CMU-VIDION: Modified BLIP with Audio for Video to Text Description

Laura Yao, Juncheng Li, Florian Metze
Carnegie Mellon University

Video Description Generation

- Automatic generation of natural language descriptions for videos
- Process various input modalities that include both visual and auditory components

a yellow bird in a parking lot, with music playing in the background

Background

- SOTA vision-language pre-trained (VLP) models like BLIP are demonstrating impressive performance for zero-shot predictions of captions based on visual inputs [1]

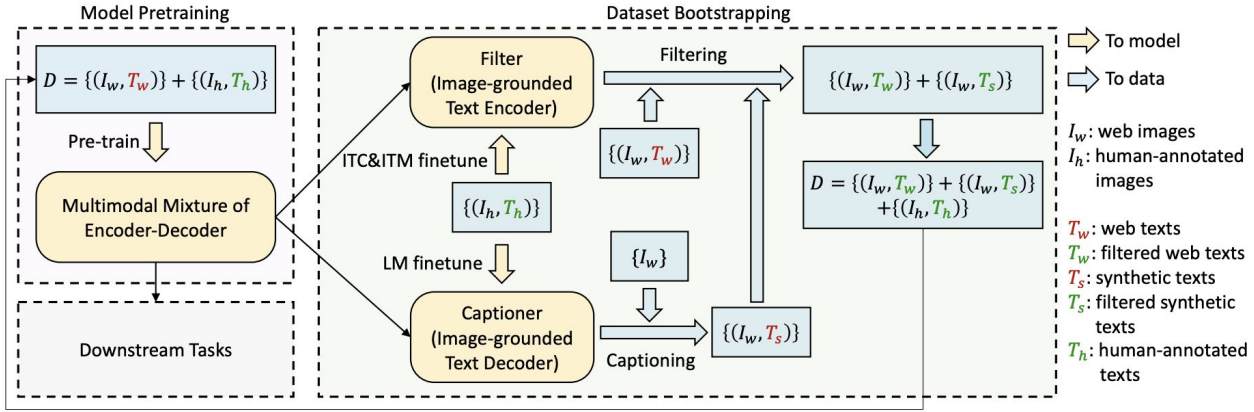


Figure 3. Learning framework of BLIP. We introduce a captioner to produce synthetic captions for web images, and a filter to remove noisy image-text pairs. The captioner and filter are initialized from the same pre-trained model and finetuned individually on a small-scale human-annotated dataset. The bootstrapped dataset is used to pre-train a new model.

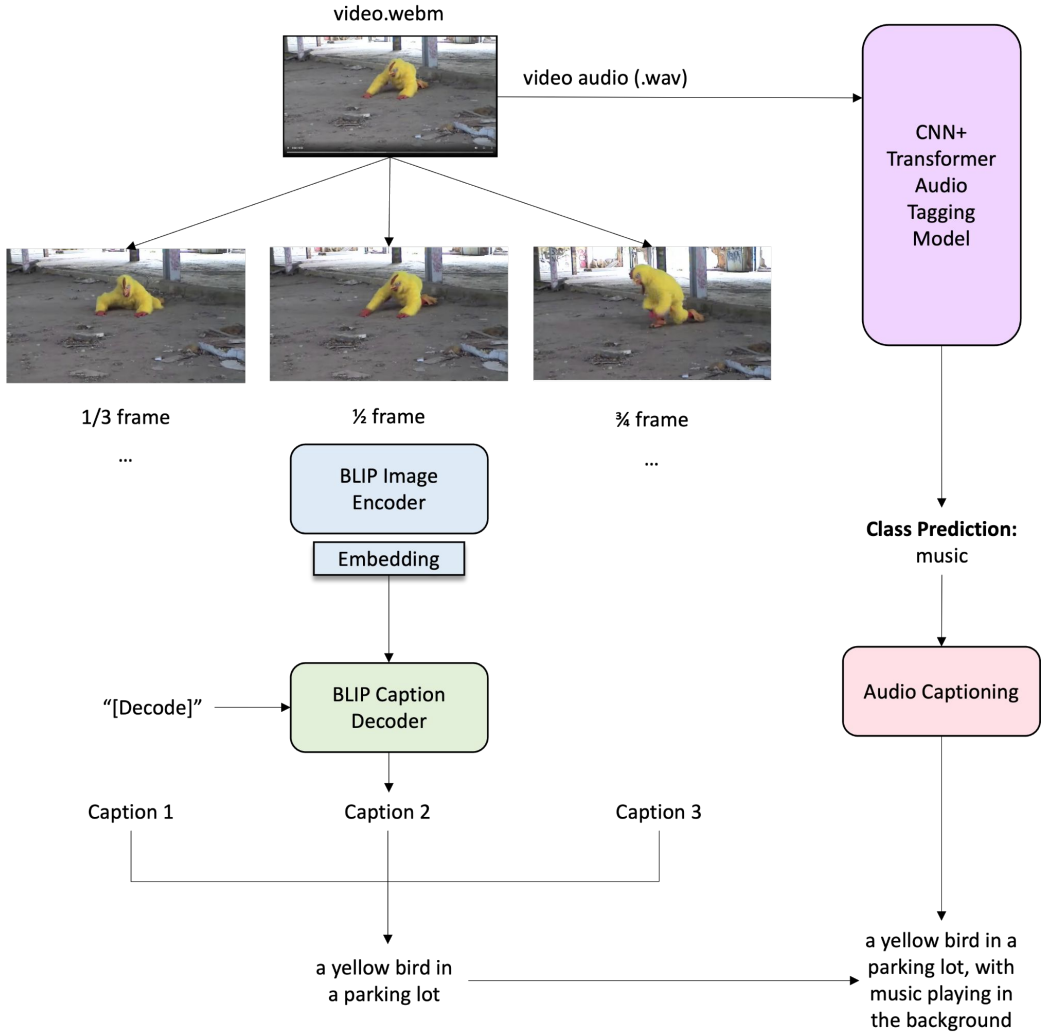
[1] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” arXiv preprint arXiv:2201.12086, 2022.

Our Model

- Current SOTA models do not leverage all 3 modalities (audio, visual, text)
- Our model leverages BLIP combined with a model trained on AudioSet [2]
    - Adds audio-contextual details to the captions
- We used an image-text model instead of a video-text model
    - Scarcity of existing video-text annotations
    - BLIP [1] demonstrated that they could outperform VideoCLIP [3]

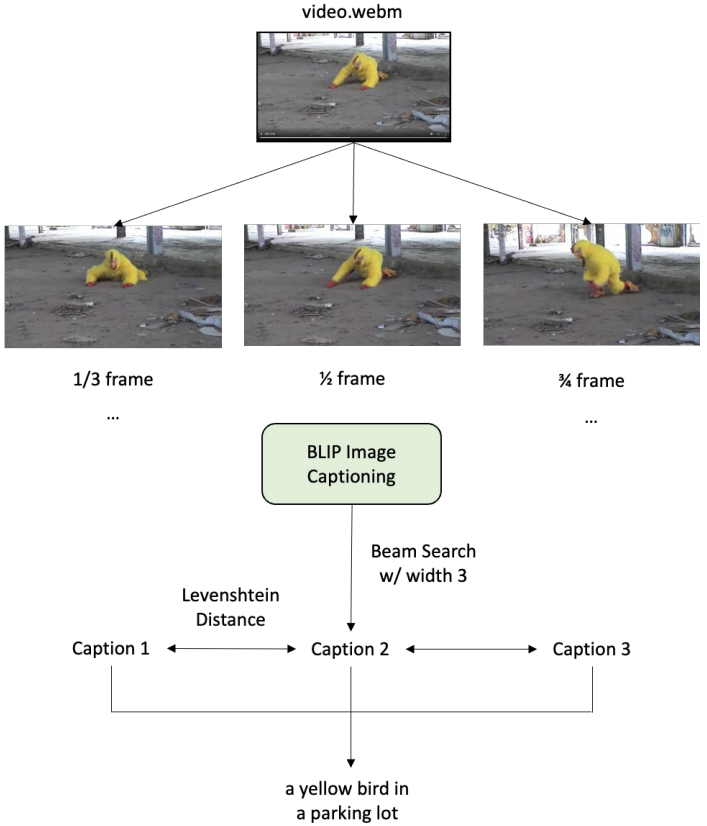
[2] Juncheng B Li, Shuhui Qu, Florian Metze, et al., “Audiotagging done right: 2nd comparison of deep learning methods for environmental sound classification,” arXiv preprint arXiv:2203.13448, 2022.

[3] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer, “Videoclip: Contrastive pre-training for zero-shot video-text understanding,” arXiv preprint arXiv:2109.14084, 2021.


video.webm
video audio (.wav)
CNN+ Transformer Audio Tagging Model
1/3 frame
...
½ frame
¾ frame
...
BLIP Image Encoder
Embedding
"[Decode]"
BLIP Caption Decoder
Caption 1
Caption 2
Caption 3
a yellow bird in a parking lot
Class Prediction: music
Audio Captioning
a yellow bird in a parking lot, with music playing in the background

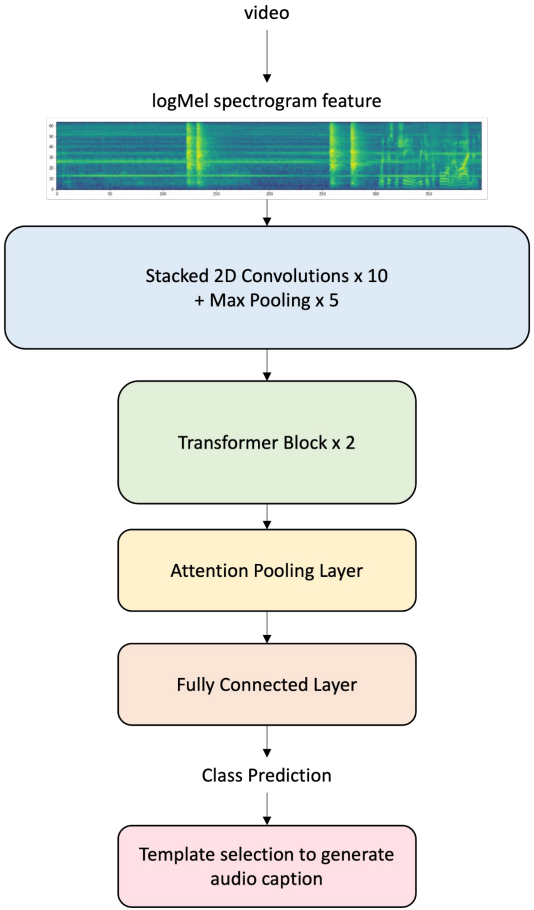
Implementation Details

- Use zero-shot BLIP image captioning
- Beam search to decode the captions with highest probability
- Removal of duplicate words that appear in the BLIP-decoded captions
- Levenshtein distance to combine text captions in order to get the “best” caption



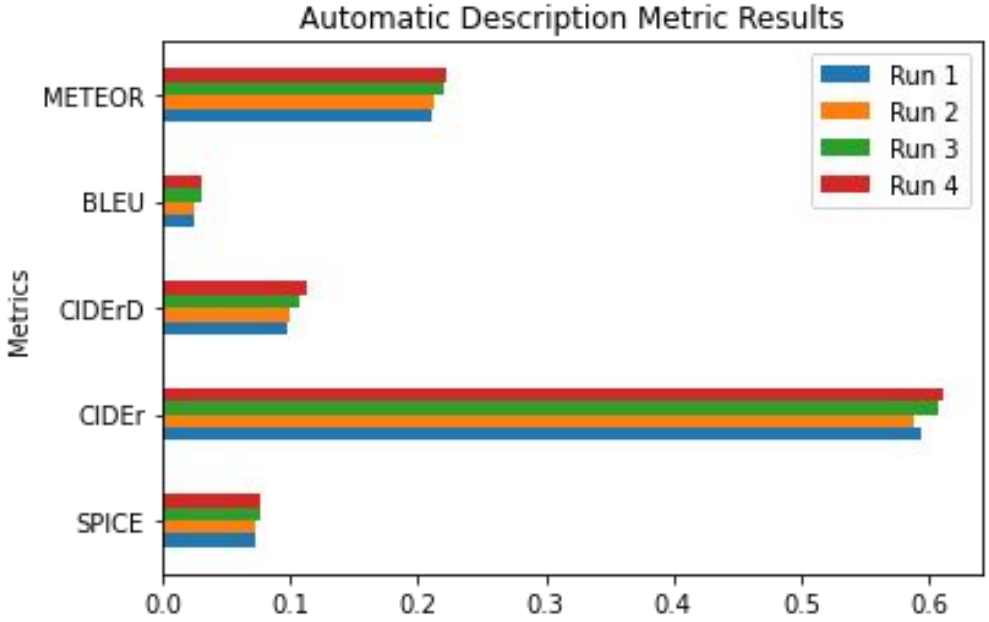
Implementation Details

- Generates audio caption based on a couple custom templates:
  - who/what is making these sounds
  - whether it is background music
  - what instrument is playing
- Uses resampled 16kHz audio which we extract logMel spectrogram features from [2]



Results

- Run 1: middle frame (½)
- Run 2: all three frames (⅓, ½, ¾)
- Run 3: middle frame and audio analysis
- Run 4 (our primary run): all three frames and audio analysis



Generated Captions

"a person sitting on the ground"

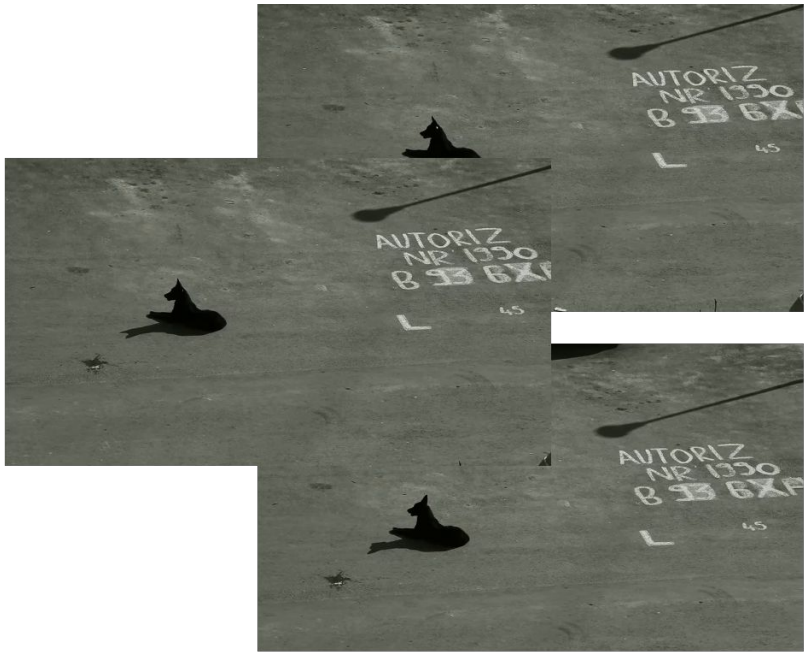
Ground Truth Caption: "young woman in a black vest and pink tights and top sitting on the curb in front of a blocked up red brick building on a sunny day."

Generated Captions

“a bride and groom walking down the aisle”

Ground Truth Caption: “A white woman in a bridal gown walking on a grass lawn near trees and a stage and seating on a sunny day.”

Generated Captions



“a dog on the ground, with music playing in the background”

Ground Truth Caption: “With eerie music playing at dusk, a black dog sits majestically on a concrete area with white lettering. and then shadow of a lamp post.”

Generated Captions

“a group of people running on the beach”

Ground Truth Caption: “Two men in white shirts are reaching the finish line in a running race, as others are still running, and photographers are taking pictures on a sunny day on the beach.”

Generated Captions

Generated Caption	Ground Truth
a person sitting on the ground	young woman in a *black vest and pink tights and top* sitting on the *curb* in front of a blocked up red brick building on a *sunny day*.
a bride and groom walking down the aisle	A white woman in a bridal gown walking on a *grass lawn* near trees and a stage and seating on a *sunny day*.
a dog on the ground, with music playing in the background	With *eerie* music playing at dusk, a *black* dog sits majestically on a *concrete area* with white lettering. and then shadow of a lamp post.
a group of people running on the beach	Two men in *white shirts* are reaching the finish line in a running race, as others are still running, and *photographers* are taking pictures on a *sunny day* on the beach.

	STS 1	STS 2	STS 3	STS 4	STS 5
Run 1	0.3967	0.3897	0.3939	0.3909	0.3903
Run 2	0.3984	0.3868	0.3928	0.3930	0.3889
Run 3	0.4065	0.3947	0.3996	0.3986	0.3955
Run 4	**0.4062**	**0.3952**	**0.3999**	**0.3985**	**0.3953**

Summary

- Audio is important for boosting caption generation similarity to ground truth
- Our current model only loosely joins audio and video/visual input
- For future works, we plan on further investigating how to better add audio into CLIP-style models
    - Joint embedding spaces could be an effective way to combine all three types of input