第00卷 第0期 计 算 机 学 报 Vol. 00 No. 0
2024年12月 CHINESE JOURNAL OF COMPUTERS 000. 2024

基于微调BERT模型的中文新闻摘要生成研究

陆文韬[1)]

[1)]东南大学，人工智能学院，南京市，中国，215500

摘 要 中文新闻摘要任务，旨在从长篇新闻文本中提取出核心信息，并生成简短、准确的摘要。在本篇研究中，我们采用BERT模型来解决中文新闻摘要的生成问题，BERT是一个预训练的Transformer模型，在许多自然语言处理的任务中都取得了非常好的效果，我们在BERT模型的基础上添加三种不同的摘要提取层：Classifier、Transformer和RNN神经网络。在提供的数据集上，ROUGE-2分别达到了16.89、16.81和16.77.

关键词 中文新闻摘要；BERT模型；Transformer模型；循环神经网络

Research on Chinese News Summarization Based on Fine-Tuning BERT Model

NAME Wentao Lu[1)]

[1)](Department of Artificial Intelligence, Southeast University, Nanjing 215500, China)

Abstract

The Chinese news summarization task aims to extract key information from long news texts and generate concise and accurate summaries. In this study, we use the BERT model to address the problem of Chinese news summary generation. BERT is a pre-trained Transformer model that has achieved outstanding results in many natural language processing tasks. We add three different summarization layers on top of the BERT model: a Classifier, a Transformer layer, and an RNN network. On the provided dataset, the ROUGE-2 scores achieved were 16.89, 16.81, and 16.77, respectively.

Keywords Chinese news summarization; Bidirectional Encoder Representations from Transformers model; Transformer model; Recurrent Neural Network

1 简介

中文新闻摘要任务，旨在从长篇新闻文本中提取出核心信息，并生成简短、准确的摘要。这一任务可以通过抽取式或生成式的方法实现：抽取式方法从原文中挑选出最重要的句子进行组合，而生成式方法则通过理解新闻内容，用自然语言生成新的摘要。中文新闻摘要任务需要准确捕捉新闻中的关键信息和主题，确保摘要简洁且涵盖主要内容。该任务在新闻聚合、信息检索和新闻推荐系统中有广泛应用，能够帮助用户快速获取新闻要点，提升信息获取效率。

在本研究中，我们着眼于抽取式方法，通过研究BERT模型的变体来提升模型在数据集上的性能。

2 相关工作

2.1 BERT模型

BERT（Bidirectional Encoder Representations from Transformers）是由Google AI开发的预训练语言模型。与传统的单向文本处理模型不同，BERT设计为能够同时从左到右和从右到左捕捉上下文信息，因此具有双向性。这使得BERT能够更好地理解单词在特定上下文中的含义，从而在各种自然语言处理（NLP）任务中取得了显著的表现。

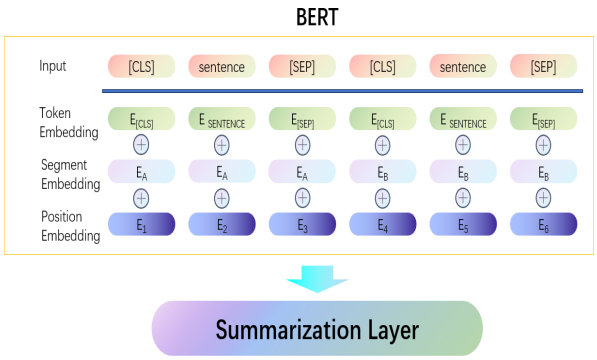


Fig. 1 BERT模型结构示意图

BERT模型基于Transformer架构，该架构采用自注意力机制（Self-Attention）来并行处理输入序列，而非按顺序处理。这种并行处理能力使得BERT能够更有效地处理文本中的长距离依赖关系。BERT通过在大规模语料库上进行预训练，主要通过以下两项任务进行训练：

- **掩码语言模型（Masked Language Model，MLM）：** 在该任务中，输入句子中的部分词语会被随机掩码，模型需要基于周围上下文来预测这些被掩码的词语。
- **下一个句子预测（Next Sentence Prediction，NSP）：** BERT训练时需要预测一个句子是否紧跟另一个句子，从而学习句子级别的关系。

完成预训练后，BERT可以通过添加任务特定的层，并利用标注数据集进行微调，以适应特定的下游任务，如文本分类、问答和摘要生成等。BERT在多项NLP基准测试中都取得了最先进的表现，成为该领域最广泛使用的模型之一。

2.2 Transformer模型

Transformer模型是由Vaswani等人在2017年提出的一种全新的神经网络架构，旨在解决序列到序列的学习任务，特别是在机器翻译中表现出色。Transformer模型的核心思想是完全基于自注意力机制（Self-Attention），不再依赖传统的循环神经网络（RNN）或长短期记忆网络（LSTM）。这种设计使得Transformer能够并行处理输入序列中的所有元素，从而大大提高了训练效率和模型性能。 Transformer模型由两大主要部分组成：编码

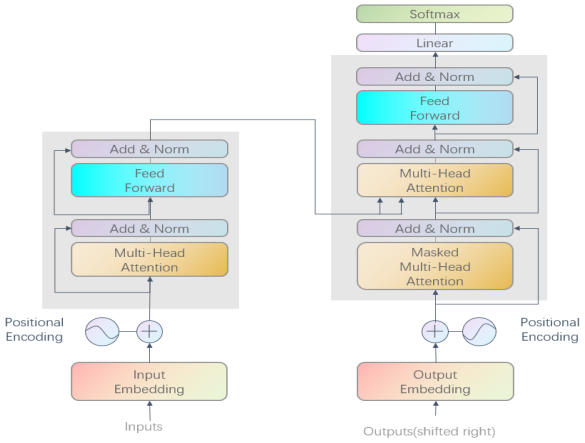


Fig. 2 Trsnformer模型结构示意图

器（Encoder）和解码器（Decoder）。每个部分都由多个相同的层堆叠而成，通常包含以下几个关键组件：

- **自注意力机制（Self-Attention）：** 自注意力机制允许模型在处理当前词语时，考虑输入序列中的所有其他词语，从而捕捉长距离的依赖关系。在每一层，输入的词向量会通过计算与其他词的相关性来加权处理，从而得到更丰富的上下文信息。

- **前馈神经网络（Feed-Forward Neural Network）：** 每个自注意力层之后都会接一个前馈神经网络，用于进一步处理信息。该网络通常由两个全连接层组成，并应用了激活函数。
- **位置编码（Position Encoding）：** 由于Transformer模型没有使用传统的序列结构（如RNN或LSTM），因此需要添加位置信息来保留词语在序列中的顺序。位置编码是通过在输入的词向量中加入位置信息来实现的。
- **多头注意力机制（Multi-Head Attention）：** 为了捕捉输入序列中不同子空间的信息，Transformer使用了多头注意力机制，即同时使用多个自注意力机制进行计算，并将其结果进行拼接。

与传统的循环神经网络（RNN）和长短期记忆网络（LSTM）不同，Transformer的并行计算特性使得它在处理长序列时表现出更高的效率和更强的建模能力。因此，Transformer模型被广泛应用于各种自然语言处理任务，如机器翻译、文本生成、文本分类等。

Transformer的成功激发了许多基于Transformer架构的变体，例如BERT和GPT，它们在各类NLP任务中取得了突破性进展。

2.3　循环神经网络（RNN）

循环神经网络（Recurrent Neural Network，简称RNN）是一类用于处理序列数据的神经网络模型。与传统的前馈神经网络不同，RNN能够通过隐藏状态（Hidden State）将序列中的信息传递到当前时刻，以捕捉时间步之间的依赖关系。因此，RNN广泛应用于时间序列预测、语音识别、自然语言处理等任务中。RNN的基本思想是通过

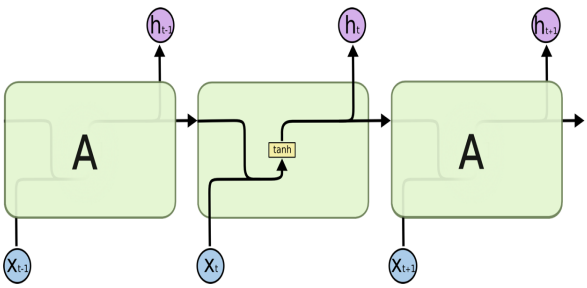


Fig. 3　RNN架构示意图

在网络中引入循环结构，使得网络能够在每个时间步上保留先前时刻的信息。在标准的RNN模型中，每个输入序列的每个元素都会依次输入到网络中，并且每个元素的输出不仅取决于当前的输入，还与前一时刻的隐藏状态有关。该结构可以通过以下公式表示：

$$h_t = \sigma(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

$$y_t = W_{hy}h_t + b_y$$

其中，h_t是当前时刻的隐藏状态，x_t是当前输入，y_t是当前输出，W_{hh}、W_{xh}和W_{hy}分别是权重矩阵，b_h和b_y是偏置项，σ是激活函数（通常使用tanh或ReLU）。

虽然RNN能够有效地处理序列数据，但其在长序列训练时存在梯度消失和梯度爆炸的问题，这使得网络难以捕捉到长期依赖关系。为了解决这一问题，出现了一些改进的RNN变体，如长短期记忆网络（LSTM）和门控循环单元（GRU），它们通过引入门控机制来控制信息流，从而能够更好地捕捉长期依赖。

- **长短期记忆网络（LSTM）：** LSTM是一种特殊类型的RNN，它通过引入遗忘门、输入门和输出门来调节信息的保留与更新，有效地解决了标准RNN中的梯度消失问题。

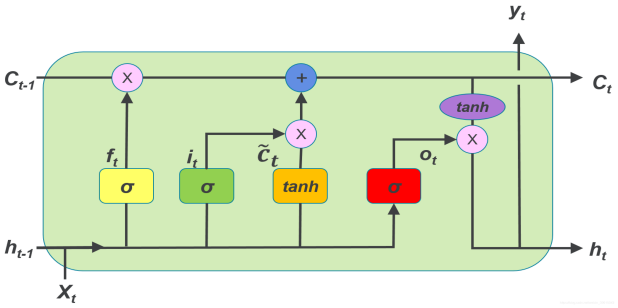


Fig. 4　LSTM架构示意图

- **门控循环单元（GRU）：** GRU是另一种改进的RNN模型，它将LSTM中的遗忘门和输入门合并为一个更新门，并简化了网络的结构，具有较少的参数，计算效率较高。

RNN及其变体在许多应用中取得了显著的成绩，尤其是在自然语言处理和语音识别等任务中。尽管近年来Transformer架构逐渐成为主流，但RNN和其变体仍在一些特定的应用场景中得到广泛使用。

2.4 ROUGE评估指标

ROUGE（Recall-Oriented Understudy for Gisting Evaluation）是一组用于自动评估文本摘要质量的指标，广泛应用于文本摘要生成、机器翻译等自然语言处理任务中。ROUGE的核心思想是通过比较生成的摘要与参考摘要之间的重合度来衡量摘要的质量。ROUGE指标主要包括基于词汇的精确度、召回率以及F1分数等度量方式。

ROUGE包括多个不同的子指标，常见的有以下几种：

- **ROUGE-N：** 计算生成摘要和参考摘要中n-gram（n元组）的重合度。常见的ROUGE-N包括ROUGE-1（单词重合）和ROUGE-2（2元组重合）。
- **ROUGE-L：** 计算生成摘要和参考摘要之间的最长公共子序列（Longest Common Subsequence, LCS）的重合度。ROUGE-L越高，说明生成摘要与参考摘要之间的结构越接近。
- **ROUGE-W：** 是ROUGE-L的一种加权版本，它通过对LCS进行加权处理，考虑了较长的公共子序列的质量。
- **ROUGE-S：** 计算生成摘要和参考摘要之间的2-gram（跳跃2-gram）重合度，能够捕捉到更多的语义信息。
- **ROUGE-SU：** 结合了ROUGE-S和ROUGE-Unigram（单元词）的评估方法，进一步加强了对语义信息的捕捉。

ROUGE评估指标通过计算生成摘要与参考摘要之间的n-gram、LCS等的重合情况，能够提供关于生成摘要的质量、信息覆盖和准确性的定量评估。它被广泛应用于机器翻译、摘要生成、文本生成等任务中，尤其是在自动摘要生成中，ROUGE是最常用的评估指标之一。

然而，ROUGE也存在一些局限性，主要体现在它主要依赖于词汇匹配，未能充分考虑语义相似性。为了克服这些问题，近年来的研究开始采用更为复杂的评估方法，如基于语义的评估方法和人类评估等。

3 实验方法

3.1 使用BERT提取摘要

在具体的实验操作中我们发现，由于BERT模型本身是作为一个掩码语言模型训练的，输出的向量是基于token的，而不是基于句子的。因此，为了使BERT能够用于抽取式摘要，我们需要对输入的序列和嵌入曾进行修改，以下是详细的解释：

3.1.1 BERT的输出和输入

BERT本身是一个掩码语言模型，它的输出是基于token（即单词或词元）的，而不是整个句子的表示。因此，BERT并不会直接为每个句子生成一个整体的向量表示。BERT在处理输入时，采用的是两句子格式（句子A和句子B），并使用两个句子标记（Sentence A 和Sentence B），但这不能满足抽取式摘要任务的要求。为了能够处理多个句子，且每个句子有独立的表示，需要对输入的序列和嵌入进行修改。

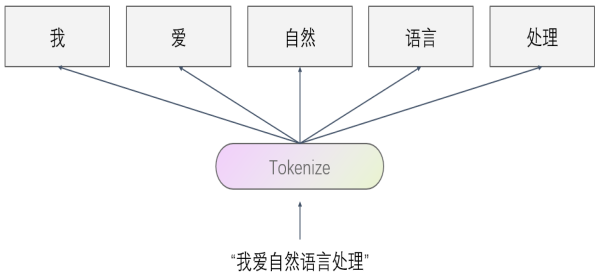


Fig. 5 分词示意图

3.1.2 修改BERT输入序列

为了使BERT能够为每个句子生成独立的表示，首先我们修改了BERT的输入格式。在标准的BERT中，输入序列通常是以一个[CLS]标记开始，后跟句子A和句子B，最后以[SEP]结束。为了处理多个句子，我们在每个句子的前面插入一个[CLS]标记，在每个句子的后面插入[SEP]标记。例如，对于多个句子，我们将输入格式修改为：

[CLS] [sentence 1] [SEP]
[CLS] [sentence 2] [SEP]
[CLS] [sentence 3] [SEP], . . .

这样，模型可以为每个句子生成独立的表示。

3.1.3 [CLS]标记的使用

在传统的BERT中，只有一个[CLS]标记用于聚合整个输入的特征（无论是单个句子还是句子对）。但是在此修改后，每个句子都会有自己的[CLS]标记，并且每个[CLS]标记对应一个句子的向量表示。这些[CLS]标记将被用来提取每个句子的特征，使得每个句子都有独立的表示。

3.1.4 段落嵌入（Segment Embeddings）

BERT使用段落嵌入（Segment Embeddings）来区分输入中的不同句子。在标准BERT中，只

有两个段落嵌入（Sentence A 和Sentence B），这适用于两个句子的输入格式。对于多句子的输入，我们为每个句子分配一个段落嵌入。为了区分不同的句子，我们采用交替的段落嵌入（例如$E_A \Delta E_B$）。例如，对于五个句子：[sentence 1, sentence 2, sentence 3, sentence 4, sentence 5]，我们将段落嵌入分配为：

$$[E_A, E_B, E_A, E_B, E_A]$$

。

3.1.5　句子表示的生成

通过修改输入格式和段落嵌入，我们使得BERT能够为每个句子生成独立的表示。对于第i个[CLS]标记，来自BERT模型的最上层的输出向量将作为该句子的表示。例如，第i个[CLS]标记对应的输出向量将作为第i个句子的特征向量。

通过上述修改，BERT模型能够处理多个句子的输入，并为每个句子生成独立的表示。这个方法使得BERT能够有效地应用于抽取式摘要任务，从而在每个句子层面生成摘要。这一修改使得BERT不仅适用于单句子的任务，还能够适应多句子、多个句子之间语境关系的抽取任务。

3.2　微调摘要层

在获得了来自BERT的句子向量之后，我们在BERT输出的基础上构建了几个针对摘要任务的层，用于捕获文档级别的特征，以便进行摘要提取。对于每一个句子，我们将计算最终预测的分数。整个模型的损失是每一个句子的分数和标签之间的二分类交叉熵。这些摘要层与BERT一起进行联合微调。

3.2.1　Classifier

简单分类器（Simple Classifier）如同原始的BERT论文中所描述的，是在BERT输出的基础上添加了一个线性变换，并通过Sigmoid函数来为每个句子生成预测分数。这种方法适用于需要对每个句子进行分类或评分的任务，如抽取式摘要生成。其公式如下：

$$Y_i = \sigma(W_o T_i + b_o)$$

其中：

- W_o 是权重矩阵，它将输入向量T_i 映射到输出分数。W_o 的大小取决于输入和输出的维度。
- T_i 是句子sent_i 的输入向量，它是从BERT模型中获得的输出。
- b_o 是偏置项，它是一个可学习的参数。
- σ 是Sigmoid函数，它将线性输出映射到0和1之间的概率值。

Sigmoid函数的定义为：

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

在这个模型中，Sigmoid函数用于为每个句子sent_i 生成一个预测分数Y_i，表示该句子是否应当被选入摘要。分数Y_i 越接近1，表示该句子越有可能被选为摘要的一部分。

3.2.2　Transformer

Transformer 是一种替代简单Sigmoid 分类器的方法，它通过在句子表示的基础上堆叠Transformer层，捕获文档级别的特征，最终通过线性层和Sigmoid 函数生成每个句子的得分。该方法相比于简单的分类器，能够更有效地提取文档级特征以更好地完成摘要任务。其主要思想是利用多头注意力机制和前馈神经网络提取句子间的全局关系。具体内容如下：

1. 基本结构　Inter-sentence Transformer 在BERT 输出的句子表示上应用多个Transformer 层，如图所示，每一层由两部分组成：

- **多头注意力机制（Multi-Head Attention, MHAtt）**：用于捕获句子间的依赖关系。

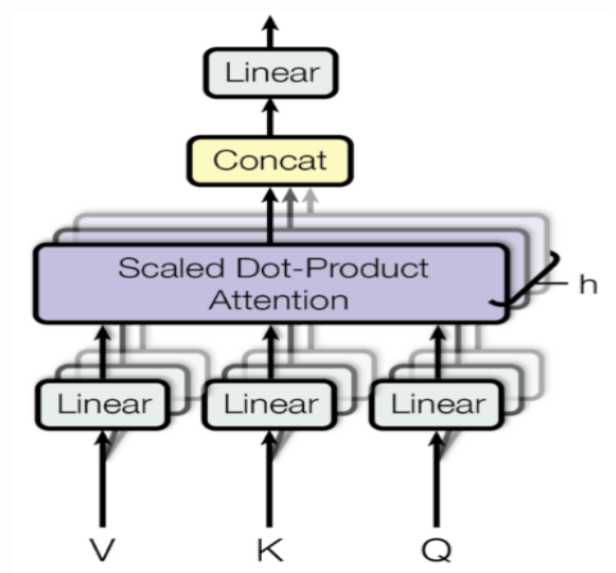


Fig. 6　多头注意力机制示意图

- **前馈神经网络（Feed-Forward Network, FFN）**：进一步处理注意力机制的输出。

每一层的处理过程如下：

(1) 首先，通过多头注意力机制和层归一化得到中间表示：

$$\tilde{h}^l = \text{LN}(h^{l-1} + \text{MHAtt}(h^{l-1}))$$

其中，h^{l-1} 是第$l-1$ 层的输出，LN 表示层归一化操作。

(2) 然后，通过前馈神经网络和层归一化得到当前层的最终输出：

$$h^l = \text{LN}(\tilde{h}^l + \text{FFN}(\tilde{h}^l))$$

其中，FFN 是前馈神经网络。

2. 输入表示

- 初始输入h^0 由BERT 输出的句子向量加上位置嵌入（Positional Embedding）得到：

$$h^0 = \text{PosEmb}(T)$$

其中，T 表示输入的句子向量，PosEmb 用于加入句子的位置信息。

- l 表示Transformer 层的深度。

3. 最终输出层 经过L 层Transformer 处理后，Inter-sentence Transformer 的最终输出通过一个线性层和Sigmoid 函数得到句子的预测分数：

$$\hat{Y}_i = \sigma(W_o h_i^L + b_o)$$

其中：

- h_i^L：表示第L 层输出的第i 个句子的向量。
- W_o：权重矩阵。
- b_o：偏置项。
- σ：Sigmoid 函数，用于将输出分数映射到区间[0, 1]。

3.2.3 循环神经网络（Recurrent Neural Network）

尽管Transformer模型在多项任务上取得了优异的表现，但循环神经网络（RNN）依然具有其独特优势，尤其是在与Transformer技术结合使用时表现更为突出[**?**]. 因此，我们在BERT输出的基础上，添加了一层长短期记忆网络（LSTM）来学习摘要任务的特定特征。由此，模型能够学习句子间的依赖关系。分层归一化的引入使得训练更加稳定，最终通过线性层和Sigmoid函数生成每个句子的预测分数，用于摘要任务。

1. 训练稳定性：分层归一化 为了稳定训练过程，我们在每个LSTM单元中引入了分层归一化（Layer Normalization），该方法由Ba 等人提出[**?**]。分层归一化有助于加速训练并提升模型稳定性。

2. LSTM单元的计算过程 在时间步i，LSTM层的输入是BERT的输出向量T_i，输出为隐藏状态h_i。LSTM单元的计算包括以下三个步骤：

(1) **计算门控机制：** 遗忘门F_i、输入门I_i、输出门O_i 和隐藏向量G_i：

$$\begin{pmatrix} F_i \\ I_i \\ O_i \\ G_i \end{pmatrix} = \text{LN}_h(W_h h_{i-1}) + \text{LN}_x(W_x T_i) \quad (5)$$

其中，W_h 和W_x 分别是权重矩阵，LN_h 和LN_x 为分层归一化操作。

(2) **更新记忆单元：** 通过遗忘门和输入门对记忆单元C_i 进行更新：

$$C_i = \sigma(F_i) \odot C_{i-1} + \sigma(I_i) \odot \tanh(G_i) \quad (6)$$

其中，σ 表示Sigmoid函数，$\tanh$ 表示双曲正切函数，$\odot$ 表示元素级乘法。

(3) **计算隐藏状态：** 使用输出门O_i 和当前记忆单元C_i 计算隐藏状态h_i：

$$h_i = \sigma(O_i) \odot \tanh(\text{LN}_c(C_i)) \quad (7)$$

其中，LN_c 为对记忆单元的分层归一化操作。

3. 最终输出层 经过LSTM层的处理后，最终输出通过一个线性层和Sigmoid函数计算句子的预测分数$\hat{Y}_i$：

$$\hat{Y}_i = \sigma(W_o h_i + b_o) \quad (8)$$

其中：

- W_o：权重矩阵；
- b_o：偏置项；
- σ：Sigmoid函数，将输出分数映射到$[0, 1]$ 区间。

4. 符号说明

- F_i, I_i, O_i：分别表示遗忘门、输入门和输出门；
- G_i：隐藏向量；

- C_i：记忆单元；
- h_i：隐藏状态（输出向量）；
- $\mathrm{LN}_h, \mathrm{LN}_x, \mathrm{LN}_c$：分层归一化操作。

4 实验过程

4.1 实验设置

本实验基于BERT 模型的”bert-base-chinese”版本进行中文新闻摘要任务的训练，训练过程中，批量大小设置为3000，累积梯度更新次数为2，学习率初始值设为2e-3，并使用Noam 学习率衰减方法。同时，为了稳定训练过程，引入了10000 步的学习率预热（Warmup Steps）。模型训练总步数为30000，训练过程中每50 步报告一次结果，并在每1000 步保存模型检查点。此外，Dropout 概率设置为0.1，以防止模型过拟合。

训练在单个GPU（RTX 4060Ti) 上进行，GPU 可见设备编号设置为1，gpu_ranks 设置为0，确保了高效的单卡训练。同时，为了进一步提升性能，实验中使用了间隔嵌入机制（Use Interval），并启用了三元组块重复检测功能（Block Trigram），以降低生成摘要中的冗余现象。此外，训练过程中启用ROUGE 评价指标，对摘要质量进行评估。

4.2 实验步骤

4.2.1 数据格式转换

原始的语料数据集应该包含两列，一列为原始的新闻文章段落内容，一列为文章所对应的摘要内容，将原来的数据集转换为json格式的文件。

4.2.2 分词以及分割文件

对json格式的文件进行切分，首先按照符号“。”，“！”，“？”进行分句，如果得到的句子数量少于两句，则用“，”，“；”进行进一步分句。

在分词之后，如果训练集的文件过大，可以考虑到分割成小文件便于后期训练。再进行分割之后，每个文件包含不超过16000条记录。

4.2.3 句子标注

对数据集中的句子进行预处理，根据ROUGE指标找出原始段落中与参考摘要最接近的n句话，标注为真。

4.2.4 模型训练

在顺利完成对数据的一系列处理之后，我们可以开始对模型的训练，我们可以任意选择三种编码器中的一种进行训练，同时还可以对模型的各项参数进行个性化的调整以更好适应不同的任务。

在本实验中，训练过程中的参数设置包括模型训练参数、训练环境参数和其他控制参数，具体如下：

1. 模型训练参数

- **Dropout 概率**（`-dropout`）：用于防止模型过拟合。可调范围如$0.1, 0.2, 0.3$ 等。
- **学习率**（`-lr`）：控制模型参数更新的步幅。可调范围如2e-3, 1e-3, 5e-4 等。
- **学习率衰减方法**（`-decay_method`）：控制学习率的衰减方式，可选值包括`noam`, `linear`, `none` 等。
- **训练步数**（`-train_steps`）：训练的总步数，决定模型的训练轮数。可调范围如$10000, 30000, 50000$ 等。
- **学习率预热步数**（`-warmup_steps`）：用于逐渐增加学习率，确保训练的稳定性。可调范围如$8000, 10000, 12000$ 等。
- **批量大小**（`-batch_size`）：影响训练速度与显存消耗。可调范围如$1000, 2000, 3000$ 等。
- **梯度累积次数**（`-accum_count`）：模拟更大的批量训练。可调范围如$1, 2, 4$ 等。
- **模型保存频率**（`-save_steps`）：模型检查点保存的步数间隔。可调范围如$500, 1000, 2000$ 等。
- **结果汇报频率**（`-report_every`）：控制训练过程中结果的打印频率。可调范围如$50, 100, 200$ 等。

2. 训练环境参数

- **可见GPU 设备**（`-visible_gpus`）：指定可用的GPU 编号，可调值如`0`, `1`, `0,1` 等。
- **GPU 排名**（`-gpu_ranks`）：定义GPU 编号与计算排名。可调范围如`0`, `0,1`。
- **世界大小**（`-world_size`）：表示训练所使用的GPU 数量。可调范围如1（单卡）、2（多卡）等。

3. 其他参数

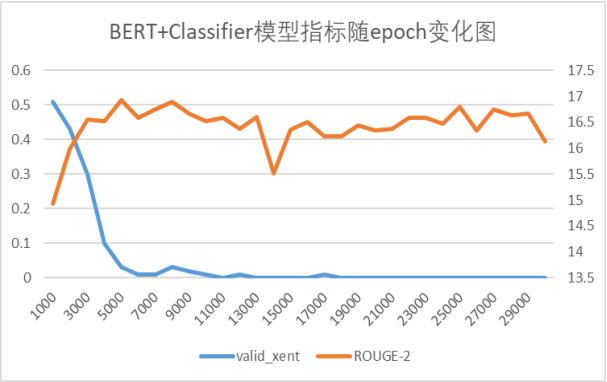
- **间隔嵌入机制**（`-use_interval`）：控制是否启用间隔嵌入机制，取值为`true`或`false`。
- **日志文件**（`-log_file`）：控制训练日志的输出路径，可根据实验不同配置进行修改。

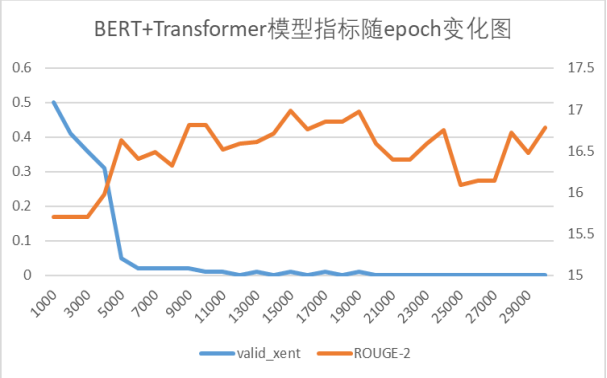
通过以上参数的调节，可以灵活控制模型的训练过程，包括学习率、训练步数、批量大小等核心参数，同时结合GPU 配置和训练策略，以优化中文摘要任务的模型表现。

5 实验结果

在模型训练完成之后，我们可以对训练好的模型进行测试，观察模型的性能，我们将分宏观和微观两部分对模型的性能进行评述。

5.1 模型指标随训练步数变化分析

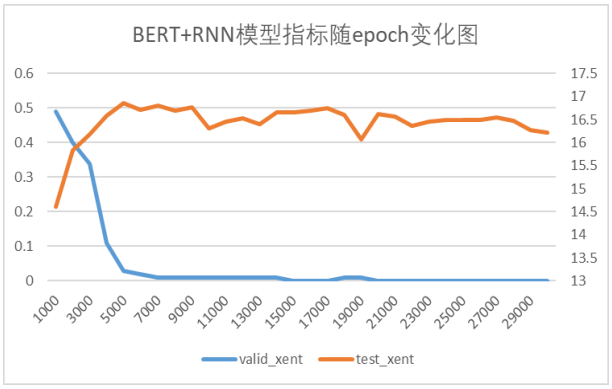




本实验对三种不同的模型架构在训练过程中进行了指标跟踪与分析，具体表现如下：

5.1.1 BERT+Classifier 模型

- 验证集交叉熵损失（`valid_xent`）在训练初期迅速下降，并在5000 步之后趋于稳定。



- ROUGE-2 得分在初期快速上升，最终稳定在16.5 左右，表现出收敛趋势。
- 在15000 步附近存在轻微波动，但整体趋势稳定。

5.1.2 BERT+Transformer 模型

- 验证集交叉熵损失与BERT+Classifier 类似，迅速下降并趋于平稳。
- ROUGE-2 得分稳定在16.5 ∼ 17.0 之间，表现优于BERT+Classifier。
- 训练过程中曲线波动较小，训练稳定性较好。

5.1.3 BERT+RNN 模型

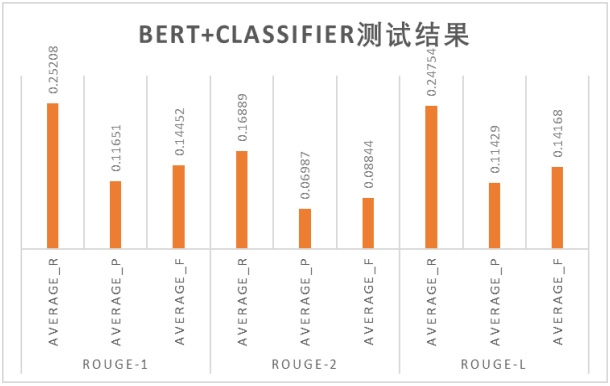
- 验证集交叉熵损失和测试集交叉熵损失（`test_xent`）均迅速下降，并在早期达到稳定状态。
- ROUGE-2 得分在训练初期提升较快，最终维持在16.5 左右。
- 验证损失和测试损失同步收敛，说明模型在训练和测试阶段表现一致。

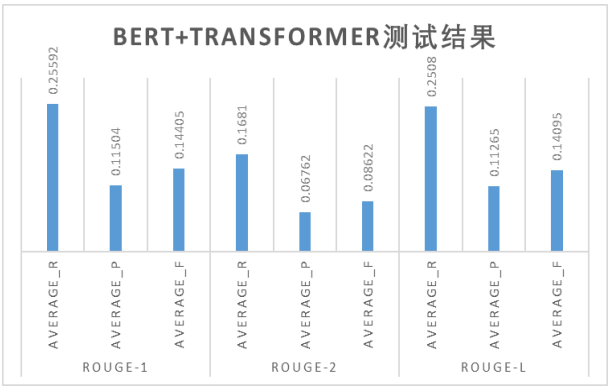
5.1.4 总结

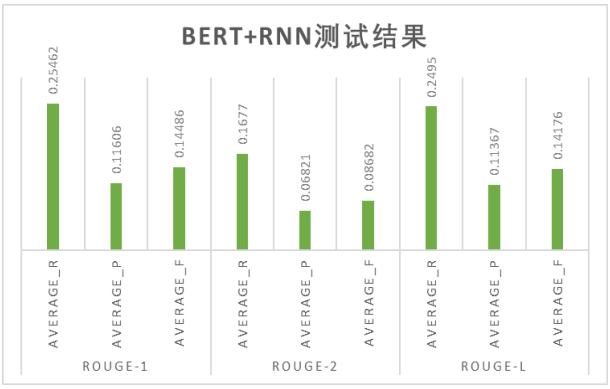
三种模型的验证损失均在训练初期迅速下降，表现出较好的收敛特性：

- BERT+Classifier 模型在训练初期提升较快，但中期略有波动。
- BERT+Transformer 模型的ROUGE-2 得分表现最好，训练曲线最为稳定。
- BERT+RNN 模型的验证损失与测试损失收敛一致，ROUGE-2 表现稳定。

总体来看，BERT+Transformer 模型在指标表现和训练稳定性方面略优于其他两种模型。







5.2 模型最佳性能

本实验针对三种不同架构的BERT模型（BERT+Classifier、BERT+Transformer和BERT+RNN）进行了对比分析，评估指标包括ROUGE-1、ROUGE-2 和ROUGE-L 的平均召回率（AVERAGE_R）和平均F1分数（AVERAGE_F）。具体结果总结如下：

5.2.1 ROUGE-1 指标

BERT+Transformer 在平均召回率上表现最好，达到0.25592，略高于BERT+Classifier和BERT+RNN。BERT+RNN 在平均F1分数上取得最优表现，得分为0.14486。

5.2.2 ROUGE-2 指标

BERT+Classifier 在平均召回率上表现最佳，得分为0.16889。BERT+RNN 和BERT+Transformer在平均召回率和F1 分数上的表现接近，但略低于BERT+Classifier。

5.2.3 ROUGE-L 指标

BERT+Transformer 在平均召回率上表现最优，得分为0.2508。BERT+RNN 在平均F1分数上略优于其他两个模型，取得了0.14176 的得分。

5.2.4 综合分析

通过以上结果分析可以看出：

- BERT+Classifier 在ROUGE-2 平均召回率上表现最优，说明该模型在捕捉句子之间的深层关系方面具有一定优势。
- BERT+Transformer 在ROUGE-1 和ROUGE-L 平均召回率上表现较好，体现了其在整体信息覆盖方面的优势。
- BERT+RNN 在ROUGE-1 和ROUGE-L 平均F1分数上稍有优势，表明其在摘要结果的精确性与召回率平衡方面表现较为稳定。

5.2.5 总结

三种模型在不同评估指标上的表现各有侧重，具体如下：

- BERT+Classifier 在ROUGE-2 平均召回率上最优。
- BERT+Transformer 在ROUGE-1 和ROUGE-L 平均召回率上表现最佳。
- BERT+RNN 在ROUGE-1 和ROUGE-L 平均F1分数上略胜一筹。

总体而言，三种模型在不同指标上的表现接近，各具优势，说明不同结构在摘要任务中对性能表现的侧重有所不同。BERT+Transformer 更适合追求信息覆盖率，BERT+RNN 在平衡精确率与召回率方面表现优异，而BERT+Classifier 在局部关系捕捉中具有较好的性能。

5.3 部分样例

表1 表说明模型预测部分样例

标签	模型预测结果
三亚村干部嫁女摆50 多桌宴席遭举报超标了	记者暗访发现，虽是在村里自办酒席，但婚宴酒桌数目在50桌以上。
“ 威马逊” 加强为强台风	今年第9 号台风“ 威马逊” 1 5 日加强为强台风。
央视春晚主持人曝光！张国立徐帆入选朱军缺席	距离马年春晚还有20 多天，主持人名单终于曝光！
另类“ 猪一样的队友”	广西南宁一男子乘公交车时，目睹一矮个猥琐男子成功性骚扰一女子后心生嫉妒，他如法炮制，不料被车上乘客制服。
一部手机引发的“ 监狱桃色风云”	“ 有些鸟儿是关不住的。

在上方的表格中，我们列举了在模型预测的结果中相对而言有代表性的一些结果，从这六个例子中，我们可以清晰的了解到，由于我们模型的范式为抽取式摘要，是通过对原始新闻段落中的内容进行选择和拼接来实现摘要，因此当样本中的原始段落中存在总结性的语句或者与标签相似度高的语句的时候，模型就能很好地总结文段，但是当样本的标签语句没有在原始段落中直接出现，或者标签语句采用了一些诸如比喻等手法进行修辞的情况下，模型的性能就会出现下降，摘要的效果就会很不理想，例如以上表格中的”另类’猪一样的队友‘“和”一部手机引发的’监狱桃色风云‘“等。

6 总结

本文针对中文新闻摘要生成任务，提出了基于微调BERT 模型的解决方案。通过在预训练的BERT 模型基础上，分别添加了三种不同的摘要提取层：线性分类器层Classifier、Transformer 层和循环神经网络RNN层，系统地探讨了不同模型架构在该任务上的表现。主要研究成果总结如下：

- **实验方法**：修改了BERT 输入序列，使其能够独立为每个句子生成特征向量。通过在BERT 输出的基础上添加不同的摘要提取层，捕获了文档级别的特征，并利用联合微调实现摘要生成。
- **实验结果**：通过在中文新闻数据集上进行实验，三种模型在ROUGE-1、ROUGE-2 和ROUGE-L 三个指标上均取得了较好的效果，具体结果如下：
  - BERT+Classifier 模型在**ROUGE-2 平均召回率** 上表现最优，得分为16.89。
  - BERT+Transformer 模型在**ROUGE-1 和ROUGE-L 平均召回率** 上表现最佳，分别达到25.592 和25.08。
  - BERT+RNN 模型在**ROUGE-1 和ROUGE-L 平均F1 分数** 上表现略优，分别达到14.486 和14.176。
- **综合分析**：三种模型在不同评估指标上表现各有侧重，具体表现如下：
  - BERT+Classifier 模型在捕捉句子之间的深层关系方面具有优势。
  - BERT+Transformer 模型在整体信息覆盖和训练稳定性方面表现更优。
  - BERT+RNN 模型在精确率与召回率的平衡上表现较好。
- **问题与局限性**：实验表明，模型性能在很大程度上取决于输入文本的质量。当原文段落中存在总结性语句时，模型能够较好地提取摘要。然而，当标签语句未直接出现在原文中，或者采用修辞手法（如比喻、隐喻）时，模型的表现有所下降。

未来工作：未来的研究将继续优化模型架构，并尝试引入语义匹配和生成式摘要方法，以进一步提升中文新闻摘要的准确性与流畅性。此外，结合领域特定的数据集，探索模型在其他任务场景中的应用潜力。

参 考 文 献

[1] homas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, et al.. 2020. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

[2] Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805, 2019. Available: https://arxiv.org/abs/1810.04805.

[3] Liu Yang, Lapata Mirella. Text Summarization with Pretrained Encoders. arXiv preprint arXiv:1908.08345, 2019. Available: https://arxiv.org/abs/1908.08345

[4] Hermann Karl Moritz, Kočiský Tomáš, Grefenstette Edward, Espeholt Lasse, Kay Will, Suleyman Mustafa, Blunsom Phil. Teaching machines to read and comprehend. arXiv preprint arXiv:1506.03340, 2015. Available: https://arxiv.org/abs/1506.03340.

[5] Vaswani Ashish, Shazeer Noam, Parmar Niki, Uszkoreit Jakob, Jones Llion, Gomez Aidan N., Kaiser Lukasz, Polosukhin Illia. Attention is all you need. arXiv preprint arXiv:1706.03762, 2023. Available: https://arxiv.org/abs/1706.03762.

Background

This paper addresses the problem of Chinese news summarization in the field of Natural Language Processing (NLP). News summarization condenses long articles into shorter summaries while preserving key information. Traditional methods, such as rule-based or extractive approaches, struggle to capture deep semantic relationships. With the rise of deep learning, especially transformer-based models like BERT, significant advances have been made in various NLP tasks, including summarization. However, challenges remain in applying these models to Chinese text due to linguistic and structural differences.

Internationally, the research community has made significant progress using neural networks, especially pre-trained models like BERT, GPT, and T5. While these models have achieved high performance in summarization tasks, they often focus on extractive summarization or struggle with generating coherent abstractive summaries in languages like Chinese. To address this, methods combining multiple neural network architectures, such as linear layers, Transformer layers, and RNN networks, have shown promise in improving summarization quality.

This paper proposes enhancing Chinese news summarization by integrating BERT with a linear classification layer, a Transformer layer, and an RNN network, aiming to generate more accurate and fluent summaries. The project is part of ongoing research to advance deep learning for natural language understanding, with the potential to improve news summarization systems in domains like media, search engines, and content aggregation platforms.