JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

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

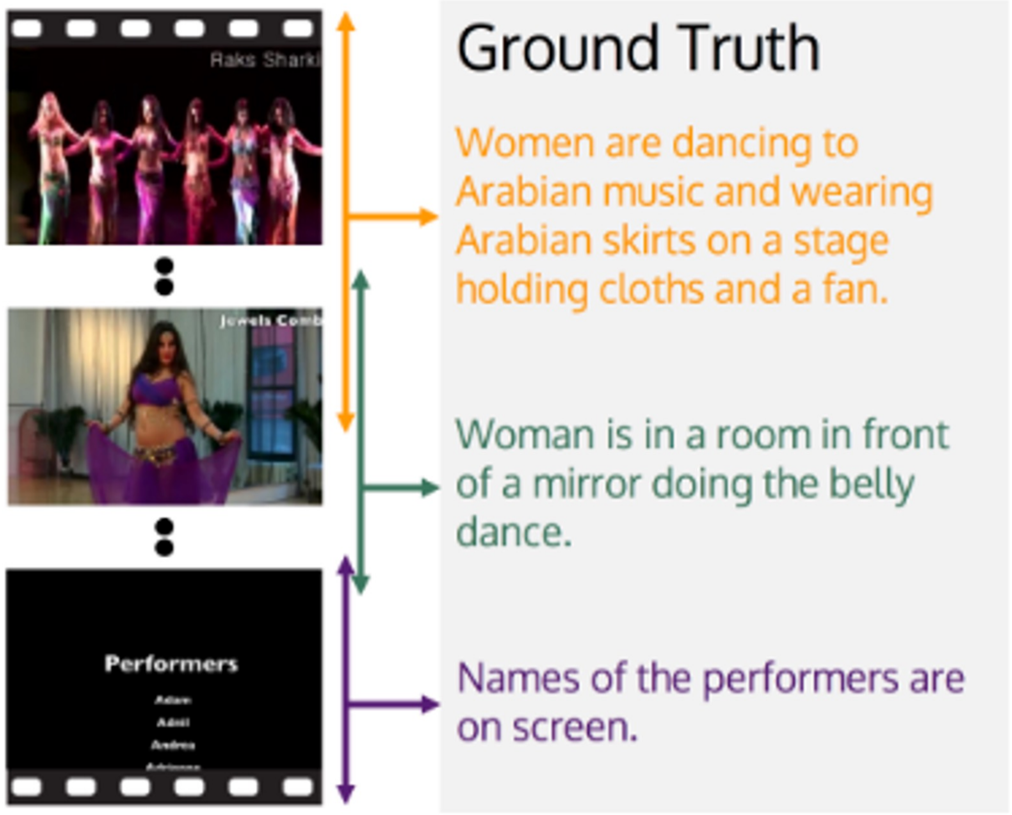
Project page: https://antoyang.github.io/vid2seq.html

Paper: https://arxiv.org/abs/2302.14115

Google Inria ENS | PSL CTU CZECH TECHNICAL UNIVERSITY IN PRAGUE

Dense Video Captioning

- **Task:** generate temporally localized captions for all events in an untrimmed minutes-long video.
- **Prior approaches (e.g. [Wang 2021]):** are task specific and trained only on manually annotated datasets.

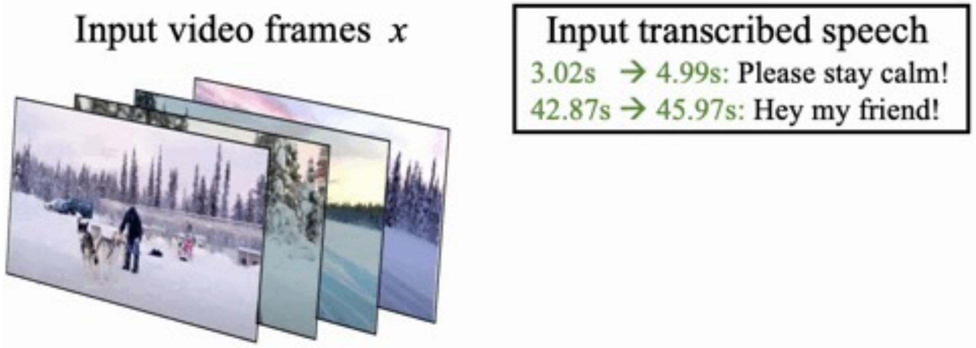


Example from the ActivityNet-Captions dataset [Krishna 2017].

[Krishna 2017] Dense-Captioning Events in Videos, Ranjay Krishna et al, ICCV 2017.
[Wang 2021] End-to-End Dense Video Captioning with Parallel Decoding, Teng Wang et al, ICCV 2021.

The Vid2Seq model

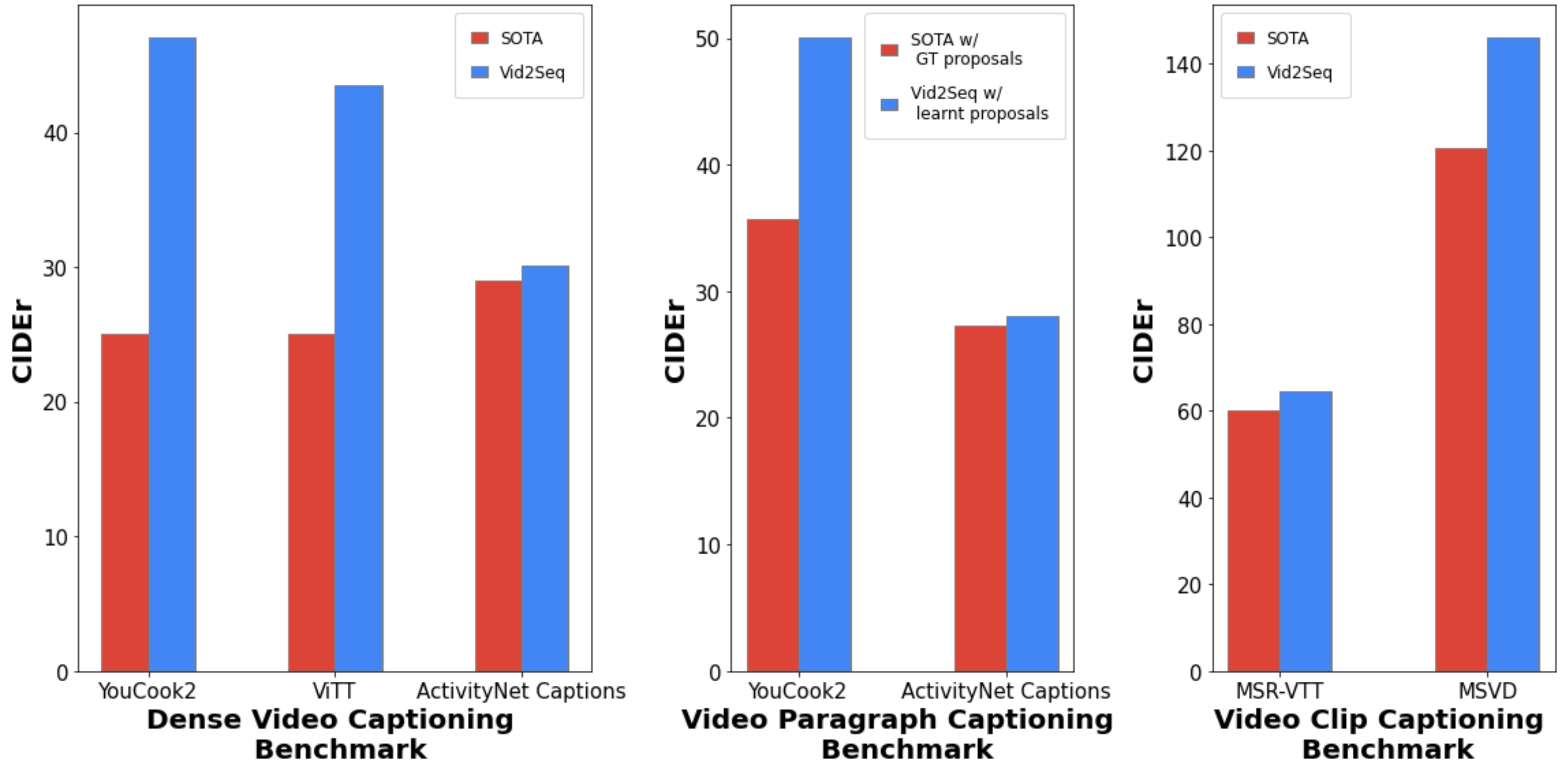
- Formulates dense video captioning as a sequence-to-sequence problem.
- Time is quantized and jointly tokenized with the text.
- **Model architecture:** visual encoder, text encoder and text decoder.



Pretraining Vid2Seq on untrimmed narrated videos

- Speech is also cast as a single sequence of text and time tokens.
- **Generative objective:** given visual inputs, predict speech.
- **Denoising objective:** given visual inputs and noisy speech, predict masked tokens.

Vid2Seq improves the SoTA on various video captioning tasks.



[Wang 2021] End-to-End Dense Video Captioning with Parallel Decoding, Teng Wang et al, ICCV 2021.
[Zhu 2022] End-to-end Dense Video Captioning as Sequence Generation, Wanrong Zhu et al, COLING 2022.
[Lei 2020] MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning, Jie Lei et al, ACL 2020.
[Seo 2022] End-to-end Generative Pretraining for Multimodal Video Captioning, Paul Hongsuck Seo et al, CVPR 2022.
[Lin 2022] SwinBERT: End-to-End Transformers with Sparse Attention for Video Captioning, Kevin Lin et al, CVPR 2022.

Benefits of pretraining on untrimmed narrated videos with time tokens

Pretraining input		YouCook2			ActivityNet Captions		
Untrimmed	Time tokens	SODA	CIDEr	F1	SODA	CIDEr	F1
✗	✗	4.0	18.0	18.1	5.4	18.8	49.2
✓	✗	5.5	27.8	20.5	5.5	26.5	52.1
✓	✓	**7.9**	**47.1**	**27.3**	**5.8**	**30.1**	**52.4**

Effect of pretraining losses and modalities

The visual inputs only model benefits from the generative objective.

The denoising objective helps the model with visual+speech inputs.

Finetuning Input		Pretraining losses		YouCook2			ActivityNet Captions		
Visual	Speech	Generative	Denoising	SODA	CIDEr	F1	SODA	CIDEr	F1
✓	✗	No pretraining		3.0	15.6	15.4	5.4	14.2	46.5
✓	✓	No pretraining		4.0	18.0	18.1	5.4	18.8	49.2
✓	✗	✓	✗	5.7	25.3	23.5	**5.9**	**30.2**	51.8
✓	✓	✓	✗	2.5	10.3	15.9	4.8	17.0	48.8
✓	✓	✓	✓	**7.9**	**47.1**	**27.3**	5.8	30.1	**52.4**

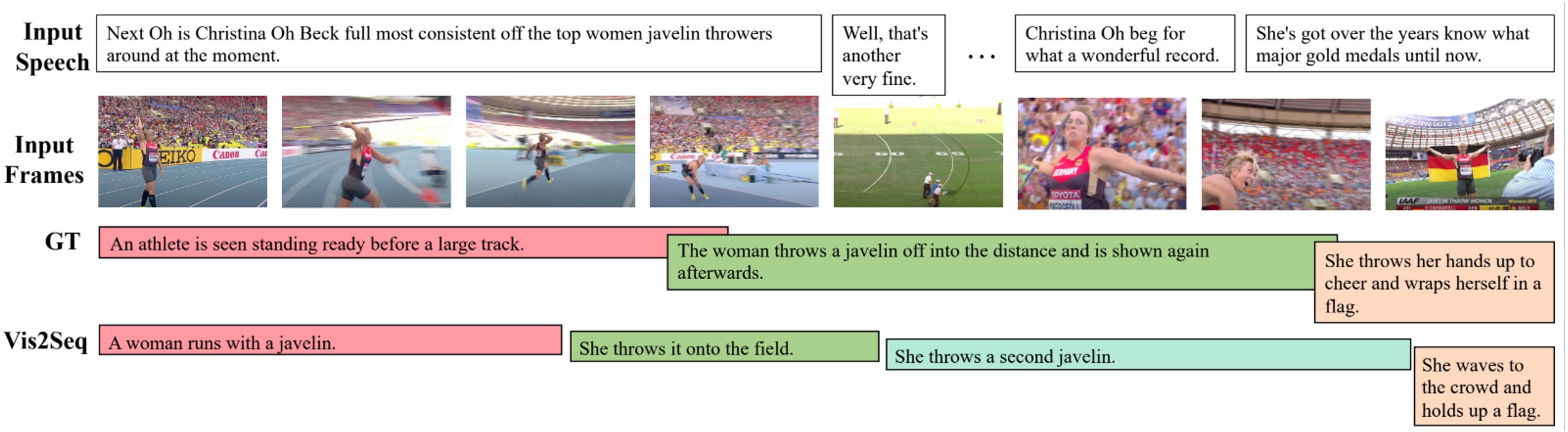
Vid2Seq generalizes well to few-shot settings.

We find that pretraining is crucial for few-shot generalization.

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**

Qualitative dense video captioning results

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



Conclusion

- We introduce Vid2Seq, a new visual language model for dense video captioning.
- We show how Vid2Seq can be effectively pretrained on unlabeled narrated videos at scale.
- The pretrained Vid2Seq model improves the SoTA on 3 dense video captioning datasets, 2 video paragraph captioning datasets and 2 video clip captioning datasets, and generalizes well to few-shot setting.