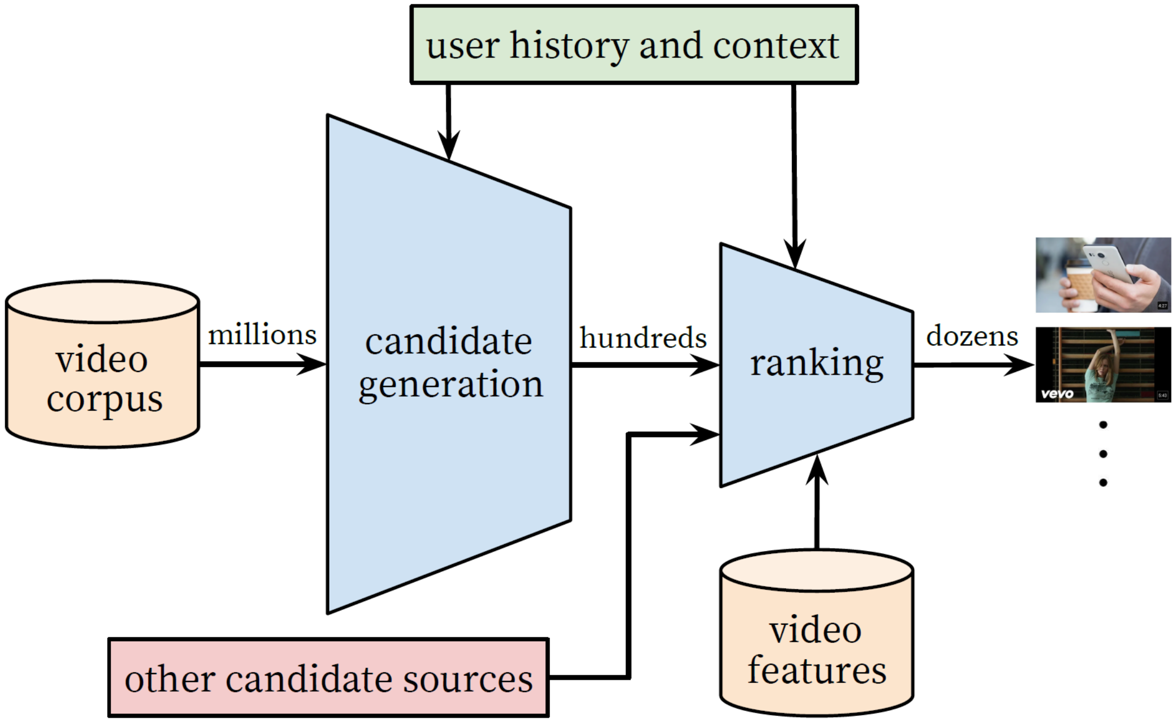

user history and context
video
corpus
millions
candidate
generation
hundreds
ranking
dozens
other candidate sources
video
features

1 多路召回


召回层
排序层
兴趣标签
兴趣Topic
兴趣实体
协同过滤
热门
Top K1
Top K2
Top K3
Top K4
Top K5
排序模型
(LR/FM/WDL/DeepFM...)

2 Embedding 召回

2.1 YouTube DNN

Deep Neural Networks for YouTube Recommendations (Google 2016)

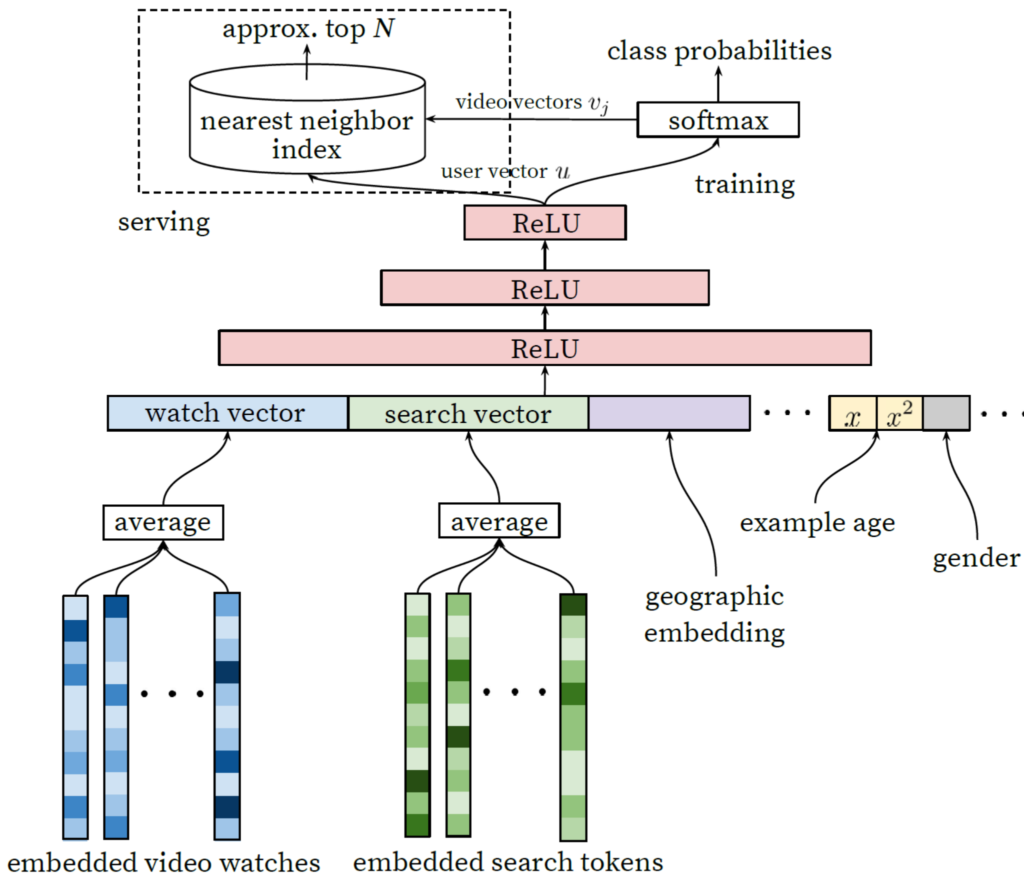


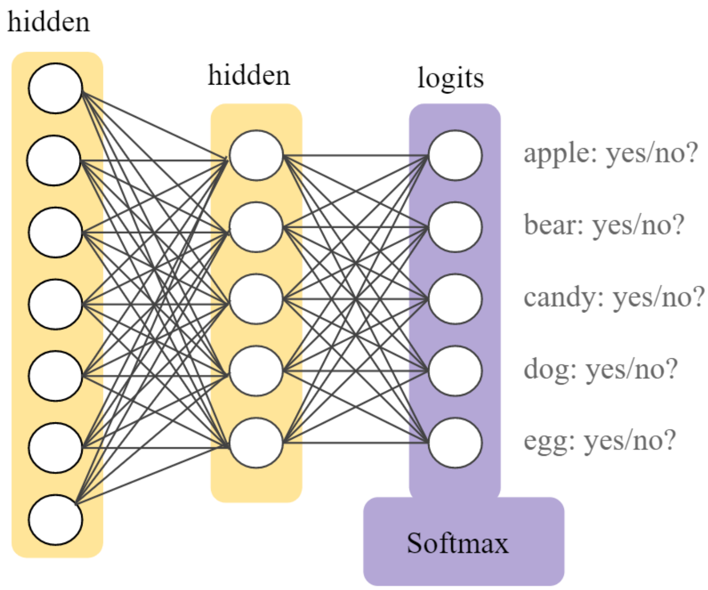
Figure 3: Deep candidate generation model architecture showing embedded sparse features concatenated with dense features. Embeddings are averaged before concatenation to transform variable sized bags of sparse IDs into fixed-width vectors suitable for input to the hidden layers. All hidden layers are fully connected. In training, a cross-entropy loss is minimized with gradient descent on the output of the sampled softmax. At serving, an approximate nearest neighbor lookup is performed to generate hundreds of candidate video recommendations.

Training 阶段

- 负采样 (candidate sampling)
  - 使用 Softmax 函数针对所有正样本计算概率，但对于负样本则仅针对其随机样本计算概率，且只针对部分隐藏层权重进行小范围更新，从而提高训练效率
  - 例如，假设某个样本的标签为“小猎犬”和“狗”，则针对“小猎犬”和“狗”类别输出以及其他类别的随机子集计算预测概率和相应的损失项
  - 负采样基于的假设是，只要正样本始终得到适当的正增强，负样本就可以从频率较低的负增强中进行学习，从而最大化正样本的概率，同时最小化负样本的概率

Serving 阶段

- video vectors v_j – Softmax 前全连接层对应的权重



2.2 双塔结构 DSSM

Learning Deep Structured Semantic Models for Web Search using Clickthrough Data (Microsoft 2013)

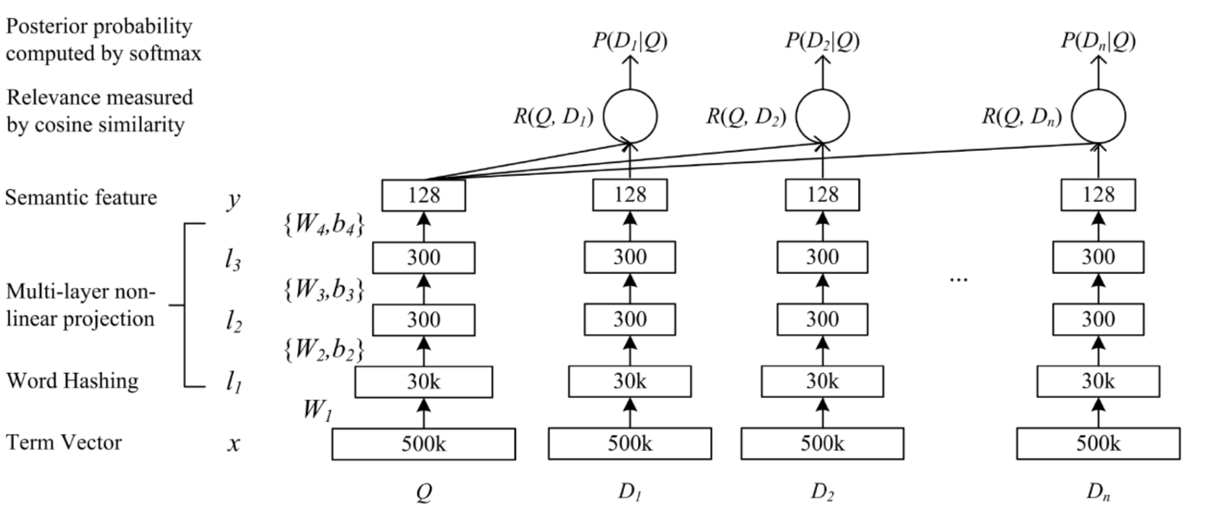


Figure 1: Illustration of the DSSM. It uses a DNN to map high-dimensional sparse text features into low-dimensional dense features in a semantic space. The first hidden layer, with 30k units, accomplishes word hashing. The word-hashed features are then projected through multiple layers of non-linear projections. The final layer's neural activities in this DNN form the feature in the semantic space.

输入层

中间层

匹配层

Sampling-Bias-Corrected Neural Modeling for Large Corpus Item Recommendations (Google 2019)

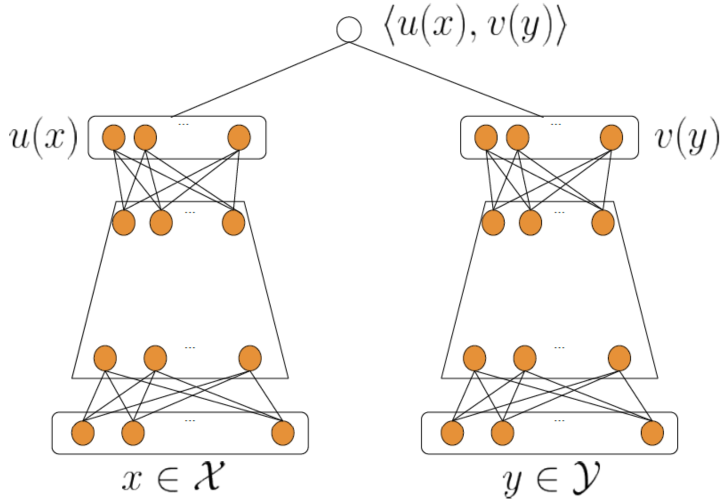


Figure 1: A two-tower DNN model for learning query and candidate representations.

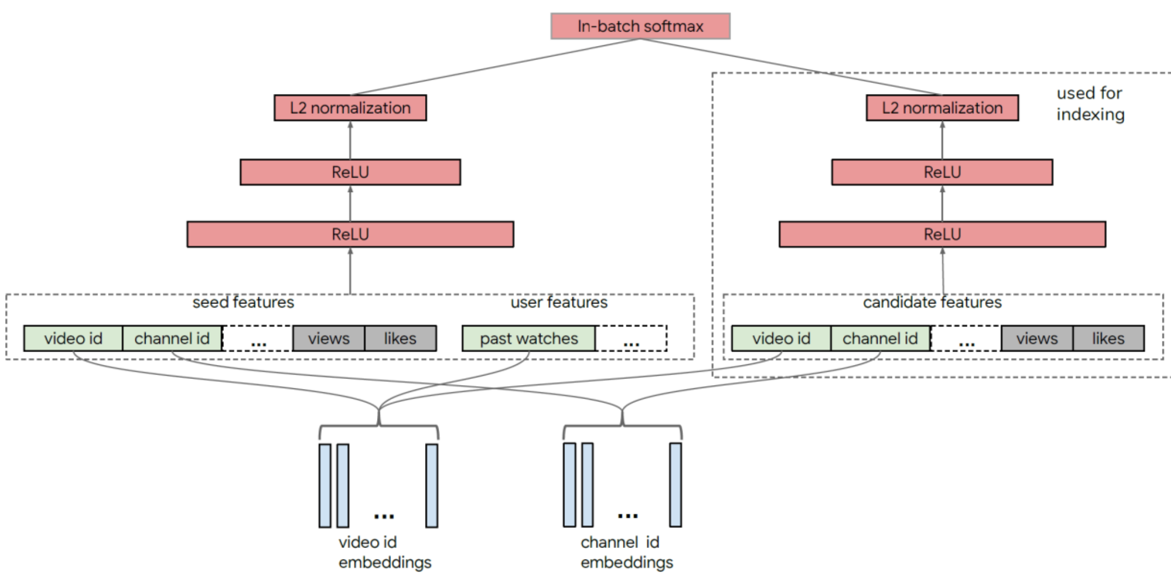


Figure 2: Illustration of the Neural Retrieval Model for YouTube.

2.3 Item2Vec

2.4 Airbnb Embedding

2.5 FM/FFM/DeepFM