IT Peer Network

SOLVING VIDEO DISCOVERABILITY USING CLOUD-SCALE DEEP LEARNING

Written by Joe Spisak | August 15, 2016

Co-written by Andres Rodriguez of Intel, Ravi Panchumarthy of Intel, Hendrik van der Meer of Vilynx, and Juan Carlos Riveiro of Vilynx

Setting the Stage

problems than ever before. However, new challenges are surfacing just as quickly as new opportunities. One such challenge/opportunity is the deluge of video content that is dominating the Internet and how to make sense of all of it. Per the Cisco 2016 Visual Networking Index [http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html] , 75% of the world's mobile data traffic will be video by 2020 and mobile video will increase 11-fold between 2015 and 2020.[1] [#_ftn1]

A parallel trend to the explosion of video traffic is the growth of machine learning in all of its forms including the use of deep convolutional neural networks (CNNs) for highly accurate image recognition tasks. The release of the seminal network topology 'AlexNet'[2] [#_ftn2] , architected by Alex Krizhevsky, Ilya Sutskever and Geoff Hinton at the University of Toronto, served as a major milestone in computer vision history by soundly beating traditional methods in the 2012 Large Scale Image Recognition Challenge (LSVRC 2012 [http://www.image-net.org/challenges/LSVRC/2012/]). This topology had and continues to have a profound influence in computer vision and image recognition deployed in many state of the art applications today. Modern neural networks contain amazing levels of representational power and accuracy due to their deep hierarchical approach to feature learning and abstraction. However, this accuracy requires a significant increase in the compute power required to train these deep networks — making cloud-based computing an ideal solution.

The confluence of cloud-based compute, machine learning and pervasive video content sets the stage for an interesting problem to solve. How can advanced computing systems automatically extract the salient points within video content to make them more easily discoverable, drive better user engagement, and ultimately be monetized? To that end, Intel and Vilynx* are working together to create a solution. As part of this effort, our engineering teams have jointly created a reference architecture that can be leveraged by the community of developers looking to tap into the power of cloud infrastructure to solve similar problems.

The Problem Statement: Video Discoverability is Broken

minute [http://www.reelseo.com/youtube-300-hours/] . Today, viewers are required to endure a painful and time-consuming process to search and discover interesting videos. This can sometimes include watching up to 30 seconds of pre-roll advertisements before being able to view the video and scrub for relevant content. A better method for previewing videos is needed so viewers can quickly find what they are looking for – and skip over what they aren't.

Vilynx [http://www.vilynx.com/] has developed a way for mobile and PC viewers to watch an automatically curated 5 second preview of a video's most interesting scenes with just a mouse over or a finger swipe. Viewers can quickly preview a video before deciding to watch it. It's a similar idea to watching a movie trailer, but thanks to machine learning, the preview can be easily applied to all videos. Publishers also benefit from this technology, as the preview can generate higher click through rates and longer engagement times. The same technology works for both social media and branded websites. More video views equal more branding opportunities for publishers and advertisers – and ultimately, more revenue.

The need to automatically extract the most interesting clips from each video in a time-sensitive manner requires heavy-duty computing capabilities. Vilynx has been working closely with Intel to improve the performance and efficiency of machine learning and deep learning algorithms to enable these automated video searches.

Why use CPUs and Cloud-Scale Infrastructure?

Deep learning has shown great promise in many practical applications, ranging from speech recognition and visual object recognition to text processing. This has been accompanied by an increase in the training sequence size and/or the parameter set size to greatly improve classification accuracy. While these advances are exciting, their use is currently limited to a small number of companies and research groups due to the high cost of the hardware infrastructure required to support them.

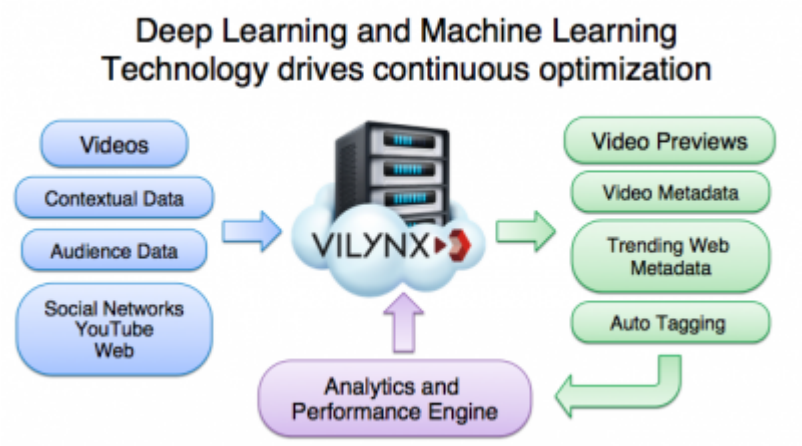
The use of high performance (and high cost) GPUs was supposed to facilitate the training of modestly sized deep networks. However, a known limitation of the GPU approach is that the

to-GPU transfers are not a significant bottleneck. While data and parameter reduction works well for small problems, they offer a poor response to problems with a large amount of data and dimensions, such as analysis of high-resolution images.

For startup organizations and the developer community, training deep networks quickly without access to dedicated hardware readily available to industry leaders like Facebook*, Baidu*, Google* and others is a challenging task. Moreover, building a grid of servers with very powerful GPUs and CPUs is expensive and only accessible to those companies with deep pockets, a challenge that is solved through the use of cloud infrastructure.

Increasing Computational Costs

Few workloads are more compute-intensive than video processing today. Adding deep learning into the mix, the level of complexity and computation is increased to a point beyond what can be achieved using a GPU alone. Within the Vilynx stack, video processing and machine learning are used to select the relevant moments from hours of videos and store them in a long-term memory. Once stored, deep learning algorithms use audience preferences to select and display video clips. Finally, semi-supervised machine learning algorithms are fed with matching keywords, metadata, social networks and web data to obtain the most relevant set of key words for a specific video.



without a high-cost, dedicated hardware solution.

This solution is not only able to learn about user preferences and display content automatically selected as most relevant, but it also enables video discovery and search via a rich set of keywords that are matched to internal moments. In short, for the first time, we are enabling automation of full search functionality inside a video.

Solution Stack

The below graphic shows the various components and hierarchy of this solution, from canonical AWS services such as Amazon Elastic Compute Cloud* (EC2) and Amazon Elastic Block Store* (EBS), to Intel-optimized software libraries and open source components all the way to the end application deployed at scale by video publishers.

Steps to Setup and Run Yourself:

1. Download and launch the AWS CloudFormation* template [https://s3.amazonaws.com/caffecfncluster/1.0/intelcaffe_cfncluster.template]
2. Connect to your Amazon EC2 instance (In our case C4.8xlarge, powered by the Intel® Xeon® Processor E5 v3 Family)

More detailed instructions can be found here [https://software.intel.com/en-us/articles/distributed-training-of-deep-networks-on-amazon-web-services-aws] .

Very special thanks to Thomas 'Elvis' Jones at AWS for his support and collaboration. Cheers!

[1] [#_ftnref1] Source: Cisco 2016 Visual Networking Index [http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html]

[2] [#_ftnref2] Source: AlexNet - NIPS2012 [http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf]

August 15, 2016 [https://itpeernetwork.intel.com/solving-video-discoverability-using-cloud-scale-deep-learning/]

Joe Spisak Machine Learning Amazon Web Services, Vilynx, Xeon Processor

ABOUT JOE SPISAK

Director of Strategy and Business Development for Machine Learning at Intel

View all posts by Joe Spisak →