



1. Highlight

Background:

- Words are represented as numerical vectors [1]
- Sentences are composed of variant-length word tokens
- VLAD algorithm in image processing [2,3]
- Goal:** Encode sentence into fixed size vector

Motivations:

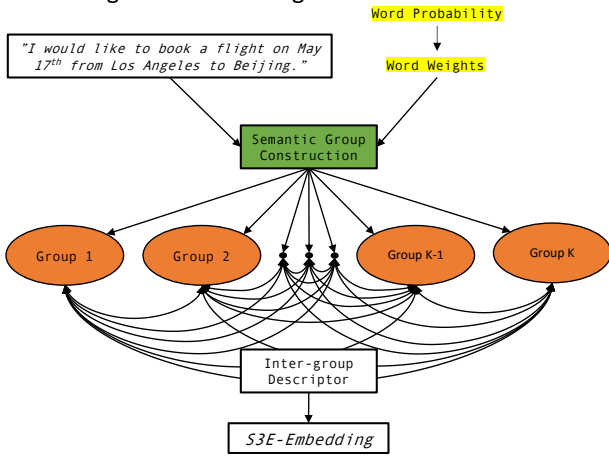
- Efficiency: encoder large amounts of sentences
- Semantic grouping property of word embedding

2. Methodology

Step 1: Semantic Group Construction

- Leveraging Semantic grouping property of word vectors
- Handling variant-length input
- Weighted k-means algorithm

$$\text{weight}(w) = \frac{\epsilon}{\epsilon + p(w)}$$



Step 2: Intra-group Descriptor

- Assign words into semantic groups
- Compute group center

$$g_i = \frac{1}{|G_i|} \sum_{w \in G_i} \text{weight}(w) v_w$$

- Find discriminate representation for current group

$$v_i = \sum_{w \in S \cap G_i} \text{weight}(w) (v_w - g_i)$$

- Matrix representation of a sentence

$$\Phi(S) = \begin{bmatrix} v_1^T \\ v_2^T \\ \vdots \\ v_K^T \end{bmatrix} = \begin{pmatrix} v_{11} & \dots & v_{1d} \\ v_{21} & \dots & v_{2d} \\ \vdots & \ddots & \vdots \\ v_{K1} & \dots & v_{Kd} \end{pmatrix}_{K \times d}$$

Step 3: Inter-group Descriptor

- Model interaction between groups as sentence representation

$$C = [C_{i,j}]_{K \times K} = \frac{1}{d} (\Phi - \mu_\Phi)(\Phi - \mu_\Phi)^T \in \mathbb{R}^{K \times K}$$

$$C = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1K} \\ \sigma_{12} & \sigma_{22}^2 & \dots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1K} & \sigma_{2K} & \dots & \sigma_{KK}^2 \end{pmatrix} \quad C_{i,j} = \sigma_{i,j} = \frac{(v_i - \mu_i)^T (v_j - \mu_j)}{d}$$

- Vectorization

$$v(S) = \text{vect}(C) = \begin{cases} \sqrt{2}\sigma_{ij}, & \text{if } i < j, \\ \sigma_{ii}, & \text{if } i = j. \end{cases}$$

3. Application and Analysis

- Sentence Similarity

Model	Dim	STS12	STS13	STS14	STS15	STS16	STSB	SICK-R	Avg.
Parameterized models									
skip-thought [5]	4800	30.8	24.8	31.4	31.0	-	-	86.0	40.80
InferSent [6]	4096	58.6	51.5	67.8	68.3	70.4	74.7	88.3	68.51
ELMo [23]	3072	55.0	51.0	63.0	69.0	64.0	65.0	84.0	64.43
Avg. BERT [24]	768	46.9	52.8	57.2	63.5	64.5	65.2	80.5	61.51
SBERT-WK [10]	768	70.2	68.1	75.5	76.9	74.5	80.0	87.4	76.09
Non-parameterized models									
Avg. GloVe	300	52.3	50.5	55.2	56.7	54.9	65.8	80.0	59.34
SIF [11]	300	56.2	56.6	68.5	71.7	-	72.0	86.0	68.50
p-mean [14]	3600	54.0	52.0	63.0	66.0	67.0	-	86.0	65.71
S3E (GloVe)	355-1575	59.5	62.4	68.5	72.3	70.9	75.5	82.7	69.59
S3E (FastText)	355-1575	62.5	67.8	70.2	76.1	74.3	77.5	84.7	72.64
S3E (L.F.P.)	955-2175	61.0	69.3	73.2	76.1	74.4	78.6	84.7	73.90

- Cosine similarity between sentences

- Classification Tasks

Model	Dim	MR	CR	SUBJ	MPQA	SST	TREC	MRPC	SICK-E	Avg.
Parameterized models										
skip-thought [5]	4800	76.6	81.0	93.3	87.1	81.8	91.0	73.2	84.3	83.54
FastSent [22]	300	70.8	78.4	88.7	80.6	-	76.8	72.2	-	77.92
InferSent [6]	4096	79.3	85.5	92.3	90.0	83.2	87.6	75.5	85.1	84.81
Sent2Vec [21]	700	75.8	80.3	91.1	85.9	-	86.4	72.5	-	82.00
USE [7]	512	80.2	86.0	93.7	87.0	86.1	93.8	72.3	83.3	85.30
ELMo [23]	3072	80.9	84.0	94.6	91.0	86.7	93.6	72.9	82.4	85.76
SBERT-WK [10]	768	83.0	89.1	95.2	90.6	89.2	93.2	77.4	85.5	87.90
Non-parameterized models										
GloVe(Ave)	300	77.6	78.5	91.5	87.9	79.8	83.6	72.1	79.0	81.25
SIF [11]	300	77.3	78.6	90.5	87.0	82.2	78.0	-	84.6	82.60
p-mean [14]	3600	78.3	80.8	92.6	89.1	84.0	88.4	73.2	83.5	83.74
DCT [15]	300-1800	78.5	80.1	92.8	88.4	83.7	89.8	75.0	80.6	83.61
VLAWE [18]	3000	77.7	79.2	91.7	88.1	80.8	87.0	72.8	81.2	82.31
S3E (GloVe)	355-1575	78.3	80.4	92.5	89.4	82.0	88.2	74.9	82.0	83.46
S3E (FastText)	355-1575	78.8	81.4	92.9	88.5	83.5	87.0	75.7	81.4	83.65
S3E(L.F.P.)	955-2175	79.4	81.4	92.9	89.4	83.5	89.0	75.6	82.6	84.23

- Cosine similarity between sentences

Complexity Analysis

Model	CPU inference time (ms)
InferSent	53.07
SBERT-WK	179.27
GEM	26.54
SIF	1.56
Proposed S3E	0.69

- Word clusters can be pre-computed
- Low time complexity with CPU
- Suitable for large scale inference (comparing with deep learning models)

4. Conclusions and Future Work

- S3E (ours) is very competitive among non-parameterized sentence embedding models
- Low time complexity

Future Work:

- With modularized design of S3E, we can try stronger clustering and correlation descriptors including subspace clustering, non-linear correlation computation with different kernel functions

Bibliography:

- [1]: Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." NIPS. 2013.
- [2]: Ionescu, Radu Tudor, and Andrei M. Butnaru. "Vector of locally-aggregated word embeddings (VLAWE): A novel document-level representation." NAACL, 2019
- [3]: Arandjelovic, Relja, and Andrew Zisserman. "All about VLAD." CVPR. 2013.

Code: <https://github.com/BinWang28/Sentence-Embedding-S3E>