Zero-Shot Temporal Action Detection via Vision-Language Prompting

Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, Tao Xiang

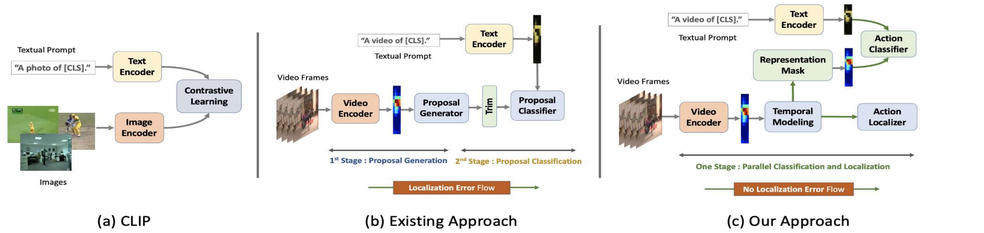
s.nag@surrey.ac.uk https://github.com/STALE

Introduction

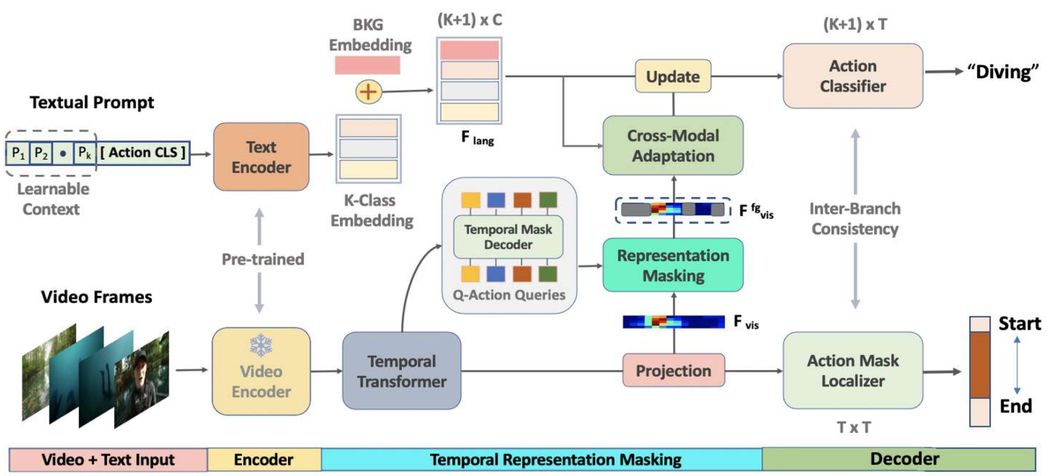
Background: Temporal action detection (TAD) aims to identify the temporal interval (i.e., the start and end points) and the class label of all action instances in an untrimmed video.

Motivation: Collecting and annotating a large training set for each class of interest is costly and hence unscalable. Zero-shot TAD (ZS-TAD) resolves this obstacle by enabling a pre-trained model to recognize any unseen action classes. Inspired by the success of zero-shot image classification aided by vision-language (ViL) models such as CLIP, we aim to tackle the more complex TAD task. An intuitive method is to integrate an off-the-shelf proposal detector with CLIP style classification. However, due to the sequential localization and classification design, it is prone to localization error propagation. To overcome this problem, in this paper we propose a novel zero-Shot Temporal Action detection model via Vision-LanguagE prompting (STALE). Such a novel design effectively eliminates the dependence between localization and classification by breaking the route for error propagation in-between.

Contributions: (1) We investigate the under-studied yet critical problem of how to capitalize large pretrained ViL models for zero-shot temporal action localization (ZS-TAD) in untrimmed videos.; (2) We present a novel one-stage model, STALE, featured with a parallel classification and localization design interleaved by a class-agnostic representation masking for zero-shot transfer to unseen classes; (3) SOTA performance on ActivityNet and THUMOS.



Model Architecture



Vision-Language for Dense Detection

Given an image, and a textual prompt like "A photo of [CLS]" one can then use CLIP to compute the similarity scores for the classification. However extending this to dense detection task is non-trivial. (1) **Firstly**, how to leverage the visual-language pre-trained model in dense prediction tasks is a barely studied problem especially in zero-shot setup (2) **Secondly**, transferring the knowledge from CLIP to dense prediction is more difficult than classification tasks, due to the substantial task discrepancy involved. The pretraining focuses on global representation learning of both images and texts, which is incompatible for local pixel-level outputs as required in dense downstream tasks.

Language Guided Temporal Action Detection

Visual-Language Embedding: We use frozen CLIP pre-trained image encoders for extracting video frame features. To add temporal consistency, we add a transformer encoder on top to obtain *Video Features*. For textual embedding, we use a standard CLIP pre-trained Transformer with learnable prompt. Since detection involves background we learn a separate embedding along with text embedding.

Representation Masking: We introduce a novel class agnostic representation masking concept for enabling the usage of a Vision-Language model for ZS-TAD. This module masks out the foreground feature representation.

Cross-Modal Adaption : Intuitively, integrating the descriptions of visual contexts is likely to enrich the text representation. For example, "a video of a man playing kickball in a big park" is richer than "a video of a man playing". This motivates us to investigate how to use visual contexts to refine the text features. Specifically, we leverage the contextual-level visual feature to guide the text feature to adaptively explore informative regions of a video.

TAD Decoders: Our TAD head is featured with parallel classification and mask prediction as detailed below

(a) `Action Classifier:` In the standard training process of CLIP, the global feature is normally used during contrastive alignment. In general, it estimates the snippet-text score pairs by taking the average pooling of snippet features and then uses it with the language features. However, this formulation is unsuitable for dense classification tasks like TAD. Hence, we use unpooled textual and visual features for snippet-level classification.

(b) `Action Localizer:` In parallel to the classification stream, this stream predicts 1-D masks of action instances across the whole temporal span of the video. Since the 1-D masks are conditioned on the temporal location, we exploit dynamic convolution to model this.

Main Results

ZS-TAD (Open-Set) Results on ActivityNetv1.3 and THUMOS14

Train Split	Methods	THUMOS14						ActivityNet v1.3			
		0.3	0.4	0.5	0.6	0.7	Avg	0.5	0.75	0.95	Avg
75% Seen 25% Unseen	B-II	28.5	20.3	17.1	10.5	6.9	16.6	32.6	18.5	5.8	19.6
	B-I	33.0	25.5	18.3	11.6	5.7	18.8	35.6	20.4	2.1	20.2
	EffPrompt	39.7	31.6	23.0	14.9	7.5	23.3	37.6	22.9	3.8	23.1
	STALE	**40.5**	**32.3**	**23.5**	**15.3**	**7.6**	**23.8**	**38.2**	**25.2**	**6.0**	**24.9**
50% Seen 50% Unseen	B-II	21.0	16.4	11.2	6.3	3.2	11.6	25.3	13.0	3.7	12.9
	B-I	27.2	21.3	15.3	9.7	4.8	15.7	28.0	16.4	1.2	16.0
	EffPrompt	37.2	29.6	**21.6**	**14.0**	**7.2**	21.9	32.0	19.3	2.9	19.6
	STALE	**38.3**	**30.7**	21.2	13.8	7.0	**22.2**	**32.1**	**20.7**	**5.9**	**20.5**

Ablation Studies

Analysis of localization error propagation

Metric	mAP 0.5	mAP Avg
Baseline-I(B-I)		
GT proposals	56.2	47.1
Predicted proposals	34.8	19.9
STALE		
GT masks	53.6	42.3
Predicted masks	38.2	24.9

Analysis of Representation-Masking on 75% seen split on ActivityNet

Masking Method	# Queries	mAP 0.5	mAP Avg
No Mask	-	21.7	10.9
GT-Mask	-	54.3	35.8
1-D CNN	-	33.5	20.1
Maskformer	5	37.5	24.4
	20	**38.2**	**24.9**
	100	39.0	25.1

Importance of finetuning *text encoder* in STALE using CLIP vocabulary on 75% train split on ActivityNet dataset

Text encoder	Fine-tune	mAP 0.5	mAP Avg
No encoder	-	6.9	11.5
CLIP-text	✗	36.9	22.8
CLIP-text	✓	38.2	24.9