MTRec: Multi-Task Learning over BERT for News Recommendation

摘要

现有的新闻推荐方法通常只基于新闻标题来学习新闻表征。为了充分地利用其他领域的新闻信息，如类别和实体，一些方法将每个领域作为一个额外的特征，并将不同的特征向量与attentive pooling相结合。随着像BERT这样的大型预训练模型在新闻推荐中的采用，上述纳入多字段信息的方法可能会遇到挑战：压缩类别和实体信息的浅层特征编码与深层BERT编码不兼容。

本文提出了一个多任务学习框架，将多领域信息纳入BERT，从而提高其新闻编码能力。此外，本文使用gradient surgery技术减少不同任务之间的梯度冲突，这进一步提高了模型的性能。

在MIND新闻推荐基准上进行的大量实验表明 MIND新闻推荐基准的广泛实验表明此方法的有效性。

1. Introduction

传统的新闻推荐模型侧重于对特征的相互作用进行建模。以前的神经网络通常只根据新闻标题来学习新闻表征向量，然后通过顺序或注意力模型聚合之前浏览过的新闻来学习用户表征。虽然有效，但这些方法忽略了其他信息，如类别或实体，我们称之为多领域信息。如图所示，传统方法通常将每个领域的信息（Title/category/entities)转化为特征向量，并通过多领域学习将这些特征结合起来。

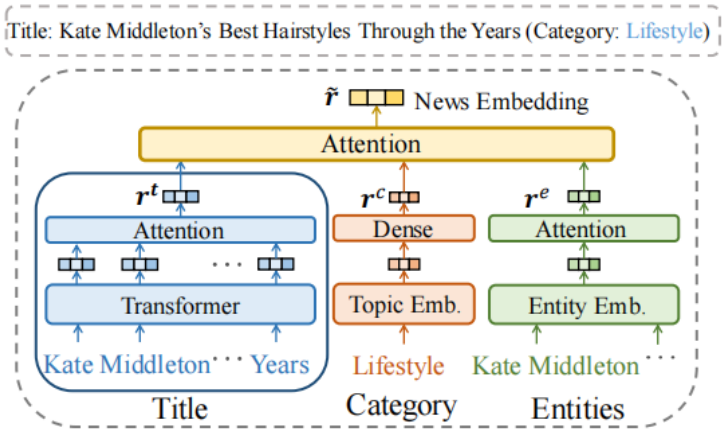
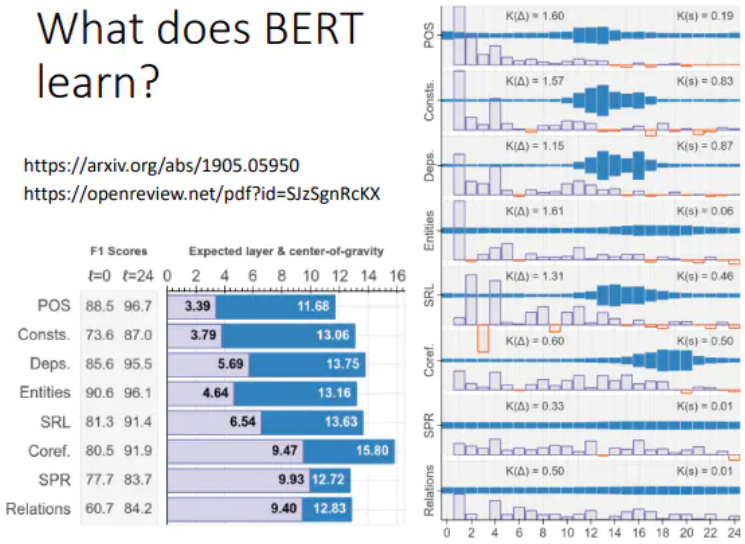


Figure 1: Traditional way to incorporate multi-field news information with attentive multi-field learning.

BERT作为预训练模型被越来越广泛地使用。（如把标题编码为图中蓝色方框）然而，可能出现的问题是：压缩类别和实体信息的浅层特征编码与标题的深度BERT编码不兼容。在新闻推荐中使用大型预训练模型时，会出现对多领域信息的无效适应。

> BERT
>
> BERT（Bidirectional Encoder Representations from Transformers）是一种Transformer的双向编码器，旨在通过在左右上下文中共有的条件计算来预先训练来自无标号文本的深度双向表示。因此，经过预先训练的BERT模型只需一个额外的输出层就可以进行微调，从而为各种自然语言处理任务生成最新模型。
>
> BERT任务的本质是理解语言。比如在文本分类中，分类器之前加一个BERT，可以实现输入句子输入类别；阅读理解增加全连接层(BERT+Softmax)，输入问题和文章，输出答案的位置。

> BERT两种训练方式：1. 掩码语言模型（MLM）：在将单词序列输入 BERT 之前，每个序列中 15% 的单词被替换为 [MASK] 标记。然后，该模型尝试根据序列中其他非掩码单词提供的上下文来预测掩码单词的原始值。2. 下一句预测（NSP）：在 BERT 训练过程中，模型接收成对的句子作为输入，并学习预测该对中的第二个句子是否是原始文档中的后续句子。在训练期间，50% 的输入是一对句子，其中第二个句子是原始文档中的后续句子，而在另外 50% 的输入中，从语料库中随机选择一个句子作为第二个句子，并假设该随机句子与第一句不相连。
>
> BERT不同层的意义
>
> 
>
>
> 通过上图我们可以看到随着NLP任务难度的加深，BERT越深层的权重越高(重要性)，越简单的任务，BERT越浅层的权重越小。（浅层和文本有关，深层和语义有关）。这就是为什么文章中说压缩类别和实体信息的浅层特征编码与标题的深度BERT编码不兼容

为解决这个问题，本文提出一个为MTRec的多任务学习框架，在BERT上进行新闻推荐，以有效纳入多领域信息。具体是使用BERT将新闻标题编码为新闻嵌入，并在BERT上设计两个辅助任务，即类别分类和命名实体识别（NER）。这两个辅助任务是和主要的新闻推荐任务一起训练的，可以帮助BERT更好捕捉新闻语义。为了进一步提升性能，本文采用`gradient surgery`技术，能在多任务训练中消除不同任务之间的梯度冲突。本文研究的是标题、类别、和实体，是异质的，可以从不同角度提供有价值的信息。

本文还发现，将多任务学习和传统的多领域学习结合可以提升模型性能。

在MIND新闻推荐基准上进行的大量实验表明 MIND新闻推荐基准的广泛实验表明此方法的有效性。

2. 方法

给定一个用户的历史点击新闻$N^h = [n_1^h, n_2^h \dots]$和一组候选新闻$N^c$，目标是根据用户历史行为计算候选新闻的interest score s_j，并把最高得分推荐给用户。对于每个新闻，有title text *T*, category label p^c, and entity set E

2.1 新闻推荐框架

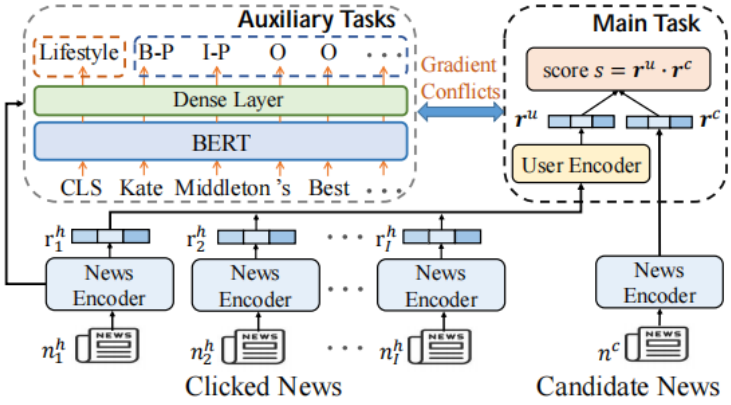


Figure 2: The overall framework of *MTRec*. We employ BERT as the news encoder and additive attention as the user encoder. In addition to the main task of news recommendation, we design two auxiliary tasks (i.e., category classification and NER) to further incorporate the category and entity information.

3个组成部分：a *news encoder*, a *user encoder*, and a *click predictor*

News encoder

对于每条新闻n，用预训练的BERT对标题进行编码。具体是将标签化文本T送入BERT模型，并采用[CLS]标记嵌入作为新闻代表r。将历史点击新闻N^h和候选新闻N^c的编码向量表示为R^h和R^c

User Encoder

为了从历史点击新闻的表征中获得用户表征，现有的方法通常采用顺序（或注意力模型）。在本文中，我们采用additive attention作为用户编码器来压缩历史信息R^h。然后，用户表征r^u被表示为...
q和W是可训练参数

Click Predictor

对于每个候选新闻，通过点成匹配候选新闻向量和用户表征$r_j^c \cdot r^u$作为兴趣得分sj。损失函数L是负的log 所有正向例子...s^+正向，L负向

$$\mathcal{L}_{Main} = -\sum_{i=1}^{|\mathcal{D}|} \log \frac{\exp(s_i^+)}{\exp(s_i^+) + \sum_{j=1}^{L} \exp(s_i^j)},$$

2.2 多领域信息

除了新闻内容，新闻类别还有其他有价值的信息，如类别标签和实体注释，这就是多领域信息。

如图1，每个领域信息最先通过嵌入查找和注意力机制转换为向量。然后表征R(三元组，标题、类别、实体)通过多领域学习被合并为最终新闻表征r。其中q和W都是可训练参数。

$$\widetilde{\mathbf{r}} = \sum_{\mathbf{r}_i \in \mathcal{R}} w_i \mathbf{r}_i, \quad w_i = \mathrm{softmax}(\mathbf{q}^T \cdot \tanh(\mathbf{W}^T \mathbf{r}_i)),$$

这里存在的问题：虽然对传统的文本编码有效，但多领域学习对深层的BERT编码可能效果不佳。因为压缩类别和实体信息的浅层特征编码可能与深层BERT编码不在同一个特征空间，直接将它们结合在一起可能会造成不兼容的问题。因此不能有效地利用多字段信息。

2.4 多任务学习

为了有效利用BERT的news encoder的多领域信息，本文提出使用在BERT基础上采用多任务学习，即**类别分类**和**命名实体识别**，如图2

Category Classification类别分类

为了纳入新闻类别信息，在BERT上增加一个分类任务，用[CLS]嵌入来预测新闻n_i的类别分布

$$\hat{\mathbf{p}}_i^c = \text{softmax}(\mathbf{W}^c\mathbf{r}_i + \mathbf{b}^c),$$

b和W可训练参数。类别分类的损失函数：

$$\mathcal{L}_{\text{Category}} = -\frac{1}{I}\sum_{i=1}^{I}\sum_{k=1}^{K^c} p_{i,k}^c \log(\hat{p}_{i,k}^c)),$$

K^c可训练参数。

Named Entity Recognition命名实体识别

在BERT顶层设计了一个NER任务，模型可以识别标题中的重要实体，能更好匹配感兴趣新闻。具体是根据精准匹配来定位新闻标题中的给定实体，并用B来表示起始词，I表示内部词，其他非实体词表示为O。然后根据BERT输出嵌入进行标签预测。 $\hat{\mathbf{p}}_{t_i}^n = \text{softmax}(\mathbf{W}^n\mathbf{r}^{t_i} + \mathbf{b}^n)$,r是第i个token的输出嵌入，b和W是可训练参数。这一部分的损失函数

$$\mathcal{L}_{\text{NER}} = -\frac{1}{I}\sum_{i=1}^{I}\sum_{l=1}^{l_i}\sum_{k=1}^{K^n} p_{l,k}^n \log(\hat{p}_{l,k}^n)),$$

Kn是所有NER tags的总数，l_i是第i个新闻的标题长度。

同时优化主任务、类别分类和NER任务的损失函数，最终loss为

$$\mathcal{L}_{\text{MTRec}} = \mathcal{L}_{\text{Main}} + \mathcal{L}_{\text{Category}} + \mathcal{L}_{\text{NER}}.$$

Multi-Task Learning with Gradient Surgery有GS的多任务学习

多任务学习中不同任务之间可能存在梯度冲突。不同任务的梯度方向夹角大于90度，不利于模型的表现。

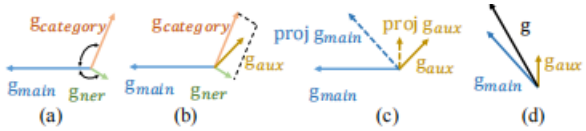


Figure 3: Illustration of the Gradient Surgery (GS).

gradient surgery, 将第i个任务的梯度g_i投影到另一个冲突任务的梯度g_j的法线上。

$$\mathbf{g}_i = \mathbf{g}_i - \frac{(\mathbf{g}_j \cdot \mathbf{g}_i)}{\|\mathbf{g}_j\|^2} \cdot \mathbf{g}_j.$$

这个方法在一定程度上是有效的，但我们的任务与普通多任务学习有些不同：我们的目的是利用辅助任务来提升主任务的性能，而不是平等对待它们。因此，要对主任务进行较少的梯度修饰，需要稍微修改上面的式子。首先合并辅助任务的梯度，然后采用系数λ来调整它们(图3(b))。λ根据经验设置为0.3

$$\mathbf{g}_{aux} = \lambda(\mathbf{g}_{category} + \mathbf{g}_{ner}),$$

然后再主任务和合并的辅助任务的梯度之间应用GS（图3 c），最终得到梯度g(图3d)

3. 实验

3.1 数据集和设置

在真实的新闻推荐数据集MIND上评估上述方法。使用小版本进行快速实验。本实验使用最近50次点击作为历史行为，每个正面新闻与4个负面新闻配对。实验结果与NAML等几个表现比较好的baseline进行比较。虽然上述方法都采用浅层文本编码，本实验采用BERT作为新闻编码器作为BERT的baseline。再现两种表现较好的基于BERT的方法，为LSTUR+BERT和NRMS+BERT。本实验同时结合多领域学习，将多领域信息和BERT baseline和MTRec相结合，分别表示为BERT+AMF和MTRec+AMF。

3.2 结果

	MIND-small			
Methods	AUC	MRR	nDCG@5	nDCG@10
NAML	66.12	31.53	34.88	41.09
LSTUR	65.87	30.78	33.95	40.15
NRMS	65.63	30.96	34.13	40.52
HieRec	67.95	32.87	36.36	42.53
BERT (baseline)	68.26	32.52	35.89	42.33
LSTUR+BERT	68.28	32.58	35.99	42.32
NRMS+BERT	68.60	32.97	36.55	42.78
BERT+AMF	68.96	33.42	37.10	43.27
MTRec	69.43	33.79	37.64	43.74
MTRec+AMF	**69.51**	**34.06**	**38.05**	**44.03**

Table 1: Performance of different methods. *MTRec* is our proposed multi-task method and "AMF" denotes attentive multi-field learning.

1. 利用BERT作为新闻编码器时，推荐新闻的表现明显更好。如LSTUR+BERT和NRMS+BERT。在LSTUR/NRMS中，只用BERT替换新闻编码器，表现就超过了他们的浅层版本。
2. BERT+AMF比BERT baseline更好，证明多领域信息的价值。不同用户喜欢不同新闻类别和实体，这些信息有利于系统做出个性化建议。
3. MTRec表现明显好于BERT+AMF，表现多任务学习策略的有效性。值得注意，多领域学习引用的是G和E的嵌入来向量化类别和实体的信息，这些特征编码可能与深层BERT编码不在同一特征空间，从而导致BERT+AMF中的多场信息使用不足。
4. MTRec+AMF取得最好的成果。R提出，多任务学习可以看做一种正则化，本文推断多领域学习直接增强了news representation（新闻代表性），不会与MTRec的多任务学习冲突。

3.3 Ablation Study消融研究

Auxiliary Tasks辅助任务

首先分别去掉类别分类和NER任务以探索它们的影响。如图4所示，当只引入一个辅助任务时，模型的性能有不同程度的下降。但它们的性能仍然优于BERT baseline，这证明了这两个辅助任务对BERT贡献了额外的信息。Wu等人（2019c）只利用了标题和类别，在图4中表示为w/o NER。

删除类别分类任务时，性能下降最多，这可能是由于类别是文档级标签，包含比实体更丰富的信息。

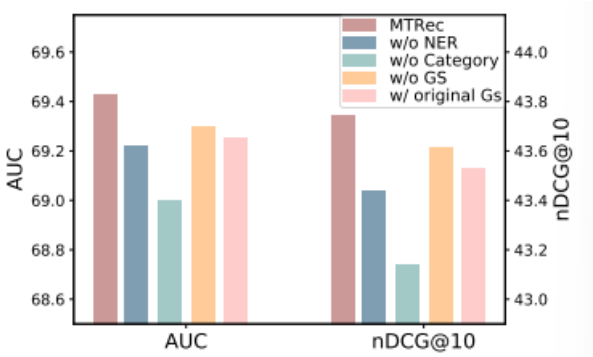


Figure 4: Ablation study to show the effectiveness of auxiliary tasks and gradient surgery (GS).

Gradient Surgery

进一步去掉MTRec中的GS。可以看到，图4中性能下降非常显著，验证了梯度下降在多任务学习中减少梯度冲突的作用。

当我们应用Yu等人(2020)的原始GS时，性能甚至更差。原因是，我们的目的是利用辅助任务来提高主任务的性能，而不是平等对待它们。附录B中有记录并绘制了主任务和辅助任务之间在训练过程中的梯度的余弦相似度。

4. 结论

本文在BERT上提出了一个新的名为MTRec的多任务学习框架，用于新闻推荐，它可以有效地纳入多领域信息。本文还修改了`gradient surgery`技术以减少梯度冲突并进一步提高模型性能。最后，本文发现将多任务学习与传统的多领域学习结合起来能达到最佳效果。在MIND数据集上的大量实验表明了我们方法的有效性。本文展望未来可以将把MTRec与更先进的用户建模方法结合起来（Li等人，2022）。