CAP6412, Spring 2024

• **Instructor**: Dr. Mubarak Shah
• **Email**: shah@crcv.ucf.edu
• **Office**: HEC 245
• **Phone**: 4078235077
• **Co-Instructor**: Dr. Kumar Nayak <GauravKumar.Nayak@ucf.edu>
• **GTA**: Swetha Sirnam <Swetha.Sirnam@ucf.edu>
• **Time**: Mondays and Wednesdays 3:00 to 4:15PM
• **Location**: BA1 O213/Zoom
• **Office Hours**: Mondays 2:00 to 3:00PM; Wednesdays 4:15 to 5:00PM; Fridays 1:00 to 2:00; and by appointment
• **Pre-requisite**: CAP5415
• **Course webpage**: [https://www.crcv.ucf.edu/courses/cap6412-spring-2024/](https://www.crcv.ucf.edu/courses/cap6412-spring-2024/)
Course Objective

• To expose graduate students to the cutting-edge research in Computer Vision.

  • We will discuss research papers on Visual-Language Models (VLM).
Students Learning Outcomes

• Read and understand a research paper.
• Write a comprehensive review of the paper.
• To identify strong and weak points of the paper.
• To generate own ideas to solve the same problem.
• To work on research project and write a research paper.
Grading Policy

Reports (individually, you have to do only 50% of the papers) 10%
Replications of papers (5 papers) 15%
Paper Presentations (by a group, roughly two) 25%
Attendance (students in virtual section, pleas turn on your video, be on time) 5%
Group Project 45%

Each group member will be evaluated by others; final grade on presentations and project will be scaled accordingly
No Mid Term, Final, Homework

• Late Policy
  • 0 for late reports
  • Projects
    • 20% off per day
    • up to 4 days
Project Timeline

• Three in-class presentations (see class schedule)
  • Project ideas / proposal [2%]
  • Update 1 [5%]
  • Update 2 [8%]
  • Final presentation/Paper [30%]
Programming/ GPU Cluster

• Pytorch, TensorFlow
• Each student will get an account on UCF HPC Newton
• Watch video: https://www.youtube.com/watch?v=niQ5hvABvKg&list=PLd3hlSJsXlkmm5i1HgmDB_z62BeoikFX&index=19
Papers/ Reports

- We will discuss one paper in each class
- All students will read the assigned paper before the class and write a report
- One group of students will make presentation and all students will participate in discussion
- Presentation should not be more than 30 minutes
- You can select the paper you want to present from the list on the course webpage
- Reports will be due just before the class meeting through Web Courses

Schedule Table: [https://www.crcv.ucf.edu/courses/cap6412-spring-2023/schedule/](https://www.crcv.ucf.edu/courses/cap6412-spring-2023/schedule/)
Reports (one page)

• Parts of Report
  • Very short Summary
  • Good points
  • Weak points
  • Questions
  • Ideas

• Reports will not be graded

• We will follow the honor system, by submitting report you will pledge that you have read the paper, have written report yourself, have not copied word by word from the paper and from any other student or leverage resources like ChatGPT etc
Statement of Academic Integrity

• The UCF Golden Rule (http://goldenrule.sdes.ucf.edu/) will be observed in the class. Plagiarism and Cheating of any kind on an examination, quiz, or assignment will result at least in an "F" for that assignment (and may, depending on the severity of the case, lead to an "F" for the entire course) and may be subject to appropriate referral to the Office of Student Conduct for further action. I will assume for this course that you will adhere to the academic creed of this University and will maintain the highest standards of academic integrity. In other words, don't cheat by giving answers to others or taking them from anyone else. I will also adhere to the highest standards of academic integrity, so please do not ask me to change (or expect me to change) your grade illegitimately or to bend or break rules for one person that will not apply to everyone.
Paper presentation review/Rehearsal Schedule

• For Monday presentation
  • Slide Review: Wednesday 4:15 a week before the scheduled presentation
  • Rehearsal: Friday a week before the scheduled presentation 1:00PM during Office hours

• For Wednesday presentation
  • Slide Review: A week before the scheduled presentation: Friday 1:00PM during Office hours
  • Rehearsal: A week of presentation on Monday 2:00PM during Office hours
Textbook

There is no textbook for this class. We will discuss recent research papers.

A good deep learning textbook:

  https://www.deeplearningbook.org/
• Alex Smola, Introduction to Machine Learning
  https://alex.smola.org/drafts/thebook.pdf
Other Resources

• Survey paper

• https://huggingface.co/blog/vision_language_pretraining
• https://medium.com/@navendubrajesh/vision-language-models-an-introduction-37853f535415
• https://www.kaggle.com/code/jhoward/getting-started-with-llms/
• models/?utm_campaign=mv_pr_lvm&utm_medium=organicsocial&utm_source=linkedIn&utm_content=pr_lvm
• https://www.promptingguide.ai/techniques/rag
• https://huggingface.co/blog/vision_language_pretraining
Research Paper

• Each paper has following parts
  1. Title
  2. Abstract
  3. Introduction
  4. Rest of the paper
    • Related Work
    • Method
    • Results
    • Conclusion

• Each part is equally important (25% each!)
How to read a research paper?

• You must read the paper several times to understand it.
  • When you read the paper first time,
    • if you do not understand something do not get stuck,
    • keep reading assuming you will figure out that later.
  • When you read it the second time, you will understand much more, and the third time even more ...

• Read the abstract first then look at the figures with captions and then conclusion
How to read a research paper?

• Try first to get a general idea of the paper
  • What problem is being solved?
  • What are the main steps?
  • How can I implement the method?,
  • Even though I do not understand why each step is performed the way it is performed

• Try to relate the method to other methods you know, and conceptually find similarities and differences.
How to read a research paper?

• In the first reading it may be a good idea to skip the related work.

• Do not use dictionary to just look up the meaning of technical terms.

• Try to understand each concept in isolation, and then integrate them to understand the whole paper.
Useful Blogs about how to read a paper?

• [https://web.stanford.edu/class/ee384m/Handouts/HowtoReadPaper.pdf](https://web.stanford.edu/class/ee384m/Handouts/HowtoReadPaper.pdf)

• [https://blogs.lse.ac.uk/impactofsocialsciences/2016/05/09/how-to-read-and-understand-a-scientific-paper-a-guide-for-non-scientists/](https://blogs.lse.ac.uk/impactofsocialsciences/2016/05/09/how-to-read-and-understand-a-scientific-paper-a-guide-for-non-scientists/)
Visual-Language Models: Short Introduction

• Material from
  • A Dive into Vision-Language Models:  
    https://huggingface.co/blog/vision_language_pretraining
  • Beginner’s Guide to Large Language Models | by Digitate | Medium  
    https://medium.com/@igniobydigitate/a-beginners-guide-to-large-language-models-e5e9e63d84a
  • Introduction to Visual-Language Model | by Navendu Brajesh | Medium  
    https://medium.com/@navendubrajesh/vision-language-models-an-introduction-37853f535415
  • https://huggingface.co/blog/vision_language_pretraining
Computer Vision Tasks

- Object Classification
- Object Detection
- Object Segmentation
- Instant Segmentation
- Object Retrieval
- Semantic Segmentation
- Action Classification
- Object Tracking
- ....
Limitation

• Computer Vision techniques output images, bounding boxes, classes,..

• They don’t communicate through text

• Humans are good at communicating with language and text
Computer Vision Tasks requiring language

- Images
  - Image Captioning
  - Visual Question & Answering
  - Image-to-Text Retrieval
  - Text-to-Image Retrieval
  - Text-guided image generation

- Video
  - Video Captioning
  - Video Q&A
  - Video-text Retrieval
  - ...
Image Captioning

O Vinyals et al. *Show and Tell: A Neural Image Caption Generator*, 2014
Text to Video Retrieval

“Animated comic scene of guy cutting up food for dinner”

“a woman holding a ribbon”

“mix ingredients refrigerate”

“add mutton pan”

Ours

EAO [11]

Sirnam, Swetha; Rizve, Mamshad Nayeem; Kuhne, Hilde; Shah, Mubarak
Preserving Modality Structure Improves Multi-Modal Learning, ICCV, 2023
<table>
<thead>
<tr>
<th>Text-to-Image Synthesis on LAION</th>
</tr>
</thead>
<tbody>
<tr>
<td>'A street sign that reads &quot;Latent Diffusion&quot;'</td>
</tr>
<tr>
<td>'A zombie in the style of Picasso'</td>
</tr>
<tr>
<td>'An image of an animal half mouse half octopus'</td>
</tr>
<tr>
<td>'An illustration of a slightly conscious neural network'</td>
</tr>
<tr>
<td>'A watercolor painting of a chair that looks like an octopus'</td>
</tr>
<tr>
<td>'A shirt with the inscription: &quot;I love generative models!&quot;'</td>
</tr>
</tbody>
</table>
Text-to-Image

Natural Language Processing (NLP)

- Search engines
- Spam filtering
- Machine translation
- Sentiment analysis
- Document summarization
- .....
Natural Language Processing (NLP)

• Limitations
  • Not able to decode visual cues

• Linguistic Ambiguities

• Verifying interpretations against real-world visual references
Natural Language Processing (NLP)

• Limitations
  • Not able to decode visual cues
  • Linguistic Ambiguities
  • Visual Interpretations

• Exhibit Flair of text analytics and generation they fall short in decoding visual cues.
• grapple with linguistic ambiguities and are handicapped when it comes to verifying their interpretations against real-world visual references,
Language Models

• Understand and Generate text
• Learn from raw text
• Transformer architecture

Source: Beginner’s Guide to Large Language Models | by Digitate | Medium
Large Language Models

• Pre-trained on large datasets

• They have large number of parameters
LLM Datasets

• Common Crawl consists of ~60% of training data.

• WebText2 (Open AI, from Reddit) consists of ~22% of training data.

• Books1 consists of ~8% of training data.

• Books2 consists of ~8% of training data.

• Wikipedia consists of ~3% of training data.

Source: Beginner’s Guide to Large Language Models | by Digitate | Medium
LLMs

- **LaMDA**: Developed by Google, trained on 1.56 trillion words of public dialog data. It powers the BARD chatbot!
- **LLaMA**: Developed by Meta, a relatively small model (7B parameters) yet accurate as compared to GPT3.
- **BLOOM**: Open source and multilingual model, trained data from 46 natural languages and 13 programming languages.
- **Galactica**: Developed by Meta, can store, combine, and reason about scientific knowledge.
- **Codex**: Model that powers GitHub Copilot. Proficient in more than a dozen programming languages, Codex can now interpret simple commands in natural language and execute them.
- **PaLM-E**: Developed by Google, a LLM focused on robot sensor data.
- **Chinchilla**: Developed by Deepmind, considerably simplifies downstream utilization because it requires much less computer power for inference and fine-tuning.

Source: Beginner’s Guide to Large Language Models | by Digitate | Medium
GPT (Generative Pre-trained Transformer)

Source: Beginner’s Guide to Large Language Models | by Digitate | Medium
Pre-Training

• Self-Supervised

• Auto-regressive

• Unidirectional

• It understands the relationship between various words in the given context

Source: Beginner’s Guide to Large Language Models | by Digitate | Medium
Limitations of LLMs

• LLMs are large

• LLMs are Black box

• LLMs can have bias

• LLMs can do hallucinations

• LLMs may have IP issue

Source: Beginner’s Guide to Large Language Models | by Digitate | Medium
Applications of LLMs

• Code generation
• Content generation tools
• Copywriting
• Conversational tools
• Educational tools
• Enterprise search
• Information retrieval

Source: Beginner’s Guide to Large Language Models | by Digitate | Medium
Visual Language Models

• Vision systems are fundamental to understanding our world
• However, humans are good in communicating through language
• Complex relations between objects and their locations can be better described in human language (text)
• Visual-Language models bridge the gap between vision and language
• VLMs understand both images and text
• The output of VLM can be modified through human-provided prompts, e.g.,
  • segmenting a particular object by providing a bounding box,
  • having interactive dialogues by asking questions about an image or video scene
  • manipulating the robot’s behavior through language instructions
Visual Language Models

• **Vision systems** are fundamental to understanding our world
• The complex relations between objects and their locations can be better described in human language (text)
• **Visual-Language models** bridge the gap between vision and language
• The output of VLM can be modified through human-provided prompts, e.g.,
  • segmenting a particular object by providing a **bounding box**, 
  • having interactive dialogues by asking **questions** about an image or video scene 
  • manipulating the robot’s behavior through **language instructions**
• Model that understand both image and text
Language and Senses

• Humans are the only known species where much of knowledge learning happens symbolically through language.

• In addition to information received directly from Five senses.
  • Vision
  • Hearing
  • Touch
  • Taste
  • Smell

Large Multi-model Models (LMMs)

- Image
- Video
- Text
- Audio (speech, music)
- Physiological signals
- ....
Visual-Language Tasks

- Image retrieval from natural language text
- Phrase grounding, i.e., performing object detection from an input image and natural text (example: A **young person** swings a **bat**).
- Visual question answering, i.e., finding answers from an input image and a question in natural language
- Generate a caption for a given image
- Detection of hate speech from social media content involving both images and text modalities
- Visual-Language Navigation

Credit: A Dive into Vision-Language Models
https://huggingface.co/blog/vision_language_pretraining
Contrastive Learning Image Pre-training (CLIP)

Input Image

Image Representation

Input Text

Text Representation

\[
\text{maximize} \left( \frac{\vec{H}_i \cdot \vec{H}_t}{||\vec{H}_i|| \times ||\vec{H}_t||} \right)
\]

A dog lying in grass

Learning Transferable Visual Models From Natural Language Supervision
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever
Contrastive Learning Image Pre-training (CLIP)

Learning Transferable Visual Models From Natural Language Supervision
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever
Contrastive Language Image Pre-training (CLIP)
Contrastive Language Image Pre-training (CLIP)
### Example of GPT-4 visual input:

**User:** What is funny about this image? Describe it panel by panel.

**GPT-4:**

The image shows a package for a "Lightning Cable" adapter with three panels.

**Panel 1:** A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

**Panel 2:** The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

**Panel 3:** A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

Source: [https://cdn.openai.com/papers/gpt-4.pdf](https://cdn.openai.com/papers/gpt-4.pdf)
Mini-GPT4: Enhancing Vision-Language Understanding with Advanced Large Language Models, Deyao Zhu, Jun Chen, Xiaqian Shen, Xiang Li, Mohamed Elhoseiny
Visual Instruction Tuning
Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee
In the image, a young woman is leaning on a man in front of a beautiful sunset, with both of them standing on a boat. They appear to be enjoying a romantic moment together. The scene suggests that they could be depicted as the characters Jack and Rose from the popular movie Titanic. In the movie, their romance blossoms aboard the fictional ocean liner Titanic, with the pair ultimately sharing a passionate kiss as the ship sinks into the ocean. The ending of the movie is known for its emotional and heartfelt scenes, where the characters are forced to say goodbye to each other as they are separated by the sinking ship. The image evokes a sense of romance and nostalgia, paying tribute to the iconic love story portrayed in the movie Titanic.

Visual Instruction Tuning
Haotian Liu, Chunyuan Li, Qingyang Wu, Yong Jae Lee
The video is funny because it shows a baby girl who is learning to read. She is sitting on a bed and looking at a book, and she is trying to read it. The baby girl is wearing glasses, and she is holding the book with both hands. The baby girl is trying to read the book, and her actions are adorable and endearing. The video captures the innocence and curiosity of a young child as she learns to read, which is a joyful and amusing sight for the viewer.
Video ChatGPT

What is this video about?
The video is about a woman applying sunscreen on her face and body.

Generate a creative advertisement feature about this product that can be used on the product website.
“Say goodbye to sunburn and hello to a glowing, healthy complexion with our new sunscreen product. With its lightweight formula and SPF 50+ protection, it’s perfect for everyday use. Apply it daily to your face and neck for a long-lasting barrier against harmful UV rays. Don’t let sun damage ruin your skin - choose our sunscreen and enjoy the benefits of a healthy, radiant complexion. Get your glow on with our sunscreen today!”

Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models
Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan
1/10/2024
The strange aspect of this video is that the lions are seen playing with a man in the wild. It is not common to see lions interacting with humans in the wild, as they are wild animals and can be unpredictable. The man is seen petting the lions and playing with them, which is quite unusual and could potentially be dangerous for both the man and the lions.
PG-Video-LLaVA: Pixel Grounding Large Video-Language Models
Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, Fahad Khan
Can you suggest a title for this video clip based on what you see and hear?

**Without Audio**
Based on the video, I would suggest "The Future of Technology and Communication" as a title for the video clip.

**With Audio**
Based on the video, the title could be "The Future of AI: A Discussion on GPT-4 and Chat GPT."

**Audio Transcript:** ... just how amazing Chat GPT turn out to be and then the improvement in GPT for. So GPT 4 has not surprised me at all in terms of reception there. Chat GPT surprised us a little bit, but I still ....... You know, they're like, oh, it's better than 3.5, but I thought it was going to be better than 3.5. ...
Thankyou