**Wide & Deep Learning for Recommender Systems**

Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra,

Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil,

Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, Hemal Shah

∗ Google Inc.

# ABSTRACT(摘要)

Generalized linear models with nonlinear feature transformations are widely used for large-scale regression and classification problems with sparse inputs. Memorization of feature interactions through a *wide* set of cross-product feature transformations are effective and interpretable, while generalization requires more feature engineering effort. With less feature engineering, *deep* neural networks can generalize better to unseen feature combinations through low-dimensional dense embeddings learned for the sparse features. However, deep neural networks with embeddings can over-generalize and recommend less relevant items when the user-item interactions are sparse and high-rank. In this paper, we present Wide & Deep learning—jointly trained wide linear models and deep neural networks—to combine the benefits of memorization and generalization for recommender systems. We productionized and evaluated the system on Google Play, a commercial mobile app store with over one billion active users and over one million apps. Online experiment results show that Wide & Deep significantly increased app acquisitions compared with wide-only and deep-only models. We have also open-sourced our implementation in TensorFlow.

arXiv:1606.07792v1 [cs.LG] 24 Jun 2016

广义线性模型和非线性变换广泛用于稀疏输入的大规模回归和分类问题上。通过广泛的向量积特征变换来**记忆**特征相互影响是有效和可解释的，而**泛化**需要更多的特征工程。对于较少的特征工程，深层神经网络可以为稀疏特征通过低维密集嵌入归纳学习更好的未知特征组合。然而，当user和item交互稀疏和高排名时，具有嵌入的深层神经网络可能会过度泛化并推荐较少的相关项目。

在这篇论文中，我们提出广深学习训练广泛线性模型和深度神经网络，为推荐系统结合**记忆**和**泛化**的好处。我们在Google play上实践和评估了该系统，Google Play是一个商业移动应用商店，拥有超过10亿活跃用户和超过一百万个应用。在线实验结果表明，广深结合和单纯的广度以及单纯深度模型相比，显著提高了app购买。我们也会开源我们的实践在TensorFlow上。

# CCS Concepts（CCS概念）

•**Computing methodologies->Machine learning;** *Neural networks; Supervised learning;* •**Information systems** –> *Recommender systems;*

# Keywords（关键词）

Wide & Deep Learning, Recommender Systems.

# INTRODUCTION（介绍）

A recommender system can be viewed as a search ranking system, where the input query is a set of user and contextual information, and the output is a ranked list of items. Given a query, the recommendation task is to find the relevant items in a database and then rank the items based on certain objectives, such as clicks or purchases.

推荐系统可以被视为搜索排名系统，其中输入query是一组user和上下文信息，并且输出是排列的item列表。给出query，推荐任务是在数据库中查找相关item，然后根据某些目标（如点击或购买）对item进行排名。

One challenge in recommender systems, similar to the general search ranking problem, is to achieve both *memorization* and *generalization*. Memorization can be loosely defined as learning the frequent co-occurrence of items or features and exploiting the correlation available in the historical data. Generalization, on the other hand, is based on transitivity of correlation and explores new feature combinations that have never or rarely occurred in the past. Recommendations based on memorization are usually more topical and directly relevant to the items on which users have already performed actions. Compared with memorization, generalization tends to improve the diversity of the recommended items. In this paper, we focus on the apps recommendation problem for the Google Play store, but the approach should apply to generic recommender systems.

推荐系统中一个挑战是类似传统搜索排名问题，就是实现**记忆**和**泛化**。

**记忆**可以被宽松的定义为：在历史数据中学习频繁同现的item或特征和可利用的相关性。

另一方面来说，**泛化**是基于相关性传递和探索未发生和很少发生的新特征组合。推荐基于记忆通常更局部和在user已经执行的操作直接相关。与**记忆**相比，**泛化**倾向于改善推荐项目的多样性。在本文中，我们专注于Google Play商店的应用推荐问题，但该方法应适用于通用推荐系统。

For massive-scale online recommendation and ranking systems in an industrial setting, generalized linear models such as logistic regression are widely used because they are simple, scalable and interpretable. The models are often trained on binarized sparse features with one-hot encoding. E.g., the binary feature “ user\_installed\_app=netflix” has value 1 if the user installed Netflix. Memorization can be achieved effectively using cross-product transformations over sparse features, such as AND(user\_installed\_app=netflix, impression\_app=pandora”), whose value is 1 if the user installed Netflix and then is later shown Pandora. This explains how the co-occurrence of a feature pair correlates with the target label. Generalization can be added by using features that are less granular, such as AND(user\_installed\_category=video, impression\_category=music), but manual feature engineering is often required. One limitation of cross-product transformations is that they do not generalize to query-item feature pairs that have not appeared in the training data.

对于工业环境中的大规模在线推荐和排名系统，诸如逻辑回归之类的广义线性模型被广泛使用，因为它们简单、可扩展、可解释。这些模型经常用二进制稀疏特征进行One-Hot编码训练。例如，如果用户安装了Netflix，二进制功能“user\_installed\_app = netflix”的值为1。

**记忆**可以通过向量积变换之上的稀疏特征（如AND, user\_installed\_app=netflix, impression\_app=pandora”）有效的实现，如果用户安装了Netflix，然后稍后显示Pandora，其值为1。这说明了特征对共现和目标标签的相关性。

**泛化**可以通过使用较少粒度的特征添加进来。（如 AND, user\_installed\_category=video, impression\_category=music）,但是需要手工特征工程。向量积变换的一个限制是它们不能泛化到没有在训练集上出现的query-item特征对。

Embedding-based models, such as factorization machines [5] or deep neural networks, can generalize to previously unseen query-item feature pairs by learning a low-dimensional dense embedding vector for each query and item feature, with less burden of feature engineering. However, it is difficult to learn effective low-dimensional representations for queries and items when the underlying query-item matrix is sparse and high-rank, such as users with specific preferences or niche items with a narrow appeal. In such cases, there should be no interactions between most query-item pairs, but dense embeddings will lead to nonzero predictions for all query-item pairs, and thus can over-generalize and make less relevant recommendations. On the other hand, linear models with cross-product feature transformations can memorize these “exception rules” with much fewer parameters.

基于嵌入的模型，例如因式分解机或深度神经网络，可以通过为每个query-item特征对学习低维密集嵌入向量，用较少的特征工程推广到以前未知的query-item特征对。

然而，当底层query-item的矩阵稀疏和高评分时候，例如有特殊偏爱或冷门的小众的项目的用户，就难以为query们和item们学习有效的低维表示。在这种情况下，大多数query-item对之间不应该有相互影响，但是密集嵌入将导致为所有的query-item对预测非零，因此可能会过度泛化和减少相应推荐建议。另一方面，具有向量积特征变换的线性模型可以将这些“例外规则”**记忆**在更少的参数中。

In this paper, we present the Wide & Deep learning framework to achieve both memorization and generalization in one model, by jointly training a linear model component and a neural network component as shown in Figure 1. The main contributions of the paper include:

|  |
| --- |
| Wide Models  Deep Models  Wide & Deep Models  Hidden Layers （隐藏层）  Sparse Features （稀疏特征）  Output Units （输出单元）  Dense  Embeddings (密集嵌入)  **图 1: The spectrum of Wide & Deep models.（广深模型范围图）** |

* The Wide & Deep learning framework for jointly training feed-forward neural networks with embeddings and linear model with feature transformations for generic recommender systems with sparse inputs.
* The implementation and evaluation of the Wide & Deep recommender system productionized on Google Play, a mobile app store with over one billion active users and over one million apps.
* We have open-sourced our implementation along with a high-level API in TensorFlow[[1]](#footnote-1).

在本文中，我们将通过广深学习框架在一个模型中**记忆**和**泛化**两者同时实现，通过共同训练线性模型组件和神经网路组件，如图1所示。本文的主要贡献包括：

* 广深学习框架用于联合训练具有嵌入的前馈神经网络和具有稀疏输入的通用推荐系统的特征变换线性模型。
* 在Google Play上生产的广深推荐系统的实施和评估，这款商店具有超过10亿活跃用户和超过一百万个应用。
* 我们在TensorFlow中开源了我们实现的高级API。

While the idea is simple, we show that the Wide & Deep framework significantly improves the app acquisition rate on the mobile app store, while satisfying the training and serving speed requirements.

虽然这个想法很简单，但我们展示了广深框架显著提高了移动应用商店的应用获取率，同时满足了训练和服务速度要求。

# RECOMMENDER SYSTEM OVERVIEW（推荐系统概述）

An overview of the app recommender system is shown in Figure 2. A query, which can include various user and contextual features, is generated when a user visits the app store. The recommender system returns a list of apps (also referred to as impressions) on which users can perform certain actions such as clicks or purchases. These user actions, along with the queries and impressions, are recorded in the logs as the training data for the learner.

Item 1

Item 2

Item 3

...

Database

Query

Items

Learner

Model

Ranked

O(10) items

Logs

User Actions

Retrieval(检索)

O(100) items

Ranking

Recommendation System

All items

**Figure 2: Overview of the recommender system.**

应用程序推荐系统的概述如图2所示。当用户访问应用商店时，会生成一个包含各种用户和上下文功能的query。推荐系统会返回用户可以执行某些操作（例如点击或购买）的应用列表（也称为印象）。 这些用户操作以及query和印象都记录在日志中作为学习者的训练数据。

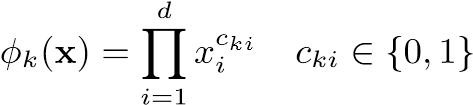
Since there are over a million apps in the database, it is intractable to exhaustively score every app for every query within the serving latency requirements (often *O*(10) milliseconds). Therefore, the first step upon receiving a query is *retrieval*. The retrieval system returns a short list of items that best match the query using various signals, usually a combination of machine-learned models and human-defined rules. After reducing the candidate pool, the *ranking* system ranks all items by their scores. The scores are usually *P*(*y*|**x**), the probability of a user action label *y* given the features **x**, including user features (e.g., country, language, demographics), contextual features (e.g., device, hour of the day, day of the week), and impression features (e.g., app age, historical statistics of an app). In this paper, we focus on the ranking model using the Wide & Deep learning framework.

由于数据库中有超过一百万个应用程序，因此在服务延迟要求（通常为O（10）毫秒）内，为每个query全面评分每个应用程序是不切实际的。因此，收到query的第一步是检索。 检索系统使用各种信号（通常是机器学习模型和人为定义的规则的组合）返回与query最匹配的item的简短列表。在减少候选池后，排名系统按其分数排列所有item。分数通常使用P（y|x），求用户动作标签y的概率在给定特征x下，包括用户特征（例如，国家、语言、人口统计），上下文特征（例如，设备、一天的哪个小时、一周的哪天 ），以及印象特征（例如app年龄，app历史统计资料）。 在本文中，我们将重点放在使用广深学习框架的排名模型中。

# WIDE & DEEP LEARNING（广深学习）

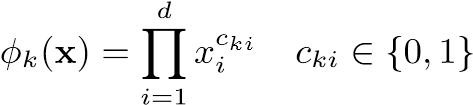
## The Wide Component（广度组件）

The wide component is a generalized linear model of the form *y* = **w***T***x** + *b*, as illustrated in Figure 1 (left). *y* is the prediction, **x** = [*x*1*,x*2*,...,xd*] is a vector of *d* features, **w** = [*w*1*,w*2*,...,wd*] are the model parameters and *b* is the bias. The feature set includes raw input features and transformed features. One of the most important transformations is the *cross-product transformation*, which is defined as:

 (1)

where *cki* is a boolean variable that is 1 if the *i*-th feature is part of the *k*-th transformation *φk*, and 0 otherwise. For binary features, a cross-product transformation ( e.g., “AND(gender=female, language=en)”) is 1 if and only if the constituent features (“gender=female” and “language=en”) are all 1, and 0 otherwise. This captures the interactions between the binary features, and adds nonlinearity to the generalized linear model.

广度组件是一个形式为*y* = **w***T***x** + *b*广义线性模型,如图一左。y表示预测，x=[*x*1*,x*2*,...,xd*]是一个特征向量，**w** = [*w*1*,w*2*,...,wd*]是模型的参数，b是偏置项。特征集包括原始输入特征和变换特征。最重要的变换是向量积变换，它被定义为：

 (1)

其中*cki*是一个布尔变量，如果第i个特征是第k个变换的一部分，则为1，反之0。对于二进制特征，当且仅当构成特征（“gender = female”和“language = en”）全为1时，向量积变换（例如“AND（gender = female，language = en）”）为1，否则为0。这捕获了二元特征之间的相互作用，并且增加了非线性为广义线性模型。

## The Deep Component（深度组件）

The deep component is a feed-forward neural network, as shown in Figure 1 (right). For categorical features, the original inputs are feature strings (e.g., “language=en”). Each of these sparse, high-dimensional categorical features are first converted into a low-dimensional and dense real-valued vector, often referred to as an embedding vector. The dimensionality of the embeddings are usually on the order of *O*(10) to *O*(100). The embedding vectors are initialized randomly and then the values are trained to minimize the final loss function during model training. These low-dimensional dense embedding vectors are then fed into the hidden layers of a neural network in the forward pass. Specifically, each hidden layer performs the following computation:

*a*(*l*+1) = *f*(*W*(*l*)*a*(*l*) + *b*(*l*)) (2)

where *l* is the layer number and *f* is the activation function, often rectified linear units (ReLUs). *a*(*l*), *b*(*l*), and *W*(*l*) are the activations, bias, and model weights at *l*-th layer.

深度组件是前馈神经网络，如图1右所示。 对于分类特征，原始输入是特征字符串（例如，“language = en”）。这些每一个稀疏、高维度分类特征首先被转换成低维和密集的实值向量，通常称为嵌入向量。嵌入的维数通常为O（10）到O（100）的数量级。嵌入向量被随机初始化，然后训练这些值用以最小化模型的最终损失函数。然后将这些低维密集嵌入向量在正向传播中反馈到神经网络的隐藏层中。具体来说，每个隐藏层执行以下计算：

*a*(*l*+1) = *f*(*W*(*l*)*a*(*l*) + *b*(*l*)) (2)

其中*l*是层数，*f*是激活函数，通常使用ReLUs。*a*(*l*), *b*(*l*)和*W*(*l*)分别是激活函数、偏置项和模型权重在第*l*层。

## Joint Training of Wide & Deep Model（广深联合训练模型）

The wide component and deep component are combined using a weighted sum of their output log odds as the prediction, which is then fed to one common logistic loss function for joint training. Note that there is a distinction between *joint training* and *ensemble*. In an ensemble, individual models are trained separately without knowing each other, and their predictions are combined only at inference time but not at training time. In contrast, joint training optimizes all parameters simultaneously by taking both the wide and deep part as well as the weights of their sum into account at training time. There are implications on model size too: For an ensemble, since the training is disjoint, each individual model size usually needs to be larger (e.g., with more features and transformations) to achieve reasonable accuracy for an ensemble to work. In comparison, for joint training the wide part only needs to complement the weaknesses of the deep part with a small number of cross-product feature transformations, rather than a full-size wide model. Joint training of a Wide & Deep Model is done by backpropagating the gradients from the output to both the wide and deep part of the model simultaneously using mini-batch stochastic optimization. In the experiments, we used Followthe-regularized-leader (FTRL) algorithm [3] with *L*1 regularization as the optimizer for the wide part of the model, and AdaGrad [1] for the deep part.

广度组件和深度组件组合使用加权和，输出记录几率作为预测，让其给共同逻辑的损失函数联合训练。请注意，联合训练和合奏有区别。在合奏中，单独的模型是分开训练而不知道彼此，而他们的预测仅在推理时间而不是训练时候组合。相比之下，联合训练通过在训练时同时考虑广、深部分以及其它们权重和优化所有参数。

对模型大小也有影响：对于合奏，由于训练不相交，因此每个单独的模型通常更大（例如更多的特征和变换）来实现合奏的合理和准确性。相比之下，对于联合训练，广度部分只需要通过少量的向量积特征变换补充深度部分短处，而不是全量模型。

通过小批次随机优化、从广深两部分模型输出反向传播梯度进行联合训练广深模型。在试验中，我们使用FTRL（Followthe-regularized-leader）算法和L1正则化作为广度模型优化器和深度部分AdaGrad。

The combined model is illustrated in Figure 1 ( center ). For a logistic regression problem, the model’s prediction is:

) (3)

where *Y* is the binary class label, *σ*(·) is the sigmoid function, *φ*(**x**) are the cross product transformations of the original features **x**, and *b* is the bias term. **w***wide* is the vector of all wide model weights, and **w***deep* are the weights applied on the final activations *a*(*lf*).

图1中间为组合模型。对于逻辑回归，模型的预测是:

 (3)

其中Y是二进制类标签，σ(·)是sigmoid函数，*φ*(**x**)是元素特征X的变换向量积，b是偏置项，**w***wide*是广度模型权重向量，**w***deep*是应用最终激活*a*(*lf*)的权重。

# SYSTEM IMPLEMENTATION（系统实现）

The implementation of the apps recommendation pipeline consists of three stages: data generation, model training, and model serving as shown in Figure 3.

应用程序推荐流程的实现包括三个阶段：数据生成，模型训练和模型服务，如图3所示。

User Data

App

Impression

Data

Training Data

Generation

Vocabulary

Generator

Model

Trainer

Model Verifier

Model Servers

Model Serving

Apps

Recommendation

Engine

Previous Models

Data Generation

Model Training

**Figure**

**3:**

**Apps**

**recommendation**

**pipeline**

**overview.**

## Data Generation（数据生成）

In this stage, user and app impression data within a period of time are used to generate training data. Each example corresponds to one impression. The label is *app acquisition*: 1 if the impressed app was installed, and 0 otherwise.

Vocabularies, which are tables mapping categorical feature strings to integer IDs, are also generated in this stage. The system computes the ID space for all the string features that occurred more than a minimum number of times. Continuous real-valued features are normalized to [0*,*1] by mapping a feature value *x* to its cumulative distribution function *P*(*X* ≤ *x*), divided into *nq* quantiles. The normalized value i for values in the *i*-th quantiles. Quantile boundaries are computed during data generation.

在这个阶段，使用一段时间内的用户和app印象数据来生成训练数据。 每个示例对应于一个印象。 该标签是app获取：如果印象的app已安装，则为1，否则为0。

在这个阶段也是生成映射分类特征字符串到整数ID的表。 系统计算所有发生超过最小次数的字符串特征的ID空间。 通过将特征值x映射到其累积分布函数*P*(*X* ≤ *x*)，将连续实值特征归一化为[0*,*1]，分为*nq*个分位数。 归一化值 是针对第i个分位数中的值。 在数据生成期间计算分位数边界。

## Model Training(模型训练)

The model structure we used in the experiment is shown in Figure 4. During training, our input layer takes in training data and vocabularies and generate sparse and dense features together with a label. The wide component consists of the cross-product transformation of user installed apps and impression apps. For the deep part of the model, A 32dimensional embedding vector is learned for each categorical feature. We concatenate all the embeddings together with the dense features, resulting in a dense vector of approximately 1200 dimensions. The concatenated vector is then fed into 3 ReLU layers, and finally the logistic output unit.

ReLU (1024)

Logistic Loss

Embeddings

ReLU (512)

ReLU (256)

User Installed

App

Impression

App

User

Demographics

Device

Class

...

Age

#App

Installs

#Engagement

sessions

...

Cross Product

Transformation

Embeddings

Embeddings

Embeddings

Concatenated Embeddings (~1200 dimensions)

Continuous Features

Categorical Features

**图 4: apps推荐的广深模型结构**

我们在实验中使用的模型结构如图4所示。在训练期间，我们的输入层接收训练数据和词汇，并产生稀疏和密集的特征以及标签。广度组件由用户安装的app和印象app向量积变换构成。模型的深度部分为每个分类特征学习32维嵌入向量。我们将所有的嵌入和密集特征连接一起，产生大约1200维的密集向量。然后将通过3层ReLU层，最后逻辑输出链接向量。

The Wide & Deep models are trained on over 500 billion examples. Every time a new set of training data arrives, the model needs to be re-trained. However, retraining from scratch every time is computationally expensive and delays the time from data arrival to serving an updated model. To tackle this challenge, we implemented a warm-starting system which initializes a new model with the embeddings and the linear model weights from the previous model.

广深模型在超过5000亿例子进行训练。每当一组新的训练数据到达，模型将需要重新训练。然而，每次全新训练在计算上是昂贵的，并且延迟了从数据到达到更新服务模型。为了解决这个挑战，我们实现了一个热启动系统，它从以前的模型中用嵌入和线性模型权重初始化一个新的模型。

Before loading the models into the model servers, a dry run of the model is done to make sure that it does not cause problems in serving live traffic. We empirically validate the model quality against the previous model as a sanity check.

在将模型加载到模型服务之前，可以对模型进行干运行，以确保在运行实时流量时不会出现问题。 我们根据先前的模型经验验证模型质量作为理性检查。

## Model Serving（模型服务）

Once the model is trained and verified, we load it into the model servers. For each request, the servers receive a set of app candidates from the app retrieval system and user features to score each app. Then, the apps are ranked from the highest scores to the lowest, and we show the apps to the users in this order. The scores are calculated by running a forward inference pass over the Wide & Deep model.

一旦模型被训练和验证，我们将其加载到模型服务中。 对于每个请求，服务器从应用检索系统和用户功能中收到一组app候选项，以评分每个app。 然后，app的排名从高分到低分，我们按照这个顺序向用户显示应用程序。分数通过在广深模型上运行前向推理计算。

In order to serve each request on the order of 10 ms, we optimized the performance using multithreading parallelism by running smaller batches in parallel, instead of scoring all candidate apps in a single batch inference step.

为了以10ms为每个请求提供服务，我们通过并行运行较小的批次来优化使用多线程并行性的性能，而不是在单个批次推理步骤中对所有候选app进行评分。

# EXPERIMENT RESULTS（实验结果）

To evaluate the effectiveness of Wide & Deep learning in a real-world recommender system, we ran live experiments and evaluated the system in a couple of aspects: app acquisitions and serving performance.

为了评估广泛深度学习在真实世界推荐系统中的有效性，我们进行了实时实验，并在几个方面对系统进行了评估：app购买和服务性能。

## App Acquisitions(app购买)

We conducted live online experiments in an A/B testing framework for 3 weeks. For the control group, 1% of **Table 1: Offline & online metrics of different models.**

我们进行3周线上实验按A/B测试框架中。对于控制组，表一的1%：不同模型的离线和线上矩阵。

**control.**

**Gain**

**Acquisition**

**Online**

**the**

**to**

**relative**

**is**

Model

Offline

AUC

Online

Acquisition

Gain

%

0

0.726

)

control

(

Wide

0.722

+2.9

%

Deep

Wide

&

Deep

0.728

+3.9

%

users were randomly selected and presented with recommendations generated by the previous version of ranking model, which is a highly-optimized wide-only logistic regression model with rich cross-product feature transformations. For the experiment group, 1% of users were presented with recommendations generated by the Wide & Deep model, trained with the same set of features. As shown in Table 1, Wide & Deep model improved the app acquisition rate on the main landing page of the app store by +3.9% relative to the control group (statistically significant). The results were also compared with another 1% group using only the deep part of the model with the same features and neural network structure, and the Wide & Deep mode had +1% gain on top of the deep-only model (statistically significant).

用户随机选择并呈现由先前版本的排名模型生成的建议，该模型是具有丰富的向量积特征变换的高度优化的只有广度逻辑回归模型。对于实验组，1％的用户被呈现了由广深模型生成的建议，训练用相同的功能组。如表1所示，广深模型将应用商店主页面上的app获取率提高了3.9％，相对控制组（统计显著）。

Besides online experiments, we also show the Area Under Receiver Operator Characteristic Curve (AUC) on a holdout set offline. While Wide & Deep has a slightly higher offline AUC, the impact is more significant on online traffic. One possible reason is that the impressions and labels in offline data sets are fixed, whereas the online system can generate new exploratory recommendations by blending generalization with memorization, and learn from new user responses.

除了在线实验，我们还显示了（AUC）在离线展示集。 虽然广深模型的离线AUC稍高，但对在线流量的影响更为显著。 一个可能的原因是离线数据集中的印象和标签是固定的，而在线系统可以通过将泛化与记忆混合来产生新的探索性建议，并从新的用户响应中学习。

## Serving Performance（服务表现）

Serving with high throughput and low latency is challenging with the high level of traffic faced by our commercial mobile app store. At peak traffic, our recommender servers score over 10 million apps per second. With single threading, scoring all candidates in a single batch takes 31 ms. We implemented multithreading and split each batch into smaller sizes, which significantly reduced the client-side latency to 14 ms (including serving overhead) as shown in Table 2.

**threads.**

**size**

**batch**

**vs.**

**latency**

**Serving**

**2:**

**Table**

**and**

Batch

size

Number

)

ms

(

Latency

Serving

Threads

of

31

200

1

100

2

17

50

4

14

服务于大流量和低延迟是我们的商业移动应用商店面临的高水平运输的挑战。 在高峰流量时，我们的推荐服务器每秒可获得超过1000万个应用。 使用单线程，在一个批次中对所有候选人进行评分需要31 ms。 我们实现了多线程，并将每个批处理分成较小的大小，这大大降低了客户端延迟到14 ms（包括服务开销），如表2所示。

# RELATED WORK（相关工作）

The idea of combining wide linear models with cross-product feature transformations and deep neural networks with dense embeddings is inspired by previous work, such as factorization machines [5] which add generalization to linear models by factorizing the interactions between two variables as a dot product between two low-dimensional embedding vectors. In this paper, we expanded the model capacity by learning highly nonlinear interactions between embeddings via neural networks instead of dot products.

将广度线性模型与向量积特征变换和具有密集嵌入的深层神经网络相结合的想法受到以前的工作的启发，如因式分解机[5]，其通过将两个变量之间的相互作用因子分解为两个点之间的点积来增加线性模型的泛化 低维嵌入载体。 在本文中，我们通过学习通过神经网络而不是点产品的嵌入之间的非线性相互作用来扩展模型容量。

In language models, joint training of recurrent neural networks (RNNs) and maximum entropy models with *n*-gram features has been proposed to significantly reduce the RNN complexity (e.g., hidden layer sizes) by learning direct weights between inputs and outputs [4]. In computer vision, deep residual learning [2] has been used to reduce the difficulty of training deeper models and improve accuracy with shortcut connections which skip one or more layers. Joint training of neural networks with graphical models has also been applied to human pose estimation from images [6]. In this work we explored the joint training of feed-forward neural networks and linear models, with direct connections between sparse features and the output unit, for generic recommendation and ranking problems with sparse input data.

在语言模型中，通过学习输入和输出之间的直接权重，已经提出了复现神经网络（RNN）和具有n-gram特征的最大熵模型的联合训练来显着地降低RNN复杂度（例如，隐藏层大小）[4]。 在计算机视觉中，深度残差学习[2]已被用于减少训练更深层次的难度，并通过跳过一层或多层的快捷连接来提高准确性。 神经网络与图形模型的联合训练也被应用于人物姿态估计从图像[6]。 在这项工作中，我们探索了前馈神经网络和线性模型的联合训练，其中稀疏特征与输出单元之间具有直接连接，用于通用推荐和稀疏输入数据的排名问题。

In the recommender systems literature, collaborative deep learning has been explored by coupling deep learning for content information and collaborative filtering (CF) for the ratings matrix [7]. There has also been previous work on mobile app recommender systems, such as AppJoy which used CF on users’ app usage records [8]. Different from the CF-based or content-based approaches in the previous work, we jointly train Wide & Deep models on user and impression data for app recommender systems.

在推荐系统文献中，通过将评估矩阵[7]的内容信息和协同过滤（CF）的深度学习相结合，探索了协作深度学习。 以前还有一些移动应用推荐系统的工作，例如AppJoy，它使用CF对用户的应用使用记录[8]。 与以前的工作中基于CF或基于内容的方法不同，我们共同研究适用于应用推荐系统的广深模型的用户和印象数据。

# CONCLUSION（结论）

Memorization and generalization are both important for recommender systems. Wide linear models can effectively memorize sparse feature interactions using cross-product feature transformations, while deep neural networks can generalize to previously unseen feature interactions through lowdimensional embeddings. We presented the Wide & Deep learning framework to combine the strengths of both types of model. We productionized and evaluated the framework on the recommender system of Google Play, a massive-scale commercial app store. Online experiment results showed that the Wide & Deep model led to significant improvement on app acquisitions over wide-only and deep-only models.

记忆和泛化对于推荐系统都是重要的。 广度线性模型可以有效地记忆使用向量积特征变换的稀疏特征相互作用，而深层神经网络可以通过低维嵌入推广到先前未见的特征相互作用。 我们提出了广深学习框架，结合两种模式的优势。 我们制作并评估了Google Play推荐系统的框架，这是一个大规模的商业应用商店。 在线实验结果表明，广深模型在广泛而深入的模式下，对应用程序的收购进行了显着的改进。

# REFERENCES

1. J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, July 2011.
2. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
3. H. B. McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and l1 regularization. In *Proc. AISTATS*, 2011.
4. T. Mikolov, A. Deoras, D. Povey, L. Burget, and J. H. Cernocky. Strategies for training large scale neural network language models. In *IEEE Automatic Speech Recognition & Understanding Workshop*, 2011.
5. S. Rendle. Factorization machines with libFM. *ACM Trans. Intell. Syst. Technol.*, 3(3):57:1–57:22, May 2012.
6. J. J. Tompson, A. Jain, Y. LeCun, and C. Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q.

Weinberger, editors, *NIPS*, pages 1799–1807. 2014.

1. H. Wang, N. Wang, and D.-Y. Yeung. Collaborative deep learning for recommender systems. In *Proc. KDD*, pages 1235–1244, 2015.
2. B. Yan and G. Chen. AppJoy: Personalized mobile application discovery. In *MobiSys*, pages 113–126, 2011.

1. See Wide & Deep Tutorial on http://tensorflow.org. [↑](#footnote-ref-1)