SimCSE

2021年6月10日　22:15

摘要

提出一种简单的对比学习框架在显著提升了sentence embedding task的SOTA

首先提出了一个无监督方法，使用dropout构建正例，batch内负采样，进行对比学习，效果和之前的监督方法打平。

然后使用NLI数据集中的有标注的数据，构建有监督方法，在sentence embedding task上达到SOTA

最后作者表明，对比学习理论上使预训练嵌入的各向异性空间更加均匀，并且在有监督信号的情况下能更好地对齐。

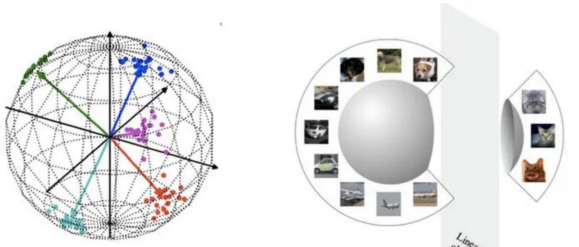
对比学习

Henderson et al., 2017): let $\mathbf{h}_i$ and $\mathbf{h}_i^+$ denote the representations of x_i and x_i^+, for a mini-batch with N pairs, the training objective for (x_i, x_i^+) is:

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

where τ is a temperature hyperparameter and $\mathrm{sim}(\mathbf{h}_1, \mathbf{h}_2)$ is the cosine similarity $\frac{\mathbf{h}_1^\top \mathbf{h}_2}{\|\mathbf{h}_1\| \cdot \|\mathbf{h}_2\|}$. In

为什么在计算相似度时我们需要对句子向量做L2正则？张俊林：对比学习研究进展精要给出了一个非常形象的解释，这样做的目的是将所有的句子向量映射在一个半径为1的超球体上，一方面我们将所有向量统一至单位长度，去除了长度信息是为了让模型的训练更加稳定；另一方面如果模型的表示能力足够好，能够把相似的句子在超球面上聚集到较近区域，那么很容易使用线性分类器把某类和其它类区分开（参考下图）。当然在图像领域上很多实验也证明了，增加L2正则确实能提升模型效果。



单位超球面及线性可分

为什么$sim(h_i, h_i^+)/r$中 温度超参 r 的作用：笔者调研到了两种解释：1. 如果直接使用sim相似度计算得到的值作为softmax的logist输入，由于其取值范围仅在[-1,1]之间，$loss(logist = [-1, 1, -1], y = [0, 1, 0]) != 0$这显然是不合理的，且logist范围太小导致softmax对正负样本无法给出足够大的差距，模型训练不充分，所以我们需要对相似度值进行修正，除以一个足够小的温度参数进行放大。2. 温度参数会将模型更新的重点，聚焦到有难度的负例，并对它们做相应的惩罚，难度越大，也即是与h_i距离越近，则分配到的惩罚越多。这其实也比较好理解，我们将$sim(h_i, h_j^+)/r$相当于同比例放大了负样本的logist值，如果r足够小，那么那些sim(h_i,h_j^+)越靠近1的负样本，其放大后会占主导（负样本间绝对差距在变大）。

训练框架

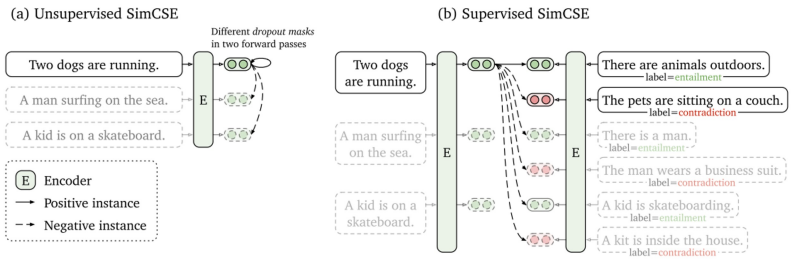


Figure 1: (a) Unsupervised SimCSE predicts the input sentence itself from in-batch negatives, with different dropout masks applied. (b) Supervised SimCSE leverages the NLI datasets and takes the entailment (premise-hypothesis) pairs as positives, and contradiction pairs as well as other in-batch instances as negatives.

Alignment and uniformity

为了更好的分析SimCSE，作者使用Wang and Isola (2020)提出的分析工具，分析正例之间的alignment和所有样本空间的uniformity，发现SimCSE有效的“flattens”向量空间的奇异值，改善向量的分布。

作者发现无监督的SimCSE本质上改善了uniformity，同时避免了通过丢失噪声退化alignment，从而大大提高了表示的表达能力。同时发现NLI训练信号能够明显改善同类之间的alignment。

tations. Given a distribution of positive pairs p_{pos}, *alignment* calculates expected distance between embeddings of the paired instances (assuming representations are already normalized),

$$\ell_{\mathrm{align}} \triangleq \mathop{\mathbb{E}}_{(x, x^+) \sim p_{\mathrm{pos}}} \|f(x) - f(x^+)\|^2. \qquad (2)$$

On the other hand, *uniformity* measures how well the embeddings are uniformly distributed:

$$\ell_{\text{uniform}} \triangleq \log \underset{x,y \overset{i.i.d.}{\sim} p_{\text{data}}}{\mathbb{E}} e^{-2\|f(x)-f(y)\|^2}, \quad (3)$$

where p_{data} denotes the data distribution. These

alignment计算了句子对距离的期望，而uniformity则用来衡量向量的分布是否一致

无监督 SimCSE

~~最核心的创新点就是使用droupout来对文本增加噪音，从而构造一个正样本对，而负样本对则是在batch中选取的其它句子。其实对于图像任务来说，做数据增强其实非常简单，有各种的手段。但是对于NLP任务来说，~~**~~传统的方法有词替，裁剪以及回译，但是作者发现这些方法都没有简单的dropout效果好。~~**

无监督的SimCSE使用dropout构造正样本：相同的句子使用模型编码两次（dropout随机种子不同）。这种方法比常用的数据增强方法（删除、替换）效果更好。通过仔细的分析，发现dropout本质上是作为最小的数据扩充，而删除它会导致表示collapse。

$$\ell_i = -\log \frac{e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z_i'})/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z_j'})/\tau}}, \quad (4)$$

训练目标：

Data augmentation			STS-B
None			**79.1**
Crop	*10%*	*20%*	*30%*
	75.4	70.1	63.7
Word deletion	*10%*	*20%*	*30%*
	74.7	71.2	70.2
Delete one word			74.8
w/o dropout			71.4
MLM 15%			66.8
Crop 10% + MLM 15%			70.8

Table 2: Comparison of different data augmentations on STS-B development set (Spearman's correlation). *Crop* k%: randomly crop and keep a continuous span with 100-k% of the length; *word deletion* k%: randomly delete k% words; *delete one word*: randomly delete one word; *MLM* k%: use $\text{BERT}_{\text{base}}$ to replace k% of words. All of them include the standard 10% dropout (except "w/o dropout").

Training objective	f_θ	$(f_{\theta_1}, f_{\theta_2})$
Next sentence	66.8	67.7
Next 3 sentences	68.7	69.7
Delete one word	74.8	70.4
Unsupervised SimCSE	**79.1**	70.7

Table 3: Comparison of different unsupervised objectives. Results are Spearman's correlation on the STS-B development set using $\text{BERT}_{\text{base}}$, trained on 1-million pairs from Wikipedia. The two columns denote whether we use one encoder f_θ or two independent encoders f_{θ_1} and f_{θ_2} ("dual-encoder"). *Next 3 sentences*: randomly sample one from the next 3 sentences. *Delete one word*: delete one word randomly (see Table 2).

p	*0.0*	*0.01*	*0.05*	*0.1*
STS-B	64.9	69.5	78.0	**79.1**
p	*0.15*	*0.2*	*0.5*	*Fixed 0.1*
STS-B	78.6	78.2	67.4	45.2

默认的dropout比例效果最好。

Table 4: Effects of different dropout probabilities p on the STS-B development set (Spearman's correlation, $\text{BERT}_{\text{base}}$). *Fixed 0.1*: use the default 0.1 dropout rate but apply the same dropout mask on both x_i and x_i^+.

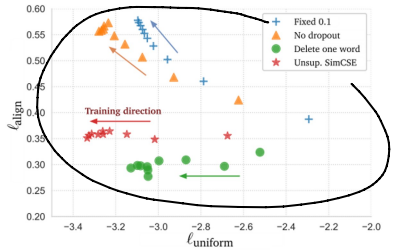


Figure 2: ℓ_{align}-ℓ_{uniform} plot for unsupervised SimCSE, "no dropout", "fixed 0.1" (same dropout mask for x_i and x_i^+ with $p = 0.1$), and "delete one word". We visualize checkpoints every 10 training steps and the arrows indicate the training direction. For both ℓ_{align} and ℓ_{uniform}, *lower numbers are better*.

所有的方法能够改善uniformity，但是alignment在FIXED 0.1和No dropout上明显降低。

有监督 SimCSE

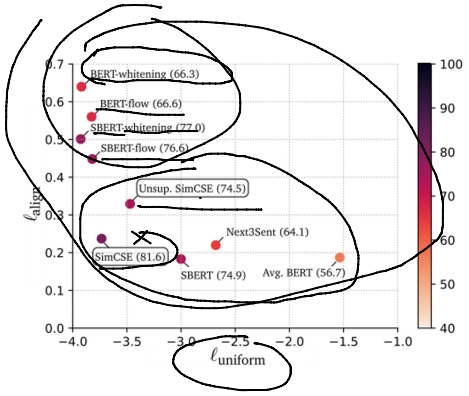
有监督的SimCSE基于NLI数据集，使用entailment关系的句子对表示正样本，contradiction关系的句子对表示负样本。简单的使用使SimCSE达到的sota,并且对比其他的标注的数据集发现NLI效果最好。

$$-\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} \left(e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau} + e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^-)/\tau} \right)}. \quad (5)$$

Dataset	sample	full
Unsup. SimCSE (1m)	-	79.1
QQP (134k)	81.8	81.8
Flickr30k (318k)	81.5	81.4
ParaNMT (5m)	79.7	78.7
SNLI+MNLI		
entailment (314k)	**84.1**	**84.9**
neutral (314k)[3]	82.6	82.9
contradiction (314k)	77.5	77.6
SNLI+MNLI		
entailment + hard neg.		**86.2**
+ ANLI (52k)	-	85.0

Table 5: Comparisons of different supervised datasets as positive pairs. Results are Spearman's correlation on the STS-B development set using $\text{BERT}_{\text{base}}$. Numbers in brackets denote the # of pairs. *Sample*: subsampling 134k positive pairs for a fair comparison between datasets; *full*: using the full dataset. In the last block, we use entailment pairs as positives and contradiction pairs as hard negatives (our final model).

作者对比了在不同数据集上进行训练后在STS-B上的验证表现，并且实验结果表明引入hard neg能提高模型的效果。并且对于有多个contradiction的premise，作者只随机抽取了一个作为hard neg，使用多个hard neg对结果并没有提升。同时添加额外的数据集ANLI没有带来提升。



作者将不同方法得到的Sentence Embeddings空间的Alignment和Uniformity指标进行比对（两个指标均为越小越好），通过下图我们可以得到相比于直接使用预训练的Bert，SimCSE方法较大程度提升了uniformity，这与我们之前的论证符合。与Bert-flow和Bert-whitening这类线性变换方法相比，SimCSE则通过拉近相似句子之间的空间距离，在Alignment上有较大的优势。总体来说SimCSE通过无监督/有监督的对比训练方法在保持Alignment的同时提高了句子向量在特征空间分布的均匀性。

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
∗ SimCSE-$\text{BERT}_{\text{base}}$	**66.68**	**81.43**	**71.38**	**78.43**	**78.47**	**75.49**	**69.92**	**74.54**
$\text{RoBERTa}_{\text{base}}$ (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
∗ SimCSE-$\text{RoBERTa}_{\text{base}}$	**68.68**	**82.62**	**73.56**	**81.49**	**80.82**	**80.48**	**67.87**	**76.50**
∗ SimCSE-$\text{RoBERTa}_{\text{large}}$	**69.87**	**82.97**	**74.25**	**83.01**	**79.52**	**81.23**	**71.47**	**77.47**
Supervised models								
InferSent-GloVe♣	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder♣	64.49	67.80	64.61	76.83	73.18	74.92	76.69	71.22
$\text{SBERT}_{\text{base}}$♣	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
$\text{SBERT}_{\text{base}}$-flow	69.78	77.27	74.35	82.01	77.46	79.12	76.21	76.60
$\text{SBERT}_{\text{base}}$-whitening	69.65	77.57	74.66	82.27	78.39	79.52	76.91	77.00
∗ SimCSE-$\text{BERT}_{\text{base}}$	**75.30**	**84.67**	**80.19**	**85.40**	**80.82**	**84.25**	**80.39**	**81.57**
$\text{SRoBERTa}_{\text{base}}$♣	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
$\text{SRoBERTa}_{\text{base}}$-whitening	70.46	77.07	74.46	81.64	76.43	79.49	76.65	76.60
∗ SimCSE-$\text{RoBERTa}_{\text{base}}$	**76.53**	**85.21**	**80.95**	**86.03**	**82.57**	**85.83**	**80.50**	**82.52**
∗ SimCSE-$\text{RoBERTa}_{\text{large}}$	**77.46**	**87.27**	**82.36**	**86.66**	**83.93**	**86.70**	**81.95**	**83.76**

在STS任务上，SimCSE可以说是全面超越之前的方法，个别的甚至提升了接近十个点。而且仅仅使用无监督的方法，就已经超越了之前的一些有监督的方法。

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
Unsupervised models								
GloVe embeddings (avg.)♣	77.25	78.30	91.17	87.85	80.18	83.00	72.87	81.52
Skip-thought♡	76.50	80.10	93.60	87.10	82.00	92.20	73.00	83.50
Avg. BERT embeddings♣	78.66	86.25	94.37	88.66	84.40	**92.80**	69.54	84.94
BERT-[CLS]embedding♣	78.68	84.85	94.21	88.23	84.13	91.40	71.13	84.66
IS-$\text{BERT}_{\text{base}}$♡	**81.09**	**87.18**	**94.96**	**88.75**	**85.96**	88.64	74.24	**85.83**
∗ SimCSE-$\text{BERT}_{\text{base}}$	80.41	85.30	94.46	88.43	85.39	87.60	71.13	84.67
w/ MLM	80.74	85.67	94.68	87.21	84.95	89.40	**74.38**	85.29
∗ SimCSE-$\text{RoBERTa}_{\text{base}}$	79.67	84.61	91.68	85.96	84.73	84.20	64.93	82.25
w/ MLM	**82.02**	**87.52**	**94.13**	**86.24**	**88.58**	**90.20**	**74.55**	**86.18**
∗ SimCSE-$\text{RoBERTa}_{\text{large}}$	80.83	85.30	91.68	86.10	85.06	89.20	**75.65**	84.83
w/ MLM	**83.30**	**87.50**	**95.27**	**86.82**	**87.86**	**94.00**	75.36	**87.16**
Supervised models								
InferSent-GloVe♣	81.57	86.54	92.50	90.38	84.18	88.20	75.77	85.59
Universal Sentence Encoder♣	80.09	85.19	93.98	86.70	86.38	93.20	70.14	85.10
$\text{SBERT}_{\text{base}}$♣	**83.64**	**89.43**	94.39	**89.86**	**88.96**	**89.60**	76.00	**87.41**
∗ SimCSE-$\text{BERT}_{\text{base}}$	82.69	89.25	**94.81**	89.59	87.31	88.40	73.51	86.51
w/ MLM	82.68	88.88	94.52	89.82	88.41	87.60	**76.12**	86.86
$\text{SRoBERTa}_{\text{base}}$	84.91	90.83	92.56	88.75	90.50	88.60	**78.14**	87.76
∗ SimCSE-$\text{RoBERTa}_{\text{base}}$	84.92	**92.00**	**94.11**	**89.82**	91.27	88.80	75.65	88.08
w/ MLM	**85.08**	91.76	94.02	89.72	**92.31**	**91.20**	76.52	**88.66**
∗ SimCSE-$\text{RoBERTa}_{\text{large}}$	88.12	92.37	95.11	90.49	92.75	91.80	76.64	89.61
w/ MLM	**88.45**	**92.53**	**95.19**	**90.58**	**93.30**	**93.80**	**77.74**	**90.23**

作者还在下游任务上对该模型进行了测试，可以发现在迁移任务上该方法并没有做到最好，不过添加了MLM任务效果有一定的提升，同时这也证明了作者的说法，句子级别的目标可能并不会有益于下游任务的训练，训练好的句子向量表示模型也并不是为了更好的适应下游任务。