Gridspace

GRIDSPACE IAP 2024 LECTURE 6
TOOLS FOR LLM PERCEPTION

Phoebe Piercy, MIT ‘20

January 25, 2024

Exercises:

- What are some aspects of NLP that is human-like? What are aspects that aren't?
    - Similarities: Semantic and context understanding
    - Differences: Limited world knowledge, lack of common sense
- How do humans combine different sensory information? How do ML models do so?
    - Humans: Cross-Modal Pathways and Regions (superior colliculus), Temporal and Spatial Coincidence
    - Models: Multimodal models, transfer learning, feature fusion

PLANNING

MEMORY

PERCEPTION

LANGUAGE &
SYMBOLIC
REASONING

PLANNING

MEMORY

LANGUAGE &
SYMBOLIC
REASONING

TODAY'S ROADMAP

- “Perception”
- Approach 1: Description
- Approach 2: Annotation/Categorization
- Approach 3: NL translation
- Approach 4: Multimodality
- Detour into State Tracking
- Exercises

TODAY'S ROADMAP

- “Perception”
- Approach 1: Description
- Approach 2: Annotation/Categorization
- Approach 3: NL translation
- Approach 4: Multimodality
- Detour into State Tracking
- Exercises

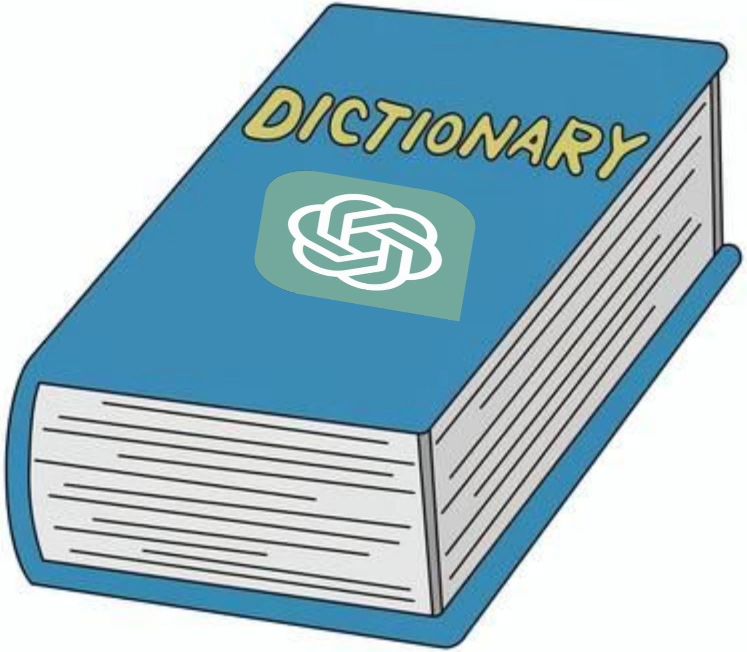
Perception

noun

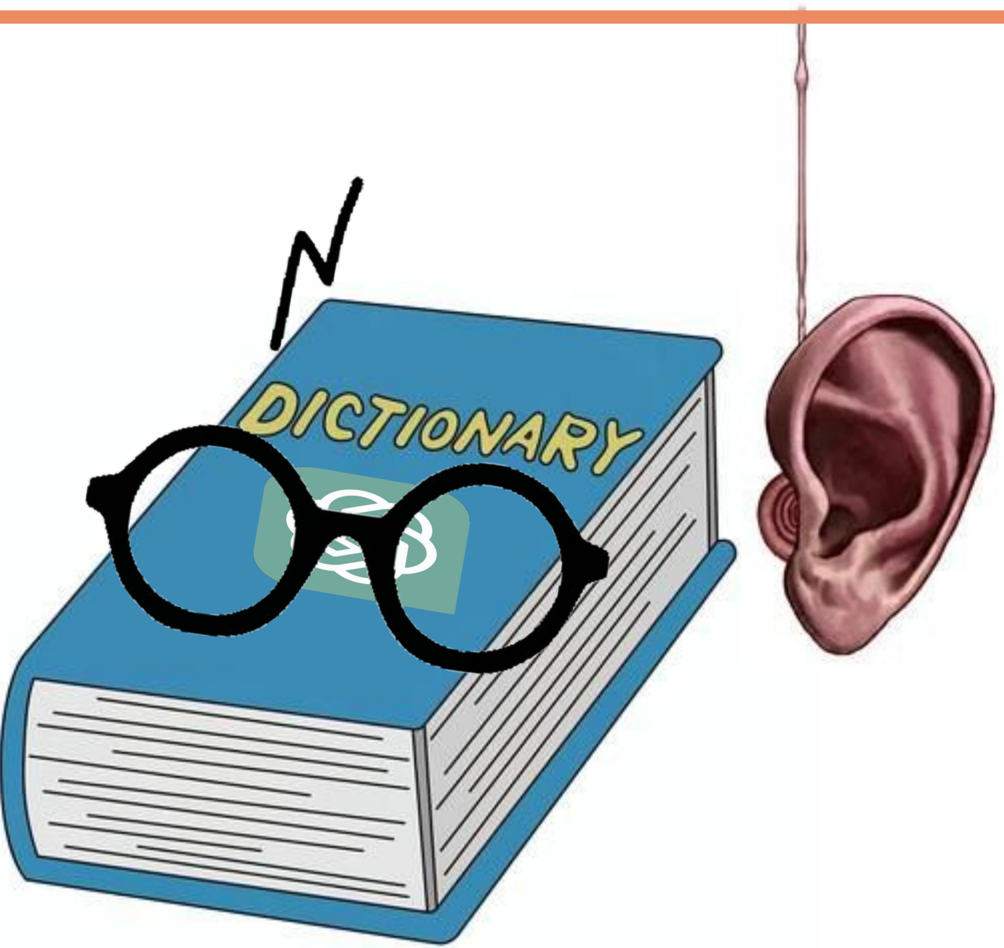
per·cep·tion /pərˈsepSH(ə)n/

1. ~~*the ability to see, hear, or become aware of something through the senses.*~~
2. *the organization, identification, and interpretation of sensory information*

Tools for LLM Perception



Tools for LLM Perception



Sensory Input

- **Sight, Sound** -> Cameras, light sensors, microphones

Sensory Input

- **Sight, Sound** -> Cameras, light sensors, microphones
- **Smell, Taste** -> Chemical spectrometers

Sensory Input

- **Sight, Sound** -> Cameras, light sensors, microphones
- **Smell, Taste** -> Chemical spectrometers
- **Touch** -> Haptic feedback

Sensory Input

- **Sight, Sound** -> Cameras, light sensors, microphones
- **Smell, Taste** -> Chemical spectrometers
- **Touch** -> Haptic feedback
- **Vestibular sensation** -> Spatial orientation and Balance - accelerometers, geomagnetic sensors, distance measurements

Sensory Input

- **Sight, Sound** -> Cameras, light sensors, microphones
- **Smell, Taste** -> Chemical spectrometers
- **Touch** -> Haptic feedback
- **Vestibular sensation** -> Spatial orientation and Balance - accelerometers, geomagnetic sensors, distance measurements
- **Proprioception** -> Position of self (bones, limbs etc) - e.g. motor speed, wheel load, robot arm joint angles, battery voltage

Sensory Input -> Perception

Raw Samples

Feature Sequence Representation
STFT [4][6]
MFCC [7][8]
Timbre, Rhythm and Pitch [1]
Chroma [9][18]
Pitch [5] [19]

Feature extraction/processing

"Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum."

Textual features and semantic grounding

perceived meaning, emotion, inference

TODAY'S ROADMAP

- "Perception"
- Approach 1: Description
- Approach 2: Annotation/Categorization
- Approach 3: NL translation
- Approach 4: Multimodality
- Detour into State Tracking
- Exercises

Tools for Physical Perception

Tools for Physical Perception

Tools for Physical Perception

Tools for Physical Perception

Write a response

Tools for Physical Perception

Write a response

"Of course! I'd be happy to help you. Could you please provide more details or context about what you'd like a response for?"

1. Provide "sound"

My mother said I have to wash the dishes. Write a response.

1. Provide “sound”

My mother said I have
to wash the dishes.
Write a response.

"Thanks for letting me know. I'll take care of the dishes as soon as I can.”

2. Provide tone

My mother said I have to wash the dishes in an angry tone. Write a response.

2. Provide tone

My mother said I have to wash the dishes in an angry tone. Write a response.

"Fine, I'll wash the dishes. It's not like I have anything better to do."

3. Provide sight

My mother said I have to wash the dishes in an angry tone, but the dishes already look clean. Write a response.

3. Provide sight

My mother said I have to wash the dishes in an angry tone, but the dishes already look clean. Write a response.

"Hey Mom, I noticed the dishes are already clean, but I'm happy to help with any other chores or tasks you may need assistance with. Just let me know what you'd like me to do!”

4. Provide context

My mother said I have to wash the dishes. She sounds angry but appears to be sad and upset. I feel extremely tired and my head is pounding. Looking around the kitchen I can see huge piles of dishes but the dishwasher is empty. The kitchen smells of curry. The calendar says its January 25th and the clock shows 6pm in the evening. I can hear my brother and father laughing in the room next door. Write a response to my mother.

"Hey Mom, I understand you're upset about the dishes, and I'm sorry if I've added to your stress. I'm feeling really tired, and my head is hurting. If it's okay, I can try to help out once I've rested a bit. I appreciate your understanding. Also, the kitchen smells like curry; did you cook something delicious?”

What are we actually doing here?

- Providing Natural Language that *describes* what the LLM should be able to perceive

Pros: - Control over all aspects of the perception. Plays to LLM strengths.
Cons: - Labor-heavy, reliant on our ability to describe *in words* things we are used to perceiving without using language. Often overly-verbose and unstructured/hard to fine tune around or reproduce responses. Very sensitive to prompt changes.

TODAY'S ROADMAP

- “Perception”
- Approach 1: Description
- Approach 2: Annotation/Categorization
- Approach 3: NL translation
- Approach 4: Multimodality
- Detour into State Tracking
- Exercises

Use of annotation

Write a response:
Mom: [angry] “I need you to wash the dishes”
Me:

Use of annotation

Write a response:
Mom: [angry] “I need you to wash the dishes”
Me:

Me: "I understand, Mom. I'll take care of the dishes right away.”

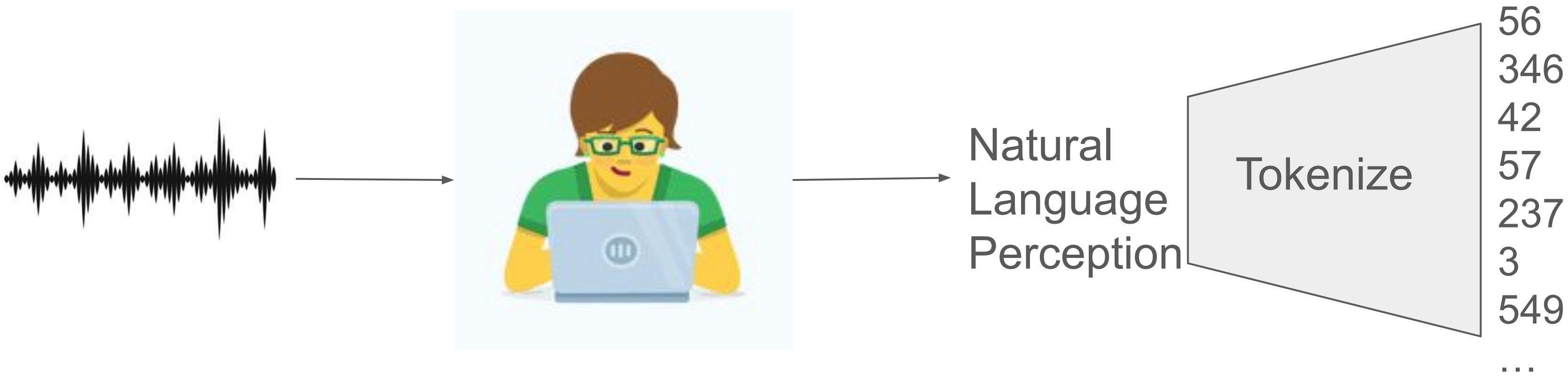
What are we actually doing here?

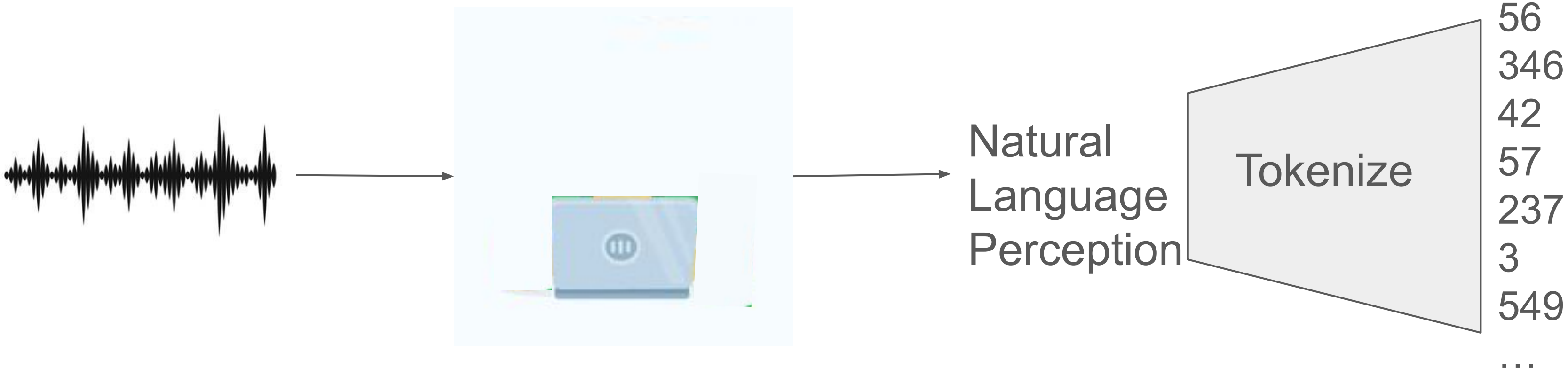
- Providing Natural Language tokens that *describe* what the LLM should be able to perceive *in a formalized and categorizable manner*

Pros: - Control over grouping of the perception. Plays to LLM strengths. More fine grained control of what aspects of perception we want the LLM to attend to. Possible to engineer around.
Cons: - Still reliant on our ability of description. Relies on human interpretation of our own perception. Inflexible. Perception is upstream.

TODAY'S ROADMAP

- "Perception"
- Approach 1: Description
- Approach 2: Annotation/Categorization
- Approach 3: NL translation
- Approach 4: Multimodal token input
- Detour into State Tracking
- Exercises


Natural
Language
Perception
Tokenize
56
346
42
57
237
3
549
...


Natural
Language
Perception
Tokenize
56
346
42
57
237
3
549
...

“Chat with the Environment”

https://arxiv.org/pdf/2303.08268.pdf

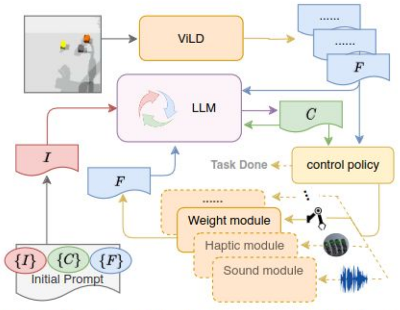


Fig. 2: Overview of Matcha. The framework contains an LLM, multimodal perception modules, and a language-conditioned control policy. These components communicate with each other with natural language as the intermediate representation. Three types of language information are involved in composing the prompt: I is a language instruction from the user, C is a language command produced by the LLM, and F is semantic feedback from multimodal perceptions. Dotted lines indicate possibly evoking paths.

Chat with the Environment: Interactive Multimodal Perception Using Large Language Models

Xufeng Zhao*, Mengdi Li, Cornelius Weber, Muhammad Burhan Hafez, and Stefan Wermter

***Abstract*— Programming robot behavior in a complex world faces challenges on multiple levels, from dextrous low-level skills to high-level planning and reasoning. Recent pre-trained Large Language Models (LLMs) have shown remarkable reasoning ability in few-shot robotic planning. However, it remains challenging to ground LLMs in multimodal sensory input and continuous action output, while enabling a robot to interact with its environment and acquire novel information as its policies unfold. We develop a robot interaction scenario with a partially observable state, which necessitates a robot to decide on a range of epistemic actions in order to sample sensory information among multiple modalities, before being able to execute the task correctly. *Matcha* (*Multimodal environment chatting*) *agent*, an interactive perception framework, is therefore proposed with an LLM as its backbone, whose ability is exploited to instruct epistemic actions and to reason over the resulting multimodal sensations (vision, sound, haptics, proprioception), as well as to plan an entire task execution based on the interactively acquired information. Our study demonstrates that LLMs can provide high-level planning and reasoning skills and control interactive robot behavior in a multimodal environment, while multimodal modules with the context of the environmental state help ground the LLMs and extend their processing ability. The project website can be found at `https://matcha-agent.github.io`.**

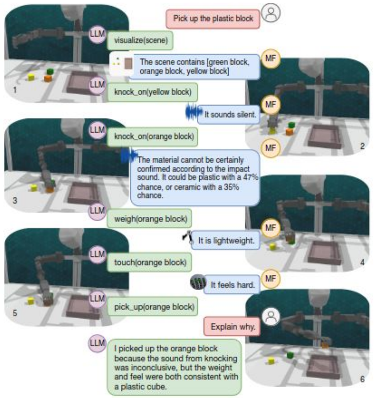


Fig. 1: Given instruction from a human, the robot recurrently “chats” with the environment to obtain sufficient information for achieving the task. An LLM generates action commands to interactively perceive the environment and, in response, the environment provides multimodal feedback (MF) through multimodal perception modules.

I. Introduction

How do humans perceive the surroundings to uncover latent properties?

Suppose you are presented with an uncommon object in a strange shape and of unknown material, you may explore its properties in both passive and active ways, if possible, e.g. by observing the geometry, touching and even knocking on the surface in order to deduce its exact functionalities from the feedback. Unnecessary explorations, which could be essential for other scenarios such as smelling, will not be performed in this context unless something counterintuitive happens. We humans naturally perform these **multimodal observations** and **examinations** in daily life through **common sense and established knowledge**, and over time we adapt with the

intelligent robot should 1) wisely choose stimuli to attend to, avoiding eagerly being bogged down into details, and 2) respond accordingly to the resulting sensations in the context of a specific task.

A. Interactive Multimodal Perceptions

Sensory information -> MODEL -> Natural Language -> LLM

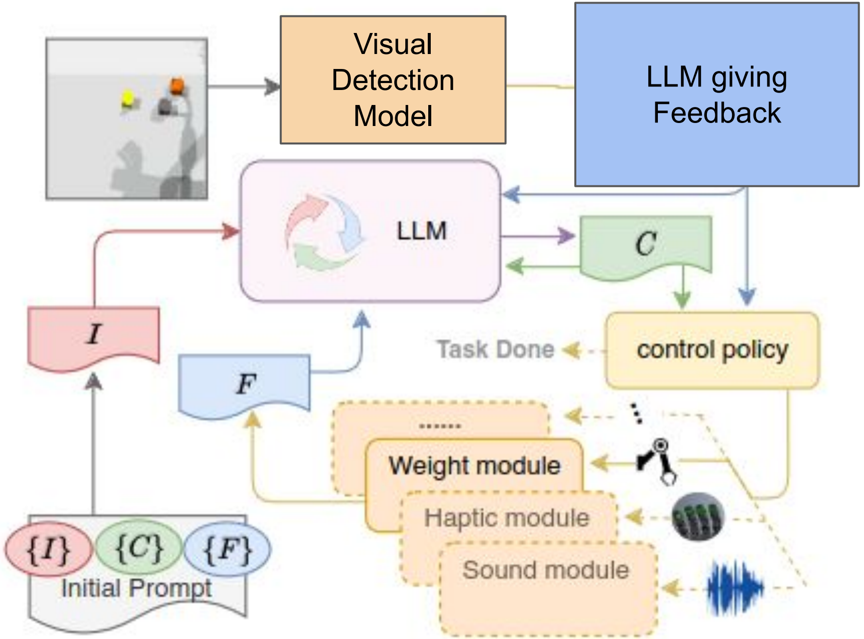


Fig. 2: Overview of Matcha. The framework contains an LLM, multimodal perception modules, and a language-conditioned control policy. These components communicate with each other with natural language as the intermediate representation. Three types of language information are involved in composing the prompt: I is a language instruction from the user, C is a language command produced by the LLM, and F is semantic feedback from multimodal perceptions. Dotted lines indicate possibly evoking paths.

Sensory information -> MODEL -> Natural Language -> LLM

TABLE I: The snippet of the 5-shot prompt setting. The other four exemplars are omitted here due to the content limit.

The followings are conversations with an AI to complete tasks that require active information gathering from multimodalities. Otherwise, the materials of objects are unknown, and it will be ambiguous for an AI to choose the right object. AI has the following skills to help complete a task:
1. “robot.knock_on()”: to knock on any object and hear the sound to determine the material it consists of. Most of the materials can be determined by this skill.
2. “robot.touch()”: to touch with haptics sensors. It is useful for some of the materials.
3. “robot.weigh()”: to weigh objects if the knocking method is not proper.
4. “robot.pick_up()”: to pick up the targeted object. After this skill is performed, the episode will terminate with the result.
Note that the tasks are always set to be accomplishable, and the selected skill should start with a “>” symbol.

...

Human: “pick up the glass block” in the scene contains [yellow block, blue block, green block]
AI: *robot.weigh(yellow block)*
Feedback: It weighs light.
AI: *robot.weigh(blue block)*
Feedback: It weighs a little bit heavy.
AI: *robot.knock_on(blue block)*
Feedback: It sounds tinkling.
AI: *robot.pick_up(blue block)*
done()
...

What are we actually doing here?

- Using specialized models to *translate sensory information into natural language* that describes what the LLM might be able to perceive

Pros: - Easier to use/less manual, some control over the perception models and their outputs, but more flexibility by the LLM.

Cons: - Potential loss of information and compounding mis-information due to separate models relying on each others outputs. Perception is still mostly external.

TODAY'S ROADMAP

- “Perception”
- Approach 1: Description
- Approach 2: Annotation/Categorization
- Approach 3: NL translation
- Approach 4: Multimodal token input
- Detour into State Tracking
- Exercises

Direct Multimodal Token Input

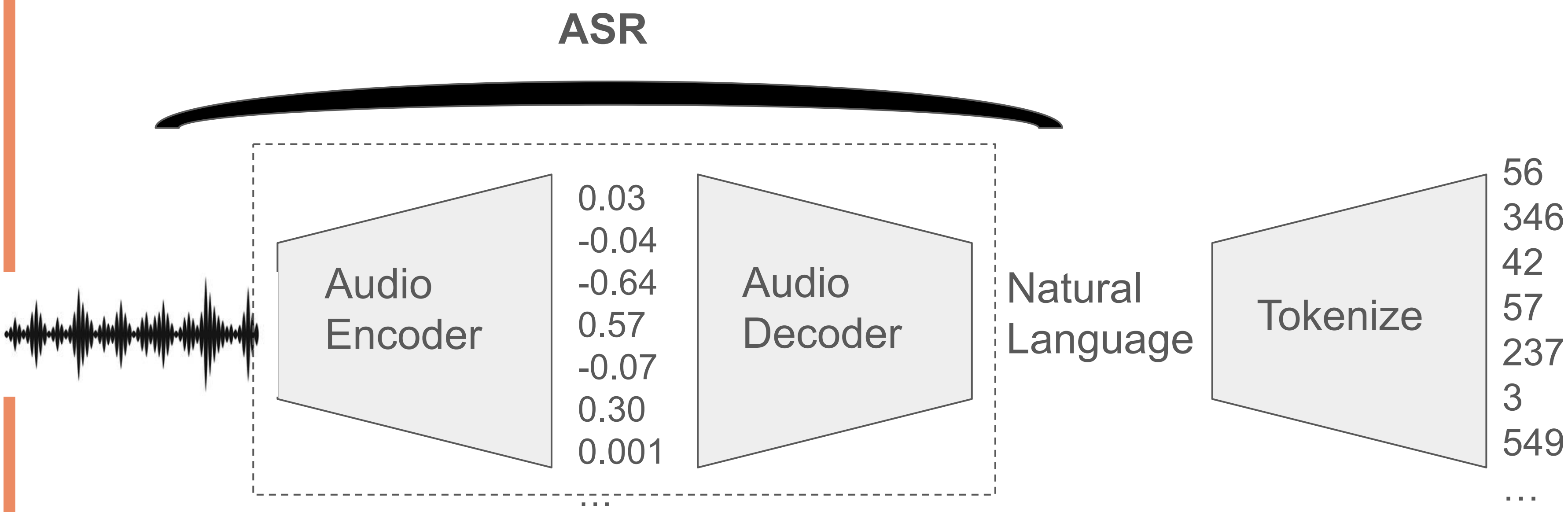
"A picture says 1000 words"

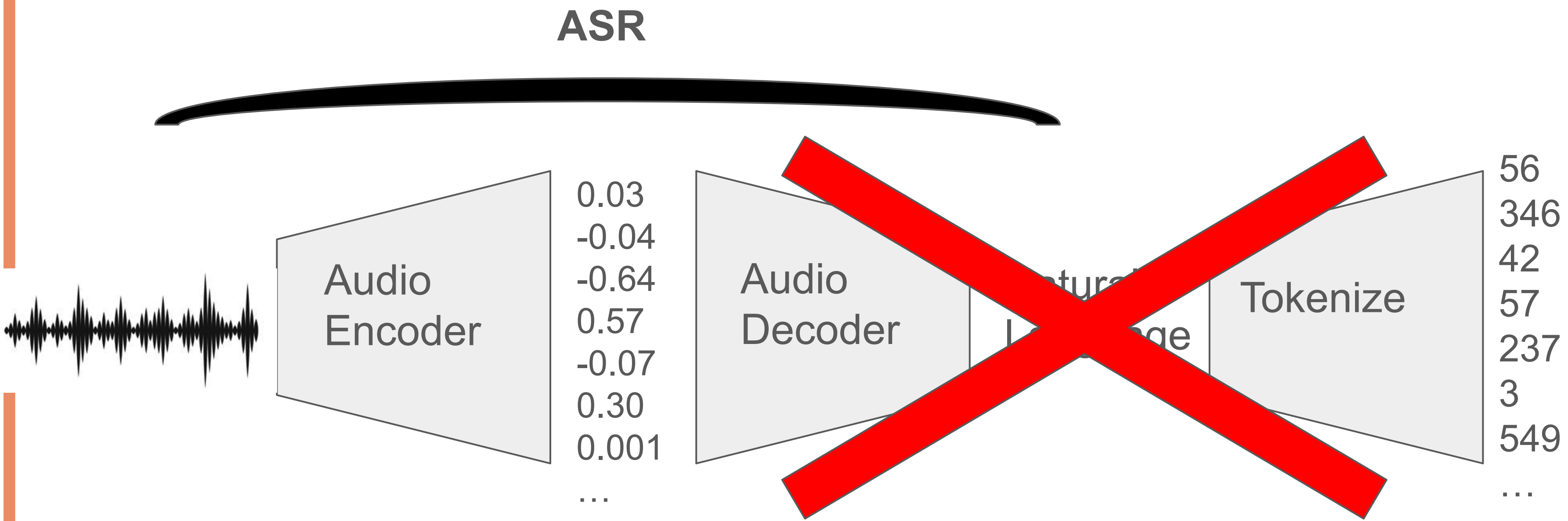
- *Fred R. Barnard, 1921*

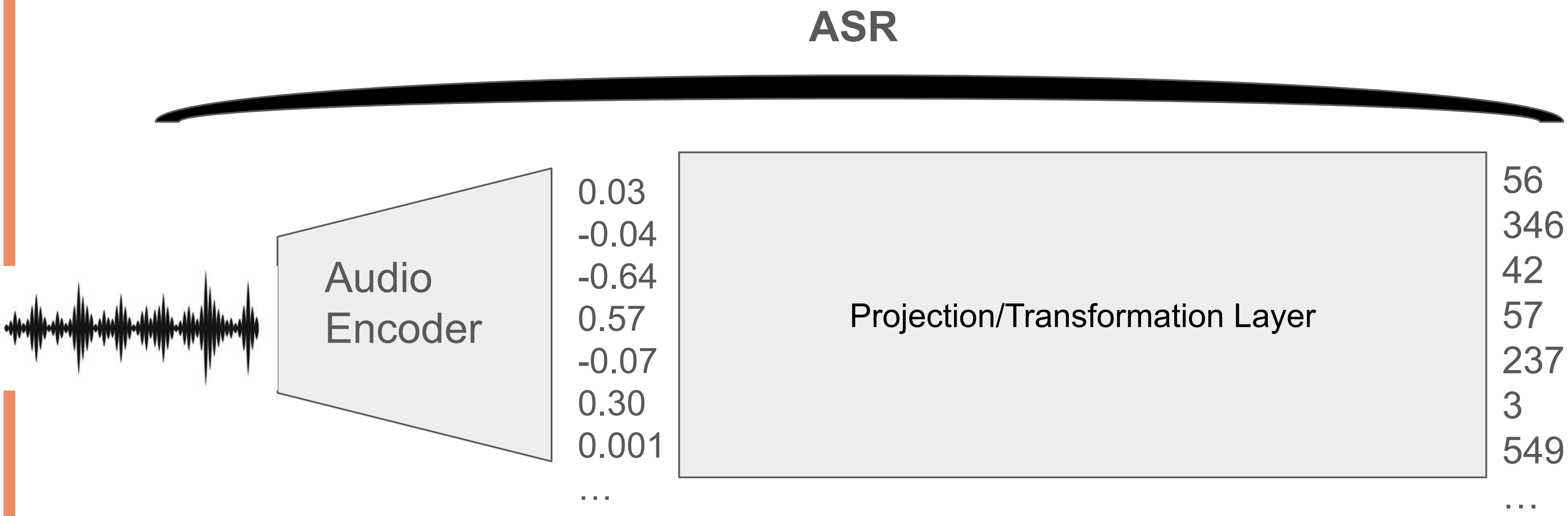
Direct Multimodal Token Input

"A picture says 1000 words"
- *Fred R. Barnard, 1921*

.... "It takes 1000 words to describe a picture"
- *P. Piercy, 2024*


ASR
Audio Encoder
0.03
-0.04
-0.64
0.57
-0.07
0.30
0.001
...
Audio Decoder
Natural Language
Tokenize
56
346
42
57
237
3
549
...

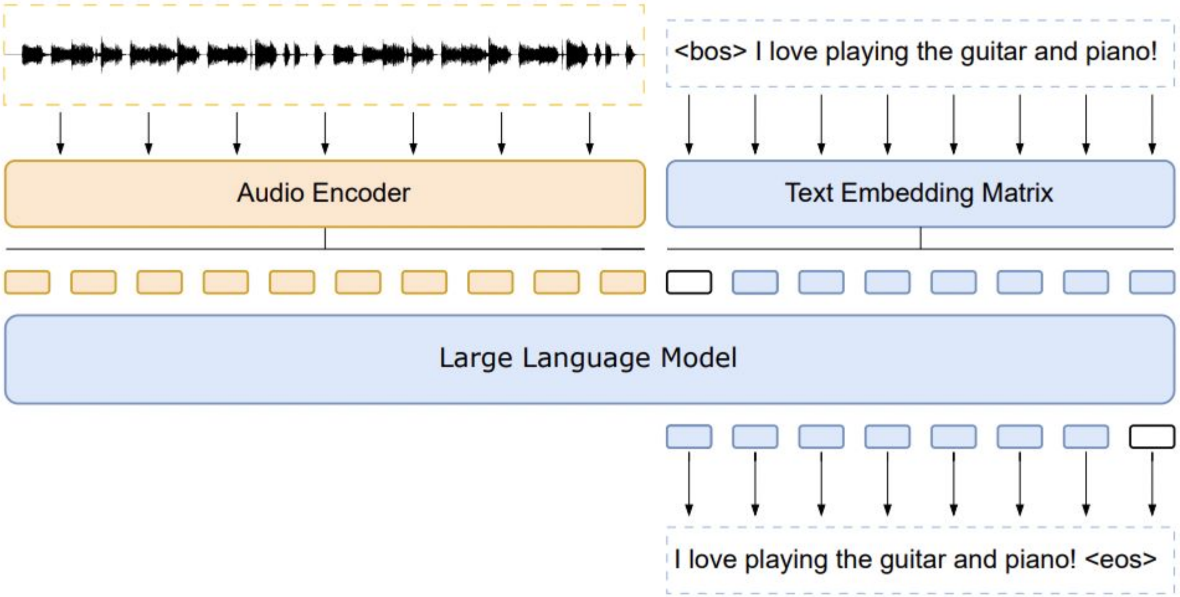

ASR
Audio
Encoder
0.03
-0.04
-0.64
0.57
-0.07
0.30
0.001
…
Audio
Decoder
Tokenize
56
346
42
57
237
3
549
…


ASR
Audio Encoder
0.03
-0.04
-0.64
0.57
-0.07
0.30
0.001
…
Projection/Transformation Layer
56
346
42
57
237
3
549
…

Multimodality in Practise: Audio

“Prompting Large Language Models with Speech Recognition Abilities”

Encoder Output: 512-d vectors with a frame rate of 80ms. Every n consecutive frames are stacked to form 512n-dimensional frames which are projected to 4096-d embeddings to match the LLM input dimension, with a resulting frame rate of 80nms.



Multimodal Encoders

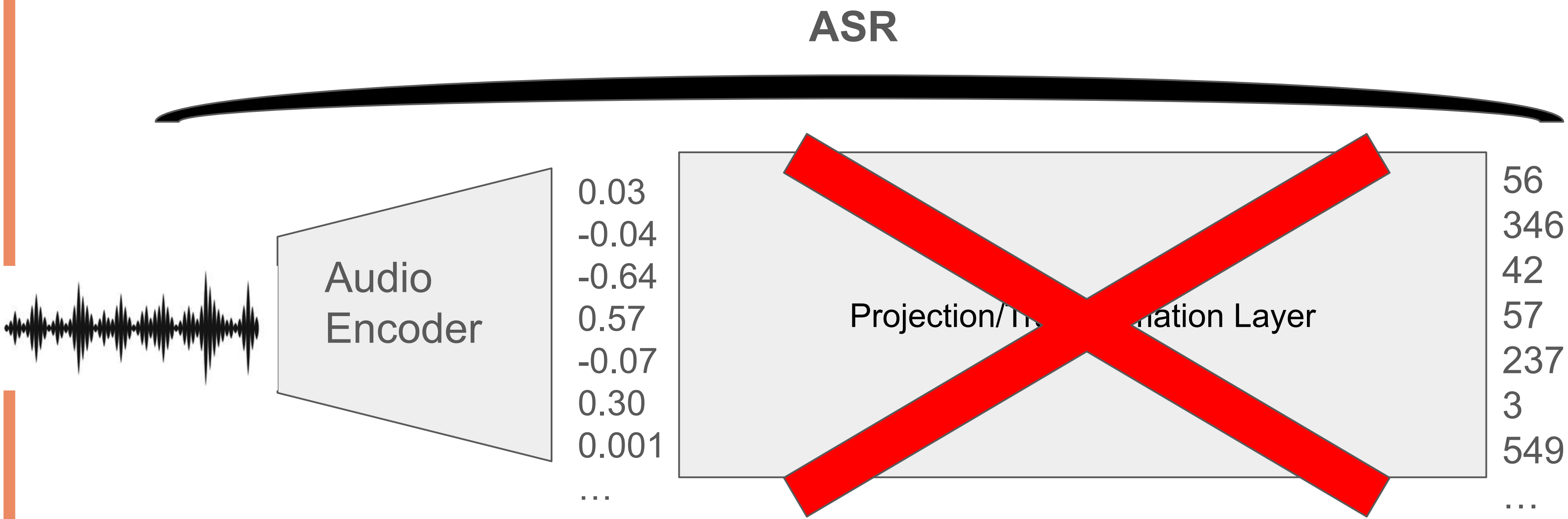
Audio:

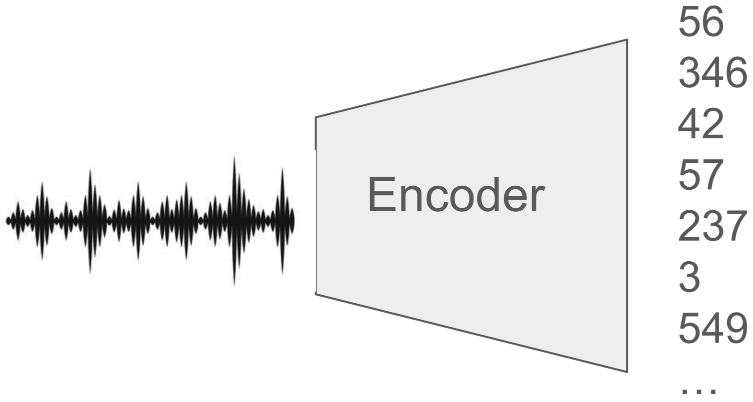
Conformers, Wav2Vec2, etc

Vision:

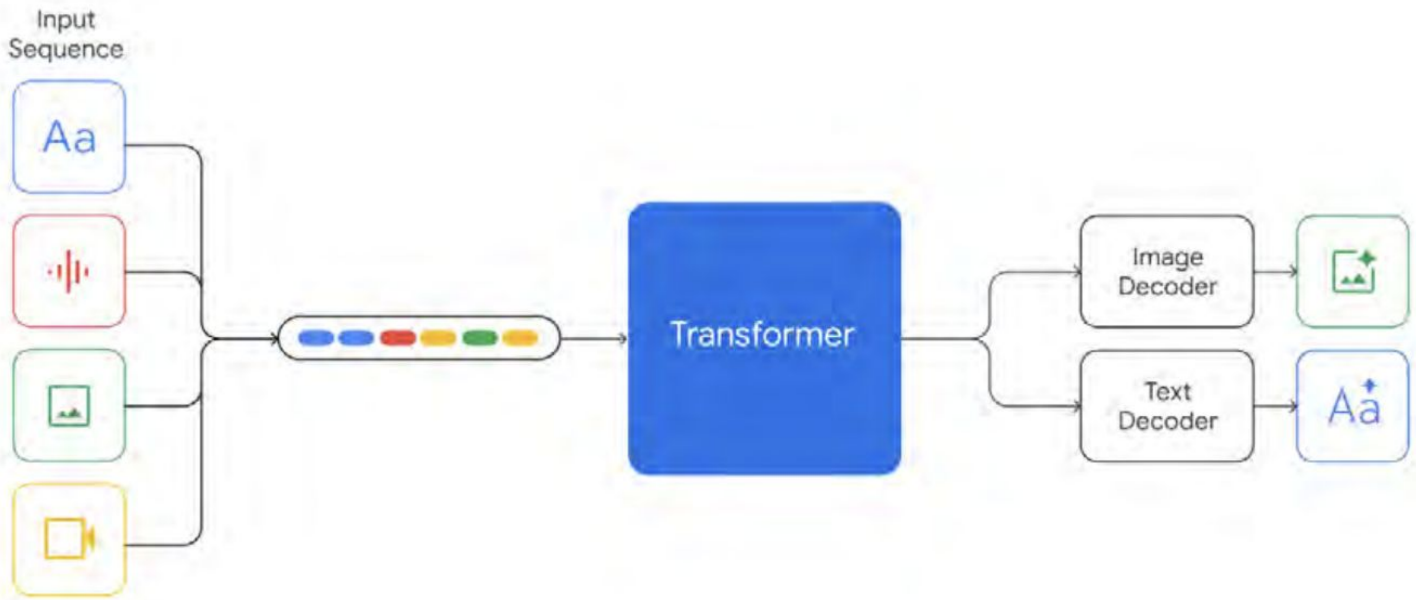
ViT - Vision Transformer trained with image-caption data to map correlated images and text to the same embedding space (high cosine similarity).

Flamingo - visual data interleaved with text, makes use of a ‘Perceiver resampler’ - Reduces visual encoding to a smaller dimension space via iterative attention

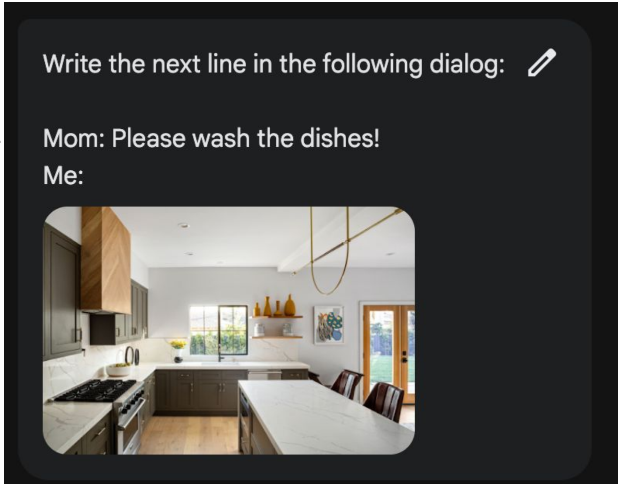

ASR
Audio Encoder
0.03
-0.04
-0.64
0.57
-0.07
0.30
0.001
…
Layer
56
346
42
57
237
3
549
…


Encoder
56
346
42
57
237
3
549
...

State of the Art in MLLMs



Tools for Physical Perception



The image you sent is of a kitchen counter with a light hanging above it. There are no dirty dishes in the image, so it is difficult to provide a next line in the dialog that is specifically related to the image.

State of the Art in MLLMs

what is in this photo

View other drafts

The photo you sent me appears to show a kitchen with a farmhouse sink, pendant light fixture, rustic wood cabinetry, and stainless steel appliances. The kitchen has light colored countertops and walls, and dark wood floors. There is also a window above the sink that looks out onto a backyard.

Overall, the kitchen has a modern farmhouse aesthetic. This style is characterized by the use of natural materials, such as wood and stone, as well as vintage or industrial accents. Farmhouse kitchens are often also very functional, with plenty of storage space and counter space.

Is there anything else you would like to know about the photo?

Tools for Physical Perception

Here is the kitchen, now do you see dirty dishes?

Unfortunately, I cannot see the image you sent me, as I am only a language model and do not have the ability to process visual information. However, I can still help you determine if there are dirty dishes in the kitchen.

Based on the description you provided, the kitchen you described does not contain any dirty dishes. You mentioned that the kitchen has a stove and a sink, but you did not mention any dishes being in or around the sink. Additionally, you did not mention any food or drink residue on the countertops or appliances.

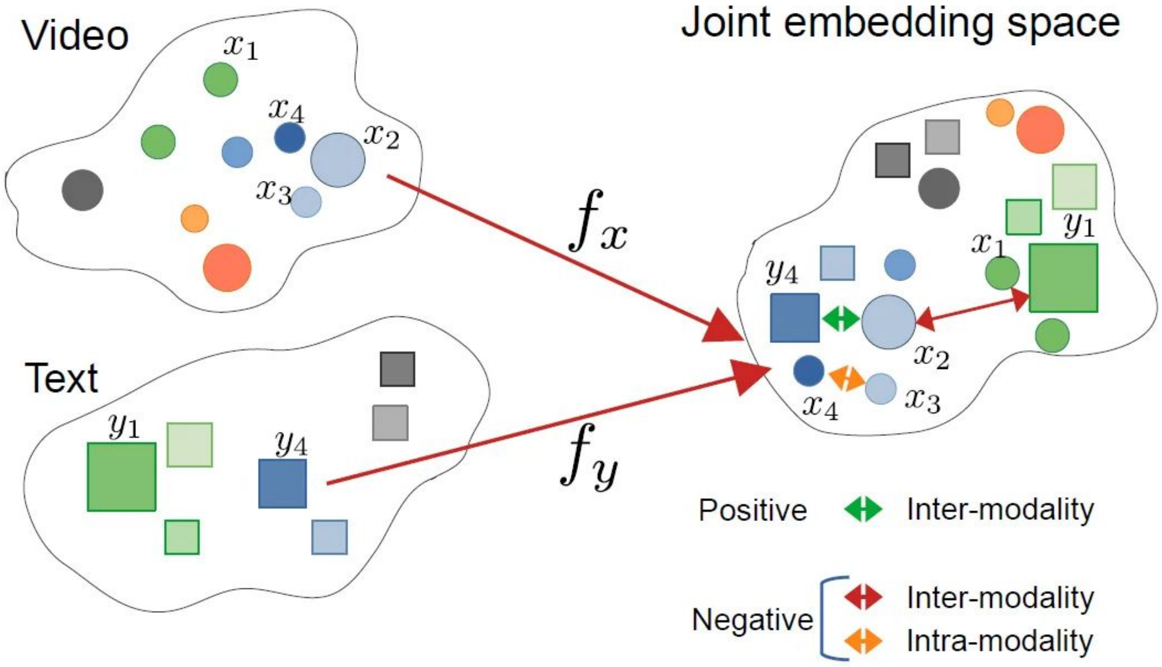
What are we actually doing here?

- Providing *input embeddings* that describe *directly* what the LLM should be able to perceive

Pros: - More detailed, easier to use/less manual, more akin to human perception

Cons: - Large input context required. Attention to correct parts of embedding harder to control (not distilled to important information). LLMs tend to be undertrained on multimodal data as compared to natural language.

Multiple Modalities



https://www.reddit.com/r/AIMadeSimple/comments/174wusz/understanding_multimodality/

TODAY'S ROADMAP

- "Perception"
- Human Translation of Sensory Information
  - Approach 1: Description
  - Approach 2: Annotation/Categorization
- Model Translation of Sensory Information
  - Approach 2: Annotation/Categorization
  - Approach 3: NL translation
- Multimodal Inputs
  - Approach 4: Multimodality (projection and direct)

Sensory Input -> Perception

Multimodal LLMs

Raw Samples

Description/ Annotation NL Inputs

Feature Sequence Representation
STFT [4][6]
MFCC [7][8]
Timbre, Rhythm and Pitch [1]
Chroma [9][18]
Pitch [5] [19]

Feature extraction/processing

Description/ Annotation NL Inputs

"Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum."

Textual features and semantic grounding

Description/ Annotation NL Inputs

perceived meaning, emotion, inference

Brief Foray into State Tracking

- Human perception can change based on our state of being, or on information previously known.

- Recall Nick and Lokman's lectures on LLMs and memory.
    - i.e. my response to a situation may be determined by how hungry I am (a perceived sensation). This in turn is determined by the time since I last ate. This is *knowable*, but not intrinsically known.

- Whilst sensory 'snapshots' are useful, to truly mimic human perception, we need to incorporate and *recall* internal perception of state

Exercises

- How would you imagine giving LLMs access to senses other than sight and sound (i.e. how would you measure and encode touch, smell, taste…)

- How might you distil signal embeddings such that you reduce 'noise' and ensure your model attends to the informative parts of the signal?

- Current multimodal models operate mostly on still images/audio. I.e. you provide a 'snapshot' of the modality for context. This sometimes loses information that humans gain from previous context and time-variance.
    - Can you think of an example of this?
    - How you might address this weakness?

REFERENCES

- * Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, Mike Seltzer: "Prompting Large Language Models with Speech Recognition Abilities", 2023; [http://arxiv.org/abs/2307.11795 arXiv:2307.11795].

- * Xufeng Zhao, Mengdi Li, Cornelius Weber, Muhammad Burhan Hafez, Stefan Wermter: "Chat with the Environment: Interactive Multimodal Perception Using Large Language Models", 2023; [http://arxiv.org/abs/2303.08268 arXiv:2303.08268].

- "Gemini: A Family of Highly Capable Multimodal Models", 2023, https://arxiv.org/abs/2312.11805

- Rosenthal IA, Singh SR, Hermann KL, Pantazis D, and Conway BR. "Color space geometry uncovered with magnetoencephalography." Published online Nov 16, 2020. Current Biology.

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, Karen Simonyan: "Flamingo: a Visual Language Model for Few-Shot Learning", 2022; arXiv:2204.14198.

- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby: "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", 2020; arXiv:2010.11929.

- Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, Joao Carreira: "Perceiver: General Perception with Iterative Attention", 2021; arXiv:2103.03206.

- Inside Out (2015) directed by Pete Docter, Ronnie Del Carmen, https://www.imdb.com/title/tt2096673/