VidChapters-7M: Video Chapters at Scale

Antoine Yang[†], Arsha Nagrani[§], Ivan Laptev[†], Josef Sivic[¶], Cordelia Schmid[†]

[†]Inria Paris and Ecole Normale Supérieure, PSL [§]VGG, University of Oxford [¶]CIIRC, CTU in Prague

NEURAL INFORMATION PROCESSING SYSTEMS

Overview

Video chapter generation

➤ Automatically segment a long video into segments and generate a chapter title for each.

Contributions

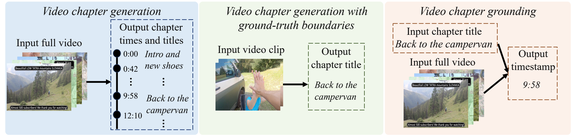
➤ **VidChapters-7M:** a dataset of 817K user-chaptered videos including 7M chapters in total.

➤ Benchmarks with various baselines and models for the tasks of video chapter generation with or without GT boundaries and video chapter grounding.

➤Video chapter generation models trained on VidChapters-7M transfer well to dense video captioning in both zero-shot and finetuning settings, improving the SoTA on YouCook2 and ViTT.

➤ **Data:** https://antoyang.github.io/vidchapters.html

➤ **Code:** https://github.com/antoyang/VidChapters



References

[1] R. Zellers, et. al., MERLOT Reserve: Neural Script Knowledge through Vision and Language and Sound. In CVPR, 2022.
[2] A. Radford, et. al., Robust Speech Recognition via Large-Scale Weak Supervision. In ICML2023.
[3] A. Radford, et. al., Learning Transferable Visual Models From Natural Language Supervision. In arXiv, 2021.
[4] H. Touvron, et al., LLaMA: Open and Efficient Foundation Language Models. In arXiv 2023.
[5] J. Li, et. al., BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In ICML 2023.
[6] T. Wang, et al., End-to-End Dense Video Captioning with Parallel Decoding. In ICCV 2021.
[7] A. Yang, et. al., Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning. In CVPR 2023.
[8] J. Devlin, et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In NAACL-HLT 2019.
[9] J. Lei, et al., QVHighlights: Detecting Moments and Highlights in Videos via Natural Language Queries. In NeurIPS 2021.

Data collection and processing

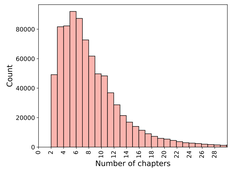
➤ **A scalable collection process:**

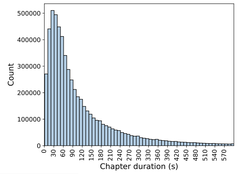
1. Start from a large pool of YouTube videos (92M [1]).
2. Download their description and check the presence of chapters.
3. Keep the videos which have chapters (817K).

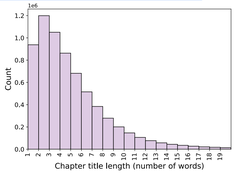
➤ **Data processing:** Whisper [2] for ASR, CLIP [3] for visual features.

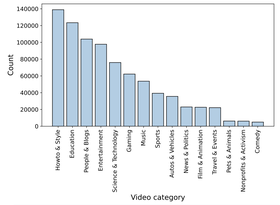
VidChapters-7M

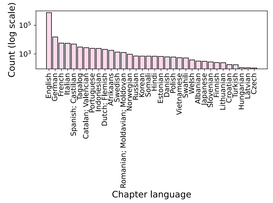
Dataset	# vids	# min / vid	# labels	Annotations
HowTo100M	1M	7	136M	ASR
YT-Temporal-1B	**19M**	6	900M	ASR
ActivityNet Captions	20K	3	100K	Dense captions
Ego4D	10K	**23**	4M	Dense captions
VidChapters-7M	817K	**23**	**7M**	**Chapters + ASR**











Type of chapters	Count (on a random sample of 100 videos)
Speech and visual	49
Audio and visual	2
Speech-only	26
Visual-only	3
Audio-only	3
Structure-only	14
Unrelated	3

New benchmarks

➤ **Full video chapter generation:**

Model	Modalities	PT Data	FT on VC	SODA	CIDEr	METEOR	R@3s	P@3s	R@0.7	P@0.7
Text tiling + LLaMA [4]	T	Text mix	No	0.2	0.5	0.3	5.8	7.9	8.9	8.8
Shot detect + BLIP-2 [5]	V	129M img-txt	No	0.6	0.2	0.6	27.4	29.7	12.5	8.7
PDVC [6]	V	None	Yes	6.8	35.8	9.4	17.8	**40.2**	22.5	**26.9**
Vid2Seq [7]	T	C4+HTM	Yes	10.5	50.7	8.7	**28.9**	23.3	27.2	24.8
Vid2Seq [7]	V+T	C4	Yes	10.6	51.3	8.8	28.6	23.8	26.9	24.9
Vid2Seq [7]	V+T	C4+HTM	Yes	**11.4**	**55.7**	**9.5**	28.5	24.0	**28.5**	26.4

➤ **Video chapter grounding:**

Model	Modalities	PT Data	FT on VC	R@3s	R@0.7
BERT [8]	T	Text mix	No	5.2	0.1
CLIP [3]	V	400M img-txt	No	3.7	2.3
Moment-DETR [9]	V	None	Yes	**12.4**	**17.6**

➤ See our paper for **video chapter generation with ground-truth boundaries**.

Transfer to dense video captioning

➤ **Fully-supervised setting:**

Model	Modalities	PT Data	YouCook2			ViTT		
			SODA	CIDEr	METEOR	SODA	CIDEr	METEOR
SoTA [7]	T+V	C4+YTT	7.9	47.1	9.3	13.5	43.5	8.5
PDVC [6]	V	None	4.8	28.8	5.8	9.4	40.6	**16.5**
PDVC [6]	V	VidChap	5.9	34.7	7.5	10.1	41.5	16.1
Vid2Seq [7]	T+V	C4+HTM	8.6	53.2	10.5	14.1	44.8	8.7
Vid2Seq [7]	T+V	C4+HTM+ 10% VidChap	9.9	63.9	12.1	14.5	47.4	9.2
Vid2Seq [7]	T+V	C4+HTM+ VidChap	**10.3**	**67.2**	**12.3**	**15.0**	**50.0**	9.5

➤ **Strong zero-shot results:** Vid2Seq achieves 3.9S/13.3C on YouCook2 & 9.0S/28.0C on VITT.

➤ **Qualitative examples:** see our paper.