论文分数	
指导教师	

中南财经政法大学

本科生学年论文

论文题目	：	基于对比学习的文章类型判别
姓　　名	：	石锴文
学　　号	：	201821130123
班　　级	：	信管 1801 班
专　　业	：	信息管理与信息系统
学　　院	：	信息与安全工程学院
指导教师	：	刘勘
完成时间	：	2021 年 8 月 29 日

摘　要

文章类型判别是信息流领域的核心问题，提升文章类型判别的准确率是提升信息流质量和精准推送的核心技术点。同一类型的文章当中可能涉及不同的话题，相比以往的文章主题判别，文章类型判别变得更加复杂。高效而准确的文章类型判别模型能够为内容质量运营平台降本增效。

本文从对比学习的角度出发，按照“减小正样本之间的距离，增大负样本之间的距离”的思路，讨论了对比学习在文本分类当中的原理及应用。本文使用 BERT 模型为编码器，构建了基于对比学习的文章类型判别模型。首先对训练数据进行采样，获取正样本与负样本；之后使用对比学习损失函数对模型进行预训练；最后把分类标签引入微调训练当中，用监督学习的方法在下游分类数据集上进行训练。实验结果表明，引入对比学习预训练的模型在下游任务中性能更好，并且随着对比学习预训练的轮数增加，样本表示向量之间的距离会随之改变。

关键词：文本分类；对比学习；预训练

目　录

引　言

（一）背景与意义

在各大以文章内容运营为主要业务的媒体平台中，优质文章的精准推送是保障用户流量的关键所在。在媒体平台中，系统需要向用户推送与用户兴趣相关的文章。为了保证推送的多样性，推荐系统要避免重复推送同一话题下的内容。借助文章类型标签，除了把握用户所感兴趣的话题以外，还能了解用户对文章类型的偏好。

当用户量与日俱增，用户对文章推送质量的要求越来越高，运营成本越来越大，高效准确的文章类型判别模型就显得非常重要。由于同一类型的文章当中可能涉及不同的话题，相比以往的文章主题判别，文章类型判别对于人类来说会更加复杂。本文把关注点聚焦在“如何生成有效的文本表示”这一问题上，希望通过对比表示学习来解决这一复杂问题。

（二）国内外现状

国外最早的对比学习研究起源于 Becker 和 Hinton 于 1992 年和 Bromley 等人在 1993 年的研究。Becker 和 Hinton[1]提出通过最大化样本之间的互信息来学习数据中抽象稳定的特征，Bromley 等人[2]将孪生网络应用于度量学习。这两项研究首次提出通过样本的对比来学习数据中的信息。

近年，Chen 等人[3]在图像表示的工作中使用无监督对比学习的方法，逼近了监督学习模型的性能，由此引发了对比学习的新浪潮。Oord 等人[4]提出 infoNCE 损失函数，有效改善了模型处理序列数据的性能。Gao 等人[5]在句子嵌入的研究中，利用 dropout 构造无监督样本中的正类样本，在多项任务中显著提升了模型的性能。经过一段时间的探索，研究人员开始关注对比学习改善模型性能的原理。Wang 等人[6]通过分析对比损失函数引出了同一性和容忍性的概念，指明了对比损失函数对难分的负样本具有自适应的敏感性。Wang 等人的工作指对比损失函数旨在将样本分布到超球面上，同时拉近正样本间的距离，拉开负样本间的距离，从而使得线性超平面能够对这些数据进行分隔。

国内有关于对比学习在自然语言处理领域的应用研究较少，大多数的研究关注在计算机视觉问题上。郭东恩等人[7]将对比学习应用于遥感图像场景分类中，通过引入有监督的对比学习损失，将不同场景样本的距离拉大，再使用预训练的 Inception V3 网络对样本表示进行分类训练，有效提升了模型的判别能力。孙浩等人[8]在图像分类问题中引入对比学习方法，通过改变不同标签样本之间的

距离，提升深度网络模型对不同样本的健壮性。杨少杰[9]将对比学习应用在安卓恶意文件检测问题中，通过提取文件的基础特征，利用对比损失函数改善样本的表示向量。

（三）本文组织结构

本文后续安排如下：第一节介绍对比学习和文本分类的主要理论和技术，第二节介绍文章类型判别模型的基本思路和构建过程，第三节围绕第二节构建的模型展开实验和分析。

一、理论与技术简介

（一）文本分类

文本分类技术通过挖掘文本中的隐含特征，来对文本的某一性质进行判别，往往会运用到监督学习的各种方法。传统的机器学习方法注重于挖掘文本的统计特征，使用诸如支持向量机、朴素贝叶斯、逻辑斯蒂回归等方法对特征进行分类判别。随着词嵌入技术（Word Embedding）的发展，现代深度学习方法通常使用循环神经网络（Recurrent Neural Network, RNN）或者卷积神经网络（Convolution Neural Network, CNN）自动提取文本的上下文语义特征，通过非线性变换对特征进行判别。近年的大规模语料预训练模型则重新定义了自然语言处理领域的学习范式：通过自监督学习的方式，以及预先设计好的预训练任务伪标签对模型进行训练，使得模型能够学习到自然语言当中复杂的语义知识；之后通过微调的方式，在下游任务当中引入相关任务的标签，对预训练好的模型进行训练，使得带有先验知识的预训练模型能够适应下游任务。

在各大预训练语言模型中，BERT（Bi-directional Encoder Representation from Transformer, BERT）的使用频率非常高。BERT 模型包含词嵌入、两个位置信息编码以及多个 Transformer 编码器。其预训练包含两个任务：输入随机遮罩的句子，预测遮罩位置的词；输入两个短句子，判断两者是否为同一句话。经过预训练之后，可以利用 BERT 中“[CLS]”token 的输出向量执行下游的分类任务。

（二）对比学习

表示学习是指将原始组织结构的数据映射到一个向量空间的学习过程，这个向量空间能够有效捕获原始数据中所包含的抽象信息，从而能够改善下游任务的表现。Le-Khac 等人[10]归纳了优良表示的三个特性：分布式（Distributed）、

抽象不变性（Abstraction and Invariant）、松耦合表示（Disentangled Representation）。分布式的特性要求表示向量能够从原始的、高维的稀疏空间中捕获到有效信息，并在低维的稠密空间中进行表达。抽象不变性要求表示向量能够发掘样本中稳定的抽象信息，即对某些类型的样本共有的特征作出归纳，当这些样本发生轻微改变的时候，表示向量依然能够保持稳定性。松耦合表示要求表示向量在其向量空间中较为分散，以保证学习到的表示能够在下游任务中具有通用的泛化性能，同样也为表示向量的可解释性提供保障。

对比学习则是一类能够满足上述优良表征学习的方法。不同于端到端的训练模式，通过对原始数据进行挖掘，构造数据之间的比较样本对，使得模型能够有效地学习到优良的表征。对比学习是在监督学习和自监督学习任务中派生出来的：监督学习中，对比学习利用现成标签构造样本对；在自监督学习中，利用基于相似度的伪标签构造样本对。对两者进行抽象，如图 1 所示，可以发现对比学习的整体思路。

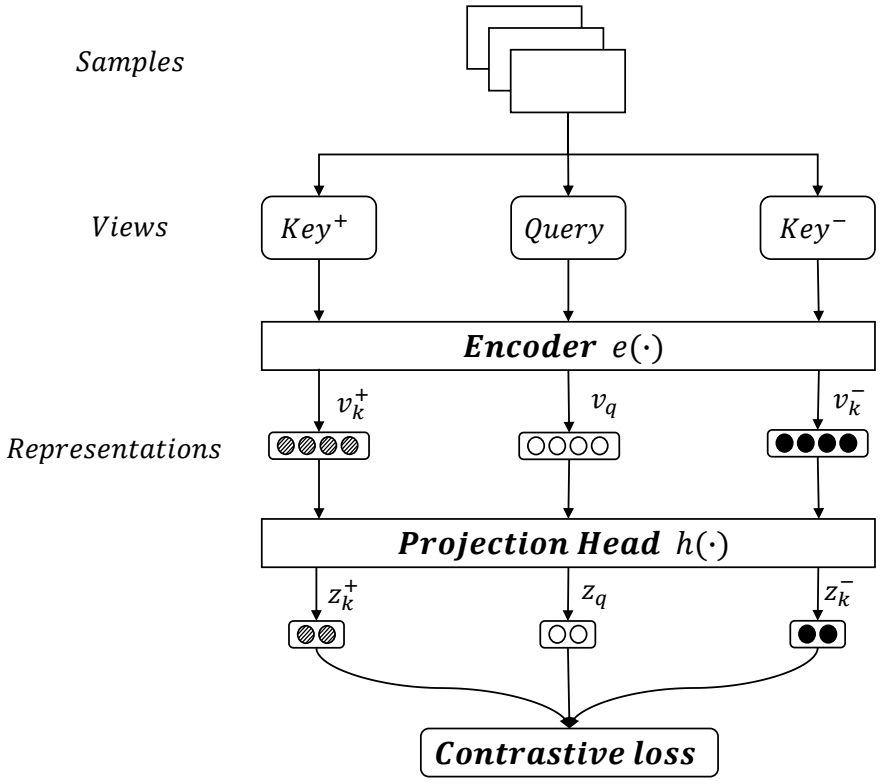

图 1 对比学习的抽象框架

在进行对比学习时，首先对原始数据进行采样，$Query$为随机采样，Key^+为正样本，Key^-为负样本。在监督学习中，Key^+采样于$Query$的同类样本，Key^-采样于$Query$的不同类样本；在自监督学习中，Key^+采样于$Query$的增强样本，Key^+采样于$Query$以外的样本。采样之后，使用编码器对样本进行编码，得到样

本的嵌入表示。之后对样本表示进行非线性映射，再计算对比损失。

直观上，在损失函数中，相似度高的正样本能够减小损失，相似度高的负样本会增大损失，故对比损失函数可以简单地表示如下。

$$\mathcal{L}_{simple} = -sim\left(z_q, \mathrm{z}_k^+\right) + \lambda sim(z_q, \mathrm{z}_k^-)$$

其中$sim\left(z_q, \mathrm{z}_k^+\right)$和$sim\left(z_q, \mathrm{z}_k^-\right)$分别表示正样本和负样本与$Query$编码后的相似度，$\lambda$为正常数。

优良的对比损失函数应该能够有效地拉近正样本之间的距离，拉大负样本之间的距离。常用的对比损失函数如表 1 所示。

表 1 常用的对比损失函数

名称	损失函数
NT-Xent	$sim(z, z_k^+)/\tau - \log \sum_{z_k \in \{z_k^+, z_k^-\}} e^{sim(z, z_k)/\tau}$
NT-Logistic	$\log \sigma(sim(z, z_k^+)/\tau) + \log \sigma(sim(z, z_k^-)/\tau)$
Margin Triple	$-max(sim(z, z_k^+) - sim(z, z_k^-) + m, 0)$
infoNCE	$-log \frac{e^{sim(x_i, x_i')/\tau}}{e^{sim(x_i, x_i')/\tau} + \sum_{k \neq i} e^{sim(x_i, x_k)/\tau}}$

本文使用实践中效果较好的 infoNCE 完成后续探究。

（三） infoNCE 原理

对训练样本$X = \{x_1, x_2, \ldots, x_N\}$，有 infoNCE 对比损失函数如下。

$$\mathcal{L}(x_i) = -log \left[\frac{e^{sim(x_i, x_i')/\tau}}{e^{sim(x_i, x_i')/\tau} + \sum_{k \neq i} e^{sim(x_i, x_k)/\tau}} \right]$$

其中$sim(\cdot)$表示样本相似度，x_i'是从x_i中采样的增强数据，τ表示温度系数（temperature）。可以发现，样本x_i与x_j之间的相似概率可以如下表示。

$$P_{i,j} = \frac{e^{sim(x_i,x_j)/\tau}}{e^{sim(x_i,x_j)/\tau} + \sum_{k\neq i} e^{sim(x_i,x_k)/\tau}}$$

分别对正样本和负样本进行分析，对$sim(x_i, x_i')$和$sim(x_i, x_k)$求偏导。

$$\frac{\partial \mathcal{L}(x_i)}{\partial sim(x_i, x_i')} = -\frac{1}{\tau}\sum_{k\neq i} P_{i,j}$$

$$\frac{\partial \mathcal{L}(x_i)}{\partial sim(x_i, x_k)} = \frac{1}{\tau} P_{i,j}$$

从负样本所产生的梯度来看：当温度系数τ较小时，相似度大的负样本的$P_{i,j}$分子会更大，那么负样本所带来的梯度会较大，从而使得相似负样本之间的表示向量更加分离。

对$\mathcal{L}(x_i)$的极限状态进行分析。

$$\lim_{\tau\to 0^+} \mathcal{L}(x_i) = \lim_{\tau\to 0^+} -log\left[\frac{e^{sim(x_i,x_i')/\tau}}{e^{sim(x_i,x_i')/\tau} + \sum_{k\neq i} e^{sim(x_i,x_k)/\tau}}\right]$$

$$= \lim_{\tau\to 0^+} log\left[1 + \sum_{k\neq i} e^{\frac{sim(x_i,x_k)-sim(x_i,x_i')}{\tau}}\right]$$

$$= \lim_{\tau\to 0^+} log\left[1 + \sum_{sim(x_i,x_k)\geq sim(x_i,x_i)} e^{\frac{sim(x_i,x_k)-sim(x_i,x_i')}{\tau}}\right]$$

当温度系数τ趋近于0^+时，损失函数只对相似的负样本敏感，相似度较低的负样本所带来的损失值在极限状态下为 0。

$$\lim_{\tau\to +\infty} \mathcal{L}(x_i) = \lim_{\tau\to +\infty} -log\left[\frac{e^{sim(x_i,x_i')/\tau}}{e^{sim(x_i,x_i')/\tau} + \sum_{k\neq i} e^{sim(x_i,x_k)/\tau}}\right]$$

$$= \lim_{\tau\to +\infty} -\frac{1}{\tau} sim(x_i, x_i') + log\sum_k e^{sim(x_i,x_k)/\tau}$$

$$= \lim_{\tau\to +\infty} -\frac{1}{\tau} sim(x_i, x_i') + log\left[\frac{1}{N}\left(\sum_k e^{\frac{sim(x_i,x_k)}{\tau}} - 1\right) + 1\right] + logN$$

$$= \lim_{\tau\to +\infty} -\frac{1}{\tau} sim(x_i, x_i') + \frac{1}{N}\left(\sum_k e^{\frac{sim(x_i,x_k)}{\tau}} - 1\right) + logN$$

$$= \lim_{\tau\to +\infty} -\frac{1}{\tau} sim(x_i, x_i') + \frac{1}{N\tau}\sum_k sim(x_i, x_k) + logN = logN$$

当温度系数τ趋近于$+\infty$时，损失函数为常数，即不同相似程度的样本产生的损失是一样的。也就是说，温度系数较大时，相似度较高的负样本很难被分离。

综合上述分析，温度系数τ在对比损失函数中可以表达对难学习样本（相似度高的负样本）的敏感性。较小的温度系数对相似度高的负样本更加严格，而正样本也会受到其影响变得更加分散。这在样本的语义表示空间中能够起到一定作用：即使样本间差异非常小，但他们的语义向量依然能够将两者分离。较大的温度系数在学习困难样本时性能可能不足，但是对相似度高的样本更有容忍度。在判别式的下游任务中，判别模型可能更容易理解不同类别的样本，但由于相似样本之间非常紧凑，判别模型也可能会因此失去一部分泛化性能。

（四） infoNCE 的实现

上一小节中介绍过两个样本之间的相似概率$P_{i,j}$，infoNCE 实质上是在计算每一对样本的相似概率的负对数。那么这一过程可以通过构造标签，用交叉熵函数来实现计算。

$$CrossEntropy(p,q) = -\sum_{x} p(x) \cdot \log q(x)$$

infoNCE 计算过程如图 2 所示。

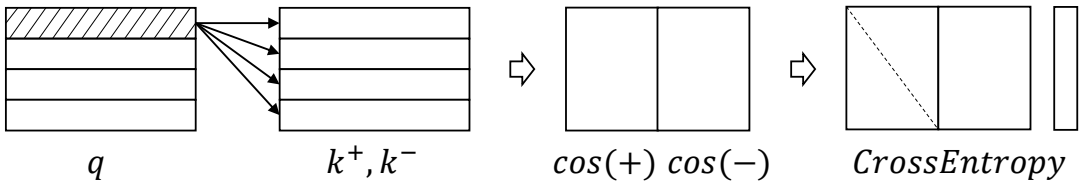

图 2 infoNCE 实现过程

首先计算q与 k^+和k^-之间的相似度矩阵，可以得到两个$N \times N$的方阵，N为样本量；随后对两个相似度矩阵进行横向拼接，得到$N \times 2N$的矩阵；构造临时标签，得到$N \times 1$的向量，该向量表示每一个样本对为 1 个类别；最后通过交叉熵函数，计算拼接后相似度矩阵的$softmax$并取负对数，左半部的方阵对角线元素之和即为当前所有样本的 infoNCE 损失。

二、基于对比学习的文章类型判别

（一）基本思路

文章类型包含“深度事件”“行业解读”等，共 10 个类型。可以发现，这些质量标签并不等同于文章内容的分类标签，例如在“深度事件”类型的文章中可能出现金融领域的事件分析，也可能出现法律领域的事件分析。而掌握大量语义知识的预训练模型较难对此类文章的标签进行分类，而这样的样本正适合用对比的方法进行学习。

本文则使用基于监督学习的对比学习方法，构建文章类型判别模型。构建样本对时，使用相同标签的样本作为正样本，不同标签的样本作为负样本，采样时按照训练集中标签的比例分布进行概率抽样。

（二）模型构建

首先使用对比学习进行文本表示模型的预训练。利用 BERT 作为文本编码器，使用“[CLS]”的 token 输出向量作为文本表示，在非线性变换之后计算样本之间的对比损失。采样阶段中，为了防止出现标签泄露，仅对下游任务的训练数据进行采样。随后在带标签的数据集上，利用表示模型的输出加上新的非线性层进行微调训练。模型结构如图 3 所示。

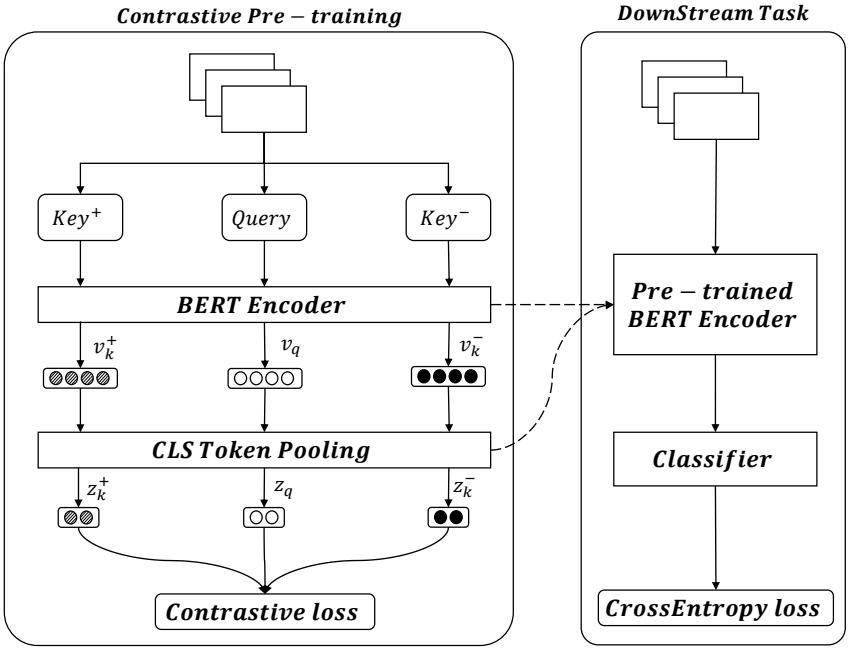

图 3 基于对比学习的文章类型判别模型结构

三、实验与结果分析

（一）探索性数据分析

本文使用的数据集来自于华为 2021 年 DIGIX 算法大赛的赛题二：基于多模型迁移预训练文章质量判别。数据集包含 76454 条标注数据。

数据集包含 10 种文章类型标签：“人物专栏”“情感解读”“科普知识文”“攻略文”“物品评测”“治愈系文章”“推荐文”“深度事件”“作品分析”“行业解读”，标签的分布情况如图 4 左所示。

本文使用文章的标题部分进行分类，如图 4 右所示，标题的长度集中分布在 30 左右，最大长度为 61。

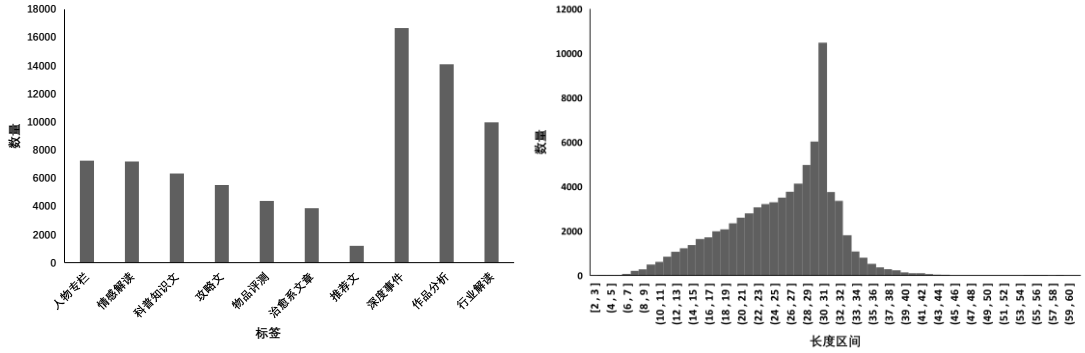

图 4 标签分布情况（左），文本长度分布（右）

（二）实验过程

本文设置两组对照实验：直接使用 BERT 模型与经过对比学习的 BERT 模型的对照、不同轮数对比学习 BERT 模型之间的对照。

数据预处理：首先使用在文本的首尾分别加上“[CLS]”和“[SEP]”，之后 BertTokenizer 对文本进行分词，并获取对应词项的 token。因为文本最大长度为 61，小于 BERT 模型的输入限制长度，故可以直接将文本的 token 列表输入到模型中。

实验在 Unbuntu18.04 操作系统上完成，使用一张 NVIDIA GeForce RTX 3090 显存为 24GB 的显卡训练模型。实验中模型的主要参数参考了 SimCSE[5]的相关实验设置，如表 2 所示。

表 2 实验参数设置

参数	参数值	备注
Epoch	1	预训练与微调时的参数保持一致
Batch size	64	
Dropout	0.5	
Learning rate	5e-5	
Weight decay	0.1	
Sampling times	140000	对比学习采样次数
Temperature	0.07	温度系数

（三）结果分析

训练集与测试集的划分比例为 0.85：0.15，对比学习的采样在训练集上完成。实验结果如表 3 所示。

表 3 实验结果

Method	Model	Precision	Recall	F1
Original BERT	BERT	0.72	0.73	0.72
Contrastive BERT	ConBERT(1)	**0.78**	0.76	**0.76**
	ConBERT(3)	0.76	**0.77**	**0.76**
	ConBERT(5)	0.77	0.75	0.75

从原始 BERT 与对比学习的 BERT 来看，经过对比学习预训练的模型在原有的基础性能上提升明显。由于微调阶段的训练轮数是一致的，说明模型在对比学习的预训练阶段中学习到了样本之间的关系，从而在微调时取得了更好的泛化性能。做出微调过程中模型的 loss 曲线，如图 5 所示。可以发现原始 BERT 模型的 loss 曲线收敛更慢，且收敛程度不如经过对比学习预训练的模型。由不同预训练轮数的实验结果，可以发现训练 1 轮和 3 轮的模型性能相近，而区别在于 1 轮模型的准确率更高，3 轮模型的召回率更高。说明在预训练过程中，模型的召回能力增强，间接地表明样本表示间的距离随训练轮数的增加而增大。

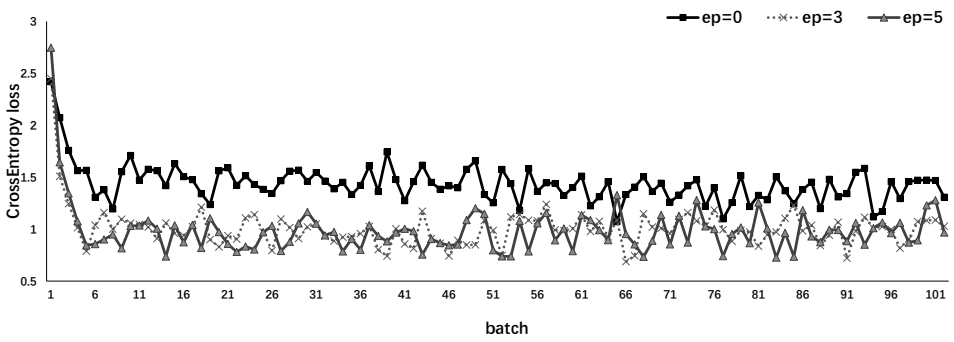

图 5 不同预训练轮数的 loss 曲线

结　语

本文主要讨论了对比学习在文本分类当中的原理和应用，并构建了基于对比学习的文章类型判别模型。本文设置的实验初步说明了对比学习对于文本表示学习的有效性，而实验中还存在一些未调整的超参数。在后续的研究当中，可以继续探究对比损失函数中温度系数对样本表示的影响，以及对比学习预训练轮数对下游任务的影响。

除了模型参数的探究，未来还可以深入探究对比表示学习的通用性。以文章类型判别为例，是否可以通过对比学习构造样本的表示，从而能够将文本表示同时应用在文章类型判别和文章主题判别任务中。也就是说，不同类型的文章间的距离会拉开，同一类型中不同话题的文章距离会拉开。再引申一步，是否存在一种通用的样本表示，能够同时适应更多的下游任务？如果问题得到解决，对比表示学习的可解释性会进一步增强。

主要参考文献

[1] Becker S, Hinton G E. Self-organizing neural network that discovers surfaces in random-dot stereograms[J]. Nature, 1992, 355(6356):161.

[2] Bromley J, Guyon I, LeCun Y, et al. Signature Verification using a “Siamese” Time Delay Neural Network[C]. //International Journal of Pattern Recognition and Artificial Intelligence. World Scientific Publishing Company, 1993:669-669.

[3] Chen T, Kornblith S, Norouzi M, et al. A Simple Framework for Contrastive Learning of Visual Representations[J]. arXiv preprint arXiv:2002.05709, 2020.

[4] Oord A, Li Y, Vinyals O. Representation Learning with Contrastive Predictive Coding[J]. arXiv preprint arXiv:1807.03748, 2018.

[5] Gao T, Yao X, Chen D. SimCSE: Simple Contrastive Learning of Sentence Embeddings[J]. arXiv preprint arXiv:2104.08821, 2021.

[6] Wang F, H Liu. Understanding the Behaviour of Contrastive Loss[J]. arXiv preprint arXiv:2012.09740, 2020.

[7] 郭东恩,夏英,罗小波,丰江帆.基于有监督对比学习的遥感图像场景分类[J].光子学报,2021,50(07):87-98.

[8] 孙浩,徐延杰,陈进,雷琳,计科峰,匡纲要.基于自监督对比学习的深度神经网络对抗鲁棒性提升[J].信号处理,2021,37(06):903-911.

[9] 杨少杰. 基于特征向量与对比学习的安卓恶意文件检测[D].兰州大学,2021.

[10] Le-Khac P H, Healy G, Smeaton A F. Contrastive Representation Learning: A Framework and Review[J]. IEEE Access, 2020, 8:193907-193934.