Building Scalable Video Captioning Models

Sarthak Mohanty, Nitin Vegesna, Aditya Chandaliya

Presentation Link:
https://drive.google.com/file/d/16ufs1MBLvk4bzQaLix-bT21RK4whw9xJ/view?usp=sharing

Introduction

- Large-scale foundational models → video captioning
- Typically use well-performing image to text models (CLIP, ALIGN)
- General theory:
  - Video is a sequence of images
  - Images can be processed separately
  - Then a sequence model is applied
  - Captures characteristics of each frame + temporal aspect
- Our modification: applying transformers as the sequence modeling

Related Work

- CLIP-Hitchhiker (Bain et al., 2022) - takes the mean of image embeddings
- Two-Stream LSTMs (Zhao et al., 2021) - encodes audio and visual data separately, then uses fusion LSTM to combine data into same latent space
- Transformers for Images (Dosovitskiy et al., 2021) - extends self-attention from 2D image space to 3D temporal space

Dataset

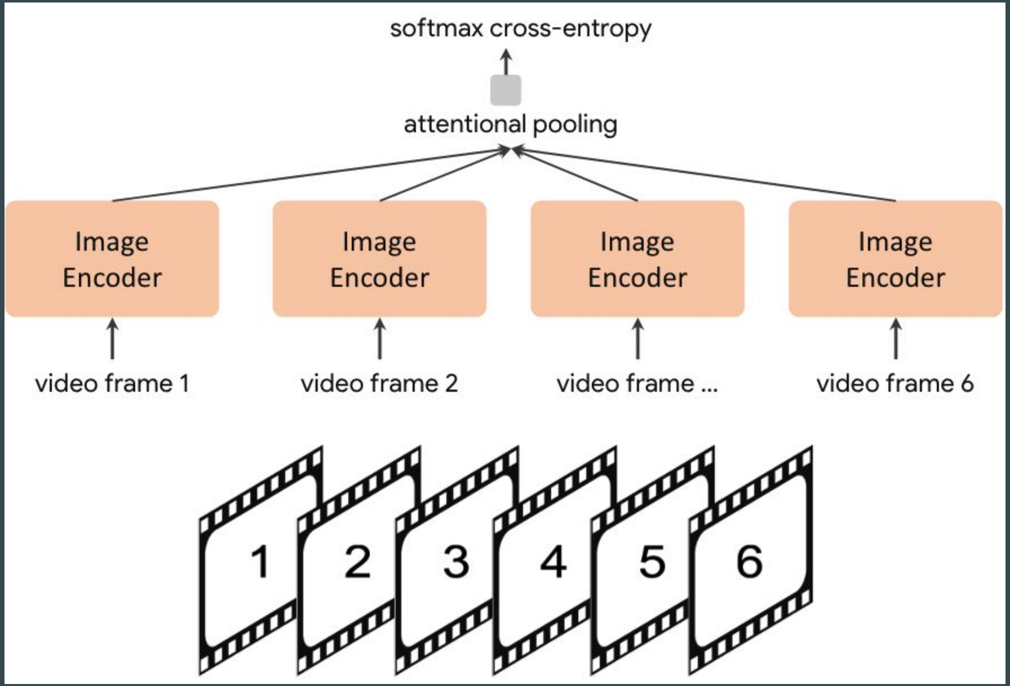
- We used the Microsoft Research Video to Text (MSR-VTT) dataset
- Consists of 10, 000 clips from 20 categories
- Each clip has 20 annotated “captions”
    - Sampled one annotation per clip for ground truth
- Due to storage limits, only used every *40th* frame
- Training-validation-test split of 70-15-15
- Captions preprocessed and tokenized using BERT

Approach 1: Zero-Shot Contrastive Captioning (CoCa)

- CoCa - pre-existing image captioning model
- Uses a cascaded decoder design
  - Bottom half - encodes text content with causally masked self-attention
  - Top half - multimodal decoder using cross-attention to align image with text
- Video captioning - apply pre-trained CoCa on each frame and average results

Approach 2: Fine-tuned CoCa

- Simple CoCa → not guaranteed to capture temporal information
- Add a pooler on top of spatial sequence tokens to attend to temporal sequence patterns



Approach 3: ResNet + Transformer

- ResNet-18 model - used to extract information from independent images
  - Use frozen convolutional blocks
  - Encode each frame of the video
  - Project onto feature space of size 768, matching captions

Approach 3: ResNet + Transformer

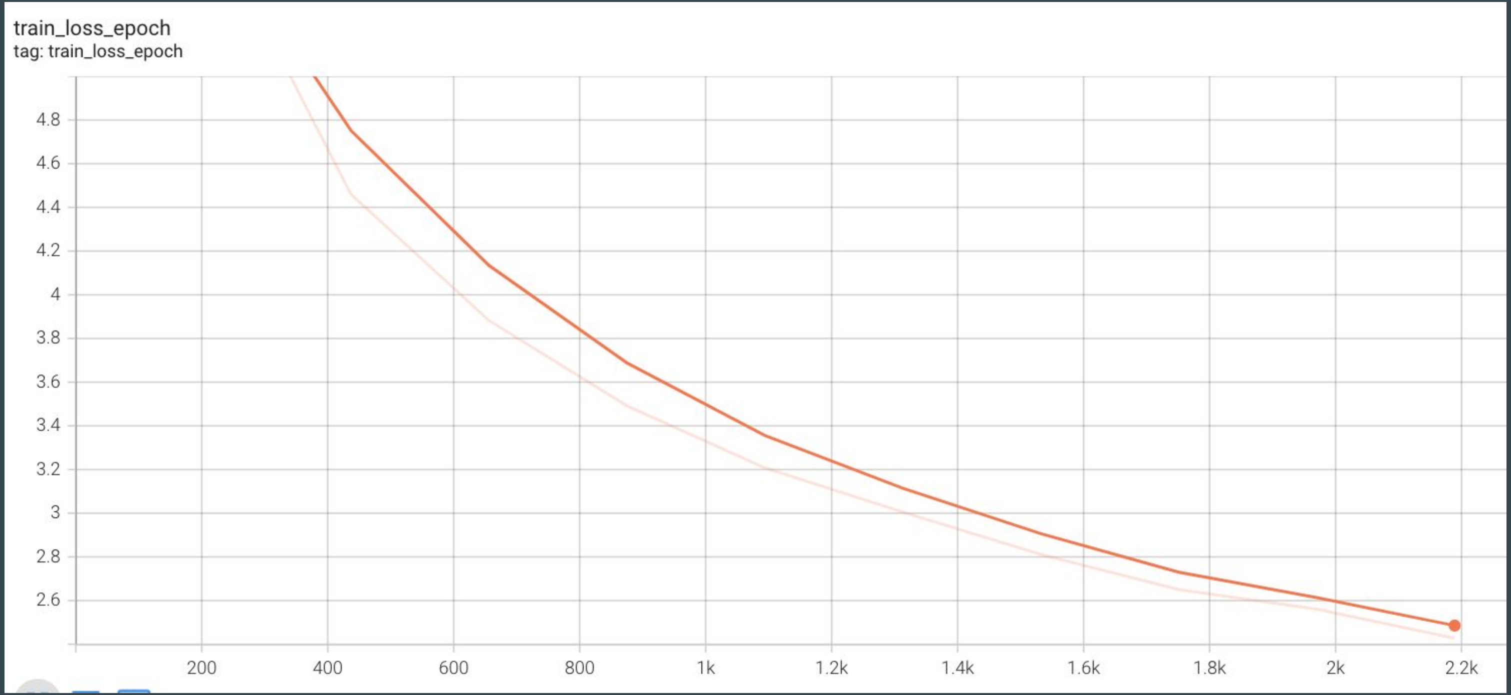
- Transformer - encoder-decoder model used to model sequential (temporal) dependencies
  - Add relative positional bias and absolute positional embeddings
  - Multi-head attention with *n* layers and *h* heads
  - No mask applied to encoder - want to capture dependencies
  - Lower triangular square mask for decoder - hides future tokens to allow parallelism
  - Linear layers used to project onto BERT's vocab space

Training

- Primarily focused on approach 3 for our final results
- Used cross-entropy loss with stochastic optimizer Adam
- Learning rate ranged from 1e-3 to 1e-6, with scheduler added in some experiments
- Trained model for 15 epochs
- Performed grid search for hyperparameters
    - Optimal combination: LR = 1e-4 with scheduler, n=6 layers, h=8 heads, norm_first = True

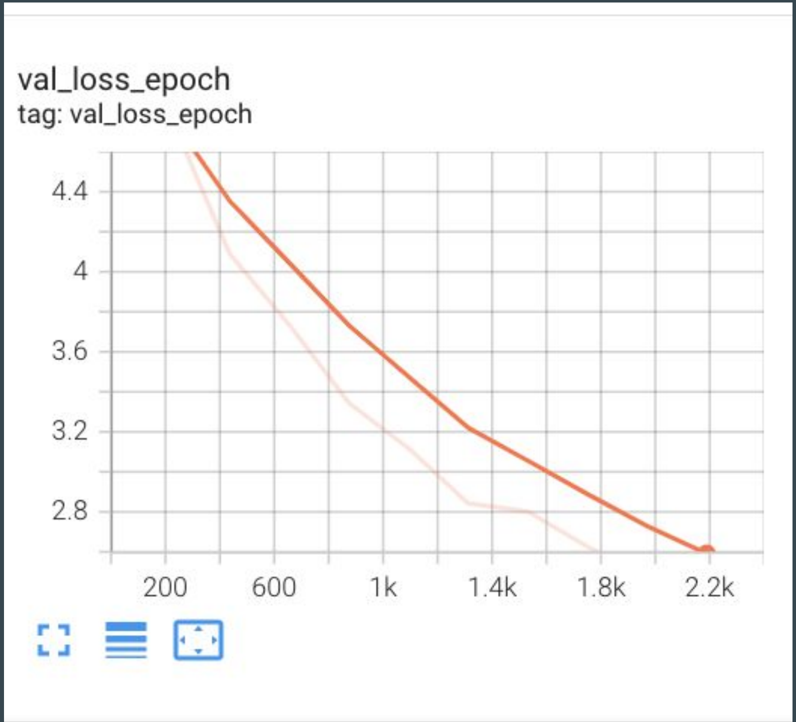
Training Loss

- Training loss steadily decreasing - indicates model hasn't converged yet



Validation Loss

- Validation loss achieved minimum (over 15 epochs) at 1.99



Sample Captions

Predicted Caption	**Ground Truth Caption**
A man in a purple shirt and red hat is playing guitar while playing on a microphone	A man in a blue shirt and blue shirt is playing guitar and singing into a microphone
A person is trying up through from the	A person is carrying someone away from danger
A group of the sports car are very at at a busy race show	A series of convertible sports cars are lined up at a large car show

Table 1: Comparison of Predicted and Ground Truth Captions

Evaluation Metrics

- Test loss achieved minimum of 2.35

Evaluation Metrics			
	Recall	**Precision**	**F1**
ROUGE-1	0.6860699710113917	0.6493109267925568	0.6618847461242713
ROUGE-2	0.45838117034278275	0.2990833003742873	0.35434326640328534
ROUGE-L	0.6722172025069564	0.6365854570425724	0.6487584559993945
BLEU		0.3899576889070564	

Table 2: Combined Evaluation Metrics (ROUGE & BLEU Scores)

Analysis & Future Work

- Results indicate that the general architecture is promising
- With limited dataset and compute power, we achieved moderately good ROUGE and BLEU scores
- More training and high-quality data needed to achieve better results
- Future work: adding audio data from video - currently not included due to memory limits on Colaboratory

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2022. A clip-hitchhiker's guide to long video retrieval.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5288–5296.

Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. 2023. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners.

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models.

Bin Zhao, Maoguo Gong, and Xuelong Li. 2021. Audiovisual video summarization.

Thank You!