UNIVERSITÀ DEGLI STUDI DI NAPOLI FEDERICO II

Artificial Intelligence and Intelligent Systems — CINI National Lab

PICUS lab — PATTERN ANALYSIS AND INTELLIGENT COMPUTATION FOR MULTIMEDIA SYSTEMS

Responsible and Reliable AI at PICUS Lab

Narendra Patwardhan[1], Lidia Marassi[1], Michela Gravina[1], Antonio Galli[1], Monica Zuccarini[1], Tannistha Maiti[2], Tarry Singh[2], Stefano Marrone[1], and Carlo Sansone[1]

[1] University of Naples Federico II, Naples, Italy
[2] Deepkapha AI, Assen, Netherlands

Ital-IA 2023: 3rd National Conference on Artificial Intelligence May 29-31, 2023, Pisa, Italy
Workshop AI Responsabile e Affidabile

Dual Perspective of Artificial Intelligence

- Artificial intelligence (AI) has made significant progress in recent years, yielding promising results in various downstream tasks

Sustainability

AI models often rely on massive computing and data, raising concerns due to high energy consumption and carbon footprint

Hominis Project

Sociological Concerns

The increasing use of generative models for fake news creation is posing serious ethical and sociological concerns

FEAD-D

What are foundation models?

- Artificial intelligence (AI) has emerged as a transformative force in modern society, with generative modelling serving as a key driver behind its rapid advancements
- Models, particularly those based on transformer architectures, have achieved remarkable performance in domains such as NLP, Computer Vision and Robotics
- Foundation models are large-scale machine learning models, pretrained on vast amounts of diverse data, that serve as a backbone for various downstream applications through fine-tuning and adaptation

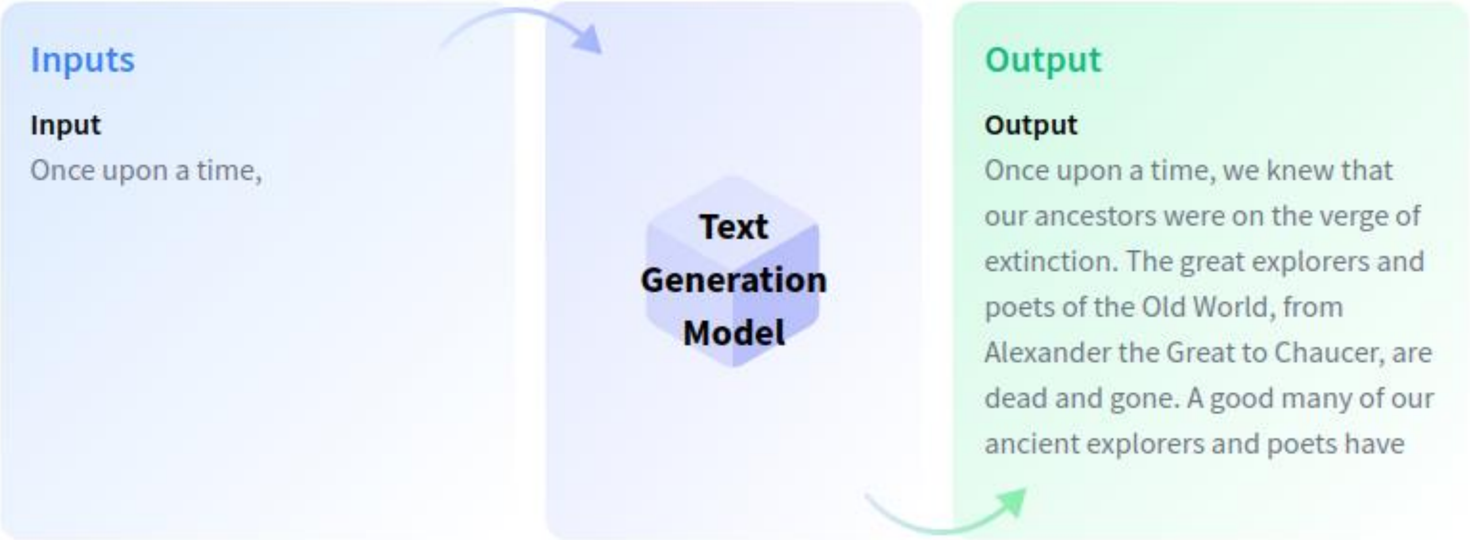
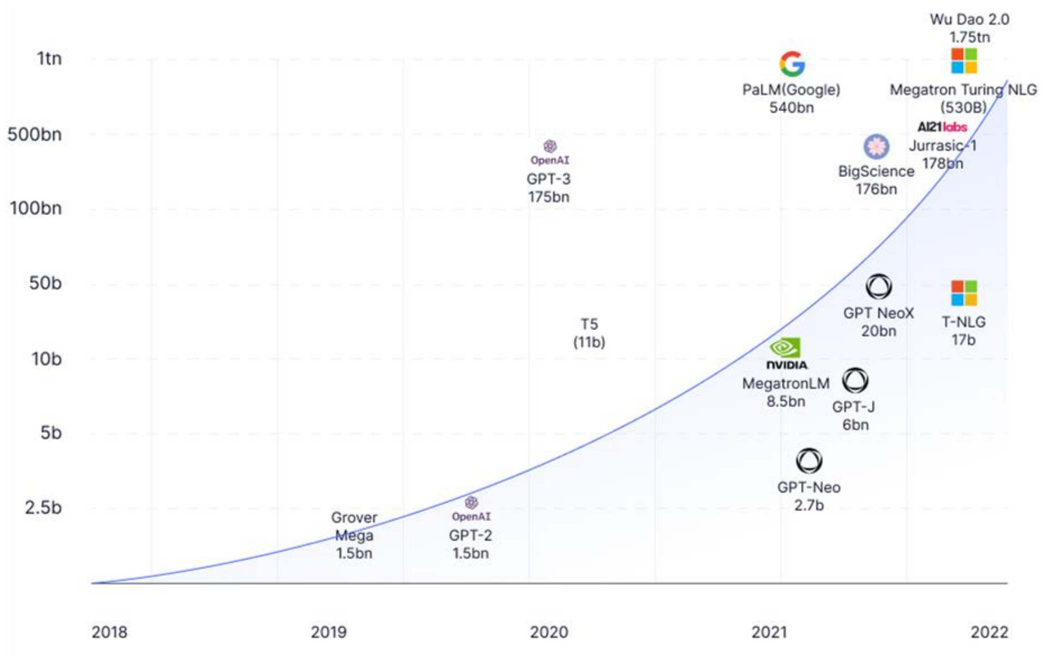


*Image taken from https://huggingface.co/tasks/text-generation

The need for Sustainability

- As foundation models grow in size and complexity, the search for optimal hyperparameters becomes increasingly challenging and resource-intensive

- Solely relying on scaling up can lead to overfitting and may result in diminished returns on model performance improvements
  - ✓ Smaller models such as Galactica/Chinchilla outperforming larger models such as GPT-3 solely based on data



- The high parameter count of foundation models presents challenges for inference on standard hardware, restricting them from a broader audience

- This trend if continued, may contribute to a digital divide between those who can afford to deploy and utilize advanced AI systems and those who cannot

Sustainable Modifications

- Top-performing Large Language Models (LLMs) are based on the Transformer architecture
- The architecture deals with tokenization by representing each input token as an embedding vector, which captures its semantic meaning, and processes these token embeddings through self-attention and feedforward layers to generate contextualized representations of the tokens
  - ✓ We are working on innovative strategies to make this process more effective and generalizable
- The attention mechanism plays a crucial role in the transformer architecture, enabling the model to focus on relevant features in the input data
  - ✓ We are experimenting with new variants explicitly designed for sustainability and fairness
- The Transformer architecture consists of an encoder and a decoder, both of which are composed of multiple layers
  - ✓ We are designing a novel structure to allow an easy conditioning processes while requiring constant time O(1) per token
- Finally, while our aim for Project Hominis is to optimize for inference and reusability, we are also working to reduce the financial and environmental expenditures during the training phase

Sustainable Sourcing

- Hominis aims to unify publicly available and community-vetted sources, including scientific papers, permissible licensed codes, reference materials, and knowledge bases, to create a high transfer-value dataset
- This unification process will also involve automated and **pessimistic filtering** of known large-scale datasets, such as common crawl, to ensure the removal of biased or harmful content
- Additionally, the project will focus on creating smaller datasets with the help of synthetic tasks to facilitate alignment via instruction tuning

TIME

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

Deepfakes

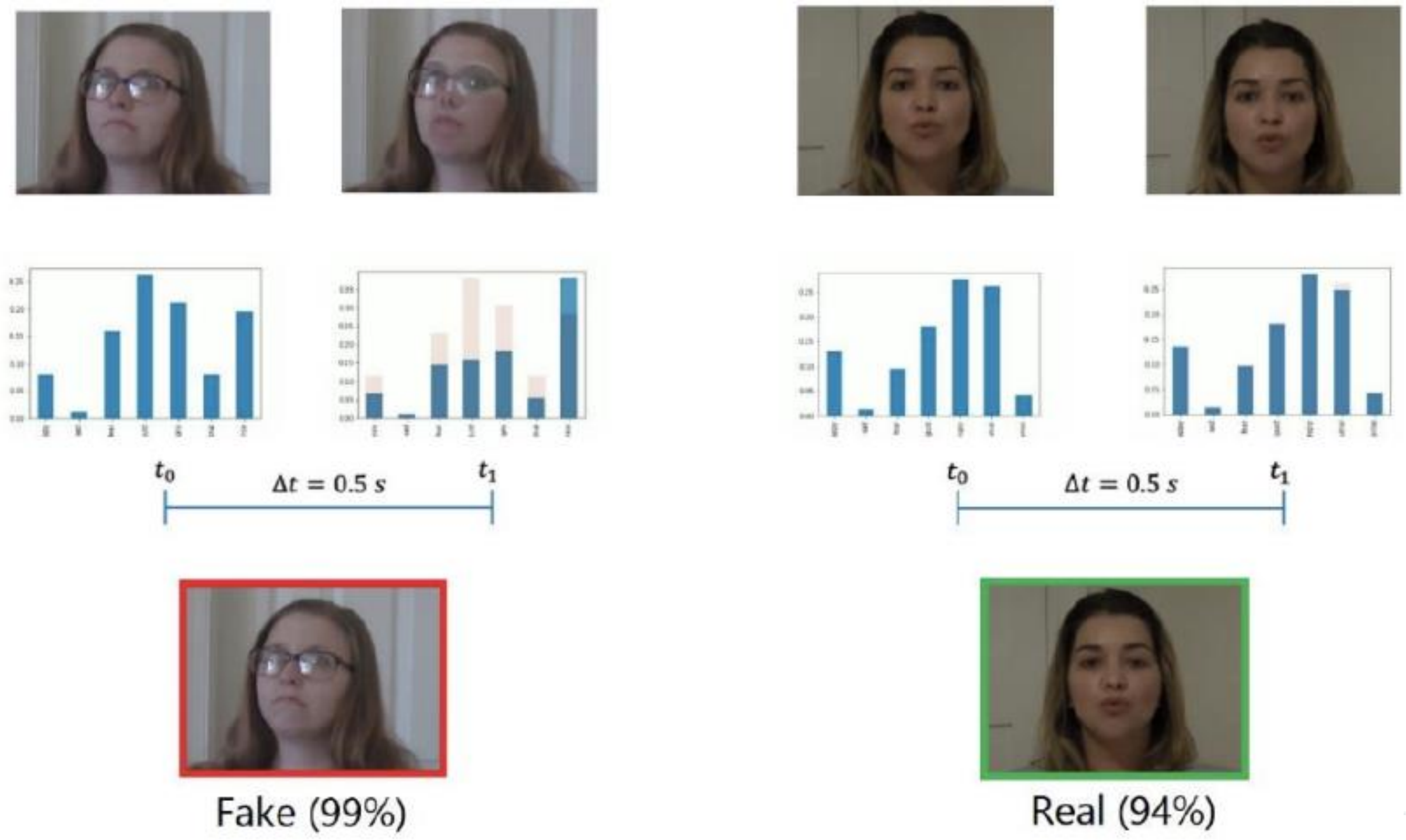
- Deepfakes refer to synthetic media, including images and videos, that are generated using Artificial Intelligence (AI) techniques to alter the appearance or speech of real individuals

- As Deep Learning (DL) approaches have advanced, deepfakes have become increasingly realistic, raising concerns about their potential to spread misinformation and manipulate public opinion
  - ✓ the ability to accurately detect deepfakes has become a critical issue

- One of the main challenges in this field is to identify effective and robust features that can distinguish between real and fake videos.

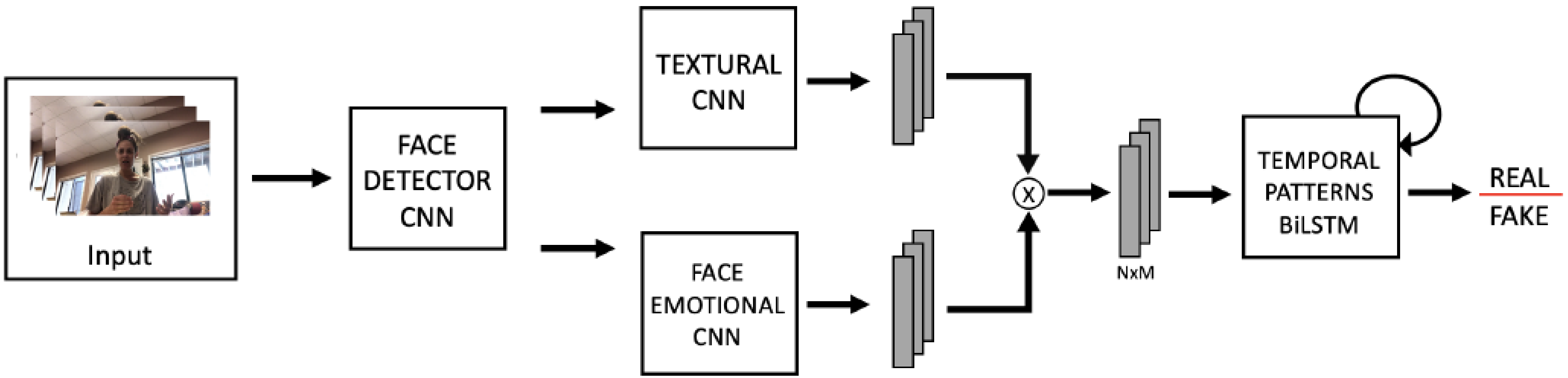
Emotion-based Deepfake detection

- Emotions have been emerging as a valuable feature for deepfake detection due to the difficulty of synthesizing realistic emotional expressions, which remains a major limitation of deepfake creation algorithms



FEAD-D

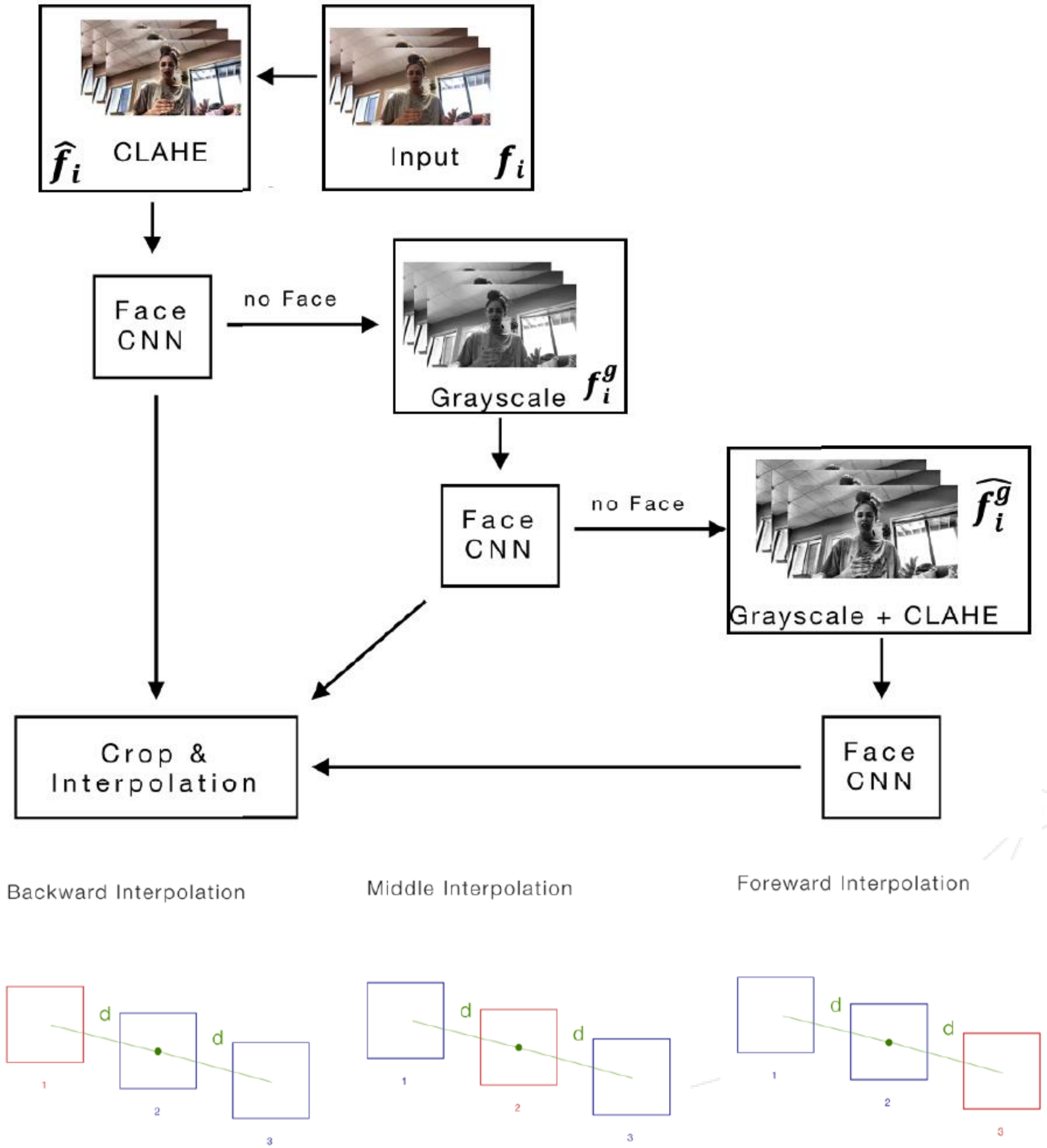
- Facial Expression Analysis in Deepfake Detection (FEAD-D, Iscra-C project) aims at exploiting the unnatural variation in the facial expressions introduced by the artefacts generated during the video creation
- The system has been trained and tested on data coming from the Deepfake Detection Challenge (DFDC)
- It exploits Convolutional Neural Networks (CNNs) as features extractors and a bidirectional Long Short-Term Memory (BiLSTM) network to analyse the temporal patterns



Face detector

- Face detection is a crucial stage, as in recent deep fakes the subject's head pose (and thus face) can change a lot during the video and/or the subject moves in the scene (e.g., walks)

- The developed algorithm also mitigates the failure of recognition by implementing different preprocessing operations on the input image



Features Extraction

Emotional

- A CNN specifically trained for emotion recognition is used as features extractor
- The network is trained considering data coming from the Facial Expression Recognition (Fer2013) challenge
  - ✓ Seven emotional categories (anger, disgust, fear, happiness, sadness, surprise, and neutral)

Textural

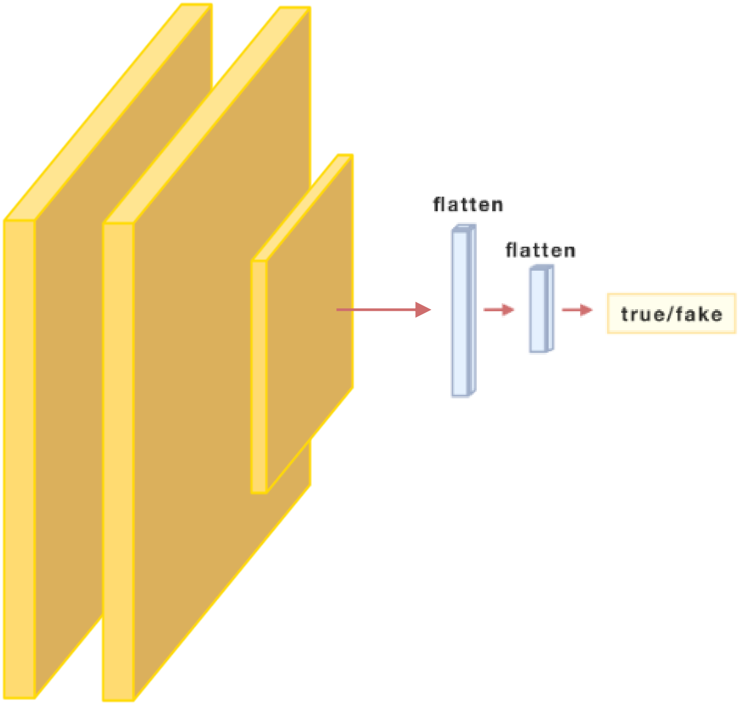
- CNN pre-trained on ImageNet dataset as a textural features extractor
- The aim is to use the textural characteristics for recognizing the artifacts related to the contextual information

- A feature vector is obtained by concatenating the representations produced by the associated CNNs

Features Temporal Analysis

- The features extracted in the previous stages are analyzed together in a cross-frame fashion to spot incoherent and unnatural patterns in the emotional evolution of the target subject

3 BiLSTM layers



- The resulting system can process a video in two minutes

- It is worth noting that although emotional analysis is a promising approach, it presents challenges related to variations across individuals, cultures, and contexts, and the possibility of creating algorithms specifically designed to mimic emotional expressions

Questions?