

摘要

由于新冠肺炎 (COVID-19) 疫情的特殊性，与之相关的文本资料呈现出短时间内规模爆炸性增长且涉及内容多样化、变化大的特点。因此为了对其文本资料进行全面的研究，有助于更好地把握病毒的发展趋势和变异方向，自动摘要生成技术起着关键性的作用。本报告中将经过了微调的 BERT(bidirectional encoder representations from transformers) 模型应用于基于公开文献资料数据生成对应的医学文本摘要任务。本模型框架主要由编码器模块、解码器模块和训练模块组成。实现摘要生成的核心部件包括含有 BERT-Embedding 和 BiLSTM 的编码器和解码器结构以及连接不同模块的多层神经网络，其使模型具有多任务学习的能力。模型中还应用了不同的注意力机制。和其他模型相比，本模型在创建的数据集上展现出了最强的性能。因此，本模型能够较好地适用于 COVID-19 相关资料的文本摘要生成任务。

目	录
1 概述	4
2 相关工作	4
3 模型	5
3.1 BERT	5
3.2 摘要生成模型构建	7
3.2.1 编码器	7
3.2.2 解码器	8
4 实验	9
4.1 数据集	9
4.2 评价指标	10
4.3 实验设置	11
4.4 实验结果	11
5 总结	12

1 概述

自新冠肺炎 (COVID-19) 病毒出现以来，病毒毒株随着时间的推移不停地产生变异，同时由于对病毒的研究持续推进，人们对该病毒的认识与理解也在不断增加。为了更好地对新冠病毒相关的研究进行一个系统的了解与回顾，掌握新冠肺炎相关信息的推演和变化，我们需要对新冠肺炎相关的文献进行理解。为了更高效地研读文献，我们可以使用自动文本摘要生成 (Automatic text summarization, ATS) 技术。

作为信息抽取中的一个重要类别，自动文本摘要生成是一项具有一定难度的自然语言处理 (NLP) 任务 [1]。它通过生成不会损失原文本核心含义的摘要完成快速阅读和较长文本的信息检索任务。这使得使用者可以快速而准确地在大规模数据中找到所需要的信息，解决了大规模低密度数据的存在和使用者对于高效准确的信息获取能力的需求之间的矛盾。在新冠肺炎相关的医学领域，文本类数据呈现着在短时间内迅速增多的爆发型趋势，因此高质量的文本摘要生成方法在数据分析中有着不可或缺、无法替代的地位。尤其是当涉及到中文文本摘要生成时，现有的高性能摘要生成模型均会出现一定的局限性 [2]。

本报告中使用 BERT 为基础创建了一个拥有较高性能的自动文本摘要生成模型。针对中文文本，模型中使用了一种基于 BiLSTM 的中文文本编码方法。模型将语料数据通过预处理输入 BERT 模型中进行预训练，可以最大程度上利用文本中的信息来强化词向量中的语意表征，从而获得更好的词向量表征。生成的词向量通过多层编码器和多层解码器的处理后，最终形成文本摘要。我们使用 ROUGE 和 BLEU 指标对生成的文本摘要进行评估，实验结果表明，本模型在通过给定的文献摘要生成对应的文献标题时展现出了较强的摘要生成能力，其生成结果能较好地符合真实参考文本并覆盖原文本中的大部分核心信息。

2 相关工作

自动文本摘要生成的方式通常可分为抽取式和生成式两类。抽取式文本摘要主要分为三个部分：(1) El-Kassas 等人 (2020) 提出的输入文本预处理 [3]。(2) 文本处理步骤。首先是 Joshi 等人 (2018) 提出的为输入文本创建合适的表征，以便于文本分析 [4]；然后是 Nenkova and McKeown (2012) 提出的根据输入文本表征对文本中的句子进行打分并排序 [5]；接着是 Shuai 等人 (2017) 提出的对其中最重要的句子进行筛选与拼接以生成一个新的摘要文本 [6]。(3) 对生成的摘要文本进行分析与处理，如对其中的句子进行重排等。近些年，抽取式文本生成被应用于许多领域并取得了好的效果。Mao 等人 (2019) 提出了一种单文本的摘要抽取方法，其结合了监督学习和无监督学习 [7]。Wang 等人 (2021) 针对长文本摘要抽取的问题，提出了一种用于长中文文档的抽取方式 [8]。他们的研究表明，基于 BERT 的模型在处理长中文文档时提高了正确率并减少了文本冗余。但是，和生成式摘要生成相比，抽取式摘要生成存在着抽取结果冗余较多、语意确实、拼接生硬和关键信息不能完全覆盖等问题。

生成式文本摘要对原文本深入分析的要求更高，它是在理解给定文本中语意的基础上，

在字符层面进行压缩和精炼，并用更少的字符和更精确的语言生成摘要的 (Al-Abdallah and Al-Taani,2017)[9]。它能够生成全新的句子而不是只在原文本中选择句子进行组合。随着深度学习的进步，越来越多的深度学习模型被用于生成式文本摘要。Sutskever(2014) 首先提出了 Seq2seq 的结构 [10]。这一结构由一个编码器和一个解码器组成，两者都是用神经网络来实现的。这一阶段中，生成式文本摘要任务主要依靠的都是 Seq2seq 的结构。Nallapati 等人 (2016) 将注意力机制融入模型，使其可以对每一时刻输入的向量表征赋予不同的权重 [11]。Hu 等人 (2015) 使用新浪微博短文本的大规模中文文本摘要数据集提出了一种带有注意力机制的基于 RNN 的 Seq2seq 模型并取得了较好的结果，为短文本的摘要生成研究奠定了基础 [12]。Siddiqui 和 Shamsi(2018) 提出了用于 Seq2seq 的部分注意力 (local attention) 机制并在解决重复词问题中取得了较好的效果 [13]。Shashi Narayan 等人 (2018) 使用了基于卷积神经网络的 Seq2seq 结构来自动生成 BBC 线上文章的摘要 [14]。Liang 等人 (2020) 提出了一种选择性强化的 seq2seq 注意力模型来生成社交媒体上文本的摘要并结合了交叉熵损失函数和其他强化学习的策略来优化模型的 ROUGE 指标 [15]。

3 模型

3.1 BERT

BERT 是一个能够使用大量文本进行无监督预训练的语言模型，是本摘要生成模型的基础。BERT 含有一个双向 Transformer 编码层，可以通过有条件的预处理过程对无标签数据的深层双向表示进行预训练，从而更好地捕捉输入句子中的双向联系。

为了应对不同人物中对输入的需求，BERT 模型可以仅输入一个句子或者是将两个句子拼接在一起。BERT 的输入被分为三层，分别是字符编码层，分块编码层和位置编码层，如图1所示。BERT 中的字符编码层把每个词转换为一个 768 维的向量表示。在首个文本被输入到字符编码层前会进行分词操作，分词方法为 WordPiece。分块编码层的作用是区分不同的句子。在掩码语言模型 (masked language model) 外，预训练操作也被应用于判断任意两个句子之间顺序的分类任务，其中顺序较前的句子中每个字符都被表示为 0，顺序较后的句子中每个字符都被表示为 1。位置编码层是在 BERT 中通过训练获得的。

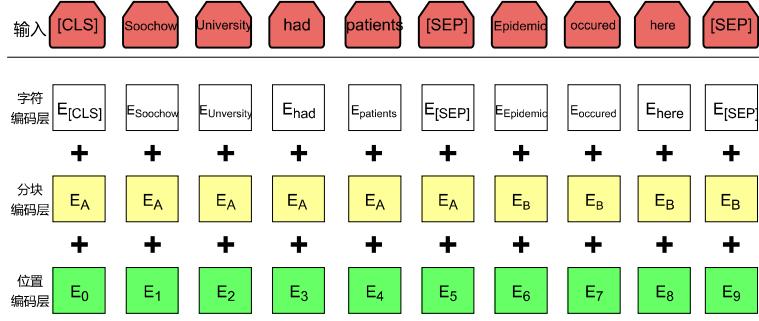


图 1: BERT 输入结构

作为一个无监督的自然语言处理模型，BERT 的一个核心部分是其自注意力机制。自注意力机制和位置编码一同解决了文本内时间关联的问题。这一机制能够在信息向前传播时动态计算其权重，实现步骤如下：

1. x^1, x^2, x^3 和 x^4 代表 4 个输入的句子，每个句子都经过与编码矩阵相乘得到 a^1, a^2, a^3 和 a^4 。

$$a^i = W a^i \quad (1)$$

2. 每个 a^i 都与三个变化矩阵分别相乘得到三个向量 q, k 和 v ，这三个向量的维度相同。

$$q^i = W^q a^i \quad (2)$$

$$k^i = W^k a^i \quad (3)$$

$$v^i = W^v a^i \quad (4)$$

3. 每个 q 都被视作对每个对应 k 的注意力，在自注意力算法中使用了带缩放的点积。

$$a_{1,i} = q^1 \bullet k^i / \sqrt{d} \quad (5)$$

其中 d 是 q 和 k 的维度。4. $a_{1,1}$ 到 $a_{1,4}$ 通过 softmax 层进行归一化得到 $\hat{a}_{1,1}$ 到 $\hat{a}_{1,4}$ 。 v^1 到 v^4 会被相乘并加至 $\hat{a}_{1,1}$ 到 $\hat{a}_{1,4}$ 来得到第一个输出向量 b_1 。之后，重复上述步骤来得到输出向量 b_2, b_3 和 b_4 。此外，BERT 中还使用了多头自注意力机制，公式如下：

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \quad (6)$$

$$\text{Multi-head}(Q, K, V) = \text{Concat}_i(\text{head}_i)W^0 \quad (7)$$

3.2 摘要生成模型构建

本报告中提出的摘要生成模型结构如图2所示，其中包含两个核心模块：编码器（Encoder）和解码器（Decoder）。

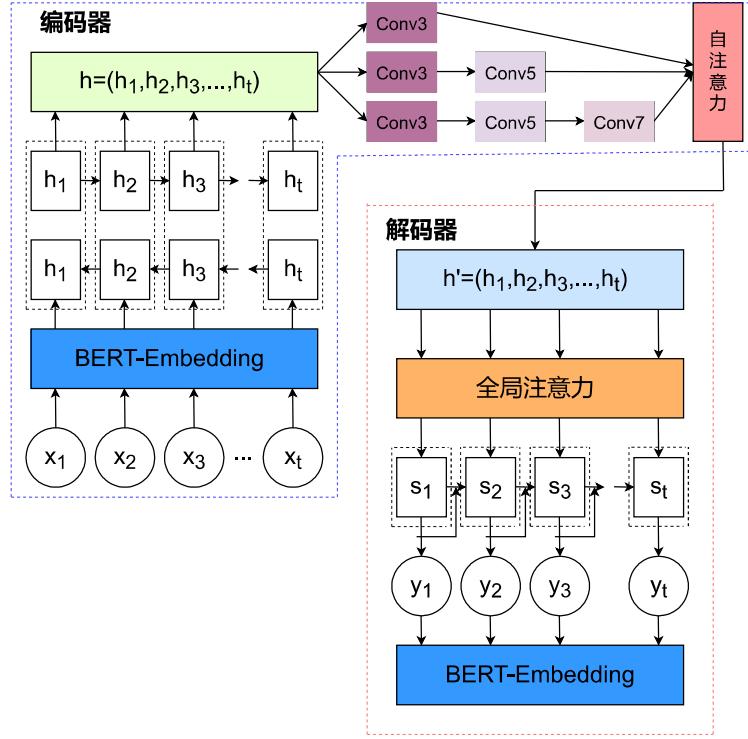


图 2: 摘要生成模型结构

3.2.1 编码器

如图2所示，编码器主要由四个部分构成：一个经过微调的 BERT-Embedding 层、一个 BiLSTM 层、一系列卷积门控单元和一个自注意力机制层。BERT-Embedding 层的作用是基于词嵌入的方式对输入序列进行初始化，之后，BiLSTM 层接收前一层的输入并对其进行编码。卷积门控单元旨在根据每个时间步中前一层的输出对其中的核心信息进行重训练。自注意力机制层的目的是探寻信息之间彼此的联系并由此进一步强化全局信息。

假设 $X_T = (x_1 \oplus x_2 \oplus x_3 \dots \oplus x_T) \subseteq R^{1 \times T}$ 是长度为 T 的输入序列，其中 x 代表该序列的基本单元， \oplus 代表拼接方式。源文本输入后，其中的基本单元会被通过词嵌入操作转换为向量，这一过程是使用预训练的 BERT 模型进行初始化的。在本报告中，使用的是 BERT-Base Chinese 模型作为初始化模型。

假设 $h = GLU(h_1, h_2, h_3, \dots, h_t) \subseteq R^{T \times dim}$ 是 BiLSTM 层对输入进行双向编码后输出的结果。GLU 代表门控线性单元 h_i 表示时间步 i 中编码的隐状态，可表示为 $h_i = [\vec{h}_i; \overleftarrow{h}_i] \in$

$$R^{T \times 2dim}.$$

模型在 BiLSTM 层之后应用了一系列的卷积操作来捕捉语句之间的内在连接，特别是 n-gram 特征。此外，过滤器和感受野也在考虑了输入序列平均长度的基础上被应用于捕获更丰富的语句内在连接。之后， $q \in R^{|q|}$ 和 $w^k \in R^k$ 两个向量会通过卷积形成一个总的特征表示 $m \in R^{|q|-k+1}$ ，计算过程为：

$$m_i = \text{ReLU}(w^k * q)_i = \text{ReLU}(w^k * q_{[i:i+k-1]}) = \text{ReLU}\left(\sum_{j=1}^{i+k-1} w_j^k q_j\right) \quad (8)$$

接收卷积操作得到的输出后，一个自注意力机制层被用于挖掘和捕捉其中的全局连接，这会使模型可以对长程依赖进行学习，且不过多地消耗算力资源。自注意力机制在每个时间步中对全局信息和相关标注建立连接的公式为：

$$\text{self-Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}V\right) \quad (9)$$

其中 Q 和 V 表示卷积操作中生成的矩阵，K 表示一个可学习的矩阵。

卷积模块用于获取输入序列中的 n-gram 特征，自注意力机制用于捕获输入序列中的长程依赖，所以在二者之后使用了一个门控单元来进行全局编码，来生成编码器最终的输出，用公式表示为：

$$\tilde{h} = h \odot \sigma(g) \in R^{T \times dim} \quad (10)$$

3.2.2 解码器

编码器输出的文本向量已对整个输入序列进行编码，接下来解码器将对这一向量中包含的信息进行解码来生成输出序列。如图2所示，解码器主要由三个部分构成：全局注意力层、BiLSTM 层和 BERT-Embedding 层。

全局注意力层对所有编码器中的隐状态进行考虑并由此在每一个解码步骤中定义一个基于注意力的文本向量，从而保证编码器能够在每个时间步将序列编码成不同的文本向量且在解码时能够组合这些不同的文本向量进行解码从而获取更准确的输出结果。BiLSTM 层可以对编码器输出的句子进行解码并根据输入的文本向量和当前输出序列对下一个词进行预测，同时获取最终的输出序列。BERT-Embedding 层可以有效地捕捉句子之间的语意联系，并在每一刻都生成一个完整的对应文本，使最终生成的摘要更加精确。

全局注意力层的核心思想是对所有编码器中的隐藏层状态进行考虑。 a^t 是一个长度可变的对齐向量，其长度和对应的时间序列中编码器部分的长度一致，是通过将当前解码器隐藏状态 h^t 隐藏层状态相比较得到的，公式为：

$$a_t(s) = \text{align}(h_t, \bar{h}_s) = \frac{\exp(score(h_t, \bar{h}_s))}{\sum_{s'} \exp(score(h_t, \bar{h}_{s'}))} \quad (11)$$

其中 score 是一种打分机制，其公式如下：

$$score = v_a^T \tanh(W_a[h_t; \bar{h}_s]) \quad (12)$$

所有 $a_t(s)$ 的值都被集成进一个权重矩阵得到最终的 W_a ，并接着求出 a_t ：

$$score = v_a^T \tanh(W_a[h_t; \bar{h}_s]) \quad (13)$$

对 a_t 进行加权平均操作即可获得对应的文本向量 c_t 。解码器中 BiLSTM 的计算过程如下：

1. 计算遗忘门并选择需要被遗忘的信息。这一步的输入是前一个时刻中的隐藏层状态 h_{t-1} 和当前时刻的输入词 x_t ，输出的是遗忘门的值 f_t 。

$$f_t = \sigma(W_f \bullet [h_{t-1}, x_t] + b_f) \quad (14)$$

2. 计算输入门并选择需要被记忆的信息。这一步的输入是前一个时刻中的隐藏层状态 h_{t-1} 和当前时刻的输入词 x_t ，输出的是输入门的值 f_t 和新细胞记忆 \tilde{C}_t 。

$$i_t = \sigma(W_i \bullet [h_{t-1}, x_t] + b_i) \quad (15)$$

$$\tilde{C}_t = \tanh(W_c \bullet [h_{t-1}, x_t] + b_C) \quad (16)$$

3. 计算当前时刻的细胞状态。这一步的输入是遗忘门的值 f_t 、输入门的值 i_t 、新细胞记忆 \tilde{C}_t 和前一个时刻的细胞状态 C_{t-1}

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (17)$$

4. 计算输出门以及当前时刻的隐藏层状态。这一步的输入是前一个时刻的隐藏层状态 h_{t-1} ，输入词 x_t 和当前时刻的细胞状态 C_t ，输出的是输出门的值 o_t 和当前时刻隐藏层状态 h_t 。

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (18)$$

$$h_t = o_t * \tanh(C_t) \quad (19)$$

4 实验

4.1 数据集

本报告中使用的数据集主要是从包括《中华医学杂志》、《南方医科大学学报》、《北京大学学报·医学版》等在内的出版文献数据中使用关键词 COVID-19 检索得到的。本报告从这

些出版文献中收集每篇文章的摘要和标题并且通过数据预处理对收集得到的数据进行清洗，之后使用 Jieba 中文分词库将中文原文本重组为词序列便于深入分析。在数据预处理的过程中，去除了标题长度在 [10,70] 之外且对应摘要长度在 [200,1100] 之外的数据，同时将数据中和文献作者、出版刊物等相关的和摘要生成任务无关的信息全部过滤。

4.2 评价指标

本报告中使用了 ROUGE 指标对文本摘要生成的性能进行评估。这一方法计算模型生成的摘要 (\hat{Y}) 和原摘要 (Y) 之间的相似程度并进行评估。ROUGE-1、ROUGE-2、ROUGE-L 三个指标被用于评估本模型的性能。

ROUGE-1 用于计算 \hat{Y} 和 Y 之间的一元语言重合度，ROUGE-2 用于计算 \hat{Y} 和 Y 之间的二元语言重合度。ROUGE-1 和 ROUGE-2 的召回率计算公式为：

$$ROUGE-n = \frac{\sum_{S \in Y} \sum_{gram_n} Count_{match}(gram_n)}{\sum_{S \in Y} \sum_{gram_n \in S} Count(gram_n)} \quad (20)$$

ROUGE-L 的计算利用了 \hat{Y} 和 Y 的最长公共子序列，计算公式为：

$$R_{lcs} = \frac{LCS(Y, \hat{Y})}{|Y|} \quad (21)$$

$$P_{lcs} = \frac{LCS(Y, \hat{Y})}{|\hat{Y}|} \quad (22)$$

$$F_{lcs} = \frac{(1 + \beta^2)R_{lcs}P_{lcs}}{R_{lcs} + \beta^2P_{lcs}} \quad (23)$$

其中 $LCS(Y, \hat{Y})$ 表示 \hat{Y} 和 Y 的最长公共子序列， β 表示一个默认的超参数， F_{lcs} 表示 R_{lcs} 和 P_{lcs} 之间的 F 值，即 ROUGE-L 的得分。

此外，BLEU 指标也被用于评估本模型的性能。这一指标是用于评估模型生成的句子和实际句子的差异的指标。它的取值范围在 0.0 到 1.0 之间，如果两个句子完美匹配，那么 BLEU 是 1.0，反之，如果两个句子完美不匹配，那么 BLEU 为 0.0。BLEU 方法的实现是分别计算模型生成的句子和实际句子的 N-grams 模型，然后统计其匹配的个数来计算得到的，公式如下：

$$BLEU = BP \times \exp\left(\sum_{n=1}^N w_n \log P_n\right) \quad (24)$$

$$BP = \begin{cases} 1 & if \quad c > r \\ e^{1-r/c} & if \quad c \leq r \end{cases} \quad (25)$$

4.3 实验设置

所有的实验都是在一个 NVIDIA GTX 2080TI GPU 上的 Linux 操作系统中完成的。在本实验中，单个句子的最长长度设定为 256，批大小 (batch-size) 为 16。BiLSTM 的隐单元维度为 512，和单向 LSTM 隐单元的维度设置是一致的。实验中使用了 Adam 优化器对损失函数进行优化，学习率为 1e-5。编码器和解码器中词嵌入操作的维度都是 768。

为了进行对比实验，报告中选择了两个常用于摘要生成的模型一同进行实验和评估：

RNN-context: 该模型将基础的 RNN 网络和基于文本的 RNN 网络相结合，使用 Seq2seq 的基本结构去完成文本摘要生成任务。和基础的 RNN 网络不同的是，基于文本的 RNN 网络使用了对文本内容的注意力机制来加强模型性能。

Super-AE: Super-AE 模型是一个使用了辅助监督机制来捕捉语意表征并由此生成更合理、更高质量的文本摘要的自编码器 (autoencoder) 模型。

4.4 实验结果

如表1所示，本模型的 ROUGE-1 值达到了 0.785，ROUGE-2 值达到了 0.731，从这两组数据可以得出模型生成的摘要中的大部分词都和原文中的摘要相同，但可能含有部分冗余。本模型的 ROUGE-L 值达到了 0.811，表示模型对最终生成的摘要进行了语法上的优化。综合以上结果，可以得出本模型拥有优于 RNN-context，且和 Super-AE 相仿的摘要生成能力，模型性能较好。这是由于基于 BERT 的模型可以在无监督的预训练中学习到大部分的语言信息，且可以在语意层面更好地对输入序列进行表征。

表 1: 模型评估结果

模型	ROUGUE-1	ROUGUE-2	ROUGUE-L	BLEU
RNN-context	0.712	0.683	0.759	0.617
Super-AE	0.782	0.735	0.798	0.628
本模型	0.785	0.731	0.801	0.640

实验中通过模型根据给定的文献摘要部分生成对应的文献标题来展现模型的摘要生成能力。训练后的模型在测试集上的生成结果和真实的文献标题之间的对照案例如表2所示。从表中的结果可以发现，本模型生成的摘要可以覆盖原文本中的大部分核心信息，与真实标题符合率较高，进一步证明了本模型的有效性。

表 2: 模型生成示例

案例 1: 归纳患者的人口学资料、疫苗接种情况、临床症状、生化及影像学检查、治疗措施与结局预后；根据患者既往病史，分为既往感染 COVID-19 者和无 COVID-19 感染史者。得到新冠肺炎疫苗相关心肌炎好发于男性（91.5 %），多见于接种第二针 mRNA 疫苗后（81.7 %），平均起病时间为 3 (1,25) d，主要临床表现为胸痛（94.4 %）、发热（45.1 %）、肌肉疼痛（26.8 %）、气短（16.9 %）等症状。几乎所有患者均出现肌钙蛋白的异常，并伴有 C 反应蛋白升高。非甾体类抗炎药和秋水仙碱被广泛用于临床治疗，约 1/5 患者给予对症支持后症状即得以缓解，多数患者住院时长为 1 ~ 2 周。既往感染 COVID-19 患者接种疫苗后心肌炎发生风险显著增加，且多见于接种第一针疫苗后（58.3 %），其临床症状与转归与无 COVID-19 感染史者存在一定差异。

真实: 新型冠状病毒肺炎疫苗相关心肌炎临床特征与机制探究

本模型生成: 通过分类实验探索新冠病毒疫苗导致心肌炎的临床表现和致病原因

案例 2: 新型冠状病毒肺炎（COVID-19）疫情暴发以来，由于该病毒具有极强的传染性，所导致的感染人数与死亡人数持续增加。筛查疑似患者和早期诊断 COVID-19 是防止疫情恶化的重要措施之一。通过核酸检测和人工检查等方法在感染早期诊断出 COVID-19 是防止其在社会中暴发的最佳途径。然而核酸检测效率低下，仅仅依靠放射科专家诊断 X 射线图像和 CT 扫描图像存在耗时长且易出现诊断误差等问题。研究人员相继提出了基于迁移学习的计算机辅助诊断算法，可以最大程度地减少传统诊断方法所产生的问题，但目前关于迁移学习在新冠肺炎成像中的应用综述较少，因此总结和分析了当前国内外基于迁移学习技术诊断 COVID-19 的研究成果。针对模型类型进行分类讨论，分别从数据集来源、数据预处理方法、基于迁移学习的诊断模型、模型可视化、评价指标以及模型性能 6 个角度进行分析和比较。并指出了当前所面临的挑战和未来的发展方向，为今后进一步的研究工作奠定了基础。

真实: 基于迁移学习的新冠肺炎计算机辅助诊断算法综述

本模型生成: 总结与分析应用于新冠肺炎的迁移学习技术诊断方法

5 总结

本报告中提出了一种用于 COVID-19 相关文献资料的文本摘要生成模型，其基础是经过了微调的 BERT 模型。本模型的框架主要由三个部分构成：编码器模块、解码器模块和训练模块。实现摘要生成的核心部件包括含有 BERT-Embedding 和 BiLSTM 的编码器和解码器结构以及连接不同模块的多层次神经网络。此外，模型中还针对不同序列状态使用了全局注意力和自注意力两种注意力机制。报告中使用了 ROUGE 和 BLEU 指标在开发集上对模型性能进行了评估，并将其与其他模型进行对比实验，得出本模型的摘要生成性能较强，能够胜任 COVID-19 的医学文本摘要生成任务。但是模型仍具有许多可改善的部分。在未来，可以通过加强对多源数据进行分析的能力进一步增强模型性能。

参考文献

- [1] Narayan S , Cohen SB, Lapata M (2018a) Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization[J]
- [2] Qiu Q, Xie Z, Wu L, Wenjia L (2018a) DGeoSegmenter: A dictionary-based Chinese word segmenter for the geoscience domain. Comput Geosci 121.
- [3] El-Kassas WS, Salama CR, Rafea AA et al (2020) Automatic text summarization: a comprehensive survey[J]. Expert Syst Appl 113679
- [4] Joshi M , Hui W, Mclean S (2018) Dense semantic graph and its application in single document summarisation. Emerging Ideas on Information Filtering and Retrieval
- [5] Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, et al. (2018) Deep contextualized word representations. Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long papers):2227–37
- [6] Shuai W , Xiang Z , Bo L , et al. (2017) Integrating extractive and abstractive models for long text summarization[C]// 2017 IEEE international congress on big data (BigData congress). IEEE
- [7] Mao X, Yang H, Huang S et al (2019) Extractive Summarization Using Supervised and Unsupervised Learning[J]. Expert Syst Appl 133:173–181
- [8] Wang C, Hazen R, Cheng Q, Stephenson M, Zhou C, Fox P, Shen S, Oberhänsli R, Hou Z, Ma X, Feng Z, Fan J, Ma C, Hu X, Luo B, Wang J (2021) The deep-time digital earth program: data-driven discovery in geosciences. Natl Sci Rev.
- [9] Al-Abdallah RZ, Al-Taani AT (2017) Arabic single-document text summarization using particle swarm optimization algorithm[J]. Procedia Comput Sci 117:30–37
- [10] Sutskever I , Vinyals O , Le Q V (2014) Sequence to sequence learning with neural networks[J]. NIPS
- [11] Nallapati R, Xiang B , Zhou B (2016a) Sequence-to-sequence RNNs for text summarization[J]
- [12] Hu B Chen Q, Zhu F (2015) LCSTS: A Large Scale Chinese Short Text Summarization Dataset.

- [13] Siddiqui T , Shamsi J A (2018) Generating abstractive summaries using sequence to sequence attention model[C]// Frontiers of information technology. IEEE Comput Soc
- [14] Narayan S , Cohen SB, Lapata M (2018a) Don't give me the details, just the summary! Topic-aware convolutional neural networks for extreme summarization[J]
- [15] Liang Z , Du J , Li C (2020) Abstractive social media text summarization using selective reinforced Seq2Seq attention model[J]. Neurocomputing, 410