MS-TCT: Multi-Scale Temporal ConvTransformer for Action Detection
Rui Dai, Srijan Das, Kumara Kahatapitiya, Michael S. Ryoo, François Brémond
Inria, Université Côte d’Azur, Stony Brook University

Introduction

- **Objective**: To capture both short-term and long-term temporal information efficiently.

Contribution

- **An effective & efficient** ConvTransformer for modeling complex temporal relations in untrimmed videos.
- **A new branch** to learn the position relative to action instance-center.
- **Improve** state-of-the-art on three densely-labelled action datasets: Charades, TSU, MultiTHUMOS.

Objective

- **Problem**: Understanding long untrimmed videos for the task of action detection.
- **Challenges**: Actions may occur densely in real-world. Videos contain complex temporal relations, including composite or co-occurring actions.

Framework

1. **Visual Encoder** encodes a preliminary video representation.

2. **Temporal Encoder**
   - models the temporal relations at different temporal scales.
   - Multi-Head Attention models global temporal relation.
   - Temporal Convolution models local temporal relation.

3. **Temporal Scale Mixer** combines multi-scale temporal representations.

4. **Classification module**
   - Multi-label classification for each token (BCE loss)
   - Instance center location prediction for each token (Focal loss).

Relative Temporal Position

- **Constructed by a Gaussian filter** (instance center and its duration).
- Indicates the relative temporal position w.r.t. to the action instance center at any given time.

Experiment

- **Qualitative Study**
  - Detection Visualization
  - MS-TCT predicts action instances more precisely compared to PDAN [1].

- **Ablation Study**
  - Each component in MS-TCT is instrumental for the task of action detection.

- **Comparaison to SoTA methods**
  - Frame-level mAP
  - Action conditional metrics

Reference: