SPOTLIGHT

Support-Set Bottlenecks for Video-Text Representation Learning

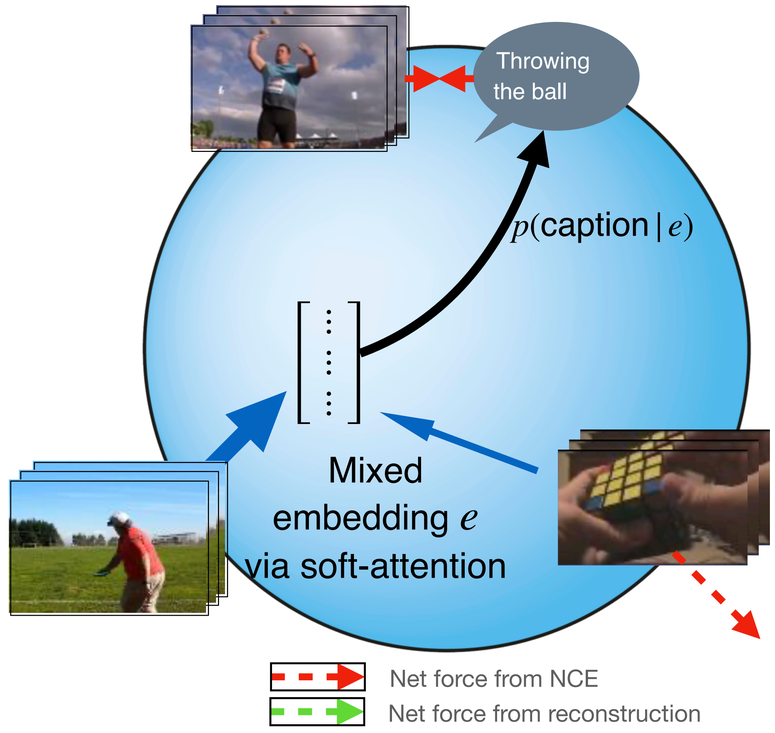
Mandela Patrick*, Po-Yao Huang*, Yuki M. Asano*,
Florian Metze, Alexander Hauptmann,
Joao Henriques, Andrea Vedaldi

FACEBOOK · UNIVERSITY OF OXFORD · CARNEGIE MELLON UNIVERSITY PITTSBURGH PENNSYLVANIA 1900 · ICLR

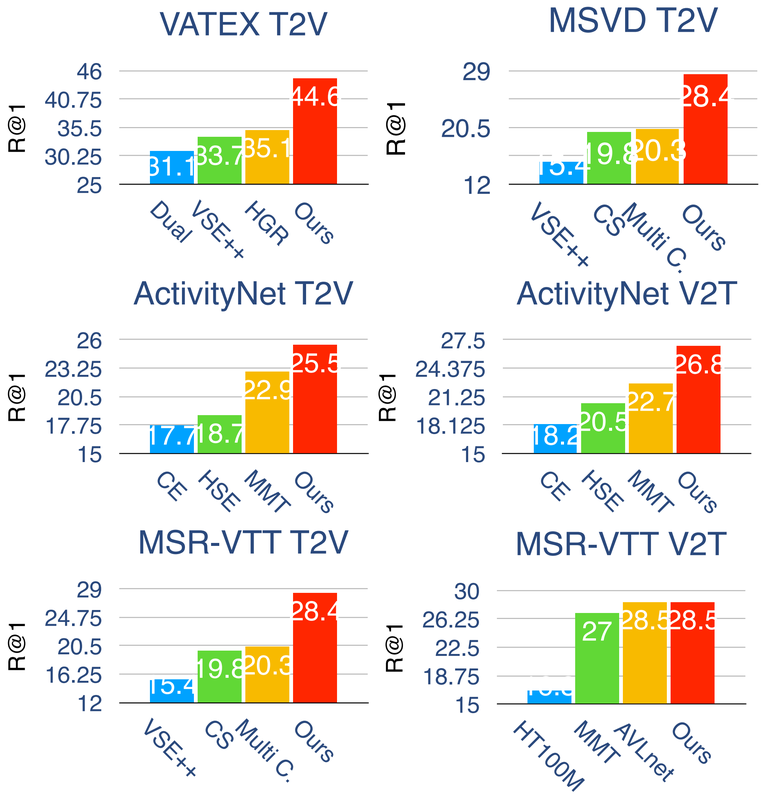
TL;DR: We improve the contrastive loss for video-text learning by using a generative loss.

Abstract. The dominant paradigm for learning video-text representations -- noise contrastive learning -- increases the similarity of the representations of pairs of samples that are known to be related, such as text and video from the same sample, and pushes away the representations of all other pairs. We posit that this last behaviour is too strict, enforcing dissimilar representations even for samples that are semantically-related -- for example, visually similar videos or ones that share the same depicted action. In this paper, we propose a novel method that alleviates this by leveraging a generative model to naturally push these related samples together: each sample's caption must be reconstructed as a weighted combination of other support samples' visual representations. This simple idea ensures that representations are not overly-specialized to individual samples, are reusable across the dataset, and results in representations that explicitly encode semantics shared between samples, unlike noise contrastive learning. Our proposed method outperforms others by a large margin on MSR-VTT, VATEX and ActivityNet, and MSVD for video-to-text and text-to-video retrieval.

Support-set based reconstruction loss alleviates this

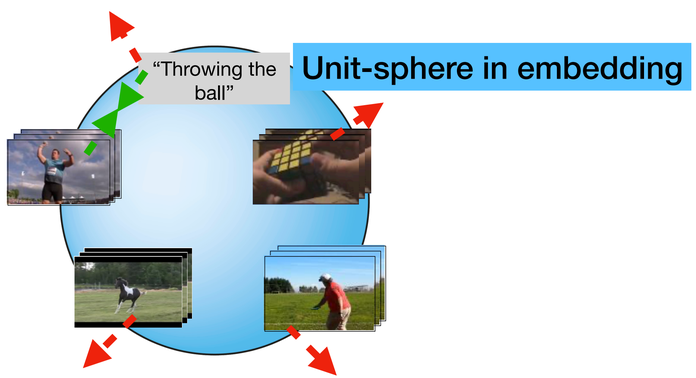


State of the art video-to-text (v2t) and text-to-video (t2v) retrieval performance across many datasets

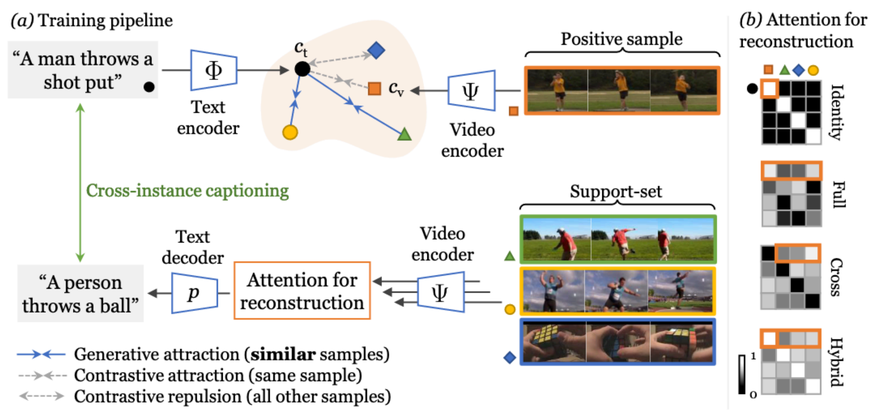


Contrastive Learning has false negatives

Contrastive learning works by treating a video and its caption as a positive pair that are learned to be close in embedding space. Each noise sample is treated as a negative and repelled, leading to potentially faulty embeddings that while sharing semantics, are far in the embeddings space (top-left and bottom-right video)



Overall method:



(a) Our cross-modal framework with the discriminative (contrastive) objective and the generative objective. The model learns to associate video-text pairs in a common embedding space with text and video encoders (top). Meanwhile, the text must also be reconstructed as a weighted combination of video embeddings from a support-set (bottom), selected via attention, which enforces representation sharing between different samples. **(b)** Weights matrices (attention maps) used in each cross-captioning objective. "Cross" works best, as requires the model to obtain more information from the support-set.

Soft-attention successfully focusses on relevant support-set items