DEEP NEURAL NETWORKS FOR YOUTUBE RECOMMENDATIONS

Paul Covington, Jay Adams, Emre Sargin

Felipe del Río

Pontificia Universidad Católica de Chile

INTRODUCTION

YOUTUBE RECOMMENDATIONS

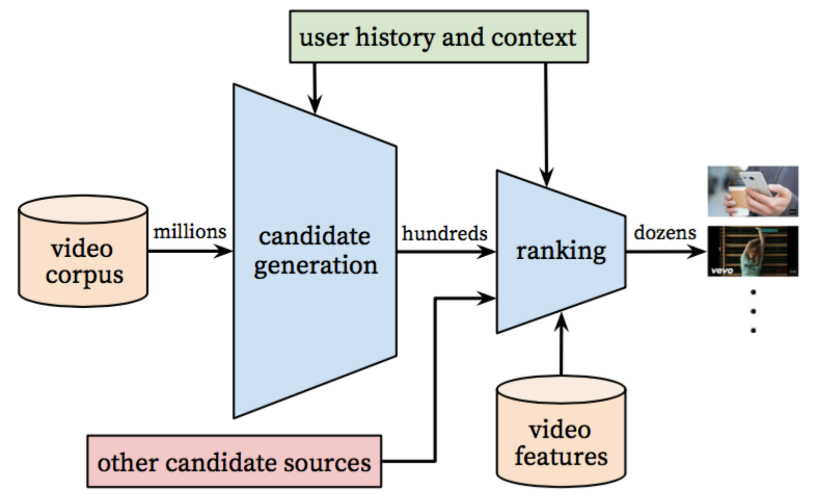
YouTube is the world's largest platform for creating, sharing and discovering video content.

Recommending YouTube videos is extremely challenging:

- Scale
- Freshness
- Noise

SYSTEM OVERVIEW

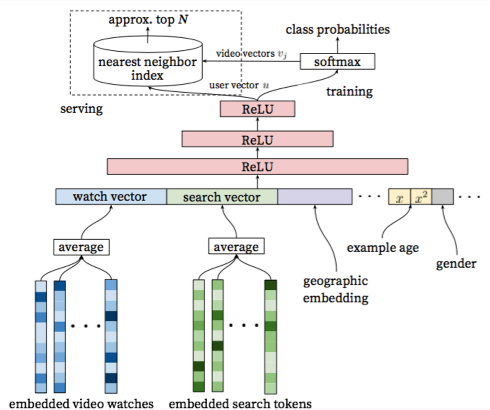
SYSTEM OVERVIEW



CANDIDATE GENERATION

CANDIDATE GENERATION

YouTube corpus is reduced down to hundreds of videos that may be relevant to the user.



CANDIDATE GENERATION: RECOMMENDATION AS A CLASSIFICATION

Recommendation as extreme multiclass classification.

Accurately classifying a specific video watch w_t at time t among millions of videos i (classes) from a corpus V based on a user U and context C

$$P(w_t = i|U, C) = \frac{e^{v_i u}}{\sum_{j \in V} e^{v_j u}}$$

Where u,v_i and v_j are embeddings learned by the deep neural network.

Use the implicit feedback [Oard et al., 1998] of watches to train the model, where a user completing a video is a positive example.

Sample negative classes and then correct via importance weighting.

MODEL ARCHITECTURE

Inspired by continuous bag of words language models [Mikolov et al., 2013].

Learn high dimensional embeddings for each video and feed these embeddings into a feedforward neural network.

User's watch history is represented by a variable-length sequence of sparse video IDs which is mapped to a dense vector representation via the embeddings.

Embeddings are learned jointly with all other model parameters.

Features are concatenated into a wide first layer.

HETEROGENEOUS SIGNALS

Search history is treated similarly to watch history, each query is tokenized into unigrams and bigrams and each token is embedded.

The user's geographic region and device are embedded and concatenated.

Simple binary and continuous features such as the user's gender, logged-in state and age are input directly into the network as real values normalized to [0, 1].

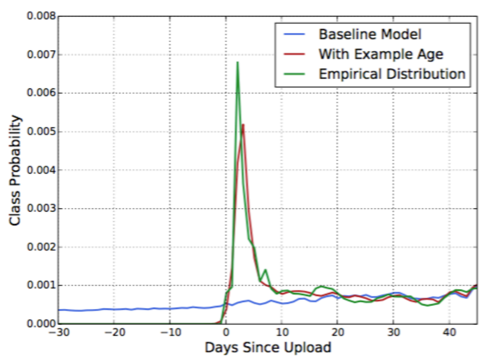
HETEROGENEOUS SIGNALS: “EXAMPLE AGE” FEATURE

Users prefer fresh content.

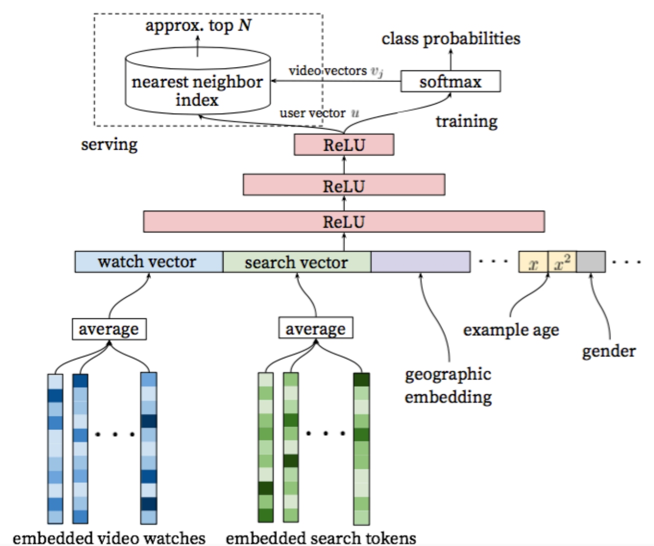
Critical secondary phenomenon of bootstrapping and propagating viral content [Jiang et al., 2014].

Feed age of the training example as a feature during training.

At serving time, this feature is set to zero (or slightly negative).



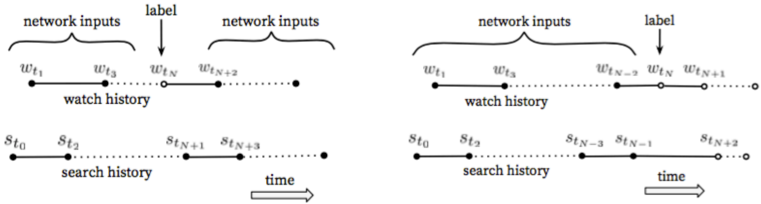
CANDIDATE GENERATOR MODEL



LABEL AND CONTEXT SELECTION

Divide test set by consider the influence of time and context.

Better performance predicting the user's next watch, rather than predicting a randomly held-out watch.



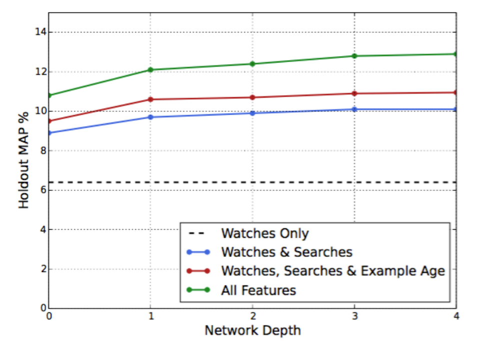
(a) Predicting held-out watch

(b) Predicting future watch

EXPERIMENTS WITH FEATURES AND DEPTH

Experimented with 1M videos and 1M search tokens embedded with 256 floats each in a maximum bag size of 50 recent watches and 50 recent searches.

Trained until convergence over all YouTube users.



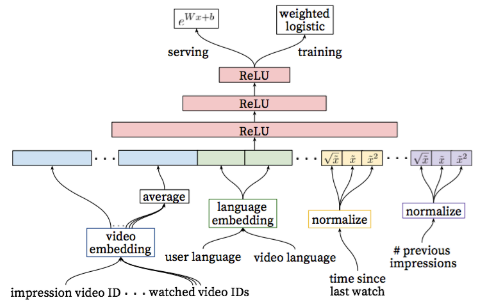
RANKING

RANKING

The primary role of ranking is to use impression data to specialize and calibrate candidate predictions for the particular user interface.

Access to many more features describing the video and the user's relationship to the video.

Final ranking objective depends A/B testing results but is a simple function of expected watch time per impression.



FEATURE REPRESENTATION

Different kinds of features:

- Categorical and ordinal features.
- Categorical features can be binary while others have millions of possible values.
- Contributes only a single value (“univalent”) or a set of values (“multivalent”).
- Describes properties of the item (“ impression”) or properties of the user/context (”query”).

FEATURE REPRESENTATION: FEATURE ENGINEERING

The nature of the raw data does not easily lend itself to be input directly into feedforward neural networks.

The main challenge is in representing a temporal sequence of user actions and how these actions relate to the video impression being scored.

User's previous interaction with the item or others:

- How many videos has the user watched from this channel?
- When was the last time the user watched a video on this topic?

Information from candidate generation into ranking:

- Which sources nominated this video candidate?
- What scores did they assign?

FEATURE REPRESENTATION: EMBEDDING AND NORMALIZING

Embeddings:

- Use embeddings to map sparse categorical features to dense representations suitable for neural networks.
- The embedding dimension that increases approximately proportional to the logarithm of the number of unique values.

Normalization:

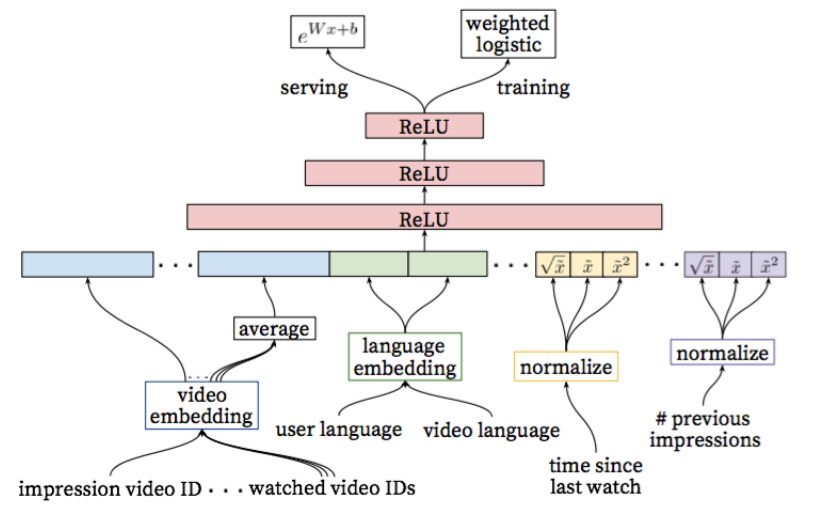
- Continuous features are quantile normalized.
- In addition to the raw normalized feature $\tilde{x}$, input powers $\tilde{x}^2$ and $\sqrt{\tilde{x}}$ are given.

MODELING EXPECTED WATCH TIME

The goal is to predict expected watch time given training examples that are either positive (video impression was clicked) or negative (not clicked).

Positive (clicked) impressions are weighted by the observed watch time on the video and negative (unclicked) impressions all receive unit weight.

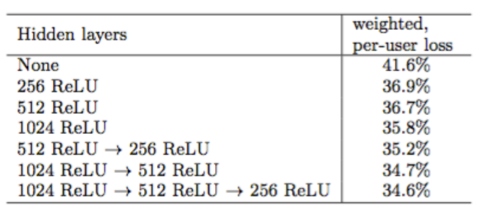
RANKING ARCHITECTURE



EXPERIMENTS WITH HIDDEN LAYERS

The value shown for each configuration ("weighted, per-user loss") was obtained by considering both positive (clicked) and negative (unclicked) impressions shown to a user on a single page.

The trade-off is server CPU time needed for inference.



Hidden layers	weighted, per-user loss
None	41.6%
256 ReLU	36.9%
512 ReLU	36.7%
1024 ReLU	35.8%
512 ReLU → 256 ReLU	35.2%
1024 ReLU → 512 ReLU	34.7%
1024 ReLU → 512 ReLU → 256 ReLU	34.6%

CONCLUSIONS

CONCLUSIONS

The deep collaborative filtering model is able to outperform previous matrix factorization approaches used previously at YouTube.

Using the age of the training example removes an inherent bias towards the past and allows the model to represent the time-dependent behavior of popular of videos.

Deep learning approach outperformed previous linear and tree-based methods for watch time prediction.

The weighted logistic regression approach performed much better on watch-time weighted ranking evaluation metrics compared to predicting click-through rate directly.

REFERENCES

Jiang, L., Miao, Y., Yang, Y., Lan, Z., and Hauptmann, A. G. (2014).
Viral video style: a closer look at viral videos on youtube.
In Proceedings of International Conference on Multimedia Retrieval, page 193. ACM.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013).
Distributed representations of words and phrases and their compositionality.
In Advances in neural information processing systems, pages 3111–3119.

Oard, D. W., Kim, J., et al. (1998).
Implicit feedback for recommender systems.
In Proceedings of the AAAI workshop on recommender systems, pages 81–83.

Questions?