A Training Details

We provide hyper-parameters of our models in Table A.1. During the training, our models are evaluated on the STS-B validation set and best models are finally tested on testing sets of the 7 STS tasks.

<table>
<thead>
<tr>
<th>Model</th>
<th>GPU</th>
<th>Text Bsz</th>
<th>Text LR</th>
<th>Img/Audio Bsz</th>
<th>Img/Audio LR</th>
</tr>
</thead>
<tbody>
<tr>
<td>VisualCSE-BERT-base</td>
<td>RTX-6000</td>
<td>64</td>
<td>$3 \times 10^{-5}$</td>
<td>48</td>
<td>$1 \times 10^{-5}$</td>
</tr>
<tr>
<td>AudioCSE-BERT-base</td>
<td>RTX-6000</td>
<td>64</td>
<td>$3 \times 10^{-5}$</td>
<td>48</td>
<td>$2 \times 10^{-7}$</td>
</tr>
<tr>
<td>VisualCSE-RoBERTa-base</td>
<td>RTX-6000</td>
<td>128</td>
<td>$2 \times 10^{-5}$</td>
<td>48</td>
<td>$5 \times 10^{-7}$</td>
</tr>
<tr>
<td>AudioCSE-RoBERTa-base</td>
<td>RTX-6000</td>
<td>128</td>
<td>$2 \times 10^{-5}$</td>
<td>48</td>
<td>$2 \times 10^{-7}$</td>
</tr>
<tr>
<td>VisualCSE-RoBERTa-large</td>
<td>RTX-A6000</td>
<td>256</td>
<td>$1 \times 10^{-5}$</td>
<td>48</td>
<td>$5 \times 10^{-7}$</td>
</tr>
<tr>
<td>AudioCSE-RoBERTa-large</td>
<td>RTX-A6000</td>
<td>256</td>
<td>$1 \times 10^{-5}$</td>
<td>48</td>
<td>$5 \times 10^{-6}$</td>
</tr>
</tbody>
</table>

Table A.1: Hyper-parameters used for training our VisualCSE and AudioCSE. All the batch sizes (Bsz) and learning rates (LR) are selected based on validation on STS-B.

B Learning Sentence Embeddings with Audio Supervision

Learning AudioCSE using audio is very similar to VisualCSE, with the only difference being in pre-processing of the raw data. The raw audio data has the shape $B \times 1024 \times 128$. We use publicly available code from Gong et al. [2] (called SSAST) to process the data so that each example is represented by a tensor of shape $1 \times 211 \times h$. This tensor is then passed into the Transformer encoder during the training.

Different from image learning that allows us to leverage common data augmentations in Computer Vision, we use Dropout augmentation (the same strategy in SimCSE) for AudioCSE. For an audio example, the two views of the input $(y_i', y_i''')$ are passed into the embedding layer of SSAST. The

---

*Contributed as co-first author.

SupCon contrastive loss is applied at the two output representations \((f'_i, f''_i)\) at the [CLS] position:

\[
L_{\text{SupCon}}^\text{audio} = \sum_{i=1}^{N} - \log \frac{e^{\text{sim}(f'_i, f''_i) / \tau}}{\sum_{y_i \text{ and } y_j \text{ from same class}} e^{\text{sim}(f'_i, f''_j) / \tau}} + \sum_{y_i \text{ and } y_j \text{ from different class}} e^{\text{sim}(f'_i, f''_j) / \tau}
\]

(1)

Or, the SimCLR loss equivalent:

\[
L_{\text{SimCLR}}^\text{audio} = \sum_{i=1}^{N} - \log \frac{e^{\text{sim}(f'_i, f''_i) / \tau}}{\sum_{j=1}^{N} e^{\text{sim}(f'_i, f''_j) / \tau}}
\]

(2)

Same as with images, the audio dataset is not paired with the text data and \(L_{\text{audio}}\) does not make any assumption or have any constraints on the audio data that is used for training, meaning that datasets other than LibriSpeech can be used by our framework as well.

For learning with \(L_{\text{audio}}\), we use train-clean-100 split from the LibriSpeech dataset [5]. We use a subset of LibriSpeech \(S_{\text{audio}}\) comprising of 29K examples from 50 classes.

C Qualitative Analysis

We compare unsup-SimCSE and unsup-VisualCSE on a small scale retrieval test. We use 150K captions in Flicker30k as the database and retrieve the top-3 most similar sentences of a query by the two models. As shown in Table C.1, VisualCSE generally retrieves qualitatively different sentences than SimCSE.

<table>
<thead>
<tr>
<th>Query</th>
<th>unsup-SimCSE-BERT-base</th>
<th>unsup-VisualCSE-BERT-base</th>
</tr>
</thead>
<tbody>
<tr>
<td>#1</td>
<td>A double decker red bus moving along the street.</td>
<td>A yellow van driving down crowded city street.</td>
</tr>
<tr>
<td>#2</td>
<td>A bus with people on it.</td>
<td>A red double-decker bus in Europe.</td>
</tr>
<tr>
<td>#3</td>
<td>The blue bus passes pedestrians on a busy city street.</td>
<td>A city street is shown with people and double-decker buses.</td>
</tr>
<tr>
<td>Query: A couple posts for a picture in front of a bus.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#1</td>
<td>A couple are paying their fare at the front of a bus.</td>
<td>A couple are paying their fare at the front of a bus.</td>
</tr>
<tr>
<td>#2</td>
<td>Two women are walking in a crosswalk with a bus ...</td>
<td>A young couple riding on a bus with the boy’s arm around the girl.</td>
</tr>
<tr>
<td>#3</td>
<td>A man and woman pose in front of some traffic.</td>
<td>A man and a woman sit on a bus, but not together.</td>
</tr>
<tr>
<td>Query: A cat sits inside of a paper airplane toy.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>#1</td>
<td>A cat is watching a girl construct a Lego airplane.</td>
<td>A cat is watching a young girl playing with a toy airplane.</td>
</tr>
<tr>
<td>#2</td>
<td>A small child plays with her airplane as a cat looks on.</td>
<td>A small child plays with her airplane as a cat looks on.</td>
</tr>
<tr>
<td>#3</td>
<td>A house cat is sitting on top of an old white car.</td>
<td>A cat is watching a girl construct a Lego airplane.</td>
</tr>
</tbody>
</table>

Table C.1: Examples of retrieved top three examples by SimCSE and VisualCSE from the Flickr30k dataset (150k sentences) using random queries.

D Transfer Tasks

Though downstream classification tasks are not the main goal of sentence representation learning, we still evaluate VisualCSE/AudioCSE in the following transfer tasks: MR [7], CR [3], SUBJ [6], MPQA [10], SST-2 [8], TREC [9], and MRPC [1] using Facebook’s publicly available code base of SentEval. Following default settings from SentEval, a linear classifier is trained on top of the frozen sentence embeddings of different models, i.e., SimCSE and VisualCSE.

Table D.1 shows that VisualCSE and AudioCSE generally outperform SimCSE in the transfer benchmarks, though some improvements are marginal. These findings show that the representations learned by our framework can be successfully applied to downstream tasks.

E Ablation on \(L_{\text{audio}}\)

Following what was done in Section 4.4, we study AudioCSE with two different contrastive losses (SimCLR and SupCon). The results in Table E.1 show a similar trend to the results for VisualCSE, shown in Table 4 i.e., changing SupCon to SimCLR does not greatly affect the performance.
Table D.1: Testing accuracy on transfer tasks using different models. ♠: results from SimCSE’s publicly available model.

<table>
<thead>
<tr>
<th>Model</th>
<th>MR</th>
<th>CR</th>
<th>SUBJ</th>
<th>MPQA</th>
<th>SST</th>
<th>TREC</th>
<th>MRPC</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsupervised models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SimCSE-BERT&lt;sub&gt;base&lt;/sub&gt; ♠</td>
<td>81.18</td>
<td>86.46</td>
<td>94.43</td>
<td>88.87</td>
<td>85.50</td>
<td>89.80</td>
<td>74.49</td>
<td>85.82</td>
</tr>
<tr>
<td>VisualCSE-BERT&lt;sub&gt;base&lt;/sub&gt;</td>
<td>82.29</td>
<td>87.18</td>
<td>94.85</td>
<td>89.34</td>
<td>86.55</td>
<td>89.60</td>
<td>74.32</td>
<td>86.30</td>
</tr>
<tr>
<td>AudioCSE-BERT&lt;sub&gt;base&lt;/sub&gt;</td>
<td>81.10</td>
<td>86.76</td>
<td>94.65</td>
<td>88.78</td>
<td>86.00</td>
<td>89.40</td>
<td>74.43</td>
<td>85.87</td>
</tr>
<tr>
<td>SimCSE-RoBERTa&lt;sub&gt;base&lt;/sub&gt; ♠</td>
<td>81.04</td>
<td>87.74</td>
<td>93.28</td>
<td>86.94</td>
<td>86.60</td>
<td>85.20</td>
<td>73.68</td>
<td>84.84</td>
</tr>
<tr>
<td>VisualCSE-RoBERTa&lt;sub&gt;base&lt;/sub&gt;</td>
<td>81.84</td>
<td>87.60</td>
<td>92.76</td>
<td>87.14</td>
<td>86.77</td>
<td>85.20</td>
<td>74.43</td>
<td>85.11</td>
</tr>
<tr>
<td>AudioCSE-RoBERTa&lt;sub&gt;base&lt;/sub&gt;</td>
<td>80.76</td>
<td>87.18</td>
<td>93.39</td>
<td>87.29</td>
<td>86.64</td>
<td>86.60</td>
<td>75.07</td>
<td>85.42</td>
</tr>
<tr>
<td>SimCSE-RoBERTa&lt;sub&gt;large&lt;/sub&gt; ♠</td>
<td>82.74</td>
<td>87.87</td>
<td>93.66</td>
<td>88.22</td>
<td>88.58</td>
<td>92.00</td>
<td>69.68</td>
<td>86.11</td>
</tr>
<tr>
<td>VisualCSE-RoBERTa&lt;sub&gt;large&lt;/sub&gt;</td>
<td>83.44</td>
<td>88.13</td>
<td>93.96</td>
<td>88.43</td>
<td>88.63</td>
<td>91.60</td>
<td>73.86</td>
<td>86.86</td>
</tr>
<tr>
<td>AudioCSE-RoBERTa&lt;sub&gt;large&lt;/sub&gt;</td>
<td>83.49</td>
<td>88.64</td>
<td>93.86</td>
<td>88.82</td>
<td>88.63</td>
<td>92.40</td>
<td>69.97</td>
<td>86.54</td>
</tr>
</tbody>
</table>

Table E.1: Comparison of AudioCSE (base) using SupCon and SimCLR as $L_{image}$. 

F VisualCSE with CIFAR images

We further carry out experiments on substituting ImageNet with a lower quality CIFAR [4] dataset for VisualCSE. CIFAR images have a shape of 32x32 and we intentionally resize (enlarge) them to 224x224 to be encoded by the ViT embedding layer. This interpolation causes the CIFAR images to become blurry and lower quality. As shown in Table F.1, our framework improves over SimCSE even with this lower quality dataset (results shown below).

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SimCSE-BERT&lt;sub&gt;base&lt;/sub&gt;</td>
<td>76.25</td>
</tr>
<tr>
<td>VisualCSE-BERT&lt;sub&gt;base&lt;/sub&gt; (CIFAR)</td>
<td>76.96</td>
</tr>
<tr>
<td>VisualCSE-BERT&lt;sub&gt;base&lt;/sub&gt; (ImageNet)</td>
<td>77.50</td>
</tr>
<tr>
<td>SimCSE-RoBERTa&lt;sub&gt;base&lt;/sub&gt;</td>
<td>76.57</td>
</tr>
<tr>
<td>VisualCSE-RoBERTa&lt;sub&gt;base&lt;/sub&gt; (CIFAR)</td>
<td>77.71</td>
</tr>
<tr>
<td>VisualCSE-RoBERTa&lt;sub&gt;base&lt;/sub&gt; (ImageNet)</td>
<td>77.87</td>
</tr>
</tbody>
</table>

Table F.1: Comparison of VisualCSE using CIFAR and ImageNet

G Pre-Trained Models & Datasets for Other Languages

For our experiments on German, French, Russian and Chinese, we use the following pre-trained models from bert-base-german-cased, dbmdz/bert-base-french-europeana-cased, DeepPavlov/rubert-base-cased, and bert-base-chinese.

The evaluations of SimCSE and VisualCSE for these languages are done on validation sets of the multi-lingual STS-B from HuggingFace, stsb_multi_mt.
References


