Deep Learning for Computer Vision

Vision-Language Models: Introduction and History

Vineeth N Balasubramanian

Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad

भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Computer Vision: Tasks and Limitations

Computer Vision Tasks

- Object Classification
- Object Detection
- Instance Segmentation
- Semantic Segmentation

Computer Vision: Tasks and Limitations

Computer Vision Tasks

- Object Classification
- Object Detection
- Instance Segmentation
- Semantic Segmentation

Limitations

Computer vision techniques output class labels, bounding boxes, masks, images, etc. However, humans communicate through language and text, which vision models lack.

Tying in Language into Computer Vision Tasks

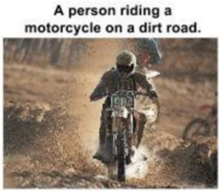

A person riding a motorcycle on a dirt road.

Figure 1: Image Captioning[a]

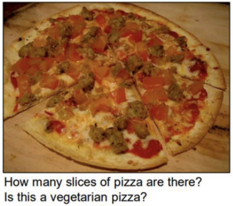

How many slices of pizza are there?
Is this a vegetarian pizza?

Figure 2: Visual Question Answering[a]

---

[a] Vinyals et al. "Show and tell: A neural image caption generator", CVPR 2014

[a] Agrawal et al, "VQA: Visual Question Answering", IJCV 2015

Tying in Language into Computer Vision Tasks

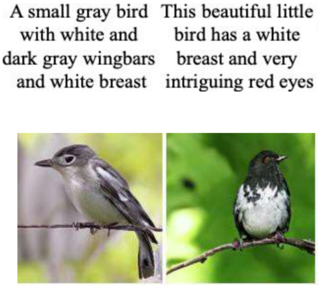

A small gray bird with white and dark gray wingbars and white breast
This beautiful little bird has a white breast and very intriguing red eyes

Figure 3: Text-to-image generation[a]

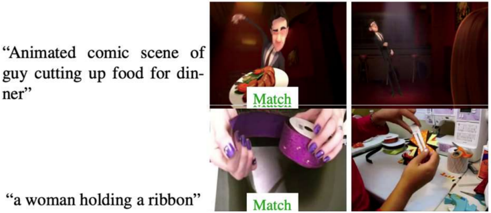

"Animated comic scene of guy cutting up food for dinner"
Match
"a woman holding a ribbon"
Match

Figure 4: Text-to-Video retrieval[a]

---

[a] Gu et al, "Vector quantized diffusion model for text-to-image synthesis", CVPR 2022

[a] Sirnam et al, "Preserving Modality Structure Improves Multi-Modal Learning", ICCV 2023

Natural Language Processing (NLP): Tasks and Limitations

NLP Tasks

- Search engines
- Spam filtering
- Machine translation
- Sentiment analysis

Natural Language Processing (NLP): Tasks and Limitations

NLP Tasks

- Search engines
- Spam filtering
- Machine translation
- Sentiment analysis

Limitations

Exhibit flair in analyzing and generating text; however, cannot process visual cues or verify interpretations against real-world visual references, especially when there are linguistic ambiguities

From Language Models to LLMs

Language Models

- Understand and generate text
- Based on Transformer architecture
- Learn from raw text

From Language Models to LLMs

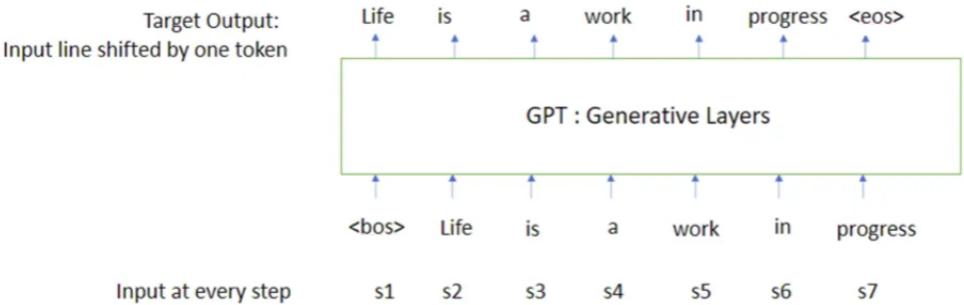
Language Models

- Understand and generate text
- Based on Transformer architecture
- Learn from raw text

Large Language Models (LLMs)

- Pre-trained on large datasets
- Large number of parameters
- Datasets used to train are: Common Crawl (60%), WebText2 (22%), Books1 (8%), Books2 (8%), Wikipedia (3%)

LLMs: Generative Pretrained Transformer (GPT) [1]


Target Output:
Input line shifted by one token
Life is a work in progress <eos>
GPT : Generative Layers
<bos> Life is a work in progress
Input at every step
s1 s2 s3 s4 s5 s6 s7

---

[1] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

GPT Training[2]

Learning via Self-supervision

The model learns from raw sentences with a target sequence shifted by one token, enabling it to grasp word relationships for accurate output prediction.

---

[2] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

GPT Training[2]

Learning via Self-supervision

The model learns from raw sentences with a target sequence shifted by one token, enabling it to grasp word relationships for accurate output prediction.

Auto-regressive

Words derive context from all preceding words. Each generated token is added to the input sequence, forming the input for the next step.

---

[2] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

GPT Training[2]

Learning via Self-supervision

The model learns from raw sentences with a target sequence shifted by one token, enabling it to grasp word relationships for accurate output prediction.

Auto-regressive

Words derive context from all preceding words. Each generated token is added to the input sequence, forming the input for the next step.

Unidirectional

The GPT model learns context strictly from left to right (earlier models like BERT used a bidirectional approach, which considers context from both directions)

---

[2] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

Other LLMs[3]

- **LaMDA**: Developed by google, trained on 1.56 trillion words of public dialog data. It powered the BARD chatbot, and a lightversion of it led to the Gemini!

---

[3] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

Other LLMs[3]

- **LaMDA**: Developed by google, trained on 1.56 trillion words of public dialog data. It powered the BARD chatbot, and a lightversion of it led to the Gemini!
- **LLaMA**: Developed by Meta, a relatively small model (7B parameters) yet accurate as compared to GPT3

---

[3] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

Other LLMs[3]

- **LaMDA**: Developed by google, trained on 1.56 trillion words of public dialog data. It powered the BARD chatbot, and a lightversion of it led to the Gemini!
- **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

---

[3] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

Other LLMs[3]

- **LaMDA**: Developed by google, trained on 1.56 trillion words of public dialog data. It powered the BARD chatbot, and a lightversion of it led to the Gemini!
- **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

---

[3] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

Other LLMs[3]

- **LaMDA**: Developed by google, trained on 1.56 trillion words of public dialog data. It powered the BARD chatbot, and a lightversion of it led to the Gemini!
- **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

---

[3] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

Other LLMs[3]

- **LaMDA**: Developed by google, trained on 1.56 trillion words of public dialog data. It powered the BARD chatbot, and a lightversion of it led to the Gemini!
- **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

---

[3] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

Other LLMs[3]

- **LaMDA**: Developed by google, trained on 1.56 trillion words of public dialog data. It powered the BARD chatbot, and a lightversion of it led to the Gemini!
- **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

---

[3] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

LLMs: Applications and Limitations[4]

Applications

- Code generation
- Content generation tools
- Copywriting
- Conversational tools
- Educational tools

---

[4] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

LLMs: Applications and Limitations[4]

Applications

- Code generation
- Content generation tools
- Copywriting
- Conversational tools
- Educational tools

Limitations

- LLMs are large, and compute-intensive to train
- LLMs can have bias
- LLMs can hallucinate

---

[4] *Source: Beginner's Guide to Large Language Models — by Digitate — Medium*

Vision-Language Models[5]

- Vision systems are fundamental to understanding our world, however, lack the ability to communicate with humans naturally

---

[5] Awais et al. "Foundational Models Defining a New Era in Vision: A Survey and Outlook", arXiv 2023

Vision-Language Models[5]

- Vision systems are fundamental to understanding our world, however, lack the ability to communicate with humans naturally
- Complex relations between objects and their locations can be better described in human language (text)

---

[5] Awais et al. "Foundational Models Defining a New Era in Vision: A Survey and Outlook", arXiv 2023

Vision-Language Models[5]

- Vision systems are fundamental to understanding our world, however, lack the ability to communicate with humans naturally
- Complex relations between objects and their locations can be better described in human language (text)
- Vision-Language models (VLMs) bridge the gap between vision and language, understanding both

---

[5] Awais et al. “Foundational Models Defining a New Era in Vision: A Survey and Outlook”, arXiv 2023

Vision-Language Models[5]

- Vision systems are fundamental to understanding our world, however, lack the ability to communicate with humans naturally
- Complex relations between objects and their locations can be better described in human language (text)
- Vision-Language models (VLMs) bridge the gap between vision and language, understanding both
- The output of a 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

---

[5] Awais et al. "Foundational Models Defining a New Era in Vision: A Survey and Outlook", arXiv 2023

Vision-Language Tasks

- Image retrieval from natural language text

Vision-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)

Vision-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

Vision-Language Tasks

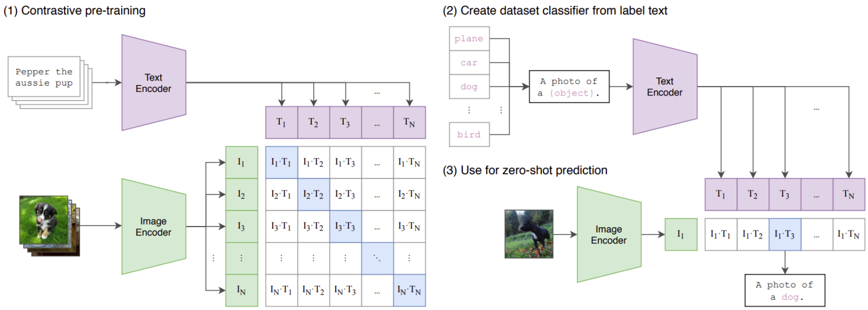
- Image retrieval from natural language text
- Phrase grounding, i.e., performing object detection from an input image and natural text (example: A young person swings a bat)
- Visual question answering, i.e., finding answers from an input image and a question in natural language
- Caption generation for a given image

Vision-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
- Caption generation for a given image
- Detection of hate speech from social media content involving both images and text modalities

Glimpse of Topics in this Module

Contrastive Language Image Pre-training (CLIP) - the pivot![6]

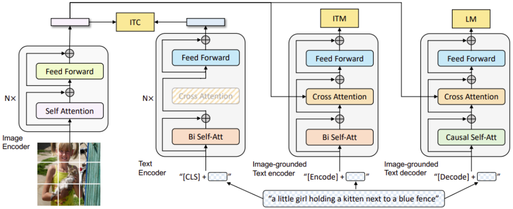

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ⋮ I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
T_1 T_2 T_3 ... T_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

---

[6] Radford et al, "Learning Transferable Visual Models From Natural Language Supervision", ICML 2021

Glimpse of Topics in this Module

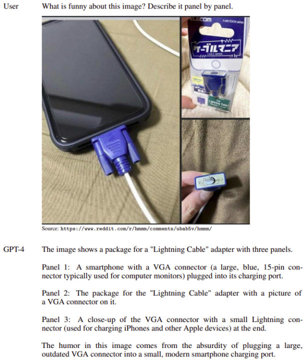
Bootstrapping Language-Image Pretraining (BLIP)[a]



---

[a] Li "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation", ICML 2022

GPT-4[a]



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

Source: https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/

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.

---

[a] Achiam et al, "GPT-4 Technical Report", OpenAI 2023