SimCSE

Simple Contrastive Learning of Sentence Embeddings，其实就是把对比学习引入了SBERT，达到了**句子相似度**SOTA。SBERT本身并不复杂，仅仅是一个基于BERT的孪生网络而已，想要在SBERT的基础上改进指标，只能从**训练目标**下手。

1. 对比学习概念

对比学习的思想很简单，即拉近相似的样本，推开不相似的样本，一种常用的对比损失是基于mini-batch采样负样本的交叉熵损失，假设我们有一个数据集$\mathcal{D}=\left\{\left(x_i, x_i^{+}\right)\right\}_{i=1}^m$，其中$x_i$和$x_i^{+}$是语义相关的，则在大小为$N$的mini batch内，$\left(x_i, x_i^{+}\right)$的训练目标为

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

其中$\operatorname{sim}\left(\mathbf{h}_1, \mathbf{h}_2\right)=\frac{\mathbf{h}_1^{\top} \mathbf{h}_2}{\left\|\mathbf{h}_1\right\| \cdot\left\|\mathbf{h}_2\right\|}$，$\mathbf{h}_i$和$\mathbf{h}_i^{+}$是$x_i$和$x_i^{+}$的编码表示，$\tau$为softmax的温度超参。

分子是真正的正样本，分母是正样本+所有负样本，这个其实就是个交叉熵损失。

1.1 怎么构造正样本

使用对比损失最关键的问题是如何构造$\left(x_i, x_i^{+}\right)$，对比学习最早起源于CV领域的原因之一就是图像的$x_i^{+}$非常容易构造，**裁剪、翻转、扭曲和旋转**都不影响人类对图像语义的理解，因此可以直接作为正样本。而结构高度**离散**的自然语言则很难构造语义一致的x_i^{+}，前人采用了一些数据增强方法来构造x_i^{+}，比如**替换、删除、重排**，但这些方法都是离散的操作，很难把控，容易引入负面噪声，模型也很难通过对比学习的方式从这样的样本中捕捉到语义信息，性能提升有限。

1.2 句子embedding的好坏评判标准：Alignment & uniformity

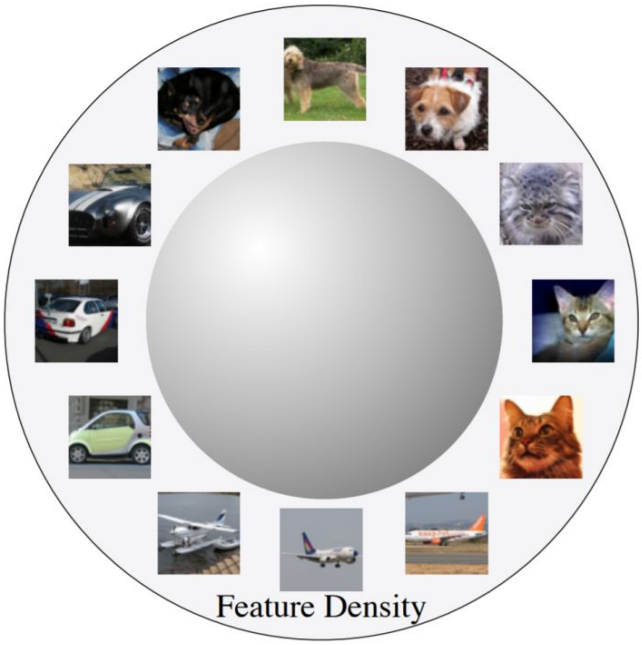
- alignment计算x_i和x_i^{+}的平均距离：

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

自然是希望正样本和正样本的距离越近越好。

- uniformity计算向量整体分布的**均匀**程度：

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



Uniformity: Preserve maximal information.

我们希望语义向量要尽可能地**均匀分布在超球面上**，因为均匀分布**信息熵最高**，分布越均匀则保留的信息越多。

“拉近正样本，推开负样本”实际上就是在优化这两个指标。

2. SimCSE

2.1 无监督的SimCSE

本文作者提出可以通过dropout 来生成正样本x_i^+，回想一下，在标准的Transformer中，dropout mask被放置在**全连接层和注意力**操作上。由于dropout mask是**随机**生成的，所以在**训练阶段**，将同一个样本分两次输入到同一个编码器中，我们会得到两个不同的表示向量z, z'，将z'作为正样本，则模型的训练目标为

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

这种通过改变dropout mask生成正样本的方法可以看作是**数据增强**的最小形式，因为原样本和生成的正样本的语义是完全一致的，只是**生成的embedding不同而已**。所以，其实SimCSE生成正样本的方式就是把样本过两次预训练好的BERT，用dropout来获得两个不一样的embedding作为正例对；负样本做mini-batch采样....对，就这。

2.2 有监督的SimCSE

在SBERT原文中，作者将NLI数据集作为一个**三分类**任务来训练(entailment, neutral, contradiction)，这种方式忽略了正样本与负样本之间的交互，而**对比损失**则可以让模型学习到更丰富的细粒度语义信息。

构造训练目标其实很简单，直接将数据集中的正负样本拿过来用就可以了，将NLI数据集中的entailment作为正样本，contradiction作为负样本，加上原样本premise一起组合为

$$\left(x_i, x_i^+, x_i^-\right)$$

，并将损失函数改进为

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