SimCSE: Simple Contrastive Learning of Sentence Embeddings

Tianyu Gao, Xingcheng Yao, Danqi Chen

摘要

本次论文改进所参考的文章是 SimCSE[1]，一个简单的对比学习框架，其极大地提高了最先进的句子 embedding 方式。文章作者首先描述了一种无监督的方法，仅使用标准 dropout 作为噪声。这种简单方法的表现与以前的监督方法相当。此次复现在标准语义文本相似性（STS）任务上对模型进行了评估，使用 BERT 预训练模型后的无监督模型实现了百分之 74.7 的 Spearman 相关性。可以看出，对比学习将预训练 embedding 的各向异性空间正则化为更均匀，从而为解决 bert 各向异性问题提供了一个全新的思路。本次论文改进主要针对文中的无监督学习的数据增强部分进行了一些改进，通过对文本添加 noise 的方式对文中单一的 dropout 方式进行了丰富，并尝试将其与 dropout 的方法相结合，使原文模型在中文数据集中的无监督学习的情况下得到一些微小的提升。

关键词：对比学习；数据增强

1 引言

本次改进的论文是 SimCSE，一个用于 sentence embedding 的对比学习框架，作者描述了一种无监督的方法，即只使用标准的 dropout 作为噪声进行数据增强，并将其方法用于解决 bert 预训练的各向异性问题。本次复现，我主要针对文中的无监督学习的数据增强部分进行了一些改进，主要是通过对文本添加 noise 的方式对文中数据增强的方式进行丰富，并将其与 dropout 的方法相结合。

2 相关工作

2.1 对比学习

对比学习旨在通过将拉进语义接近的内容，拉远语义不相近的内容来学习有效的表示。它假设有一组成对的示例 $\mathcal{D}=\left\{\left(x_i, x_i^{+}\right)\right\}_{i=1}^{m}$, 其中 x_i 和 x_i^{+} 是语义相关的。本文遵循的对比框架，是采用交叉熵目标函数：令 $\mathbf{h}_i$ 和 $\mathbf{h}_i^{+}$ 作为 x_i 和 x_i^{+}, 的表示。则在以 N 对为一组的训练过程中，对于 $\left(x_i, x_i^{+}\right)$ 的训练目标如下所示:

$$\ell_i=-\log \frac{e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{h}_i^{+}\right) / \tau}}{\sum_{j=1}^{N} e^{\operatorname{sim}\left(\mathbf{h}_i, \mathbf{h}_j^{+}\right) / \tau}},$$

其中 τ 是温度超参数，$\operatorname{sim}\left(\mathbf{h}_1, \mathbf{h}_2\right)$ 是余弦相似度 $\frac{\mathbf{h}_1^{\top} \mathbf{h}_2}{\left\|\mathbf{h}_1\right\| \cdot\left\|\mathbf{h}_2\right\|}$。在本次实验中，使用了预训练的语言模型 bert 对输入句子进行编码 $\mathbf{h}=f_\theta(x)$，然后使用对比学习目标微调所有参数。

而在对比学习中最重要的两个属性：即 Alignment 和 uniformity，本次实验使用它们来衡量表征的质量。给定正对 p_{pos} 的分布，Alignment 计算成对实例 embedding 之间的期望距离：

$$\ell_{\text{align}} \triangleq \underset{(x, x^{+}) \sim p_{\text{pos}}}{\mathbb{E}}\left\|f(x)-f\left(x^{+}\right)\right\|^2 .$$

另一方面，用 uniformity 度量 embedding 的均匀分布程度：

$$\ell_{\text{uniform}} \triangleq \log \underset{x,y \overset{i.i.d.}{\sim} p_{\text{data}}}{\mathbb{E}} e^{-2\|f(x)-f(y)\|^2},$$

其中 p_{data} 代表数据分布。这两个指标与对比学习的目标是一致的，即正例应该保持接近，而实例的 embedding 应该均匀分散在超球面上。

2.2 基于 dropout 的对比学习

SimCSE 将 dropout 方式视为数据增强的最小形式：即正对采用完全相同的句子，它们的嵌入只在 dropout 掩模中有所不同。为了进一步理解 dropout 噪声在无监督 SimCSE 中的作用，作者在图一中尝试了不同的 dropout 率，分别是“没有 dropout”($p = 0$) 和“固定 dropout=0.1”，以及一个简单的数据增强模型“删除一个词”。如图所示，所有的模型都大大提高了均匀性。“删除一个词”改进了对齐，但在均匀性度量上获得了较小的增益，最终的性能低于无监督的 SimCSE。

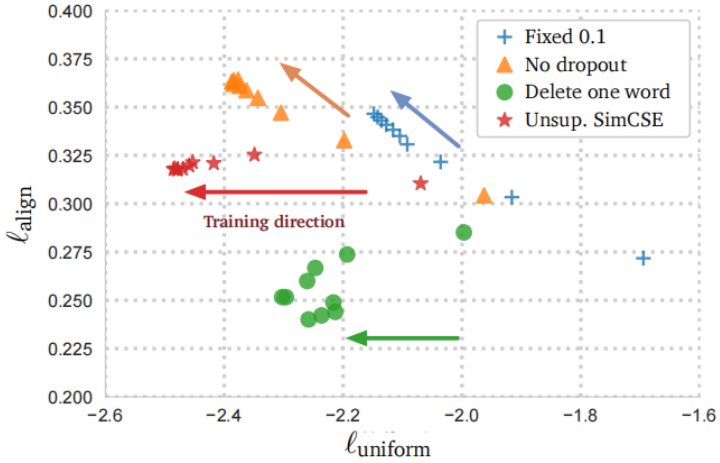

图 1: 无监督 SimCSE

上图是关于无监督 SimCSE 的绘图，其中有“无 dropout”、“dropout0.1”和“删除一个单词”三种方式进行数据增强。每 10 个 epoch 可视化检查点，箭头指示训练方向。对于 ℓ_{align} and ℓ_{uniform}，而言，越小的数值越好。

3 本文方法

3.1 本文方法概述

本次实验复现的无监督 SimCSE，利用 dropout 作为数据增强方式，从而改进 bert 模型的各向异性问题。通过改变 dropout mask 生成正样本的方法可看成数据增强的最小形式，因为原样本和生成的正样本的语义一致，只是生成的 embedding 不同而已。换句话说，将相同的句子传递给预先训练过的编码器两次：通过应用标准 dropout 两次，我们可以得到两个不同的嵌入作为“正对”，然后我们将同一小批中的其他句子作为“负例”。主要方法如图 2 所示：

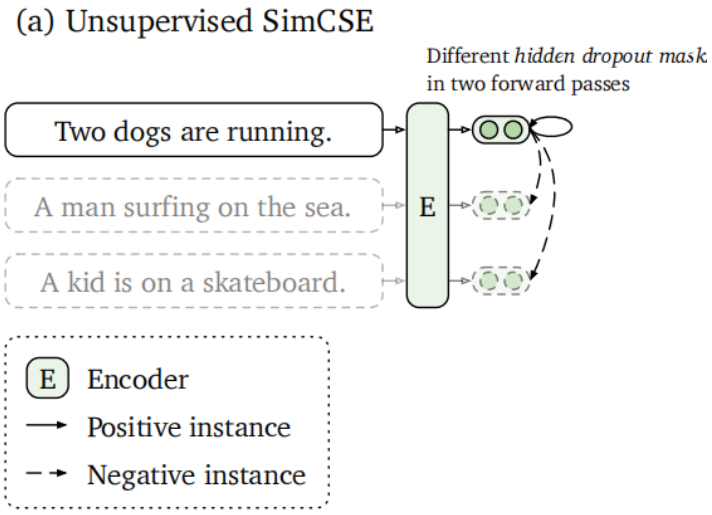

图 2: 方法示意图

3.2 目标函数定义

本文进行无监督学习的概念非常简单：取一个句子的集合 $\{x_i\}_{i=1}^m$，并使 $x_i^+ = x_i$。关键是通过对 x_i 和 x_i^+ 使用独立采样的 dropout。在 transformer 的标准训练中层上设置了 dropout（默认为 $p = 0.1$）。表示为 $\mathbf{h}_i^z = f_\theta(x_i, z)$，其中 z 是一个随机的 dropout 掩模。只需将相同的输入输入给编码器两次，就可以得到两个具有不同的 dropout 掩模的 embeddingz, z'，本文训练的目标函数为：

$$\ell_i = -\log \frac{e^{\text{sim}\left(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z_i'}\right)/\tau}}{\sum_{j=1}^N e^{\text{sim}\left(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z_j'}\right)/\tau}},$$

4 复现细节

4.1 与已有开源代码对比

本次论文的参考代码如下网址可见：

https://github.com/princeton-nlp/SimCSE；

https://github.com/shuxinyin/SimCSE-Pytorch

本次实验对文本添加了 noise 也就是离散的数据增强（如单词删除和替换）的方式进行处理，并将其与单一的 dropout 方法相结合。此次相较原文代码新增核心代码如下所示，另因使用中文数据集，故而代码细节改变较多，不逐一列举。

```
if args.un_supervise:
    train_data_source = load_sts_data_unsup(train_path_unsp)
    sentence = [data[0] for data in train_data_source]
    if args.addreverse:
            train_sents = random_swap_word(sentence, 0.1)
    else:
        if args.addremove:
            train_sents = random_delete_word(sentence, 0.1)
        else:
            train_sents = [data[0] for data in train_data_source]
    train_dataset = TrainDataset(train_sents, tokenizer, max_len=args.max_length)
else:
    train_data_source = load_data_sup(train_path_sp)
    # train_sents = [data[0] for data in train_data_source] + [data[1] for data in train_data_source]
    train_dataset = TrainDataset_sup(train_data_source)
```

图 3: 本次实验新增核心代码

```
def random_swap_word(sentence, prob):
    if random.random() > prob:
        return sentence
    else:
        words = list(jieba.cut(sentence))
        if len(words) == 1:
            return sentence
        index1 = random.randint(0, len(words)-1)
        index2 = random.randint(0, len(words)-1)
        while index2 == index1:
            index2 = random.randint(0, len(words)-1)
        words[index1], words[index2] = words[index2], words[index1]
        sentence = "".join(words)
    return sentence

def random_delete_word(sentence, prob):
    if random.random() > prob:
        return sentence
    else:
        words = list(jieba.cut(sentence))
        delete_index = random.randint(0, len(words)-1)
        del words[delete_index]
        sentence = "".join(words)
    return sentence
```

图 4: 本次实验新增核心代码

4.2 实验环境搭建

本次实验所需环境如下：

$torch = 1.8.2$

$transformers = 4.12.3$

$tqdm$

$scikit-learn$

$scipy >= 1.5.4, < 1.6$

$numpy >= 1.19.5, < 1.20$

setuptools

4.3 下游任务展示

本次实验利用语义标准语义文本相似性（STS）任务进行了评估，即输入一个句子，对其进行分析，并选出语义相似的 topk 个句子，下游任务展示图如图 5 所示（并未完善）。

图 5: 下游任务展示图

4.4 创新点

本次实验复现的无监督 SimCSE，与原文中只使用 dropout 作为数据增强方式相比，对文本添加了 noise 也就是离散的数据增强（如单词删除和替换）的方式进行处理，并将其与单一的 dropout 方法相结合。通过改变 dropout mask 生成正样本的方法可看成数据增强的最小形式，因为原样本和生成的正样本的语义一致，生成的 embedding 不同而已。换句话说，将相同的句子传递给预先训练过的编码器两次：通过应用标准 dropout 两次，我们可以得到两个不同的嵌入作为“正对”。然后我们将同一小批中的其他句子作为“负例”，并且对于句子采用随机概率进行单词的删除以及替换。

5 实验结果分析

本次实验是在中文数据集当中所使用的基于 BERT-base 的无监督模型。图 5 所示为在同等环境下，增加 noise 与不增加 noise 的结果比较，在同等环境下增加 noise 的基于 BERT-base 的无监督模型在验证集以及测试集中分别获得了百分之 74.9 以及百分之 69.4 的 Spearman 相关性。而不增加 noise 的模型的 Spearman 相关性为百分之 73.7 以及百分之 68.1。图 6 为原文结果数据展示。

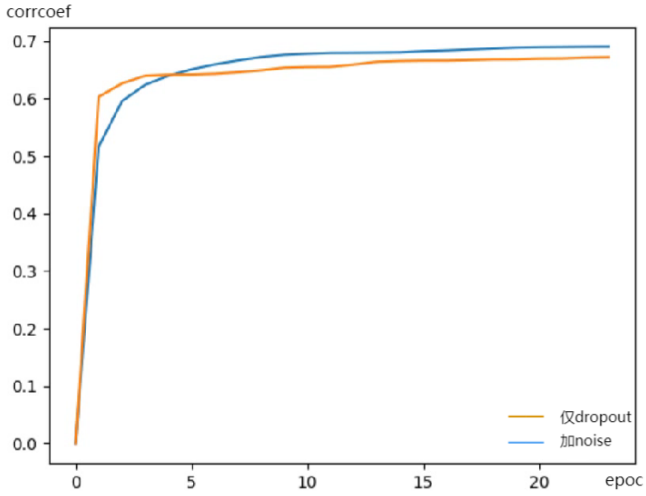

图 6: 实验结果示意图

Model	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
Unsupervised models								
GloVe embeddings (avg.)♣	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
$\text{BERT}_{\text{base}}$ (first-last avg.)	39.70	59.38	49.67	66.03	66.19	53.87	62.06	56.70
$\text{BERT}_{\text{base}}$-flow	58.40	67.10	60.85	75.16	71.22	68.66	64.47	66.55
$\text{BERT}_{\text{base}}$-whitening	57.83	66.90	60.90	75.08	71.31	68.24	63.73	66.28
IS-$\text{BERT}_{\text{base}}$♡	56.77	69.24	61.21	75.23	70.16	69.21	64.25	66.58
CT-$\text{BERT}_{\text{base}}$	61.63	76.80	68.47	77.50	76.48	74.31	69.19	72.05
∗ SimCSE-$\text{BERT}_{\text{base}}$	**68.40**	**82.41**	**74.38**	**80.91**	**78.56**	**76.85**	**72.23**	**76.25**
$\text{RoBERTa}_{\text{base}}$ (first-last avg.)	40.88	58.74	49.07	65.63	61.48	58.55	61.63	56.57
$\text{RoBERTa}_{\text{base}}$-whitening	46.99	63.24	57.23	71.36	68.99	61.36	62.91	61.73
DeCLUTR-$\text{RoBERTa}_{\text{base}}$	52.41	75.19	65.52	77.12	78.63	72.41	**68.62**	69.99
∗ SimCSE-$\text{RoBERTa}_{\text{base}}$	**70.16**	**81.77**	**73.24**	**81.36**	**80.65**	**80.22**	68.56	**76.57**
∗ SimCSE-$\text{RoBERTa}_{\text{large}}$	**72.86**	**83.99**	**75.62**	**84.77**	**81.80**	**81.98**	**71.26**	**78.90**

图 7: 原文实验结果

6 总结与展望

由于预训练好的 BERT 模型并不能完美适应所有下游任务，BERT 的输出结果如果不进行微调，这些向量会坍缩在一个小区域内，从而损害 BERT 效果。本次实验的 SimCSE 采用了对比学习来解决这个问题，利用 dropout 作为噪声对比学习解决了 BERT 各向异性问题。SimCSE 这篇文章中利用对比学习以及加入噪声解决各向异性问题的这些思想具有很强的启发性。而本次实验进行的改进仅仅是增加了一些文本的噪声，改进并不明显，未来仍有待加强。

参考文献

[1] GAO T, YAO X, CHEN D. SimCSE: Simple Contrastive Learning of Sentence Embeddings[J]. arXiv e-prints, 2021, arXiv:2104.08821: arXiv:2104.08821. arXiv: 2104.08821 [cs.CL]. DOI: 10.48550/arXiv.2104.08821.