Demo: On-Device Video Analysis with LLMs

Vishnu Jaganathan, Deepak Gouda, Kriti Arora, Mohit Aggarwal, Chao Zhang
{vjaganathan3,deepakgouda,kriti,mohit7,chaozhang}@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Abstract

We present a new on-device pipeline that efficiently summarizes lecture videos and provides relevant answers directly from a smartphone. We utilize widely accessible tools like OCR and Vosk speech-to-text, coupled with powerful large language models (LLMs), to identify crucial sentences and generate summaries. By harnessing the capabilities of LLMs and the computational power of mobile devices, we fine-tune and quantize BERT and GPT-2 to achieve efficient lecture video summarization and question answering on consumer-grade smartphones like the Pixel 8 Pro. Notably, this approach eliminates the need for cloud APIs, ensuring enhanced user privacy and minimal mobile data usage.
https://www.youtube.com/shorts/zwGdONlKays

CCS Concepts: • **Computing methodologies → Video summarization**; • **Human-centered computing → Smartphones**.

Keywords: LLM, on-device ML, video understanding

ACM Reference Format:
Vishnu Jaganathan, Deepak Gouda, Kriti Arora, Mohit Aggarwal, Chao Zhang. 2024. Demo: On-Device Video Analysis with LLMs. In *The 25th International Workshop on Mobile Computing Systems and Applications (HOTMOBILE '24), February 28–29, 2024, San Diego, CA, USA.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3638550.3643052

1 Introduction

Finding specific information in lengthy lecture videos can be challenging and time-consuming for students, especially during revision. Recent deep learning techniques for video summarization that require high end GPUs are inaccessible to many due to hardware limitations. Additionally, using online video analysis services can be costly, risk exposure of proprietary content, and suffer from high latency. Our objective is to design an efficient pipeline suitable for high-end smartphones. These recent smartphones are equipped with advanced chipsets, enabling them to handle LLMs with hundreds of millions of parameters. After using OCR and Vosk to process the video input, our approach integrates BERT-QA for answering questions and employs a version of GPT-2 [3] we fine tune and quantize for concise summarization.

2 Design

Our pipeline uses a combination of off-the-shelf and customized components. As shown in Figure 1, we first build a text context to use with the LLMs. We use an open-source android library Vosk to transcribe the lecture, and use OCR to get text from the slides. For choosing which frames to read, we use a pixelwise L2 loss to detect when the slide changes significantly, and extract this new text. Concatenating these two modes together gets us the full context. For the Q&A task, we use a quantized BERT-QA model, which is BERT [1] fine-tuned to answer questions within a limited context. The model iteratively processes text chunks that fit this context, extracting sentences that answer the question, and then evaluates the best match for the question amongst these answer candidates. For the summarization task, we fine tune a small version of GPT-2 on an article-summary dataset. This enables GPT-2 to produce one sentence summaries of the given context. We quantize our fine-tuned model and deploy it on device with the same hierarchical strategy as BERT-QA to avoid context length limits.

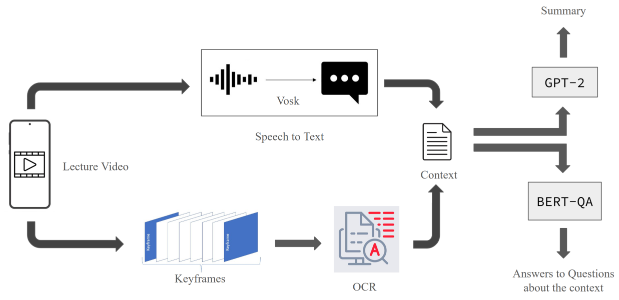

Figure 1: Pipeline used to process videos and extract summary/QA info on android app.

3 Related Work

To the best of our knowledge, this is the first work to demonstrate video summarization and/or Q&A on a mobile device. However, there are similar works such as [2] that utilize multimodal features for summarization with desktop GPUs.

4 Demo

We plan to show a live demo on a Pixel 8 Pro smartphone, allowing users to choose a lecture video of their choice to analyze. They can summarize the key points of the video or choose to ask questions about the content, and compare the results to the actual video.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).
[2] Haopeng Li, Qiuhong Ke, Mingming Gong, and Tom Drummond. 2023. Progressive Video Summarization via Multimodal Self-supervised Learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision.* 5584–5593.
[3] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

HOTMOBILE '24, February 28–29, 2024, San Diego, CA, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0497-0/24/02.
https://doi.org/10.1145/3638550.3643052