Developing Motion Code Embedding for Action Recognition in Videos
Maxat Alibayev, David Paulius, and Yu Sun

Department of Computer Science & Engineering, University of South Florida, Tampa, FL, USA.

Introduction
- We propose a new embedding of manipulations that encodes actions based on its salient features known as motion codes.
- Such features are mechanical characteristics of manipulation that are relevant to robotics, including contact and trajectory descriptors.
- By including motion codes in a verb classification framework, we can improve their overall performance.
- We demonstrate this improvement on the baseline I3D model.

Motion Taxonomy
- Motion taxonomy – hierarchical structure of salient features that are relevant to robotics.
- We can build a motion classifier to classify these features from a given input demonstration video.

Methodology
- The verb classification framework combines a probability distribution for verb classes with motion code components into a single feature vector that is passed through an MLP to output a verb prediction.
- Optionally, the framework can use extracted semantic features using objects-in-action used in manipulation.
- Special objective function is defined to train predictors of individual motion code components, which is given as:

\[ L_N = \sum_{k=1}^{K} \sum_{l=1}^{L} m_l \log(f_l^{(k)}(v)) \]

- \( f_l^{(k)} \): classifier for \( k \)-th motion code component
- \( \lambda_k \): constant weight
- \( m_l \): \( l \)-th element of ground-truth vector one-hot vector for \( k \)-th motion code component

Experimental Results
- Our objective is to demonstrate the significant improvement obtained once motion codes are integrated in the verb classification pipeline.
- We evaluate the following models:
  - \( V'_k \): I3D with visual features only
  - \( V'_k, z \): I3D with visual features and nouns
  - \( \psi(V'_k, M'_k) \): our model (I3D + Motion code embedding)
  - \( \psi(V'_k, M'_k, z) \): our model with nouns (I3D + Motion code embedding)
- Our testing results show that motion code prediction improves verb classification as compared to baseline model.
- Our validation results also show that verb classification can be improved, granted that motion code prediction is accurate.

Implementation
- For visual feature extraction, we used **Inflated 3D ConvNet** (or I3D) trained on Kinetics dataset [Carreira et al. 2017].
  - Both RGB and optical flow were used.
- Video segments were obtained from EPIC-KITCHENS dataset [Damen et al. 2018] dataset:
  - 3,528 video segments – 2,742 training, 786 validation
  - Sampled 1,517 test videos with 33 verb classes and 32 unique motion codes.
- Training details and parameters:
  - Trained for 50 epochs with Adam optimizer
  - Convolutional layers frozen for first 3 epochs
  - Learning rate set to 0.0003, decreasing by 40% per 5 epochs
  - MLP trained with learning rate of 0.0005 for 200 epochs
  - \( \psi(M'_k) \): classifier
  - Video frames sampled to 6 frames / sec
  - 12 consecutive frames randomly sampled, cropped and horizontally flipped as done by NTU-CML-MIRA in EPIC-KITCHENS 2019 challenge.

Acknowledgement
This material is based upon work supported by the National Science Foundation under Grant Nos. 1812933 and 1910040.