How will Deep Learning Change Internet Video Delivery?

Hyunho Yeo
KAIST

Sunghyun Do
KAIST

Dongsu Han
KAIST

1 INTRODUCTION

Internet video has experienced tremendous growth over the last few decades and is still growing at a rapid pace. Internet video now accounts for 73% of Internet traffic and is expected to quadruple in the next five years [9, 41]. Augmented reality and virtual reality streaming, projected to increase twenty-fold in five years [9], will also accelerate this trend.

From content delivery networks (CDNs) [35] to HTTP adaptive streaming [3, 21, 37] and data-driven optimization for quality of experience [22], the networking community has brought fundamental advancements in Internet video delivery. However, video delivery still leaves large room for improvement. First, the video delivery infrastructure has largely been agnostic to the video content it delivers, treating it as a stream of bits. Second, the basis of how we represent video has remained as an unexplored topic within the networking community. In fact, the fundamental basis of video encoding has largely remain the same. In particular, the practice of video encoding is to use signal processing techniques (e.g., discrete cosine transform and inter-frame prediction) to spacial and temporal redundancies that occur at short time-scales (e.g., within a frame or a group of pictures).

This paper shows that advancement in deep neural networks present new opportunities that can fundamentally change Internet video delivery. In particular, deep neural networks allow content delivery network to easily capture the content of video and thus enable content-aware video delivery. Based on the observation, we explore new design space for content-aware video delivery networks.

First, video contains large amounts of redundancy that occur at large timescales. For example, a basketball game video shares similar background throughout the video. Moreover, series of games, episodes, and streams often share common features. For example, streams of the same game from Twitch share large amounts of redundancy. While binge-watching—practice of watching a series of shows in one sitting—is common [34, 40], tradition video delivery does not take advantage of redundancy that occur at large timescales. As a result, when the network is congested, video quality degrades drastically, even though similar footage is being played. To tackle the problem, we design a content-aware solution that leverages redundancy across videos. In particular, our design leverages image super-resolution using deep neural networks and use client computation to enhance the video quality. The content-aware delivery network classifies videos and generates a small super-resolution network (~ 7.8 MB) using images from similar videos. We show that the content-aware video delivery achieves better quality using the same amount of bandwidth. We believe this has far-reaching implications on dynamic adaptive streaming and quality of experience optimization.

Second, a video frame contains many objects that show up frequently throughout the video, but traditional streaming cannot capture this because it cannot capture common features from these objects. Deep neural networks provide an alternative to encode object representations. To demonstrate this, we leverage Generative Adversarial Networks (GANs), known to synthesize images that look authentic to human [17], for synthesizing objects within a video. We use GAN trained using similar videos to synthesize a high-quality video from an alternative form that contains much less information.

To demonstrate the feasibility of the approach, we prototype the system and quantify benefits and costs of the approach. We articulate how different parts of the video delivery infrastructure should change to accommodate the design. In summary, this paper takes a first attempt to answer the following question: *how will advances in deep neural networks change Internet video delivery?* In answering the question, we find deep learning opens up large design space and has far-reaching implications in the video delivery ecosystem. Finally, we call upon the networking community to embrace recent advances in deep learning and to rethink Internet video delivery in the context of what the new technology enables.

2 MOTIVATION AND INTUITION

Limitations of conventional adaptive streaming: The traditional approach to improving video stream quality includes designing new bitrate selection algorithms [3, 19, 21], choosing better servers and CDN [2, 26, 48], and utilizing a centralized control plane [27, 33]. These approaches focus on how we could fully utilize the given network resources. However, there are two significant limitations.

First, recent devices including mobile devices have significant computational power. Market reports [36] show around

HotNets-XVI, November 30–December 1, 2017, Palo Alto, CA, USA

ACM ISBN 978-1-4503-5569-8/17/11…$15.00
https://doi.org/10.1145/3152434.3152440

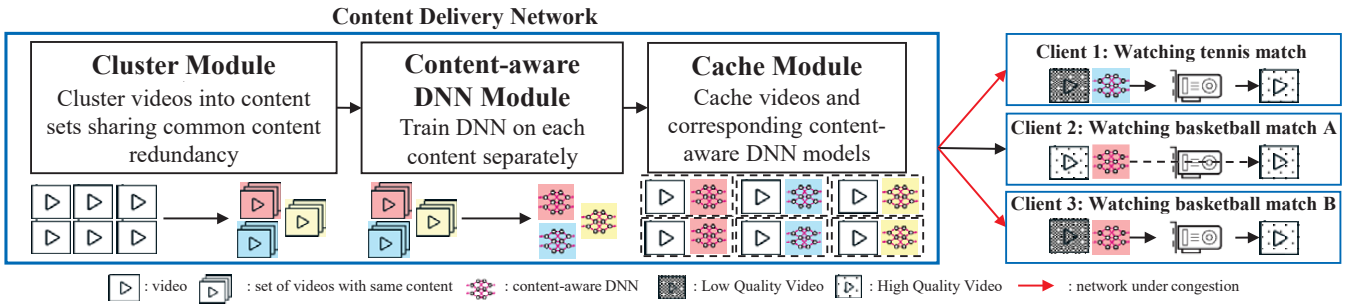


Figure 1: High-level vision of a DNN-Based Content-Aware Content Distribution Network

50% of users watch video on PCs, which have large computation power. Mobile devices that account for the rest are also equipped with power-efficient mobile graphic processing units (GPUs) "whose performance exceed that of older-generation game consoles" (e.g., XBox 360) [14]. The popularity of mobile games and the advent of emerging media, such as virtual and augmented reality, will accelerated this trend. This leaves a great opportunity for trading off computation for reduced bandwidth under network congestion. However, the current video delivery infrastructure does not offer any way to utilize client's computational power. Thus, when the network is congested, the stream quality suffers directly. With client's growing computational capacity and ever increasing demand for bandwidth, we envision a video delivery system in which clients take an active role in improving the video quality using their own computational power.

Second, video contains large amount of redundancy that occur at large timescales, and its high-level features contain valuable information that can be leveraged for video coding. For example, meaningful objects recognized by human, such as sport player, stadium and score board, reappear frequently. However, standard video coding, such as MPEG and H.26x, only captures two kinds of redundancy and lacks any mechanisms to leverage motion picture's semantics. Spatial redundancy exploits pixel-level similarity within a picture [47]. The intra-frame coding compresses a picture using discrete cosine transform (DCT), quantization, and entropy encoding [18]. Temporal redundancy represents similarities between successive frames. Inter-frame coding encodes the difference between adjacent frames to compresses a motion picture [15].

Imagine a popular sports game (e.g., NBA finals) watched by millions of people. Same objects, such as balls and players, and scenery, such as the court and background, show up repeatedly quarter after quarter with a small variation. Similarly, such redundancy is also found within episodes of each TV show, games in sport leagues, videos from the same YouTube or Twitch streamers. Leveraging such redundancy is an unexplored opportunity. However, capturing this using pixel-level processing on successive frames is very difficult if not impossible. We envision a video delivery network that exploits redundancy by capturing semantically meaningful objects rather than encoding video purely at the pixel level.

Key approach: Fortunately, deep neural networks provide a mechanism to abstract meaningful features from images. DNN is a computational model with multiple layers of hierarchy, each of which processes the input in a non-linear fashion and delivers its output to the upper layer. It is designed to learn high-level abstract features from a complex low level representation of data [5]. However, developing and utilizing a generic model that works well across all videos is too expensive for practical purposes, given the amount and diversity of Internet video—the size of DNN generally has a positive correlation with its expressive power. In addition, capturing all objects in a single network amounts to devising a DNN-based generic video compression algorithm, which is non-trivial. Even the quality of state-of-the-art DNN-based image compression is only as good as JPEG2000 only when the compression ratio is set very high [44].

Instead, this paper takes a content-aware approach. The content distribution network clusters videos of similar nature and generates DNN models for each cluster. The model contains abstract representations of video by capturing high-level features rather than a pixel-level encoding. In the next two sections, we explore a concrete design realizing the vision with various examples of DNN models.

3 DNN-BASED CONTENT-AWARE CDN

Leveraging DNN and utilizing content redundancy are two core components of our design that mark a drastic takeoff from traditional video delivery. Figure 1 presents a high-level architecture of a DNN-based content-aware video delivery that realize our vision.

Server-side: When new content is registered to providers, such as Netflix, Twitch and YouTube, the clustering module classifies the video into groups of similar videos. The video may already have metadata to aid clustering; e.g., 2017 NBA finals game 2. Many platforms (e.g., Twitch and YouTube) also provide the categories of videos being streamed (e.g., a Starcraft game). The video clustering can also be done utilizing DNNs by extending image classification [4, 46, 49]. If the new content shares large redundancy with an existing cluster, we find the nearest neighbor and categorize the content into the cluster. If not, the video belongs to a cluster of its own.

For each cluster of a significant size, the CDN creates a DNN model which is an abstract representation of the cluster. It then associates each video stream in the cluster with the abstract representation by recording it on the video meata-data (i.e., a video manifest file). According to what the DNN

model captures, the CDN manipulates the original video to create an alternative version that is much smaller in size. The alternative version is then recorded in the manifest file. Thus, the manifest file lists abstract representations and the corresponding version of video. Note that there can be multiple alternate abstract representations and video versions.

Client-side: Clients receive manifest files that contain multiple alternative ways to stream the video. The manifest files prescribe how each model should be applied to the corresponding representation of the video. Clients choose abstract representations they understand considering their computational power. Then, they apply the abstract representation to their video chunks to enhance the quality of the video. Note that the framework accommodates a variety of DNN models and is agnostic to the video encoding format. In the following section, we show three examples of such models and quantify their costs and benefits to demonstrate the viability.

4 CASE STUDIES

We present three different examples of DNN models that can be used to improve video quality. Each of the models improves an orthogonal aspect of user QoE. We take existing DNNs, but reinterpret their power in the context of Internet video streaming. We highlight the benefit of content-aware video delivery and discuss its impact.

4.1 Content-Aware Super-resolution

Image and video super-resolution [12, 42] can recover a high resolution image from a low resolution media. Motivated by this, we explore the possibility of using content-aware super-resolution for video delivery. We show leveraging content-aware super resolution provides an alternative to adaptive streaming and delivers enhanced and more stable quality. Finally, we quantify its network and computational overhead.

DNN model: We use a very deep convolution network called VDSR [23] for image super-resolution. The model consists of 20 convolution layers with 64 filters, and its network footprint is about 7.8 MB uncompressed. To create a content-aware model, we obtain video episodes from a series. We use four datasets: a basketball game from the 2012 London Olympics available on YouTube [6]; 100 m and 200 m men's final from the 2012 London Olympics on YouTube [7, 8]; multiple plays of a computer game (Starcraft); and the Conan monologue episodes from the official YouTube channel of late night host Conan O'Brien [10, 11]. For basketball, we trained using images from the first half and use the second half as the test set. For others, we use one video in the series for training and another for testing.

To compare the approach with a content-agnostic one, we use the same DNN model but train it on a benchmark dataset widely used for super-resolution [23–25, 43]. Finally, we use, as the baseline of comparison, bicubic, a frequently used up-scaling heuristics.

Quality: Figure 2 presents the average video quality in Peak Signal-to-Noise Ratio (PSNR) of content-aware DNN (awDNN), content-agnostic DNN (agDNN), and bicubic. The resolution of the original video is 720p, and we show the resulting quality of super-resolution from lower resolutions (x-axis). The PSNR value is calculated against the original video. Content-aware DNN delivers much better quality. As shown in Figure 2(a), content-aware super-resolution (e.g., 43.36 dB with 360p to 720p up-scaling) shows much better quality than bicubic (34.28 dB) and content-agnostic super resolution (36.71 dB).

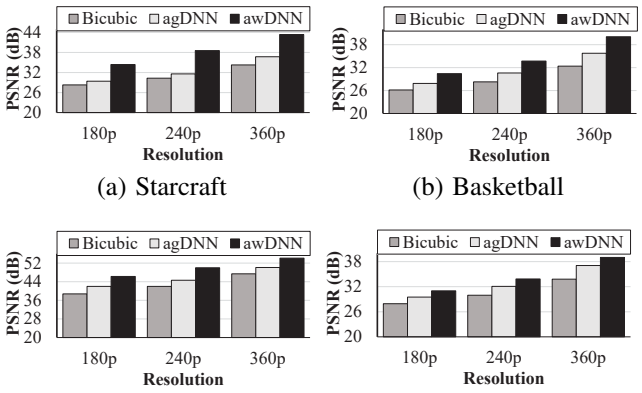


Figure 2: Video quality comparison on various resolution utilizing DNN super resolution

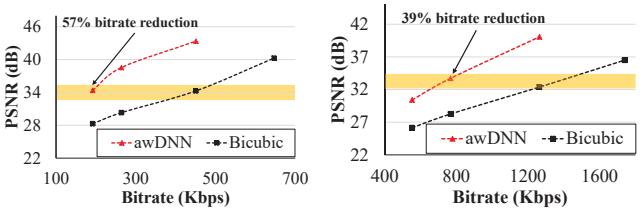
Figure 3 shows part of the images to fit the page, including the original 720p image and up-scaled images from 180p. The content-aware approach shows visible differences. Redundant objects, such as the Olympic symbol, the "Command Center" building, and even text from the game image, are only successfully recovered in the content-aware approach.

Bitrate vs. quality: Our approach allows us to deliver same quality of videos using less bandwidth. Figure 4 plots the bitrate and quality of video delivered using content-aware super-resolution. As shown in Figure 4(a), a 192 Kbps video with content-aware super-resolution shows better quality than a 451 Kbps video with bicubic, delivering more than 57% reduction in network bandwidth. Note, the bandwidth shown excludes the overhead of the DNN-model (7.8 MB). The overhead depends on the video length, because this overhead gets amortized over the duration of video. This demonstrates the benefit of using client computation and applying content-aware processing.

Bandwidth and computation overhead: We quantify the additional network bandwidth and computation required at the client. DNN neural network model is 7.8 MB in size. Figure 5(a) and 5(b) show the data usage for delivering the same video quality (~34 dB and ~33 dB respectively). The cost of transferring the DNN model gets amortized by 242 seconds in Figure 5(a) and 126 seconds in Figure 5(b). The time difference comes from the fact that model size is same (7.8MB), but 5(b) (-221 Kbps) reduces more bitrate over 5(a) (-196 Kbps) compared to baseline approach, bicubic. The result suggests that viewers who watch similar content for an extended period [34, 40] can benefit significantly.

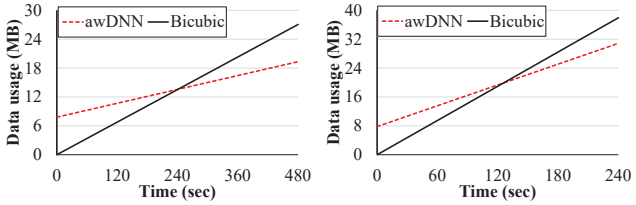
(a) Original (b) Content-aware DNN (c) Congent-agnostic DNN (d) Bicubic

Figure 3: Super resolution result images recovered from 180p to 720p (Video Source: [6, 10, 11])



(a) Starcraft (b) Basketball

Figure 4: Video quality versus bitrate



(a) Starcraft (b) Basketball

Figure 5: Overall data usage over time

We also quantify the initialization time and the frame-rate of the super-resolution network using a desktop GPU (NVIDIA Titan Xp). The initialization time of DNN is very fast; it takes 0.031 seconds on average to load all parameters on memory. Its average feed forward time (or the time to recover a single high resolution image) is 0.127 seconds, resulting in 7.87 frames per second to recover a 720p resolution image. The larger the target image size, the longer it takes. This indicates real-time super-resolution is a challenge.

However, we believe, performance will not be a significant barrier in the long run for the following reasons: First, DNNs often allow trading run-time for quality. For example, Table 1 shows different parameter settings (# layers and # filters per convolution layer) for VDSR and the speed-quality tradeoff. It shows VDSR can achieve 31 frames per second while delivering better quality (31.68 dB) than bicubic (28.28 dB). Second, one can rely on opportunistic super-resolution and client scheduling. Clients keep a reasonable amount of buffer. Similar to how the rate adaptation algorithm considers the buffer size and available bandwidth, clients can take account for the time to perform super-resolution and apply super-resolution on buffered content whenever possible. Third, advances in deep neural networks will reduce the overhead. Super-resolution networks smaller than the ones we use have shown to work at real time [13]. Furthermore, recent work [51] achieves real-time performance without modifying the super resolution network of prior work [12] through propagating a super-resolution result of a frame over multiple successive frames. Finally, computational power of GPU has been growing at a tremendous rate [31]. This will increase room for our approach.

Resolution	(# Layer, # Filter)	Frames per second	Quality
180p	(20, 64)	7.87	34.40 dB
	(15, 48)	13.88	33.41 dB
	(10, 32)	31.05	31.68 dB
240p	(20, 64)	7.87	38.55 dB
	(15, 48)	13.88	37.19 dB
	(10, 32)	31.05	35.13 dB
360p	(20, 64)	7.87	43.36 dB
	(15, 48)	13.88	41.62 dB
	(10, 32)	31.05	41.20 dB

Table 1: Run-time and accuracy trade-off inside VDSR (Content: Starcraft)

Implications on QoE optimization: So far, we have explored the feasibility of leveraging client computation in video delivery and shown the benefit of content-aware processing. This allows us to trade client computation for video quality, which is great for improving QoE. However, it complicates the QoE optimization problem because it opens up a new axis. Here, we discuss the implications.

Traditional QoE optimization assumes bandwidth is the only resource constraint and optimizes the resource use leveraging a global view [22]. Now, client computation is an additional critical factor affecting video quality. At the same time, it also serve as a resource constraint and is heterogeneous, similar to bandwidth. Moreover, when the client-side processing cannot be done at real time, it has interactions with buffer size, which is a function of bandwidth availability and rate adaptation. Many factors interact with each other, making the problem even more challenging. In-depth research is required to explore its full potential.

4.2 Content-Aware Video Generation

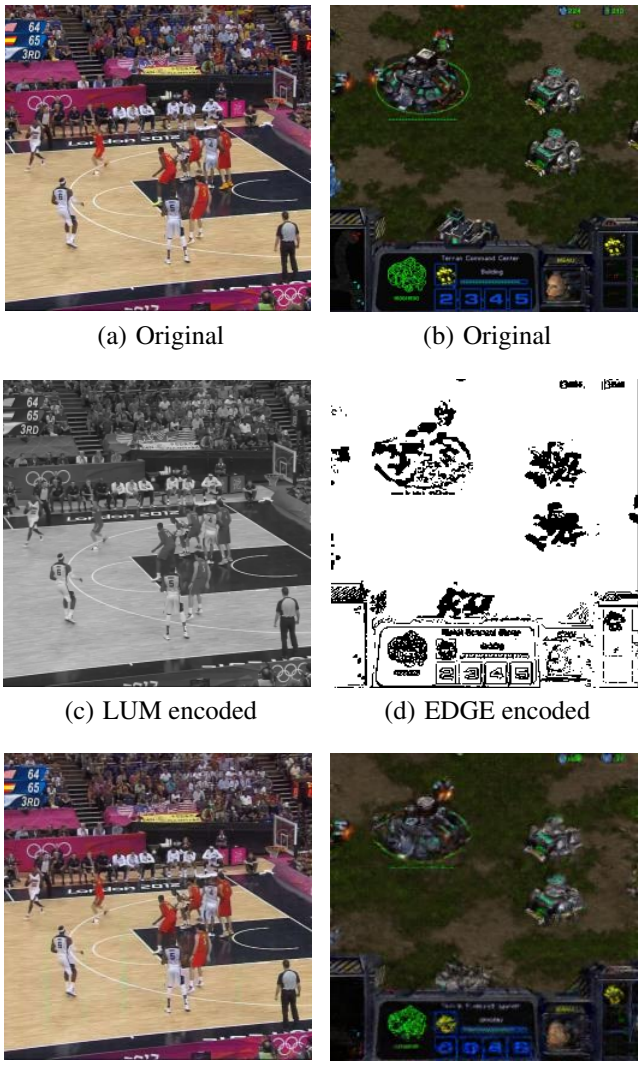
Generative adversarial networks (GANs) can synthesize images that are indistinguishable from real images [17] given a simple description of the image. Motivated by this, we explore generating a high-quality video from a compact alternative representation of video that contain less redundancy. As proof-of-concept, we explore two alternative video representations, LUM and EDGE. LUM removes chroma in YCbCr color space and only contains luminance (Y) from the original video. For EDGE, we extract the boundaries of objects by applying an edge detection algorithm for each frame and produce a black-and-white image using 1-bit quantization. Figure 6(c) and 6(d) respectively show example frames for LUM and EDGE applied on Basketball and Starcraft video. Note they contain much less information than the original.

We use GAN developed by Isola et al. [20] that performs image-to-image translation. It uses 16 layers for the generator and 6 for its discriminator, and the network footprint is 654 MB. For training, we pre-process videos in our dataset to produce LUM and EDGE representations and train the network to generate the original from them. For example, for LUM, the network synthesizes the original form (including chroma) from luminance values. For a preliminary evaluation, we use a low resolution image (256x256) because the network only takes 256x256 input and outputs the same size.

Image size and quality: We compare the image size of LUM and EDGE representations with JPEG images of similar quality. Table 2 shows the resulting image size excluding the GAN model. LUM (20.33 KB) delivers similar quality image using data less than 11% compared to JPEG (22.84 KB). Figure 6(e) shows a recovered image of LUM. The generated color well matches with the original. It suggest that chroma is one of the long-term redundancy that can be captured using a DNN model, and the content-aware approach can exploit this. EDGE (3.65 KB) also uses less data for delivering a similar quality image than JPEG (9.29 KB). Figure 6(f) shows a decoded image of EDGE. The generated color well matches with the original except some distortion happens on the object's outline. It indicates a black and white image composed of edges has enough information for a DNN model to restore an original image when long-term redundancy exists on a video. EDGE and LUM shows a trade-off between the details of representation and the quality of decoded images.

We also quantify the initialization time and the frame rate of the image generation network using NVIDIA Titan Xp. The initialization takes 2.08 seconds. The average feed forward time is 0.068 second, resulting 14.7 frames per second.

Challenges in video generation: We re-encode the LUM and EDGE images into video using H.264 and compare its size against a video of similar quality. The result in Table 2, however, shows that LUM and EDGE representations are not compressed well using H.264. While our image-based preliminary study suggests the approach is potentially promising, extending the result to video and making it practical require solving many challenges. First, finding new types of



(a) Original (b) Original

(c) LUM encoded (d) EDGE encoded

(e) LUM decoded (f) EDGE decoded

Figure 6: Sample 256x256 images of new representations EDGE and LUM (Video source: [6])

representation that efficiently utilize common features from video is challenging. Leveraging luminance and edges are just two naive ways, but capturing features is more complex, considering an object can have variations in its appearance throughout a cluster of videos. Second, even though alternative representations contain much less information, they often do not compress well using standard video codecs. For example, H.264 encoding of LUM results in 404 Kbps, 11% smaller than the original. But compared to a video of same quality, it is 80% larger, even excluding the model size. Further research is required to design efficient algorithms for compressing alternative representations. Finally, real-time processing must be enabled for the approach to be used for streaming. This is especially challenging because real-time processing and supporting large resolutions are at odds. When the challenges are properly addressed, we envision this can be used for bandwidth-efficient video upload, download, and storage for large videos that share similarity.

Method	PSNR	Image Size	Video Size
LUM	29.67 dB	20.33 KB	2445 KB
JPEG/H.264	29.98 dB	22.84 KB	1351 KB
EDGE	21.78 dB	3.65 KB	500 KB
JPEG/H.264	21.35 dB	9.29 KB	74 KB

Table 2: Efficiency of new representations

4.3 Content-Aware Frame Interpolation

Many studies [28, 30, 39] suggest deep learning is also effective on predicting future frames that look natural to human eyes. While they are limited in the number of frames they can predict [30], their predictive power offers a new way for frame interpolation that can be used to enhance the video frame rate, which is another key metric for user QoE [52]. This offers yet another way of trading computation for quality. In addition, traditional signal processing based frame interpolation often shows visible artifacts and make movies look unnatural [32]. We believe content-awareness is a promising way to overcome this defect, because content redundancy implies small varieties of objects and movements are repeated frequently.

5 NEW RESEARCH DIRECTIONS

This paper only scratches the surface of how deep neural networks can be used to enhance video delivery. Fully realizing its potential requires solving many challenges. We discuss several new research directions worth exploring.

Why should the networking community care? Internet video delivery is inherently an interdisciplinary research that requires expertise and understanding in networking, distributed systems, human behavior, data science, and signal processing. We have witnessed that a disruptive technology in one domain (e.g., large-scale data analytics [50]) often bears fruit to innovations in the entire ecosystem (e.g., data-driven optimization [22]). The deep learning community mainly focuses on developing new DNNs, but does not explore issues integrating them into the existing video delivery infrastructure.

Interaction with adaptive streaming: DNN-based content-aware streaming provides a mechanism to trade client computation for bandwidth in video delivery. Thus, this has interactions with adaptive streaming. How one effectively uses the mechanisms in combination is a new direction to explore. For example, the super-resolution can be used opportunistically with rate adaptation to mask the effect of bandwidth fluctuations. Frame interpolation may better cope with bandwidth fluctuations that occur at short timescales than the current adaptation. In addition, bitrate adaptation using deep reinforcement learning [29] would provide a way of integrating the computing resource in an end-to-end fashion.

Heterogeneous clients: DNN-based content-aware streaming enables clients (or edge resources) to take an active role. Because it uses clients' computing resources and video often requires real-time playback, one has to consider the computational power at the clients' side that are heterogeneous in nature. Thus, similar to how we adapt video quality to screen size, the video delivery system may have to adapt to client's computational power and power availability. The challenge lies in accommodating widely varying resources and dynamically adapting to the power availability.

QoE optimization: DNN-based content-aware streaming provides multiple ways to enhance metrics, such as frame rate, resolution, and image quality, which impact QoE. However, the community is still at the early stage in understanding how these metrics impact QoE. In addition, traditional quality metrics, such as PSNR, do not reflect human perception. For example, in motion pictures, objects that occur frequently often bear more meaning (e.g., players and balls in a ball game as oppose to individual spectators). DNN-based enhancements are likely to be more effective. Metrics that reflect human perception and real-world measurements on human behavior are needed for advancement in video delivery.

Video representation, metadata, and indexing: System design issues arise in implementing content-awareness. HTTP adaptive streaming changed how clients interact with the CDNs. For example, CDNs distribute manifest files and clients download video data in chunks. Similarly, DNN model now becomes part of metadata and clients also take more active roles. Also, allowing a more fine-grained adaption may mean that the CDN may have to generate data dynamically from its internal representation. How to store and distribute video and metadata, how the CDN manages DNN models, and how the CDN and client work together to support fine-grained dynamic adaptation are new avenues for research.

Application to emerging media: Bandwidth is a major bottleneck for virtual reality (VR) and augmented reality (AR) streaming [16, 38, 45]. To avoid motion sickness, it must stream images from multiple cameras at a very high resolution (e.g., 4k to 8k) with up to 36 bits per pixel at 60 to 120 frames per second [45]. It is estimated that a 720p VR game will require 50 Mbps (and a 4K game 500 Mbps) [1]. This is because current video streaming can only send video signals, not their semantics. We believe that a fundamental paradigm shift is required. We envision a DNN model that creates intermediate representations of objects, which are more compact. The intermediate representation is transferred over the network, and clients then render images from the representation. We believe the our DNN-based content-aware video delivery is a first step towards this.

6 CONCLUSION

In this paper, we argue that client's increasing computation power and advancement in deep neural networks allow us to take advantage of long-term redundancy found in videos, which leads to quality improvement at lower bandwidth cost for Internet video delivery. Our goal is to highlight the potentials of the approach and identify new research directions it opens up. Through case studies, we explore preliminary design examples of a deep neural network based content-aware video delivery and demonstrate their benefits. We hope this paper offers insights into improving the quality of experience in the era of Internet video and emerging media.

ACKNOWLEDGMENTS

This work was supported by Institute for Informatio & communications Technology Promotion (IITP) grant funded by the Korea government (No.2015-0-00164)

REFERENCES

[1] 2016. ARRIS Gives Us a Hint of the Bandwidth Requirements for VR. http://www.onlinereporter.com/2016/06/17/arris-gives-us-hint-bandwidth-requirements-vr/. (June 2016).

[2] V. K. Adhikari, Y. Guo, F. Hao, V. Hilt, Z. L. Zhang, M. Varvello, and M. Steiner. 2015. Measurement Study of Netflix, Hulu, and a Tale of Three CDNs. *IEEE/ACM Transactions on Networking* 23, 6 (Dec 2015), 1984–1997. https://doi.org/10.1109/TNET.2014.2354262

[3] Saamer Akhshabi, Ali C Begen, and Constantine Dovrolis. 2011. An experimental evaluation of rate-adaptation algorithms in adaptive streaming over HTTP. In *Proc. ACM conference on Multimedia systems*. ACM, 157–168.

[4] Sean Bell and Kavita Bala. 2015. Learning visual similarity for product design with convolutional neural networks. *ACM Transactions on Graphics* 34, 4 (2015), 98.

[5] Yoshua Bengio et al. 2009. Learning deep architectures for AI. *Foundations and trends in Machine Learning* 2, 1 (2009), 1–127.

[6] Olympic Channel. 2012. Basketball - USA vs Spain - Men's Gold Final | London 2012 Olympic Games. https://www.youtube.com/watch?v=l9wUr-CK1Y4&t=8394s/. (Aug. 2012).

[7] Olympic Channel. 2012. Usain Bolt Wins 200m Final | London 2012 Olympic Games. https://www.youtube.com/watch?v=LWZQAVtkMBo. (Aug. 2012).

[8] Olympic Channel. 2012. Usain Bolt Wins Olympic 100m Gold. https://www.youtube.com/watch?v=2O7K-8G2nwU. (Aug. 2012).

[9] Cisco. 2017. Cisco Visual Networking Index: Forecast and Methodology, 2016âĂŞ2021. http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/complete-white-paper-c11-481360.pdf. (July 2017).

[10] Team Coco. 2012. Monologue 02/15/12 - CONAN on TBS. https://www.youtube.com/watch?v=DSzCNC4ypiA&list=PLXi05HpOWLE6eLdY$_E$pYooY$_o$NtO4DZ-w&index=3. (Feb. 2012).

[11] Team Coco. 2012. Monologue 03/19/12 - CONAN on TBS. https://www.youtube.com/watch?v=rULPBI9O8d4&index=4&list=PLXi05HpOWLE6eLdY$_E$pYooY$_o$NtO4DZ-w. (Feb. 2012).

[12] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2014. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision*. Springer, 184–199.

[13] Chao Dong, Chen Change Loy, and Xiaoou Tang. 2016. Accelerating the super-resolution convolutional neural network. In *European Conference on Computer Vision*. Springer, 391–407.

[14] K. Fatahalian. 2016. The Rise of Mobile Visual Computing Systems. *IEEE Pervasive Computing* 15, 2 (Apr 2016), 8–13. https://doi.org/10.1109/MPRV.2016.35

[15] F. H. P. Fitzek and M. Reisslein. 2001. MPEG-4 and H.263 video traces for network performance evaluation. *IEEE Network* 15, 6 (Nov 2001), 40–54. https://doi.org/10.1109/65.967596

[16] MATTIAS Fridström. 2016. The bandwidth problem: 5 issues the VR industry must resolve. https://venturebeat.com/2017/05/06/the-bandwidth-problem-5-issues-the-vr-industry-must-resolve/. (May 2016).

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proc. NIPS*. 2672–2680.

[18] Fred Halsall. 2000. *Multimedia Communications* (1st ed.). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[19] Te-Yuan Huang, Nikhil Handigol, Brandon Heller, Nick McKeown, and Ramesh Johari. 2012. Confused, Timid, and Unstable: Picking a Video Streaming Rate is Hard. In *Proc. IMC*. 225–238.

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2016. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004* (2016).

[21] Junchen Jiang, Vyas Sekar, and Hui Zhang. 2012. Improving Fairness, Efficiency, and Stability in HTTP-based Adaptive Video Streaming with FESTIVE. In *Proc. ACM CoNEXT*.

[22] Junchen Jiang, Shijie Sun, Vyas Sekar, and Hui Zhang. 2017. Pytheas: Enabling Data-Driven Quality of Experience Optimization Using Group-Based Exploration-Exploitation.. In *Proc. USENIX NSDI*.

[23] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proc. IEEE CVPR*. 1646–1654.

[24] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Deeply-recursive convolutional network for image super-resolution. In *Proc. IEEE CVPR*. 1637–1645.

[25] Wei-Sheng Lai, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. 2017. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. *arXiv preprint arXiv:1704.03915* (2017).

[26] Hongqiang Harry Liu, Ye Wang, Yang Richard Yang, Hao Wang, and Chen Tian. 2012. Optimizing cost and performance for content multihoming. In *Proc. ACM SIGCOMM*. 371–382.

[27] Xi Liu, Florin Dobrian, Henry Milner, Junchen Jiang, Vyas Sekar, Ion Stoica, and Hui Zhang. 2012. A case for a coordinated internet video control plane. In *Proc. ACM SIGCOMM*. 359–370.

[28] William Lotter, Gabriel Kreiman, and David Cox. 2016. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv preprint arXiv:1605.08104* (2016).

[29] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *Proc. ACM SIGCOMM*. ACM, New York, NY, USA, 197–210. https://doi.org/10.1145/3098822.3098843

[30] Michael Mathieu, Camille Couprie, and Yann LeCun. 2015. Deep multi-scale video prediction beyond mean square error. *arXiv preprint arXiv:1511.05440* (2015).

[31] Khalid Moammer. 2015. Nvidia : Pascal Is 10x Faster Than Maxwell, Launching in 2016 On 16nm - Features 3D Memory, NV-Link and Mixed Precision. http://wccftech.com/nvidia-pascal-gpu-gtc-2015/. (March 2015).

[32] Tim Moynihan. 2014. WTF Just Happened: My New HDTV Makes Movies Look Unnaturally Smooth. https://www.wired.com/2014/08/wtf-just-happened-soap-opera-effect/. (Aug. 2014).

[33] Matthew K. Mukerjee, David Naylor, Junchen Jiang, Dongsu Han, Srinivasan Seshan, and Hui Zhang. 2015. Practical, Real-time Centralized Control for CDN-based Live Video Delivery. In *Proc. ACM SIGCOMM*. 311–324.

[34] Netflix. [n. d.]. Netflix & Binge: New Binge Scale Reveals TV Series We Devour and Those We Savor. https://media.netflix.com/en/press-releases/netflix-binge-new-binge-scale-reveals-tv-series-we-devour-and-those-we-savor-1. ([n. d.]). Last accessed: July 2017.

[35] Erik Nygren, Ramesh K. Sitaraman, and Jennifer Sun. 2010. The Akamai Network: A Platform for High-performance Internet Applications. *SIGOPS Oper. Syst. Rev.* 44, 3 (Aug. 2010), 2–19.

[36] Ooyala. [n. d.]. Ooyala GLOBAL VIDEO INDEX Q1 2017. http://go.ooyala.com/rs/447-EQK-225/images/Ooyala-Global-Video-Index-Q1-2017.pdf. ([n. d.]). Last accessed: July 2017.

[37] Roger Pantos and William May. 2017. *HTTP Live Streaming*. Internet-Draft draft-pantos-http-live-streaming-23. IETF Secretariat. http://www.ietf.org/internet-drafts/draft-pantos-http-live-streaming-23.txt

[38] Adrian Pennington. 2017. The realities of making VR and AR streaming a reality. https://knect365.com/media-networks/article/98d3564c-2e6b-43f7-8728-dda3038f818a/the-realities-of-making-vr-and-ar-streaming-a-reality. (July 2017).

[39] MarcAurelio Ranzato, Arthur Szlam, Joan Bruna, Michael Mathieu, Ronan Collobert, and Sumit Chopra. 2014. Video (language) modeling: a baseline for generative models of natural videos. *arXiv preprint*

arXiv:1412.6604 (2014).

[40] Reuters. [n. d.]. How Network TV Figured Out Binge-Watching. http://fortune.com/2016/03/11/netflix-changing-game-network-tv/. ([n. d.]). Last accessed: July 2017.

[41] Sandvine. [n. d.]. 2016 Global Internet Phenomena Report: North America and Latin America. https://www.sandvine.com/downloads/general/global-internet-phenomena/2016/global-internet-phenomena-report-latin-america-and-north-america.pdf. ([n. d.]). Last accessed: July 2017.

[42] Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[43] Ying Tai, Jian Yang, and Xiaoming Liu. 2017. Image Super-Resolution via Deep Recursive Residual Network. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[44] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. 2017. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395* (2017).

[45] Valley Voices. 2017. Why The Internet Pipes Will Burst When Virtual Reality Takes Off. https://www.forbes.com/sites/valleyvoices/2016/02/09/why-the-internet-pipes-will-burst-if-virtual-reality-takes-off/#7049e4583858. (July 2017).

[46] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1386–1393.

[47] John Watkinson. 2004. *The MPEG Handbook: MPEG-1, MPEG-2, MPEG-4*. Taylor & Francis.

[48] Patrick Wendell, Joe Wenjie Jiang, Michael J. Freedman, and Jennifer Rexford. 2010. DONAR: Decentralized Server Selection for Cloud Services. In *Proc. ACM SIGCOMM*.

[49] Sergey Zagoruyko and Nikos Komodakis. 2015. Learning to compare image patches via convolutional neural networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4353–4361.

[50] Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J Franklin, Scott Shenker, and Ion Stoica. 2012. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proc. USENIX NSDI*.

[51] Zhang, Zhengdong and Sze, Vivienne. 2017. FAST: A Framework to Accelerate Super-Resolution Processing on Compressed Videos. In *CVPR Workshop on New Trends in Image Restoration and Enhancement*.

[52] T. Zinner, O. Hohlfeld, O. Abboud, and T. Hossfeld. 2010. Impact of frame rate and resolution on objective QoE metrics. In *Proc. International Workshop on Quality of Multimedia Experience (QoMEX)*. 29–34. https://doi.org/10.1109/QOMEX.2010.5518277