

# 基于特征加权与自动交互的点击率预测模型

陈海青, 蔡江辉, 杨海峰, 贺艳婷

(太原科技大学 计算机科学与技术学院, 山西 太原 030024)

**摘要:**在大数据时代的点击率(Click-Through Rate, CTR)预测任务中,输入数据不仅数量多而且特征维度很高,在特征选择时容易出现信息干扰或丢失,在进行特征交互时不同的交互方式也会影响预测性能。针对该问题,文中提出了一种基于特征加权与自动交互的预测模型,用于学习原始特征权重并进行自动交互。首先,引入ECANet模块提出一种不降维的特征加权方法,该方法可以通过对 $k$ 个相邻特征进行一维卷积有效实现。然后,分别用多头自注意网络和深度神经网络(DNN)去自动学习显式和隐式的特征交互。最后,将两者相结合进行预测,弥补了单一模型的缺陷。一方面,它可对输入特征进行重要性选择;另一方面,它能同时以显式和隐式的方式自动学习任意低阶和高阶的特征交互。通过在四个真实数据集上的实验,验证了其比以往的预测模型获得了更好的准确度。

**关键词:**点击率预测;特征交互;特征加权;深度神经网络;多头自注意网络

中图分类号:TP391

文献标识码:A

文章编号:1673-629X(2023)11-0196-06

doi:10.3969/j.issn.1673-629X.2023.11.029

## Click-through Rate Prediction Model Based on Feature Weighting and Automatic Interaction

CHEN Hai-qing, CAI Jiang-hui, YANG Hai-feng, HE Yan-ting

(School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)

**Abstract:** In the click-through rate (CTR) prediction task in the era of big data, the input data is not only large in quantity but also has a high feature dimension, which is prone to information interference or loss during feature selection. Different interaction modes during feature interaction will also affect the prediction performance. To solve this problem, a prediction model based on feature weighting and automatic interaction is proposed, which is used to learn the original feature weights and interact automatically. Firstly, we introduce an ECANet module and propose a feature-weighting method without dimensionality reduction, which uses one-dimensional convolution of  $k$  adjacent features to learn feature weights. Then, multi-head self-attention network and deep neural network (DNN) are used to automatically learn explicit and implicit feature interactions. Finally, the two are combined to predict, solving the defects of a single model. On the one hand, it can select the importance of input features. On the other hand, it can automatically learn arbitrary low- and high-order feature interactions in both explicit and implicit ways. And experimental results on four real data sets have proved that the prediction model is more accurate than the previous model.

**Key words:** click through rate prediction; feature interaction; feature weighting; deep neural network; multi-head self-attention network

## 0 引言

点击率(CTR)预测,旨在预测用户点击推荐项目(如音乐、广告)的概率,在推荐系统中发挥着重要作用。在CTR预测任务中,原始输入特征起着重要作用。由于直接使用原始特征很少能获得理想的预测结果,数据科学家一般会对原始特征进行交互获得更加全面的数据信息,以生成最佳预测系统。但并非所有的特征交互都对预测任务有用,无用的特征交互可能

引入噪声信号,进而导致过度拟合。同时,模型的输入特征大多都是类别特征,而且数量在几十到数百,不同的特征对预测目标影响不同。

在早期阶段,Polynomial-2 (Poly2)<sup>[1]</sup>模型将输入特征进行两两组合并为每一个交叉项设置一个权重参数。但由于数据稀疏,模型无法很好地学习参数。于是,Steffen<sup>[2]</sup>提出了因子分解机(Factorization Machine, FM),它在具备高效计算性能的同时又具备

收稿日期:2022-12-22

修回日期:2023-04-25

基金项目:国家自然科学基金项目(U1931209)

作者简介:陈海青(1996-),男,硕士,研究方向为机器学习和数据挖掘;通信作者:蔡江辉(1978-),男,教授,博士,CCF高级会员(74390S),研究方向为机器学习和数据挖掘。

良好的可解释性,但主要缺点是,它对所有特征交互进行建模,包括有用和无用的组合。与无用特征的交互可能会引入噪声并降低性能。近年来,随着神经网络的发展,出现了许多基于深度学习的预测模型,如 Factorization - Machine Supported Neural Networks (FNN)<sup>[3]</sup>、Product-based Neural Networks (PNN)<sup>[4]</sup>、Higher-Order Factorization Machines (HOFM)<sup>[5]</sup> 等。这些模型的主要缺点是,它们更多地关注高阶特征交互,而很少捕捉低阶交互。上述模型都利用深度神经网络(Deep Neural Network, DNN)来学习高阶特征交互,然而 DNN 以隐式建模来模拟高阶特征交互,因此缺乏关于哪些特征组合是有意义的良好解释。而 Attentional Factorization Machines (AFM)<sup>[6]</sup>、深度因子分解机 DeepFM<sup>[7]</sup>、极深因子分解机 XdeepFM<sup>[8]</sup>、Automatic feature interaction learning via self-attentive neural networks (AutoInt)<sup>[9]</sup> 等模型,使用显式建模来寻求不同阶的特征组合,将整个特征表示到低维空间中,同时提供良好的模型解释性。但是显式特征交互的缺点是:(1)对非线性的建模能力有限;(2)很难扩展到更高阶的特征交互;(3)数据稀疏性很大时,模型训练比较困难;(4)对所有特征交叉项一视同仁,可能会限制模型的表达能力。为了弥补显式和隐式模型的不足,该文用多头自注意网络<sup>[10]</sup>和深度神经网络去自动学习显式和隐式的特征交互,它能同时以显式和隐式的方式自动学习任意低阶和高阶的特征交互。

同时,在广告推荐场景中,输入数据不仅数量多而且特征维度很高,在特征选择时容易存在信息干扰或丢失等问题,限制了模型的表达能力。虽然 AFM 将 attention 机制引入到特征交叉模块进行加权,进而学习每种交互特征的重要性。但这种方式没有考虑到在特征交互时会引入干扰信息并且只考虑了特征两两交互的情况,因此预测准确度的提升十分有限。而 Feature Importance and Bilinear feature Interaction (FiBi-NET)<sup>[11]</sup>利用了挤压提取网络来获取特征重要性。但这种网络在挤压信息时会造成大量特征信息的丢失,进而影响模型性能。

综上,以往方法在对特征进行选择时存在信息干扰和丢失问题,另外在进行特征交互时不同的交互方式也会影响预测性能。为此,该文提出了一种基于特征加权与自动交互的点击率预测模型(Automatic feature interaction prediction model based on feature weighting, FWAI)。首先,引入 Efficient Channel Attention for Deep Convolutional Neural Networks (ECANet)模块<sup>[12]</sup>对输入特征进行加权,该模块通过一维卷积有效地提取特征权重,解决了在特征选择时存在信息干扰和丢失的问题。然后,分别用多头自注

意网络和深度神经网络(DNN)去自动学习显式和隐式的特征交互,从而有效挖掘有用的交互信息。

## 1 相关工作

### 1.1 点击率预测

CTR 预测任务中的关键挑战是如何有效地建模特征交互。在该领域已经提出了许多模型<sup>[13-15]</sup>,如逻辑回归(LR)<sup>[16]</sup>、多项式-2(Poly2)、DeepFM、XdeepFM、AutoInt 等模型。随着深度学习在计算机视觉和自然语言处理等许多研究领域的巨大成功,近年来提出了许多基于深度学习和神经网络的 CTR 模型,如:DeepFM、DCN<sup>[17]</sup>、AFM。

### 1.2 ECANet 模块

ECANet 是 2020 年在 CVPR 上提出的一种通道注意力模块,并在很多任务中被引用。例如, Gera 等人<sup>[18]</sup>利用 ECANet 来消除面部识别过程中的冗余特征,显著提高了模型的鲁棒性。基于该模块在特征重要性提取上的优势,该文提出的 FWAI 通过 ECANet 对输入特征进行动态加权,以提高模型的准确率。在后文将重点介绍该模块的理论和实现机制等。

### 1.3 多头自注意网络和残差网络

提出的模型利用了深度学习文献中的最新技术:注意力机制<sup>[19]</sup>和残差网络<sup>[20]</sup>。注意力最早是在视觉图像领域提出的,并在各种任务中被证明是有效的,如问答、文本摘要和推荐系统。Vaswani 等人进一步提出了多头自注意模型,以模拟机器翻译中单词之间的复杂依赖关系。该文利用它来建模不同特征之间的交互。残差网络在当年的 ImageNet 竞赛中取得了最优的表现。它在解决神经网络中的梯度消失、梯度爆炸以及网络退化等问题中具有突出作用。

## 2 模型介绍

本节介绍 FWAI 模型,模型结构如图 1 所示。

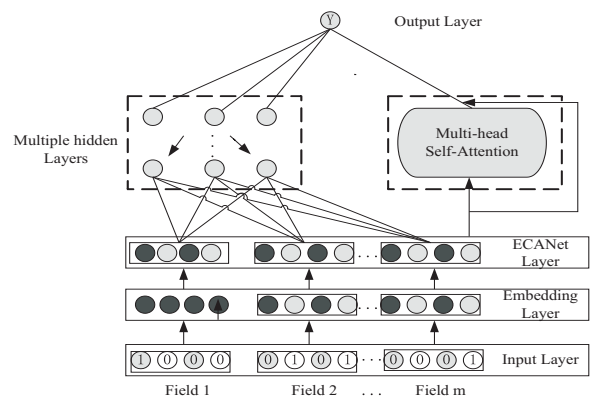


图 1 基于特征加权与自动交互的点击率预测模型

FWAI 主要包含以下部分:输入层(Input Layer)、嵌入层(Embedding Layer)、ECANet 层(ECANet

Layer)、交互层(Interacting Layer)和输出层(Output Layer)。接下来,介绍该方法的细节。

## 2.1 Input Layer

在CTR预测任务中,数据中的特征通常分为类别特征和数值特征。将所有特征输入用稀疏向量表示:

$$\mathbf{x} = [x_1; x_2; \dots; x_m] \quad (1)$$

其中,  $m$  是总特征字段的数量,  $x_i$  是第  $i$  个字段的特征表示。如果第  $i$  个字段是分类的,  $x_i$  是一个热编码向量(one-hot vector)。如果第  $i$  个字段是数值的,  $x_i$  是一个标量值。

## 2.2 Embedding Layer

由于分类特征都是稀疏高维的,所以采用最常见的 Embedding 方法,将它们表示到低维空间。用低维向量表示每个分类特征,即:

$$\mathbf{e}_i = \mathbf{V}_i \cdot \mathbf{x}_i \quad (2)$$

其中,  $\mathbf{V}_i$  是场  $i$  的嵌入矩阵,  $\mathbf{x}_i$  是一个热编码向量。在真实数据中,分类特征可能拥有多个值,  $\mathbf{x}_i$  是一个多值热编码向量(multi-hot vector)。为了与多值输入兼容,对式(2)进行修改,并将多值特征表示为相应特征嵌入向量的平均值:

$$\mathbf{e}_i = \frac{1}{n} \mathbf{V}_i \cdot \mathbf{x}_i \quad (3)$$

其中,  $n$  是第  $i$  个特征字段的值的数量,  $\mathbf{x}_i$  是该字段的多个热向量表示。因为数据中也会存在数值特征,为了允许分类特征和数值特征之间的相互作用,必须将数值特征也表示在同一低维特征空间中,即:

$$\mathbf{e}_i = \mathbf{V}_i \cdot \mathbf{x}_i \quad (4)$$

其中,  $\mathbf{V}_i$  是字段  $m$  的嵌入向量,  $\mathbf{x}_i$  是标量值。

最后,嵌入层的输出是多个特征嵌入向量的串联,  $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_m]$ 。

## 2.3 ECANet Layer

在真实应用场景和数据中,不同的特征对于预测任务的重要性是不同的。比如在预测一条音乐广告是否会被点击时,显然演唱者这个特征比其他特征更重要。受 ECANet 在计算机视觉领域取得重大成功的启发,将 ECANet 机制引入到模型中。对于特定的 CTR 预测任务,可以通过 ECANet 机制动态增加重要特征的权重,降低干扰特征的权重。用嵌入特征作为输入,ECANet 为嵌入特征生成权重向量  $\mathbf{A} = \{a_1, \dots, a_i, \dots, a_m\}$ ,然后用向量  $\mathbf{A}$  新缩放原始嵌入  $\mathbf{E}$ ,得到新的嵌入  $\mathbf{V} = [v_1, \dots, v_i, \dots, v_m]$ ,其中  $a_i \in \mathbb{R}$  是一个标量,表示第  $i$  个特征的  $\mathbf{V}_i$  的权重。如图2所示,ECANet 由三个步骤组成:特征压缩、权重提取、特征加权。

(1)特征压缩(FGAP)。

第一步用于获取每个特征向量的统计信息,使用的方法是平均池化,将嵌入特征  $\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_m]$  转化

为统计向量  $\mathbf{Z} = [z_1, \dots, z_m]$ ,其计算方法如式(5)所示:

$$z_i = F(\mathbf{e}_i) = \frac{1}{k} \sum_{t=1}^k e_i^{(t)} \quad (5)$$

其中,  $e_i^{(t)}$  表示第  $i$  个输入向量的第  $t$  维信息,  $t = 1, 2, \dots, k$ ,  $k$  是嵌入向量的维度。

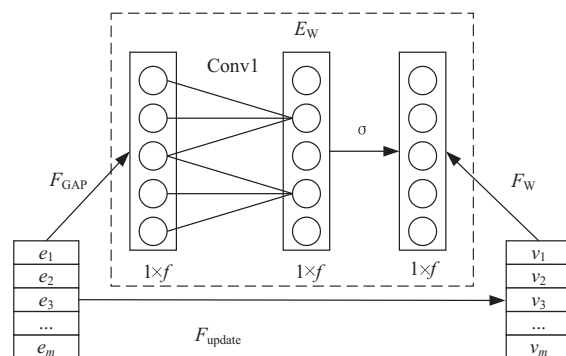


图2 ECANet 层

(2)权重提取(Ew)。

第二步用来学习基于统计向量  $\mathbf{Z}$  的每个字段嵌入的权重。通过考虑每个特征及其  $k$  个邻居来捕获局部跨通道交互信息。可以通过大小为  $k$  的 1D 卷积来有效实现,其中卷积核大小为  $k$  代表了局部跨信道交互的覆盖率,即,该特征附近有多少邻居特征参与了这个特征的注意力预测。形式上,字段嵌入的权重可以计算如下:

$$\mathbf{A} = \sigma(\text{conv1D}_k(\mathbf{Z})) \quad (6)$$

其中, conv1D 表示一维卷积,  $\mathbf{A} \in \mathbb{R}^m$  是一个向量,  $K$  表示卷积核的大小,  $\sigma$  是激活函数。

(3)特征加权(Fw)。

ECANet 最后一步是特征加权,这一步将原始嵌入向量  $\mathbf{E}$  和学习到的权重向量  $\mathbf{A}$  进行相乘,得到加权后的嵌入向量  $\mathbf{V} = [v_1, \dots, v_i, \dots, v_m]$ 。计算过程如公式(7):

$$\mathbf{V} = F(\mathbf{A}, \mathbf{E}) = [a_1 \cdot e_1, \dots, a_m \cdot e_m] = [v_1, \dots, v_m] \quad (7)$$

其中,  $a_i \in \mathbb{R}$ ,  $e_i \in \mathbb{R}^k$ ,  $v_i \in \mathbb{R}^k$ ,  $i$  表示第  $i$  个特征域,  $k$  表示嵌入向量的维度。

## 2.4 Interacting Layer

### 2.4.1 隐式高阶交互

文中在特征嵌入向量  $\mathbf{v}$  上使用前馈神经网络对隐式特征交互进行建模,其计算公式为:

$$x^1 = \sigma(W^{(1)}v + b^1) \quad (8)$$

$$x^n = \sigma(W^{(n)}v^{(n-1)} + b^n) \quad (9)$$

其中,  $n$  表示神经网络层的深度,  $\sigma$  是激活函数,  $x^n$  是第  $n$  层神经网络的输出。

### 2.4.2 显式高阶交互

文中用多头自注意力网络进行显式特征交互。以特



征  $j$  为例,将解释如何识别涉及特征  $j$  的多个有意义的高阶特征。首先定义特定注意头  $h$  下特征  $j$  和特征  $k$  之间的相关性,如下:

$$a_{j,k}^{(h)} = \frac{\exp(\psi^{(h)}(v_j, v_k))}{\sum_{l=1}^M \exp(\psi^{(h)}(v_j, v_l))} \quad (10)$$

$$\psi^{(h)}(v_j, v_k) = \langle W_{\text{Query}}^{(h)} v_j, W_{\text{Key}}^{(h)} v_k \rangle \quad (11)$$

其中,  $\psi^{(h)}(\cdot, \cdot)$  是一个注意函数,它定义了特征  $j$  和  $k$  之间的相似性。接下来,通过组合由系数  $a_{j,k}^{(h)}$  引导所有的相关特征来更新子空间  $h$  下特征  $j$  的表示:

$$v_j^{(h)} = \sum_{k=1}^M a_{j,k}^{(h)} (W_{\text{Value}}^{(h)} v_k) \quad (12)$$

$v_j^{(h)} \in R^d$  是特征  $j$  及其相关特征(在空间  $h$  下)的组合,它代表了一种新的组合特征。此外,一个特征也可能涉及不同的组合特征,通过使用多个 head 来实现这一点,通过创建不同的子空间来学习不同的特征交互。收集了在所有子空间中学习到的组合特征,如下所示:

$$v_j = v_j^{(1)} \oplus v_j^{(2)} \oplus \cdots \oplus v_j^{(H)} \quad (13)$$

其中,  $\oplus$  是串联运算符,  $H$  是总头数。

为保留以前学习到的组合特征和原始特征,在网络中添加了标准的残差网络,即:

$$v_j^{\text{Res}} = \text{ReLU}(v_j + W_{\text{Res}} v_j) \quad (14)$$

其中,  $W_{\text{Res}}$  表示残差部分, ReLU 表示激活函数。通过这样一个交互层,每个特征的表示将更新为新的特征表示  $v_j^{\text{Res}}$ ,这是高阶特征的表示。通过将多个这样的层叠加,可以对任意顺序的组合特征建模。

## 2.5 Output Layer

模型 FWAI 将隐式与显式高阶特征交互结合起来。其模型输出为:

$$\hat{y} = \sigma(W^T(v_1^{\text{Res}} \oplus v_2^{\text{Res}} \oplus \cdots \oplus v_M^{\text{Res}}) + w_{\text{dnn}}^T X_{\text{dnn}}^n + b) \quad (15)$$

其中,  $\sigma$  是 sigmoid 函数,将值转化为用户单击概率。 $X_{\text{dnn}}^n$  是 DNN 的输出,  $W^T$  和  $W_{\text{dnn}}^T$  是可学习的权重参数,  $b$  是偏差。

## 2.6 Training

文中损失函数是交叉熵损失,定义如下:

$$\text{Logloss} = -\frac{1}{N} \sum_{j=1}^N (y_j \log(\hat{y}_j) + (1 - y_j) \log(1 - \hat{y}_j)) \quad (16)$$

其中,  $y_j$  和  $\hat{y}_j$  分别给出了用户点击的真实情况和估计的 CTR,  $j$  为训练样本索引,  $N$  为训练样本总数。

## 3 实验

本节将介绍实验数据以及参数设置,通过在四个真实的公开广告数据集上进行实验,对 FWAI 模型的

预测性能进行评价。实验内容包括:

- (1) FWAI 与目前常见的模型对比;
- (2) ECANet 层对模型性能的提升;
- (3) 不同交互方式对模型性能的影响;
- (4) 超参数设置对模型性能的影响。

### 3.1 实验数据集及设置

#### 3.1.1 实验数据集

使用 4 个公开的真实数据集。表 1 总结了数据集的统计数据。在二值化过程中,将评分小于 3 的样本视为负样本。Criteo、Avazu 和 KDD12 原始数据集的规模都超过了 4 000 万条,由于实验条件限制,该文对每个数据集都只选取了前 1 000 万条数据,然后通过均匀随机抽样方式将数据集按 8 : 1 : 1 划分为三部分:80% 用于训练,10% 用于验证,10% 用于测试。

表 1 实验数据集统计

Data	Samples	Fields
Criteo	10 000 000	39
Avazu	10 000 000	23
KDD12	10 000 000	13
Movielens-1M	739 012	7

#### 3.1.2 实验超参数设置

所有方法都在 TensorFlow 中实现。对于所有深度模型,根据经验将嵌入维度  $d$  设置为 16, batch size 设置为 1 024, Drop Out Ratio 设置为 0.7, 学习率设置为 0.000 1 层深度设置为 3, 最后,使用 Adam 优化所有基于深度神经网络的模型。

#### 3.1.3 评价指标

AUC: ROC 曲线下的面积(AUC),衡量的是 CTR 预测因子对随机选择的正样本项目的评分高于随机选择的负样本项目的概率。AUC 越高,性能越好。

LogLoss: 是基于概率的最重要的分类度量,也是二分类问题中经常使用的评价标准,用来表示预测值与真实值之间的差距。当 LogLoss 的下限为 0 时,表示预测值与真实值完全相同。在实验中,较小的 LogLoss 值表示预测模型具有更好的性能。

### 3.2 FWAI 模型与基线模型性能对比

表 2 总结了在四个数据集上 5 次不同运行的平均结果。观察实验结果可以看出:(1) 二阶特征组合和高阶特征交互的模型在所有数据集上的表现都大大优于 LR,这表明单个特征在 CTR 预测中是不够的,挖掘高级特征交互对于模型至关重要。(2) FWAI 模型在四个数据集上的性能始终优于其他模型。这是因为模型引入了注意力机制对输入特征进行加权,而且分别用多头自注意网络和深度神经网络(DNN)去自动学习显式和隐式的特征交互。使得 FWAI 能够通过加权对输入特征进行重要性选择,使得模型可以有选择性

地学习重要特征,并抑制不太有用的特征,同时以显式和隐式的方式自动学习高阶特征交互,从而有效挖掘特征间的交互信息,显著提升了模型的预测能力。

表 2 四个数据集上不同模型的性能对比

Model Class	Model	Criteo		Avazu		KDD12		MovieLens-1M	
		AUC	LogLoss	AUC	LogLoss	AUC	LogLoss	AUC	LogLoss
First-order	LR	0.776 0	0.475 8	0.757 7	0.384 9	0.736 1	0.168 4	0.771 6	0.442 4
Second-order	FM	0.778 7	0.474 3	0.766 2	0.381 6	0.775 9	0.157 3	0.825 2	0.399 8
	AFM	0.789 5	0.464 3	0.766 5	0.380 2	0.765 9	0.159 1	0.822 7	0.404 8
	FNN	0.792 3	0.461 5	0.765 7	0.382 2	0.771 5	0.159 3	0.838 6	0.395 5
High-order	NFM	0.794 6	0.456 2	0.770 2	0.379 5	0.751 5	0.163 1	0.825 7	0.388 3
	DCN	0.795 1	0.455 2	0.771 9	0.376 8	0.777 3	0.157 2	0.796 8	0.426 6
	HOFM	0.798 6	0.453 4	0.771 3	0.375 4	0.770 7	0.158 6	0.830 4	0.401 3
	AutoInt	0.798 8	0.450 7	0.771 5	0.372 9	0.778 3	0.158 7	0.845 6	0.388 0
	FWAI	0.801 7	0.448 1	0.776 5	0.372 4	0.783 8	0.154 6	0.852 2	0.386 1

### 3.3 ECANet 层对模型性能的提升

本节将讨论 ECANet 机制对模型性能的影响。为了方便起见,用 NO-ECA 表示从模型中删除 ECANet 层。实验结果如表 3 所示。

表 3 ECANet 层对模型性能的提升

Data Sets	Models	AUC	LogLoss
Criteo	FWAI	0.801 7	0.448 1
	NO-ECA	0.800 2	0.449 3
Avazu	FWAI	0.776 5	0.372 4
	NO-ECA	0.773 1	0.374 3
KDD12	FWAI	0.783 8	0.154 6
	NO-ECA	0.775 1	0.163 3
MovieLens-1M	FWAI	0.852 2	0.386 1
	NO-ECA	0.836 7	0.389 4

从实验数据可以发现,当模型加入 ECANet 机制后,预测效果有很大提升。这说明,ECANet 层对于提升模型的性能是十分必要的,它通过对输入特征加权来进行重要性选择,能够有效增强模型能力并使基线模型获得更好的性能。

### 3.4 不同交互方式对模型性能的影响

本节讨论了显式和隐式交互方式对模型性能的影响。使用 E 和 I 来表示使用哪种交互方式。“E”表示使用显式特征交互(该文使用的是多头自注意网络),“I”表示使用隐式特征交互(即深度神经网络),而“EI”表示同时使用显式和隐式交互。在四个不同数据集上进行了实验,并将结果总结在表 4 中。根据实验结果可以观察到,将显性特征交互和隐性特征交互相结合的效果优于单个交互方式,证明了模型的有效性。

### 3.5 超参数设置对模型性能的影响

本节讨论了 FWAI 模型的一些超参数对算法性能的影响,重点放在嵌入层和 DNN 部分。具体包括:嵌入维度;网络深度。

表 4 不同交互方式的模型性能

Data Sets	Models	AUC	LogLoss
Criteo	FWAI-E	0.789 5	0.450 7
	FWAI-I	0.799 5	0.455 2
	FWAI-EI	0.801 7	0.448 1
	FWAI-E	0.771 5	0.372 9
Avazu	FWAI-I	0.770 4	0.376 7
	FWAI-EI	0.776 5	0.372 4
	FWAI-E	0.781 4	0.158 9
KDD12	FWAI-I	0.780 9	0.156 7
	FWAI-EI	0.783 8	0.154 6
	FWAI-E	0.849 6	0.387 8
MovieLens-1M	FWAI-I	0.850 3	0.386 4
	FWAI-EI	0.852 2	0.386 1

表 5 嵌入维度对模型 AUC 和 LogLoss 的影响

Data Sets		16	32	64
Criteo	AUC	0.801 7	0.802 2	0.799 5
	LogLoss	0.458 1	0.446 2	0.446 9
Avazu	AUC	0.776 5	0.777 9	0.774 6
	LogLoss	0.372 5	0.370 5	0.374 2
KDD12	AUC	0.783 8	0.787 5	0.781 6
	LogLoss	0.154 6	0.153 9	0.154 3
Moviel	AUC	0.852 2	0.857 7	0.849 5
	LogLoss	0.386 1	0.384 3	0.384 5

(1)研究了嵌入层的输出维度  $d$  对 FWAI 预测性能的影响。该文将嵌入维度大小按照 16,32,64 进行实验并在表 5 总结了实验结果。通过实验结果观察到了以下信息:从四个数据集上可以看到,随着维度大小的增加,性能不断提高。但当维度达到 32 后,性能开始下降。原因是当参数太多时,模型拟合过度。

(2)网络深度对模型预测性能的影响。

表 6 展示了隐藏层数量对模型性能的影响。可以观察到,FWAI 的性能在开始时随着网络深度的增加而增加。然而,当网络深度大于 3 时,模型性能会下降。这是由于过度拟合造成的。

表 6 网络深度对模型 AUC 和 LogLoss 的影响

Data Sets		2	3	4
Criteo	AUC	0.795 4	0.801 7	0.793 5
	LogLoss	0.453 7	0.448 1	0.456 2
Avazu	AUC	0.765 7	0.776 5	0.763 4
	LogLoss	0.378 2	0.372 5	0.379 6
KDD12	AUC	0.773 6	0.783 8	0.772 1
	LogLoss	0.163 5	0.154 6	0.164 8
Moviel	AUC	0.849 6	0.852 2	0.846 7
Ens-1M	LogLoss	0.390 1	0.386 1	0.389 4

#### 4 结束语

该文提出了一种新的基于特征加权与自动交互的点击率预测模型(FWAI),该模型旨在动态学习特征重要性并自动构建有效的特征交互。一方面,FWAI能够通过加权对输入特征进行重要性选择;另一方面,它能同时以显式和隐式的方式自动学习高阶的特征交互。在四个真实数据集上的实验结果证明了该模型的有效性。对于未来的工作,将继续对模型进行优化,以降低模型的时间复杂度,并改进其在线推荐的性能。

#### 参考文献:

- [1] CHANG Y W, HSIEH C J, CHANG K W, et al. Training and testing low-degree polynomial data mappings via linear SVM[J]. Journal of Machine Learning Research, 2010, 11 (11): 1471-1490.
- [2] RENDLE S. Factorization machines[C]//The 10th IEEE international conference on data mining. Sydney: IEEE, 2010: 995-1000.
- [3] ZHANG W, DU T, WANG J. Deep learning over multi-field categorical data[C]//European conference on information retrieval. Padua: Springer, 2016: 45-57.
- [4] QU Yanru, CAI Han, REN Kan, et al. Product-based neural networks for user response prediction[C]//2016 IEEE 16th international conference on data mining (ICDM). Auckland: IEEE, 2016: 1149-1154.
- [5] BLONDEL M, FUJINO A, UEDA N, et al. Higher-order factorization machines[C]//Proceedings of the 30th international conference on neural information processing systems (NIPS'16). Red Hook: Curran Associates Inc., 2016: 3359-3367.
- [6] XIAO J, YE H, HE X, et al. Attentional factorization machines: learning the weight of feature interactions via attention networks[C]//Proceedings of the 26th international joint conference on artificial intelligence. Melbourne: [s. n.], 2017: 3119-3125.
- [7] GUO Huifeng, TANG Ruiming, YE Yunming, et al. DeepFM: a factorization-machine based neural network for CTR prediction[C]//Proceedings of the 26th international joint conference on artificial intelligence. Melbourne: [s. n.], 2017: 1725-1731.
- [8] LIAN Jianxun, ZHOU Xiaohuan, ZHANG Fuzheng, et al. xDeepFM: combining explicit and implicit feature interactions for recommender systems[C]//Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery and data mining. London: ACM, 2018: 1754-1763.
- [9] SONG Weiping, SHI Chence, XIAO Zhiping, et al. AutoInt: automatic feature interaction learning via self-attentive neural networks[C]//Conference on information and knowledge management. Beijing: ACM, 2019: 1161-1170.
- [10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. Long Beach: [s. n.], 2017: 6000-6010.
- [11] HUANG T, ZHANG Z, ZHANG J. FiBiNET: combining feature importance and bilinear feature interaction for click-through rate prediction[C]//Proceedings of the 13th ACM conference on recommender systems. Copenhagen: ACM, 2019: 169-177.
- [12] WANG Q, WU B, ZHU P, et al. ECA-Net: efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Seattle: IEEE, 2020: 11531-11539.
- [13] 宋永强, 王红, 王露潼. 基于多模态深度融合模型的广告点击率预估[J]. 小型微型计算机系统, 2019, 40(12): 2539-2544.
- [14] 蒋兴渝, 黄贤英, 陈雨晶, 等. 基于特征增强聚合的融合广告点击率预测模型[J]. 计算机工程, 2022, 48(1): 312-320.
- [15] 郝晓培, 朱建生, 单杏花. 基于出行关系的广告点击率预测模型的研究[J]. 计算机技术与发展, 2022, 32(5): 136-140.
- [16] MCMAHAN H B, HOLT G, SCULLEY D, et al. Ad click prediction: a view from the trenches[C]//ACM SIGKDD international conference on knowledge discovery & data mining. Chicago: ACM, 2013: 1222-1230.
- [17] YING S, HOENS T R, JIAN J, et al. Deep crossing: web-scale modeling without manually crafted combinatorial features[C]//Proceedings of the 22nd ACM SIGKDD international conference. San Francisco: ACM, 2016: 55-262.
- [18] GERA D, BALASUBRAMANIAN S. Landmark guidance independent spatio-channel attention and complementary context information based facial expression recognition[J]. Pattern Recognition Letters, 2021, 145: 58-66.
- [19] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[C]//International conference on learning representations. San Diego: [s. n.], 2015.
- [20] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Amsterdam: IEEE, 2016: 770-778.