Non-Linguistic Supervision for Contrastive Learning of Sentence Embeddings

Yiren Jian
Department of Computer Science
Dartmouth College
yiren.jian.gr@dartmouth.edu

Chongyang Gao*
Department of Computer Science
Northwestern University
chongyanggao2026@u.northwestern.edu

Soroush Vosoughi
Department of Computer Science
Dartmouth College
soroush.vosoughi@dartmouth.edu

Abstract

Semantic representation learning for sentences is an important and well-studied problem in NLP. The current trend for this task involves training a Transformer-based sentence encoder through a contrastive objective with text, i.e., clustering sentences with semantically similar meanings and scattering others. In this work, we find the performance of Transformer models as sentence encoders can be improved by training with multi-modal multi-task losses, using *unpaired* examples from another modality (e.g., sentences and unrelated image/audio data). In particular, besides learning by the contrastive loss on text, our model clusters examples from a non-linguistic domain (e.g., visual/audio) with a similar contrastive loss at the same time. The reliance of our framework on unpaired non-linguistic data makes it language-agnostic, enabling it to be widely applicable beyond English NLP. Experiments on 7 semantic textual similarity benchmarks reveal that models trained with the additional non-linguistic (images/audio) contrastive objective lead to higher quality sentence embeddings. This indicates that Transformer models are able to generalize better by doing a similar task (i.e., clustering) with *unpaired* examples from different modalities in a multi-task fashion. The code is available at https://github.com/yiren-jian/NonLing-CSE.

1 Introduction

Learning semantically meaningful sentence embeddings is an important task in Natural Language Processing (NLP), with applications for semantic retrieval and search among other things. The embeddings of sentences are such that semantically similar sentences are located closer to each other in the embedding space. The current state-of-the-art method for learning sentence embeddings is SimCSE [13], which uses contrastive learning. Unsupervised contrastive learning clusters the representations of same examples under different augmentations. SimCSE's main contribution is the use of Dropout augmentation, which they show outperforms other traditional text augmentation methods (e.g., random deletion, swapping, etc.) for learning sentence embeddings. To improve the results further, supervised SimCSE leverages a labeled Natural Language Inference (NLI) dataset during the learning. Each entry of the NLI dataset is a triplet of sentences, a source sentence and its positive and negative pairs (e.g., *"Two dogs are running"* (src), *"There are animals outdoors"* (pos),

*Contributed as co-first author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

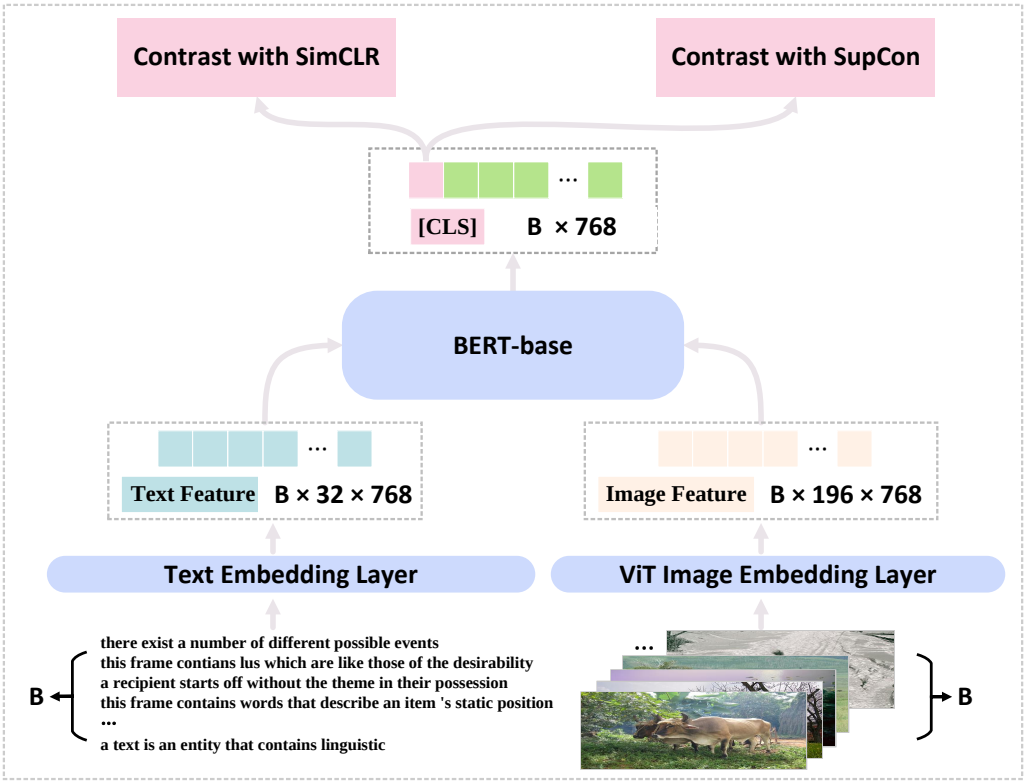

Figure 1: Overview of our method VisualCSE. The mini-batch contains two batches, one with sentences from Wikipedia and the other with images from ImageNet. The text batch is processed by a BERT embedding layer and the image batch is processed by a Vision Transformer (ViT) [11] embedding layer separately. The resulted text and image features are fed into the Transformer encoder and we take the outputs at `[CLS]` position as the representations of examples. Like in SimCSE, the text representations are supervised by the SimCLR loss.

"The pets are sitting on a couch" (neg)). Supervised SimCSE uses hard-negative and strong-positive terms in the contrastive learning objective based on the negative and positive examples. Because of the high quality of negative and positive pairs in the NLI dataset, the learning signal from these terms has been shown to be very effective in learning sentence embeddings [13].

However, as we show in this paper, the benefits of a labeled dataset (such as NLI) are highly dependent on its size and the annotation quality. The need for such large and high quality datasets limits the applicability of the method to high-resource languages. For instance, the dataset used by SimCSE is a combination of the SNLI [2] and MNLI [57] datasets which contain 570K and 433K manually annotated examples, respectively. Constructing such a dataset is costly and may be impossible in low-resource languages. In fact, it has been shown that the performance of supervised SimCSE is highly dependent on the specific labeled dataset used. Specifically, training SimCSE with QQP [43], ANLI [37], Flickr30k [62], and ParaNMT [56] was shown to lead to significant performance drops [13]. Such specific reliance on an NLI dataset is especially problematic as constructing it in a new language can be very difficult as it requires annotators to understand each sentence and write a seemingly related, but semantically different sentence. Further supporting this claim, Zhang et al. [64] show that if the distribution of the test set differs significantly from the NLI dataset used for training, the quality of the semantic embeddings is degraded.

The aforementioned issues show that learning sentence embeddings from *unlabeled text data* is a crucial task that requires further exploration. In this paper, we investigate the addition of non-linguistic supervision (provided by easily obtainable public datasets) to contrastive-learning-based sentence embedding learning. Specifically, we propose a Transformer-based multi-modal multi-task framework

which jointly learns two independent contrastive tasks (i.e., one textual and one non-textual) using *unpaired* examples. Our method (shown in Figure 1) is inspired by Lu et al. [35] which shows the ability of Transformers to transfer knowledge from one modality to another (e.g., they show that a Transformer model pre-trained on text can be fine-tuned on downstream visual tasks). In addition to generating higher-quality sentence embeddings, our investigations shed light on the capability of Transformers to learn from mixed modalities when trained on similar tasks. This is complementary to the findings of Lu et al., as those concern the cross-modal transfer ability of Transformers.

Though our method can be used with any modality, we use vision as the modality of choice to describe our method and later show the generalizability of our method to audio. As shown in Figure 1, our method, VisualCSE (or AudioCSE when learning with audio), has a multi-task loss with two terms: one contrastive learning loss for text, which is the same as the one used in SimCSE, and an additional contrastive learning loss to cluster data from images concurrently. Note that the two learning processes are independent of each other, thus allowing us to leverage *unpaired* text and image/audio datasets, e.g., sentences from Wikipedia, images from ImageNet [9], and audio from LibriSpeech [39]. Our method is different than other multi-modal representation learning methods (such as CLIP [44]) in that it does not require paired examples across modalities. This crucial feature makes our method easy to deploy for low-resource languages. In addition, our method also uses a single model with all parameters shared for tasks in both modalities, highlighting the ability of Transformers to learn from non-parallel mixed modalities when trained on similar tasks.

The contributions of this work are as follows:

(1) We propose a novel framework for introducing non-linguistic supervision for contrastive-learning-based sentence embedding learning. Crucially, our framework is language-agnostic, does not rely on a specific modality, and does not require paired examples across modalities. We show that the supervision provided by other modalities can improve unsupervised sentence embedding learning, enabling sentence embedding learning in low-resource languages. We improve SimCSE in several semantic textual similarity benchmarks with 3 different pre-trained language models (LMs).

(2) Complimentary to the findings of Lu et al. [35], which demonstrate the capability of Transformers to transfer knowledge from language to vision, we show that Transformer models can generalize better by learning a similar task (i.e., clustering) with multi-task losses using non-parallel examples from different modalities.

2 Related Work

Sentence Embedding Learning. Early works on learning sentence embeddings include training a model to reconstruct the nearby sentences of an encoded sentence [30]. Sent2Vec [38] devises an efficient loss for training the sentence encoder, inspired by the word embedding algorithm Word2Vec. Recent works in this area are based on contrastive learning: through taking two different views of examples [14], data augmentation [59], or by passing examples into different encoders [4, 28, 64]. Post-processing methods also exist for improving the quality of embeddings [23, 32, 47]. The current state-of-the-art sentence embeddings method, SimCSE [13] passes the text into an encoder twice and uses Dropout as the augmentation to create multi-views of examples for contrastive learning. All these methods (and recent works such as DiffCSE [8], DebiasedCSE [65], PromptBERT [25] and many others [3, 16, 22, 26, 33, 42, 45, 51, 52]) solely use text datasets for training their sentence encoders.

A concurrent work, MCSE [63] is the first attempt to learn sentence embedding using text with another modality (image). But MCSE requires *aligned* image-text pairs (i.e., images and captions from Flickr), limiting its applicability to other languages and modalities. We are the first to show that the embedding quality of a sentence encoder can be improved by jointly training with unpaired examples from another modality.

Contrastive Learning. Contrastive learning [7, 19, 49, 58, 60] is a form of self-supervised training using a loss derived from noise contrastive estimation [50]. The seminal method SimCLR [5] clusters the same examples under two sets of augmentation (positive pairs), and pulls away different examples in a mini-batch (negative pairs). To alleviate the hardware requirements of SimCLR, MoCo [18] designs a moving average memory bank for storing negative samples, which simulates large number of negative samples for effective contrastive learning. CMC [49] further extends SimCLR from

two-view to multi-view contrastive learning. While the contrastive loss was originally designed for unsupervised pre-training, SupCon [27] finds that additionally contrasting examples based on label information is beneficial when labels of examples are available. In NLP, contrastive learning mainly applies to sentence embedding learning [4, 38, 28, 59, 64], but other works [17, 24] have shown it to be beneficial to few-shot learning [12] as well.

The contrastive objective perfectly matches the goal of sentence embeddings [13], i.e., semantically similar sentences should have comparable representations. Our method for learning sentence embeddings contrasts examples from text and examples from another modality (e.g., image or audio) simultaneously. The image and audio datasets we used come with *"free"* labels, allowing us to apply a loss based on SupCon, in addition to SimCLR.

Learning Representations With Multiple Modalities. Using paired image-text instances, CLIP [44] learns aligned representations of text and images by two separate encoders, achieving superior results in several downstream tasks. Building upon CLIP, SLIP [36] further improves the representations of examples by leveraging a self-supervised contrastive objective. UniT [21] learns a multi-modal Transformer model with a set of aligned tasks from visual and language domains. In video learning, XDC [1] learns an encoder by cross pseudo-labeling between parallel RGB frames and audio signals.

The data used for training such models includes images and their associated captions (CLIP and SLIP), aligned visual-language tasks (UniT), or video clips that come with RGB frames and the corresponding audio (XDC). These methods, as well as many other works on multi-modal learning [6, 29, 34, 48, 61], learn better representations by leveraging information shared among different modes of the same instances (i.e., paired examples). Our method also learns better representations of sentences from examples of different modalities. However, the datasets used for training our method do not require parallel examples. Our method performs similar contrastive tasks in two independent domains at the same time, and generalizes better as a sentence encoder by becoming a better clustering model.

3 Technical Method

3.1 Overview

Given a text dataset $\mathcal{S}_{\text{text}}$ (e.g., Wikipedia or NLI), a dataset from a different modality (e.g., an image dataset $\mathcal{S}_{\text{image}}$ (e.g., ImageNet), or an audio dataset $\mathcal{S}_{\text{audio}}$ (e.g., LibriSpeech)), and a pre-trained LM $\mathcal{M}$ (e.g., BERT), our goal is to fine-tune $\mathcal{M}$ with $\mathcal{S}_{\text{text}}$ and $\mathcal{S}_{\text{image}}/\mathcal{S}_{\text{audio}}$ using contrastive losses such that the resulting model can produce high-quality sentence representations.

In the following sections, we use images as the example to show how non-linguistic supervision can improve the sentence embedding learning (VisualCSE). A similar approach is used to adapt our learning algorithm to audio (AudioCSE), as shown in Appendix B. Our method trains the sentence encoder with a multi-task loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{text}} + \omega_{\text{image}}\mathcal{L}_{\text{image}} \tag{1}$$

where ω_{image} is a weighting factor. We provide further details of $\mathcal{L}_{\text{text}}$ and $\mathcal{L}_{\text{image}}$ in Sections 3.2 and 3.3, respectively.

3.2 Learning from Text

We train the sentence encoder with text using the method laid out by SimCSE. This allows us to directly investigate the contribution of additional non-linguistic supervision for training sentence encoders. Here, we briefly summarize how SimCSE works.

SimCSE learns the sentence encoder by a contrastive loss, using Dropout as a form of data augmentation. For a single unlabeled example x_i of $\mathcal{S}_{\text{text}}$ (Wikipedia) in a mini-batch, SimCSE passes x_i into the Transformer encoder $\mathcal{M}$ twice. The corresponding output hidden states at `[CLS]` position (which are used as sentence embeddings) are denoted as $\mathbf{h}_i^{z_i}$ and $\mathbf{h}_i^{z_i'}$, where z_i and z_i' are two different random masks by Dropout layers of the Transformer model. SimCSE applies the SimCLR loss to contrast examples in a mini-batch, i.e., the unsupervised loss based on data augmentation is as

follows:

$$\mathcal{L}_{\text{text}}^{\text{unsup}} = \sum_{i=1}^{N} -\log \frac{e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_i^{z_i'})/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{h}_i^{z_i}, \mathbf{h}_j^{z_j'})/\tau}} \tag{2}$$

where j denotes another example in the mini-batch, N is the size of mini-batch, `sim` is the operation for computing cosine similarity, and τ is a temperature scaling factor. The key to SimCSE is using Dropout as a form of augmentation, the rest of its contrastive learning follows SimCLR.

If $\mathcal{S}_{\text{text}}$ is a labeled NLI dataset that comes with (x_i, x_i^+, x_i^-) (i.e., source, entailment, contradiction), corresponding to the hidden states (h_i, h_i^+, h_i^-), its supervised loss is defined as:

$$\mathcal{L}_{\text{text}}^{\text{sup}} = \sum_{i=1}^{N} -\log \frac{e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_i^+)/\tau}}{\sum_{j=1}^{N} (e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^+)/\tau} + e^{\text{sim}(\mathbf{h}_i, \mathbf{h}_j^-)/\tau})} \tag{3}$$

Following SimCSE, we use $\tau = 0.05$ for both losses.

3.3 Learning from Images

We train our encoder with images using SupCon (main results) and SimCLR (in ablation). For an image y_i in $\mathcal{S}_{\text{image}}$, we apply two sets of data augmentations in the input RGB space. The resulting views of the input (y_i', y_i'') are passed into a ViT embedding layer and further encoded by the same Transformer encoder $\mathcal{M}$ from Section 3.2. The SupCon contrastive loss is applied on the two output representations $(\mathbf{f}_i', \mathbf{f}_i'')$ at the `[CLS]` position:

$$\mathcal{L}_{\text{image}}^{\text{SupCon}} = \sum_{i=1}^{N} -\log \frac{e^{\text{sim}(\mathbf{f}_i', \mathbf{f}_i'')/\tau} + \sum\limits_{y_i \text{ and } y_j \text{ from same class}} e^{\text{sim}(\mathbf{f}_i', \mathbf{f}_j'')/\tau}}{\sum\limits_{y_i \text{ and } y_j \text{ from different class}} e^{\text{sim}(\mathbf{f}_i', \mathbf{f}_j'')/\tau}} \tag{4}$$

We use the SupCon default $\tau = 0.07$. Also, we can substitute SupCon loss with the SimCLR:

$$\mathcal{L}_{\text{image}}^{\text{SimCLR}} = \sum_{i=1}^{N} -\log \frac{e^{\text{sim}(\mathbf{f}_i', \mathbf{f}_i'')/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{f}_i', \mathbf{f}_j'')/\tau}} \tag{5}$$

As we show in Section 4.4, SimCLR and SupCon produce in similar performances. This is because we use a very small batch size (48) due to the limitations of our GPU memory. Thus, the second term in the numerator of $\mathcal{L}_{\text{text}}^{\text{SupCon}}$ is nearly non-existent, and Equation 4 reduces to Equation 5.

Note that $\mathcal{L}_{\text{image}}$ does not make any assumptions on the image dataset. Thus, we can leverage the easily obtainable ImageNet images. Furthermore, $\mathcal{L}_{\text{image}}^{\text{SimCLR}}$ can be applied to any unlabeled images from the Internet, at no cost. This allows our framework to be easily deployed for learning sentence embeddings in a new domain or a new language where NLI, or other high-quality labeled datasets are not available.

3.4 PyTorch Style of Pseudo-Code

We provide the pseudo-code of our algorithm VisualCSE in the style of PyTorch in Algorithm 1. Learning AudioCSE is similar except that we do not apply `transform()` to the audio data, but instead pass the audio data into the model twice (i.e., augmentation is provided through Dropout).

3.5 Why Does It Work?

As we can see, $\mathcal{L}_{\text{text}}$ and $\mathcal{L}_{\text{image}}$ share a similar (if not the same) form, with the only difference being the input data. We speculate that the Transformer model becomes a better contrastive learner by jointly training with similar but independent contrastive objectives in different modalities.

Algorithm 1: Visual contrastive sentence embedding learning (VisualCSE)

```
# Max_Step:         Number of training steps
# LM:               The language model
# D_text:           Dataloader for text training set
# D_image:          Dataloader for image training set
# transform:        A set of data transformations
# SimCLR:           SimCLR loss
# SupCon:           SupCon loss
# opt_text:         Optimizer for learning from text
# opt_image:        Optimizer for learning from image
for i in Max_Step:
    ### Retrieve two augmented views of a same sent/img
    sent = next(D_text)
    sent_0, sent_1 = sent, sent
    img, y = next(D_image)
    img_0, img_1, = transform(img), transform(img)
    ### Learning from text
    L_text = SimCLR(LM(sent_0), LM(sent_1))
    L_text.backward()
    opt_text.step()
    ### Learning from images
    L_img = SupCon(LM(img_0), LM(img_1), y)
    L_img.backward()
    opt_image.step()
return LM
```

4 Experiments

4.1 Setup

Evaluation. Following standard practice for sentence embedding, we evaluate our approach on 7 semantic textual similarity (STS) tasks: STS 2012-2016, STS-Benchmark and SICK-Relatedness. For fair comparison, we use the same evaluation protocol as SimCSE [13], and use Spearman's correlation as the metric.

Models and Training Datasets. We test our method with three different pre-trained Transformer models used in SimCSE: BERT-base-uncased, RoBERTa-base and RoBERTa-large. We fine-tune the pre-trained models with our learning objective, shown in Eq. 1.

For learning with $\mathcal{L}_{\text{text}}$, we use 10^6 sentences down-sampled from the Wikipedia English dataset for unsupervised sentence embedding learning (Eq. 2). For supervised sentence embedding learning (Eq. 3), we (and SimCSE) use a combined NLI dataset with 314K sentences with paired examples labeled as entailment, neutral, and non-entailment. SimCSE uses entailment examples as strong-positives and non-entailment examples as hard-negatives.

For learning with $\mathcal{L}_{\text{image}}$, both unsupervised and supervised sentence embedding settings use a down-sampled ImageNet dataset $\mathcal{S}_{\text{image}}$. Images from $\mathcal{S}_{\text{image}}$ are chosen *randomly* from 60 classes with each class having 500 images[2]. The images have a shape of $B \times 3 \times 224 \times 224$, thus the image features processed by the ViT embedding layers are tensors of size $B \times 197 \times h$, where B is the batch size and h is the hidden dimension of the LM. Learning with audio is described in Appendix B.

Optimization Details. Following SimCSE [13], we train unsupervised models with AdamW for one epoch, and supervised models for 3 epochs. We then search batch sizes and learning rates from $\{64, 128, 256\}$ and $\{1e^{-5}, 2e^{-5}, 3e^{-5}\}$ for $\mathcal{L}_{\text{text}}$. Moreover, we use a fixed batch size of 48 for $\mathcal{L}_{\text{image}}$ (and $\mathcal{L}_{\text{audio}}$) and search learning rates among $\{5e^{-6}, 2e^{-6}1e^{-6}, 5e^{-7}, 2e^{-7}, 1e^{-7}\}$. The models are selected based on the validation set of the STS-Benchmark.

All the unsupervised base LMs are trained on 24GB Nvidia RTX-6000 GPUs, while supervised and large models are trained on 48GB Nvidia RTX-A6000 GPUs. We use pytorch-1.10 with CUDA 11.3,

[2]It takes SimCSE and VisualCSE thousands of steps to converge. Thus, we construct a smaller subset of ImageNet so that the model goes over the dataset for a few epochs.

torchvision-0.11.3, torchaudio-0.10.2, and Huggingface transformers-4.5.0 for our implementation. Training with a different version of software and hardware may lead to different results. We use the default random seed 42 (same as SimCSE) of Huggingface for all of our major experiments for reproducibility.

The validated hyper-parameters can be found in Appendix A.

4.2 Main Results

We show results of our method learning with visual supervision (VisualCSE) in the unsupervised text setting in Table 1. Our method outperforms baseline SimCSE that learns only from unlabeled text with three different pre-trained LMs. Similarly, we show in Table 2 that learning with audio supervision can improve over SimCSE in almost all tasks.

The improvement in performance can be attributed to the fact that SimCSE only clusters text examples, while VisusalCSE or AudioCSE simultaneously cluster text and examples from another modality, making them better contrastive learners. This could indicate that Transformer-based models are able to generalize better by doing the same task (e.g., clustering) with independent examples from different modalities in a multi-task fashion (i.e, with a multi-task loss). Importantly, this learning setup does not require any paired examples across modalities.

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

Table 1: Spearman's correlation of different sentence embedding methods using unsupervised text. Comparing to SimCSE that only leverages text, our VisualCSE takes addition supervision from ***images***. ♠: results from the SimCSE paper and reproduced by officially released models.

Model	**STS12**	**STS13**	**STS14**	**STS15**	**STS16**	**STS-B**	**SICK-R**	**Avg.**
Unsupervised models								
SimCSE-BERT$_{\text{base}}$ ♠	68.40	82.41	74.38	80.91	78.56	76.85	**72.23**	76.25
AudioCSE-BERT$_{\text{base}}$	**71.65**	**84.27**	**76.69**	**83.22**	**78.69**	**79.94**	70.49	**77.85**
SimCSE-RoBERTa$_{\text{base}}$ ♠	**70.16**	81.77	73.24	81.36	80.65	80.22	68.56	76.57
AudioCSE-RoBERTa$_{\text{base}}$	68.44	**83.96**	**75.77**	**82.38**	**82.07**	**81.63**	**70.56**	**77.83**
SimCSE-RoBERTa$_{\text{large}}$ ♠	**72.86**	83.99	75.62	84.77	81.80	81.98	71.26	78.90
AudioCSE-RoBERTa$_{\text{large}}$	72.10	**84.30**	**76.74**	**85.11**	**82.51**	**82.94**	**72.45**	**79.45**

Table 2: Spearman's correlation of different sentence embedding methods using unsupervised text. Comparing to SimCSE that only leverages text, our AudioCSE takes addition supervision from ***audio***.

Though not being the main goal of sentence embedding, we evaluate our models on transfer tasks in Appendix D, showing that the sentence representations learned using our framework can be well-suited to downstream tasks. We also show example sentence retrieval experiments in Appendix C and VisualCSE with lower quality CIFAR images in Appendix F.

4.3 Repeated Experiments

SimCSE uses a fixed random seed (42) for training their models, selects models using the STS-B validation set and reports results on all 7 testing sets of the benchmarks. For direct comparison, we follow this setup in Section 4.2. To investigate the statistical significance of our results, we further run repeated experiments for unsupervised VisualCSE and SimCSE using 5 different random seeds. Because the best validated model is saved at one point during the whole learning process, using different random seeds for training VisualCSE/SimCSE can be thought of as simulating scenarios

where different subsets of the text and image datasets are used for training the models. As shown in Table 3, we see a performance drop for both SimCSE and VisualCSE using BERT-base. However, VisualCSE, using either model, is shown to be statistically significantly better than SimCSE.

4.4 Ablation on $\mathcal{L}_{\text{image}}$

By default we use a variant of SupCon (Eq. 4) as our contrastive objective for images when learning with visual supervision. This allows us to leverage the label information available in the ImageNet dataset. Besides clustering different views of examples in a mini-bacth, SupCon also tries to pull together examples in a mini-batch that share the same label.

However, in practice, due to the limitations of our GPU memory, we train VisualCSE with a batch size of 48. As we have 60 image classes, each with 500 images, in our down-size ImageNet dataset used for training, it is not likely that multiple examples from the same class are contained in one mini-batch. Thus, the SupCon loss in practice degenerates to the SimCLR loss. As we can see from Table 4, changing SupCon to SimCLR does not greatly affect the performance (ablation on $\mathcal{L}_{\text{audio}}$ shows similar trends as shown in Appendix E).

Model	**Avg.**	p
SimCSE-BERT$_{\text{base}}$	$75.74_{\pm 0.90}$	0.0235
VisualCSE-BERT$_{\text{base}}$	$\mathbf{76.96}_{\pm 0.38}$	
SimCSE-RoBERTa$_{\text{base}}$	$76.36_{\pm 0.18}$	<0.0001
VisualCSE-RoBERTa$_{\text{base}}$	$\mathbf{77.74}_{\pm 0.25}$	

Table 3: 5 repeated experiments of unsupervised experiments. We report means and standard deviations for average Spearman's correlation of the 7 STS benchmarks.

The results with SimCLR show that our framework can be successfully deployed even when only unlabeled non-linguistic datasets are available. Nevertheless, in this paper, since we use labeled image and audio datasets, we use SupCon as our default loss to possibly harvest the label information presented in our datasets (though this is unlikely given our batch size).

	STS12	**STS13**	**STS14**	**STS15**	**STS16**	**STS-B**	**SICK-R**	**Avg.**
BERT (SimCLR)	71.81	82.57	75.40	82.58	79.34	78.96	70.83	77.36
BERT (SupCon)	71.16	83.29	75.13	81.59	80.05	80.03	71.23	77.50
RoBERTa (SimCLR)	71.02	83.19	74.47	82.59	81.95	81.94	69.87	77.86
RoBERTa (SupCon)	70.41	83.51	74.87	82.79	81.67	81.89	69.95	77.87

Table 4: Comparison of VisualCSE (base) using SupCon and SimCLR as $\mathcal{L}_{\text{image}}$.

4.5 VisualCSE on Different Languages

One of the main advantages of our proposed framework is its ability to generate higher quality sentence embeddings than unsupervised SimCSE, without the need for additional labeled datasets (as opposed to supervised SimCSE, which requires a large, high-quality, and difficult to annotate NLI dataset, see Section 4.6). This is especially important in low-resource languages where NLI datasets are hard to acquire. Here, we show the generalizability of our framework to other languages. Specifically, we experiment on German, French and Russian, using BERT-base models from HuggingFace (see Appendix G). We follow the procedure in SimCSE to construct 1 million sentences from Wikipedia of German, French, Russian and Chinese as training datasets. The evaluations are done on validation sets of the multilingual STS-B from HuggingFace. As shown in Table 5, VisualCSE can improve SimCSE in all three languages by leveraging additional visual supervision, with Russian seeing an improvement of 4.98 in the Spearman's correlation.

Language	**Model**	**Spearman's**
German	SimCSE	67.34
	VisualCSE	**69.87**
Chinese	SimCSE	67.98
	VisualCSE	**70.05**
French	SimCSE	70.31
	VisualCSE	**72.52**
Russian	SimCSE	72.50
	VisualCSE	**77.48**

Table 5: SimCSE and VisualCSE with different languages.

4.6 Supervised Sentence Embedding

Lastly, we investigate whether supervision from other modalities can improve *supervised* sentence embedding. In this setting, the text contrastive loss, $\mathcal{L}_{\text{text}}$, uses entailment pairs as strong positives in the numerator and has an extra term in the denominator as the hard-negative learning signal to push away non-entailment pairs in the NLI dataset (see Eq. 3). Since in our image dataset $\mathcal{S}_{\text{image}}$ (or audio dataset $\mathcal{S}_{\text{audio}}$) we do not have such labeled examples to form the hard-negative pairs, we apply the same unmodified $\mathcal{L}_{\text{image}}$ (or $\mathcal{L}_{\text{audio}}$) to the supervised sentence embedding learning. In Table 6, we observe that Visual/AudioCSE can still outperform supervised SimCSE in all the tasks, but with much smaller margins. This shows that the learning signal from the high-quality negative and positive pairs of the NLI dataset are very strong (leading to a 5.32 improvement over unsupervised SimCSE) and cannot be supplemented by supervision from other modalities.

Model	**STS12**	**STS13**	**STS14**	**STS15**	**STS16**	**STS-B**	**SICK-R**	**Avg.**
Supervised models								
SimCSE-BERT$_{\text{base}}$ ♠	75.30	84.67	80.19	85.40	80.82	84.25	80.39	81.57
VisualCSE-BERT$_{\text{base}}$	75.39	84.69	**80.34**	85.76	**81.60**	**84.85**	80.39	**81.86**
AudioCSE-BERT$_{\text{base}}$	**75.73**	**84.92**	80.25	**85.90**	81.00	84.36	**80.75**	81.84

Table 6: Spearman's of different sentence embedding learning methods in the supervised setting.

5 Analysis

5.1 Visual Supervision Versus NLI Supervision

Given the strong performance of supervised SimCSE with the NLI dataset, natural questions to ask are: (1) How much NLI data can be compensated for through visual supervision of unsupervised VisualCSE? (2) How sensitive is supervised SimCSE to the quality of the labeled dataset used for the hard negatives? Or in other words, at what level of noise in the NLI dataset, does the unsupervised VisualCSE's performance surpass that of supervised SimCSE?

To answer the first question, we down-size the NLI dataset used in supervised SimCSE into different sizes and train SimCSE with these sub-NLI datasets. As shown in Figure 2, our unsupervised VisualCSE works roughly equal to supervised SimCSE with 30K NLI examples.

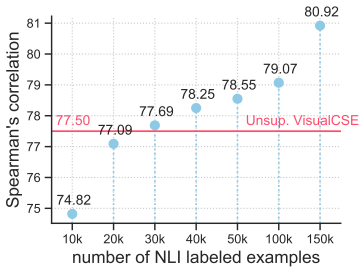

Figure 2: Supervised SimCSE with different subsets of the NLI dataset.

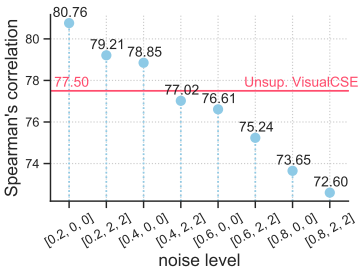

Figure 3: Supervised SimCSE with different levels of noise. E.g., [0.3,2,2] is applying 30% random deletion, and 2 insertions and swaps.

As Gao et al. show, the performance of supervised SimCSE is sensitive to the dataset used for supervision in Eq. 3 [13]. For instance, they show that replacing the NLI dataset with ParaNMT can reduce the performance by 6 points (i.e., degrading to the performance of unsupervised SimCSE). Furthermore, they show that using only contradiction without entailment of the NLI also drops the performance by 7 points (they also show performance drops with other datasets, such as QQP, Flickr30k, and ANLI). This indicates that the performance of contrastive-learning-based supervised models for sentence embedding relies heavily on the high quality of the NLI dataset.

To explore this, we further run experiments on supervised SimCSE with a noisy NLI dataset (at different levels of noise) and observe the point at which our unsupervised VisualCSE starts outperforming

supervised SimCSE. We add noise through deletion, insertion, and swapping. As shown in Figure 3, the performance of supervised SimCSE clearly goes down as the quality of the data is degraded, going below unsupervised VisualCSE at around 40% random deletions.

5.2 Uniformity and Alignment of the Sentence Embeddings

The quality of learned representations can also be measured by *Uniformity* and *Alignment* [53]. *Alignment* is computed based on the expected distance between representations of positive pairs. Lower *Alignment* indicates close representations between positive pairs: $\texttt{Align} = \mathbb{E}_{(x,x^+)\sim p_{\text{pos}}} |f(x) - f(x^+)|^2$. *Uniformity*, on the other hand, measures how well representations are uniformly scattered across the representation space: $\texttt{Uniform} = \mathbb{E}_{(x,y)\sim p_{\text{data}}} e^{-2|f(x)-f(y)|^2}$.

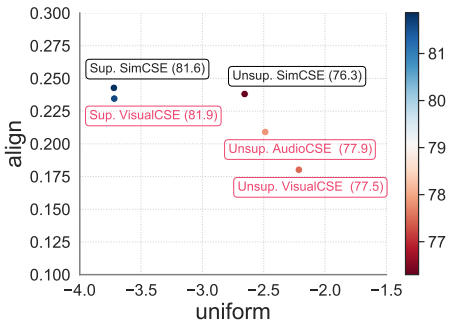

Figure 4: *Alignment* v.s. *Uniformity* of models based on $\text{BERT}_{\text{base}}$ (ours in red).

We show the *Uniformity* and *Alignment* plot of our models evaluated on the STS-B in Figure 4. Our VisualCSE and AudioCSE exhibit better *Alignment* than the text-only SimCSE models, both in supervised and unsupervised settings. However, note that the unsupervised Visual/AudioCSE suffer in *Uniformity*, this is especially true of VisualCSE, which though has better *Alignment* than AudioCSE, has a lower performance due to its worse *Uniformity*. These findings indicate that the non-linguistic supervision in our models make them better clusterers (which is basically what *Alignment* measures).

6 Discussion and Limitations

We have shown that non-linguistic supervision provided by unpaired data from the visual and audio domain can improve the performance of contrastive-learning-based sentence embedding learners in *unsupervised* settings. Though this extra supervision does not fully compensate for the presence of the large high-quality labeled dataset used in *supervised* sentence embedding, we show that it can indeed compensate for smaller, noisier versions. Our findings are important from both a technical point of view and their implications for broader impacts. Our findings indicate that Transformer models can generalize better as contrastive-learners by learning to cluster with *independent* data from different modalities. Future work could investigate whether this behavior is true for tasks other than clustering (e.g., classification). In terms of broader impact, as we have shown, our framework is language-agnostic, enabling it to be deployed for low-resource languages where large, high-quality, and hard-annotated labeled datasets may be hard to obtain. This is possible because our method does not require *paired* datasets (i.e., the same image/audio dataset can be used for any language).

A limitation of our method is the reduction in *Uniformity* of the sentence embeddings. This means that the representations learned may be grouped together in regions of the representation space. Future work could investigate the reason behind this phenomenon and propose methods to alleviate it. Another limitation of our method is the increase in the training time, as it takes twice as long to train our models since we require an additional forward/backward pass for the non-linguistic data.

7 Conclusion

In this paper, we find that sentence representation learning can be improved by contrasting *unpaired* examples from non-linguistics domains, in addition to text. Our framework for incorporating such non-linguistic supervision for sentence embedding is language- and modality-agnostic. Specifically, we show that both visual and audio supervision from unpaired examples can improve unsupervised (and to a lesser extent, supervised) contrastive-learning based sentence representation learners across various languages. Our models, Visual/AudioCSE achieve superior performances in standard sentence embedding benchmarks. Due to its reliance on unpaired data from non-linguistic modalities and its language-agnostic nature, our framework has the potential to be used for low-resource languages.

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[2] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

[3] Lele Cao, Emil Larsson, Vilhelm von Ehrenheim, Dhiana Deva Cavalcanti Rocha, Anna Martin, and Sonja Horn. PAUSE: Positive and annealed unlabeled sentence embedding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10096–10107, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[4] Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*, 2020.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 13–18 Jul 2020.

[6] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[7] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[8] Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States, July 2022. Association for Computational Linguistics.

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[10] William B Dolan and Chris Brockett. Automatically constructing a corpus of sentential paraphrases. In *Proceedings of the Third International Workshop on Paraphrasing (IWP2005)*, 2005.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[12] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Association for Computational Linguistics (ACL)*, 2021.

[13] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[14] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 879–895, Online, August 2021. Association for Computational Linguistics.

[15] Yuan Gong, Cheng-I Jeff Lai, Yu-An Chung, and James Glass. Ssast: Self-supervised audio spectrogram transformer. *arXiv preprint arXiv:2110.09784*, 2021.

[16] Koustava Goswami, Sourav Dutta, Haytham Assem, Theodorus Fransen, and John P. McCrae. Cross-lingual sentence embedding using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9099–9113, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[17] Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. Supervised contrastive learning for pre-trained language model fine-tuning. In *International Conference on Learning Representations*, 2020.

[18] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[19] Olivier Henaff. Data-efficient image recognition with contrastive predictive coding. In *International Conference on Machine Learning*, pages 4182–4192. PMLR, 2020.

[20] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[21] Ronghang Hu and Amanpreet Singh. Transformer is all you need: Multimodal multitask learning with a unified transformer. *CoRR*, abs/2102.10772, 2021.

[22] James Y. Huang, Kuan-Hao Huang, and Kai-Wei Chang. Disentangling semantics and syntax in sentence embeddings with pre-trained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1372–1379, Online, June 2021. Association for Computational Linguistics.

[23] Junjie Huang, Duyu Tang, Wanjun Zhong, Shuai Lu, Linjun Shou, Ming Gong, Daxin Jiang, and Nan Duan. WhiteningBERT: An easy unsupervised sentence embedding approach. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 238–244, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[24] Yiren Jian, Chongyang Gao, and Soroush Vosoughi. Contrastive learning for prompt-based few-shot language learners. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5577–5587, Seattle, United States, July 2022. Association for Computational Linguistics.

[25] Ting Jiang, Shaohan Huang, Zihan Zhang, Deqing Wang, Fuzhen Zhuang, Furu Wei, Haizhen Huang, Liangjie Zhang, and Qi Zhang. Promptbert: Improving bert sentence embeddings with prompts. *arXiv preprint arXiv:2201.04337*, 2022.

[26] Subhradeep Kayal. Unsupervised sentence-embeddings by manifold approximation and projection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1–11, Online, April 2021. Association for Computational Linguistics.

[27] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18661–18673. Curran Associates, Inc., 2020.

[28] Taeuk Kim, Kang Min Yoo, and Sang-goo Lee. Self-guided contrastive learning for BERT sentence representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2528–2540, Online, August 2021. Association for Computational Linguistics.

[29] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021.

[30] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. *Advances in neural information processing systems*, 28, 2015.

[31] A Krizhevsky. Learning multiple layers of features from tiny images. *Master's thesis, University of Tront*, 2009.

[32] Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online, November 2020. Association for Computational Linguistics.

[33] Che Liu, Rui Wang, Jinghua Liu, Jian Sun, Fei Huang, and Luo Si. DialogueCSE: Dialogue-based contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2396–2406, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[34] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10437–10446, 2020.

[35] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 2021.

[36] Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. SLIP: self-supervision meets language-image pre-training. *CoRR*, abs/2112.12750, 2021.

[37] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[38] Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. Unsupervised learning of sentence embeddings using compositional n-gram features. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 528–540, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[39] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[40] Bo Pang and Lillian Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, pages 271–es, 2004.

[41] Bo Pang and Lillian Lee. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 115–124, 2005.

[42] Nina Poerner, Ulli Waltinger, and Hinrich Schütze. Sentence meta-embeddings for unsupervised semantic textual similarity. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7027–7034, Online, July 2020. Association for Computational Linguistics.

[43] QQP. https://www.quora.com/q/quoradata/. In *Quora*, 2022.

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[45] Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online, November 2020. Association for Computational Linguistics.

[46] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[47] Jianlin Su, Jiarun Cao, Weijie Liu, and Yangyiwen Ou. Whitening sentence representations for better semantics and faster retrieval. *arXiv preprint arXiv:2103.15316*, 2021.

[48] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

[49] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 776–794, Cham, 2020. Springer International Publishing.

[50] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. What makes for good views for contrastive learning. In *NeurIPS*, 2020.

[51] Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. DefSent: Sentence embeddings using definition sentences. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 411–418, Online, August 2021. Association for Computational Linguistics.

[52] Kexin Wang, Nils Reimers, and Iryna Gurevych. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[53] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020.

[54] Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7, 2019.

[55] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210, 2005.

[56] John Wieting and Kevin Gimpel. Paranmt-50m: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, 2018.

[57] Adina Williams, Nikita Nangia, and Samuel R Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2018*, pages 1112–1122. Association for Computational Linguistics (ACL), 2018.

[58] Zhirong Wu, Yuanjun Xiong, X Yu Stella, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[59] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*, 2020.

[60] Yuwen Xiong, Mengye Ren, and Raquel Urtasun. Loco: Local contrastive representation learning. *Advances in neural information processing systems*, 33:11142–11153, 2020.

[61] Xu Yang, Chongyang Gao, Hanwang Zhang, and Jianfei Cai. Auto-parsing network for image captioning and visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2197–2207, 2021.

[62] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78, 2014.

[63] Miaoran Zhang, Marius Mosbach, David Adelani, Michael Hedderich, and Dietrich Klakow. MCSE: Multimodal contrastive learning of sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5959–5969, Seattle, United States, July 2022. Association for Computational Linguistics.

[64] Yan Zhang, Ruidan He, Zuozhu Liu, Kwan Hui Lim, and Lidong Bing. An unsupervised sentence embedding method by mutual information maximization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1601–1610, Online, November 2020. Association for Computational Linguistics.

[65] Kun Zhou, Beichen Zhang, Xin Zhao, and Ji-Rong Wen. Debiased contrastive learning of unsupervised sentence representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6120–6130, Dublin, Ireland, May 2022. Association for Computational Linguistics.

A Training Details

We provide hyper-parameters of our models in Table C.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.

Model	GPU	Text Bsz	Text LR	Img/Audio Bsz	Img/Audio LR
VisualCSE-BERT-base	RTX-6000	64	$3e^{-5}$	48	$1e^{-6}$
AudioCSE-BERT-base	RTX-6000	64	$3e^{-5}$	48	$2e^{-7}$
VisualCSE-RoBERTa-base	RTX-6000	128	$2e^{-5}$	48	$5e^{-7}$
AudioCSE-RoBERTa-base	RTX-6000	128	$2e^{-5}$	48	$2e^{-7}$
VisualCSE-RoBERTa-large	RTX-A6000	256	$1e^{-5}$	48	$5e^{-7}$
AudioCSE-RoBERTa-large	RTX-A6000	256	$1e^{-5}$	48	$5e^{-6}$

Table C.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. [15] (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 SupCon contrastive loss is applied at the two output representations $(\mathbf{f}_i', \mathbf{f}_i'')$ at the `[CLS]` position:

$$\mathcal{L}_{\text{audio}}^{\text{SupCon}} = \sum_{i=1}^{N} -\log \frac{e^{\text{sim}(\mathbf{f}_i', \mathbf{f}_i'')/\tau} + \sum\limits_{y_i \text{ and } y_j \text{ from same class}} e^{\text{sim}(\mathbf{f}_i', \mathbf{f}_j'')/\tau}}{\sum\limits_{y_i \text{ and } y_j \text{ from different class}} e^{\text{sim}(\mathbf{f}_i', \mathbf{f}_j'')/\tau}} \tag{6}$$

Or, the SimCLR loss equivalent:

$$\mathcal{L}_{\text{audio}}^{\text{SimCLR}} = \sum_{i=1}^{N} -\log \frac{e^{\text{sim}(\mathbf{f}_i', \mathbf{f}_i'')/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(\mathbf{f}_i', \mathbf{f}_j'')/\tau}} \tag{7}$$

Same as with images, the audio dataset is not paired with the text data and $\mathcal{L}_{\text{audio}}$ 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 $\mathcal{L}_{\text{audio}}$, we use `train-clean-100` split from the LibriSpeech dataset [39]. We use a subset of LibriSpeech $\mathcal{S}_{\text{audio}}$ 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 D.1, VisualCSE generally retrieves qualitatively different sentences than SimCSE.

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

Table D.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 [41], CR [20], SUBJ [40], MPQA [55], SST-2 [46], TREC [54], and MRPC [10] 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 E.1 shows that VisualCSE and AudioCSE generally outperfom 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.

Model	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	Avg.
Unsupervised models								
SimCSE-BERT$_{\text{base}}$ ♠	81.18	86.46	94.43	88.87	85.50	89.80	**74.49**	85.82
VisualCSE-BERT$_{\text{base}}$	**82.29**	**87.18**	**94.85**	**89.34**	**86.55**	**89.60**	74.32	**86.30**
AudioCSE-BERT$_{\text{base}}$	81.10	86.76	94.65	88.78	86.00	89.40	74.43	85.87
SimCSE-RoBERTa$_{\text{base}}$ ♠	81.04	**87.74**	93.28	86.94	86.60	84.60	73.68	84.84
VisualCSE-RoBERTa$_{\text{base}}$	**81.84**	87.60	92.76	87.14	86.77	85.20	74.43	85.11
AudioCSE-RoBERTa$_{\text{base}}$	80.76	87.18	**93.39**	**87.29**	**87.64**	**86.60**	**75.07**	**85.42**
SimCSE-RoBERTa$_{\text{large}}$ ♠	82.74	87.87	93.66	88.22	88.58	92.00	69.68	86.11
VisualCSE-RoBERTa$_{\text{large}}$	83.44	88.13	**93.96**	88.43	**88.63**	91.60	**73.86**	**86.86**
AudioCSE-RoBERTa$_{\text{large}}$	**83.49**	**88.64**	93.86	**88.82**	**88.63**	**92.40**	69.97	86.54

Table E.1: Testing accuracy on transfer tasks using different models. ♠: results from SimCSE's publicly available model.

E Ablation on $\mathcal{L}_{\text{audio}}$

Following what was done in Section 4.4, we study AudioCSE with two different contrastive losses (SimCLR and SupCon). The results in Table G.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.

	STS12	STS13	STS14	STS15	STS16	STS-B	SICK-R	Avg.
BERT (SimCLR)	70.98	83.03	75.74	83.03	78.10	78.33	69.74	77.00
BERT (SupCon)	71.65	84.27	76.69	83.22	78.69	79.94	70.49	77.85
RoBERTa (SimCLR)	68.88	83.74	74.69	82.46	82.02	81.52	70.65	77.71
RoBERTa (SupCon)	68.44	83.96	75.77	82.38	82.07	81.63	70.56	77.83

Table G.1: Comparison of AudioCSE (base) using SupCon and SimCLR as $\mathcal{L}_{\text{image}}$.

F VisualCSE with CIFAR images

We further carry out experiments on substituting ImageNet with a lower quality CIFAR [31] 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 H.1 our framework improves over SimCSE even with this lower quality dataset (results shown below).

Model	Avg.
SimCSE-BERT$_{\text{base}}$	76.25
VisualCSE-BERT$_{\text{base}}$ (CIFAR)	76.96
VisualCSE-BERT$_{\text{base}}$ (ImageNet)	77.50
SimCSE-RoBERTa$_{\text{base}}$	76.57
VisualCSE-RoBERTa$_{\text{base}}$ (CIFAR)	77.71
VisualCSE-RoBERTa$_{\text{base}}$ (ImageNet)	77.87

Table H.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`.