ISCAS 中国科学院软件研究所学术年会’2021 暨计算机科学国家重点实验室开放周 学术论文

ICANN 2021

Short Text Clustering with A Deep Multi-Embedded Self-Supervised Model

Kai Zhang, Zheng Lian, Jiangmeng Li, Haichang Li, Xiaohui Hu
Institute of Software Chinese Academy of Sciences, Beijing, 100190, China
Phone: 13981924856 Email: zhangkai2020c@iscas.ac.cn

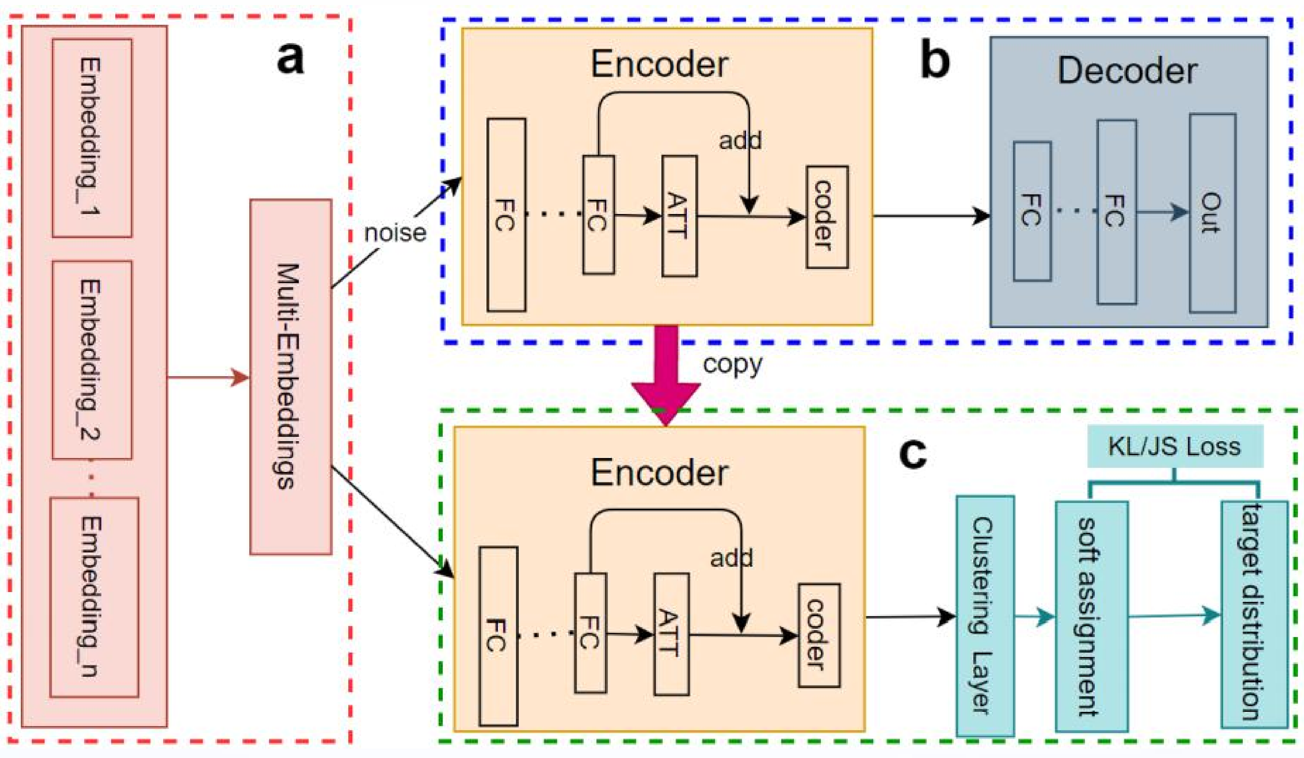
Abstract

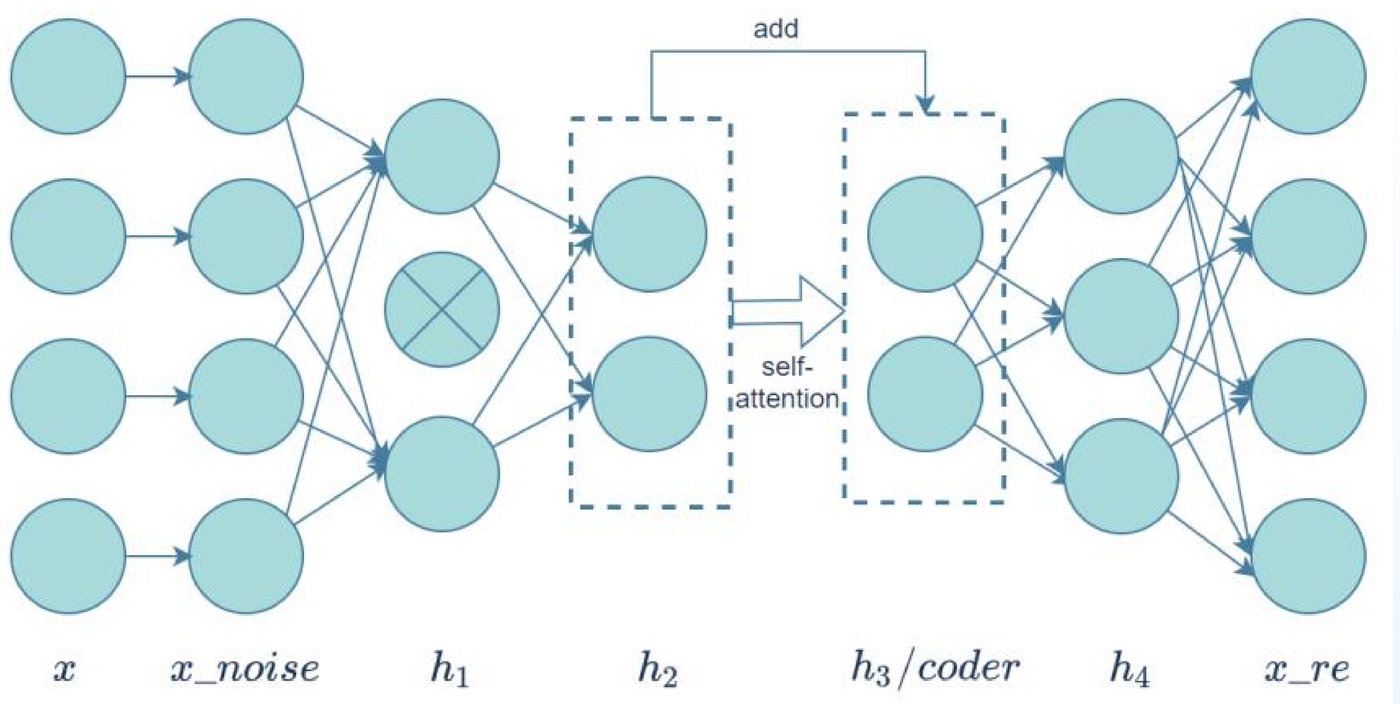
Short text clustering is challenging in NLP. In this paper, fused multi-embedded features are employed. Then, a denosing autoencoder(DAE) with an attention layer is adopted to extract low-dimensional features. Furthermore, we propose a novel distribution estimation to better fine-tune the encoder. Combining the above work, we propose a deep multi-embedded self-supervised model (DMESSM). Our method outperforms the state-of-the-art methods on **4** benchmark datasets.

Introduction

- Traditional clustering algorithm such as Kmeans can be applied on short text vector representations. Besides, topic models and neural networks are recently widely used in short text clustering.
- We focus on neural networks. **Static embedding and dynamic embedding** are fused to express short texts better. We **add an attention block** on DAE, thus the model can output the important low-dimensional features. We propose **a new target distribution** which can better preserve the order of soft assignment than before and enhance the clustering.

Methods





- (a)**Combine** many different embeddings into a multi-embeddings to express short texts.
- (b)**Pretrain** a denoising autoencoder with an attention block.
- (c)Copy the encoder and do **self-supervised** clustering.

Given a multi-embeddings *x*, we add White Gaussian Noise to get *x_noise* as the input. The encoder which includes **a FNN and a self-attention layer** maps *x_noise* to a low-dimensional representation *coder*. Then the decoder reconstructs an input *x_re*. We choose the MSE as the loss function.

Self-Supervised Iterative Clustering

First, we compute a soft assignment between the embedded points and cluster centroids.

$$q_{ij} = \frac{\left(1+\|z_i-u_j\|^2\right)^{-1}}{\sum_j\left(1+\|z_i-u_j\|^2\right)^{-1}}$$

Second, the adjust fuction g and target distribution p are:

$$g(q) = \frac{\sqrt[3]{2q-1}+1}{2}$$

$$p_{ij} = \frac{g^2(q_{ij})/\sum_{i'} q_{i'j}}{\sum_{j'}\left(g^2(q_{ij'})/\sum_{i'} q_{i'j'}\right)}$$

Third, construct a loss function between two distributions using KL or JS divergence:

$$KL(P||Q) = \sum_i\sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

Results

method	Stackoverflow		SearchSnippets		Tweet89		20ngnews	
	ACC	NMI	ACC	NMI	ACC	NMI	ACC	NMI
TF	13.5±2.2	7.8±2.5	24.7±2.2	9.0±2.3	52.9±1.3	70.3±1.7	13.1±2.1	7.4±1.2
TF-IDF	20.3±4.0	15.6±4.7	33.8±3.9	21.4±4.4	53.1±2.3	76.2±3.4	20.7±2.4	18.8±1.6
Word2vec	38.1±2.4	36.5±1.5	67.5±0.1	51.5±0.1	48.9±0.9	77.2±1.5	28.1±0.2	28.4±0.8
SIF	48.5±1.3	45.8±1.6	66.8±0.2	50.6±0.1	49.1±1.2	76.8±0.7	29.1±0.6	30.2±0.7
SBERT	63.2±2.5	60.5±2.2	67.2±0.5	48.5±0.5	51.8±0.7	80.1±1.0	31.3±1.1	31.5±0.6
STCC	51.1±2.9	49.0±1.5	77.0±4.1	62.9±1.7	-	-	-	-
SIF-Auto.	59.8±1.9	54.8±1.0	77.1±1.1	56.7±1.0	54.5±3.3	74.6±3.2	28.2±1.8	28.6±1.3
DMESSM	**79.9±0.3**	**70.7±0.2**	**83.3±0.2**	**65.0±0.2**	**77.3±2.2**	**85.8±2.5**	**38.5±0.6**	**37.7±0.4**

Our model has achieved **the best results** on datasets of different sizes and categories, showing its superiority.

Conclusion

- Our model DMESSM starts from an unsupervised method using SIF and SBERT, then does iterative clustering by using a denoising autoencoder and a clustering layer.
- We improve the target distribution of short text clustering.
- The experimental study shows that our model can reach the most advanced level on multiple datasets.