

物品冷启动：聚类召回

王树森

ShusenWang@xiaohongshu.com

<http://wangshusen.github.io/>



聚类召回

基本思想

- 如果用户喜欢一篇笔记，那么他会喜欢内容相似的笔记。
- 事先训练一个神经网络，基于笔记的类目和图文内容，把笔记映射到向量。
- 对笔记向量做聚类，划分为 1000 cluster，记录每个 cluster 的中心方向。(k-means 聚类，用余弦相似度。)

聚类召回

聚类索引

- 一篇新笔记发布之后，用神经网络把它映射到一个特征向量。
- 从 1000 个向量（对应 1000 个 cluster）中找到最相似的向量，作为新笔记的 cluster。
- 索引：

cluster → 笔记ID列表（按时间倒排）

聚类召回

线上召回

- 给定用户ID，找到他的 last-n 交互的笔记列表，把这些笔记作为种子笔记。
- 把每篇种子笔记映射到向量，寻找最相似的cluster。
(知道了用户对哪些 cluster 感兴趣。)
- 从每个 cluster 的笔记列表中，取回最新的 m 篇笔记。

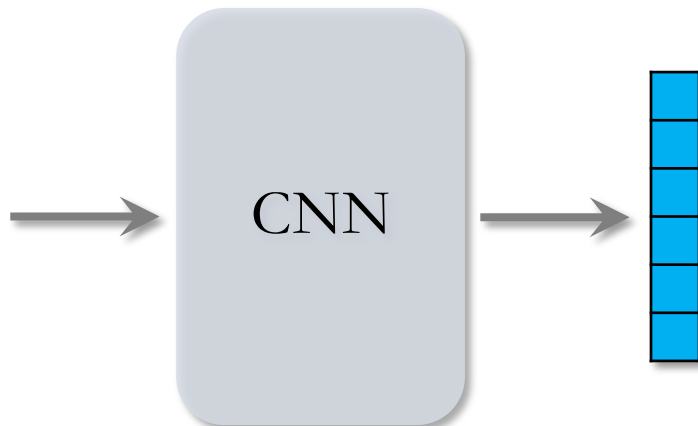
聚类召回

线上召回

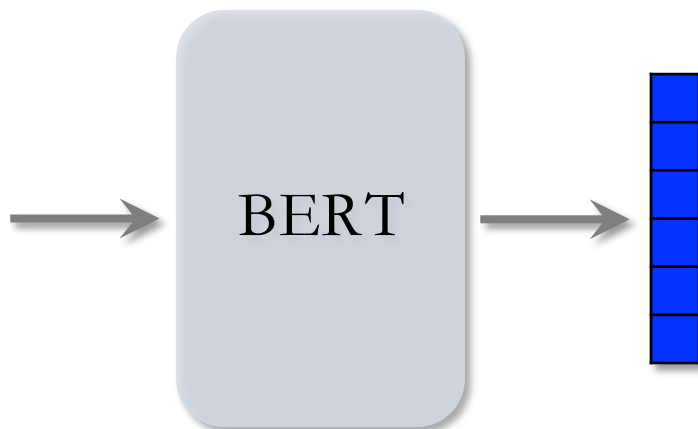
- 给定用户ID，找到他的 last- n 交互的笔记列表，把这些笔记作为种子笔记。
- 把每篇种子笔记映射到向量，寻找最相似的cluster。
(知道了用户对哪些 cluster 感兴趣。)
- 从每个 cluster 的笔记列表中，取回最新的 m 篇笔记。
- 最多取回 mn 篇新笔记。

内容相似度模型

提取图文特征



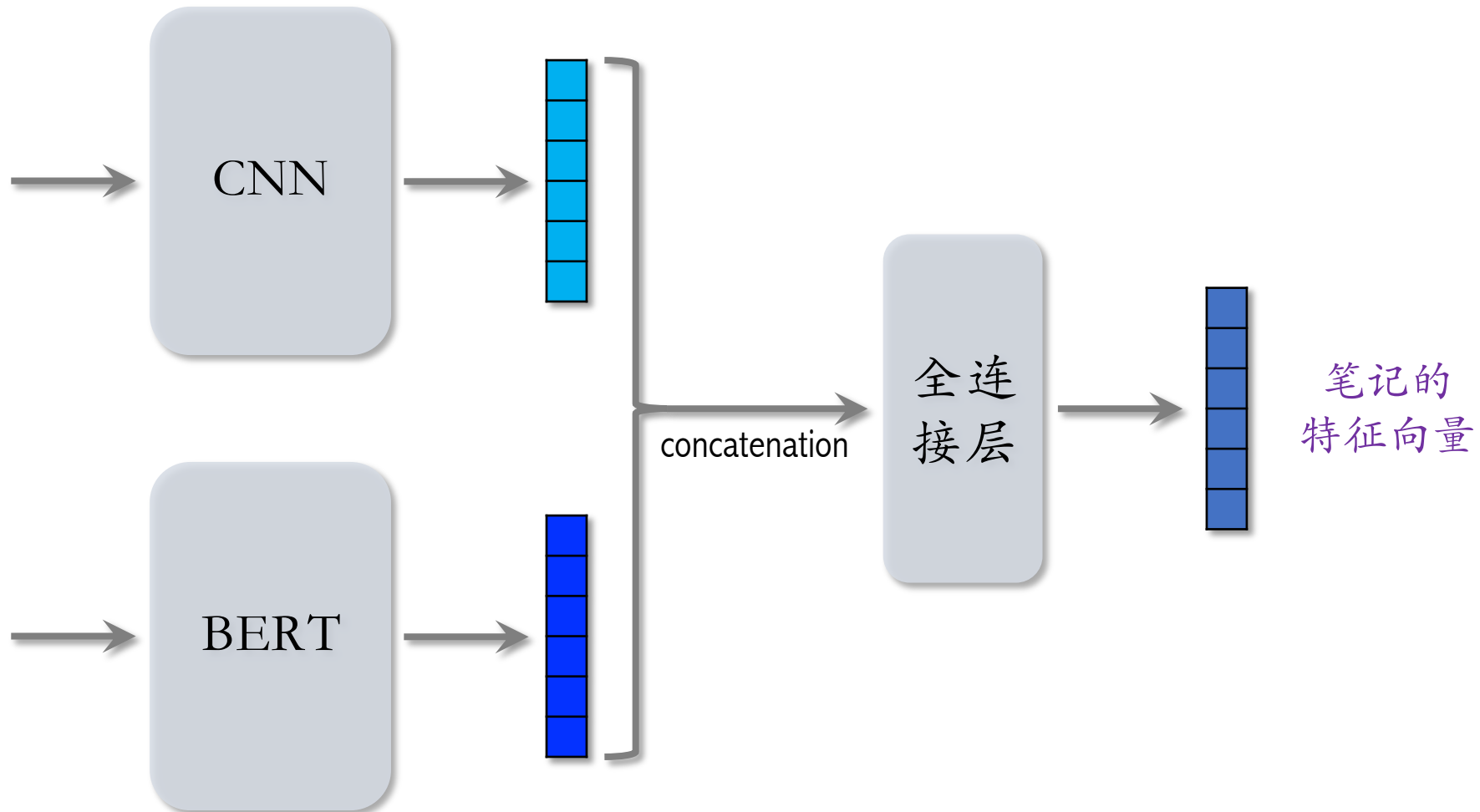
在日本买柴犬最详细解说
我是店长八谷。我是日本人，今年是我在东京开宠物店的第16年。
很多中国客人特意来日本买柴犬。有客人说大阪某宠物店60万日元（约4w人民币）一条柴犬。
日本柴犬真那么贵嘛？怎么可能💣💣💣
这是八谷为过往客人办理柴犬的费用💰
中国的朋友只看这一篇就可以了解大致的行情📊



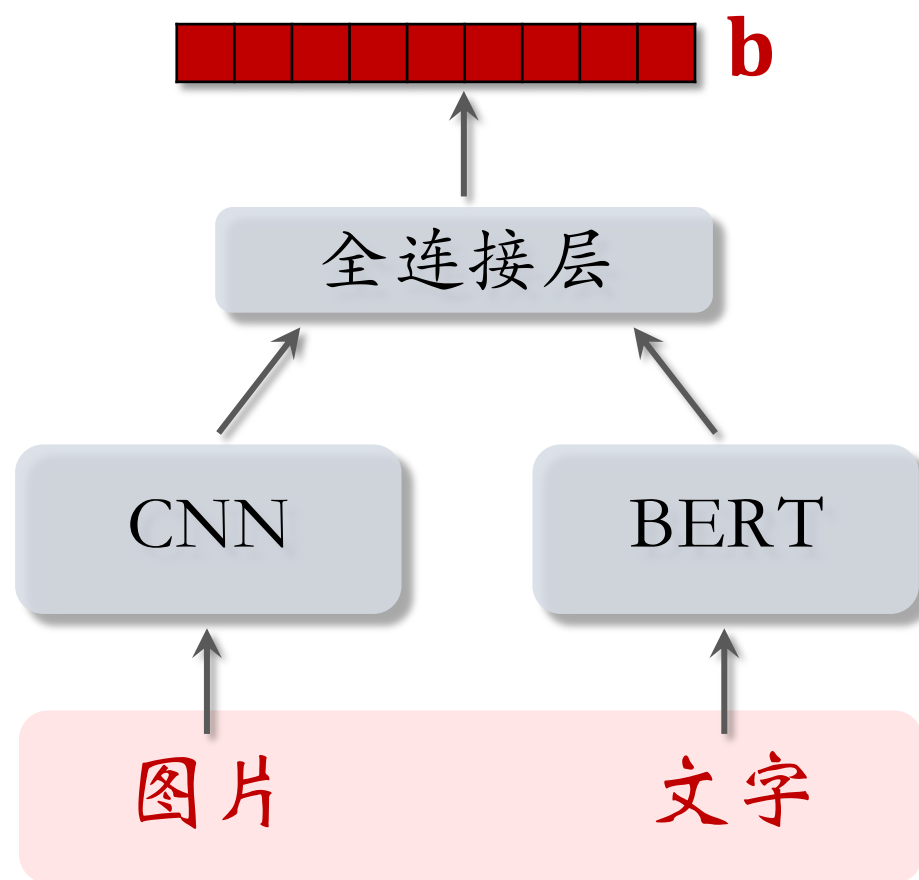
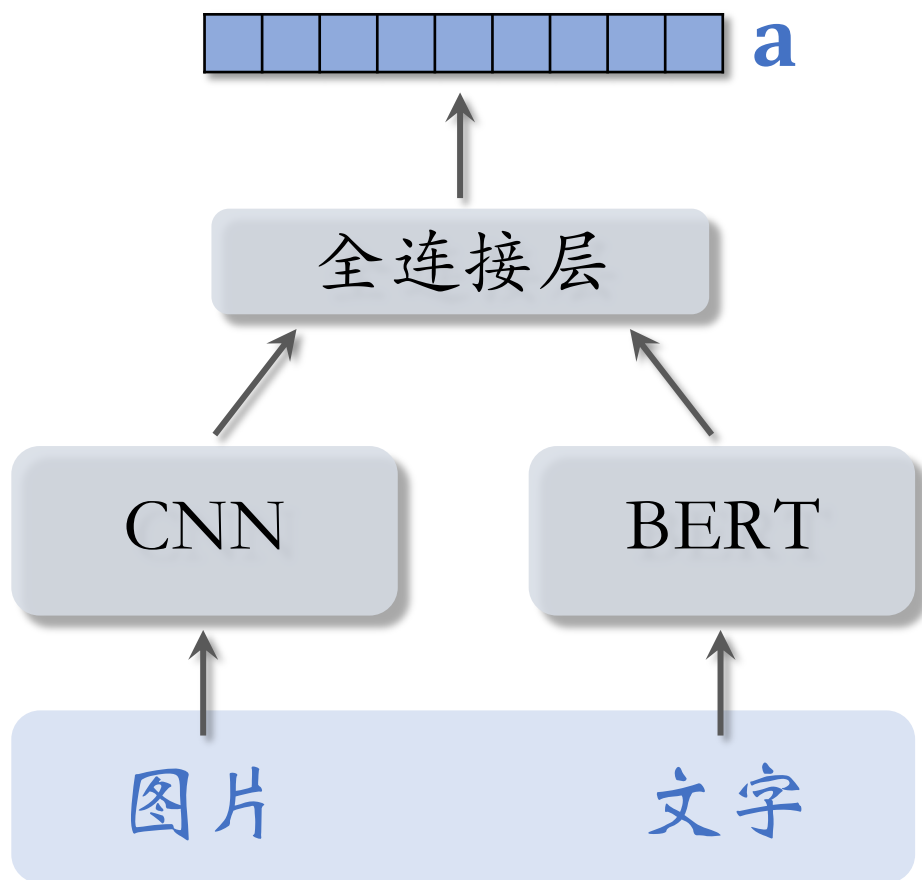
提取图文特征



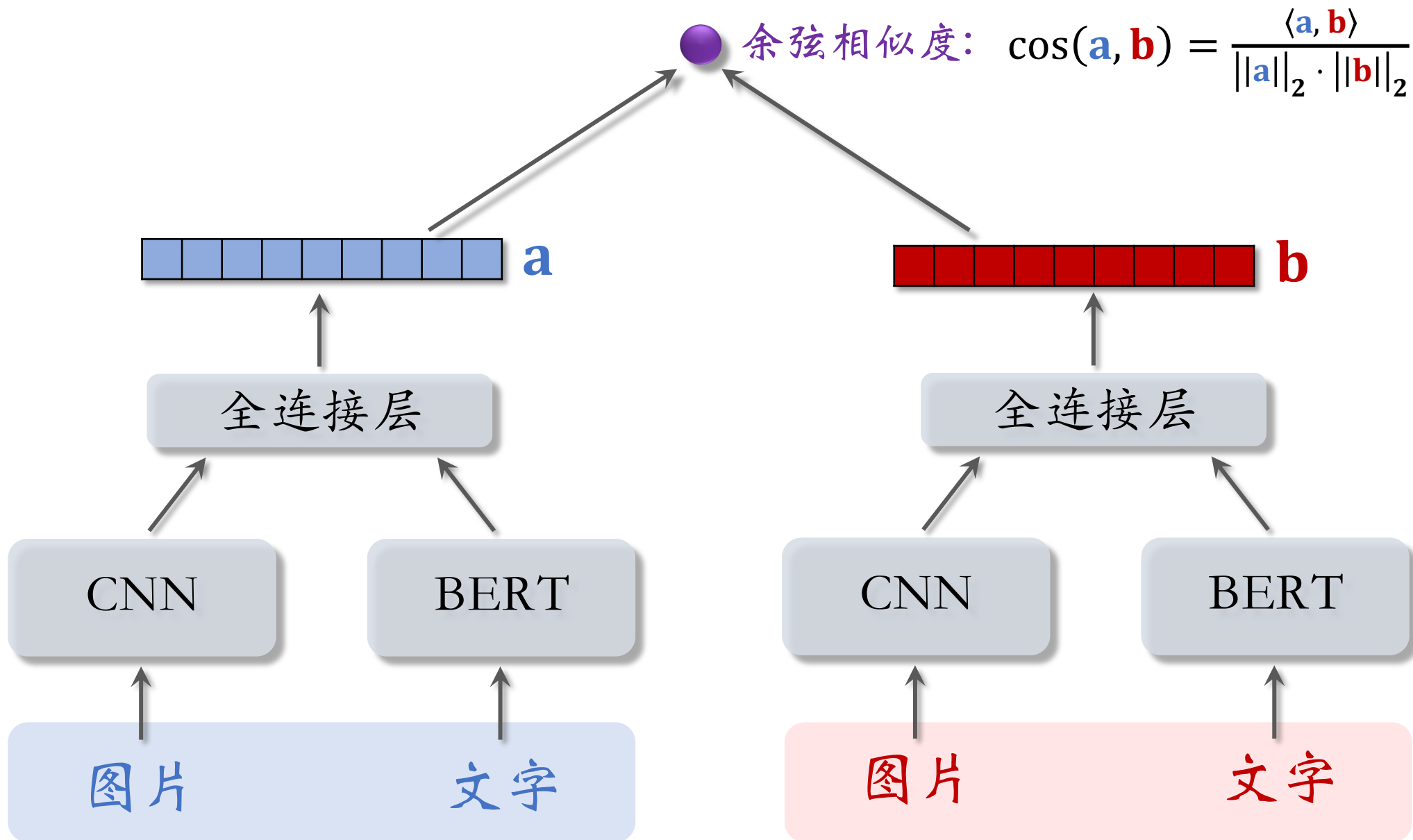
在日本买柴犬最详细解说
我是店长八谷。我是日本人，今年是我在东京开宠物店的第16年。
很多中国客人特意来日本买柴犬。有客人说大阪某宠物店60万日元（约4w人民币）一条柴犬。
日本柴犬真那么贵嘛？怎么可能💣💣
💣
这是八谷为过往客人办理柴犬的费用💰
中国的朋友只看这一篇就可以了解大致的行情📊



两篇笔记内容相似度



两篇笔记内容相似度



训练内容相似度模型

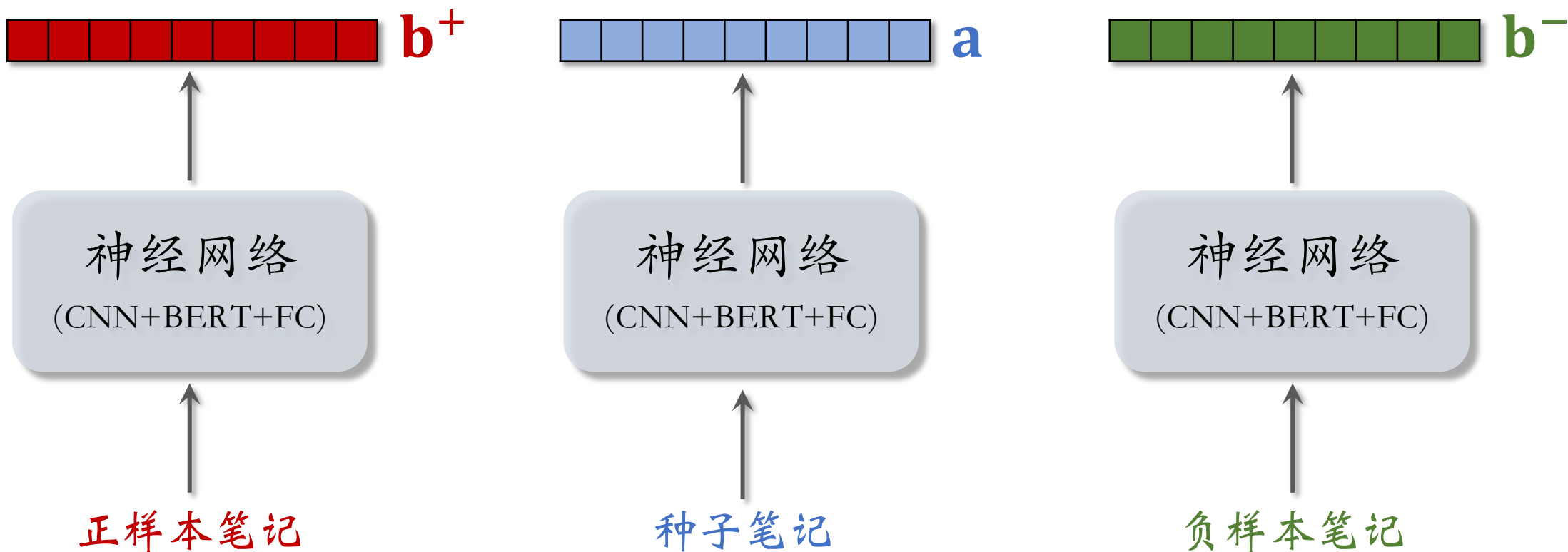
模型的训练

正样本笔记

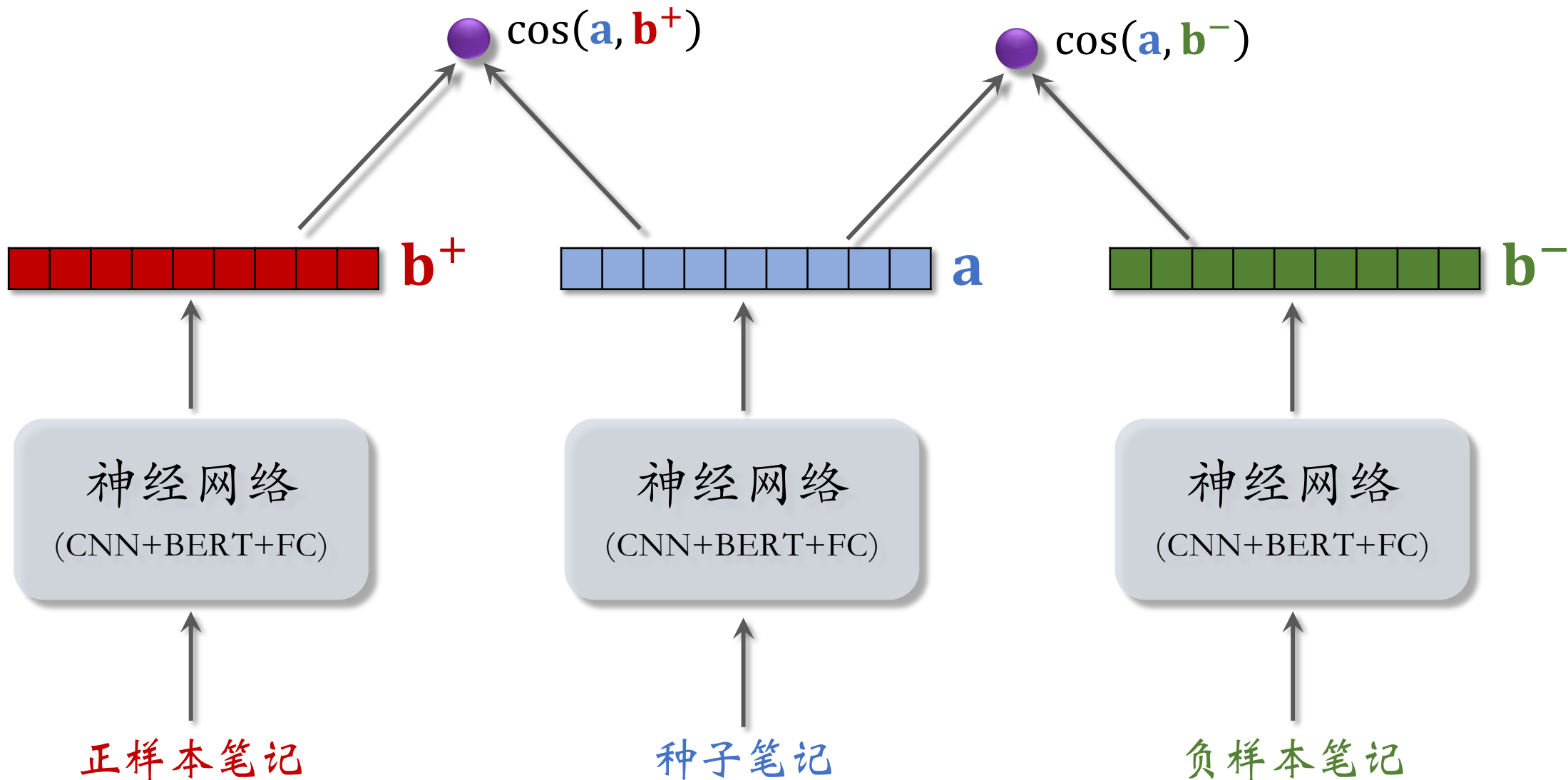
种子笔记

负样本笔记

模型的训练



模型的训练



模型的训练

基本想法：鼓励 $\cos(\mathbf{a}, \mathbf{b}^+)$ 大于 $\cos(\mathbf{a}, \mathbf{b}^-)$

Triplet hinge loss:

$$L(\mathbf{a}, \mathbf{b}^+, \mathbf{b}^-) = \max\{0, \cos(\mathbf{a}, \mathbf{b}^-) + m - \cos(\mathbf{a}, \mathbf{b}^+)\}.$$

Triplet logistic loss:

$$L(\mathbf{a}, \mathbf{b}^+, \mathbf{b}^-) = \log(1 + \exp(\cos(\mathbf{a}, \mathbf{b}^-) - \cos(\mathbf{a}, \mathbf{b}^+))).$$

<种子笔记， 正样本>

方法一：人工标注二元组的相似度

方法二：算法自动选正样本

- 筛选条件：
 - 只用高曝光笔记作为二元组（因为有充足的用户交互信息）。
 - 两篇笔记有相同的二级类目，比如都是“菜谱教程”。
- 用 ItemCF 的物品相似度选正样本。

<种子笔记， 负样本>

- 从全体笔记中随机选出满足条件的：
 - 字数较多（神经网络提取的文本信息有效）。
 - 笔记质量高，避免图文无关。

聚类召回总结

- 基本思想：根据用户的点赞、收藏、转发记录，推荐内容相似的笔记。
- 线下训练：多模态神经网络把图文内容映射到向量。
- 线上服务：

用户喜欢的笔记 → 特征向量 → 最近的Cluster → 新笔记

Thank You!

<http://wangshusen.github.io/>

长期招聘优秀的算法工程师

- 部门：小红书社区技术部。
- 方向：搜索、推荐。
- 职位：校招、社招、实习。
- 地点：上海、北京。
- 联系方式：`ShusenWang@xiaohongshu.com`