基于 BERT 模型的中文剧本情感识别

李霄龙[1] 刘昀鑫[1] 杨佳豪[1]

[1]（北京航空航天大学北京学院 北京 100191）

Emotion recognition of Chinese script based on BERT model

Li Xiaolong[1], Liu Yunxin[1], and Yang Jiahao[1]

[1]（*Schoolof Beijing, Beihang University, Beijing* 100191）

Abstract Mining the rich emotional semantic information hidden in the text is one of the important tasks of data mining. However, different from general text data, the data in the script text exhibits strong contextual relevance and role restrictions, which makes the existing text emotion recognition methods unable to be well adapted to the downstream tasks of script emotion recognition. At present, there are relatively many researches on sentiment analysis for different data sets, but they can still be optimized in the practice of special downstream tasks. Aiming at the problem of script emotion recognition, this paper proposes an emotion recognition method based on the BERT model, and through a series of targeted model optimization methods such as preprocessing (confrontation training), adjustment of hyperparameters, model soft voting, etc. The experimental results are obtained in the labeled data set from IQIYI. The experimental results show that the model can get good scores under the condition of limited hardware resources and proper parameter adjustment, that is, the Chinese script emotion recognition method proposed in this paper has reached the advanced level in terms of accuracy and fit.

Key words Text emotion recognition; BERT model; Hyperparameter optimization; Confrontation training

摘要 挖掘隐藏在文本中丰富的情感语义信息是数据挖掘的重要任务之一。然而，不同于一般文本数据，剧本文本中的数据呈现出具有强上下文关联，以及角色限制性等特点，这使得现有的文本情感识别方法无法较好地适应于剧本情感识别的下游任务。目前，针对不同数据集的情感分析研究工作相对较多，但在特殊任务的实践过程中依然可以优化。本文针对剧本情感识别问题，提出一种基于 BERT 模型的情感识别方法，并通过预处理(对抗训练)、调整超参数、模型软投票等等一系列针对性的模型优化方式在爱奇艺官方给出的标签数据集中得到了实验结果。实验结果表明，该模型能够在有限的硬件资源与适当调参的条件下得到不错的效果，即本文提出的中文剧本情感识别方法在准确度和拟合度方面都达到先进水平。

关键词 文本情感识别；BERT 模型；超参数优化；对抗训练

中图法分类号 TP391

文本情感分析，也称为意见挖掘、倾向性分析等。即对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。在数据挖掘领域，文本情感分析已经得到快速发展和广泛的研究。

相对于具有一般性的文本情感识别来讲，剧本的文本情感识别具有上下文关联，以及角色限制等特点。另外，剧本文的行文风格和通常的新闻类语料差别较大，更加口语化。

因此，准确而快速地提取不同剧本语句的情感，就需要使用自然语言处理(Neutral Language Processing, NLP)来进行数据挖掘。

为了进一步提高剧本情感识别效果，本文提出一种基于 Bert 模型的情感识别方法，该方法主要将预训练语言模型 Macbert-base 等 base(或 Large)模型的 12 层 Transformer(Large 为 24 层 Transformer)与每种情感设立的全连接层连接，计算 loss 后进行反向传播并输出作为基本方法，并对部分步骤进行特定的优化，得到最终输出。

本文的主要贡献如下：

1) 我们提出了一种基于 BERT[1]模型的中文剧本情感识别方法，并通过预处理(对抗训练等)、基于 CBAM 的特征提取、调整超参数以及模型软投票进行

优化；

2) 应用于竞赛中的 A、B 测试集的实验结果表明我们提出的基于 BERT 模型的中文情感识别方法，通过微调等优化操作后有很好的实验集识别效果。

1 相关工作

文本语言情感分析已经作为底层研究出现在 NLP 领域，且随着时代的发展，不同的模型不断地涌现。接下来，我们会概述一下已有的语言模型以及在不同条件下的应用。

第一个神经语言模型是 Bengio 等人[2]在 2001 年提出的前馈神经网络，以词作为输入向量表征。这种向量被现在的学者们称为“词嵌入”(word2vec)。这些词嵌入级联后被输入到一个隐藏层中，该隐藏层的输出又被输入到 SoftMax 层。

Collobert 和 Weston 在 2008 年将多任务学习[3]首次应用于 NLP 的神经网络。在他们的模型中，单词嵌入矩阵在两个接受不同任务训练的模型之间共享。

Mikolov 等人在 2013 年提出的创新技术[4] [5]是通过去除隐藏层，逼近目标，进而使这些单词嵌入的训练更加高效。虽然这些技术变更本质上很简单，但它们与高效的 word2vec 配合使用，便能使大规模的词嵌入训练成为可能。

NLP 问题在 2014 年左右开始引入神经网络模型。使用最广泛的三种主要的神经网络是：循环神经网络(RNN)、卷积神经网络(CNN)和递归神经网络(Recursive Neural Network)。

2014 年，Sutskever 等提出了 sequence-to-sequence 模型[6]。这是一个使用神经网络将一个序列映射到另一个序列的通用框架。在该框架中，编码器神经网络逐符号处理一个句子，并将其压缩为一个向量表示；然后，一个解码器神经网络根据编码器状态逐符号输出预测值，并将之前预测的符号作为每一步的输入。

2015 年，注意力机制(Bahdanau 等[7])是神经网络机器翻译(NMT)的核心创新之一，也是使 NMT 模型胜过经典的基于短语的 MT 系统的关键思想。sequence-to-sequence 模型的主要瓶颈是需要将源序列的全部内容压缩为一个固定大小的向量。注意力机制通过允许解码器回头查看源序列隐藏状态来缓解这一问题，然后将其加权平均作为额外输入提供给解码器。

预训练语言模型于 2015 年被首次提出(Dai & Le, 2015[8])，它们被证明在各种任务中效果很不错。语言模型嵌入可以作为目标模型中的特征(Peters 等，2018[9])，或者使用语言模型对目标任务数据进行微调(Howard & Ruder, 2018[10])。

本文提出的基于 BERT 模型的剧本情感识别方法，通过文本数据增强、不同参数对预训练模型微调以及不同投票策略的方式，不仅增强了模型泛化性能，弱化了部分输出所具有的特异性，也很大程度上保留了强上下文关系的情感，很好的提高了模型的鲁棒性。该方法可以根据不同的数据集进行特定的超参优化，也可根据不同的预训练模型进行针对性优化。

2 基于 BERT 模型定义的情感分析架构

在本节中，我们主要介绍对于剧本情感分析中自定义的 BERT 模型架构。下面会从基于 BERT 的模型架构，Multi-Attention 机制，基于 CBAM[11]的特征提取以及其余模块进行详述。

2.1 基于 BERT 的模型架构

BERT (Bidirectional Encoder Representations from Transformer) 模型的目标是利用大规模无标注语料训练、获得文本的包含丰富语义信息的语义表示，然后将文本的语义表示在特定 NLP 任务中作微调，最终应用于该 NLP 任务。

BERT 是基于 Transformers[12]的，而 Transformers 模型主要包括两大部分：Encoder、Decoder。

在基于深度神经网络的 NLP 方法中，文本经过分词后，经过 Encoder 可以把自然语言翻译成高维矩阵的形式，该向量既可以随机初始化，也可以利用 Word2Vector 等进行预训练以作为初始值。我们通常希望语义相近的词向量在特征空间上的距离也比较接近，因此由词向量转换而来的文本向量也能够包含更为准确的语义信息。

在此基础上，神经网络会将文本中各个词的一维词向量作为输入，经过一系列映射变换后，输出一个一维词向量作为文本的语义表示。

得到的输出经过 Decoder，可以把变换过的高维矩阵翻译回自然语言，这也和传统的 Seq2Seq 模型是一致的。由于是为了预训练服务，所以设计中只有 Encoder 部分。

特别的，对于文中提到的情感分析任务，BERT 模型在文本前插入一个[CLS]符号，并将该符号对应的输出向量作为规定长度文本的语义表示。

2.2 Multi-head-Attention 机制

首先，明确 Multi-head Self-Attention 机制的主要作用是让神经网络把“注意力”放在一部分输入上，即区分输入的不同部分对输出的影响。在这一步当中，一共需要区分 3 层操作，即 Attention、Self-Attention，以及 Multi-head-Attention。

Attention 主要涉及到三个概念：Query、Key 和

Value。对于一段文本来讲，目标词作为 Query、其上下文的各个分词作为 Key，并将 Query 与各个 Key 的相似性作为权重，把上下文各个词的 Value 融入原始 Value 中。Attention 机制将目标和上下文各个字词的语义向量表示作为输入，通过线性变换获得目标的 Query 向量表示、上下文各个分词的 Key 向量表示以及目标与上下文各个词的原始 Value 表示，然后计算 Query 向量与各个 Key 向量的相似度作为一层 Attention 权重，并将目标的 Value 向量和各个上下文分词的 Value 向量加权融合，整体作为 Attention 的输出。

对于 Self-Attention，需要对其中的每个词分别增强语义向量表示。即 Attention 机制中的输入文本分别将每个字作为 Query，加权融合文本中所有分词的语义信息，得到各个分词的增强语义向量，如图 1 所示。在这种情况下，Query、Key 和 Value 的向量表示均来自于同一输入文本。

为了增强 Attention 的多样性，利用不同的 Self-Attention 模块获得文本中每个词在不同语义空间下的增强语义向量，并将每个词的多个增强语义向量进行线性组合，从而获得一个最终与原始词向量长度相同的增强语义向量，这就是本论文架构中使用到的 Multi-head-Attention。

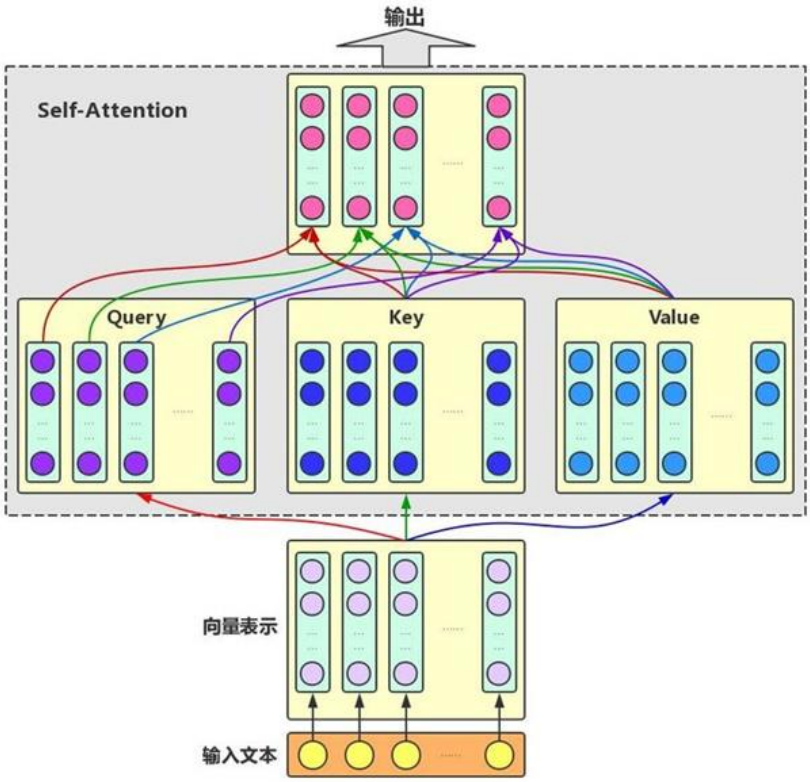

Fig. 1 Self-Attention overall structure.

图 1 Self-Attention 机制内部架构

2.3 基于 CBAM 的特征提取

上述架构组装好后，是一个 Transformers 堆叠的 BERT 模型，模型的输出端本质为 max_len*隐藏层宽度(Base 为 768，Large 为 1024)维度的向量组。对于下游任务为一个 6 分类的情感回归任务，前述模型的输出端需要进行全局语句的特征提取。

已知的 CBAM 原理简单有效，用于前向卷积神经网络。在 CV 领域中，CBAM 分别从通道和空间维度顺序提供注意力图，并应用于中间的特征图。CBAM 的特征提取方法主要应用了对于不同通道的平均池化层和最大池化层。对于前述模型，我们可移植类似的原理，但我们只需要做 1D 池化并将其连接在末端，作为近似的卷积注意力模块，进行特征提取。

我们取前述模型输出的向量组最后一层中各个[CLS] 对应的向量作为句子向量，对每个向量中的词语义向量通道进行类似于 CBAM 的赋权，即使用两个全连接层并进行非线性变换，这样可以更好的进行特征提取。插入该模块后，模型输出端将获得单向量维度为隐藏层宽度的向量组。

2.4 其余模块

在前述架构的前端，需要对文本进行预处理，即分词、词向量嵌入、以及对抗训练，得到长度为

max_len 的序列。

对于末端的处理，由于对应的是前述的基于 CBAM 的特征提取模块，因此针对我们需要进行的下游文本情感分析，应根据每一种情感建立一个全连接层，计算 Loss 并反向传播，训练完成后输出预测答案。经过一轮训练后，我们将经过微调的模型重新作为预训练模型进行重新训练。经过多次训练后，对于剧本文本或其它特定数据集的特征已经更多的遗留在微调好的模型中，因此这时选择输出预测，可以得到较好的效果。

另外，根据不同的预训练模型，或不同的超参数配置的训练模式，经过上述训练过程后，会得到不一样的、丰富的预测结果，我们这时可以进行软投票，以增加模型的鲁棒性。

3 情感分析架构的实现

3.1 预处理

在剧本情感识别中，训练数据相较于普通的文本情感分析更特殊，因为其中带有特定对象。因此文本输入的预处理很重要。输入的预处理中至少需要操作两部分：剧本文本与角色名。在 BERT 模型的输入端要把两部分拼起来。例如，我们采取的拼接模式是“原句 + ‘角色：’+ 角色名”。这样就能让 BERT 完全看到目标与原句，进而提取特征。

添加对抗样本，也是增加泛化性能的一种处理。在这里我们增加了部分扰动，它们呈现出的特点有：微小、使模型犯错。例如从输入的语料入手，用中文数据增强工具做随机同义词替换；也可以尝试使用同义句生成、回译等策略产生丰富的语料。之后用对应的同义部分替换原部分，作为后面的第 i 个 Epoch 所使用的输入。

同时，考虑到剧本人物情感受前后文影响较大，还采取了把句子按剧本段落拼接后作为单句输入的做法，长度为 len(input_sequence)，小于 max_len。

3.2 预训练语言模型

BERT 规定了两种规格的预训练模型，参数大小分别为 Base：12 * 768 * 12 = 110M；Large：24*1024*16 = 330M。因此一般设备 (显存 16GB 及以下) 想要做勉强又比较客观的训练，则只能使用 Base 模型，因此硬件限制也对最终结果产生着影响。对于中文来说，不同的预训练模型在同样的数据集上微调后输出的预测可能会差别很大，因此比较推荐的预训练模型有：MacBert、Roberta-wwm-ext、bert-base-Chinese、Chinese_Simbert 等。如果设备条件允许，可以选用 Large 预训练模型，但效果提升比较有限，甚至会出现下滑。

3.3 超参数调整

对于深度学习模型来说，模型调参过程可以说是必不可少的。由于在训练一个模型时只可以针对部分应用场景(受限于数据集和数据分布)，因此这一步做的时候要较为小心。

整个模型架构中，主要有以下超参数可以调整：

1. learning rate(学习率)，可以分开设置 Bert 部分和下游模型的学习率，以避免 Bert 过度学习导致不可逆遗忘或下游模型欠拟合的情况。这个超参数主要参考 Bert 模型的训练者提供的文档，一般量级为 1e-5。

2. Epoch(训练轮次)，少了会欠拟合，多了容易过拟合。例如对于 1e5 级别的数据 Epoch 不宜超过 2，否则就会带来过拟合。另外可以采用 early stopping (测试集的 loss 在几轮训练中没有下降则停止训练)的策略来自动停止训练。

3. BatchSize，大一点效果会更好，可以更好的对抗噪声。也不能太小，太小会导致收敛慢，甚至可能不收敛。

4. Max_len(语句最大长度)，可用范围是 0~512。对于单句预测，越长越好。对于剧本情感分析，拼接后长度最好设置在 300~400 效果最佳。

5. Warm_Up_Step | Ratio (训练预热步数/比例)，即在刚开始的一定步数或比例内提升学习率，可以提升学习效果和速度，但是很容易导致模型不可逆遗忘。学习率预热可以从一个极小的学习率开始，保证模型的预训练参数不会出现过大的波动。从效果上看，有一定提升，但要小心。

6. Drop_Out (神经元随机失活比例) 这个比较难调，效果也不明显，因此不建议调整该参数。

3.3 模型融合与软投票

模型的投票追求和而不同。不同的预训练模型训练时都会有不同的语料库，或大或小；以及可能会有不同的模型架构、训练方法，例如 BERT 和 RoBerta 训练架构是不同的，Roberta 甚至删除了 NSP 的任务。把它们作为不同的预训练模型，训练出来的效果可能会有的基本相同，有的差异却会很大。

另外，不同的拼接长度作为不同的样本进行训练得到的效果也不同，因此可以采用不同的拼接长度训练出来的结果进行投票。

由于训练的模型对应的是数据的测试集，且评测标准为 RMSE(均方根误差)，进行模型融合和预测结果软投票是必要的。软投票，也称之为加权平均概率投票，其加权计算方式有很多种。由于无法确定不同预训练模型导致的类概率判定影响的大小，因此最终处理的时候可以使用最简单的算术平均值作为输出。

软投票后的模型鲁棒性提升，预测效果会有相当

一部分的提升。该处理也不容易导致针对数据集性质的过拟合情况发生。

4 实验与结果

在本节中，我们使用本文提出的技术构建了一个针对中文剧本情感识别的模型，并且在 Data Fountain 网站，爱奇艺官方数据集上得到了比较好的效果，最终排名 A 榜 90，B 榜 75 名。

4.1 测试标准

算法评分采用常用的均方根误差（RMSE）来计算评分，按照“文本内容+角色名”识别出的 6 分类情感对应的情感值来统计，最终的得分为 1/(1+RMSE)。

4.2 数据集

本文所提出的方法仅在 Data Fountain 网站爱奇艺官方给出的 1 个数据集上进行了测试。这个数据集的描述与通过数据集需要完成的任务为：数据来源于一部分电影剧本作为训练集，训练集数据已由人工进行标注。参赛队伍需要对剧本场景中每句对白和动作描述中涉及到的每个角色的情感从多个维度进行分析和识别。

观察这个数据集，发现一共存在 36782 条数据，其中大部分的语句情感均为 0，而训练数据的输入为一个 6 维向量，标记情感粒度的程度。

验证集包括 21376 条数据，因此可以看得出来，训练集数据量没有远远大于验证集，这也就意味着任务是建立在数据不充足的基础上的，当然也需要更好的解决过拟合的问题。

4.3 基于 CBAM 的特征提取效果

基于 CBAM 的特征提取模块引入的 2 个全连接层实现了特征的提取效果，这里给出引入全连接后的示例效果：其中，图 2 中的[CLS]标识的第一原句为：“b2 哭的很伤心而 a1 却哈哈大笑。角色: b2”，图 3 中的[CLS] 标识的第二原句为：“b2 哭的很伤心而 a1 却哈哈大笑。角色: a1”。max_len 设置长度为 25。

可以看出，对于不同的角色，第一原句赋予了“哭”更高的权重；而第二原句赋予了“哈哈大笑”部分更高的权重。

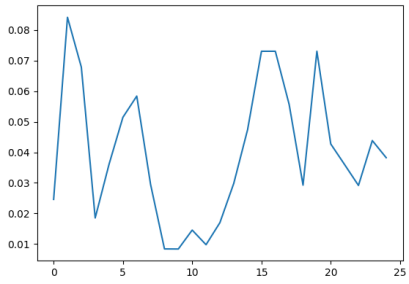

Fig. 2 Feature extraction vector of the first original sentence

图 2 第一原句的特征提取向量

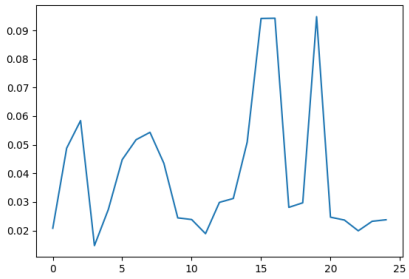

Fig. 3 Feature extraction vector of the second original sentence

图 3 第二原句的特征提取向量

4.4 实验和结果

训练集数据量没有远远大于验证集，这也就意味着任务是建立在数据不充足的基础上的，当然也需要更好的解决过拟合的问题。我们通过前述模型的描述，不断改进自己的模型架构，并获得了以下的实验结果。作为对比，实验前 4 组为 DataFountain 官方基于 Paddle 框架的 Baseline 试验结果。

Table 1 The Score of Proposed Method on IQIYI Dataset

表 1 提出的训练方法在爱奇艺数据集上的部分实验得分

Methods	*Max_len*	*Pretrained Model*	*Adversarial*	*Soft voting*	*Use CBAM*	*Score(A)*	*Score(B)*	*Rank(/562)*
Paddle Baseline 1	---	Roberta-wwm-ext	No	No	No	0.6740	---	379
Paddle Baseline 2	---	Macbert-large-chinese	No	No	No	0.6736	---	386
Paddle Baseline 3	---	Macbert-base-chinese	No	No	No	0.6738	---	383
Paddle Baseline opt	---	Macbert-base-chinese	No	Yes	No	0.6768	---	352
Bert Random Forest	---	Bert-Base-Chinese	No	No	No	0.6796	---	297
Bert Random Forest	---	Roberta-wwm-ext	No	No	No	0.6792	---	300
Bert Fine-tuning	128	Roberta-wwm-ext	No	No	Yes	0.6948	---	150

Bert Fine-tuning	128	Roberta-wwm-ext	Yes	No	Yes	0.6949	---	146
Bert Fine-tuning	128	Macbert-Chinese-Base	Yes	No	Yes	0.6958	---	123
Bert Fine-tuning	256	Macbert-Chinese-Base	Yes	Yes	Yes	**0.6996**	---	90
Bert Fine-tuning	256	Macbert-Chinese-Large	Yes	Yes	Yes	---	**0.7019**	**75**

5 总结

本文提出了一种基于 BERT 模型的中文剧本情感识别方法，并通过预处理（对抗训练等）、基于 CBAM 的特征提取、调整超参数、模型软投票等等一系列针对性的优化模型方式，在爱奇艺官方给出的标签数据集中得到了实验结果。实验结果表明，本文提出的中文剧本情感识别方法在准确度和拟合度方面达到先进水平。因此，本文所提到的方法是可行的，且效果是较好的。

参 考 文 献

[1] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

[2] Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A Neural Probabilistic Language Model. The Journal of Machine Learning Research, 3, 1137–1155.

[3] Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing. Proceedings of the 25th International Conference on Machine Learning - ICML '08, 20(1), 160–167.

[4] Mikolov, T., Corrado, G., Chen, K., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of the International Conference on Learning Representations (ICLR 2013), 1–12.

[5] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. NIPS, 1–9.

[6] Sutskever, I., Vinyals, O. and Le, Q.V. (2014). Sequence to sequence learning with neural networks. Adv. Neur. In. pp. 3104-3112.

[7] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[8] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," arXiv preprint arXiv:1511.01432, 2015.

[9] M. E. Peters, M. Neumann, M. Iyyer et al., "Deep contextualized word representations," 2018. arXiv preprint arXiv:1802.05365.

[10] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 328–339.

[11] Sanghyun Woo, Jongchan Park, Joon-Young Lee and In So Kweon, "CBAM: Convolutional block attention module", Proc. of the European Conf. on Computer Vision, pp. 3-19, Sep. 2018.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017.

Li Xiaolong, born in 2000.10. He is reading the third year of Beihang University. His main research interests include data mining, machine learning.

Liu Yunxin, born in 2002.10. He is reading the third year of Beihang University. His main research interests include data mining, machine learning.

Yang Jiahao, born in 2001.7. He is reading the third year of Beihang University. His main research interests include data mining, machine learning.

三人贡献比例：**33%**、**34%**、**33%**。