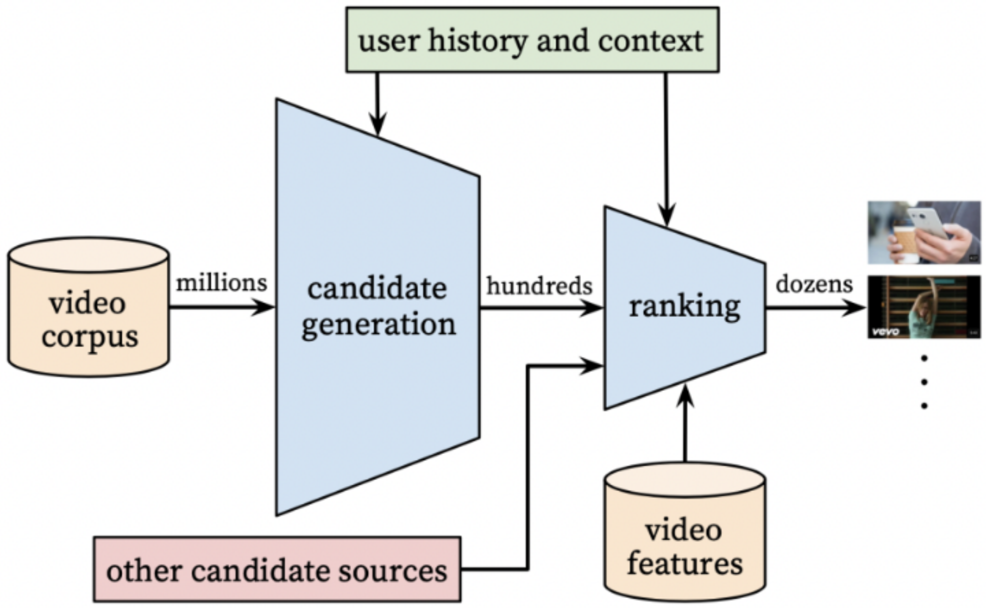
Assignment 3 “Deep Neural Networks for YouTube Recommendations”

Module 3 “Neural Networks–From Market Structure Analysis to Prescriptive Marketing” Learning from Big Data



Summary

The following review paper presents one of the most known deep learning recommender systems in the world - Youtube. The authors discuss the platform as an industrial solution for solving large-scale video recommendation systems, where multiple challenges, as well as the high dimensionality of several factors, should be taken into consideration.

The first challenge the Youtube team faces while training the model on hundreds of billions of examples is scalability. Such a massive platform as Youtube with a high number of DaU (Daily Active Users) requires unconventional methods for finding efficient rather than practical solutions. Secondly, constantly uploaded content on the platform requires a freshness of the platform and non-stop updates in calculating, and interpreting the results. Another factor to bear in mind is that from time to time people randomly click on the video in search of something new, which creates additional noise as an untraceable variety factor that makes decision-making more complicated but helps to prevent overfitting.

On top of that, it is harder to interpret the implicit feedback which is the case when the person watched the video but the model is unaware of the person’s reaction. On the other hand, it becomes easier with explicit feedback, which is a person providing his opinion towards a video by pressing a like-button or writing a comment.

The key idea presented for solving described challenges is the use of two neural networks: one for candidate generation and one for ranking. Candidate Generation aggregates billions of videos as input, taking features such as video ID, tokenized query search, and demographics into a small corpus. This method not only helps to provide broad personalizations via filtering but also accounts for the extreme cost of rather calculating the ratings for a massively large corpus. Moreover, with high precision, the model filters for relevant videos allow small but not crucial errors. The following ranking neural network assigns the scores to each video according to the objective function using a rich set of features describing the video and a user. With just a few dozen relevant videos it’s easier to engineer more variety to the features (both univalent and multivalent) as accessing thumbnail quality, rather than implementing such features to the hundreds of billions of videos. With the help of Recall, videos are then ranked from highest to lowest scores, generating the most relevant and fresh content possible. Since the opportunity cost of missing the video is extremely high, the usage of Recall perfectly fits the problem.

The evaluation of the two-stage approach is conducted online and offline via A/B testing. Since the author mentions that offline experiments are not correlated with online ones, additional features such as watch time and many other metrics should be kept in mind for the assessment.

Figure 4. represents an essential pre-computed expert feature of “Age” which accounts for a major improvement in the model compared to the baselines. The idea of giving the model sense of an “Age” or better said the information about Days since the upload both reduces the cost and the Neural Network picks up the feature instantaneously rather than figuring it out itself. The Youtube team makes minor sacrifices in the relevance of the recommended video at an expense of the freshness. As the author mentions “Without the feature, the

model could predict approximately the average likelihood over the training window”. Furthermore, the second factor to this feature is the phenomenon of bootstrapping and propagating the “viral” content.

One of the key ideas was labeling the context selection which is essentially the trade-off between predicting the next watch vs predicting a randomly held-out watch. The reasoning for choosing to predict the next watch is counterintuitive. When watching a series people tend to watch episodes one after another in a sequential manner. In contrast, making a prediction on what the person watched 4 weeks ago shows poor performance. By combining all features and depth the author achieves the lowest Mean Average Precision.

The target variable is predicting the expected watch time given a positive impression, the video was clicked, or negative, the video was not clicked. After the user clicks on the video the main task of predicting the expected watch time is executed using the weighted logistic regression, while negative impressions receive unit weight. The choice towards targeting the expected watch time rather than the click-through rate directly performs significantly better with the lowest weighted loss per user.

Paper contribution

The contribution the Youtube team has made with their model is immense. Although the paper doesn’t introduce a type of neural network or derivative new insights, it does provide a creative, efficient, and industrial solution. First, it was one of the first recommended systems that operate on that scale. Although the concept of recommending videos was not new, it showed that executing such a heterogenous system in practice is possible. Furthermore, comprising two neural networks together where one filters a huge amount of Youtube’s history activity using precision and afterward implementing the second neural network generating the scores to the most relevant one using recall is a simple, but brilliant idea.

Such video recommender systems are today implemented almost on every subscription streaming platform. The main page of such services uses just several features as input to generate the most relevant but not precise recommendations. For example, creating a new account on Netflix, the platform first asks what region the user is from and what series and shows the person watched in the past.

Strength and weaknesses.

Although the paper shows a new approach in deep neural networks, there are several critiques that we can shed a light on.

Trade-off: Relevance vs Freshness

It is clear why the Youtube team prioritizes freshness at an expense of relevance with an "Age" feature. On the contrary, this feature even today feels forced from the platform to its users rather than giving them the freedom of choice. Different users have different preferences. Some would prefer to watch only fresh content at a sacrifice of relevance and others would prefer the relevance of some channels at the expense of the freshness of less relevant videos.

The great trade-off between both surprisingly shows Twitter. The social media platform gives its users a choice between what they value the most, see **Figure 1.** This might be a great addition to the Youtube policy of implementing a such feature and giving its users more mobility and freedom of choice.

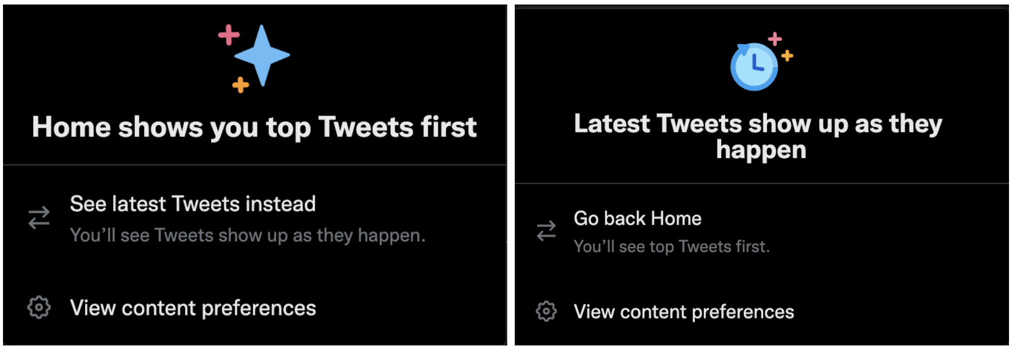


Figure 1.

Transparency of features

Although the paper describes in detail the features and technical aspects of normalizing and assigning weights - it definitely lacks transparency on what other features the Youtube algorithm takes into consideration or what insights from A/B testing they implement into the model. For example, the most popular Youtube channel - Mr.Beast with 110 Mio. Subscribers and every video tops over 50 Mio. views mentions the complexity of the algorithms for promoted videos. He mentions that the resolution of your thumbnail or even the color palette you choose does play a role in gaining higher scores and eventually being promoted by the algorithm. (Reference: Joe Rogan x Mr.Beast, 9:50 - 15:00)

Practical lesson

The Youtube paper has made me think of an even more complex problem with a lot of heterogeneity in play and an even larger amount of data that as well should be constantly interpreted, , scored and evaluated online and offline. Financial institutions around the world face similar challenges as Youtube dealing with the scale, freshness, and noise of capital markets.

Scale.

Every day from East to West the stock exchanges around the world generate a great amount of data about the economy, business industries, and politics. In order to process the magnitude of constantly upcoming information and evaluate the probabilities, for example, of what stock in the portfolio manager should short or long - the use of highly specialized algorithms and efficient serving systems are essential for handling the information that can have an influence on the markets.

Freshness.

Capital markets are highly competitive and dynamic places, where the information flow can be priced in a matter of minutes, sometimes even seconds with the use of HFT. (High-Frequency Trading). The algorithms should be responsive enough for constant evaluation of events that can play a role in the portfolios of the institutions. As in Youtube balancing between the existing and newly given information is key.

Noise

Noise in financial markets is often understood as retail investors or an average person. She or he may be buying BTC because a friend recommended it, or selling a company when the rest of the market sells.

Recommending videos ≠ Recommending stocks

We can recommend the next video, but can we recommend stocks?

Youtube two staged neural network approach can be helpful in recommending and optimizing the stock portfolio to institutional investors with a different portfolio risk profile. First, the similar deep neural network **"Stock" Generator** is applied to the portfolio using past holding positions as input in order to understand the risk profile but also some additional features such as wished return, a timeline of the portfolio, and demographics would help to retrieve a small subset of (hundreds) of stocks with a similar profile. These candidates are picked with the use of high precision and also personalized via filtering depending on the wished exposure of the portfolio manager. The **ranking** of stocks would allow assigning stocks the probabilities of achieving wished return-risk profile of the portfolio. The risk of recommending the wrong stocks is a problem of even greater magnitude than recommending the wrong front-page video, thus, the high recall would play a big part. Building some additional features during ranking such as liquidity, trading volume, price, macro sentiment, and technical analysis would help to make the model more efficient since only the dozens of stocks should be evaluated.

In the world of investing the financial health of the company and the sentiment of the markets are only one part of the story. Political decisions, legal rulings, social hype, and so on could have different pillars for making the decision towards buying security. Thus, as in Youtube, an offline evaluation of the model final-rated stocks should be conducted sorely and weighted accordingly. It's hard to say what weights should be assigned to a professional stock picker (Manager) and a model for evaluating the final recommended stock choice. Both managers and models can be biased in their own way. For example, the model analyzing the past market data can recommend stocks solely from the technology industry for the reason that the best return-risk profile of the last decade was the technology sector, however, every manager has his/her own diversification profile. On the other hand, a manager that focuses on diversification a lot can make a wrong judgment on a final judgment of a recommended stock choice missing a potentially recommended alpha.

The proposed research might be interesting since a lot of current algorithmic trading models focuses either on High-Frequency Trading or buying or selling stock in a matter of second reacting to political events. The cooperation between the physical person and algorithm is crucial in order to have a better scope and understanding of the economical environment. Such a stock recommendation system can be then proposed as a SaaS product and used by a different variety of institutional investors.

Reference:

1. https://open.spotify.com/episode/5lokpznqvSrJO3gButgQvs?si=dcfec81fc5ff4283