Dissecting embedding method: learning higher-order structures from data

Liubov Tupikina, Kathuria Hritika

Dissecting embedding method: learning higher-order structures from data

Liubov Tupikina[1 2 3] and Hritika Kathuria[2]

[1] Nokia Bell labs, France, Paris, France,
[2] LPI, University Paris Descrartes, Learning Planet Institute, INSERM, Paris, France
[3] MIPT, Moscow Insitute of Physics and Technology, Moscow, Russia

1 Introduction and problem statement

Active area of research in AI is the theory of manifold learning and finding lower-dimensional manifold representation on how we can learn geometry from data for providing better quality curated datasets [1]. There are however various issues with these methods related to finding low-dimensional representation of the data, the so-called curse of dimensionality [2]. Geometric deep learning methods for data learning often include set of assumptions on the geometry of the feature space. Some of these assumptions include pre-selected metrics on the feature space, usage of the underlying graph structure, which encodes the data points proximity. However, the later assumption of using a graph as the underlying discrete structure, encodes only the binary pairwise relations between data points, restricting ourselves from capturing more complex higher-order relationships, which are often often present in various systems [1], [5]. These assumptions together with data being discrete and finite can cause some generalisations, which are likely to create wrong interpretations of the data and models outputs. Hence overall this can cause wrong outputs of the embedding models themselves, while these models being quite and trained on large corpora of data, such as BERT, Yi and other similar models.
The objective of our research is twofold, first, it is to develop the alternative framework to characterize the embedding methods dissecting their possible inconsistencies using combinatorial approach of higher-order structures which encode the embedded data. Second objective is to explore the assumption of the underlying structure of embeddings to be graphs, substituting it with the hypergraph and using the hypergraph theory to analyze this structure. We also demonstrate the embedding characterization on the usecase of the arXiv data.

2 Data

The arXiv dataset encompasses scientific articles published online between 1992 and 2018, which span 175 scientific fields [3]. Each article has the ground-truth information (metadata) stored as a category or a tag, e.g. *cond-mat.dis-nn* and others.

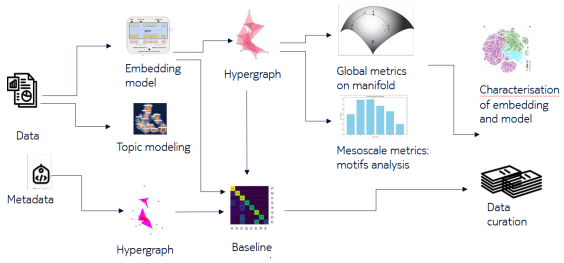

Fig. 1. Scheme of the methodological framework.

3 Methodological framework

Our methodological framework consists of the following steps, **Fig. 1**. **First**, we apply the embedding model to the datapoints D_i with the metadata m_i to get the coordinates of the embedded datapoints. **Second**, we cluster the datapoints D_i in the embedding using some clustering method (e.g. k-means clustering method) and apply ML metrics (e.g. confusion matrix, accuracy scores) to deduce scores of the embedding model, which produced the vector representation of datapoints D_i using the metadata labels of clusters m_i. **Third,** we test the discripancies in the resulted confusion matrix by constructing the *neighborhood hypergraph* H from datapoints $\{D_i\}_{i=1}^n$ and calculate hypergraph motifs in this hypergraph H, corresponding to the embedding. Then we analyze the hypergraphs constructed from the metadata m_i and sub-hypergraphs, which contain the misclassified tags based on the metrics from the second step of the method. Motifs in hypergraphs are subhypergraphs structures. By studying them one can dissect the most popular of the higher-order intricate properties of data structures [7].

Neighborhood hypergraph construction method. The *neighborhood hypergraph* H_n is constructed from data points $\{D_i\}_{i=1}^n$ based on the idea of estimating the relations between the datapoints. If datapoints $(D_{i_1},...,D_{i_k})$ are close to each other, they form a hyperedge of a hypergraph in H_n of k-arity. Depending on the nature of data there are different ways to characterize closseness of datapoints, which can be based on their geometrical or semantical proximity. Geometrical proximity of points $\{D_i\}$ is estimated using typical $l2$ metrics or cosine similarity applied to coordinates of each processed datapoint D_i. Semantical proximity of points is encoded by labelled similarity of datapoints, for instance, in case of arXiv data it means that if paper $\{D_1\}$ has tags a,b,c, paper D_2 has tags a,b and paper D_3 has tags a,d, then these papers are encoded as nodes of the hypergraph forming the hyperedge with the field a. One can also construct other encoding of the higher-order structure from data using *metadata* $\{m_i\}_{i=1}^m$ information H_m which encodes relations between tags themselves and can characterize the fields closseness instead of using the standard binary graph representation of data as co-tags networks [3]: when tags a,b,c are nodes of hypergraph, these nodes share the hyperedge, iff there is at least one paper with all these tags, like paper D_1 in this case. Our intuition behind using the *neighborhood hypergraph* representation of data is based on the observation that in many textual data similarities between documents

are not just based on one measure of closeness but rather on high-order relations like in words: word *France* is close to *Italy* and *Slovenia* in different way than *France* to the word *Paris*.

The embedding models. In this short paper we choose BERT as **the embedding model**, general autoencoder architecture, which provides the embedding of the textual data [2]. Our framework can be applied to other embedding models.

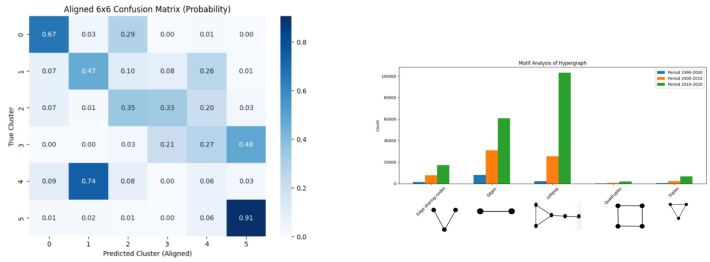

Fig. 2. (Left) Confusion matrix is calculated for the clustered datapoints from BERT embedding method to compare them to the clusters from ground truth labels (metadata). We consider subsample of data from (2000-2011). (Right) Characterisation of persistent hypergraph motifs in the hypergraphs constructed from abstracts datapoints D_i for each decade.

Analytical background of the method and general relation to the manifold learning. Our general motivation comes from the manifold regularisation method, which allows one to learn the properties of data limiting the space of solutions to the manifold. The central idea of the manifold learning is to form a neighborhood graph G_n [1]. We aim further to extend the idea of considering the neighborhood hypergraph H_n instead of the neighborhood graph as the underlying learned structure from data. Usually we assume that the feature space of data has the standard $l2$ metric, which induces the Hilbert space. The neighborhood graphs (hypergraphs) are constructed such that one chooses k nearest points as the neighborhood or points at the maximum distance r for the neighborhood. Manifold regularization further simplifies assumptions using a binarized metric or heat kernel. The problem of the manifold regularisation with the hypergraph H_n then can be further formalized as:

$$f^* = \min_{f \in h_k}(L(f)) + \lambda (f^T L(H_n) f),$$

where $L(H_n)$ is the generalized hypergraph Laplacian. $(f, L(H_n)f) = f^T L(H_n) f$ is a scalar product, λ is the regularization parameter [1], f^T is the transpose of a vector, where $\min_{f \in h_k}$ is minimum found from the set of possible hypothesis e.g. hypothesis on spaces search. Finding the underlying hypergraph can then extend the general method for manifold learning from data, yet it has some drawbacks, such as increasing complexity of the computations.

4 Results

We cluster datapoints D_i with coordinates estimated by BERT method. Then we estimate the confusion matrix, **Fig. 2 (left)**, to characterize the potential discrepancy in accuracy of the embedding method, verifying the clustered datapoints according to the labels of metadata m_i. Most of the misclassified datapoints belong to the cluster with *condmat* metadata field label, which is also known to be interdisciplinary field [3]. Majority of papers in *condmat* typically share common tags with other fields, which correspond to its propensity to be involved in various hyperedges of the *neighrborhood hypergraph* reconstructed from the data. For exploring this aspect we analyze the hypergraph, which encodes the proximity of abstracts (our datapoints D_i). For this we study the higher-order motifs in the whole hypergraph and its subhypergraphs, **Fig. 2 (right)**. We studied both, hypergraph H_n constructed from all datapoints D_i, as well as subhypergraphs, which consists of only nodes with *condmat* as their tags. In those subhypergraphs we found higher-order motifs, such as *triples, lollipops* in both structures, and we were able quantitatively characterize them. Moreover, we see the persistent growth in number of such higher-order motifs from the first decade of arXiv data to the present one, and specifically lollipop motifs, which are the most persistent patterns in these hypergraphs.

5 Conclusions

In our work we develop the framework to dissect the geometrical properties of embeddings based on the higher-order structures. This approach can be used in addition to traditional combinatorics and information theory based approaches to characterize embeddings [4]. In particular, in our approach we characterize possible inconsistencies in embedded data estimating the prevailing motifs in subhypergraphs, associated with the misclassified data. We plan to incorporate hypergraph embedding methods [6] further to investigate their applicability to data curation for LLMs training, as well as to dimensionality reduction methods. Currently we studied only BERT embedding model, yet our methodology has the potential to be further extended to differentiate between distinct embeddings, when we construct corresponding hypergraphs to each separate embedding method and then compare embeddings using the hypergraphs theory [7].

References

1. L. Melas-Kyriazi, Mathematical Foundations of Manifold Learning, Lecture notes (2020)
2. C. de Dampierre, A. Mogoutov, N. Baumard, Towards Transparency: Exploring LLM Trainings Datasets through Visual Topic Modeling and Semantic Frame arXiv. (2024)
3. C. Singh, L. Tupikina, M. Starnini, M. Santolini, Charting mobility patterns in the scientific knowledge landscape, EPJ D.S. 13,12(2024)
4. M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. (arXiv:2203.05794). URL http://arxiv.org/abs/2203.05794
5. F. Battiston, G. Cencetti, I. Iacopini, V. Latora, M. Lucas, A. Patania, J.-G. Young, G. Petri, Networks beyond pairwise interactions: Structure and dynamics, Phys.Rep., Vol.874, (2020)

6. F. Shi, J. Evans, Surprising combinations of research contents and contexts are related to impact and emerge with scientific outsiders from distant disciplines. Nat Commun 14, 1641 (2023)
7. C. Zapata-Carratala, et al., Hypermatrix Algebra and Irreducible Arity in Higher-Order systems, Adv. Compl. Sys. (2023)