TubeDETR: Spatio-Temporal Video Grounding with Transformers

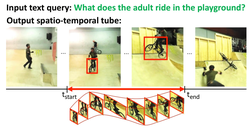
Antoine Yang[1,2], Antoine Miech[3], Josef Sivic[4], Ivan Laptev[1,2], Cordelia Schmid[1,2]

[1]Inria Paris [2]Ecole Normale Supérieure, PSL [3]DeepMind [4]CIIRC, CTU in Prague

Overview

Spatio-Temporal Video Grounding Task

➤ Given an input video, localize a *spatio-temporal tube* corresponding to an input natural language query



Motivation

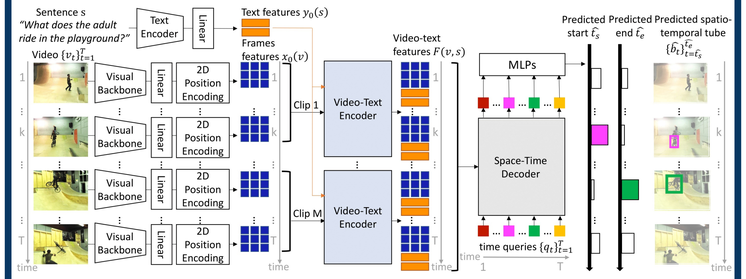
➤ Existing approaches rely on object proposals [2], tube proposals [3] or upsampling layers [4], hence use *different* representations for spatial and temporal localization

➤ Spatio-Temporal Video Grounding requires *efficient* modeling of temporal, spatial and visual-linguistic interactions

Contributions

➤ A **space-time decoder** that reasons about *time queries*, which are used for both spatial and temporal localization

➤ A *dual-stream* **video-text encoder** that efficiently encodes spatial and multi-modal interactions

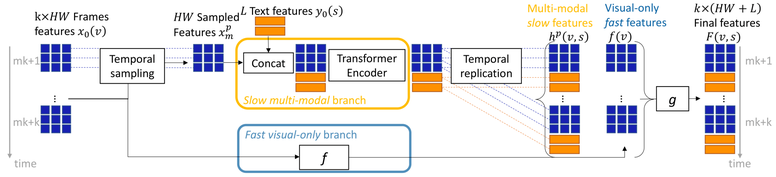


Links

➤ **Code and models:** *https://github.com/antoyang/TubeDETR*

➤ **Online Demo:** *http://stvg.paris.inria.fr/*

Video-Text Encoder

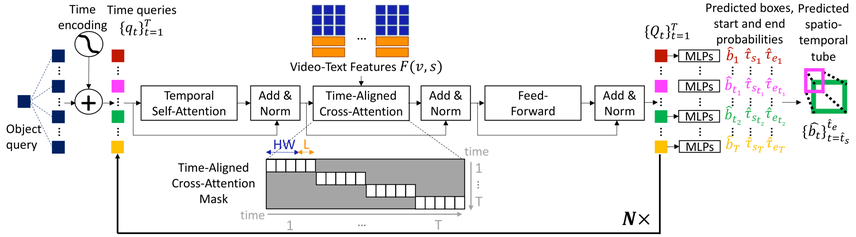


➤ Efficiently computes spatial and visual-linguistic interactions

➤ **Slow multi-modal branch:** samples video features from one from every k frames, and computes visual-linguistic interactions using a transformer encoder

➤ **Fast visual-only branch:** processes features from all frames but without any attention layers / with no backpropagation to the backbone for increased efficiency

Space-Time Decoder



➤ Models temporal interactions and predicts the tube using the video-text features

➤ **Inputs:** time encodings called *time queries*, one per frame, and video-text features

➤ **Architecture:** Transformer decoder with block-diagonal cross-attention, *ie* succession of *temporal self-attention*, *time-aligned cross-attention*, *feed-forward layers*

➤ **Prediction heads:** Multi-layer perceptrons on top of the contextualized time queries

Training

➤ **Losses:** Spatial (L_1 + $gIoU$) + Temporal (KL divergence + guided attention)

➤ **Initialization:** from MDETR [1] pretrained on Visual Genome, COCO and Flickr, and transferring its spatial attention weights to the temporal attention in our decoder

Ablation studies (VidSTG)

Space-Time Decoder

➤ Time encoding helps, especially for temporal localization (+22.0% tIoU)

➤ Temporal self-attention helps (+4.5% vIoU@0.3)

Initialization

➤ MDETR [1] initialization helps esp. for spatial localization (+9.2% m_sIoU)

➤ Transferring spatial attention to temporal attention helps (+3.3% vIoU@0.3)

Video-Text Encoder

➤ Dual-stream enables using higher spatial resolutions at the same GPU memory, e.g. 352 pixels with $k = 4$ instead of 224 with $k = 1$.

➤ Fast branch helps (+2.1% vIoU@0.3) at low GPU memory cost (+6.3GB)

Comparison to the state of the art

➤ State-of-the-art results on VidSTG and HC-STVG

Method	Pretraining data	VidSTG Declarative vIoU	VidSTG Declarative vIoU@0.3	VidSTG Interrogative vIoU	VidSTG Interrogative vIoU@0.3	HC-STVG1 vIoU	HC-STVG1 vIoU@0.3
STGRN [2]	Visual Genome	19.8	25.8	18.3	21.1	-	-
STGVT [3]	Visual Genome + Conceptual Cap.	21.6	29.8	-	-	18.2	26.8
STVGBert [4]	ImageNet + Visual Genome + Conceptual Cap.	24.0	30.9	22.5	26.0	20.4	29.4
Ours	ImageNet + Visual Genome + COCO + Flickr	**30.4**	**42.5**	**25.7**	**35.7**	**32.4**	**49.8**

Qualitative results

References

[1] A. Kamath et al., MDETR -- Modulated Detection for End-to-End Multi-Modal Understanding. In ICCV 2021.
[2] Z. Zhang et al., Spatio Temporal Video Grounding for Multi-Form Sentences. In CVPR, 2020.
[3] Z. Tang, et. al., Human-centric Spatio-Temporal Video Grounding With Visual Transformers. In TCSVT, 2021.
[4] R. Sui et. al., STVGBert: A Visual-linguistic Transformer based Framework for Spatio-temporal Video Grounding. In ICCV, 2021..