

#### 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.

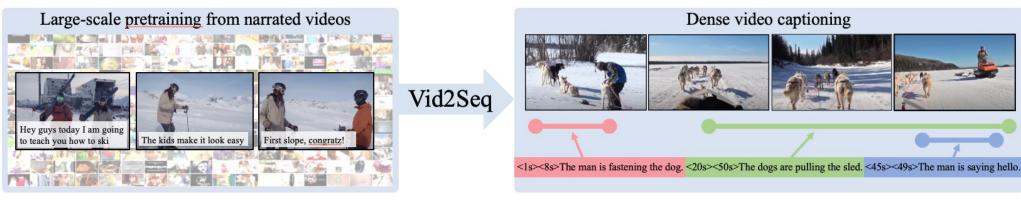
### Contributions

> 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.

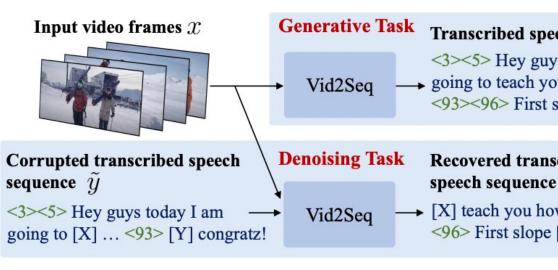
#### Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning Antoine Yang<sup>†\*</sup>, Arsha Nagrani<sup>§</sup>, Paul Hongsuck Seo<sup>§</sup>, Antoine Miech<sup>#</sup>, Jordi Pont-Tuset<sup>§</sup>, Ivan Laptev<sup>†</sup>, Josef Sivic<sup>¶</sup>, Cordelia Schmid<sup>§</sup> <sup>§</sup>Google Research <sup>†</sup>Inria Paris and Ecole Normale Supérieure, PSL <sup>#</sup>DeepMind <sup>¶</sup>CIIRC, CTU in Prague <sup>\*</sup>Work done as Google intern. The Vid2Seq Model Comparison to SoTA Visual token $0.08s \rightarrow 49.70s$ : The dogs are pulling the sle Dense video captioning benchmarks. Time token Text token $-\underbrace{y_2^t}_2 - \underbrace{y_3^t}_3 - \underbrace{y_4^t}_4 - \underbrace{y_5^t}_5 - \underbrace{y_6^t}_6 - \underbrace{y_7^t}_7 - \underbrace{y_8^t}_8 - \underbrace{y_9^t}_9 - \underbrace{y_s^t}_8$ ↑ ··· ↑ YouCook2 Model Language Modeling Head $h^l$ CIDEr SODA Femporal Encoder f Transformer Text Encoder $g^t$ SODA METEOR Transformer Text Decoder $h^t$ 4.4 [1] 25.0 [3] 4.7 [1] SoTA $\boldsymbol{\chi}_2^{\boldsymbol{S}} \cdots \boldsymbol{\chi}_{F-1}^{\boldsymbol{S}}$ 1 Vid2Sec 47.1 13.5 7.9 9.3 Spatial Encoder Model Event localization performance. Franscribed speech sequence Text + Time Tokenization stay calm man is hey <0><1>20.7 [3 SoTA $\mathbf{t}_{\text{start1}} = \left\lfloor \frac{\mathbf{S}_1 \times \mathbf{N}}{\mathbf{T}} \right\rfloor = \left\lfloor \frac{3.02 \times 100}{49.70} \right\rfloor = 6$ Video paragraph captioning $3.02s \rightarrow 4.99s$ : Please stay calm! $42.87s \rightarrow 45.97s$ : Hey my friend! Vid2Seq 27.9 Input transcribed speech benchmarks. > Sequence construction: both the transcribed speech input and the Video clip captioning benchmarks. **ActivityNet Captions** Model YouCook2 dense event captioning annotations are cast as a sequence of text CIDEr METEOR CIDEr METEOR sentences interleaved with time tokens grounding the text in the video. SoTA w/ GT 35.7 [5] 15.9 [5] 27.3 [1] 16.6 [6] proposals > Architecture: visual encoder, text encoder and text decoder. 28.0 17.0 Vid2Seg w/ learnt 50.1 24.0 > Initialization: CLIP visual backbone and T5 language model. proposals Pretraining Vid2Seq on narrated videos-Qualitative results **iput** Next Oh is Christina Oh Beck full most consistent off the top women javelin throwers Christina Oh beg for around at the moment another what a wonderful record. very fine. ► More at https://www.youtube.com/watch? athlete is seen standing ready before a large track. The woman throws a javelin off into the distance and is shown again > Pretraining is done using *untrimmed* videos by exploiting speech v=3oEHSU5Exsl. Vis2Seq A woman runs with a javelin. She throws it onto the field. She throws a second javel Pretraining Dataset: YT-Temporal-1B (18 million narrated videos). Few-Shot Dense Video Captioning **Generative objective:** predict $\blacktriangleright$ New setting using a small fraction of the downstream dataset for finetuning. Transcribed speech sequence <3><5> Hey guys today I am → going to teach you how to... YouCook2 Data <93><96> First slope, congratz CIDEr METEOR SODA SODA **Recovered transcribed Corrupted transcribed speech** speech sequence $\overline{U}$ 10.1 2.0 2.4 3.3 1% <3><5> Hey guys today I am $\rightarrow$ [X] teach you how to ski [Y] → Vid2Seq going to [X] ... <93> [Y] congratz! <96> First slope [Z] 3.8 18.4 10.7 5.2 10% 32.1 12.5 50% 6.2 7.6 38 Finetuning on various tasks is done with a language modeling loss. 100% 47.1 13.5 7.9 9.3

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

timestamps with *time tokens*  $\rightarrow$  crucial to performance.

speech given visual inputs.

> **Denoising objective:** predict masked tokens given noisy speech and visual inputs  $\rightarrow$ benefits multi-modal reasoning.

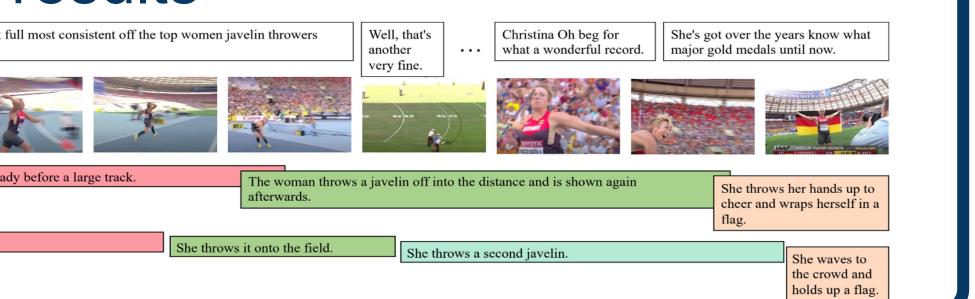






ViTT		ActivityNet Captions			
MET	FEOR	SODA	CIDEr	METEOR	
8.	1 [3]	5.5 [4]	29.0 [1]	8.0 [1]	
8	8.5	5.8	30.1	8.5	
ouCook2 Vi		iTT	ActivityNet Captions		
Precision	Recall	Precision	Recall	Precision	
20 6 [2]	20 0 [0]	20 1 [2]	<b>50 0</b> [4]	<b>CO 2</b> [4]	
20.6 [3]	32.2 [3]	32.1 [3]	<b>59.0</b> [4]	<b>60.3</b> [4]	
	ME 8. 8. 6 0 6 0 k2 Precision	METEOR 8.1 [3] 8.5 00k2 V Precision Recall	METEOR SODA   8.1 [3] 5.5 [4]   8.5 5.8   ok2 ViTT   Precision Recall Precision	METEOR SODA CIDEr   8.1 [3] 5.5 [4] 29.0 [1]   8.5 5.8 30.1   ok2 ViTT ActivityNe   Precision Recall Precision Recall	

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



ViTT		ActivityNet Captions			
DEr	METEOR	SODA	CIDEr	METEOR	
.4	1.9	2.2	6.2	3.2	
8.6	6.0	4.3	20.0	6.1	
8.8	7.8	5.4	27.5	7.8	
3.5	8.5	5.8	30.1	8.5	