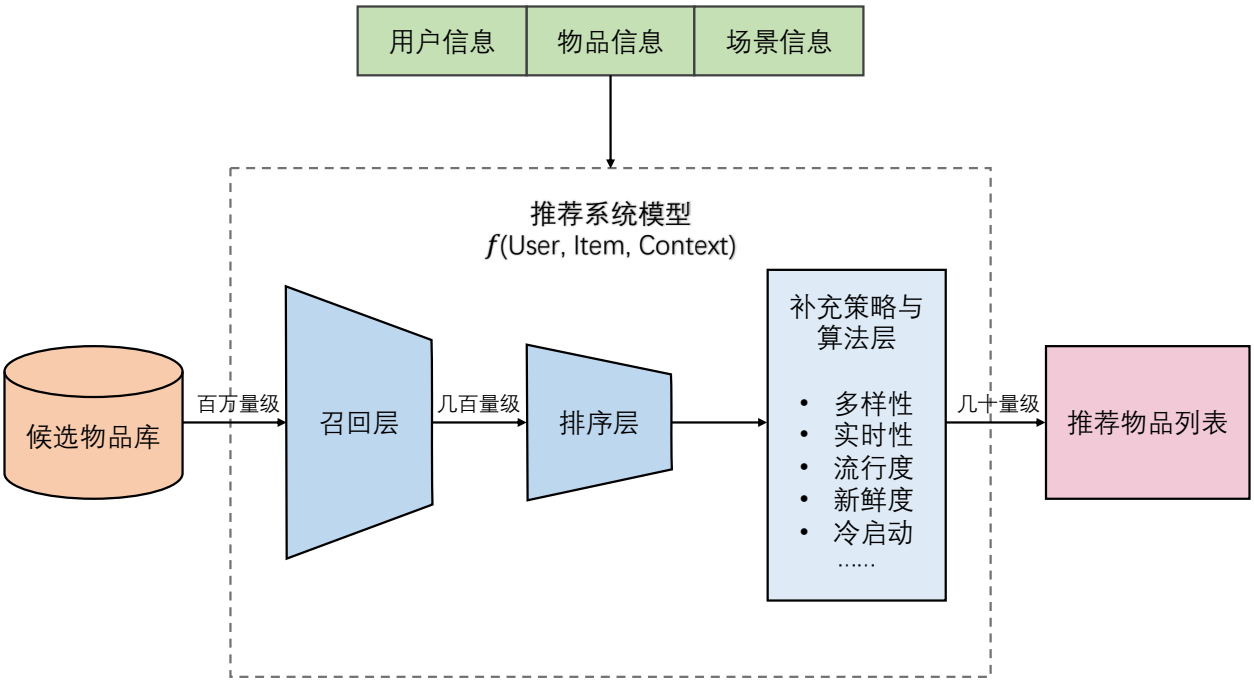
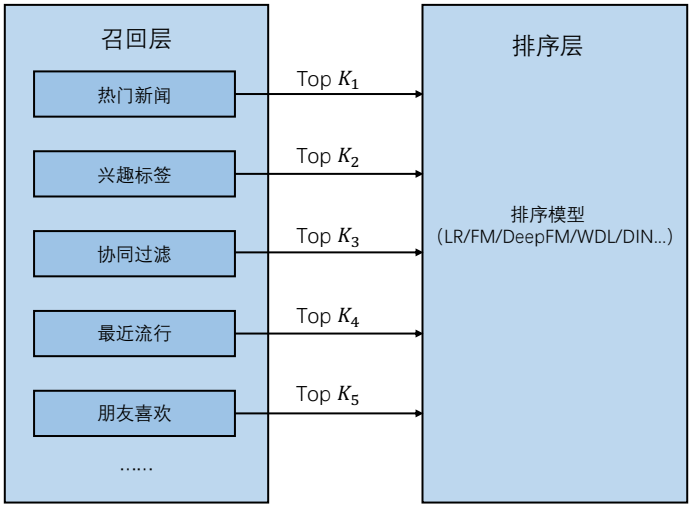


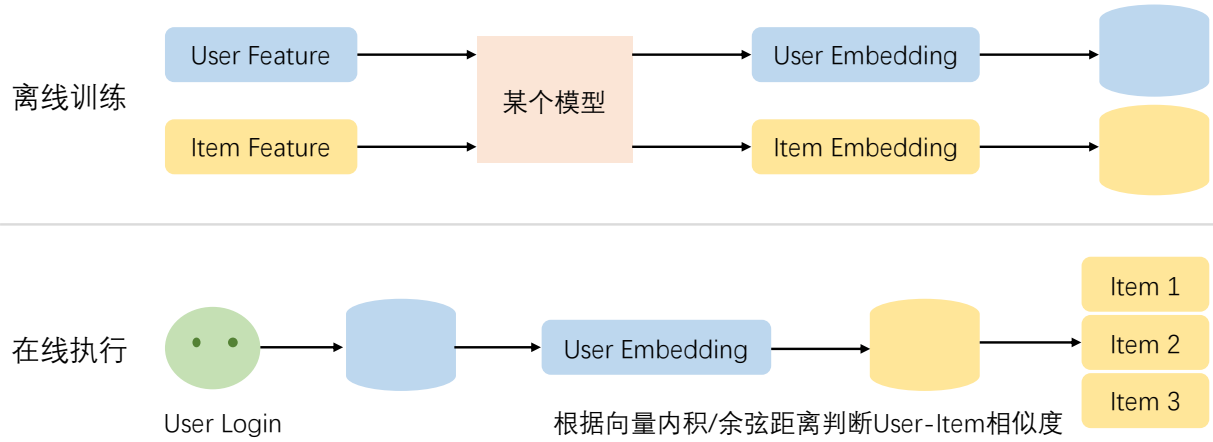
1 多路召回策略



采用不同的策略、特征或简单模型分别召回一部分候选集，然后将候选集混合在一起供后续排序模型使用。

- **优点**：各简单策略能保证候选集的快速召回；召回策略的选择与业务强相关，从不同的角度设计的策略能使得召回率接近理想的状态
- **缺点**：从召回策略的选择到候选集大小参数K到调整都需要人工参与；策略之间的信息也是割裂的，无法综合考虑不同策略对同一个物品的影响

2 基于 Embedding 的召回方法



2.1 YouTube DNN

Deep Neural Networks for YouTube Recommendations (Google 2016)

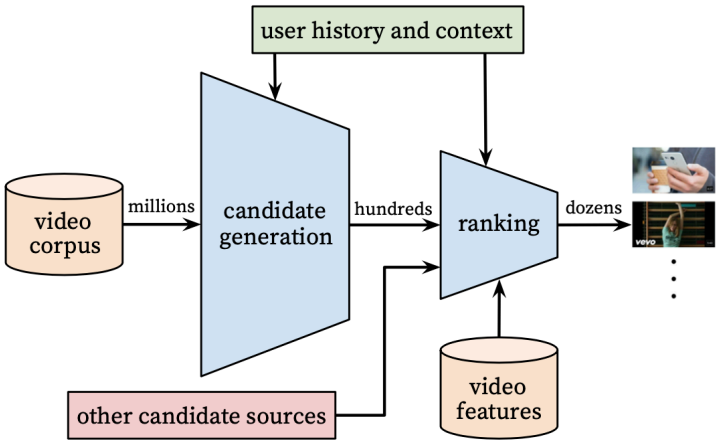


Figure 2: Recommendation system architecture demonstrating the "funnel" where candidate videos are retrieved and ranked before presenting only a few to the user.

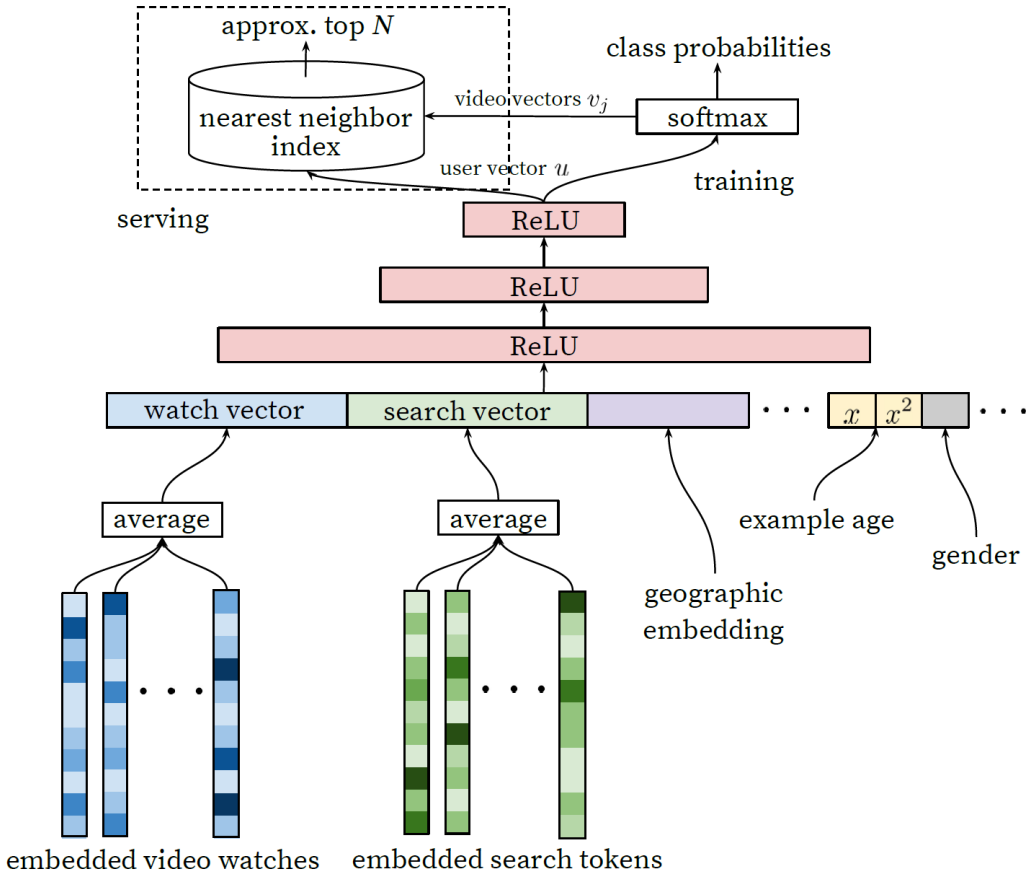


Figure 3: Deep candidate generation model architecture showing embedded sparse features concatenated with dense features. Embeddings are averaged before concatenation to transform variable sized bags of sparse IDs into fixed-width vectors suitable for input to the hidden layers. All hidden layers are 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.

总的来说，YouTube 推荐系统的候选集生成模型（candidate generation model）是一个标准的利用 Embedding 预训练特征的深度神经网络模型。

输入层

底层输入包括用户的历史观看视频 Embedding 和搜索词 Embedding，以及用户的地理属性特征 Embedding、性别、年龄等。

- 为了生成视频 Embedding 和搜索词 Embedding，YouTube 利用用户的观看序列和搜索序列，采用 Word2Vec 方法对视频和搜索词作做 Embedding，再作为候选集生成模型的输入
- 除了进行预训练 Embedding，还可以直接在深度学习网络中增加 Embedding 层，与上层 DNN 一起进行端到端训练，但 Embedding 层的存在往往会拖慢整个神经网络的收敛速度
  - Embedding 层输入向量往往维度很大，导致整个 Embedding 层的参数数量巨大，大部分训练时间和计算开销都被 Embedding 层占据
  - 由于输入向量过于稀疏，在随机梯度下降过程中，只有与非零特征相连的 Embedding 层权重才会被更新，进一步降低 Embedding 层的收敛速度

输出层

模型训练（training）

YouTube 把选择候选视频集看作给用户推荐下一次观看视频（next watch），是一个多分类问题，模型最终的输出是一个在所有候选视频上的概率分布，采用 Softmax 函数作为输出层。

$$P(y=j|\mathbf{x})=\frac{\exp\left(\mathbf{w}_j^{\mathrm{T}}\mathbf{x}+b_j\right)}{\sum_{k\in K}\exp\left(\mathbf{w}_k^{\mathrm{T}}\mathbf{x}+b_k\right)}$$

模型服务（serving）

在模型训练完成后，逐个输入所有用户的特征向量到模型中，得到所有用户的 Embedding 向量导入线上 Embedding 数据库。在预测某用户的视频候选集时，先得到该用户对应的 Embedding 向量，再在视频 Embedding 向量空间中利用局部敏感哈希（Locality Sensitive Hashing, LSH）等最近邻搜索算法搜索该用户 Embedding 向量的 Top$K$近邻，即可得到K个候选视频

- 视频 Embedding 向量　最终输出 Softmax 层的参数假设是一个$m \times n$维的矩阵，其中m指的是最后一层 ReLU 层的维度，n指的是类别数目，即 YouTube 所有视频的总数，那么这个$m \times n$维的列向量就是视频 Embedding
- 用户 Embedding 向量　输入特征向量都是用户相关的特征，在某用户u的特征向量所谓模型输入时，最后一层 ReLU 层的输出向量就可以作为该用户的 Embedding

2.2 双塔结构 DSSM

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

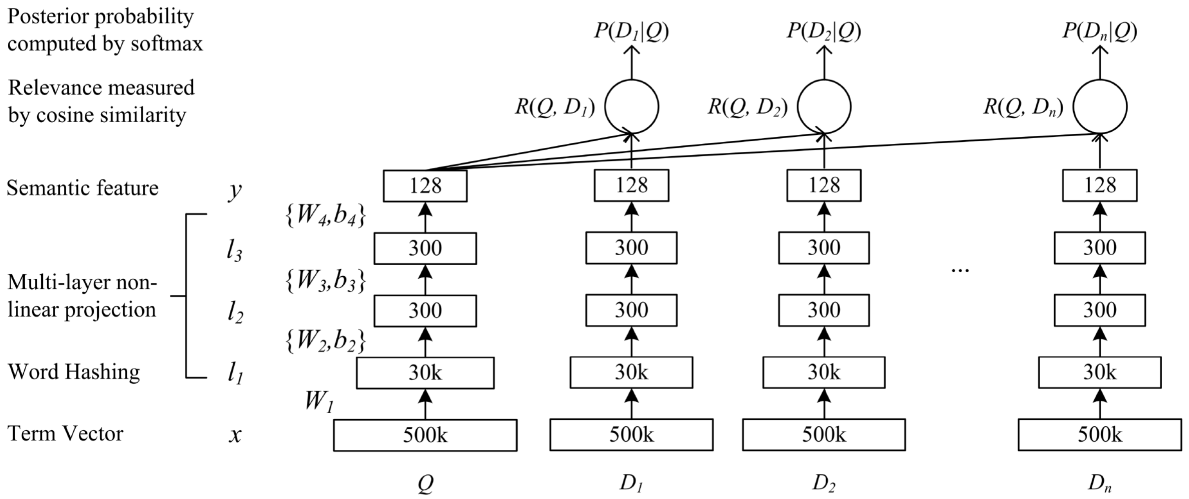


Figure 1: Illustration of the DSSM. It uses a DNN to map high-dimensional sparse text features into low-dimensional dense features in a semantic space. The first hidden layer, with 30k units, accomplishes word hashing. The word-hashed features are then projected through multiple layers of non-linear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

DSSM（Deep Structured Semantic Model）是最早提出双塔结构的模型，由于用到了点击数据，相比 LDA 等隐语义模型，DSSM 是有监督等学习。原文中单词 query 和文档 doc 等匹配和 doc 等匹配对应到推荐系统中是 user 和 item 的匹配。

$$R(Q,D) = \text{cosine}(y_Q, y_D) = \frac{y_Q^T y_D}{\|y_Q\| \|y_D\|}$$

$$P(D|Q) = \frac{\exp(\gamma R(Q,D))}{\sum_{D' \in \boldsymbol{D}} \exp(\gamma R(Q,D'))}$$

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

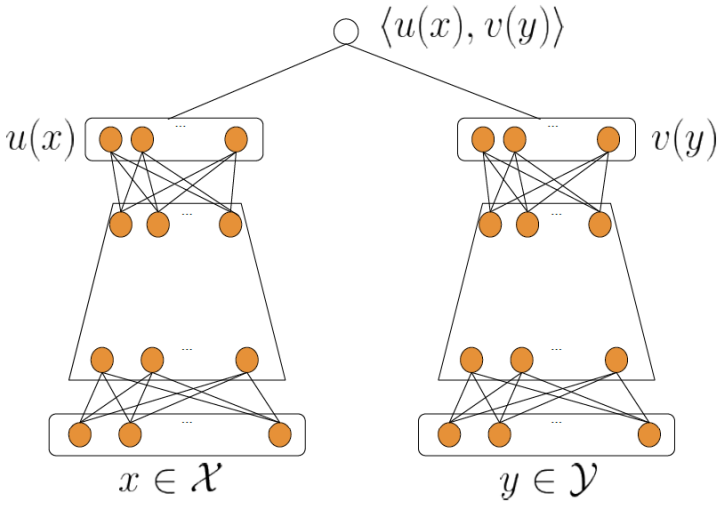


Figure 1: A two-tower DNN model for learning query and candidate representations.

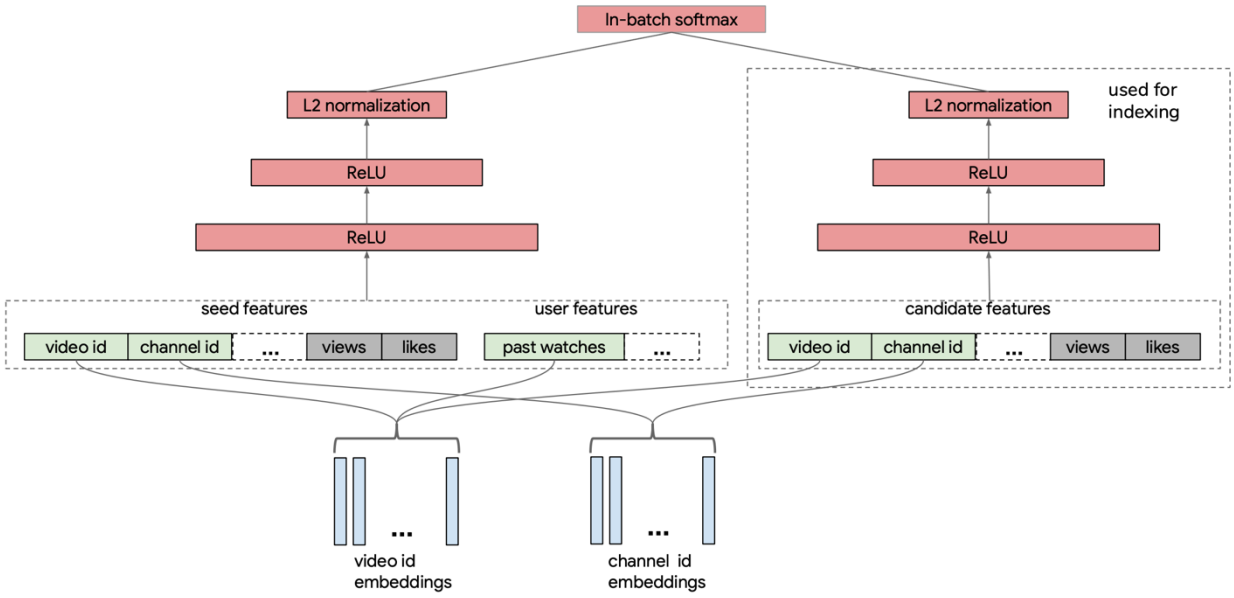


Figure 2: Illustration of the Neural Retrieval Model for YouTube.

YouTube 双塔结构召回模型：左塔输出是包含 user 信息的 embedding 向量，右塔输出是 item embedding 向量，均使用 DNN 来学习，两塔联合训练，确保 user embedding 和 item embedding 在同一向量空间，内积才有意义。

2.3 FM 召回模型

FM 模型——隐向量特征交叉

$$\hat{y}(x) := w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \langle v_i, v_j \rangle x_i x_j$$

$$\langle v_i, v_j \rangle := \sum_{f=1}^{k} v_{i,f} \cdot v_{j,f}$$

FM 模型在 LR 模型的基础上引入任意两两特征的交叉项，对每个特征学习一个大小为k的一位向量，特征x_i和x_j交叉的权重系数通过特征对应向量v_i和v_j的内积$\langle v_i, v_j \rangle$来表示，本质上是对特征进行 embedding 化表示。

如何利用 FM 模型做统一召回模型

Step1 对于每个用户，查询离线训练好的 FM 模型对应的用户特征 embedding 向量，然后将用户特征 embedding 累加，得到用户 embedding 向量；类似地，查询物品特征的 embedding 向量并累加，可得到物品 embedding 向量，存入在线数据库。

Step2 当用户登陆或刷新页面时，可根据用户 ID 取出其对应的 embedding 向量，与物品 embedding 向量做内积运算，按得分由高到低取 Top$K$作为召回结果。

3 统一名单召回策略

- 类比推荐系统，普惠可看成唯一“用户”，提交客户名单看成“推荐产品”的过程
  - 策略输入部分可看成“召回层”，主要负责名单的初步筛选，目标是尽可能圈中更多放款客户
  - 智能诊断部分可看成“排序层”，利用意向/资质模型打分提高名单预约率/征信通过率/转化率等指标
- **假设与放款客户在某些特征上相似的候选客户未来更有可能放款**，可在全量客户中寻找与放款客户相似的 Top$K$客户作为候选集。即每个标杆可带来K个候选客户
  1. 计算所有客户的 Embedding 向量
  2. 对于每个标杆客户，利用最近邻搜索算法寻找与之相似的 Top$K$非标杆客户加入候选集，最后对候选集进行去重