Ego-centric Videos and Natural Language Queries

Leone Aurora s334258 | Narese Michele s329892 | Racca Riccardo s315163

Abstract

This report aims to tackle the task of action recognition and identification in egocentric videos, that is, identifying a timeframe where the answer to a natural language query can be seen within the provided video and then exploiting this knowledge in order to build a textual answer.
Different architectures were implemented and trained on various pre-extracted features of Ego4D dataset [1] and multiple metrics were used to compare the performance of different models with the benchmark, in order to obtain a robust model that is employable in the next step.
Subsequently, a video question-answering pipeline was built by leveraging the results provided by the previous architectures and using them as input for a VLM (Video Language Model) to not only retrieve a time interval, but also provide a textual answer. Finally, an ensemble method was introduced to obtain better predictions while exploiting the, partially wasted, computational effort needed to build and compare different models.

References

[1] Eugene Byrne Zachary Chavis Antonino Furnari Rohit Girdhar Jackson Hamburger Hao Jiang Miao Liu Xingyu Liu et al Kristen Grauman, Andrew Westbury. Ego4d: Around the world in 3,000 hours of egocentric video. 2022. 1