Using Video-Language models to Generate Audio Descriptions for Movies

Andrew Zisserman

January 2024

CVPR 2023, ICCV 2023: Research with Max Bain, Tengda Han, Arsha Nagrani, Gül Varol, Weidi Xie
Introduction: Video Understanding

• What is happening in the video?
• Who is in the video?
• What are they doing?
• What is the scene?
• Where is it?
• What is the story?

Buster Keaton `Cops' (1922)
What is movie Audio Description?

- Narration describing visual elements in the movie to aid the visually impaired

Movie clip from `Out of Sight` (1998) with Audio Description
What is movie Audio Description (AD)?

Narration describing visual elements in movies to aid the visually impaired:

- Complementary to the raw audio track (no need to describe the audio)
- Aim is storytelling: includes character names, emotion, actions, ...
- Dense descriptions over time (previous context very important)

- "A Dataset for Movie Description", A Rohrbach, M Rohrbach, N Tandon, B Schiele, CVPR 2015
- Large Scale Movie Description Challenge (LSMDC) 2015-2021
He takes the seat opposite, then places his lighter on the table.

- A new way to evaluate movie understanding abilities
  - Long-form videos; multi-modal; fine-grained recognition

- Societal impact:

  "Hello, I’m KT. Just wanted to say thank you for the AD that you all have made available. I’m able to enjoy lots of different films I grow up with but wasn’t able to really understand them because I am blind. So thanks again."

  - Available from AudioVault (https://audiovault.net/), provided by volunteers
Train a Visual-Language Model to generate the AD

Video frames

Visual-language model (VLM)

- A man approaches toying with a lighter.
- She turns her head, and finds Jack standing beside her.

AD context

Subtitle context

> Can I buy you a drink?
> Yeah I'd love one. Sit down.

He takes the seat opposite, then places his lighter on the table
Outline

1. Background on visual language models
   • Two types of network architecture using adapters

2. A basic AD model, data, and training
   • Adapting pre-trained vision and language models to this task

3. Improving the `who’ in generated AD
   • Supplying supplementary information on characters

4. Improving the `what’ in generated AD
   • Adapting pre-trained video-language models to this task
   • Evaluating performance
How to splice a visual encoder into a language model

Language model:
- pre-trained transformer decoder (e.g. GPT2 like)
- Pre-trained and Frozen

Visual encoder:
- Can be Convolutional or Visual Transformer (ViT)
- Pre-trained and Frozen

Visually-conditioned Language Models:
- Method 1: Prompt Tuning
- Method 2: Cross-attention
Overview of Architecture 1: Prompt-tuning GPT

- **Input**: visual data
- **Output**: free form text
- Visual encoder: pretrained and frozen
- Language model (GPT2): pretrained and frozen
- Only the mapping network (adapter) is trained

“A very serious cat


“BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models”, Junnan Li, Dongxu Li, Silvio Savarese, Steven Hoi, arXiv, 2023
Overview of Architecture 2: Cross-attention (X-Attn) GPT

- **Input**: visual data
- **Output**: free form text
- Visual encoder: pretrained and frozen
- Language model (GPT2): pretrained and frozen
- Only the mapping network and X-Attn layers (adapters) are trained

Flamingo: a Visual Language Model for Few-Shot Learning, Jean-Baptiste Alayrac et al, NeurIPS 2022
Training choices …

- Use discrete (VQ) or soft tokens for the LLM

- Architecture for mapping network, e.g.
  - Simple linear mapping
  - Transformer encoder

- Training end-to-end
2. A basic Audio Description model
Architecture – Prompt tuning GPT

Video Captioning with Long Multimodal Context

- Pretrained GPT for text generation
- All conditioning added as prompting vectors
  - Visual features (CLIP), movie subtitles, contextual AD

ClipCap, BLIP
Architecture – Prompt tuning GPT
Architecture – Prompt tuning GPT

N previous captions
Training

**Labelled Data (MADv2)**

- 488 Movies
- 316K AD captions
- 900 hrs video provided as CLIP features at 5 fps

`Complete` movie dataset: video features, AD, subtitles

MAD: A Scalable Dataset for Language Grounding in Videos From Movie Audio Descriptions, Soldan et al, CVPR 2022
Challenge: the lack of training data

- Movie data with corresponding visual, subtitles and description elements are very limited in size.

- MAD dataset has only 316k clips with AD.

- Use this for an evaluation (testing) split of 10 films, but too small to adapt foundation model to AD task.

- But there are several large datasets available …
Pretrain with **Partial Data**

- Available large scale datasets:
  - Paired *visual-textual* data (without temporal context): **CC3M, WebVid**
  - Movie AD data (without visual information): downloaded from **AudioVault**

- Use **partial data** to **pretrain** particular modules from large-scale datasets

- And then finally **finetune** the entire architecture with the complete movie dataset (MAD training)
AudioVault Dataset

Audio soundtracks for 7000+ movies containing the combined original audio and AD audio

- Process this dataset to obtain separate subtitles and AD as text
- How? WhisperX & Diarization

12,000+ hours transcribed
3.3M AD captions

MAD-v2 and AudioVault are publicly released
Partial Pre-training

Subtitle context

Can I buy you a drink?
Yeah I’d love one. Sit down.

AD context

A man approaches toying with a lighter.
She turns her head, and finds Jack standing beside her.

\[ \mathcal{L}_{\text{NLL}} \]
Partial Pre-training

- Visual captioning

Use paired visual-textual data (without temporal context): CC3M, WebVid (image-caption, video-caption datasets)
Partial Pre-training

- Text only

Use text Movie AD data (without visual information) from AudioVault
Final fine-tuning

- Complete movie data

Visual features, subtitles, and AD data from MADv2
Results: context and pretraining

- Visual context, AD context is helpful
- Partial-data pretraining is helpful
- However, subtitle input does not help

CIDEr performance measure:
Scores the similarity between the predicted and actual AD for a clip, over the MAD-Eval test set

Baseline: 4
+ visual context: 6.7
+ AD context: 12.6
+ AudioVault Pretrain: 14.1
+ WebVid Pretrain: 14.3
Qualitative Results

The Great Gatsby (2013)

Context AD: Nick and Daisy smile and Gatsby gestures towards the ballroom. Klipspringer a wild-haired young man with glasses, plays the organ.

Ground-truth AD: Gatsby reclines on cushions as Nick and Daisy dance in the ballroom, which is lit by hundreds of candles.

Prediction: A man and a woman dance in a circle.

Harry Potter and the Order of the Phoenix (2007)

Context AD: Professor Snape approaches behind Harry. Snape takes Harry down to his storeroom. Snape raises his wand. Harry body goes rigid.

Ground-truth AD: His mind fills with terrifying memories.

Prediction: His eyes widen.
Summary point & Limitations

- Developed a Prompt-tuning GPT from pre-trained foundation models using partial training
- Produces AD conditioned on visual frames of current clip and previous AD
- Does not reference character names
- Action and scene descriptions incomplete

The Great Gatsby (2013)

Context AD: Surrounded by gushing fountains and ornamental palms, they look up at the house. Gatsby looks at Daisy framed by the fountain. It's an orange-squeezing machine.

Ground-truth AD: Daisy Gatsby and Nick swim on his private beach.

Prediction: A man swims in the pool.
Outline

1. Background on visual language models
   • Two types of network architecture using adapters

2. A basic AD model, data, and training
   • Adapting pre-trained vision and language models to this task

3. Improving the `who’ in generated AD
   • Supplying supplementary information on characters

4. Improving the `what’ in generated AD
   • Adapting pre-trained video-language models to this task
   • Evaluating performance
[Who] is in the scene?

Possible characters: Jack Foley played by George Clooney; Karen Sisco played by Jennifer Lopez.
[Who] is in the scene?

- **Objective**: identify the active characters in a clip using face recognition
- Provide their names and example face images as prompts to aid AD naming
- How to achieve this objective?
  - Supply a “character bank” dataset for each film with names and face images
  - Use the character bank to identify the active characters in a clip
**[Who] is in the scene?**

**Character Bank**

- Use the Internet Movie Database (IMDb)
  - provides mapping between characters and *cast* names
  - also provides cast photos

- Use cosine similarity between CLIP features of photos and frames to select visual exemplars for each actor

---

**Out of Sight (1998)**

Top-5 exemplars in the movie:

- **Jack Foley** played by George Clooney
  - IMDb portrait

- **Karen Sisco** played by Jennifer Lopez
  - IMDb portrait

---

Top cast  
George Clooney  
Jack Foley  
Karen Sisco  
Jennifer Lopez  
Ving Rhames  
Buddy Bragg  
Steve Zahn  
Glenn Michaels
Character recognition module

Classify active characters in video clip

- Allow “face exemplars” to interact with movie frames
- Binary classification whether each character present in clip or not
- All features computed using frame-level CLIP
Character recognition module

- Allow “face exemplars” to interact with movie frames
- Binary classification whether each character present in clip or not
- All features computed using frame-level CLIP
An aside: standard face recognition approach

1. Detect Faces
2. Represent each face by a vector
3. Recognize a face from a gallery using closest distance between vectors

For this to work, need vectors to only represent identity, and not be affected by expression, pose, lighting, age, etc. Vectors obtained by deep network trained for identity
Character recognition module

- Allow “face exemplars” to interact with movie frames
- Binary classification whether each character present in clip or not
- All features computed using frame-level CLIP
Character recognition module

- Train character recognition module using MovieNet annotations
Results of [Who]

Task: predict active characters in movie clip
  ○ Baseline: CLIP feature cosine-similarity between face exemplars and frames
Character Bank Use for AD

For each new film:

– Download principal character data from IMDb (character names, actor names, portrait images)

– Obtain in-domain exemplars for each character as the character bank for that film

– Process all clips to determine “active characters”

– Provide active character names and visual exemplars as prompts for the AD for that clip

– No further training required
Generate AD with names – Prompt-tuned GPT

- Active **character names** are fed into the model as **text prompts**
- Also provides the **actor names**
- Also provides the **visual exemplars**

*Previous AD Context*: Possible characters: Jack played by George Clooney, Karen Sisco played by Jennifer Lopez. Describe <video>: 

ClipCap, BLIP
Generate AD with names – X-Attn GPT

- Active **character names** are fed into the model as **text prompts**
- Also provides the **actor names**
- Also provides the **visual exemplars**
- Use X-Attn to bring in more interactions
Architecture performance comparison

MAD-Eval test set

- CLIP-Cap
- AutoAD-I
- AutoAD-II

Prompt-tuned GPT
X-Attn GPT
Example 1: result on `Harry Potter and the Order of the Phoenix’ (not part of training)

Predicted Audio Description: “Snape points at Harry. Harry’s eyes close in horror”
Example 2: result on `Harry Potter and the Order of the Phoenix’ (not part of training)

Predicted Audio Description:

“Hermione, Ron and Luna’s eyes are fixed on Harry, who is standing in the doorway. Harry rides on the horse's back as the horse rears up in the air.”
Summary point & Limitations

- ‘Who’ performance improved significantly by introducing a Character Bank
- ‘What’ performance limited by CLIP frame descriptors
- Evaluation measures (e.g. CIDEr, Bleu) not fit for purpose

Ides of March (2011)

Prediction: Later, Stephen walks down the street with his hands in his pockets
Outline

1. Background on visual language models
   • Two types of network architecture using adapters

2. A basic AD model, data, and training
   • Adapting pre-trained vision and language models to this task

3. Improving the `who’ in generated AD
   • Supplying supplementary information on characters

4. Improving the `what’ in generated AD
   • Adapting pre-trained video-language models to this task
   • Evaluating performance
Start from pre-trained **video-language** model

- Model ingests 8 frames
- ViT spatial feature map for each frame
- Larger LM – Llama2-7B
- Video Q-Former trained on Webvid-2M

- Only train linear projection layer
- Need pixel level data to train the model …

---

The Condensed Movies Dataset

• 34,000 movie scenes from 3,600 movies
• ~10 ordered key scenes per movie, each about 2 minutes long
• Provides condensed snapshots into full-length stories

The Karate Kid 1984

#1 #2 #3 #4

#8 #7 #6 #5
time

M. Bain, A. Nagrani, A. Brown, A. Zisserman, ACCV 2020
Movie datasets with pixels: CMD-AD

movie clips from YouTube
e.g. 2 minutes, with unknow $[t_1, t_2]$

precise temporal alignment

Statistics:
Train: 1332 movies
Test: 100 movies
Duration: 477 hrs
Number of AD: 101k
Strong Vision-Language Models (VLM) with pixel inputs

Prompt: Possible characters:
- Jack Foley <Image> <Image> ;
- Karen Sisco <Image> <Image> ;
- <Video> <Video> ;

Please provide a detailed description of this movie clip.

Target: As Karen stares gloomily out of the window, Jack approaches toying with a lighter.

Architecture details:
- Visual feature extractor: EvaCLIP-L14
- Q-Former: 12-layer transformer
- LLM: OPT-2.7B or Llama2-7B
Architecture performance comparison

MAD-Eval test set

CLIP-Cap
AutoAD-I
AutoAD-II
AutoAD-III

Prompt-tuned GPT-2
X-Attn GPT-2
Movie-Llama

CIDEr
Problem of classical captioning metrics

<table>
<thead>
<tr>
<th></th>
<th>AD Sentence</th>
<th>CIDEr*</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference</strong></td>
<td>The donkeys make a smoke message in the sky which reads, we love you, Daddy.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>Prediction 1</strong></td>
<td>Donkeys use smoke to write ‘We love you daddy’ in the sky.</td>
<td>302.9</td>
<td>25.4</td>
<td>30.8</td>
<td>43.6</td>
</tr>
<tr>
<td><strong>Prediction 2</strong></td>
<td>Donkey's children write, we love you daddy, in the pale sky with their smoky breath.</td>
<td>226.5</td>
<td>18.9</td>
<td>20.3</td>
<td>25.9</td>
</tr>
<tr>
<td><strong>Prediction 3</strong></td>
<td>The young donkeys write 'Love you, daddy' in the sky.</td>
<td>187.0</td>
<td>0.0</td>
<td>25.7</td>
<td>38.6</td>
</tr>
<tr>
<td><strong>Prediction 4</strong></td>
<td>The donkey's kids use their breath to write 'We love you, Dad' in the light-colored sky.</td>
<td>112.3</td>
<td>0.0</td>
<td>23.1</td>
<td>25.2</td>
</tr>
</tbody>
</table>

*: to compute CIDEr for one sample, we use tf-idf from coco-eval

Source: Shrek The Third (2007)
New metric 1: LLM-AD-Eval

Prompts:
You are an intelligent chatbot designed for evaluating the quality of generative outputs for movie audio descriptions.

Please evaluate the following movie audio description pair:
Correct Audio Description: \{text\_gt\}
Predicted Audio Description: \{text\_pred\}

Provide your evaluation only as a matching score where the matching score is an integer value between 0 and 5, with 5 indicating the highest level of match.

Answer: 4

Perfect score = 5 for each AD


• "Can large language models be an alternative to human evaluations?", Cheng-Han Chiang, Hung-yi Lee, arXiv:2305.01937, 2023

### Problem of classical captioning metrics

<table>
<thead>
<tr>
<th>AD Sentence</th>
<th>CIDEr*</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE-L</th>
<th>LLM-AD-Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference: The donkeys make a smoke message in the sky which reads, we love you, Daddy.</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Prediction 1: Donkeys use smoke to write ‘We love you daddy’ in the sky.</td>
<td>302.9</td>
<td>25.4</td>
<td>30.8</td>
<td>43.6</td>
<td>5</td>
</tr>
<tr>
<td>Prediction 2: Donkey’s children write, we love you daddy, in the pale sky with their smoky breath.</td>
<td>226.5</td>
<td>18.9</td>
<td>20.3</td>
<td>25.9</td>
<td>5</td>
</tr>
<tr>
<td>Prediction 3: The young donkeys write ‘Love you, daddy’ in the sky.</td>
<td>187.0</td>
<td>0.0</td>
<td>25.7</td>
<td>38.6</td>
<td>4</td>
</tr>
<tr>
<td>Prediction 4: The donkey’s kids use their breath to write ‘We love you, Dad’ in the light-colored sky.</td>
<td>112.3</td>
<td>0.0</td>
<td>23.1</td>
<td>25.2</td>
<td>4</td>
</tr>
</tbody>
</table>

*: to compute CIDEr for one sample, we use tf-idf from coco-eval

Source: Shrek The Third (2007)
New metric 2: CRITIC
-- Co-Referencing In Text for Identifying Characters

Objective: evaluate whether the characters are referred correctly

Ground truth
AD reference

Predicted AD

(a)

(b)

(c)
New metric 2: CRITIC
-- Co-Referencing In Text for Identifying Characters

Objective: evaluate whether the characters are referred correctly

Ground truth
AD reference

Predicted AD

Co-referencing Cluster for each AD:
- [Rose Dewitt Bukater, Rose]; [Jack Dawson, Jack, Mr. Dawson]
- [Rose Dewitt Bukater, Rose]; [Jack Dawson, Jack, Mr. Dawson]
- [Jack Dawson, Jack, Mr. Dawson]
- [Jack Dawson, Jack, Mr. Dawson]
New metric 2: CRITIC
-- Co-Referencing In Text for Identifying Characters

Objective: evaluate whether the characters are referred correctly

Ground truth
AD reference

Predicted AD
## Quantitative results

<table>
<thead>
<tr>
<th>Method</th>
<th>CMD-AD-Eval</th>
<th>MAD-Eval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CIDEr</td>
<td>R@1/5</td>
</tr>
<tr>
<td>Video-BLIP2 [69] (no ft)</td>
<td>4.8</td>
<td>22.0</td>
</tr>
<tr>
<td>Video-Llama2 [73] (no ft)</td>
<td>5.2</td>
<td>23.6</td>
</tr>
<tr>
<td>AutoAD-I [20]</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>AutoAD-II [21]</td>
<td>13.5</td>
<td>26.1</td>
</tr>
<tr>
<td>Movie-BLIP2 (ours)</td>
<td>22.3</td>
<td>29.8</td>
</tr>
<tr>
<td>Movie-Llama2 (ours)</td>
<td>25.0</td>
<td>31.2</td>
</tr>
<tr>
<td>MM-Narrator + GPT4</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>MM-Narrator + GPT4v</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Movie-BLIP2 (from Video-BLIP)
Movie-Llama2 (from Video-Llama)

* = not trained on MAD-Train

MM-Narrator: Narrating Long-form Videos with Multimodal In-Context Learning. Zhang et al. 2023
Qualitative examples of AutoAD-III: `< [title] ground-truth || prediction >`
Summary

• Generating AD as a new task
  – Well defined task, so can be evaluated
  – Long form video understanding

• Appraisal
  – Have usable model for AD
  – Character bank significantly improves character naming in AD
  – General method for providing supplementary visual information as a prompt

• The future
  – Place and object banks/memory
  – Use of audio stream
  – Beyond AD: conversation and Q & A with model
Publications and resources

Authors: Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, Andrew Zisserman

• AutoAD: Movie Description in Context, CVPR 2023

• AutoAD II: The Sequel - Who, When, and What in Movie Audio Description, ICCV 2023

• AutoAD III: The Prequel - Back to the Pixels, On arXiv soon

• Datasets and models: https://www.robots.ox.ac.uk/~vgg/research/autoad/
  – MAD-v2: https://github.com/Soldelli/MAD
  – AudioVault: subtitles and AD for 7000+ movies