CINCE: Enhanced Sentence Meta-Embeddings for Textual Understanding

Sourav Dutta & Haytham Assem

Huawei Research, Ireland

sourav.dutta2@huawei.com

Abstract

*We propose **CINCE**, a novel sentence meta-embedding framework for* aggregating, *in a principled way*, various semantic information captured by different embedding architectures.

Introduction

- **Sentence embedding** provide dense vector representation capturing the overall contextual and semantic meaning of text.
- Diverse sentence embedding techniques captures different semantic aspects and linguistic features.
- ***"Meta-Embedding"*** study intelligent combination of sentence embedding vector space from multiple pre-trained architectures.

Contributions

- *Canonicalized Independent Component based Embeddings* (**CINCE**), a principled and effective novel sentence meta-embedding framework.
- Amalgamates different aspects of semantic information captured by diverse embedding techniques, by using *component analysis*.
- Experiments on *SentEval benchmark* tasks demonstrate the effectiveness of our framework.

Assumption: Different embedding techniques capture possibly different linguistic aspects in possibly varying degrees.

Apply Principal Component Analysis (PCA) on each embedding space

- Identify dimensions containing important information.
- Project on dimensions to eliminate non-informative aspects.

Conduct Independent Component Analysis (ICA) on each project embedding dimension

- Extract unique informative cues from different embedding.

Perform Canonical Correlation Analysis (GCCA) on independent components

- Provide a common space to project the different source embeddings.
- Group unique information captured by various encoding components.

Empirical Results

- The "*canonicalized independent components*" together form the meta-embedding.
- Using *SBERT, LASER and USE* embedding, we compare against *independent, averaging, concatenation and auto-encoding strategies*.
- On the *SentEval STS task*, CINCE achieves ***the best Pearson's correlation*** score (4% improvement) with comparable Spearman's score.
- For *the supervised classification tasks*, CINCE also **performs better** than the baselines on nearly all the datasets.

Table 1. Mean Pearson's and Spearman's Correlation (ρ) scores achieved by the approaches on the STS tasks of SentEval benchmark. (Best results are marked in **bold**, second-best results are underlined, and * indicates statistically significant results using paired bootstrap resampling).

Benchmark / Approaches	*dim*	STS-12	STS-13	STS-14	STS-15	STS-16	STS-B
		(mean Pearson's $r \times 100$ / Spearman's $\rho \times 100$)					
SBERT	512	68.84/66.98	68.97/70.00	75.77/73.27	81.30/82.05	78.65/80.02	81.83/81.91
USE	512	69.58/68.15	69.16/70.35	76.56/73.51	80.69/81.58	78.81/79.73	81.58/81.52
LASER	1024	62.90/62.30	48.69/51.61	67.83/67.04	75.03/75.38	71.78/72.32	78.15/78.11
Average	1024	70.99/69.06	68.74/69.43	77.27/74.78	81.97/82.60	79.77/80.78	82.80/82.66
Concatenation	2048	71.29/**69.38**	68.85/69.76	77.65/75.10	82.56/**83.17**	80.08/**81.10**	83.63/**83.77***
Auto-Encoder	1024	67.65/67.22	61.19/63.99	74.18/72.27	80.63/81.56	78.27/79.62	80.88/81.22
GCCA	1024	68.41/68.09	65.26/66.64	77.32/75.93	79.01/80.00	76.63/80.22	79.52/79.61
CINCE	1024	**71.56**/69.23	**75.35*/75.48***	**79.26*/76.50***	**82.66**/82.71	**80.12**/80.92	**83.91**/82.06

Table 2. SentEval classification task accuracy.

Tasks / Methods	CR	MR	MPQA	TREC	SUBJ	SST-5
SBERT	82.22	74.25	86.97	91.60	91.67	44.75
USE	79.66	71.50	86.44	92.40	90.91	42.53
LASER	82.01	74.50	87.95	92.40	91.94	45.52
Average	81.11	74.82	88.40	93.40	92.47	44.48
Concatenation	83.21	**76.09**	**88.60**	**95.00**	92.93	46.79
Auto-Encoder	80.76	72.96	87.86	91.40	91.59	44.43
GCCA	78.15	74.68	87.86	94.60	92.65	44.62
CINCE	**83.28**	75.62	**88.60**	**95.00**	**93.04**	**46.92**