I. Title

Feature Extraction for Video Understanding in Group Interaction Scenario using Transformer based Architectures

II. General Objective

Feature extraction is a challenging computer vision problem which targets extracting relevant information from raw data in order to reduce dimensionality and capture meaningful patterns. When this needs to be done in a dataset and task invariant way, it is referred to as general feature extraction. This is a crucial step in machine learning pipelines and popular methods like VideoSwin and VIdeoMAE work well for the task of action recognition and video understanding. However, these works and also the datasets that they are tested on, like Something-Something and Kinetics, fail to capture information about interactions in daily life.

Towards this research direction, several methods [] have been proposed to model these complex fine grained interactions using datasets like UDIVA, MPII Group Interactions and Epic-Kitchen. Those datasets encompassing real-world challenges share the following characteristics: Firstly, there is rich multimodal information available where each modality provides important information relevant to the labels. Secondly, there is a lot of irrelevant information that has to be ignored as deep learning models easily identify patterns that are coincidental (local minima). For example, the colour of the T-shirt could be used to assign a certain personality score to someone if by coincidence the majority of the extrovert people are wearing warm colours. Lastly, the videos in these datasets are generally very long.

So, the main question is how to extract general features from multimodal data with a lot of noise in the form of irrelevant information?

Typical situations that we would like to monitor are daily interactions, responses and reactions and analyse cause and effect in behaviour (it could be human-human interaction or human-object interaction).

The system we want to develop will be beneficial for all tasks requiring focus on interactions. Specifically, healthcare for psychological disorders - general feature extraction will allow deep learning models to assist in various subtasks involved in the diagnosis process.

III. PhD objective

In this work, we would like to go beyond existing computer vision deep learning models and introduce ways to extend them to utilise information from new modalities. Also, to identify ways to focus on relevant information for interactions in the input. The system should also take into account the long temporal duration of videos in the datasets in this domain. These have to be done in a flexible way, so that there is minimal change to the original model and hence the original model's trained weights are useful too.

Existing methods [] have mostly focused on modelling the variation of visual cues pertinent to the classes provided for video classification tasks. Though they perform these tasks well, changes in the recording setting or addition of noise in the form of irrelevant background information makes it hard for these models to perform well. So, for obtaining a general feature extractor, the models have to be modified to accommodate for these shortcomings.

In this work, we focus on two things - first, in group interaction scenarios, utilising all available information to obtain relevant features for multiple downstream tasks while ignoring irrelevant background information. The second, efficient transfer learning for a new recording paradigm. This can include new modalities, change in recording settings, and different downstream tasks. The first objective can be tackled by forcing attention in transformers to attend to relevant parts of the input and having more specific architectures for modelling interactions. The second objective caters to a more general problem of parameter efficient transfer learning which has benefited from works like adapters, prefix tuning and prompt tuning [refs for all three]. These have worked well for the field of NLP and have been adapted to computer vision, but work only for specific cases. The theory behind these techniques can be utilised to develop new methods that serve the second objective of this work.

Large pretrained vision models and their architectures can be used as the backbone for this work.

This work will be conducted under the GAIN project which aims to further research in Georgia along with France. The evaluation of proposed framework should be performed on public datasets which contain group interactions with additional background information like UDIVA, MPIIGroupInteractions and also popular video classification public datasets with everyday activities like Charades and Epic-Kitchen.