Embeddings store thesis proposal

Theofilos Papapanagiotou

October 2021

1 Introduction

Continuous machine learning pipelines are successful due to having event-driven retraining [2]. That event or trigger depends on the drift detected in the distribution of feature values across multiple batches of data [4]. Furthermore, skew detection between training and serving also depends on such metrics [5], using L_∞ distance for numerical or shannon distance for categorical features.

Feature stores are traditionally key value stores like feast redis and hopsworks NDB [11]. Add Michelangelo, Bootleg, if needed. They provide low latency retrieval of features based on key. Businesses (cite) usually encode in that key some extra information such as the model version used to calculate a particular version of the embedding. Any further retrieval is based on that and other metadata stored around that key hashmap.

Due to the growth of language models and the reusability of the vector representations of word/sentences meanings (embeddings), production feature store systems which store embeddings require retrieval capabilities that a KV cannot offer. Examples are the similarity search and the monitoring of drift detection.[12]

This thesis aims to look at the later, and contribute a set of metrics to be integrated with a vector database to enable such monitoring capabilities.

2 Problem Statement

2.1 What is the fundamental research problem that you aim to solve?

Enrich the embedding store capabilities in the space of monitoring. Identify which metrics are optimal for shift detection of words and sentenses meaning in language models.

2.2 Why is the problem important?

Statistically significant linguistic shifts in the meaning [3] and usage of words.

In continuous training pipelines, we need an event or trigger to retrain on drift [2].

2.3 Why is the problem non-trivial?

Since embeddings computed in different semantic spaces are not directly comparable, time related representations are usually made comparable either by aligning different semantic spaces through a transformation matrix [9], [1], [7] or by initializing the embeddings at $t_i + 1$ using those computed at t_i [].

Because embedding spaces are different and have different dimensionality structures, we cannot compare the vectors of a word in two different spaces directly.

2.4 Why is the problem not solved by the current state-of-the-art

Tools in the space of vector databases and embedding stores recently flourish, but focus only on the similarity search.

Meaning shift/Drift detection in the embedding space is left due to the difficult nature of measuring semantic shifts in unstructured data.

3 Proposed Approach & Contributions

4 Example

5 Experimental Evaluation

- word stability [1]
- change point detection in time series to assign significance of change scores to each word [9]
- conformity: the rate of semantic change scales with an inverse power-law of word frequency [7]
- innovation: independent of frequency, words that are more polysemous have higher rates of semantic change [7]
- linear transformation of vectors across embedding spaces [1]
- graph-based node similarity[1]
- combination of the above [1]
- distance-based distributional time series of word meanings [9]
- Measuring semantic displacement (cosine similarity) [7]
- Word Comparisons over time (cosine similarity) [8]

6 Related Work

- Short-Term Change in Word Representation [14]
- Semantic shifts [1]
- Short-Term Meaning Shift [6]
- Meaning in word embeddings [10]
- Embedding store [12]
- Statistically significant linguistic shifts in the meaning and usage of words [9]
- Temporal evolution of natural language [9]: frequency, part-of-speech tag distribution, and word co-occurrence
- CheckList [13]

7 Open Questions

8 Next Steps

References

[1] Hosein Azarbonyad et al. *Words Are Malleable: Computing Semantic Shifts in Political and Media Discourse.* Nov. 15, 2017. arXiv: `1711.05603 [cs]`. URL: `http://arxiv.org/abs/1711.05603` (visited on 10/29/2021).

[2] Denis Baylor et al. "Continuous Training for Production {ML} in the TensorFlow Extended ({TFX}) Platform". In: 2019 {USENIX} Conference on Operational Machine Learning (OpML 19). 2019, pp. 51–53. ISBN: 978-1-939133-00-7. URL: `https://www.usenix.org/conference/opml19/presentation/baylor` (visited on 11/03/2021).

[3] Emily M. Bender and Alexander Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.* ACL 2020. Online: Association for Computational Linguistics, July 2020, pp. 5185–5198. DOI: `10.18653/v1/2020.acl-main.463`. URL: `https://aclanthology.org/2020.acl-main.463` (visited on 11/03/2021).

[4] Eric Breck et al. "Data Validation for Machine Learning". In: (), p. 14.

[5] Emily Caveness et al. "TensorFlow Data Validation: Data Analysis and Validation in Continuous ML Pipelines". In: *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data.* SIGMOD '20. New York, NY, USA: Association for Computing Machinery, June 11, 2020, pp. 2793–2796. ISBN: 978-1-4503-6735-6. DOI: `10.1145/3318464.3384707`. URL: `https://doi.org/10.1145/3318464.3384707` (visited on 11/03/2021).

[6] Marco Del Tredici, Raquel Fernández, and Gemma Boleda. "Short-Term Meaning Shift: A Distributional Exploration". In: (Sept. 10, 2018). URL: https://arxiv.org/abs/1809.03169v3 (visited on 10/29/2021).

[7] William L. Hamilton, Jure Leskovec, and Dan Jurafsky. "Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change". In: (May 30, 2016). URL: https://arxiv.org/abs/1605.09096v6 (visited on 11/03/2021).

[8] Yoon Kim et al. "Temporal Analysis of Language through Neural Language Models". In: (May 14, 2014). URL: https://arxiv.org/abs/1405.3515v1 (visited on 11/03/2021).

[9] Vivek Kulkarni et al. "Statistically Significant Detection of Linguistic Change". In: *Proceedings of the 24th International Conference on World Wide Web*. WWW '15. Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, May 18, 2015, pp. 625–635. ISBN: 978-1-4503-3469-3. DOI: 10.1145/2736277.2741627. URL: https://doi.org/10.1145/2736277.2741627 (visited on 11/03/2021).

[10] Tomas Mikolov et al. "Distributed Representations of Words and Phrases and Their Compositionality". In: *Advances in Neural Information Processing Systems*. Vol. 26. Curran Associates, Inc., 2013. URL: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html (visited on 10/29/2021).

[11] Alexandru A Ormenis. "Horizontally Scalable ML Pipelines with a Feature Store". In: (), p. 2.

[12] Laurel Orr et al. *Managing ML Pipelines: Feature Stores and the Coming Wave of Embedding Ecosystems*. Aug. 11, 2021. arXiv: 2108.05053 [cs]. URL: http://arxiv.org/abs/2108.05053 (visited on 10/13/2021).

[13] Marco Tulio Ribeiro et al. "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList". In: (May 8, 2020). URL: https://arxiv.org/abs/2005.04118v1 (visited on 11/03/2021).

[14] Ian Stewart et al. "Measuring, Predicting and Visualizing Short-Term Change in Word Representation and Usage in VKontakte Social Network". In: (), p. 12.