Exploring Conversational Agents and Continual Learning in Artificial Intelligence

Lucian Gruia
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
luciancgruia@gmail.com

Bogdan Ionescu
AI Multimedia Lab
National University of Science and Technology POLITEHNICA Bucharest
Bucharest, Romania
bogdan.ionescu@upb.ro

***Abstract*—This article presents an overview of my research activities at the AI Multimedia Lab, focusing on continual learning techniques and generative artificial intelligence (GenAI) as applied to conversational agents. The primary aim is to enhance conversational agent capabilities through adaptive learning methodologies and to explore intersections with broader AI technologies, such as neural networks, video processing, and digital avatar creation. This paper summarizes key achievements, including secondments, course development, and applied research, as well as ongoing efforts to advance conversational technologies.**

***Index Terms*—Conversational AI, Continual Learning, Generative AI, Neural Networks, Video Processing, Digital Avatars, Knowledge Retrieval, RAG Architecture, Multimodal Interfaces, Self-awareness in AI, Agentic AI**

I. Introduction

The landscape of conversational AI has transformed in recent years, with the rise of increasingly sophisticated chatbots and conversational agents. Initially, chatbots were limited by rule-based programming, where interactions followed predefined scripts. Today, advances in deep learning, particularly neural networks and language models, have enabled conversational agents to achieve human-like interaction and contextual awareness. However, there remain significant challenges in creating agents that learn continually, adapt to new information, and handle nuanced human conversation effectively. My research focuses on addressing these challenges by applying continual learning methodologies and exploring the role of generative models in achieving responsive, adaptive conversational agents. This article summarizes my activities over the past year in line with this goal, providing an overview of specific research projects, educational contributions, and technological advancements within the AI Multimedia Lab.

II. Literature Review

The literature on conversational AI has rapidly evolved, highlighting the potential and limitations of AI-driven chatbots. Current research emphasizes machine learning, natural language processing, and reinforcement learning as core techniques for creating intelligent agents. Language models such as GPT and BERT have enabled conversational agents to achieve impressive results in understanding context, emotion, and user intent. However, continual learning in conversational AI remains a developing area, with significant barriers to real-time adaptability and personalization.

Continual learning is particularly relevant in conversational AI, as it allows an agent to integrate new knowledge and adapt its responses without forgetting previously learned information. Techniques such as knowledge distillation, dynamic memory networks, and generative pre-trained transformers (GPTs) have been proposed as solutions for creating adaptive AI systems. However, challenges like catastrophic forgetting, ethical considerations, and resource constraints persist. This research seeks to address these gaps, exploring scalable solutions that enhance adaptability in conversational agents and broaden their application in real-world scenarios.

III. Research Objectives

The main objective of this research is to design a conversational agent that is not only adaptive but also capable of learning continually, enabling it to maintain relevance over extended interactions. The objectives include:

- Developing techniques for enhancing the agent's contextual retention and knowledge updating mechanisms, enabling responses that reflect accumulated understanding.
- Examining the relationship between neural network structure, data dimensionality, and processing capacity in generative language models, particularly in relation to conversation quality.
- Investigating ethical considerations around AI adaptability, autonomy, and the agent's potential self-awareness, contributing to ongoing discussions around responsible AI usage.

These objectives are realized through diverse projects within the AIM lab, each contributing distinct insights and practical advancements in the field of conversational AI.

IV. Methodology and Project Activities

A. AI4Media Secondment

During my secondment at AI4Media, I had the opportunity to engage with advanced research in neural networks and transformer architectures under expert guidance. This experience included designing and implementing a neural network

1st Workshop on Artificial Intelligence for Multimedia, AI Multimedia Lab, CAMPUS Research Institute, POLITEHNICA Bucharest, Bucharest, Romania, 8 November, 2024

framework that allows for flexible configuration of layers, activation functions, and learning parameters. The framework serves as a tool for experimenting with various network architectures, providing insights into how specific configurations impact language model performance.

Key activities included authoring a research article for the doctoral symposium, conducting a 6-hour course at the AI Doctoral Academy, and exploring graph database applications in conversational agents. These engagements allowed me to deepen my understanding of AI models, especially in the context of retrieval-augmented generation (RAG) architectures, and reinforced my foundation in neural network optimization. A major output of this secondment was the development of a customizable chatbot that integrates RAG with continual learning, capable of retaining and applying knowledge from past interactions in a simulated environment.

B. Lecturer at AIDA (AI Doctoral Academy)

As part of my contribution to AIDA, I developed and presented a course on context-aware conversational agents, exploring the underlying architectures and challenges in creating scalable, adaptable AI systems. The course covered:

- Key components of conversational architecture, including natural language understanding (NLU), dialogue management, and response generation.
- Retrieval-Augmented Generation (RAG) frameworks, where conversational agents are augmented with external knowledge bases.
- Techniques for embedding and semantic search, enabling the agent to retrieve and use relevant information from large datasets effectively.

This course aimed to provide participants with a comprehensive understanding of the complexity of building conversational agents, addressing issues such as scalability, privacy, and data security in enterprise settings.

C. Participation in the Doctoral Symposium on Electronics, Telecommunications & Information Technology

In October 2023, I participated in the 1st Doctoral Symposium on Electronics, Telecommunications, and Information Technology (SDETTI), hosted by the National University of Sciences and Technologies Politehnica Bucharest. The symposium served as a platform to discuss advanced research topics within the field and provided an opportunity to share insights and developments with fellow researchers.

Paper Abstract: My paper, titled "Neural Networks and the Emergence of Learning," introduces a tri-component framework designed to facilitate the study of neural network behaviors, focusing particularly on the phenomenon of emergent learning. The framework is constructed using a combination of Java, Python, and JavaScript, allowing for comprehensive visualization and analysis of network dynamics.

The study employs a single-neuron model to explore key mathematical functions and learning algorithms, such as the Sigmoid activation function and backpropagation. Through this approach, the paper examines and critiques reductionist methodologies, advocating for integrative methods to better understand the intricate dynamics that characterize both artificial and natural learning systems.

Key Contributions:

- Development of a flexible framework that integrates Java, Python, and JavaScript components for cross-functional neural network analysis.
- Examination of emergent learning properties through single-neuron models, enabling targeted study of specific learning algorithms and activation functions.
- Advocacy for holistic approaches in studying neural networks, challenging reductionist views and supporting the need for multi-disciplinary perspectives.

Through this work, I aim to encourage further research into the fundamental aspects of neural networks and to explore how emergent learning can provide insights into both artificial intelligence and biological cognition.

D. ContinualBot Framework

ContinualBot is a conversational AI framework developed to support continuous learning and retrieval-augmented generation. Key features include:

- Impersonation capabilities allowing the bot to adapt its personality based on user preferences.
- Integration with various Large Language Models (LLMs) and vector databases, including Qdrant and ChromaDB, to support diverse use cases.
- Customizable configurations for applications in customer service, research, and more, with advanced user management to maintain privacy and personalization.

ContinualBot stands out for its modularity and flexibility, supporting integration with external plugins and real-time adaptation through continual learning algorithms. Its user management system ensures that knowledge and interaction histories are compartmentalized, thus supporting a range of personalized experiences.

E. Video Search Proof of Concept (PoC)

At FrancoTech Paris 2024, I introduced an innovative video-processing module as part of the ContinualBot framework. This system allows users to query video content interactively through text-based questions, leveraging multimodal AI techniques to produce a seamless, searchable video interface. The core of this module lies in its ability to segment videos into key frames, generate detailed descriptions for each frame using language models, and link these frames to precise timestamps, providing a robust framework for conversational video retrieval.

1) System Overview and Key Frame Extraction: The system starts by downloading the video from a given YouTube URL and processing it on an application server. An image processing module extracts video frames at configurable intervals, such as every 5 seconds, and applies a Distinct Key Frame Extraction algorithm. This algorithm identifies frames with substantial visual changes using metrics like Structural Similarity Index Measure (SSIM) and Histogram Difference,

ensuring that only unique, representative frames are retained. This approach reduces redundancy and isolates the most informative frames, which serve as pivotal reference points for subsequent querying and analysis.

2) Multimodal Analysis and Image-to-Text Conversion: Once key frames are identified, each frame undergoes a multimodal analysis using Generative AI neural networks based on Vision Transformer (ViT) and CLIP (Contrastive Language-Image Pretraining) architectures. This neural network generates high-dimensional embeddings that capture intricate features such as object presence, texture, and semantic content. Using contrastive learning, the model ensures that similar frames have closely positioned embeddings, which enhances retrieval accuracy when responding to user queries.

Alongside embeddings, an Image-to-Text Conversion module utilizes encoder-decoder transformer models to produce both concise and detailed descriptions for each frame. These descriptions are structured around key attributes—such as objects, actions, and scene context—using a predefined schema. This dual representation (embeddings and text descriptions) forms a comprehensive data layer that enables the system to respond to varied queries with high precision.

3) Storage in Vector Database and Retrieval Mechanism: The generated embeddings and frame descriptions are stored in a vector database optimized for similarity search, such as Qdrant. Each database entry is linked with the timestamp of the respective frame, ensuring temporal accuracy when retrieving specific moments. This database supports efficient querying through similarity metrics like cosine similarity or Euclidean distance, allowing users to retrieve frames that best match their queries. Temporal anchoring allows users to directly access relevant frames at precise points in the video, making the system highly responsive and reliable for interactive video querying.

4) User Interaction, Retrieval-Augmented Generation (RAG), and Response Generation: Users interact with the system through natural language prompts, which are processed by a Retrieval-Augmented Generation (RAG) framework. The framework first maps the user's query into the embedding space of the stored frames. Through vector similarity search, the system identifies frames whose content aligns with the query. These relevant frame descriptions are then combined with the user's prompt in a Prompt Augmentation module, ensuring that the generated responses are contextually accurate and coherent.

The augmented prompt is sent to a Large Language Model (LLM) trained for multimodal input, which generates a response grounded in the video content. Techniques like attention masking guide the LLM to focus on the most relevant sections of the prompt, providing users with precise answers accompanied by the timestamp of the matching frame. This direct link to the timestamp allows users to validate the response by viewing the exact video moment.

5) Applications and Potential Use Cases: This Video Search PoC demonstrates the feasibility and versatility of integrating video content with conversational AI. Its applications extend across various fields:

- **Surveillance and Security:** Rapid identification and retrieval of specific events or objects from extensive video footage, enhancing situational awareness and incident response.
- **Medical Imaging:** Supporting diagnostics by identifying key frames within sequences of medical imagery, aiding in analysis and treatment planning.
- **Manufacturing Quality Control:** Monitoring production line footage to detect defects or anomalies in real-time, ensuring high standards of quality and efficiency.
- **Sports Analytics:** Allowing coaches and analysts to query game footage for specific plays, formations, or player actions, enabling detailed analysis and strategy development.
- **Educational Content Search:** Assisting educators and learners by enabling quick access to relevant segments in instructional videos, enhancing the learning experience.

Through the integration of advanced multimodal AI, image processing, and RAG methodologies, this project sets a new standard for video-based conversational AI. The system's ability to transform static video content into a dynamic, interactive, and searchable format broadens the scope of conversational agents, positioning this technology as a powerful tool for applications that demand precision, flexibility, and high-level video understanding.

F. Video Frame-Based Information Retrieval Algorithm

Building on the Video Search PoC, this project aims to advance techniques for embedding individual video frames as searchable units, enabling efficient querying and interaction with video content. This project is central to expanding the capabilities of multimodal conversational AI, allowing agents to retrieve relevant video frames based on text-based questions and to interact fluidly with video and textual content alike.

1) Frame Processing and Embedding Generation: The process begins by segmenting videos into frames at configurable intervals. Each frame is analyzed to generate embeddings—high-dimensional vector representations that capture semantic information, such as objects, actions, and background context. These embeddings are generated using advanced image-to-text models, including encoder-decoder transformers, which produce concise yet descriptive representations of the frame's content. To minimize storage and enhance retrieval efficiency, the embeddings are compressed and stored with timestamps.

One of the primary challenges in frame processing is selecting the most relevant frames for embedding. A significant issue in video analysis is data redundancy, where multiple frames may contain similar or identical information, increasing storage requirements and reducing retrieval efficiency. To tackle this, a Distinct Key Frame Extraction algorithm is used, applying metrics such as Structural Similarity Index Measure (SSIM) and Histogram Difference. This process ensures that

only frames with significant visual changes—those that provide new information—are stored as key frames, enhancing the system's ability to identify relevant content rapidly.

2) Storage and Retrieval Using NoSQL Databases: Given the volume and complexity of video data, NoSQL databases are ideal for storing video frame embeddings and metadata. In this system, a vector-based NoSQL database, such as Qdrant or ChromaDB, is used to manage embeddings efficiently. NoSQL databases support high-dimensional vector storage and are optimized for similarity-based searches, which allows for fast querying of relevant frames based on user input. Each frame embedding is stored with its corresponding timestamp and a descriptive metadata entry, linking it to specific moments in the video.

Storing embeddings in a NoSQL database provides several advantages:

- **Scalability:** NoSQL databases handle large datasets and can scale horizontally to accommodate growing video libraries, essential for high-volume applications.
- **Flexible Schema:** NoSQL's flexible schema allows for diverse data types, accommodating video metadata, frame descriptions, and vector embeddings within the same database.
- **Optimized Retrieval:** NoSQL databases support similarity-based retrieval mechanisms, such as cosine similarity or Euclidean distance, which are essential for matching user queries with the most relevant video frames.

However, using NoSQL databases introduces challenges, particularly in efficiently indexing and querying high-dimensional embeddings. Vector searches, while effective, can be computationally expensive, particularly as the database grows. Indexing techniques like approximate nearest neighbor (ANN) search help address these issues, reducing retrieval times by approximating matches instead of exhaustive searches. This trade-off between retrieval speed and accuracy is an ongoing area of optimization.

3) Challenges in Identifying Relevant Frames: Identifying the most relevant frames in response to user queries is a nuanced task. A user might ask a question that requires locating frames with specific actions, objects, or contextual details. Since a video contains thousands of frames, matching queries with exact frame content is challenging due to semantic gaps between the user's language and the visual content.

The system addresses this by generating embeddings that capture frame semantics, allowing for indirect matching based on similarity. By embedding user queries in the same vector space as the frames, the system can perform a similarity search to find frames that align closely with the query context. However, this approach requires careful tuning of embedding models to ensure that they capture subtle context cues, such as distinguishing between similar-looking scenes with different actions.

Additional techniques like Retrieval-Augmented Generation (RAG) are also used, where retrieved frames serve as context to refine user queries before sending them to a large language model (LLM) for a response. This metaprompting approach enhances the accuracy of frame selection, enabling the LLM to generate answers based on specific frames that match the user's intent.

4) Applications and Future Directions: The ability to embed and search video frames with high precision has applications across multiple fields:

- **Surveillance and Security:** Efficient retrieval of events or objects from surveillance footage based on specific queries.
- **Sports Analytics:** Searching game footage for tactical formations or key plays, enabling in-depth analysis and coaching insights.
- **Medical Imaging:** Locating critical frames in diagnostic videos, aiding medical professionals in identifying key moments relevant to patient care.

Future work will focus on improving embedding techniques to enhance context retention, as well as optimizing NoSQL database indexing to ensure faster and more accurate retrieval. This project represents a significant step in making video data as accessible and searchable as text, broadening the potential for conversational AI in multimedia-rich applications.

G. Exploring AI Avatars and Digital Personas

In a related project, I developed an AI avatar with capabilities for real-time interaction, combining speech synthesis, voice cloning, and lip-sync animation. This avatar functions as a digital representation capable of impersonating specific personas, with potential applications in both social and professional contexts. Such avatars have far-reaching implications, from virtual meetings to personal branding, creating new possibilities for individuals and organizations to present dynamic, customizable digital identities.

As generative AI technologies advance, we approach a pivotal shift where creating highly realistic digital avatars is accessible to a broader audience. Currently, synthetic media tools are primarily the domain of developers, but rapid advancements in productization will make it possible for anyone to generate lifelike avatars, modify videos, and enhance photos in a matter of minutes. This technology will enable users to control how they appear in virtual spaces, interact on social media, and participate in various online environments, potentially replacing live appearances with AI-driven avatars.

1) Challenges in Authenticity and Verification: The ease of creating synthetic media raises critical challenges in verifying authenticity. When avatars can replicate real individuals with remarkable precision, discerning between a genuine presence and a digital impersonation becomes increasingly complex. This blurring of lines between real and synthetic content is likely to redefine the concept of authenticity in digital spaces. Social media platforms, content creators, and audiences may need to adopt new practices and verification tools to ensure that digital interactions reflect reality, fundamentally transforming the relationship between truth and virtual representations.

2) Digital Persona Development Process: The development of a fully realized AI avatar involves multiple stages, each demanding high fidelity and consistency. Starting with advanced editing of real photographs, the avatar is designed to maintain a distinct visual identity across varied settings. Key techniques in this process include:

- **Identity Embedding:** Establishing core attributes, or "identity embeddings," that define the avatar's unique appearance and personality, ensuring they persist across images.
- **Feature Extraction and Reinforcement:** Applying feature-extraction techniques to capture essential traits like facial symmetry, eye color, and hairstyle, with contrastive learning reinforcing these features.
- **Conditional Generation and Style Transfer:** Implementing style transfer techniques that allow the avatar to adapt to different contexts while retaining its core identity traits.

These techniques enable the creation of a consistent digital persona, capable of maintaining recognizable characteristics even in varying environments. The goal is to develop an AI-driven "digital actor" with a coherent identity, one that can transition fluidly across images and scenes without losing its essence.

3) Toward Interactive Digital Beings: The next phase involves integrating this avatar with the ContinualBot framework to introduce dynamic knowledge access and conversational capabilities in video format. This integration enables the avatar not only to appear lifelike but also to interact in real-time, responding to users with contextual awareness and personality. By combining photorealistic visuals, synthesized voice, and AI-driven behavioral responses, the avatar transcends static images, evolving into an immersive, interactive digital persona.

Future enhancements include adding emotional expression and personalized voice modulation, making it possible for the avatar to display a range of emotions and conversational tones. This step aims to bridge the gap between visual realism and human-like interaction, allowing digital personas to engage audiences on a more profound level. Potential applications span virtual influencers, digital brand ambassadors, and personalized customer service representatives, where a highly interactive, consistent digital presence is valuable.

4) Ethical Considerations and Future Directions: The development of photorealistic AI avatars brings both opportunities and ethical dilemmas. With widespread access to synthetic media tools, questions of accountability and responsible use arise. For instance, how should digital impersonation be regulated, especially when avatars can convincingly replicate real individuals? Additionally, this technology could impact social norms, as virtual identities become indistinguishable from physical presence.

To address these concerns, ongoing research must explore methods for verifying authenticity in synthetic content. New standards for transparency and disclosure may be necessary, particularly as avatars are used in professional and social contexts. Ensuring that users understand the nature of these digital interactions will be crucial in preserving trust and integrity in an increasingly synthetic media landscape.

As we progress toward a future where digital personas are as interactive and nuanced as real individuals, we enter a new era of human-AI collaboration. This exploration of AI avatars marks only the beginning of what is possible, hinting at a future where virtual and real worlds intersect seamlessly. This ongoing work in digital persona development serves as a foundation for future innovations in AI-driven identity, immersive media, and human-computer interaction.

H. GraphRAG Survey

My current work on the GraphRAG survey focuses on integrating graph databases with Retrieval-Augmented Generation (RAG) methods to enhance knowledge retrieval in conversational AI. This research explores the application of graph-based architectures as an alternative to traditional vector-based retrieval systems, especially for handling complex queries that require relational context.

The survey includes a comprehensive review of foundational concepts, such as chatbots, multimodal interfaces, and continual learning, along with key terminology relevant to GraphRAG. By categorizing existing methods, applications, and architectural frameworks, this survey establishes a taxonomy of approaches shaping conversational AI. Additionally, I am examining the potential impact of quantum computing on GraphRAG's efficiency and scalability, with an eye toward future advancements that could significantly enhance knowledge retrieval capabilities.

I. Participation in SOL Projects

In addition to my primary research, I am also contributing to three SOL projects aimed at advancing artificial intelligence applications:

- **SOL5/2024:** Advanced Integrated System for Vehicle Identification Using Multiple Recognition/Confirmation Elements Based on Artificial Intelligence (AI).
- **SOL10/2024:** Toolset for Processing and Linguistic Analysis for the Romanian Language (RoNLP).
- **SOL12/2024:** Detection of Relationships Between Entities in Unstructured and Structured Data Sets (DeteRel).

V. Results and Future Work

The accomplishments include a neural network framework, open-source contributions, and a doctoral symposium article. Future work will involve refining continual learning techniques within conversational AI, exploring novel applications for RAG, and addressing ethical challenges in AI impersonation. Integration of video analysis with graph-based retrieval methods will be further explored, with potential use cases in customer service, education, and content management.

Acknowledgments

Special thanks to the AI Multimedia Laboratory, CAMPUS Research Institute, CEA (Commissariat à l'énergie atomique et aux énergies alternatives) in France, AI4Media, the University of Medicine Carol Davila Bucharest, and the Doctoral

School of Electronics, Telecommunications, and Information Technology at the National University of Science and Technology Politehnica Bucharest (formerly known as University Politehnica of Bucharest) for their invaluable support.

I am especially grateful to my PhD supervisor, Prof. Dr. Eng. Bogdan Ionescu, for his guidance and mentorship, and to Prof. Dr. Eng. Adrian Popescu, my Senior Fellow from CEA/Paris-Saclay University, for his invaluable insights and support throughout my research journey.

REFERENCES

[1] K. Clark, "Computing Neural Network Gradients," unpublished, 2017.
[2] R. Salakhutdinov, "Deep learning," in *KDD*, 2014, p. 1973.
[3] N. Bacaër, "Verhulst and the logistic equation (1838)," in *A short history of mathematical population dynamics*, Springer, 2011, pp. 35–39.
[4] C. Nwankpa, W. Ijomah, A. Gachagan, and S. Marshall, "Activation functions: Comparison of trends in practice and research for deep learning," *arXiv preprint arXiv:1811.03378*, 2018.
[5] P. Clayton, "Conceptual Foundations of Emergence Theory," in *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion*, Oxford University Press, 2008.
[6] M. Buscema, "Back Propagation Neural Networks," *Substance use & misuse*, vol. 33, pp. 233-70, Feb. 1998.
[7] M. A. Nielsen, *Neural networks and deep learning*, Determination Press, San Francisco, CA, USA, 2015.
[8] G. Chaudhary, "Artificial Intelligence: The Personhood Conundrum," *Artificial Intelligence and Law*, 2021.
[9] D. Lewis, "General semantics," in *Montague grammar*, Elsevier, 1976, pp. 1–50.
[10] A. W. Huttunen, G. K. Adams, and M. L. Platt, "Can self-awareness be taught? Monkeys pass the mirror test—again," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 114, no. 13, pp. 3281–3283, 2017.
[11] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *arXiv preprint arXiv:2302.00487*, 2023.