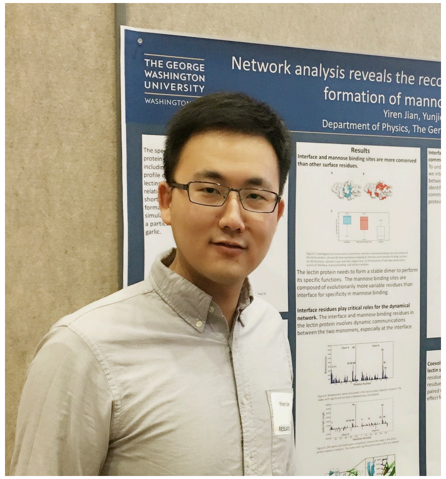
Non-Linguistic Supervision for Contrastive Learning of Sentence Embeddings

Yiren Jian


THE GEORGE WASHINGTON UNIVERSITY
Network analysis reveals the reco
formation of mann
Yiren Jian, Yunji
Department of Physics, The Ge
Results

Chongyang Gao*

Soroush Vosoughi

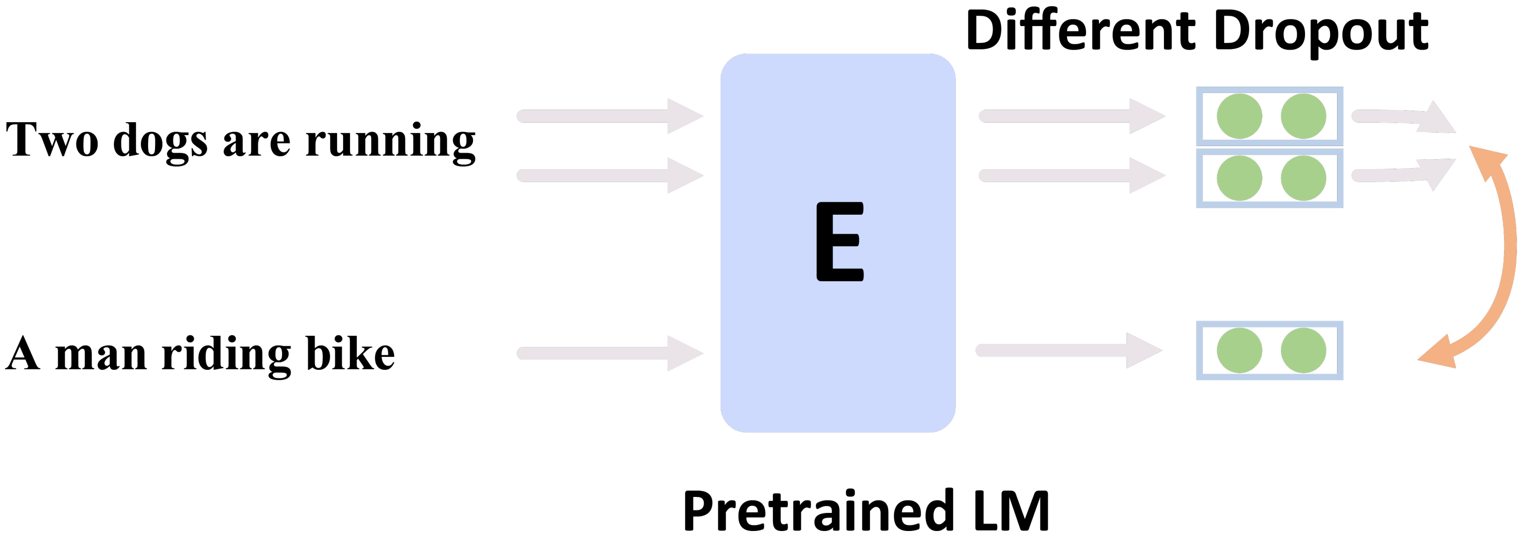
DARTMOUTH
Department of Computer Science

Northwestern | McCORMICK SCHOOL OF ENGINEERING
Computer Science

Background: Sentence Embedding Learning

Goal: Semantic similar sentences should have “close” embeddings

Solution: Contrastive learning (SimCSE)



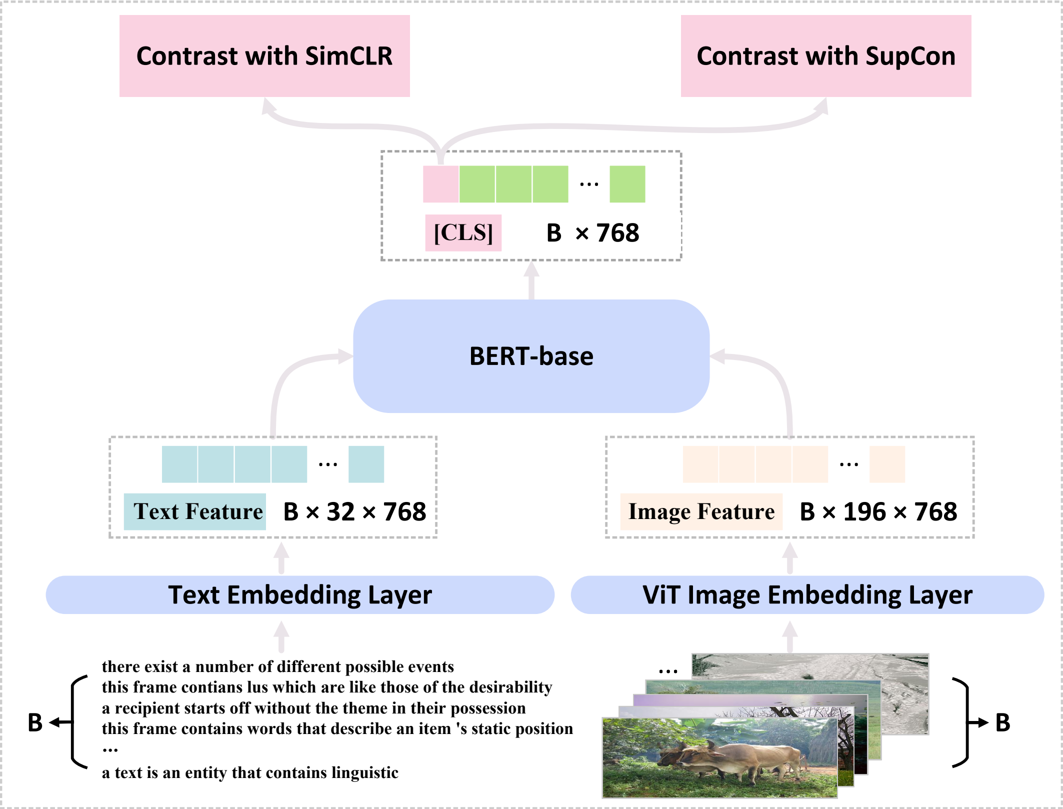
Gao T, Yao X, Chen D. Simcse: Simple contrastive learning of sentence embeddings[J]. arXiv preprint arXiv:2104.08821, 2021.

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

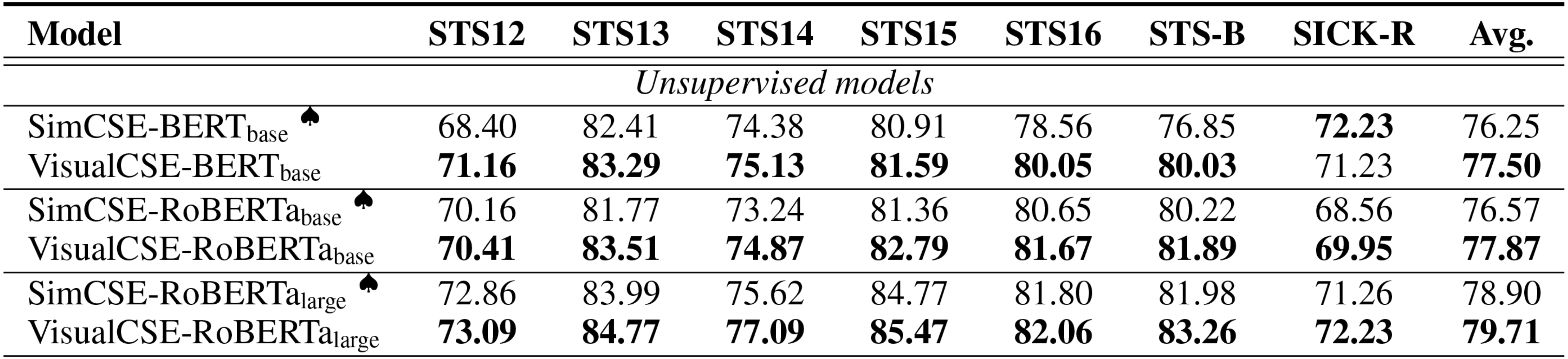


there exist a number of different possible events
this frame contians lus which are like those of the desirability
a recipient starts off without the theme in their possession
this frame contains words that describe an item 's static position
...
a text is an entity that contains linguistic

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

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

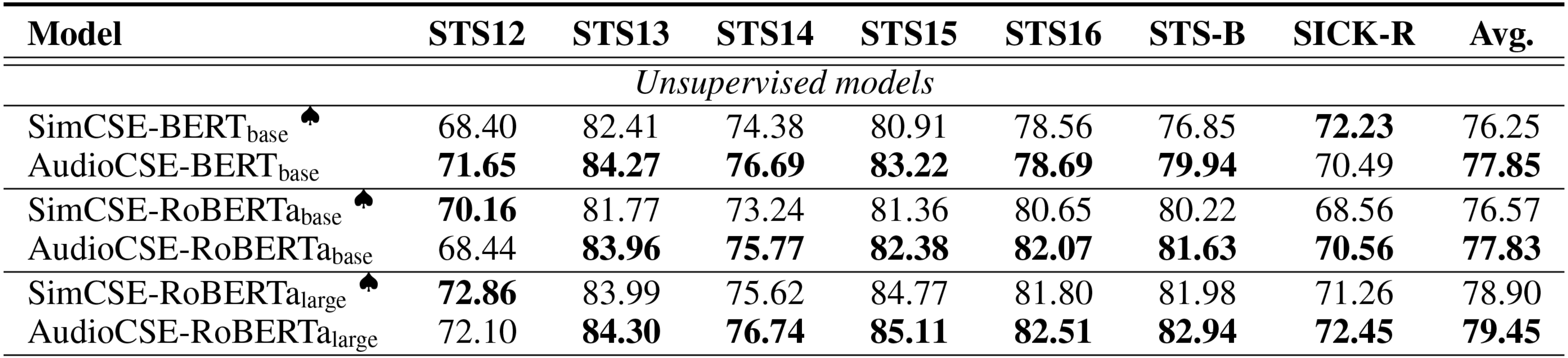
Results of VisualCSE



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**

Results of AudioCSE

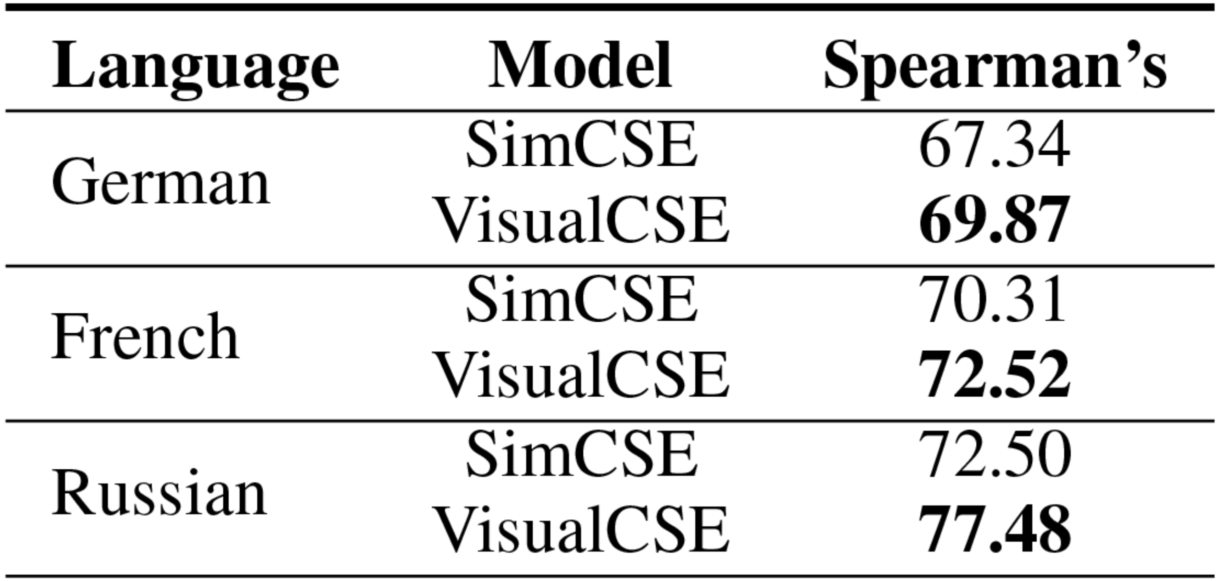
Replacing images with audios



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**

Evaluating on other languages

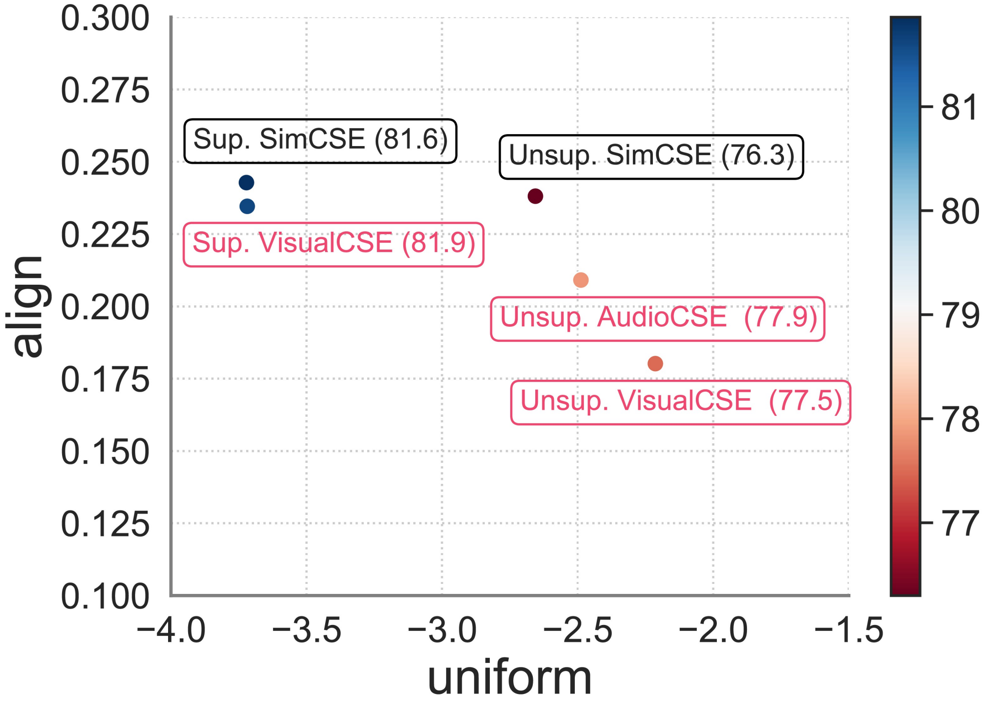
A key advantage of our Non-linguistic CSE is that it does not require aligned (paired) examples, allowing us to apply them to different languages.



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.