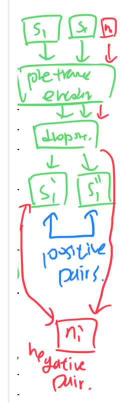


we find that dropout acts as minimal "data augmentation" of hidden representations while removing it leads to a representation collapse

$sim(h_1,h_2) = \frac{h_1^T h_2}{|h_1||h_2|}$ same. τ: temperature

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

different.

$f_\theta(x_i) = h_i$ $\quad f_\theta(x_i^+) = h_i^+$

f : pretrained language model.
BERT & RoBERTa.

$h = f_\theta(x)$

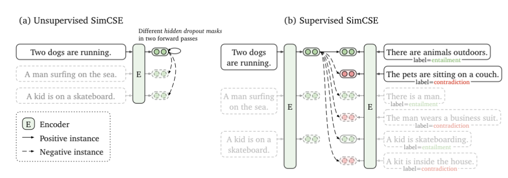


Figure 1: (a) Unsupervised SimCSE predicts the input sentence itself from in-batch negatives, with different hidden dropout masks applied. (b) Supervised SimCSE leverages the NLI datasets and takes the entailment (premise-hypothesis) pairs as positives, and contradiction pairs as well as other in-batch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

To better understand the strong performance of SimCSE, we borrow the analysis tool from Wang and Isola (2020), which takes *alignment* between semantically-related positive pairs and *uniformity* of the whole representation space to measure the quality of learned embeddings. Through empirical analysis, we find that our unsupervised SimCSE essentially improves uniformity while avoiding degenerated alignment via dropout noise, thus improving the expressiveness of the representations. The same analysis shows that the NLI training signal can further improve alignment between positive pairs and produce better sentence embeddings. We also draw a connection to the recent findings that pre-trained word embeddings suffer from anisotropy (Ethayarajh, 2019; Li et al., 2020) and prove that—through a [proof] spectrum perspective—the contrastive learning objective "flattens" the singular value distribution of the sentence embedding space, hence improving uniformity.

Key 1 alignment.

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

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

Key 2 uniformity.

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

why i.i.d?

3 Unsupervised SimCSE

$\{x_i\}_{i=1}^m$ use $x_i^+ = x_i$

$\boxed{\text{Dropout}}$ $h_i^z = f_\theta(x_i; z)$.

z is random mask for dropout

Training objective	f_θ	$(f_{\theta_1}, f_{\theta_2})$
Next sentence	67.1	68.9
Next 3 sentences	67.4	68.8
Delete one word	75.9	73.1
Unsupervised SimCSE	**82.5**	80.7

Single encoder. f_θ $\quad$ Dual encoder. $f_{\theta_1}, f_{\theta_2}$

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

∂ negative pair & positive pair.
term 의 영향력을 어떻게 조절하는지

p	*0.0*	*0.01*	*0.05*	*0.1*
STS-B	71.1	72.6	81.1	**82.5**
p	*0.15*	*0.2*	*0.5*	*Fixed 0.1*
STS-B	81.4	80.5	71.0	43.6

same mask

Table 3: Effects of different dropout probabilities p on the STS-B development set (Spearman's correlation, $\text{BERT}_{\text{base}}$). *Fixed 0.1*: default 0.1 dropout rate but apply the same dropout mask on both x_i and x_i^+.

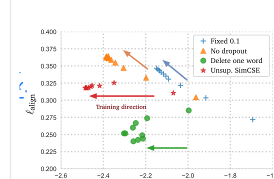


Figure 2: ℓ_{align}-ℓ_{uniform} plot for unsupervised SimCSE, "no dropout", "fixed 0.1", and "delete one word". We visualize checkpoints every 10 training steps and the arrows indicate the training direction. For both ℓ_{align} and ℓ_{uniform}, *lower numbers are better*.

5 Connection to Anisotropy

$$-\frac{1}{\tau}\mathop{\mathbb{E}}_{(x,x^+)\sim p_{\text{pos}}}\left[f(x)^\top f(x^+)\right] + \mathop{\mathbb{E}}_{x\sim p_{\text{data}}}\left[\log \mathop{\mathbb{E}}_{x^-\sim p_{\text{data}}}\left[e^{f(x)^\top f(x^-)/\tau}\right]\right], \quad (6)$$

minimize

$\sum_{i=1} \ell_i = -\sum_{i=1} \log e^{\sin(h_i, h_i^+)/\tau} + \sum_{i=1} \log \sum_{j=1}^N e^{\sin(h_i, h_j^+)/\tau}$

$= -\frac{1}{\tau} \mathop{E}_{(x_i,x_i^+)\sim p_{pos}} \left[f(x_i; z_i)^T f(x_i^+; z_i) \right]$

$+ \mathop{E}_{x\sim p_{data}} \left[\log \mathop{E}_{x^-\sim p_{data}} \left[e^{f(x)^T f(x^-)/\tau} \right] \right]$

p_{data} is uniform over finite samples $\{x_i\}_{i=1}^m$, with $\mathbf{h}_i = f(x_i)$, we can derive the following formula from the second term with Jensen's inequality:

non-convex 이므로. 일차가고있다.

$$\mathop{\mathbb{E}}_{x\sim p_{\text{data}}}\left[\log \mathop{\mathbb{E}}_{x^-\sim p_{\text{data}}}\left[e^{f(x)^\top f(x^-)/\tau}\right]\right] = \frac{1}{m}\sum_{i=1}^m \log\left(\frac{1}{m}\sum_{j=1}^m e^{\mathbf{h}_i^\top \mathbf{h}_j/\tau}\right) \geq \frac{1}{\tau m^2}\sum_{i=1}^m\sum_{j=1}^m \mathbf{h}_i^\top \mathbf{h}_j. \quad (7)$$

W : sentence embedding matrix.
corresponding to $\{x_i\}_{i=1}^m$
i-th row of W

$[W]_i = h_i$

$\text{Sum}(WW^T) = \sum_{i=1}^m \sum_{j=1}^m h_i^T h_j$

$\forall i,\ |h_i| = 1. \Rightarrow \text{diag}(WW^T) = 1$

WW^T : symmetric matrix.

$p(t) = \det(A - tI) = 0$ if, get eigenvalue.

$= c(t-\lambda_1) \cdots (t-\lambda_n)$.

$= (t^n \cdots \sum_{i=1}^n \lambda_i - \prod_{i=1}^n \lambda_i)$

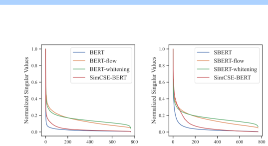
$\frac{\partial \det(A)}{\partial t} = a \sum_{i=1}^n \lambda_i$

$A = PDP^{-1}$
$tr(A) = tr(D) =$ eigenvalue.

symmetric. ⇒ 1. real value eigenvalue.
2. $x_i x_j = \delta_{ij}$
3. diagonizable.

$(WW^T)_{ij} > 0$.

F Distribution of Singular Values



Unsupervised models $\quad$ Supervised models

Figure F.1: Singular value distributions of sentence embedding matrix from sentences in STS-B. We normalize the singular values so that the largest one is 1.

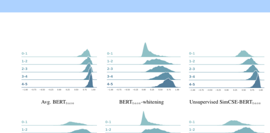


Figure G.1: Density plots of cosine similarities between sentence pairs in STS-B. Pairs are divided into 5 groups based on ground truth ratings (higher means more similar) along the y-axis, and x-axis is the cosine similarity.