

Video Temporal Analysis Techniques

Ryan Amundson

<https://www.linkedin.com/in/ryanamundson1>

<https://github.com/ryanamundson1>

ryanamundson1@yahoo.com

Northwestern University School of Professional Studies

339 EAST CHICAGO AVENUE, CHICAGO, IL

Abstract

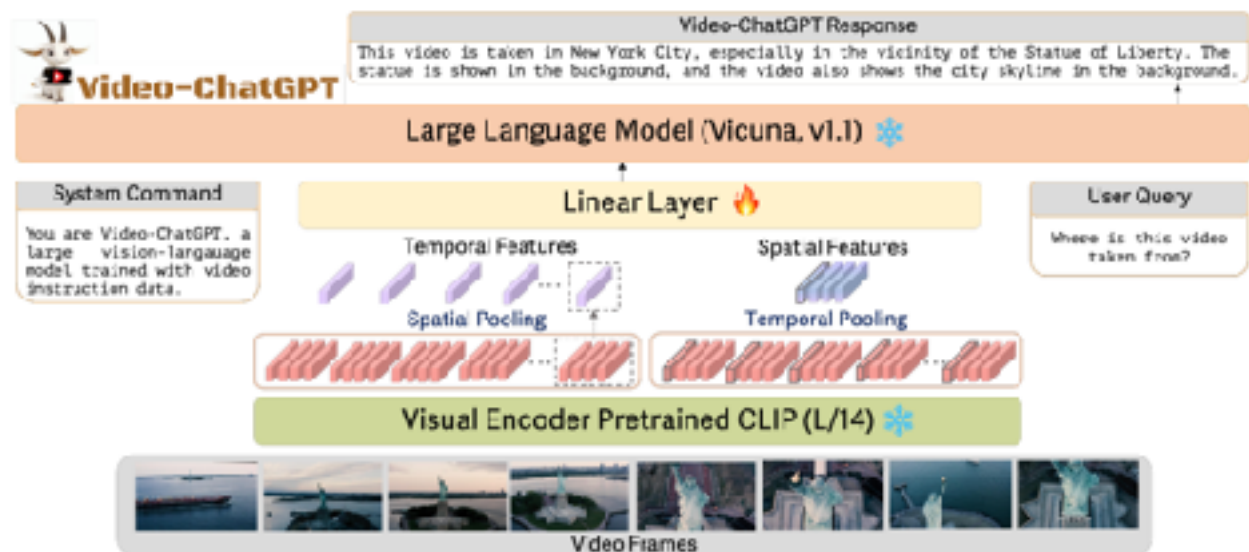
The goal for this report is to investigate methods to infer temporal data for video understanding. This is useful in the media and entertainment industries for many use cases, including script breakdowns for ad placement, video summaries for content, and context specific relationships in the video data. The need for useful methods of understanding video content is growing. Video accounted for 82.5% of global internet traffic in 2022 and the number of people consuming video traffic is constantly growing (Muuga 2024). As of February, 2024 there are almost 4 billion videos on YouTube alone (Broz 2024). The need for ways to analyze, summarize and ask questions of video content could not be greater.

Introduction

Video analysis is a pervasive problem in the media industry. Because videos are often relatively large sized and can consist of image content, text data from closed captioning, and synchronized audio data, video is often complex and computationally expensive to deal with. There are not many standardized tools and AI models to handle this multimodal data present in video. An additional complexity is the temporal nature of the video data, as the meaning in the current frame is very dependent on what happened previously. There is still quite a lot of manual work involved in annotating and/or summarizing video data to outline scene data to determine what characters are present and what is happening in the scene, which is costly, time consuming and prone to human error.

Literature Review

An initial algorithm that was investigated for semantic video understanding and question-answering was Video-CHATGPT (Maaz 2023). Video-CHATGPT combines Vicuna, a fine-tuned self-attention based transformer large language model (Zheng 2023) with a CLIP model. The CLIP model transforms images into instructive and/or descriptive text. It's similar to the LLaVA, Large Language and Vision Assistant (Liu 2023), but the transformer model is fine-tuned for temporal video data. The core model used is based on the original pre-trained LLaMA (Touvron 2023) set of models. The figure below is the architecture overview of Video-CHATGPT.



(Figure 1: Maaz 2023, architecture diagram of Video-ChatGPT)

Another way video data can be analyzed is to extract the dialog content and see if the text could be used to summarize the action and dialog. The Whisper

(Radford 2022) transformer model has been shown to be very effective at audio to text conversions. Once the dialog is extracted, the resulting text can be given to OpenAI's ChatGPT (OpenAI 2023) service for a plot and character summary.

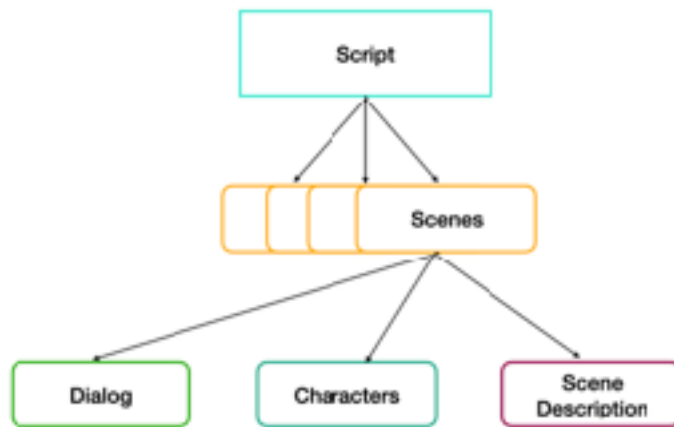
Most of the state of the art current architectures for video analysis are built on top of the original Transformer model presented in the "Attention Is All You Need" (Vaswani 2017) paper, but there are interesting newer architectures that might be better suited for the challenges video presents, like Mamba structured state space models (Gu 2023) and Mirasol3B (Noble 2023).

Data

The TV video long-form episode data for this project was sourced from the Internet Archive to see if effective summaries of the videos could be produced using the current generation video models. Some additional data was sourced from the TVQA+ dataset. This is a dataset that consists of television episodes and asks questions that require temporal understanding. As TVQA+ uses TV media data, this is perfect for many use cases that would occur in the TV media and film industry. The ideal use case would be to determine if a script could be generated from the video data.

This is inherently difficult because it would need to describe the frames along with understanding the spoken dialog.

Here is the simplified ontological breakdown of script data:



(Figure 1: Ontological breakdown of script characteristics)

You can see in the ontology of a script how it requires data from both the audio track for dialog and the video images for scene descriptions and characters. Scripts and video summaries are often produced after the fact by third-party vendors because the show's working script might not contain many of the ad-libbed or live-action lines.

The pulled data consists of 5 episodes of "Parks and Rec", which is a popular comedy series that ran from 2009-2015 on NBC. The videos are comprised of a typical 22 minute 30fps comedy series in 853×480 pixel mp4 format, which is

approximately 34,000 frames with an accompanying synchronized audio track. Another video used was from the show Farscape. This is a science fiction drama series with a longer runtime of 50 minutes, which is typical for an hour long television episode.

Methods

The first method of video understanding and summary was to use Video-CHATGPT (Maaz 2023). The code was updated to allow it to run locally using the MPS architecture for Apple silicon.

The second method of video analysis is to extract the dialog from the MP4 video file using the Whisper (Radford 2022) model and then use the resulting dialog in a prompt to the ChatGPT service.

Online cloud services exist that can extract information from video data. One interesting one in particular is Microsoft Azure's AI Video Indexer. It is a black-box cloud service that runs a variety of AI models on both the image frames and audio track and will index them for various analytics.

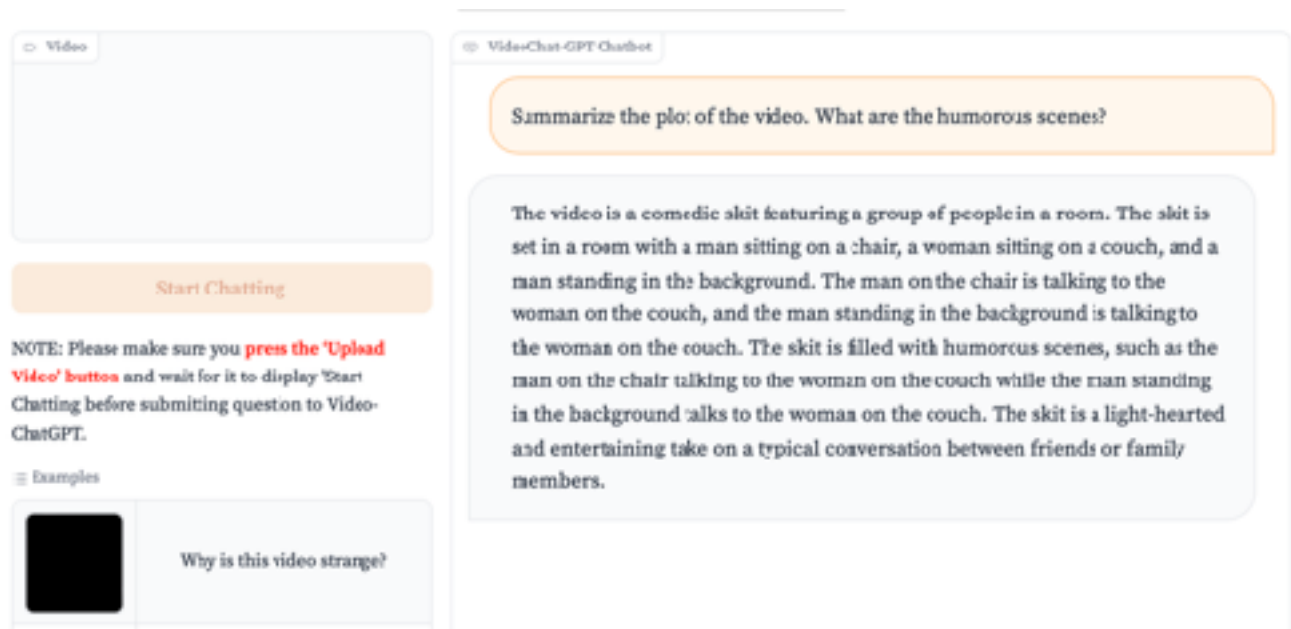
One last method that can also be used to extract information and do video Q&A is to pull individual frames and summarize the action and content in each one. The videos will be sampled at a rate of ~1 frame per second to reduce the

processing, each image frame will then be passed to Salesforce's Bootstrapping Language-Image Pre-training (Li 2023) BLIP model, which is similar to CLIP to produce a series of captions with timecode information. The dialog content for that time slice will also be included the using the Whisper model. All of the resulting data and the image content will then be indexed in a vector database so that a RAG (Retrieval Augmented Generation) based chatbot using a multimodal LLM can be created using the dialog and captions for context retrieval. OpenAI's GPT-4V(ision) model will be used to generate the final content for the multimodal chat system.

Results

There were some significant downsides and limitations to using Video-CHATGPT. One limitation is performance and CUDA only support. Some issues with performance and memory could be addressed by code refactoring and minor modifications. Other limitations seen are that the model will only look at 100 frames maximum, so those are extracted evenly across the video. In a 22 minute television video, at 30 fps, that gives approximately 34k frames, so 100 frames does not give a very good picture of what happens in a typical television episode. Also, using only the frames and not the dialog fails to really describe what happens in the episode

The figure below shows the result of asking for a plot summary of Season 7, episode 2 of the show “Parks and Rec”, which isn’t really on point.



(Figure 3: Video-ChatGPT gradio output)

“The video is a comedic skit featuring a group of people in a room. The skit is set in a room with a man sitting on a chair, a woman sitting on a couch, and a man standing in the background. The man on the chair is talking to the woman on the couch, and the man standing in the background is talking to the woman on the couch. The skit is filled with humorous scenes, such as the man on the chair talking to the woman on the couch while the man standing in the background talks to the woman on the couch. The skit is a light-hearted and entertaining take on a typical conversation between friends or family members.”

This really exemplifies the problems with trying to understand long-form video content with the current technologies, with strictly selective images and descriptive language, it doesn’t give a complete picture of long-form content.

A second method of using easily accessible technology is to extract the dialog from the MP4 video file using the Whisper (Radford 2022) model. The resulting dialog can be passed to OpenAI's ChatGPT-4 (OpenAI 2023) for summarization. The text below demonstrates the result of that summary, which is much improved from the summary derived by only using the video frames:

"In this TV show episode, a character named Leslie is desperately trying to persuade her former City Councilmate Jeremy Jam to vote against rezoning land for commercial use. Jeremy, manipulated by his girlfriend Tammy, is initially planning to vote in favor. Leslie, with the help of Ron, manages to convince Jeremy to break up with Tammy, therefore influencing his final vote. Meanwhile, another character named Tom recruits his ex-girlfriend Lucy for his expanding business. Lastly, April, disillusioned with her career, explores a potential job at a mortuary, which quickly takes a comedic turn. The episode blends comedy, politics, and personal growth, highlighting the eccentricity of each character and the lengths they are willing to go to achieve their goals."

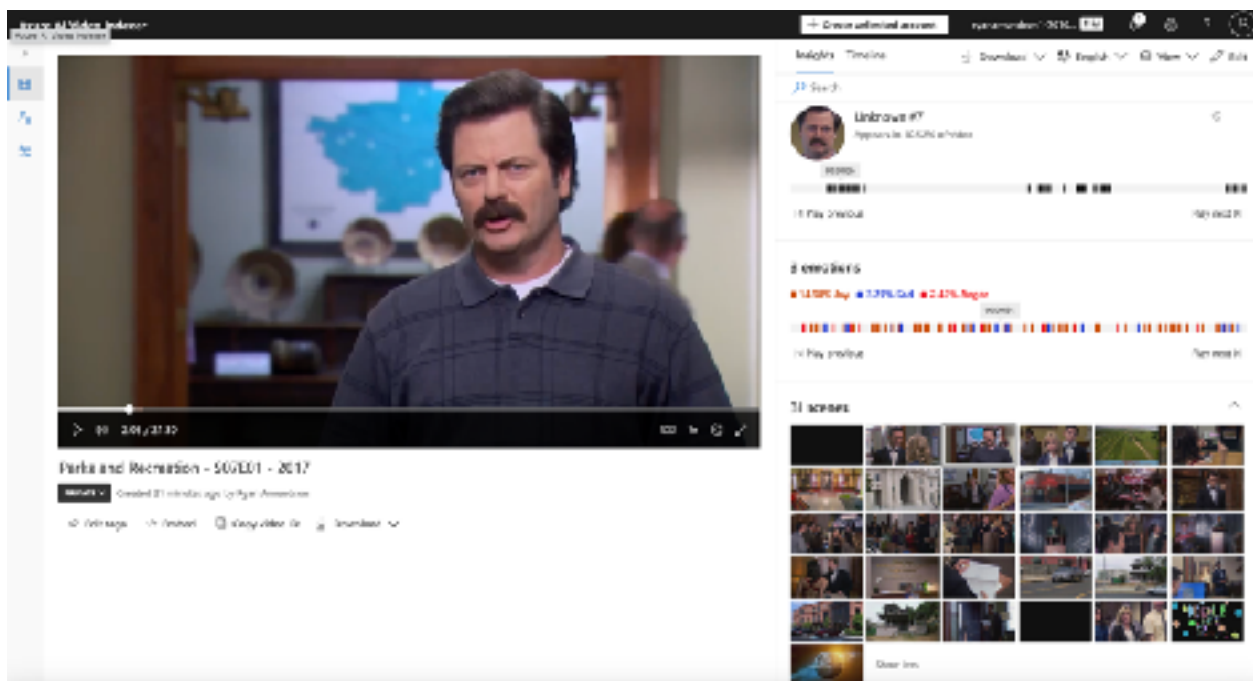
This summary is quite reasonable. It was possible to also obtain the character list from the episode using the given dialog and GPT-4, which is very usable as a character list/analysis for the episodic content:

- "1. Leslie: She is a very determined and ambitious woman who works for the city council. She is in a bitter rivalry over a zoning issue with Ron Swanson and Jeremy Jam.
2. Jeremy Jam (also known as Jam or Jammy): He is a member of the city council, who Leslie has a contentious relationship with. Currently manipulated by Tammy his girlfriend.
3. Tammy: Jeremy's manipulative girlfriend, she is hell-bent on controlling Jeremy and tries to manipulate the zoning vote in her favor.
4. Joan: A relatively eccentric TV show host who loves her job and gives a speech that makes April question her own career.
5. April: An assistant at the city council, who is uncertain about her job and is considering quitting after reflecting on Joan's speech.

6. Ron Swanson: Leslie's antagonist and opponent in the zoning issue. He is also known for having a tumultuous relationship with Tammy in the past.
7. Tom: An entrepreneur who is considering hiring his former girlfriend, Lucy, in his company.
8. Lucy: Tom's ex-girlfriend who resides in Chicago, and Tom is considering getting back together with her.
9. Andy: A friend and confidant of Tom, he encourages Tom to pursue Lucy and accompanies him to Chicago.
10. Donna: A friend of April's who tries to help April find a career path she could enjoy. She seems content at her job and desires to assist April in finding her own happiness.
11. Ben: He along with April visits a mortician's place for a possible job interest for April.

From the context, there are a few more characters like Commissioner Gordon, Buddy Holly, Victoria Herzog, and others, however, they don't present during the direct dialogue, but are only mentioned in passing or indirectly. The ones listed above are primary characters with significant roles."

In investigating some of the cloud service options, Microsoft Azure's Video Index proved to be a useful technology. It is a tool that will separately process the image and audio content and provide indexed results. Here is a screen shot of some of the features it provided:



(Figure 4: Microsoft Azure video indexer output)

The video indexer is capable of properly discerning scenes and shots, along with several other analytics capabilities, like identifying which actors are in what shots, along with sentiment analysis. As a black-box solution, it seems very useful, however, it is not customizable or capable of video Q&A or summary analysis.

For the final method of creating a RAG system using the BLIP model for captioning, using Whisper for dialog text extraction and then storing the models output in a vector database for a retrieval based context to send, along with the corresponding frame images, to a multi-modal hosted model, in this case, GPT-4V. This allowed for a question and answer chat system that seems to have good potential to analyze and understand components of the video content. For instance, the text below displays some output to the question “What are some location establishment shots? Where are the locations?”. In this case, the retrieval system sent the proper establishing shot images with captions and dialog to GPT-4V, which was able to further elucidate information about those shots based on context.

“The image provided appears to be a night shot of the front entrance to a hotel. The visual elements include the illuminated facade of the building, with a large arched window and a well-lit entrance that suggests a welcoming atmosphere. The architecture seems to be of a classic style, and the street is quiet, indicating that the photo was taken at a time when there is little to no pedestrian or vehicular traffic. There are no people visible in the image, so it serves as an establishing shot, setting the scene for a location where events in the film or TV show will unfold.

The caption provided with the timecode indicates that this image is likely from a scene at approximately 7 minutes and 40 seconds into the film or show. The caption reads "the front entrance to a hotel at night," which aligns with the visual content of the image. The dialog, "And on the phone to Washington, call in every favor that we have," suggests a scene involving characters who are likely to be involved in some form of political or high-stakes situation where they need to leverage their connections.

The type of shot used here is an establishing shot, which is commonly used in film and television to set up the context for a scene, showing the audience where the action is taking place. Establishing shots are often used at the beginning of a new sequence to establish the setting and to transition the audience to a different location.

The other captions and dialog provided seem to be from different scenes, each with their own timecodes:

- At 0:15:39.449524, a car is parked in front of an old house, accompanied by dramatic music, which could indicate a suspenseful or significant moment related to that location.
- At 0:15:48.790649, the entrance to the city hall building in San Francisco is shown, with the dialog "What up player?" This could be a casual greeting between characters, possibly in a more light-hearted or comedic context.
- At 0:16:59.516235, there is an office building located in a parking lot, with the dialog "Jessica, Mr. Nelson." This suggests an interaction between characters, likely taking place in a professional setting.

Each of these descriptions and dialogs corresponds to a different location and scene, and they are all examples of location establishment shots that provide the audience with information about where the subsequent action will take place. The connections between the images and the related text are that the text serves to give context to the visual elements, indicating the setting and sometimes the mood or tone of the scene through dialog or music."



(Figure 5: Question & answer chatbot output)

Conclusions

The video models are very bleeding edge technology right now, and that comes with a significant number of pains in adopting the technology. It seems like they are testing the limits of the self-attention transformer models. OpenAI's ChatGPT service is capable of returning very usable results, but this requires calling out to a remote service and communicating with a closed source service, which can be a non-starter in some cases.

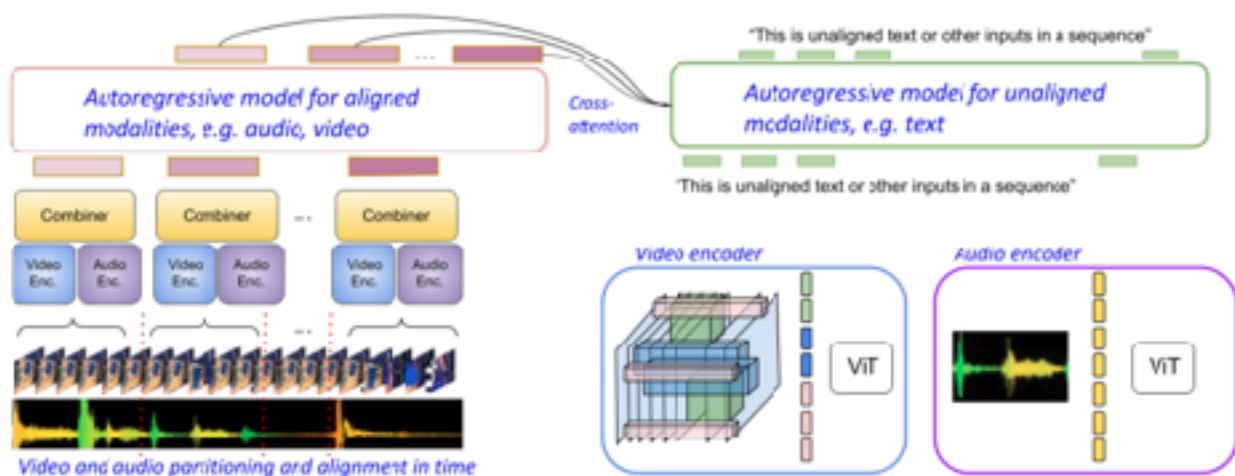
There are a number of potential advances in the horizon, both with Mirasol3B (Noble 2023) and Mamba SSM (Gu 2023), but there is not yet an openly available pre-trained multi-modal model using those architectures. Hopefully those will be

available soon, as those architectures could really advance video understanding models to the next level.

Future Directions

Some future directions that the project could take would be to update the transformer based architecture models with the Mamba SSM model (Gu 2023). The much more scalable and increased token length of Mamba would make for a much more effective base model than the current-generation transformer models.

Another direction to take this video analysis project would be to look into the multi-modal Mirasol3B model. This model combines temporal audio and video data.



(Figure 6: Noble 2023, architecture diagram of Mirasol3b)

From the research papers, this looks like an effective solution to dealing with long-form video content analysis.

An additional research avenue would be to investigate Google's newest multi-modal modal, Gemini (Pichai 2023), which could potentially be used as a replacement for OpenAI GPT-4V.

Appendix/appendices

Figure 1: Maaz 2023, architecture diagram of Video-ChatGPT from <https://github.com/mbzuai-oryx/Video-ChatGPT>

Figure 2: Ontological breakdown of script characteristics

Figure 3: Local gradio output of Video-ChatGPT analysis of episode 2, season 7 of Parks and Recreation

Figure 4: Microsoft Azure video indexer output of episode 1, season 7 of Parks and Recreation

Figure 5: Question & answer dialog extraction and BLIP captioning to GPT-4V chatbot output of episode 1, season 7 of Parks and Recreation

Figure 6: Noble 2023, architecture diagram of Mirasol3b from <https://blog.research.google/2023/11/scaling-multimodal-understanding-to.html>

Modified Video-CHATGPT project code with updated to run on Apple MPS:
<https://github.com/ryanamundson1/Video-ChatGPT>

Chatbot code for dialog extraction and BLIP captioning to GPT-4V: https://github.com/ryanamundson1/MSDS/blob/main/MSDS-498/final_project_llm_chatbot.ipynb

References

- Zheng, Lianmin, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, et al. “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena.” arXiv.org, December 24, 2023. <https://arxiv.org/abs/2306.05685>.
- Frangidis, Paschalis, Konstantinos Georgiou, and Stefanos Papadopoulos. “Sentiment Analysis on Movie Scripts and Reviews.” SpringerLink. Springer International Publishing, May 29, 2020. https://link.springer.com/chapter/10.1007/978-3-030-49161-1_36.
- Maaz, Muhammad, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. “Video-CHATGPT: Towards Detailed Video Understanding via Large Vision and Language Models.” arXiv.org, June 8, 2023. <https://arxiv.org/abs/2306.05424>.
- Mourchid, Y., Renoust, B., Cherifi, H., El Hassouni, M. (2019). Multilayer Network Model of Movie Script. In: Aiello, L., Cherifi, C., Cherifi, H., Lambiotte, R., Lió, P., Rocha, L. (eds) Complex Networks and Their Applications VII. COMPLEX NETWORKS 2018. Studies in Computational Intelligence, vol 812. Springer, Cham. https://doi.org/10.1007/978-3-030-05411-3_62
- Wang, Shijie, et al. “Vamos: Versatile Action Models for Video Understanding.” *arXiv.Org*, 22 Nov. 2023, arxiv.org/abs/2311.13627.
- Lei, Jie, et al. “TVQA+: Spatio-Temporal Grounding for Video Question Answering.” *arXiv.Org*, 11 May 2020, arxiv.org/abs/1904.11574.
- Vaswani, Ashish, et al. “Attention Is All You Need.” *arXiv.Org*, 2 Aug. 2023, arxiv.org/abs/1706.03762.
- OpenAI “Clip: Connecting Text and Images.” CLIP: Connecting text and images. Accessed January 29, 2024. <https://openai.com/research/clip>.
- Liu, Haotian, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. “Visual Instruction Tuning.” arXiv.org, December 11, 2023. <https://arxiv.org/abs/2304.08485>.

Internet archive: Digital Library of Free & Borrowable Books, movies, Music & Wayback Machine. Accessed January 29, 2024. https://archive.org/details/classic_tv_1990s.

Gu, Albert, and Tri Dao. “Mamba: Linear-Time Sequence Modeling with Selective State Spaces.” arXiv.org, December 1, 2023. <https://arxiv.org/abs/2312.00752>.

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. “Robust Speech Recognition via Large-Scale Weak Supervision.” arXiv.org, December 6, 2022. <https://arxiv.org/abs/2212.04356>.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, et al. “GPT-4 Technical Report.” arXiv.org, December 19, 2023. <https://arxiv.org/abs/2303.08774>.

Noble, Isaac, and Anelia Angelova. “Scaling Multimodal Understanding to Long Videos.” – Google Research Blog. Accessed February 4, 2024. <https://blog.research.google/2023/11/scaling-multimodal-understanding-to.html>.

Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, et al. “Llama: Open and Efficient Foundation Language Models.” arXiv.org, February 27, 2023. <https://arxiv.org/abs/2302.13971>.

Muuga, Erkki. “10 Important Video Marketing Statistics to Help You in 2024.” influno, January 4, 2024. <https://influno.com/video-marketing-statistics/#:~:text=2.,That's%20a%20lot%20of%20traffic>.

Broz, Matic. “The Latest Mindblowing YouTube Video Numbers of February 2024.” How many videos are on YouTube?, February 18, 2024. <https://photutorial.com/how-many-videos-on-youtube/#:~:text=As%20of%20February%202024%2C%20there,of%20which%20are%20YouTu be%20Shorts.&text=YouTube%20is%20the%20most%20popular,when%20you%20use%20our%20data>.

Jin, Peng, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. “Chat-Univi: Unified Visual Representation Empowers Large Language Models with Image and Video Understanding.” arXiv.org, November 14, 2023. <https://arxiv.org/abs/2311.08046>.

Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi. “Blip-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models.” arXiv.org, June 15, 2023. <https://arxiv.org/abs/2301.12597>.

Microsoft. “Azure AI Video Indexer – Video Analyzer for Media: Microsoft Azure.” Video Analyzer for Media | Microsoft Azure. Accessed March 4, 2024. <https://azure.microsoft.com/en-us/products/ai-video-indexer>.

OpenAI. “GPT-4V(ISION) System Card.” GPT-4V(ision) system card. Accessed March 4, 2024. <https://openai.com/research/gpt-4v-system-card>.

Pichai, Sundar. “Introducing Gemini: Our Largest and Most Capable AI Model.” Google, December 6, 2023. <https://blog.google/technology/ai/google-gemini-ai/>.