计算机科学
COMPUTER SCIENCE

基于注意力机制的可解释点击率预估模型研究

杨斌, 梁婧, 周佳薇, 赵梦赐

引用本文

杨斌, 梁婧, 周佳薇, 赵梦赐.基于注意力机制的可解释点击率预估模型研究[J]. 计算机科学, 2023, 50(5): 12-20.

YANG Bin, LIANG Jing, ZHOU Jiawei, ZHAO Mengci. Study on Interpretable Click-Through Rate Prediction Based on Attention Mechanism [J]. Computer Science, 2023, 50(5): 12-20.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000032

基于注意力机制的可解释点击率预估模型研究

杨 斌[1] 梁 婧[2] 周佳薇[2] 赵梦赐[3]
1 中国联合网络通信有限公司研究院 北京 100048
2 北京邮电大学计算机学院 北京 100876
3 北京邮电大学人工智能学院 北京 100876

摘 要 在推荐系统研发中,点击率(Click-Through Rate,CTR)预估是非常重要的工作,点击率预估精度的提升直接影响到整个推荐系统的收益,对其性能和解释性的研究有助于理解系统决策的机理,同时还能帮助优化需求和系统设计。当前点击率预估深度模型多基于线性特征交互和深度特征提取进行设计。由于深度模型的黑盒特点,该类模型在解释性方面存在局限性,并且在先前的研究中,对点击率预估模型的解释性研究非常少。因此,文中基于多头自注意力机制,对该类模型的解释性进行研究,通过多头注意力机制对特征嵌入、线性特征交互和深度部分进行增强和解释,在深度部分设计了两种模型,即注意力增强的深度神经网络和注意力叠加的深度模型,通过计算每个模块的注意力得分对其进行解释。所提方法在多个真实数据集上进行了大量实验,结果表明所提方法能够有效提升模型效果,并且模型自身带有一定的解释性。

关键词:推荐系统;点击率预估;多头自注意力机制;特征交互;模型解释性
中图法分类号 TP391

Study on Interpretable Click-Through Rate Prediction Based on Attention Mechanism

YANG Bin[1],LIANG Jing[2],ZHOU Jiawei[2] and ZHAO Mengci[3]
1 China Unicom Research Institute,Beijing 100048,China
2 School of Computer Science,Beijing University of Posts and Telecommunications,Beijing 100876,China
3 School of Artificial Intelligence,Beijing University of Posts and Telecommunications,Beijing 100876,China

Abstract Click-Through Rate(CTR) prediction is critical to recommender systems. The improvement of CTR prediction can directly affect the earnings target of the recommender system. The performance and interpretation of the CTR prediction algorithm can guide developers to understand and evaluate recommender system accurately. That's also helpful for system design. Most existing approaches are based on linear feature interaction and deep feature extraction,which have poor model interpretation in the outcomes. Moreover,very few previous studies were conducted on the model interpretation of the CTR prediction. Therefore,in this paper,we propose a novel model which introduces multi-head self-attention mechanism to the embedding layer,the linear feature interaction component and the deep component,to study the model interpretation. We propose two models for the deep component. One is deep neural networks(DNN) enhanced by multi-head self-attention mechanism,the other computes high-order feature interaction by stacking multiple attention blocks. Furthermore,we calculate attention scores and interpret the prediction results for each component. We conduct extensive experiments using three real-world benchmark datasets. The results show that the proposed approach not only improves the effect of DeepFM effectively but also offers good model interpretation.

Keywords Recommender system,Click-Through Rate prediction,Multi-head self-attention mechanism,Feature interaction,Model interpretability

1 引言

随着互联网、软件技术的发展,推荐系统已经成为各种网络应用中不可或缺的核心组成部分,如在电商、社交网络、电影音乐等推荐场景中,推荐系统发挥着重要的作用[1]。近年来基于大数据和机器学习驱动的智能推荐系统研发已成为主流趋势。由于智能推荐系统不仅属于系统研发,还涉及到相关智能算法,因此其软件研发涉及到软件工程、数据科学、机器学习等领域。随着技术的发展,该方向的发展逐渐呈现出新的趋势:1)软件研发自动化,从数据采集到评估监控的研发,均可通过自动化的流水线实现;2)数据来源多样化,为了能精确描述用户和推荐物品的特征和关系,研发人员会尽可能引入不同领域的数据,由于不同数据对推荐效果的贡献不同,因此多源数据的引入会增加模型决策的复杂性,带来模型

到稿日期:2022-10-05 返修日期:2023-02-26
通信作者:杨斌(researcher_yang@outlook.com)

解释性的困难；3）模型算法复杂化，为了提升模型效果，推荐模型从机器学习演进到深度学习，当前基于神经网络的推荐模型日趋复杂，更加不可解释。因此，对模型决策原理的理解变得非常必要。

如图 1 所示，推荐系统的研发主要包括需求设计、数据采集与存储、候选集召回、候选集排序、推荐列表生成、评估监控、展示系统等[2]。其中，点击率预估（见图 1 中的绿色区域）是候选集召回、排序、推荐列表生成等阶段中非常重要的实现手段，其主要目的是预估用户对推荐商品或广告的点击概率。在推荐系统中，商家通过提升点击率预估模型的精度，来达到提升整体推荐系统收益的目的[3]。

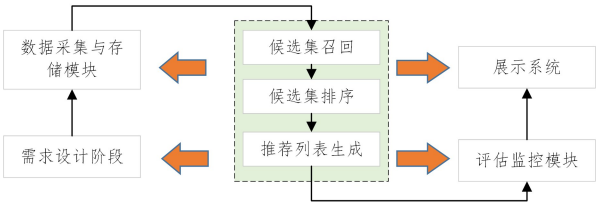

图 1　推荐系统研发流程示意图（电子版为彩图）

Fig. 1　Framework of recommend system

从图 1 可以看出，点击率预估对其他模块的研发具有重要的影响。1）优化需求调研设计。需求调研设计是推荐系统研发生命周期中的初始模块。如果能在该阶段有准确的点击率预估结果用于参考，那么就能更精准地理解和评估推荐系统的目标和性能要求。研发人员可根据点击率预估精度和实际业务需求来优化系统设计。2）指导数据采集与存储设计。通过点击率预估模型解释性研究，可得到特征的相对重要性，从而可优先保证重要数据的采集和加工任务的完成度和时效性。3）指导监控评估模块和展示模块的设计。依据模型解释性，研发人员可了解模型决策原理，理解数据和特征的重要性，对重要特征和模型结果进行评估和监控，也可对模型决策原理进行展示。

由于点击率预估在推荐系统中具有重要价值，因此学者们对这个方向进行了大量的研究。最初主要应用传统的机器学习方法进行预估，如逻辑回归（Logistic Regression，LR）、梯度提升决策树（Gradient Boosting Decision Tree，GBDT）、集成学习等[4-6]。其中，以因子分解机（Factorization Machine，FM）和其变体为代表的算法为学者们提供了通过特征交互方式提升模型效果的思路[7-9]。这些算法多用线性模型和树模型，其模型效果略逊于深度学习算法。但其由于原理简单且易理解，具有很好的解释性；同时还具有易并行、易部署等优势，被学术界和工业界广泛应用。

随着深度学习在各个领域的应用，CTR 预估领域通过深度学习建模的应用越来越多，如 Deep Neural Network（DNN），Wide&Deep，Deep Cross Network（DCN），DeepFM，xDeepFM 等[10-13]。这些模型多为经典模型，其中 DeepFM 模型凭借其性能优势，还被广泛应用于其他二分类建模场景，如风控、医疗、质量评估等[14-16]。随着注意力机制在自然语言处理（Natural Language Processing，NLP）和计算机视觉（Computer Vision，CV）领域被广泛应用[17-18]，其也被引入到点击率模型中。Xiao 等提出了 Attentional Factorization Machine（AFM）模型，该算法将注意力机制应用于不同特征组合的权重计算[19]。Song 等提出了 Automatic Feature Interaction Learning（AutoInt）模型，该模型基于多头注意力机制预估点击率[20]。

已有的工作存在解释性方面的局限性。由于传统预估方法基于具有优秀解释性的线性模型或树模型，因此该类预估方法自身具有较好的解释性，易被建模人员理解和接受。但基于特征交互和深度模型的点击率预估模型的解释性相对较差。如 DeepFM 模型，主要是因为其中特征交互部分为特征两两交互，算法自身并没有考虑交互特征贡献的差异性。同时，深度部分的 DNN 模型属于一种黑盒模型，不具备自解释能力。另一方面，引入注意力机制的模型[19-20]没有展开对点击率预估场景中模型解释性的探索。同时，通过调研发现，基于特征交互和深度模型的结构是众多点击率预估模型的基础[21-23]，且该类模型在其他众多领域被广泛应用。因此，对 DeepFM 这类非常有代表性的模型进行解释性研究，有助于更好地理解模型的决策原理，帮助建模人员更好地挖掘模型的价值。

考虑到注意力机制得分大小能代表输入信息的重要性，因此注意力机制可以被用于理解模型解释性。当前，基于注意力机制对模型解释性的研究在 CV 和 NLP 等领域[17,24-26]显示出了其优越性，但对 CTR 预估模型具有解释性研究还非常少。因此，本文尝试将注意力机制引入到 CTR 预估场景中，选择经典的特征交互和深度模型结合类模型 DeepFM 进行解释性研究。在特征交互部分和 DNN 模型部分均增加注意力机制，将注意力得分作为判断特征重要性的标准，以此进行模型解释。由于 DNN 模型的黑盒属性，通过多层全连接后，无法通过增加注意力机制很好地解释模型的决策机制。因此，本文还提出通过多次注意力机制叠加的方式来代替 DNN 模型，使其深度部分具有解释性。

整体上，本文通过注意力机制对 DeepFM 模型的解释性进行研究，针对模型中的不同模块分别进行解释，包括特征嵌入模块、线性交互模块和深度特征提取模块。首先对特征嵌入模块增加注意力机制，计算其注意力得分；然后对线性特征交互部分增加注意力机制，对交互特征进行解释。对深度部分的解释性研究采用两种方式：1）在输出位置增加注意力机制，通过注意力机制得分进行解释；2）提出了一种基于注意力叠加方式来替换 DeepFM 模型中的深度模型，使其具备自身可解释性。本文的主要贡献如下：

（1）提出了一个新模型，即基于注意力机制增强的 DeepFM 模型。该模型对特征嵌入、线性特征交互、深度部分均通过多头注意力机制进行增强，提升了模型效果和模型解释性。

（2）提出了一种解释性更好的线性特征和深度特征结合的模型。该模型可通过叠加多个注意力机制来实现模型的解释性，可将其作为 DNN 模型的替代部分。

（3）在多个真实数据上进行了大量的实验。实验结果显示，本文方法能够有效提升模型的效果，并且能够增强 CTR 预估模型的解释性。

本文第 2 节主要介绍了相关工作；第 3 节主要介绍了本文方法；第 4 节进行了实验并对结果进行了分析；最后总结全文。

2 相关工作

2.1 点击率预估模型

点击率模型不仅主要被应用于推荐广告系统，一些经典点击率预估算法还被广泛应用于其他决策类场景，如风控、医疗、质量评估等[14-16]。最初的研究主要以矩阵分解为代表[27]，后来引入浅度机器学习方法[4-6]。这一阶段的研究多集中在如何通过特征工程获取表征能力更强的高阶特征。Facebook 的研究人员提出通过 GBDT 将特征编码到隐空间，以获得更具表征能力的高阶特征，然后将高阶特征输入到 LR 进行最终决策[9]。FM 和 FFM 等算法的出现也是为了能够获取更有价值的交互特征[7-8]，这类算法学习每个独立特征之间的交互关系，得到二阶交互特征，用于提升模型效果。

随着深度学习的发展，从学术界到工业界都将深度学习引入到 CTR 预估的场景展开了大量的研究。Wide & Deep 模型[11]是谷歌研究人员首次提出的，其将线性模型和深度模型相结合来获取特征。后续的 DeepFM 模型[3]为了减轻在特征工程中人工选择特征，以 Wide & Deep 模型为基础，用 FM 替换线性部分。Neural Factorization Machines（NFM）[28]则基于嵌入向量的内积，提出了一种串行方式的特征向量线性组合的点击率预估模型。综上可以发现，由于交互特征和深度特征结合的方式在模型中的效果表现优异，后续很多 CTR 预估算法都借鉴了该思想[21-23]。本文模型将以 DeepFM 为基础，探索将多头注意力机制分别应用于模型的不同部分时对模型性能的影响及其模型解释性。

2.2 注意力机制

注意力机制主要用于学习特征的重要性，其最早是在机器翻译领域[29]提出的，被用于给源语言和目标语言更相关的词赋予更大的权重，使得这些相关的词在后续计算中发挥更大作用。由于注意力机制能够凸显重要信息，在模型决策的过程中发挥更重要的作用，因此注意力机制被广泛地应用到众多领域[30-32]。

由于注意力机制能够揭示特征表征的显著性分布，因此注意力机制可以被用于模型理解。其在 CV、NLP 和健康医疗等领域被应用于模型解释性工作[32-34]。在 CTR 预估模型中，解释性好的模型多为线性模型和树模型，因为这类算法具有天然的解释性。LR 算法主要是因为其特征权重[4]而具有很好的解释性。Xgboost 则主要通过计算特征重要性被用于模型预测结果的解释性[5]。尽管注意力机制被应用到点击率预估中，但当前很少有工作对其解释性进行探索研究。如 AFM[19]考虑到 FM 算法没有区分不同特征组合的权重，为了区别重要性不同的特征组合，用注意力网络生成交互特征权重。例如 Song 等提出的 AutoInt 则是基于多头注意力机制的模型，其模型效果提升显著[20]。

在点击率预估场景中，对生产实践中被广泛应用的经典模型 DeepFM 在解释性方面的研究很少。由于模型本身具有黑盒属性，其自身并不具备解释性。因此，本文主要针对 DeepFM 模型，通过注意力机制进行模型解释性研究。

3 注意力增强的点击模型

本文基于 DeepFM 模型，引入多头自注意力机制对模型进行解释。如图 2 所示，模型的整体架构主要包括 4 个部分：输入层、特征嵌入层、线性交互和深度层、输出层。分别为特征嵌入层、线性交互部分和深度部分配备多头自注意力机制进行增强，以学习不同子空间中特征交互的多义性。其中深度部分，除了基于注意力机制增强的深度模型，我们还提出通过叠加注意力机制的方式来获得具有解释性更好的深度模型。

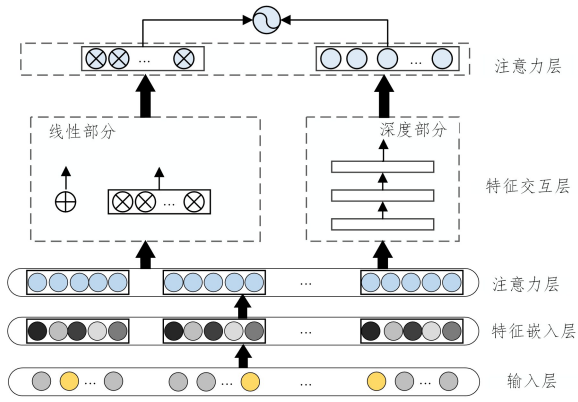

图 2 模型的结构示意图

Fig. 2 Structure illustration of the model

3.1 多头自注意力机制

多头自注意力机制（Multi-Head Self Attention，MHSA）能够在不考虑单词顺序和距离的情况下，学习句子内或跨句子的单词对的文本语义协同作用，具有很好的解释性。

图 3 为多头自注意力机制的示意图。其中，多头自注意力机制使用多个头来创建不同的子空间，分别学习特征在不同子空间下的特征相关性。其中，$head_i$ 代表多头自注意力机制下学习到的第 i 个子空间下的特征[17]，将计算 $head_i$ 的输入矩阵简写为 $\boldsymbol{X}$。本文中，由于多头自注意力机制被用于增强不同的模块，因此在不同模块下对应的输入矩阵 $\boldsymbol{X}$ 也不同，分别为不同模块的输出。首先，将输入矩阵 $\boldsymbol{X}$ 映射到 3 个不同的矩阵：$\boldsymbol{Q}_i$（Query），$\boldsymbol{K}_i$（Key）和 $\boldsymbol{V}_i$（Value）。

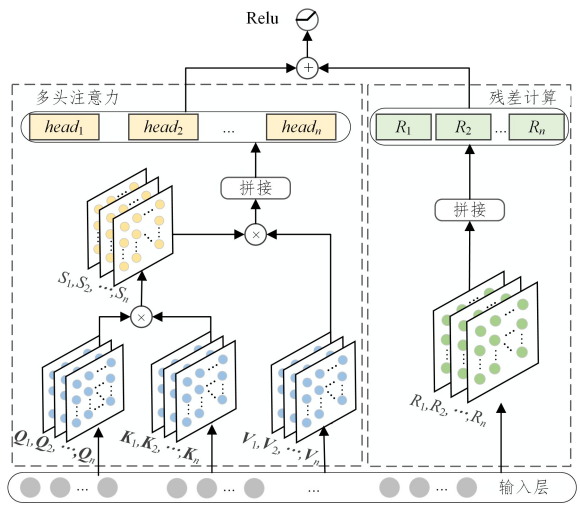

图 3 多头自注意力机制示意图

Fig. 3 Illustration of multi-head self-attention mechanism

$$\begin{aligned}\boldsymbol{Q}_i &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{Q^i}\\ \boldsymbol{K}_i &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{K^i}\\ \boldsymbol{V}_i &= \boldsymbol{X}^{\mathrm{T}}\boldsymbol{W}^{V^i}\end{aligned} \tag{1}$$

其中，投影矩阵 $\boldsymbol{W}^{Q^i}\in R^{d_k\times d}$，$\boldsymbol{W}^{K^i}\in R^{d_k\times d}$，$\boldsymbol{W}^{V^i}\in R^{d_k\times d}$，$d_K$，表示注意力因子。如式(2)所示，为了防止点积太大，我们令 Query 和所有 Key 的点积除以 $\sqrt{d_K}$，并使用 Softmax 函数进行标准化，获得 Values 的权重 $\boldsymbol{S}_i$(Score)。然后，如式(3)所示，将权重 $\boldsymbol{S}_i$ 与 Value 相乘得到 $head_i$：

$$Score_i=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_i\boldsymbol{K}_i^{\mathrm{T}}}{\sqrt{d_K}}\right) \tag{2}$$

$$headi=Score_i\boldsymbol{V}_i \tag{3}$$

将多个头在不同子空间下学习到的不同的特征组合在一起，如式(4)所示。在计算多头注意力的同时，为了保留输入向量的初始信息，我们结合残差网络，将原始输入信息也保留，输出如式(5)所示：

$$Head=Concat(head_1,head_2,\cdots,head_n) \tag{4}$$

$$MHSA(\boldsymbol{X})=\mathrm{ReLU}(Head+\boldsymbol{W}_{\mathrm{Res}}\boldsymbol{X}) \tag{5}$$

3.2　注意力增强的特征嵌入

在点击率预估中，一个点击记录包括一组特征和一个表示是否点击的二进制标签。特征部分包括两类，即类别特征或数值特征，不同的特征被定义为不同的值。由于类别特征无法直接用于数值计算，为了表征这类特征，我们采用独热编码将特征向量转换为高维稀疏向量。考虑到高维稀疏特征难以处理，通常采用特征嵌入将高维稀疏特征嵌入到低维稠密的向量空间中，以获得特征嵌入向量 $\boldsymbol{E}$。如式(6)所示，M 表示特征字段总数，e_i 表示第 i 个嵌入特征向量。

$$\boldsymbol{E}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_M] \tag{6}$$

为了探究多头自注意力机制在模型不同部分的可解释性，我们在特征嵌入层输出后接入多头自注意力机制。特征嵌入层的最终输出如下：

$$\boldsymbol{E}_{\mathrm{att}}=MHSA(\boldsymbol{E}) \tag{7}$$

3.3　注意力增强的线性交互

在 DeepFM 模型中，线性部分使用内积来建模特征交互。但由于其计算结果是一个数值，不便于多头自注意力机制计算特征权重，因此采用 AFM 模型中提出的成对交互层(Pairwise Interaction Layer)来取代内积。具体地，将 M 个向量拓展为 $M(M-1)/2$ 个相互作用的向量，其中每个相互作用向量是两个不同向量元素的乘积，输出表示为：

$$F_{PI}=\{(v_i\odot v_j)x_ix_j\}_{(i,j)\in R_x} \tag{8}$$

其中，设特征向量中非零特征的集合为 $\boldsymbol{X}$，嵌入层的输出特征向量为 $e_i=\{v_ix_i\}_{i\in\boldsymbol{X}}$，符号⊙表示两个向量元素级乘积，$\boldsymbol{R}_x$ 表示 $\{(i,j)\}_{i\in X,j\in X,j>i}$。

为了探索多头自注意力机制对线性部分的解释性，我们将成对交互层的输出接入多头自注意力机制。最终注意力增强后的线性交互特征为：

$$PI_{\mathrm{att}}=MHSA(F_{PI}) \tag{9}$$

3.4　深度部分

由于 DeepFM 模型的深度部分 DNN 是一个黑盒模型，不具备模型解释性，因此，本文在深度部分基于注意力机制提出了两个模型。

3.4.1　模型 1：注意力增强的 DNN

如图 4 中的模型 1 所示，全连接的深度神经网络(DNN)作为模型深度部分，隐式地捕获高阶交互特征，其前馈过程可表示为：

$$a^{(l+1)}=\sigma(\boldsymbol{W}^{(l)}\cdot a^{(l)}+b^{(l)}) \tag{10}$$

其中，l 为 DNN 网络深度，σ 为激活函数，$W^{(l)}$，$a^{(l)}$ 和 $b^{(l)}$ 分别是第 l 层的模型权重、输出和偏置。模型的初始输入为带有多头自注意力机制的特征嵌入层的输出，即 $a^{(0)}=E_{\mathrm{att}}$。假设经过 $|H|$ 个全连接隐层后，得到一个稠密的特征向量 $\boldsymbol{F}_{\mathrm{DNN}}$：

$$\boldsymbol{F}_{\mathrm{DNN}}=\boldsymbol{W}^{|H|+1}\cdot a^{|H|}+b^{|H|+1} \tag{11}$$

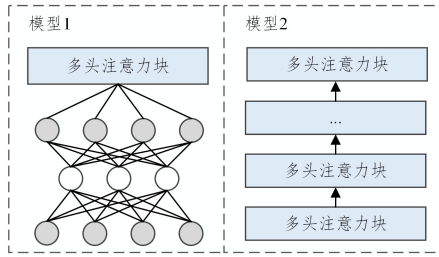

图 4　深度部分模型结构图

Fig. 4　Framework of deep component

与多头自注意力机制增强其他模块相似，我们将深度部分的输出也接入多头自注意力机制，输出结果如下：

$$DNN_{\mathrm{att}}=MHSA(\boldsymbol{F}_{\mathrm{DNN}}) \tag{12}$$

3.4.2　模型 2：注意力叠加模型

由于深度神经网络 DNN 不具备解释性，因此模型 1 的解释性相对有限。虽然多头自注意力机制可以用于解释模型 1 的输出，但无法映射到输入特征上。

如图 4 中的模型 2 所示，为了能够解释模型的深度部分，我们提出了一个 DeepFM 模型的变体。受 Transformer 模型[17]的启发，考虑到注意力机制具有更好的解释性，模型 2 通过叠加多个多头自注意力层，来对高阶特征组合进行建模，最终代替 DNN 作为模型的深度部分。具体来说，特征嵌入层的结果输入到第一个多头自注意力层后，当前多头自注意力层的输出将作为下一个多头自注意力层的输入。最终的输出是：

$$Deep_{\mathrm{att}}^{(i+1)}=MHSA(Deep_{\mathrm{att}}^{(i)}) \tag{13}$$

其中，$Deep_{\mathrm{att}}^{(0)}$ 为带有多头自注意力机制的特征嵌入层的输出 $\boldsymbol{E}_{\mathrm{att}}$。

3.5　输出层

在获得了线性交互部分和深度部分的特征后，我们对线性部分和深度部分的输出结果进行非线性投影，具体如下：

$$\hat{y}_{\mathrm{FM}}=w_0+\sum_{i=1}^{n}w_ix_i+\sigma(w^{\mathrm{T}}PI_{\mathrm{att}}+b) \tag{14}$$

$$\hat{y}_{\mathrm{DNN}}=\sigma(w^{\mathrm{T}}DNN_{\mathrm{att}}+b) \tag{15}$$

$$\hat{y}_{\mathrm{Deep}}=\sigma(w^{\mathrm{T}}Deep_{\mathrm{att}}+b) \tag{16}$$

其中，$\hat{y}_{\mathrm{FM}}$ 为线性部分的输出，除了二阶特征也可以增加 1 阶特征。深度部分的输出包括模型 1 和模型 2，分别为 $\hat{y}_{\mathrm{DNN}}$ 和 $\hat{y}_{\mathrm{Deep}}$。对于模型 1，我们基于线性部分预估 $\hat{y}_{\mathrm{FM}}$ 和深度部分预估 $\hat{y}_{\mathrm{DNN}}$ 通过全连接层进行预测，最终的预测结果如式(17)所示。同理，模型 2 的最终预测结果如式(18)所示。

$$\hat{y}_{\mathrm{model1}}(x)=\sigma(w^{\mathrm{T}}(\hat{y}_{\mathrm{FM}}+\hat{y}_{\mathrm{DNN}})+b) \tag{17}$$

$$\hat{y}_{\mathrm{model2}}(x)=\sigma(w^{\mathrm{T}}(\hat{y}_{\mathrm{FM}}+\hat{y}_{\mathrm{Deep}})+b) \tag{18}$$

此外，我们使用交叉熵损失函数来训练本文模型。

$$Logloss=-\frac{1}{N}\sum_{j=1}^{N}(y_j\log(\hat{y}_j)+(1-y_j)\log(1-\hat{y}_j)) \quad (19)$$

其中，y_j 为输入样本 j 的真实类别，$\hat{y}_j$ 为样本 j 的预估概率，N 为训练样本总数。

3.6 解释性

本节主要阐述多头自注意力机制在本文模型中的可解释性。一般来说，模型的可解释性指建模人员对该模型的理解程度。由于多头自注意力机制可以通过在不同子空间分别计算查询向量(Query)和键向量(Key)乘积，得到查询向量和各个对应的键向量之间的相关性，即一个子空间中特征的注意力分数。该注意力分数可以代表不同特征的贡献，因此可以基于此对模型预测结果进行解释。

本文将在模型进行充分学习之后，通过特征分析对模型进行事后解释。我们通过多头注意力机制对 DeepFM 模型的特征嵌入部分、线性特征交互部分、深度部分均进行了增强。模型的特征嵌入部分和线性特征交互部分通过多头自注意力机制后，可以从注意力得分中找出注意力分数高于其他注意力得分的特征。这些注意力得分高的嵌入特征、一阶特征和二阶交互特征相比其他注意力得分低的特征发挥着更重要的作用。

对于模型的深度部分，由于 DeepFM 中的 DNN 网络具有不可解释性，在通过多层全连接后，难以通过仅在输出端增加注意力机制来获取对应的初始特征的注意力得分，导致解释性不理想。因此，本文使用叠加多个多头自注意力机制来代替 DNN 模型的方式，让深度部分具有解释性。同样地，我们可以根据多头自注意力机制计算的注意力分数，选出对预测结果影响较大的深度特征，用于解释模型在深度部分的决策机制。

4 实验基础

4.1 实验设置

(1)数据集

本文选择在 Criteo[1)]，Avazu[2)] 和 MovieLens[3)] 数据集上进行实验，用于评估模型的有效性和解释性。

1)Criteo：被广泛用于点击率预估的标准数据集，其包含一个月的广告点击日志，每个样本包含 13 个数值特征和 26 个类别特征。为了快速评估模型的有效性，我们选择了 500 万条数据用于对基线模型和本文模型进行实验。

2)Avazu：包含了 10 天内移动广告点击日志，每个点击数据包含 23 个功能字段，包括域名、应用设备等信息。我们选择其中 500 万条数据用于对基线模型和本文模型进行实验。

3)MovieLens-1M：包含来自 6 040 位在 2000 年加入 MovieLens 的用户对大约 3 900 部电影的评价。我们将评分为 0～2 的样本视为负样本，将评分大于等于 3 的样本视为正样本，对数据进行二值化处理。

(2)评价指标

本文采用了两个评价指标：AUC(Area Under ROC Curve)和 Logloss(Logistic loss)。

1)AUC：ROC 曲线下的面积，AUC 值越大的模型预估性能越高。

2)LogLoss：基于概率的最重要的分类度量指标，反映了样本的平均偏差，其常作为模型的损失目标来进行优化。LogLoss 值越低，意味着模型的预测性能越好。

(3)基线模型

本文选择了 6 个基线模型。

1)LR[35-36]：只对原始特征进行线性变换。

2)FM[7]：使用因子分解机对二阶特征进行建模。

3)DeepFM[3]：结合 FM 部分和深度模型提取高阶交互特征。

4)xDeepFM[13]：结合了深度模型和压缩交互网络，同时对高阶特征交互进行隐式和显式建模。

5)AFM[19]：在 FM 的基础上引入注意力机制，区分不同交互特征的重要性。

6)AutoInt[20]：使用多头自注意力机制显式地学习输入特征的高阶特征交互。

(4)超参数设置

在实验中使用 Tensorflow 实现了所有的模型。本文中，多头自注意力机制中头的数量 H 分别在 Criteo，Avazu 和 MovieLens 这 3 个数据集上，设置为 $H=2$，$H=4$ 和 $H=6$。对于所有的数据，我们将嵌入向量的维度设置为 8。对于具有 DNN 部分的模型 1，在 Criteo 和 Avazu 数据集上隐藏层的深度设置为 3，每层的神经元个数分别设置为(256，128，64)；在 MovieLens 数据集上，隐层深度设置为 2，每层的神经元个数分别设置为(128，64)。考虑到深度模型的隐层深度为 3，本文中设置模型 2 的多头自注意力机制叠加数为 3，激活函数使用 ReLU 函数。为了防止过拟合，我们将 Dropout 设置为 0.5，批处理大小为 1 024，优化器为 Adam，学习率设置为 0.001。

4.2 结果和分析

为了验证本文模型的有效性，我们在 Criteo，Avazu 和 MovieLens 这 3 个数据集上分别进行实验，实验结果如表 1 所列。

表 1 本文模型与基线模型的性能比较

Table 1 Performance comparison of the proposed model and baseline models

模型	Criteo		Avazu		MovieLens	
	AUC	Logloss	AUC	Logloss	AUC	Logloss
LR	0.7308	0.5133	0.7296	0.4059	0.7716	0.4424
FM	0.7543	0.5192	0.7394	0.4011	0.8252	0.3998
xDeepFM	0.7836	0.4703	0.7656	0.3949	0.8413	0.3298
AFM	0.7441	0.4964	0.7362	0.4337	0.8227	0.4048
AutoInt	0.7878	0.4822	0.7655	0.4070	0.8357	0.3324
DeepFM	0.7661	0.4955	0.7450	0.3984	0.8323	0.3343
模型 1	0.7781	0.4844	0.7560	0.4019	0.8124	0.3602
模型 2	0.7786	0.4752	0.7654	0.3959	0.8390	0.3306

1) https://www.kaggle.com/c/criteo-display-ad-challenge

2) https://www.kaggle.com/c/avazu-ctr-prediction

3) https://grouplens.org/datasets/movielens/

可以发现，模型 1 和模型 2 相比原始的 DeepFM 模型都有很大的提升。相比 DeepFM 模型，模型 1 的性能在数据集 Criteo 和 Avazu 上分别提升了 1.20%和 1.10%；模型 2 则提升了 1.25%和 2.04%。在 MovieLens 数据集上模型 1 基本和 DeepFM 的效果持平，但模型 2 有比较明显的提升，为 0.67%。实验结果充分说明了基于多头自注意力机制增强在 DeepFM 模型上的有效性。

通过进一步观察可以发现，在 Criteo 和 Avazu 数据集上，模型 1 和模型 2 的效果与效果最好的模型 AutoInt 和 xDeepFM 基本持平，尤其是模型 2 在 3 个数据集上的效果都比较理想。该结果充分证明了本文基于 DeepFM 的增强模型效果可媲美其他先进模型的能力。

4.3　消融实验

本文在所有数据集上设计了两个消融实验，以验证不同模块上多头自注意力机制对模型最终效果的贡献。

为了验证注意力机制和每个模块结合的效果，我们在第一个消融实验中，针对模型 1，分别只在嵌入层、线性部分或深度部分添加多头自注意力机制。其中，未添加多头自注意力机制的模型即为 DeepFM 模型。而由于模型 2 的深度部分通过叠加的多头自注意力层实现，因此在第一个消融实验中，针对模型 2 仅在深度部分进行。实验结果如表 2 所列。

表 2　消融实验 1：独立模块的实验结果

Table 2　Performance of ablation experiment 1：result of independent component experiment

模型	嵌入层	线性部分	深度部分		Criteo		Avazu		MovieLens	
			DNN	叠加层	AUC	Logloss	AUC	Logloss	AUC	Logloss
模型 1					0.7661	0.4955	0.7450	0.3984	0.8323	0.3343
	√				0.7764	0.4840	0.7568	0.3995	0.8137	0.3570
		√			0.7771	0.4722	0.7597	0.3979	0.8197	0.3437
			√		**0.7803**	**0.4772**	**0.7651**	**0.3982**	**0.8439**	**0.3329**
	√	√	√		0.7781	0.4844	0.7560	0.4019	0.8124	0.3602
模型 2				√	**0.7815**	**0.4705**	**0.7656**	**0.3939**	**0.8406**	**0.3356**
	√	√		√	0.7786	0.4752	0.7654	0.3959	0.8390	0.3306

注：√表示在对应模块添加多头自注意力机制

为了验证每个模块结合深度模型是否存在叠加效果，我们在模型 1 和模型 2 的基础上设计了第二个消融实验，分别在嵌入层、线性部分和深度部分减去 1 个多头自注意力机制，以验证不同模块组合的效果。实验结果如表 3 所列。

表 3　消融实验 2：模块组合的实验结果

Table 3　Performance of ablation experiment 2：result of combined component experiment

模型	嵌入层	线性部分	深度部分		Criteo		Avazu		MovieLens	
			DNN	叠加层	AUC	Logloss	AUC	Logloss	AUC	Logloss
模型 1	√	√	√		0.7781	0.4844	0.7560	0.4019	0.8124	0.3602
	√	√			0.7771	0.4731	0.7563	0.4018	0.8080	0.3561
	√		√		**0.7777**	**0.4745**	0.7576	0.4006	0.8146	0.3522
		√	√		0.7758	0.4727	**0.7591**	**0.3998**	**0.8339**	**0.3468**
模型 2	√	√		√	0.7786	0.4752	0.7654	0.3959	0.8390	0.3306
	√			√	0.7770	0.4772	0.7548	0.4015	0.8284	0.3400
		√		√	**0.7844**	**0.4744**	**0.7677**	**0.3950**	**0.8404**	**0.3295**

注：√表示在对应模块添加多头自注意力机制

由表 2 可知，在所有数据集上，相比 DeepFM 模型，只在深度部分 DNN 上使用多头自注意力增强后模型性能得到大幅提高。在 Criteo 和 Avazu 数据集上，消融实验表明在嵌入层、线性部分和深度部分分别添加多头自注意力机制后的模型性能相比 DeepFM 模型有提升，其中仅在 DNN 部分添加自注意力机制的模型性能最好；在 MovieLens 上仅在嵌入层或线性部分添加多头自注意力机制，与在嵌入层、线性部分和深度部分这 3 个部分都使用多头自注意力增强一样，两种方式的模型性能反而不如 DeepFM，这可能与数据集数量大小有关。

进一步对比模型 1 和模型 2 的深度部分的实验结果可以发现，这两个模型的性能效果相近，模型 2 的效果略好。这意味着将 DeepFM 的深度部分替换为多个叠加的多头自注意力后，不仅维持了模型的良好性能，还使模型具备了解释性，这样的替换是有意义的。

由表 2、表 3 可知，在 3 个数据集上，模型 2 中深度部分和线性部分的结合效果最理想。并且在模型 1 和模型 2 中，基于注意力机制的深度部分和线性部分的组合效果比其他模块的组合效果更好。这充分说明了基于注意力的方式不仅具有解释性，还可以提升模型的性能。

4.4　超参数研究

本文模型有两个关键超参数，分别为多头自注意力机制中头的数量和 DNN 网络的深度。我们分别探究了这两个超参的设置对本文模型效果的影响。

4.4.1　多头自注意力机制中头的数量

为了了解多头自注意力机制头的数量这一超参数对实验结果的影响，本文进行了调参实验。保持其他参数固定，设置头的数量分别为 2，4，6，8，实验结果如图 5 所示。可以发现当头的数量的设置从少到多时，模型性能整体有提升。但随着头的数量的增加，从 6 开始，由于过度参数化，模型性能有所下降。在 Avazu 数据集上设置头的数量在 4 以上性能达到最好，随后也有波动。实验结果证明，多头自注意力机制有益

于模型效果，且不同的数据集最好的参数也不一致，具有差异性。

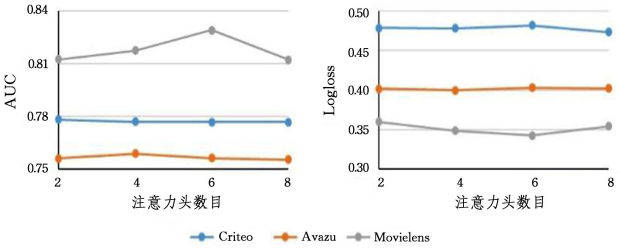

图 5 注意力头数目对模型性能的影响

Fig. 5 Effect of different number of heads on performance

4.4.2 DNN 网络的深度

图 6 给出了模型中深度部分 DNN 网络的隐藏层数量对模型性能的影响。其中，在 Criteo 和 Avazu 数据集上，当网络深度设置为 3 层时，模型的性能达到最好效果，而在 MovieLens 数据集上，当网络深度设置为 2 时模型性能最好。对于 3 个数据集而言，当 DNN 网络深度的设置大于最优值后，模型的性能衰减，有不同程度的下降，这是因网络层数过多导致模型过拟合而造成的。

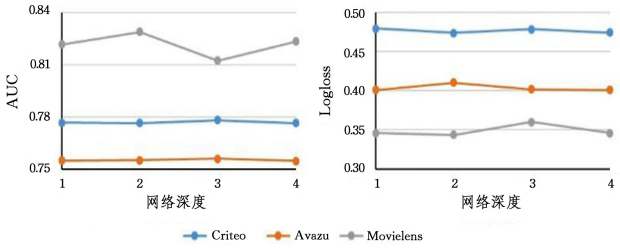

图 6 网络深度对模型性能影响

Fig. 6 Effect of different network depths on performance

4.5 可解释性

本节使用可视化的方法，对模型中学习到的重要的嵌入特征、线性特征和高阶特征进行解释。由于 Criteo 数据集和 Avazu 数据集为了保护用户隐私，对用户点击记录的实际内容进行了加密编码处理，导致我们无法直接使用这两个数据集对模型进行可视化解释。但 MovieLens 数据集包含来自 6040 位于 2000 年加入 MovieLens 的用户对大约 3900 部电影的评价，且每一个数据样本包括用户信息（用户年龄、性别、职业等）、电影发行年份、所属类型等电影信息。这些特征信息有助于对模型决策的理解，因此选择基于 MovieLens 数据进行可视化模型解释。

我们绘制了特征嵌入部分、线性交互部分和深度部分的输出特征的多头自注意力权重热图，如图 7—图 10 所示。图 7—图 10 中，特征对应单元格的颜色越深，表示其注意力得分越高，特征就越重要，模型最终决策对该特征的依赖越强；反之，颜色越浅代表得分越低，特征重要性越低，模型最终决策对该特征的依赖越小。

4.5.1 特征嵌入层

图 7 给出了 1990 年上映的电影《Dances with Wolves》在嵌入层的多头自注意力得分，这个电影所属类型为冒险、剧情和西部。从图片中可以看出，冒险（Adventure）、剧情（Drama）和西部（Western）对应的单元格颜色明显比其他单元格的颜色更深，符合电影所属类型。同时，与电影主题类型不太相符的类别标签（如动画（Animation）、科幻（Sci-Fi）和音乐（Musical）类型）的得分就相对较低。

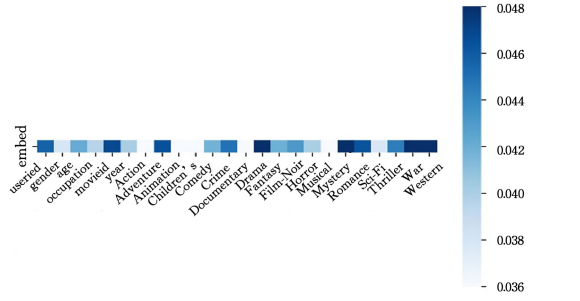

图 7 嵌入特征注意力权重示意图

Fig. 7 Attention weights diagram of embedding features

4.5.2 线性部分

图 8 给出了一部作为剧情类型获得较高声誉的电影——《True Lies》的二阶交互特征的重要性得分。我们可以观察到，在二阶交互特征中，用户编号（Userid）和爱情（Romance）的交互特征以及冒险（Adventure）和悬疑（Mystery）的交互特征获得了较高的注意力分数。这意味着，该部影片的用户群体会更愿意选择爱情类型电影。并且电影《True Lies》是一部兼具冒险和悬疑元素的电影，电影的剧情内容也比较符合这一特点。同样地，和电影类型差别较大的类别标签的组合（如冒险（Adventure）、少儿（Children's）及奇幻（Fantasy）的组合）的得分都很低。

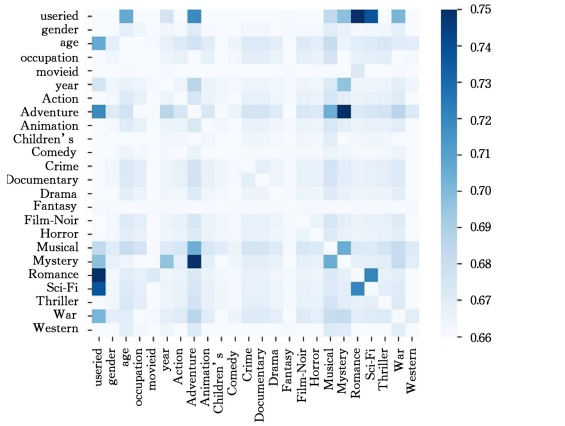

图 8 一个线性特征交互注意力权重示意图

Fig. 8 Attention weights diagram of linear feature interaction

4.5.3 深度部分

图 9 和图 10 分别为深度部分为 DNN 或多个叠加多头自注意力机制时，生成的深度高阶特征的注意力权重热图。

图 9 给出了电影《North by Northwest》的 DNN 部分输出的子注意力得分。可以看出，大部分高阶特征对应的颜色深浅很接近，这意味着经过 3 层 DNN 网络后的输出特征的重要性没有明显区别。这主要是因为 DNN 没有考虑到特征之间的相对重要性，且由于深度神经网络的算法具有复杂性和不透明性，因此我们无法得知每一个最终输出的高阶特征所对应数据集中的初始特征，无法通过获得初始特征的重要

性分布来进行模型解释。

由于深度神经网络 DNN 不具备解释性，我们选择通过叠加多个多头自注意力层来作为模型的深度部分。为了便于解释，将叠加多头自注意力得到的特征称为高阶特征。图 10 给出了一个叠加多个多头自注意力机制的例子，描述了电影《North by Northwest》经过嵌入层和高阶交互层之后特征的重要性。可以观察到，对于嵌入层而言，剧情(Drama)这个特征的得分较高；而对于高阶特征而言，电影在惊悚(Thriller)和西部(Western)方面的属性被捕捉到，符合该电影作为口碑较高的惊悚悬疑片。以上 3 个主要特征综合影响了这部电影的评分结果。

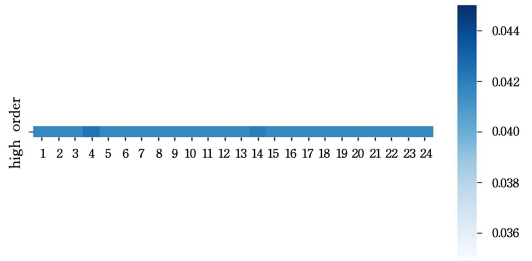

图 9　模型 1 深度部分特征注意力权重示意图

Fig. 9　Attention weights diagram of deep features in model 1

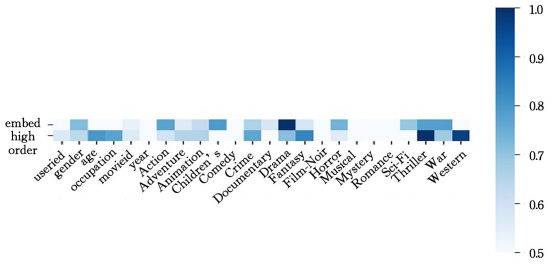

图 10　模型 2 深度部分特征注意力权重示意图

Fig. 10　Attention weights diagram of deep features in model 2

结束语　点击率预估的准确性和解释性对推荐系统的研发有正向的指导作用，能够用于优化数据存储与计算，以及评估与监控等系统的设计。考虑到学术界对深度点击率预估模型解释性的研究较少，因此本文针对该场景在模型解释性方面的不足，提出了基于注意力机制增强的 DeepFM 方法，用于提升模型的精度和解释性。基于多头自注意力机制增强的 DeepFM 模型对特征嵌入层、线性特征交互部分、深度部分均通过多头注意力机制进行增强。为了得到解释性更好的模型，我们还提出通过叠加多个注意力机制来替代深度部分 DNN 模型的方法。

本文在多个数据集上进行了实验验证，结果表明，本文模型可以充分学习特征交互，并且可以给模型带来更好的解释能力。这说明基于多头自注意力机制的方式进行深度点击率预估模型解释是可行的。未来的工作中，我们计划探索将多头自注意力机制和更多的深度点击率预估模型结合，使更多的深度模型具备可解释性。

参 考 文 献

[1] XIANG L. Recommender system practice [M]. People Post Press, 2012.

[2] WANG Z. Deep Learning Recommender System[M]. Electronic Industry Press, 2020.

[3] GUO H, TANG R, YE Y, et al. DeepFM: a factorization-machine based neural network for CTR prediction[J]. arXiv: 1703.04247, 2017.

[4] LIU M J, ZENG G C, YUE W, et al. Review on click-through rate prediction models for display advertising [J]. Computer Science, 2019, 46(7): 38-49.

[5] CHEN T, GUESTRIN C. Xgboost: A scalable tree boosting system[C]// Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining. 2016: 785-794.

[6] HE X J, PAN W J, CHENG H. An advertisement click-through rate prediction model based on ensemble learning[J]. Computing Engineering & Science, 2019, 41(12): 2278-2284.

[7] RENDLE S. Factorization machines [C]// 2010 IEEE International Conference on Data Mining. IEEE, 2010: 995-1000.

[8] JUAN Y, ZHUANG Y, CHIN W S, et al. Field-aware factorization machines for CTR prediction[C]// Proceedings of the 10th ACM Conference on Recommender Systems. 2016: 43-50.

[9] HE X, PAN J, JIN O, et al. Practical lessons from predicting clicks on ads at facebook[C]// Proceedings of the Eighth International Workshop on Data Mining for Online Advertising. 2014: 1-9.

[10] ZHANG W, DU T, WANG J. Deep learning over multi-field categorical data[C]// European Conference on Information Retrieval. Cham: Springer, 2016: 45-57.

[11] CHENG H T, KOC L, HARMSEN J, et al. Wide & deep learning for recommender systems [C] // Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. 2016: 7-10.

[12] WANG R, FU B, FU G, et al. Deep & cross network for ad click predictions[M]// Proceedings of the ADKDD' 17. 2017: 1-7.

[13] LIAN J, ZHOU X, ZHANG F, et al. xdeepfm: Combining explicit and implicit feature interactions for recommender systems [C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2018: 1754-1763.

[14] NING T, MIAO D Z, DONG Q W, et al. Wide and deep learning for default risk prediction[J]. Computer Science, 2021, 48(5): 197-201.

[15] DING Y, LEI X, LIAO B, et al. MLRDFM: a multi-view Laplacian regularized DeepFM model for predicting miRNA-disease associations [J]. Briefings in Bioinformatics, 2022, 23(3): bbac079.

[16] CAO B Q, XIAO Q X, ZHANG X P, et al. An API service recommendation method via combining self-organization map-based functionality clustering and deep factorization machine-based quality prediction [J]. Chinese Journal of Computers, 2019, 42(6): 1367-1383.

[17] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems, 2017: 6000-6010.

[18] HAN K, WANG Y, CHEN H, et al. A survey on vision trans-

former[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45(1): 87-110.
[19] XIAO J, YE H, HE X, et al. Attentional factorization machines: Learning the weight of feature interactions via attention networks[C]// Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017: 3119-3125.
[20] SONG W, SHI C, XIAO Z, et al. AutoInt: Automatic feature interaction learning via self-attentive neural networks[C]// Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019: 1161-1170.
[21] 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. 2019: 169-177.
[22] YU R, YE Y, LIU Q, et al. Xcrossnet: Feature structure-oriented learning for click-through rate prediction[C]// Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer, 2021: 436-447.
[23] WANG R, SHIVANNA R, CHENG D, et al. Dcn v2: Improved deep & cross network and practical lessons for web-scale learning to rank systems[C]// Proceedings of the Web Conference 2021. 2021: 1785-1797.
[24] CHEFER H, GUR S, WOLF L. Transformer interpretability beyond attention visualization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 782-791.
[25] HAO Y, DONG L, WEI F, et al. Self-attention attribution: Interpreting information interactions inside transformer[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(14): 12963-12971.
[26] LIN Z, FENG M, SANTOS C N, et al. A structured self-attentive sentence embedding[C]// Proceedings of the International Conference on Learning Representations. 2017.
[27] TU D D, SHU C C, YU H Y. Using unified probabilistic matrix factorization for contextual advertisement recommendation[J]. RuanJian Xue Bao/Journal of Software, 2013, 24(3): 454-464.
[28] HE X, CHUA T S. Neural factorization machines for sparse predictive analytics[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017: 355-364.
[29] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv: 1409.0473, 2014.
[30] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.
[31] LIANG B, LIU Q, XU J, et al. Aspect-based sentiment analysis based on multi-attention CNN[J]. Journal of Computer Research and Development, 2017, 54(8): 1724-1735.
[32] WANG W G, SHEN J B, JIA Y D. Review of visual attention detection[J]. Ruan Jian Xue Bao/Journal of Software, 2019, 30(2): 416-439.
[33] PARK D H, HENDRICKS L A, AKATA Z, et al. Multimodal explanations: Justifying decisions and pointing to the evidence [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 8779-8788.
[34] CHOI E, BAHADORI M T, SUN J, et al. Retain: An interpretable predictive model forhealthcare using reverse time attention mechanism[C]// Advances in Neural Information Processing Systems. 2016: 3512-3520.
[35] LEE K, ORTEN B, DASDAN A, et al. Estimating conversion rate in display advertising from past erformance data[C]// Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2012: 768-776.
[36] REN K, ZHANG W, RONG Y, et al. User response learning for directly optimizing campaign performance in display advertising [C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. 2016: 679-688.

YANG Bin, born in 1986, Ph. D, is a member of China Computer Federation. His main research interests include recommended algorithm and natural language processing.

(责任编辑:喻藜)