



对NLP以及推荐系统未来发展趋势的看法



张俊林
你所不知道的事

已关注

318 人赞同了该文章

(注：这是InfoQ的访谈记录，欢迎大家参加QCon2019上海站)

Q:首先请您做一下简单的自我介绍，包括您的工作经历和学习经历。

A:我目前在微博工作，担任微博机器学习团队AI Lab的负责人，主要是推动业界先进技术在微博的信息流推荐业务落地。之前在阿里担任资深技术专家并负责新技术团队，也在百度凤巢做过。博士毕业于中科院软件所，主要的专业兴趣集中在自然语言处理及推荐搜索等方向上，喜欢新技术并乐于做些技术分享，之前写过两本书《这就是搜索引擎》以及《大数据日知录》，这两本书的读者评价还可以。

Q:老师现在主要研究的领域是自然语言处理（NLP）和推荐系统，能介绍一下NLP技术吗？它有哪些落地场景？

A: NLP是个非常大的技术方向，包含很多子领域，比如文本分类、文本摘要、机器翻译、阅读理解、对话机器人.....等等，很多子方向。不过不同子领域面临的核心目标是一致的：就是如何让机器真正理解自然语言文本。这个目标难度还是很高的，不说是AI的终极目标估计也差太远。当然，目前的技术水准，距离这个目标明显也差得很远。

至于落地场景，其实有许多，只是它作为技术支撑，比较靠后，不太容易被察觉出来。典型的例子比如搜索、推荐、信息流等相对成熟的应用领域，也有最近比较热的应用领域，比如对话机器人、智能客服、阅读理解等很多应用领域。一般大公司都有专门的NLP团队，这说明了NLP的应用普遍性，只是普通用户有时候很难直观体会到它的存在。

Q:似乎目前NLP领域很少出现“独角兽”企业，也有人说，自然语言处理是个大坑，那么您认为NLP落地有哪些难点？

A: NLP各种任务更偏向后台技术平台的形式，是一种能力输出，往往不会是一种业务形态，它一般是支撑某些业务形态的支撑技术。总体而言，跟用户能够直接感知到的产品形态比较，整体靠后一些，一般终端用户很难感知到它的存在。能够在产品形态方面直接感知到NLP技术的业务领域，其实很少，比如对话机器人，智能客服等这些算是比较靠前的业务形态，大多数都偏靠后一些。NLP是个技术领域，和产品形态本来也没必然的关系。所

已赞同 318

22 条评论

分享

喜欢

收藏



角兽”企业这种说法，如果说是某个独角兽企业，大量应用了NLP技术，可能会更准确一些。当然，这种独角兽企业确实也很少，因为独角兽企业本来就很少，摊上用NLP技术的更少，所以也正常。

要说落地难点的话，其实本质上就一个：就是技术不够成熟，不足以支撑好的应用体验。当然，这个分具体的子领域，有些领域，比如文本分类聚类、NER命名实体识别等一些简单任务，其实落地效果已经不错了，也被普遍应用。而对于很多高难度的应用领域，其它问题都是表面问题，本质问题其实是技术不够成熟。

上面说的这个问题有长期性，不是短期内能够解决的。如果说NLP落地短期难点的话，比如Transformer和Bert这种效果很好的新技术，可能因为模型比较重，高效的分布式大规模训练系统以及快速的在线服务，可能是影响大规模应用目前的短期障碍，但是我相信很快会有大公司开源出一些好用的系统的，所以其实问题也不大。

Q:发展到现在，自然语言处理领域出现了很多令人激动的新想法与新工具，比如号称最强 NLP 预训练模型的BERT，它的出现对NLP领域有哪些重大影响？

A: Bert的出现不仅仅是自然语言处理领域的里程碑，即使放到人工智能这个大的领域，也是非常重要的里程碑性的工作。对我而言，我觉得Bert的出现比把深度学习引入NLP还要重要些，主要原因是Bert的两阶段模式效果远远超出预期。另外，它给NLP研发者趟通并指明了一条光明大道：就是通过自监督学习，把大量非监督的文本充分利用起来，并将其中的语言知识编码，对各种下游NLP任务产生巨大的积极作用。这是非常了不起的贡献，在Bert之前，其实很多工作希望能够达到这个目的，但是应该说路并没有被趟通，而Bert证明了这条路是能走通的。

而且我的个人判断是：Bert+Transformer有可能在未来几年替代掉NLP各个不同应用领域之前五花八门的具体技术，有可能成为一个NLP领域的大一统模型，这在NLP历史上也是没有出现过的情形。总而言之，Bert的重要性再强调也不为过。当然，我这个人一向在技术路线上比较激进，所以上述观点纯属个人判断。

Q:中英文语言有相当大的差异，那么 BERT 在处理中文语言时，会不会比处理英文语言更难？他们有哪些异同？

我不认为NLP的技术在处理英文和中文的时候有什么本质的区别，也包括Bert在内。大家一般的传统意识会认为中文更难一些，因为中文是表意文字，而且中文还需要分词等等。我觉得这个不是什么重要区别，如果我们具备一个强大的特征抽取器，那么理论上语义作为特征，能够被编码到参数里，所以是中文是英文无所谓。至于中文分词，我觉得并不是必须存在的，如果特征抽取器足够强，中文NLP任务也应该以字作为输入，至于哪些连续的字应该作为一个单词是合理的，这应该作为内部特征，由特征抽取器来自动做，人并不需要看到这些单词。目前很多NLP任务，都是先分词，再往后走，以前是这样，但是目前技术发展足够成熟了，现在，包括以后，我不觉得这是必须具备的步骤。

总体而言，我觉得之所以之前大家普遍认为中文比较难，是因为技术发展不够成熟，没有一个强大的特征抽取器导致的。所以需要很多人工的工作，或者需要把NLP划分为多个阶段，把一个难题分而治之，但是也会带来相应的问题，比如错误累计和放大等问题等。随着技术的发展，我觉得这些都不是必要的，不同语言不是问题，所有任务都应该End to End的被解决，包括分词等，我感觉也不是必须单独存在的。

Q:许多推荐系统产品都离不开NLP，NLP在新浪微博有哪些业务形态？做推荐系统需要对NLP有很深的研究吗？

A:NLP在微博主要是作为推荐业务的后台支持方向存在的，也有比较多的适用场景，比如要深入理解用户的兴趣，需要能够理解用户发的或者互动过的微博内容，而对微博内容的理解，以及用户兴趣的建模等，都离不开先进的NLP技术积累。目前，我们也大规模地应用了Bert等模型，并取得了较好的业务效果。

对于推荐系统算法工程师来说，对NLP的了解不需要特别深入，但是也应该具备相关知识和经验。一方面来说，文本内容往往是很多推荐场景的重要特征，而想要把这块做好，选择先进的NLP技术对于知识储备还是有要求的；另外一方面，文本特征可能只是很多类型特征中的一种，而且推荐的知识体系和NLP还是区别较大（当然，以后也许会被统一到Transformer也不一定，不过目前还是区别较大），所以貌似也不需要对于NLP有特别深入的





Q: 新浪微博的信息推荐系统有哪些技术栈?

A: 因为微博的推荐场景是个典型的信息流业务, 所以包含了很多环节。从偏后台的物料质量评估、用户兴趣建模、微博内容理解、图片视频理解以及多模态融合, 到业务前台的推荐系统的召回、粗排以及精排等推荐环节, 以及离线及在线的大规模机器学习模型训练及服务。算是一个标准的信息流推荐系统。

Q: 为打造更精准的推荐系统, 新浪微博做了哪些工作? 有计划引入一些新的技术吗? 或者说, 有在关注哪些新技术?

A: 近两年来, 在推荐系统的召回和排序两个环节, 微博机器学习团队都进行了大规模的技术升级, 并获得了明显的业务效果。召回层面, 我们目前已经实现了大规模FM统一召回模型, 正在逐步替代传统的多路召回模型, 并在各项指标取得了非常明显的业务效果; 排序层面, 我们经过了LR、大规模FM、FM+FTRL在线模型等不断的模型升级, 每次大的模型升级都取得了收益, 目前也小流量了以DeepFM为基础的深度学习排序模型。在物料比如微博的理解方面, 目前也在尝试多模态的技术路线, 并取得了一定进展。

在业务团队将新技术不断升级落地的同时, 微博AI Lab也不断尝试些推荐领域重要方向的新技术, 并推动这些新技术在业务落地。比如我们在18年提出的双线性FFM模型, 在获得和FFM模型相近效果的情况下, 参数规模减少了几十倍, 也提出了几个新的深度学习排序模型, 比如FibiNet以及FAT-FFM模型, 这些工作陆续发表在Recsys2019以及ICDM2019等国际会议上。

我个人比较关注统一召回模型技术以及新型的CTR模型上。感觉这两块在业界来说, 还有很大的优化空间, 是值得深入探索的方向。

Q: 2019年已经过半, 您认为在未来, 推荐系统和NLP领域还有哪些方向值得研究?

A: 工业化的推荐系统经过最近几年的发展, 目前已经进化到了深度学习时代, 代表性的模型有很多, 比如Wide&Deep或者DeepFM模型。但是跟NLP或者图像领域相比, 很明显深度学习在实用化的推荐系统应用场景还未能发挥符合人们期待的效果。并没有哪个深度学习推荐模型能够相对传统模型有非常大的性能提升, 这里面原因可能比较复杂。从模型角度看, 我比较看好将Transformer模型应用在推荐领域, 大家知道, Transformer在NLP里面已经大放异彩, 而且从机制上, 它是比较适合用来做推荐或者CTR模型的。但是可能仍然需要针对推荐领域的特点做些改造, 直接应用效果也没有太体现出来, 但是总体而言, 我个人比较看好这个模型。当然, 从实用化的推荐引擎角度, 还有很多值得关注的点, 比如多模态融合 / 多目标、多任务与Transfer Learning以及AutoML的应用等, 都是很有前景和值得探索的方向。刚才讲的是排序模型和推荐机制, 至于推荐领域的另外一个重要环节: 召回阶段, 我觉得采用模型统一召回代替传统的多路召回是个比较明显的趋势。

至于NLP领域, 自从Bert的出现, 应该说NLP各个应用领域进入了一个新时代。Bert在相当多的NLP应用领域使用后, 取得了目前最佳的效果, 而且不少领域有大幅度的性能提升。这预示着大量无监督数据开始被NLP真正使用起来, 这个如果能用好威力无穷。而且目前看, Bert+Transformer有统一NLP各个子领域的能力和趋势, 这也是非常好的一个进展。当然, 因为Bert和Transformer的出现时间还比较短, 所以大家其实对这两者理解还不够深入, 对其中真正发挥作用的因子还没有整理清楚, 所以深刻理解两者的作用机理, 以及针对性的对两者进行改进, 是非常值得深入探索的方向。比如如何将Bert应用在文本生成领域, 如何融合多模态的信息等很多方向都具备很大的挑战。

Q: 你觉得深度学习在推荐系统的应用方面存在哪些问题?

A: 就像上面说的, 尽管推荐领域目前也进入了深度学习时代, 但是明显技术前进的步伐缓慢, 应该说深度学习在推荐领域还没有发挥出特别大的作用。一方面, 并没有哪个DNN推荐模型相对传统模型有幅度巨大的效果提升, 包括从模型深度来说, 图像领域比如Resnet已经做到了几百层, NLP领域的Transformer也能够做到了40到70层, 而你会发现DNN版本的推荐系统目前还只能做到3到5层, 深度做不起来。说深度貌似没说点到点上, 其实不然。我们回顾一下, 在Resnet出现之前, 图像领域最多只能做到20多层, NLP也只能做到不超过10层, Resnet和Transformer突破了这种情况, 把深度做起来了, 效果也跟着深度做起来了。其实深度和效果是有比较紧密的正相关作用的, 所以我个人觉得, 目前推荐模型, 如何把深度做起来可能是比较核心的问题。如果深度做不起来, 效果可能也很难获得很大的提升, 背后的原理, 因为是我个人的猜测, 所以就不细说了, 避免造成误导。

其实, 之所以推荐模型感觉进步慢, 还有一个隐藏的
真实训练数据, 尽管网上可以下到个别规模比较大的

已赞同 318

22 条评论

分享

喜欢

收藏



还是严重缺乏的。这其实是比较拖技术发展后腿的一个缺陷。如果一个新技术是在小规模数据上验证的，并没有任何保证能够在大规模数据集上这种技术优势仍然存在；而很多新技术的论文使用的数据规模比较大，但是都是私有数据，其他人无法复现或者在现有技术改进。这严重制约了技术发展步伐。如果能够有几个高质量、大规模的推荐数据，那么大家就可以在这些数据上不断尝试新技术，明显技术迭代速度会快很多。一个很好的参照就是NLP的阅读理解领域，从早期的小规模阅读理解数据，到大规模但是难度不太高的Squad 1.1，再到更难Squad 2.0，每次数据集规模的扩大及难度的提高，都带来很多新的有效模型的提出。两者道理其实是类似的，而推荐领域这块明显落后太多。



Q:在10月举办的QCon上海2019，您将担任“人工智能应用与实践”专题的出品人，可否讲讲您的策划思路？

A:因为我参加过多次InfoQ举办的会议，也担任过多次人工智能专题的出品人，在这方面多少还是有些经验。总体而言，我觉得应该重视以下几点：首先，专题里的讲座应该具备技术前沿性和先进性，同时具备落地的实用性，这一直是我作为专题出品人对专题内讲座的一个基本期待；其次，在话题方面，考虑人工智能领域覆盖的全面性，我们会邀请各个大的互联网公司技术专家，在搜索推荐、自然语言处理、图像与视频以及人工智能工程架构等多个方面进行经验分享；再次，我一般对分享嘉宾以及分享内容有些比较高的期待，嘉宾希望是有经验的，包括专业实践经验以及比较丰富的演讲经验，分享内容则希望干货尽可能多一点。所以，我会按照上述几个原则去策划这个专题，尽可能给大家一个能够充分交流和学习的机会。

发布于 2019-08-25

自然语言处理 推荐系统 深度学习 (Deep Learning)

文章被以下专栏收录



深度学习前沿笔记

已关注

推荐阅读

Youtube推荐系统最大提升：RL Off policy + 推荐系统

分享一下谷歌的一篇强推荐系统&强化学相结合的论文，据说获得了Youtube近两年单次上线的最高收益：Top-K Off-Policy Correction for a REINFORCE Recommender System谷歌这篇...
魔鬼吊儿郎

Table 1: Reproducible works on deep learning algorithms for top-n recommendation per conference series from 2015 to 2018.

Conference	Rep. / Non-rep.	Reproducible
KDD	3/4 (75%)	[17], [23], [48]
RecSys	1/7 (14%)	[53]
SIGIR	1/3 (30%)	[10]
WWW	2/4 (50%)	[14], [24]
Total	7/18 (39%)	

Non-reproducible: KDD: [43], RecSys: [41], [6], [38], [44], [21], [45], SIGIR: [32], [7], WWW: [42], [11]

最近三年四大顶会深度推荐系统上的18篇论文

深度传送门 发表于深度推荐系...

揭开YouTube Deep Serving之谜

王喆

22 条评论

切换为时间排序

写下你的评论...

张大大，能否写个专栏简单讲讲你们是怎么用FM替代多路召回的？

7 个月前

7 个赞

已赞同 318

22 条评论

分享

喜欢

收藏

张俊林 (作者) 回复 梵高斯基

7 个月前

嗯，好，下篇发这个

 10

梵高斯基 回复 张俊林 (作者)

7 个月前

欧克

 赞

查看全部 7 条回复

summer

7 个月前

最近做召回用的简单的分类模型。打分排序这样子。想在这方面做点改进。

 赞

全宇宙的希望

7 个月前

大神，科大讯飞不算独角兽么

 1

RayAg

7 个月前

公开超大规模的真实训练数据的话，我之前看到阿里那也有部分人有这个想法，但是这种和KPI无关的事情，真推动起来恐怕还很难。

 赞

驯猫师

7 个月前

个人觉得微博的推荐从用户体验上讲效果很差，作用很弱

 1

刘博简 回复 驯猫师

7 个月前

看个人吧，我就刷到根本停不下来

 赞

驯猫师 回复 刘博简

7 个月前

对比下抖音和头条就能看出差距了

 3

展开其他 1 条回复

Ingale

7 个月前

nlp近几年的创业机会多吗？

 赞

李木子

7 个月前

大佬，催更Bert第三篇



 赞

littlefucker

7 个月前

推荐的model做不深做不复杂，在线inference时latency的限制是很重要的因素。

 赞

江湖少年

7 个月前

我在思考nlp真正的春天是什么时候？现在我们这些nlp工作者很难真正地攻克掉这些难题，实践过程中或多或少出现各种问题

 赞

已赞同 318

22 条评论

分享

喜欢

收藏



故事尾音 回复 江湖少年

6 个月前



思考一下，那种情况下你觉得还有要我们的必要吗？

👍 赞



朱文文

6 个月前

张大大，我想请教下如何用bert做文本摘要生成任务[飙泪笑]

👍 赞



高腾

5 个月前

Youtube在2016年那篇用神经网络做的召回，效果怎么样啊？（预测阶段，对<用户，上下文>通过神经网络进行编码得到user-embedding，然后在video-embedding库里召回与其内积最大的上百个做候选集）

👍 赞