International Journal of Computer Vision (2023) 131:3060–3073
https://doi.org/10.1007/s11263-023-01846-2

Universal Prototype Transport for Zero-Shot Action Recognition and Localization

Pascal Mettes[1]

Received: 26 July 2022 / Accepted: 3 July 2023 / Published online: 19 July 2023

Abstract

This work addresses the problem of recognizing action categories in videos when no training examples are available. The current state-of-the-art enables such a zero-shot recognition by learning universal mappings from videos to a semantic space, either trained on large-scale seen actions or on objects. While effective, we find that universal action and object mappings are biased to specific regions in the semantic space. These biases lead to a fundamental problem: many unseen action categories are simply never inferred during testing. For example on UCF-101, a quarter of the unseen actions are out of reach with a state-of-the-art universal action model. To that end, this paper introduces universal prototype transport for zero-shot action recognition. The main idea is to re-position the semantic prototypes of unseen actions by matching them to the distribution of all test videos. For universal action models, we propose to match distributions through a hyperspherical optimal transport from unseen action prototypes to the set of all projected test videos. The resulting transport couplings in turn determine the target prototype for each unseen action. Rather than directly using the target prototype as final result, we re-position unseen action prototypes along the geodesic spanned by the original and target prototypes as a form of semantic regularization. For universal object models, we outline a variant that defines target prototypes based on an optimal transport between unseen action prototypes and object prototypes. Empirically, we show that universal prototype transport diminishes the biased selection of unseen action prototypes and boosts both universal action and object models for zero-shot classification and spatio-temporal localization.

1 Introduction

This paper addresses the problem of recognizing actions in videos. Foundational deep network approaches performed action recognition through frame-level fusion (Karpathy et al., 2014), two-stream networks (Simonyan & Zisserman, 2014; Feichtenhofer et al., 2016), and 3D convolutional networks (Carreira & Zisserman, 2017). Building upon these approaches, recent works have shown great recognition capabilities through *e.g.,* slow-fast architectures (Feichtenhofer et al., 2019), separated 3D convolutions (Tran et al., 2019), and video transformers (Arnab et al., 2021). Such deep networks require large amounts of video material for training and efforts have been made to meet those video demands, such as ActivityNet (Caba Heilbron et al., 2015), EPIC Kitchens (Damen et al., 2018), Kinetics (Carreira & Zisserman, 2017), HowTo100M (Miech et al., 2019), and EGO4D (Grauman et al., 2022) to name a few. While such datasets increase the coverage of the action space, we seek to recognize actions even when no examples are available during training.

In zero-shot action recognition, many works have outlined approaches that mirror successes in the image domain, for example by using attributes (Liu et al., 2011; Gan et al., 2016b) or feature synthesis (Mishra et al., 2020) to transfer knowledge from seen to unseen actions. More recently, state-of-the-art results have been achieved by taking a universal learning perspective, where large-scale models are trained to map input videos to a shared semantic space occupied by both seen and unseen categories. In the first universal perspective, large-scale networks train a mapping from videos to a semantic space on hundreds of seen actions from *e.g.,* ActivityNet (Zhu et al., 2018) or Kinetics (Brattoli et al., 2020; Pu et al., 2022). For a target dataset with unseen actions, zero-shot inference is directly possible through a nearest neighbour search between the video mappings and the embeddings of unseen actions. In the second perspective, networks are trained on thousands of objects (Jain et al., 2015; Mettes et

Communicated by Dima Damen.

✉ Pascal Mettes
P.S.M.Mettes@uva.nl

[1] Universiteit van Amsterdam, Amsterdam, the Netherlands