Non-Linguistic Supervision for Contrastive Learning of Sentence Embeddings

Yiren Jian                           Chongyang Gao*                  Soroush Vosoughi
Background: Sentence Embedding Learning

Goal: Semantic similar sentences should have “close” embeddings

Solution: Contrastive learning (SimCSE)

Two dogs are running
A man riding bike

Pretrained LM

Different Dropout

Gao T, Yao X, Chen D. Simcse: Simple contrastive learning of sentence embeddings[J]. arXiv preprint arXiv:2104.08821, 2021.
Sentence Embedding Models as General Contrastive Learners

Treating SimCSE as a contrastive learner:

• SimCSE basically contrasts test examples under different views

• We propose to learn a more generalized contrastive learner by examples from other modalities, e.g., images or audio

• It doesn’t require to aligned (paired) examples
VisualCSE: Learning CSE with Text and Image

\[ \mathcal{L}_{\text{sup}} = \sum_{i=1}^{N} - \log \frac{e^{\text{sim}(h_i, h'_i)}/\tau}{\sum_{j=1}^{N} e^{\text{sim}(h_i, h'_j)}/\tau} \]

\[ \mathcal{L}_{\text{sup}}^{\text{image}} = \sum_{i=1}^{N} - \log \frac{e^{\text{sim}(f_i, f'_i)}/\tau}{\sum_{y_i \text{ and } y_j \text{ from same class}} e^{\text{sim}(f_i, f'_j)}/\tau} + \sum_{y_i \text{ and } y_j \text{ from different class}} e^{\text{sim}(f_i, f'_j)}/\tau \]
Results of VisualCSE

| Model               | STS12  | STS13  | STS14  | STS15  | STS16  | STS-B  | SICK-R | Avg.  |
|---------------------|--------|--------|--------|--------|--------|--------|--------|-------|
| SimCSE-BERT\textsubscript{base}  | 68.40  | 82.41  | 74.38  | 80.91  | 78.56  | 76.85  | 72.23  | 76.25 |
| VisualCSE-BERT\textsubscript{base} | 71.16  | 83.29  | 75.13  | 81.59  | 80.05  | 80.03  | 71.23  | 77.50 |
| SimCSE-RoBERT\textsubscript{a base} | 70.16  | 81.77  | 73.24  | 81.36  | 80.65  | 80.22  | 68.56  | 76.57 |
| VisualCSE-RoBERT\textsubscript{a base} | 70.41  | 83.51  | 74.87  | 82.79  | 81.67  | 81.89  | 69.95  | 77.87 |
| SimCSE-RoBERT\textsubscript{a large} | 72.86  | 83.99  | 75.62  | 84.77  | 81.80  | 81.98  | 71.26  | 78.90 |
| VisualCSE-RoBERT\textsubscript{a large} | 73.09  | 84.77  | 77.09  | 85.47  | 82.06  | 83.26  | 72.23  | 79.71 |
# Results of AudioCSE

Replacing images with audios

| Model               | STS12 | STS13 | STS14 | STS15 | STS16 | STS-B | SICK-R | Avg.  |
|---------------------|-------|-------|-------|-------|-------|-------|--------|-------|
| **Unsupervised models** |       |       |       |       |       |       |        |       |
| SimCSE-BERT<sub>base</sub> 🌟 | 68.40 | 82.41 | 74.38 | 80.91 | 78.56 | 76.85 | 72.23  | 76.25 |
| AudioCSE-BERT<sub>base</sub> | **71.65** | **84.27** | **76.69** | **83.22** | **78.69** | **79.94** | 70.49  | **77.85** |
| SimCSE-RoBERT<sub>base</sub> 🌟 | 70.16 | 81.77 | 73.24 | 81.36 | 80.65 | 80.22 | 68.56  | 76.57 |
| AudioCSE-RoBERT<sub>base</sub> | 68.44 | **83.96** | **75.77** | **82.38** | **82.07** | **81.63** | **70.56** | **77.83** |
| SimCSE-RoBERT<sub>large</sub> 🌟 | **72.86** | 83.99 | 75.62 | 84.77 | 81.80 | 81.98 | 71.26  | 78.90 |
| AudioCSE-RoBERT<sub>large</sub> | 72.10 | **84.30** | **76.74** | **85.11** | **82.51** | **82.94** | **72.45** | **79.45** |
Evaluating on other languages

A key advantage of our Non-linguistic CSE is that it does not require aligned (paired) examples, allowing us to apply them to different languages.

| Language | Model   | Spearman’s |
|----------|---------|------------|
| German   | SimCSE  | 67.34      |
|          | VisualCSE | 69.87     |
| French   | SimCSE  | 70.31      |
|          | VisualCSE | 72.52     |
| Russian  | SimCSE  | 72.50      |
|          | VisualCSE | 77.48     |
What does additional supervision improve?

Non-linguistic supervision improves the alignment of sentence embeddings.
Discussion and Conclusion

• A novel framework to learn **generalized contrastive learners** from unpair examples to improve sentence embeddings.

• A finding that knowledge transfer between language and images/audio could be transferred using “unpaired” examples.