Bandwidth Estimation for Video and Audio Transfer using A2C

Haipeng Zhang  
 Peking University  
 Beijing China  
 1800012831@pku.edu.cn

Shenhan Zhu  
Peking University  
 Beijing China  
1800012794@pku.edu.cn

ABSTRACT

In this work, the bandwidth estimation task can be formulated as a Reinforcement Learning problem. We employ A2C to implement an estimator for the challenge. This estimator is trained with the simulated network environment gym. By using Actor-Critic structure, we can let the model choose a suitable bandwidth for video/audio transfer based on various network conditions.

∗Article Title Footnote needs to be captured as Title Note

†Author Footnote to be captured as Author Note

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

*ACM MMSys’21, June, 2021, Istanbul, Turkey*

© 2021 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06...$15.00

https://doi.org/10.1145/1234567890

KEYWORDS

network performance, bandwidth estimation, reinforcement learning

1 Introduction

In this Bandwidth Estimation Challenge, we seek to deal with the problem of choosing an appropriate bandwidth for video/audio real time transfer in various network conditions. To improve the quality of experience (QoE) of users, we need to build an estimator to evaluate the network condition and choose a bandwidth to transfer our video/audio in each time slot. This estimator should take packet loss, network delay, network congestion and all the other relevant factors into account, and improve the quality of the received video/audio as much as possible.

The bandwidth estimation task can be simplified as a turn-based game, in which the agent should choose the best action (which is the bandwidth of video/audio transfer) to receive the largest reward (which is referred to as that low packet loss rate, low network delay and all the other factors that would improve the quality of the received video/audio) from the environment. In each turn, the agent should choose one action based on the current environment state, and interact with the network environment, then adjust its policy according to the reward it receives and the change in the environment state. This leads us to a machine learning (ML) based technique named Reinforcement Learning (RL). RL aims to resolve the tasks in which an agent needs to interact with the environment and learn to achieve specific goals. In most of the tasks, these goals can be measured by well-designed rewards. These rewards can be fed into RL networks and help to train an agent that can do good choice in action to interact with the environment. In this sense, we formulate this bandwidth estimation task as a RL problem.

In this task, I choose to use Advantage Actor Critic (A2C), an improved version of Actor-Critic based on RL, as the agent’s model, trained with the provided environment “gym”. Gym is an environment that can generate a state as the network condition in each step, and once the agent give a bandwidth as the video/audio input to gym, it will respond with simulated receiving rate, delay, and loss ratio, then generate the network condition in the next step. We can compute the agent’s reward based on the gym’s reaction, which enables us to train a A2C model.

In an Actor-Critic framework, we need to build an “Actor” to choose actions based on each state, and an “Critic” to evaluate Actor’s actions. Both of them can be implemented with deep neural networks. During training, Actor will improve its performance by achieving better evaluation given by the Critic, and the Critic will try to give more accurate evaluations by approximating the state’s value (which can be computed by expected rewards). Meanwhile, A2C improves the basic Actor-Critic framework by using a better function, named Advantage function, to train the Critic network. By using A2C, we have implemented an estimator for bandwidth estimation, and apply it into the provided test environment.

2 Method

In RL, usually researchers train the agent to evaluate the value of each action based on current state and choose the best one (which is called the value-based method). However, in this task, we have a continuous action space, which means there are infinite actions. Hence, instead of training the agent to evaluate different actions, we have to train the agents’ policy directly, which is called the policy-based method. In this challenge, we use Advantage Actor Critic (A2C), one of the policy-based algorithms, to train an estimator.

2.1 Model structure

In A2C, there is an Actor to choose actions, and a Critic to evaluate the action, as shown in Figure 1. Since we aim to deal with the continuous action space, the Actor should not output the value for each action and choose the best one. Instead, the Actor will generate a distribution for the action space, and in each step we sample an action according to the generated distribution. Our objective is to train the Actor to generate a distribution which assigns more possibility to better actions, and in order to achive this, we have to train the Critic to evaluate the next state after the action is applied.

When the agent interacts with the environment, the environment state will be fed into both Actor and Critic. The actor will output a distribution of actions, and then sample an action from this distribution. The Critic will output the value of this state. And the sampled action will be fed into the environment to get rewards and update the environment state.

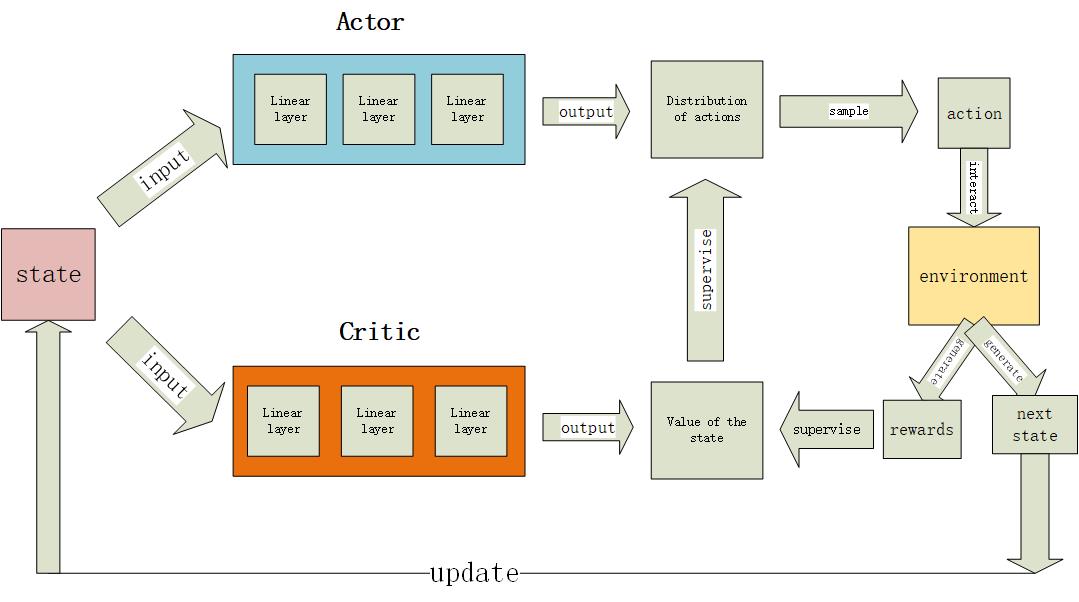


Figure 1: the pipeline for A2C interacting with the environment.

2.2 Training of the model

In A2C, we need to train both Actor and Critic to let them choose a good action corporately.

For Actor, if we represent Actor’s policy by , where represents parameters，in basic policy gradient algorithm we know to train this policy we should upgrade in each step with formula

,

where is the learning rate.

In A2C，we have

(1)

where is the state in step t, is the action taken by the agent in step t, and represents the value of action taken in state , represents the value of state . In this formula, function can be computed by the reward received by the agent, and function should be given by the Critic.

One factor differs A2C from Actor-Critic algorithm is the function , which is called “Advantage function”. It reflects how much is action better than average. Compared with the pure Actor-Critic, by using advantage function, A2C draws a baseline for the actions and hence can have both negative and positive feedbacks when updating.

For Critic, we compute its loss by

where we approximate the real value of state by , where is a parameter, and value is computed by Critic.

Our goal is to train the Critic’ output to approximate the value of every state. Therefore, by training Critic with the loss function above, we can decrease the distance between the Critic’s output and the real value of every state.

3 Results

In each training epoch, the average reward received by the agent is shown in Figure 2.

As we can see, the best average reward the agent can get is about 0.37, and after the 22nd epoch, the average reward starts to decrease, which might be the result of overfitting. So eventually the parameters in 22nd epoch is chosen.

This model along with the trained parameters has been tested by the given test AlphaRTC environment, and the output video & audio is correctly generated.

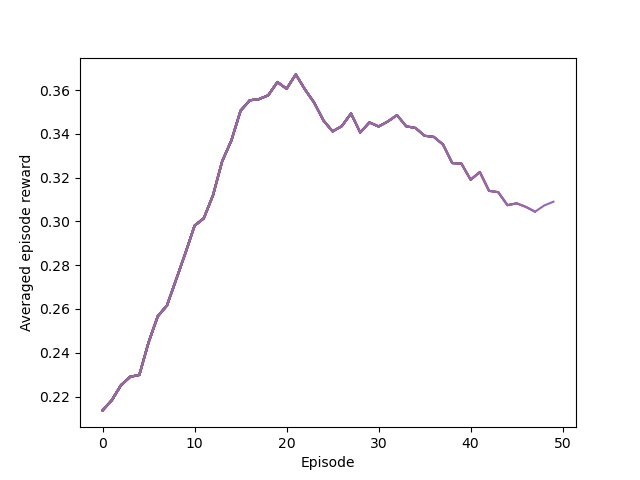


Figure 2: the average reward received by the agent in each training epoch.

4 Conclusion

This bandwidth estimation problem can be formulated as a standard Reinforcement Learning task. By using the RL framework, we can train a model who learns to choose the best bandwidth to interact with the simulated environment, aiming at getting the largest reward. Meanwhile, this model can be used as the estimator directly. Since the action space is not discrete, value-based methods are not suitable for this challenge. Instead, A2C, an RL algorithm who combines both value-based methods and policy-based methods, can be applied to this task. We can train an A2C model using the given simulated network environment “gym”. Further extension of this implementation lies in future works.

REFERENCES

[1] Barto, AG, *Neuron-like Adaptive Elements that can Solve Difficult Learning Control-Problems* (1984). Behavioural Processes. 361.Conference Name:ACM Woodstock conference

Conference Short Name:WOODSTOCK’18

Conference Location:El Paso, Texas USA

ISBN:978-1-4503-0000-0/18/06

Year:2018

Date:June

Copyright Year:2018

Copyright Statement:rightsretained

DOI:10.1145/1234567890

RRH: F. Surname et al.

Price:$15.00