Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning

Antoine Yang†*, Arsha Nagrani§, Paul Hongsuck Seo§, Antoine Miech♯, Jordi Pont-Tuset§, Ivan Laptev†, Josef Sivic¶, Cordelia Schmid§

§Google Research †Inria Paris and Ecole Normale Supérieure, PSL ♯DeepMind ¶CIIRC, CTU in Prague *Work done as Google intern.

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Overview

Dense Video Captioning

➤ Generate temporally localized captions for all events in an untrimmed minutes-long video.

Motivation

➤ Prior dense video captioning methods contain task-specific components like event counters [1].

➤ Pix2seq [2] shows that it is possible to tackle object detection via language modeling.

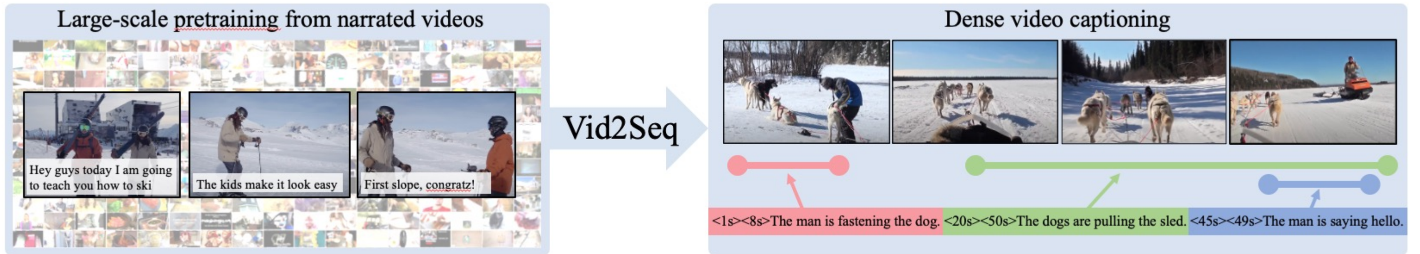
Contributions

➤ **Vid2Seq:** a visual language model that can densely caption untrimmed videos by generating a single sequence of (text and time) tokens.

➤ Vid2Seq considerably benefits from pretraining on unlabeled narrated videos at scale, by using transcribed speech sentences and corresponding timestamps as pseudo dense captioning annotations.

➤ SoTA on 3 dense video captioning datasets, 2 video paragraph captioning benchmarks, 2 video clip captioning datasets and promising few-shot results.

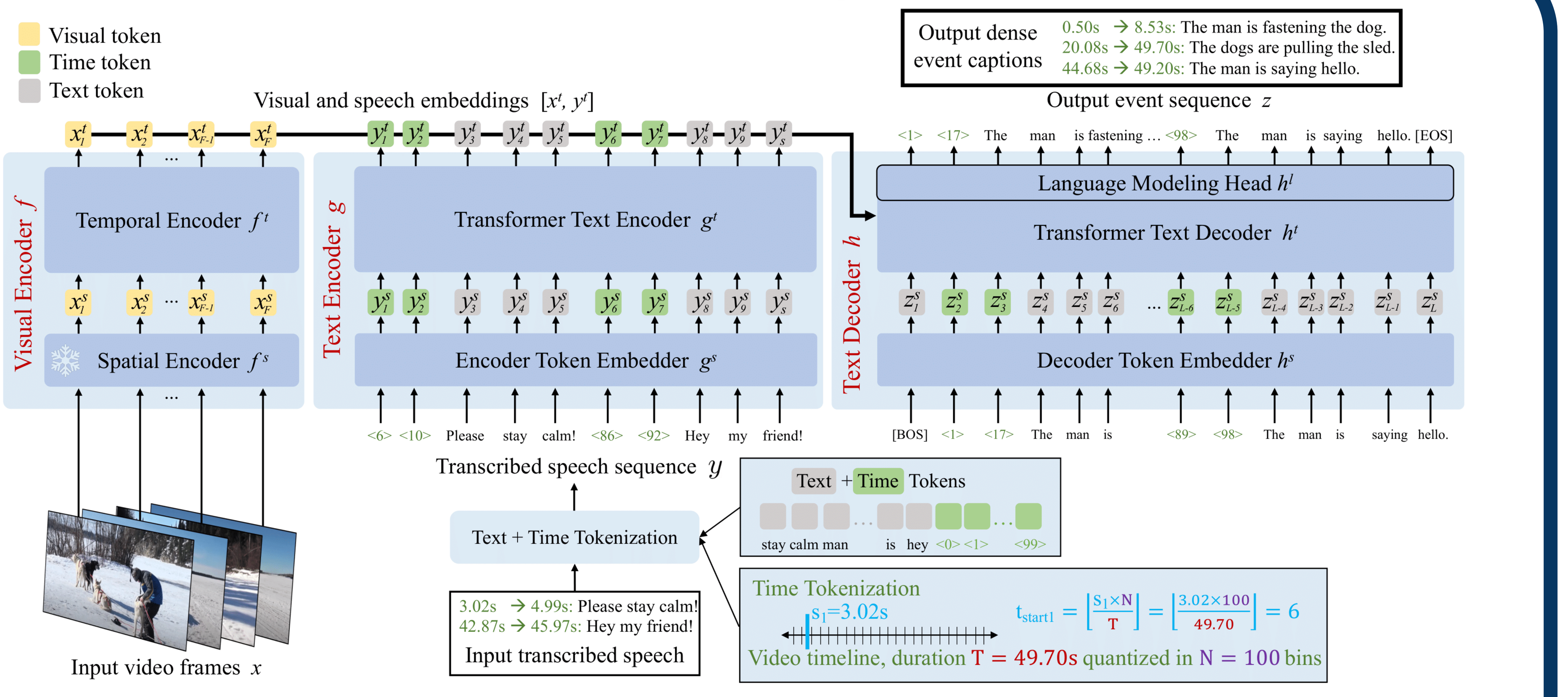
➤ **Code:** https://github.com/google-research/scenic/tree/main/scenic/projects/vid2seq



References

[1] T. Wang, et al., End-to-End Dense Video Captioning with Parallel Decoding. In ICCV 2021.
[2] T. Chen, et. al., Pix2seq: A Language Modeling Framework for Object Detection. In ICLR 2022.
[3] W. Zhu, et al., End-to-end Dense Video Captioning as Sequence Generation. In COLING 2022.
[4] Q. Zhang, et. al., Unifying Event Detection and Captioning as Sequence Generation via Pre-Training. In ECCV 2022.
[5] J. Lie, et al., MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning. In ACL 2020.
[6] J.S. Park, et al., Adversarial Inference for Multi-Sentence Video Description. In CVPR 2019.
[7] P.H. Seo, et al., End-to-end Generative Pretraining for Multimodal Video Captioning. In CVPR 2022.
[8] K. Lin, et al., SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning. In CVPR 2022.

The Vid2Seq Model



➤ **Sequence construction:** both the transcribed speech input and the dense event captioning annotations are cast as a sequence of text sentences interleaved with time tokens grounding the text in the video.

➤ **Architecture:** visual encoder, text encoder and text decoder.

➤ **Initialization:** CLIP visual backbone and T5 language model.

Pretraining Vid2Seq on narrated videos

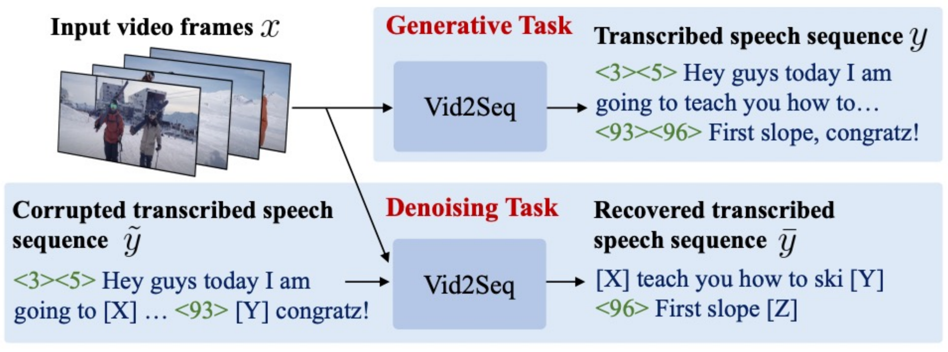
➤ We use transcribed speech sentences and corresponding timestamps as pseudo dense event captioning annotations.

➤ Pretraining is done using *untrimmed* videos by exploiting speech timestamps with *time tokens* → crucial to performance.

➤ **Pretraining Dataset:** YT-Temporal-1B (18 million narrated videos).

➤ **Generative objective:** predict speech given visual inputs.

➤ **Denoising objective:** predict masked tokens given noisy speech and visual inputs → benefits multi-modal reasoning.



➤ Finetuning on various tasks is done with a language modeling loss.

Comparison to SoTA

➤ Dense video captioning benchmarks.

Model	YouCook2			ViTT			ActivityNet Captions		
	SODA	CIDEr	METEOR	SODA	CIDEr	METEOR	SODA	CIDEr	METEOR
SoTA	4.4 [1]	25.0 [3]	4.7 [1]	-	25.0 [3]	8.1 [3]	5.5 [4]	29.0 [1]	8.0 [1]
Vid2Seq	**7.9**	**47.1**	**9.3**	**13.5**	**43.5**	**8.5**	**5.8**	**30.1**	**8.5**

➤ Event localization performance.

Model	YouCook2		ViTT		ActivityNet Captions	
	Recall	Precision	Recall	Precision	Recall	Precision
SoTA	20.7 [3]	20.6 [3]	32.2 [3]	32.1 [3]	**59.0** [4]	**60.3** [4]
Vid2Seq	**27.9**	**27.8**	**42.6**	**46.2**	52.7	53.9

➤ Video paragraph captioning benchmarks.

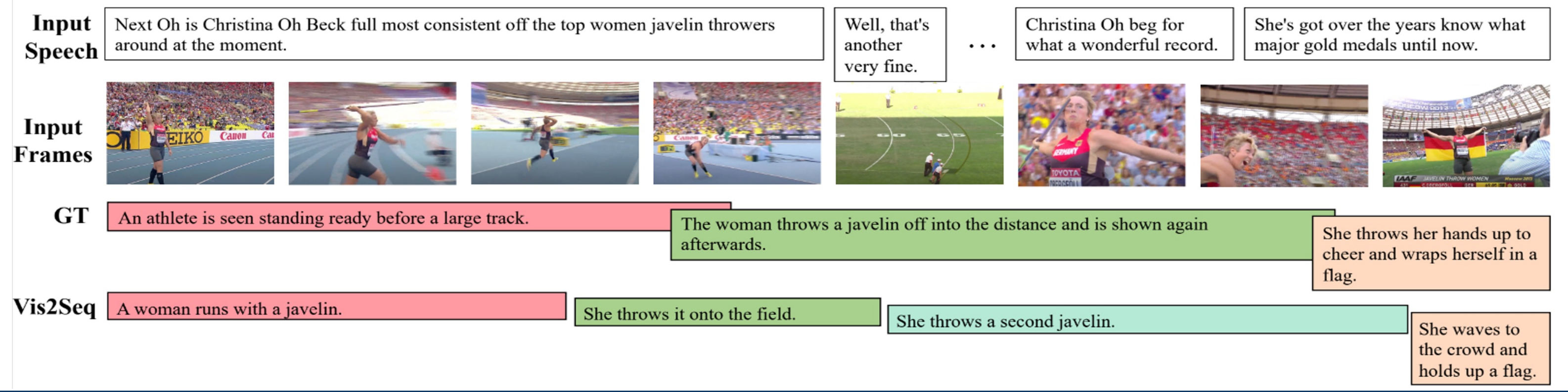
Model	YouCook2		ActivityNet Captions	
	CIDEr	METEOR	CIDEr	METEOR
SoTA w/ GT proposals	35.7 [5]	15.9 [5]	27.3 [1]	16.6 [6]
Vid2Seq w/ learnt proposals	**50.1**	**24.0**	**28.0**	**17.0**

➤ Video clip captioning benchmarks.

Model	MSR-VTT		MSVD	
	CIDEr	METEOR	CIDEr	METEOR
SoTA	60.0 [7]	29.9 [8]	120.6 [8]	41.3 [8]
Vid2Seq	**64.6**	**30.8**	**146.2**	**45.3**

Qualitative results

➤ More at https://www.youtube.com/watch?v=3oEHSU5ExsI.



Few-Shot Dense Video Captioning

➤ New setting using a small fraction of the downstream dataset for finetuning.

Data	YouCook2			ViTT			ActivityNet Captions		
	SODA	CIDEr	METEOR	SODA	CIDEr	METEOR	SODA	CIDEr	METEOR
1%	2.4	10.1	3.3	2.0	7.4	1.9	2.2	6.2	3.2
10%	3.8	18.4	5.2	10.7	28.6	6.0	4.3	20.0	6.1
50%	6.2	32.1	7.6	12.5	38.8	7.8	5.4	27.5	7.8
100%	**7.9**	**47.1**	**9.3**	**13.5**	**43.5**	**8.5**	**5.8**	**30.1**	**8.5**