MAF
MILANO AUDIOVISUAL FORUM

ULTRAHD FORUM ITALIA
ANNUAL CONFERENCE

29 | 30 OTTOBRE 2024
Centro Congressi Stella Polare
Fiera Milano Rho

IA generativa multimodale

Come i modelli linguistici di grandi dimensioni stanno cambiando pelle

Elena Baralis
Politecnico di Torino, Italy

Politecnico di Torino
1859

Multimodal AI

- Closer to human brain perception
- ***Processes at once*** information from different modalities, such as text, images, audio, and video

➡ Enhances understanding and interaction capabilities

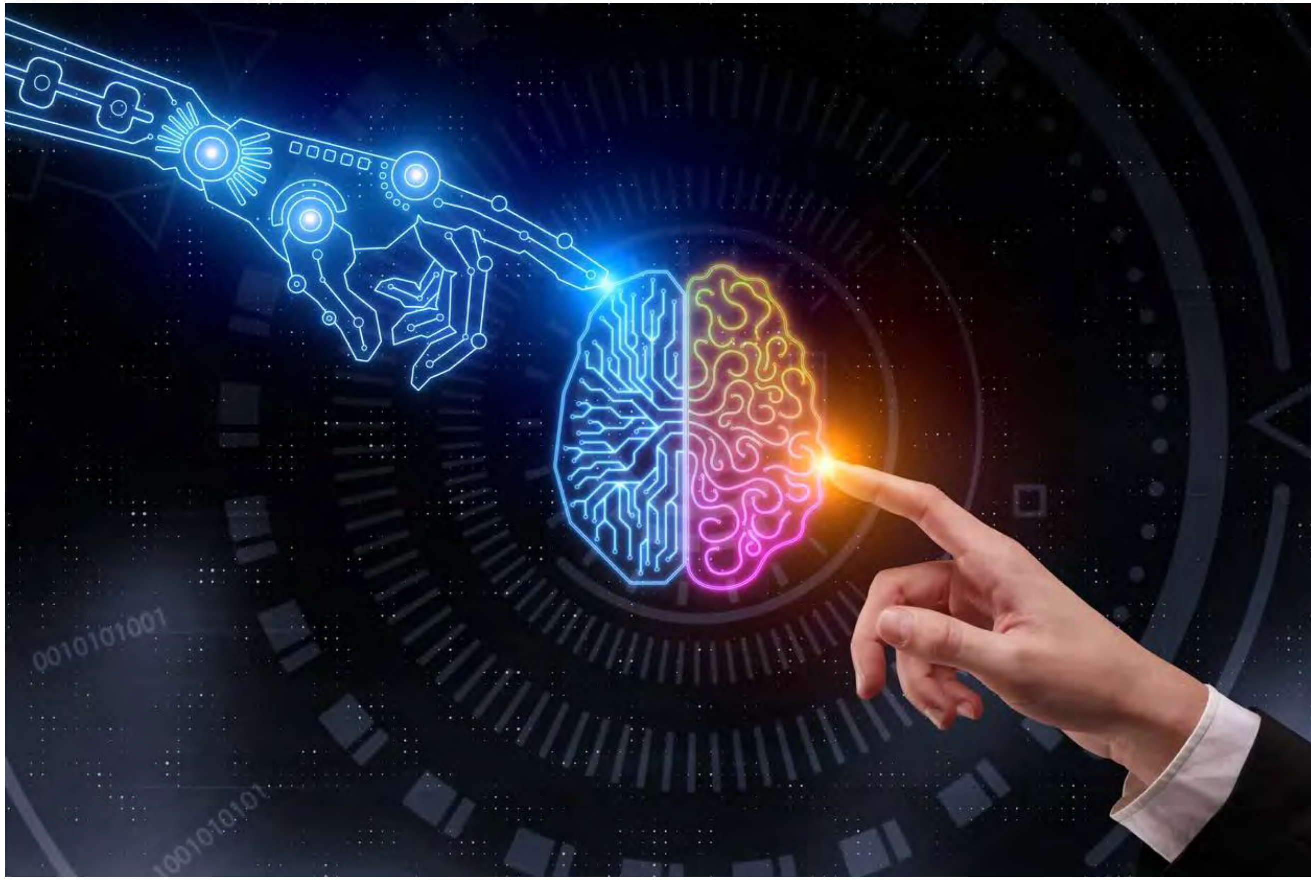
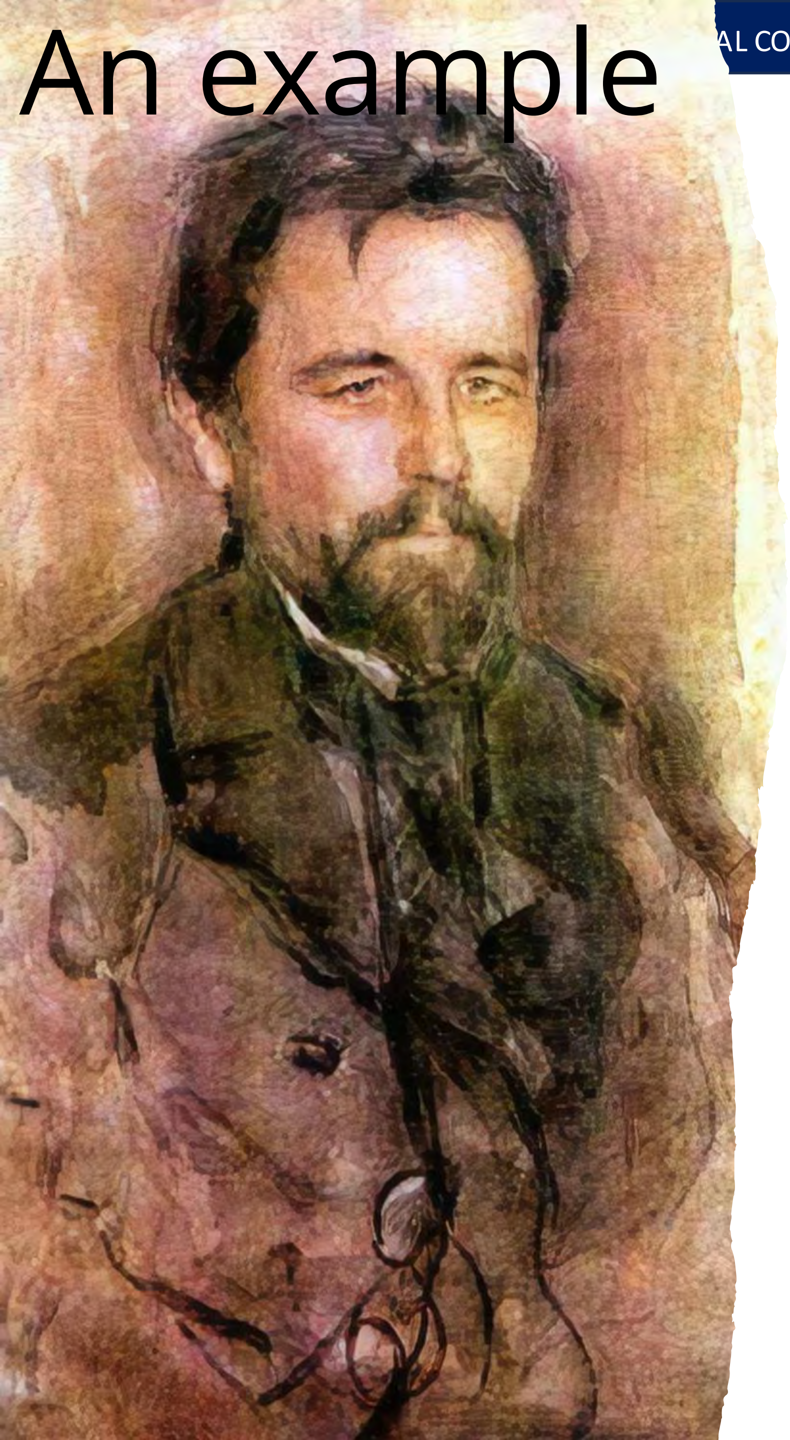


Image source: Great Learning

Types of modalities

- **Text:** Natural language processing (NLP)
- **Images:** Computer vision
- **Audio:** Speech recognition and synthesis
- **Video:** Temporal analysis and event recognition

An example

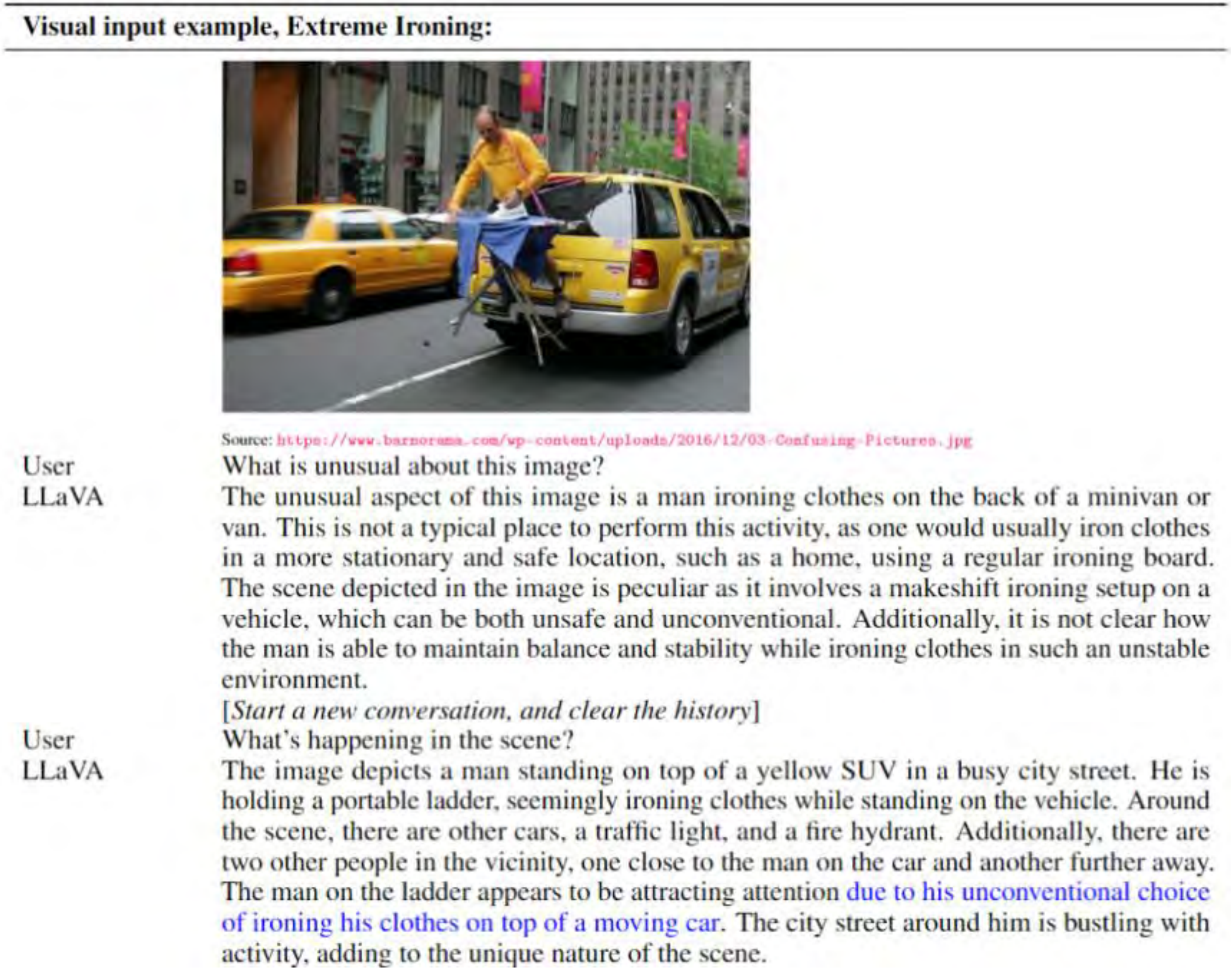


Prompt: ***Please describe this image***

- The image is a stylized portrait, likely digitally manipulated to resemble a watercolor painting, of a man with dark hair and a beard.
- Here's a breakdown of the description:
- **Subject:** The central focus is a man's head and shoulders. He appears to be middle-aged, with dark brown or black hair that is somewhat unkempt. He has a full, slightly unkempt beard and mustache. His expression is serious, almost melancholic, and his gaze is directed slightly downward and away from the viewer.
- **Style:** The style is impressionistic and painterly. It mimics the look of a watercolor painting, with visible brushstrokes, washes of color, and a soft, slightly blurry quality. The colors are muted earth tones, primarily browns, tans, and muted greens, creating a somewhat somber mood. The details are not sharply defined; instead, the focus is on capturing the overall mood and form.
- **Clothing:** The man is dressed in what appears to be a dark-colored coat or jacket, possibly with a high collar. There's a suggestion of a lighter-colored shirt or vest underneath. *A pair of eyeglasses or pince-nez is visible hanging from a cord around his neck.*
- **Background:** The background is a simple, indistinct wash of light beige or tan, which does not detract from the portrait's subject.
- **Overall Impression:** The image conveys a sense of quiet contemplation or introspection. The soft, muted colors and the slightly blurred details contribute to a feeling of nostalgia or even melancholy. The style suggests a possible artistic interpretation of a historical figure or a character from literature.

by ***gemini-1.5-flash-002***

Another example



Visual input example, Extreme Ironing:

Source: https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg

User	What is unusual about this image?
LLaVA	The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.
	[*Start a new conversation, and clear the history*]
User	What's happening in the scene?
LLaVA	The image depicts a man standing on top of a yellow SUV in a busy city street. He is holding a portable ladder, seemingly ironing clothes while standing on the vehicle. Around the scene, there are other cars, a traffic light, and a fire hydrant. Additionally, there are two other people in the vicinity, one close to the man on the car and another further away. The man on the ladder appears to be attracting attention due to his unconventional choice of ironing his clothes on top of a moving car. The city street around him is bustling with activity, adding to the unique nature of the scene.

Generative AI

Creating New Content

GenAI generates new data that resembles the input data distribution, often through the generation of novel content or creative outputs

Learning Approach

GenAI uses unsupervised learning techniques to understand and mimic the underlying patterns in data, allowing them to generate new instances similar to the original

Examples

Image generation, text generation, music composition

Large Language Models

Self-supervision

Learn language properties from data and generate human-like text

Vast Knowledge Base

Trained on extensive datasets comprising diverse text sources, allowing it to grasp a wide range of topics and concepts

Natural Language Understanding

Capable of “comprehending” nuances of human language, including context, semantics, and syntax

Applications

Natural language processing tasks such as text generation (e.g., chatbots), summarization, translation, sentiment analysis

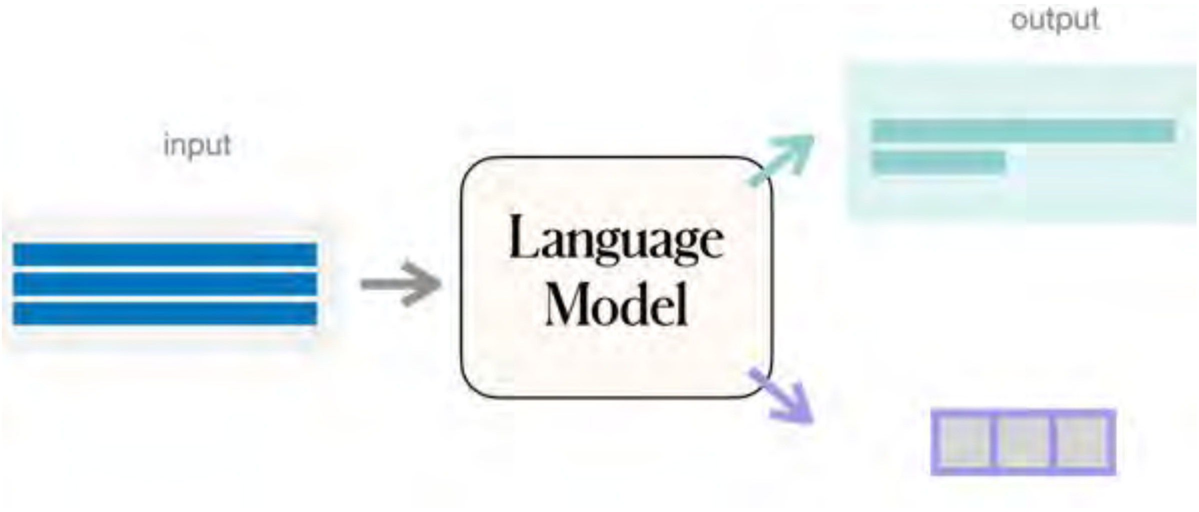
Continuous Learning

May be continuously updated and improved with new data, enhancing its capabilities over time

Next Token Prediction

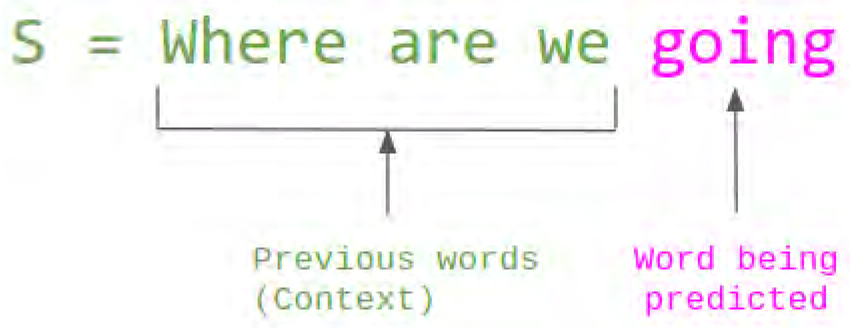
Given a sequence of words, what is the next likely term?

- Candidate words belong to a dictionary
- A large body of documents is used to learn word frequencies and relative positions of words in text sequences



A Survey of Large Language Models. 2023. Wayne Xin Zhao et al. https://arxiv.org/abs/2303.18223

Language Model

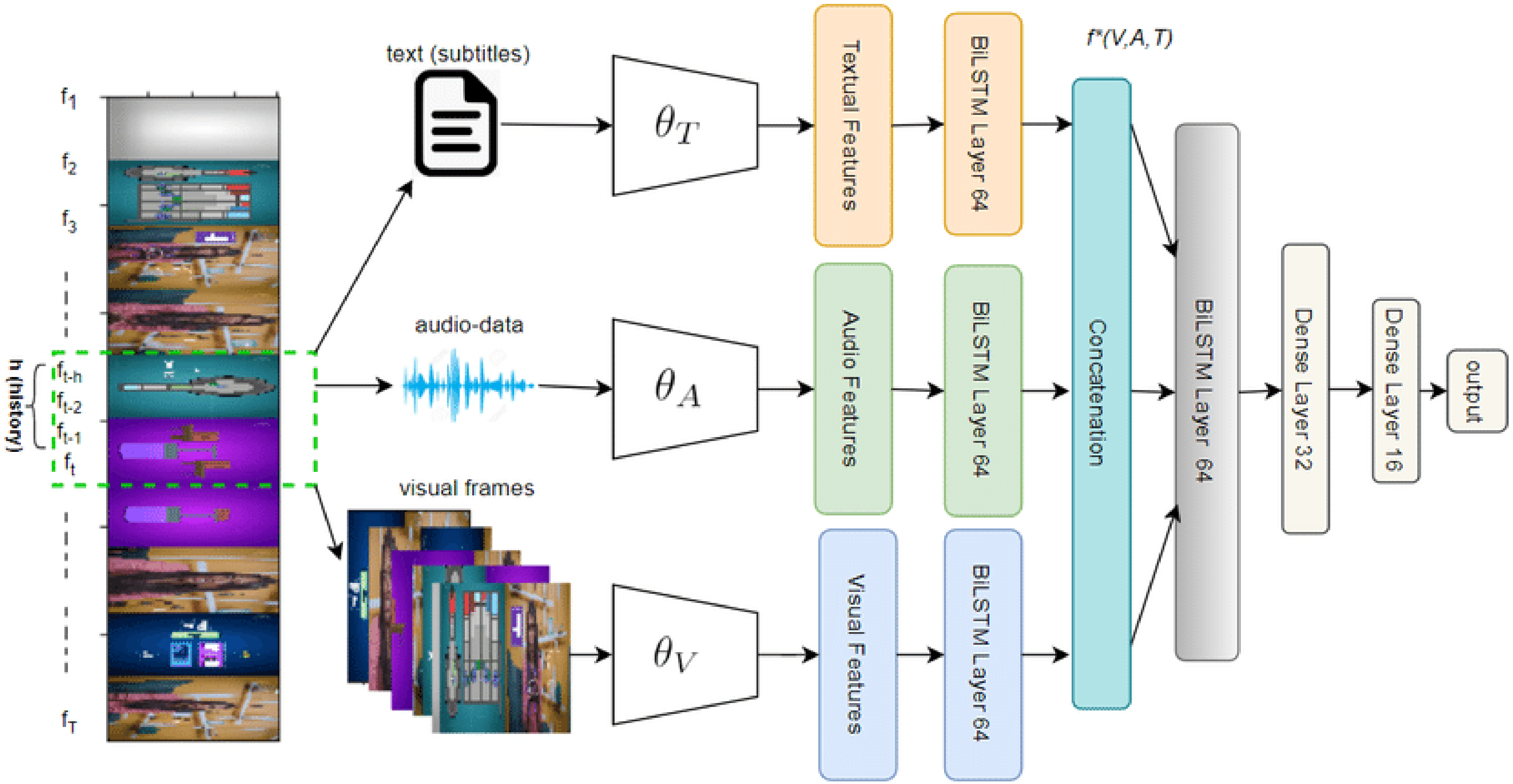

S = Where are we going

Previous words (Context)

Word being predicted

P(S) = P(Where) x P(are | Where) x P(we | Where are) x P(going | Where are we)

Fusion in multimodal AI



Tim Cvetko, “Multimodality Explained. Part I: Fusion”, in AI Advances.

Challenges in Multimodal AI

- ***Data.*** Huge quantities of data are needed to train models
- ***Scalability.*** Handling large datasets effectively
- ***Data Alignment.*** Synchronizing data from different modalities
- ***Deployment.*** Prompting may become more difficult
- ***Interpretability.*** Understanding model decisions across modalities

Issues in Multimodal AI

- ***Hallucinations***
  - Potential reputational damage or legal liability
- ***Data bias***
  - Training data that does not violate intellectual property (black box also for training data)
- ***Inappropriate use***
  - Use of ethical and copyright filters
- ***Security***
  - The number of possible attack vectors increase

Hallucination

Generation of text not grounded in factual data, or not supported by the input prompt

➡ The generated content is fictional, inaccurate, or unrelated to the provided context

- Causes
  - Data quality
  - Incomplete or ambiguous prompts
  - Overfitting to (irrelevant) training data
  - Missing fact checking in the loop

Hallucination can erode trust in LLMs as reliable sources of information or tools for natural language understanding and generation

What’s next?

Multimodal AI represents a significant leap in AI capabilities

- ***Intuitive systems.*** It empowers more sophisticated user interactions
- ***Enhanced personalization.*** It allows tailoring experiences by considering multimodal data
- ***Interdisciplinary research.*** It fosters collaboration across different research fields (e.g., neuroscience, linguistics)

Thank you!

Politecnico
di Torino