Movie Question Answering

It has happened to all of us: watching a movie and not understanding a scene, or not remembering some characters or their relations. Now imagine a system that can understand and reason about movies in the same way that humans do. This system shall be capable of understanding a movie and answer our questions regarding the plot, the connection among characters, and it can even be capable of reasoning about past events and predict upcoming ones [Kukleva20,Wu21]. Movie Question Answering systems can provide a more natural and intuitive way for users to interact with movies, making it easier and more convenient to access information and perform movie-related tasks. Most computer vision approaches focus on answering specific questions about visual data [Yang22,Chen22], typically without long-term reasoning and without taking into consideration the particularities of the domain, such as jumps in time, cuts or intentions [Brown21,Pardo21,Tapaswi16].

This project falls into the cutting-edge video understanding domain and aims at answering questions for movies. For this, we will formulate the problem of movie question answering. Specifically, given long-range video scenes, we will propose state-of-the-art methods designed to answer questions about movies, such as plot details, character information, and trivia by exploiting multimodal cues (video, audio, text). This could be done using recent advances in Large Language Models [Brown20,ChatGPT] to parse user queries and retrieve relevant visual information.

Practical information

The internship is co-supervised by Vicky Kalogeiton, Assistant Professor at the VISTA team at the Computer Science Department (LIX) of ´Ecole Polytechnique and Ivan Laptev, senior researcher and head of the WILLOW project-team at Inria Paris. The internship is co-located at Ecole Polytechnique and Inria Paris, and we are open to partial remote work.

Requirements

We are looking for strongly motivated candidates with an interest in machine learning and computer vision.

- In pursuit of Masters degree in a relevant field (CS, Informatics, ...) with strong background in applied mathematics
- Excellent programming skills and proven experience in Python
- Hands-on experience with deep learning frameworks (PyTorch) is a plus
- Experience in computer vision is a plus
- High level of innovation and motivation
- Communication skills in English

We offer

- Two highly-technical and international teams composed of professors, researcher, students, and engineers
- PhD opportunity: Depending on the candidate's qualifications, results and the mutual match, the project can lead to a joint PhD in video understanding

Starting date: March 2023
Duration: 5-6 months
Application Deadline: 21 January 2023

Application
To apply, please contact Vicky Kalogeiton at vicky.kalogeiton@polytechnique.edu and Ivan Laptev at ivan.laptev@inria.fr with '[Internship/PhD application: Movie Understanding]' in the subject line, and please provide (1) a CV, (2) your graduate/undergrad transcripts, and (3) a short statement of research interests (a couple of paragraphs). If needed, we may ask for two references. We particularly encourage applications from women, and from underrepresented groups in academia.

References

[Kukleva20] Kukleva A., Tapaswi M., Laptev I. Learning interactions and relationships between movie characters. In CVPR 2020
[Pardo21] Pardo A., Caba F., Alcázar J.L., Thabet A.K., Ghanem B. Learning to Cut by Watching Movies. In ICCV 2021
[Brown21] Brown A., Kalogeiton V., Zisserman A. Face, Body, Voice: Video Person-Clustering with Multiple Modalities. In ICCV-W 2021
[Tapaswi16] Tapaswi M., Zhu Y., Stiefelhagen R., Torralba A., Urtasun R., Fidler S. MovieQA: Understanding stories in movies through question-answering. In CVPR 2016
[Wu21] Wu C.Y., Krahenbuhl P. Towards long-form video understanding. In CVPR 2021
[ChatGPT] ChatGPT: Optimizing Language Models for Dialogue, OpenAI 2022
[Brown20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, A. Neelakantan et al. Language models are few-shot learners. In NeurIPS 2020
[Yang22] A. Yang, A. Miech, J. Sivic, I. Laptev and C. Schmid. Zero-Shot Video Question Answering via Frozen Bidirectional Language Models. In NeurIPS 2022
[Chen22] J. Chen, H. Guo, K. Yi, B. Li, M. Elhoseiny. VisualGPT: Dataefficient adaptation of pretrained language models for image captioning. In CVPR 2022