[Re] Autoencoding Keyword Correlation Graph for Document Clustering

Alireza Torabian
torabian@eecs.yorku.ca

April 2021

Paper

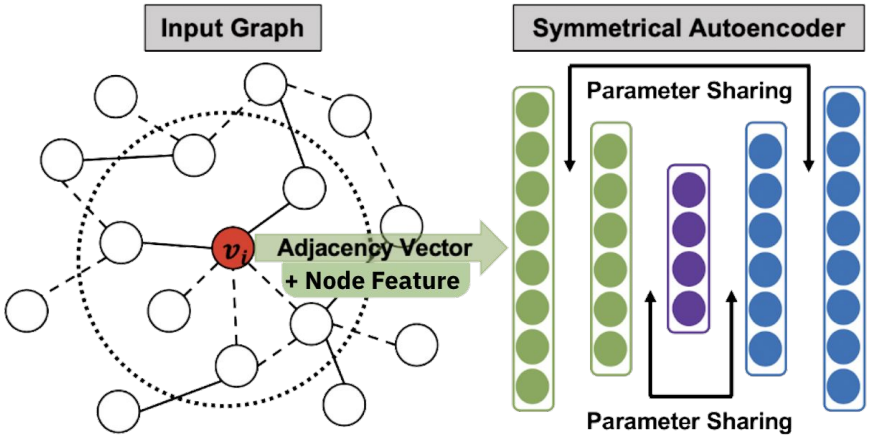
- Cluster documents using both local and global features of sentences
- Based on a graph-based representation of documents
- In three main steps

1. Keyword Correlation Graph

1. Extracting top-50 keywords to be used as the graph nodes
2. Assigning each sentence of each document to the most related keyword
3. Node feature equals to the average of the related sentences' embeddings
4. Edge weights:
    i. Sentence Similarity
    ii. Sentence Position
    iii. Word Co-occurrence

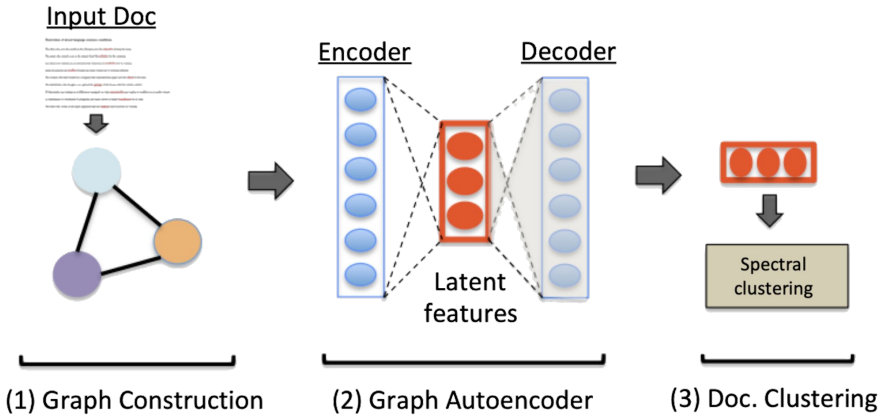
2. Graph Autoencoder

- Using Multi-Task Graph Autoencoder (MTGAE)
- Reconstruct both graph and node features using the decoder part
- Transforming each node to the latent space



3. Clustering

- Embedding of each document is the global average pooling of its related nodes' latent feature
- Clustering documents using their embeddings
- By Spectral Clustering



Scope of reproducibility

> The best results they have achieved using:

- Sentence Similarity for edge weight
- SBERT for sentence embedding
- MTGAE for graph autoencoding

Methodology

- Implementing from scratch
- Evaluation after each main step
  - What proportion of documents are not assigned to the KCG
  - Evaluating the MTGAE by splitting the KCG nodes into train and test sets
- Evaluating and comparing the clustering model using different setups
- Constructing the KCG is the most compute intensive part
- Using a hardware system with eight vCPUs and 32 GB memory provided by Google Cloud

Results - Default model

> Reproducing the results in the paper was not successful

		ACC	AMI
Reuters	Paper result	0.563	0.584
	Our result	0.352	0.125
20 Newsgroups	Paper result	0.474	0.530
	Our result	0.108	0.055

Table 1: Comparing results with the paper report

Results - Different clustering method

> Using different clustering method did not have much effect

		ACC	AMI
Reuters	Spectral Clustering	0.352	0.125
	K-Means	0.339	0.133
20 Newsgroups	Spectral Clustering	0.108	0.055
	K-Means	0.110	0.060

Table 2: Results using different clustering methods

Results - Larger graph

> Using a larger keyword correlation graph with 70 nodes

		ACC	AMI
Reuters	Spectral Clustering	0.316	0.139
	K-Means	0.302	0.200
20 Newsgroups	Spectral Clustering	0.112	0.071
	K-Means	0.119	0.075

Table 3: Results using different clustering methods on big graphs

Stretch Goal - Deep clustering

- Using a deep clustering model in the third part
- A 4-layer autoencoder to which a clustering layer is attached

			ACC	AMI
Reuters	Small graph	Spectral Clustering	0.352	0.125
		Deep Clustering	**0.386**	**0.044**
	Big graph	Spectral Clustering	0.316	0.139
		Deep Clustering	**0.326**	**0.066**
20 Newsgroups	Small graph	Spectral Clustering	0.108	0.055
		Deep Clustering	**0.100**	**0.040**
	Big graph	Spectral Clustering	0.112	0.071
		Deep Clustering	**0.106**	**0.056**

Table 5: Results of the stretch goal on small and big graphs

Conclusion

- We do not support the claims in the paper.
- Keyword correlation graph do not cover all documents.
- Clustering document embeddings part is working well.
- Keyword correlation graph cannot preserve sufficient information about documents.

References

[1] Billy Chiu, Sunil Kumar Sahu, Derek Thomas, Neha Sengupta, and Mohammady Mahdy. Autoencoding keyword correlation graph for document clustering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3974–3981, Online, July 2020. Association for Computational Linguistics.

[2] Phi Vu Tran. Learning to make predictions on graphs with autoencoders. In *5th IEEE International Conference on Data Science and Advanced Analytics*, 2018.