How will Deep Learning Change Internet Video Delivery?

Hyunho Yeo
Ph.D. Candidate@KAIST

- Work on DL-based video delivery
- Publish papers at top-tier system/network conferences

Neural Adaptive Content-aware Internet Video Delivery
Hyunho Yeo, Youngmok Jung, Jaehong Kim, Jinwoo Shin, Dongsu Han
USENIX OSDI 2018

How will Deep Learning Change Internet Video Delivery?
Hyunho Yeo, Sunghyun Do, Dongsu Han
ACM HotNets 2017
Overview

“How will Deep Learning Change Internet Video Delivery?”

1. Observation/Limitation of Current Video Delivery

2. Recent research: DL-based adaptive streaming [OSDI 18]

3. Vision of DL-based Video Delivery
Era of Internet Video Delivery

Internet video traffic has *exponentially* grown over last decade!

![Graph showing exponential growth of internet video traffic](image)

1: CISCO Visual Networking Index, 16 data was interpolated

To handle bandwidth heterogeneity over space and time, Adaptive streaming has been widely deployed.
Traditional Approaches

- **Optimizing ABR algorithms**
  Pensieve [SIGCOMM 17], MPC [SIGCOMM 15]

- **Choosing better servers, CDNs**
  Content Multihoming [SIGCOMM 12], VDN [SIGCOMM 15]

- **Leveraging centralized control plan**
  Video Control Plane [SIGCOMM 12], Pythease [NSDI 17]

Goal: Find how to best utilize the network resource
Limitation of Current Video Delivery

Video quality heavily depends on available bandwidth
Client computing power is scarcely utilized other than for decoding.
Observation on Current Video Ecosystem

Standard codecs efficiently reduce redundancy *only* inside GOP

Group of Pictures (GOP): Intra-frame coding

Video Quality

Seamless switching

Video

Standard codec (H.26x, VPx, AV1)

Compressed

I-frame P, B-frames I-frame

[Adaptive streaming]

: Intra-frame coding

: Inter-frame coding

Time

2—10 seconds
Limitation of Current Video Delivery

Standard codecs lack any mechanisms for exploiting redundancy that occurs at large timescales.
What Deep Neural Network (DNN) Can Do?

1. Utilize **client computation** to enhance video quality
What Deep Neural Network (DNN) Can Do?

1. Utilize **client computation** to enhance video quality

![Diagram showing video server, client computing device, and network congestion, with low resolution video being enhanced to high resolution using a super-resolution DNN.]
What Deep Neural Network (DNN) Can Do?

2. Trained and operate in **large timescales** (video)
What Deep Learning (DL) can Do?

Can we overcome the current limitations via DNN?

How much quality improvement can we achieve?

To answer these, let’s move to our recent research, NAS [OSDI18]

Neural Adaptive Content-aware Internet Video Delivery

Hyunho Yeo   Youngmok Jung   Jaehong Kim   Jinwoo Shin   Dongsu Han

*KAIST*
NAS: DL-based Adaptive Streaming

Apply super-resolution DNN on top of bitrate adaptation
1. Content: Video on demand (VOD)

Example

![YouTube, Netflix, Hulu, Amazon logos]

2. Computing device: NVIDIA GTX 10 series

Example

<table>
<thead>
<tr>
<th>GTX 1050 Ti (Entry-level)</th>
<th>Titan Xp (High-end)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Price</td>
<td></td>
</tr>
<tr>
<td>$139</td>
<td>$1,200</td>
</tr>
</tbody>
</table>
NAS utilizes DNN and client computation, but ...
NAS: Two Initial Challenges

⚠️ NAS utilizes **DNN** and client computation, but ...

1. DNN testing accuracy is **unreliable** for unseen/new content
   • Even worse, degradation can occur (below figure)

Unseen content  EDSR [CVPRW 17]  Quality
(Trained on DIV2K dataset)  
SSIM = 0.86  SSIM = 0.84

For the real-world deployment, DNN accuracy should be **guaranteed**
NAS: Two Initial Challenges

⚠ NAS utilizes DNN and **client computation**, but ...

2. Client must process DNN at real-time, but computing power **varies** across space and time

Client A: *Entry-level* GPU
(1.98 TFLOPS – 1050 Ti)

Client B: *High-end* GPU
(10.79 TFLOPS – Titan Xp)

Adaptation to computing power is required
Key Design 1: Content-aware DNN

Challenge: Providing reliable DNN quality

1. New video admission
2. Generates a content-aware DNN
3. Provide (video, DNN)

Video server

Content-aware DNN 1

Content-aware DNN 2

Video 1

Video 2

Content-aware DNN delivers the reliable (over-fitted) training accuracy instead of the unpredictable testing accuracy.
Training a content-aware super-resolution

1. Prepares training data

Raw high-resolution (1080p)
Compressed low-resolutions (240p—720p)

2. Updates the DNN parameters

Updates parameters
Input ➔ DNN ➔ Output ➔ Target
Content-agnostic vs. Content-aware

“PSNR 2~4 dB gain over content-agnostic”
Bicubic vs. Content-aware DNN

Average PSNR: 28.28dB

Average PSNR: 34.40dB
Bicubic vs. Content-aware DNN

Average PSNR: 26.15dB

Average PSNR: 30.42dB
Key Design 2: Multiple Quality DNNs

Challenge: Enabling real-time super-resolution on heterogeneous clients

1. Provides multiple quality DNN options

Video server

Downloads several MBs?

→ Delay video streaming

Client

- Quality: Low → High
- Size: Small (93KB) → Large (2,143KB)
- Compute: Low → High
Key Design 2: Multiple Quality DNNs

Challenge: Enabling real-time super-resolution on heterogeneous clients

1. Provides multiple quality DNN options
   - Quality: Low
     - Size: Small (93KB)
     - Compute: Low
   - High
     - Size: Large (2,143KB)
     - Compute: High

2. Delivers DNN description
   - (#Layer, #Channel)

Video server

Manifest file

Client

MPD (Media Presentation Description)

<table>
<thead>
<tr>
<th>Period</th>
<th>Adaptation Set (Video)</th>
<th>Adaptation Set (DNN) (#layer, #filter)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1080p 4.8Mb/s</td>
<td>Low-240p (20, 9)</td>
</tr>
<tr>
<td></td>
<td>720p 2.4Mb/s</td>
<td>Med.-240p (20, 21)</td>
</tr>
<tr>
<td></td>
<td>480p 1.2Mb/s</td>
<td>High-240p (20, 32)</td>
</tr>
</tbody>
</table>
Key Design 2: Multiple Quality DNNs

Challenge: Enabling real-time super-resolution on heterogeneous clients

1. Provides multiple quality DNN options
   - Quality: Low
     - Size: Small (93KB)
     - Compute: Low
   - High
     - Size: Large (2,143KB)
     - Compute: High

2. Delivers DNN description
   (#Layer, #Channel)

3. Test-runs and selects the highest-quality running at real-time
   - Computing device (GTX 1080)
   - Mock DNNs
     - 53 fps
     - 38 fps
   - Selected
     - 52 fps
     - 21 fps
NAS: Two Additional Challenges

⚠️ NAS streams video with a content-aware DNN, but ...
NAS: Two Additional Challenges

NAS streams video with a content-aware DNN, but ...

1. Takes long time to download and utilize a DNN

Video server

Ultra-high (2,145KB)

1 x

21 seconds

360p video (400Kbps)

Client

: DNN data

: Video data

Need to provide incremental benefit during downloading a DNN
NAS: Two Additional Challenges

NAS streams video with a content-aware DNN, but ...

2. A DNN competes bandwidth with video

- Aggressive download: rebuffering, low video quality
- Conservative download: low DNN benefit

Need to *carefully* decide how/when to download a DNN model
Key Design 3: Scalable DNN

Challenge: Takes a long time to utilize a DNN

1. Implement a scalable DNN (+ divide into similar-size chunks)
2. Download/Apply a partial DNN
NAS DNN Architecture

1. **Per-resolution DNN**: enable real-time processing

   - MDSR [1]
   - # Channel
   - Input: 240p
   - 240p
   - 1080p

   - NAS-MDSR (⋯: Bicubic interpolation)
   - 270p
   - 360p
   - 360p
   - 1080p
   - 1080p
   - 540p
   - 480p
   - 1080p

2. **Additional bypassing paths**: enable anytime prediction

   - Residual paths
   - 480p image
   - Bicubic Conv
   - Conv
   - ReLU
   - Conv
   - Sum
   - 1080p image
   - Conv
   - Sum
   - Subpixel
   - Conv

   - By-passing paths

Key Design 4: Integrated ABR

Challenge: How to decide when to download a DNN

💡 ABR already handles unpredictable bandwidth variations

→ Integrate DNN download decisions with existing **RL-based ABR** (Pensieve [1])

[2]: Upper right figure is from the slide of “Neural adaptive video streaming with pensieve.”,
Key Design 4: Integrated ABR

Challenge: How to decide when to download a DNN

- **Integrate** DNN download decisions with existing **RL-based** ABR (Pensieve) [1]

QoE metric = bitrate - rebuffering – smoothness

Goal: Maximize the total QoE over an entire video

Key Design 4: Integrated ABR

Challenge: How to decide when to download a DNN

- **Integrate** DNN download decisions with existing **RL-based** ABR (Pensieve) [1]

QoE metric = DNN (bitrate) - rebuffering - smoothness

Reward $r_t$

Goal: Maximize the total QoE reflecting DNN-based quality enhancement
Putting All Together: Implementation

Server → Video → DNN Processor → NAS Player (dash.js)

DNN Processor: 6.3K LOC

NAS Player (dash.js):
Δ1.7K LOC (8.8%)
Integrated ABR
5.5K LOC
Evaluation

1) How much benefit does NAS deliver?

2) What are the cost and benefit of NAS?

3) Does NAS effectively adapt to heterogeneous clients?
NAS vs. Existing Video Delivery : QoE

• **17.8 hours real-world network traces**: collected from 3G network and broadband (average bandwidth: 1.31Mbps)
• **27 YouTube videos**: 5-24 minutes, encoded at {400, 800, 1200, 2400, 4800}kbps
• **Computing device**: NVIDIA Titan Xp, **DNN quality**: Ultra-high
• **Video player**: Chromium browser, **Video server**: Apache server
Quantify user experience of video streaming

Generalized QoE model\textsuperscript{1,2,3}:
\[ q(R_n) - \mu(T_n) - (q(R_n) - q(R_{n-1})) + (\text{Quality}) - (\text{rebuffering}) - (\text{smoothness}) \]

- \( q(R_n) \): Perceptual quality of \( n \)\textsuperscript{th} video chunk bitrate \( R_n \)
- \( T_n \): Rebuffering time for downloading \( n \)\textsuperscript{th} video chunk

1: MPC-SIGCOMM15, Pensieve-SIGCOMM17, Oboe-SIGCOMM18
NAS vs. Existing Video Delivery: QoE

- **17.8 hours real-world network traces**: collected from 3G network and broadband (average bandwidth: 1.31Mbps)
- **27 YouTube videos**: 5-24 minutes, encoded at {400, 800, 1200, 2400, 4800}kbps
- **Computing device**: NVIDIA Titan Xp, **DNN quality**: Ultra-high
- **Video player**: Chromium browser, **Video server**: Apache server

NAS improves user QoE by 43.08% compared to Pensieve and 92.28% compared to BOLA using same amount of bandwidth.
When the total viewing reaches 30 hours (per minute of video), NAS CDN recoups the initial training cost.
### Heterogeneous Clients

Each GPU processes at real-time (> 30fps for all resolutions)

<table>
<thead>
<tr>
<th>DNN Quality</th>
<th>GPU Model (Price)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>GTX 1050 Ti ($139)</td>
</tr>
<tr>
<td>Medium</td>
<td>GTX 1060 ($249)</td>
</tr>
<tr>
<td>High</td>
<td>GTX 1070 Ti ($449)</td>
</tr>
<tr>
<td></td>
<td>GTX 1080 ($559)</td>
</tr>
<tr>
<td>Ultra-high</td>
<td>GTX 1080 Ti ($669)</td>
</tr>
<tr>
<td></td>
<td>Titan Xp ($1,200)</td>
</tr>
</tbody>
</table>

NAS adapts to heterogeneous devices, and a device with higher computing power receives greater benefit.
• NAS shows that applying DNN on video content utilizing client computation can significantly enhance user QoE.

• NAS accommodates four key designs: Content-aware DNN, Multiple quality DNNs, Scalable DNN, Integrated ABR.
NAS = Adaptive streaming + VoD contents + Desktop-class GPUs

What’s Next?

• Integrate DL with **various parts** in video delivery infrastructure
• Apply DL on **diverse video applications** (e.g., Live/4K/AR/VR)
• Deploy DL-based streaming on **commercial mobile devices**
Conclusion

“How will Deep Learning Change Internet Video Delivery?”

• The advance of deep learning presents unseen opportunities

- Bandwidth
- Client computation
- Content redundancy

Current streaming

DL-based streaming

Better experience

• Rethinking the video delivery infrastructure is required to take advantage of the new opportunities

Neural Adaptive Content-aware Internet Video Delivery

Hyunho Yeo  Youngmok Jung  Jaehong Kim  Jinwoo Shin  Dongsu Han

: First step toward this direction
Thank you

• Personal homepage
  http://ina.kaist.ac.kr/~hyunho/

• Lab homepage
  http://ina.kaist.ac.kr/

• Project homepage
  http://ina.kaist.ac.kr/~nas/

OSDI conference @ Carlsbad, CA, USA
Content-agnostic vs. Content-aware

- Original
- agDNN
- awDNN

PSNR (dB)

Content type

1 2 3 4 5 6 7 8 9
QoE breakdown

Bitrate utility | Rebuffer penalty | Smoothness penalty

BOLA | R-MPC | Pensieve | NAS

0.05 | 0.16 | 0.02 | 0.07

0.05 | 0.15

0.05 | 0.10
Average QoE over Content Types

Normalized average QoE

Content type

1 2 3 4 5 6 7 8 9

BOLa  R-MPC  Pensieve  NAS
Scalable DNN

- CDF of Average QoE for NAS, NAS FULL, and Pensieve.
- 17.54% improvement for NAS compared to Pensieve.
Case Study: Timeline

Play time = 20 sec
24 sec, 1st DNN
32 sec, 2nd DNN
44 sec, 3rd DNN
52 sec, 4th DNN
60 sec, full DNN

Mock DNN test
Apply 1st DNN chunk to 6th video chunk
Download 1st video chunk
Download 2nd DNN chunk
Download manifest file

DNNs are fully downloaded

Video chunk
DNN chunk