# Paper Summary
# Neural Adaptive Video Streaming with Pensieve

Sarthak | 2020CS10379

June 2024

## Problem Statement

The paper addresses the challenge of generating fine-grained Adaptive Bitrate (ABR) algorithms for video streaming, focusing on enhancing user Quality of Experience (QoE).

## Motivation

- Conflicting video QoE requirements (high bitrate, minimal rebuffering, smoothness etc.) and cascading effects of bitrate decisions
- Current ABR algorithms use fixed control rules based on simplified or inaccurate models of the deployment environment leading to coarse-grained decisions

## Key Idea

Pensieve uses modern Reinforcement Learning techniques to learn a control policy for bitrate adaptation purely through experience. It knows nothing about the task in hand at the start and then gradually learns to make better ABR decisions through reinforcement, in the form of reward signals that reflect video QoE for past decisions. It represents its control policy as a neural network that maps observations (e.g., throughput samples, playback buffer occupancy, video chunk sizes) to the bitrate decision for the next chunk.

## Framework

Uses the Asynchronous Advantage Actor-Critic (A3C) method which involves training two neural networks.

### Policy

Upon receiving the state input $s_t$ , Pensieve's RL agent takes an action at that corresponds to the bitrate for the next video chunk. It actions based on a policy, defined as a probability distribution over actions $\pi : \pi(s_t, a_t) \longrightarrow [0, 1]$. $\pi(s_t, a_t)$ is the probability that action $a_t$ is taken in state $s_t$.

In practice, there are intractably many {state, action} pairs. To overcome this, Pensieve uses a neural network (layered 1D-CNNs) to represent the policy with a manageable number of adjustable parameters, $\theta$.

### Policy Gradient Training

- Estimate the gradient of the expected total reward by observing the trajectories of executions obtained by following the policy
- Requires the advantage $A^{\pi_\theta}(s_t, a_t)$ (difference in expected reward upon deterministical picking vs following policy $\pi_\theta$) to calculate this estimate
- To compute the advantage for a given experience, we need an estimate of the value function, $v^{\pi_\theta}(s)$, the expected total reward starting at state s and following the policy $\pi_theta$
- The role of the critic network is to learn an estimate of $v^{\pi_\theta}(s)$ from empirically observed rewards using a similar (layered 1D-CNN) Neural Network

Pensieve uses parallel training methodology, where each agent sends their experiences to a central agent who then updates the model and sends it to parallel agents.

## Contributions

- Uses RL to generate ABR algorithms, learning from the performance of past decisions
- Development of a simulation environment that accurately models video streaming dynamics, allowing efficient training of the RL agent
- Proposals for modifications that enable a single model to handle multiple videos with different encoding characteristics
- Adapts to different network conditions and QoE objectives, Outperforms existing ABR algorithms in a variety of scenarios

## Strengths

- Outperforms the best state-of-the-art scheme, with improvements in average QoE of 12%–25%

- Generalizes well, outperforming existing schemes even on networks for which it was not explicitly trained
- Does not rely on pre-programmed models or assumptions about the environment
- Do not need hand-crafted features, can be applied directly to raw observation signals

## Weaknesses

- The implementation and training of RL models are complex and require significant computational resources
- Analysis focuses primarily on specific QoE metrics; broader considerations, such as power consumption and fairness across users, are not addressed
- While the paper demonstrates effectiveness in simulated environments, real-world performance may vary due to differences in network conditions and user behavior