Project Proposal: Approximating Pensieve with Classical Techniques

Thomas Schibler

Friday, April 18, 2025

Motivation

Pensieve [4] is a performant black-box model for choosing the optimal bit rate during video streaming. However, its decision-making process is relatively opaque. This project aims to better understand how Pensieve makes decisions by approximating its behavior using interpretable models. The goal is to determine whether Pensieve is performing complex reasoning beyond what standard, white-box models can do, and if so, where and why.

Problem Specification

The exact question is: to what degree can Pensieve be imitated using existing white-box models or hybrid combinations of such models? Specifically, the aim is to reproduce the output of Pensieve as closely as possible on a known dataset, using interpretable models or combinations of said models through hyperparameter tuning.

Existing Approaches

Explaining black box models through surrogate models is not new. For example, Trustee [3] is a model-agnostic approach that has been used to interpret and evaluate black-box behavior, including Pensieve. This project differs by directly leveraging domain-specific white-box models rather than taking a purely agnostic stance, with the help of Pensieve's output in hindsight.

Datasets

The primary dataset will be the one used by the authors of Pensieve, a broadband dataset from the FCC [1]. If needed, it is feasible to generate additional data by running Pensieve on other raw datasets.

Novelty

Instead of applying model-agnostic explanation tools, this project will exploit the existence of white-box models already designed for the domain; see, e.g. [2, 5, 6]. The project also includes a novel hybridization step, combining multiple models and tuning their parameters specifically to overfit to Pensieve's decisions.

Implementation Plan

- Reimplement (or ideally reuse) existing aforementioned white-box models.
- Optionally replicate Pensieve for comparison and data generation.
- Combine white-box models into a hybrid system (e.g., weighted ensemble), treating weights and other configurations as tunable hyperparameters.
- Use hyperparameter optimization tools to maximize fidelity to Pensieve's output. The key here is to minimize error (e.g. MSE) against Pensieve's raw bit rate predictions.
- Analyze where the hybrid model diverges from Pensieve. If time permits, use these insights in a closed-loop ML pipeline to refine the model or generate hypotheses about missing decision factors.

Evaluation & Success Metrics

- **Model simplicity**: Track the complexity of the approximating model(s). Ideally, the simpler the model that can mimic Pensieve, the better. Interpretability and parameter count may be used as rough complexity measures.
- **Imitation fidelity**: Determine how well the hybrid white box model imitates Pensieve, either by comparing the raw bit rate choices, or a quality of experience metric, using e.g. R^2 score. The latter (QoE metric) would reveal whether naively mimicking decisions at the level of bit rate still translates to high user satisfaction.

References

[1] Federal Communications Commission. Raw data - measuring broadband america. https://www.fcc.gov/reports-research/reports/measuring-broadband-america/raw-data-measuring-broadband-america-2016, 2016. Accessed: 2025-04-17.

[2] Te-Yuan Huang, Ramesh Johari, Nick McKeown, Matthew Trunnell, and Mark Watson. A buffer-based approach to rate adaptation: evidence from a large video streaming service. In *Proceedings of the 2014 ACM Conference on SIGCOMM*, SIGCOMM '14, page 187–198, New York, NY, USA, 2014. Association for Computing Machinery.

[3] Arthur S. Jacobs, Roman Beltiukov, Walter Willinger, Ronaldo A. Ferreira, Arpit Gupta, and Lisandro Z. Granville. Ai/ml for network security: The emperor has no clothes. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, CCS '22, page 1537–1551, New York, NY, USA, 2022. Association for Computing Machinery.

[4] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. Neural adaptive video streaming with pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, SIGCOMM '17, page 197–210, New York, NY, USA, 2017. Association for Computing Machinery.

[5] Kevin Spiteri, Rahul Urgaonkar, and Ramesh K. Sitaraman. Bola: Near-optimal bitrate adaptation for online videos. *IEEE/ACM Transactions on Networking*, 28(4):1698–1711, 2020.

[6] Xiaoqi Yin, Abhishek Jindal, Vyas Sekar, and Bruno Sinopoli. A control-theoretic approach for dynamic adaptive video streaming over http. *SIGCOMM Comput. Commun. Rev.*, 45(4):325–338, August 2015.