

Image Encoder

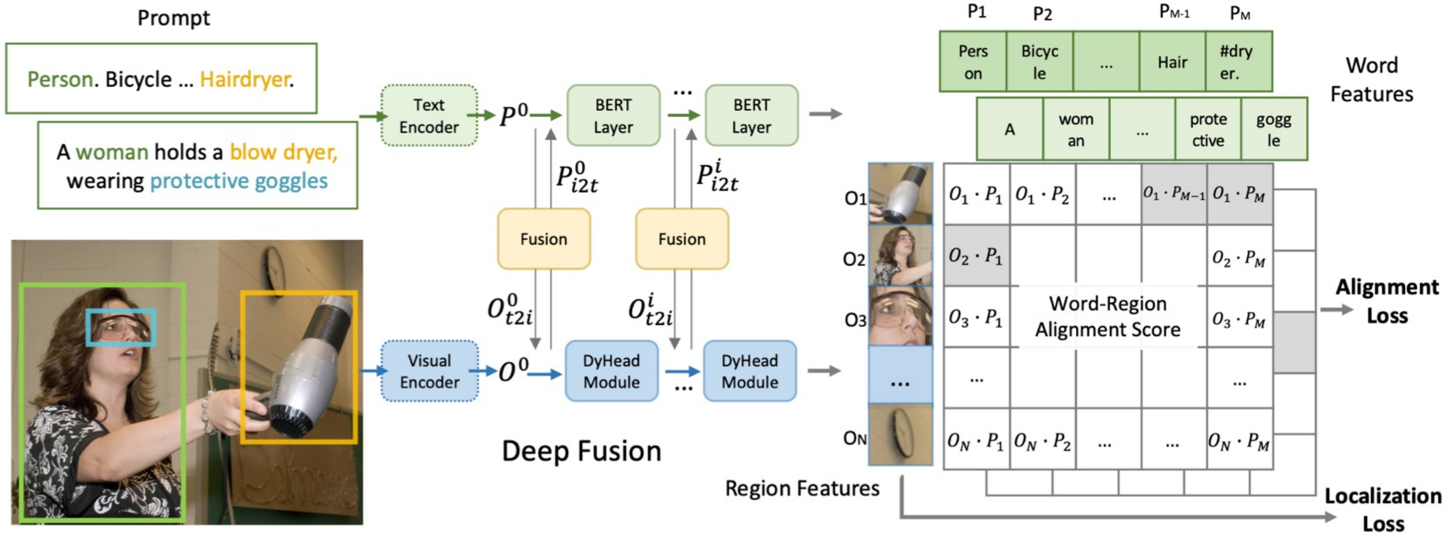
I_1

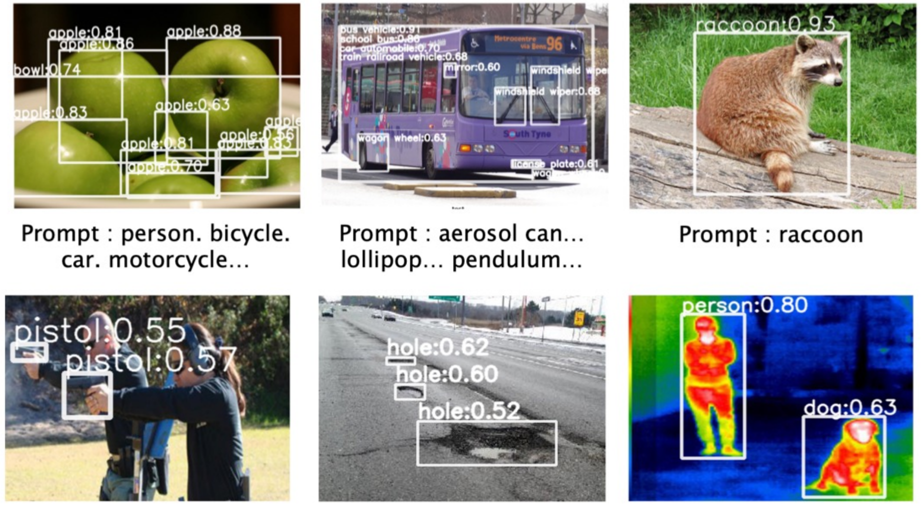
$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$

A photo of a dog.

https://github.com/openai/CLIP

GLIP





Prompt : person. bicycle. car. motorcycle…

Prompt : aerosol can… lollipop… pendulum…

Prompt : raccoon

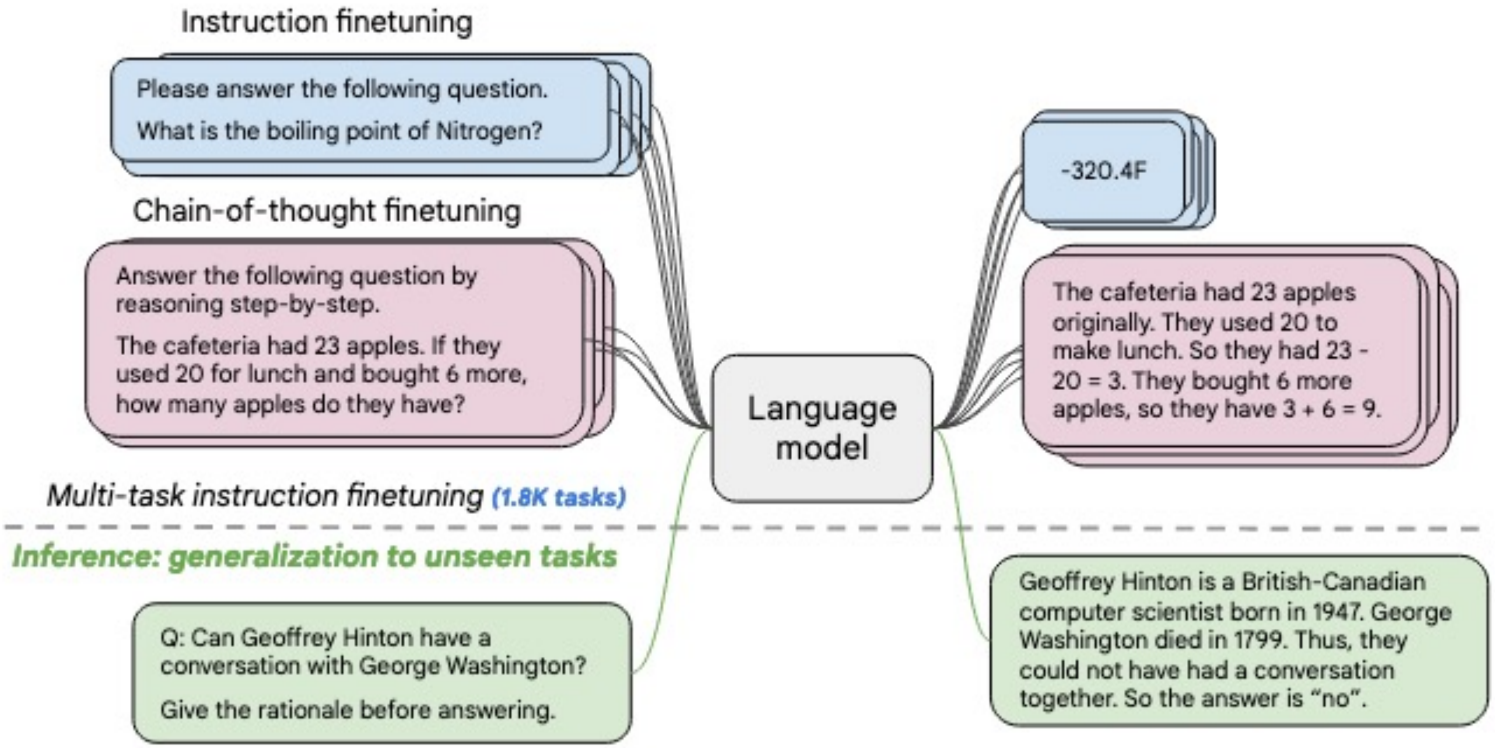
Prompt : pistol

Prompt : there are some holes on the road

Prompt : person. dog.

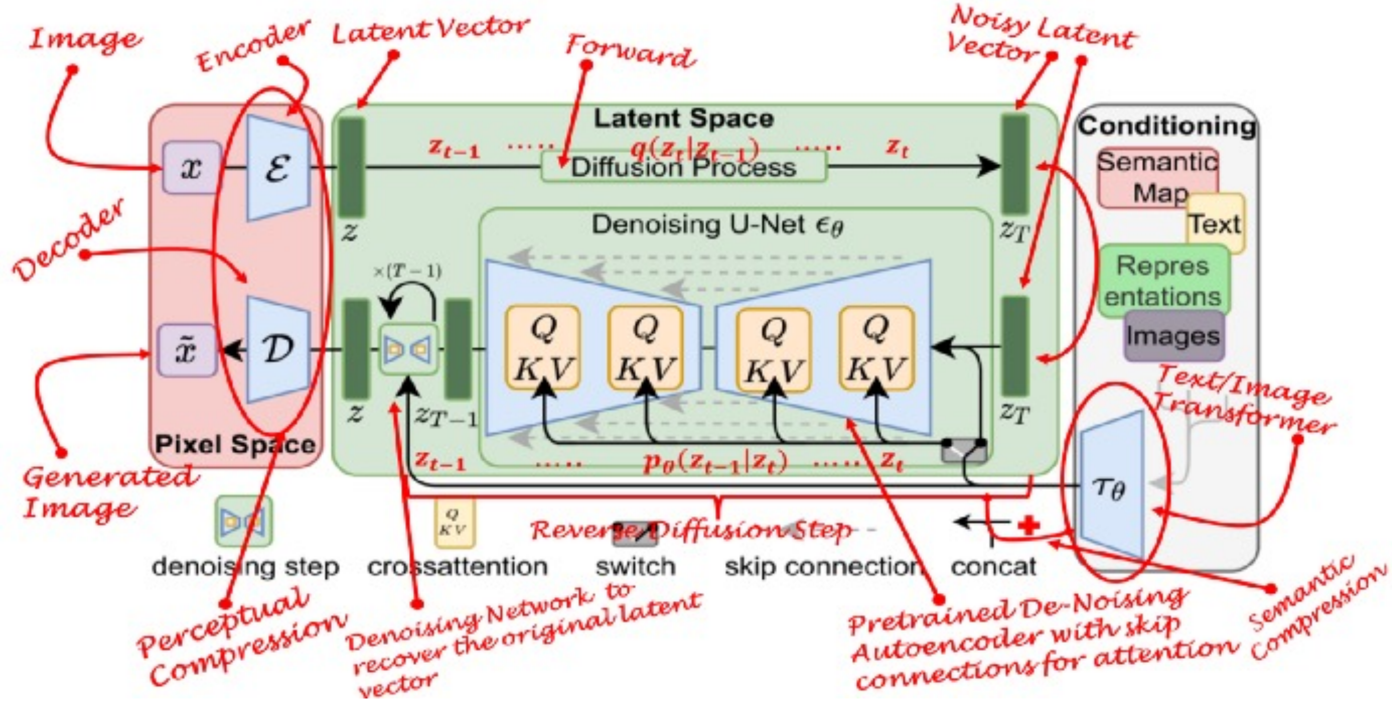
https://github.com/microsoft/GLIP

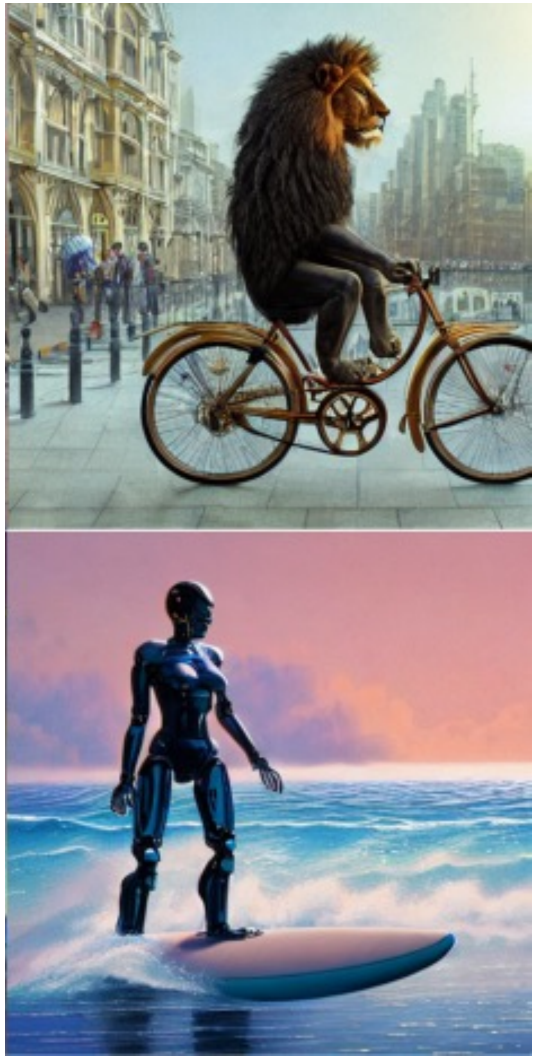
FLAN T5


Instruction finetuning
Please answer the following question.
What is the boiling point of Nitrogen?
Chain-of-thought finetuning
Answer the following question by reasoning step-by-step.
The cafeteria had 23 apples. If they used 20 for lunch and bought 6 more, how many apples do they have?
Multi-task instruction finetuning (1.8K tasks)
Inference: generalization to unseen tasks
Q: Can Geoffrey Hinton have a conversation with George Washington?
Give the rationale before answering.
Language model
-320.4F
The cafeteria had 23 apples originally. They used 20 to make lunch. So they had 23 - 20 = 3. They bought 6 more apples, so they have 3 + 6 = 9.
Geoffrey Hinton is a British-Canadian computer scientist born in 1947. George Washington died in 1799. Thus, they could not have had a conversation together. So the answer is "no".

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

Stable Diffusion v2

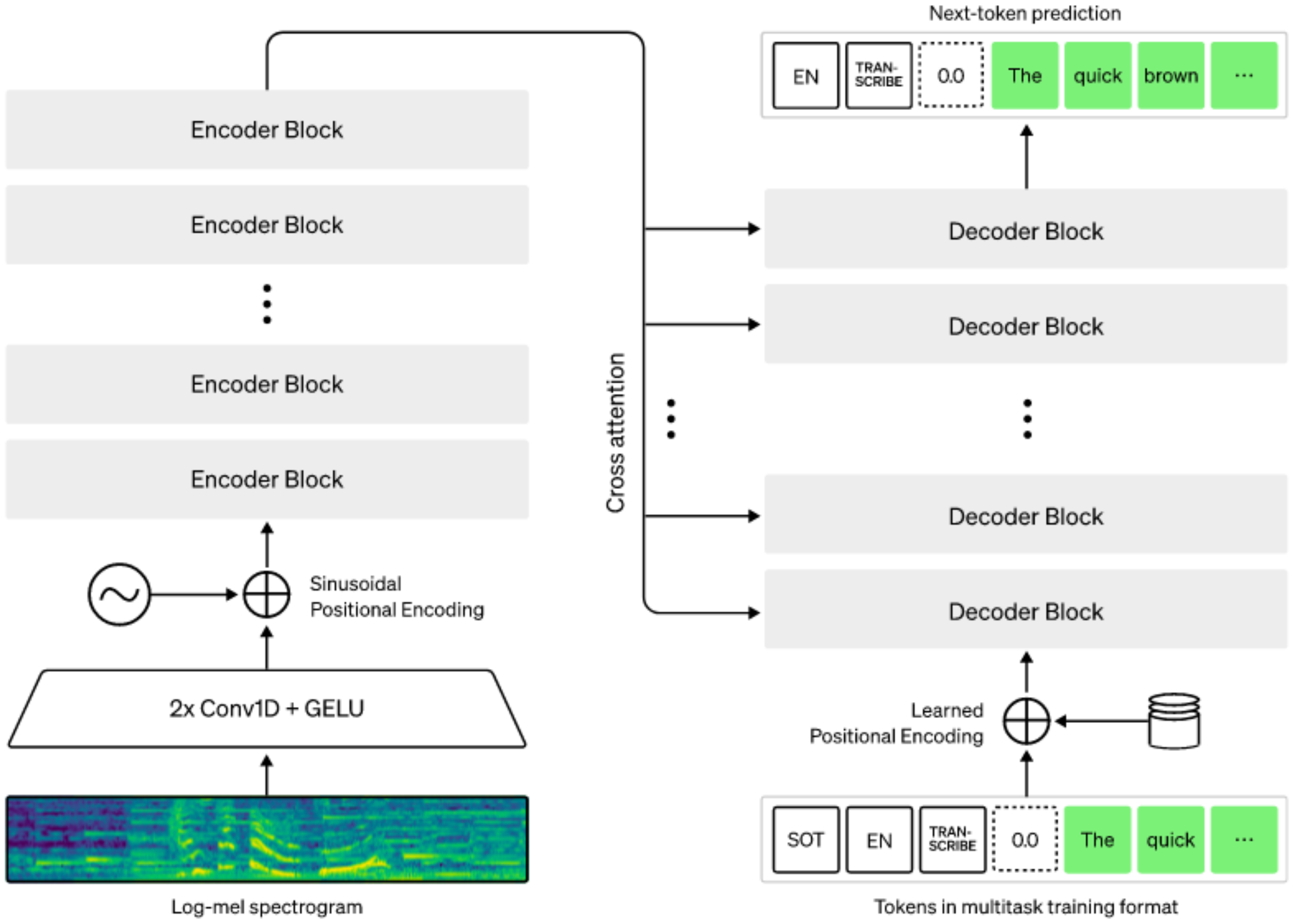




https://towardsdatascience.com/what-are-stable-diffusion-models-and-why-are-they-a-step-forward-for-image-generation-aa1182801d46

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

Whisper



https://github.com/openai/whisper

We will be using

PyTorch

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

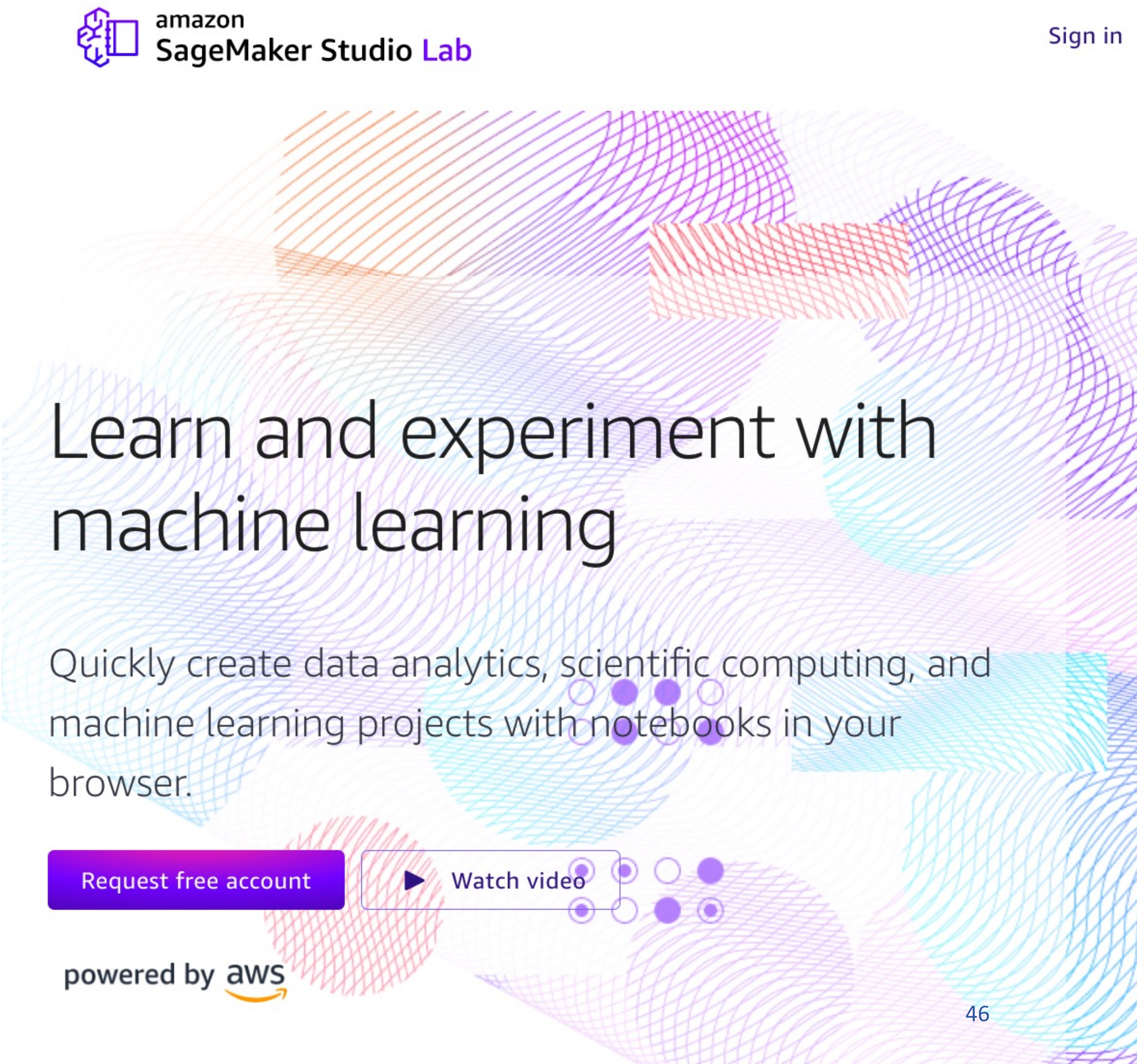
We will also be using…

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

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

Also try using:



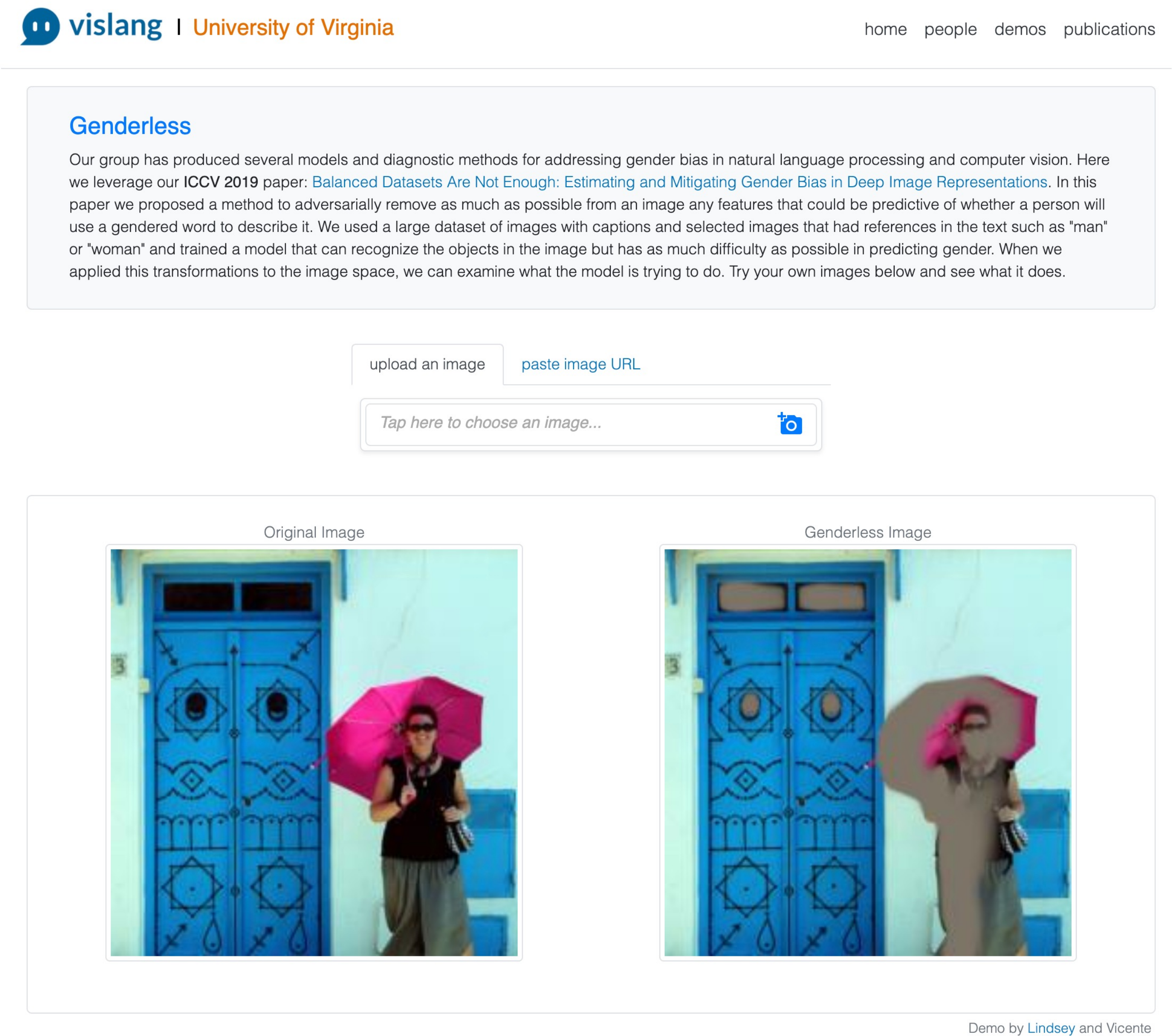
Learn and experiment with machine learning

Quickly create data analytics, scientific computing, and machine learning projects with notebooks in your browser.

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

Demos

https://vislang.ai/genderless



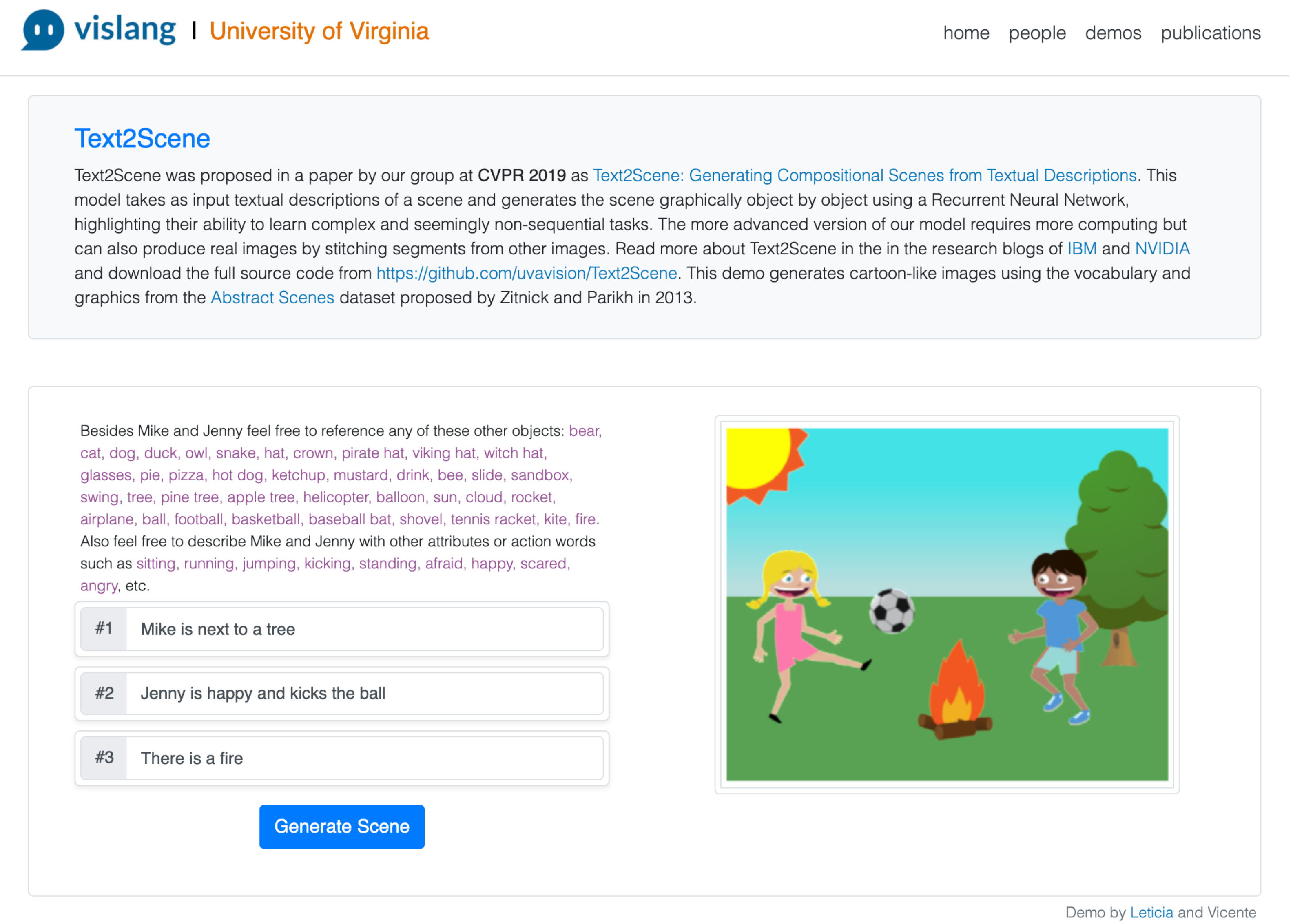
Genderless

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.

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/uvavision/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

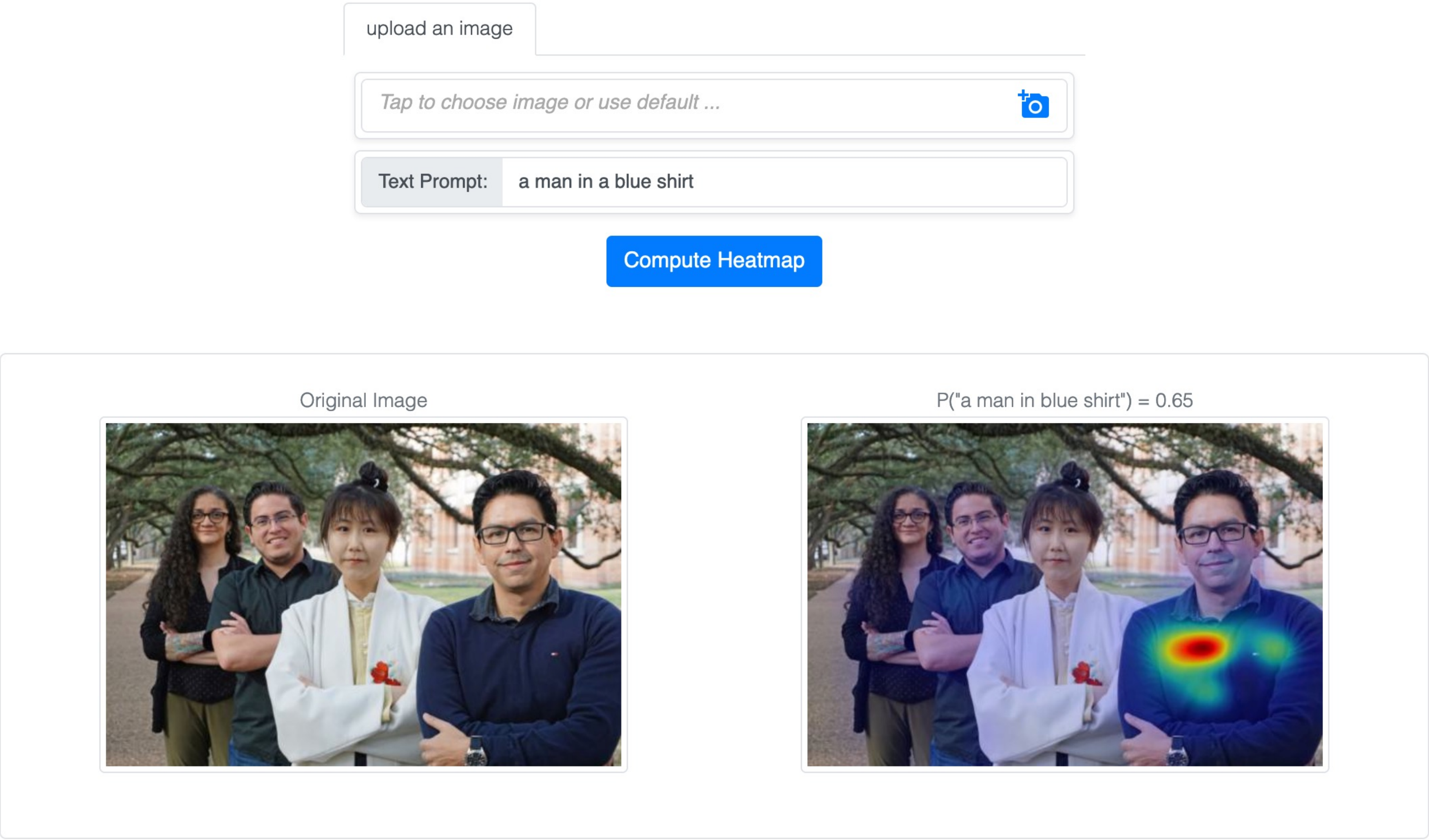
#2 Jenny is happy and kicks the ball

#3 There is a fire

Demo by Leticia and Vicente

AMC Visual Grounding

https://vislang.ai/amc



Text Prompt: a man in a blue shirt

P("a man in blue shirt") = 0.65

Demo by Ziyan 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?