Video Action Transformer Network

Rohit Girdhar[1]* João Carreira[2] Carl Doersch[2] Andrew Zisserman[2,3]
[1]Carnegie Mellon University [2]DeepMind [3]University of Oxford
http://rohitgirdhar.github.io/ActionTransformer

Abstract

We introduce the Action Transformer model for recognizing and localizing human actions in video clips. We repurpose a Transformer-style architecture to aggregate features from the spatiotemporal context around the person whose actions we are trying to classify. We show that by using high-resolution, person-specific, class-agnostic queries, the model spontaneously learns to track individual people and to pick up on semantic context from the actions of others. Additionally its attention mechanism learns to emphasize hands and faces, which are often crucial to discriminate an action – all without explicit supervision other than boxes and class labels. We train and test our Action Transformer network on the Atomic Visual Actions (AVA) dataset, outperforming the state-of-the-art by a significant margin using only raw RGB frames as input.

1. Introduction

In this paper, our objective is to both localize and recognize human actions in video clips. One reason that human actions remain so difficult to recognize is that inferring a person's actions often requires understanding the people and objects around them. For instance, recognizing whether a person is 'listening to someone' is predicated on the existence of another person in the scene saying something. Similarly, recognizing whether a person is 'pointing to an object', or 'holding an object', or 'shaking hands'; all require reasoning jointly about the person and the animate and inanimate elements of their surroundings. Note that this is not limited to the context at a given point in time: recognizing the action of 'watching a person', after the watched person has walked out of frame, requires reasoning over time to understand that our person of interest is actually looking at someone and not just staring into the distance.

Thus we seek a model that can determine and utilize such contextual information (other people, other objects) when determining the action of a person of interest. The Transformer architecture from Vaswani *et al.* [43] is one suitable model for this, since it explicitly builds contextual support for its representations using self-attention. This architecture has been hugely successful for sequence modelling tasks compared to traditional recurrent models. The question, however, is: how does one build a similar model for human action recognition?

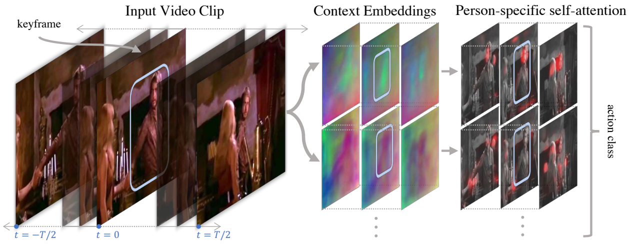

Figure 1: **Action Transformer in action.** Our proposed multi-head/layer Action Transformer architecture learns to attend to relevant regions of the person of interest and their context (other people, objects) to recognize the actions they are doing. Each head computes a clip embedding, which is used to focus on different parts like the face, hands and the other people to recognize that the person of interest is 'holding hands' and 'watching a person'.

Our answer is a new video action recognition network, the **Action Transformer**, that uses a modified Transformer architecture as a 'head' to classify the action of a person of interest. It brings together two other ideas: (i) a spatio-temporal I3D model that has been successful in previous approaches for action recognition in video [7] – this provides the base features; and (ii) a region proposal network (RPN) [33] – this provides a sampling mechanism for localizing people performing actions. Together the I3D features and RPN generate the query that is the input for the Transformer head that aggregates contextual information from other people and objects in the surrounding video. We describe this architecture in detail in section 3. We show in section 4 that the trained network is able to learn both to track individual people and to contextualize their actions in terms of the actions of other people in the video. In addition, the transformer attends to hand and face regions, which is reassuring because we know they have some of the most relevant features when discriminating an action. All of this is

*Work done during an internship at DeepMind