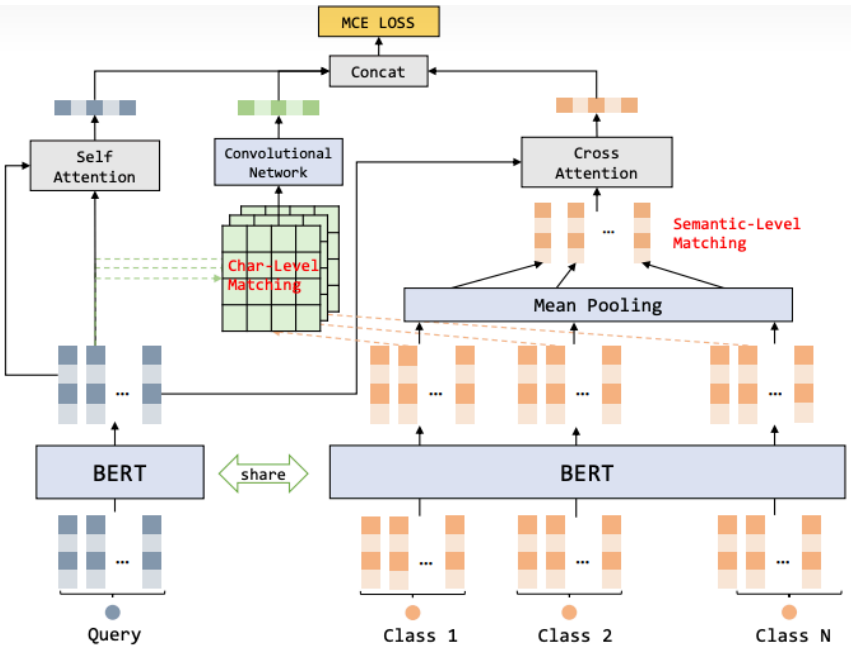
知乎



京东2023-《搜索意图分类的多粒度匹配注意力网络》论文阅读

SmartMindAI
专注搜索、广告、推荐、大模型和人工智能最新技术，欢迎关注我
已关注

8 人赞同了该文章

论文《A Multi-Granularity Matching Attention Network for Query Intent Classification in E-commerce Retrieval》

Introduction

本文讨论了在线购物+中电子商务搜索系统对查询意图分类的需求。现有的多标签分类模型在电子商务应用中效果不佳，因为查询通常很短且对词序不敏感。为了解决这个问题，我们提出了一种新的查询意图分类模型MMAN，它包括三个模块以全面提取特征，并减轻查询和类别之间的表达差距。该模型能够解决长尾查询意图分类中的挑战。本文的主要贡献如下： 1. 建立了新的数学模型，通过该模型可以预测某类现象的发生概率。 2. 通过实验验证了该模型的准确性和有效性。 3. 提出了一种新的算法，可以快速有效地处理大规模数据集。

- 提出了一种新策略，旨在通过明确地引入类别信息来缩小查询和类别之间的表达差距。该策略通过使用特定算法，将类别信息扩展到查询中，从而提高了查询的准确性。这种方法有望在各种应用中提高查询性能。
- 设计了一个模型MMAN，包含自匹配、字符级匹配和语义级匹配三个模块，旨在提高查询表示学习、增强长尾查询和消除语义歧义。该模型对输入进行分割并提取特征，再利用神经网络+进行匹配和表示学习。

Model

图展示了模型的组件，主要由四个模块组成：（1）查询和类别表示学习模块；（2）自我匹配模块；（3）字符级匹配模块；（4）语义级匹配模块。该模型通过这四个模块实现。

知乎

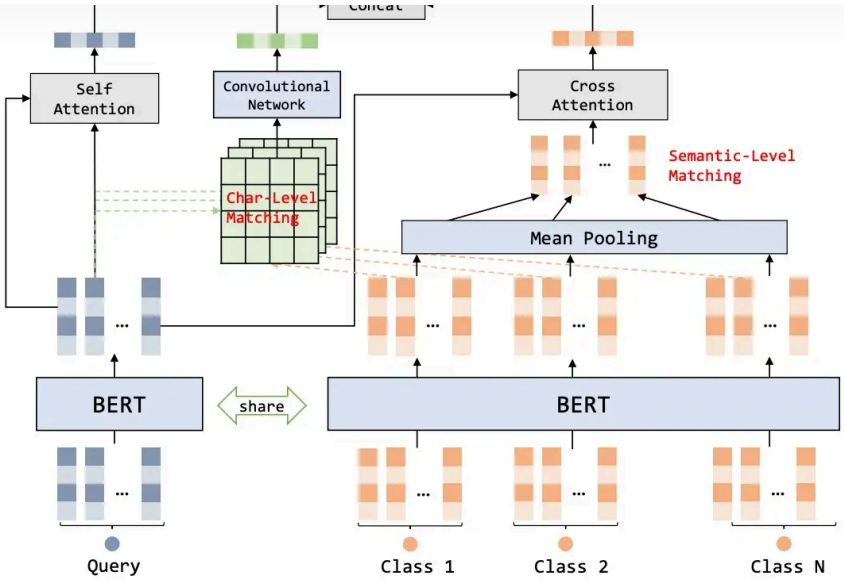


图 1: Multi-granularity Matching Attention Network.

Query and Category Representation

查询和类别表示对齐基础在于两者到同一语义空间*的映射。BERT广泛应用于工业应用中，我们使用BERT作为查询和类别的编码器。类别字符序列由两部分组成：类别名称和核心产品词。高质量产品词与类别名称拼接后输入BERT进行编码。查询和类别共享BERT模型以映射到同一语义空间。

$$\mathbf{Q}_i = \mathbf{BERT}_{\mathbf{Token}}([x_1, x_2, \ldots, x_{L_q}]),$$
$$\mathbf{C}_j = \mathbf{BERT}_{\mathbf{Token}}([n_1, n_2, \ldots, n_{L_n}, m_1, m_2, \ldots, m_{L_m}]),$$

本文研究了BERT最后一层嵌入矩阵在查询和类别令牌嵌入中的应用。其中BERT*Token不包括CLS，查询和类别令牌嵌入矩阵分别为$\mathbf{Q} * i \in \mathbb{R}^{L_q \times d}$和$\mathbf{C} * j \in \mathbb{R}^{L_c \times d}$。通过应用这些嵌入矩阵，实现了查询和类别令牌之间的映射，提高了系统的性能。

Self-matching module

研究了文本分类模型，利用自注意力机制*对查询嵌入矩阵进行概括，提取对表示查询重要的意图相关词。该模型建立在纯查询文本上，具有显著优势。

$$\mathbf{u}_i = \mathbf{v}_i \tanh\left(\mathbf{W}_q \mathbf{Q}_i^T\right),$$
$$\mathbf{q}_i = \sum_{t=1}^{L_q} \mathbf{Q}_{i,t} \mathbf{softmax}(\mathbf{u}_{i,t}),$$

基于评分函数确定组成当前查询的句子表示中单词重要性的方法。其中，$\mathbf{v}_i \in \mathbb{R}^{1 \times d}$，$\mathbf{W}_q \in \mathbb{R}^{d \times d}$，$\alpha = \mathbf{softmax}(\mathbf{u}_{i,t})$是相关变量，用于构建查询表示。通过应用评分函数，可以确定单词在句子中的重要性，进而对查询进行更有效的表示。该方法有望提高搜索和问答系统的性能。

Char-level matching module

长尾查询情况下，模型缺乏足够训练样本来精确预测用户意图。通过提取查询和类别之间的细粒度交互特征，利用点积运算*，并堆叠查询表示和类别表示在通道维度上，补充辅助知识可促进模型决策。

$$\mathbf{M}_j = \mathbf{Q}_i \mathbf{W}_{qc} \mathbf{C}_j^T,$$
$$\mathbf{M} = [\mathbf{M}_1, \mathbf{M}_2, \ldots, \mathbf{M}_C],$$

阵、特征图以及卷积⁺模块的应用，进一步提高了任务识别和分类的性能。

$$\mathbf{s}_{i,j}^{(k)} = ReLU\left(\sum_{a=0}^{r_w}\sum_{b=0}^{r_h}\mathbf{W}_{a,b}\mathbf{M}_{i+a,j+b}^{(k)} + \mathbf{b}\right),$$

研究了卷积神经网络⁺中的特征图，每个通道进行了卷积操作，随后用二维最大池化层提取最重要的特征。其中k为特征图$\mathbf{M}$的第k个通道，$\mathbf{W}_{a,b}$为一个卷积核⁺，$\mathbf{b}$为偏置向量。 卷积和最大池化都是深度学习中常用的技术，它们对于从数据中提取有用的特征以及优化网络性能至关重要。同时，该方法还提高了网络的性能和稳定性。 为了确保结果的准确性，建议进一步实验验证该方法的有效性。此外，还可以考虑将该方法与其他深度学习技术相结合，以进一步提高模型的性能。

$$\tilde{\mathbf{s}}_{i,j}^{(k)} = \underset{0\le c\le p_w}{max}\ \underset{0\le d\le p_h}{max}\ \mathbf{s}_{i+c,j+d}^{(k)},$$

对于二维最大池化在特征提取⁺中的应用，通过分析p_w和p_h对最终特征图的影响。将输出展平并通过线性变换⁺层映射到低维空间中，得到$\mathbf{Z_1} \in \mathbb{R}^{|C|\times d}$，其中包含查询和每个类别之间的细粒度⁺匹配特征。

Semantic-level matching module

对于字面匹配特征可能不足以捕获用户真实意图的问题，因为查询词可能是多义的。通过获取语义级别的类别表示，对类别表示的时间步长进行平均池化⁺，并将每个类别表示堆叠在一起，能够捕获查询和类别之间的语义相关性。这对于跨类别检索尤为重要。

$$\mathbf{c}_i = \mathbf{mean}(\mathbf{C}_i),$$
$$\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_{|C|}],$$

其中$\mathbf{C} \in \mathbb{R}^{|C|\times d}$表示所有类别的表示。通过应用交叉注意力层，查询和所有类别的表示被整合，有助于提高分类性能。

$$\mathbf{Z_2} = \mathbf{Q}_i^T\mathbf{softmax}(\mathbf{C}\mathbf{W}_{qs}\mathbf{Q}_i^T),$$

其中$\mathbf{W}_{qs} \in \mathbb{R}^{d\times d}$是可训练的权重，$\mathbf{Z_2} \in \mathbb{R}^{|C|\times d}$是查询和标签之间在语义层面上的匹配特征。通过这些特征，可以更准确地识别查询和标签之间的关系。

Training and Inference

经过上述过程，我们得到了查询自我表示$q*i$、细粒度查询-类别匹配特征$Z*1$和粗粒度匹配特征$Z*2$。这些表示用于预测用户意图，矩阵乘法⁺融合了它们。通过引入非线性变换层，特征得到了非线性变换⁺。具体定义为

$$\hat{y} = \mathbf{W}_x^T ReLU\left(\mathbf{q}_i\mathbf{W}_{qf} + [\mathbf{Z}_1, \mathbf{Z}_2]\mathbf{W}_z\right),$$

对于多标签交叉熵⁺损失在查询分类任务中的应用。我们使用线性变换矩阵⁺$\mathbf{W}_{qf}$、$\mathbf{W}_z$和$\mathbf{W}_x$进行查询分类。真实标签$y \in \mathcal{R}^{|C|}$，其中$y_i = 0, 1$表示查询是否属于类别i。框架采用多标签交叉熵损失进行训练，损失函数⁺公式如下。

$$\mathcal{L} = -\sum_{c=1}^{C} y^c \log(\sigma(\hat{y}^c)) + (1-y^c)\log(1-\sigma(\hat{y}^c)),$$

Experiment

Dataset

本文通过在两个大型真实数据集上实验验证了MMAN的有效性和通用性，数据集统计信息详见表

知乎

Statistic	Scene Data		Category Data	
	Train	Test	Train	Test
Queries	4,459,214	9,877	4,593,037	9,877
Total Labels	8	8	90	90
Avg. chars	7.63	5.00	7.69	5.00
Avg. # of labels	1.04	1.67	1.19	1.77
Min. # of labels	1	1	1	1
Max. # of labels	7	3	26	21

- 通过抽取查询和点击产品数据，评估了MMAN的性能。类别数据被用作意图分类，通过归一化点击频率并计算类别概率的累积分布函数（CDF），过滤不可靠类别。当CDF大于0.9时，低概率类别被移除。
- 通过收集八个不同领域的场景数据，如旅游、酒店预订、医疗咨询、汽车服务等，形成场景数据集。查询类别映射到领域，并由领域专家进行标注，包括查询所属的所有类别。与训练数据不同，测试数据集具有更高的准确性。

Baseline Models

本文比较了MMAN与几个强大的基线，包括广泛使用的多标签分类方法。介绍了多标签文本分类基线，如 RCNN、XML-CNN、LEAM 和 LSAN；也介绍了查询意图分类基线，如 PHC、DPHA、BERT 和 SSA-AC。最后指出使用BERT微调用户的意图是本研究的有效基线。

Experiment Settings

基于Tensorflow实现模型，提取字符级特征映射，使用Adam算法和学习率设为5e-5，最大长度为16。标签阈值设置为0.5，以评估查询意图分类的微观和宏观精确度、召回率和F1分数。

Experimental Results and Analysis

MMAN在查询意图分类和多标签分类模型比较中表现出显著优势，适用于长文本上下文建模。MMAN模型能够处理缺乏上下文信息的短查询，提高微观和宏观F1得分约3%。所有组件相互提供补充信息，是意图分类所必需的。

Online Evaluation

在生产环境中部署MMAN之前，通常在京东搜索引擎上随机部署MMAN作为测试组，并监控其性能与先前部署的模型进行比较。在线评估使用业务指标，如页面浏览量（PV）、产品点击量（Click）、总商品价值（GMV）、UV值和用户转化率（UCVR）。

与基准组相比，新模型显著改善了PV和Click指标，表明新模型召回的增量类别是用户所需的，且提高相关类别的召回率导致用户查看和点击更多产品。随着产品选择增加，转化率提高，GMV和UCVR提升（+0.351%）。

知乎

Models	Scene Data						Category Data					
	Micro			Macro			Micro			Macro		
	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
RCNN [?]	94.14	77.67	85.11	83.09	86.01	83.69	69.76	54.03	60.89	70.51	62.42	62.15
XML-CNN [?]	94.73	76.00	84.34	80.87	86.47	81.91	66.73	56.36	61.11	68.08	64.15	62.12
LEAM [?]	94.19	68.46	79.29	88.84	78.60	82.84	72.67	49.91	59.18	69.96	47.56	52.15
LSAN [?]	94.73	74.14	83.18	80.31	86.05	81.48	68.33	51.36	58.64	71.64	61.00	61.93
PHC [?]	94.63	77.93	85.47	83.17	86.62	83.74	60.12	**59.41**	59.76	64.08	64.90	60.67
DPHA [?]	95.23	77.43	85.41	82.01	84.35	82.06	71.55	54.06	61.58	75.39	54.99	61.83
SSA-AC [?]	94.82	78.15	85.68	84.15	84.26	83.92	72.36	53.20	61.32	74.38	62.19	63.38
MMAN	**95.52**	**82.26**	**88.39**	**87.26**	86.15	**85.93**	**75.64**	55.07	**63.74**	**75.77**	64.56	**66.47**
w/o self-matching	**96.03**	81.24	88.02	**88.14**	85.72	84.86	75.25	54.35	63.11	73.26	64.08	65.68
w/o char matching	95.16	80.28	87.09	82.12	**89.38**	83.74	68.72	**57.13**	62.39	72.16	62.58	65.12
w/o semantic matching	95.86	81.14	87.89	84.36	87.62	84.15	72.18	55.16	62.17	73.61	64.27	65.05
BERT [?]	95.39	79.22	86.56	81.20	88.48	83.00	65.88	56.23	60.67	68.47	**67.28**	64.53

Conclusion and Future Work

提出了一种多粒度匹配注意力网络，从查询-类别交互矩阵的字符级和语义级全面提取特征，显著改进了长尾查询，消除表达差异，A/B实验带来商业价值，未来工作将探索利用外部知识以提高模型性能。

发布于 2023-10-24 21:28 · IP 属地北京

搜索引擎 意图识别 京东