为电子商务营销寻找相似客户

Yang Peng
yang.peng@walmart.com
Walmart Global Tech

Changzheng Liu
changzheng.liu@walmart.com
Walmart Global Tech

Wei Shen
wei.shen@walmart.com
Walmart Global Tech

arXiv:2301.03147v2 [cs.LG] 2 Jul 2023

摘要

以客户为中心的营销活动为沃尔玛电子商务网站带来了很大一部分流量。随着客户数据规模不断扩大，拓展营销受众以触达更多客户，对于电子商务公司推动业务增长并为客户创造更多价值变得愈发重要。本文介绍了一个可扩展且高效的系统，用于拓展营销活动的目标受众，可处理数亿客户。我们使用基于深度学习的嵌入模型来表示客户，并采用近似最近邻搜索方法，快速找到感兴趣的相似客户。通过构建可解释且有意义的客户相似度指标，该模型能够应对各种业务兴趣。我们进行了大量实验，以证明该系统和客户嵌入模型具有出色的性能。

1 引言

在客户关系管理（CRM）系统中，获取和留存客户对营销成功至关重要。扩大目标客户群是客户关系管理系统中客户获取和留存的一项重要工作。本文探讨如何构建一个大规模营销受众扩展系统，旨在通过为客户关系管理电子邮件和推送营销活动寻找更多客户，推动电子商务业务增长。

具体而言，本文要解决的问题是：给定某项营销活动的一组现有客户（种子客户集），如何找到更多与这些种子客户相似的客户，从而增加收入并为沃尔玛电子商务业务带来更多流量。

我们面临的最大挑战之一是数据规模。仅在美国市场，沃尔玛就拥有数亿活跃客户。每位客户可能有数百甚至数千个特征，而且客户数据逐年快速增长。因此，构建一个可扩展、高效的系统来处理不断增长的大规模客户数据，是我们的首要任务。除了可扩展性之外，在寻找相似客户时，我们还需要考虑方法的可解释性。模型的可解释性不仅有助于向业务合作伙伴说明我们的方法如何运作，也为评估系统质量提供了直观的方式。

本文提出了一种可扩展且高效的营销活动受众扩展系统，可处理数亿客户。

- 我们的系统能够生成低维稠密嵌入来表示客户。
- 在客户嵌入空间中，我们使用两个客户嵌入向量之间的余弦距离来估算两个客户之间的相似性。
- 我们设计的相似度指标可以衡量不同的业务兴趣（例如购买行为、访问和互动），这些都是可解释且有意义的营销活动业务目标。
- 我们使用近似最近邻搜索，在客户嵌入空间中快速找到与种子客户相似的客户。

为了提升客户嵌入模型的质量，我们的模型会从不同数据源摄取多模态特征，例如交易、访问、互动和客户元数据。多模态融合技术通过利用不同类型数据之间的互补和相关关系，在信息检索、信息抽取和分类等任务中展现出显著优势 [4, 17, 18, 20–26]。我们的嵌入模型还可以通过融合不同类型的数据，提升受众扩展的效果。

我们的贡献如下：

- 我们提出了一种高效有效的营销受众扩展系统，可处理数亿客户。
- 我们使用深度学习模型生成客户嵌入。该深度学习模型能够处理稠密数值特征和稀疏类别特征，并通过迁移学习对位置嵌入进行编码。
- 我们的系统能够针对不同的营销活动和业务目标构建不同的相似度指标，因而具有很强的适应性。这些相似度指标既易于解释，也具有实际意义。
- 我们基于 FAISS 开发了一种可扩展且高效的近似最近邻搜索方法，能够快速找到相似客户。
- 我们利用各种数据源设计了多模态特征。
- 我们进行了大量实验，以证明系统和嵌入模型的可扩展性、效率和质量。

概述第 2 节讨论相似人群建模和受众扩展系统的相关工作。第 3 节介绍我们的系统概述。第 4 节阐述嵌入模型。第 5 节通过大量实验展示我们系统的有效性和效率。第 6 节讨论相似人群模型的结论和未来工作。

2 相关工作

本节将讨论受众扩展系统、表示模型和快速相似性搜索方法方面的先前工作。在文献综述中，我们主要关注行业解决方案中的相关工作，因为我们处理的是大规模乃至互联网规模的数据集，而这类数据集在学术研究中极少受到研究。

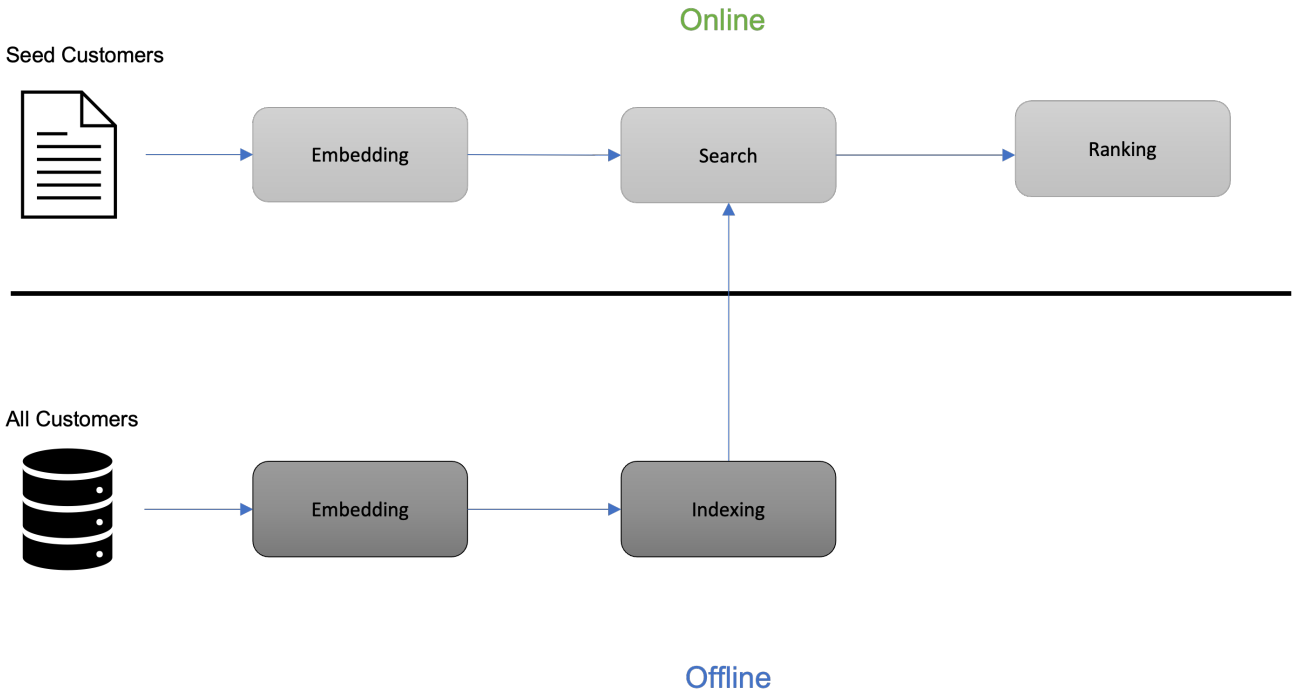

图 1：Walm 中的受众扩展系统 **art.**

2.1 受众扩展

在 [16],Ma 等人讨论了营销活动受众扩展的三类方法：基于相似性的方法、基于回归的方法和基于细分的方法。他们在雅虎广告平台提出了一种基于图的相似用户系统，兼具简单相似性方法和基于回归的方法的优点。在 [14], Liu 等人在领英开发了两种实现受众扩展的方法：基于用户属性、与营销活动无关的扩展，以及通过最近邻搜索实现的针对营销活动的扩展。在 [8], Jiang 等人讨论了用于寻找相似用户的基于规则、基于相似性和基于模型的方法，并在明略科技提出了一种用于受众扩展的深度神经网络分类模型。在 [3], deWet 等人提出了一种分两阶段的、基于嵌入的受众扩展模型，并将其部署在 Pinterest 的生产环境中。第一阶段，他们基于全站用户活动日志训练了一个全局用户嵌入模型。第二阶段，他们使用统计技术，在嵌入空间中为每位广告主创建轻量级的种子列表表示。

在我们的系统中，我们首先使用基于相似度的方法，从整个客户全集中搜索相似客户，然后根据相似度分数或单独的分类模型对这些新客户进行排序。第一步选择基于相似度的方法，是因为它具有出色的可扩展性和较低的搜索延迟。基于相似度的方法通常需要先构建用户表示。下一节中，我们将讨论用于表示客户的嵌入模型。

2.2 用户表示

嵌入模型在将客户的高维稀疏特征向量转换为低维稠密表示方面非常有用。嵌入模型已在业界得到广泛应用，例如用于沃尔玛的搜索引擎营销 [9–11], 、Airbnb 的搜索排序 [5], 以及 Pinterest 的推荐 [19]。嵌入模型可以通过训练来捕捉客户之间的相似性，使我们能够使用近似最近邻搜索方法，快速找到与种子客户相似的客户。在我们的案例中，我们使用双塔架构训练客户嵌入模型，这种架构已在以往研究中得到广泛认可。我们在用户表示方面的创新之处，是利用两个嵌入向量之间的余弦相似度对业务指标进行建模，这具有很强的可解释性。

2.3 相似度搜索

在表示用户之后，我们需要一种快速方法，从规模可能达数亿客户的庞大客户全集中搜索相似客户。扫描整个客户全集既无法扩展，也不够高效 [1, 15, 27–29]。近似最近邻搜索是以往工作中最常用的方法。例如，以往工作曾使用局部敏感哈希（LSH） [3, 14, 16]。有多种优秀的近似最近邻搜索开源工具 [7, 13, 30–34] ，例如谷歌的 ScaNN [6] 和脸书的 FAISS [12] 。我们选择 FAISS 来搜索相似客户，因为它可扩展至处理数十亿个向量，并支持多种距离度量（例如点积、余弦距离和欧几里得距离）。

3 系统概述

本节将介绍我们在沃尔玛用于电子商务营销、寻找相似客户的受众扩展系统流程。系统示意图见图 1。我们的受众扩展系统包含在线阶段和离线阶段。离线阶段中，我们为客户全集中的所有客户生成客户嵌入，然后在客户嵌入空间中建立索引。在线阶段中，种子客户会被转换为嵌入，随后我们使用离线阶段预先构建的索引搜索相似客户。找到相似客户后，我们会对其进行筛选和排序。排序方法可以基于其相似度分数，也可以采用其他分类模型。

3.1 嵌入

本文主要关注嵌入模型，并将在后续章节说明我们如何构建该模型。客户嵌入模型可为客户生成统一的稠密表示。我们的客户嵌入模型可用于许多场景。除了寻找相似客户外，客户嵌入还可作为其他客户模型的输入特征，例如购买倾向模型和客户终身价值模型。本文不重点讨论排序方法，我们将在未来的工作中对此进行研究。

3.2 索引与搜索

该系统面临的可扩展性问题是，如何从数亿名候选客户中快速搜索相似客户。为了缩小待考虑客户的范围，我们使用 FAISS [12] 构建索引。选择 FAISS 有以下几个原因。FAISS 支持多种索引，并提供 CPU 和 GPU 实现。FAISS 还支持多种距离度量，例如欧几里得距离、点积和余弦相似度。此外，我们可以利用 FAISS 中的压缩技术处理无法全部载入内存的大型数据集。本文不重点讨论索引和搜索的实现方式。如果您想进一步了解 FAISS，请访问其 GitHub 项目页面。

4 嵌入模型

本节介绍我们基于深度学习的嵌入模型，该模型能够捕捉客户之间的相似性。要设计这一模型，我们首先需要定义两个客户之间的相似度指标。以往的一些研究中的模型 [3]学习了正负客户对之间的相对相似度分数，而我们采用两个客户之间的直接相似度指标，这些指标具有可解释性、意义明确，并能反映业务指标。我们定义的相似度指标对于提升和解释营销活动效果非常有用。

我们采用双塔架构计算两个客户嵌入向量之间的余弦距离，并将余弦距离作为衡量两个客户之间相似性的相似度指标估计值。客户嵌入模型经过训练，以最小化训练数据集中余弦距离与真实相似度分数之间的总损失。

4.1 相似度指标定义

在沃尔玛，我们关注许多不同的营销活动业务指标，例如交易、网站访问和营销活动互动。为营销活动拓展受众时，找到在沃尔玛电子商务网站上行为与现有种子客户相似的新客户，能够带来业务优势。找到业务指标相似的新客户，可使营销活动在拓展后保持相似的客户分布，这对于冷启动营销活动或转化营销活动非常有利。

我们的营销活动尤其关注三类业务指标：交易、访问和互动。下面以交易为例，说明我们如何定义相似度指标。假设沃尔玛目录中有一系列产品类别，c_1，c_2，……，c_n。客户 A 在这些类别中下过采购订单，$O_A = (a_1, a_2, \ldots\ldots, a_n)$。客户 B 也在这些类别中下过采购订单，$O_B = (b_1, b_2, \ldots\ldots, b_n)$。A 和 B 之间的相似性定义如下：

$$similarity(A, B) = cosine_similarity(O_A, O_B) \quad (1)$$

我们也可以使用其他相似度距离函数，例如雅卡尔相似度和欧几里得距离。这种计算相似度指标的方法也可用于访问和互动。

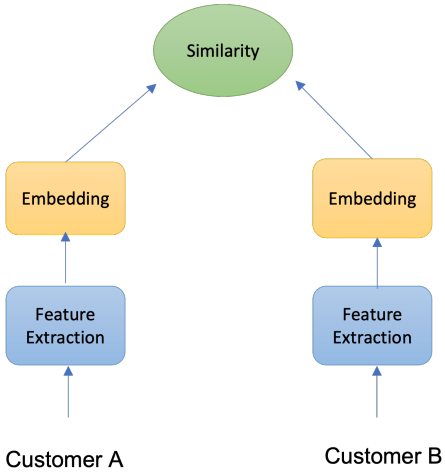

图2：双塔模型架构。

4.2 双塔架构

定义客户之间的相似度指标后，下一项任务是构建一个机器学习模型，根据客户特征预测相似度指标。我们的方法如下：

（1）提取客户的原始特征；（2）将客户特征转换为客户嵌入；（3）计算嵌入空间中客户嵌入对之间的余弦距离；（4）将余弦距离用作客户对真实相似度分数的估计值。

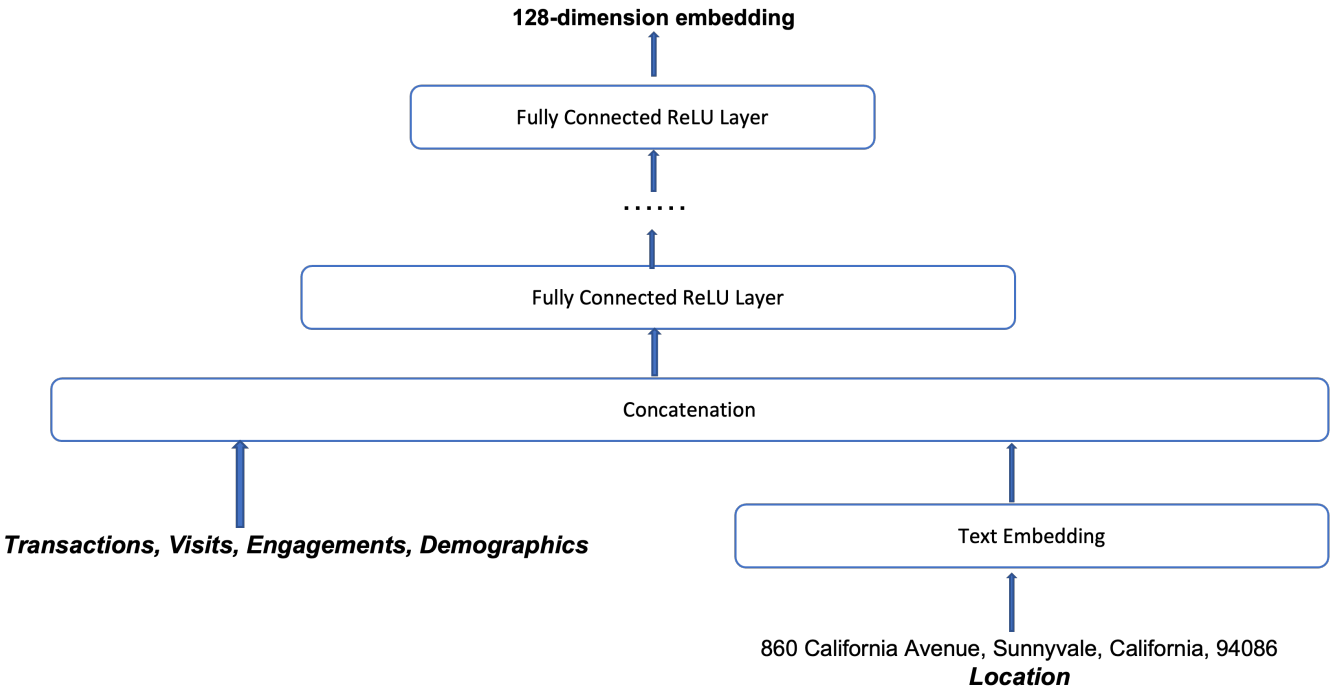

图 3：嵌入模型。

该流程如图2所示。用于优化的损失函数是余弦距离预测值与真实相似度指标之间的L1损失。

多模态客户特征从多个数据源中提取，包括交易数据、访问数据、互动数据和人口统计数据。多模态客户特征由稠密数值特征（例如订单数、商品交易总额、访问次数）和稀疏类别特征（例如性别、教育程度、职业、位置）构成。对于低维类别特征，我们可以使用独热编码将其转换为数值，例如性别和教育程度。位置特征包含街道地址、城市名称、州名称和邮政编码，因此它是一个维度非常高的类别特征，使用独热编码的效率过低。我们通过迁移学习将位置特征转换为位置嵌入，然后与其他特征拼接起来，下一部分将对此进行说明。

4.3 嵌入模型结构

嵌入模型如图3所示。我们将交易、访问、互动、人口统计和位置特征作为嵌入模型的输入。位置特征被视为文本句子，然后通过迁移学习，利用预训练文本嵌入模型将其转换为位置嵌入。随后，我们将数值特征和位置嵌入拼接起来，作为最终前馈神经网络的输入。该前馈神经网络由多个全连接ReLU层组成。前馈神经网络的输出是一个128维嵌入，作为客户表示。

4.3.1 位置嵌入。我们尝试了几种不同的方法，将位置文本转换为位置嵌入。一种方法是使用 PyTorch 中预训练的词嵌入模型（GloVe），如图4所示。另一种方法是使用先进的 BERT 模型 [2] 进行文本表示学习，如图5所示。我们还在训练过程中对预训练的BERT模型进行了微调。

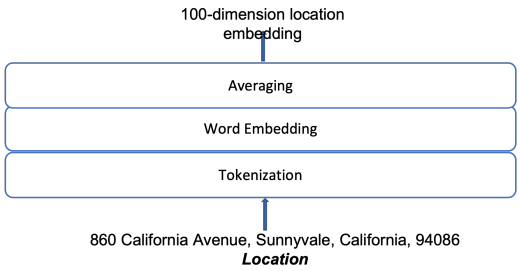

图 4：使用词嵌入生成位置嵌入。

5 实验结果

为了进行评估，我们按照不同的时间窗口设置训练、验证和测试数据集。例如，我们可以使用过去 n 年的数据作为训练数据，并使用接下来一个月或一个季度的数据作为验证和测试数据。用于评估模型质量的指标是平均绝对误差（MAE）。由于沃尔玛的隐私政策，

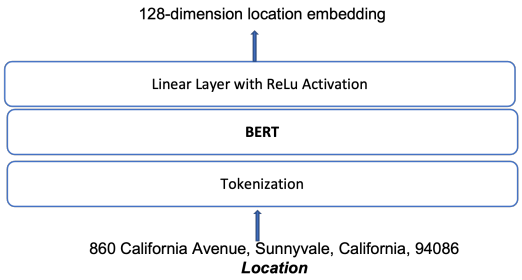

图5：使用 BERT 进行位置嵌入。

结果以相对于基线嵌入模型的百分比形式呈现。

在本节中，我们比较不同模型配置的质量和推理时间，涵盖全连接层数量的差异以及位置嵌入方法的差异。基线模型使用两个全连接ReLU层，且不包含位置嵌入。结果见表1和表2。

表1：不同配置下嵌入模型的质量（平均绝对误差）。

	无位置信息	词嵌入	BERT
2层	100%	97%	91%
3 层	94%	87%	86%

表2：不同配置下嵌入模型的推理时间。

	无位置信息	词嵌入	BERT
2 层	100%	110%	>500%
3 层	105%	115%	>500%

在表1和表2中，百分比越低，表示模型性能越好。尽管采用 BERT 的嵌入模型在模型质量方面表现最佳，但其推理延迟是基线的 4 倍，速度非常慢。考虑到我们需要离线处理数亿客户，使用 BERT 的系统总运行时间并不理想。采用词嵌入的嵌入模型在质量和推理延迟之间取得了良好的平衡。

6 结论

本文提出了一种高效、有效的营销受众扩展系统，可处理数亿客户。我们使用深度学习模型生成客户嵌入。该深度学习模型能够处理密集数值特征和稀疏类别特征，并通过迁移学习对位置嵌入进行编码。通过针对不同营销活动构建不同的相似度指标，我们的系统具有很强的适应性。这些相似度指标具有可解释性和实际意义，能够反映不同的业务指标。我们开展了大量实验，以验证系统和嵌入模型的可扩展性、效率和质量。

我们的市场营销受众扩展系统未来有几个研究方向。首先，我们可以探索结合基于相似性和基于分类的方法来搜索相似客户。其次，我们可以研究如何利用机器学习模型筛选并排序相似客户，以提高排序质量。

参考文献

[1] S Amirrahmat、KA Alshibli、MF Jarrar、B Zhang 和 RA Regueiro。2018年。基于砂土三维颗粒运动学测量的等效连续体应变计算。*International Journal for Numerical and Analytical Methods in Geome- chanics* 42 2018 999 1015 8 10 1002 2779 ( ), – 。第 期。[2] https://doi.org/ 。/nag。Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2018年。BERT：用于语言理解的深度双向变换器预训练。https://doi.org/ 10 48550 1810 04805 。/ARXIV。 。[3] Stephanie deWet 和 Jiafan Ou。 。寻找行为相似的用户：通过迁移学习扩展广告主受众。载于《第*25*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集》。2251–2259。[4] Dihong Gong、Daisy Zhe Wang 和 Yang Peng。2017年。用于网页信息抽取的多模态学习。载于《第*25*届*ACM*多媒体国际会议论文集（*MM '17*）》。计算机协会，纽约，纽约州，美国，288–296。[5] Mihajlo Grbovic 和 Haibin Cheng。2018年。利用嵌入实现 Airbnb 搜索排序的实时个性化。载于《第*24*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集》。311–320。[6] Ruiqi Guo、Philip Sun、Erik Lindgren、Quan Geng、David Simcha、Felix Chern 和 Sanjiv Kumar。2020年。利用各向异性向量量化加速大规模推理。载于国际机器学习会议。https://arxiv.org/abs/ 1908 10396 。Robert A. Jarrow、Rinald Murataj、Martin T. Wells 和 Liao Zhu。2021年。低波动率异象与自适应多因子模型。*arXiv*预印本 *arXiv:2003.08302* (2021年)。[8] Jinling Jiang、Xiaoming Lin、Junjie Yao 和 Hua Lu。2019年。基于端到端神经预测的全面受众扩展。载于*CEUR*研讨会论文集，第2410卷。CEUR研讨会论文集。[9] Cheng Jie、Zigeng Wang、Da Xu 和 Wei Shen。2022年。用于电子商务搜索引擎营销系统的基于聚类的多目标竞价算法。《大数据前沿》(2022年)，77。[10] Cheng Jie、Da Xu、Zigeng Wang 和 Wei Shen。2022年。通过基于深度学习的页面创建提升电子商务自然搜索流量。https://doi.org/10.48550/ARXIV.2209.10792[11] Cheng Jie、Da Xu、Zigeng Wang、Lu Wang 和 Wei Shen。2021年。面向电子商务的高效基于群组的搜索引擎营销系统。https://doi.org/10.48550/ARXIV.2106.12700[12] Jeff Johnson、Matthijs Douze 和 Hervé Jégou。2019年。利用图形处理器进行十亿规模的相似度搜索。《*IEEE*大数据汇刊》7, 3 (2019年)，535–547。[13] Yifei Li、Kuangyan Song、Yiming Sun 和 Liao Zhu。2021年。FrequentNet：一种新颖的可解释深度学习图像分类模型。*arXiv*预印本 *arXiv:2001.01034* (2021年)。[14] Haishan Liu、David Pardoe、Kun Liu、Manoj Thakur、Frank Cao 和 Chongzhe Li。2016年。在线社交网络广告的受众扩展。载于《第*22*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集》。165–174。[15] Huiyang Luo、Huiluo Chen、Runyu Zhang、Yao Ren、Boning Zhang、Richard A. Regueiro、Khalid Alshibli 和 Hongbing Lu。2022年。4 - 高应变率单轴应变加载下颗粒材料的本构行为。载于《实验冲击力学进展》，Bo Song（编）。爱思唯尔，99–124。https://doi.org/10.1016/B978-0-12-823325-2.00005-4[16] Qiang Ma、Musen Wen、Zhen Xia 和 Datong Chen。2016年。次线性、大规模的相似受众扩展系统：一种大规模相似受众扩展。在《*KDD 2016*第五届大数据、数据流与异构源挖掘国际研讨会：算法、系统、编程模型与应用论文集（《机器学习研究论文集》）》*style*>中，*Wei Fan*、*Albert Bifet*、*Jesse Read*、*Qiang Yang* 和 *Philip S. Yu*（编），第*53*卷。*PMLR*，旧金山，加利福尼亚州，美国，*51–67*。*https://proceedings.mlr.press/v53/ma16.htmlMorteza Shahriari Nia*、*Christan Grant*、*Yang Peng*、*Daisy Zhe Wang* 和 *Milenko Petrovic*。*2013*年。佛罗里达大学知识库加速笔记。《第二十二届文本检索会议（*TREC 2013*）(2013年)。[18] Morteza Shahriari Nia、Christan Earl Grant、Yang Peng、Daisy Zhe Wang 和 Milenko Petrovic。2014年。面向维基百科实体的流式事实抽取，

网络规模。。载于*FLAIRS*会议。[19] Aditya Pal、Chantat Eksombatchai、Yitong Zhou、Bo Zhao、Charles Rosenberg 和 Jure Leskovec。2020年。Pinnersage：Pinterest 推荐的多模态用户嵌入框架。载于《第*26*届*ACM SIGKDD*知识发现与数据挖掘国际会议论文集》。2311–2320。[20] Yang Peng。2017年。多模态融合：理论与应用。佛罗里达大学(2017年)。[21] Yang Peng。2023年。利用问答和多模态融合进行查询驱动的知识图谱构建。载于《*ACM Web Conference 2023*伴随论文集（*WWW '23*伴随会议）》。计算机协会，纽约，纽约州，美国，

1119 1126 10 1145 3543873 3587567 – 。

[22] 2022 https://doi.org/ 。/ 。Yang Peng 和 Daisy Zhe Wang。。利用基于网页的问答和多模态融合完成知识库。https://doi.org/10.48550/ARXIV.2211.07098[23] Yang Peng 和 Daisy Zhe Wang。2022年。利用多模态知识图谱上的多模态路径融合进行查询驱动的知识库补全。https://doi.org/10.48550/ARXIV.2212.01923[24] Yang Peng、Daisy Zhe Wang、Ishan Patwa、Dihong Gong 和 Chunsheng Victor Fang。2015年。用于多模态词义消歧的概率集成融合。载于《*2015*年*IEEE*多媒体国际研讨会（*ISM*）》。IEEE，172–177。[25] Yang Peng、Xiaofeng Zhou、Daisy Zhe Wang 和 Chunsheng Victor Fang。2016年。利用多模态融合实现可扩展图像检索。载于《第二十九届*FLAIRS*国际会议》。[26] Yang Peng、Xiaofeng Zhou、Daisy Zhe Wang、Ishan Patwa、Dihong Gong 和 Chunsheng Fang。2016年。用于消歧与检索的多模态集成融合。*IEEE MultiMedia* (2016年)。[27] Christopher T Senseney、Zheng Duan、Boning Zhang 和 Richard A Regueiro。2017年。结合球形多面体离散元（DE）与有限元（FE）的计算模型，用于模拟无黏性土上的竖向板加载。*Acta Geotechnica*

12 2017 593 603 10 1007 11440 016 0519 8 ( ), – 。

[28] https://doi.org/ 。/s - - - Boning Zhang、Eric B. Herbold、Michael A. Homel 和 Richard A. Regueiro。2015年。离散元颗粒破裂模型。技术报告。劳伦斯利弗莫尔国家实验室（LLNL），利弗莫尔，加利福尼亚州（美国）。[29] Boning Zhang、Richard A. Regueiro、Andrew M. Druckrey 和 Khalid Alshibli。2018年。根据砂土SMT成像构建多椭球颗粒形状，并开发一种新的离散元接触检测算法。*Engineering Computations*

35, 2 2018 733 771 10 1108 01 2017 0026 ( ), – 。

[30] 2020 https://doi.org/ 。/EC- - - Liao Zhu。。自适应多因子模型与金融市场。eCommons。[31] Liao Zhu、Sumanta Basu、Robert A. Jarrow 和 Martin T. Wells。2020年。高维估计、基准资产与自适应多因子模型。*The Quarterly Journal of Finance* 10, 04 (2020年), 2050017。[32] Liao Zhu、Robert A. Jarrow 和 Martin T. Wells。2021年。利用自适应多因子模型进行时间不变系数检验。*The Quarterly Journal of Finance* 11, 04 (2021年), 2150019。[33] Liao Zhu、Ningning Sun 和 Martin T. Wells。2021年。微观结构度量的聚类结构。*arXiv*预印本 *arXiv:2107.02283* (2021年)。[34] Liao Zhu、Haoxuan Wu 和 Martin T. Wells。2021年。一种基于新闻的自适应资产定价机器学习模型。*arXiv*预印本 *arXiv:2106.07103* (2021年)。