Text classification with word embedding regularization and soft similarity measure

Vít Novotný
witiko@mail.muni.cz

Faculty of Informatics, Masaryk University

October 29, 2019

Math information retrieval

The past, the present, and the future

- The past: DML-CZ (2005), EuDML (2010), TAČR Omega (2016).
- The present: GAČR (2020), TAČR Zéta (2020), FTIR (2020).
- The (near) future: Math Information Retrieval with NNLMs:
  - Bi-Directional Tree-Structured LSTMs,
  - Soft Cosine Measure (SCM) [1, 2, 3] with math-aware word embeddings [4, 5, 6].

Text classification as proxy for information retrieval

- We use text classification to compare to a related document similarity measure: the Word Mover's Distance (WMD) [7].
- Text classification is related, but not identical to information retrieval: topic modeling with LSI improves task performance on text classification, but not on information retrieval [8].
- Since the SCM achieved SOTA performance on the semantic text similarity task [2], we are confident in its ability to capture semantics, not just general topics.

Abstract

Situation

- Since Mikolov et al. (2013) [9], word embeddings have become the preferred word representations for many natural language processing tasks
- Document similarity measures extracted from word embeddings, such as the soft cosine measure (SCM) and the Word Mover's Distance (WMD), were reported to achieve state-of-the-art performance on the semantic text similarity and text classification.

Soft cosine similarity measure

Intuition

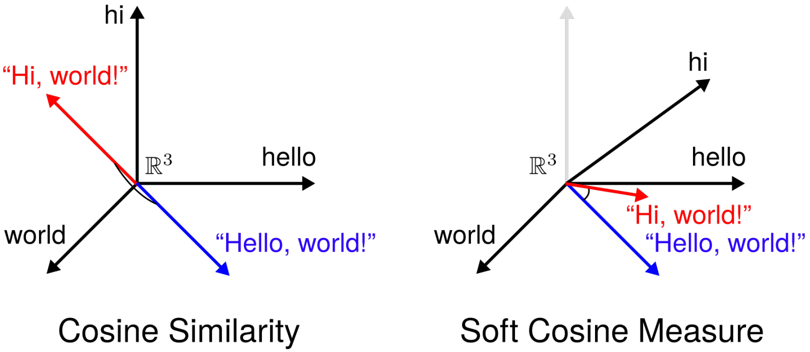


Figure: The geometric representation of the documents "Hi, world!", and "Hello, world!" in the standard VSM (left), and the soft VSM (right, [3]).

Soft cosine similarity measure

Definition

- Cosine similarity of $\mathbf{x}$ and $\mathbf{y}$ equals $\langle \mathbf{x}/\|\mathbf{x}\|_2, \mathbf{y}/\|\mathbf{y}\|_2 \rangle$, where $\langle \mathbf{x}, \mathbf{y} \rangle = \big((\mathbf{x})_\beta\big)^\mathsf{T} (\mathbf{y})_\beta$, β is an orthonormal basis, and $\|\mathbf{z}\|_2$ is the ℓ_2-norm of $\mathbf{z}$.
- Soft cosine similarity of $\mathbf{x}$ and $\mathbf{y}$ equals $\langle \mathbf{x}/\|\mathbf{x}\|_2, \mathbf{y}/\|\mathbf{y}\|_2 \rangle$, where $\langle \mathbf{x}, \mathbf{y} \rangle = \big((\mathbf{x})_\beta\big)^\mathsf{T} \mathbf{S} (\mathbf{y})_\beta$, β is a non-orthogonal normalized basis, $\|\mathbf{z}\|_2$ is the ℓ_2-norm of $\mathbf{z}$, and $\mathbf{S}$ is a word similarity matrix.

We define the word similarity matrix $\mathbf{S}$ like Charlet and Damnati (2017, [2]): $s_{ij} = \max(t, \langle \mathbf{e}_i/\|\mathbf{e}_i\|_2, \mathbf{e}_j/\|\mathbf{e}_j\|_2 \rangle)^o$, where $\mathbf{e}_i$ and $\mathbf{e}_j$ are the embeddings for words i and j, and o and t are free.

We use the implementation in the `similarities.termsim` module of Gensim [10].

The worst-case time complexity of the SCM is $\mathcal{O}(p_\mathbf{x} p_\mathbf{y})$, where $p_\mathbf{x}$ is the number of unique words in $\mathbf{x}$ and $p_\mathbf{y}$ is the number of unique words in $\mathbf{y}$.

Word mover's distance measure

Intuition

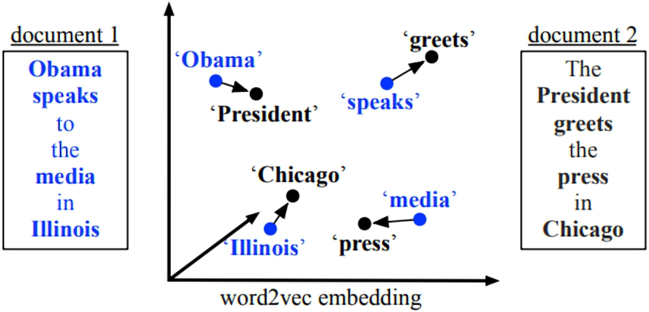


Figure: The Word mover's distance (WMD) between the VSM representations of documents "Obama speaks to the media in Illinois", and "The president greets the press in Chicago". [7]

Word mover's distance measure

Definition

The Word mover's distance (WMD) of $\mathbf{x}$ and $\mathbf{y}$ equals the minimum cumulative cost $\sum_{i,j} f_{ij} c_{ij}$ of a flow $\mathbf{F} = (f_{ij})$ subject to $\mathbf{F} \geq 0$, $\sum_j f_{ij} = (x_i)_\beta$, where the cost c_{ij} is the Euclidean distance of embeddings for words i and j.

We use the implementation in PyEMD [11, 12] with the best known average time complexity $\mathcal{O}(p_{\mathbf{xy}}^3 \log p_{\mathbf{xy}})$, where $p_{\mathbf{xy}}$ is the number of unique words in $\mathbf{x}$ and $\mathbf{y}$.

Abstract

Problem

- Despite the strong performance of the WMD on text classification and semantic text similarity, its super-cubic average time complexity is impractical.
- The SCM has quadratic worst-case time complexity, but its performance on text classification has never been compared with the WMD.
- Recently, two word embedding regularization techniques were shown to reduce storage and memory costs, and to improve training speed, document processing speed, and task performance on word analogy, word similarity, and semantic text similarity. However, the effect of these techniques on text classification has not yet been studied.

Word embedding quantization

The CBOW with negative sampling minimizes the following loss:

$$J(\mathbf{u}_o, \hat{\mathbf{v}}_c) = -\log\left(\sigma(\langle \mathbf{u}_o, \hat{\mathbf{v}}_c \rangle)\right) - \sum_{i=1}^{k} \log\left(\sigma(-\langle \mathbf{u}_i, \hat{\mathbf{v}}_c \rangle)\right),$$

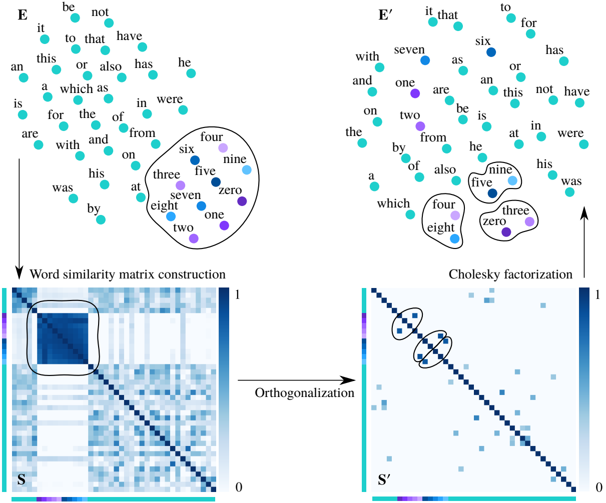
where $\hat{\mathbf{v}}_c = \frac{1}{2w} \sum_{-w+i \leq i \leq w+o, i \neq o} \mathbf{v}_i$, $\mathbf{u}_o$ is the vector of a center word with corpus position o, $\mathbf{v}_i$ is the vector of a context word with corpus position i, and the window size w and the number of negative samples k are free parameters.

Following the approach of Lam [13], we quantize the center word vector $\mathbf{u}_o$ and the context word vector $\mathbf{v}_i$ to $\pm 1/3$ during the forward and backward propagation stages of the training. Since the quantization function is non-differentiable at certain points, we use Hinton's straight-through estimator [14, Lecture 15b] as the gradient:

$$\nabla(1/3 \cdot \text{sign}) = \nabla I, \text{where } \nabla \text{ is gradient operator and } I \text{ is identity.}$$

Orthogonalized word embeddings

Intuition



Abstract

Solution

- In our work, we investigate the individual and joint effect of the two word embedding regularization techniques on the document processing speed and the task performance of the SCM and the WMD on text classification.
- The SCM has quadratic worst-case time complexity, but its performance on text classification has never been compared with the WMD.
- For evaluation, we use the *k*NN classifier and six standard datasets: BBCSPORT, TWITTER, OHSUMED, REUTERS-21578, AMAZON, and 20NEWS.

Word embedding orthogonalization

Introduction

Vít [15] shows that producing a sparse word similarity matrix $\mathbf{S}'$ that stores at most C largest values from every column of $\mathbf{S}$ reduces the worst-case time complexity of the SCM to $\mathcal{O}(p_{\mathbf{x}})$, where $p_{\mathbf{x}}$ is the number of unique words in a document vector $\mathbf{x}$.

Vít [15] also claims that $\mathbf{S}'$ improves the performance of the soft VSM on the question answering task and describes a greedy algorithm for producing $\mathbf{S}'$, which we will refer to as the orthogonalization algorithm. The orthogonalization algorithm has three boolean parameters: Sym, Dom, and Idf. Sym and Dom make $\mathbf{S}'$ symmetric and strictly diagonally dominant. Idf processes columns of $\mathbf{S}$ in descending order of inverse document frequency [16]:

$$-\log_2 \mathrm{P}(w \mid D) = \log_2 \frac{|D|}{|\{d \in D \mid w \in d\}|}, \text{ where } D \text{ are documents.}$$

Word embedding orthogonalization

Definition (Orthogonalized word embeddings)

Definition Let $\mathbf{E}, \mathbf{E}'$ be real matrices with $|V|$ rows, where V is a vocabulary of words. Then $\mathbf{E}'$ are orthogonalized word embeddings from $\mathbf{E}$, which we denote $\mathbf{E}' \leq_\perp \mathbf{E}$, iff for all $i, j = 1, 2, \ldots, |V|$ it holds that $\langle \mathbf{e}'_i, \mathbf{e}'_j \rangle \neq 0 \implies \langle \mathbf{e}'_i, \mathbf{e}'_j \rangle = \langle \mathbf{e}_i, \mathbf{e}_j \rangle$, where $\mathbf{e}_k$ and $\mathbf{e}'_k$ denote the k-th rows of $\mathbf{E}$ and $\mathbf{E}'$.

Theorem (Orthogonalization produces orthogonalized w. e.)

Let $\mathbf{E}$ *be a real matrix with* $|V|$ *rows, where* V *is a vocabulary of words, and for all* $k = 1, 2, \ldots, |V|$ *it holds that* $\|\mathbf{e}_k\|_2 = 1$. *Let* $\mathbf{S}$ *be a word similarity matrix constructed from* $\mathbf{E}$ *with the parameter values* $t = -1$ *and* $o = 1$. *Let* $\mathbf{S}'$ *be a word similarity matrix produced from* $\mathbf{S}$ *using the orthogonalization algorithm with the parameter values* $Sym = \checkmark$ *and* $Dom = \checkmark$. *Let* $\mathbf{E}'$ *be the Cholesky factor of* $\mathbf{S}'$. *Then* $\mathbf{E}' \leq_\perp \mathbf{E}$.

Abstract

Evaluation

- We show 39% average kNN test error reduction with regularized word embeddings compared to non-regularized word embeddings.
- We describe a practical procedure for deriving such regularized embeddings through Cholesky factorization.
- We also show that the SCM with regularized word embeddings significantly outperforms the WMD on text classification and is over $10{,}000\times$ faster.

Results

T-SNE document visualizations

The following figures show confusion matrices and t-SNE document visualizations [17] for the soft VSM with non-regularized word embeddings and for the soft VSM with orthogonalized and quantized word embeddings.

- OHSUMED with non-regularized w.e. and with regularized w.e.,
- BBCSPORT with non-regularized w.e. and with regularized w.e.,
- REUTERS with non-regularized w.e. and with regularized w.e.,
- AMAZON with non-regularized w.e. and with regularized w.e.,
- 20NEWS with non-regularized w.e. and with regularized w.e..

Results

Test error I

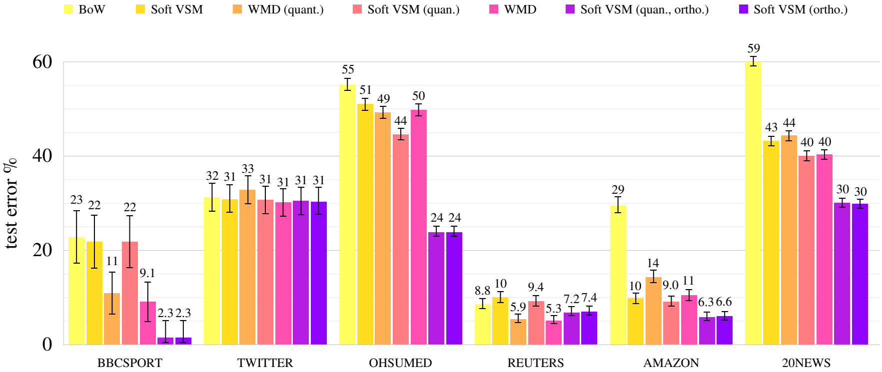


Figure: 95% interval estimates for the *k*NN test error on six text classification datasets

Results

Test error II

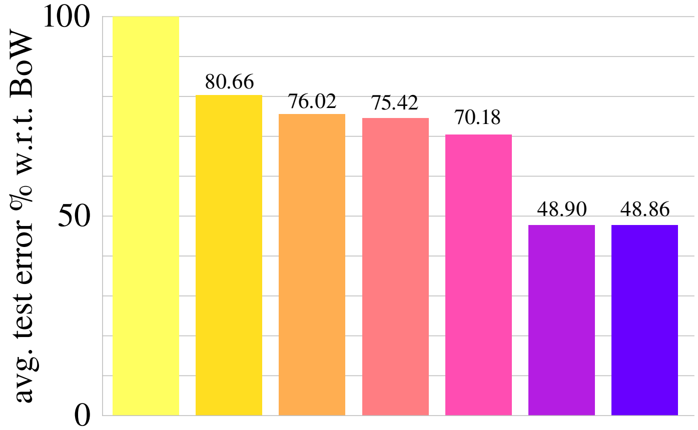


Figure: Average document processing speed on one Intel Xeon X7560 core

Results

Processing speed

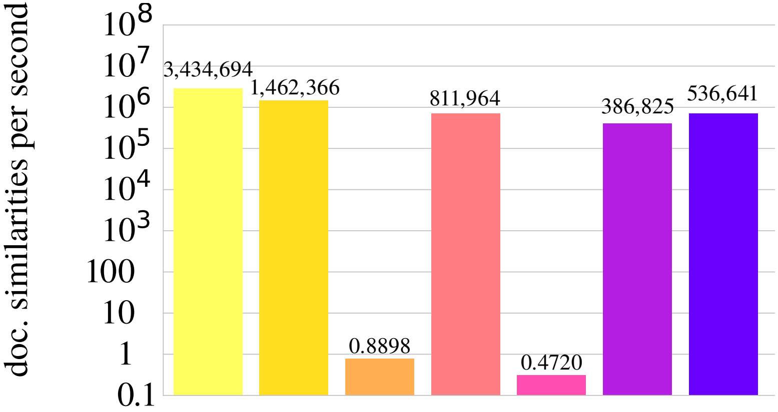


Figure: Average document processing speed on one Intel Xeon X7560 core

References I

[1] Grigori Sidorov, Alexander Gelbukh, Helena Gómez-Adorno, and David Pinto.
Soft similarity and soft cosine measure: Similarity of features in vector space model.
Computación y Sistemas, 18(3):491–504, 2014.

[2] Delphine Charlet and Geraldine Damnati.
Simbow at semeval-2017 task 3: Soft-cosine semantic similarity between questions for community question answering.
In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 315–319, 2017.

References II

[3] Vít Novotný.
Implementation notes for the soft cosine measure.
In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, pages 1639–1642, New York, NY, USA, 2018. ACM.

[4] Liangcai Gao, Zhuoren Jiang, Yue Yin, Ke Yuan, Zuoyu Yan, and Zhi Tang.
Preliminary exploration of formula embedding for mathematical information retrieval: can mathematical formulae be embedded like a natural language?, 2017.

[5] Kriste Krstovski and David M. Blei.
Equation embeddings, 2018.

References III

[6] Behrooz Mansouri, Shaurya Rohatgi, Douglas W. Oard, Jian Wu, C. Lee Giles, and Richard Zanibbi.
Tangent-cft: An embedding model for mathematical formulas.
In *ICTIR*, 2019.

[7] Matt J. Kusner, Yu Sun, et al.
From word embeddings to document distances.
In *Proceedings of the 32nd International Conference on International Conference on Machine Learning*, volume 37, pages 957–966. JMLR.org, 2015.

[8] Avinash Atreya and Charles Elkan.
Latent semantic indexing (lsi) fails for trec collections.
SIGKDD Explorations, 12:5–10, 2010.

References IV

[9] Tomas Mikolov, Kai Chen, et al.
Efficient estimation of word representations in vector space.
arXiv preprint, 2013.
Accessed 22 October 2019.

[10] Radim Řehůřek and Petr Sojka.
Software framework for topic modelling with large corpora.
In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, 2010.

[11] Ofir Pele and Michael Werman.
A linear time histogram metric for improved sift matching.
In *Proceedings of the 10th ECML: Part III*, ECCV '08, pages 495–508, Berlin, Heidelberg, 2008. Springer-Verlag.

References V

[12] Ofir Pele and Michael Werman.
Fast and robust earth mover's distances.
In *IEEE 12th International Conference on Computer Vision*, pages 460–467. IEEE, September 2009.

[13] Maximilian Lam.
Word2Bits – quantized word vectors.
arXiv preprint, 2018.
Accessed 22 October 2019.

[14] Geoffrey Hinton.
Neural networks for machine learning.
Coursera, 2012.
Accessed 23 October 2019.

References VI

[15] Vít Novotný.
Implementation notes for the soft cosine measure.
In *Proceedings of 27th ACM International Conference on Information and Knowledge Management (CIKM '18)*, pages 1639–1642. Association of Computing Machinery, 2018.

[16] Stephen Robertson.
Understanding inverse document frequency: on theoretical arguments for IDF.
Journal of documentation, 60(5):503–520, 2004.

[17] Laurens Maaten and Geoffrey Hinton.
Visualizing data using t-SNE.
Journal of machine learning research, 9(Nov):2579–2605, 2008.

MUNI
FACULTY
OF INFORMATICS