Continual Learning for Sentence Representations Using Conceptors

Tianlin Liu*, Lyle Ungar♯, João Sedoc♯

*Department of Computer Science and Electrical Engineering, Jacobs University Bremen, Bremen, Germany.
♯Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA.

Abstract

We consider a **continual sentence representation learning task**: Given a collection of corpora presented sequentially, how to train the sentence encoder with respect to the new corpus while maintaining its accuracy on the old corpora? To address this problem, we propose sentence encoders with the following desiderata:

1. **Zero-shot learning**. The initialized sentence encoder (no training corpus used) can effectively produce sentence embeddings.
2. **Resistant to catastrophic forgetting**. When the sentence encoder is adapted on a new training corpus, it retains strong performances on old ones.

Introduction

Traditional Sentence encoders:

- Trained on some *a priori* fixed corpora.
- When the trained encoder is adapted on a new training corpus (which may have very different word distributions than the old ones), it performs bad under old ones.

These traditional encoders are not suitable for open-domain NLP systems, where the environment is dynamic, training data are accumulated sequentially over time, and the distributions of training data vary with respect to external input

Our sentence encoder:

- Initialized without any corpus
- Sequentially update when a new corpus is available.
- Preserve useful features from old training corpora.

Relevant work: linear sentence encoders

Linear sentence encoders (e.g., SIF encoder [1]) usually contain two steps:

1. Do weighted sum over a collection of word vectors.
2. Remove some special directions ("common discourse features") from the weighted sum.

Algorithm 1: SIF sentence encoder.

Input : A training corpus D; a testing corpus G; parameter a, monogram probabilities $\{p(w)\}_{w\in V}$ of words

for *sentence* $s \in D$ **do**

$q_s \leftarrow \frac{1}{|s|}\sum_{w\in s}\frac{a}{p(w)+a}v_w$

end

Let u be the first singular vector of $[q_s]_{s\in D}$.

for *sentence* $s \in G$ **do**

$q_s \leftarrow \frac{1}{|s|}\sum_{w\in s}\frac{a}{p(w)+a}v_w$

$f_s^{\mathrm{SIF}} \leftarrow q_s - uu^\top q_s$.

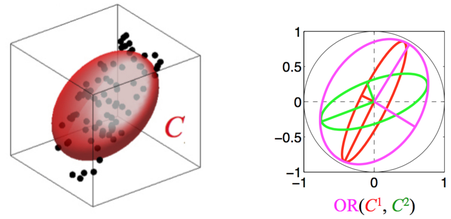
end

Output: $\{f_s^{\mathrm{SIF}}\}_{s\in G}$

Our question: What if the "common discourse features" are varying over a sequence of training corpora?

Our solution: Use conceptors [2] to dynamically characterize and update the common discourse features.

Conceptors



(a) (b)

Figure 1: (a): The red conceptor characterizes linear subspace occupied by point clouds. (b): OR operation on conceptors.

The conceptor, C, is a soft projection matrix on the linear subspace where the samples of x lie. From a set of n-dimensional points $\{x_i\}_{i\in I}$, C is defined as:

$$\arg\min_C \frac{1}{|I|}\sum_{i\in I}||x_i - Cx_i||^2 + \alpha^{-2}||C||_{\mathrm{F}}^2 \tag{1}$$

where α is a hyperparameter and $||\cdot||_{\mathrm{F}}$ is the Frobenius norm, $|I|$ is the cardinality of I. This optimization problem has a closed-form solution

$$C = R(R+\alpha^{-2}\mathbf{Id})^{-1} \quad \text{where} \quad R = \frac{1}{|I|}XX^\top \tag{2}$$

where X is a matrix with x_i as columns, and $\mathbf{Id}$ is the $n\times n$ identity matrix. Relationship between singular values of R and C:

$$\begin{aligned} R &= U\Sigma U^\top & C &= USU^\top \\ \Sigma &= \begin{bmatrix}\sigma_1 & & \\ & \ddots & \\ & & \sigma_n\end{bmatrix} & S &= \begin{bmatrix}s_1 & & \\ & \ddots & \\ & & s_n\end{bmatrix} \\ s_i &= \sigma_i/(\sigma_i+\alpha^{-2}) \in [0,1) \end{aligned} \tag{3}$$

Notation: we write $C(\{x_i\},\alpha)$ to stress the dependence on $\{x_i\}$ and α.

Proposed Conceptor-aided (CA) sentence encoder

The general idea of CA encoder is the following:

1. Initialize the common discourse features using a set of stop words.
2. Sequentially update common discourse features using OR operation of conceptors.

Algorithm 2: CA sentence encoder.

Input : A sequence of M training corpora $\mathcal{D} = \{D^1,\cdots,D^M\}$; a testing corpus G; hyper-parameters a and α; word probabilities $\{p(w)\}_{w\in V}$; stop word list Z.

$C^0 \leftarrow C([v_w]_{w\in Z},\alpha)$.

for *corpus index* $i = 1,\cdots,M$ **do**

for *sentence* $s \in D^i$ **do**

$q_s \leftarrow \frac{1}{|s|}\sum_{w\in s}\frac{a}{p(w)+a}v_w$

end

$C^{\mathrm{temp}} \leftarrow C([q_s]_{s\in D^i},\alpha)$

$C^i \leftarrow C^{\mathrm{temp}} \vee C^{i-1}$

end

for $s \in G$ **do**

$q_s \leftarrow \frac{1}{|s|}\sum_{w\in s}\frac{a}{p(w)+a}v_w$

$f_s^{\mathrm{CA}} \leftarrow q_s - C^M q_s$

end

Output: $\{f_s^{\mathrm{CA}}\}_{s\in G}$

Proposed Zero-shot CA sentence encoder

Only use **stop words** as features for common discourse – **No training corpus used!**

Algorithm 3: Zero-shot CA sentence encoder.

Input : A testing corpus G; hyper-parameters a and α; word probabilities $\{p(w)\}_{w\in V}$; stop word list Z.

$C^0 \leftarrow C([v_w]_{w\in Z},\alpha)$.

for $s \in G$ **do**

$q_s \leftarrow \frac{1}{|s|}\sum_{w\in s}\frac{a}{p(w)+a}v_w$

$f_s^{\mathrm{CA}} \leftarrow q_s - C^0 q_s$

end

Output: $\{f_s^{\mathrm{CA}}\}_{s\in G}$

Experiment

Dataset: Semantic textual similarity (STS) datasets splited into five corpora by their genre: news, captions, wordnet, forums, tweets.

Evaluation criterion: Pearson correlation coefficient (PCC) between the predicted sentence similarities and the ground-truth sentence similarities.

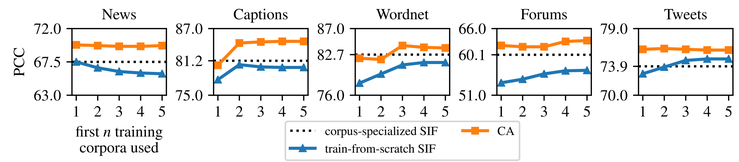


Figure 2: PCC results of STS datasets. Each panel shows the PCC results of a testing corpus (specified as a subtitle) as a function of increasing numbers of training corpora used. The setup of this experiment mimics [3, section 5.1].

	News	Captions	WordNet	Forums	Tweets
av. train-from-scratch SIF	66.5	79.7	80.3	55.5	74.2
zero-shot CA	65.6	79.8	82.5	61.5	75.2
av. CA	**69.7**	**83.8**	**83.2**	**62.5**	**76.2**

Table 1: Time-course averaged PCC of train-from-scratch SIF and conceptor-aided (CA) methods, together with the result of zero-shot CA. Best results are in boldface and the second best results are underscored.

References

[1] S. Arora, Y. Liang, and T. Ma.
A simple but tough-to-beat baseline for sentence embeddings.
In *International Conference on Learning Representations*, 2017.

[2] H. Jaeger.
Using conceptors to manage neural long-term memories for temporal patterns.
Journal of Machine Learning Research, 18(13):1–43, 2017.

[3] F. Zenke, B. Poole, and S. Ganguli.
Continual learning through synaptic intelligence.
In *International Conference on Machine Learning*, 2017.

Full paper

Codes on Github