How will Deep Learning Change Internet Video Delivery

HotNets-XVI Dialogue

Mohammad Alizadeh and Sujata Banerjee

SB: Machine learning and AI techniques are all the rage currently. It is natural to explore the possibility of solving some hard networking problems using AI/ML. Reducing the bandwidth requirements of video delivery systems by leveraging advances in deep neural networks to detect redundancies in videos seemed like a neat idea and I am excited that we will be discussing this paper at HotNets. At the same time, there is so much work on video delivery systems with so many techniques to reduce the complexity and overhead. Your work on Pensieve for example, considered the use of reinforcement learning for bit rate selection in video streaming. Do you think this paper picked the right components of video delivery systems to apply ML techniques to?

MA: That's a good question. There's indeed been a lot of recent work on optimizing video quality of experience using ML techniques. But the existing work has focused on using data-driven techniques to optimize the traditional video delivery pipeline, say for things like CDN server selection, bitrate adaptation, and so forth. What I found exciting about this paper is that it's arguing for an entirely new space of potential optimizations that exploit the video *content* to improve video delivery. The idea that, maybe by learning the characteristics of the video itself, we can make smarter decisions about what to send over the network is pretty cool.

SB: Isn't that what existing video coding and compression techniques already do to some extent? There has been prior work on leveraging the encoding structure to preferentially treat some parts of the video bits (the I-frame for example) to improve quality under challenging loss or delay conditions. I agree that using content-aware DNN models takes it further in this paper, where there is also the aspect of clustering similar videos.

MA: Yes, I must say that I was a bit surprised by the super-resolution results. The images produced by the content-aware DNN in Figure 3 look quite impressive for a 4-to-1 upscaling. Content-aware video *generation* with Generative Adversarial Networks (GANs) seems a little scary though. What is to guarantee that the GAN doesn't just hallucinate arbitrary details into video? When I'm watching a video, say of a basketball game, I generally expect it to be real. Models like GANs can't really guarantee to match reality; they just make it hard to tell if the output is real or fake. So there doesn't seem to be anything to prevent the model from, say, changing the color of player jerseys or replacing one player's number with another!

SB: I agree! Super-resolution or upscaling seems like a very worthwhile goal. However, if the video is being generated from an imperfect model, or say, down the road, from a model that has been loaded maliciously, there are some serious security-related issues to consider.

Let me ask a different question: should we trust DNNs whose workings may be hard to explain, sometimes even for ML experts? For example, the VDSR approach, which I don't know much about, has been validated in the computer vision community and works for the sample videos in the paper. Would it simply work with any content? Do we know enough to detect when it stops working and as you said before, creates "hallucinations"?

MA: These are important questions for sure with no clear answers yet. Another important question is at what granularity do we need to train these content-aware models. For example, do we need a separate

model for every basketball game? What about for different camera angles from the same game? If we need to train a new model for every new video and download it to the client, this approach becomes less practically appealing than if clients can cache a single model and use it for many similar videos.

SB: The other elephant in the room is of course the computational power required at the client for the architecture presented in this paper to work. Here I found the results mixed. The paper shows that even with a good desktop GPU, the super-resolution network could only run at 7.8 frames per second at 720p. It seems we're going to need more efficient models and techniques if we are to apply these ideas to higher resolution 4K videos. While computation capabilities are clearly increasing at the edge, we are not there yet.

MA: Yes, training time could also be a challenge. It seems unlikely that in the near term, it will be possible to train a new model in real-time for live videos but it might work for stored content. Again, a lot seems to hinge on how reusable these DNN models are.

SB: In her Athena lecture at SIGCOMM this year, Jen Rexford mentioned how programming languages and networking researchers have collaborated over the last several years to produce some significant results. It would be great to have ML and networking researchers work together on important interdisciplinary problems.

MA: Absolutely. As networking researchers, one of our challenges will be to keep up with the rapidly evolving techniques and results in the ML community. How do we assess ML-centric papers in our community? How do we make sure that these papers represent the state of the art? And how do we get more ML/AI experts involved in our community? These are all good questions to discuss at HotNets.