WHITE PAPER

EXTRACTING INTELLIGENCE FROM CORPORATE STREAMING USING MACHINE LEARNING

CONTENTS

INTRODUCTION

While video technology and video streaming have advanced considerably within the enterprise, the introduction of cloud services and machine learning to video applications promises to greatly improve productivity within these organizations. Before the internet became ubiquitous, corporate streaming was largely confined to prerecorded content accessed via taped media libraries used in trainings, sales, or advertising. Larger organizations could broadcast live video to satellite offices but this was relatively rare, a capability reserved primarily for broadcast television.

With the advent of the internet, and specifically internet protocol (IP)-enabled video, services proliferated to help organizations of any size engage in online video conversations, broadcast town-hall meetings, access on-demand videos for training, and enable other methods of collaboration. Productivity increased because video content reached a wider audience within or across organizations in shorter amounts of time.

Netflix first applied machine learning to user preference prediction and its movie recommendation engine in 2006. Ever since, organizations have been fascinated by the potential for machine learning to optimize technology workflows and algorithmic predictions. In a survey conducted by Wainhouse Research in 2017 (Source: WR Enterprise Web Communications Survey, 4Q2017), 73 percent of survey respondents believed that when applied to video data, machine learning was an important factor in influencing purchasing decisions. However, this sort of computational modeling had historically been relegated to organizations with the necessary brain power, hardware, and accompanying budgets and was therefore out of reach of the majority of enterprises. Recently, efforts to abstract the complexities of machine learning into use case-based offerings have empowered organizations of any size to take advantage of machine learning without hiring a staff of programmers or data scientists.

In this paper, we explore the impact of machine learning on corporate streaming by examining how it is changing multiple video workflows. In addition, we delve into how organizations can leverage existing machine learning technologies and provide actionable steps on how to implement those technologies to stay competitive.

BASIC BUILDING BLOCKS

Organizations such as AWS have purposely built building blocks based on certain capabilities such as speech-to-text, object identification, and video clipping which, when combined with triggers, will permit any organization to design more sophisticated workflows that function in an automated manner. This can help organizations simplify even the most complex and time-consuming tasks. While AWS provides multiple machine learning higher-level services, the most important of these functions for corporate streaming include Amazon Rekognition, Amazon Transcribe, and Amazon Translate. Each of these services are described briefly below and, when used in conjunction with AWS Elemental Media Services, allows an organization to alleviate the workload behind several tasks.

- Amazon Rekognition — intelligent image and video analysis to detect objects, people, text, scenes, and activities, as well as to detect any inappropriate content. These features make it possible for media professionals to easily extract metadata from their content and then use the metadata to create innovative solutions.
- Amazon Transcribe — automatic speech recognition (ASR) service that makes it easy to add speech-to-text capability. The service provides time stamps for every word to help easily locate the audio in

the original source by searching for the text, an ability to expand and customize the speech recognition vocabulary, and a speaker identification feature.

- Amazon Translate — neural machine translation service that delivers fast, high-quality, and affordable language translation.

TYPICAL WORKFLOWS

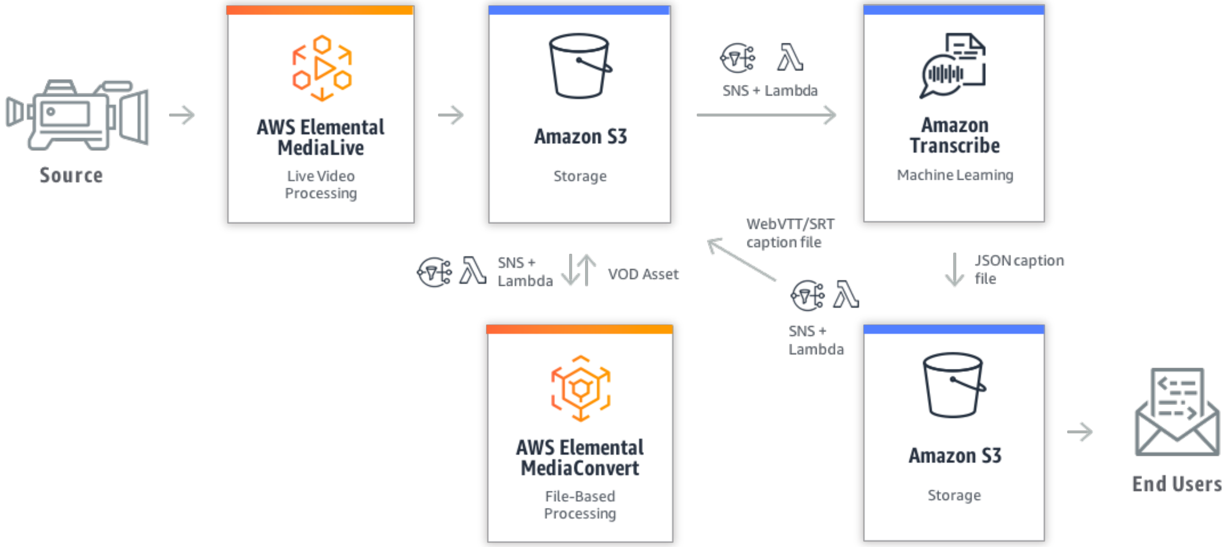
APPLICATION: AUTOMATICALLY CAPTION VIDEO MEETINGS

The productivity of meetings, including video conference calls, are enhanced when a transcript of what is said is readily available to engender commitment and avoid any confusion among participants about what was said to whom. An October 2015 survey[1] of 1000 working professionals found that 95 percent of respondents said they find it easier to remember things when they've written them down.

Transcription services have been around for some time but until the advent of speech-to-text in the cloud, they were costly. Pay-as-you-go services such as Amazon Transcribe make transcription of meeting discussions not only simple but very cost-effective. And because Amazon Transcribe can identify multiple speakers, there won't be a mix-up as to who said what.

> *"No matter how big, small, simple or complex an idea is, get it in writing... If you don't write your ideas down, they could leave your head before you even leave the room."*
> *Richard Branson, Virgin Group Founder, February 2015*[2]

Tools: AWS Elemental MediaLive, AWS Elemental MediaConvert, Amazon Simple Storage Service (S3), Amazon Transcribe



[1] https://www.prnewswire.com/news-releases/new-survey-shows-professionals-prefer-pen-and-paper-note-taking-for-increased-productivity-and-retention-300177380.html
[2] https://www.linkedin.com/pulse/importance-taking-notes-richard-branson/)

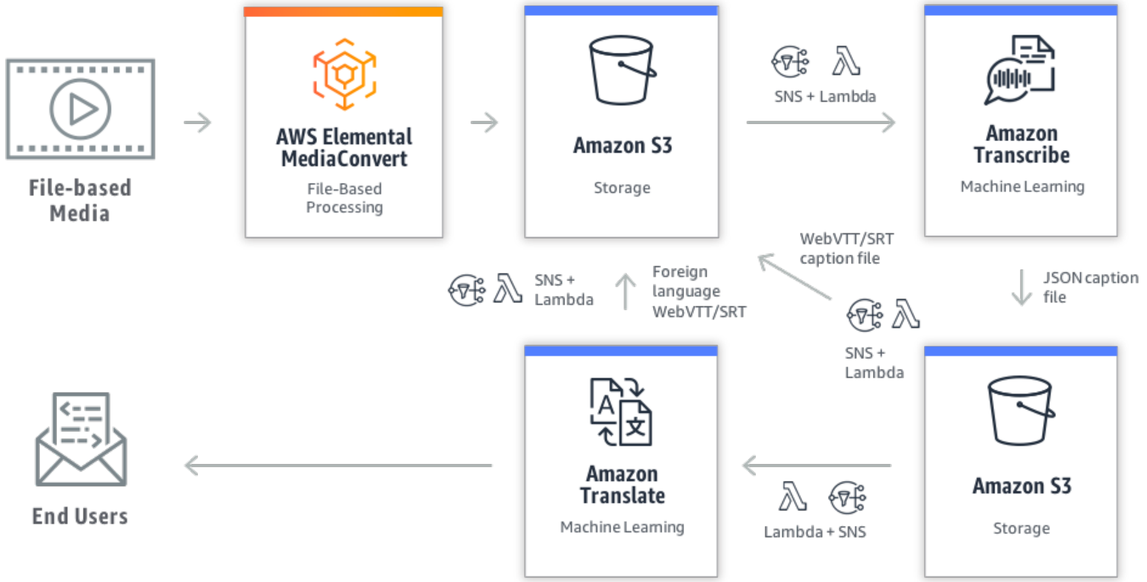
Basic Workflow: Automated transcription of live audio or video conferencing

- MediaLive deposits the live video stream into a S3 bucket (assumes defined start and stop times)
- The availability of the file triggers an AWS Lambda function for MediaConvert to convert the file into a VOD asset, depositing the file back into the S3 bucket
- The availability of the new file triggers an AWS Lambda function stripping out the audio using FFmpeg or equivalent tool and then initiating Amazon Transcribe with that audio file
- Amazon Transcribe deposits the transcription (in JSON format) into an S3 bucket, and SNS sends an email to meeting participants with the transcript
- The availability of the JSON file triggers an AWS Lambda function to convert it to WebVTT or SRT format, as appropriate, so that it might be used with the VOD asset. The Lambda function then saves the SRT or WebVTT file in the same S3 bucket and directory as the VOD asset

APPLICATION: AUTOMATED CAPTIONING AND TRANSLATION

Whether a townhall meeting, an earnings report, a lecture or sermon, any video asset will benefit from transcription (speech-to-text) services. Transcription opens doors to other capabilities ranging from the ability to search for spoken words to more complex functions such as sentiment analysis and translation of that text into foreign languages. While transcription services have been around since the written word, the innovation enabled by machine learning is simplifying the task, as well as making it cost-effective and fast. Additionally, once that transcription is available, the ability to translate that text into multiple languages lets an organization share a video's message across many different regions, countries, and cultures.

Tools: AWS Elemental MediaConvert, Amazon Simple Storage Service (S3), Amazon Transcribe, Amazon Translate



Basic Workflow: Subtitling a file-based archive of a broadcast or webinar

- A video is encoded in the correct ABR suite by MediaConvert and deposited into a S3 bucket
- The availability of the file triggers an AWS Lambda function stripping out the audio using FFmpeg or equivalent tool and then initiating Amazon Transcribe with that audio file
- Amazon Transcribe deposits the transcription (in JSON format) into an S3 bucket
- The availability of the JSON file triggers an AWS Lambda function to convert it to WebVTT or SRT format, as appropriate, so that it might be used with the VOD asset. The Lambda function then saves the SRT or WebVTT file in the same S3 bucket and directory as the VOD asset

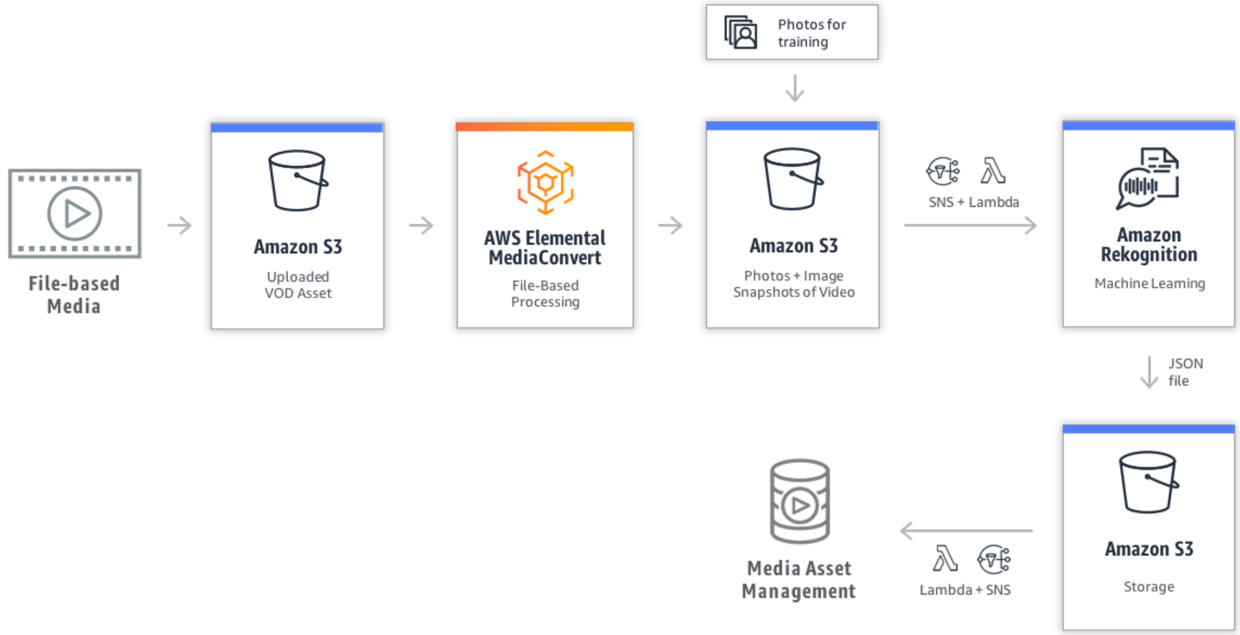
Advanced Workflow: Translation into multiple languages

- The availability of the JSON file in S3 also triggers an AWS Lambda function to initiate Amazon Translate (note that the languages for translation must be selected in advance)
- The translated JSON file triggers an AWS Lambda function to convert it to WebVTT or SRT format, as appropriate, so that it might be used as an appropriate language file with the VOD asset. The Lambda function then saves the SRT or WebVTT file in the same S3 bucket and directory as the VOD asset with the appropriate language designation

APPLICATION: AUTOMATED VISUAL RECOGNITION OF INDIVIDUALS, ACTORS, OBJECTS

Identifying where a key executive, an actor, or an object appears in videos is typically very cumbersome, and doing so at scale can be resource-prohibitive. For enterprises, the content team may need a simple method of searching for the appearance of executives or other individuals which appear in company videos. By tagging those videos with metadata indicating the appearance of specific individuals and objects — and even tagging the precise timestamp when those individuals appear — the company can enrich its video repository with robust search capabilities. Combined with a media asset management system that ingests metadata about the appearance of the recognized objects, object recognition tools can streamline search in content production and media consumption workflows.

Tools: AWS Elemental MediaConvert, Amazon Simple Storage Service (S3), Amazon Rekognition



Basic Workflow: Images to Amazon S3 to Amazon Rekognition to MAM (for combination with metadata)

- Upload the video asset into a S3 bucket
- Amazon Rekognition can already automatically recognize thousands of objects. Upload photos into a S3 bucket to quickly train Rekognition to recognize specific individuals
- Next, use MediaConvert to take static images of the video asset and deposit them into a S3 bucket, triggering an AWS Lambda function to submit the images to Amazon Rekognition (Note that Rekognition video can also be used but MediaConvert needs to convert the video to H.264, in MPEG-4 or MOV formats
- Amazon Rekognition will process and deliver JSON metadata of the timecode of recognized objects to the media asset management (MAM) system (Note that the resulting JSON should be formatted to be usable with the MAM system)

A NOTE ABOUT SECURITY AND PRIVACY

Underlying any workflow involving machine learning is a shared responsibility model for maintaining security and privacy of content and processes. AWS is responsible for the physical security of its infrastructure, detecting fraud and abuse, and responding to incidents by notifying customers. In turn, the customer is responsible for restricting the unauthorized sharing of data, enforcing compliance and governance policies, and identifying when a user misuses AWS. While AWS continues to innovate in machine learning, customers ultimately must ensure that those capabilities are used responsibly. For more on this topic, please visit the machine learning blog.[3]

FUTURE OF CORPORATE STREAMING

Machine learning provides the basic building blocks to enable organizations to quickly build much more complex solutions with simple API interfaces. AWS services, in particular, can easily be joined using AWS Lambda step functions in an “if-this-then-that” process flow, triggering additional services which can be combined to automate complex workflows. For example, the following use cases can now be simplified with machine learning:

- Corporate training looking to transcribe and segment every course in its massive library
- A townhall meeting for a global company that wants its audience to view the video recording with subtitles and audio localized in their native language
- PR team seeking all appearances of a company CEO in a catalog of corporate video

Automation of machine learning functions requires organizations to understand the basic building blocks, break down the workflow into individual sub-workflows that utilize the machine learning services, then determine the triggers that bind the pieces together. In the townhall example above, the organization combines the transcription workflow with the chaptering workflow to obtain transcribed text, segmented by individual chapters. Finally, the transcription is processed through a translation and speech-to-text workflow to be translated into multiple languages and recombined with the video as audio tracks for consumption. With traditional manual post-processing, an effort like this could take weeks and several individuals to accomplish what a carefully designed machine learning-workflow could output in a matter of hours.

[3] https://aws.amazon.com/rekognition/the-facts-on-facial-recognition-with-artificial-intelligence/

CONCLUSION

Organizations investigating the use of machine learning to optimize their corporate streaming workflows may have the misconception that they need cadres of data scientists and artificial intelligence-knowledgeable developers on staff to leverage existing technologies. In fact, real efficiencies are possible by understanding the existing services offered by today's cloud providers, identifying sub-optimal areas of video workflows which take too long or aren't cost-efficient, and then applying those cloud offerings to improve the workflow. With a fairly minimal effort, enterprises today can leverage the power of machine learning in their corporate streaming workflows.