LEARNING VIDEO REPRESENTATION FOR REDUCING NETWORK CONGESTION

Wei-Hung Weng, Schrasing Tong, Yu-An Chung
{ckbjimmy,st9,andyyuan}@mit.edu

1 INTRODUCTION

Efficient Internet video delivery has become increasingly important since people are more desperate for real-time video broadcast. Studies have shown that 73% of Internet traffic comes from Internet video and may increase more in the following years (Visual, 2016). An important issue of Internet video streaming is how to maximize the utilization of current video delivery network infrastructure to effectively send the video content without losing the stream quality too much. The bottleneck of the current infrastructure is the network bandwidth since the computational capacities of both host and client ends are growing rapidly due to the improvement of computer hardware. To minimize the effect of network bottleneck, learning a good video representation is an essential strategy to reduce the network congestion due to less efficient video delivery.

Learning meaningful low-dimensional representation helps compress the information size that needs to be transmitted through network. Using data, representation learning has driven advances in many domains including computer vision, natural language processing (NLP), automatic speech recognition, even in healthcare, to deliver various machine learning-based solutions. Representation learning has an ability to abstract data while keeping meaningful hidden information inside data (Bengio et al., 2013). However, this approach is relatively unexplored in the field of computer networking.

In this study, we propose an alternative approach to learn video representation that can directly learn the video representation from raw video. Recently, Yeo et al. utilized the power of deep neural network to learn complex low level image and video representation for adaptive streaming (Yeo et al., 2017). The authors took content-aware approach, developed content delivery network (CDN) to cluster videos into few clusters and generated representation learning models for each of them at the server-side, sent the video along with the cluster representation to the client, and used the content-aware neural networks to acquire the higher quality video at the client-site. They also applied the model to the video super-resolution and video generation tasks. However, using the CDN approach, we need to cluster videos into the appropriate cluster, which may increase the uncertainty of learning a good video representation, since the representation is learned from the cluster-based neural network after assigning the cluster. i.e. we may choose a completely inappropriate neural network for representation learning due to the incorrect cluster assignment. Additionally, in the real-world it is hard to know natural clusters of all the videos. Therefore, learning representation directly from video is needed.

2 PROPOSED APPROACH

We propose to adopt the convolutional neural network (CNN) with sequence to sequence (seq2seq) learning for video representation learning. seq2seq learning has been widely applied in the NLP and speech domain, such as machine translation or for free-from question answering problems, due to its nature of word sequence (Sutskever et al., 2014). It is a structured learning algorithm that is able to train models to convert sequences from one domain to sequences in another domain, e.g. English to French translation. Thus, seq2seq learning can learn the representation of a sentence by sequential input of words. For example, it can be used for next word prediction by learning the semantic representation of the sentence.

A video is a sequence of images. Therefore, it is also possible to utilize the characteristic of seq2seq learning to learn the representation of video by splitting a video into a sequence of image frame. The learning framework is shown as 1. After splitting a video into images, we first used a pretrained CNN autoencoder to compress the image into low-dimensional vector. Then we pass the CNN-

derived vectors into seq2seq encoder-decoder architecture (Chung et al., 2018), which is based on the variant of recurrent neural network (RNN) structure. The RNN encoder (left side) will learn a low-dimensional vector embedding which can represent the sequence of images. The RNN decoder (right side) will generate the output predicting the next frame sequentially. The encoder-decoder model will be optimized based on the loss between outputs and original images.

Once the model is trained, the encoder can be placed at the server-site, and the decoder can be placed at the client-site. Only the video representation (the blue node) will be transmitted through network. The output of decoder will be recovered to the image by using the decoder part of previous trained CNN autoencoder. Eventually, all the outputs can be pasted together back to the video.

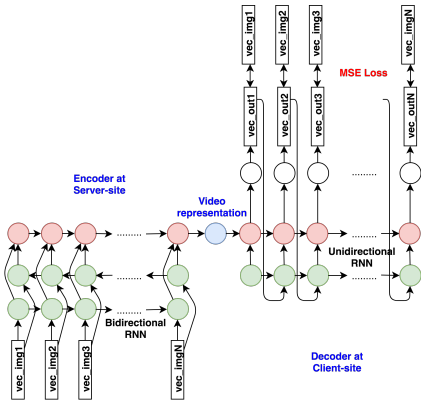

Figure 1: CNN autoencoder-Seq2seq architecture for learning video representation. The vec_img cells are the low-dimensional vectors, which represent the corresponding image, are generated by the CNN autoencoder.

3 Potential Challenges, Timeline and Resource

The potential challenges in this proposal may be (1) how to identify the ideal seq2seq architecture for sequential images representation, (2) how to reconstruct images well, and (3) how to evaluate the result in a quantitative way. Previous study used PSNR for evaluation (Yeo et al., 2017). It is, however, not really useful for deciding the quality of video reconstruction since the quality of video is usually related to other issues rather than noise.

We plan to obtain the training data before October 20, develop the neural network before November 3 and run the first experiment before November 10. We will do fast iteration to improve our model before December 1 by the performance of model. Potential solutions include adding more training data, and increasing/decreasing the complexity of model through regularization.

The resource required in this study is mainly GPU for modeling, which we have already requested from labs.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Yu-An Chung, Wei-Hung Weng, Schrasing Tong, and James Glass. Unsupervised cross-modal alignment of speech and text embedding spaces. In *NIPS*, 2018.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

Cisco Visual. Forecast and methodology, 2016-2021, white paper. *San Jose, CA, USA*, 1, 2016.

Hyunho Yeo, Sunghyun Do, and Dongsu Han. How will deep learning change internet video delivery? In *ACM Workshop on Hot Topics in Networks*, 2017.