焦宇超, 阎刚. 基于 BERT 与要素提取的相似案例匹配[J]. 智能计算机与应用,2025,15(1):130-135. DOI:10.20169/j.issn.2095-2163.250120

基于 BERT 与要素提取的相似案例匹配

焦宇超, 阎 刚
(河北工业大学 人工智能与数据科学学院, 天津 300401)

摘 要: 相似法律案件检索是一项特殊的检索任务,对于给定的查询案例,需要从给定的候选案例中搜索相似的案例。与传统的文本匹配不同,法律案件匹配具有文本较长、主题性强的特点。针对上述问题,本文提出了一种基于案件要素的相似案例检索方法。首先对 BERT 模型使用通用语料进行微调;然后采用段落聚合方法,对案件文书上下文语义信息进行编码,同时将法律文书数据融入模型。本文在 LeCaRD 数据集上进行了广泛的实验,实验结果表明,本文提出的模型优于现有模型。

关键词: 相似案例匹配; BERT; 长文本; 法律要素
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 2095-2163(2025)01-0130-06

Similar case matching based on BERT and feature extraction

JIAO Yuchao, YAN Gang
(School of Artificial Intelligence, Hebei University of Technology, Tianjin 300401, China)

Abstract: Similar legal case retrieval is a special retrieval task in which similar cases need to be searched from given candidate cases for a given query case. Unlike traditional text matching, legal case matching has the characteristics of long text and strong subjectivity. To address these issues in similar case matching in legal cases, this thesis proposes a similar case retrieval method based on case elements. This thesis first uses general corpora to fine-tune the BERT model, then encodes the context-specific semantic information of case documents using the paragraph aggregation method, and integrates legal document data into the model. Extensive experiments on the LeCaRD dataset were conducted in this paper, and the results show that the proposed model is superior to existing models.

Key words: similar case matching; BERT; long text; legal elements

0 引 言

随着智慧司法在国内的快速发展,最高人民法院实行了法律案例强制检索制度。在司法责任制改革的背景下,法律相似案件匹配机制是中国统一裁判尺度的重要举措[1]。相似案件匹配的目的,是识别与给定案件相似的案件。最高人民法院提供了一系列的指导性案例,供类似案件做审判参考。指导性案例由标题、关键词、判决要点、相关法条、案件基本事实、判决结果、判决理由以及包括有效法官和受判决人员姓名在内的信息组成[2]。

在过去的研究中,人们提出了大量的文本检索模型,特别是针对特定文本的检索模型,解决了文本数据特征维数复杂、检索困难的问题[3-5]。早期语义表示的常用方法包括向量空间模型、主题模型及其变体,如经典的 LDA 模型[6]。随着词嵌入技术的不断发展,研究人员开始使用基于深度学习模型的密集文本向量表示作为机器学习算法的输入[7-8]。传统的词袋红外模型包括 TF-IDF[9]、BM25[10]、Doc2Vec[11]等。Ponte 等[12]在计算法律文本相似度时,比较了 3 种无监督文本向量生成模型的效果,结果表明 Doc2Vec 效果最好。Vo 等[13]指出,基于词嵌入的文本语义表示在法律文本检索领域很有帮助。

相似案例匹配的问题本质上是文本相似度的研究。然而,法律案例检索任务在案例文本的长度、相关性的定义等方面与传统的文本相似度匹配问题有很大的不同。现有的文本相似度方法下,解决该问题存在以下两个主要挑战:

作者简介: 焦宇超(1997—),男,硕士研究生,主要研究方向:深度学习,自然语言处理。
通信作者: 阎 刚(1977—),男,博士,副教授,主要研究方向:图像处理,模式识别。Email:yangang@hebut.edu.cn。

收稿日期: 2023-08-31

首先，刑事案件通常是很长的文本，长度在数百到上千字符之间，这导致模型在建立文本向量表示时，无法处理所有有用的信息。目前，文本领域最常用的神经网络模型如长短期记忆网络（Long Short－Term Memory，LSTM）[14]，其记忆能力不强，在长文本中的应用效果不佳[15]。Shao 等[16]针对相似案例检索匹配过程中准确性问题提出基于段落语义的编码模型，在基于 Transformer 的双向编码器表示（Bidirectional Encoder Representation from Transformers，BERT）模型[17]对于案件文本表示的基础上，针对案件段落级别语义关系进行计算，通过聚合段落级的交互矩阵，计算两个案例之间的相关性。Hu 等[18]提出了基于法律事实的相似案例检索模型（Bidirectional Encoder Representation from Transformers－Law Former，BERT－LF），将法律主题和法律要素实体相结合，使文档表示向量更适用于法律场景。该模型采用段落分割和聚合的方式，将法律文本按照案件的逻辑顺序分成短段落，然后通过基于 BERT 的文本编码方法，表示查询候选段落对。

其次，法律案例的相似性不同于一般的文本相似性，在一定程度上也超出了主题相关性的一般定义，其需要探讨法律文书中所包含的案件事实的相似性。使用传统的文本相似度方法，确实可以学习到语义相似度，但模型并不了解法律领域的知识，因此可能无法学习到表面语义下更深层次的法律相关逻辑关系，从而导致仅使用文本相似度无法发现高度相似的法律案例。因此，识别案件在法律问题和法律程序方面的相似之处至关重要。

1　方法介绍

针对当前法律文本匹配面对的问题，本文提出了基于 BERT 与要素提取的相似案例匹配模型（Bidirectional Encoder Representation from Transformers－Paragraph Legal Elements Extraction，BERT－PLE）。模型分为 3 个模块：BERT 微调模块、法律数据提取模块、语义提取融合模块。模型整体结构如图 1 所示。

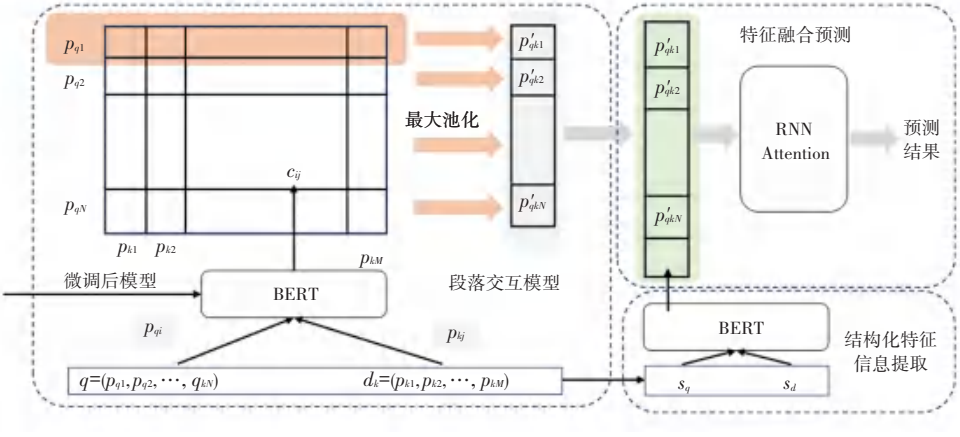

图 1　BERT－PLE 整体结构

Fig. 1　BERT－PLE general structure

1.1　BERT 微调模块

微调预训练模型相对于重新训练模型，可以省去大量计算资源和计算时间，提高计算效率。因此，本文在使用 BERT 来推断案例段落之间的相似程度之前，使用搜狐 2021 校园文本匹配算法大赛数据集，对预训练模型进行微调。该数据集中每条数据由 3 个句子构成，即样例句子 A 和候选句子 B、C，任务需要从 B、C 中选择与样例句子语义更相近的句子。为了适应后面部分训练的特点，在数据处理过程中，本文选择长度均在 64～255 之间的文本对进行训练。

通过微调使得 BERT 能够推断段落之间的语义相似程度，并将在语义提取融合模块中使用。模型以端到端方式对 BERT 在句子对相似度匹配任务上的所有参数进行微调。输入由数据集中文本对组成，整体结构如图 2 所示。文本对由［SEP］标志分隔，并在文本对之前添加一个［CLS］标志。对于 BERT 的输出模型将对第一个输入令牌（［CLS］）对应的最终隐藏状态向量输送到分类层。分类层使用全连接层来做二分类，损失函数使用的交叉熵损失函数如下所示：

$$loss = -(y\log\hat{y} + (1-y)\log(1-\hat{y})) \quad (1)$$

式中：y 为给出的文书与候选文书实际的相似度，$\hat{y}$ 为模型的预测相似度。

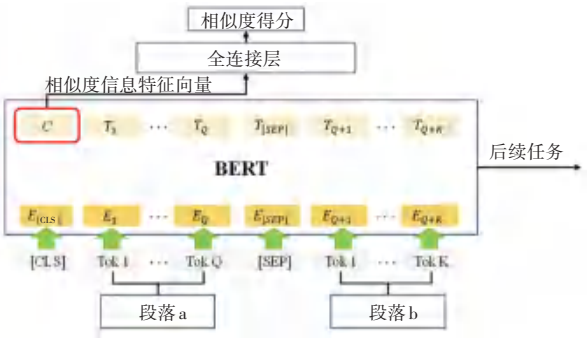

图 2 BERT 微调模块

Fig. 2 BERT fine-tuning module

1.2 法律数据提取模块

本文建立了刑事判决文书的标准化模块，以对司法文书中的共有信息进行分类，例如案件相关的罪名、涉及的法律条款以及量刑等。针对常见的刑事案件，对于伤害类案件提取了涉案凶器和被害人伤情等信息；财产类案件提取了涉案财物类型和价值等信息；对于事故类案件，提取了关键案情要素，例如是否饮酒、是否超速等信息。通过使用正则表达式，所有要素信息被提取并连接为一个长文本，作为后续处理的数据输入。

刑事判决文书的结构化特征见表 1。特征被划分为多个级别逐层细化。在特征大类方面，由于不同种类文书结构和具体要素存在一定差异，本文仅选择了几个较为典型的刑事案件进行分析。

表 1 刑事判决文书结构化特征

Table 1 Structured characteristics of criminal judgment documents

要素特征	特征维度	细粒度特征
案件元信息	判决罪名	
	相关法条	
	罚金情况	没收所得、罚金额度、赔偿情况
	判罚刑期	判罚刑期、缓刑情况
伤害类	被害人伤情	受伤情况、伤残情况
	嫌疑人手段	凶器、主观恶意
财产侵害	诈骗类	诈骗手段、诈骗金额
	盗窃抢劫	抢劫盗窃方式、涉案金额
肇事类	车辆肇事	车速情况、车辆类型、受害人伤情、是否酒驾/醉驾/毒驾

本文重点关注 3 种类型的司法文书，分别是伤害类罪名、财产侵害类罪名和肇事类罪名。在伤害罪的相关文书中，重点考虑被害人的伤情和嫌疑人实施的实时侵害手段。其中，被害人的伤情包括在案件中受到的伤害情况和伤残情况；嫌疑人侵害手段包括所使用的手段和主观恶意情况。

在财产侵害类案件中，包括诈骗类案件和抢劫盗窃类案件。对于诈骗类案件，主要考虑嫌疑人实施的实时诈骗手段，如电话诈骗、网络诈骗等。诈骗金额部分分为两类，即受害人初次遭受诈骗损失的财产情况，以及当前诈骗活动中再次(一次或者多次)损失的财产情况。对于盗窃抢劫类案件，重点考量嫌疑人的作案方式是否存在暴力行为，来区分盗窃类案件与抢劫类案件；涉案金额包含被害人遭受相关侵害之后损失的全部财物情况。

肇事类案件主要面向交通肇事类案件，包括车辆肇事时的时速情况、车辆类型、受害人伤情(和伤害类标准一致)以及驾驶员是否存在酒驾、醉驾、毒驾情形。

1.3 语义提取融合模块

为了解决长文本匹配的问题，本文将法律文书长文档切分成段落，并且在语义层面对长文档进行交互。如，查询 q 和一个候选文件 d_k 切分成段落可以表示为：

$$q = (p_{q1}, p_{q2}, \cdots, p_{qN}) \tag{2}$$

$$d_k = (p_{k1}, p_{k2}, \cdots, p_{kM}) \tag{3}$$

其中，N 和 M 分别表示 q 和 d_k 中的段落总数。

对于每一个段落 q 和段落 d_k，本文构建一个段落对 (p_{qi}, p_{kj})。其中，$1 \leqslant i \leqslant N$、$1 \leqslant j \leqslant M$，与[CLS]和[SEP]一同作为 BERT 的输入，将[CLS]的最终隐藏状态向量作为输入段落对的聚合表示。通过这种方式，可以得到最终查询 q 和一个候选文件 d_k 的段落交互图，图中由每一个代表查询段落 p_{qi} 与段落 p_{kj} 之间的匹配信息 C_{ij} 组成。对于查询的每一个段落，本文使用最大池化的方式获得与候选文档的匹配表示，从而得到一系列匹配向量，表示如下：

$$\boldsymbol{p}_{qki} = \mathrm{MaxPool}(C_{i1}, C_{i2}, \cdots, C_{iM}) \tag{4}$$

随后，使用循环神经网络（Recurrent Neural Network，RNN）模型进一步对段落匹配向量间的关系进行挖掘。在前向传播中，RNN 生成一系列隐藏状态：

$$\boldsymbol{h}_{qk} = [h_{qk1}, h_{qk2}, \cdots, h_{qkN}] \tag{5}$$

采用注意力机制，来推断每个隐藏状态的重要性。每个状态的注意力权重通过以下方式衡量：

$$\boldsymbol{\alpha}_{qki} = \frac{\exp(h_{qki} \cdot u_{qk})}{\sum_i \exp(h_{qki} \cdot u_{qk})} \tag{6}$$

其中，h_{qki} 是 RNN 给出的第 i 个隐藏状态，u_{qk} 通过以下公式生成：

$$\boldsymbol{u}_{qk} = W_u + \mathrm{MaxPool}(h_{qk}) + b_u \tag{7}$$

通过注意力聚合来获得文档级别的表示：

$$d_{qk} = \sum_i \alpha_{qki} \cdot h_{qki} \tag{8}$$

d_{qk} 通过一个全连接层，并且通过 Softmax 函数进行预测：

$$\hat{y}_{qk} = \mathrm{Softmax}(W_p \cdot d_{qk} + b_p) \tag{9}$$

b_p 表示文书相似度标签的集合，在训练过程中，采用以下函数来计算 $loss$：

$$loss = -\sum_r^R y_{qkr} \log(\hat{y}_{qkr}) \tag{10}$$

查询 q 与对应的候选文档片段 d_k 首先通过上文描述的段落级别交互匹配，得到段落级交互特征向量。在此基础上，通过案件结构化特征抽取方法构建的案件结构化特征文本，通过输入 BERT 案件结构化特征匹配向量。将案件结构化特征匹配向量与语义匹配矩阵池化后的结果拼接后，通过循环注意力网络进行整个篇章的匹配特征的计算，最终通过预测层进行结果的预测。

2　实验结果及分析

2.1　数据集

BERT 微调模块采用搜狐 2021 校园文本匹配算法大赛数据集，该数据对中存在短短文本匹配、短长文本匹配和长长文本匹配。为了适应下一模块训练时划分段落长度的特点，在数据处理过程中，选择了长度均在 64 ~ 255 之间的文本对进行训练，共 2 459 条。

本文使用的法律相似案例匹配数据集 LeCaRD[19] 是第一个基于中国法律体系的刑事类法律案例检索数据集。LeCaRD 由 107 个查询案例和 10 700 个候选案例组成，原案例选自于 43 000 多份中国刑事判决书。不同于以往工作中通过引文中的支持性案例或专家知识来识别相关案例的相似性判断，此数据集提出了一系列基于关键因素的主观性评价相结合的相关性判断标准。此标准是在中国人民最高法院公布的官方文件的指导下进行制定的，所有结果均由多名刑法法律专家进行评估。LeCaRD 的数据集概况见表 2。

表 2　LeCaRD 数据集概况

Table 2　Overview of the LeCaRD dataset

统计	数量
总查询数	107
每条查询的候选文书数	100
每条查询的平均相关文书数	10.33
查询文档的罪名数	20

2.2　评价指标

文本评价指标从结果精度和排序两个方面来评估。精度度量标准包括准确度（Precision，P）和平均准确率均值（Mean Average Precision，mAP），排序度量标准采用了归一化折损累积增益（Normalized Discounted Cumulative Gain，NDCG）。

$$P@n = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{11}$$

式中：$P@n$ 代表前 n 个结果的准确度。本文用 $y_i = \{0, 1, 2, 3\}$ 分别表示第 i 个结果的相关程度。

$$AP = \frac{1}{R}\sum_{r=1}^{R} P@r \tag{12}$$

$$mAP = \frac{1}{T}\sum_{t=1}^{T} AP_t \tag{13}$$

式中：AP 是 $P@n$ 的一个平均，本文中 $R = 10$；mAP 指对每一次实验结果计算 AP 累加后求平均，T 表示实验次数（本文中 $T = 5$）。

$$CG@T = \sum_{i=1}^{T} g_i \tag{14}$$

$$DCG@T = \sum_{i=1}^{T} \frac{g_i}{\log_2(i+1)} \tag{15}$$

$$NDCG@T = \frac{DCG@T}{IDCG@T} \tag{16}$$

式中：CG 代表累计收益，在文档排序中存在对应的相关度 g，$g_i = \{0, 1, 2, 3\}$ 表示数据集文书相关程度；DCG 中加入排名权重 i，表示排序位置，排名越靠后的文章对于指标值的影响越小，$NDCG$ 表示归一化的 DCG 值；$IDCG$ 表示使用数据集计算的理想 DCG 值。

2.3　参数设置

实验参数设置见表 3。

表 3 实验超参数设置

Table 3 Experimental hyperparameter settings

参数名	数量
查询文档切分段数	8
候选文档切分段数	16
迭代轮次	10
学习率	1.00E-05
单卡批次大小	4

实验结果见表 4,其中 BM25 与 TF-IDF 为传统的增强型布尔检索模型以及向量空间检索模型,ARC-II[20]为基于神经网络的文本匹配方案。BERT 方案中直接将查询文书文本与候选文书文本各从尾部截取长度为 255 字符的段落并拼接,之后输出预测分数。同时还选取了当前在类案匹配上效果最好的 BERT-LF 模型进行对比。与以上方法相比,本文提出的方案在效果上有着一定的提高。

表 4 实验结果

Table 4 Experimental results

模型	*P*@5	*P*@10	*mAP*	*NDCG*@10	*NDCG*@20	*NDCG*@30
BM25	0.380	0.350	0.498	0.739	0.804	0.894
TF-IDF	0.270	0.215	0.459	0.817	0.836	0.853
ARC-II	0.310	0.285	0.468	0.754	0.823	0.860
BERT	0.470	0.430	0.568	0.774	0.821	0.899
BERT-LF	0.490	0.445	**0.592**	0.816	0.864	0.891
BERT-PLE	**0.492**	**0.445**	0.555	**0.848**	**0.881**	**0.938**

为了进一步分析各个模块对于模型的影响,本文对多个模块进行了消融实验,实验结果见表 5。SEM-FE 表示仅使用 BERT 进行语义编码计算文书相似度;SEM-TE 表示仅使用法律实体提取模块计算;BERTorg 表示未经微调的 BERT 模型。同时,也对 RNN 使用 LSTM 与门控循环单元(Gated Recurrent Unit,GRU)[21]的效果进行了比较。可以看出,BERT 微调、法律信息提取、语义提取融合模块均对模型性能有一定提升。

表 5 消融实验

Table 5 Ablation study

模型	*P*@5	*P*@10	*mAP*	*NDCG*@10	*NDCG*@20	*NDCG*@30
SEM-FE(BERTorg)	0.320	0.340	0.398	0.693	0.760	0.862
SEM-FE (LSTM)	0.396	0.377	0.460	0.747	0.806	0.889
SEM-FE (GRU)	0.406	0.402	0.461	0.757	0.810	0.890
SEM-TE	0.397	0.358	0.457	0.759	0.792	0.840
BERT-PLE(LSTM)	0.510	0.430	0.547	0.803	0.832	0.909
BERT-PLE(GRU)	0.492	0.445	0.555	0.848	0.881	0.938

3 结束语

本文针对类案检索任务文本较长、主题性强的特点,提出了 BERT-PLE 模型,该模型通过两个主要方法解决了其中的难点。此外,为了提高检索排序的效果以及增强模型的健壮性和可解释性,本文构建了司法文书结构化特征,使用正则表达式对法律实体特征进行提取,并用相关实验证明了结构化特征在文书检索匹配任务上的有效性。与现有模型相比,本文提出的融合案件结构化特征的类案检索模型具有更好的性能。

参考文献

[1] 孙跃. 案例指导制度的改革目标及路径——基于权威与共识的分析[J]. 法制与社会发展, 2020, 26(6): 67-84.

[2] 高尚. 司法类案的判断标准及其运用[J]. 法律科学(西北政法大学学报), 2020, 38(1): 24-35.

[3] MAKIHARA Y, MANSUR A, MURAMATSU D, et al. Multi-view discriminant analysis with tensor representation and its application to cross-view gait recognition[C]//Proceedings of the 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). Piscataway, NJ: IEEE, 2015: 1-8.

[4] BEN X, GONG C, ZHANG P, et al. Coupled patch alignment for

matching cross－view gaits [J]. IEEE Transactions on Image Processing, 2019, 28(6): 3142-3157.
[5] BEN X, REN Y, ZHANG J, et al. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms [J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2021, 44(9): 5826-5846.
[6] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3(1): 993-1022.
[7] DONG J, LI X, XU C, et al. Dual encoding for video retrieval by text [J]. IEEE Transactions on Pattern Analysis Machine Intelligence, 2021, 44(8): 4065-4080.
[8] YANG X, WANG S, DONG J, et al. Video moment retrieval with cross－modal neural architecture search [J]. IEEE Transactions on Image Processing, 2022, 31: 1204-1216.
[9] SALTON G, BUCKLEY C. Term－weighting approaches in automatic text retrieval[J]. Information Processing & Management, 1988, 24(5): 513-523.
[10] ROBERTSON S E, WALKER S. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval[C]// Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Cham: Springer, 1994: 232-241.
[11] LE Q V, MIKOLOV T. Distributed representations of sentences and documents[C]// Proceedings of International Conference on Machine Learning. PMLR, 2014: 1188-1196.
[12] PONTE J M, CROFT W B. A language modeling approach to information retrieval[J]. ACM SIGIR Forum, 2017, 51(2): 202-208.
[13] VO N P A, PRIVAULT C, GUILLOT F. Experimenting word embeddings in assisting legal review[C]//Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law. Piscataway, NJ: IEEE, 2017: 189-198.
[14] MALHOTRA P, VIG L, SHROFF G, et al. Long Short Term Memory Networks for anomaly detection in time series [C]// Proceedings of the 23rd European Symposium on Artificial Neural Networks. Piscataway, NJ: IEEE, 2015: 2015.
[15] XIAO C, HU X, LIU Z, et al. Lawformer: A pretrained language model for chinese legal long documents[J]. AI Open, 2021, 2: 79-84.
[16] SHAO Yunqiu, MAO Jiaxin, LIU Yiquun, et al. BERT-PLI: Modeling paragraph-level interactions for legal case retrieveal[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20). Piscataway, NJ: IEEE, 2020: 3501-3507.
[17] DEVLIN J, CHANG M-W, LEE K, et al. Bert: Pretraining of deep bidirectional transformers for language understanding [J]. arXiv preprint arXiv, 1810.04805, 2018.
[18] HU W, ZHAO S, ZHAO Q, et al. BERT_LF: A similar case retrieval method based on legal facts [J]. Wireless Communications and Mobile Computing, 2022, 2022(1): 2511147.
[19] MA Y, SHAO Y, WU Y, et al. LeCaRD: A legal case retrieval dataset for Chinese law system [C]//Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, , 2021: 2342-2348.
[20] HU B, LU Z, LI H, et al. Convolutional neural network architectures for matching natural language sentences [J]. Advances in Neural Information Processing Systems, 1503.03244, 2015. DOI: 10.48550/arXiv, 1503.03244
[21] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv preprint arXiv, 1412.3555, 2014.