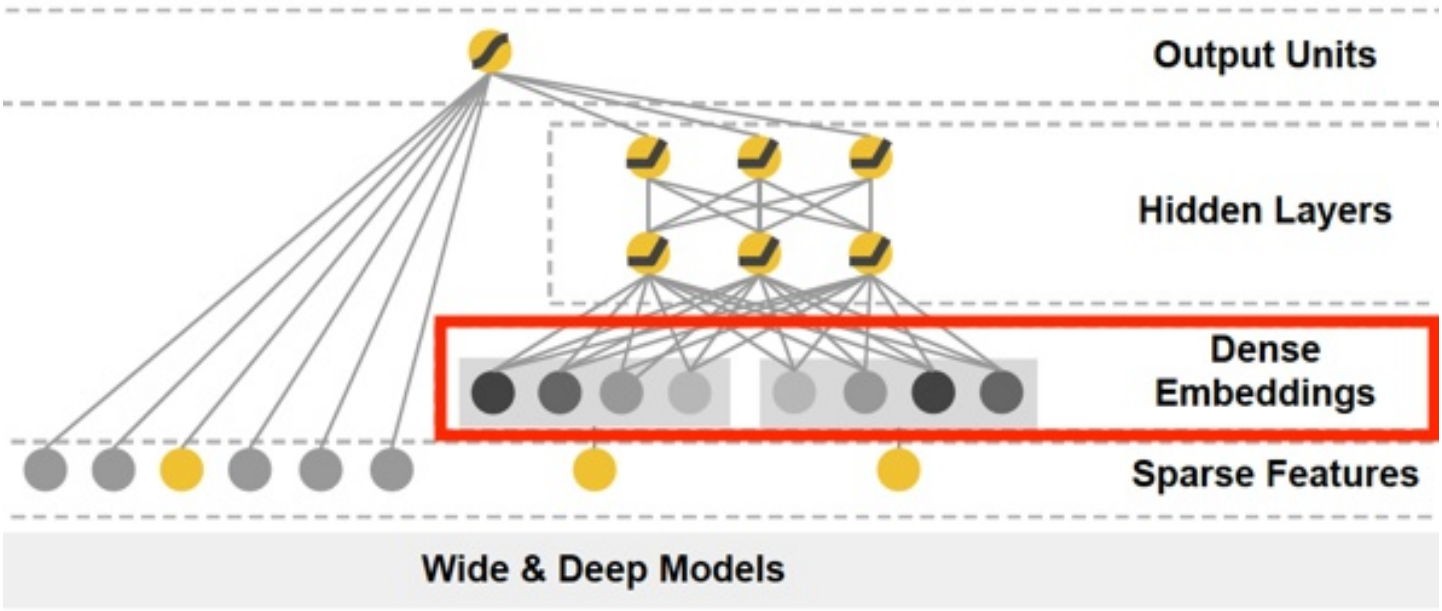
知乎 首发于 王喆的机器学习笔记



Embedding在深度推荐系统中的3大应用方向

王喆
机器学习、深度学习（Deep Learning） 话题的优秀回答者

已关注

张小磊、萧瑟、陈然、到处挖坑蒋玉成等 256 人赞同了该文章

这里是**「王喆的机器学习笔记」**的第十七篇文章，之前已经有多篇专栏文章介绍了从Word2Vec，到Graph Embedding，再到深度学习CTR模型等等Embedding相关的知识。作为深度学习推荐系统模型和CTR模型中不可或缺的“基本操作”，如何强调Embedding技术的重要性都是不为过的。

这篇文章中，我们将Embedding技术单独抽取出来进行讲解。介绍在深度学习推荐系统中，Embedding主要的三个应用方向：

1. **在深度学习网络中作为Embedding层，完成从高维稀疏特征向量到低维稠密特征向量的转换；**
2. **作为预训练的Embedding特征向量，与其他特征向量连接后一同输入深度学习网络进行训练；**
3. **通过计算用户和物品的Embedding相似度，Embedding可以直接作为推荐系统或计算广告系统的召回层或者召回方法之一。**

一、深度学习网络中的Embedding层

由于高维稀疏特征向量天然不适合多层复杂神经网络的训练，因此如果使用深度学习模型处理高维稀疏特征向量，几乎都会在输入层到全连接层之间加入Embedding层完成高维稀疏特征向量到低维稠密特征向量的转换。典型的例子是微软的D
深度部分。

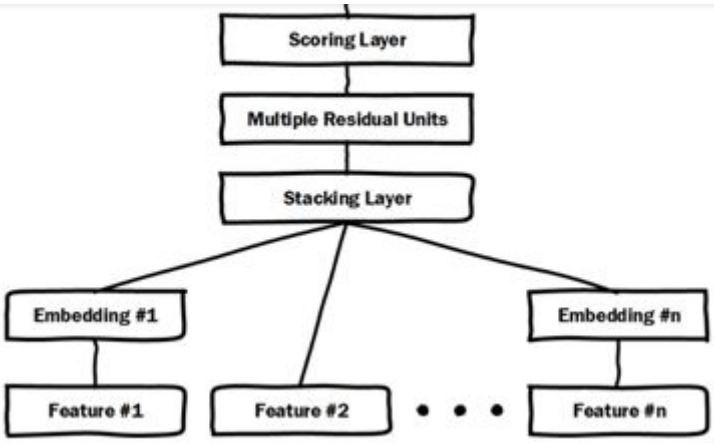


图1 微软Deep Crossing模型

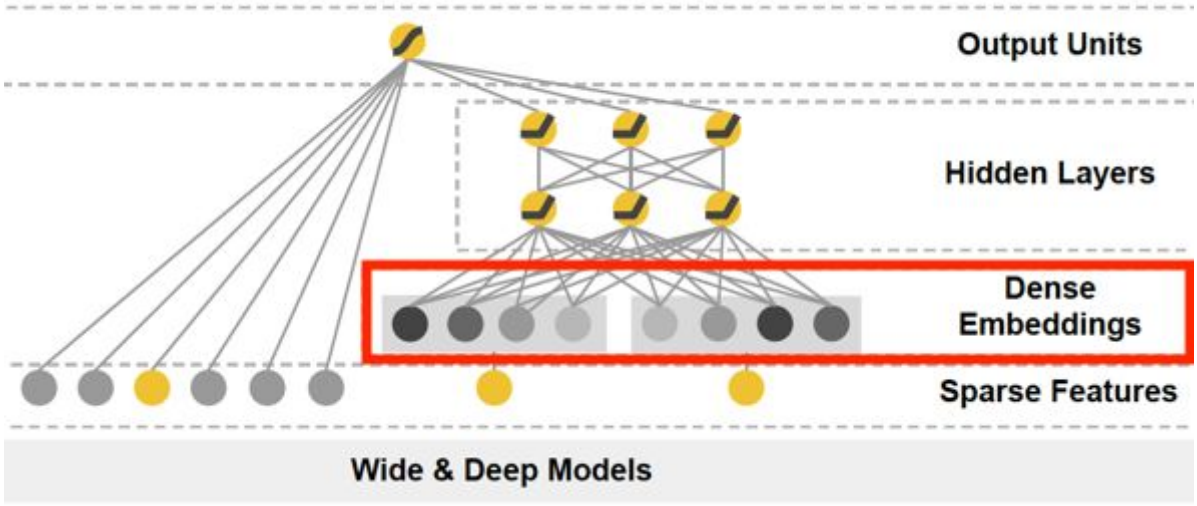
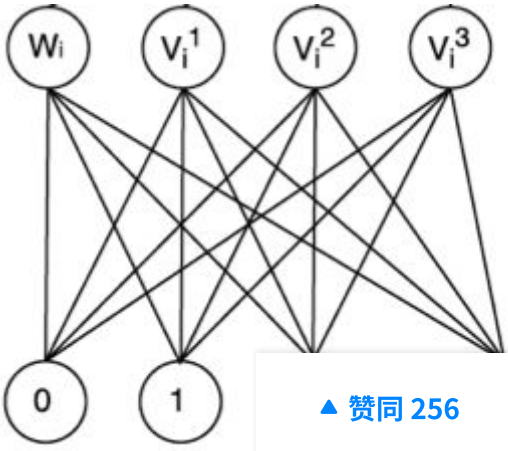


图2 Google Wide Deep模型

图1中可以清晰的看到Deep Crossing模型中的Embedding层将每一个Feature转换成稠密向量，图2Wide&Deep模型中Deep部分的Dense Embeddings层同样将稀疏特征向量进行转换。广义来说，Embedding层的结构可以比较复杂，只要完成高维向量的降维就可以了，但一般为了节省训练时间，深度神经网络中的Embedding层是一个高维向量向低维向量的直接映射（如图3）。



一般来说，推荐系统的输入向量中包含大量稀疏的one hot特征，图3展示了典型的稀疏向量向稠密embedding向量的最简单的embedding层结构。

用矩阵的形式表达Embedding层，本质上是求解一个m（输入高维稀疏向量的维度） x n（输出稠密向量的维度）维的权重矩阵的过程。如果**输入向量是one-hot特征向量的话，权重矩阵中的列向量即为相应维度one-hot特征的embedding向量**。

将Embedding层与整个深度学习网络整合后一同进行训练是理论上最优的选择，因为上层梯度可以直接反向传播到输入层，模型整体是自洽和统一的。但这样做的缺点同样显而易见的，由于Embedding层输入向量的维度甚大，Embedding层的加入会拖慢整个神经网络的收敛速度。

> 这里可以做一个简单的计算。假设输入层维度是100,000，embedding输出维度是32，上层再加5层32维的全连接层，最后输出层维度是10，那么输出层到embedding层的参数数量是32*100,000= 3,200,000，其余所有层的参数总数是 （32*32）*4+32*10=4416。那么embedding层的权重总数占比是 3,200,000 / (3,200,000 + 4416) = 99.86%。

也就是说embedding层的权重占据了整个网络权重的绝大部分。那么训练过程可想而知，大部分的训练时间和计算开销都被Embedding层所占据。正因为这个原因，Embedding层往往采用预训练的方式完成。

二、Embedding的预训练方法

通过上面对Embedding层的介绍，同学们肯定已经知道Embedding层的训练开销是巨大的。为了解决这个问题，Embedding的训练往往独立于深度学习网络进行。在得到稀疏特征的稠密表达之后，再与其他特征一起输入神经网络进行训练。典型的采用Embedding预训练方法的模型是FNN（如图4）。

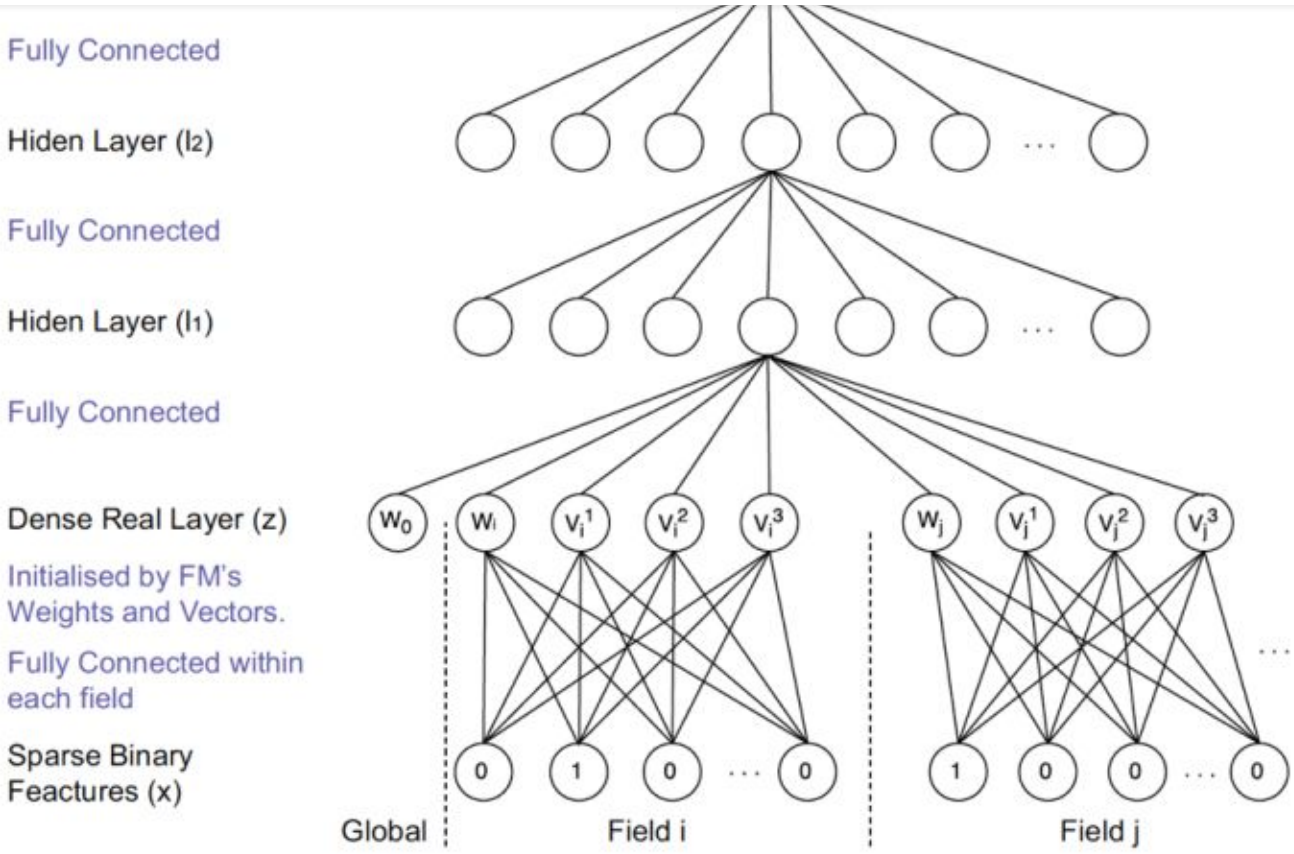


图4 FNN模型结构

FNN利用了FM训练得到的物品向量，作为Embedding层的初始化权重，从而加快了整个网络的收敛速度。在实际工程中，直接采用FM的物品向量作为Embedding特征向量输入到后续深度学习网络也是可行的办法。

再延伸一点讲，Embedding的本质是建立高维向量到低维向量的映射，而“映射”的方法并不局限于神经网络，实质上可以是任何异构模型，这也是Embedding预训练的另一大优势，就是可以采用任何传统降维方法，机器学习模型，深度学习网络完成embedding的生成。

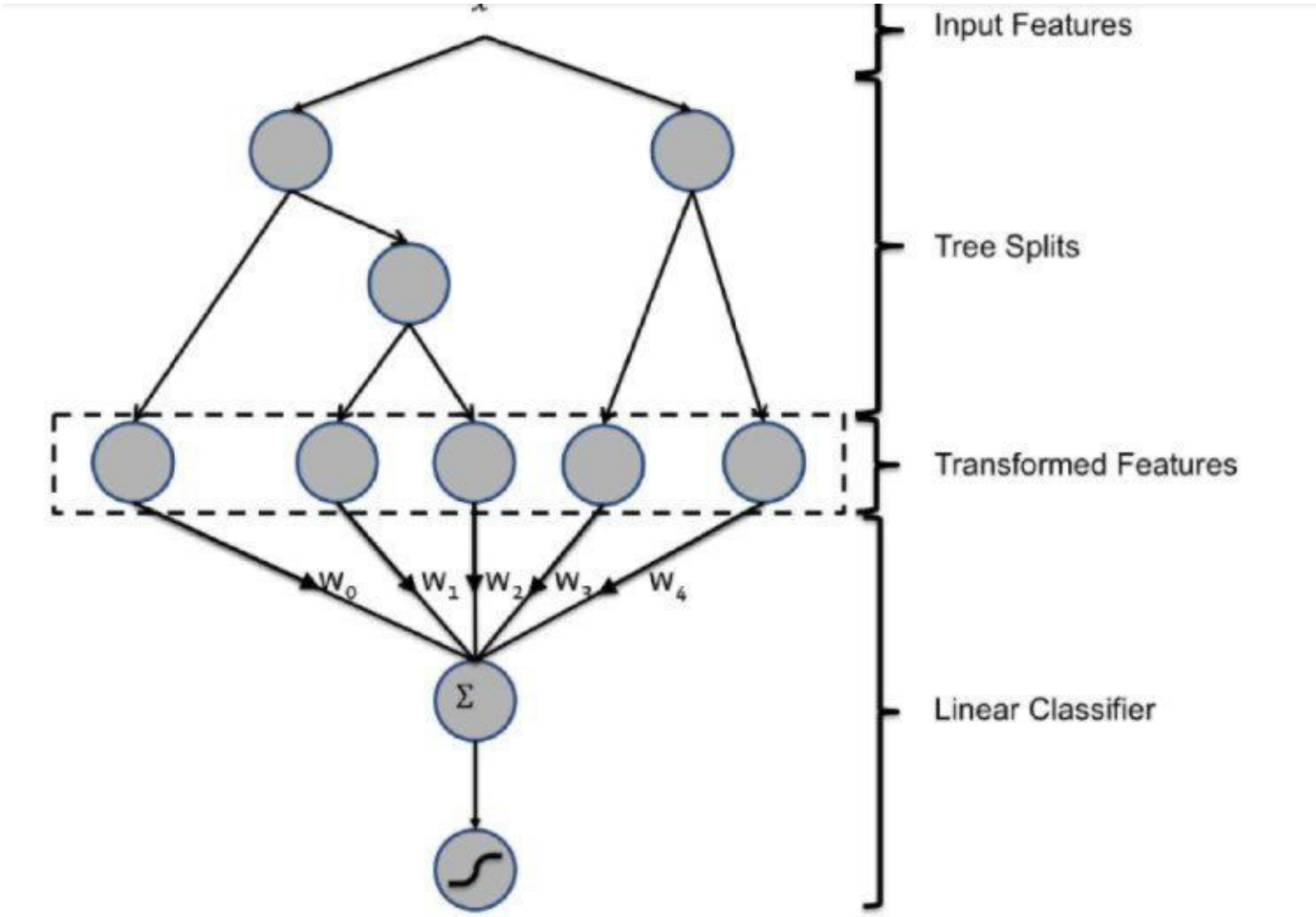


图5 GBDT+LR模型 GBDT完成Embedding过程

典型的例子是2013年Facebook提出的著名的GBDT+LR的模型，其中GBDT的部分本质上也是完成了一次特征转换，可以看作是利用GBDT模型完成Embedding预训练之后，将Embedding输入单层神经网络进行CTR预估的过程。

2015年以来，随着大量Graph Embedding技术的发展，Embedding本身的表达能力进一步增强，而且能够将各类特征全部融合进Embedding之中，这使Embedding本身成为非常有价值的特征。这些特点都使Embedding预训练成为更被青睐的技术途径。

诚然，将Embedding过程与深度网络的训练过程割裂，必然会损失一定的信息，但训练过程的独立也带来了训练灵活性的提升。举例来说，由于物品或用户的Embedding天然是比较稳定的（因为用户的兴趣、物品的属性不可能在几天内发生巨大的变化），Embedding的训练频率其实不需要很高，甚至可以降低到周的级别，但上层神经网络为了尽快抓住最新的正样本信息，往往需要高频训练甚至实时训练。使用不同的训练频率更新Embedding模型和神经网络模型，是训练开销和模型效果二者之间权衡后的最优方案。

三、Embedding作为推荐系统或计算广

召回层（如图6）的解决方案是典型的做法。

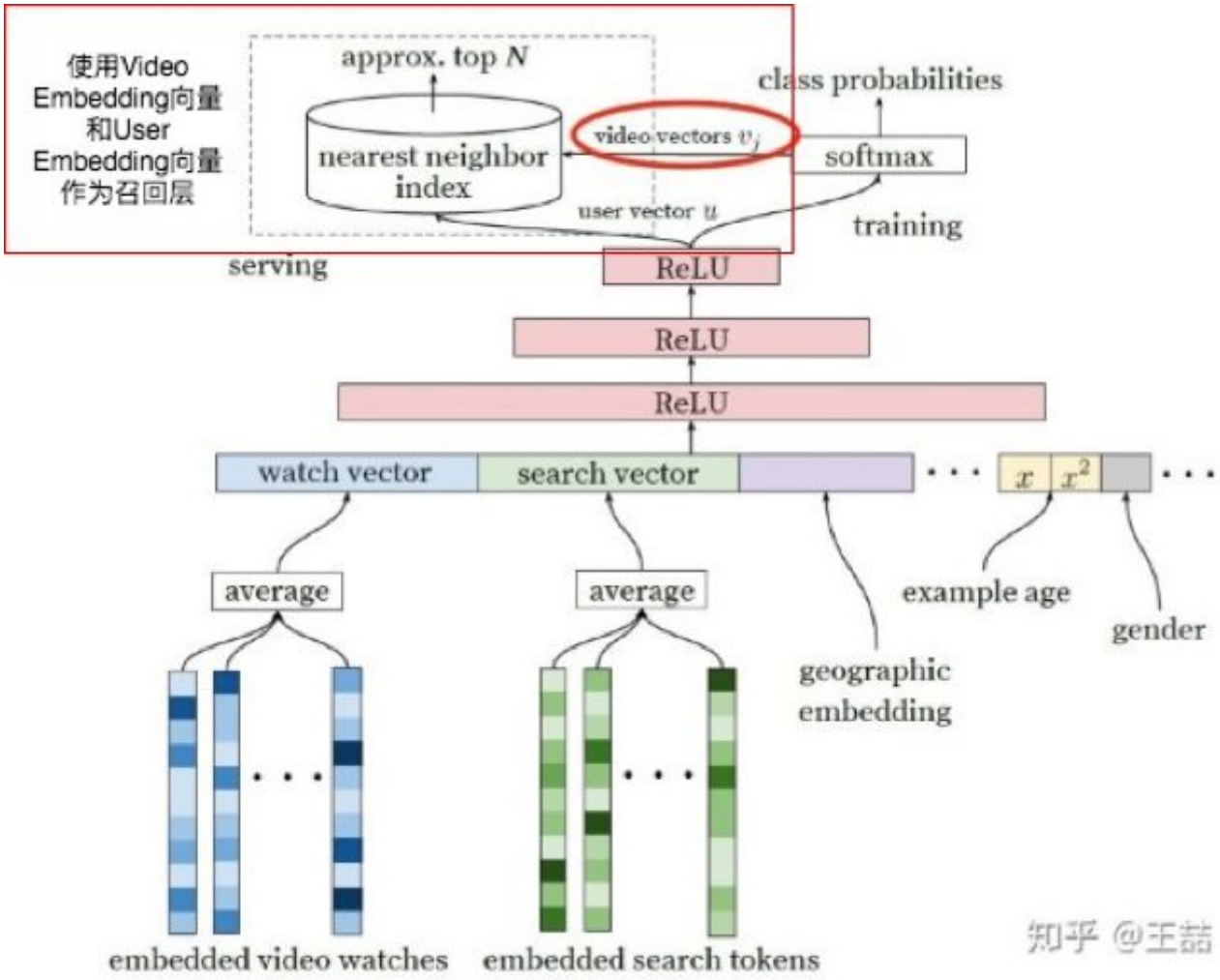


图6 Youtube采用Embedding作为推荐系统召回层

我曾经在文章《重读Youtube深度学习推荐系统论文，字字珠玑，惊为神文》中介绍过了Youtube利用深度学习网络生成Video Embedding和User Embedding的方法。利用最终的Softmax层的权重矩阵，每个Video对应的列向量就是其Item Embedding，而Softmax前一层的输出就是User Embedding。在模型部署过程中，没有必要部署整个深度学习网络来完成从原始特征向量到最终输出的预测过程，只需要将User Embedding和Item Embedding存储到线上内存数据库，通过内积运算再排序的方法就可以得到item的排名。这大大加快了召回层的召回效率。

事实上，除了上述的三种主要的Embedding应用方向，业界对于Embedding的创新性研究不仅没有停止，而且有愈演愈烈之势，阿里的EGES，Pinterest的GNN应用，Airbnb基于Embedding的搜索模型等大量表达能力非常强的Embedding方法的诞生，使Embedding本身就已经成为了优秀的CTR模型和推荐系统模型。作为计算广告和推荐系统领域的从业者，无论如何强调Embedding的重要性都不过分，也希望今后能与大家继续分享Embedding领域的前沿知识。

文章最后按惯例跟大家讨论两个问题，希望大家能分享自己的观点，讨论出真知：

的Embedding向量？大家有什么实践经验分享吗？

最后欢迎大家关注我的**微信公众号：王喆的机器学习笔记（wangzhenotes）**，跟踪计算广告、推荐系统等机器学习领域前沿。

想进一步交流的同学也可以通过公众号加我的微信一同探讨技术问题！

想查阅之前Embedding相关文章的同学也可以直接在公众号里输入关键词查询：

- 关键词: **word2vec** "万物皆Embedding，从经典的word2vec到深度学习基本操作item2vec"
- 关键词：**graph embedding** "深度学习中不得不学的Graph Embedding方法"
- 关键词：**embedding paper** “Embedding从入门到专家必读的十篇论文”
- 关键词：**airbnb embedding** “Airbnb实时搜索排序中的Embedding技巧”

编辑于 2019-05-28

深度学习（Deep Learning） 机器学习 推荐系统

文章被以下专栏收录