Streaming Dense Video Captioning — Supplementary Material

Xingyi Zhou* Anurag Arnab* Shyamal Buch Shen Yan
Austin Myers Xuehan Xiong Arsha Nagrani Cordelia Schmid
Google

We provide further training details (Sec. A) and additional qualitative results of our model (Sec. B).

A. Training hyperparameters

All our experiments are conducted using the Scenic library [3] and JAX [1]. With the GIT [9] architecture, we first pretrain on the WebLI [2] dataset for general image captioning. WebLI [2] contains 100M image-text pairs derived from alt-text from the internet. The image encoder is initialized from CLIP-L [7], and the language decoder is randomly initialized. During pretraining, we use the standard label-smoothed (factor 0.1) cross-entropy loss following GIT [9] and train for 10 epochs. We use the Adam [5] optimizer, with no weight decay. The learning rate is set to 5×10^{-5} with a batch-size of 1024, with a cosine decay schedule. Following GIT [9], we use $0.2\times$ lower learning rate for the image encoder.

When finetuning on dense-video captioning datasets [4, 6, 12], we freeze the image encoder. We again use the Adam [5] optimizer with 0 weight decay. We train for 20 epochs with batch size of 32, and use a learning rate of 10^{-5}, dropped by $10\times$ at the 16th epoch.

With Vid2Seq [10], we take the publicly released pretrained checkpoint[1], which is pretrained on the YT-Temporal dataset [11] with a denoising and a captioning objective [10]. When finetuning on dense-video captioning datasets [4, 6, 12], we follow their official training parameters. Specifically, we freeze the image encoder and pool the image tokens among the spatial dimensions to get one token per frame. The T5 [8] decoder uses a dropout rate of 0.1. We again use Adam [5] optimizer with 0 weight decay. We train for 40 epochs with batch-size 32, and use a learning rate of 3×10^{-4} with a cosine decay schedule.

For all models, we follow the standard protocol to use beam-search decoding, with a beam size of 4 and a brevity penalty of 0.6 [8]. We also emphasize that wherever applicable, all base architectures and backbones are consistent between comparisons and baselines.

*Equal contribution. {zhouxy, aarnab}@google.com

[1] https://github.com/google-research/scenic/tree/main/scenic/projects/vid2seq

B. Further qualitative results

We provide qualitative results of our model and the ground truth on ActivityNet in folder `results`. We also include a "`results.html`" to display the videos in a web browser. Our model provides accurate captions and localizations across a diverse range of events.

References

[1] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. 1

[2] Xi Chen, Josip Djolonga, Piotr Padlewski, Basil Mustafa, Soravit Changpinyo, Jialin Wu, Carlos Riquelme Ruiz, Sebastian Goodman, Xiao Wang, Yi Tay, et al. Pali-x: On scaling up a multilingual vision and language model. In *arXiv:2305.18565*, 2023. 1

[3] Mostafa Dehghani, Alexey Gritsenko, Anurag Arnab, Matthias Minderer, and Yi Tay. Scenic: A JAX library for computer vision research and beyond. In *CVPR Demo*, 2022. 1

[4] Gabriel Huang, Bo Pang, Zhenhai Zhu, Clara Rivera, and Radu Soricut. Multimodal pretraining for dense video captioning. *arXiv:2011.11760*, 2020. 1

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[6] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *ICCV*, 2017. 1

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[8] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *JMLR*, 2020. 1

[9] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. In *arXiv:2205.14100*, 2022. 1

[10] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *CVPR*, 2023. 1

[11] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *CVPR*, 2022. 1

[12] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 1