Temporal Action Localization in Untrimmed Videos via Multi-stage CNNs
Zheng Shou, Dongang Wang, Shih-Fu Chang
Columbia University

Overview:
- **Motivation:** Localizing actions in untrimmed long videos, such as those in surveillance, can save tremendous time and costs.
- **Problem definition:** Given an untrimmed video, when does each specific action instance start and end, and which action class does it belong to?

Framework of Segment-CNN:
- Generate candidate segments via multi-scale sliding window (16, 32, 64, 128, 256, 512 frames with 75% overlap in time)
- Segment-based CNNs:
  - Network architecture: C3D [Tran et al.]
  - Training: Proposal → Classification → Localization (init by the trained cls model)
  - Prediction: Proposal → Localization
  - Conduct post-processing steps such as non-maximum suppression

Training details of Segment-CNN:
- **Proposal network:**
  - Two categories (background and being action)
  - Identify candidate segments that may contain actions
- **Classification network:**
  - Train N+1 multi-classification model using Softmax Loss:
  
- **Localization network:**
  - Motivation: NMS might remove segment of small score but of large overlap with ground truth.

Experimental results on THUMOS14:
- **Comparison with state-of-the-art systems:**
  - Mainly use FV encoding of iDTF and frame-level CNN features
  - Some work leveraged video-level classifiers trained on multiple features

- **Efficiency analysis:**
  - Speed: GTX980. ~1s per batch. ~0.5s to process 1s video.
  - Storage: less than 1 GB. do not need to cache intermediate features.

- **Impact of individual networks:**
  - Proposal: S-CNN (19.0%) VS S-CNN w/o proposal (17.1%)
  - Classification: ...
  - Localization: ...

Contact: zheng.shou@columbia.edu
Code available: https://github.com/zhengshou/scnn

Experimental setup:
- **Evaluation metrics:**
  - Regard as retrieval problem and evaluate mAP
  - What is a correct action instance prediction: correct category prediction + IoU with ground truth instance larger than the evaluation threshold (set to 0.5 unless specified)
  - Redundant detections are not allowed
- **THUMOS14 temporal action detection task:**
  - 20 sports categories
  - Training data: 2,755 trimmed videos from UCF101 + 1,010 untrimmed YouTube videos of 3,007 instances
  - Test data: 213 untrimmed videos of 3,358 action instances

Examples:
- ![Example Data](image)