研究与开发

基于联合损失函数和组合对比学习的语义嵌入方法

高晓欣[1]，陆谣[1]，孔祥茂[1]，刘玉玺[1]，邓伟[1]，杨淞皓[2]
（1. 北京中电普华信息技术有限公司，北京 102209；
2. 华北电力大学控制与计算机工程学院，北京 102206）

摘　要：对比学习在语义嵌入中表现出色，能够通过捕捉数据样本间的关系提升模型的表示能力。然而，其效果主要受正样本构建和目标函数选择的影响。正样本需要精心设计，以确保模型能有效识别有意义的相似性并减少噪声干扰。为此，提出一种新方法，通过拆分、编码、聚合和投射文本来构建正样本。文本被分解为片段，编码用于提取语义内容，聚合用于突出关系，最终投射到适合学习的语义空间。此外，设计了两种监督损失函数，与标准对比损失相辅相成，以增强语义空间的区分性，从而提升模型辨别能力。实验结果表明，该方法在2个公开数据集和1个私有数据集上表现出色，显著提升了语义嵌入质量，解决了对比学习的核心挑战，并为其在自然语言处理领域的进一步应用奠定了基础。

关键词：对比学习；句子嵌入；语义相似度；文本分类；联合损失函数
中图分类号：TP183
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2025142

Semantic embedding via joint loss function and composite contrastive learning

GAO Xiaoxin[1], LU Yao[1], KONG Xiangmao[1], LIU Yuxi[1], DENG Wei[1], YANG Songhao[2]
1. Beijing China-Power Information Technology Co., Ltd., Beijing 102209, China
2. School of Control and Computer Engineering, North China Electric Power University, Beijing 102206, China

Abstract: Contrastive learning has shown excellent performance in semantic embeddings by capturing relationships between data samples to enhance model representation. However, its effectiveness largely depends on constructing positive samples and selecting appropriate objective functions. Positive samples must be carefully designed to ensure the model can identify meaningful similarities while reducing noise. To address this, a novel method that constructed positive samples by splitting, encoding, aggregating, and projecting text was proposed. The text was broken into segments, encoded to extract semantic content, aggregated to highlight relationships, and projected into a semantic space optimized for learning. Additionally, two supervised loss functions were designed, complementing the standard contrastive loss, to enhance the discriminability of the semantic space and thereby improve the model's discrimination

收稿日期：2024−11−18；修回日期：2025−05−07
基金项目：国网总部科技项目（No.5700-202318834A-4-2-KJ）
Foundation Item: The Science and Technology Project of State Grid Corporation of China (No.5700-202318834A-4-2-KJ)

ability. The experimental results show that this method performes well on two public datasets and one private dataset, significantly improving the quality of semantic embedding, solving the core challenges of contrastive learning, and laying the foundation for further applications in the field of natural language processing.

Key words: contrastive learning, sentence embedding, semantic textual similarity, text classification, joint loss function

0　引言

文本分类任务需要为每个输入文档分配适当的类别标签，这是机器学习研究中的一个重要问题。文本分类在情感分析[1-2]、问答系统[3-4]和意图识别[5-6]等多个应用领域中起着关键作用。例如，在情感分析中，文本分类用于判断用户评论或社交媒体帖子的情感是正面、负面还是中性；在问答系统中，文本分类将用户提问分为不同的主题或类别，以提高问答系统的响应速度和准确性，提供更有针对性的答案；在意图识别中，文本分类用于理解用户在对话系统中的意图，从而提升智能助手和聊天机器人的交互效果，提供更智能和个性化的服务。然而，尽管深度神经网络在文本分类方面取得了显著进展，训练高性能的神经分类器仍然需要大量的人工标注数据，这在新应用场景中尤为困难且成本高昂，特别是在新兴领域或低资源语言环境下，获取足量的标注数据成为关键挑战。为应对这一问题，研究人员聚焦于自监督预训练模型在文本分类任务中的应用研究，以期降低对人工标注数据的依赖。

自监督学习中的对比学习方法，因其能够在无标签数据上有效提取数据的潜在特征，逐渐受到研究人员的关注。对比学习方法通过构建正样本对和负样本对，并通过最小化正样本对之间的距离以及最大化负样本对之间的距离来训练模型。近期的研究表明，使用通过对比学习的视觉表示简单框架（simple framework for contrastive learning of visual representation，SimCLR）[7]中引入的无监督学习框架，可以在没有标记数据的情况下学习到良好的向量嵌入。在计算机视觉领域，SimCLR 通过噪声、剪切、变换等多种数据增强技术构建正样本对，并使用 InfoNCE（information noise contrastive estimation）损失函数减少正样本嵌入向量的距离、增加负样本嵌入向量的距离。该损失函数会聚合构建的正样本，同时也会平等地远离其他样本，这就可能导致在新的语义空间中同一类型数据的距离过远。因此，本文引入了两种监督学习[8]的损失函数来避免在聚合构建正样本的同时平等地远离其他样本。联合损失函数示意图如图 1 所示。

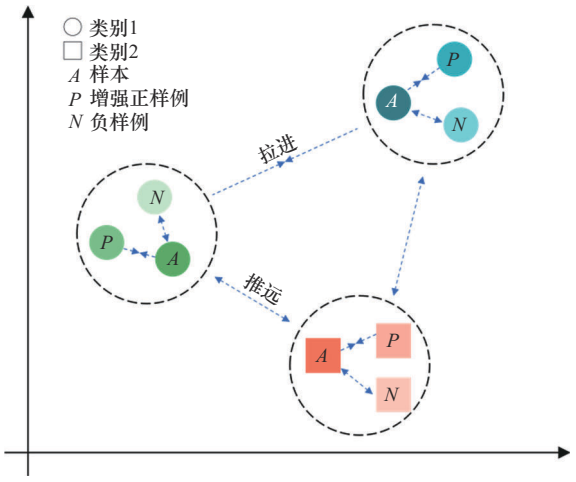


图 1　联合损失函数示意图

对比学习框架在计算机视觉中的成功应用主要归功于其多样的增强方法，这些方法能够在保留原始样本主要信息的同时，降低输入样本之间相互依赖的程度。然而，这种依赖离散增广构建正对的方法在自然语言处理（natural language processing，NLP）领域并不适用。文本数据与图

像数据有本质的不同，裁剪、同义词替换等文本增强方法往往会改变句子的语义结构，从而影响模型的学习效果。简单的对比句嵌入（simple contrastive sentence embedding，SimCSE）[9]方法针对裁剪、同义词替换等文本增强方法进行了消融实验，结果表明这些方法会损害模型在语义文本相似度（semantic textual similarity，STS）任务中的性能。相反，研究人员发现，通过设置10%的Dropout来动态生成正样本对，模型可以获得更强的表征能力。这一发现使得现有的基于对比学习的方法大多采用这种方式来构建正样本。Dropout是一种在神经网络中随机忽略一部分神经元的正则化技术，轻微的Dropout噪声可以在保持句子整体语义不变的情况下，生成略有不同的句子表示，从而构建出有效的正样本对。

尽管上述方法取得了一定的成功，但其在应对复杂多变的语言现象时，仍存在一定的局限性，即Dropout方法倾向于将长度相同或相近的句子判断为语义相似。ESimCSE[10]通过随机重复的方式改变句子长度，从而解决了Dropout正样本构建方法使模型倾向于长度相关的问题。SIFTER[11]根据下游任务的具体情况对模型和数据的信息进行缩放，通过增强句子词干和减少句子中不重要的成分来改进正样本对的构建。这些方法都是对正样本构建方法的探索，但Dropout构建正样本是否会普遍带来长度相关性，以及针对不同下游任务采用相同的信息缩放是否可行，这些都是未知的。因此，本文根据相似文本的具体情况，探索了新的数据增强方法，通过数据的拆解、编码、重构、聚合来实现正样本的构建。

总体来说，本文提出了一种用于文本分类的句子嵌入方法，即句子嵌入的联合损失函数组合对比学习（joint loss functions composition contrast learning for sentence embedding，JCCSE）方法，用于改善模型的表征能力。本文提出的方法基于语言组合的特性，通过在潜在空间中拆分组合文本，生成新的正样本对，并利用无监督和监督学习的联合损失函数来避免在学习到的语义空间中同一类型数据的距离过远。这种方法不仅简单高效，还能捕捉到更深层次的语义关系。

为了验证JCCSE方法的有效性，本文在两个公开的数据集上进行了实验。实验结果表明，JCCSE方法的表现优于现有的其他方法，显著提升了分类性能。此外，目前许多电力企业在不同专业领域的标准制定上存在分散和缺乏协调统一的问题，导致标准条款内容交叉重复或矛盾，使基层员工在开展生产工作时对条款内容理解产生歧义。因此，本文基于《电网设备技术标准差异条款统一意见》整理出了用于查找、清洗相似条款的分类数据集Contrast Clause，并在该数据集上进行实验验证。实验结果表明，JCCSE方法不仅在公开的英文文本分类数据集上表现良好，而且在本文构建的中文分类数据集Contrast Clause上也优于其他模型，证明了该模型的通用性和鲁棒性。

1 相关工作

1.1 文本分类

现有的利用深度学习进行监督分类的方法已经取得了显著的进展。除了经典的卷积神经网络（convolutional neural network，CNN）和循环神经网络（recurrent neural network，RNN）模型，文献[10]为了使模型能够给某些关键判别词赋予更高的权重，引入了注意力机制，最初用于机器翻译任务，此后，注意力机制在各种NLP任务中得到了广泛的应用和发展。文献[11]提出了分层注意力网络（hierarchical attention network，HAN），模仿句子的分层结构，同时捕捉词级和句子级的特征，进一步提高了文本分类的效果。文献[12]通过学习全局依赖关系，堆叠多个自注意力模块，以产生更稳健的句子级表示，推动了NLP模型性能的进一步提升。文献[13]结合了基于Transformer的架构和大型语料库，以最大限度地

发挥 Transformer 的能力，推出了广泛使用的 BERT（bidirectional encoder representations from transformer）模型。文献[14]将基于图的模型与 Transformer 块相结合，用以增强情感分类任务的表现，进一步拓展了 Transformer 模型在情感分析中的应用。

1.2　对比学习

对比学习[15-16]是一种度量学习方法，旨在通过拉近嵌入空间中相似输入的距离来提高模型的区分能力。目前流行且有效的对比学习方法为批对比学习（batch contrastive learning）[17-18]，可通过将相似输入（正对）和不同输入（负对）放在同一批次中进行训练，旨在同时最小化正对之间的表示距离、最大化负对之间的表示距离。对比学习中的一个关键问题是如何构建正样本。在自监督预训练[8]中，正对通常通过数据增强的方法来生成，如将旋转后的样本视为一个正样本。在监督对比学习[19]中，属于同一类的样本被视为正对。在 NLP 中，对比学习通常被用作预训练语言模型的额外自监督预训练任务，这是因为未经微调的预训练语言模型的句子嵌入不能直接用于下游任务[20]。SimCSE[8]方法通过使用 Dropout 作为最小数据增强方法，获得了最先进的无监督句子表示效果。在有监督的 SimCSE 方法中，具有蕴涵关系的句子被视为正对。数据增强的其他方法还包括句子改写[21]、反翻译[22]、双编码器[23]、语言模型损坏[24]和翻译对[25]。此外，对比学习是一种常用的神经文本检索训练算法[26]。逆完形填空任务（inverse cloze task，ICT）[27]是检索中常用的对比预训练任务，用于预测从其余文本中随机选择的句子，也可以通过文档结构来构建正对[28]。

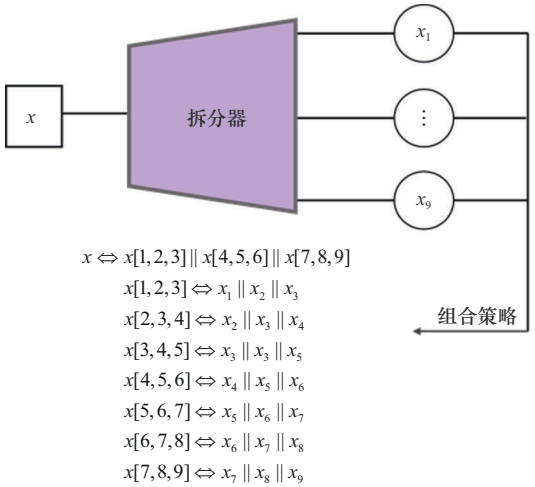
2　方法

2.1　模型结构

在一个没有标注的数据集上，将样本 x_i 在语义空间的不同组合 $x_i^+=(x_i',x_i'',x_i''')$ 视为增强的正样本。生成 x_i^+ 的一种简单有效的方法是将 x_i 按照语素分为不同语素单元 $\{x_1,\cdots,x_9\}$，这些语素单元按照如图 2 所示的数据增强方法组合形成 x_i^+。图 2 中双向箭头两侧的样本代表模型两侧的输入，双竖线“||”代表分割操作，$x[\cdot]$ 代表部分语素单元短句的组合。

来自《电网设备技术标准差异条款统一意见》构建的 Contrast Clause 数据集的一个示例可以清楚地解释本文在输入空间中分解语句的原因。该示例已被标记为具有高语义相似性。

GB/T 50832—2013《1 000 kV 系统电气装置



$\boldsymbol{x}$：GB/T 50832—2013《1 000 kV系统电气装置安装工程电气设备交接试验标准》1.0.4对于1 000 kV充油电气设备，在真空注油和热油循环后应静置不小于168 h，方可进行耐压试验
$\boldsymbol{x}_1$：GB/T 50832—2013
$\boldsymbol{x}_2$：《1 000 kV系统
$\boldsymbol{x}_3$：电气装置安装工程
$\boldsymbol{x}_4$：电气设备
$\boldsymbol{x}_5$：交接试验标准》
$\boldsymbol{x}_6$：1.0.4对于1 000 kV充油电气设备
$\boldsymbol{x}_7$：在真空注油和热油循环后
$\boldsymbol{x}_8$：应静置不小于168 h
$\boldsymbol{x}_9$：方可进行耐压试验

- $x \Leftrightarrow x' \| x'' \| x'''$
- ||：分割操作

图2　数据增强方法

安装工程电气设备交接试验标准》第1.0.4条款中：对于1 000 kV充油电气设备，在真空注油和热油循环后应静置不小于168 h，方可进行耐压试验。

GB 50835—2013《1 000 kV电力变压器、油浸电抗器、互感器施工及验收规范》第3.11.2条款中：变压器、电抗器注油完毕后，在施加电压前，静置时间不应少于120 h，且绝缘油合格。

上述两个条款的主要差异在于设备静置时间的不同。条款中有3个语义成分在起作用。

（1）条款的标准编号和标准名称。

（2）条款的主体和应用场景。

（3）条款主体的使用限制。

只有基于第（2）和第（3）个语句片段，两个文本之间才能被认为是高相似的。在主体和应用场景不同的情况下，即使条款具有同样的使用限制也不是高相似的条款，并且这与条款的标准编号和标准名称是高度不相关的。通过这些语句片段的组成（包含与舍弃），句子编码器能够学习到更精确的文本语义表示。在反向传播的迭代过程中，有效的语句片段组合会给模型带来正面反馈，无效的片段组合则不会对模型的迭代产生影响。本文使用大语言模型（large language model，LLM），并通过合适的提示词，按照语素把语句拆分成不同的单元，最后由人工审核数据集，挑选高质量拆分语句，具体的计算流程将在下面详述。

JCCSE模型结构如图3所示。本文选择BERT模型作为编码模型，利用其对输入文本进行双向编码的能力，充分捕捉文本中的上下文信息，为后续的任务提供高质量的文本表示。BERT的计算流程：首先对文本进行分词处理，同时在文本开头和结尾添加“[CLS]”和“[SEP]”标记，随后将预处理后的数据输入BERT模型中。BERT模型会对输入文本进行正向和反向的编码。在Transformer架构的每个编码层中，通过多头自注意力机制和前馈神经网络（feedforward neural network，FNN）对输入的向量进行转换，最后取输出的[CLS]张量作为文本向量编码。

JCCSE模型结构中，BERT被应用于左右两个分支，左侧分支输入锚点样本，右侧分支输入增强样本。这种设计旨在充分利用BERT的编码能

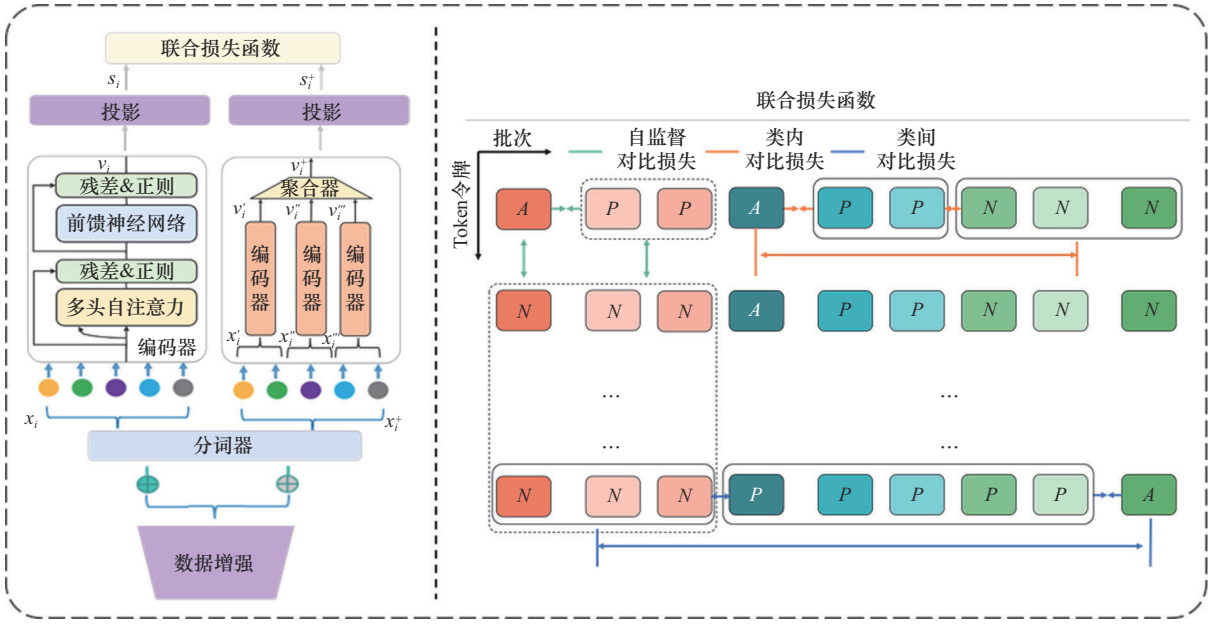


图3 JCCSE模型结构

力，为不同类型的样本生成有效的语义表示，以支持后续的处理和分析。首先，使用LLM通过提示词prompt将锚点样本按照语素分割成小短句$\{x_1,\cdots,x_9\}$，并通过图2所示的组合策略形成正样本x_i^+，如下所示：

$$(x_1,\cdots,x_9)=\mathrm{LLM}(x_i,\mathrm{prompt}) \tag{1}$$

$$x_i^+=(x_i',x_i'',x_i''')=\mathrm{Comb}(x_1,\cdots,x_9) \tag{2}$$

其中，Comb为组合策略，按照顺序每次取3个语素短句。在实际使用过程中，本文选择GPT-3.5-Turbo模型作为大语言模型拆分器，模型名称以及提示词示例见表1。

表1　模型名称以及提示词示例

模型	提示词
GPT-3.5-Turbo	请将以下文本按语素逻辑分割为恰好9个短句，遵循以下规则： （1）每个短句保留完整语义单位； （2）优先在连词/标点处拆分； （3）保持原文本顺序不变； （4）若原文不足9单元，则重复最后语素补足； （5）若超出9单元，则合并最小语义单元； （6）输出格式：数字序号+短句（每行一条）。 待处理文本：[输入文本]

接着，将构建的正样本以及锚点样本分别送到模型的左右两个分支，利用BERT的[CLS]张量表达锚点样本的语义信息v_i以及正样本的语义信息v_i',v_i'',v_i'''：

$$v_i=\mathrm{BERT}(x_i) \tag{3}$$

$$v_i',v_i'',v_i'''=\mathrm{BERT}(x_i'),\mathrm{BERT}(x_i''),\mathrm{BERT}(x_i''') \tag{4}$$

为了平衡样本和增强样本的token维度，也为了融合v_i',v_i'',v_i'''的特征，本文使用最大池化方法聚合增强样本的[CLS]张量，并采用多层感知器（multilayer perceptron，MLP）作为投影层微调[CLS]张量：

$$v_i^+=\mathrm{MaxPooling}(v_i',v_i'',v_i''') \tag{5}$$

$$s_i,s_i^+=\mathrm{MLP}(v_i),\mathrm{MLP}(v_i^+) \tag{6}$$

2.2　BERT编码器

本文所提方法使用BERT作为编码器，BERT是由谷歌团队于2018年提出的预训练语言模型，其核心基于Transformer架构，旨在通过对大规模文本的无监督学习，获取通用的语言表示。与传统的单向语言模型不同，BERT采用双向Transformer编码器，能够同时考虑文本的前文和后文信息，从而更全面地捕捉语言的语义和句法特征。BERT编码器结构如图4所示。

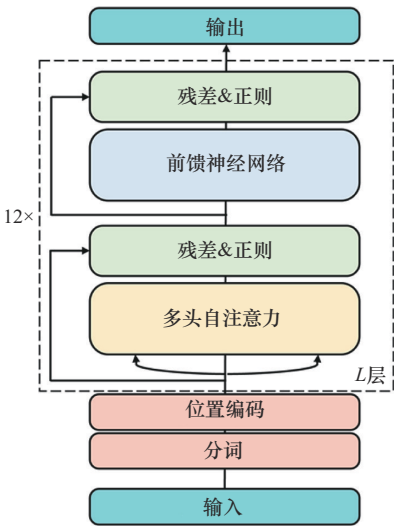


图4　BERT编码器结构

由于Transformer架构本身不具备对序列中元素位置信息的感知能力，因此引入了位置嵌入（position embedding）。位置嵌入的作用是为每个输入的词向量赋予一个表示其在序列中位置的向量，从而使模型能够捕捉到词汇的顺序信息。位置嵌入向量的生成方式是通过固定的正弦和余弦函数来计算的，其表达式为：

$$\mathbf{PE}_{(\mathrm{pos},2i)}=\sin(\mathrm{pos}/10000^{2i/d_{\mathrm{model}}}) \tag{7}$$

$$\mathbf{PE}_{(\mathrm{pos},2i+1)}=\cos(\mathrm{pos}/10000^{2i/d_{\mathrm{model}}}) \tag{8}$$

其中，pos表示词汇在序列中的位置，i表示嵌入向量中的维度索引，d_{model}表示整个模型的维度。通过这样的方式，不同位置的词汇会获得不同的位置嵌入向量，并且位置之间的相对距离也能通过嵌入向量反映出来。在实际应用中，位置嵌入向量与词嵌入向量直接相加，即$\boldsymbol{E}=\boldsymbol{E}_{\mathrm{token}}+$

$\boldsymbol{E}_{\text{position}}$，作为后续多头自注意力机制等模块的输入。其中，$\boldsymbol{E}$表示最终输入模型的向量，$\boldsymbol{E}_{\text{token}}$表示词嵌入向量，$\boldsymbol{E}_{\text{position}}$表示位置嵌入向量。BERT的核心是多头自注意力机制，其表达式为：

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

其中，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$表示输入的向量，d_k表示键向量的维度。多头自注意力机制则是将多个注意力头并行计算，然后拼接结果，表达式为：

$$\text{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{Concat}(\text{head}_1,\cdots,\text{head}_h)\boldsymbol{W}^{O} \quad (10)$$

$$\text{head}_i=\text{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}},\boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}},\boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}) \quad (11)$$

其中，h表示注意力头的数量，$\boldsymbol{W}^{O}$和$\boldsymbol{W}_i^{\boldsymbol{Q}}$、$\boldsymbol{W}_i^{\boldsymbol{K}}$、$\boldsymbol{W}_i^{\boldsymbol{V}}$表示可学习的权重矩阵。将多头自注意力的计算结果送入FFN，表达式为：

$$\text{FFN}(\boldsymbol{x})=\text{ReLU}(\boldsymbol{x}\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2 \quad (12)$$

其中，$\boldsymbol{x}$为输入向量，$\boldsymbol{W}_1$、$\boldsymbol{W}_2$为权重矩阵，$\boldsymbol{b}_1$、$\boldsymbol{b}_2$为偏置向量。本文通过FFN，对经过多头自注意力机制处理后的特征进行进一步的变换和特征提取。

2.3 损失函数

为了提供一个更清晰的训练目标并最大化利用已知信息，本文结合自监督训练目标、监督训练目标提出联合对比损失函数。损失函数包含自监督对比损失函数L_{self}、类间对比损失函数L_{inter}以及类内对比损失函数L_{intra}。自监督对比损失函数将同一样本生成的样本看作正样本，其余的样本和增强样本看作负样本。该函数将最小化与正样本之间的距离，同时最大化与负样本之间的距离作为目标，表达式为：

$$L_{\text{self}}^{(i)}=-\frac{1}{|A^*(i)|}\sum_{a\in A^*(i)}\ln\frac{\exp(\boldsymbol{z}_i\cdot\boldsymbol{z}_a)}{\sum_{j\in\bar{A}(i)}\exp(\boldsymbol{z}_i\cdot\boldsymbol{z}_j)} \quad (13)$$

其中，$A^*(i)$为样本i的增强样本，$\bar{A}(i)$为除样本i及其增强样本外的其他样本，包括增强样本，$\boldsymbol{z}$为文本嵌入向量。这种损失函数虽然可以最小化$\boldsymbol{z}_i$与$\boldsymbol{z}_a$之间的距离，但它均匀地排斥了其他所有的样本，并没有完全利用组内和组间的信息，因此，本文使用另外两种损失函数，以充分利用分类数据集的类别信息。类内对比损失函数的表达式为：

$$L_{\text{intra}}^{(i)}=-\frac{1}{|A^*(i)|}\sum_{a\in A^*(i)}\ln\frac{\exp(\boldsymbol{z}_i\cdot\boldsymbol{z}_a)}{\sum_{j\in B^*(i)}\sum_{k\in A(j)}\exp(\boldsymbol{z}_i\cdot\boldsymbol{z}_k)} \quad (14)$$

其中，$B^*(i)$为与样本i同类的样本但不包括样本i，$A(j)$为样本j及其增强样本。这种损失函数将样本自身的增强样本作为正样本，将同类中的其他样本及其增强样本作为负样本。此外，本文还设计了类间对比损失函数，其将同一类的样本视为正样本，将不同类的视为负样本。类间对比损失函数的表达式为：

$$L_{\text{inter}}^{(i)}=-\frac{1}{|P^*(i)|}\sum_{p\in P^*(i)}\ln\frac{\exp(\boldsymbol{z}_i\cdot\boldsymbol{z}_p)}{\sum_{j\in\bar{P}(i)}\exp(\boldsymbol{z}_i\cdot\boldsymbol{z}_j)} \quad (15)$$

其中，$P^*(i)$为与样本i同类的样本和增强样本，但不包括样本i，$\bar{P}(i)$表示与样本i不同类的样本和增强样本。

这3种损失函数可以互相弥补彼此的缺点，本文利用超参数α、β组合这3种损失函数：

$$L=-\frac{1}{N}\sum_{i=1}^{N}(L_{\text{self}}^{(i)}+\alpha L_{\text{inter}}^{(i)}+\beta L_{\text{intra}}^{(i)}) \quad (16)$$

3 实验与分析

3.1 实验数据集

为了评估本文所提方法在语义相似度计算中的有效性，本文使用了两个公开的英文数据集STS2015和STS2016。STS2015和STS2016是语义文本相似度评测数据集，主要用于评估两个文本在语义上的相似程度，其数据来源广泛且覆盖多种场景，为语义相关研究提供了有力支撑。此外，本文还运用了数据集Contrast Clause开展实验，该数据集是一个中文分类数据集，由电力行业标准条款中的相似条款组成，其内容源于《电

网设备技术标准差异条款统一意见》并经人工整理，涵盖了 6 000 对电力标准条款，Contrast Clause 数据集示例见表 2。在电力行业蓬勃发展的当下，多数电力企业在各专业领域的标准是独立制定的，缺乏统一的协调与规范，这使得条款出现交叉、重复和矛盾的情况，给基层的工作执行带来了极大困扰。为解决这一问题，研究人员构建了 Contrast Clause 数据集，以便训练模型进行数据清洗。此外，本文还在另一个中文数据集 THUCNews 上进行了对比试验。数据集 THUCNews 是中文文本分类数据集，由清华大学自然语言处理实验室整理，涵盖体育、娱乐、家居等 10 个类别，约 5 万个样本。

3.2　评价指标

针对语义相似度和文本分类这两种任务选择对应的评价指标。斯皮尔曼等级相关系数（Spearman’s rank correlation coefficient）用于语义相似度任务，是一种非参数统计方法，用于测量两个变量之间的单调关系。它们不假设变量之间的线性关系，而是基于变量的排序（排名）来计算其相关性，计算式为：

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{17}$$

其中，n 为观测值的数量，d_i^2 为第 i 对观测值的排名差。由于数据集 Contrast Clause 的正负样本不平衡，因此本文选择 F1 分数作为数据集 Contrast Clause 的评价指标。F1 分数是机器学习和信息检索领域常用的评价指标之一，尤其适用于不平衡分类问题。F1 分数是精确率（precision）和召回率（recall）的调和平均数，综合考虑了模型在正确识别正类样本方面的能力，其计算式为：

$$\mathrm{F1} = 2 \times \frac{P \times R}{P + R} \tag{18}$$

其中，P 为精确率，R 为召回率。P 和 R 的计算式为：

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{19}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{20}$$

其中，TP（true positive）表示正确预测为正类的样本数量，FP（false positive）表示错误预测为正类的样本数量，衡量了模型预测的正类样本中实际为正类的比例，FN（false negative）表示实际为正类但被错误预测为负类的样本数量，衡量了实际正类样本中被模型正确预测为正类的比例。在使用 F1 分数进行评价时，精确率和召回率的平衡非常重要。高精确率意味着模型在预测正类样本时有很高的准确性，但可能会忽略一些实际为正类的样本（低召回率）。相反，高召回率意味着模型能捕捉到大多数正类样本，但可能会引入更多错误的正类预测（低精确率）。F1 分数通过综合考虑这两者，提供了一种在不平衡数据集上更为合理的评价方式。

3.3　参数设置

本文使用预训练的 BERT 作为基础模型，利用 SimCSE 的代码框架搭建 JCCSE 模型。所有实验数据集均采用 7:1:2 比例划分为训练集、验证集

表 2　Contrast Clause 数据集示例

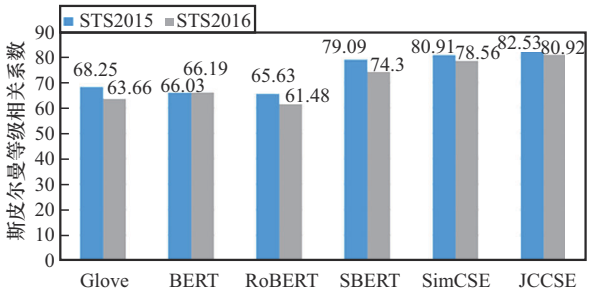
条款 1	条款 2
变压器在规定的工作条件和负载条件下运行，预期寿命应不低于 40 年	变压器在规定的工作条件和负荷条件下运行，并按照制造厂的使用说明书进行维护后，变压器的预期寿命应不低于 30 年
绝缘油中颗粒含量：过滤以后的油中直径大于 5 μm 的颗粒数不多于 1 000 个/100 mL	油中颗粒含量：过滤以后的油中大于 5 μm 的颗粒不多于 2 000 个/100 mL
分接开关长期载流的触头，在 1.2 倍额定电流下，对变压器油的稳定温升不超过 20 K	应在 1.2 倍最大额定通过电流下进行温升试验，对变压器油的稳定温升不超过 20 K
烃类气体总和的产气速率大于 0.5 mL/h、相对产气速率大于 10%/月，则认为设备有异常	烃类气体总和的产气速率大于 6 mL/d 或相对产气速率大于 10%/月，则认为设备有异常

和测试集。实验都是在训练集上迭代，并选择验证集上最优的模型在测试集上训练。使用 Adam 优化器训练模型参数，初始学习率为 1×10^{-5}，迭代次数为 100，批处理大小为 64。超参数 α、β 均选择 0.5。该模型使用 PyTorch 实现，并在 NVIDIA GeForce RTX 4060 8 GB GPU 上训练。为确保模型的稳定性，本文所有实验结果均为测试 10 次后取平均。

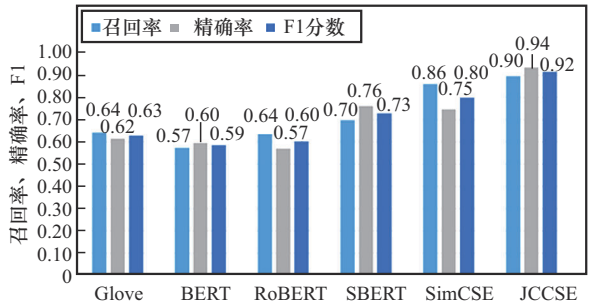
3.4 实验结果

3.4.1 算法对比

为了验证本文所提方法的有效性，实验选取了 GloVe、BERT、RoBERT、SBERT 这 4 个基线模型与 SimCSE 进行比较。JCCSE 和基线模型的性能对比如图 5 所示。



（a）STS数据集实验对比



（b）Contrast Clause数据集实验对比

图 5 JCCSE 与基线模型的性能对比

对于 STS2015、STS2016 这两个数据集，使用斯皮尔曼等级相关系数作为评价指标。对于 Contrast Clause、THUCNews 数据集，使用召回率、精确率、F1 分数作为评价指标。图 5（a）中的柱状图分别代表基线模型和 JCCSE 模型在 STS2015 和 STS2016 两个数据集上的斯皮尔曼等级相关系数指标，JCCSE 模型在 STS2015 和 STS2016 数据集上的表现优于其他对比模型，分别达到 82.53 和 80.92。图 5（b）中，JCCSE 模型在 Contrast Clause 数据集中的得分也最高，分别为 90%、94%、0.92。最后，本文还在额外的中文数据集 THUCNews 上验证了模型的通用性，THUCNews 数据集实验对比结果见表 3。由表 3 可知，本文提出的 JCCSE 模型在 3 个指标上分别为 96%、94%、0.95，相较其他基线模型，本文提出的模型在召回率和 F1 分数上均显示最优。

表 3 THUCNews 数据集实验对比结果

模型	召回率	精确率	F1 分数
GloVe	88%	86%	0.87
BERT	90%	92%	0.91
RoBERT	91%	93%	0.92
SBERT	95%	92%	0.94
SimCSE	93%	**95%**	0.94
JCCSE	**96%**	94%	**0.95**

以上实验结果充分表明，JCCSE 模型在语义相似度和文本分类任务中展现出卓越的性能，其出色表现主要归因于基于语句片段组合的正样本构建方法以及联合损失函数的运用。具体而言，JCCSE 模型采用语句片段组合的思想来构建对比学习所需的正样本。该方法通过在潜在空间中对文本片段进行拆分和重新组合，能够生成更加丰富多样的正样本，有助于模型捕捉更为细致且深层次的语义关系，从而显著提升其表征能力。此外，在训练过程中，JCCSE 模型联合了自监督和有监督的学习目标。这一联合学习目标的优势在于，它可以充分挖掘未标注数据中的潜在信息，同时借助标注数据所提供的监督信息，二者相辅相成，进而进一步优化了模型的性能。

3.4.2 消融实验

为了更好地分析 JCCSE 模型的性能，本文在 3 个数据集上进行了消融实验，JCCSE 模型消融

实验结果见表4。本文分别进行了不使用自监督学习目标JCCSE$_{w/o\ self}$、不使用类间对比损失函数JCCSE$_{w/o\ inter}$、不使用类内对比损失函数JCCSE$_{w/o\ intra}$、使用10%的Dropout来构建正样本而不是使用语句片段组合JCCSE$_{w/o\ comp}$、仅使用自监督学习目标JCCSE$_{w/\ self}$的实验。由表4可知，不使用联合损失函数的实验JCCSE$_{w/\ self}$性能有所下降，这是因为自监督损失函数均匀地排斥了所有样本，在相似性任务中不能很好地区分哪些是相似的文本，哪些是不相似的文本。由实验JCCSE$_{w/o\ self}$、JCCSE$_{w/o\ inter}$以及JCCSE$_{w/o\ intra}$可以看出，联合损失函数中的每一个学习目标都是不可或缺的。由JCCSE$_{w/o\ comp}$实验可以看出，JCCSE模型的正样本构建方法是优于随机Dropout方法的，这是因为在潜在空间的语句片段能更好地选择重要的语义信息。

表4　JCCSE模型消融实验结果

实验组	数据集		
	STS2015	STS2016	Contrast Clause
JCCSE$_{w/o\ self}$	78.86	77.28	0.899 7
JCCSE$_{w/o\ inter}$	74.08	72.55	0.845 8
JCCSE$_{w/o\ intra}$	81.19	79.59	0.926 1
JCCSE$_{w/o\ comp}$	80.54	78.58	0.865 0
JCCSE$_{w/\ self}$	72.16	70.65	0.824 2
JCCSE	**82.53**	**80.92**	**0.941 2**

4　结束语

本文结合对比学习方法提出了一个用于语义相似度和文本分类任务的模型JCCSE。该模型可以提取更细致、更深层次的语义关系，并且考虑了有监督学习目标和自监督学习目标的优点，从而提高了模型在语义相似度和文本分类任务上的整体性能。本文首先设计了正样本构建方法，将文本均分成3段，通过编码器、聚合器以及投影层得到正样本表示，然后通过联合损失函数学习目标学习语义表示，以便从潜在空间中获得语句不同组合的向量表示，并拉近真正赋予语句意义的片段的向量。JCCSE模型在公开的数据集STS2015、STS2016、THUCNews以及一个由《电网设备技术标准差异条款统一意见》整理出的中文数据集Contrast Clause上均表现最佳。本文提出的方法可以为自然语言处理领域提供新的思路。

参考文献：

[1] 欧阳梦妮, 樊小超, 帕力旦•吐尔逊. 基于目标对齐和语义过滤的多模态情感分析[J]. 计算机技术与发展, 2024, 34(6): 171-177.
OUYANG M N, FAN X C, PALIDAN T S. Multimodal sentiment analysis based on target alignment and semantic filtering [J]. Computer Technology and Development, 2024, 34(6): 171-177.

[2] 吴加辉, 加云岗, 王志晓, 等. 基于深度学习的微博疫情舆情文本情感分析[J]. 计算机技术与发展, 2024, 34(7): 175-183.
WU J H, JIA Y G, WANG Z X, et al. Sentiment analysis of weibo epidemic public opinion text based on deep learning[J]. Computer Technology and Development, 2024, 34(7): 175-183.

[3] 陈鹏,郎彬,石英,等. 基于知识图谱的电力设备缺陷问答系统研究[J]. 广西师范大学学报(自然科学版), 2024, 42(16): 149-163.
CHEN P, TAI B, SHI Y, et al. Research on defect Q&A system of power equipment based on knowledge graph [J]. Journal of Guangxi Normal University (Natural Science Edition), 2024, 42(16): 149-163.

[4] 贾鹏. 基于大语言模型的农业知识问答系统的研究与设计[D]. 秦皇岛: 河北科技师范学院, 2024.
JIA P. Research and design of agricultural knowledge question answering system based on macrolanguage model[D]. Qinhuangdao: Hebei Normal University of Science & Technology, 2024.

[5] 李东亚, 白涛, 香慧敏, 等. 基于ERNIE及改进DPCNN的棉花病虫害问句意图识别[J]. 山东农业科学, 2024, 25(16): 143-151.
LI D Y, BAI T, XIANG H M, et al. Intention recognition of cotton pest questions based on ERNIE and improved DPCNN[J]. Shandong Agricultural Sciences, 2024, 25(16): 143-151.

[6] 谌文佳, 杨琳, 李金林. 嵌入意图识别的医疗健康问答文本语义分类模型[J]. 数据分析与知识发现, 2025, 9(2): 26-38.
CHEN W J, YANG L, LI J L. Semantic classification model for healthcare Q & A texts with embedded intent recognition[J]. Data Analysis and Knowledge Discovery, 2025, 9(2): 26-38.

[7] CHEN T, KORNBLITH S, NOROUZI M, et al. A simple framework for contrastive learning of visual representations[J]. arXiv preprint, 2020: 2002.05709.

[8] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning[C]//Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020). [S.l.: s.n.], 2020: 18661-18673.

[9] GAO T Y, YAO X C, CHEN D Q. SimCSE: simple contrastive learning of sentence embeddings[J]. arXiv preprint, 2021: 2104.08821.

[10] WU X, GAO C C, ZANG L J, et al. ESimCSE: enhanced sample building method for contrastive learning of unsupervised sentence embedding[J]. arXiv preprint, 2021: 2109.04380.

[11] BAHDANAU D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint, 2014:1409.0473.

[12] YANG Z C, YANG D Y, DYER C, et al. Hierarchical attention networks for document classification[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (USAACL). [S.l.:s.n.], 2016: 1480-1489.

[13] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. arXiv preprint, 2017: 1706.03762.

[14] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint, 2018: 1810.04805.

[15] TANG H, JI D H, LI C L, et al. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (USAACL). [S. l.: s. n.], 2020: 6578-6588.

[16] HADSELL R, CHOPRA S, LECUN Y. Dimensionality reduction by learning an invariant mapping[C]//Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06). Piscataway: IEEE Press, 2006: 1735-1742.

[17] CHOPRA S, HADSELL R, LECUN Y. Learning a similarity metric discriminatively, with application to face verification[C]// Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Piscataway: IEEE Press, 2005: 539-546.

[18] HE K M, FAN H Q, WU Y X, et al. Momentum contrast for unsupervised visual representation learning[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 9729-9738.

[19] GUNEL B, DU J, CONNEAU A, et al. Supervised contrastive learning for pre-trained language model fine-tuning[J]. arXiv preprint arXiv, 2020: 2011.01403.

[20] KHOSLA P, TETERWAK P, WANG C, et al. Supervised contrastive learning[J]. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673.

[21] LI B H, ZHOU H, HE J X, et al. On the sentence embeddings from pre-trained language models[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). [S.l.:s.n.], 2020: 9119-9130.

[22] WU Z F, WANG S N, GU J T, et al. CLEAR: contrastive learning for sentence representation[J]. arXiv preprint, 2020: 2012.15466.

[23] FANG H C, WANG S C, ZHOU M, et al. CERT: contrastive self-supervised learning for language understanding[J]. arXiv preprint, 2020: 2005.12766.

[24] CARLSSON F, GOGOULOU E, YLIPÄÄ E, et al. Semantic retuning with contrastive tension[C]//Proceedings of the 2021 International Conference on Leaening Representations(ICLR). [S.l.:s.n.], 2021: 1-21.

[25] MENG Y, XIONG C Y, BAJAJ P, et al. COCO-LM: correcting and contrasting text sequences for language model pretraining[J]. arXiv preprint, 2021: 2102.08473.

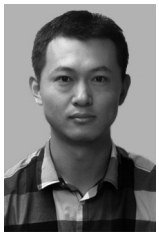
[26] WANG Y S, WU A, NEUBIG G. English contrastive learning can learn universal cross-lingual sentence embeddings[J]. arXiv preprint, 2022: 2211.06127.

[27] XIONG L E, XIONG C Y, LI Y, et al. Approximate nearest neighbor negative contrastive learning for dense text retrieval[J]. arXiv preprint, 2020: 2007.00808.

[28] LEE K, CHANG M W, TOUTANOVA K. Latent retrieval for weakly supervised open domain question answering[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (USAACL). [S.l.:s.n.], 2019: 6086-6096.

[作者简介]

高晓欣（1983– ），女，北京中电普华信息技术有限公司工程师，主要研究方向为科研体系、标准数字化。



刘玉玺（1984– ），男，博士，北京中电普华信息技术有限公司正高级工程师，主要研究方向为数据模型、人工智能、数据分析、标准数字化。

陆谣（1997– ），女，北京中电普华信息技术有限公司助理工程师，主要研究方向为标准数字化。

邓伟（1976– ），男，北京中电普华信息技术有限公司正高级工程师，主要研究方向为工业互联网、信息通信运营、标准数字化。

孔祥茂（1993– ），男，北京中电普华信息技术有限公司工程师，主要研究方向为标准数字化、电力数据分析。

杨淞皓（2001– ），男，华北电力大学控制与计算机工程学院硕士生，主要研究方向为电力标准条款差异性分析。